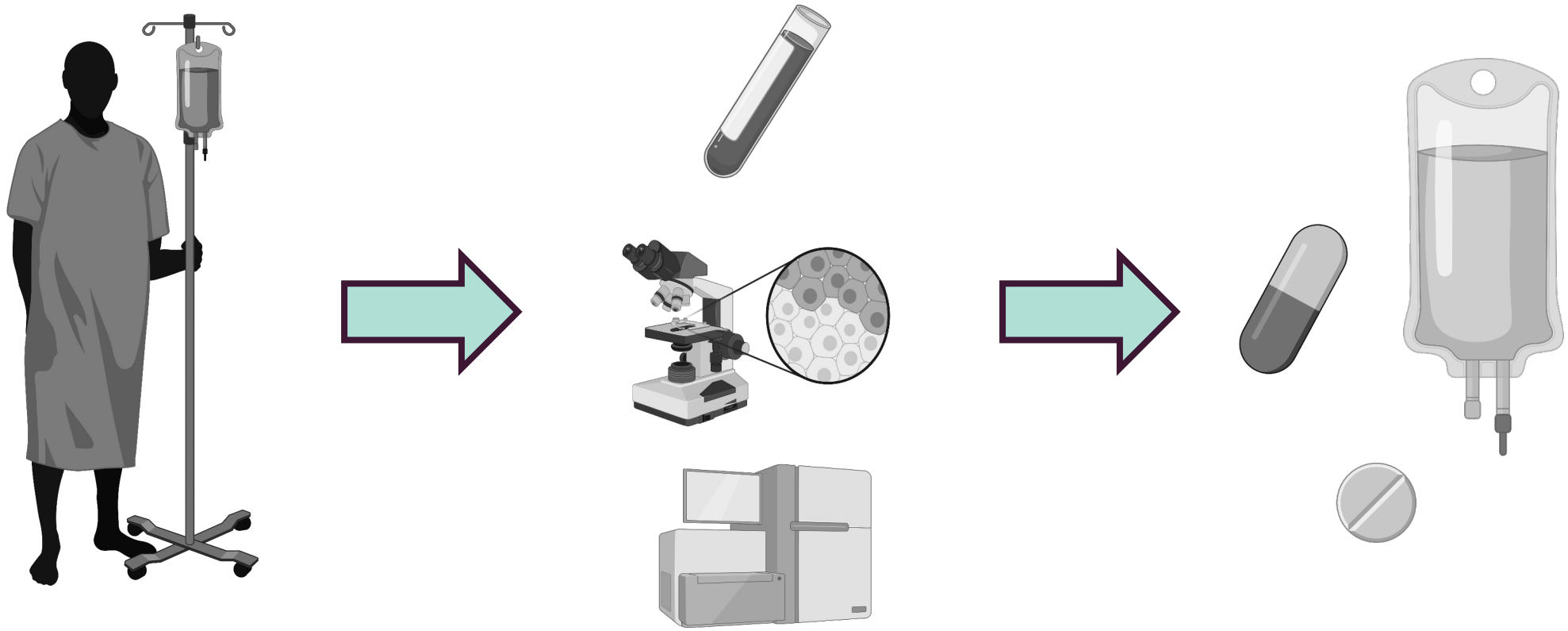
Live brain cancer classification during surgery

Basic Machine Learning Bioinformatics

09-03-2023

Personalized medicine

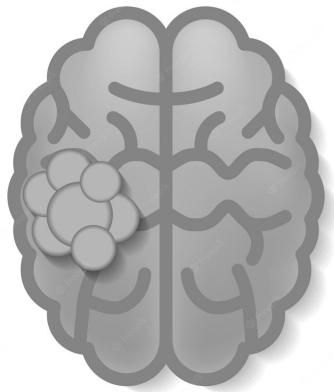


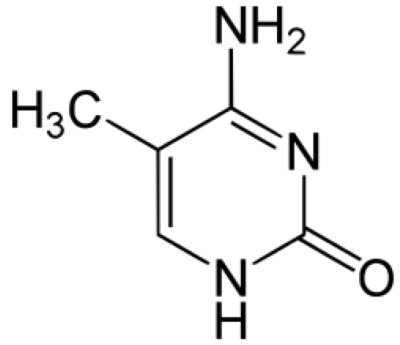
Patient

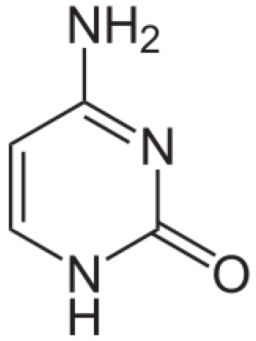
Biomarker

Treatment

Personalized medicine







Brain cancer

Methylation status

Personalized surgery

Personalized surgery

How much resection should be done?

Too much → Decrease in the quality of life

Too little → Second surgery disease progression

Personalized surgery

Peri-operative CNS classification

Surgery

Informed surgeon

TIME

90 min (approx)

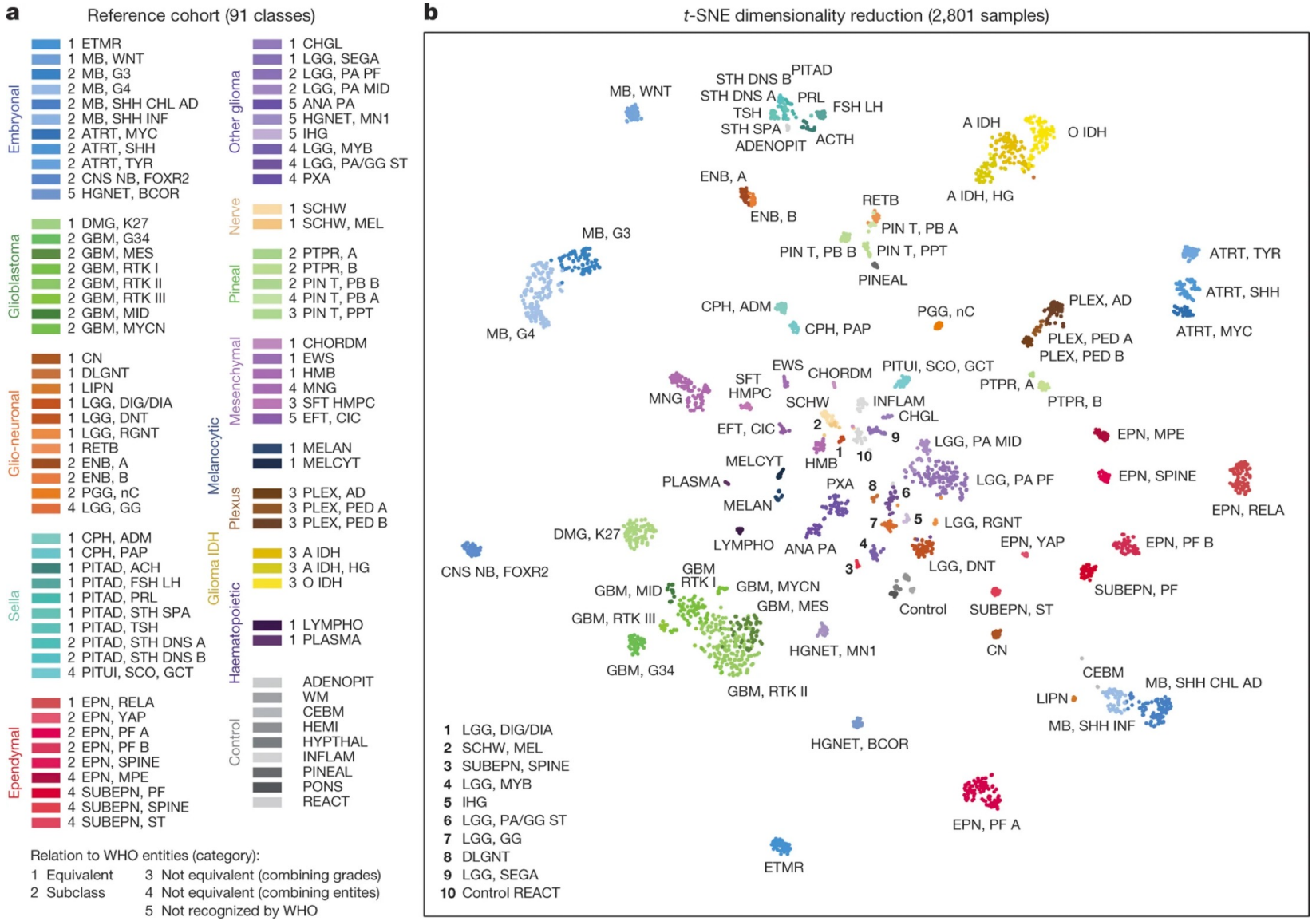
Sample preparation

Histology

Pathologist

Sometimes is not sure
Cannot distinguish between tumor subtypes
Sometimes makes mistakes
They require training

Methylation as CNS class biomarker

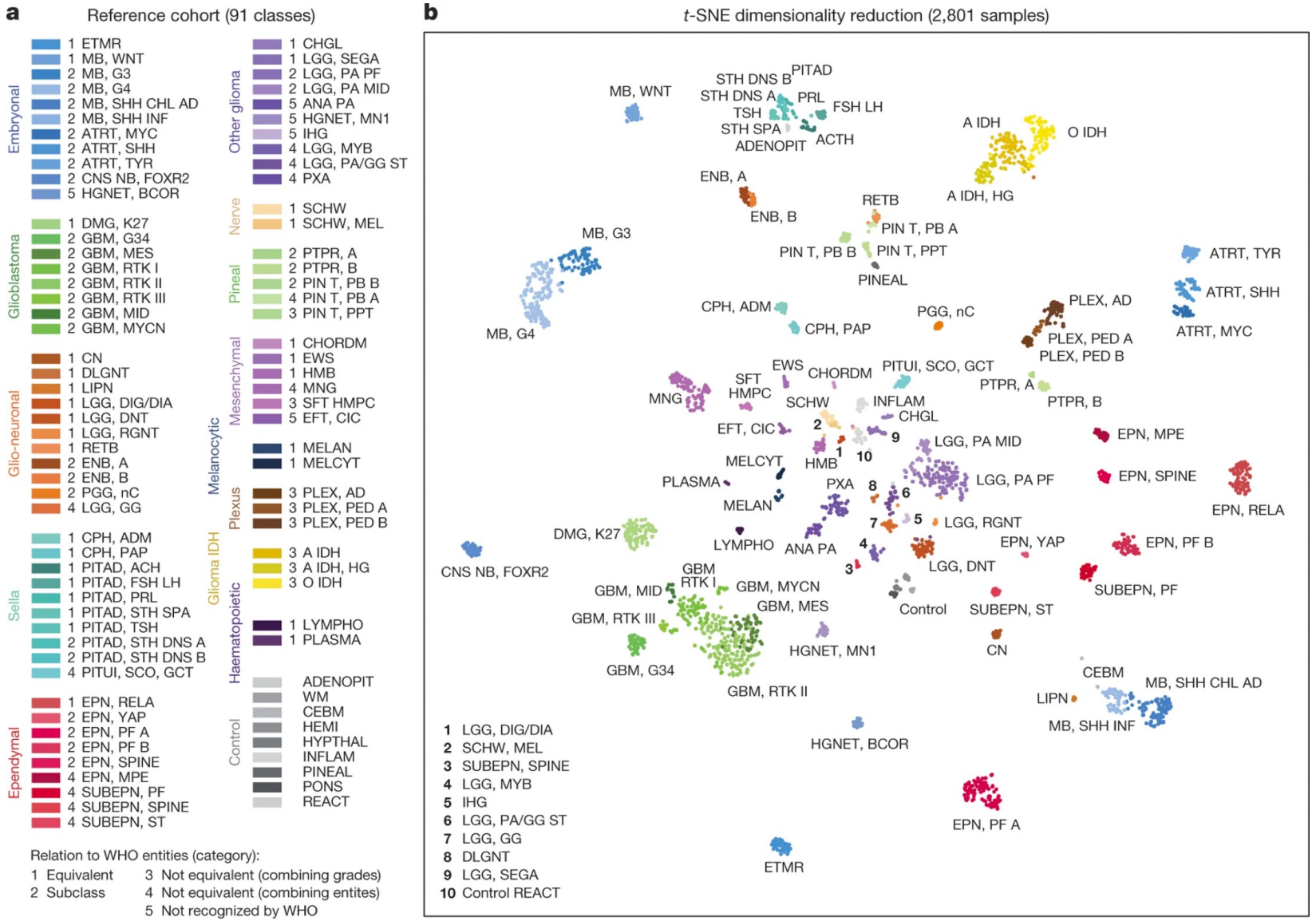


2801 CNS samples
82 tumor classes
9 control classes

Illumina Infinium array

WHAT DO YOU SEE?

Methylation as CNS class biomarker

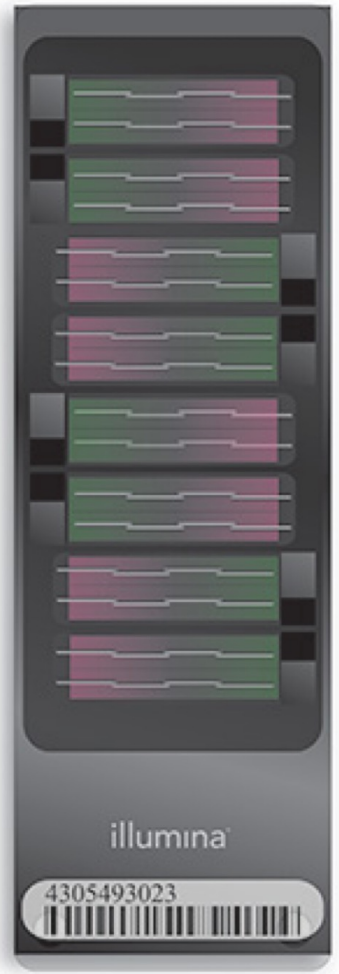


2801 CNS samples
82 tumor classes
9 control classes

Illumina Infinium array

Methylation is a really good biomarker
Unsupervised clustering already gives very clear clusters

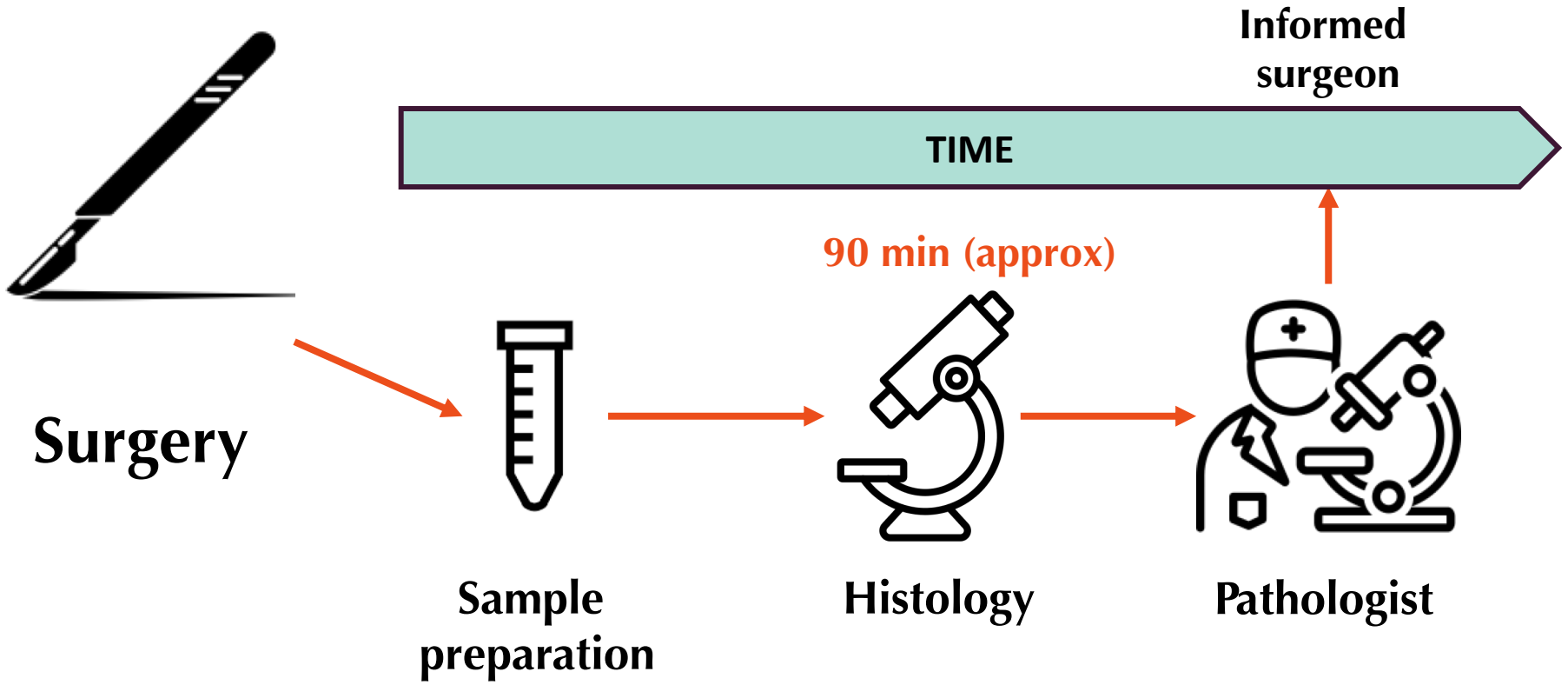
Illumina methylation arrays



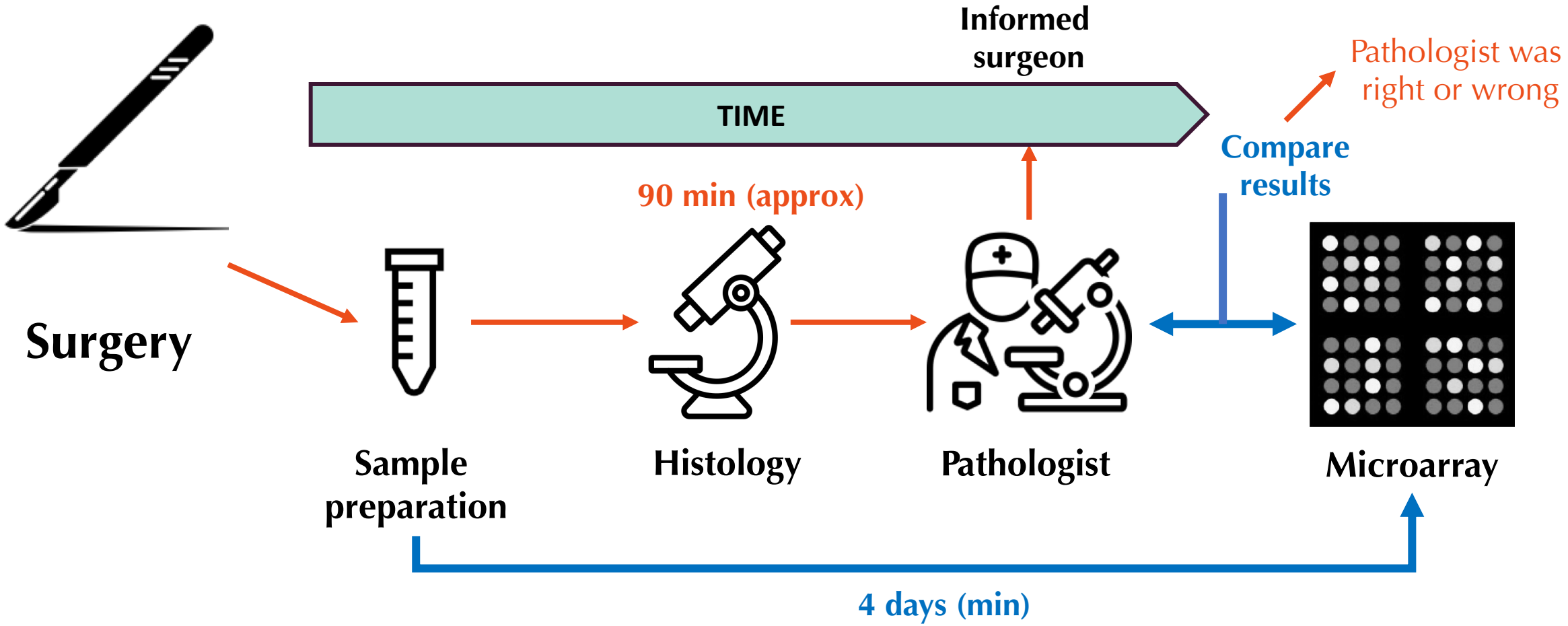
Illumina Infinium array

- 450k CpG methyation sites
- 8 samples simulateously
- Genome-wide coverage
- 4 days workflow

Peri-operative CNS classification



Peri/post-operative CNS classification

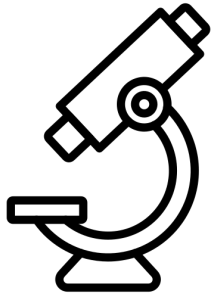


Peri/post-operative CNS classification

Surgery

Sample preparation

90 min (approx)



Histology

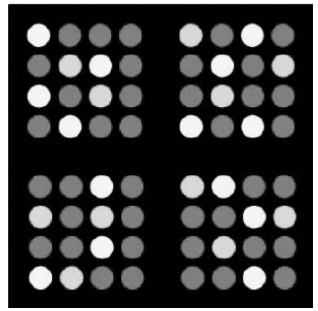
Pathologist

Informed surgeon

TIME

Compare results

Pathologist was right or wrong



Microarray

4 days (min)

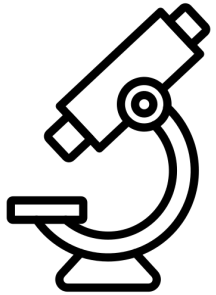
STANDARD OF CARE

Peri/post-operative CNS classification

Surgery

Sample preparation

90 min (approx)



Histology

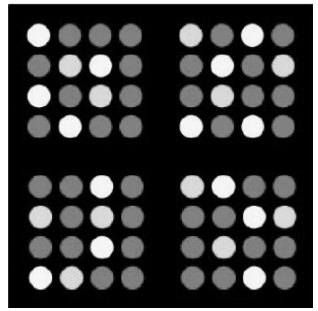
Pathologist

Informed surgeon

TIME

Compare results

Pathologist was right or wrong

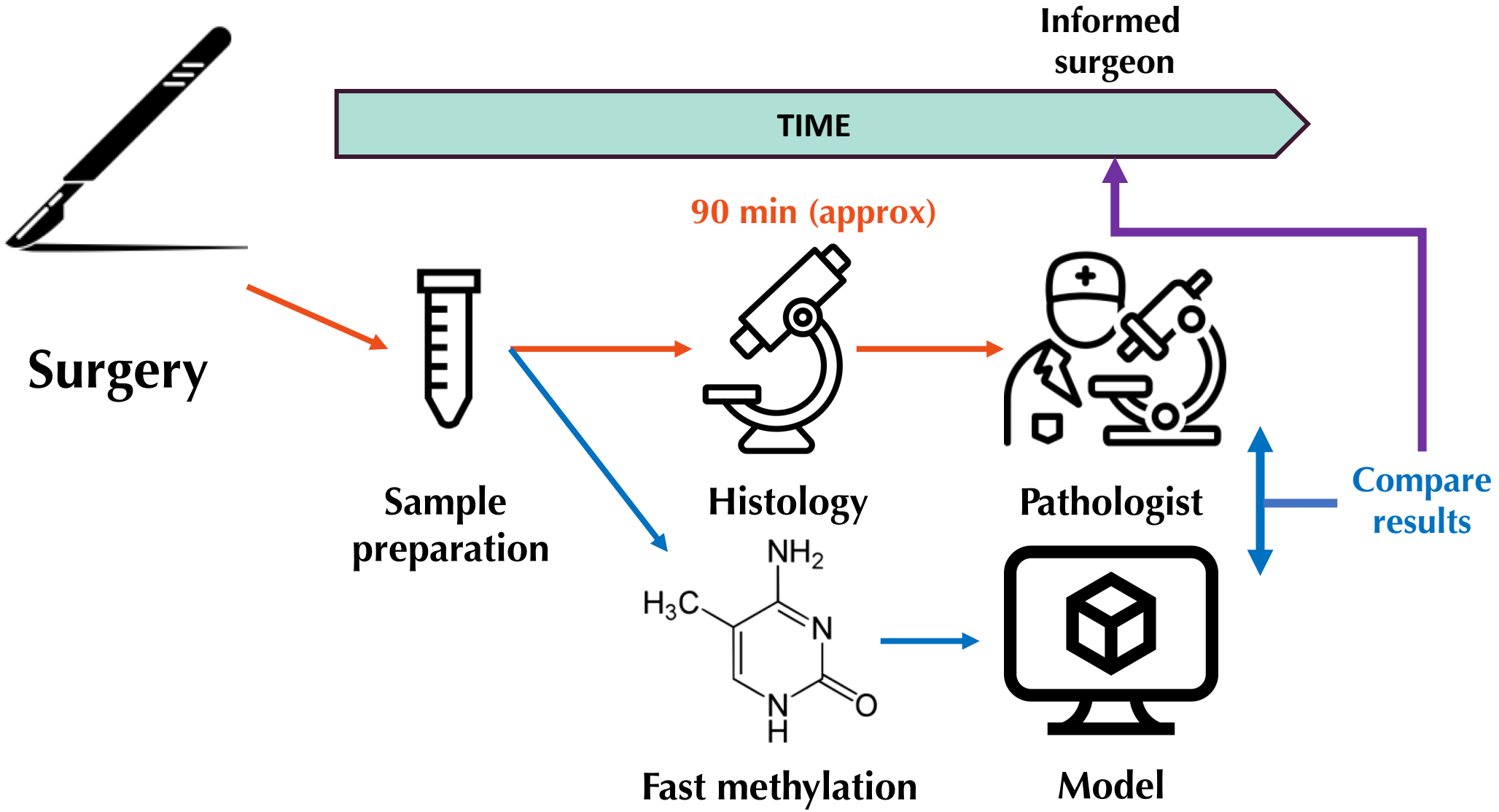


Microarray

4 days (min)

NOT FEASIBLE DURING SURGERY

Peri-operative CNS classification



Fast methylation detection

Nanopore sequencing

- Native DNA sequencing
- Potential genome-wide coverage
- 15 minutes sample preparation
- Real time data output

Sparsity problem

Microarray

0% missing values
Continuous signal

Sparsity problem

Microarray

0% missing values
Continuous signal

Nanopore

>90% missing values
Binary signal

Unknown which missing values

Sparsity problem

TRAIN

Microarray

0% missing values
Continuous signal

TEST

Nanopore

>90% missing values
Binary signal

Unknown which missing values

Sparsity problem

TRAIN

Microarray

0% missing values
Continuous signal

TEST

Nanopore

>90% missing values
Binary signal

Unknown which missing values

Can we develop a method that can handle an arbitrary random amount of missing values?

Sparsity problem

TRAIN

Microarray

0% missing values
Continuous signal

TEST

Nanopore

>90% missing values
Binary signal

Unknown which missing values

ANY IDEAS?

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology
JOURNAL OF THE BRITISH NEUROPATHOLOGICAL SOCIETY
WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology WILEY

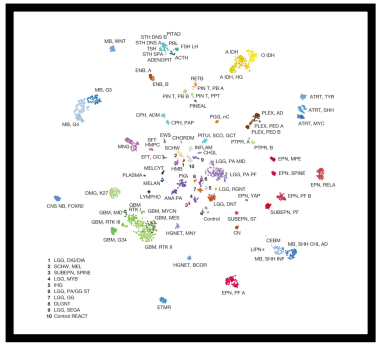
Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

Nanopore sequencing

TIME

2801x450000



Microarray data

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

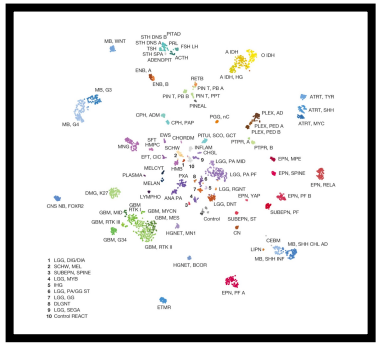
Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

Nanopore sequencing

TIME

15.000 CpG sites

2801x450000



Microarray data

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

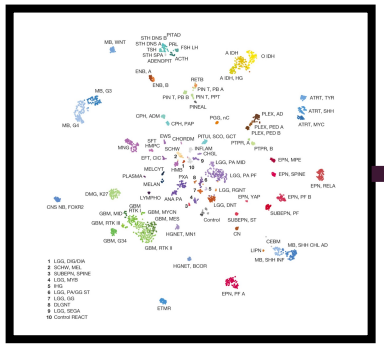
Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

Nanopore sequencing

TIME

15.000 CpG sites

2801x450000



Microarray data

Subset

2801x15000

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

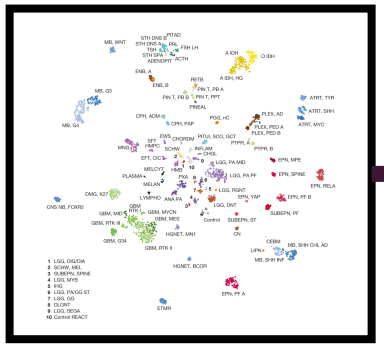
Nanopore sequencing

TIME

15.000 CpG sites

Predict

2801x450000

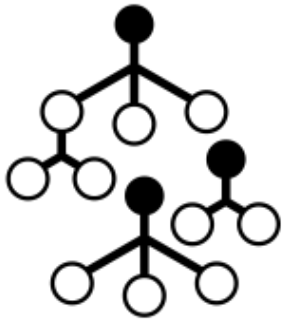


Microarray data

Subset

2801x15000

Train



Random forest

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

Nanopore sequencing

TIME

15.000 CpG sites

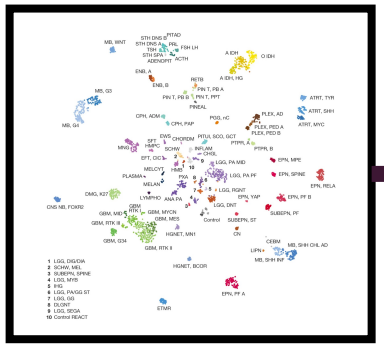
Predict

Is the score high enough?

Yes: tell surgeon

No: try again with more data

2801x450000

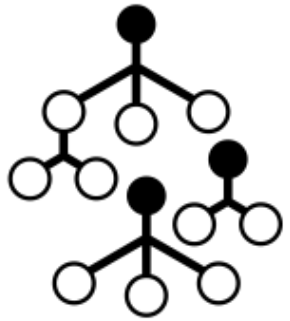


Microarray data

Subset

2801x15000

Train



Random forest

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

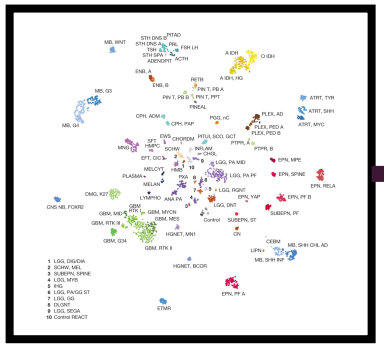
Nanopore sequencing

TIME

30.000 CpG sites

Predict

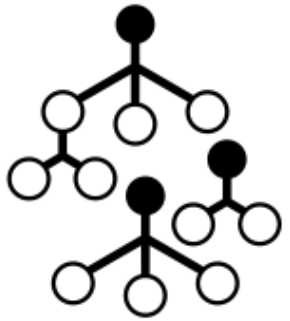
2801x450000



Subset

2801x30000

Train



Microarray data

Random forest

Existing approaches

ORIGINAL ARTICLE

Neuropathology and Applied Neurobiology WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

Nanopore sequencing

TIME

15.000 CpG sites

Predict

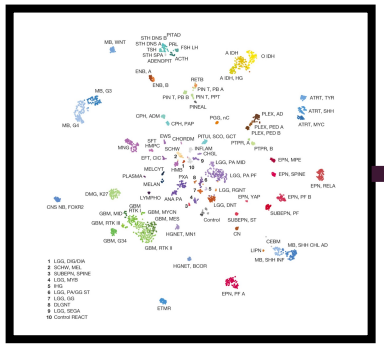
PROS

- It works

CONS

- ?

2801x450000

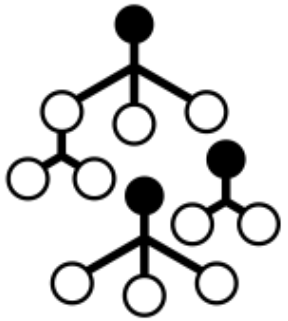


Microarray data

Subset

2801x15000

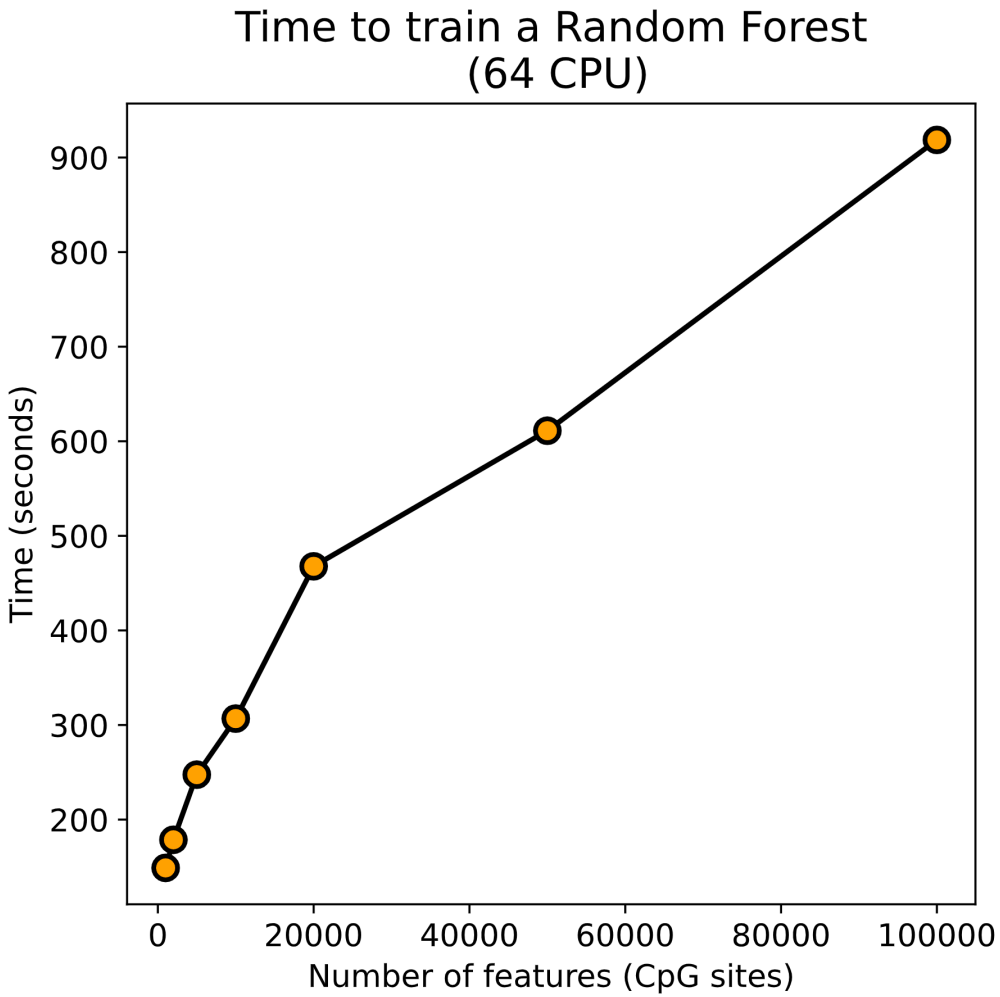
Train



Random forest

Existing approaches

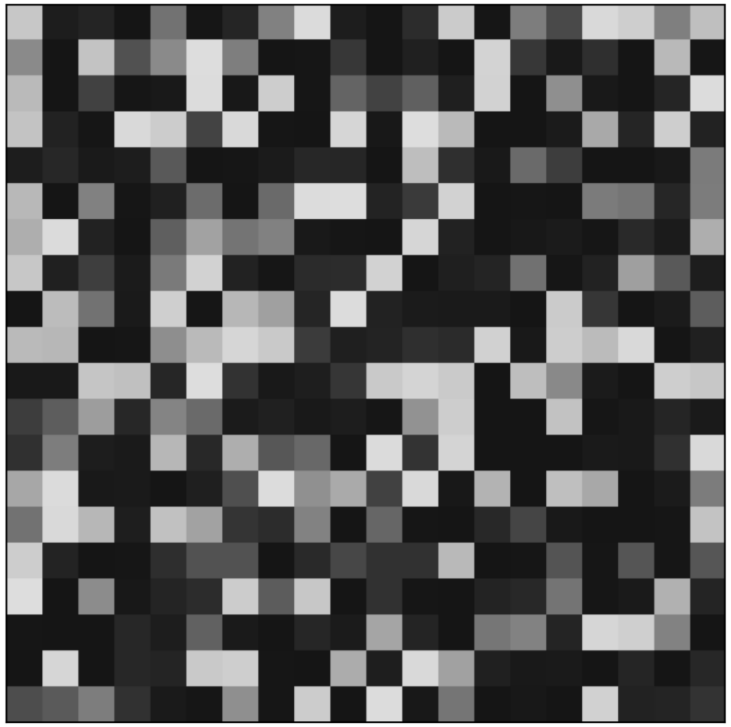
- Training takes valuable time
- Patient specific models
- Requires on-the-fly validaiton
- Requires access to data
- Poor scalability



Research question

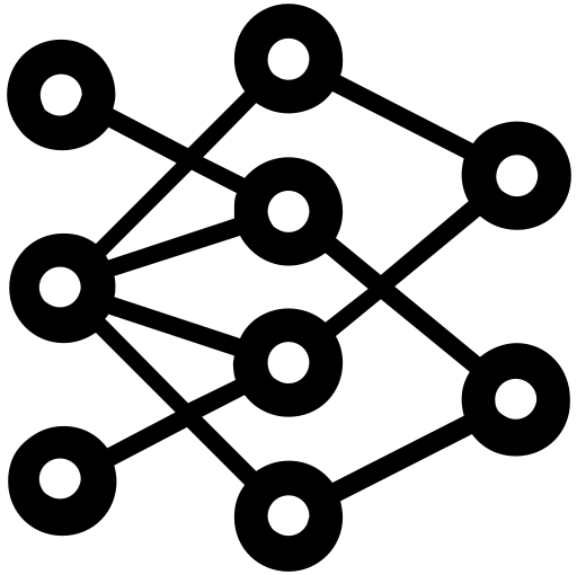
Can we develop a model that can handle an arbitrary random amount of missing values?

Our approach: Sturgeon



Micoarray data
2801 samples

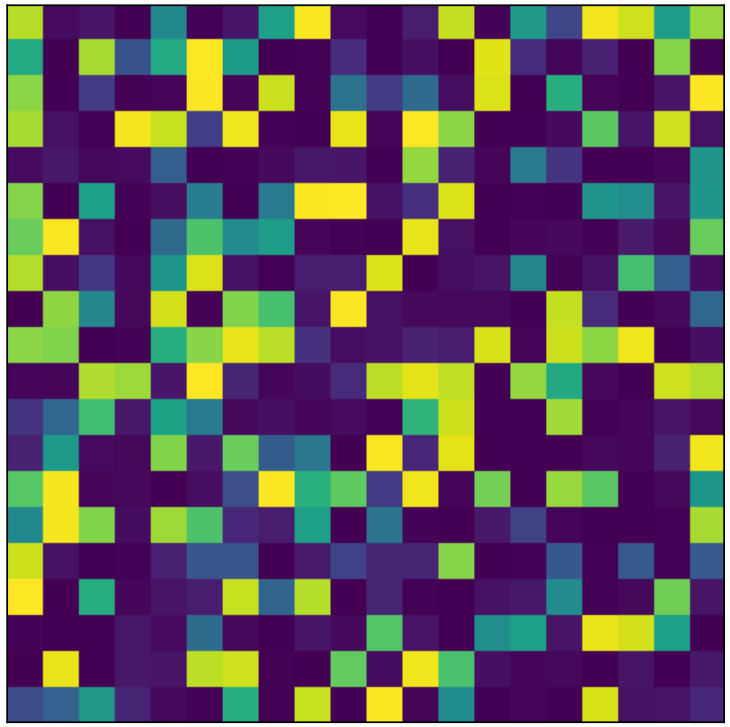
Simulate Nanopore data
> 36.000.000 simulations



Neural networks

Data simulation

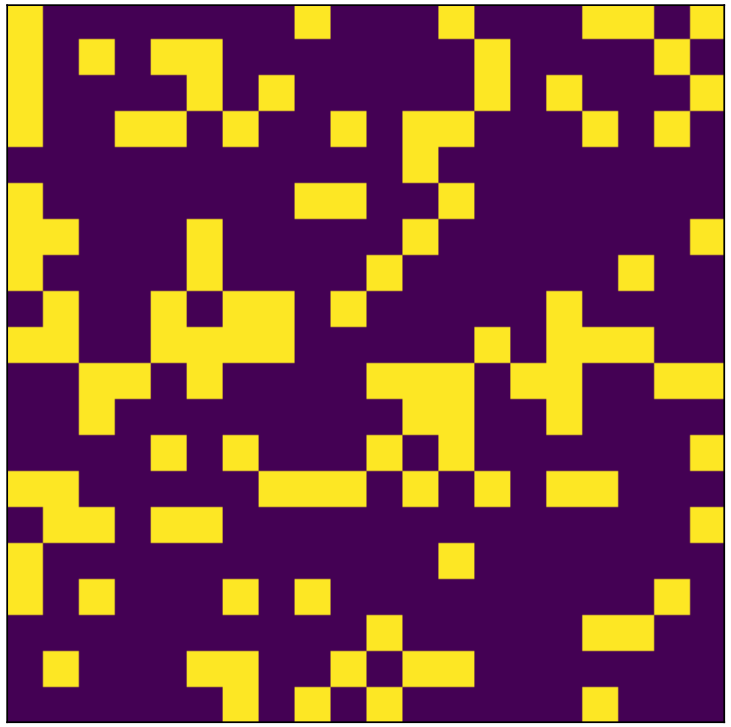
Microarray



Continuous measurement
Fraction of methylated reads

Data simulation

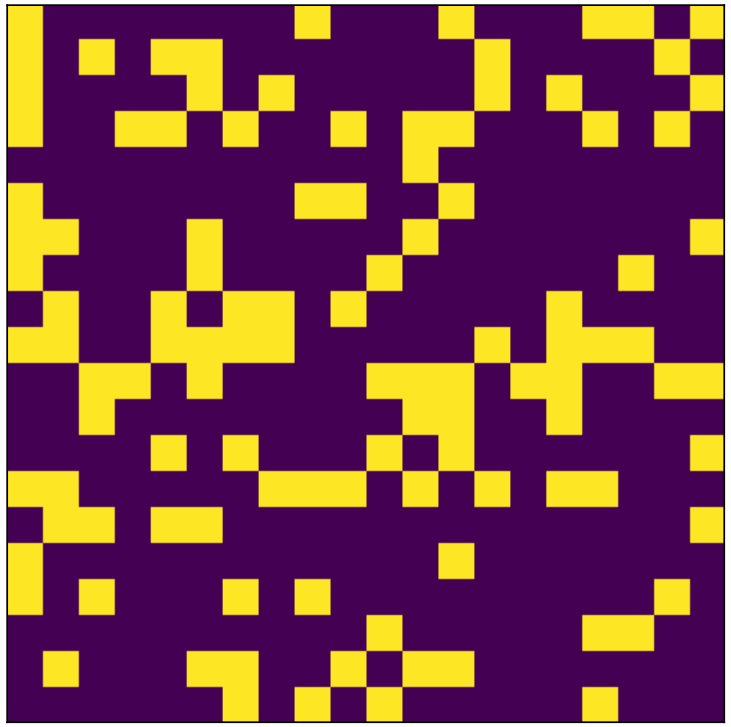
Microarray



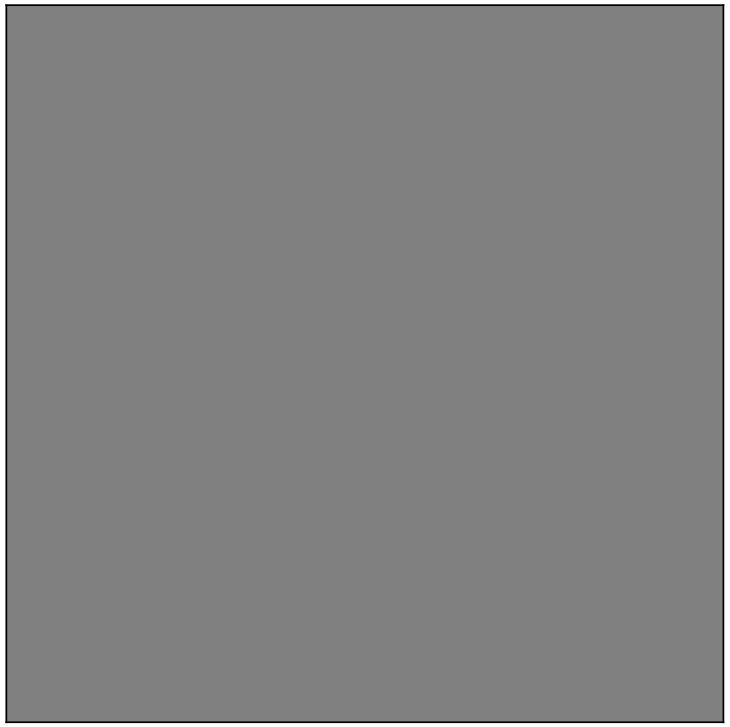
Binarized measurement
In shallow Nanopore sequencing
a site is methylated or not,
coverage of a site is very unlikely >1

Data simulation

Microarray

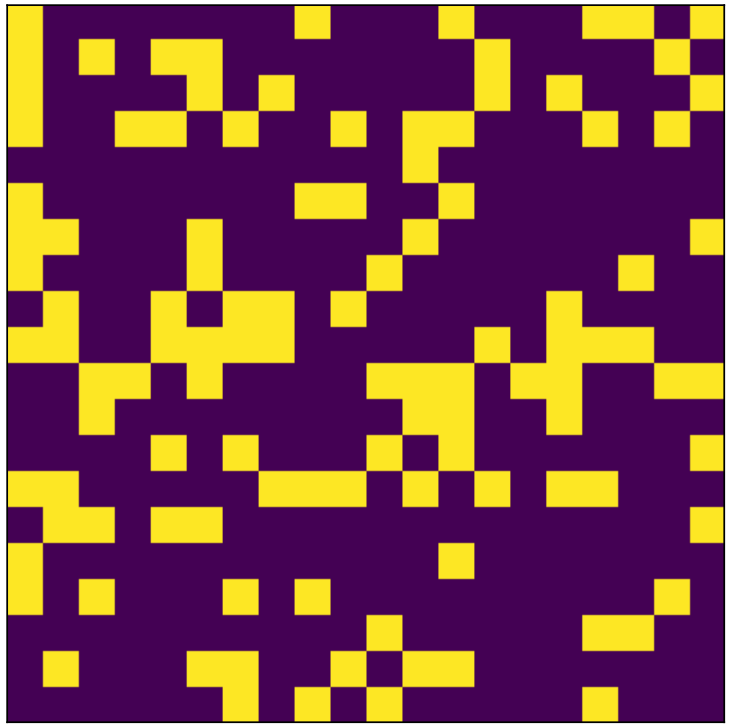


Nanopore



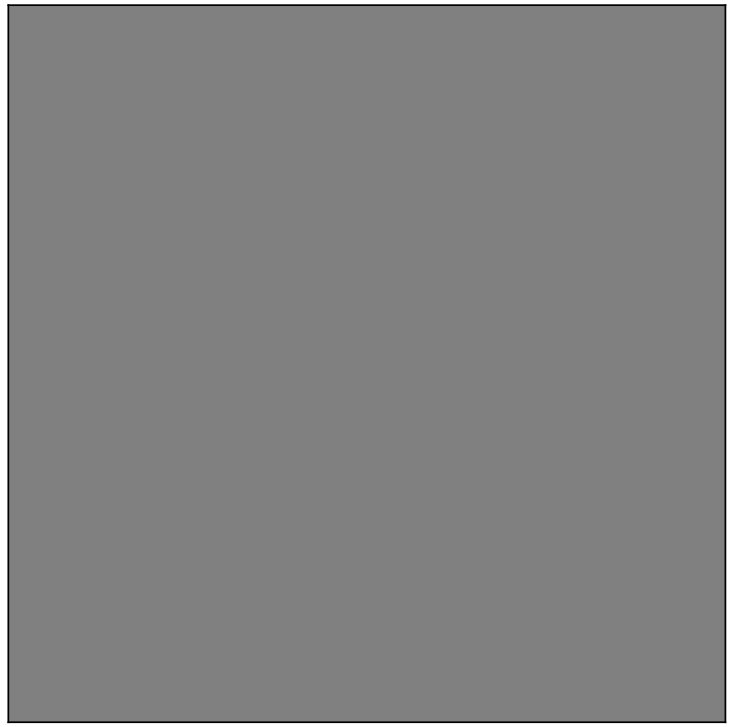
Data simulation

Microarray



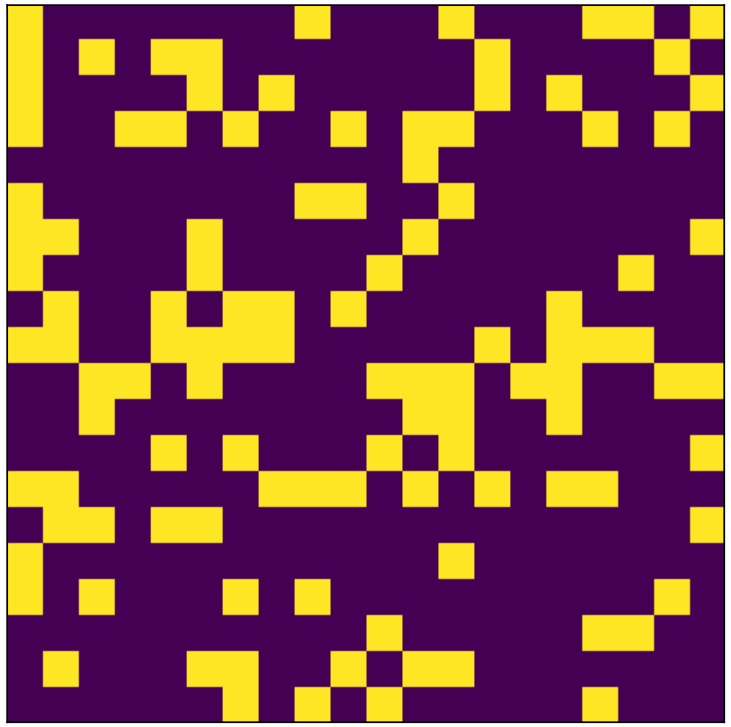
?

Nanopore



Data simulation

Microarray



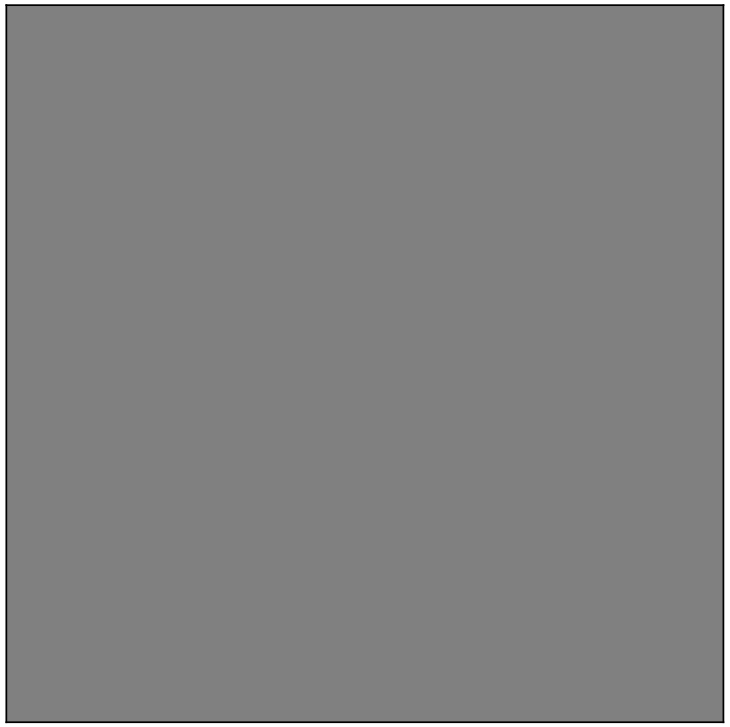
Read simulation

Genome position
[3 billion]

Strand
[Fwd, Rev]

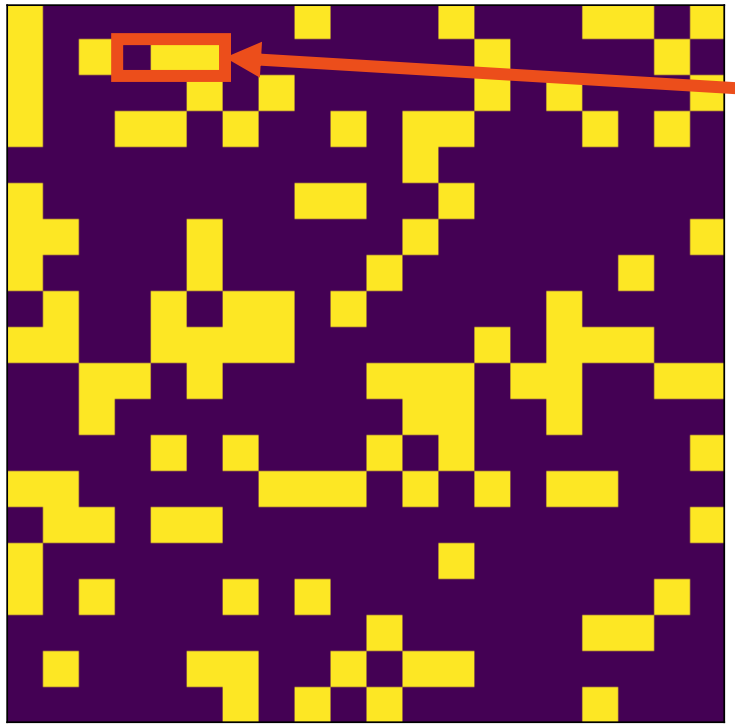
Read length
[Distribution]

Nanopore



Data simulation

Microarray



Chr3
45000–46000
Fwd

x1

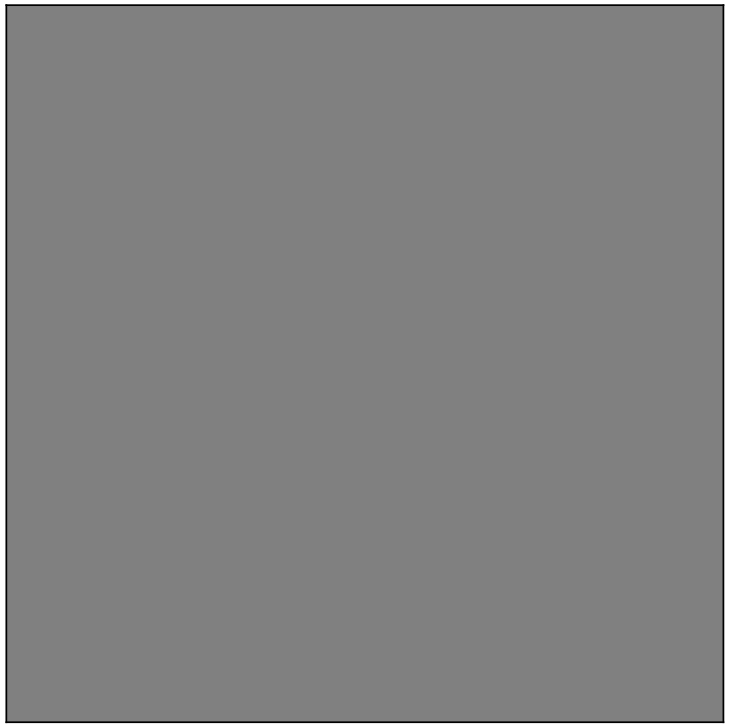
Read simulation

Genome position
[3 billion]

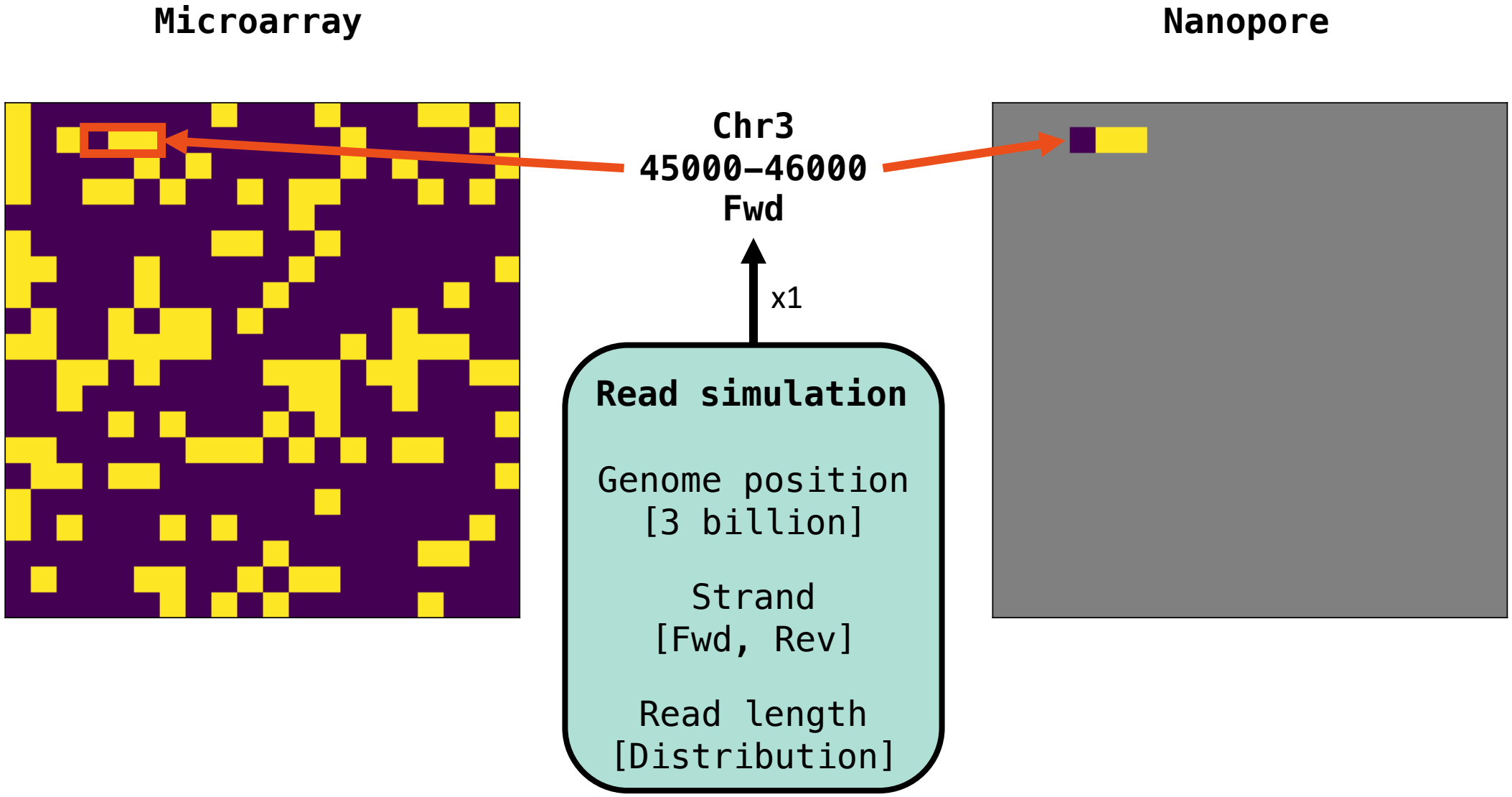
Strand
[Fwd, Rev]

Read length
[Distribution]

Nanopore

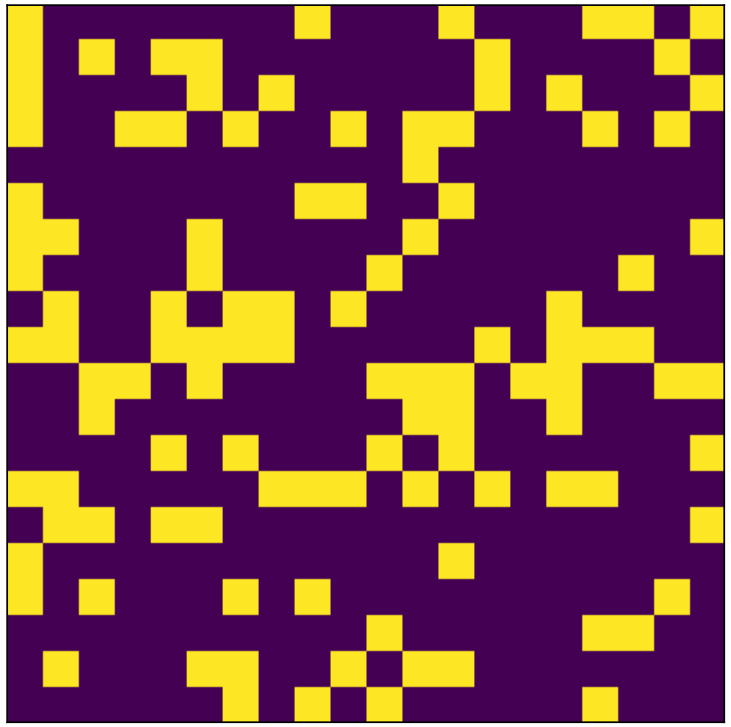


Data simulation



Data simulation

Microarray



Chr17
30000–50000
Rev

x2

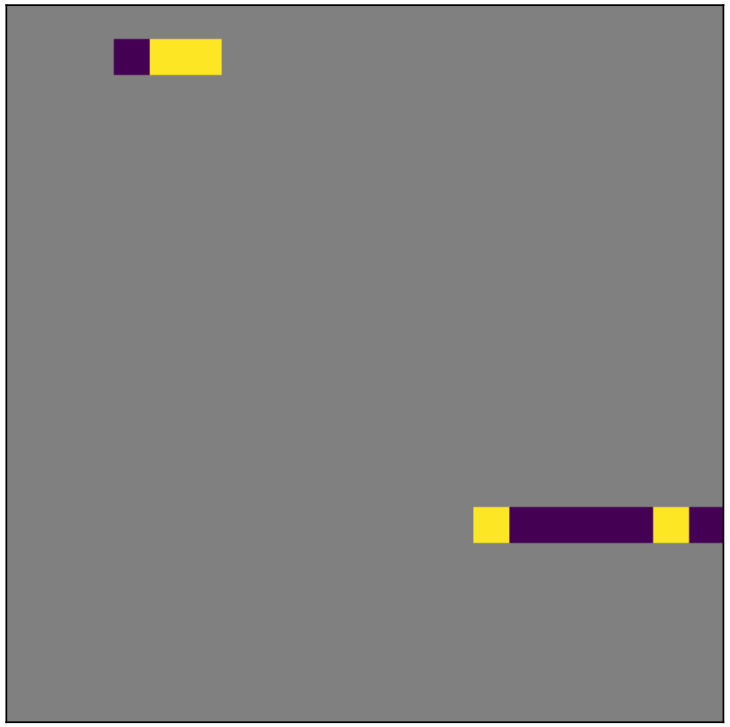
Read simulation

Genome position
[3 billion]

Strand
[Fwd, Rev]

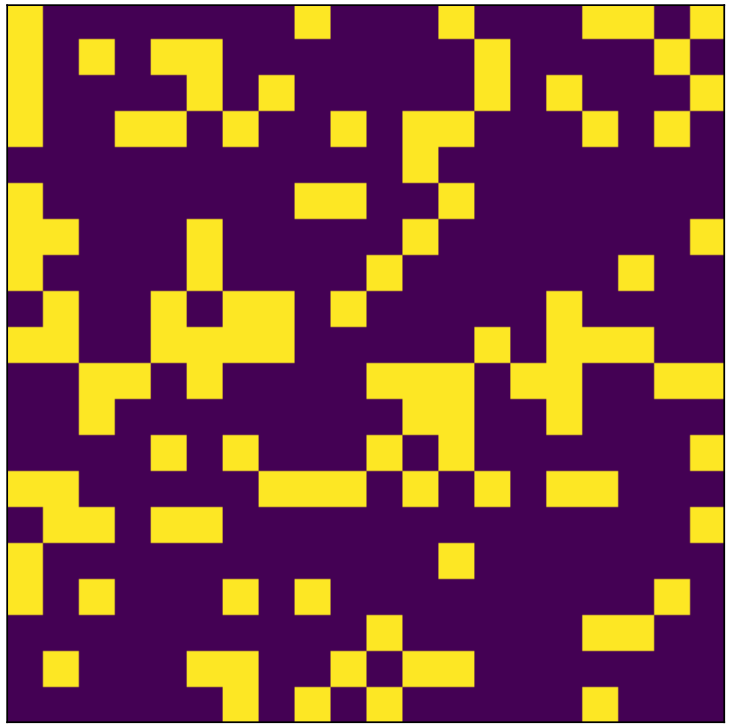
Read length
[Distribution]

Nanopore



Data simulation

Microarray



Chr18
2000–10000
Rev

x3

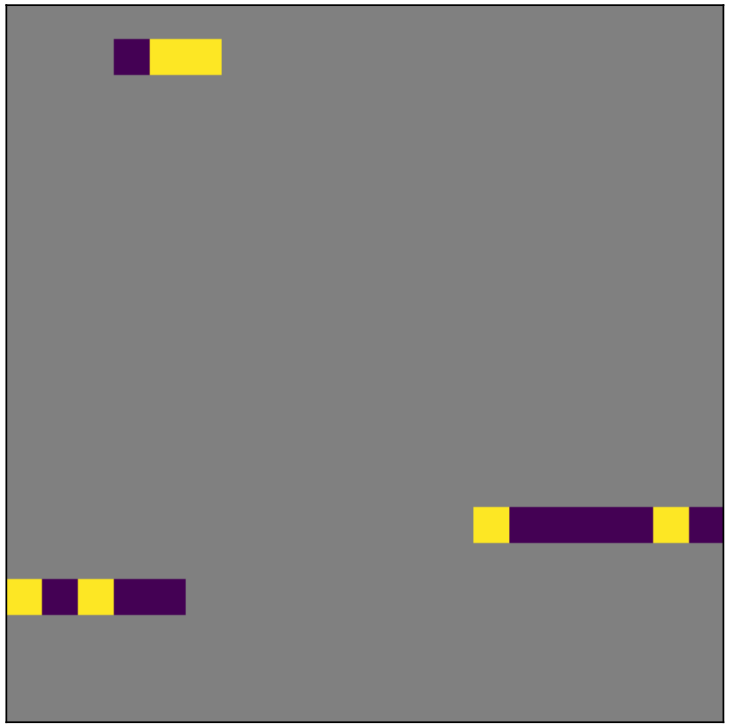
Read simulation

Genome position
[3 billion]

Strand
[Fwd, Rev]

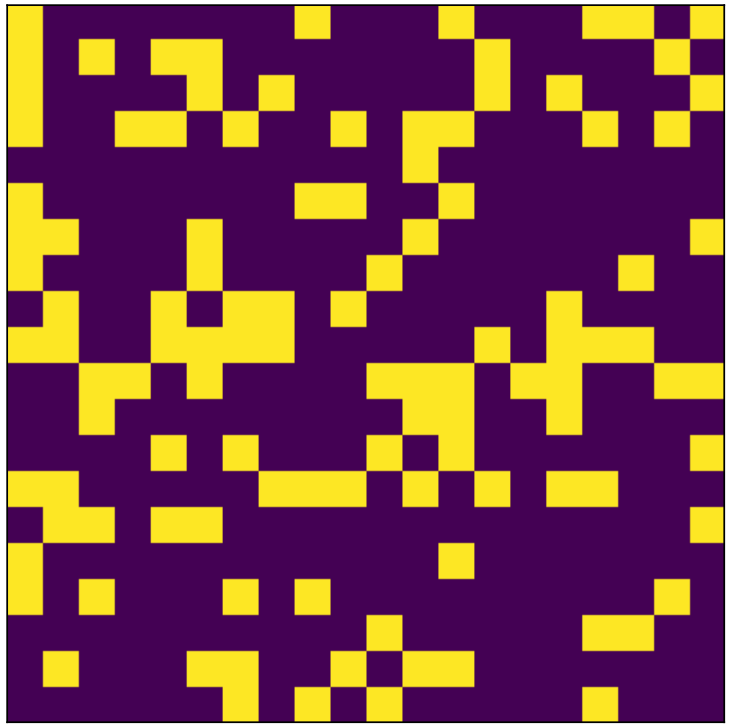
Read length
[Distribution]

Nanopore



Data simulation

Microarray



Chr10
12000–20000
Rev

x4

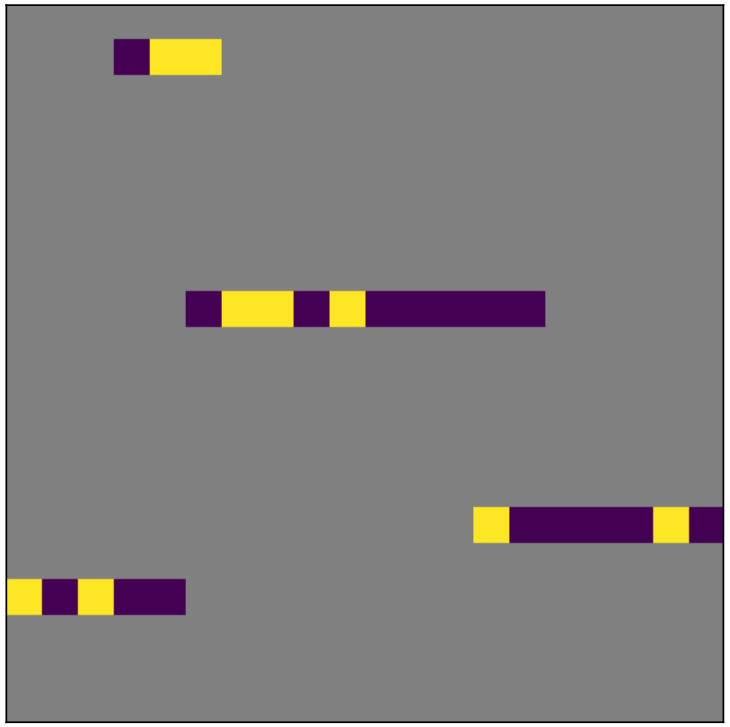
Read simulation

Genome position
[3 billion]

Strand
[Fwd, Rev]

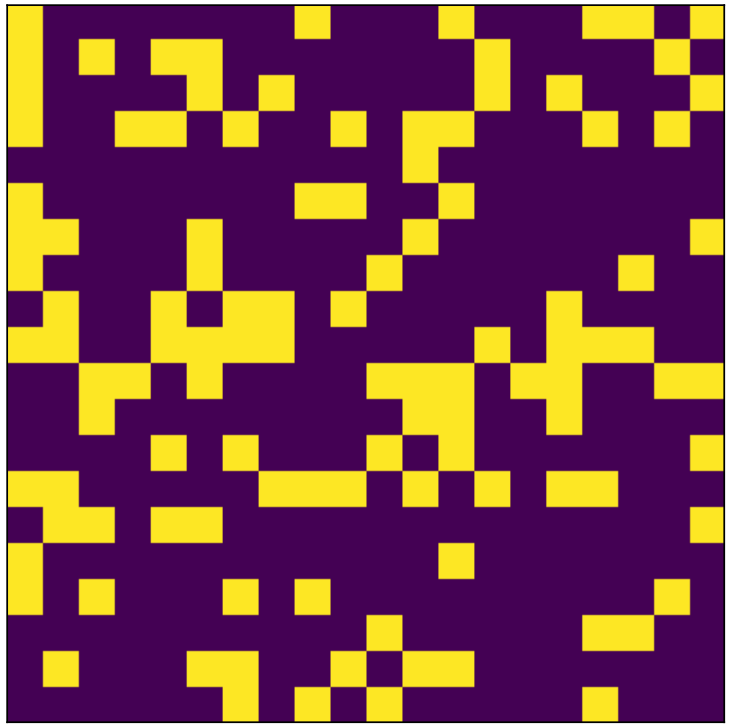
Read length
[Distribution]

Nanopore



Data simulation

Microarray



Chr8
25000–45000
Fwd

x5

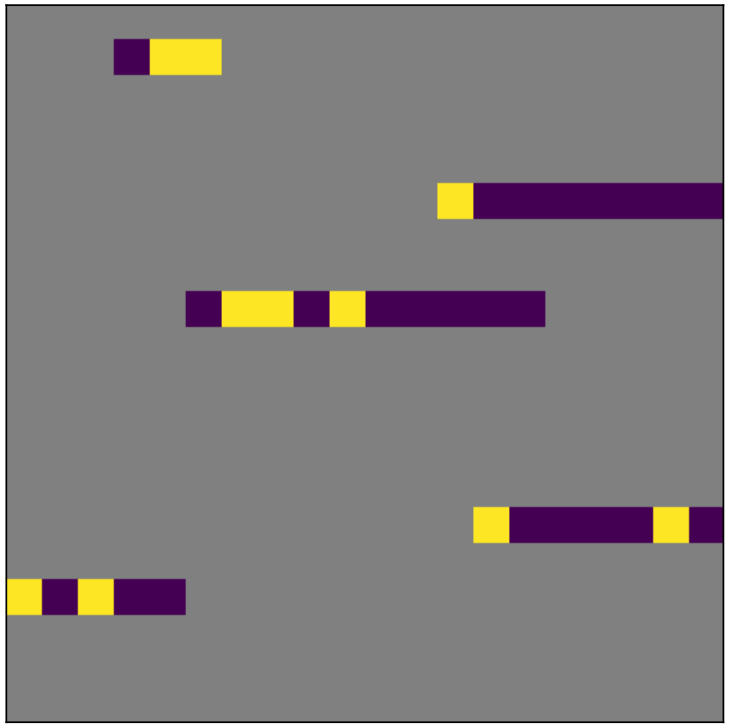
Read simulation

Genome position
[3 billion]

Strand
[Fwd, Rev]

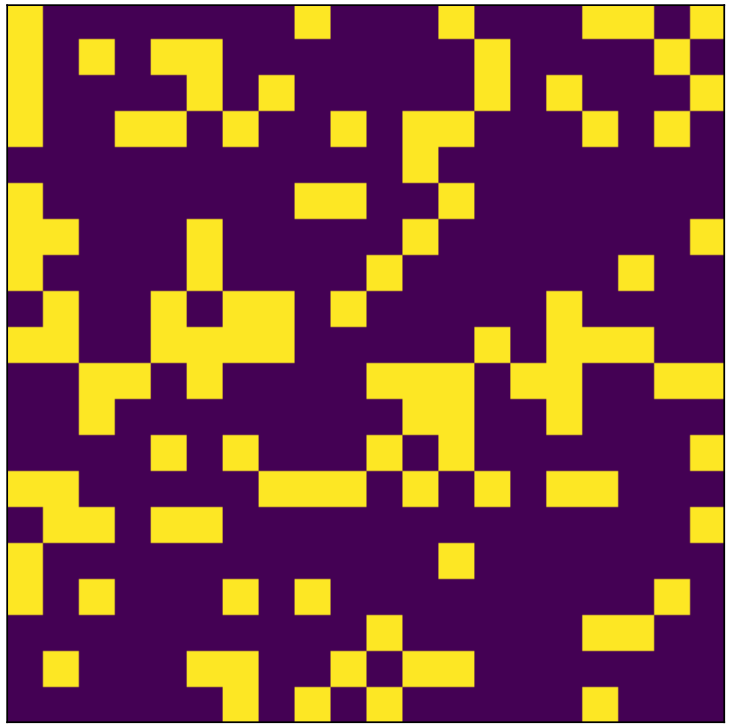
Read length
[Distribution]

Nanopore



Data simulation

Microarray



Chr8
25000–45000
Fwd

x5

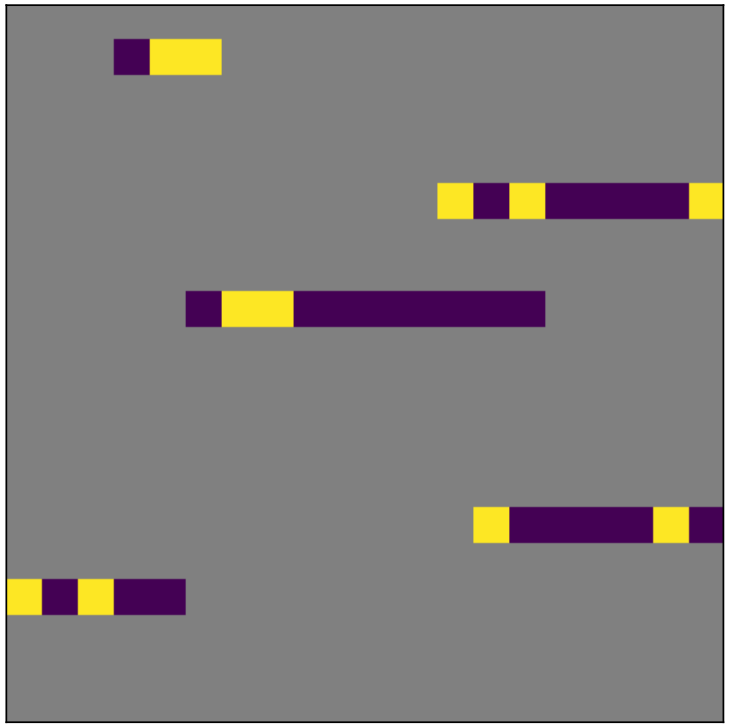
Read simulation

Genome position
[3 billion]

Strand
[Fwd, Rev]

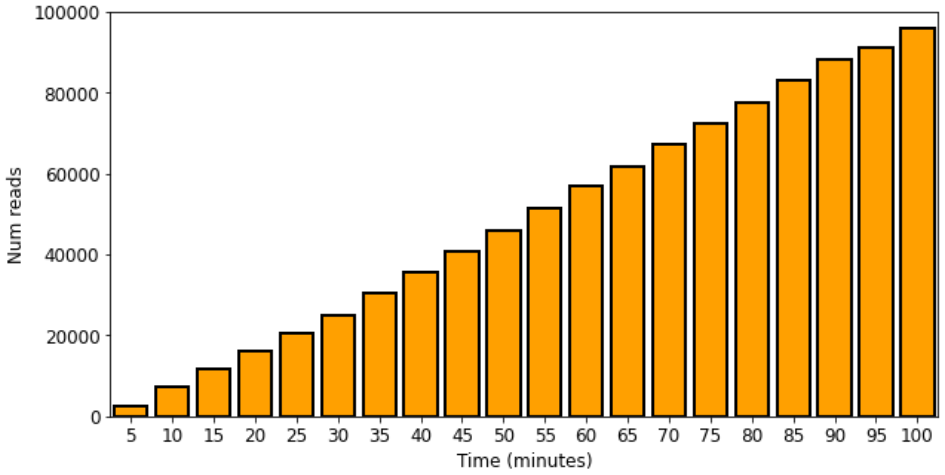
Read length
[Distribution]

Nanopore



+10% noise
Flip methylation status

Data simulation



How many reads do we simulate?
How much time,
do we want to sequence?

ChrN
Start–End
Strand

xN

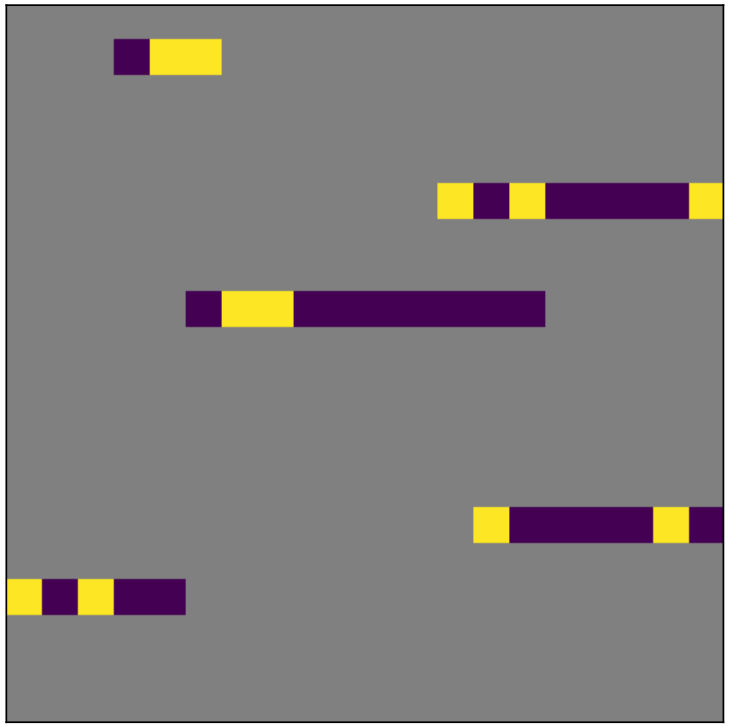
Read simulation

Genome position
[3 billion]

Strand
[Fwd, Rev]

Read length
[Distribution]

Nanopore



+10% noise
Flip methylation status

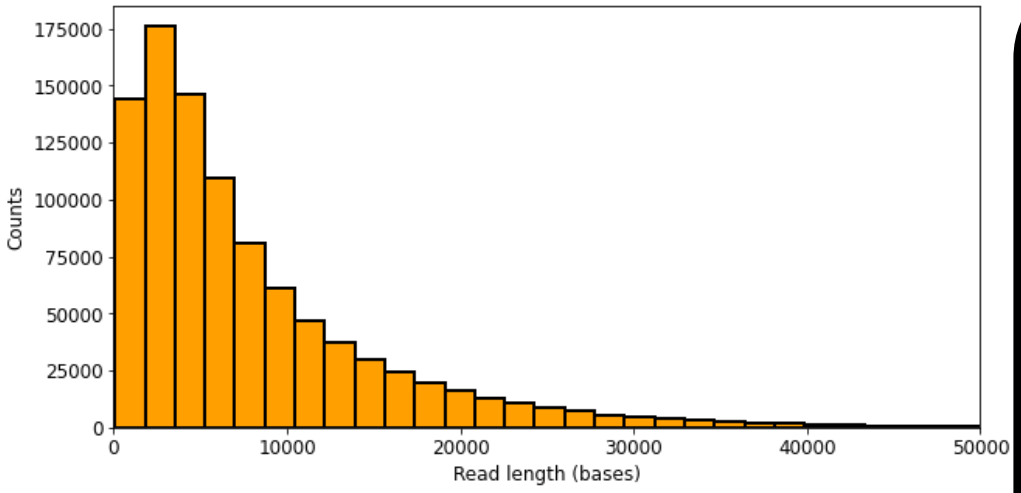
Data simulation

Nanopore

ChrN
Start–End
Strand

xN

How long should be the simulated reads?

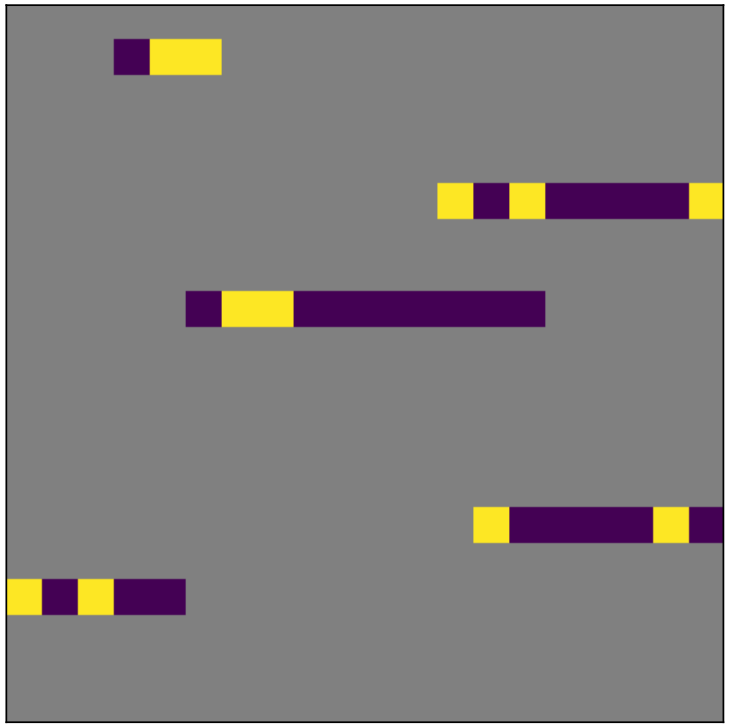


Read simulation

Genome position
[3 billion]

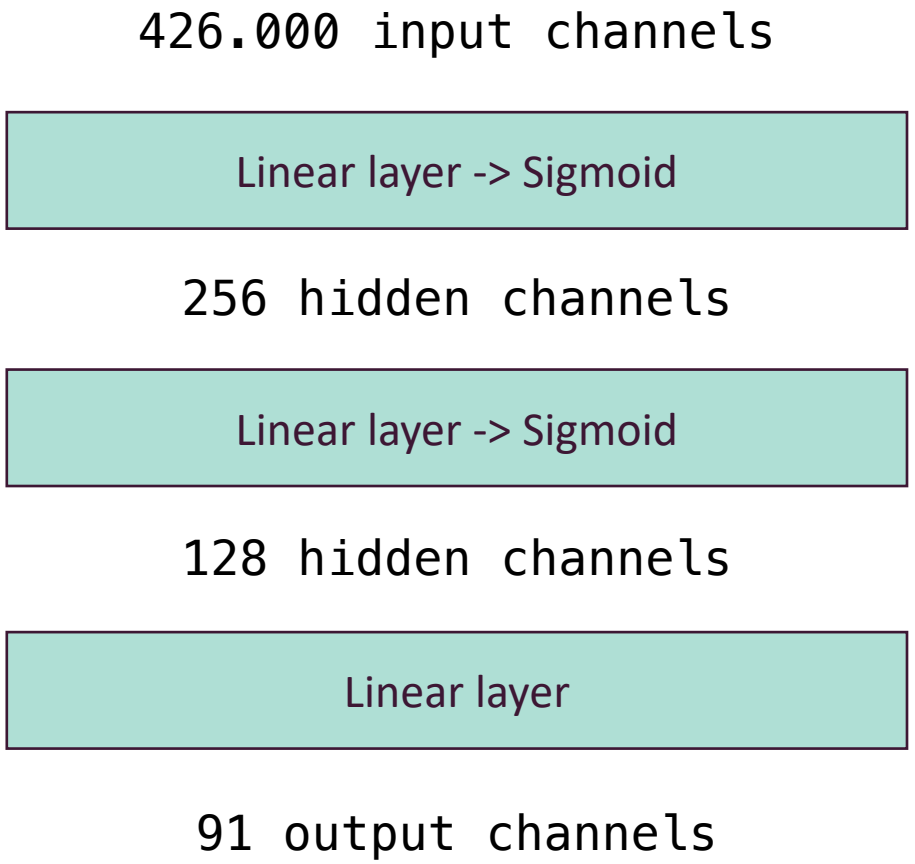
Strand
[Fwd, Rev]

Read length
[Distribution]



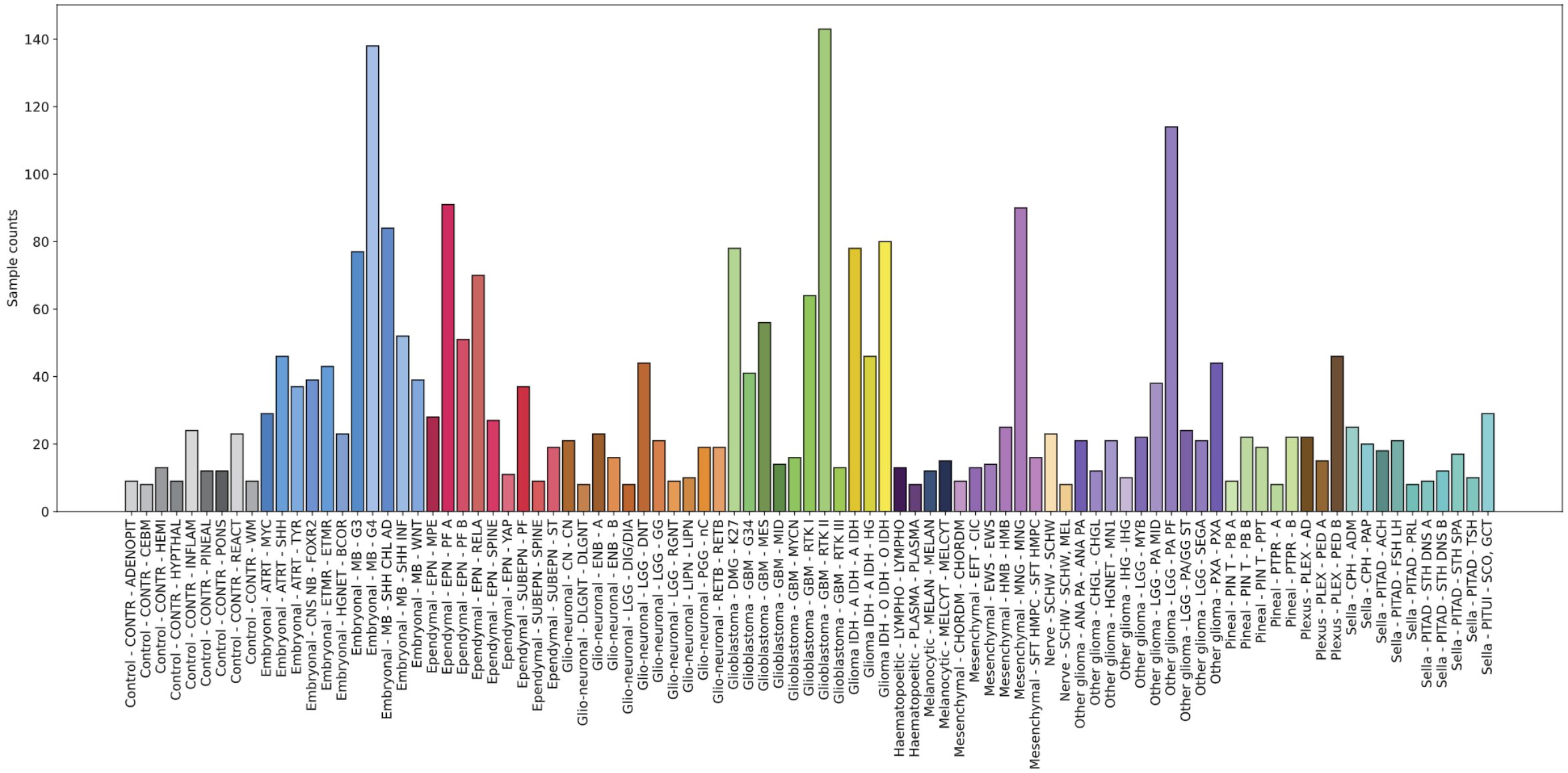
+10% noise
Flip methylation status

Neural network

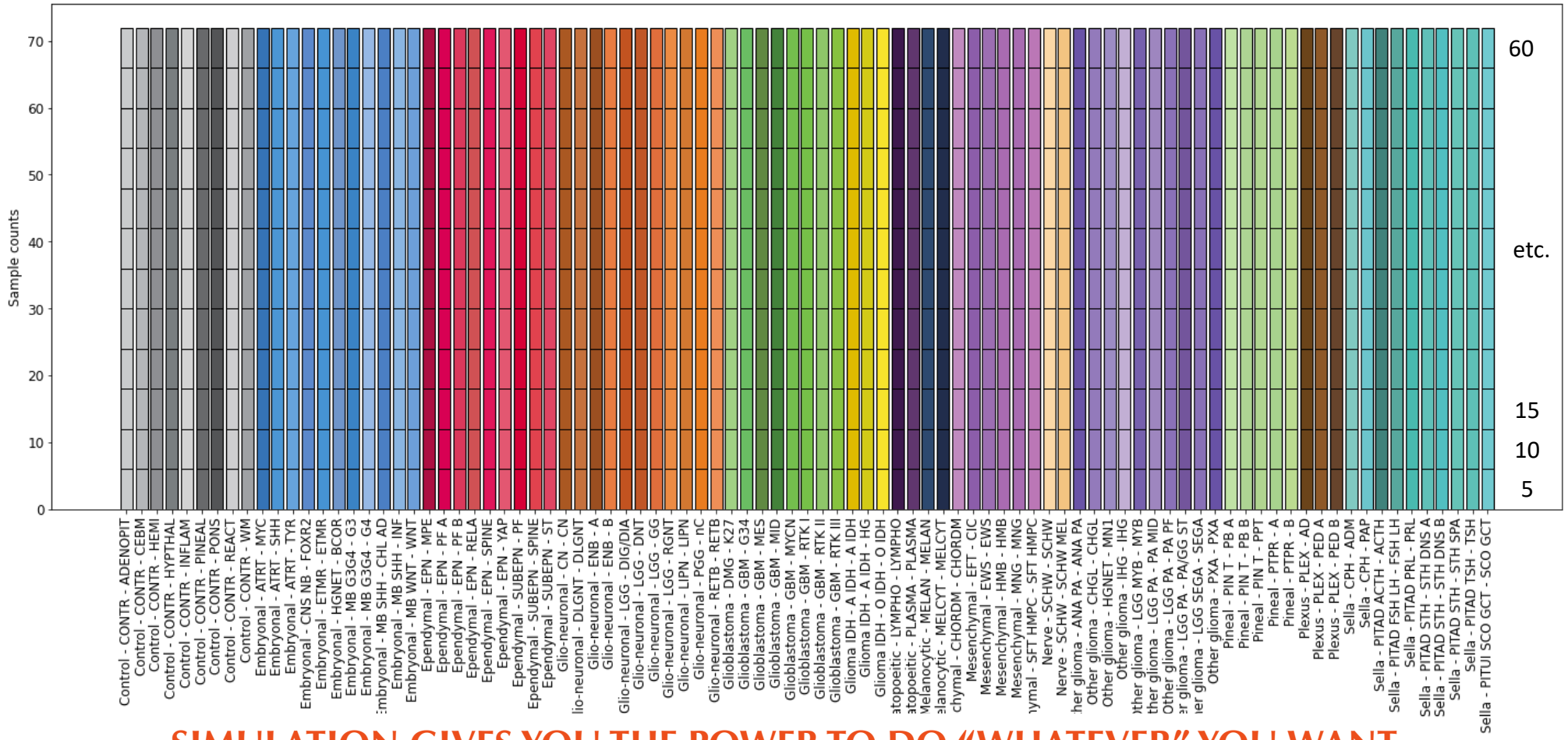


Extremely simple feed forward network because the data has no structure

Cross-validation



Class balancing via up-sampling



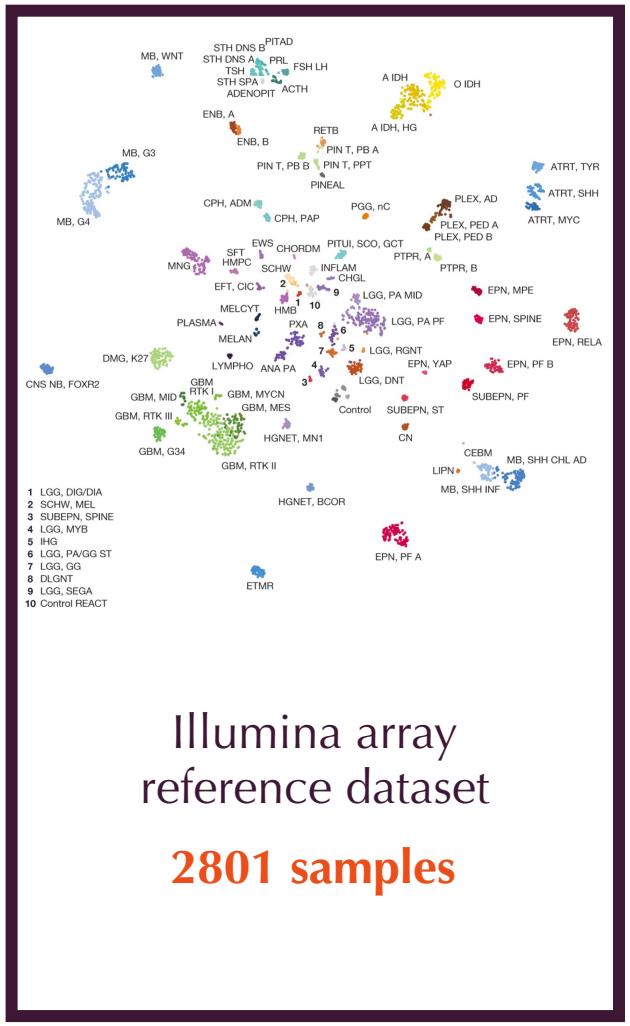
Simulation time (minutes)

SIMULATION GIVES YOU THE POWER TO DO "WHATEVER" YOU WANT

Upsample all classes to the number of samples of the most represented class
Equal simulation times for each class and number of samples

Cross-validation

Keep balanced class proportions

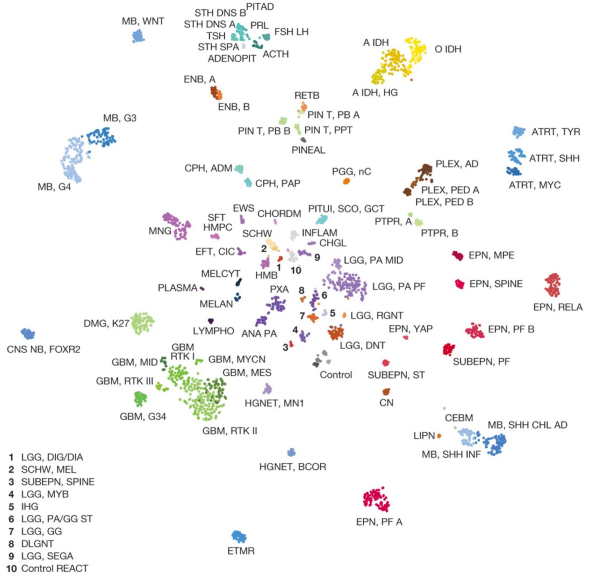


Illumina array reference dataset

2801 samples

25%	**Train**	
25%		
25%	**Validation**	Stop training Model calibration
25%	**Test**	Final model performance

Cross-validation



Illumina array reference dataset

2801 samples

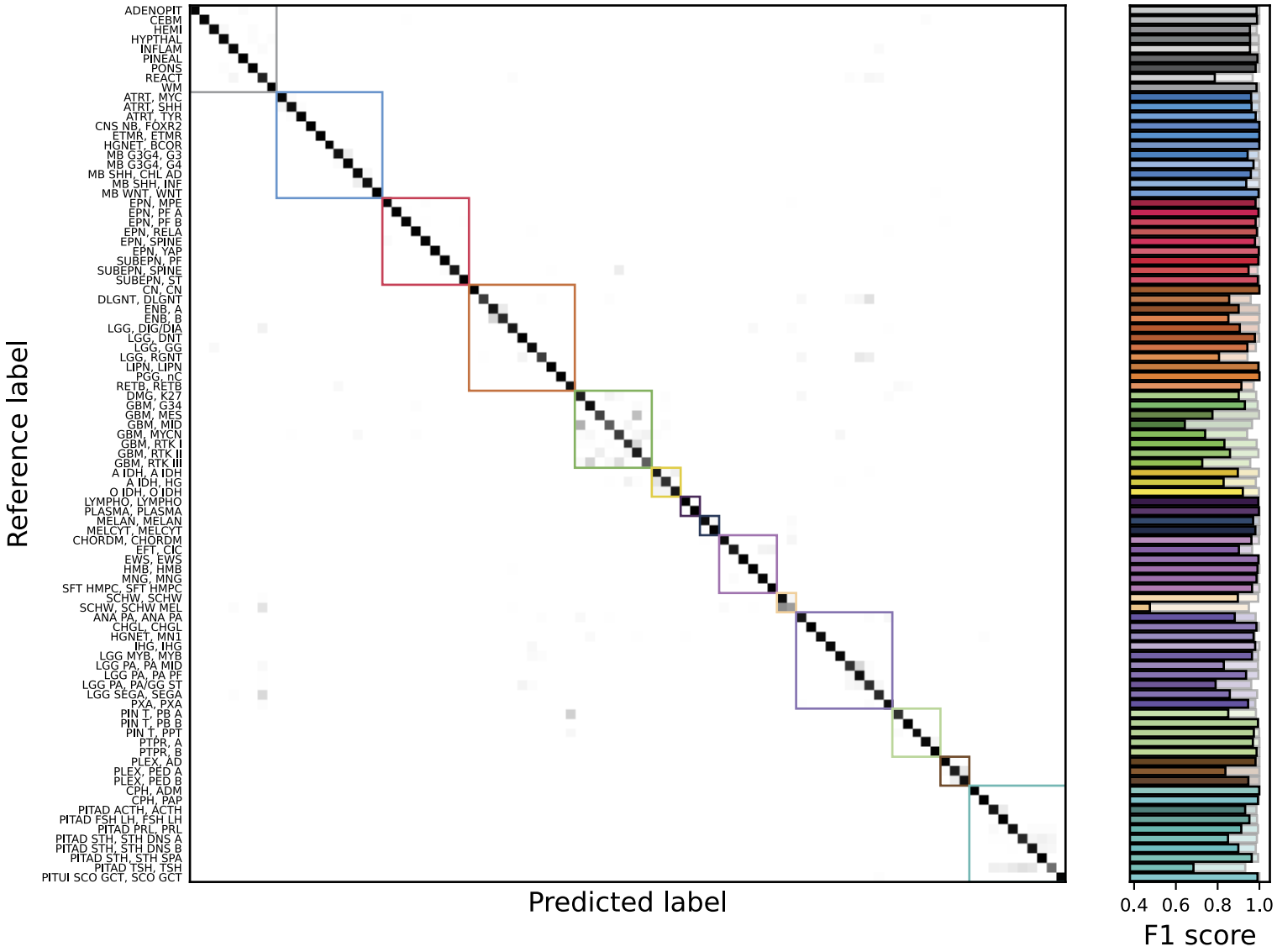
Keep balanced class proportions

25%	**Train**
25%	
25%	**Validation**
25%	**Test**

Simulation seeds (non-overlapping)

1.000.000 (36.8 million)
500 (4.2 million)
500 (4.2 million)

Performance on simulations

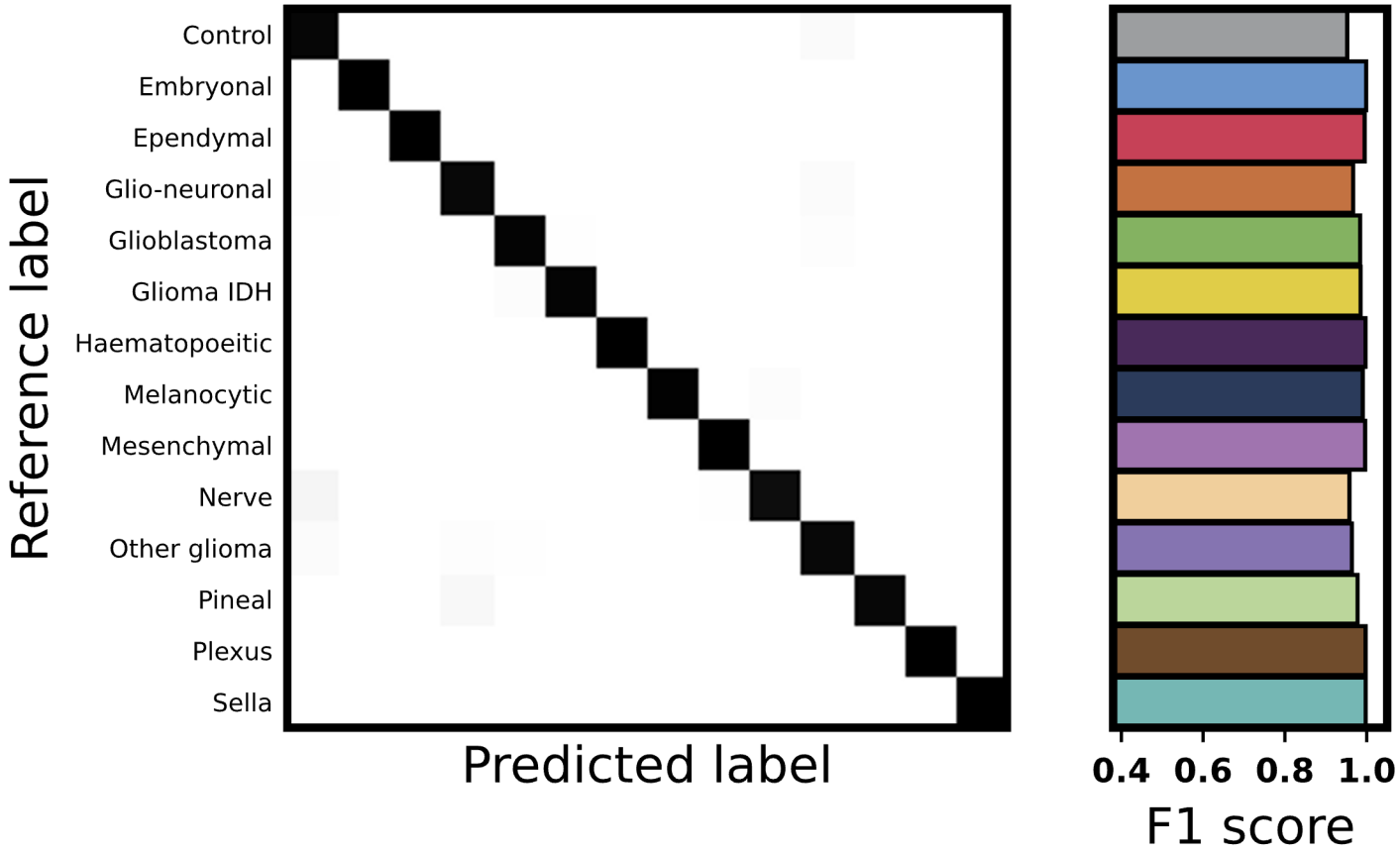


Performance at **20 minutes** simulations

F1-Score
Top1: 0.926
Top3: 0.990

$$F1 = \frac{2TP}{2TP + FP + FN}$$

Performance on simulations

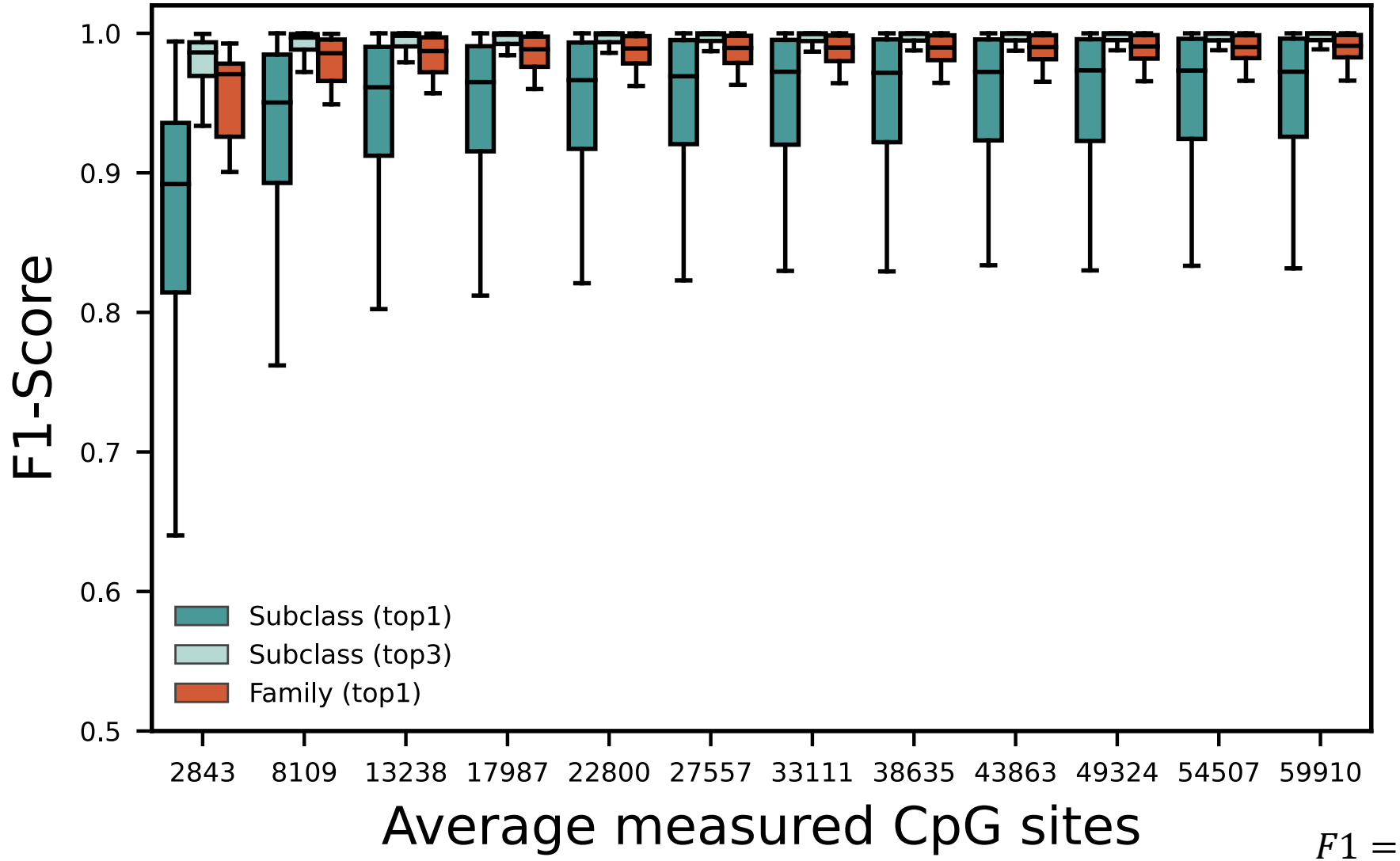


Performance at **20 minutes** simulations

F1-Score
Top1: 0.982

$$F1 = \frac{2TP}{2TP + FP + FN}$$

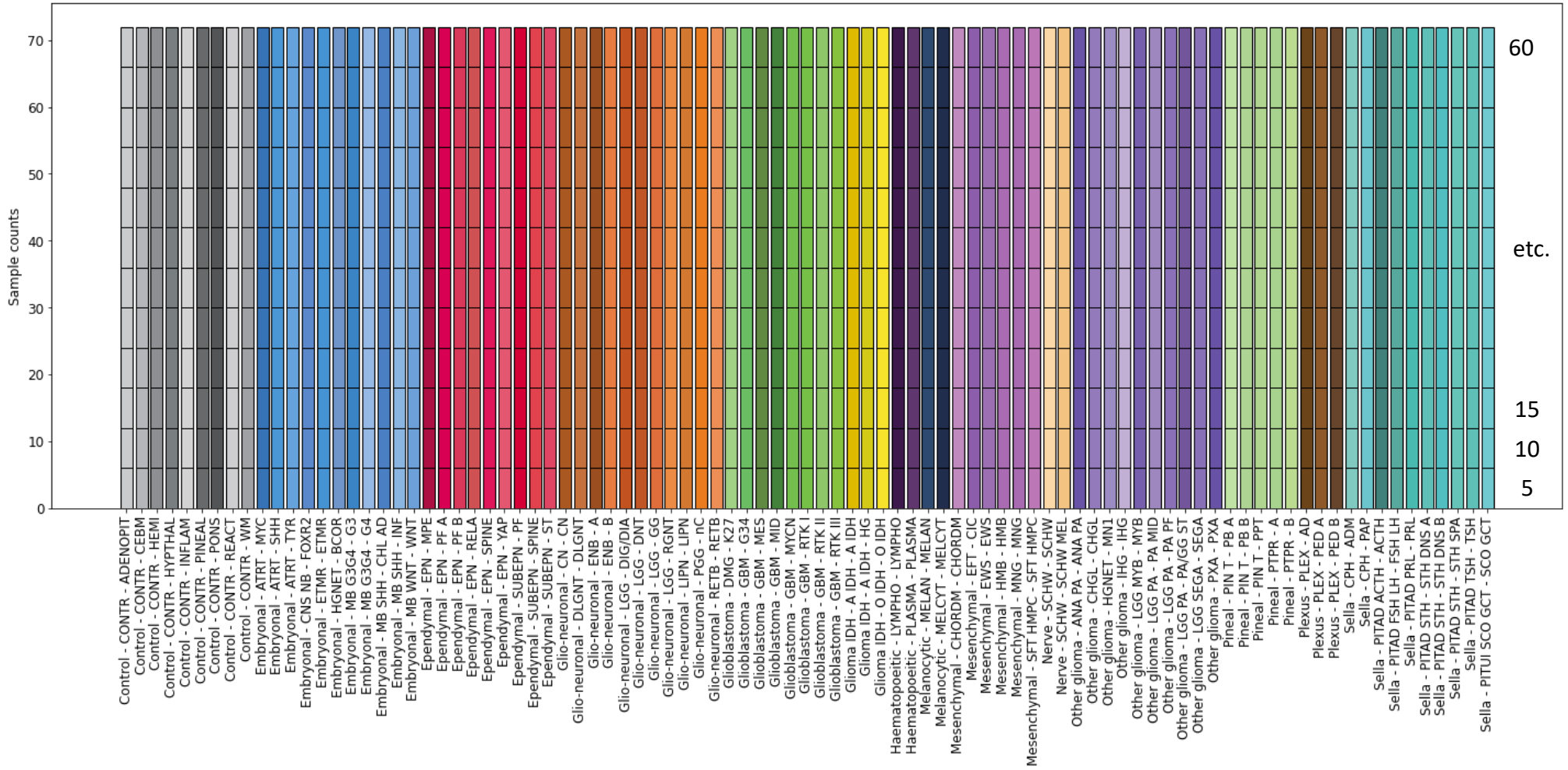
Performance on simulations



$$F1 = \frac{2TP}{2TP + FP + FN}$$

Adaptive sampling

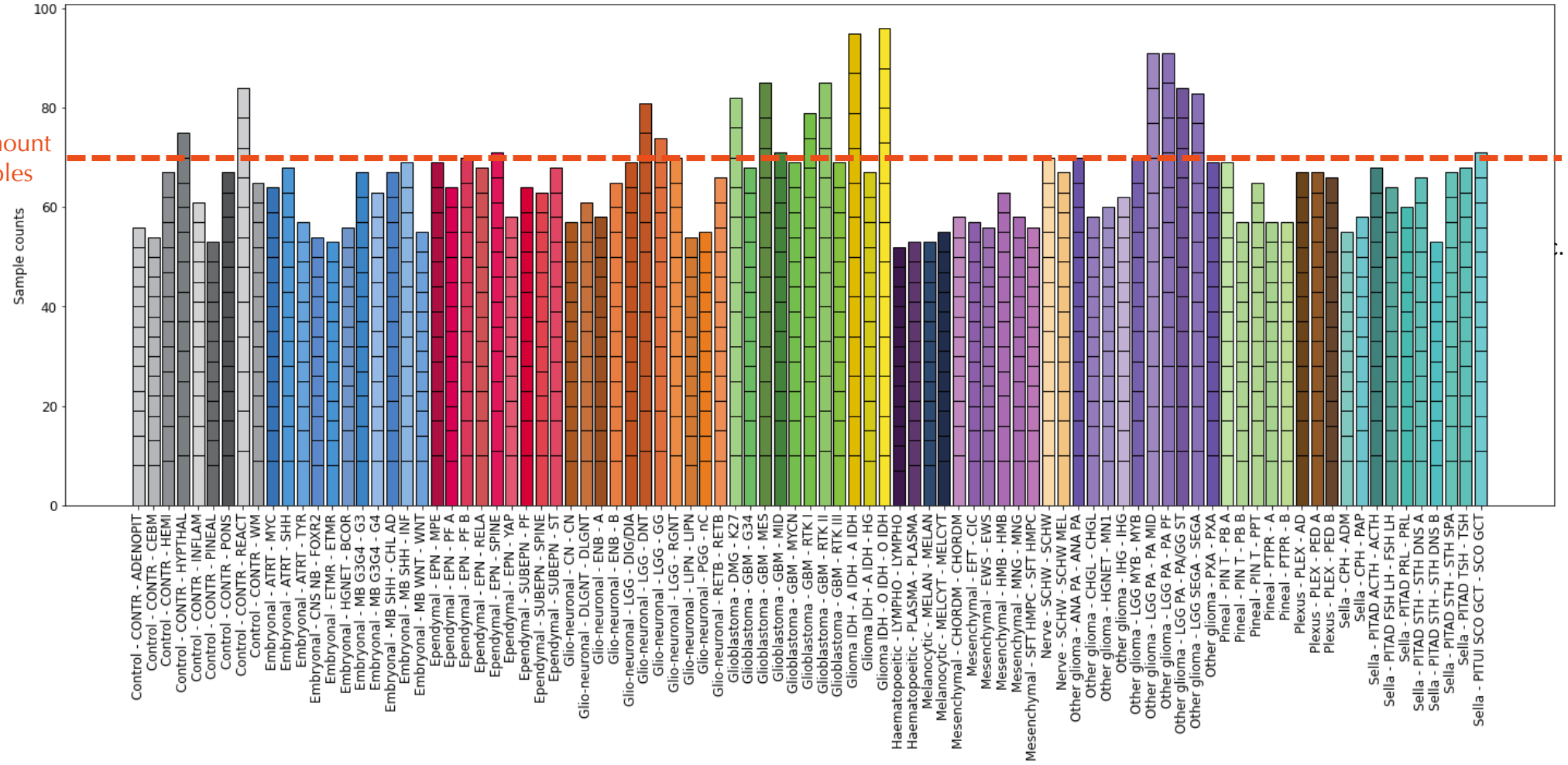
Epoch 1



Simulation time (minutes)

Give homework on the classes and simulation times that the model does worse

Adaptive sampling



Simulation time (minutes)

Give homework on the classes and simulation times that the model does worse

Adaptive sampling

5 min

60 min

Adaptive sampling
improves performance
at lower simulation times

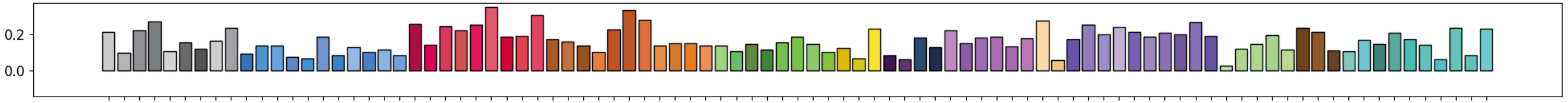
Adaptive sampling

5 min

60 min

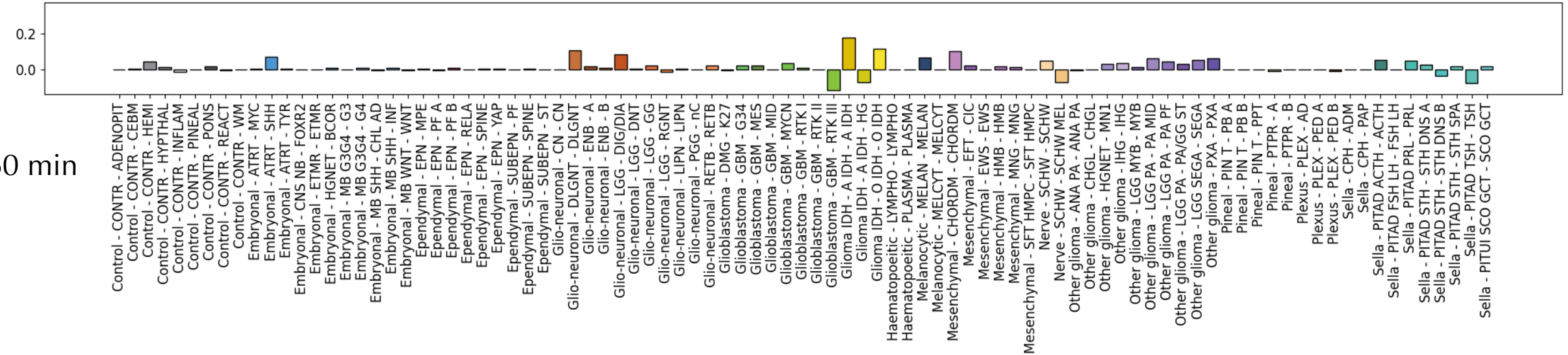
Adaptive sampling
improves performance
at lower simulation times

Adaptive sampling



5 min

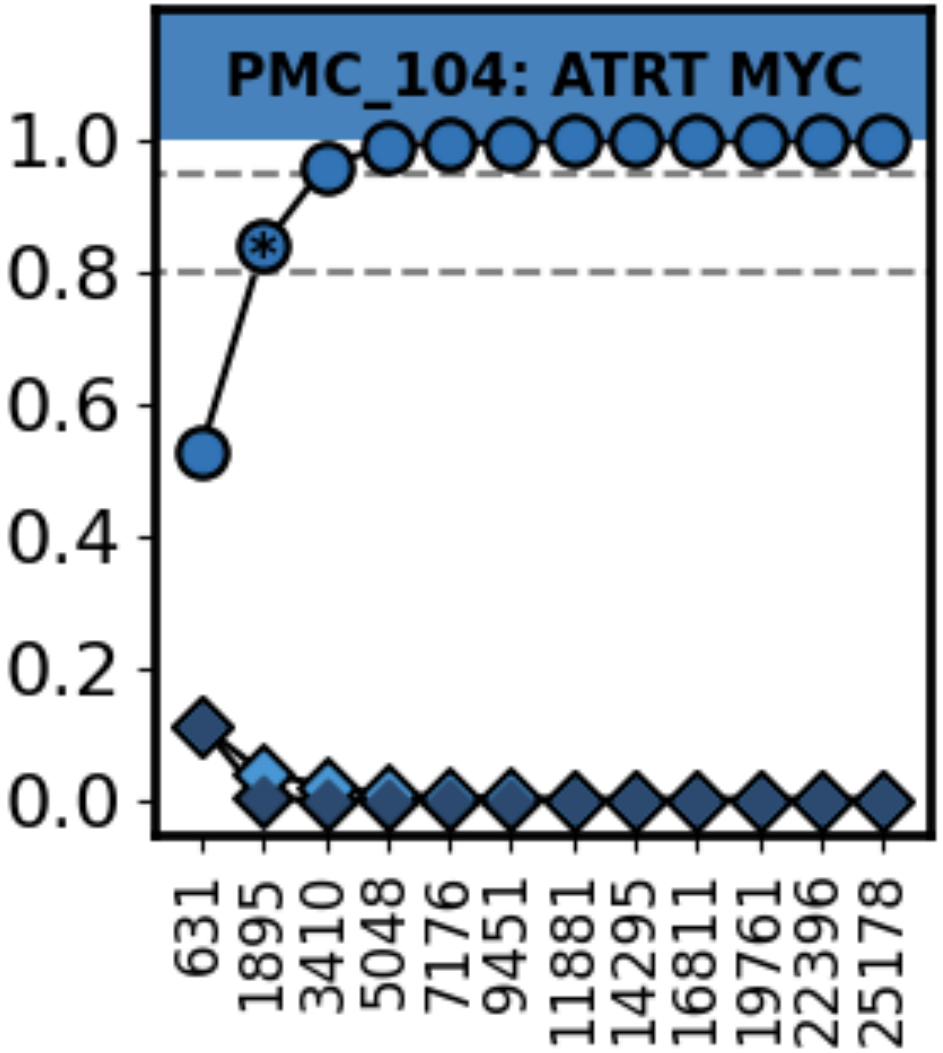
60 min



Nanopore

All this is nice, but it's all simulations
Does it actually work on Nanopore data

PMC Nanopore

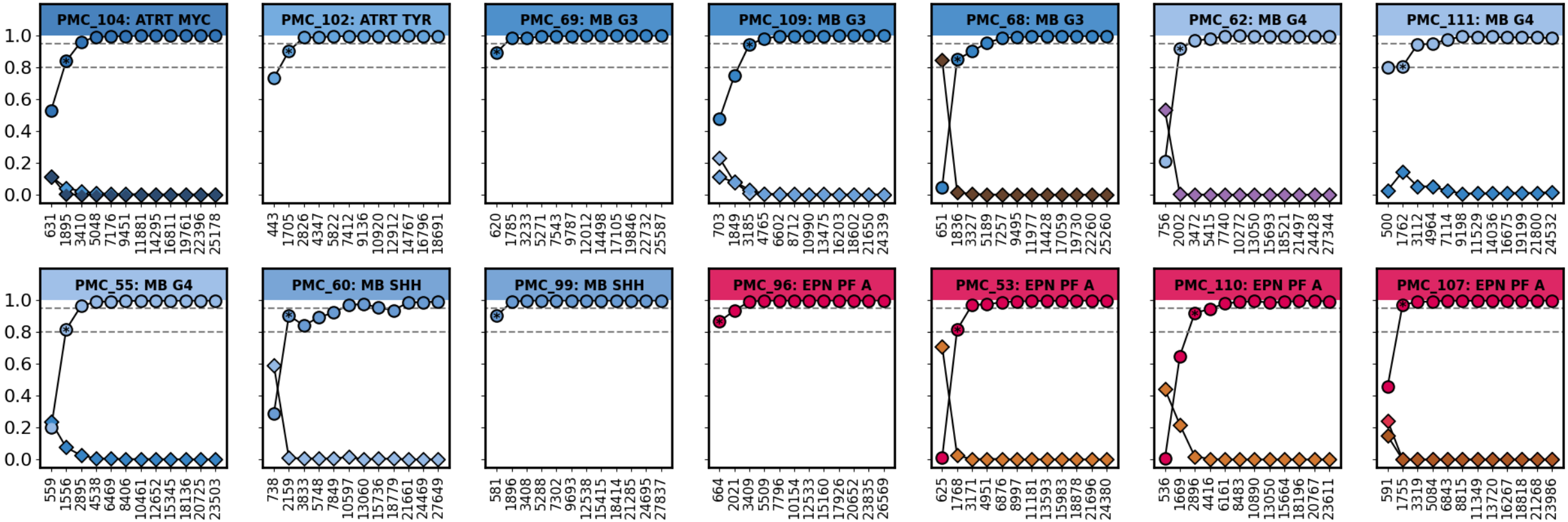


Circles = correct class

Rombes = incorrect class

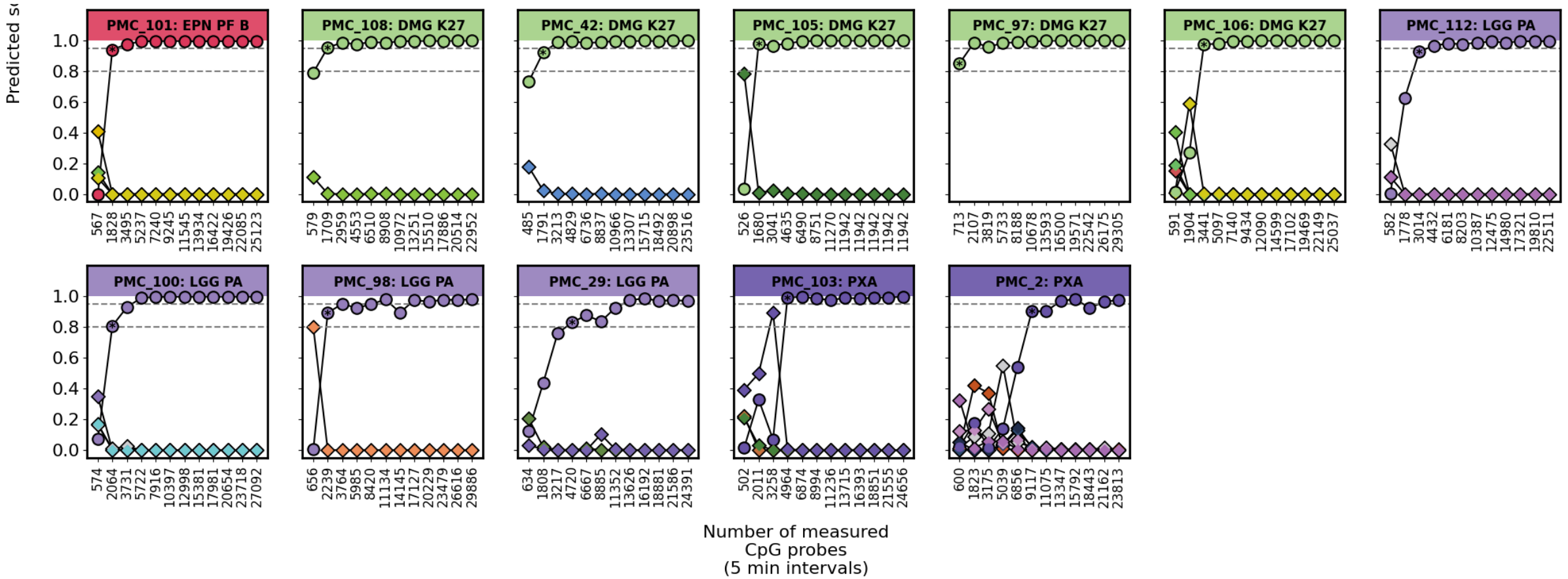
Each tick is 5 minutes of sequencing

PMC Nanopore

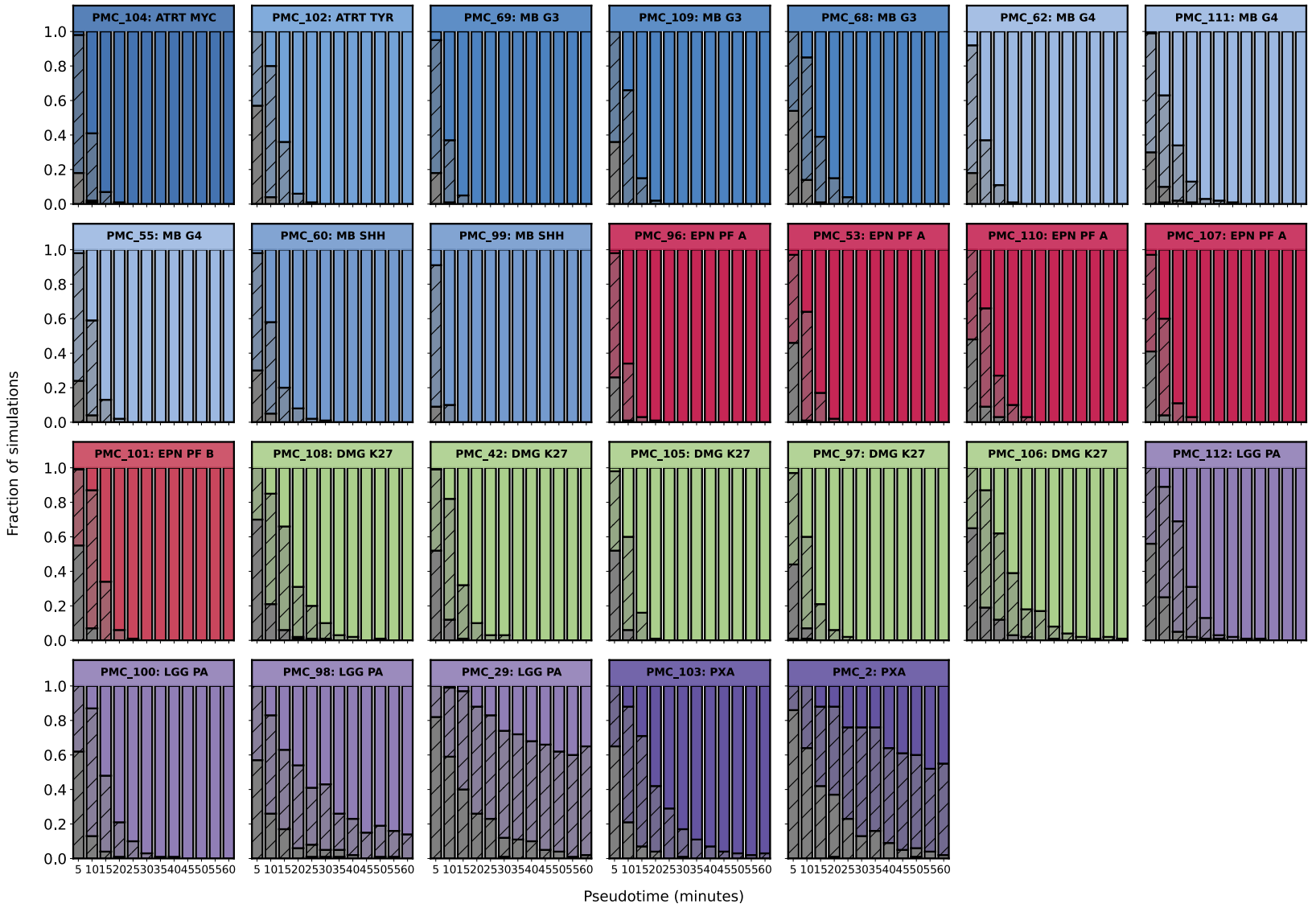


core

PMC Nanopore



Simulation power



Oslo Nanopore

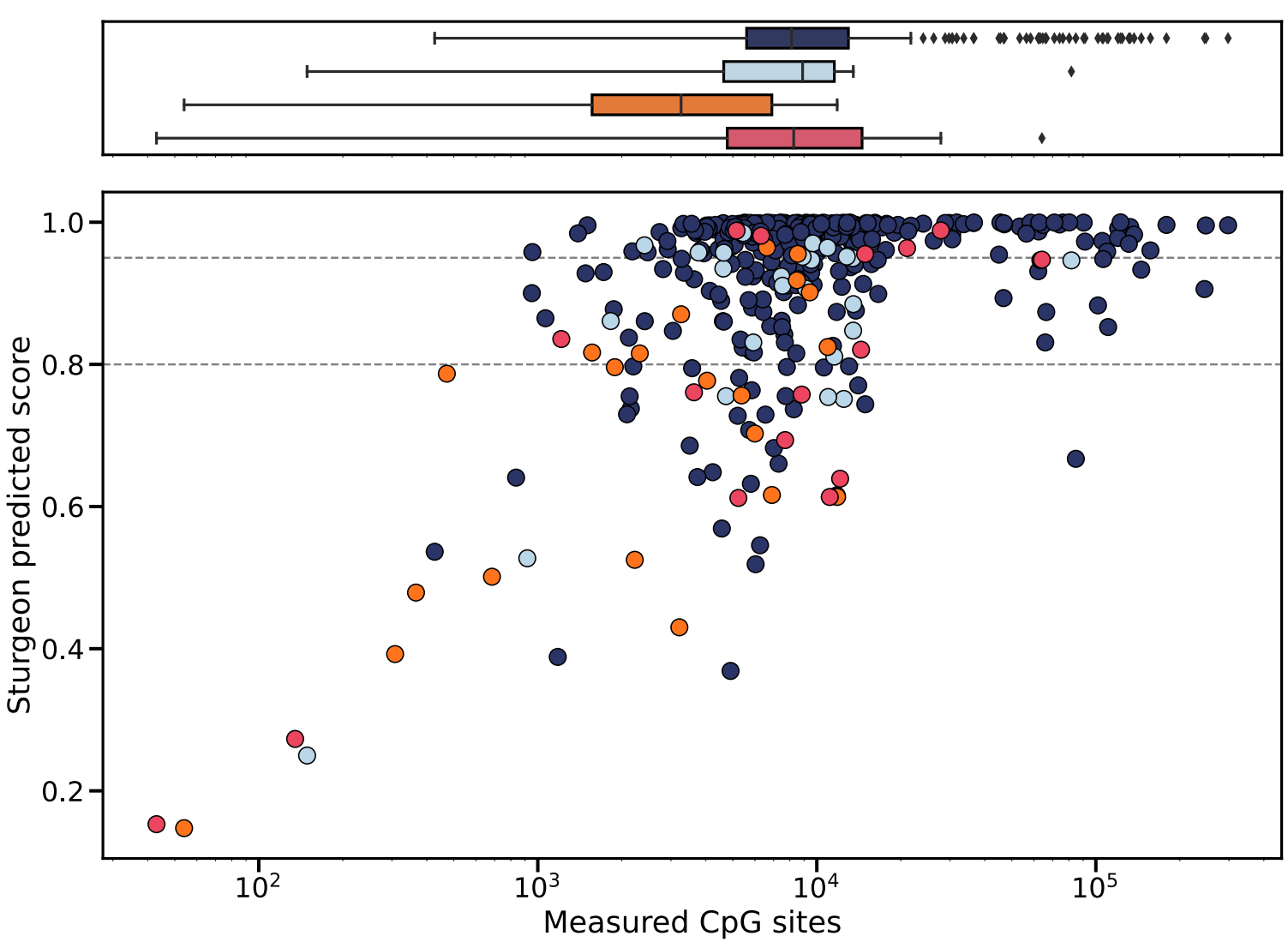
ORIGINAL ARTICLE

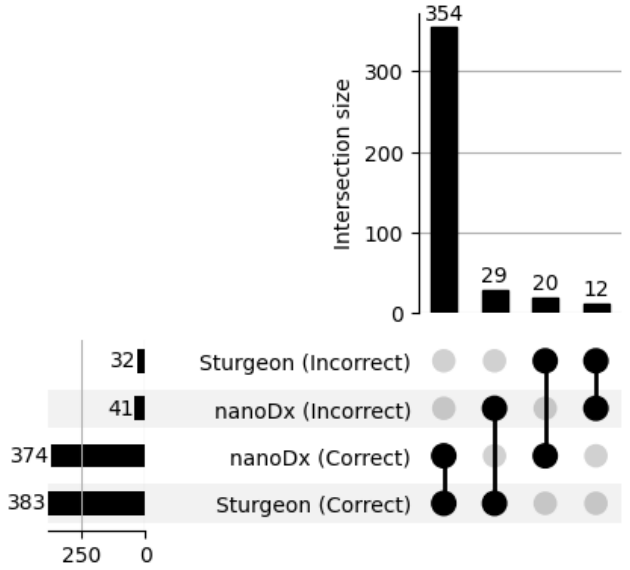
Neuropathology and Applied Neurobiology
JOURNAL OF THE BRITISH NEUROPATHOLOGICAL SOCIETY
WILEY

Robust methylation-based classification of brain tumours using nanopore sequencing

Luis P. Kuschel[1] | Jürgen Hench[2] | Stephan Frank[2] | Ivana Bratic Hench[2] | Elodie Girard[3] | Maud Blanluet[3] | Julien Masliah-Planchon[3] | Martin Misch[4] | Julia Onken[4] | Marcus Czabanka[4] | Dongsheng Yuan[1,5] | Sören Lukassen[5] | Philipp Karau[5] | Naveed Ishaque[5] | Elisabeth G. Hain[6] | Frank Heppner[6] | Ahmed Idbaih[7] | Nikolaus Behr[1] | Christoph Harms[1,8] | David Capper[6,9] | Philipp Euskirchen[1,9]

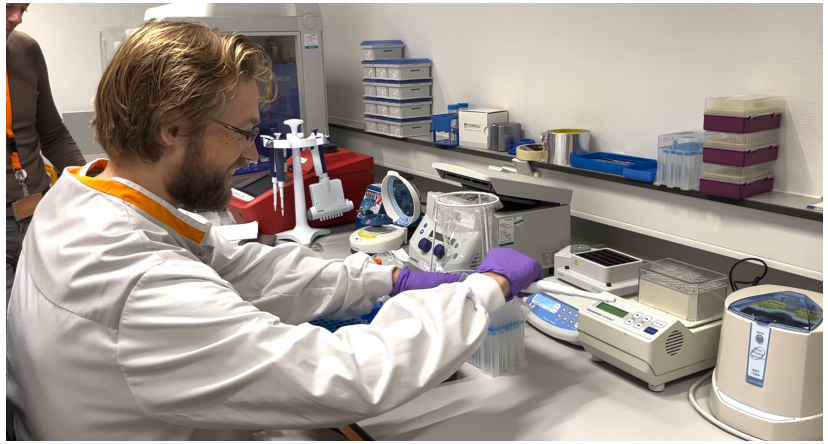
Oslo Nanopore



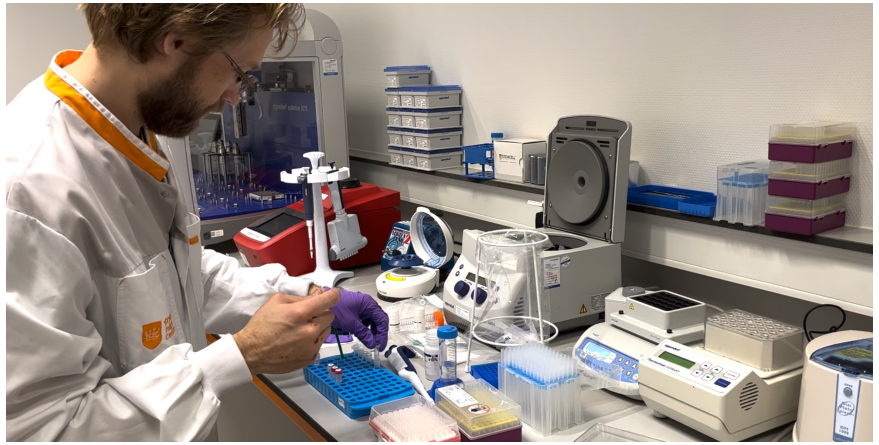


Both correct
Both incorrect
nanoDx correct
Sturgeon correct

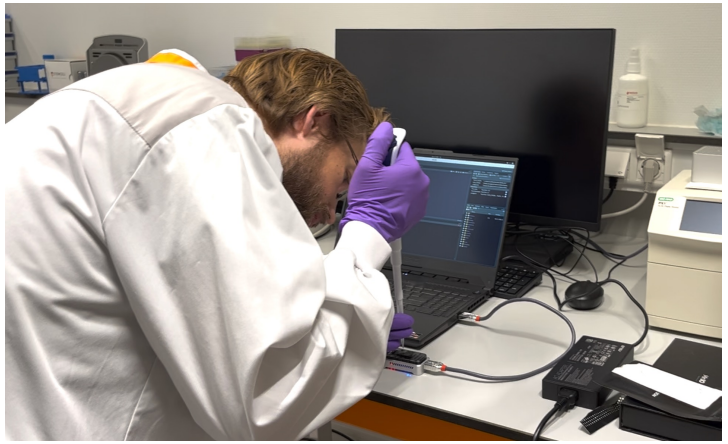
Real live run



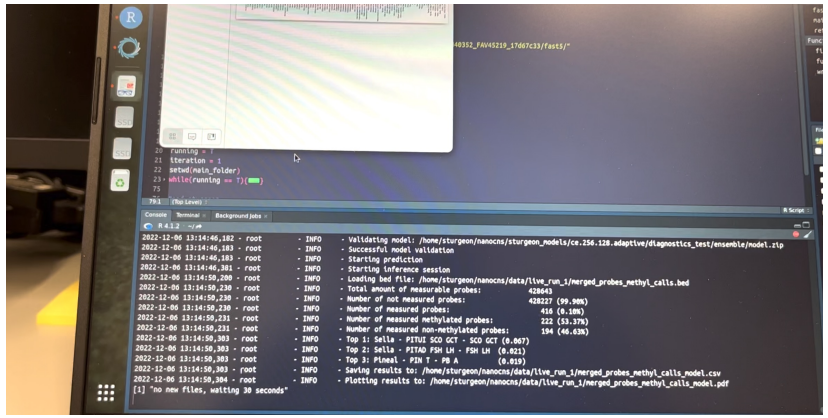
Start
00:00



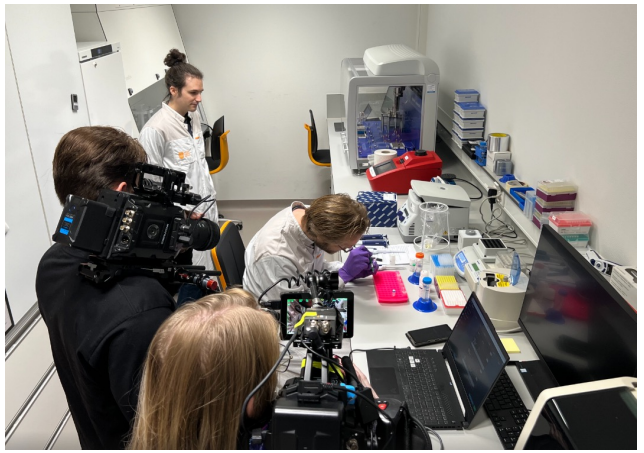
Finish DNA extraction
17:45



Start sequencing
26:00



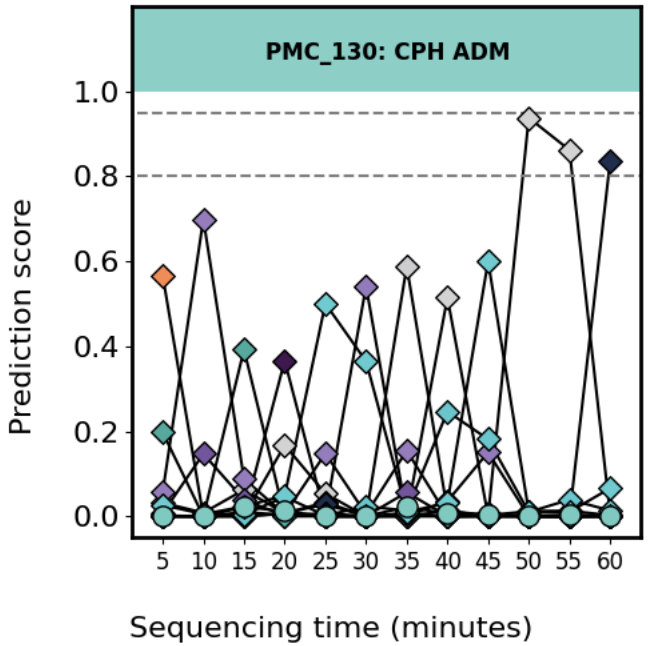
First prediction result
40:25



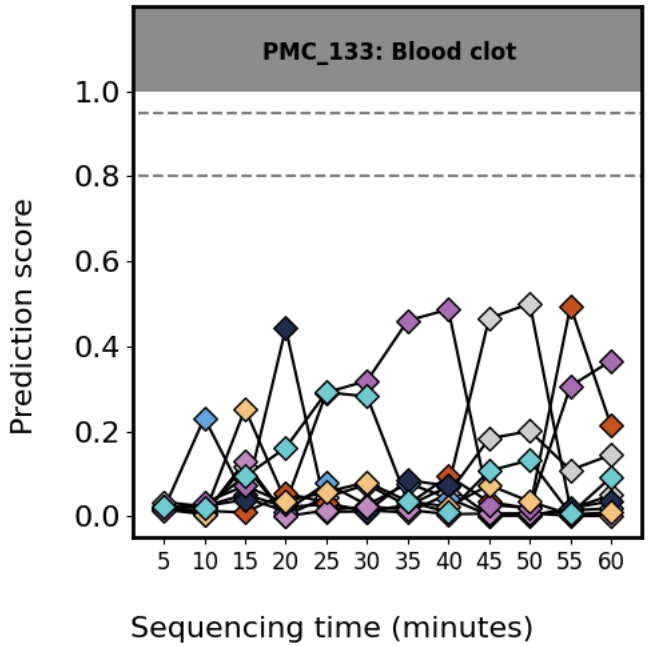
Doable in one hour

Real live run

5 % sample purity

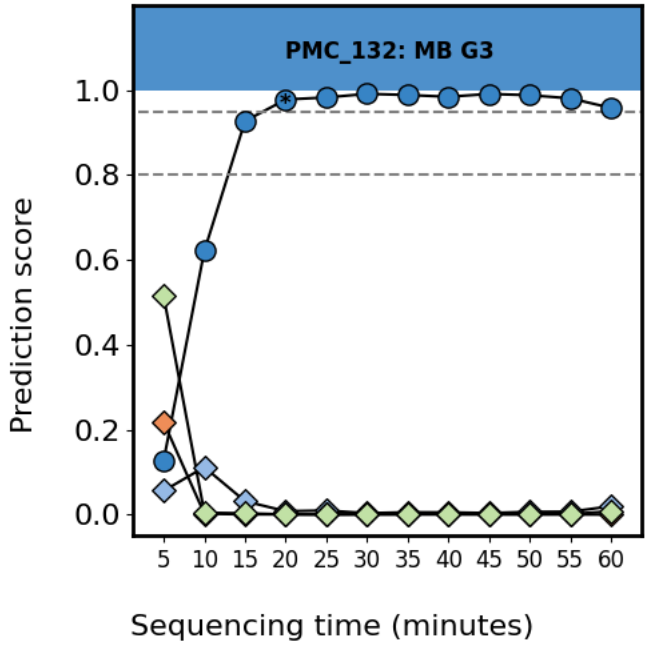


Blood clot

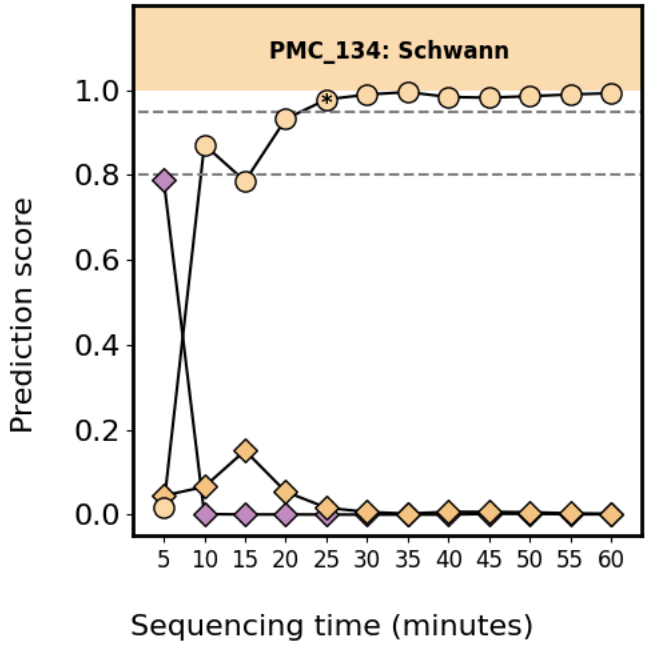


Real live run

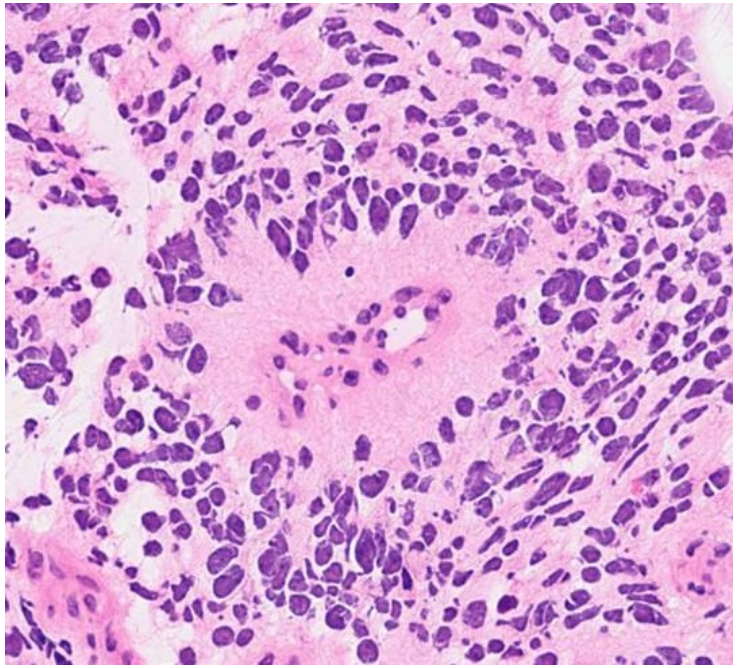
Medulloblastoma G3

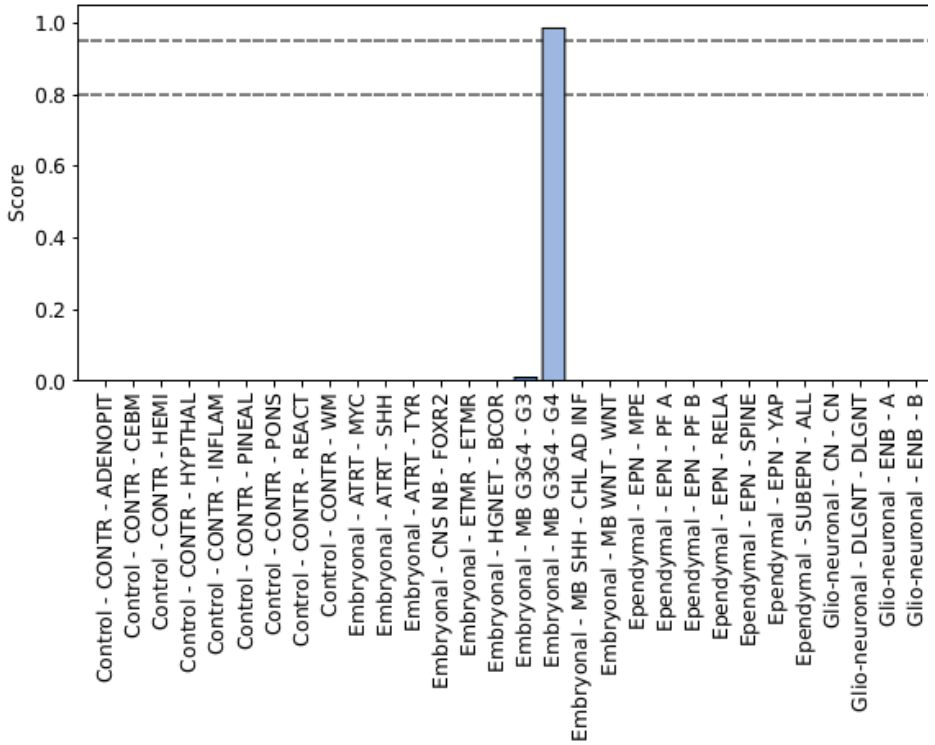


Schwanomma



Real live run





Histology: Ependymal Rosettes

Sturgeon after 10 minutes of sequencing: **Medulloblastoma G4**

Ependymoma: radical resection
One week later: methylation classifier says it's a Medulloblastoma G4

Conclusions

- We can train a neural network that can handle an arbitrary amount of random missing values and accurately predict brain cancer subtypes

- Data simulation allows us to
  - Massively increase sample size
  - Upsample to balance and also to improve on worse performing cases
  - Translate from microarray to Nanopore sequencing data
  - “Infinitely validate”

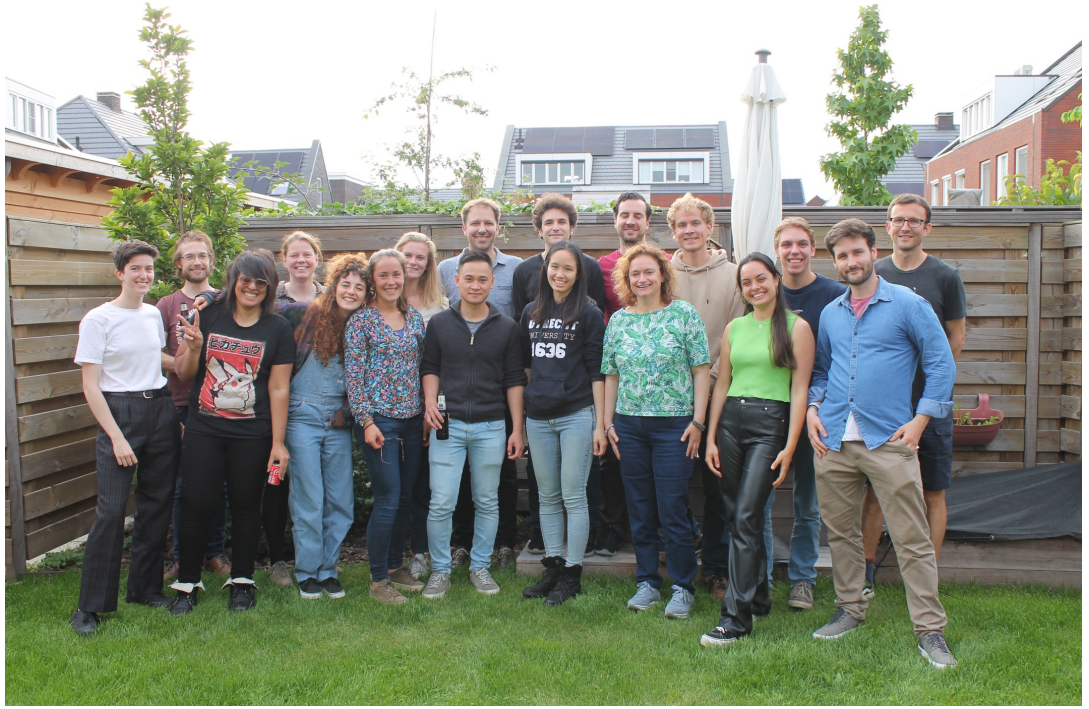
- All together is fast enough to be done and influence surgical strategy

Acknowledgements

de Ridder group
Jeroen de Ridder
Carlo Vermeulen
Adrien Melquiond
Dieter Stoker
Emmy Wesdorp
Inez den Hond
Joanna Wolthuis
Li-Ting Chen
Luca Santuari
Lucía Barbadilla
Nicolle Besselink
Roy Straver

Princess Maxima Center
Bastiaan Tops
Lennart Kester
Eric Strengman
Peter Wesseling
Eelco Hoving
Mariëtte Kranendonk
Kirsten van Baarsen
Jasper van der Lugt

Universitätsklinikum Münster
Christian Thomas
Anne Albers



UMC Utrecht
prinses máxima centrum
Oncode Institute

Preprint available

de Ridder group
Jeroen de Ridder
Carlo Vermeulen
Adrien Melquiond
Dieter Stoker
Emmy Wesdorp
Inez den Hond
Joanna Wolthuis
Li-Ting Chen
Luca Santuari
Lucía Barbadilla
Nicolle Besselink
Roy Straver

Princess Maxima Center
Bastiaan Tops
Lennart Kester
Eric Strengman
Peter Wesseling
Eelco Hoving
Mariëtte Kranendonk
Kirsten van Baarsen
Jasper van der Lugt

Universitätsklinikum Münster
Christian Thomas
Anne Albers

UMC Utrecht
prinses máxima centrum
Oncode Institute