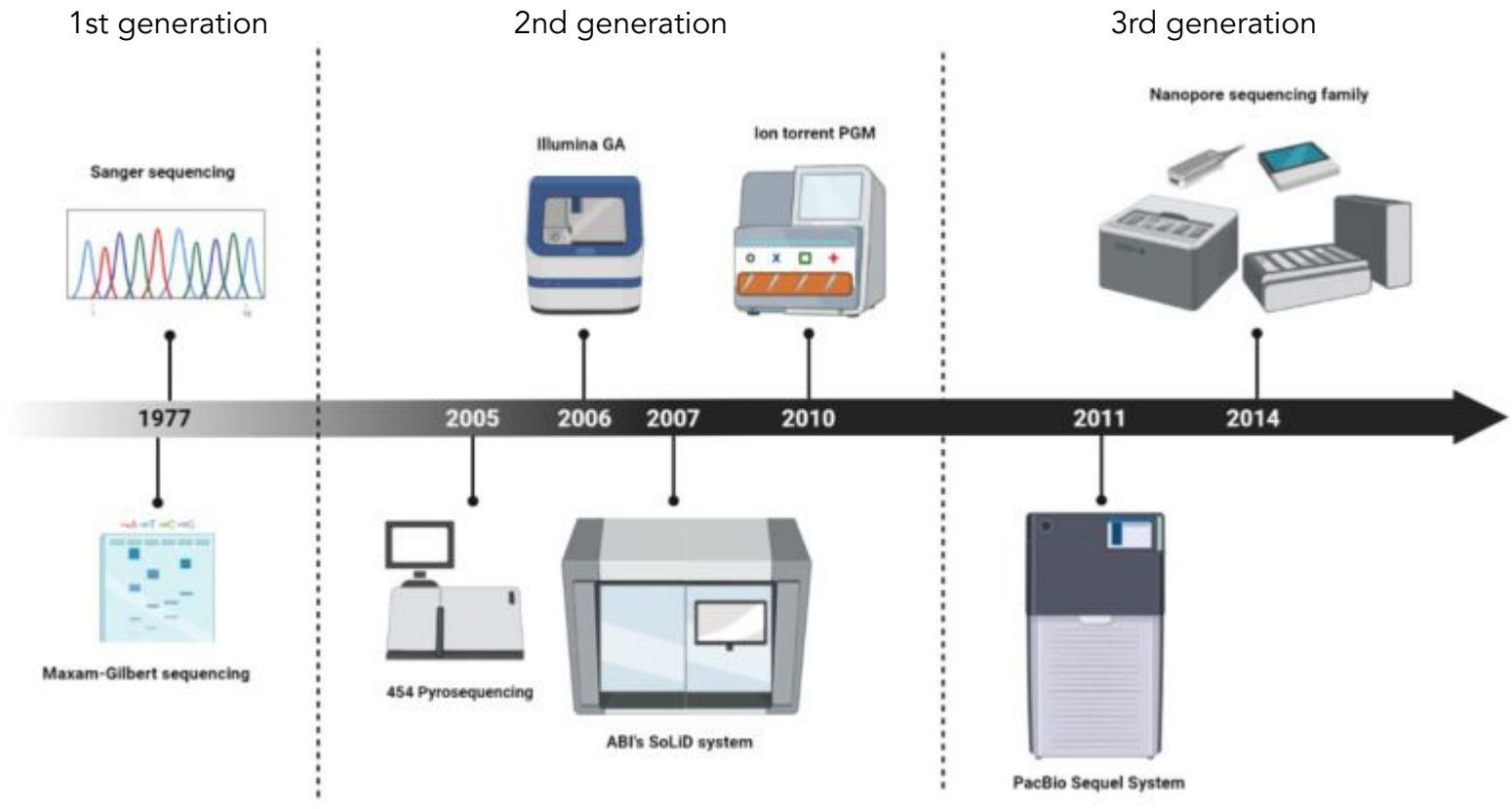
Single-molecule sequencing

Quantitative Biology 2022
10/28/22

Doctoral Program in
Cell, Molecular, Developmental Biology, and Biophysics

JOHNS HOPKINS UNIVERSITY

The history of sequencing



Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life (Basel). 2021 Dec 26;12(1):30.

Single-molecule sequencing (SMS)

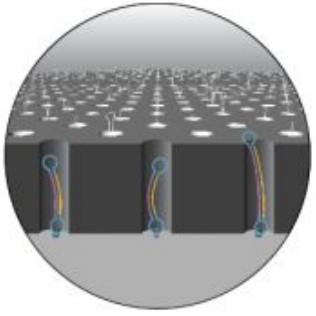
Advantages

- No PCR
- Small input sample size
- Real-time sequencing
- More uniform genome coverage
- Longer read lengths (Kb to Mb)
- Faster data production

Disadvantages

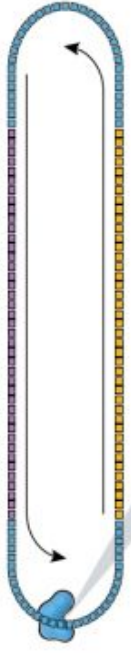
- Higher error rates
- Fresh samples and careful handling needed to preserve ultralong reads
- Lack of database/analysis tools

PacBio - sequencing



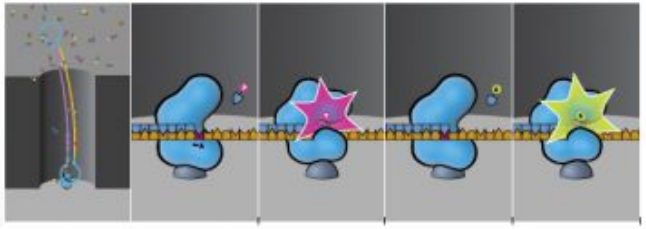
SMRT Cells contain millions of zero-mode waveguides (ZMWs)

SMRTbell® templates enable repeated sequencing of circular template with real-time detection of base incorporation

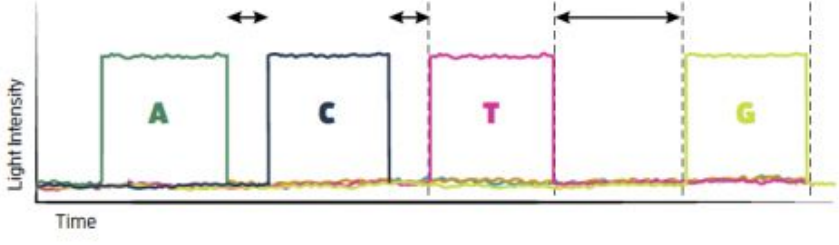


A single molecule of DNA is immobilized in each ZMW

+ Phospholinked nucleotides



As anchored polymerases incorporate labeled bases, light is emitted



Directly detect DNA modifications during sequencing

Nucleotide incorporation kinetics are measured in real time

https://wp.unil.ch/gtf/technology/

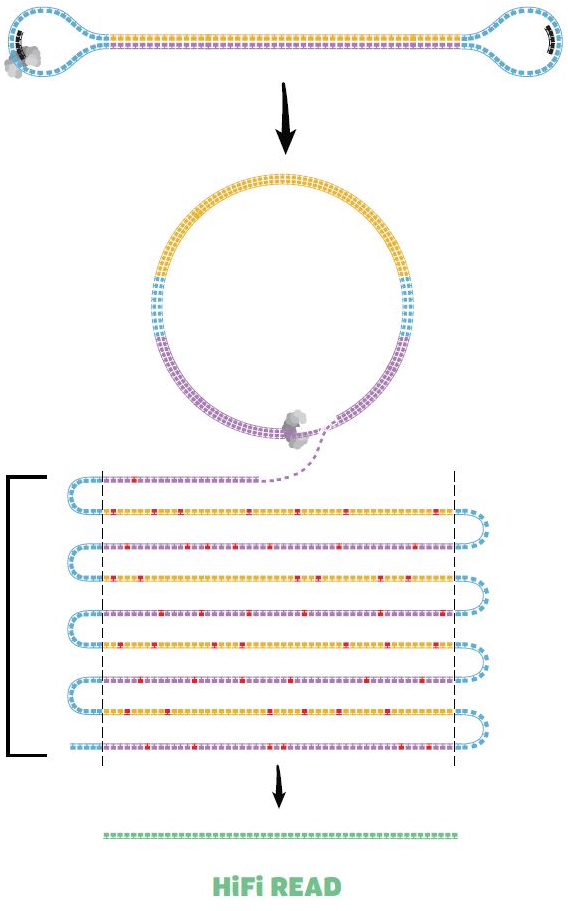
PacBio - basecalling

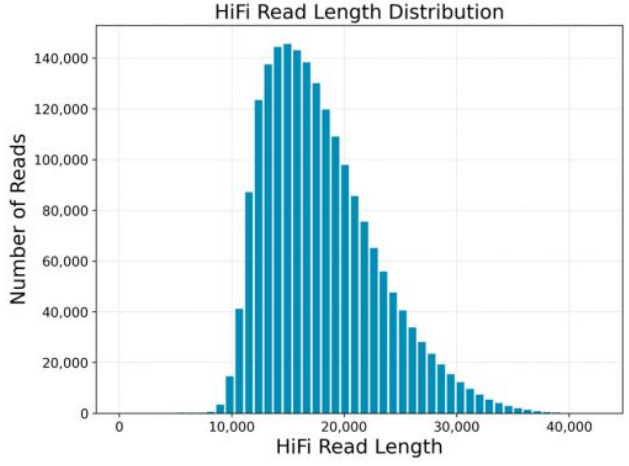
Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus is called from subreads





https://wp.unil.ch/gtf/technology/

PacBio

Advantages

- No PCR
- Small input sample size
- Real-time sequencing
- More uniform genome coverage
- Longer read lengths (up to 300Kb, ~15Kb average)
- Faster data production

Disadvantages

- Higher error rate (~90% accurate, HiFi ~99.9% accurate)
- Requires fresh samples
- Immature database/analysis tools

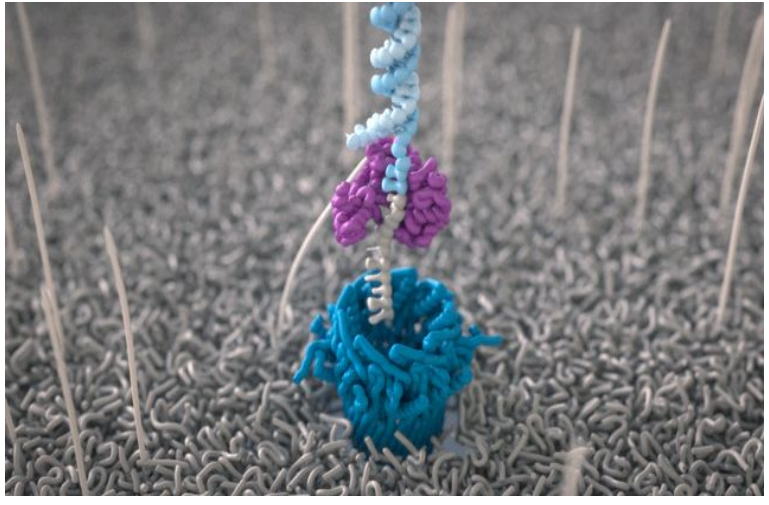
PacBio

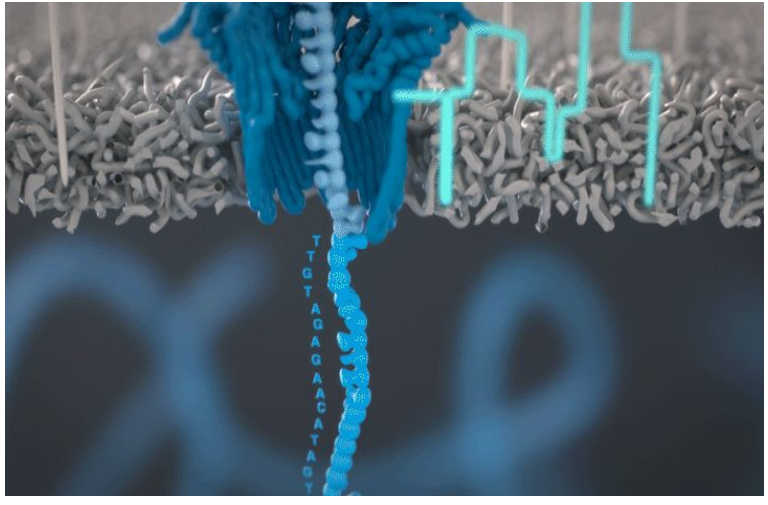
A single run on a PacBio Sequel IIe (~$500K/machine)

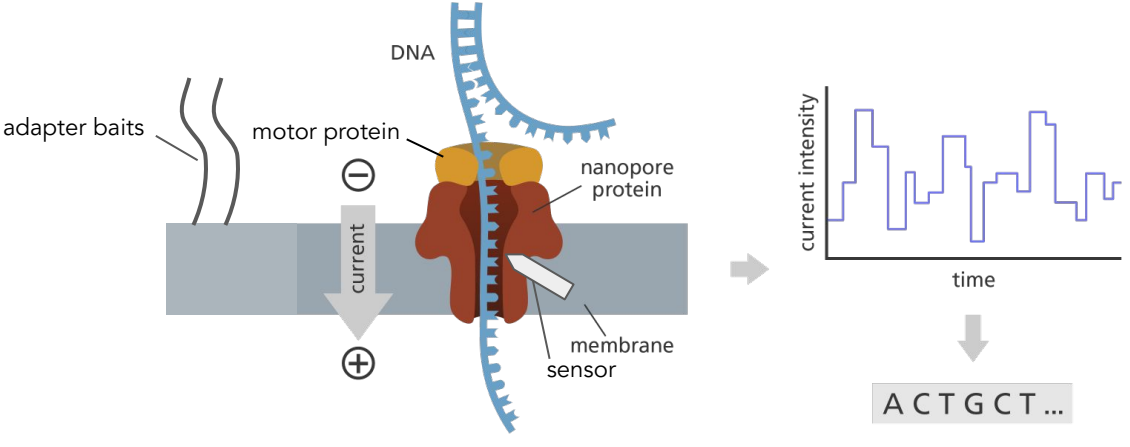
- Yields ~4 million reads
- Takes about 30 hours
- Sequences fragments with an average size of 15Kb
- Generates ~60Gbp of sequence
- Costs ~$2-3K

For comparison, an Illumina run for the same amount of data would cost ~$1-1.5K

Oxford Nanopore Technologies (ONT)







https://nanoporetech.com/how-it-works

https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/

Nanopore sequencing

Advantages

- No PCR
- Portable
- Low equipment cost
- Can directly sequence RNA
- Small input sample size (~order magnitude less than PacBio)
- Real-time sequencing
- More uniform genome coverage
- Longer read lengths (up to 4Mb, ~50-100Kb average)
- Faster data production
- Reusable flowcells

Disadvantages

- Higher error rate (90-95% accurate)
- Requires fresh samples
- Immature database/analysis tools

Nanopore Sequencing

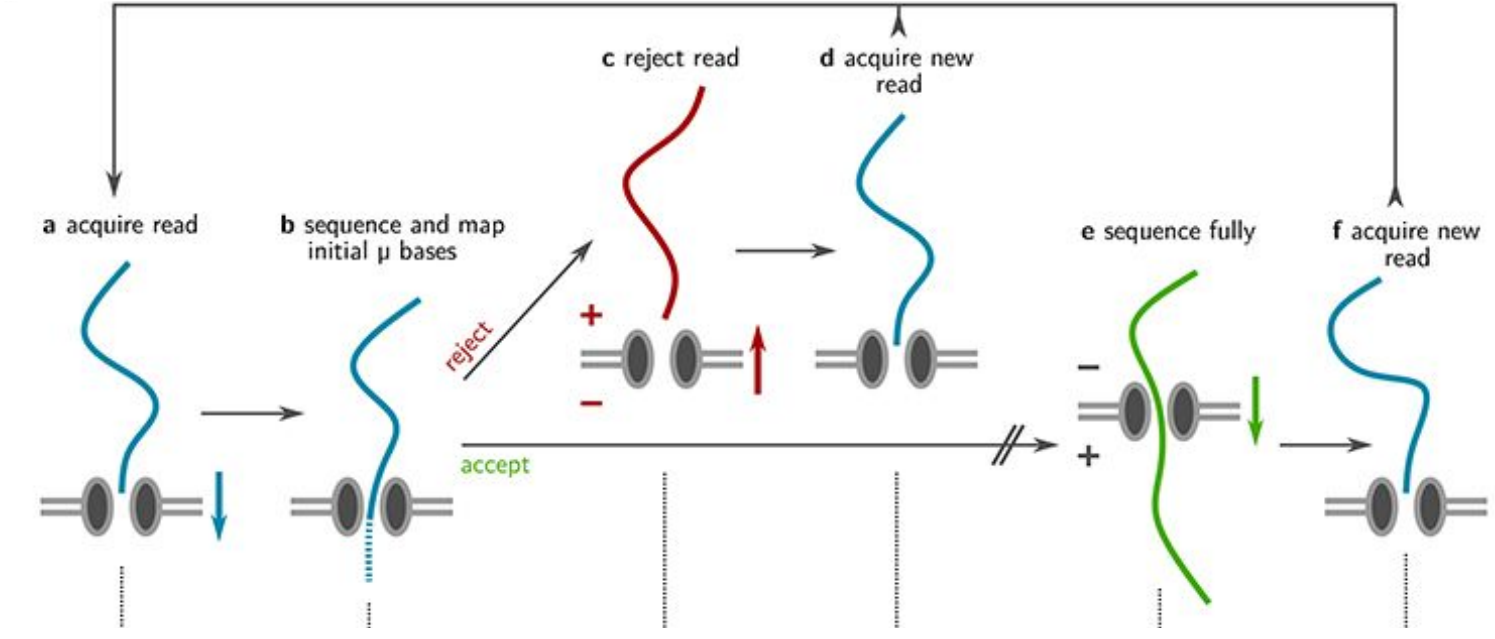
A single run on a MinIon (~$1K/machine)

- Yields ~200 thousand reads
- Takes about 24-72 hours
- Sequences fragments with an average size of 50-100Kb
- Generates ~10-20Gbp of sequence
- Costs ~$500

For comparison, an Illumina run for the same amount of data would cost ~$300

ONT advances

“Read until” Adaptive Read Selection

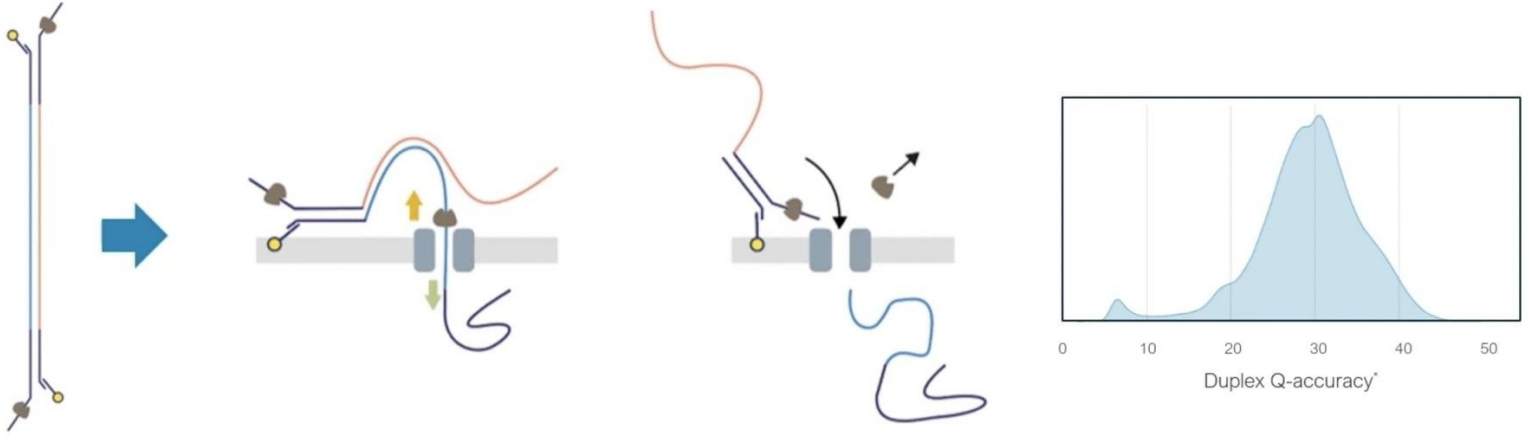


Sequence-based read acceptance/rejection allows enrichment of target templates at the time of sequencing

Weilguny L, De Maio N, Munro R, et al. Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design. bioRxiv; 2020.

ONT advances

Duplex Sequencing

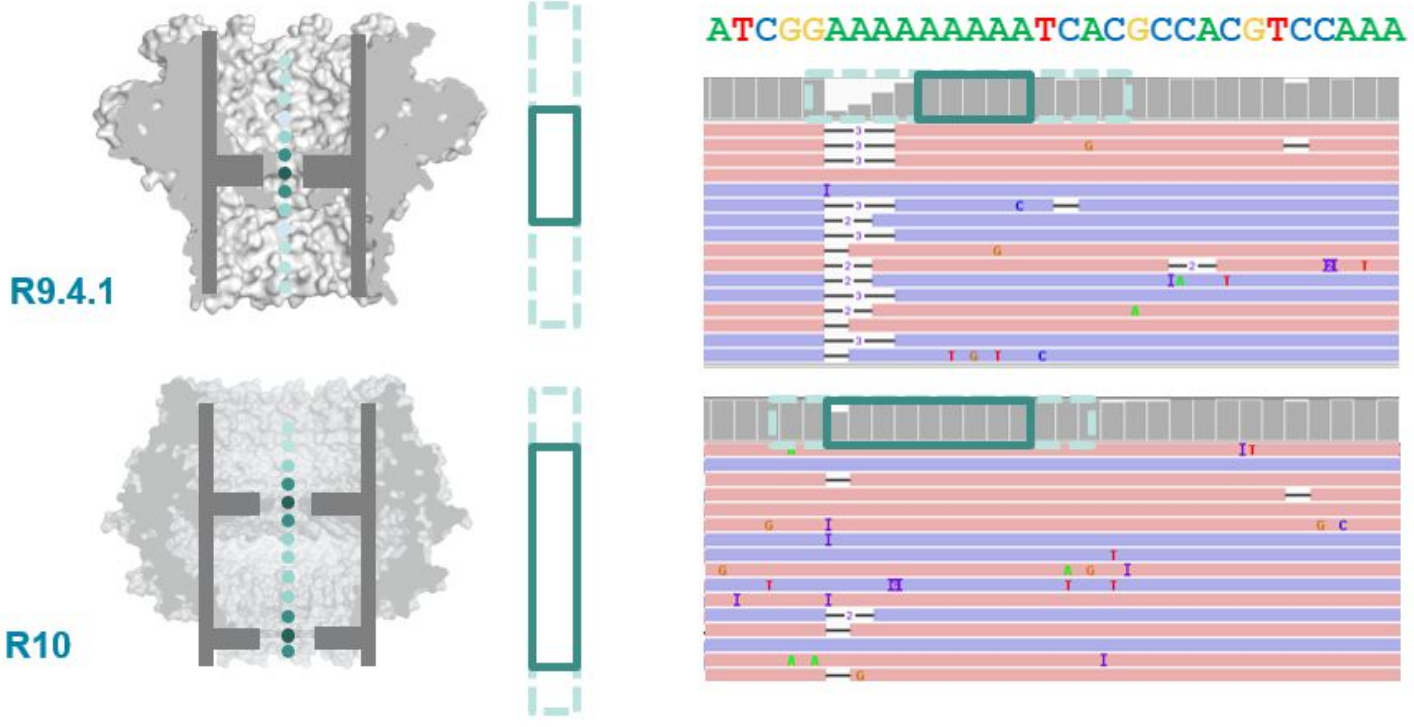


Forward-reverse sequencing allows joint basecalling, vastly improving read accuracy (~99.9% accurate on average)

https://nanoporetech.com/resource-centre/clive-brown-ncm-update-2021

ONT advances

Dual Pore Sensors

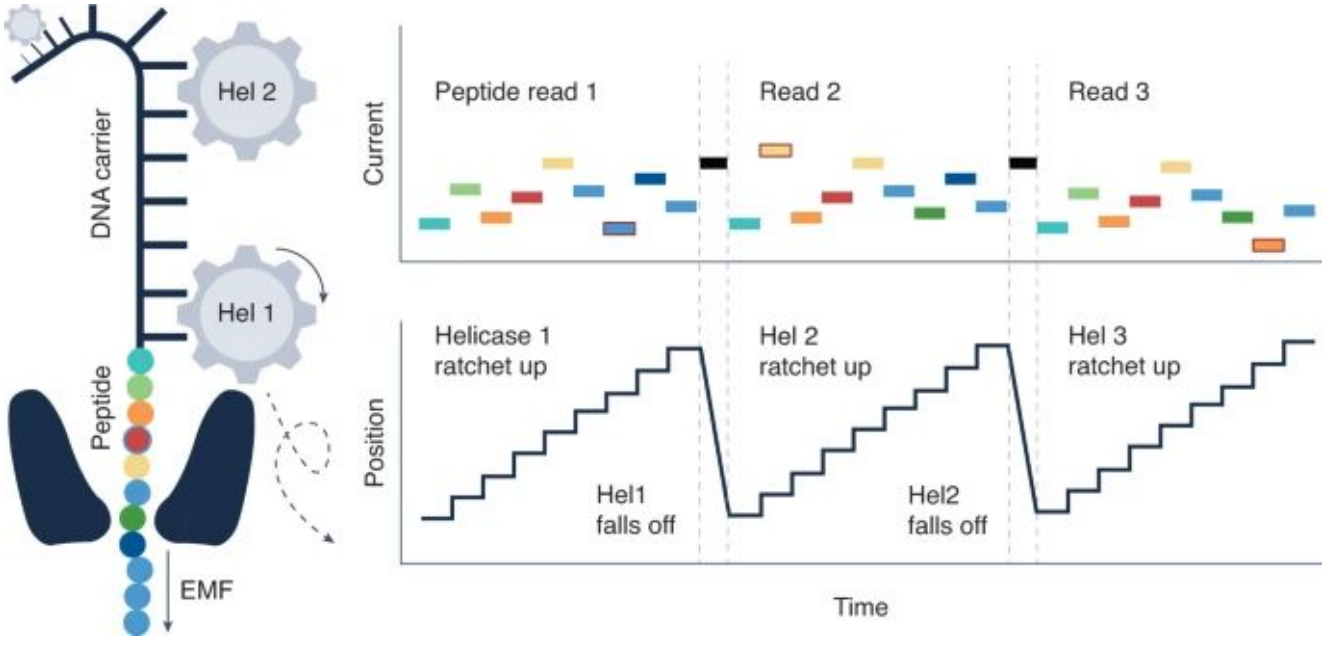


Multiple offset sensors improve resolution of homopolymers

https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store

ONT advances

Direct peptide sequencing

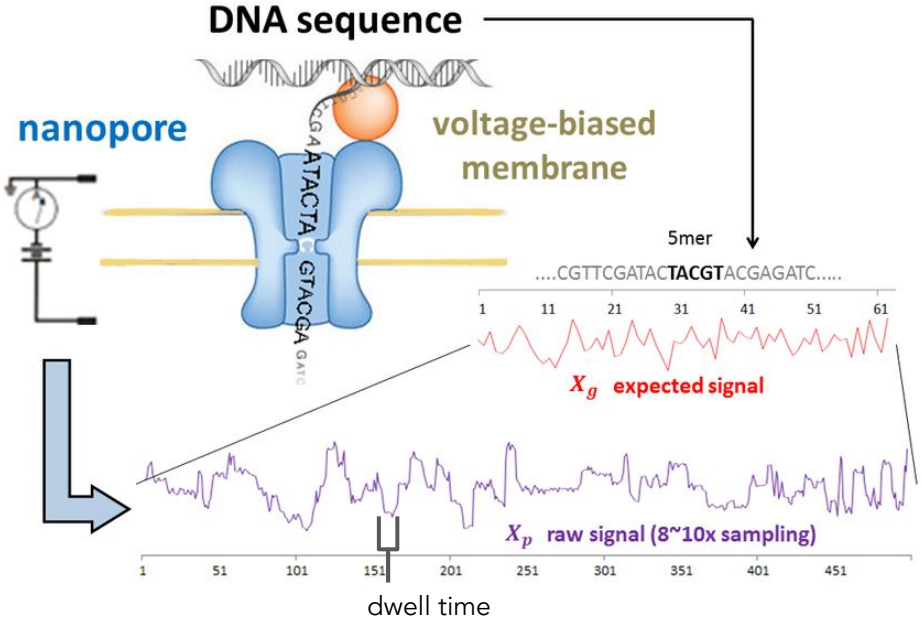


Helicases ratchet peptides peptides through pores repeatedly for a high quality consensus protein sequence

Wanunu, M. Back and forth with nanopore peptide sequencing. *Nat Biotechnol* 40, 172–173 (2022).

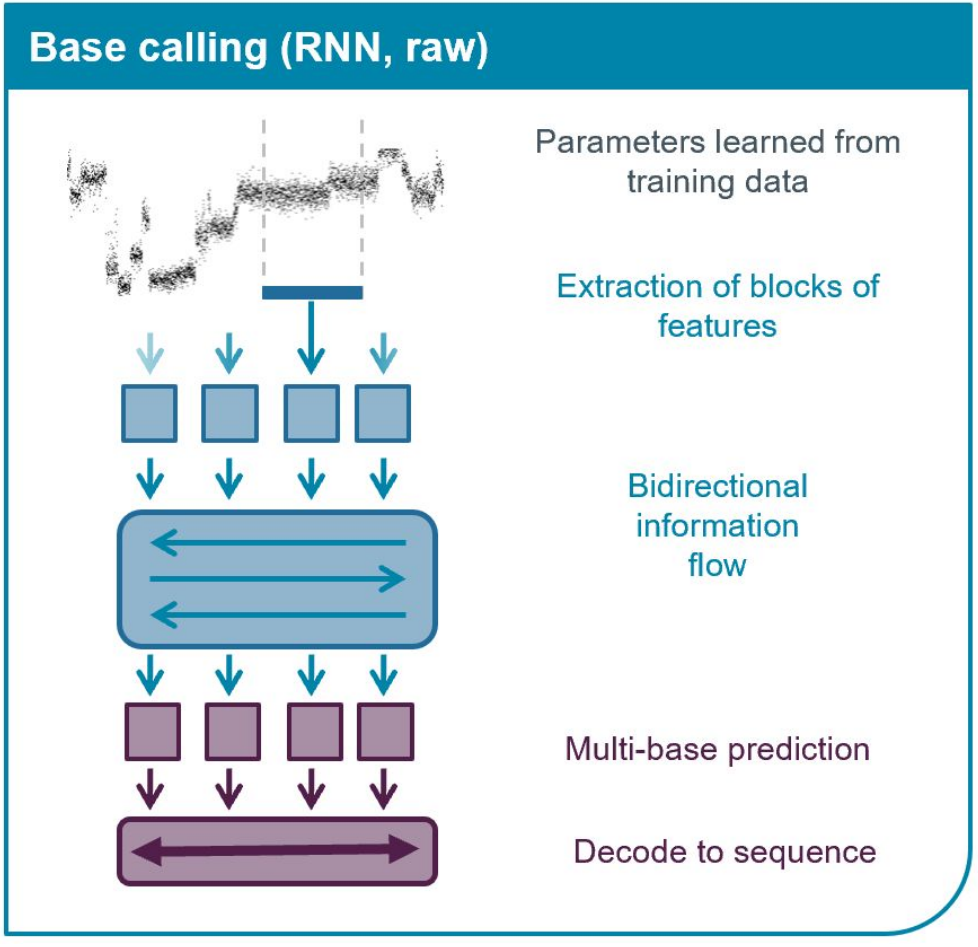
From voltage to base calls

- Voltage is determined by 5 bases currently in pore
- Each 5-mer yields (semi-)unique voltage
- Dwell time in pore is variable



Han, Renmin & Li, Yu & Wang, Sheng & Gao, Xin. (2017). An accurate and rapid continuous wavelet dynamic time warping algorithm for unbalanced global mapping in nanopore sequencing.

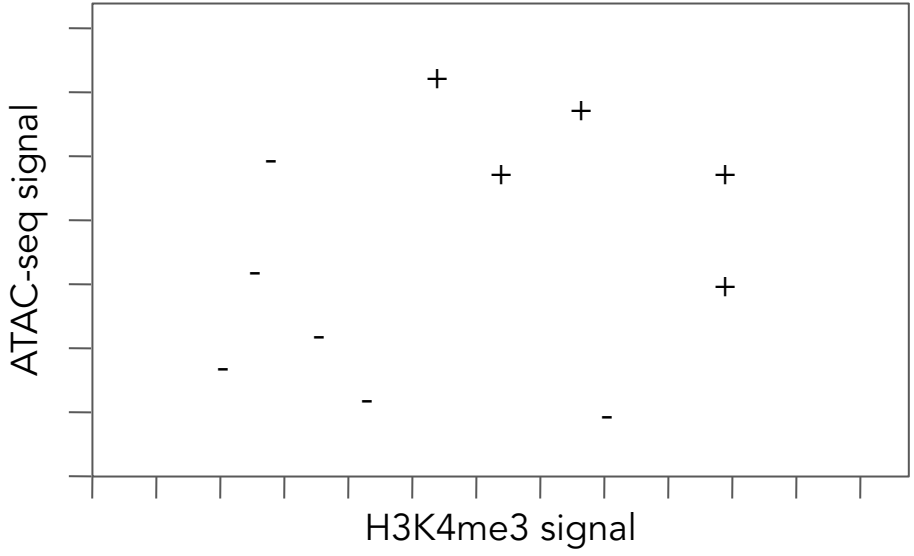
Guppy - a recurrent neural network basecaller



https://nanoporetech.com/how-it-works/basecalling

What is a neural network

Let's take a step back and look at the simple case of 2 category classification

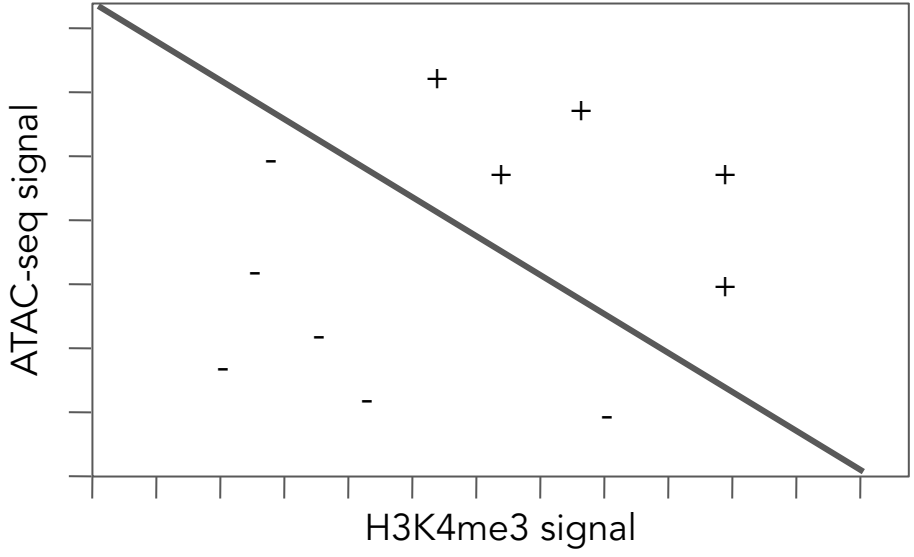


We want to determine if a likely active using H3K4me3 and ATAC-seq signal

We start with known examples

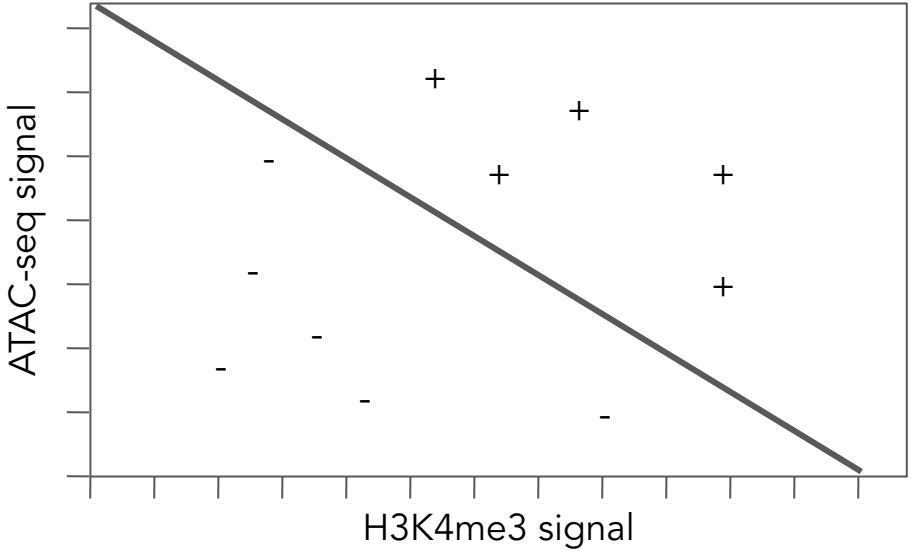
What is a neural network

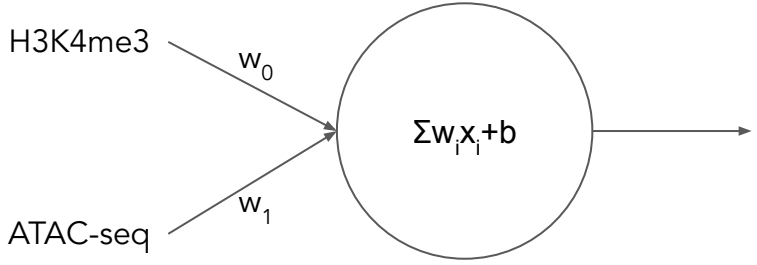
This is easy to classify with a simple linear relationship



What is a neural network

This is easy to classify with a simple linear relationship

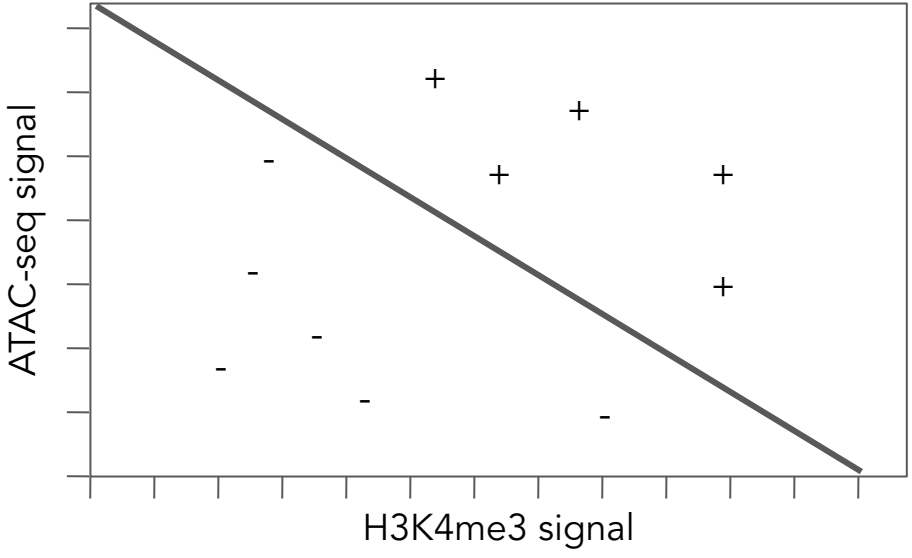


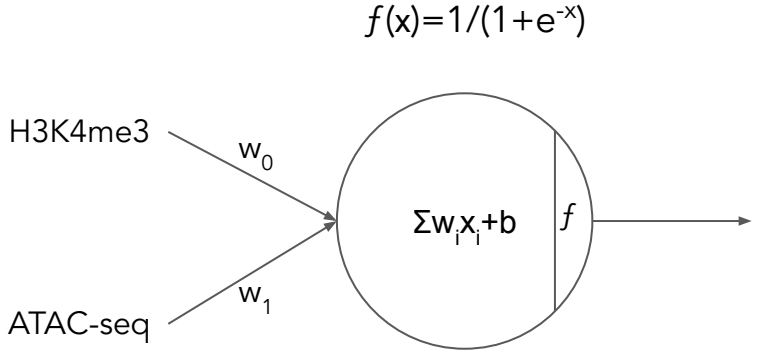


We can create a simple linear combination (regression) of the inputs to define the line

What is a neural network

This is easy to classify with a simple linear relationship



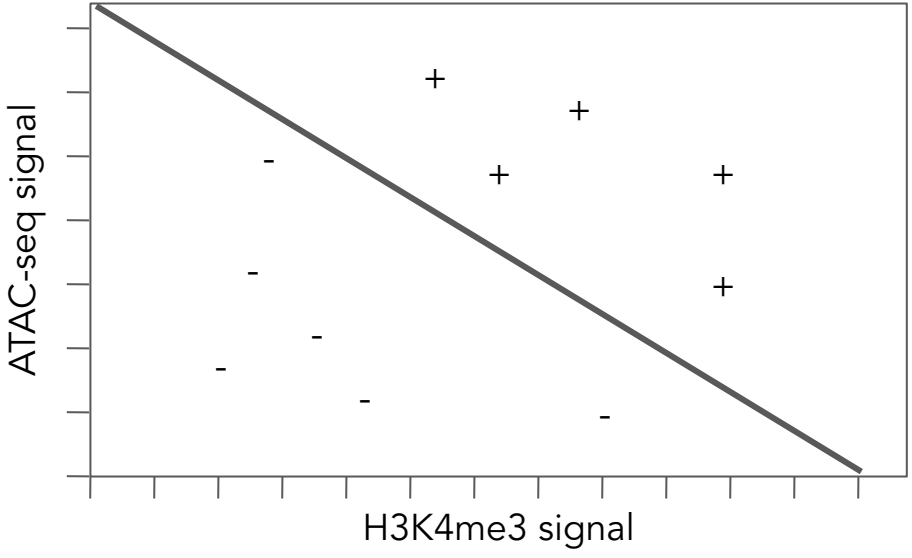


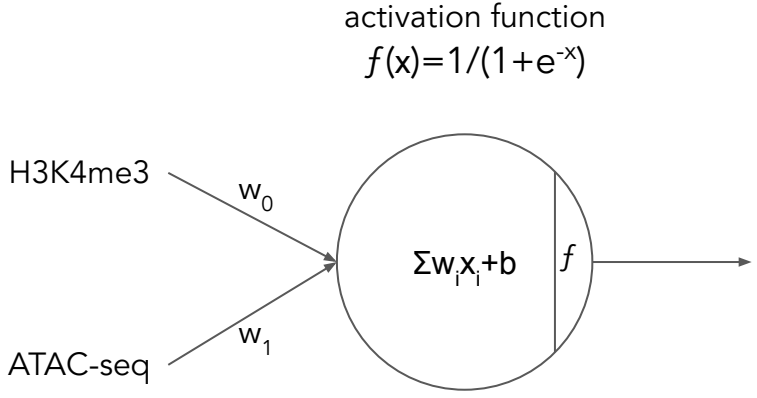
We can create a simple linear combination (regression) of the inputs to define the line

We can add a sigmoid function to change the output to 0-1. This turns it into a logistic regression.

The perceptron

This is easy to classify with a simple linear relationship

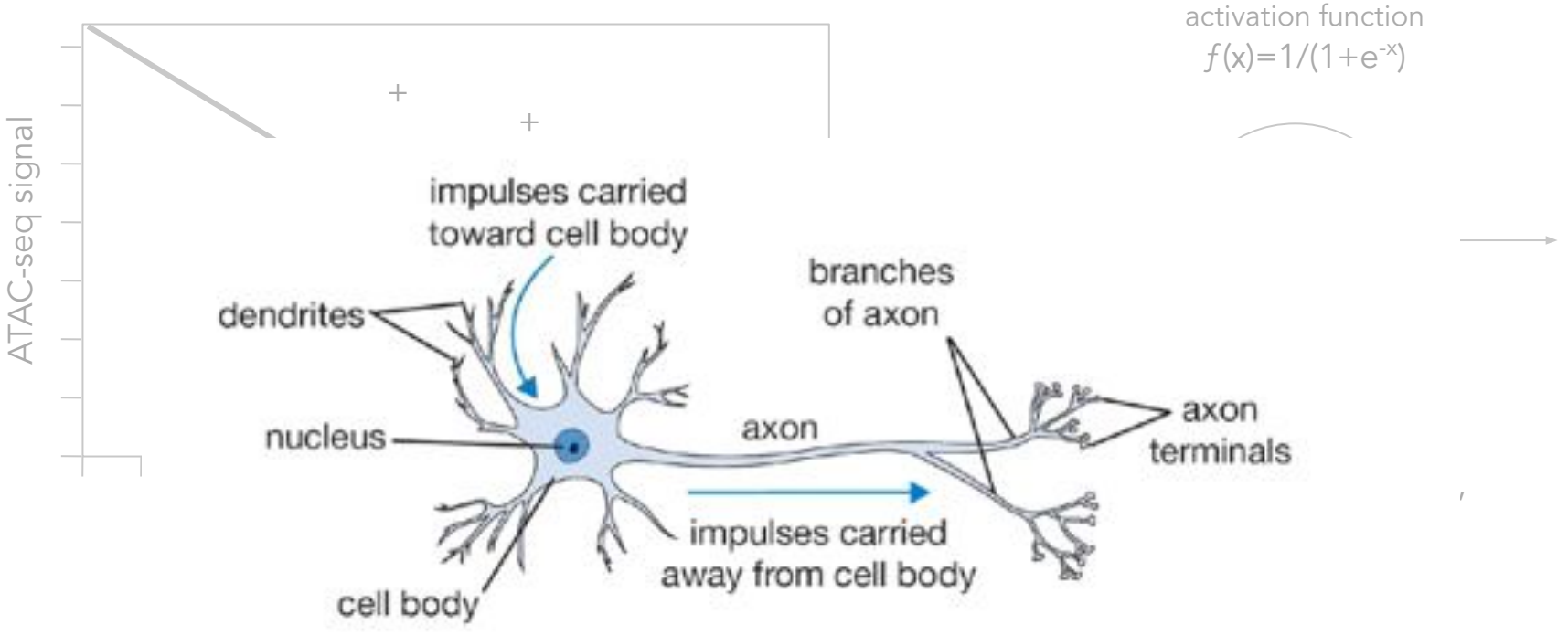




We have just defined a "perceptron"

The “neural” in neural network

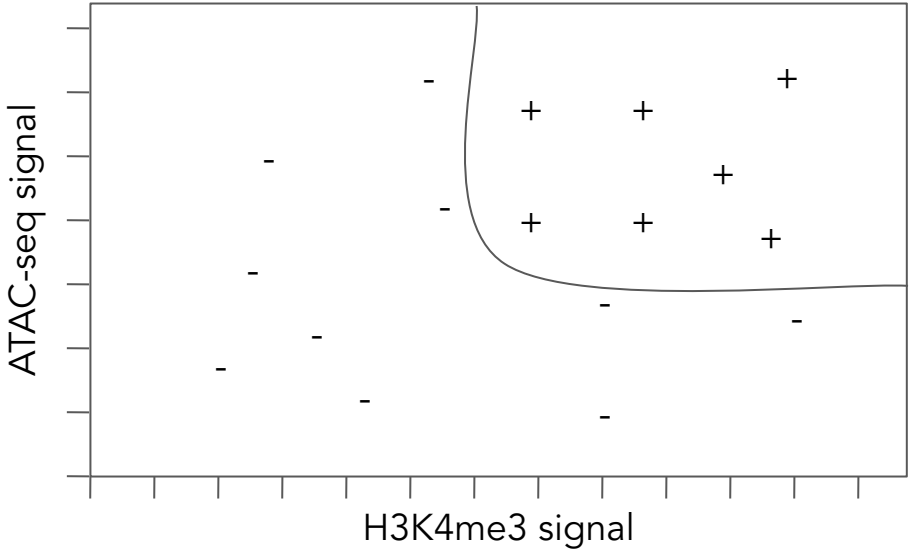
This is easy to classify with a simple linear relationship

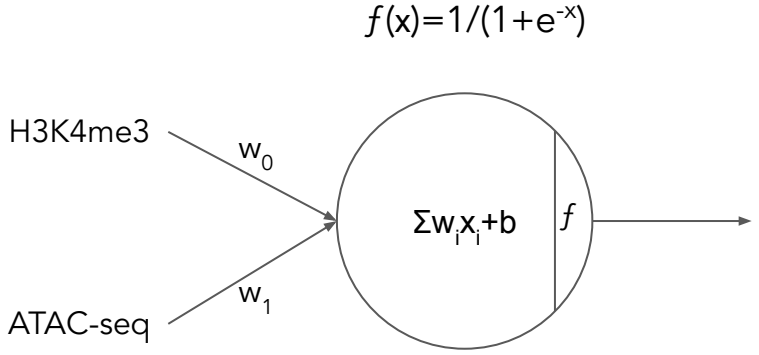


http://cs231n.github.io/neural-networks-1/

When a perceptron isn't enough

What if it's not a linear relationship?

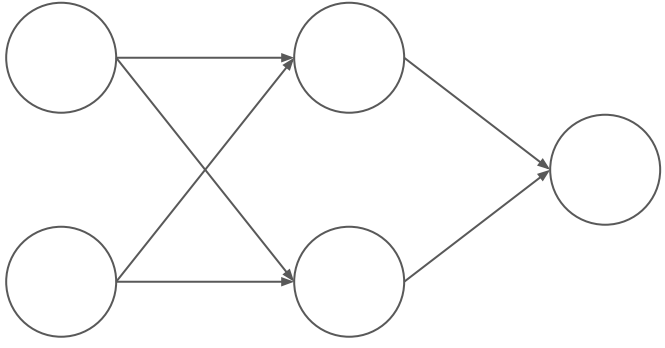




A simple neural network

By combining multiple perceptrons, we can create non-linear functions.

This is a 2-layer network (inputs are not counted)

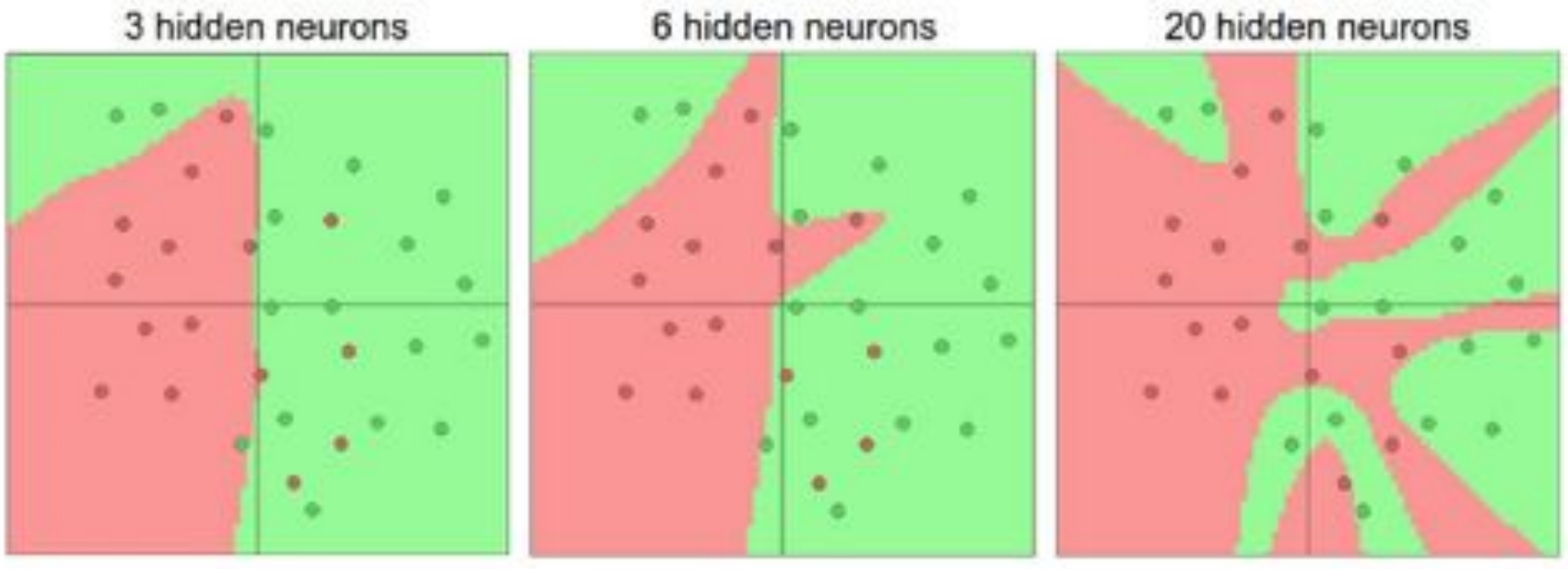


Inputs

Hidden Layer

Output

A universal approximator

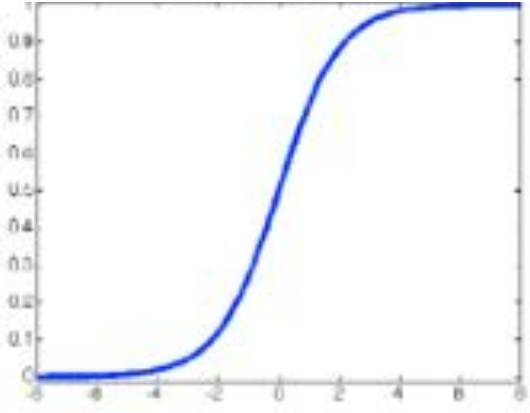


Given enough hidden nodes (neurons), any nonlinear function can be approximated

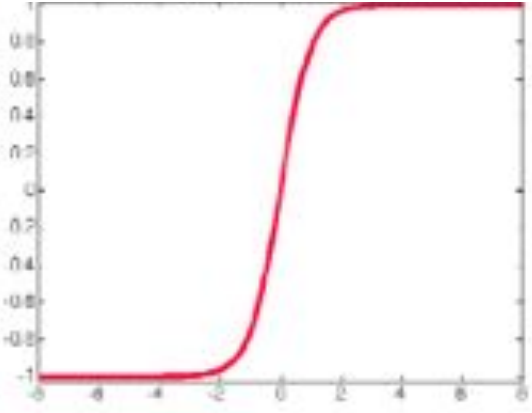
https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/10_nn1.pdf

Activation functions

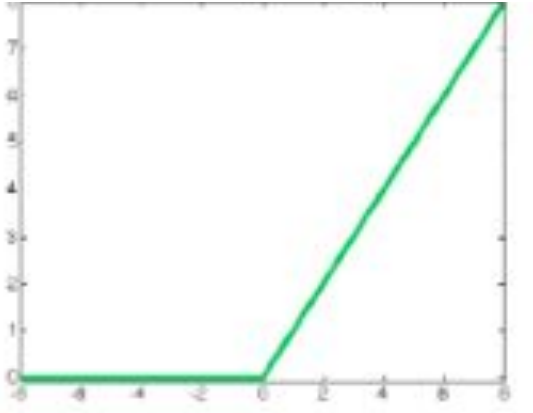
Sigmoid: $f(x)=1/(1+e^{-x})$



Tanh: $f(x)=(e^{x}-e^{-x})/(e^{x}+e^{-x})$



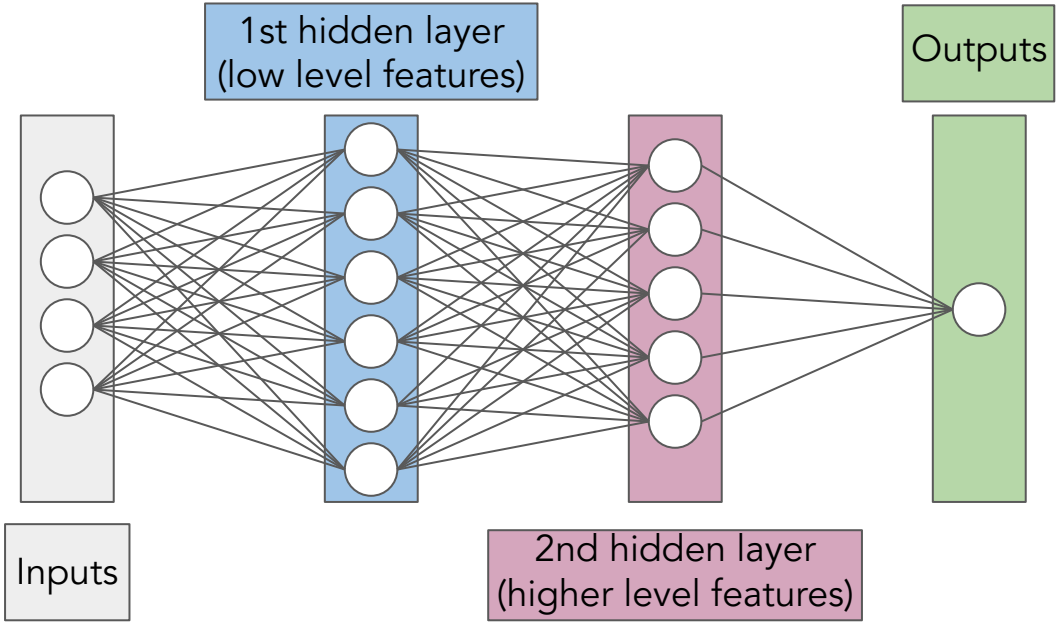
ReLU (Rectified Linear Unit):
$\mathrm{ReLU}(x)=\max(0, x)$



https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/10_nn1.pdf

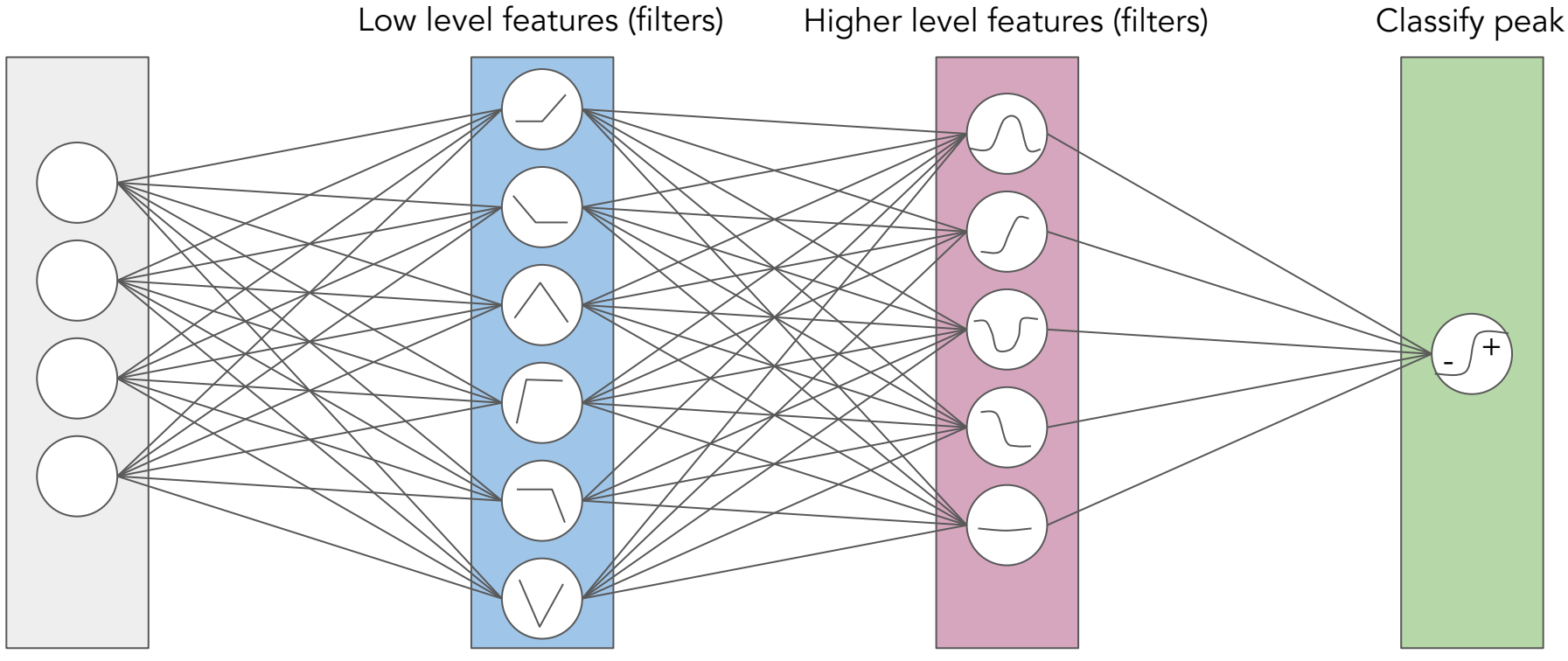
A “deep” neural network

- Each node in a hidden layer can be thought of as learning a specific feature
- Nodes closer to the inputs learn lower level features
- Nodes closer to the outputs learn more complex features



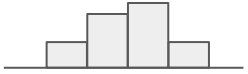
Features (filters)

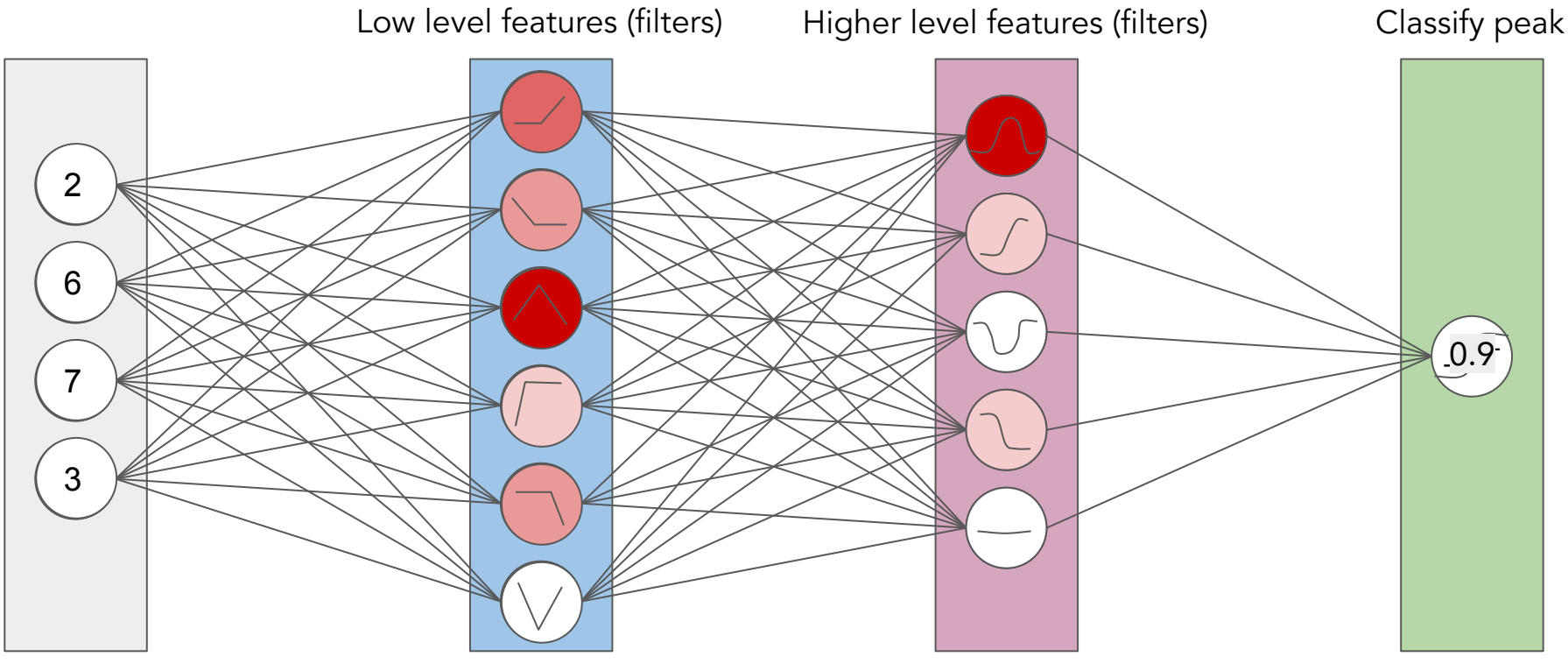
Let's say our inputs are signal intensity across a window of base positions and we want to classify peaks



Features (filters)

Given the input signal

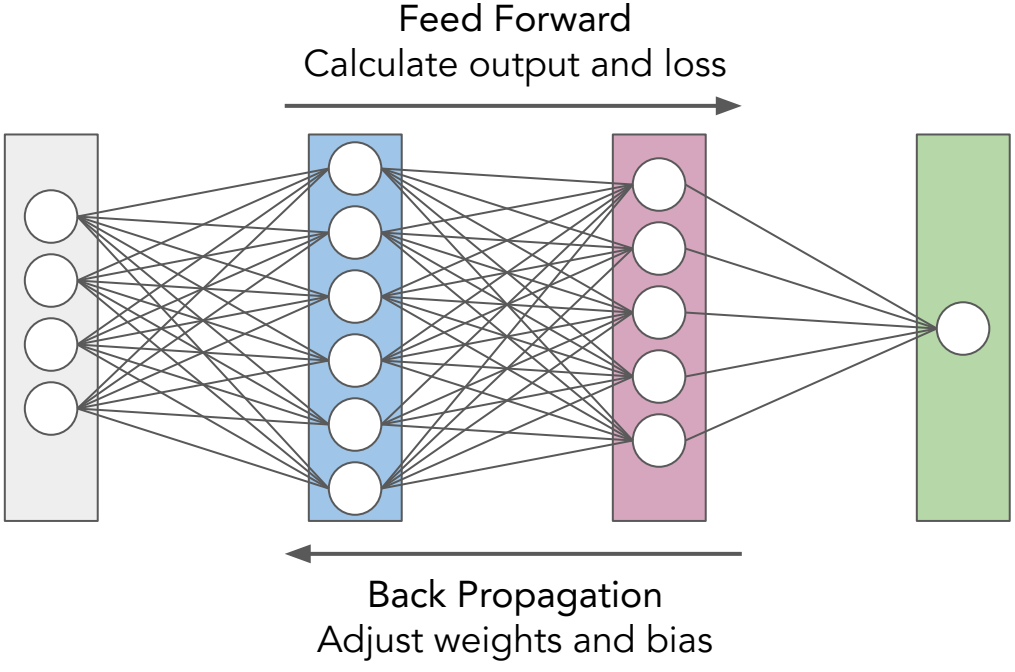




Training a neural network

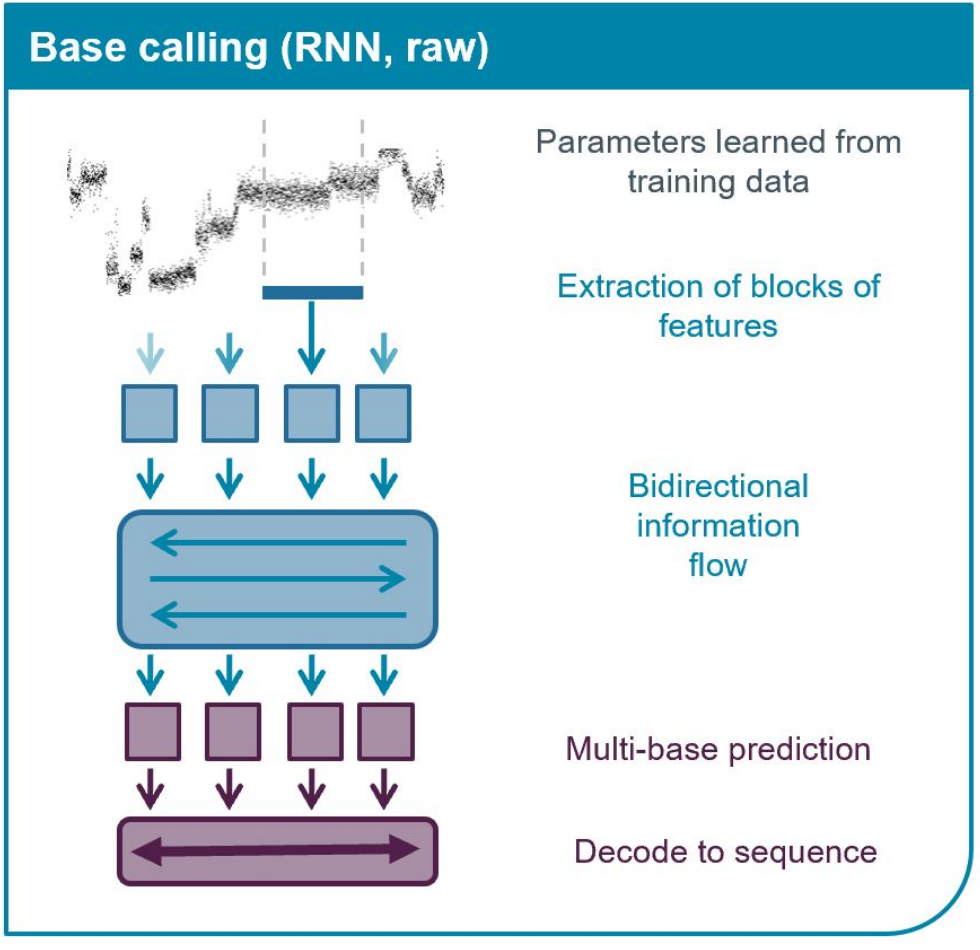
- Requires a large high-quality training set of ground-truth labels/values
- Need to define a loss function (how well is each prediction?)

Information only flows forward
(each node in a layer is independent of every other node in that layer)



Guppy - a recurrent neural network basecaller

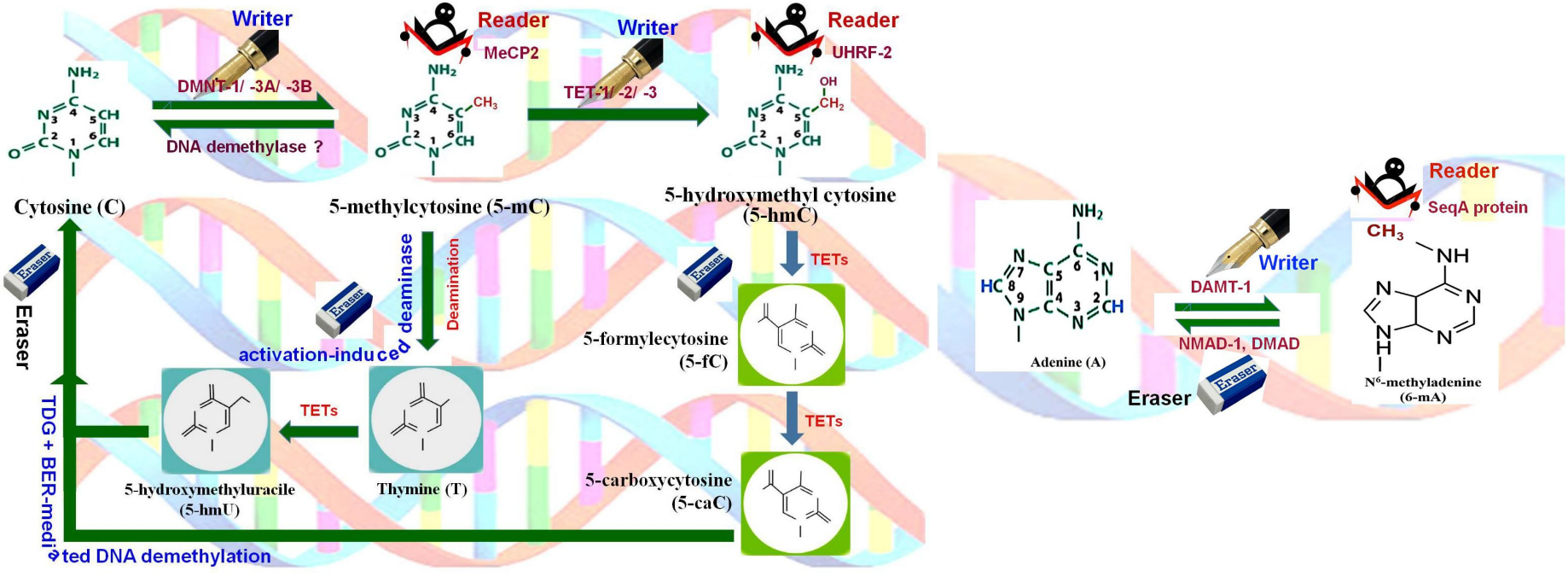
- Different low-level features are captured in the first layer across multiple windows
- Higher level features are compiled from low-level features
- Information is shared between high-level features
- Base predictions (probabilities) output for multiple positions
- All probabilities for a given base position contribute to determining base call



https://nanoporetech.com/how-it-works/basecalling

DNA/RNA base modification

Although DNA and RNA have a simple 4 letter alphabet, both have a number of possible chemical modifications that impact regulation and function



Kumar S, Chinnusamy V and Mohapatra T (2018) Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front. Genet.* 9:640.

5mC

- The most common DNA base modification is methylation of the fifth carbon of cytosine (5mC)

- ~1% of the human genome is composed of 5mC

- 5mC has multiple functions
  - Silencing retroviral elements
  - Tissue-specific gene regulation
  - X-inactivation
  - Genomic imprinting

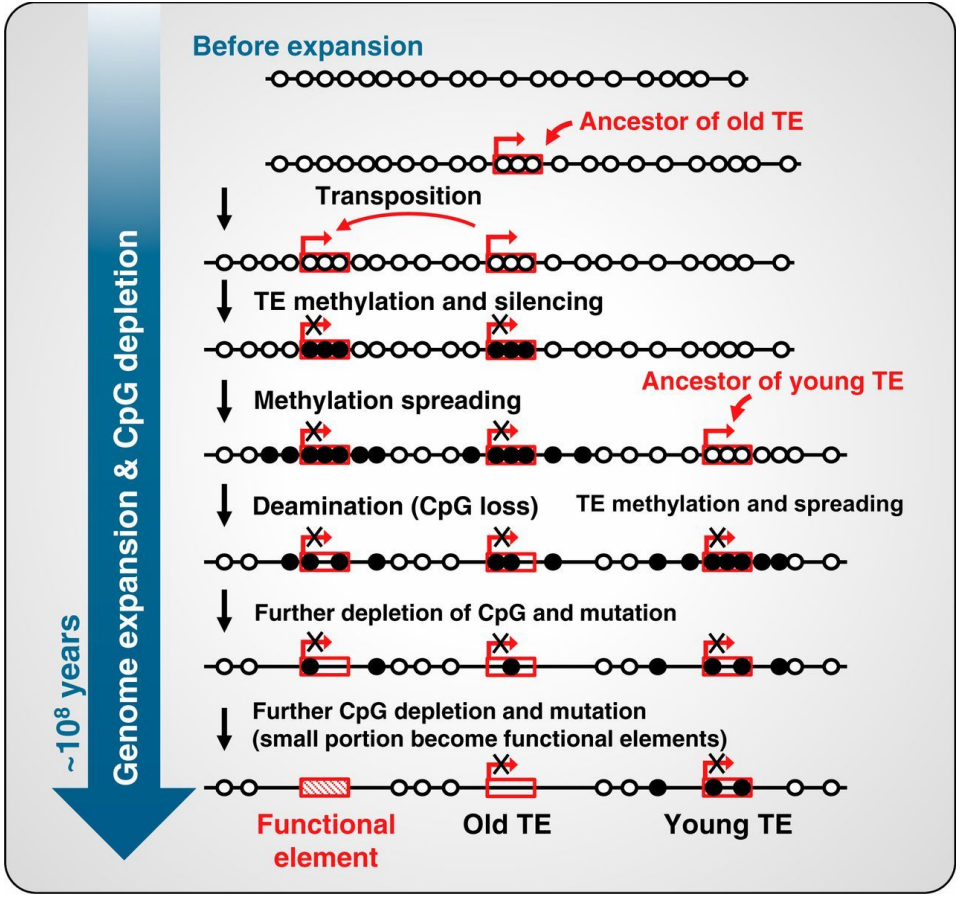
Suppression of Transposable Elements (TEs)

TEs are mobile repetitive sequences that can replicate and integrate into the genome

Two classes:

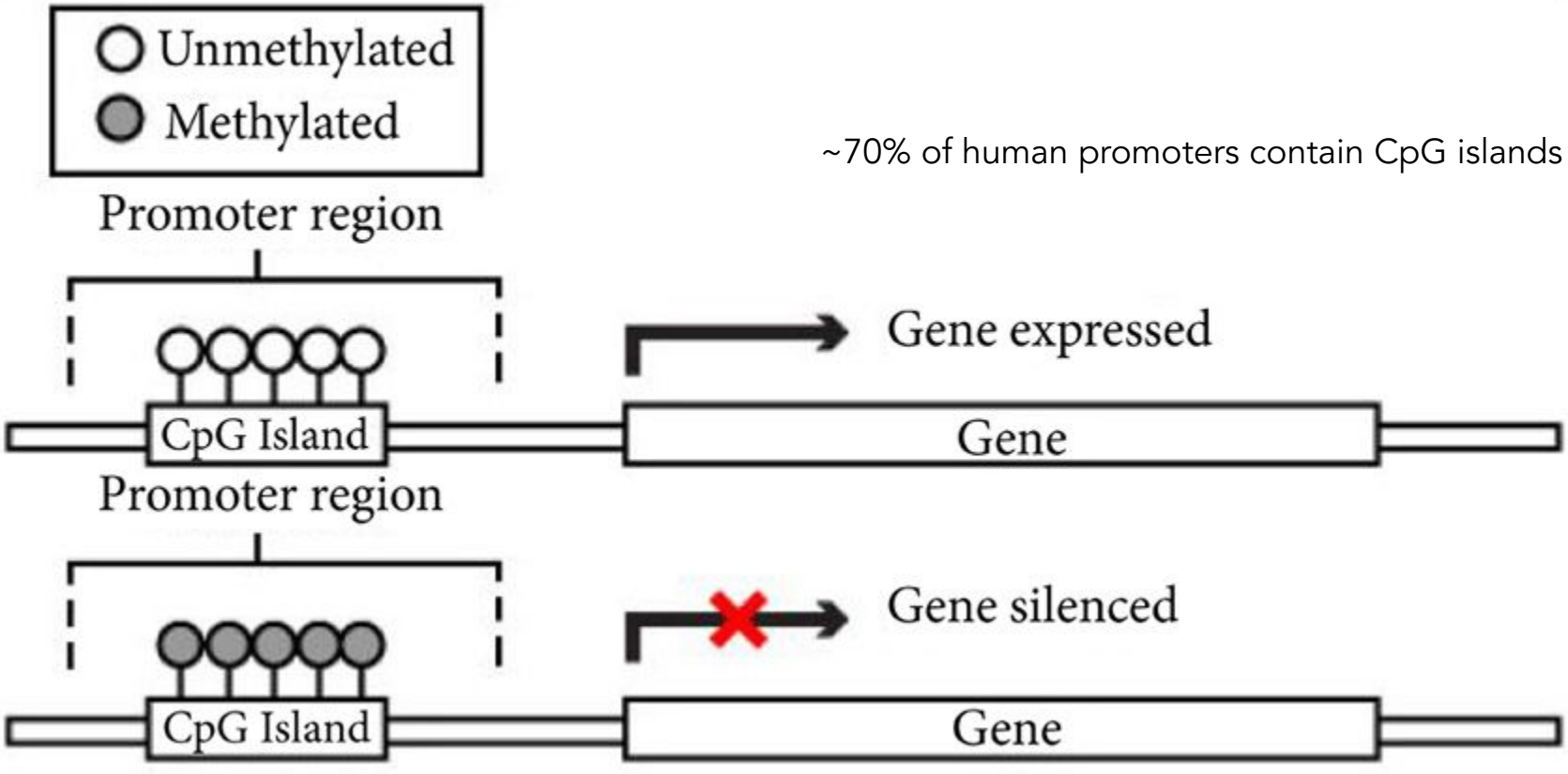
- Class 1 - RNA intermediate
- Class 2 - DNA intermediate

Make up ~45% of the human genome



Wanding Zhou, Gangning Liang, Peter L. Molloy, Peter A. Jones. DNA methylation enables transposable element-driven genome expansion. Proceedings of the National Academy of Sciences Aug 2020, 117 (32) 19359-19366

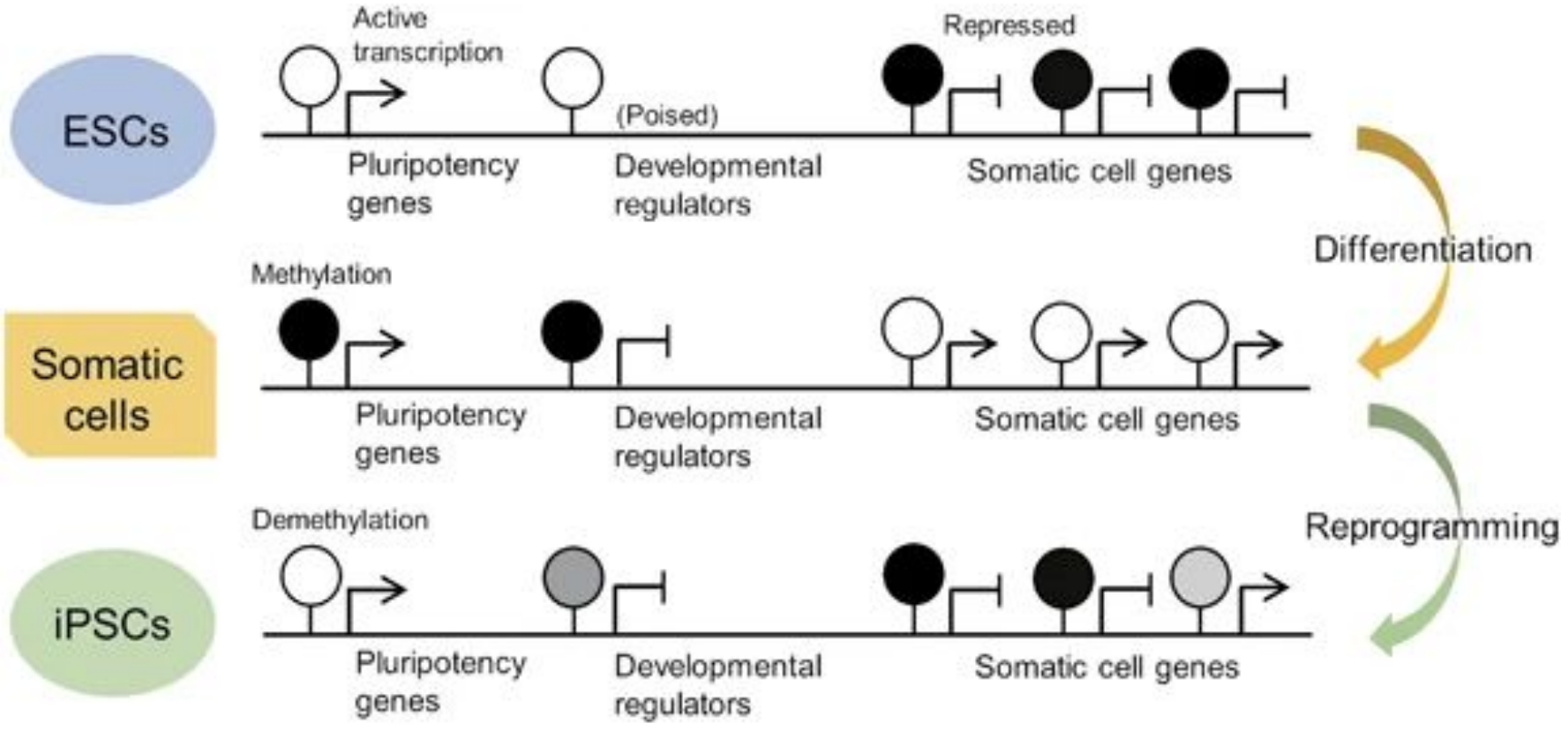
CpG Methylation



~70% of human promoters contain CpG islands

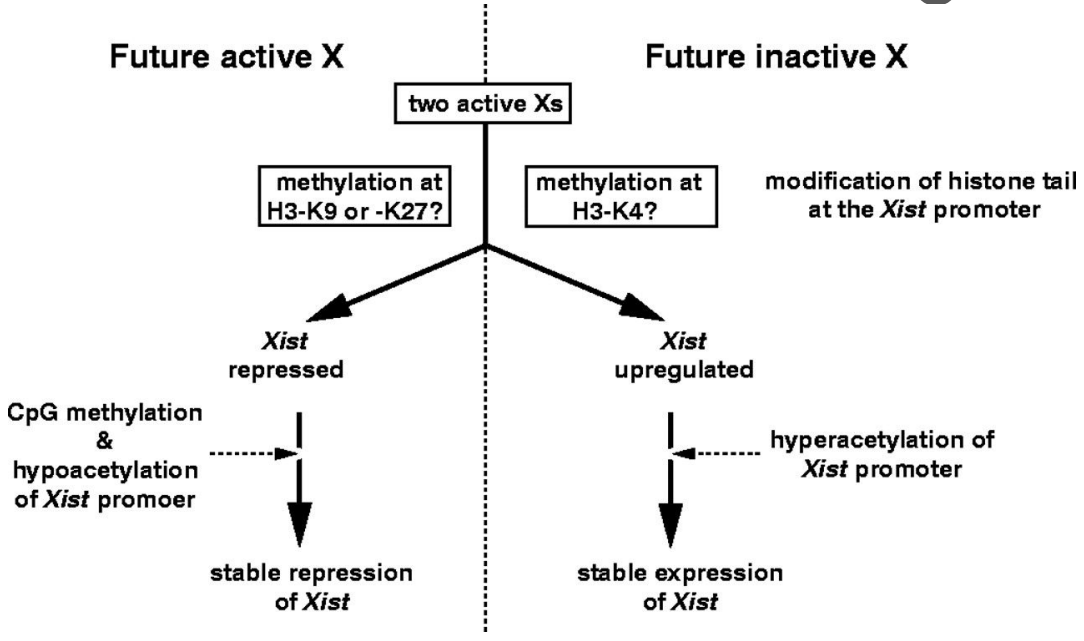
Peters et al. *Oncology Discovery* 2014 2:3

Methylation reinforces differentiation



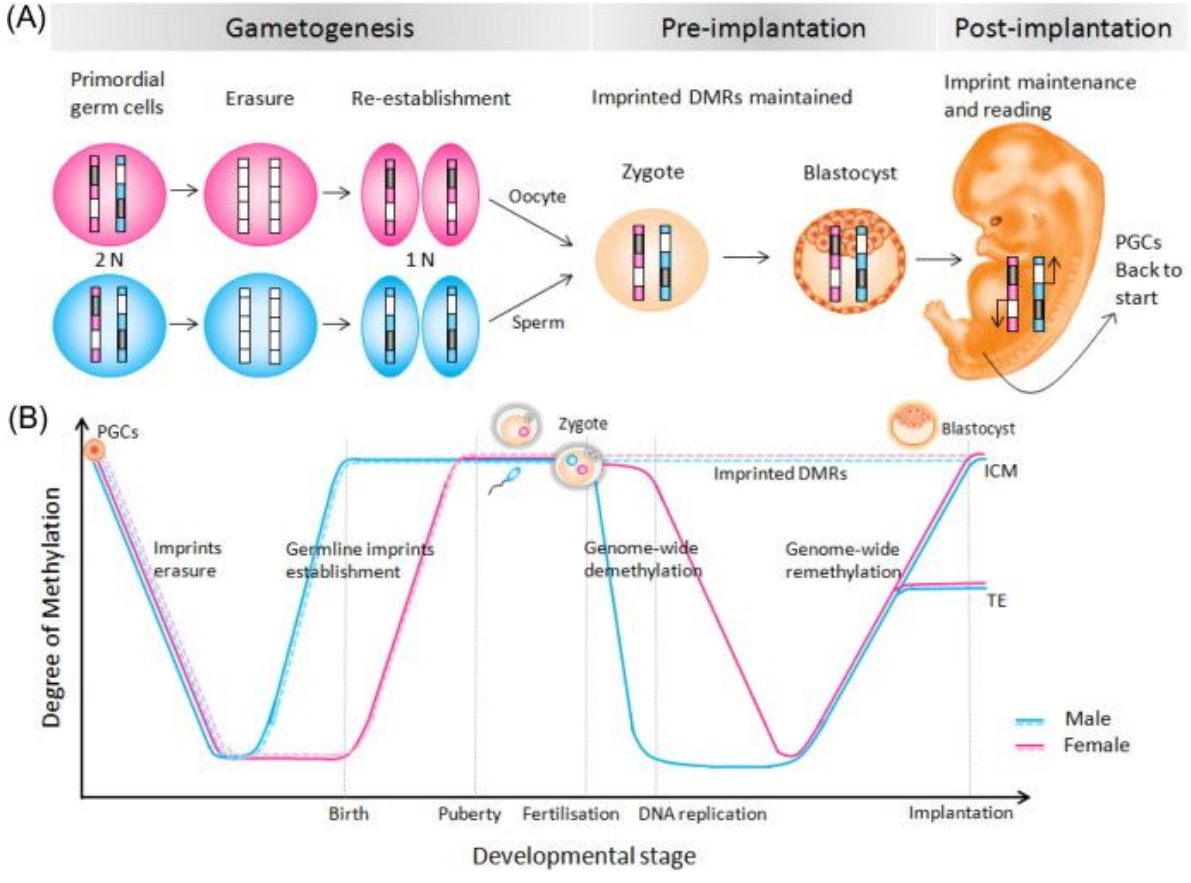
Kim M, Costello J. DNA methylation: an epigenetic mark of cellular memory. Exp Mol Med. 2017 Apr 28;49(4):e322.

X Chromosome Inactivation



Sado T, Okano M, Li E, Sasaki H. De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. Development. 2004 Mar;131(5):975-82.

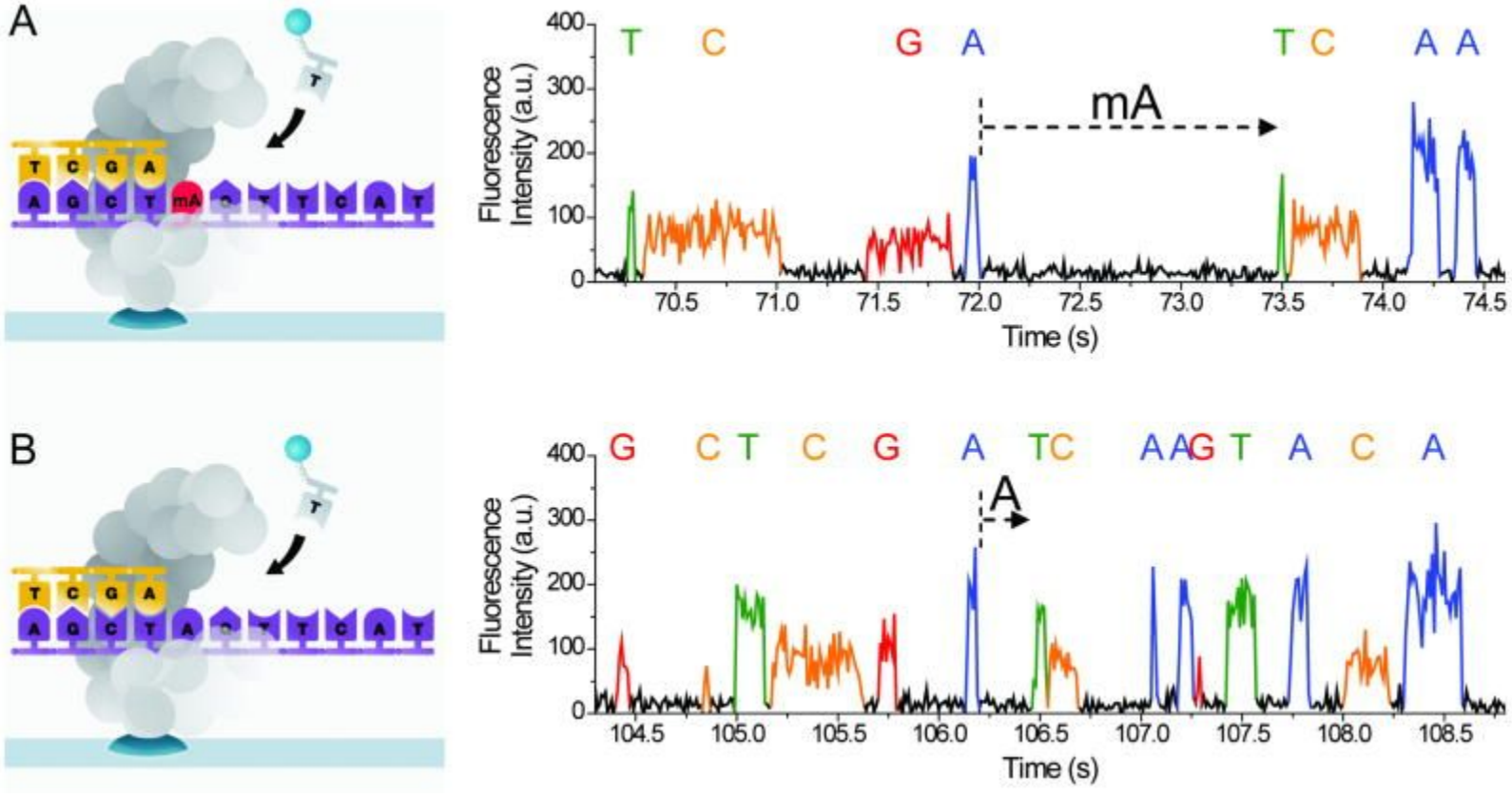
Imprinting



Ishida, Miho & Moore, Gudrun. (2012). The role of Imprinted genes in humans. Molecular Aspects of Medicine. 34. 10.1016/j.mam.2012.06.009.

Direct sequencing of DNA methylation

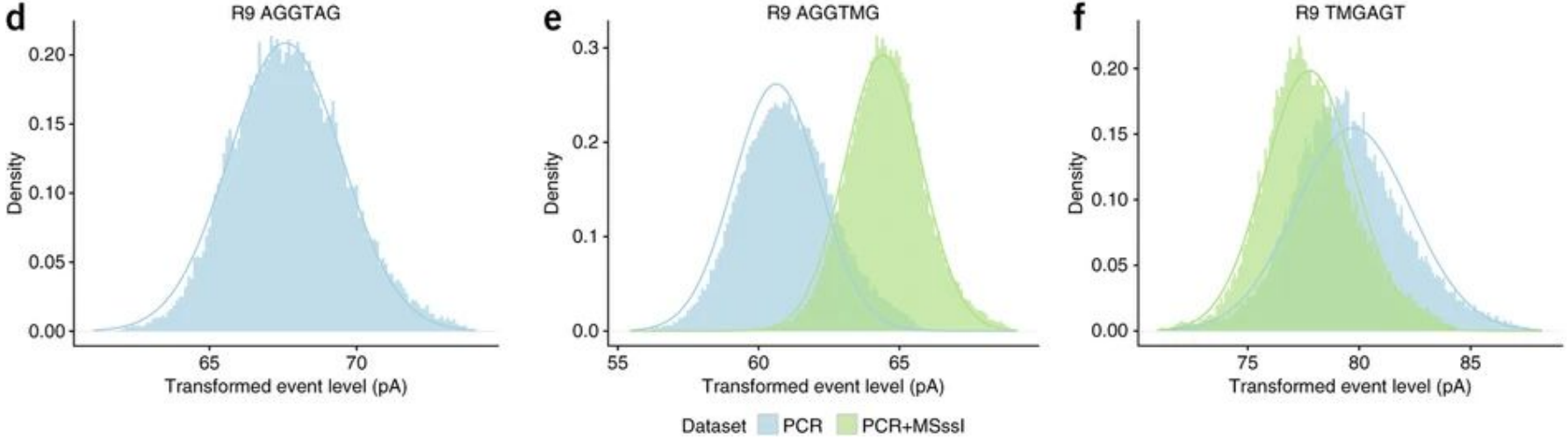
PacBio



Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010 Jun;7(6):461-5. doi: 10.1038/nmeth.1459. Epub 2010 May 9.

Direct sequencing of DNA methylation

Nanopore



M = 5mC
MSssI = CpG Methyltransferase

Simpson, J., Workman, R., Zuzarte, P. *et al*. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 14, 407–410 (2017).