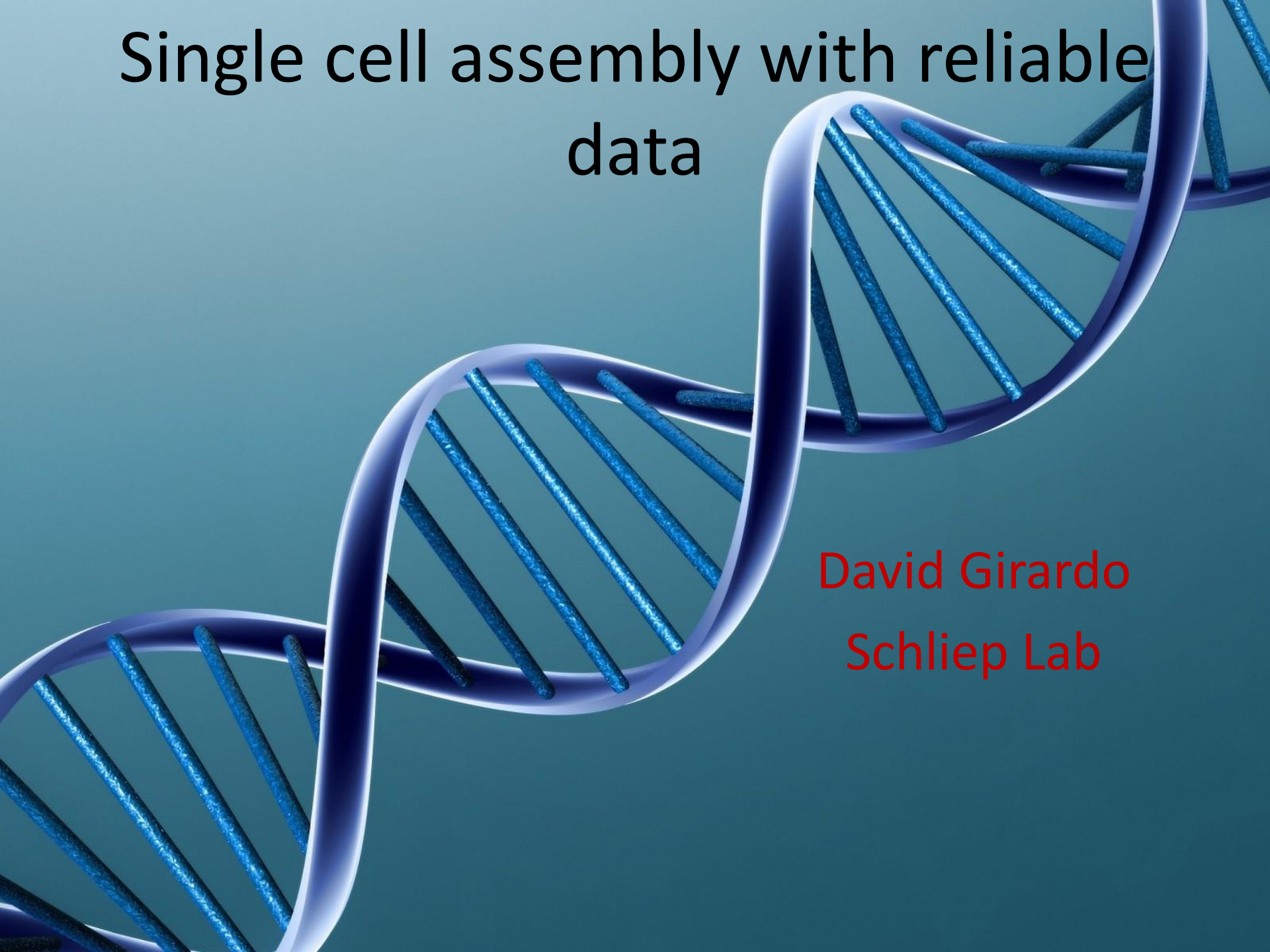


Single cell assembly with reliable data

David Girardo
Schliep Lab



Some Quick Genomics

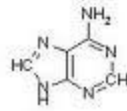


Yellow < A

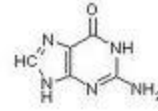
Green > T

Blue < C

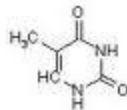
Red > G



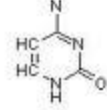
Adenine



Guanine



Thymine



Cytosine

Nucleotides
or
“Bases”



tctttcatagttggaacaagattt

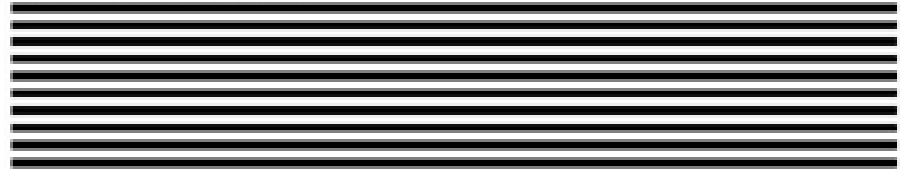
agaaagtatcaaccttggttctaaa



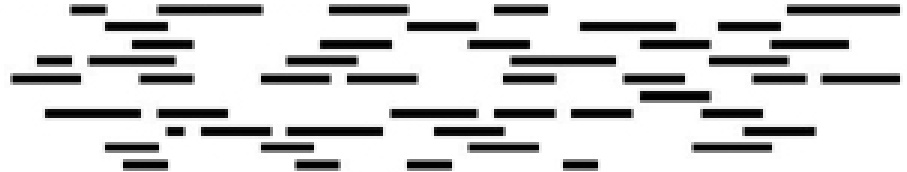
Reverse Complement

Shotgun Sequencing

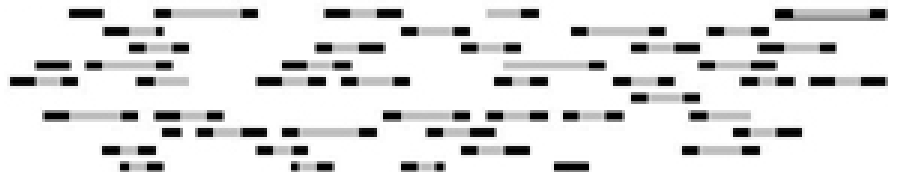
Multiple copies of genome



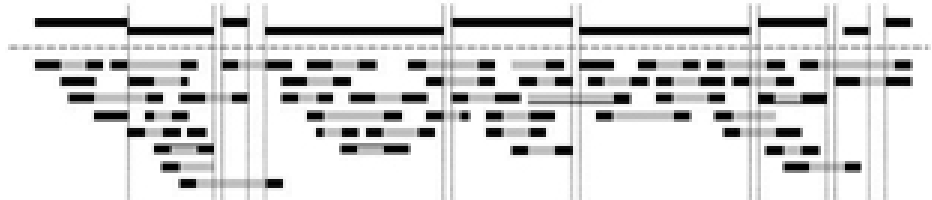
Sheared random fragments



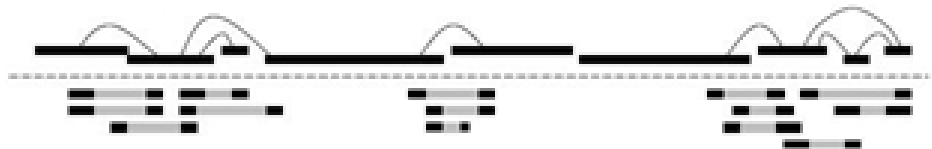
Sequenced Reads



Contig Assembly

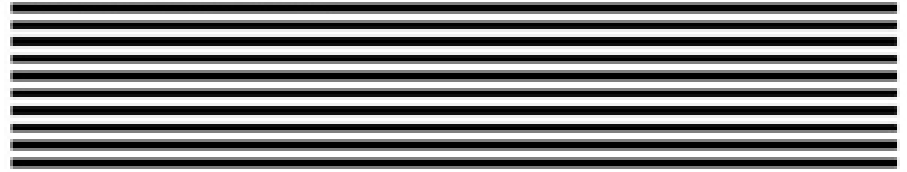


Scaffold Assembly



Shotgun Sequencing

Multiple copies of genome

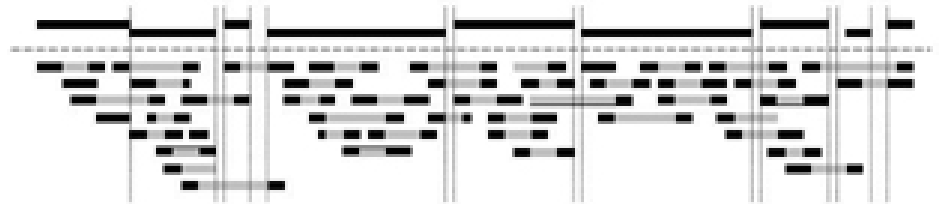


Sheared

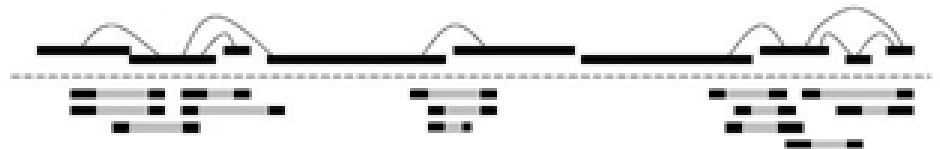
Se

BIG DATA

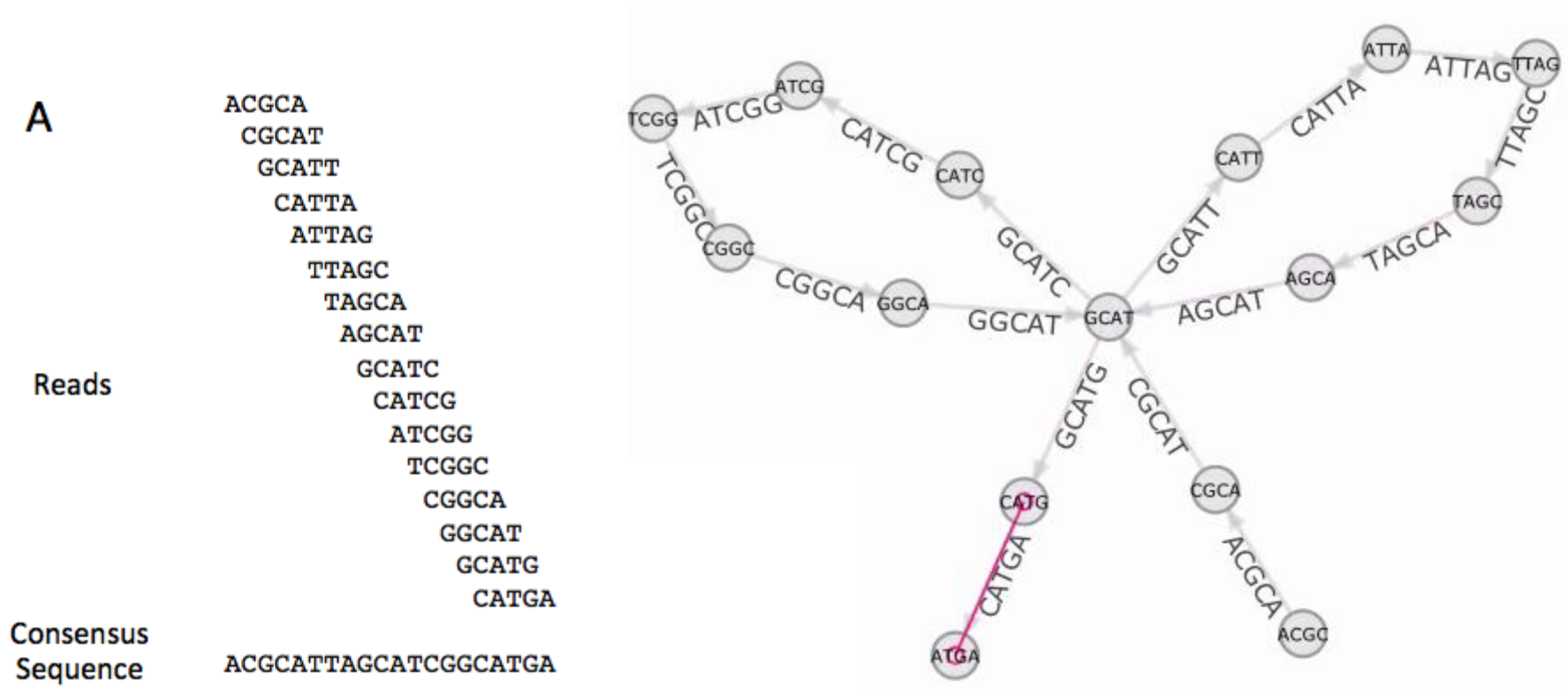
Contig Assembly



Scaffold Assembly



De Bruijn Graph assembly



Assembly difficulties

biological sequence

ATTATATGCGCTATGAGCT

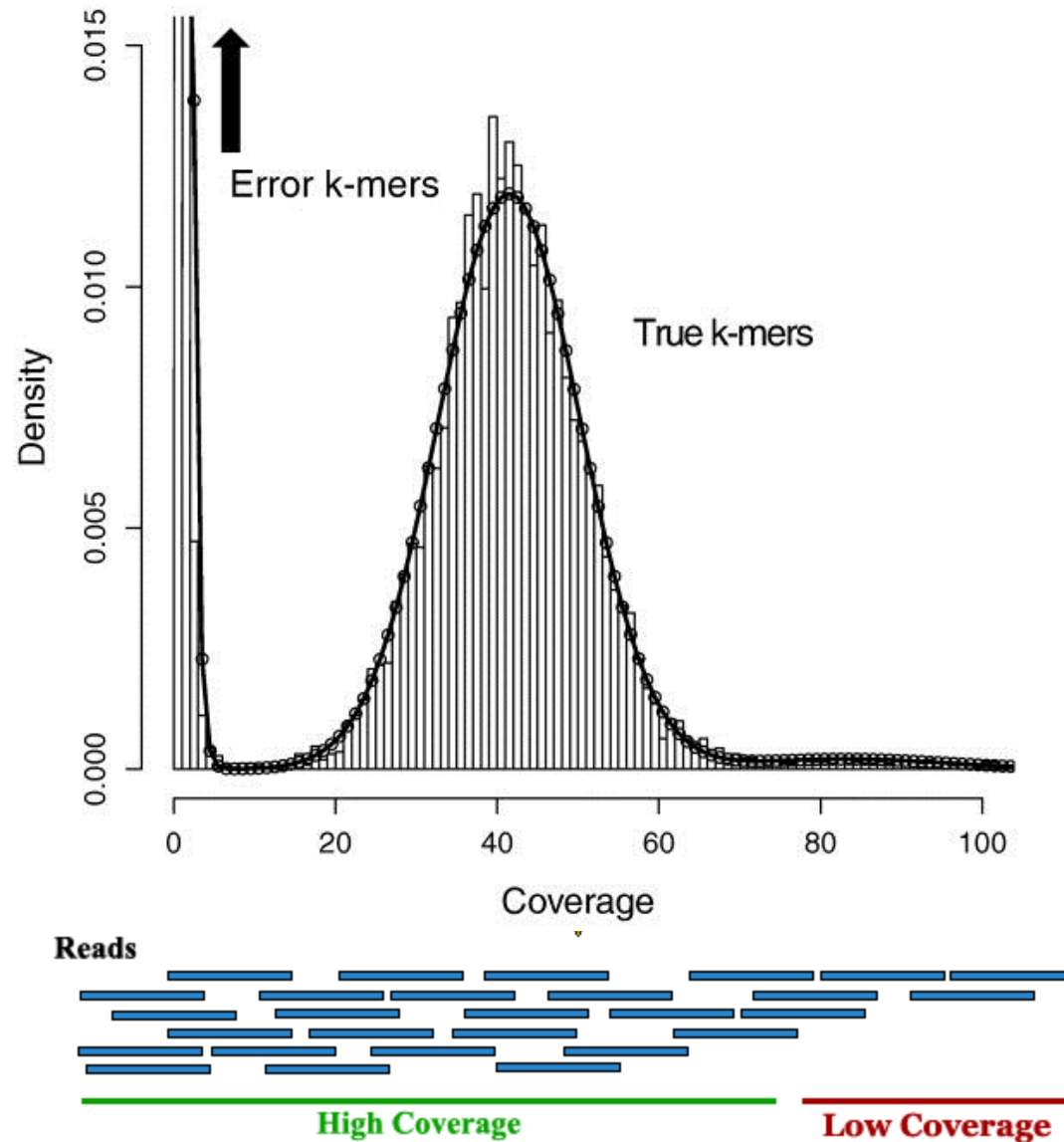
base calling errors

ATTATATGCGCTACGAGAT

read contaminants

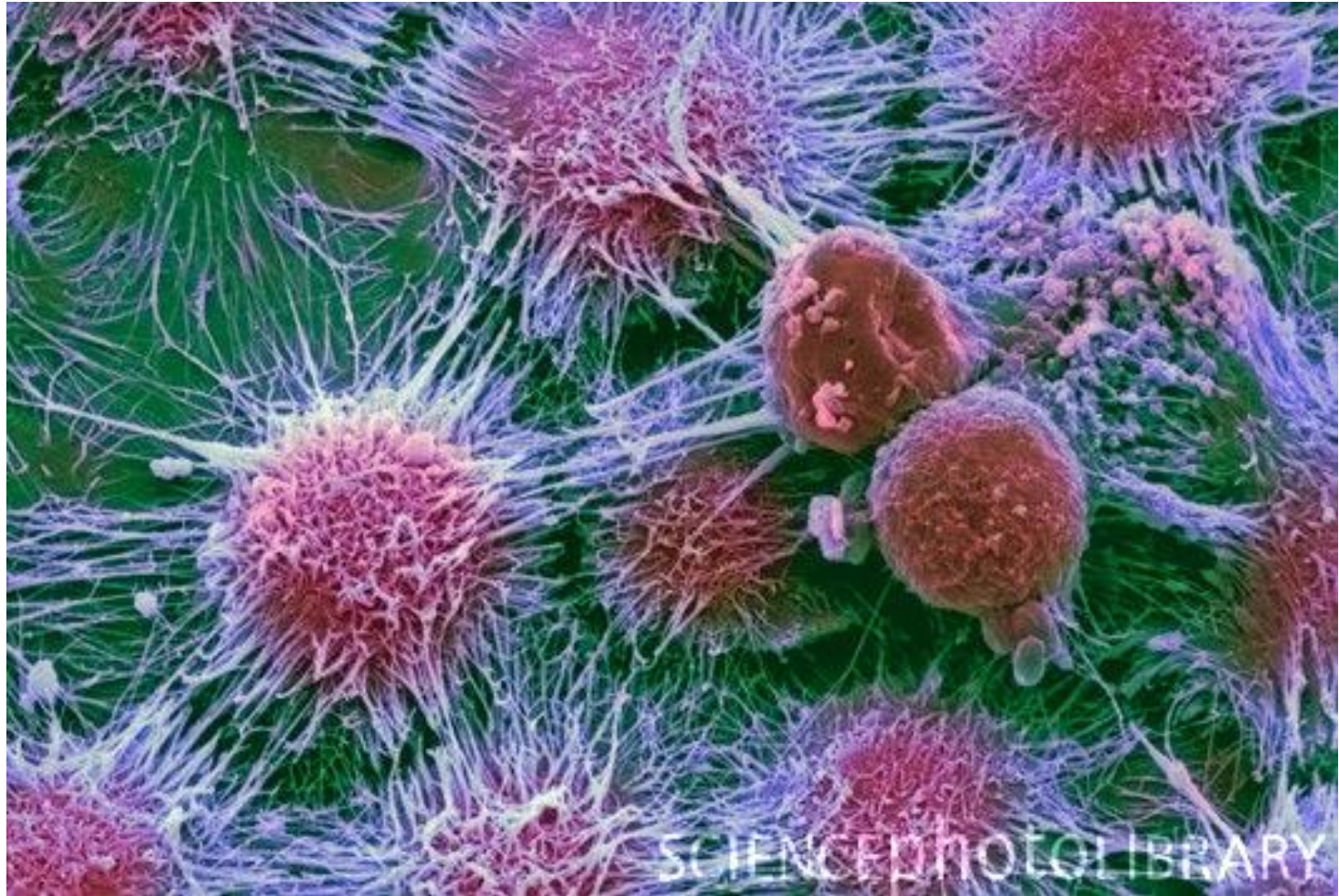
ATTATATGCGCTAAATCTAC

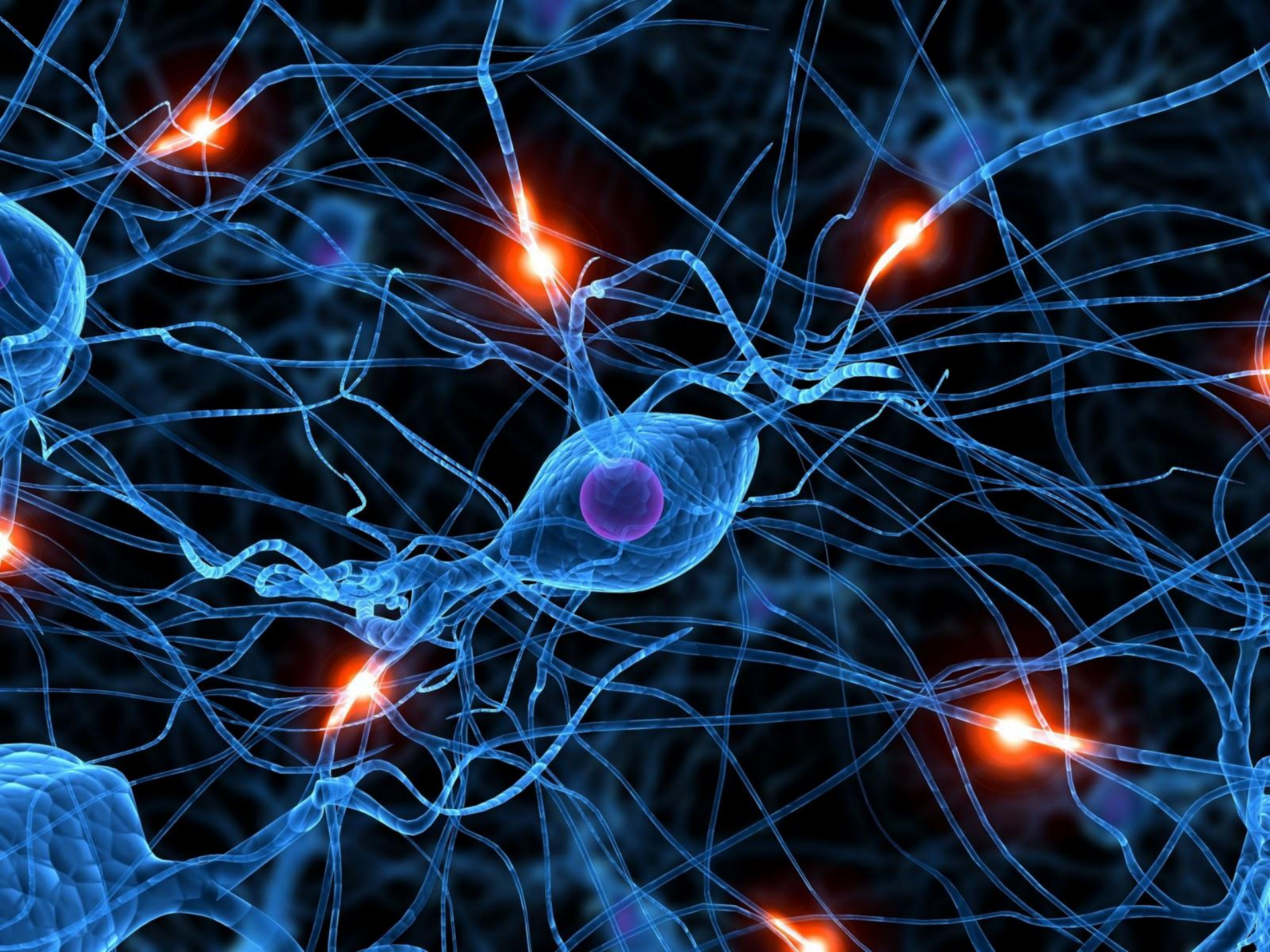
Base call error filtering



Single Cell Genomics

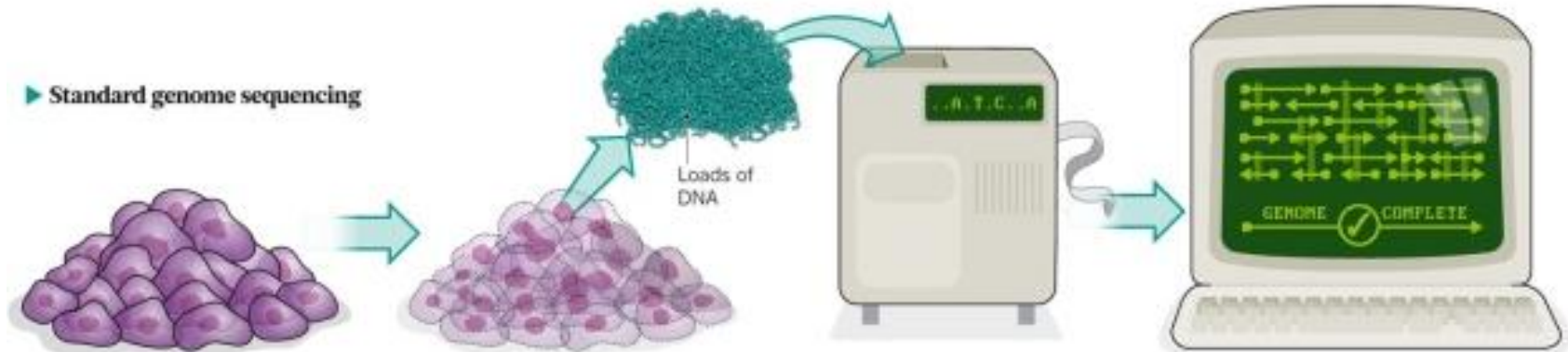




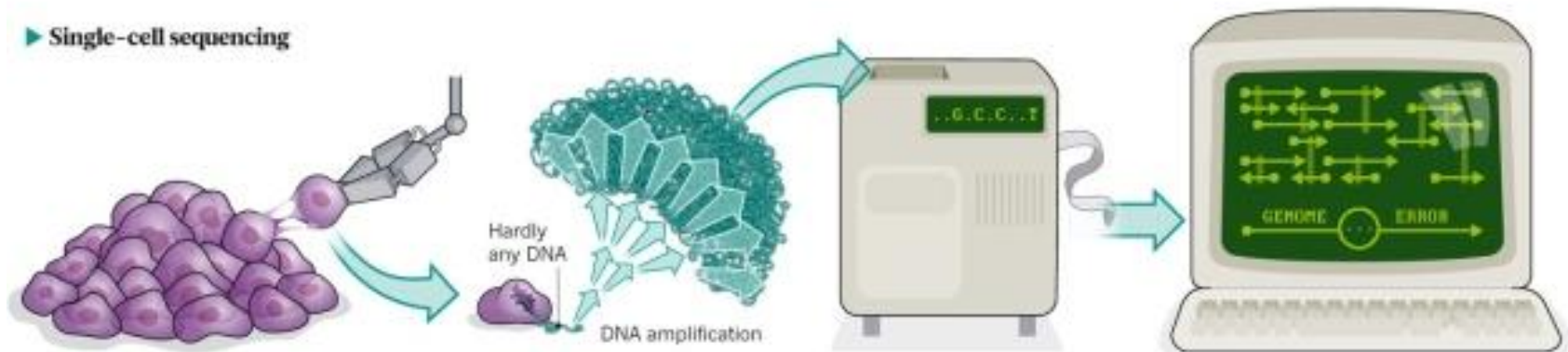


Single cell problems

► Standard genome sequencing

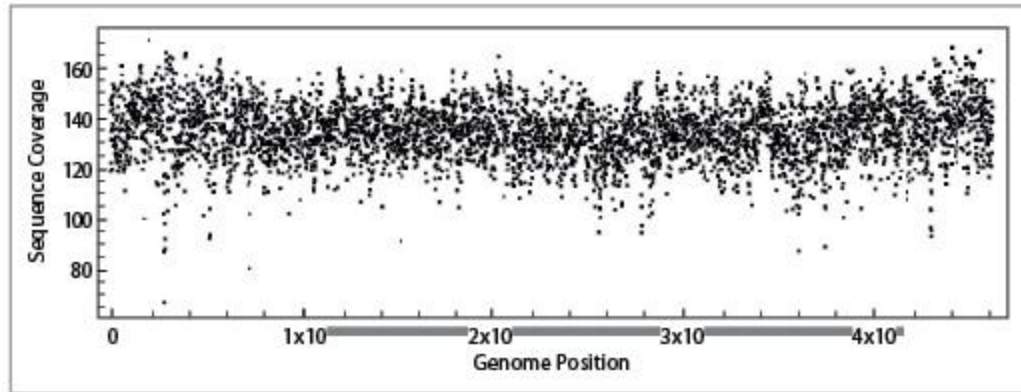


► Single-cell sequencing

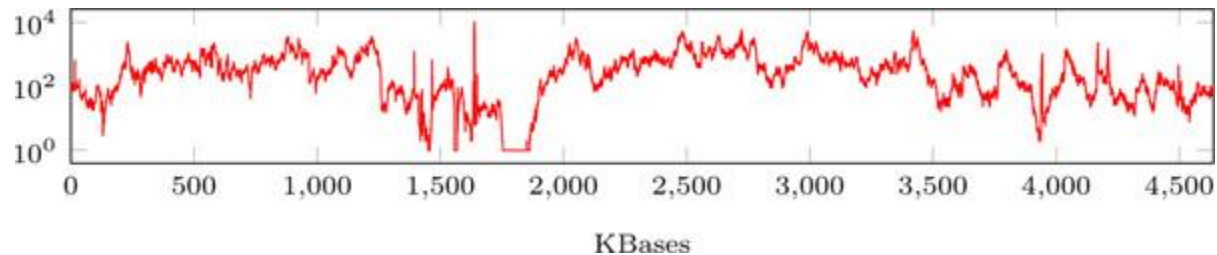


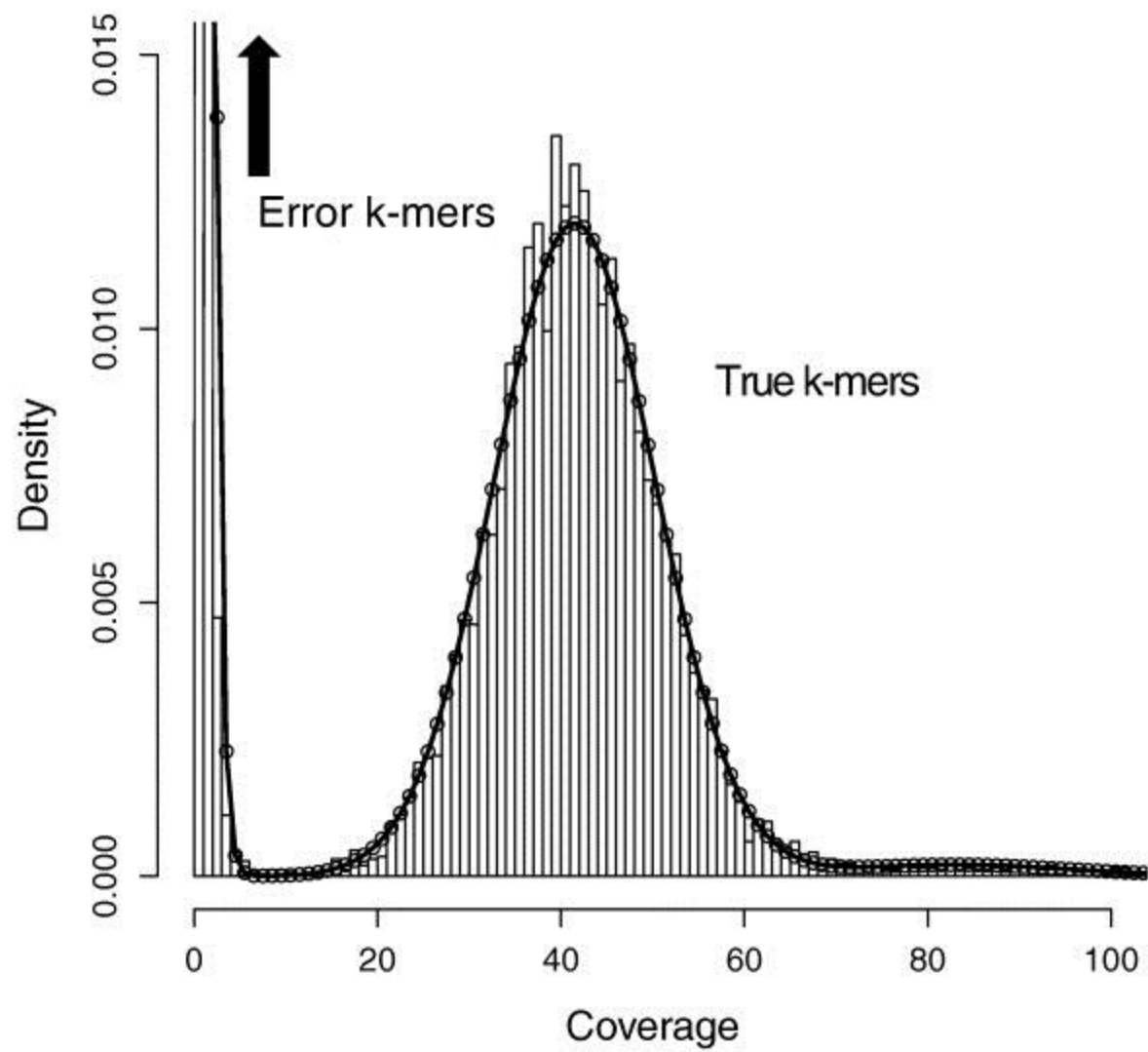
Single cell problems

Heterogeneous Sample Coverage



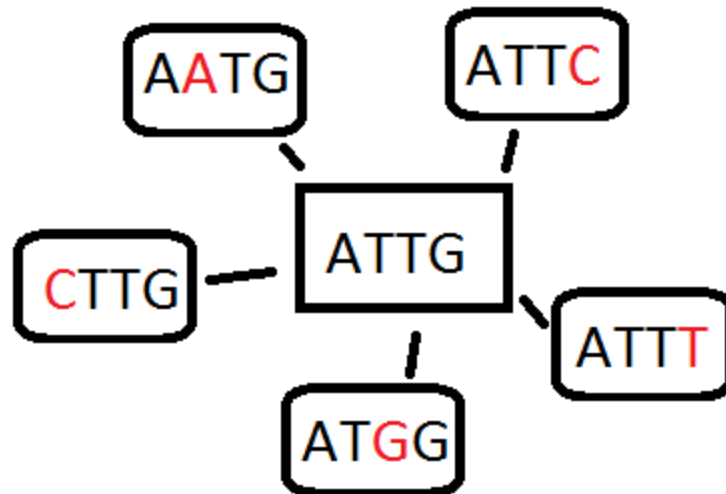
Amplified Single Cell Sample Coverage





A possibility

- Collapse hamming distance balls around most frequent K-mers



How can this be improved?

- First: Does this really work? How well?
 - How many real K-mers do we lose?
 - What kind of data performs best?
- More biologically sophisticated heuristics
 - Ex. Maximum Likelihood/Maximum entropy
- Find pairs of reliable K-mers with reliable distance estimate

