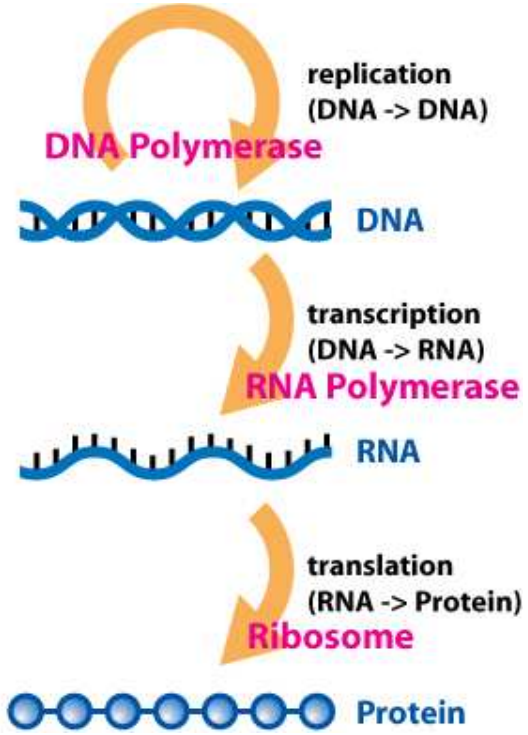
An Overview of Illumina's Sequencing Technology and its Applications

Dr. Epameinondas Fritzilas
Computational Biology Group
Illumina Cambridge

University of Primorska
4 March 2011

illumina®

What is DNA sequencing?



Central dogma of molecular biology

- We read the DNA: the primary piece of information, the letters of the book.
- We can get (almost) all letters of the book, but this doesn't mean that we understand the meaning of everything that is written there.

illumina®

More and more organisms are getting completely sequenced

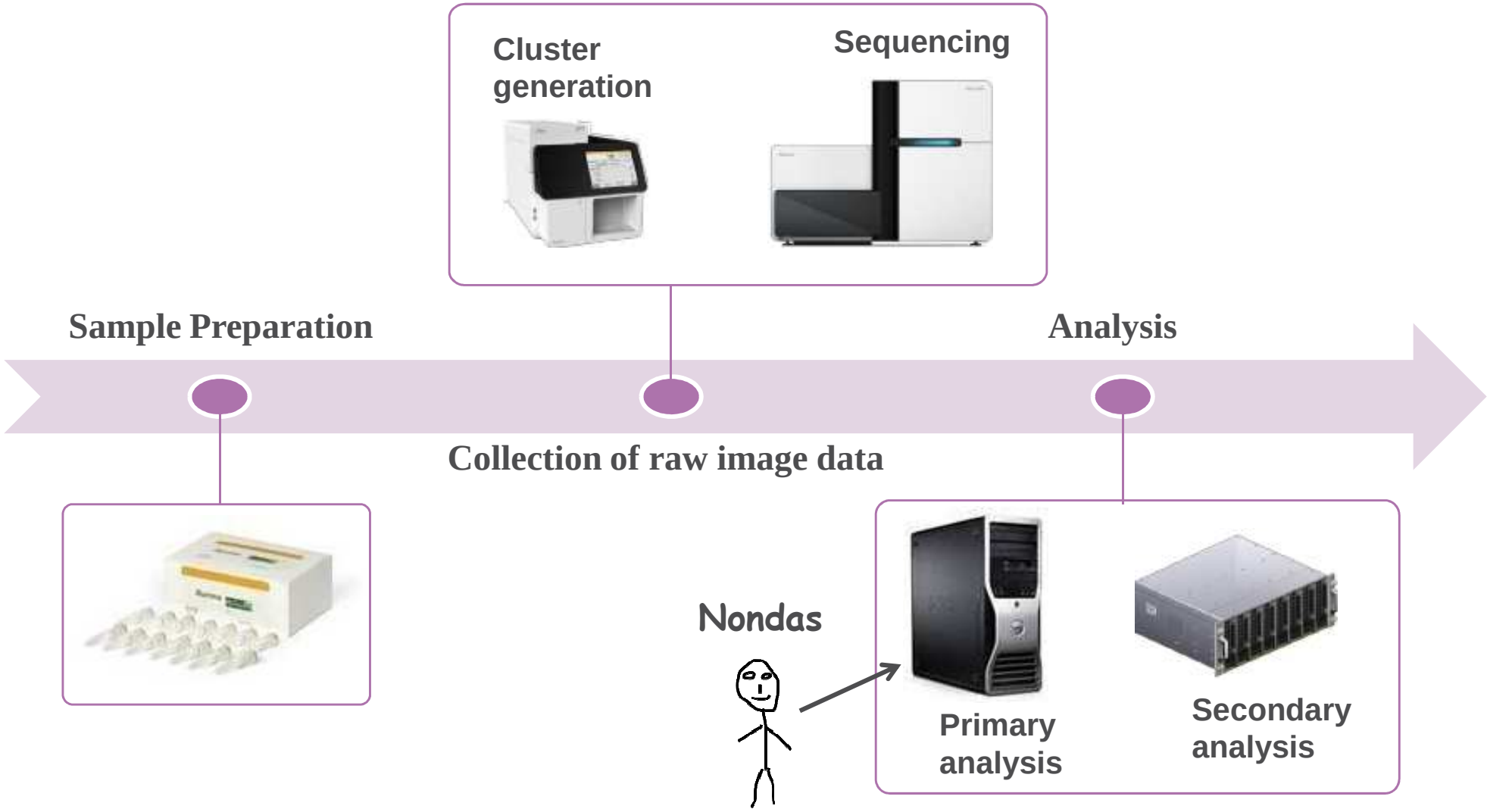
illumina®

Who is Illumina?

- A company based in San Diego (California, USA) with sites in Singapore, Hayward (California) and Chesterford (near Cambridge, UK)
- Illumina started as a company making microarrays.
- The sequencing technology was invented at Cambridge University and developed in a spin-off company called Solexa Ltd.
- Illumina bought out Solexa in 2006.
- Other companies in the high-throughput sequencing business: Life Technologies, 454/Roche, Helicos BioSciences, Complete Genomics, Pacific Biosciences, Oxford Nanopore Technologies

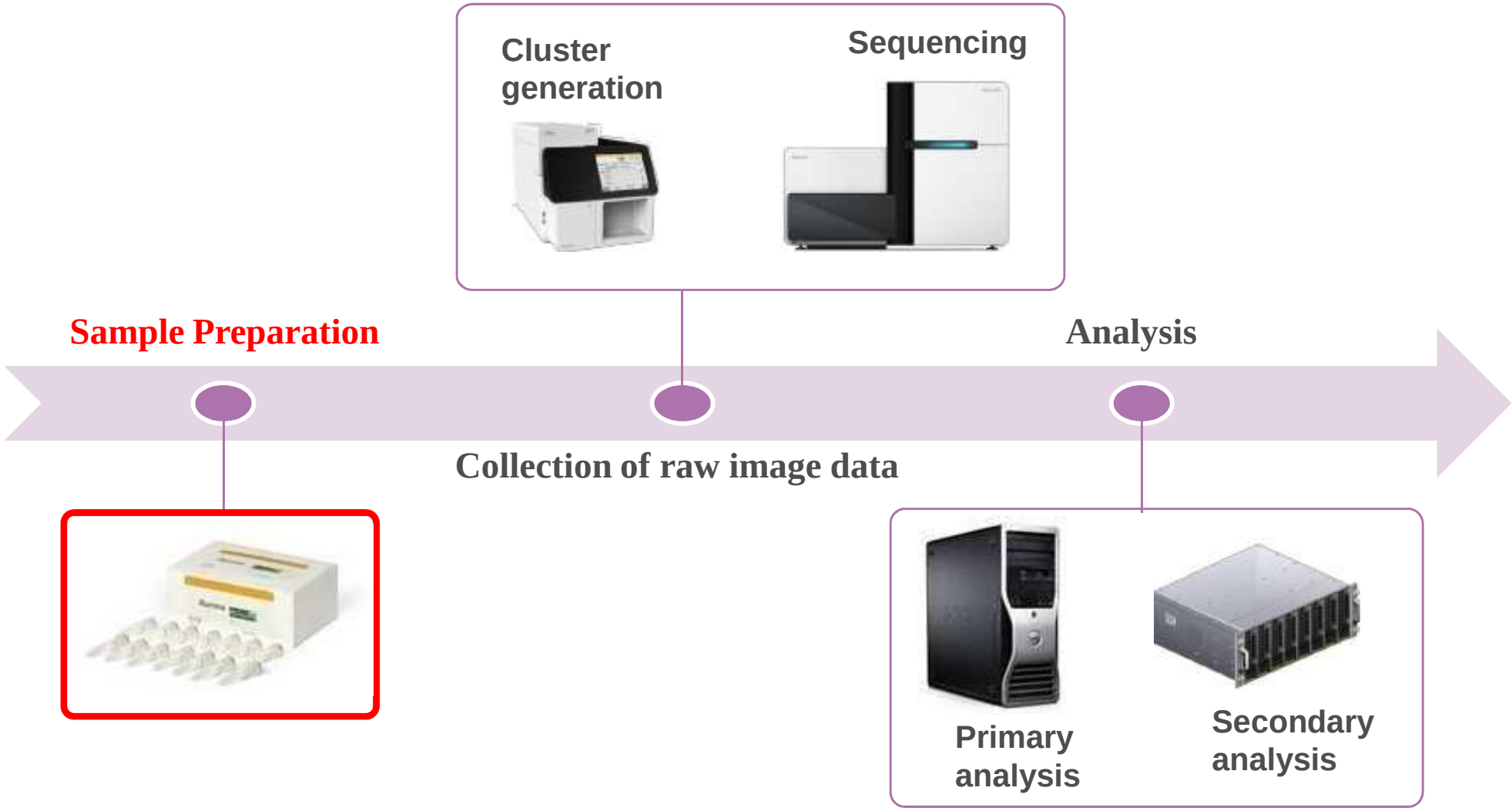
illumina®

Today's topic: Illumina's sequencing workflow



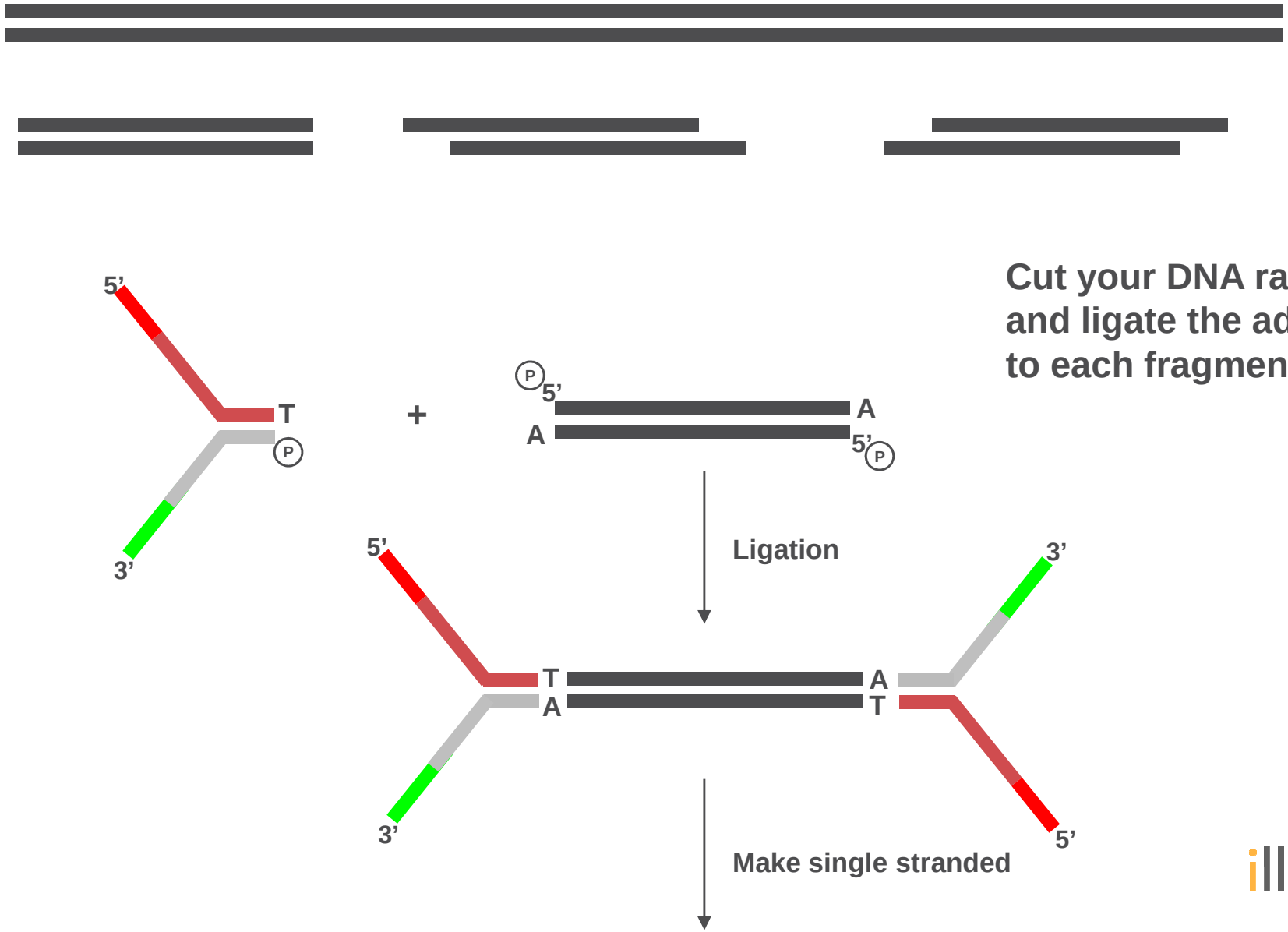
illumina®

Sequencing workflow



illumina®

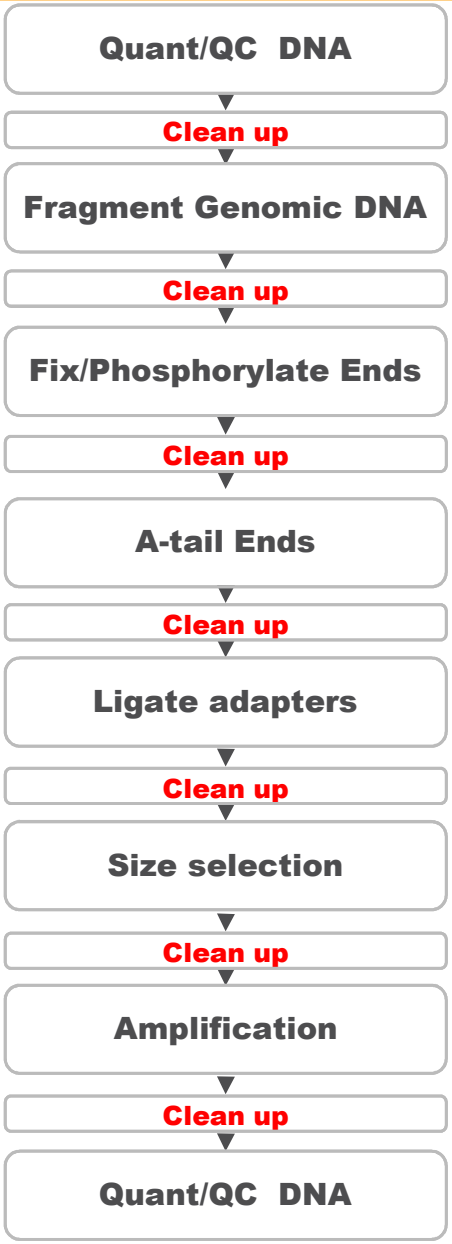
Essence of the sample preparation



Cut your DNA randomly and ligate the adapters to each fragment

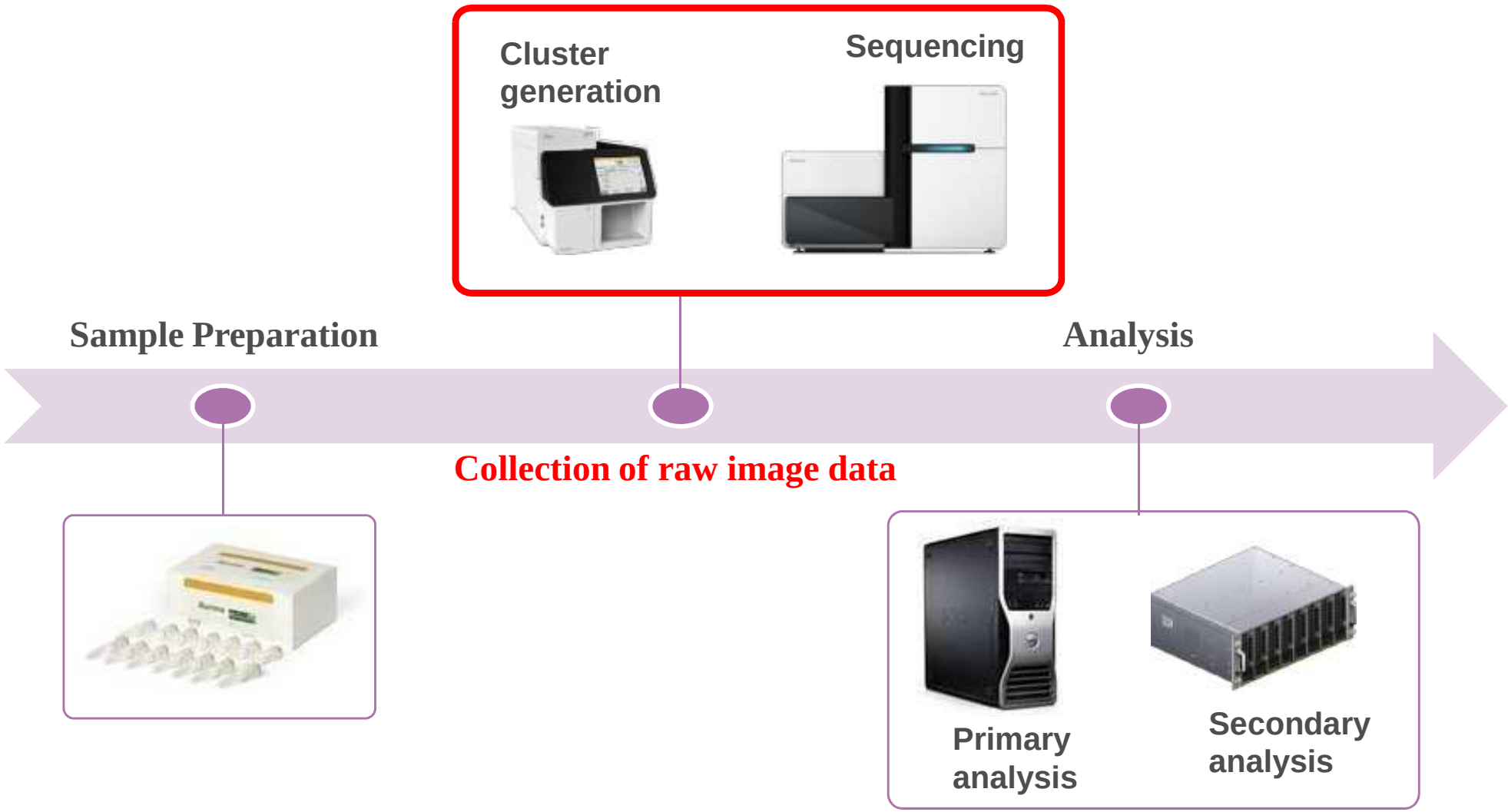
illumina®

In practice many steps are involved



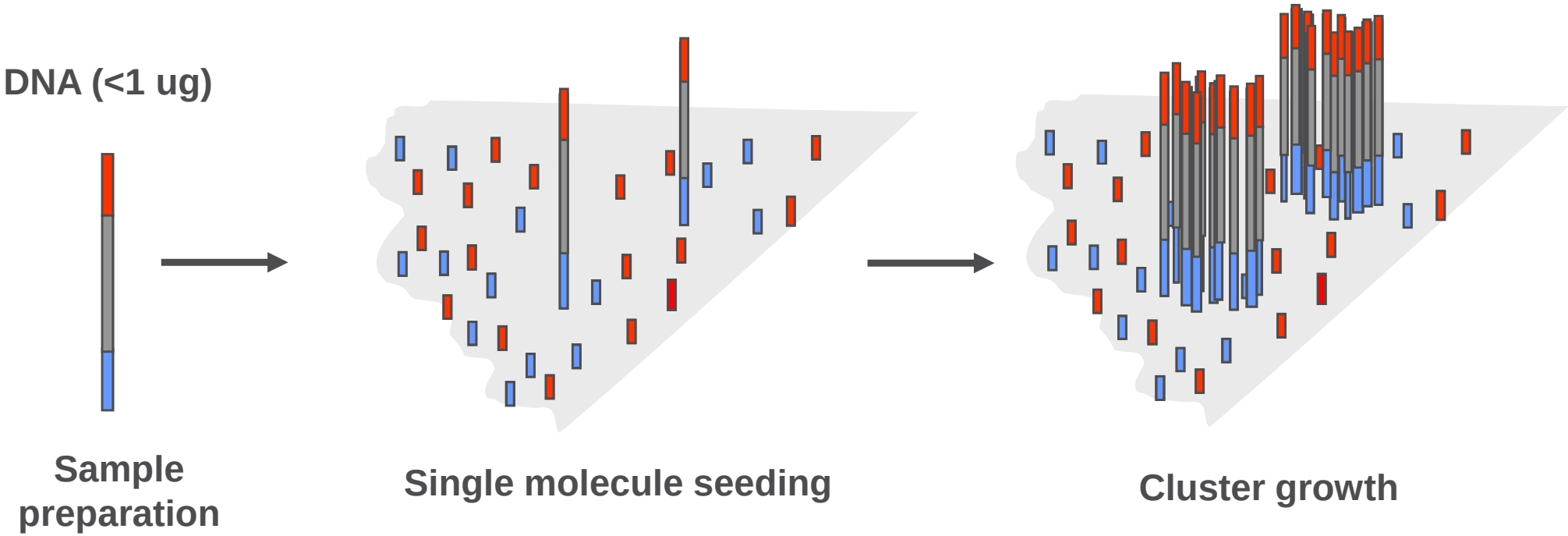
illumina®

Sequencing workflow



illumina®

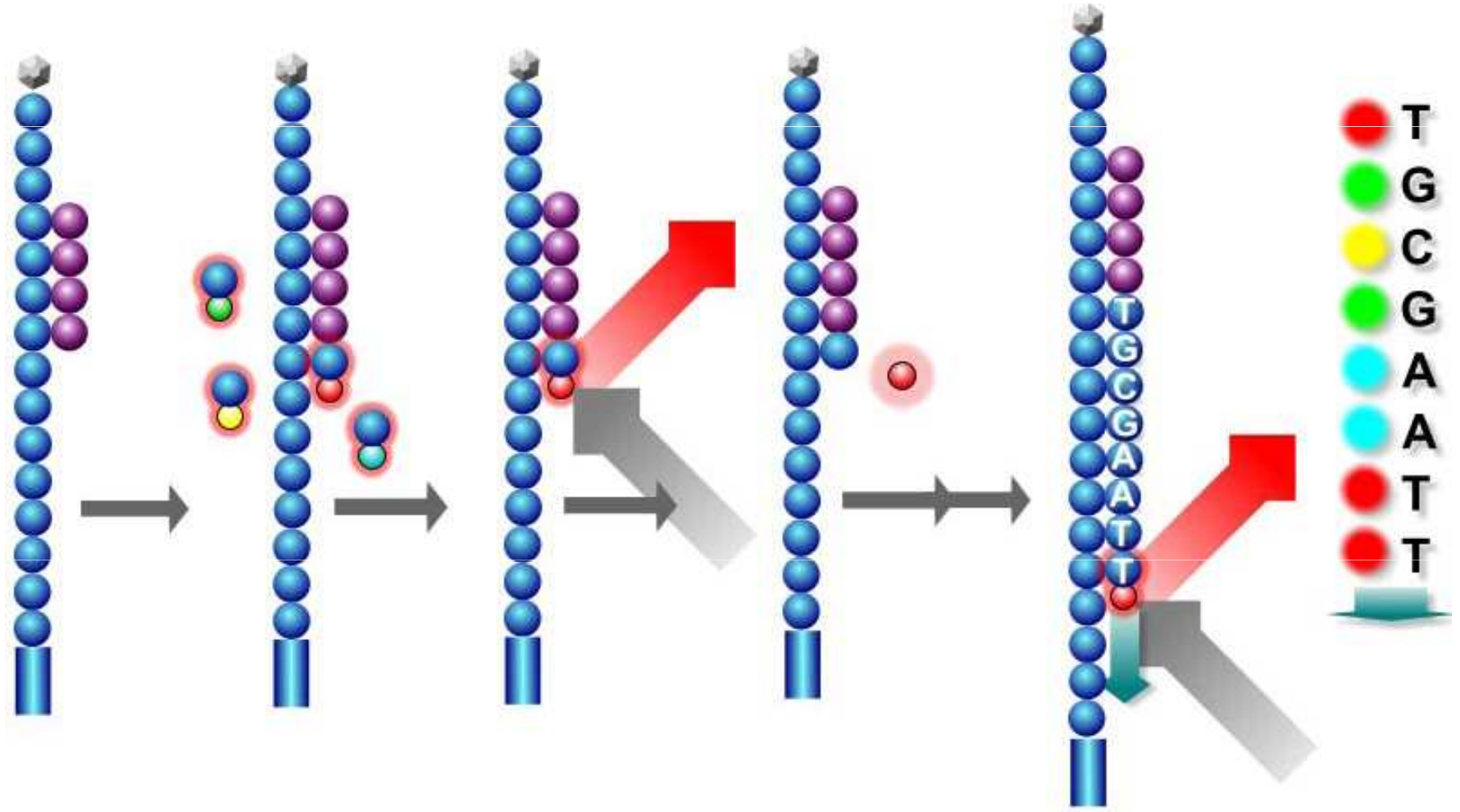
Step 1: Cluster generation on the surface



Each cluster is a colony with many copies of the same fragment. We need many copies in order to get a detectable signal.

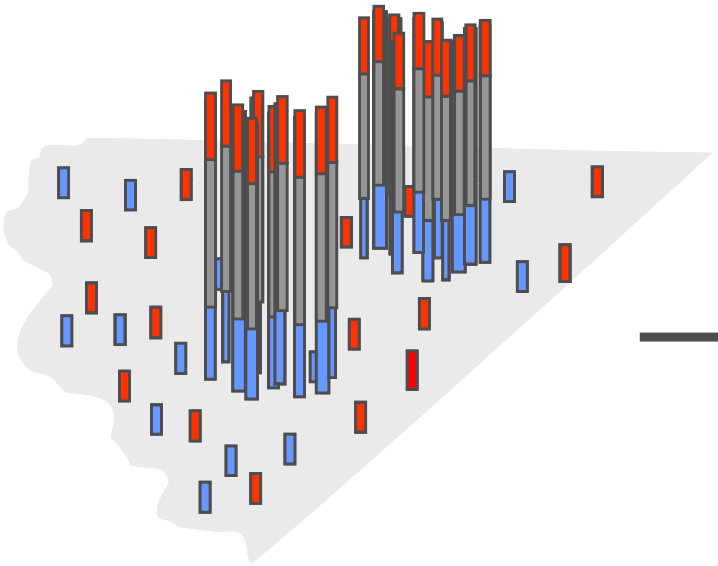
illumina®

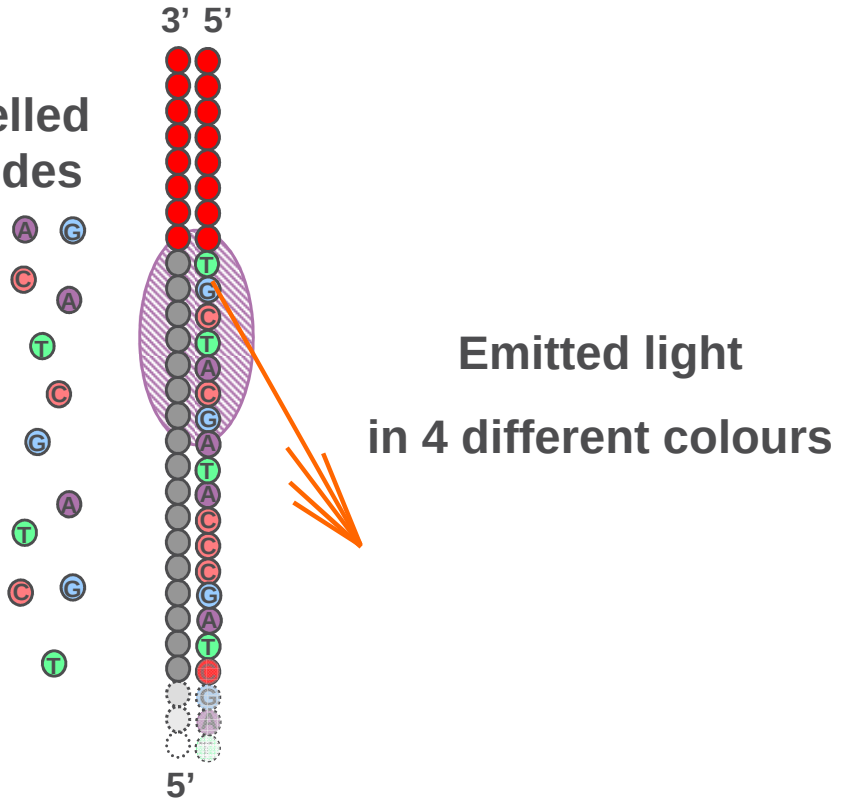
Step 2: Sequencing by Synthesis

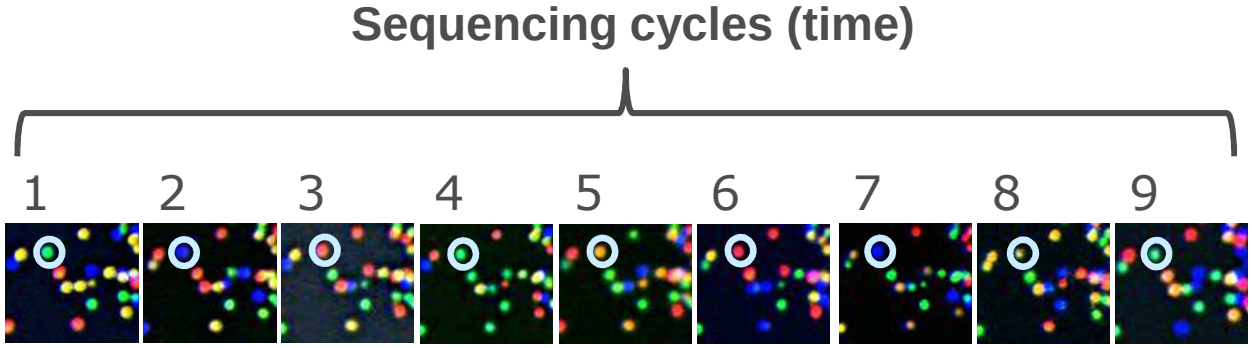


illumina®

Step 2: Sequencing by Synthesis



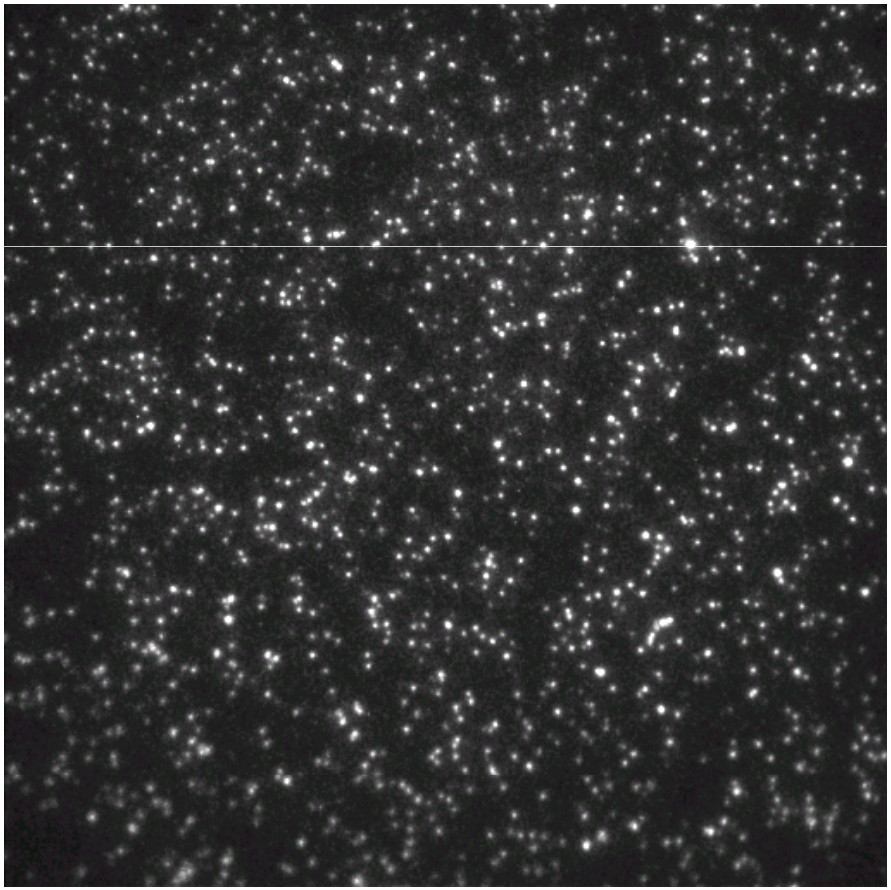




Base calling

illumina®

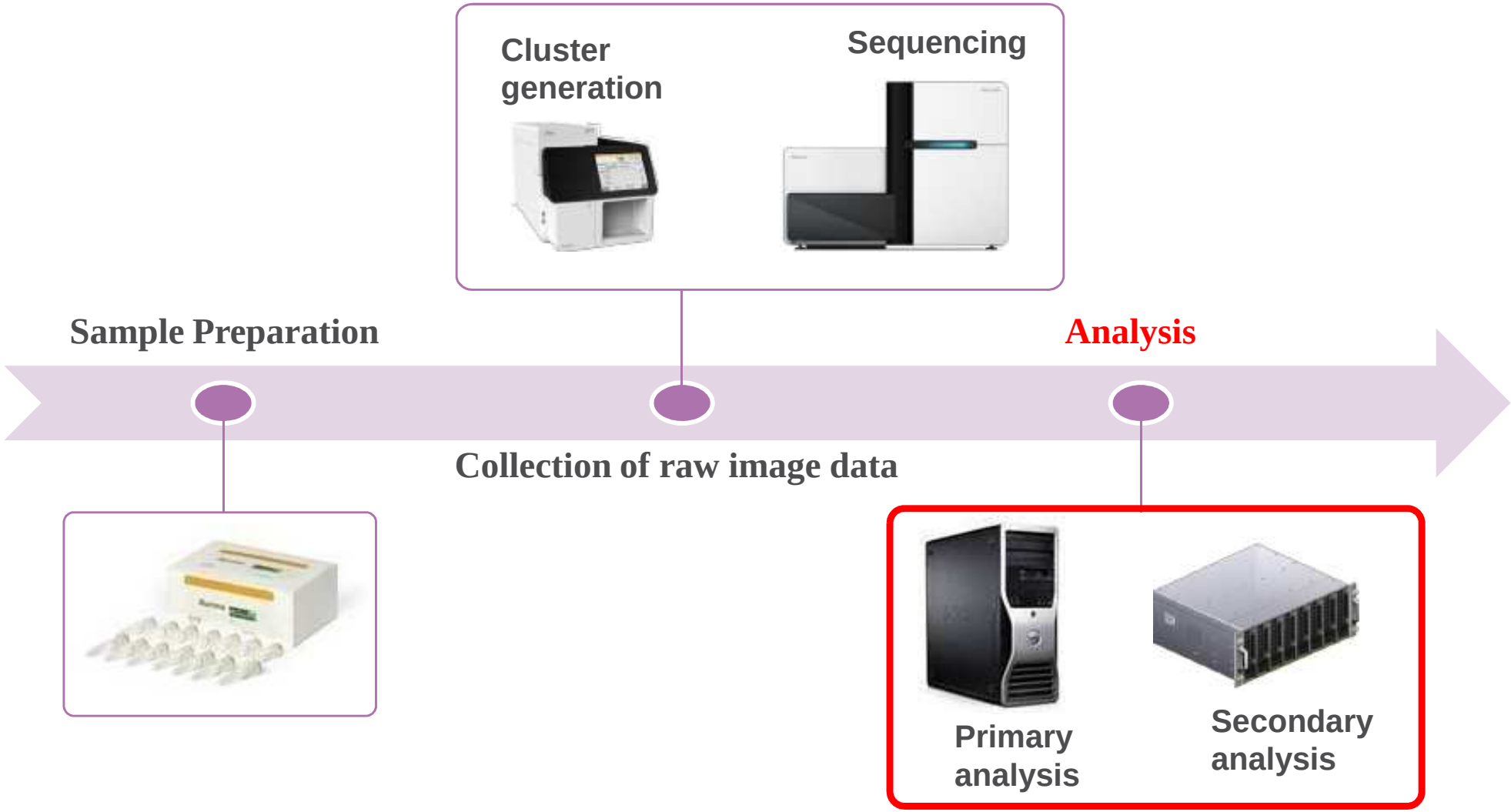
How do the real images look like?



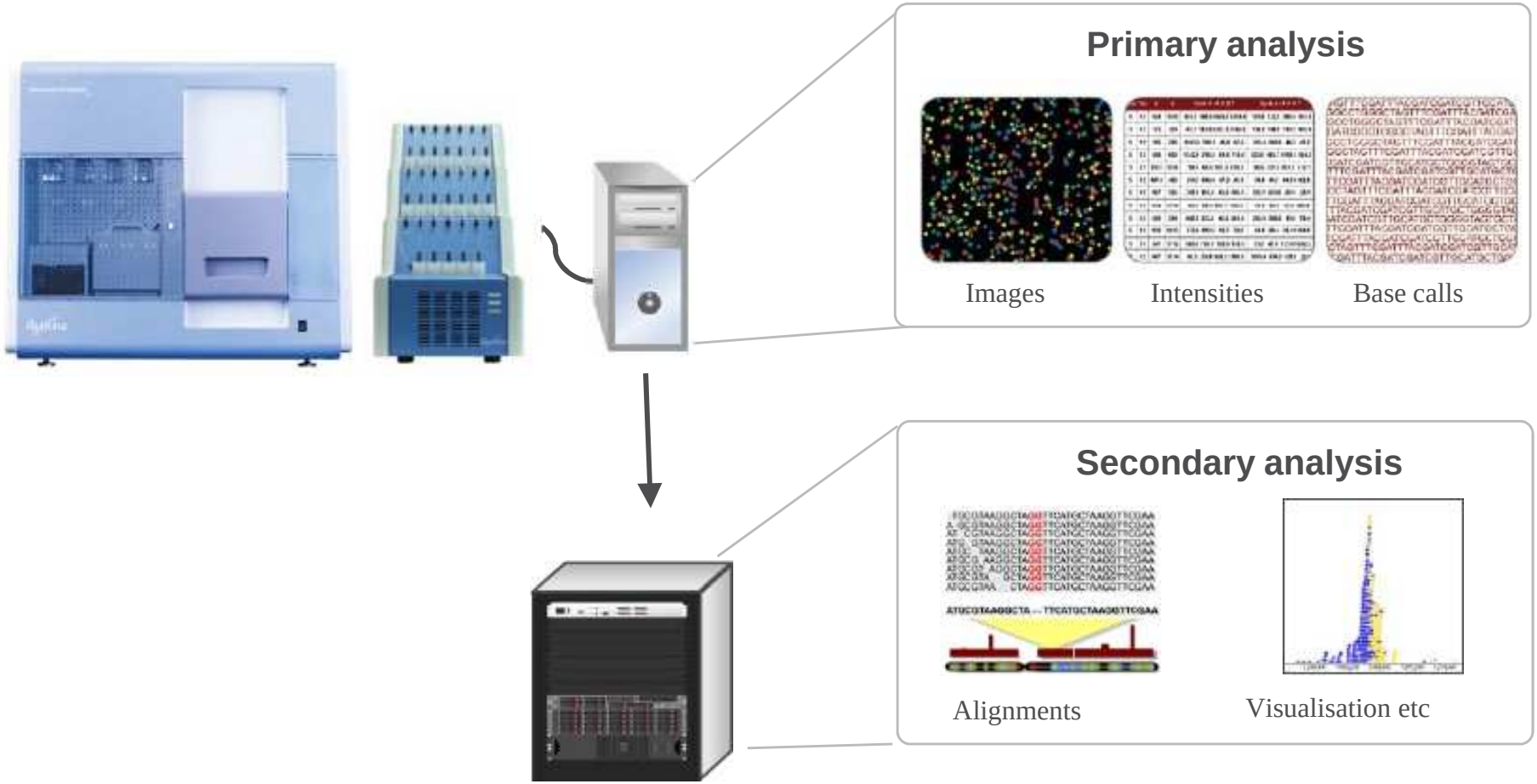
- Approximately 1 million spots / mm^2
- For each sequencing cycle we get 4 such images, one for each base colour.

illumina®

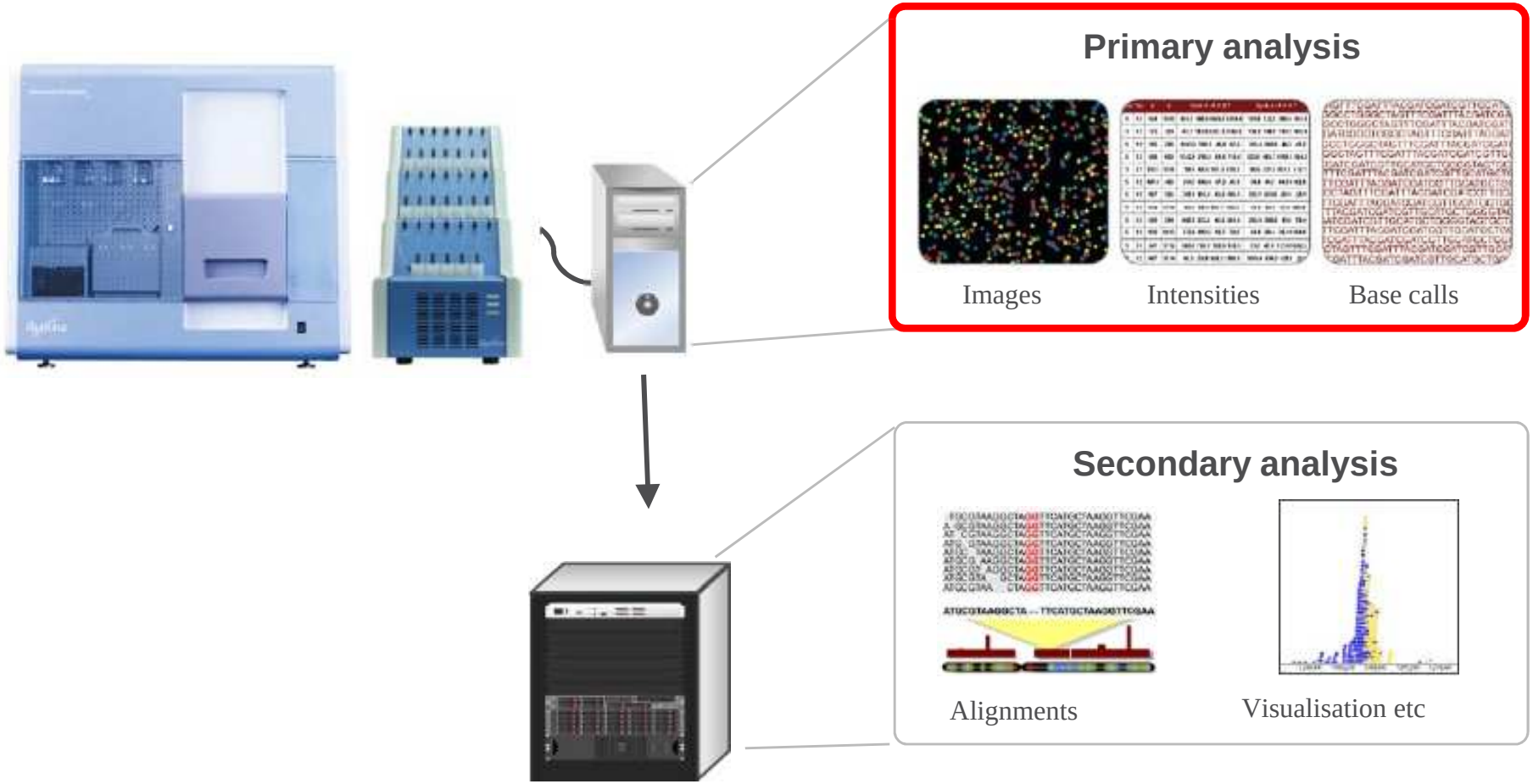
Sequencing workflow



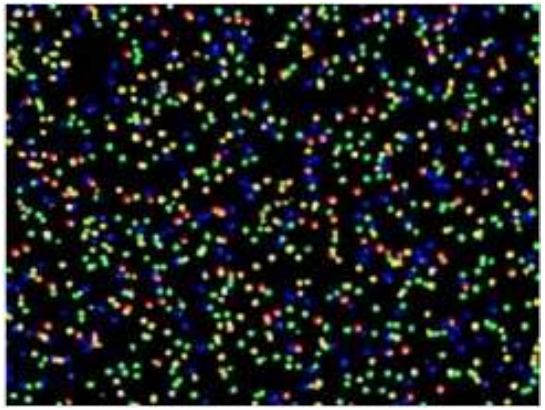
Analysis workflow (100% informatics)



Analysis workflow



From the images to the intensities



x	y	t	A	C	G	T
17	23	3	97	2	10	5
17	25	18	3	4	76	1
...	...	...	...	...	...	...
1001	1234	50	5	100	20	7

1. Detection: Find all clusters on the image
2. Registration: Track clusters over multiple sequencing cycles
3. Extraction: Give intensity estimates for clusters in a given image

illumina®

Base-calling

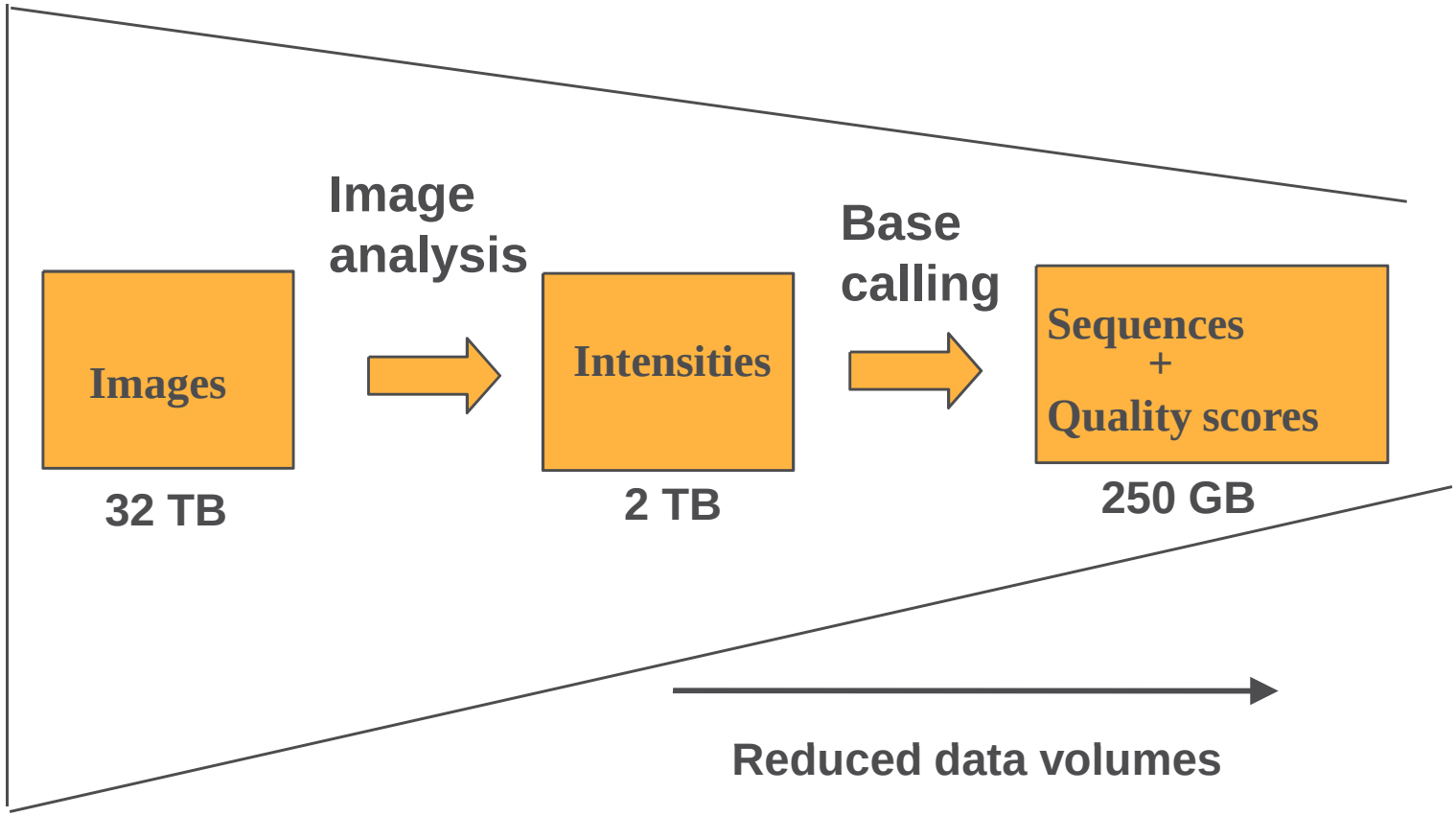
- Conversion of intensity data into sequences and quality scores.

x	y	t	A	C	G	T
17	23	3	97	2	10	5
17	25	18	3	4	76	1
…	…	…	…	…	…	…
1001	1234	50	5	100	20	7

- Essentially a classification problem that can be attacked with machine learning. But it has to be solved very fast.
- We need to output not only a base-call, but also a confidence score for the correctness of the call.

illumina®

Data reduction in primary analysis is crucial

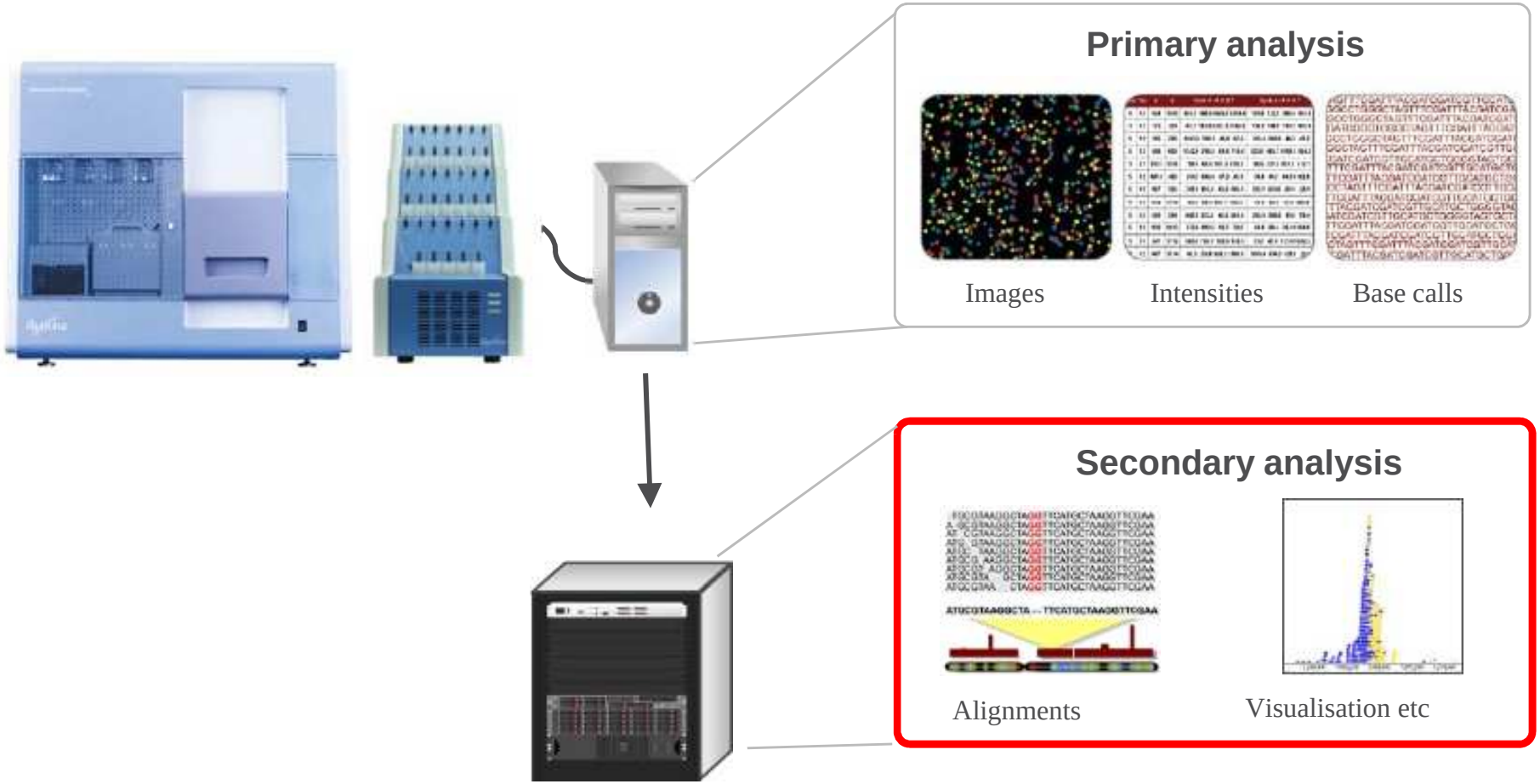


- Data volumes are shown for a HiSeq run that outputs 200 billion bases.
- Massive reduction in data volumes
- Image analysis and base-calling are done on the instrument PC.

  Only the sequences are transferred to a remote analysis server.

illumina®

Analysis workflow



OK, we got 1 billion reads from the instrument. And now what? ...

- Remember that the reads are randomly sampled short sequences across the whole genome.
- 1 billion reads x 100 bases per read = 100 billion bases

  Human genome = 3 billion bases

  So, every position of the genome is covered 33 .3 times on average.

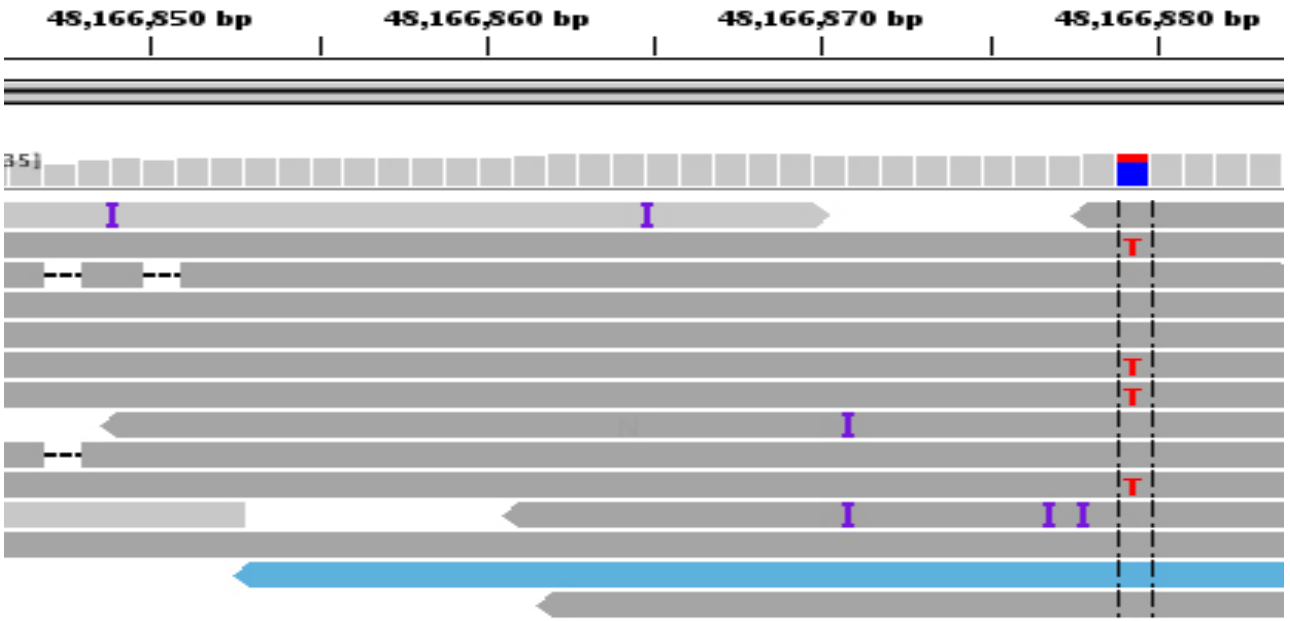
  More precisely, we use Poisson statistics for the coverage distribution.

- We can use the reads to solve two completely different tasks:

  re-sequencing and de-novo assembly

illumina®

Application I: Re-sequencing



Goal:

- Align sequences to approximately known reference sequence, allowing for small number of differences (approximate pattern matching)
- Look for consistent differences between reference and sample

illumina®

Fundamental task
Alignment of the reads against the reference

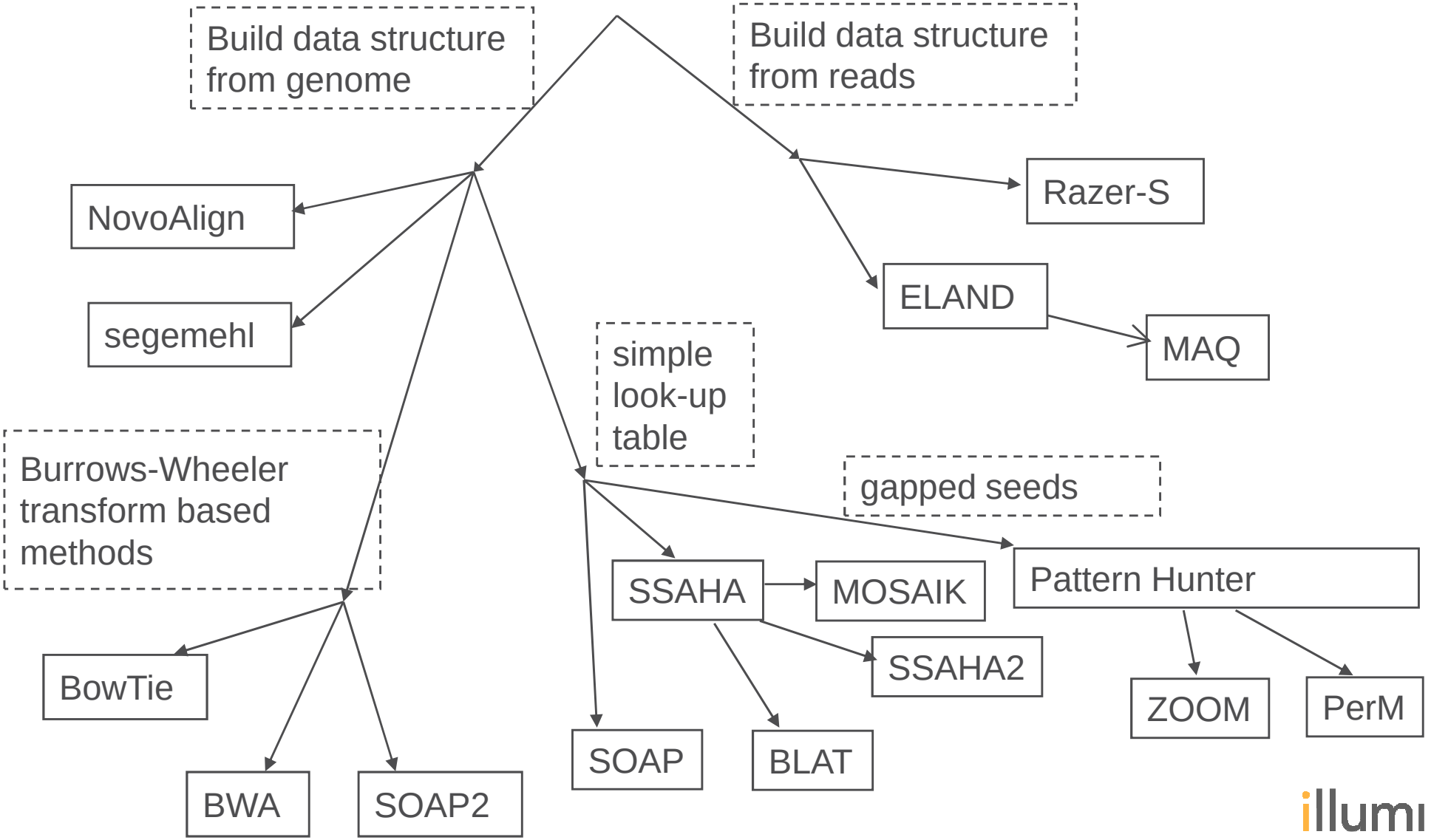
- Need to work reasonably fast for very large number of reads.

  For example, we need to align 1 billion reads (each 100 bases long) against the Human reference (3 billion bases long) in a few hours.

- We can’t afford to use exhaustive dynamic programming algorithms from the beginning.
- First we need a very fast filtering approach (with some kind of indexing) to identify perfect-match candidates.
- Then we can use a more sensitive (and time-consuming) algorithm to work out the local details.

illumina®

Large amount of existing research



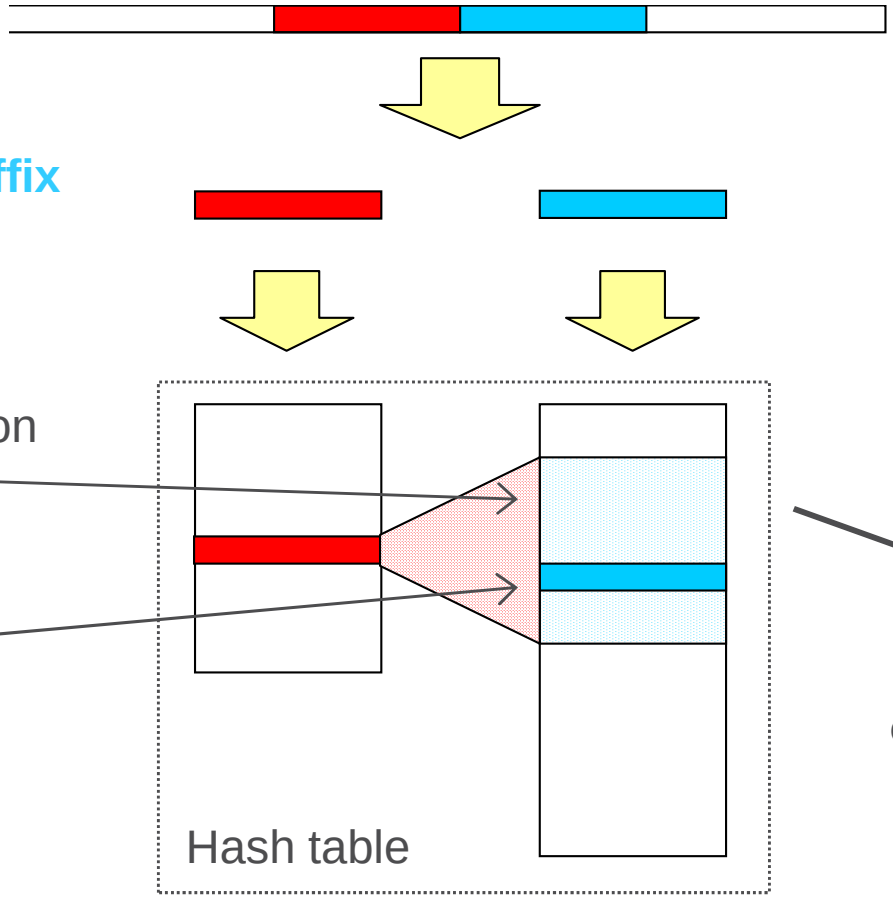
illumina®

Hash-based algorithm to solve multiple exact matching problem (Kim/Kim 1999[1])

Problem: Find all exact occurrences of a set of sequences in the reference genome

For all *k*-mers in genome

Split into **prefix** and **suffix**

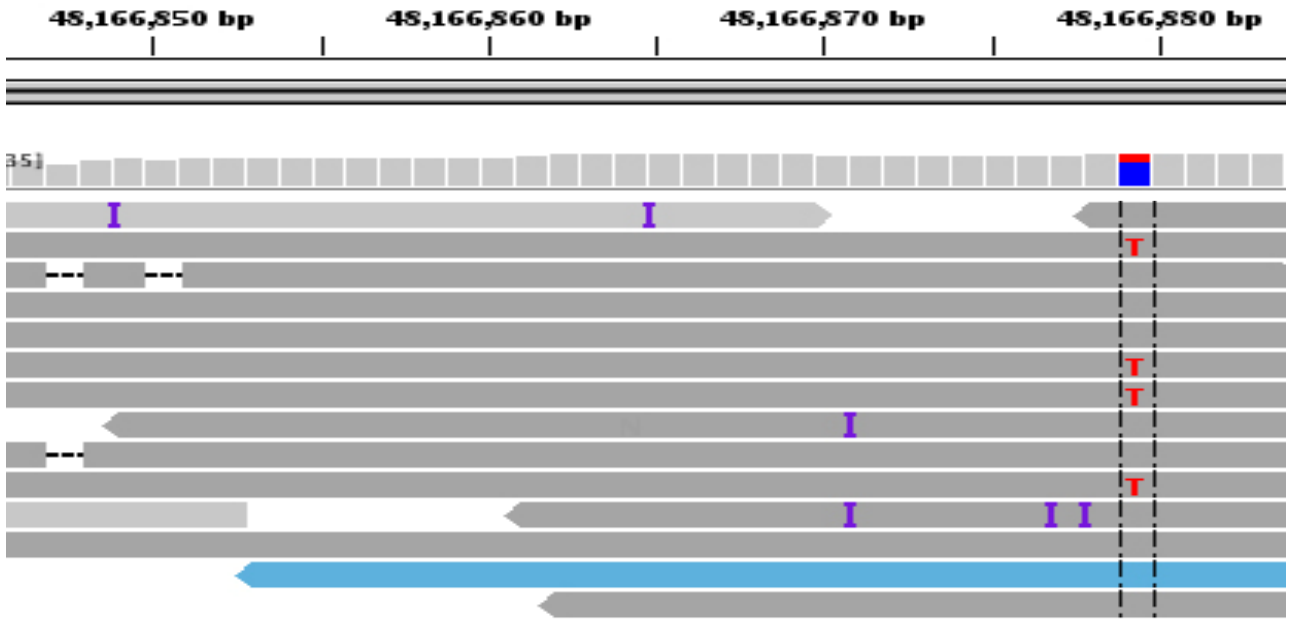


This hashtable is constructed from the reads

[1]*Proc.17th AoM/IAoM International Conference on Computer Science, May 1999.*

illumina®

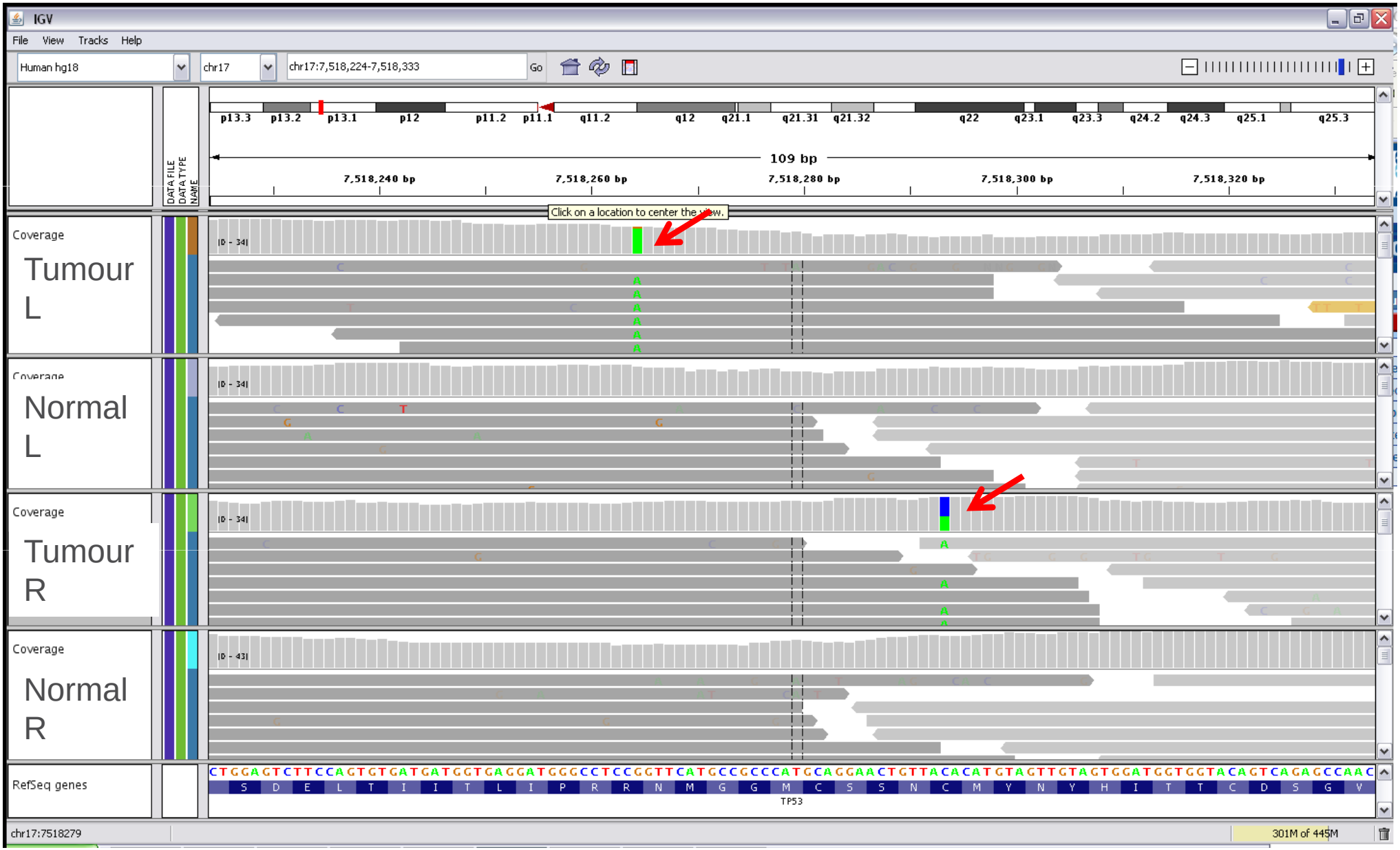
What are we doing with the aligned reads?



- Look for consistent differences between reference and sample

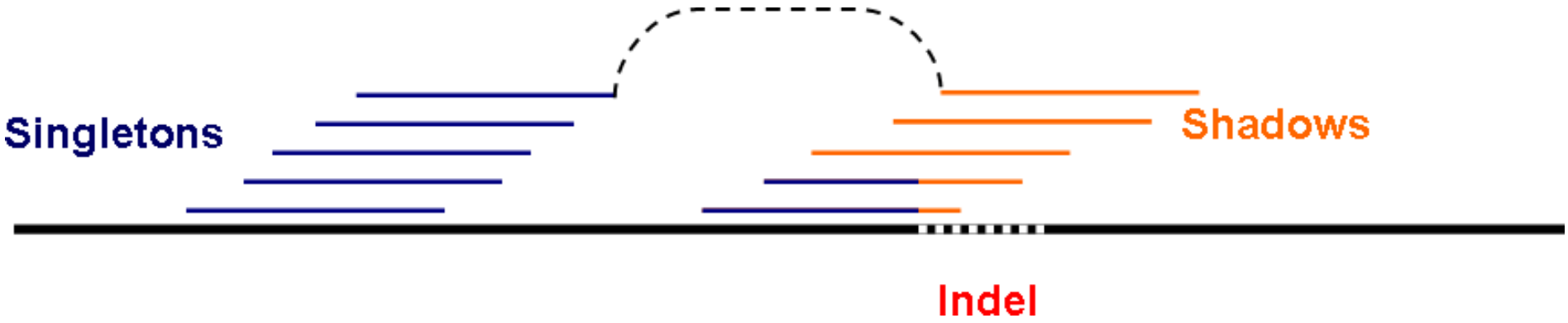
illumina®

I. Single Nucleotide Polymorphisms (SNPs)



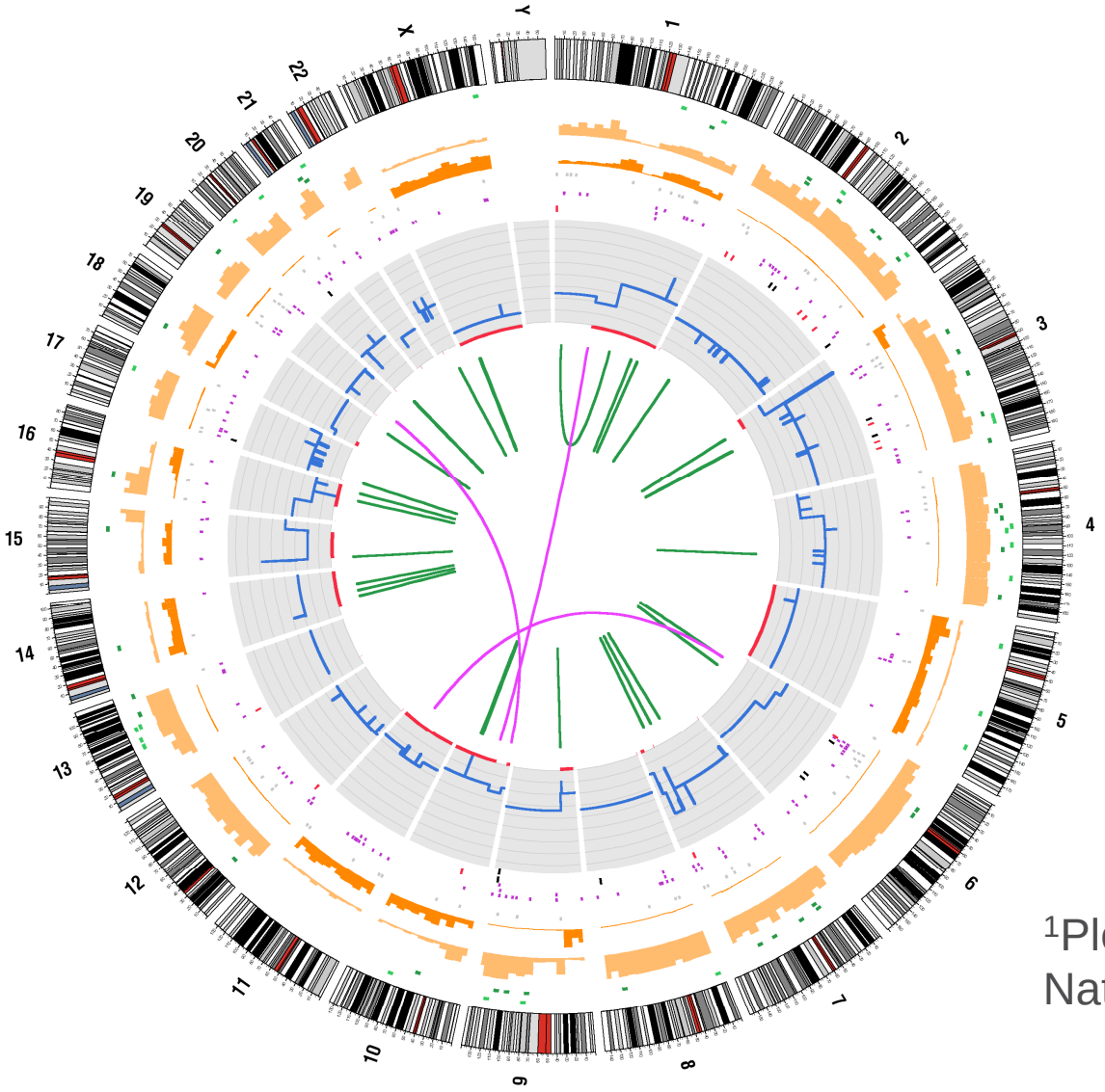
II. Structural variants

- Look for consistent differences between reference and sample beyond the single nucleotide level, for instance: larger insertions/deletions, inversions
- Each variant has a read pair signature
- This is an example for the case of deletions



illumina®

Beyond a single genome: Differences between samples



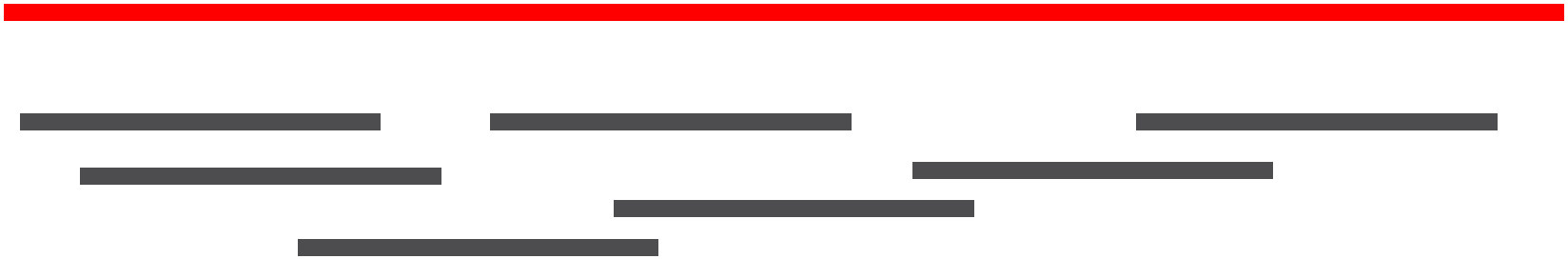
Ideograms / val indels / sn_{10M} / cdsn / cn /loh / sv

- 33,345 Single base substitutions
  - 286 coding
- 1018 small indels
  - 14 coding
- 37 Structural rearrangements
  - 34 intrachromosomal:
    - 25 deletions
    - 6 insertions
    - 2 duplications
    - 1 complex
  - 3 interchromosomal
  - 19 breakpoints in genes
- 198 changes in copy number

[1]Pleasance, Cheetham *et al.* 2009, Nature 463:191-6

illumina®

We got 1 billion reads from the instrument. Now what? ... Application II: De-novo assembly



- Remember that the reads are randomly sampled short sequences across the whole genome.
- De-novo assembly: computational reconstruction of a genome sequence from the short reads.
- Assembly is possible, because if we have high enough coverage, the reads are partially overlapping.
- Outcome of a de-novo assembly: Contiguous reconstructed pieces of the genome.

illumina®

How does it work?

- The most common model for assembly are de Bruijn graphs
- Split reads into overlapping k-mers (k is an adjustable parameter)
- Elegant theoretical model
- Does not deal well with sequencing errors and repeats
- Widely used assemblers: Velvet, SOAPdenovo, AllPaths, ABySS

Nicolaas G. de Bruijn

illumina®

Toy example
read length = 5 and k = 3

AGACTCCTG Unknown genome

- We want to reconstruct the unknown genome from the reads.

illumina®

Toy example
read length = 5 and k = 3

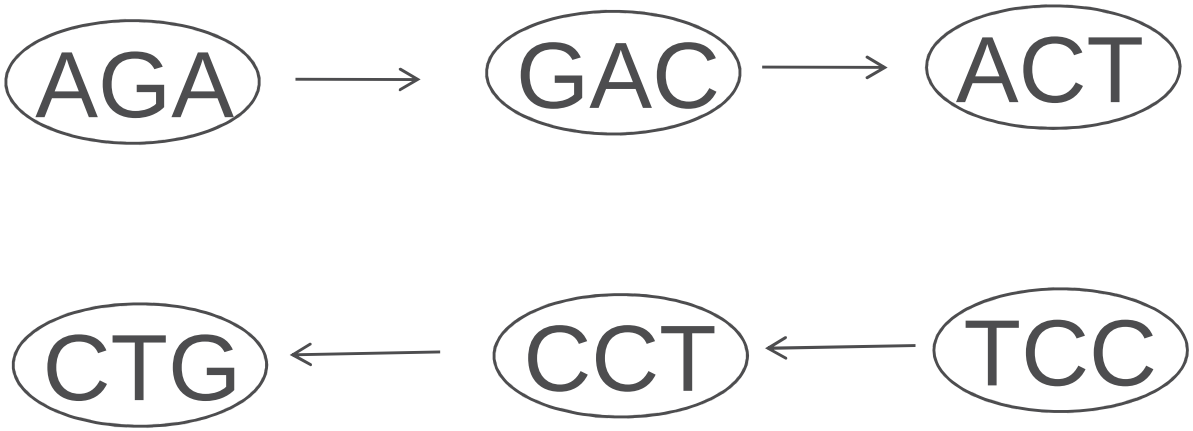
1st read **AGACTCCTG** Unknown genome

- Slide a window of size k = 3 over the read.
- For each k-mer draw a vertex.
- For adjacent k-mers draw an edge.

illumina®

Toy example
read length = 5 and k = 3

2nd read **AGACTCCTG** Unknown genome

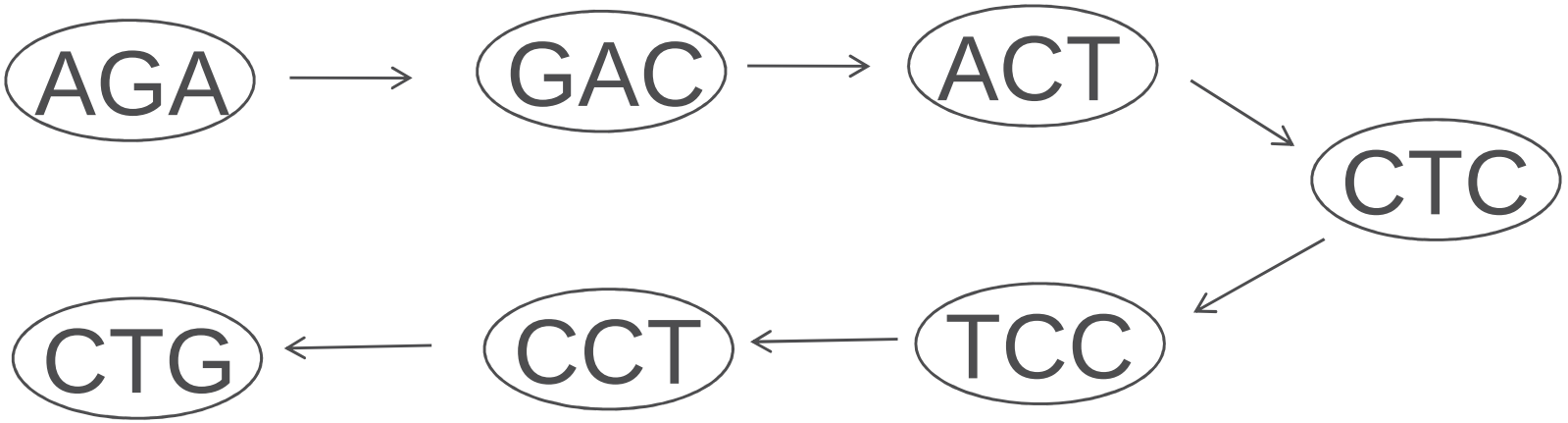


- Slide a window of size k = 3 over the read.
- For each k-mer draw a vertex.
- For adjacent k-mers draw an edge.

illumina®

Toy example
read length = 5 and k = 3

3rd read **AGACTCCTG** Unknown genome

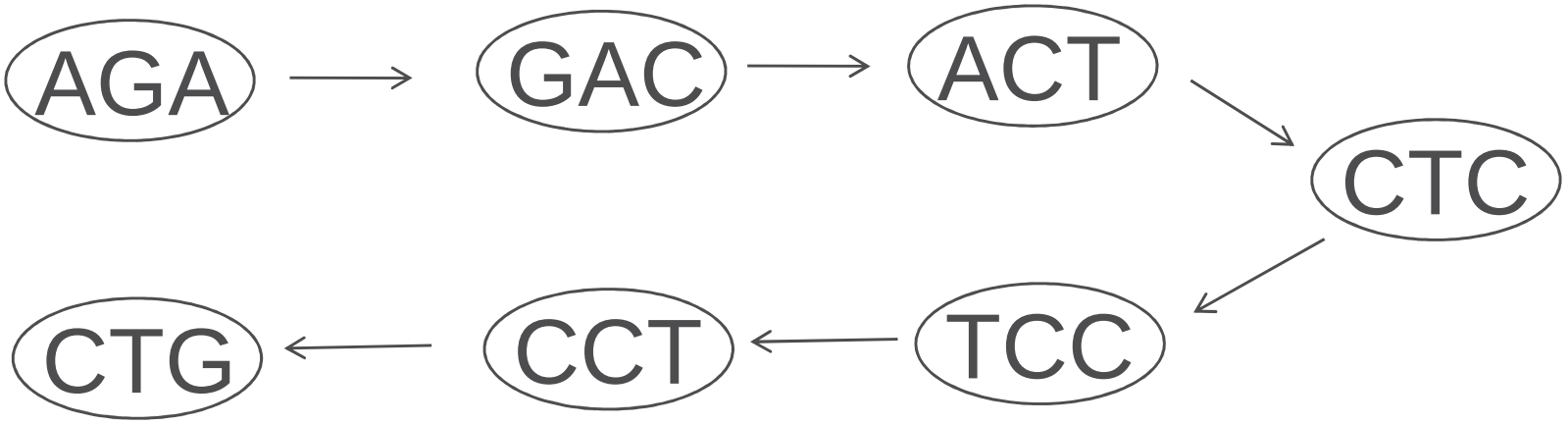


- Slide a window of size k = 3 over the read.
- For each k-mer draw a vertex.
- For adjacent k-mers draw an edge.

illumina®

Toy example
read length = 5 and k = 3

AGACTCCTG Unknown genome

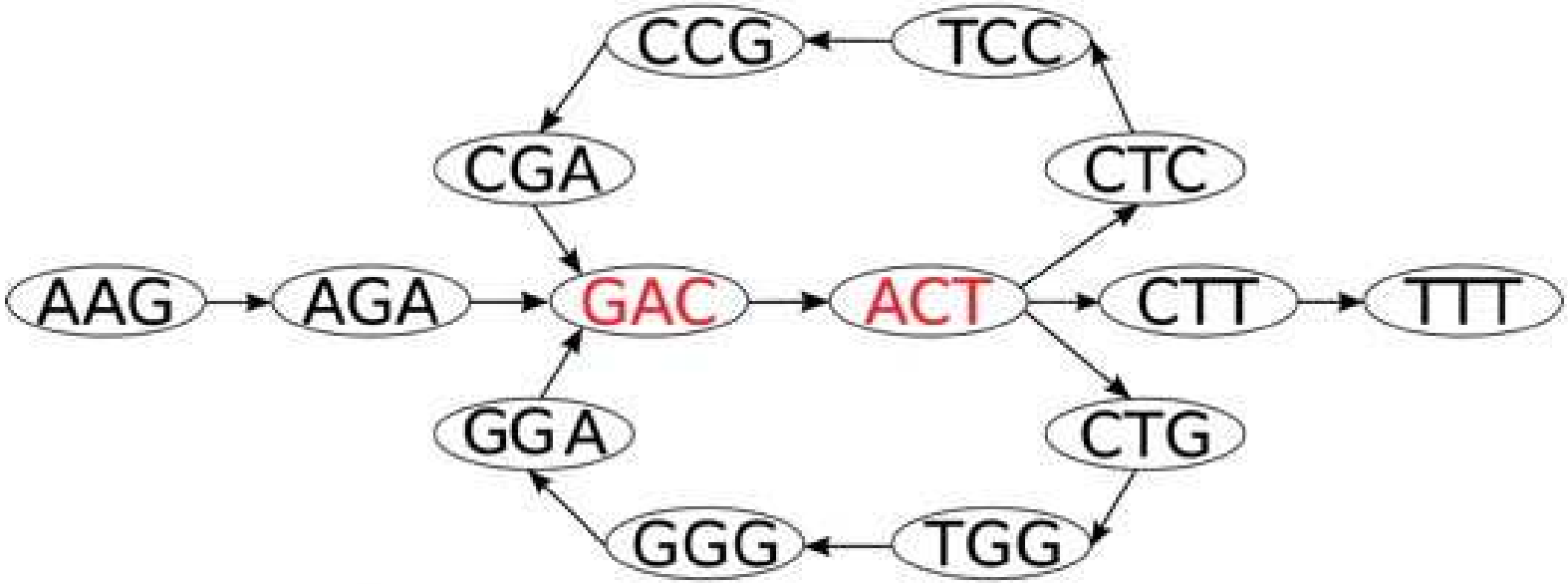


- Remember that the graph was constructed from the reads.
- Now, we can uncover the unknown genome by walking along the graph.

illumina®

Real life is cruel: Repeats introduce ambiguities

AAGACTCCGACTGGGACTTT

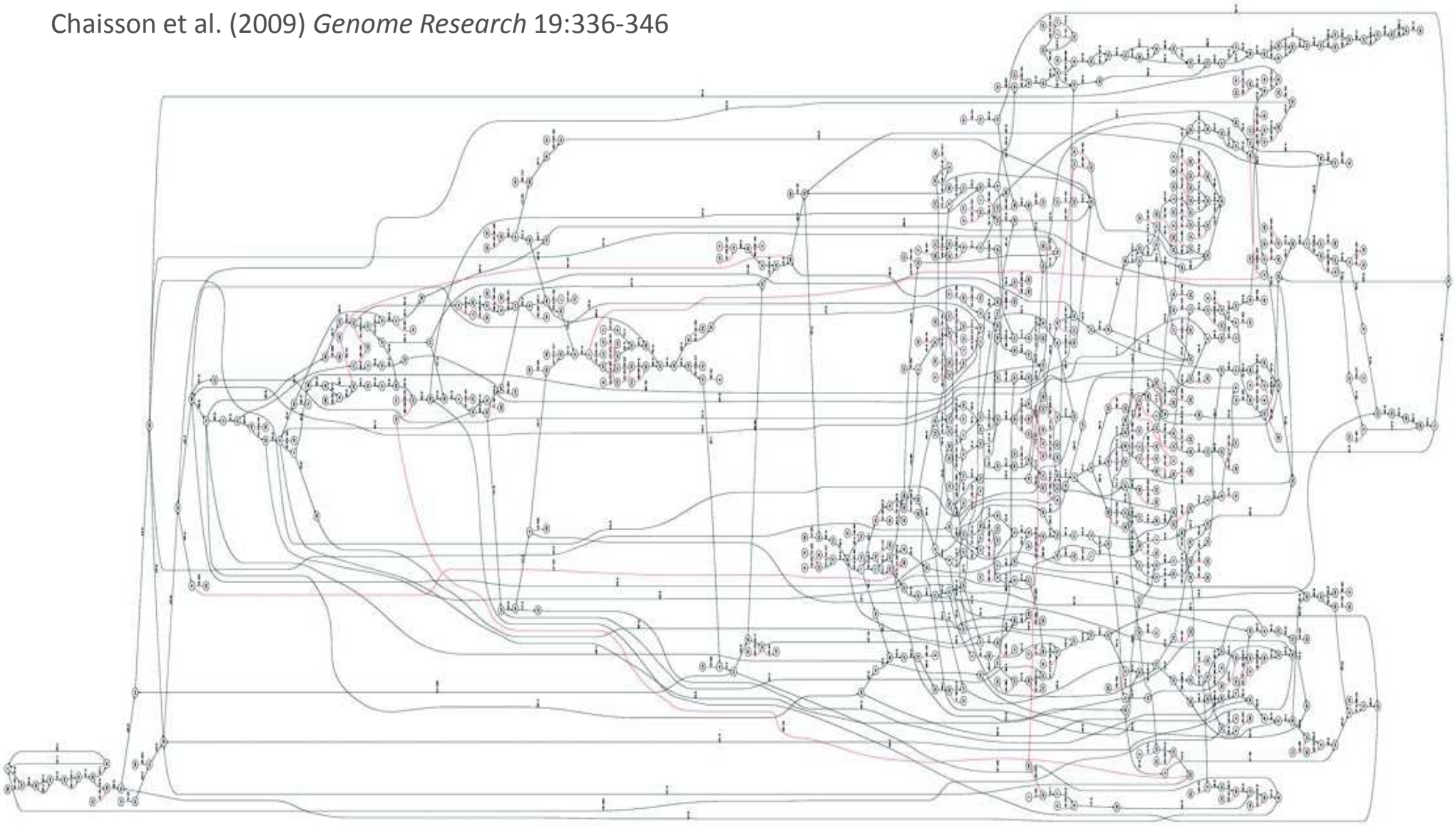


Chaisson et al. (2009) *Genome Research* 19:336-346

illumina®

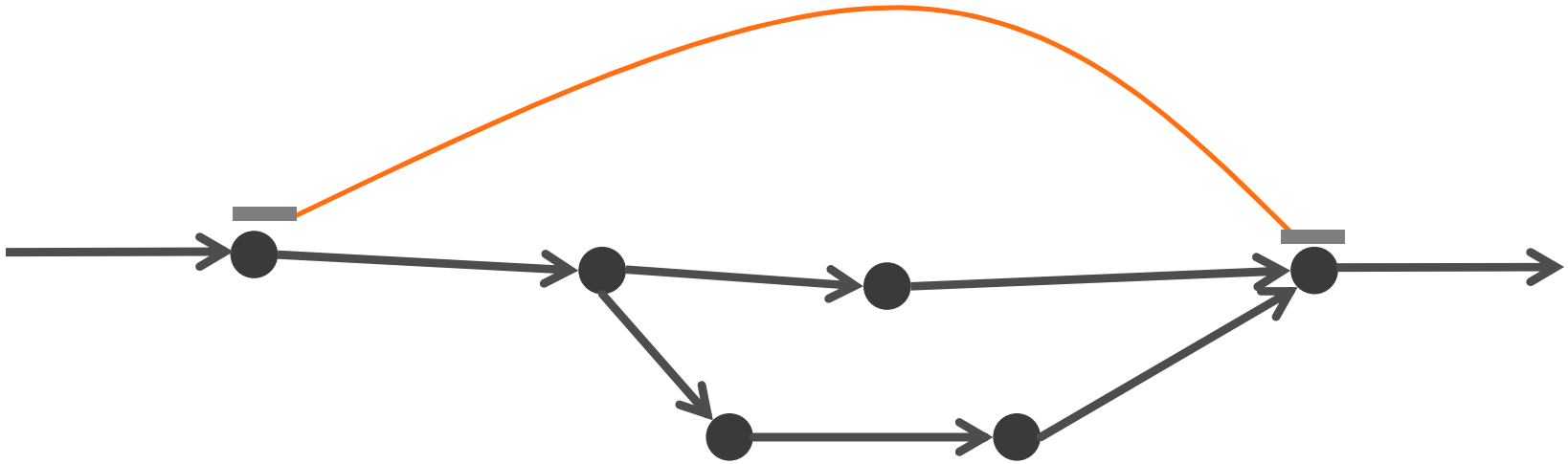
A real de Bruijn graph for E. coli

Chaisson et al. (2009) *Genome Research* 19:336-346



Using read pairs to build scaffolds

This corresponds to mapping (threading) reads onto the graph and joining contigs connected by read pairs into scaffolds.



illumina®

Other applications of high-throughput sequencing

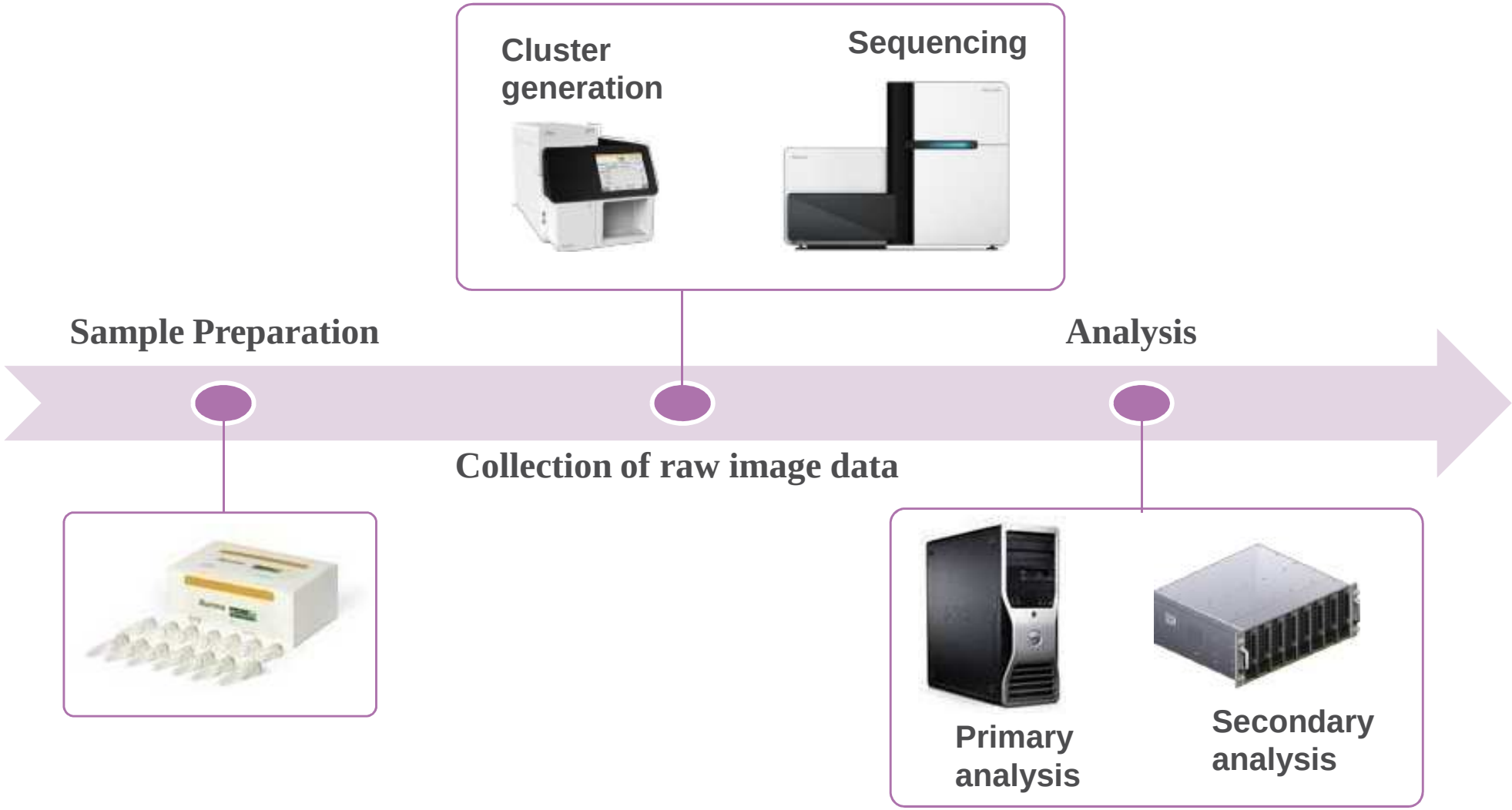
- Whole genome re-sequencing
- *De-novo* assembly
- Targeted sequencing (regions, genes, exomes)
- Whole transcriptome sequencing
- miRNA discovery and profiling
- DNA Methylation
- Histone Modification
- DNA-protein interaction
- ....

illumina®

Some messages to take home

illumina®

The sequencing workflow is a collaborative effort between chemists, physicists, biologists, engineers and computer scientists



Conflicting variables need to be optimized simultaneously

- Library diversity
- Amount of DNA starting material
- Simplicity of sample prep
- Robustness of instrument
- Versatility
- Hands-on time
- Time to result
- Accuracy
- Overall yield
- Yield per day
- Number of reads
- Read length
- Error profile
- Cost per experiment
- Cost per base
- Cost of the instrument

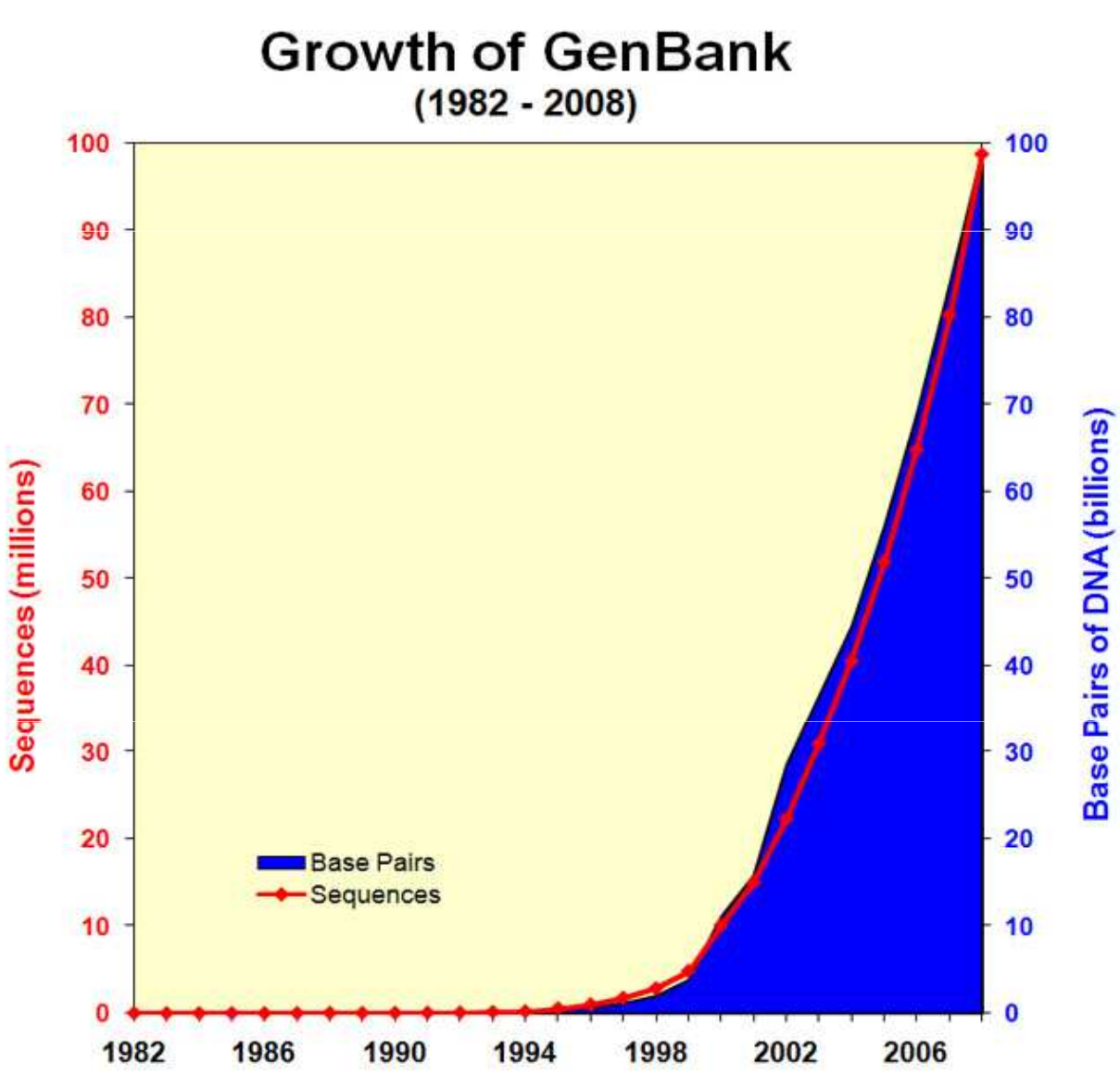
illumina®

Improvement of the instruments

Feature	GA$_{IIx}$	HiSeq2000
Flowcells	1	2
Surface imaging	Single	Double
Read length	2 x 100 (2 x 150)	2 x 100
Yield per run (PF data)	50 Gb (95 Gb)	200 – 350 Gb
Raw Data Quality (v5 chemistry, 2 x 100)	>90% bases are >99.9% accurate	>90% bases are >99.9% accurate
Runtime (2x100)	10 days	8 days
Data Rate	5 Gb / day	25 – 40 Gb / day

illumina®

Increasing data volumes are good news for scientists ...



GenBank Data		
Year	Base Pairs	Sequences
1982	680,338	606
1983	2,274,029	2,427
1984	3,368,765	4,175
1985	5,204,420	5,700
1986	9,615,371	9,978
1987	15,514,776	14,584
1988	23,800,000	20,579
1989	34,762,585	28,791
1990	49,179,285	39,533
1991	71,947,426	55,627
1992	101,008,486	78,608
1993	157,152,442	143,492
1994	217,102,462	215,273
1995	384,939,485	555,694
1996	651,972,984	1,021,211
1997	1,160,300,687	1,765,847
1998	2,008,761,784	2,837,897
1999	3,841,163,011	4,864,570
2000	11,101,066,288	10,106,023
2001	15,849,921,438	14,976,310
2002	28,507,990,166	22,318,883
2003	36,553,368,485	30,968,418
2004	44,575,745,176	40,604,319
2005	56,037,734,462	52,016,762
2006	69,019,290,705	64,893,747
2007	83,874,179,730	80,388,382
2008	99,116,431,942	98,868,465

... but make bioinformaticians struggle like donkeys

illumina®

Many thanks

to Martin for the invitation …

… and to my colleagues for the slides

- Klaus Maisinger
- David Townley
- Markus Bauer
- Ole Schulz-Trieglaff
- Niall Gormley

illumina®