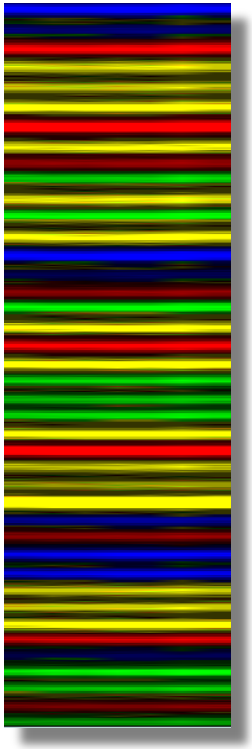
WS2018/2019

„Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik- “

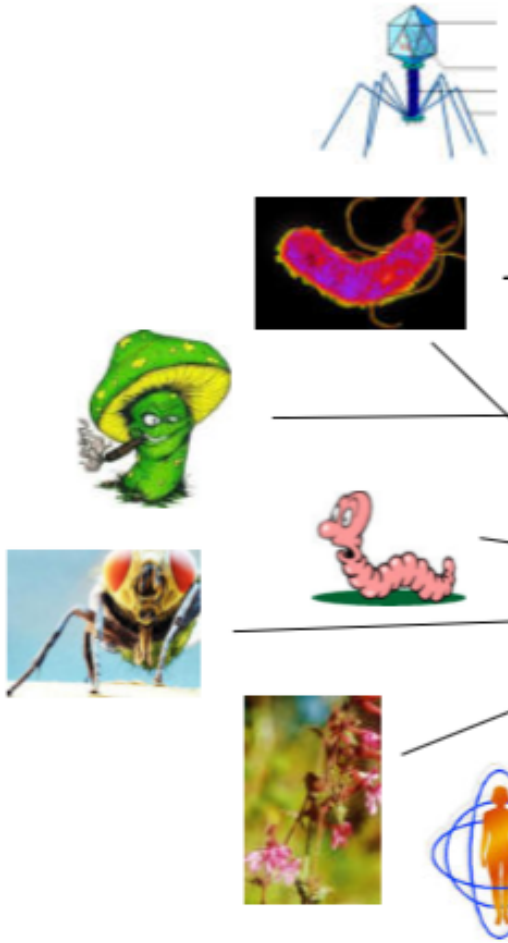
Thomas Hankeln



Methodik der Genomsequenzierung

NGS-Technologie

Meilensteine



Phi X 174	1977	5.386 bp
λ– Phage	1982	48.502 bp
M. genitalium	1995	580.000 bp
H. influenzae	1995	1.830.000 bp
M. jannaschii	1996	1.660.000 bp
S. cerevisae	1997	12.500.000 bp
E. coli	1997	4.654.000 bp
C. elegans	1998	97.000.000 bp
D. melanog.	1999	116.000.000 bp
A. thaliana	2000	115.000.000 bp
H. sapiens	2001	2.693.000.000 bp

2012 > 1000genomes project (human)

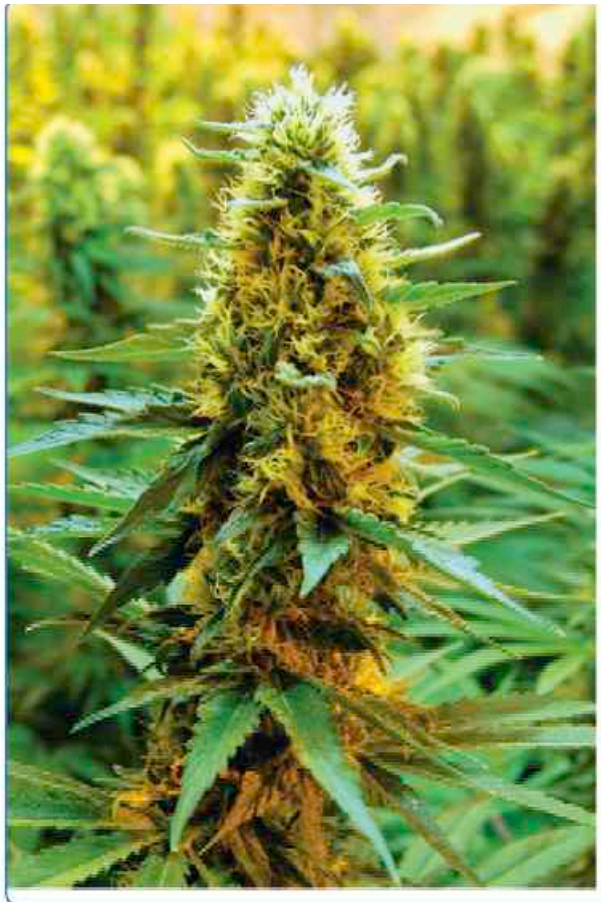
soon > Genome 10K (vertebrates)

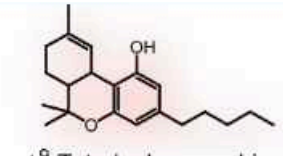
05.11.13 | **Erbgutanalyse**

Liegt das Talent für Mathematik in den Genen?

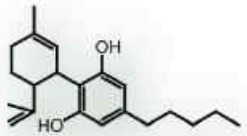
Ist ein Talent für Mathematik angeboren oder liegt es an der Erziehung? Ein US-Unternehmer will die Frage nun klären: Beim "Projekt Einstein" wird das Erbgut von 400 Mathematik-Genies analysiert. *Von Norbert Lossau*

How hemp got high: Cannabis genome sequenced

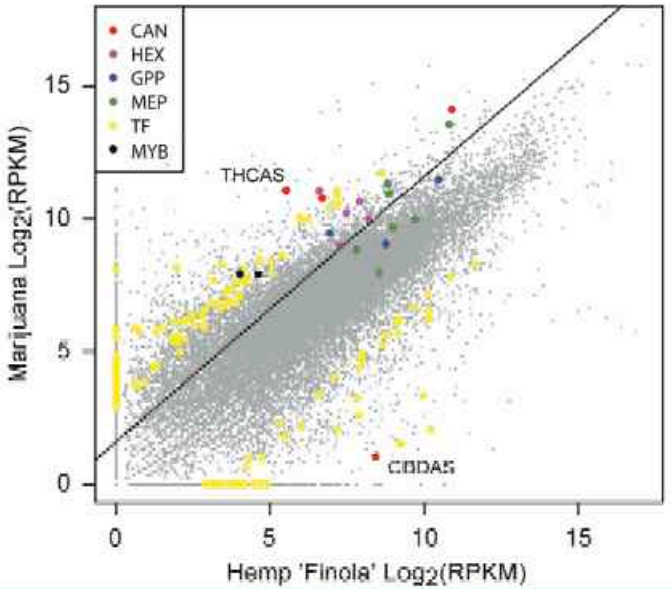




Δ^9-Tetrahydrocannabinol



Cannabidiol



van Bakel et al., Genome Biology 2011

Celebrity genomics

Science confirms the Neanderthal in Ozzy Osbourne

Among the findings that will presented Friday: Osbourne's genes reveal a probability of alcohol dependency "six times higher than the average person."

"All this is big news for blokes everywhere, I think: If the Neanderthals could get laid, there's hope for us all."

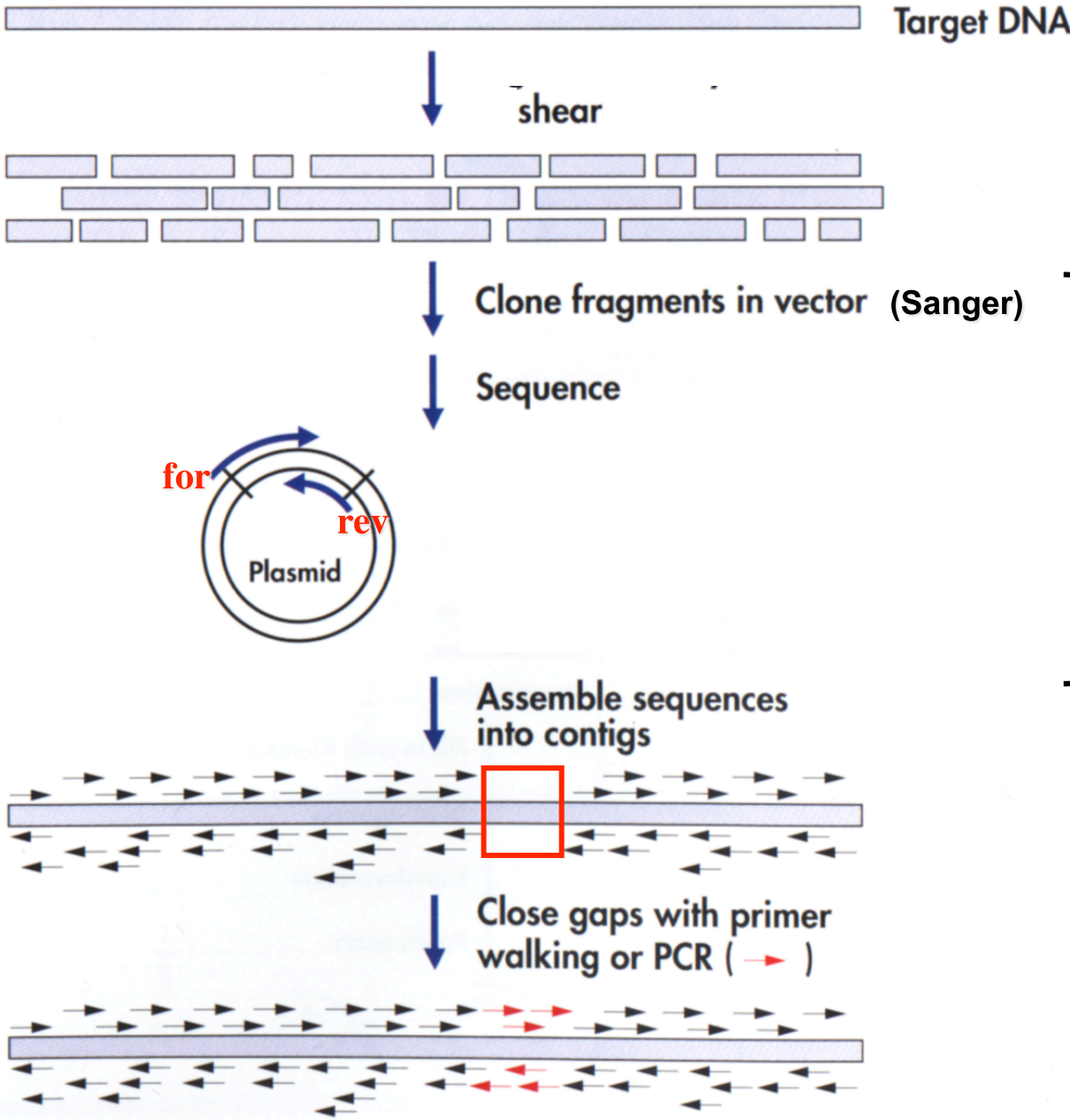
Methodik der Genomsequenzierung

- zu sequenzierende DNA muss zunächst in „handliche“ Abschnitte zerlegt und dann vermehrt werden

  (beim Sanger-Verfahren durch Klonierung oder bei Next-Generation Sequencing (NGS) meist durch PCR; es gibt aber auch schon NGS-Verfahren, die ohne Vermehrung einen einzelnen DNA-Strang sequenzieren)

- DNA kann mit dem klassischen Sanger-Verfahren in Teilabschnitten von 600-1000 Bp („long reads“) sequenziert werden.

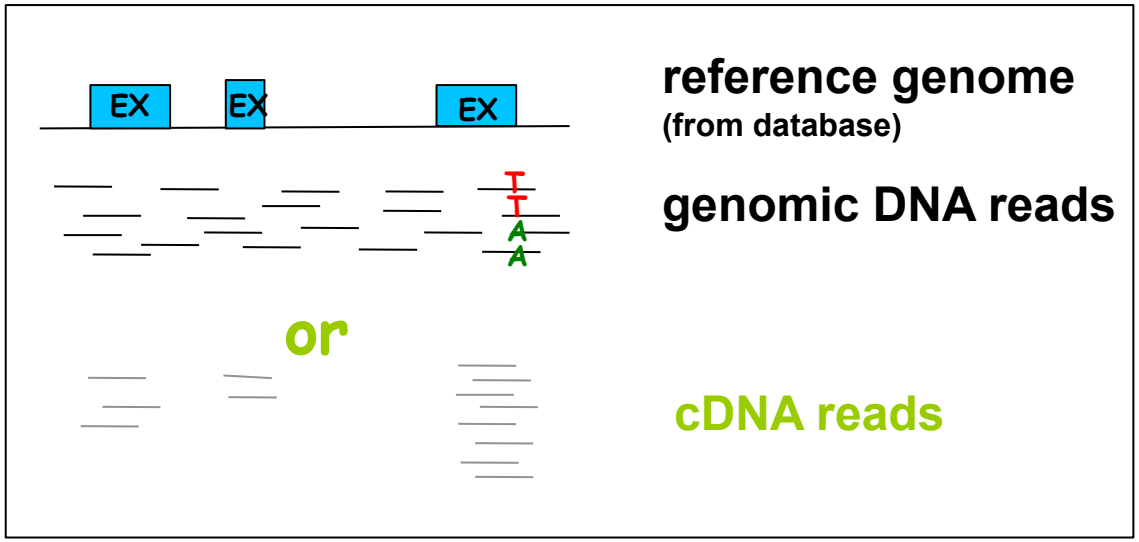
- NGS-Verfahren lesen derzeit meist kurz (z.B. 50 bis 250 Bp; = short reads“), aber auch mittlerweile sehr lang (>> 5000 Bp; long reads)



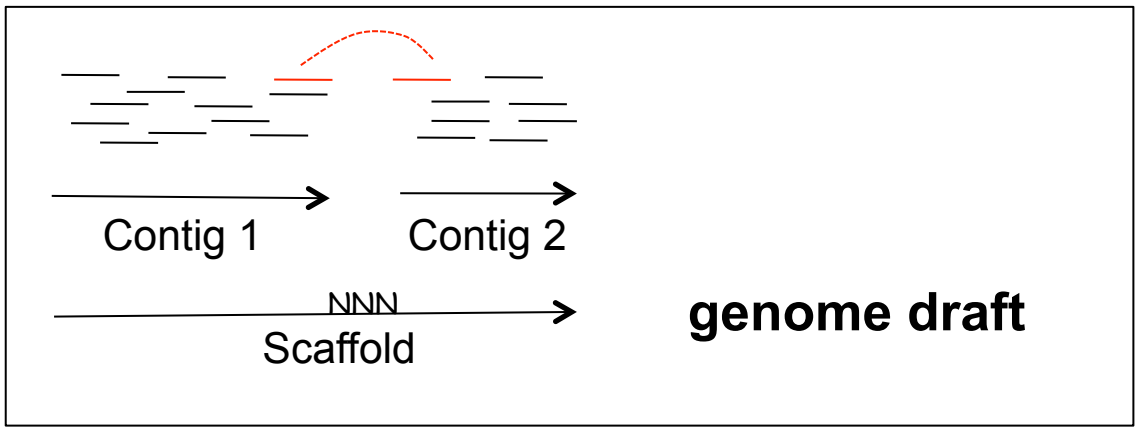
no cloning, but PCR and sequencing in NGS

Shotgun-Strategie

Genome sequencing:
Re- or *de novo*



- variant detection
- differential gene expression (RNA-Seq)

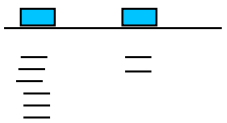


- de novo genome
- gene discovery
- genome evolution

Re- vs. de novo-sequencing

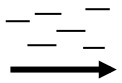
...require completely different bioinformatics algorithms

- Re-Seq



> short reads ok
> very high redundancy (SNP detection!)
> no assembly, but **alignment to existing reference sequence (*„Mapping“*)**
> possible on normal PCs (64bit, quadcore, 16 GB RAM)

- de novo



> longer reads much better (due to repeats)
> **assembly** (OLC, de Bruijn graph)
> extremely RAM intensive (HPC needed)

Martin JA, Wang Z, Nature Reviews Genetics 2011

NGS-Bioinformatik

Applikationen von Re-Sequencing

- Wo ist unser Genom variabel? Wo sind Mutationen?
  > Gesamtgenom-Resequenzierung

- Wo sind unsere Exons variabel? Mutationen?
  > Exome-Seq

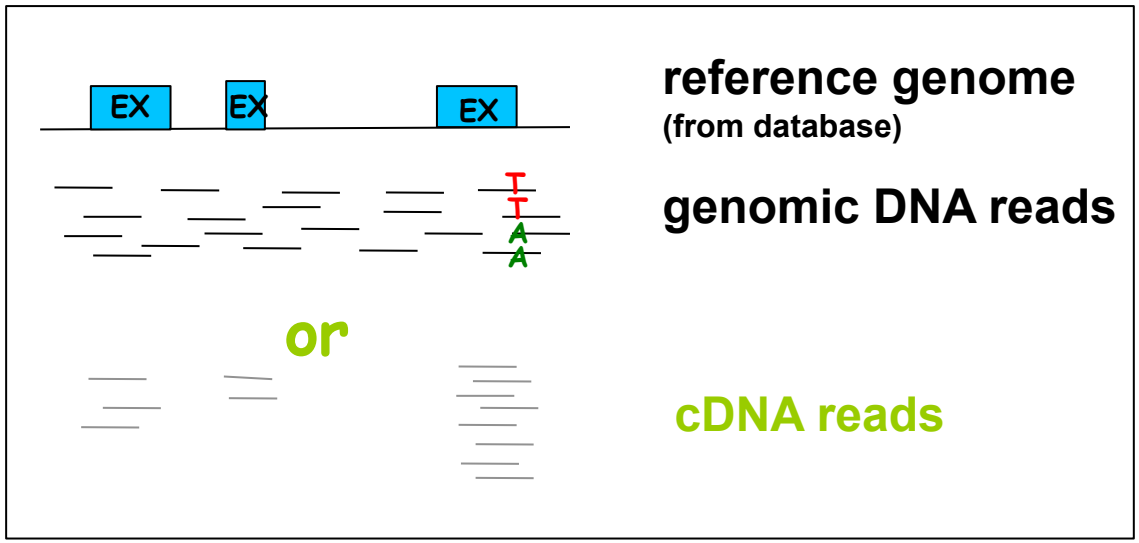
- Welche Gene werden transkribiert? Wie stark?
  > RNA-Seq

- Welchen Chromatin-Status hat das Genom?
  > ChIP-Seq (u.v. a.)

- Welchen Methylierungsstatus hat das Genom?
  > Methyl-Seq, Bisulfite-Seq (u.a.)

...entdecke die Möglichkeiten!

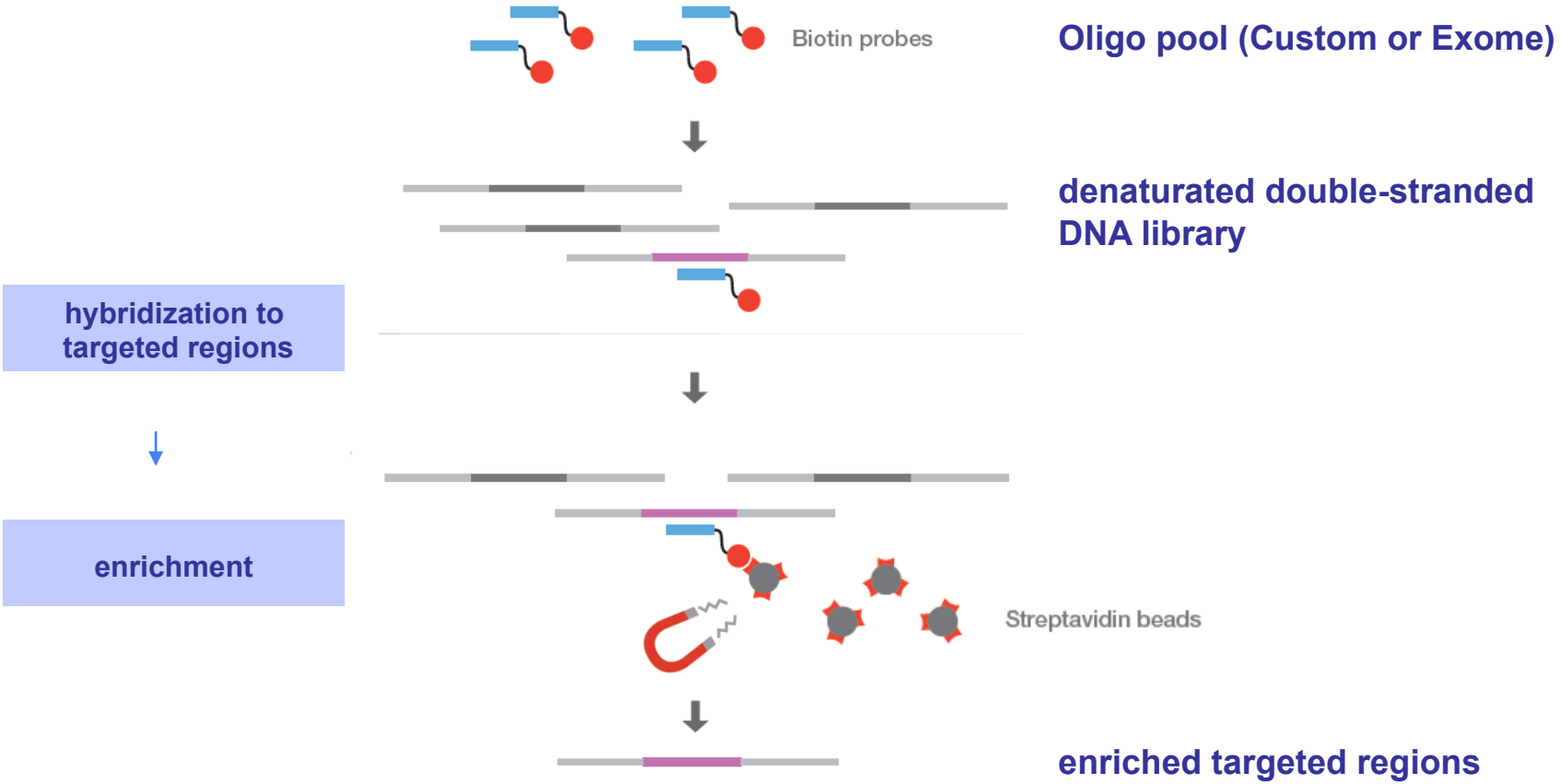
Gemeinsames Prinzip bei Re-Sequencing



Wichtigster Schritt: Zuordnung der Reads zu den passenden Regionen im Genom („Mapping“)

Exome-Seq

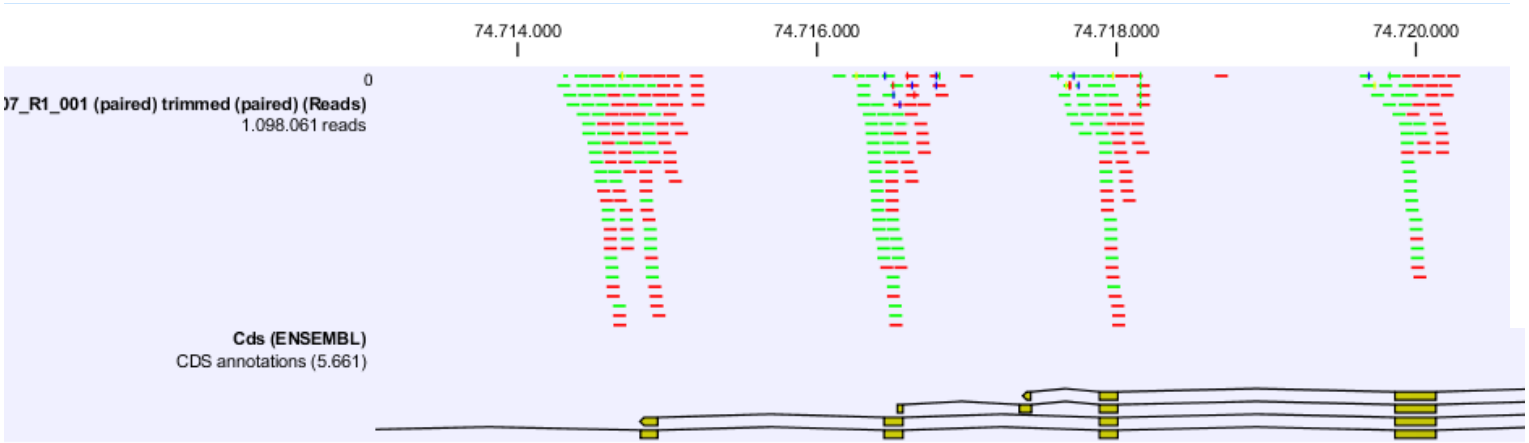
- **requires *a priori* knowledge of target to be sequenced** (e.g. exons)
- **target enrichment by hybridization**



Exome-Seq

Exome Enrichment:

Target region size	62 Mb
Number of target genes	20,794
Number of target exons	201,121
Number of probes	340,427



Exome-Seq

Ergebnis der Exome-Seq ist eine Liste von Nt-Veränderungen, durch die sich das sequenzierte Exom von einem Referenzgenom unterscheidet.

SNVs : single nucleotide variants

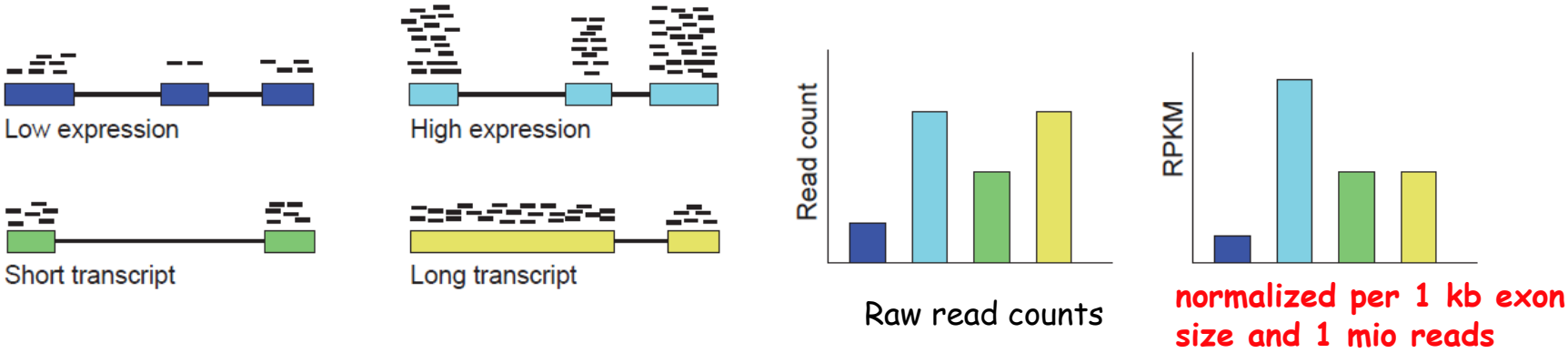
SNPs : single nucleotide polymorphisms*

cSNPs : coding SNPs

* mind. zu 1% als Allel in Human-Population vorhanden

Transcriptome analysis (RNA-Seq)

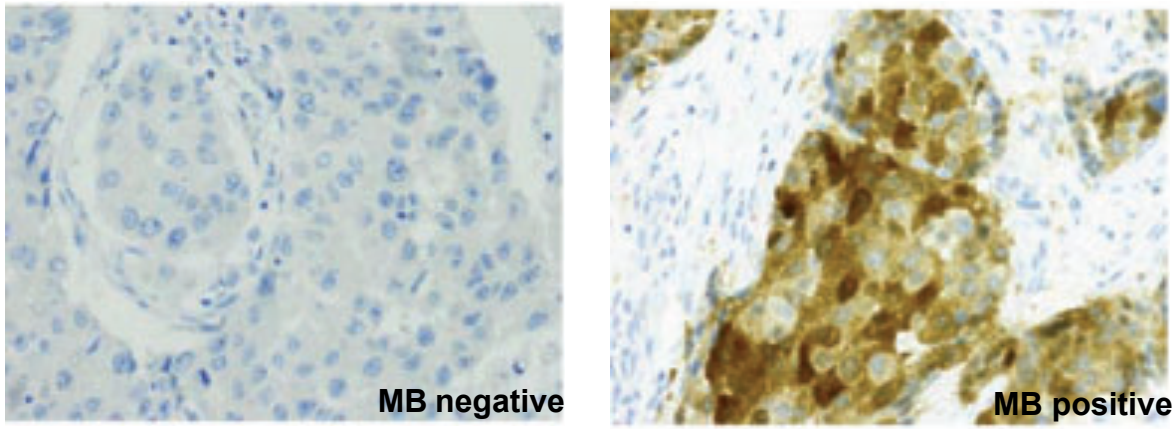
- RNA isolation > cDNA production & fragmentation
- high-throughput sequencing
- mapping of fragments to reference genome
- read-counting as measure of gene transcription



A typical measure is “RPKM“
= reads per Kb of exon sequence and 1 mio reads in the dataset

Beispiel im Kurs:
Was macht Myoglobin in Brustkrebs-Zellen?

MB immunostaining on breast tumors



RNA-Seq

Differenziell exprimierte Gene
= Hinweis auf molekulare Veränderungen

Gorr et al, 2011

Transcriptome analysis (RNA-Seq)

Ergebnis der RNA-Seq ist eine Liste von Genen, die zwischen den betrachteten Datensätzen statistisch signifikant differenziell reguliert sind.

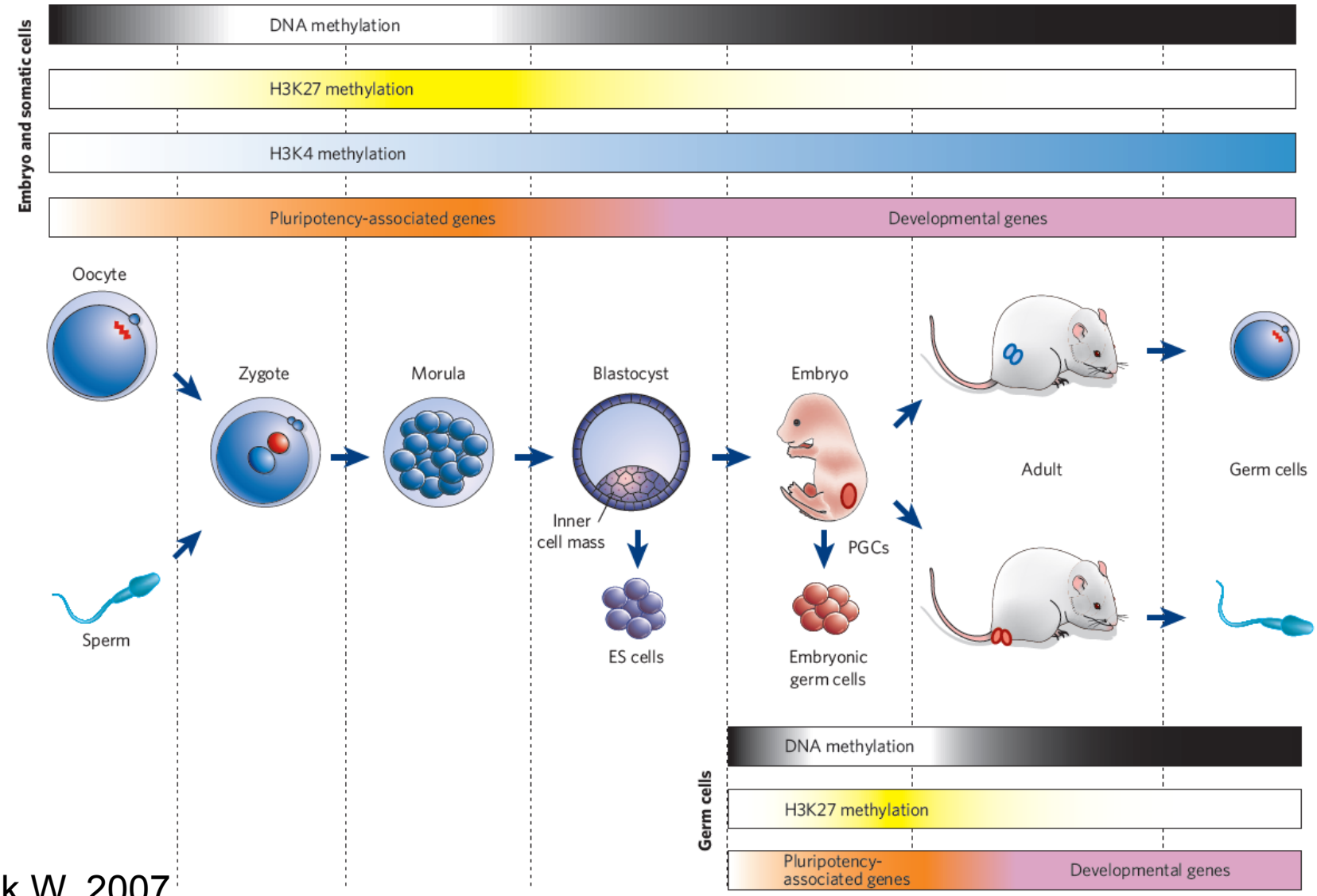
Manchmal enthalten diese Listen Hunderte oder Tausende von Genen.

Dies macht es erforderlich, die Gene funktionell zu kategorisieren (GeneOntology-Vokabular, KEGG Pathways).

Sind bestimmte funktionelle Kategorien in den differenziell regulierten Genen über- oder unterrepräsentiert?

Was sagt das über die „Biologie“ der verglichenen Proben?

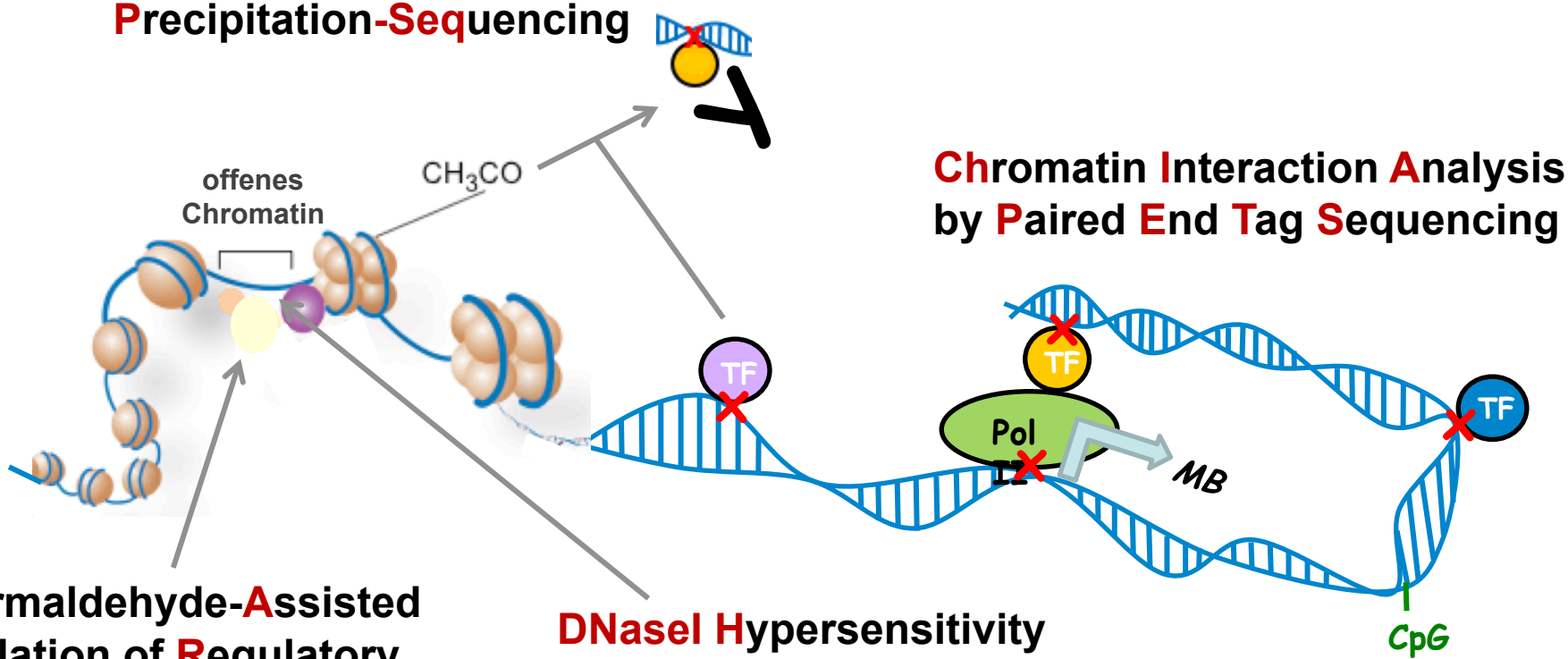
Epigenetische Genommodifikationen



Reik W, 2007

Genomweite Detektion von Chromatin-Modifikationen durch NGS-Methoden

Chromatin Immuno-Precipitation-Sequencing



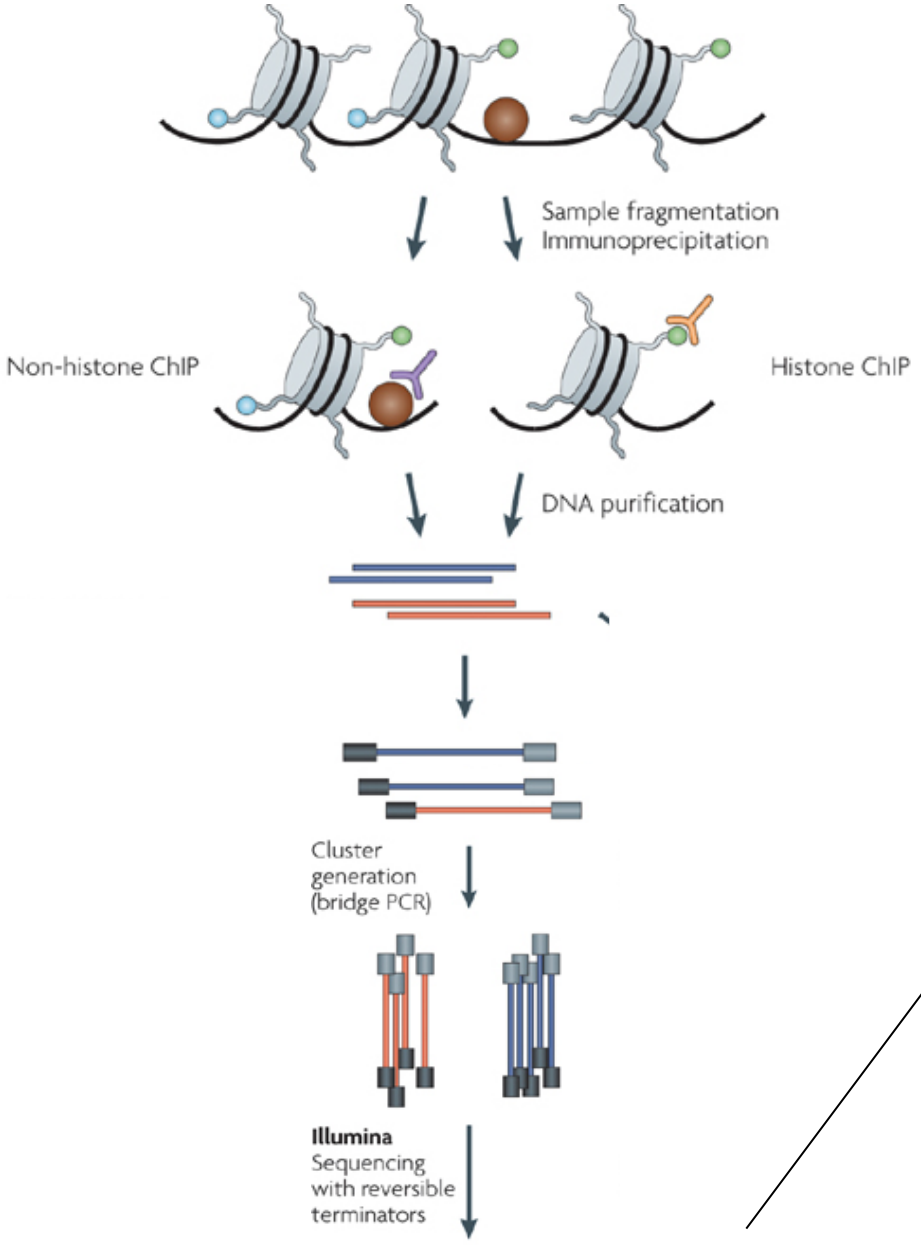
Chromatin Interaction Analysis by Paired End Tag Sequencing

Formaldehyde-Assisted Isolation of Regulatory Elements-Sequencing

DNaseI Hypersensitivity Sites-Sequencing

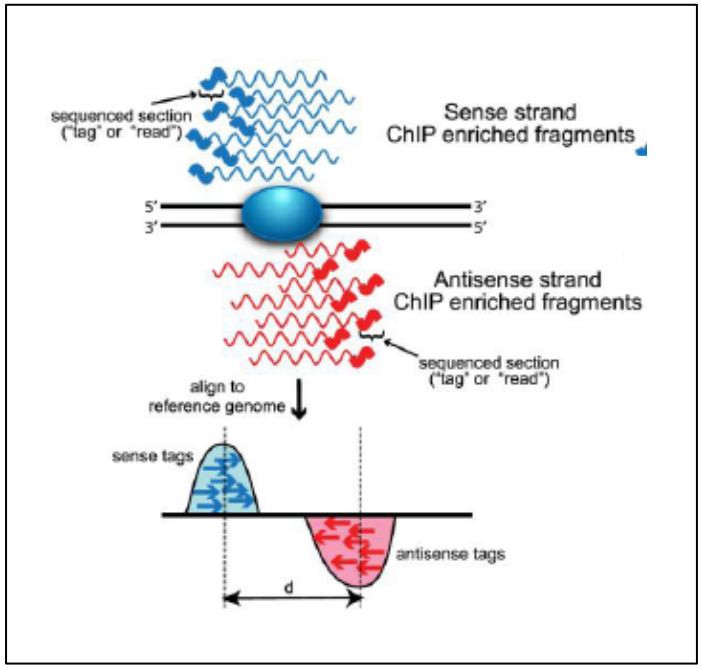
Bisulfite-Sequencing

ChIP-Seq



Reads > Mapping to reference genome

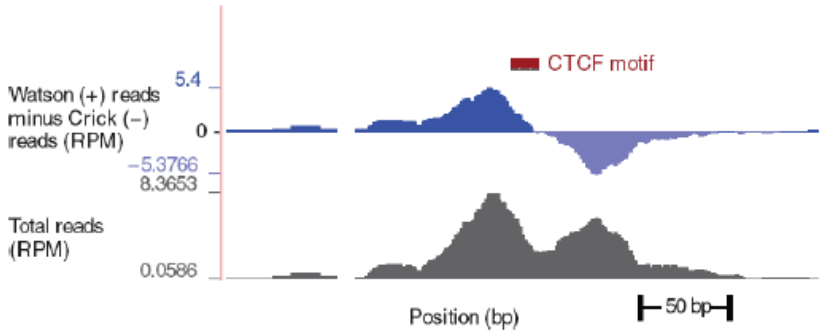
„peak calling“



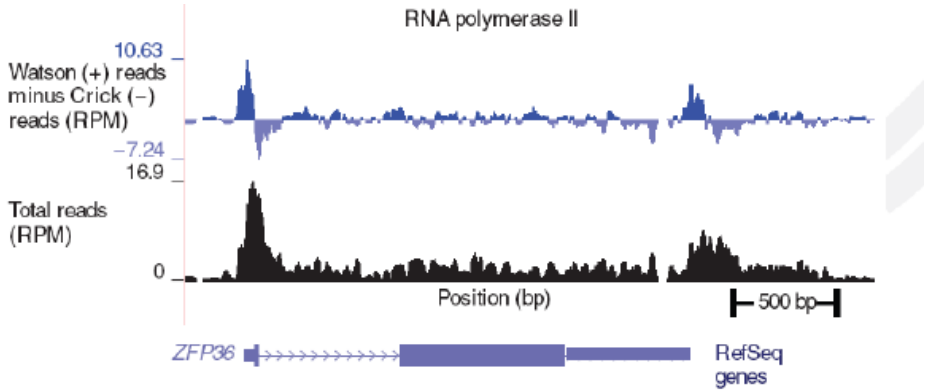
➢peaks zeigen Orte, wo TF gebunden hat

Nature Reviews | Genetics

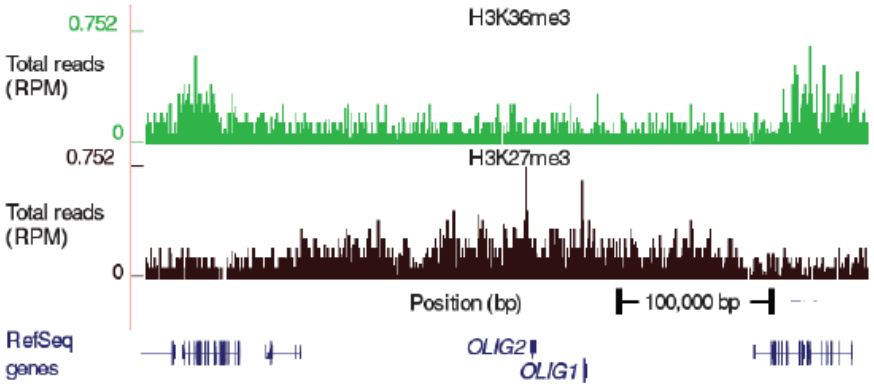
Chip-Seq: Ergebnisbeispiele



Trancriptional regulation protein (CTCF)



RNA polymerase II



Epigenetics: Repressive mark for H3K27me

ChIP Seq peak types from different experiments

Computation for ChIP-seq and RNA-seq studies
Shirley Pepke, Barbara Wold & Ali Mortazavi
Nature Methods 6, **S22 - S32 (2009)**

NGS-Kursprogramm Modul 7A

- Trio-Analyse:

Identifizierung von Krankheitsgenen durch **Exome-Seq** von Familien (Tag 1 & 2)

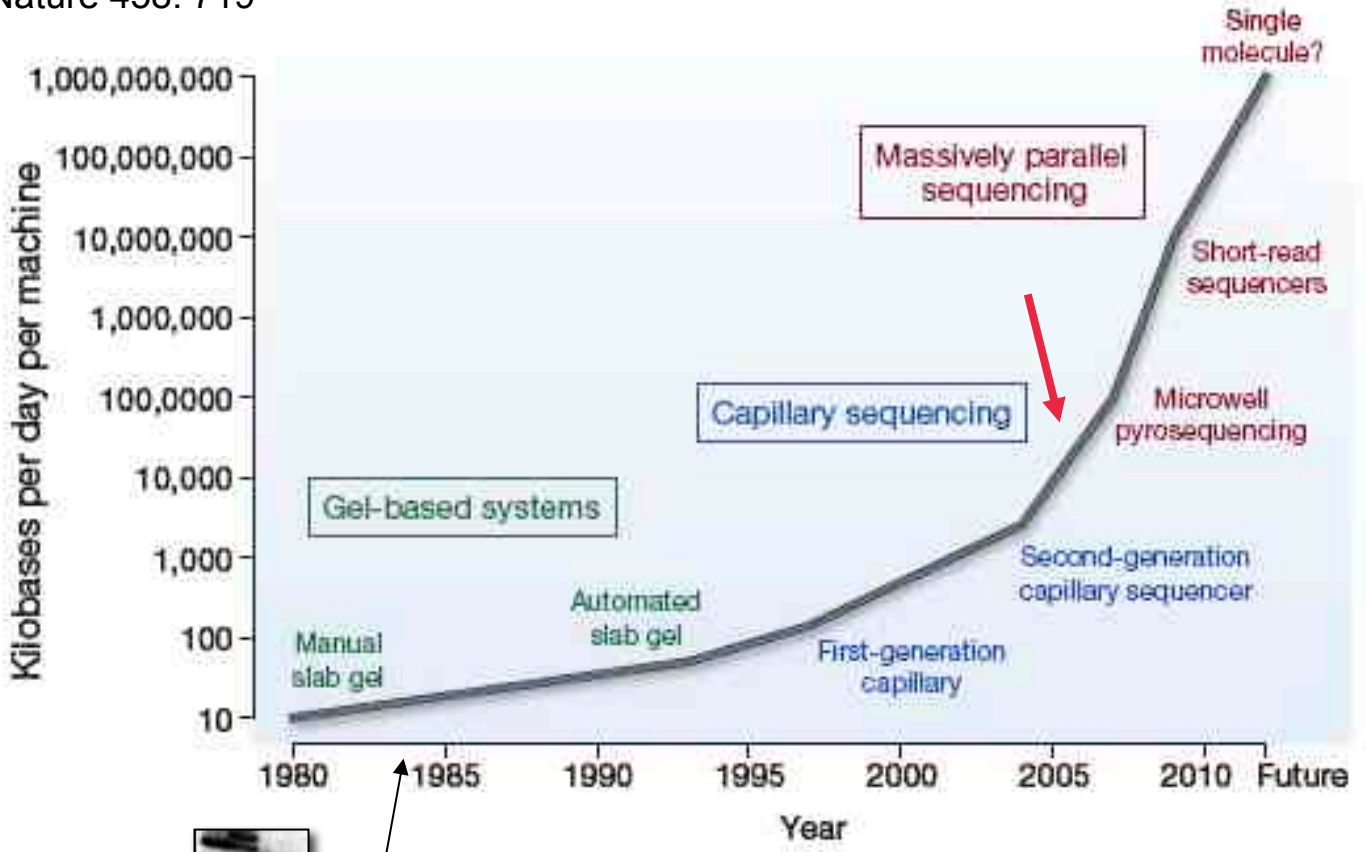
- Transkriptomanalyse:

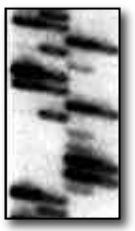
Der Einfluss der Myoglobinexpression auf das Gesamt-Transkriptom von Krebszellen (**RNA-Seq**; Tag 3 & 4)

Q: Wie funktioniert NGS technisch?

Sequencing technology:
A million-fold improvement!

Nature 458: 719





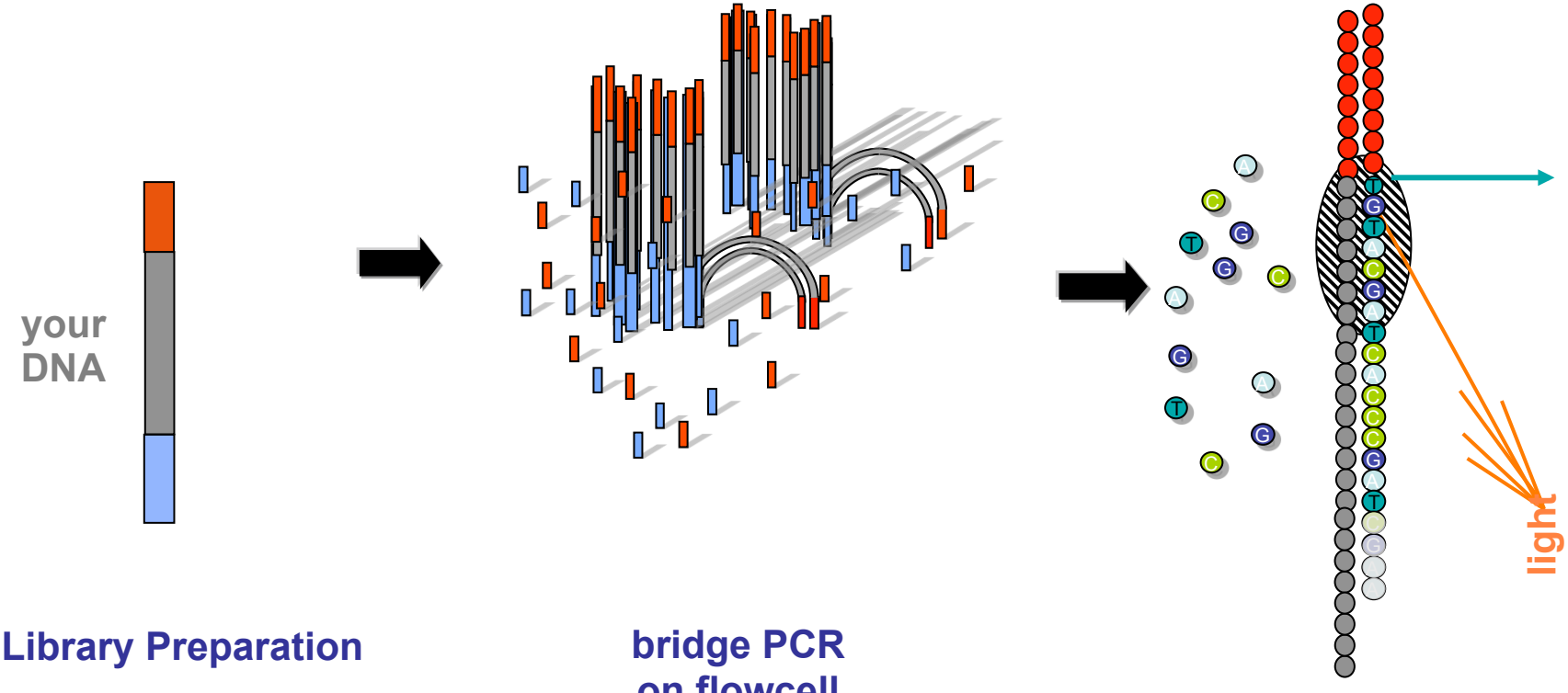
my diploma thesis: 1kb Maxam-Gilbert, 4 weeks (day & night in the isotope lab)

NGS technology: How to...

tedious cloning
high chemical costs
slow electrophoresis

PCR or even single molecules
extreme miniaturisation
massively-parallel read-out

Illumina Sequencing - principle



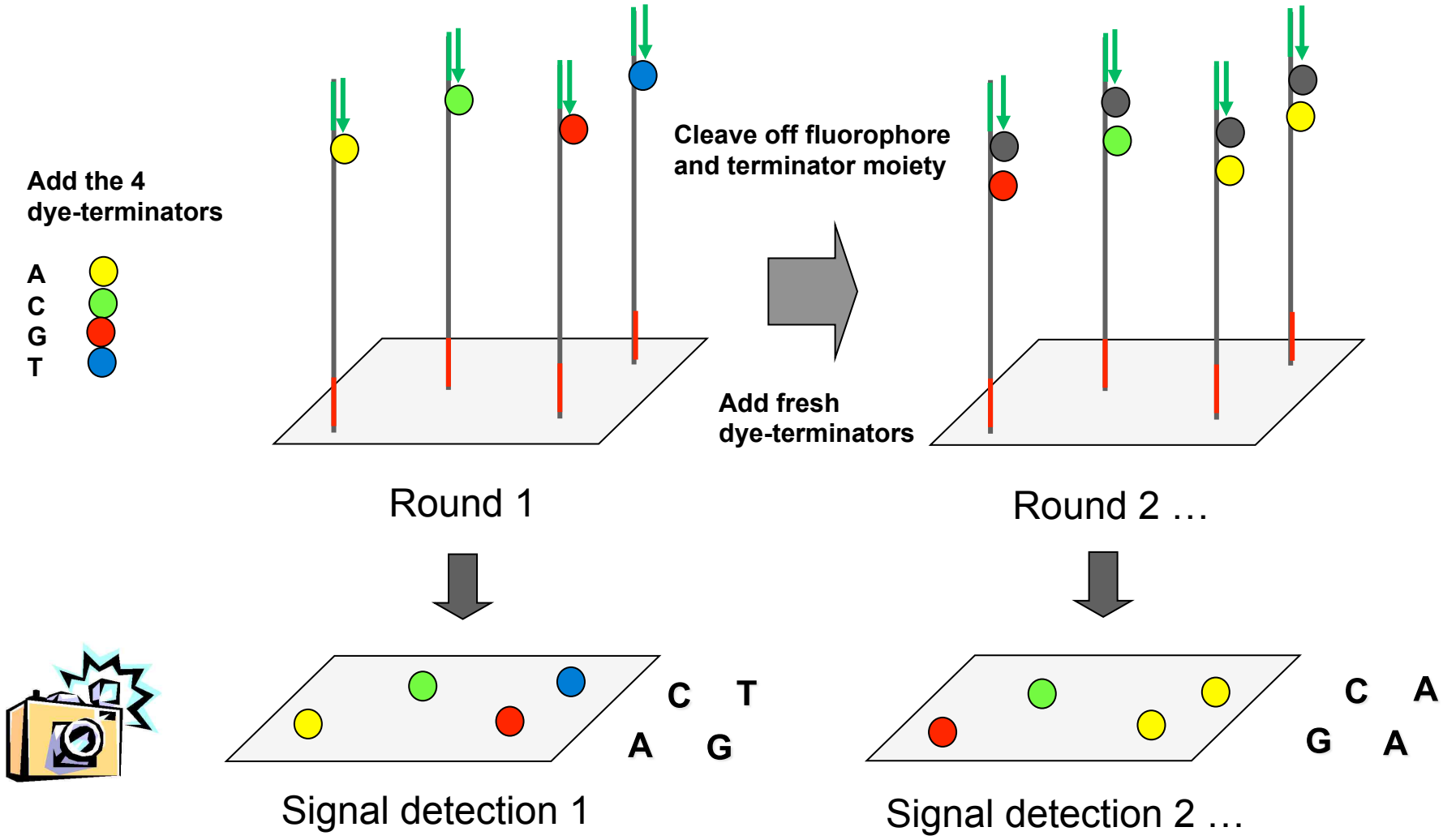
Library Preparation

bridge PCR
on flowcell

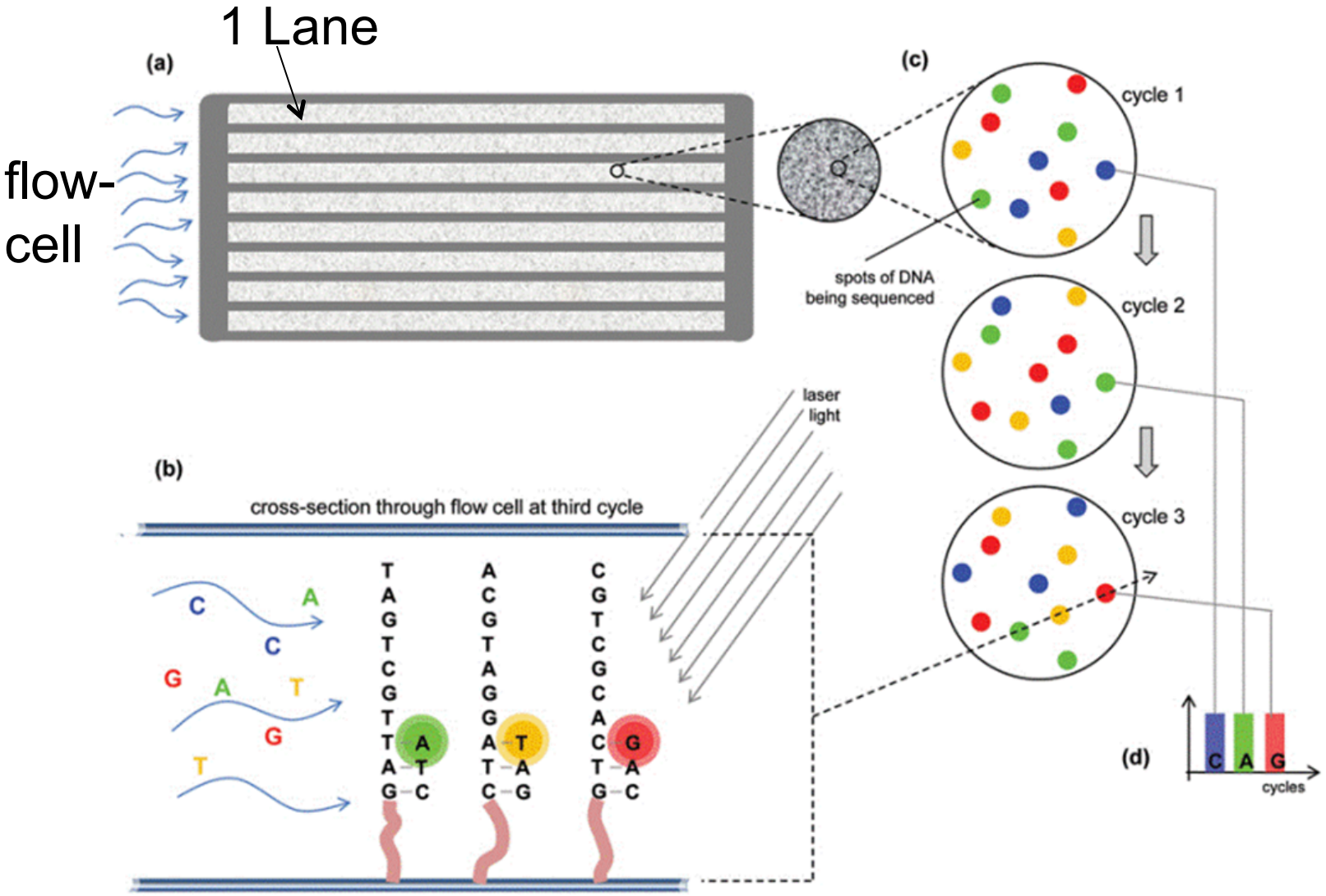
Seq-by-synthesis

Illumina Sequencing

- „Sequencing-by-synthesis“ using reversible dye-terminators

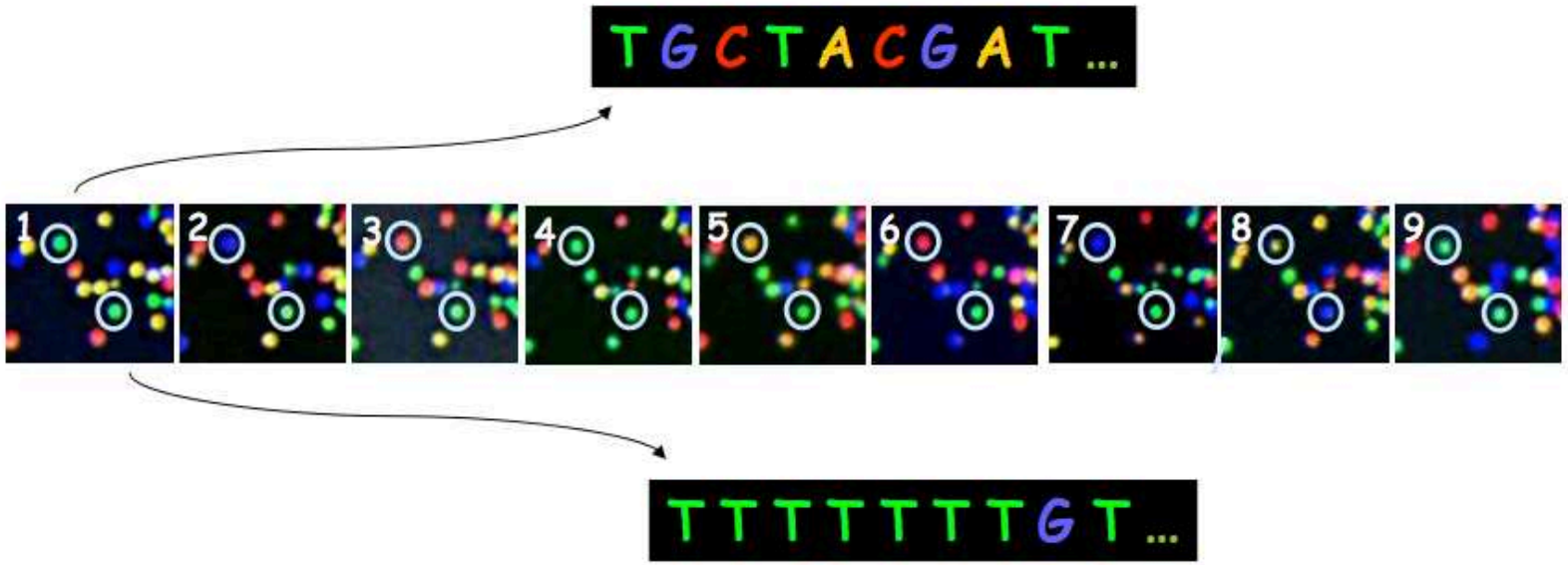


Illumina-Sequenzierung



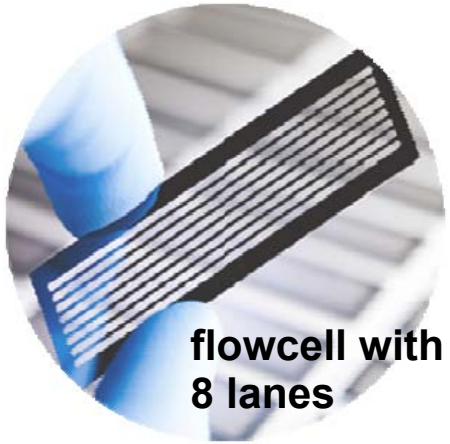
Illumina-Sequencing

9 cycles shown, two cluster position marked...



The sequence of each cluster is determined by consecutive images.

Illumina HiSeq 2500



- 600 Gbp / run = 6 Billion Reads x 100 Bp!
- 600 Gbp / 3 Gbp / 30 x coverage = 6 human genomes
- run time 11 days (high-throughput mode; rapid mode possible)

Illumina Library-Preparation

gDNA, cDNA, plasmids, amplicons ...

↓

DNA fragmentation

↓

end-repair

↓

A-tailing

↓

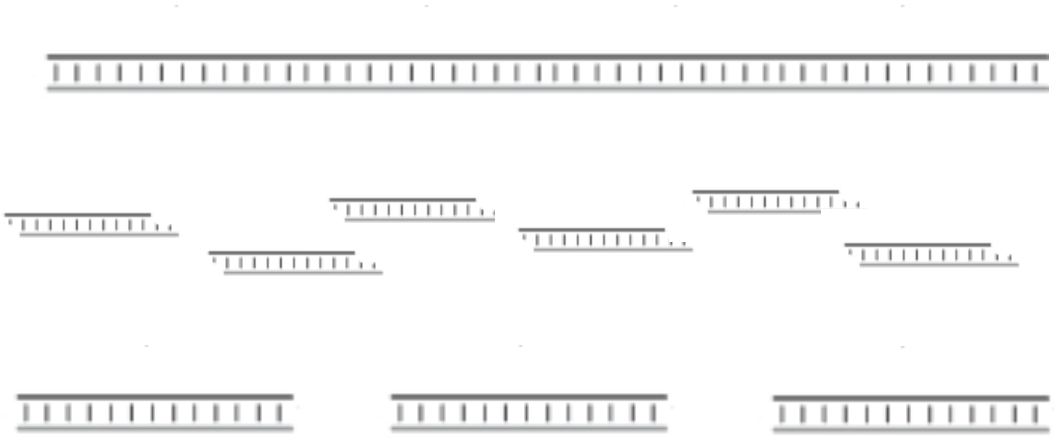
adapter-ligation

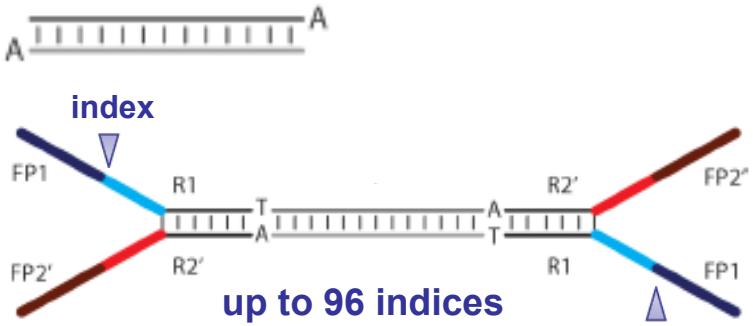
↓

size selection

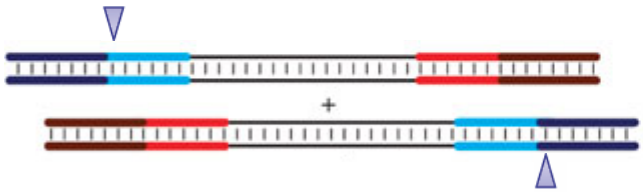
↓

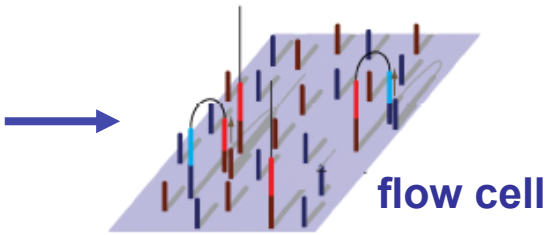
enrichment





e.g. 200-400 bp





Illumina Library-Preparation

Genom-DNA oder cDNA

```
ATTGCGTAGCATCGCGATACGACGTGCTAGATGACTGATCGTACGACGATGATGATCGAGTAGCATGCTCATTGCGTAGCATCGCGATACGACGTGCTAGATGACTGATCGT
TAACGCATCGTAGCGCTATGCTGCACGATCTAGTGACTAGCTAGCTGCTACTACTAGCTCATCGTACGAGTAACGCATCGTAGCGCTATGCTGCACGATCTAGTGACTAGCA
```

Fragmentierung

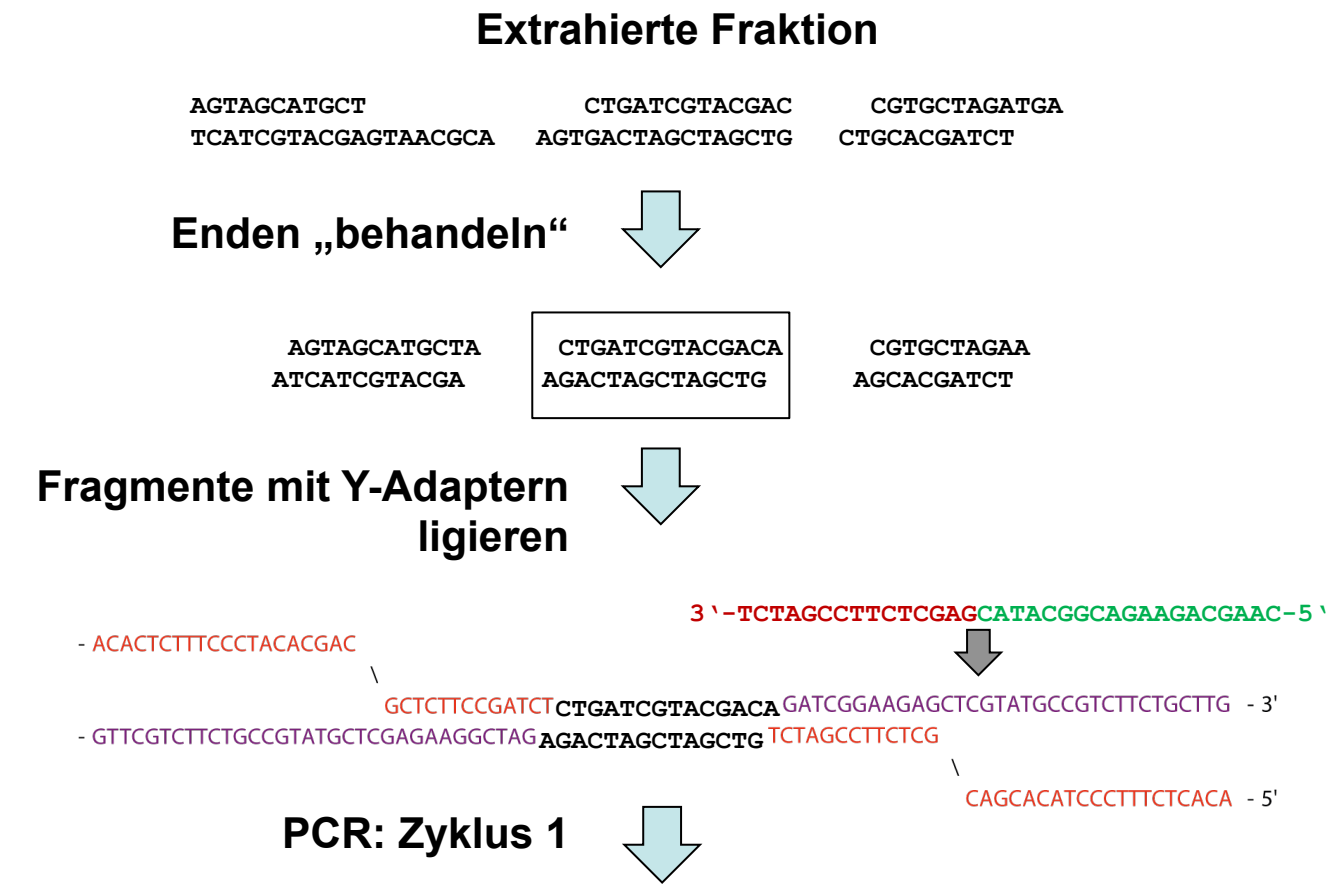
```
ATTGCGTAGCATCGCGATACGA          CTGATCGTACGAC          AGTAGCATGCT
TAACGCATCGTAGCGCTATG          AGTGACTAGCTAGCTG          TCATCGTACGAGTAACGCA

                    CGTGCTAGATGA          GATGATGATCG          CATTGCGTAGCATCGCGATACGACGTGCTAGATGACTGATCGT
                  CTGCACGATCT          CTACTACTAGC                    TCGTAGCGCTATGCTGCACGATCTAGTGACTAGCA
```

Nach Größe sortieren

```
CATTGCGTAGCATCGCGATACGACGTGCTAGATGACTGATCGT
        TCGTAGCGCTATGCTGCACGATCTAGTGACTAGCA

ATTGCGTAGCATCGCGATACGA
TAACGCATCGTAGCGCTATG

AGTAGCATGCT
TCATCGTACGAGTAACGCA

   CTGATCGTACGAC
AGTGACTAGCTAGCTG

  CGTGCTAGATGA
CTGCACGATCT

GATGATGATCG
CTACTACTAGC
```

gewünschte Fraktion isolieren (oft 200 bis 500 bp, richtet sich nach Leselänge)

Illumina Library-Preparation



DNA-Strang mit 2 unterschiedlichen Enden
(das reziproke Produkt des anderen Strangs ist nicht gezeigt)

Illumina Library-Preparation

DNA-Strang mit 2 unterschiedlichen Enden

ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATCGTACGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAGACTAGCTAGCTGTCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC

PCR: Zyklen 2+

AATGATACGGCGACCACCGAGAACACTCTTTCCCTACACGACGCTCTTCCGATCT >
AATGATACGGCGACCACCGAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATCGTACGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG

TTACTATGCCGCTGGTGGCTCTTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAGACTAGCTAGCTGTCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC
< TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC

PCR: Zyklen 2+

einzigartiger Bereich des Adapters ROT

zu sequenzierender Bereich

einzigartiger Bereich des Adapters GRÜN

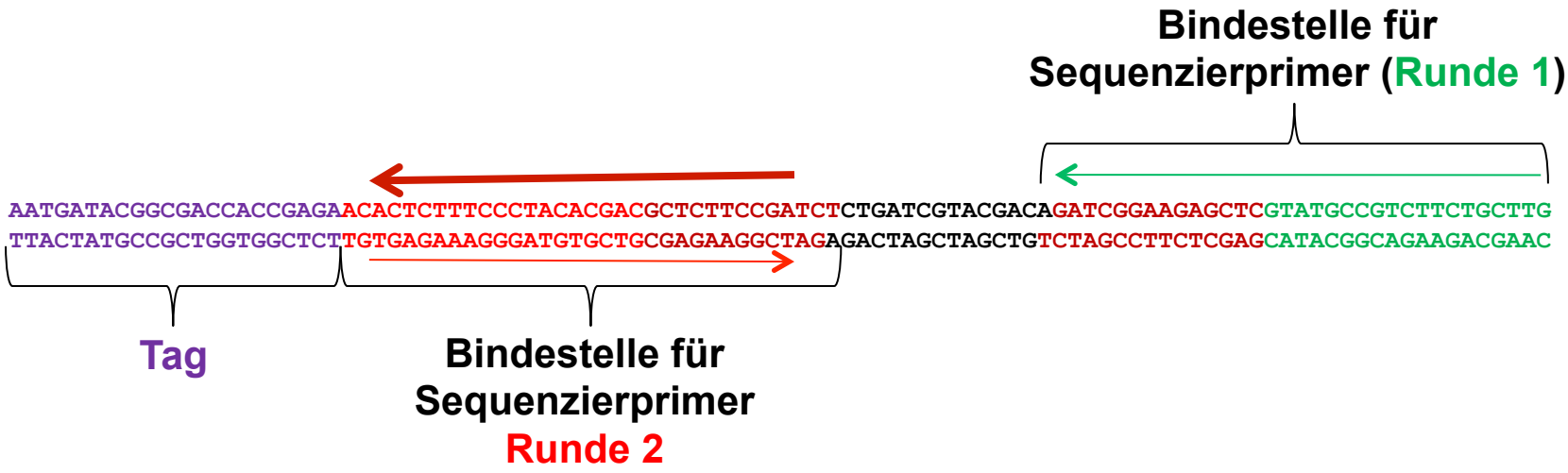
AATGATACGGCGACCACCGAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATCGTACGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAGACTAGCTAGCTGTCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC

durch PCR angehängt
(enthält Multiplex-Tag)

von Y-Adapter

Sequenz, die in beiden Adaptern identisch ist

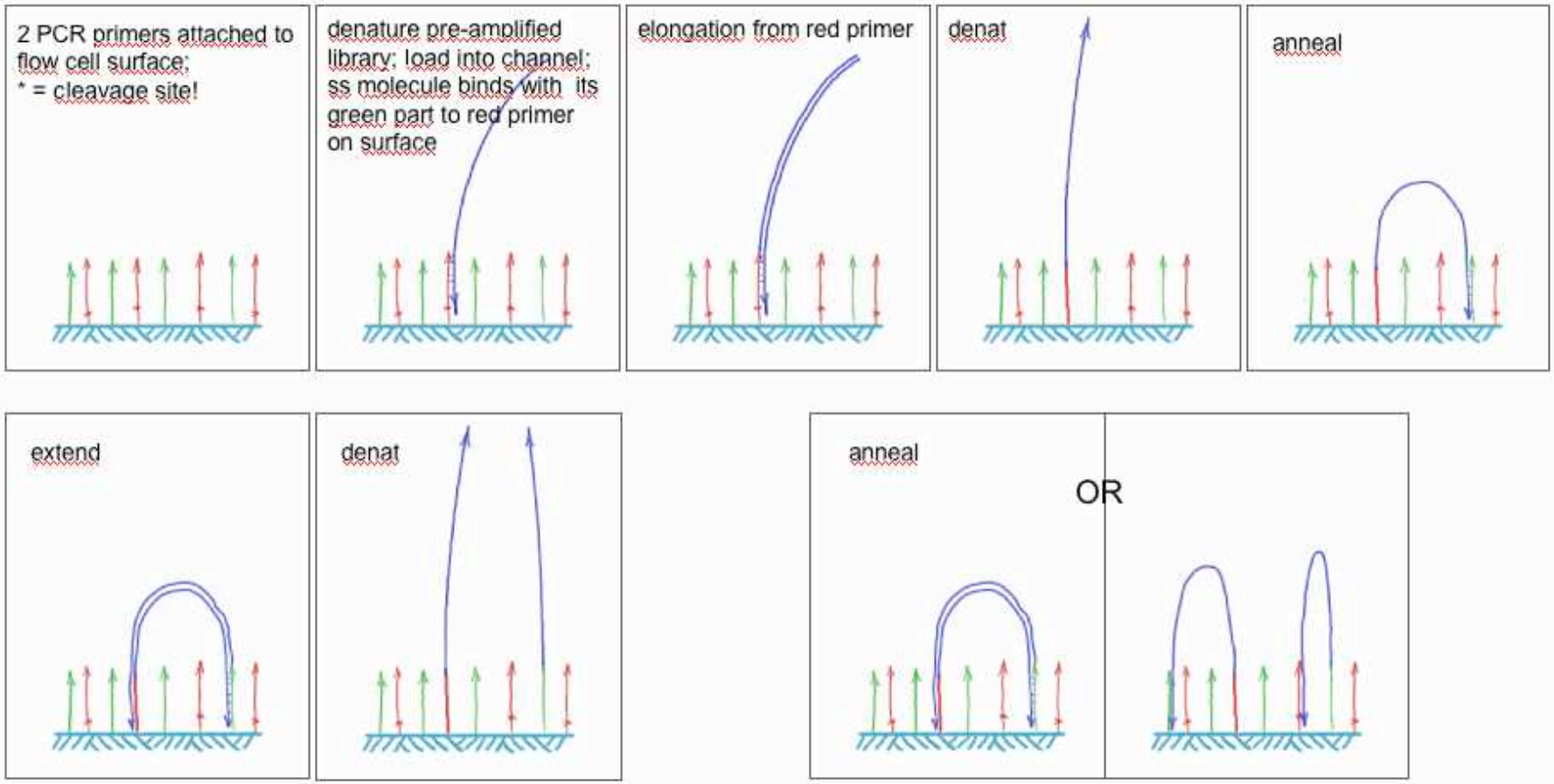
Illumina Library-Preparation



- Multiplex-Tag ist also nicht in den Sequenzen enthalten, die von den Primern der Runde 1 und 2 ausgehen
- Für Tags gibt es daher einen separaten kleinen Sequenzierlauf ausgehend vom Primer DUNKELROT
- die Sequenzierungen von Primer 1 und Primer 2 aus nennt man PAIRED-ENDS!

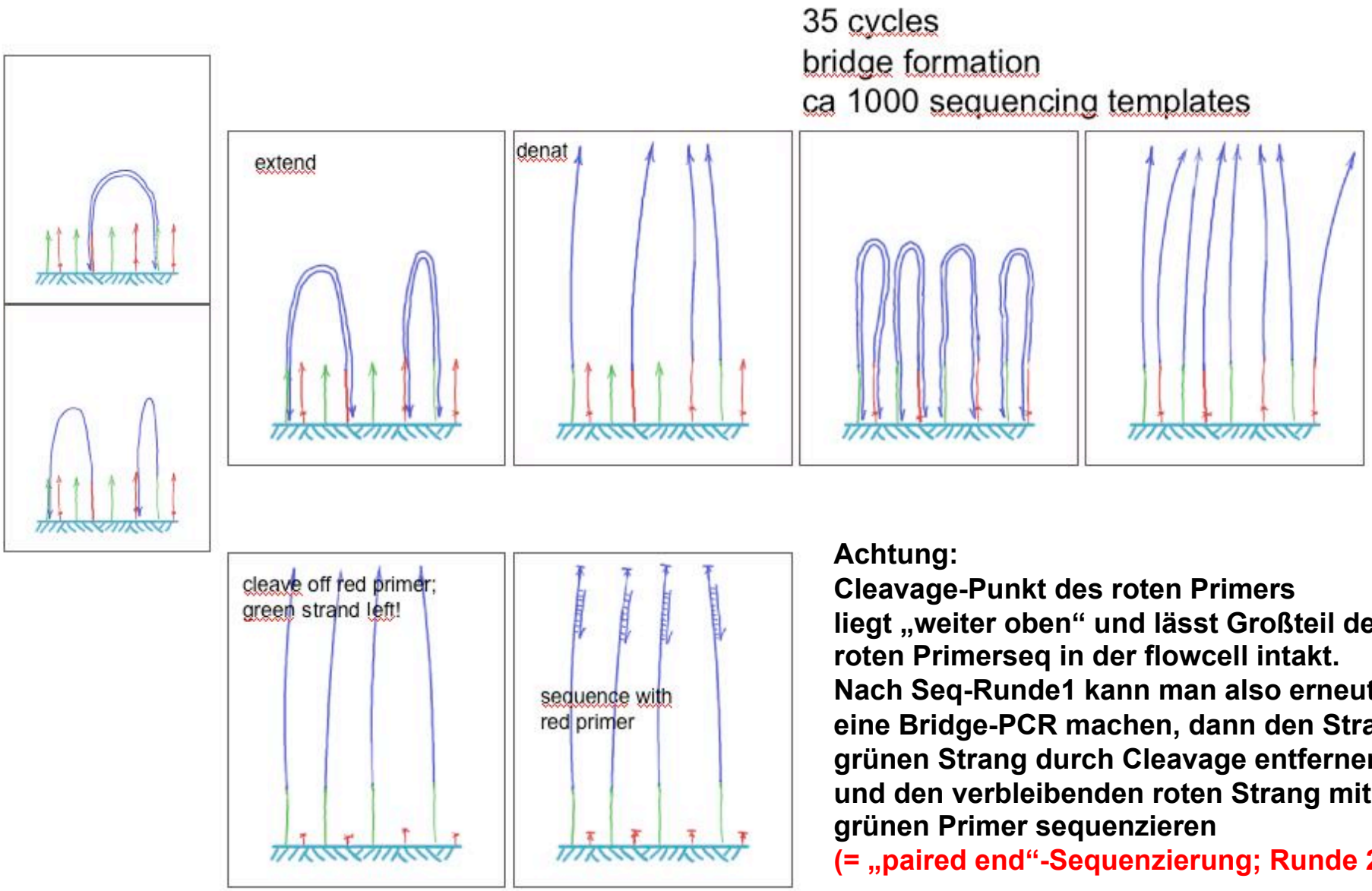
Illumina template prep: bridge PCR

Shear DNA > size fractionate > attach adapters > pre-amplify by PCR with red and green primers > then: bridge PCR on flowcell



p://seq.molbiol.ru/sch_clon_ampl.html

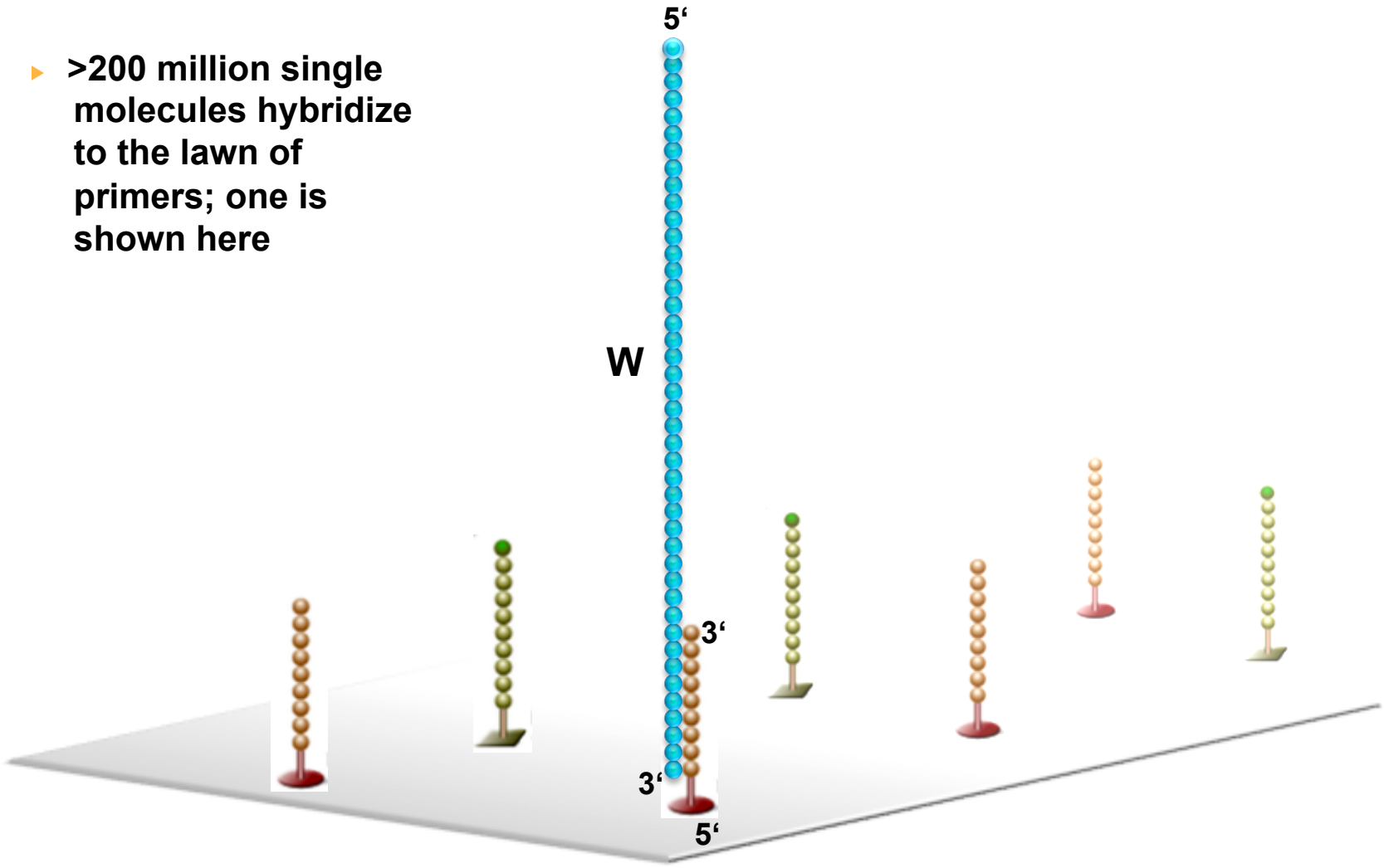
Illumina template prep: bridge PCR



Achtung:
Cleavage-Punkt des roten Primers liegt „weiter oben“ und lässt Großteil der roten Primerseq in der flowcell intakt.
Nach Seq-Runde1 kann man also erneut eine Bridge-PCR machen, dann den Strang grünen Strang durch Cleavage entfernen und den verbleibenden roten Strang mit dem grünen Primer sequenzieren
(= „paired end“-Sequenzierung; Runde 2)

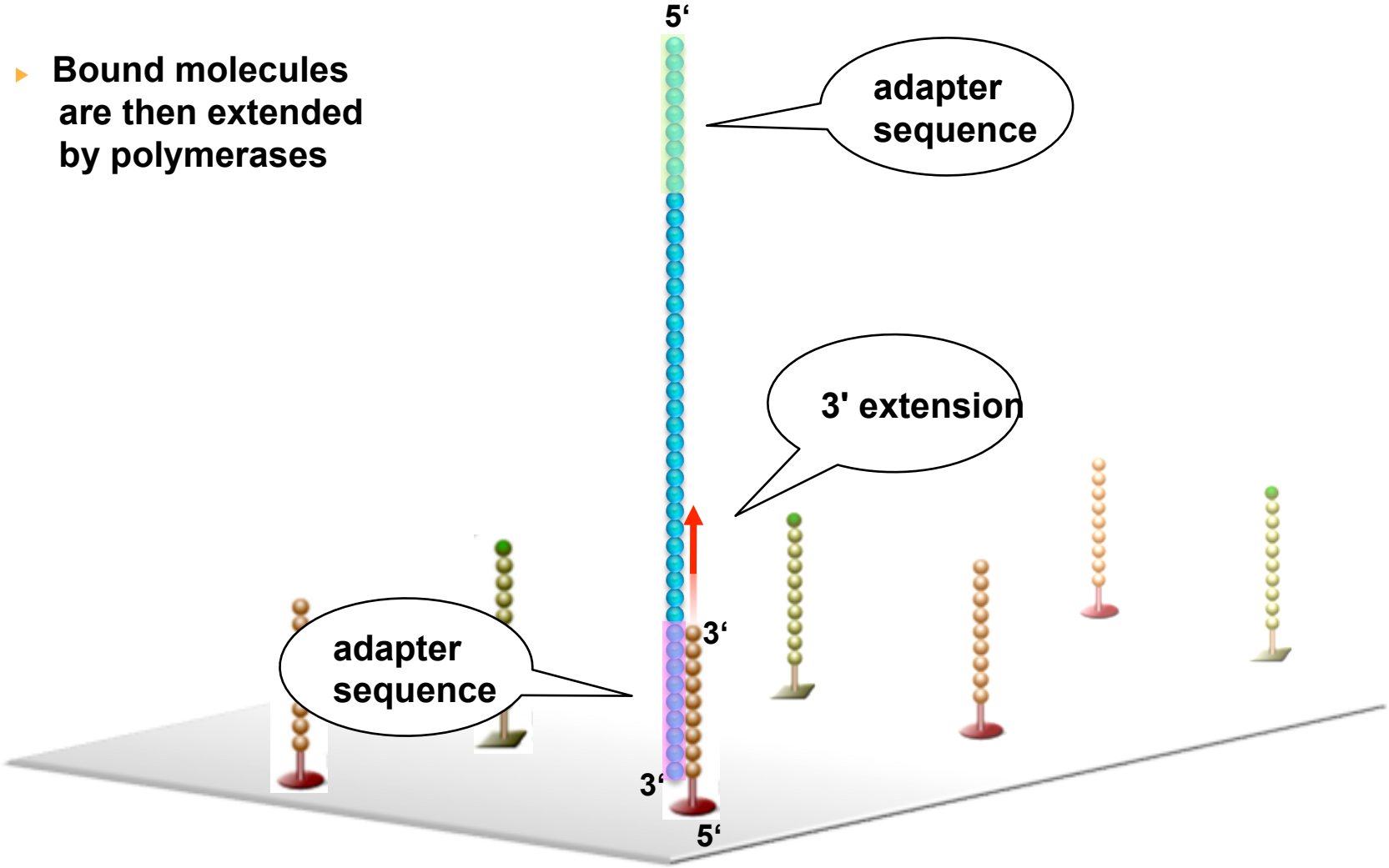
Illumina Cluster Generation

- >200 million single molecules hybridize to the lawn of primers; one is shown here



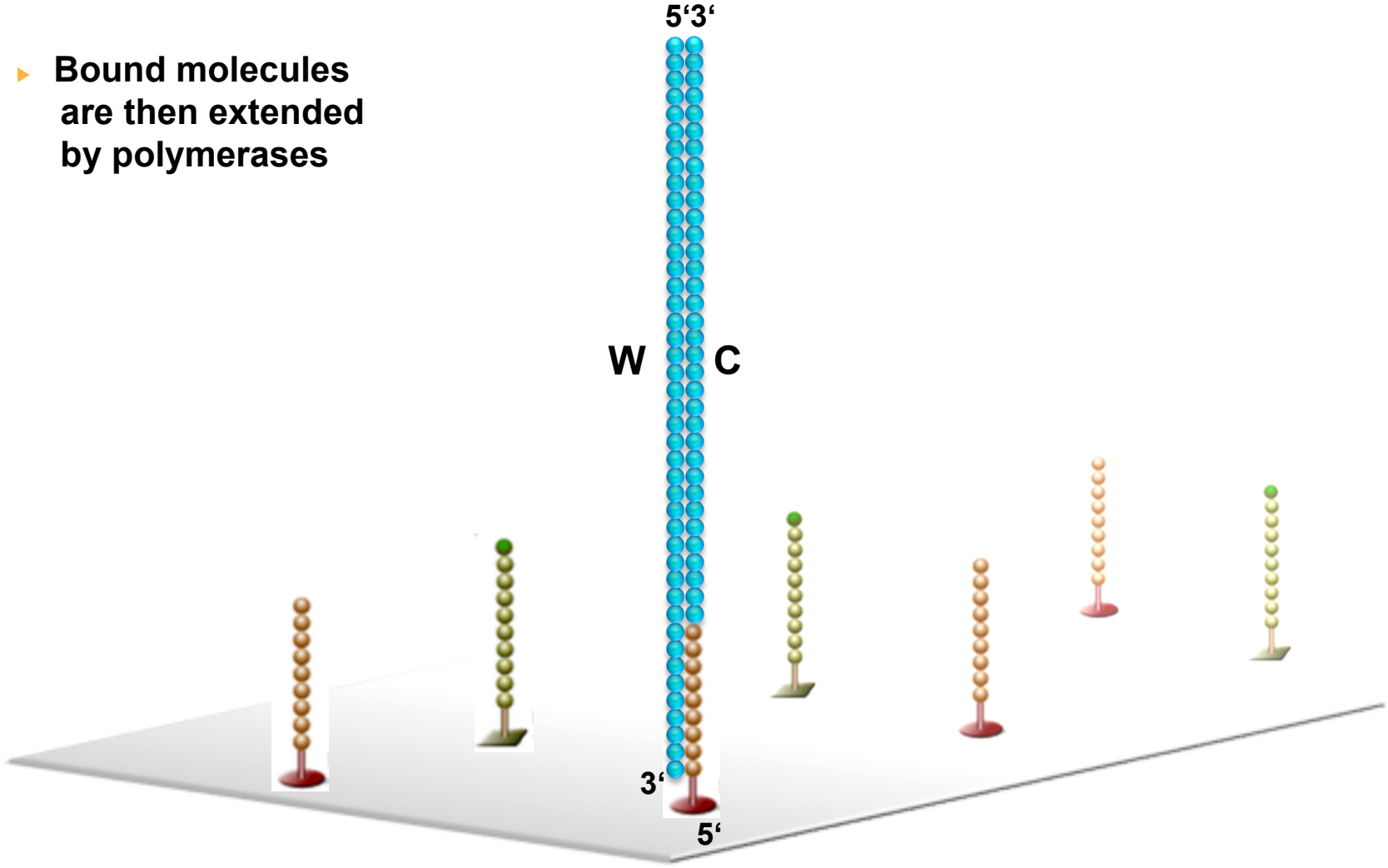
Illumina Cluster Generation

- Bound molecules are then extended by polymerases



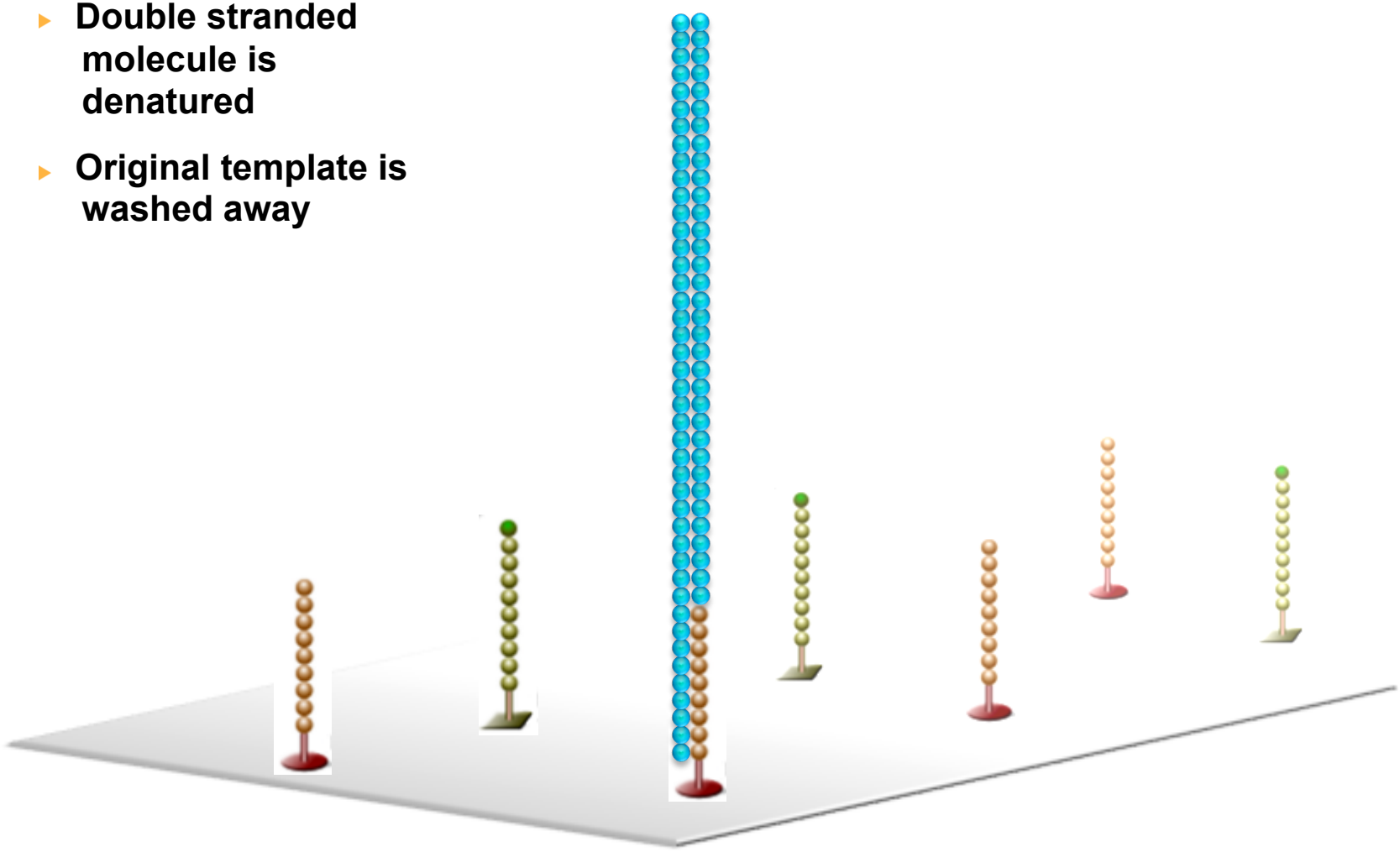
Illumina Cluster Generation

- Bound molecules are then extended by polymerases



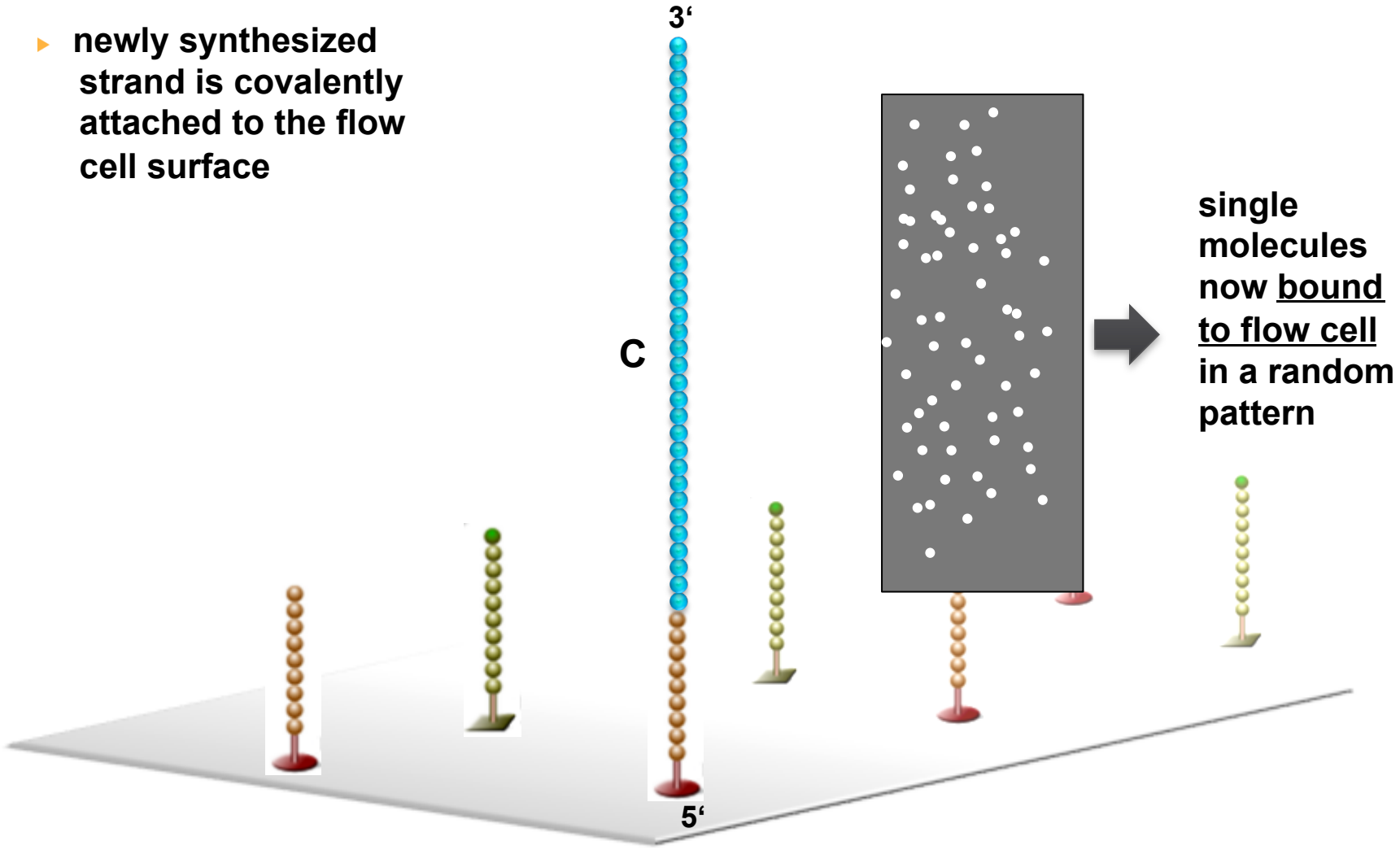
Illumina Cluster Generation

- Double stranded molecule is denatured
- Original template is washed away



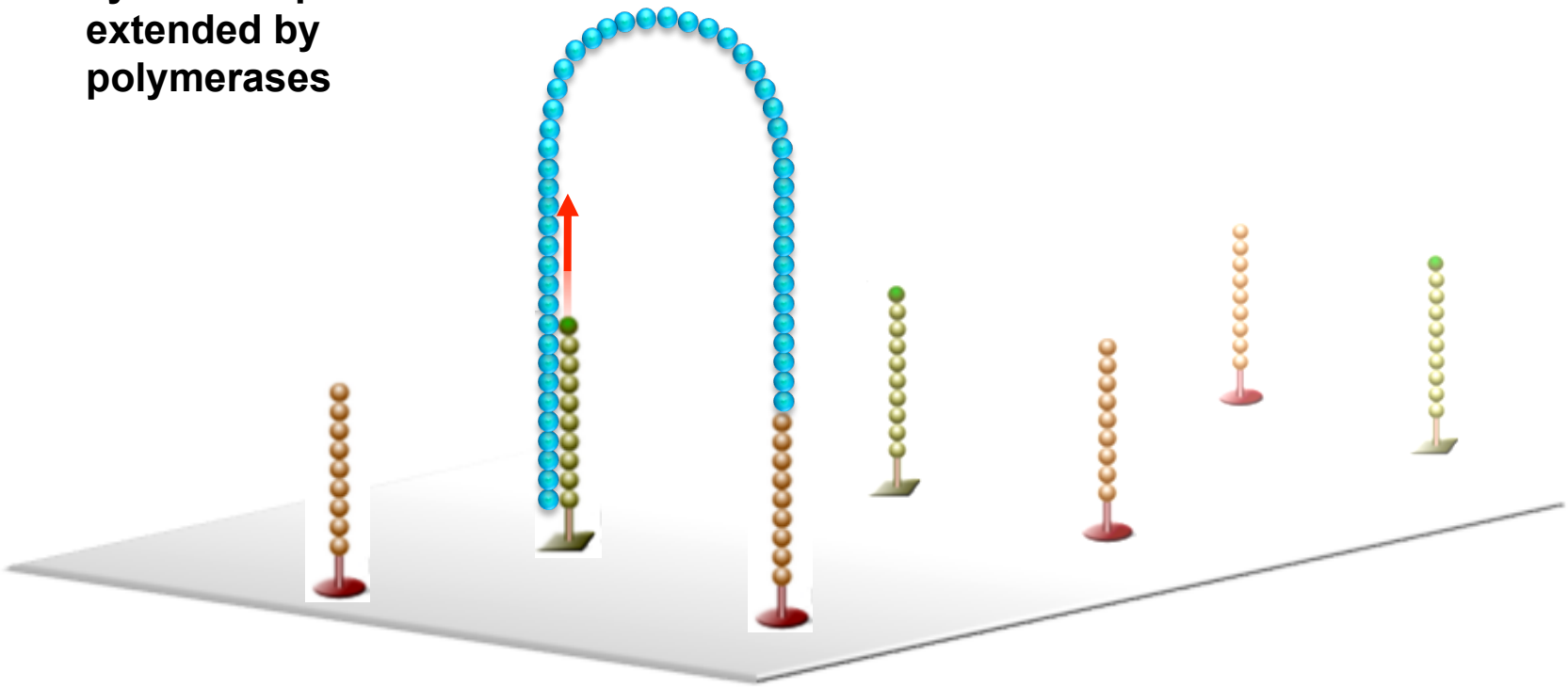
Illumina Cluster Generation

- newly synthesized strand is covalently attached to the flow cell surface



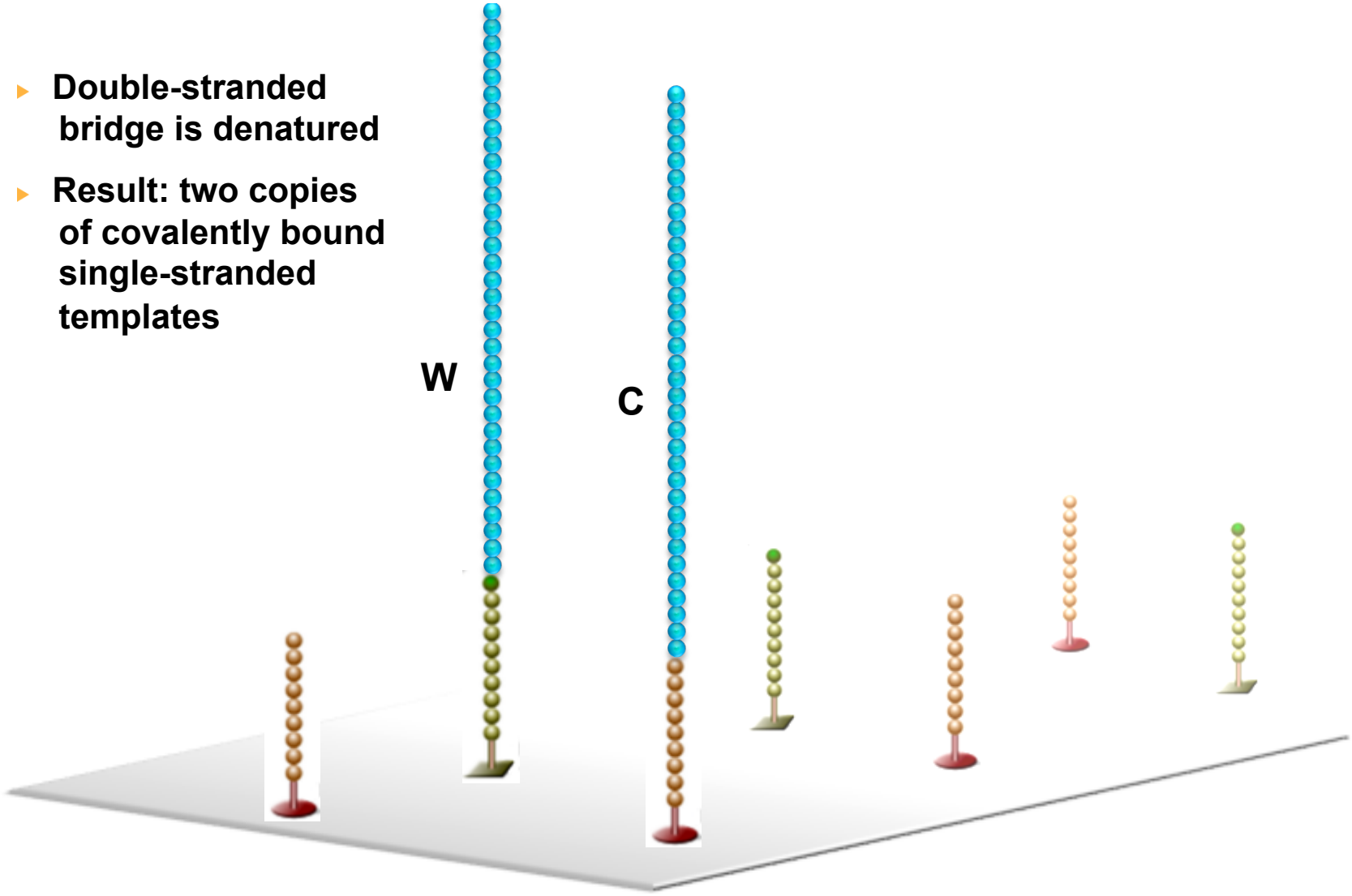
Illumina Cluster Generation

- Single-strand flips over to hybridize to adjacent oligos to form a bridge
- Hybridized primer is extended by polymerases



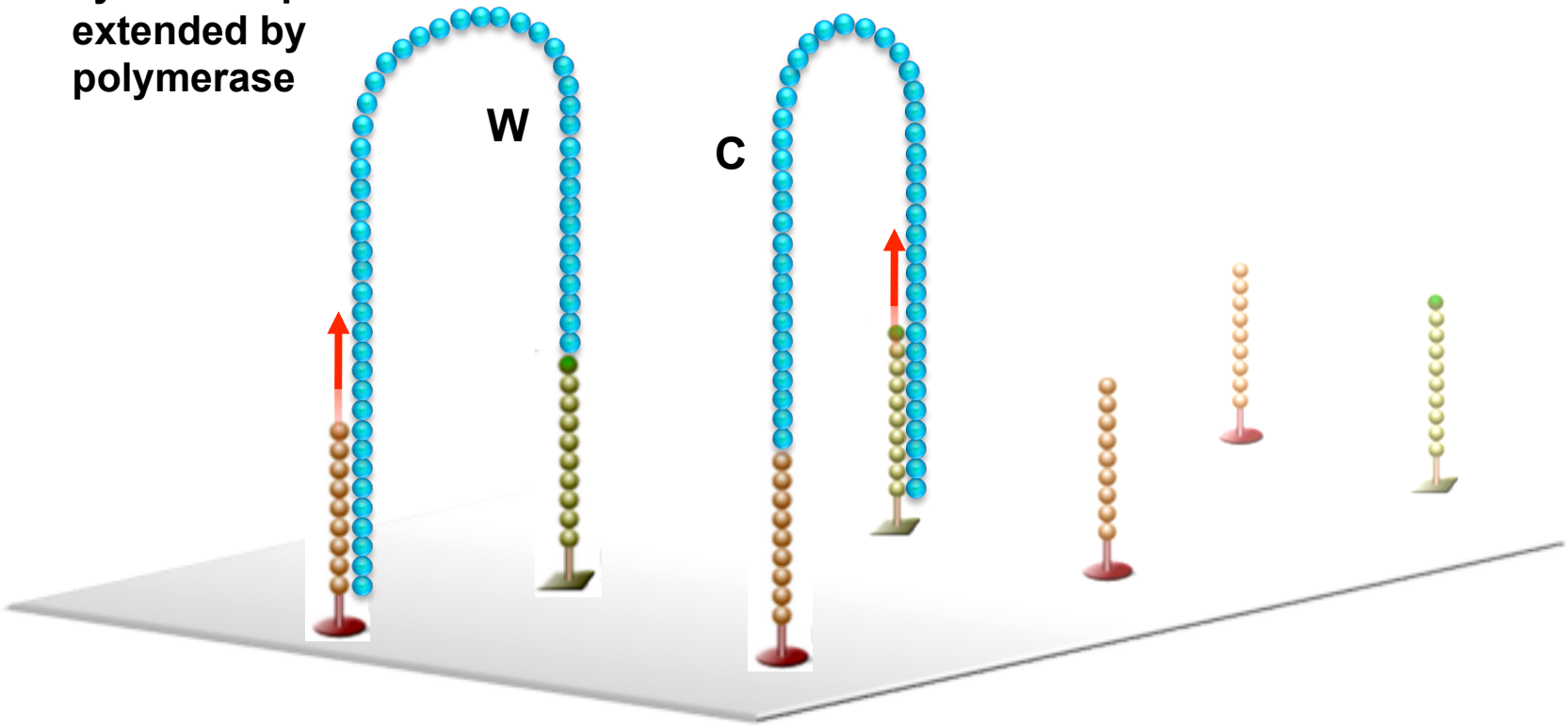
Illumina Cluster Generation

- Double-stranded bridge is denatured
- Result: two copies of covalently bound single-stranded templates



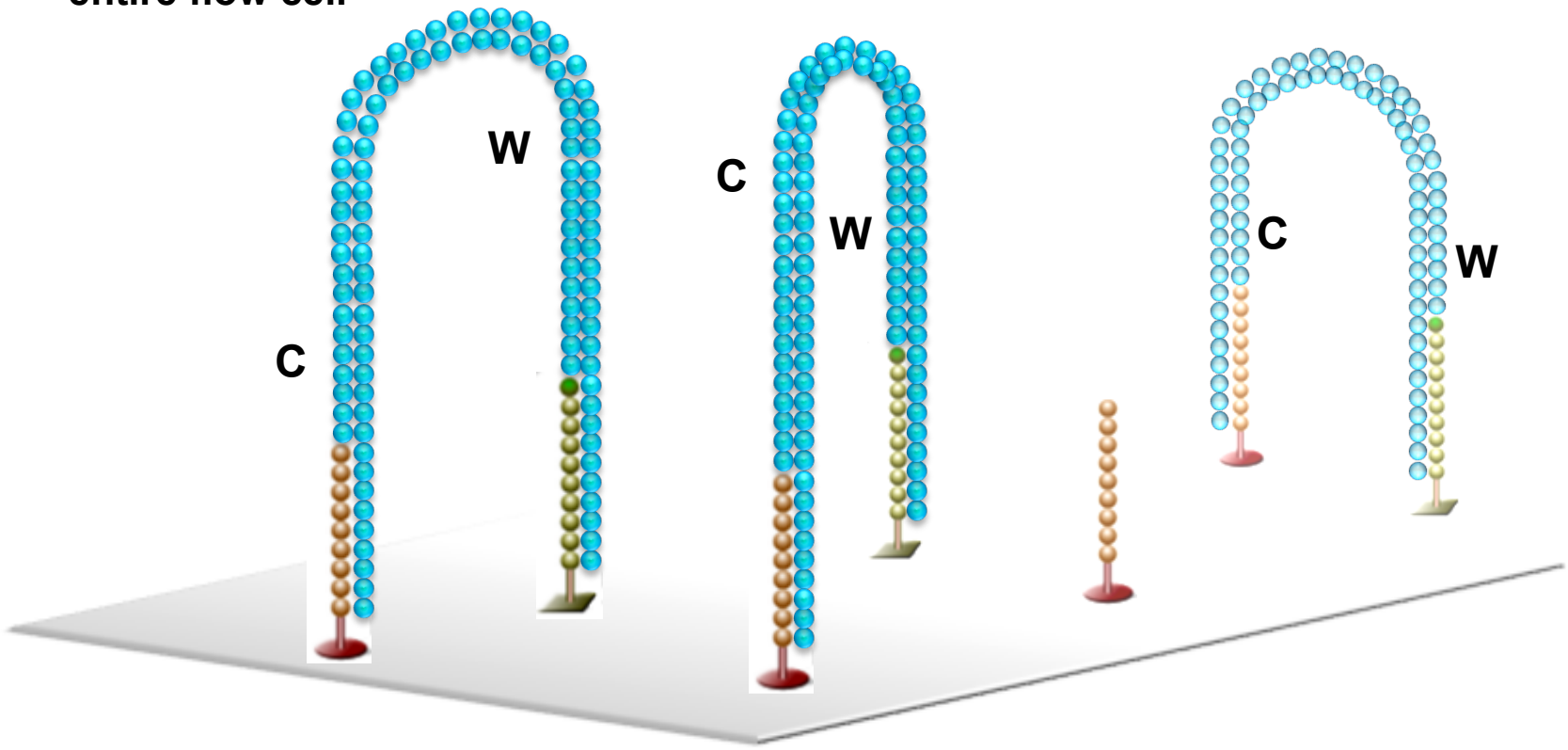
Illumina Cluster Generation

- Single-strands flip over to hybridize to adjacent oligos to form bridges
- Hybridized primer is extended by polymerase



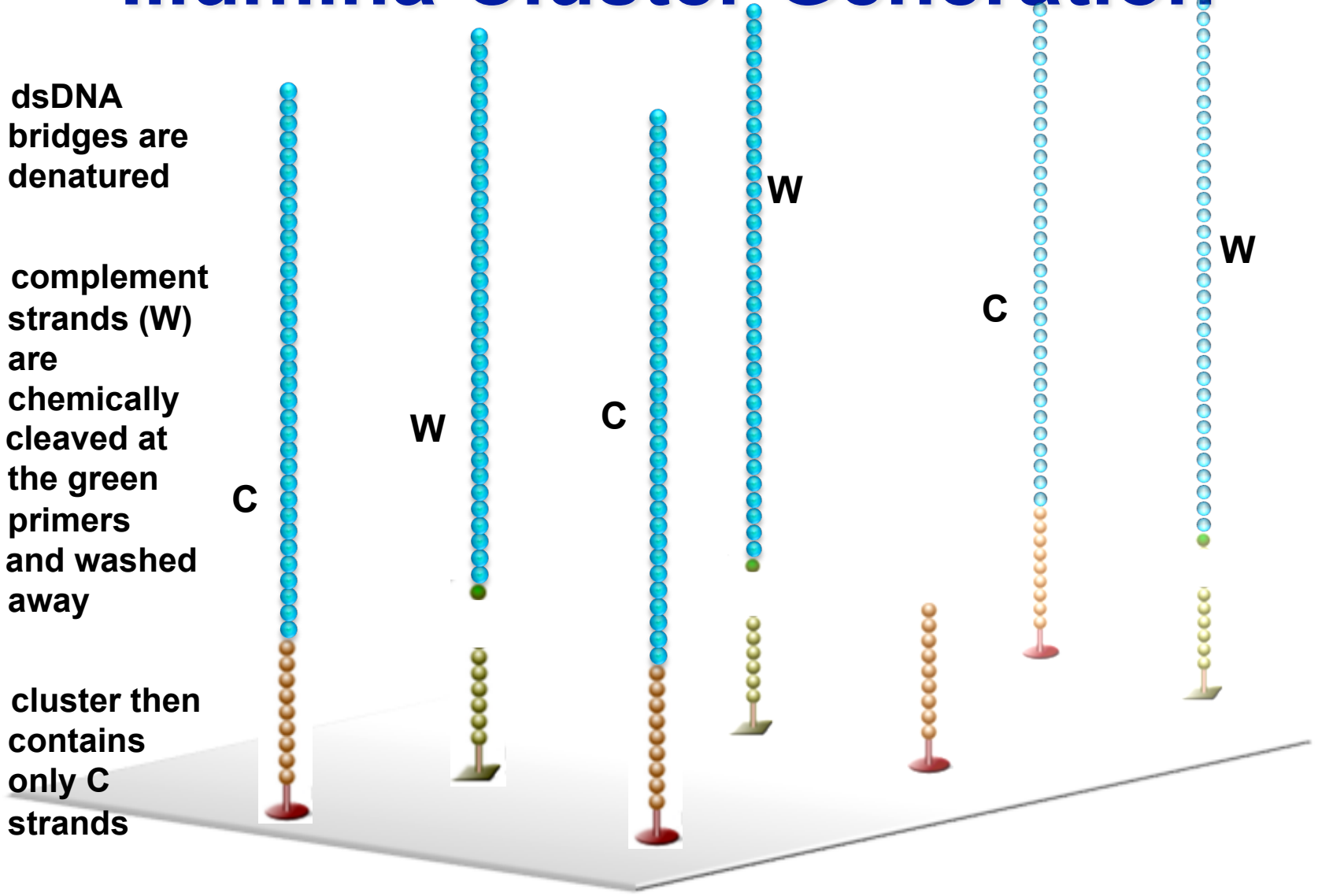
Illumina Cluster Generation

- **Bridge amplification cycle repeated until multiple bridges are formed across the entire flow cell**



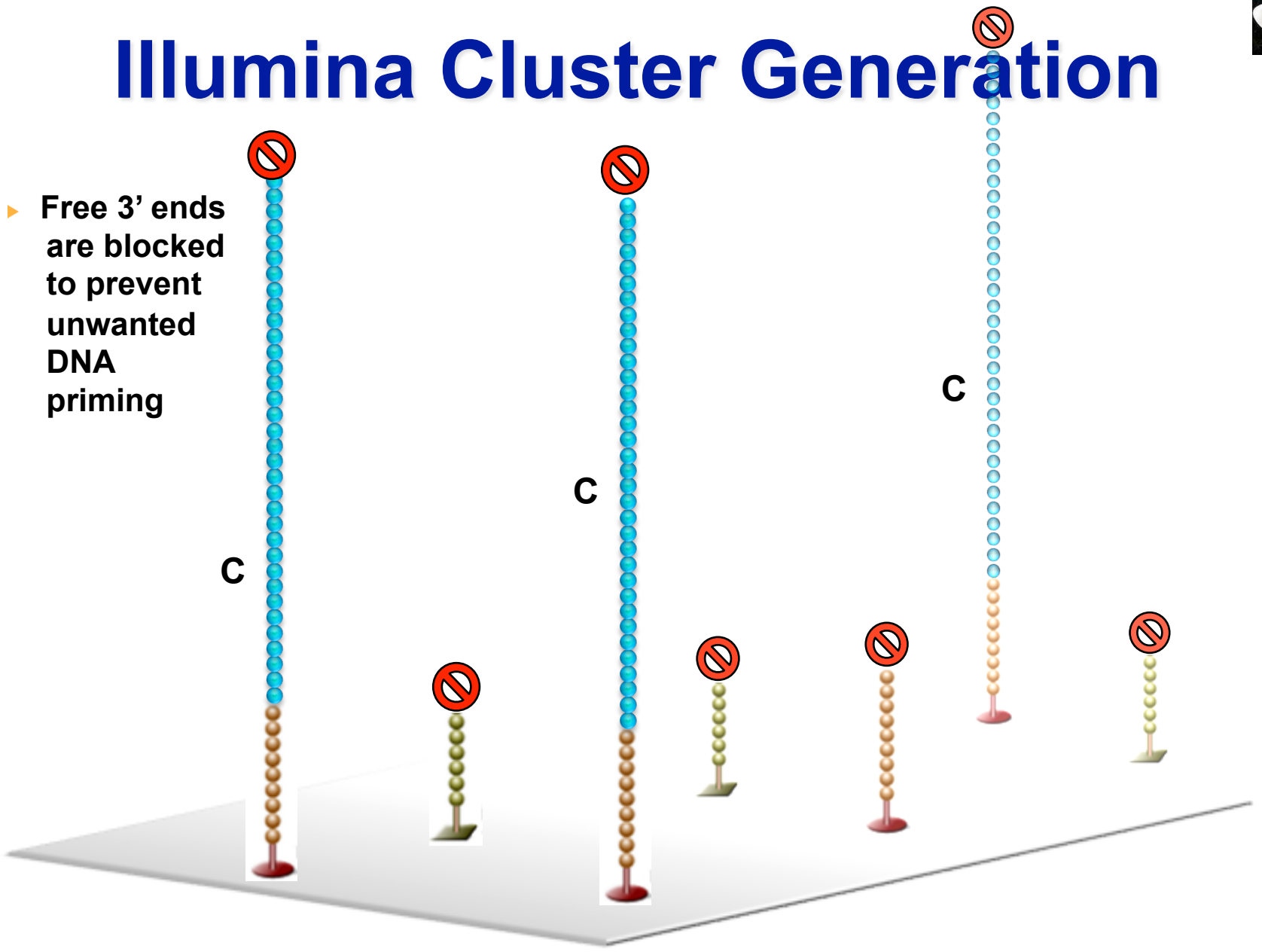
Illumina Cluster Generation

- dsDNA bridges are denatured
- complement strands (W) are chemically cleaved at the green primers and washed away
- cluster then contains only C strands



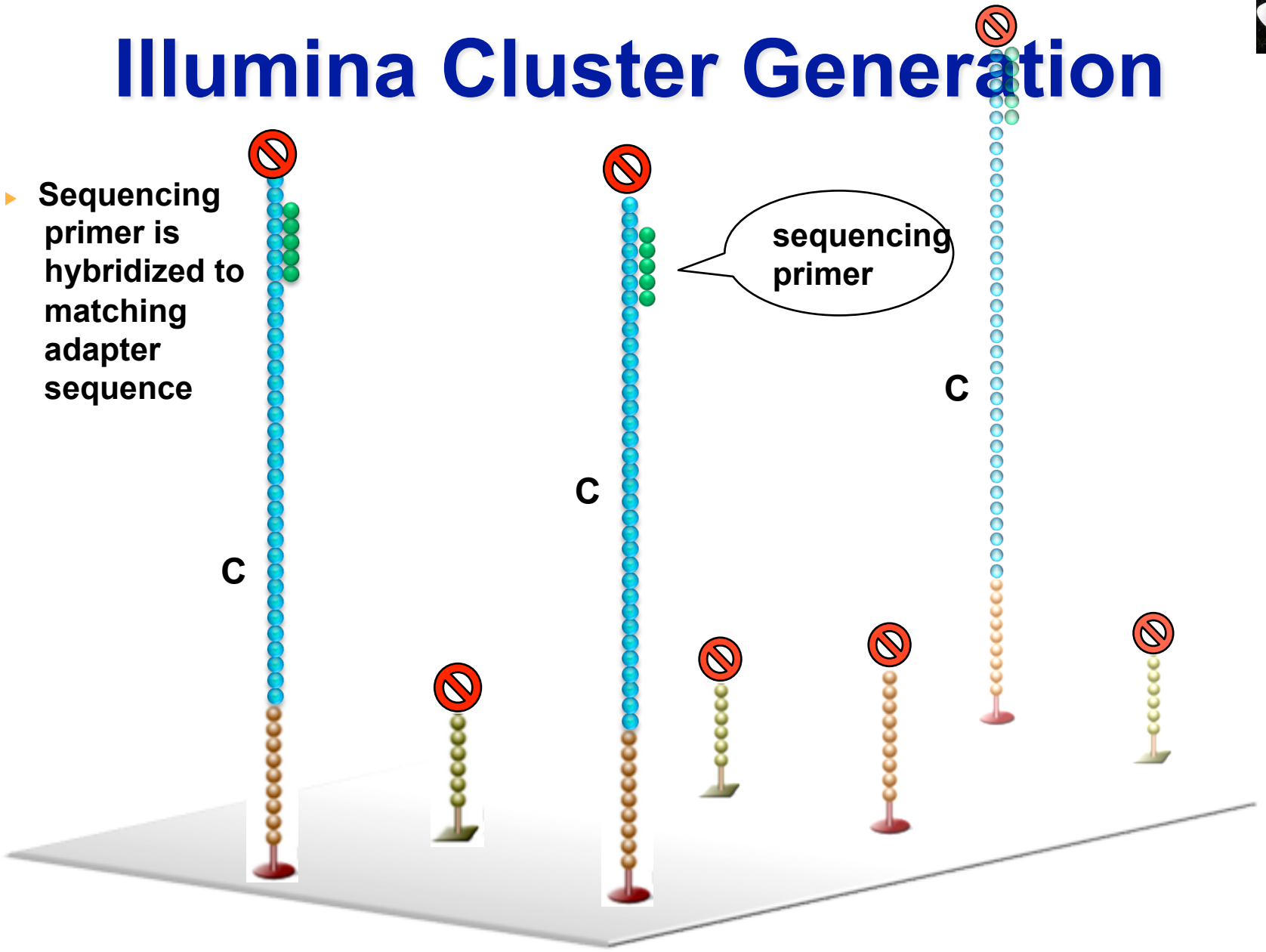
Illumina Cluster Generation

- Free 3’ ends are blocked to prevent unwanted DNA priming

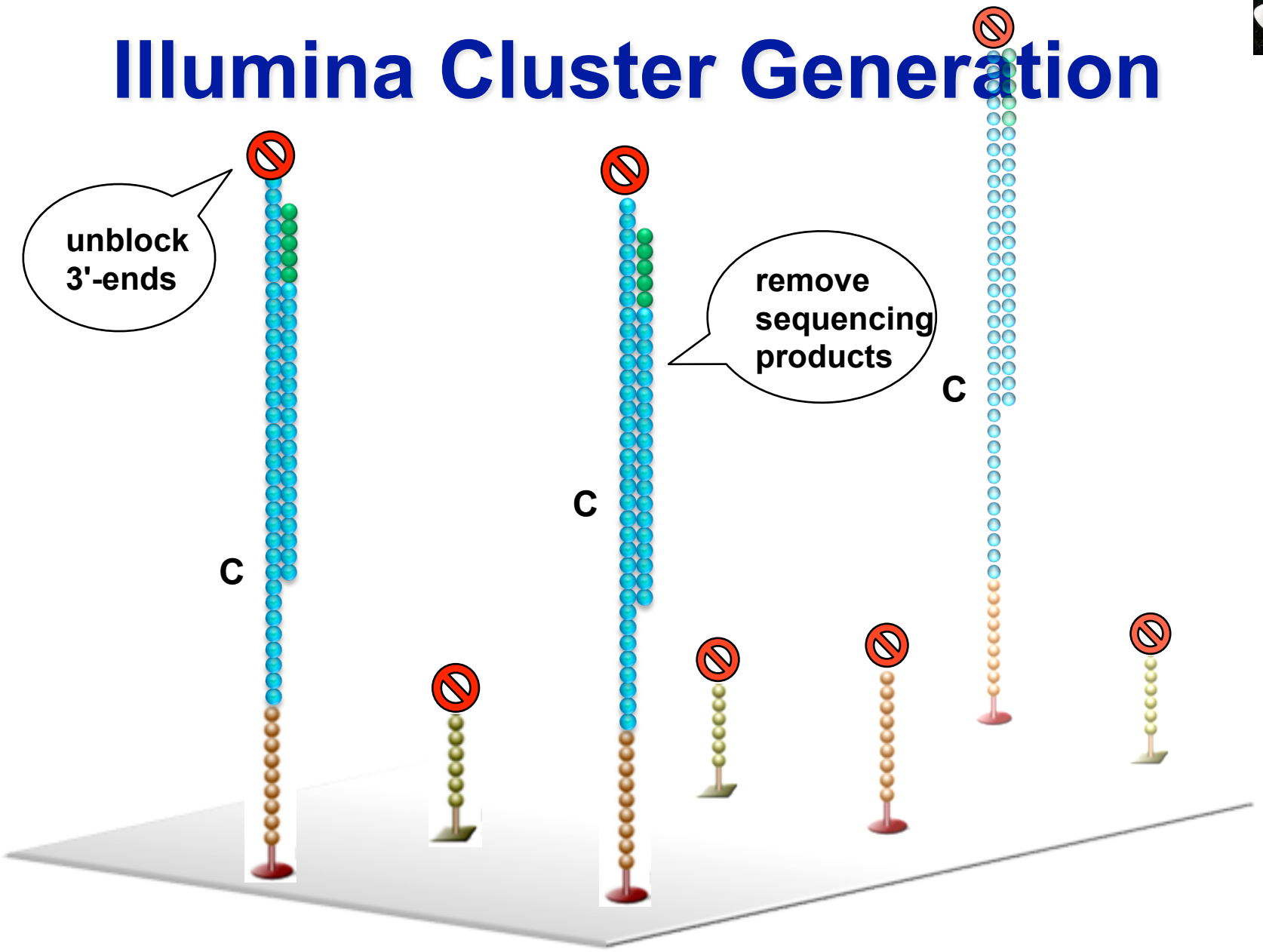


Illumina Cluster Generation

- Sequencing primer is hybridized to matching adapter sequence



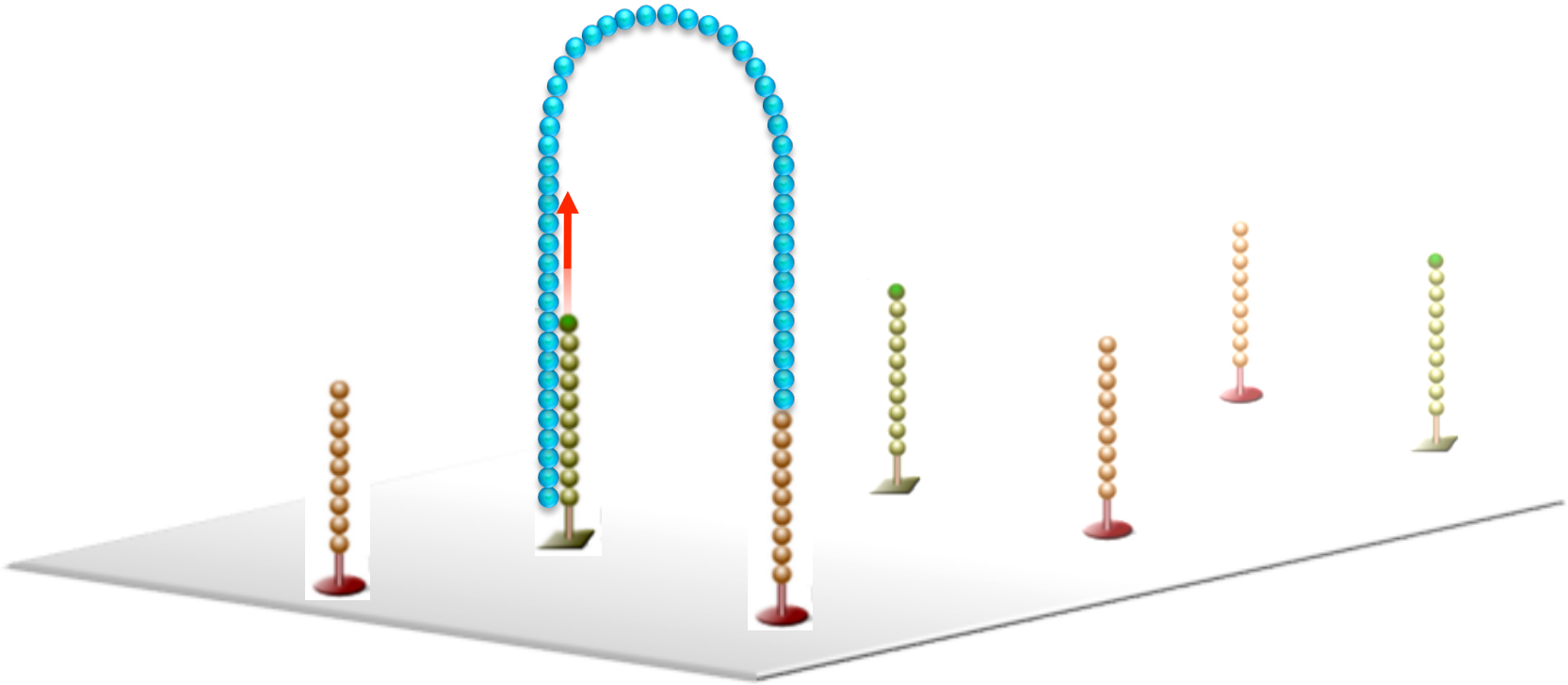
Illumina Cluster Generation



Sequence 2nd strand! (Paired-end sequencing) > W strand needed

Illumina Cluster Generation

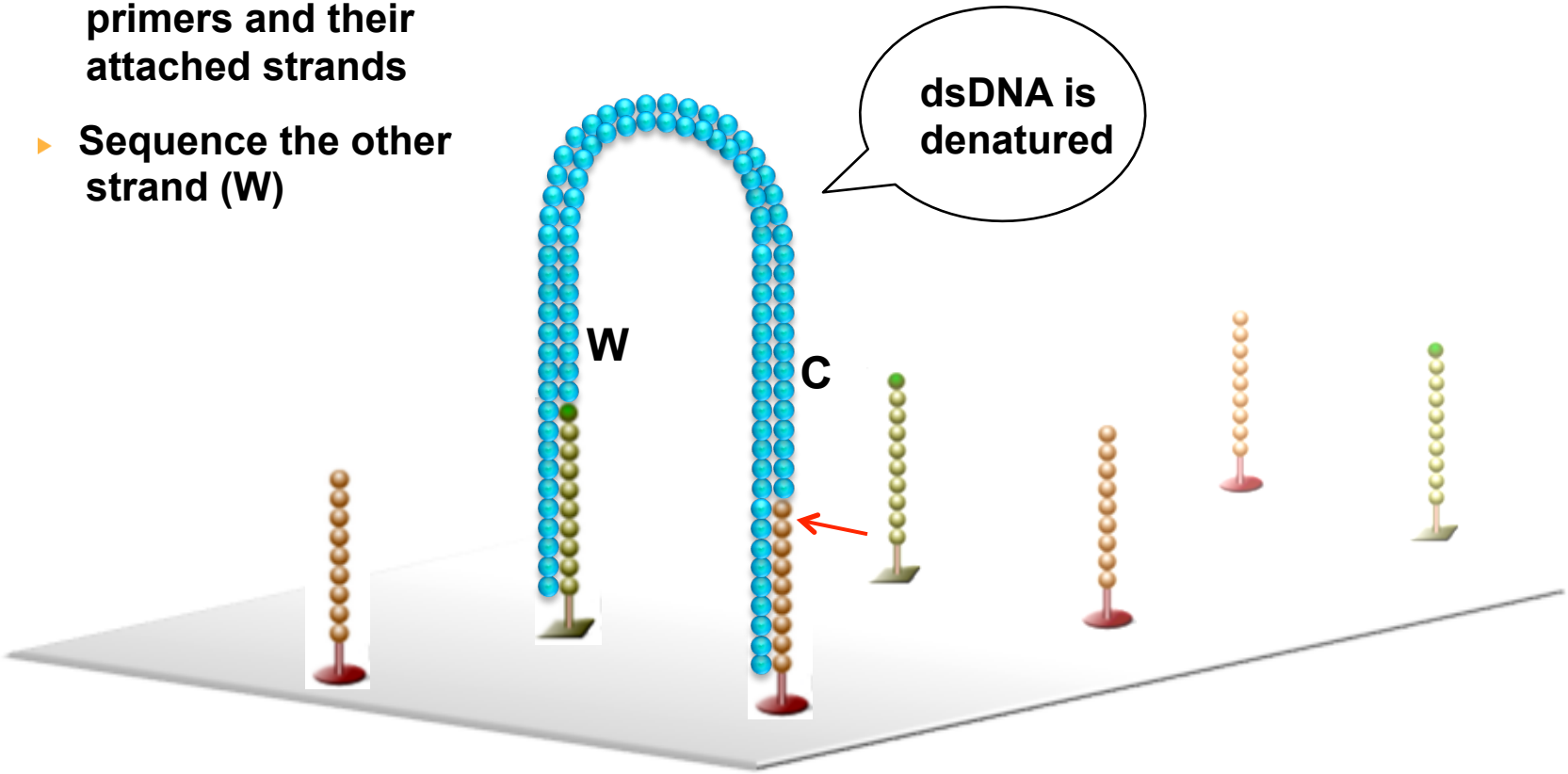
- **Bridge formation and 3' extension**



Do it again! New bridge PCR! Then remove C strand…

Illumina Cluster Generation

- Bridge PCR Round 2
- Cleave off red primers and their attached strands
- Sequence the other strand (W)



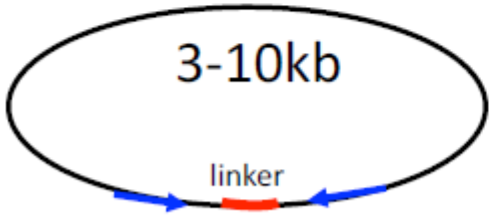
> Paired-end sequencing

Sequenzierstrategien

Illumina/SOLiD

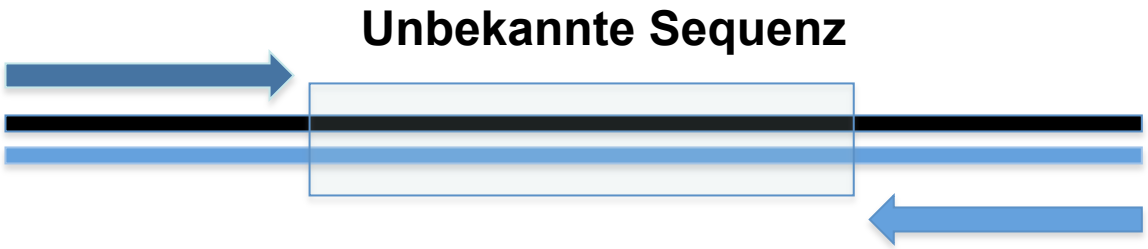
Single end (SE)

Paired-end (PE), short fragment ends



Mate-pair (MP), another type of paired-end, circularization (cloning)

Paired End Sequencing



„inward-facing reads“, Abstand gering und definiert

Zusätzlich brauchen wir „long-distance“-Sequenzinformation, um Contigs relativ zueinander anzuordnen...

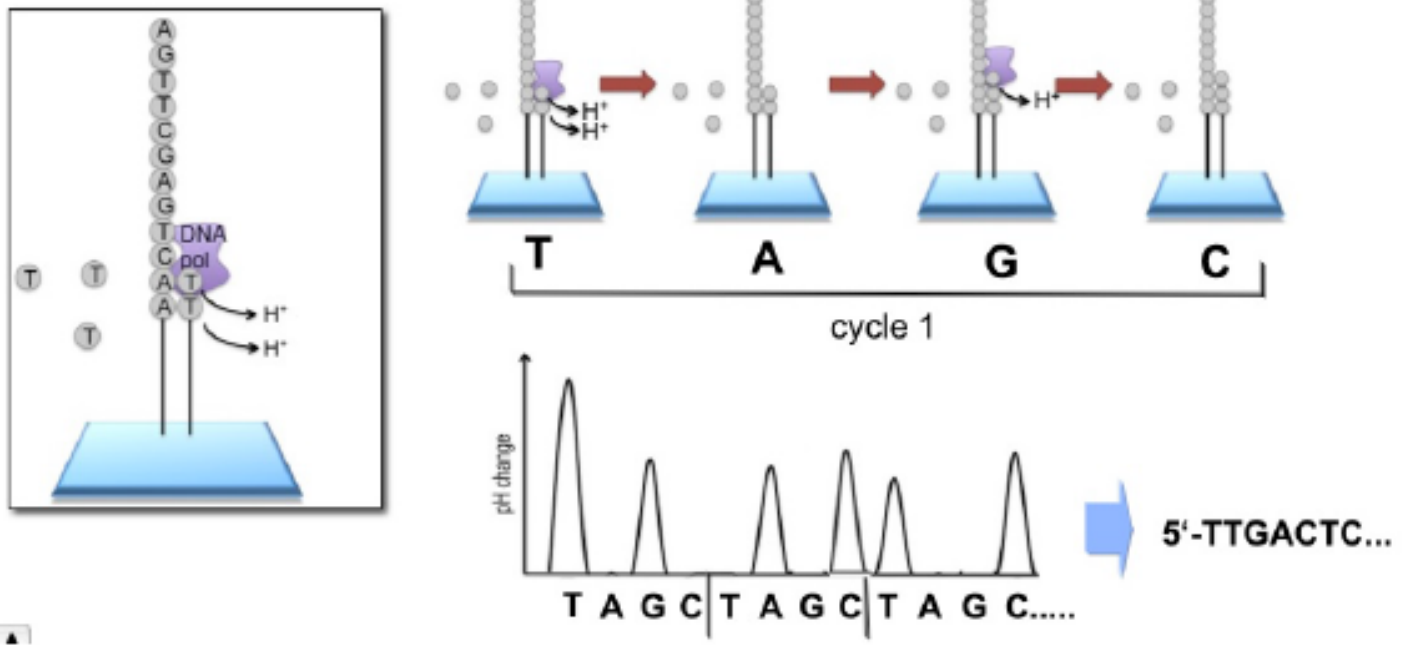
Vergleich von Sequenziermethoden

(2nd generation NGS)

	454 (Roche)	Ion Torrent	Illumina	Complete Genomics	Sanger
DNA matrix	Emulsion PCR	Emulsion PCR	Bridge PCR	amplification: DNA nano balls	Plasmids Clones PCR
Sequencing Method	seq-by-synth: Pyrosequencing	seq-by-synth: Proton release	seq-by-synth: reversible Dye-Terminators	Seq-by-ligation	seq-by-synth: Dye Terminator 96 capillaries
Read length	av. 600 bp (up to 1000)	Up to 700	2 x 100 bp (up to 2 x 300)	70 bp	1000 bp
Data	600 Mbp	1 Gbp	Up to 1.5 Tbp	20-60 Gbp	0,1 Mbp
Runtime	10 hrs	90 min	2-10 Days	?	2 hrs

Ion Torrent (Life Tech)

- nukleotide incorporation > proton release > change in pH



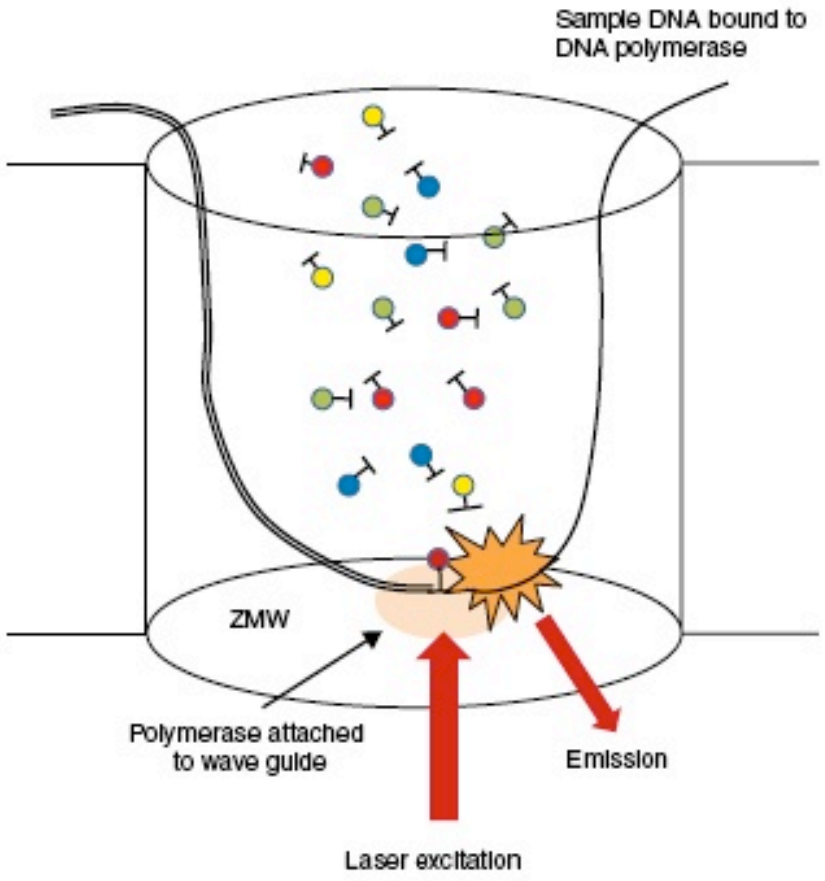
Vergleich von Sequenziermethoden *(3rd generation NGS)*

	Pacific Bioscience	Oxford Nanopore
DNA matrix	Single-mol	Single-mol
Sequencing Method	seq-by-synth: labeled hexaphosphate Nt‘s	direct sequencing in nanopores
Read length	20 kbp (mean) 90 kbp max	20 kbp (mean) Up to 1.2 Mbp
Data	5-8 Gbp /SMRT Cell (16 Cells)	10-20 Gbp / Flow Cell
Runtime	30 min – 10 hrs	Realtime (48 hrs max)

DNA sequencing giant Illumina just bought rival Pac Bio for $1.2 billion — here's why

- Illumina just paid $1.2 billion for Pacific Biosciences, to help it retain its dominant position in the DNA sequencing space, biotech experts say.
- Illumina, which is valued at more than $45 billion, makes the machines that companies from 23andMe to Ancestry rely on for their sequencing.

Pacific Biosystems:
single molecule long-range sequencing

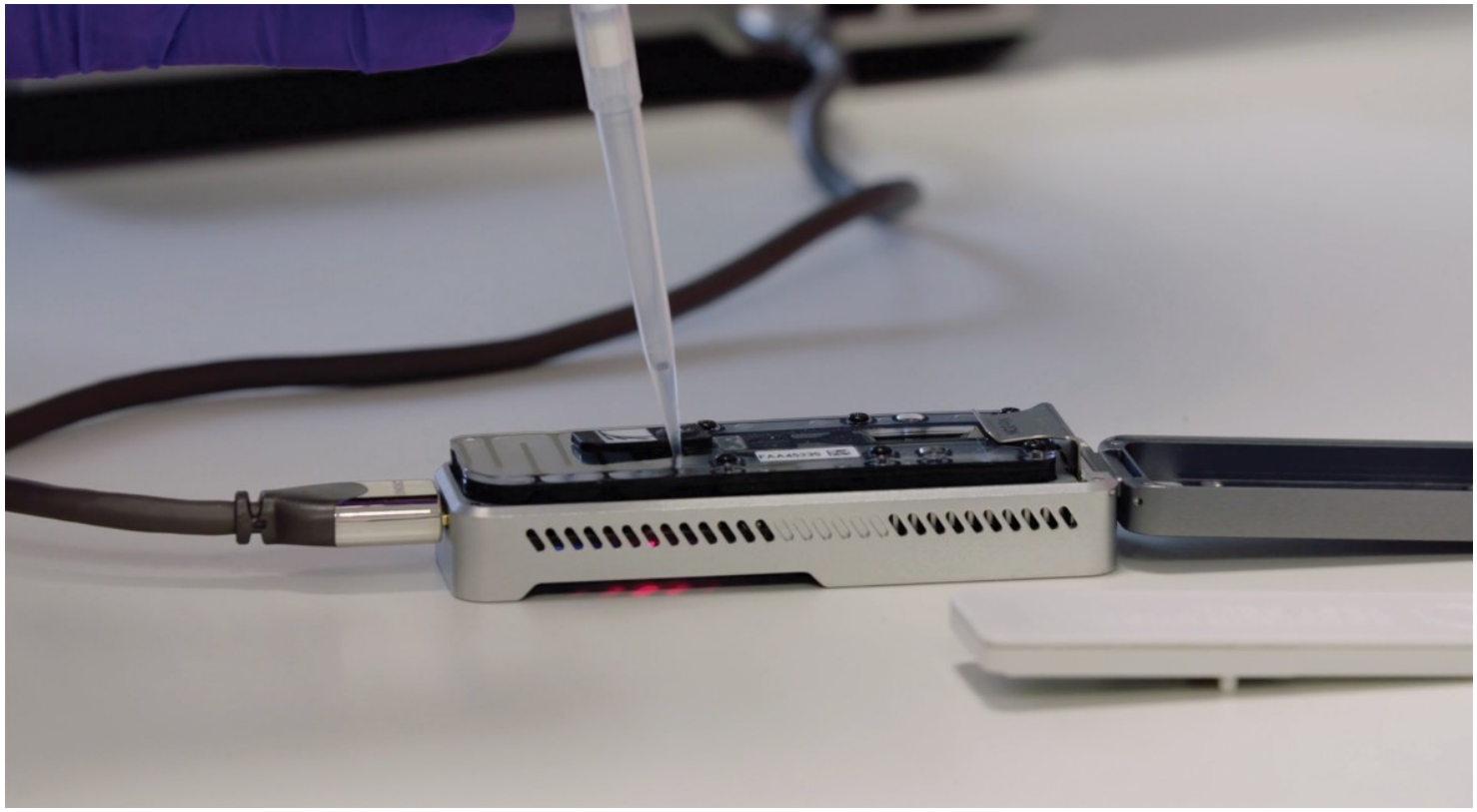


- **read length of > 5000 Bp !!**
- **long-range sequence information for de novo genome sequnencing !!**

but

- **expensive machine**
- **low to medium throughput**
- **high error rate (up to 20%)**

Nanopore sequencing: towards single molecule detection



Oxford Nanopore „MinION“

Oxford Nanopore MinION

Pro

- Small (90g only)
- Cheap (1000 $)
- Very long reads (up to 1 MBp)
- Can be used anywhere!
- $1D^2$ Reads (old = 2D)
- direct RNA sequencing

Con

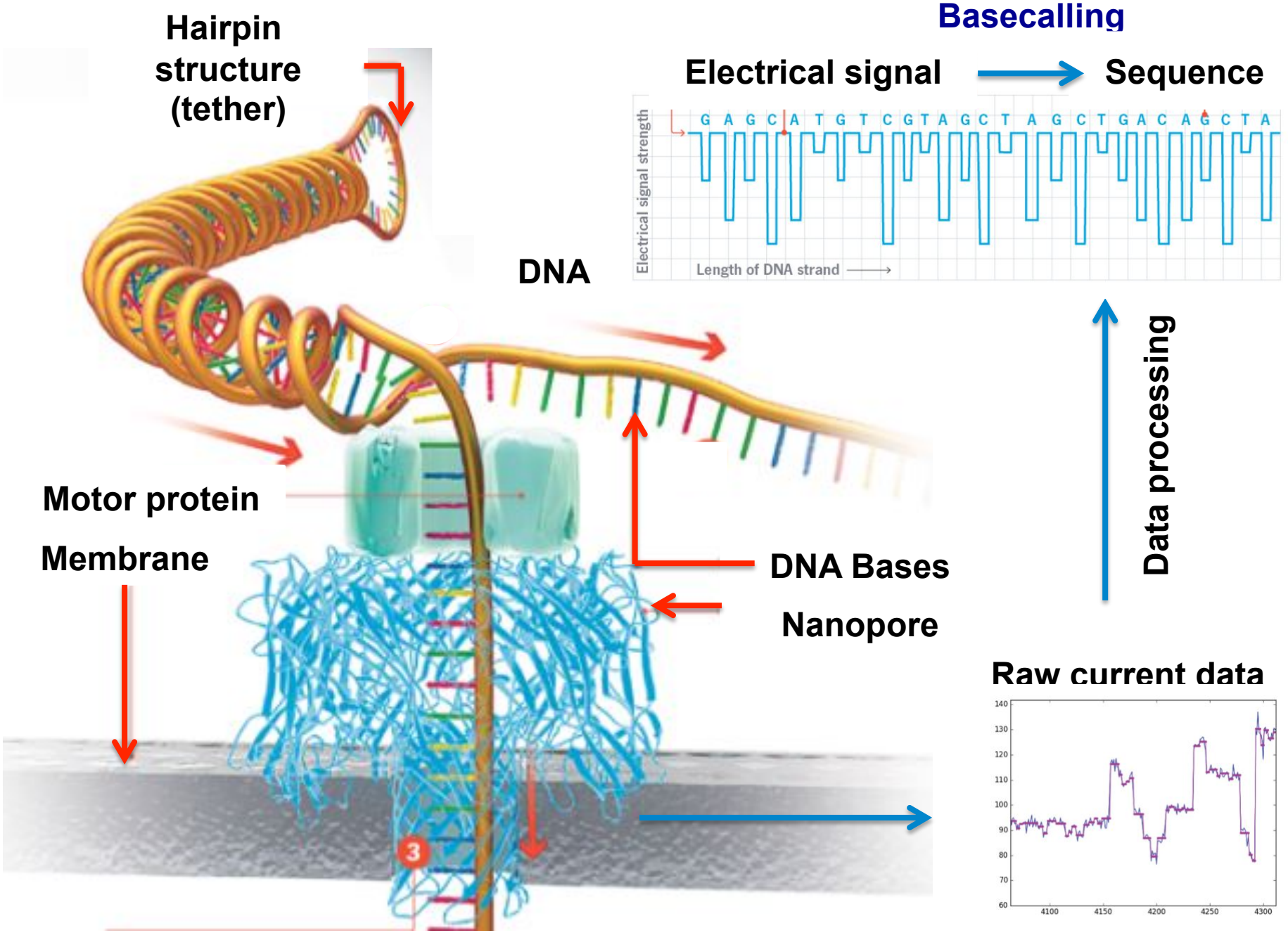
- Low throughput
- High error rate

Ebola research in Guinea

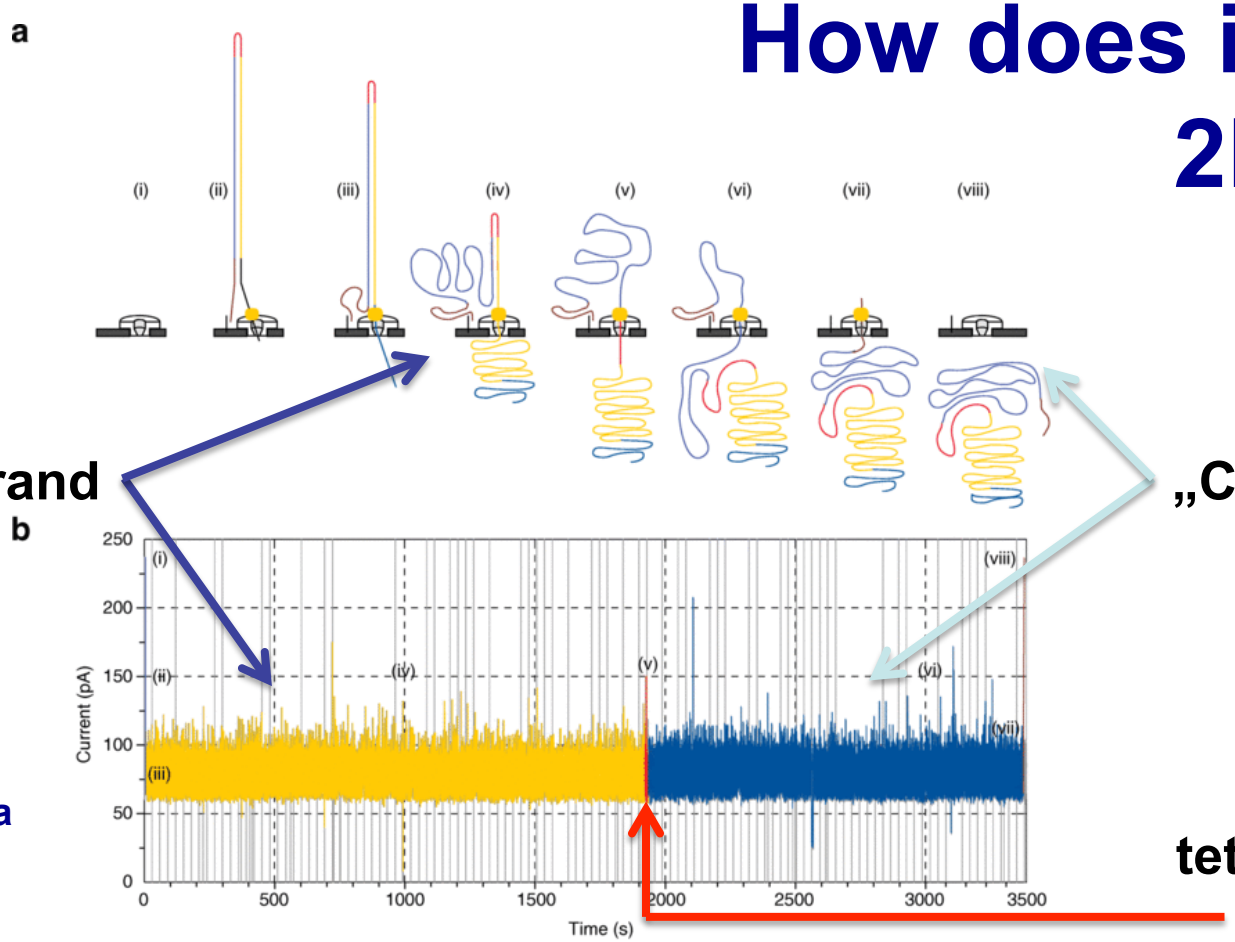
By MUSE/Science Museum of Trento.

First sequencing run in space

Astronaut Kate Rubins (2016)


Hairpin structure (tether)
Basecalling
Electrical signal
Sequence
G A G C A T G T C G T A G C T A G C T G A C A G C T A
Electrical signal strength
Length of DNA strand
DNA
Data processing
Motor protein
Membrane
DNA Bases
Nanopore
Raw current data
140
130
120
110
100
90
80
70
60
4100
4150
4200
4250
4300

How does it work?
2D-reads



„Watson“ strand

„Crick“ strand

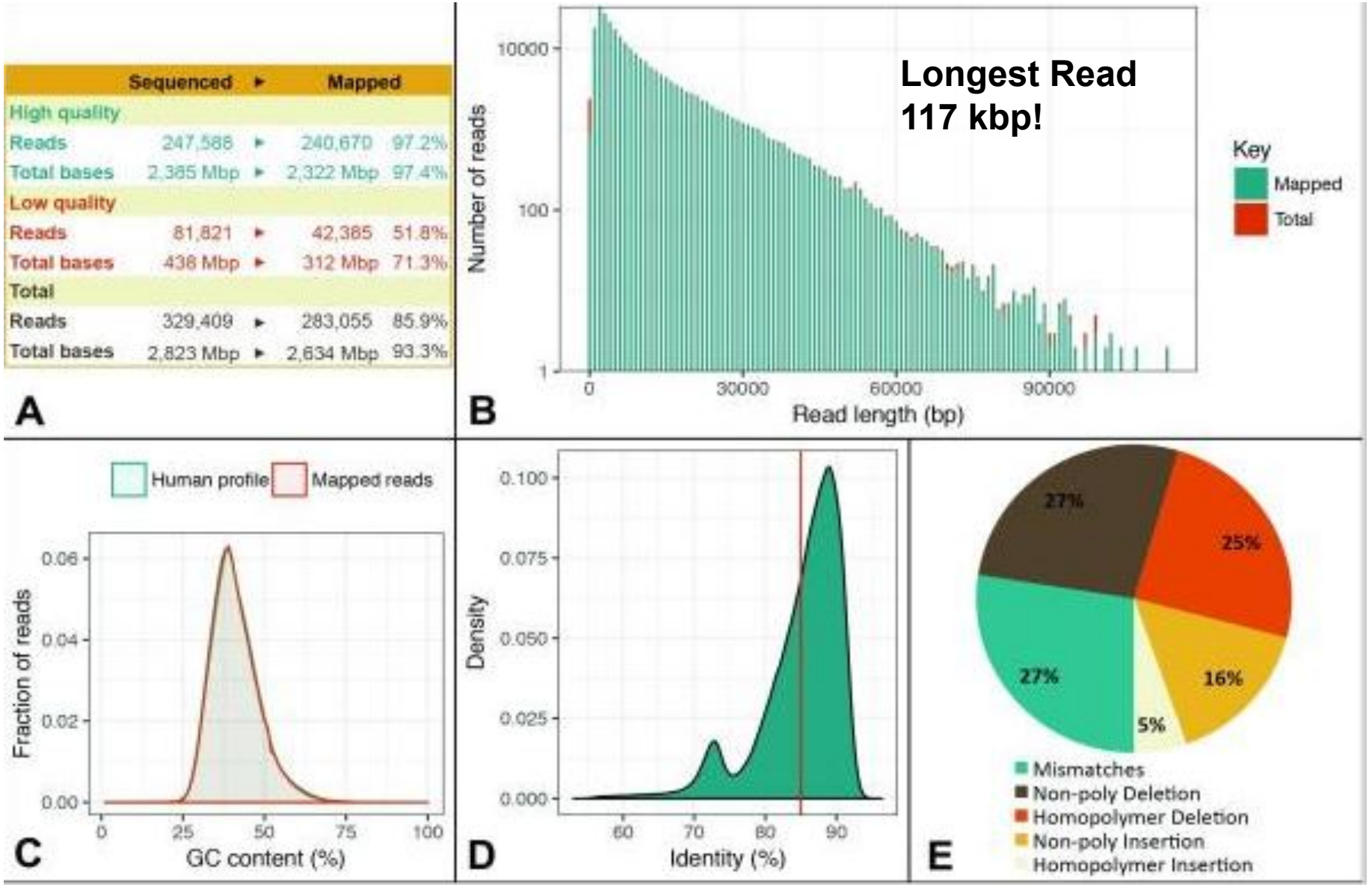
Raw current data
(squiggle data)

tether

processed electrical
signal

Jain et al, 2016

Statistics of a human* genome assembly

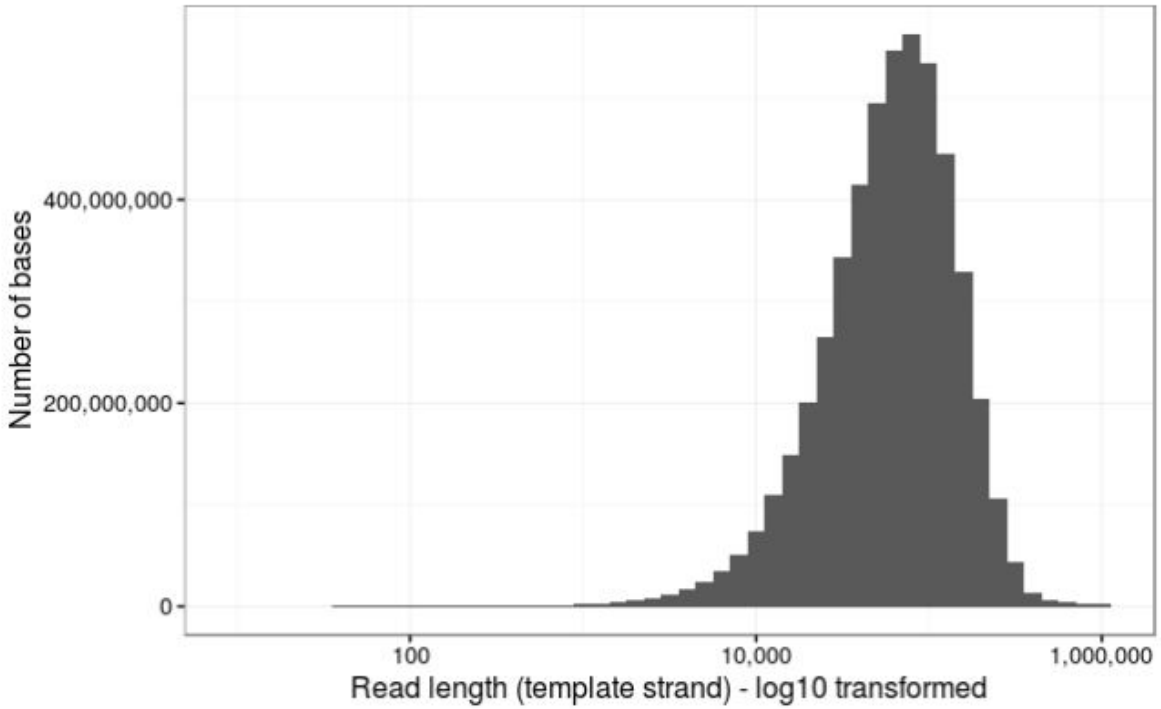


*HAP1 cells (haploid) | *Robust long-read native DNA sequencing using the ONT CsgG Nanopore system 2017*

Nanopore records

- **1x coverage of *E.coli* chromosome (4.6 Mbp) with the 7 longest reads**
- **longest read 882 kbp**
- **1/6 of *E.coli* genome with one read**
- **N50 = 63 747 bp**

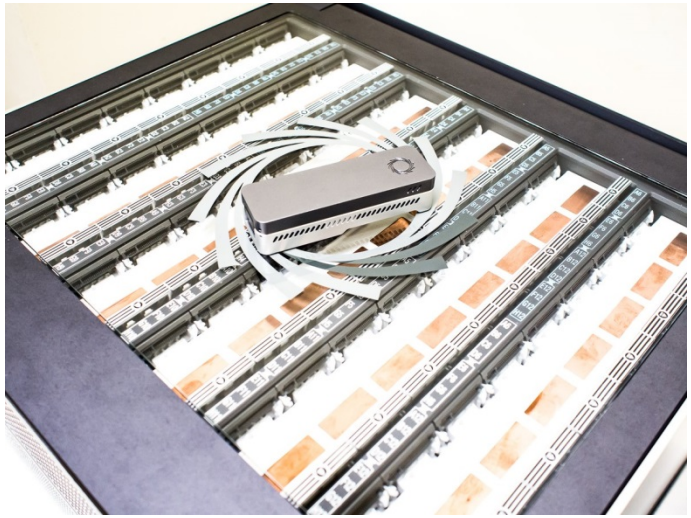
(Loman Labs 2017)



The future...

- Output = ?
- Mobile DNA analysis for everyone and anywhere

PromethION



SmidgION

- 48 Flowcells with 2048 Channels (= 192 MinIONs)
- 6-11 Tbp Output / 24h
- Direct RNA sequencing?
- Direct 5-mC sequencing