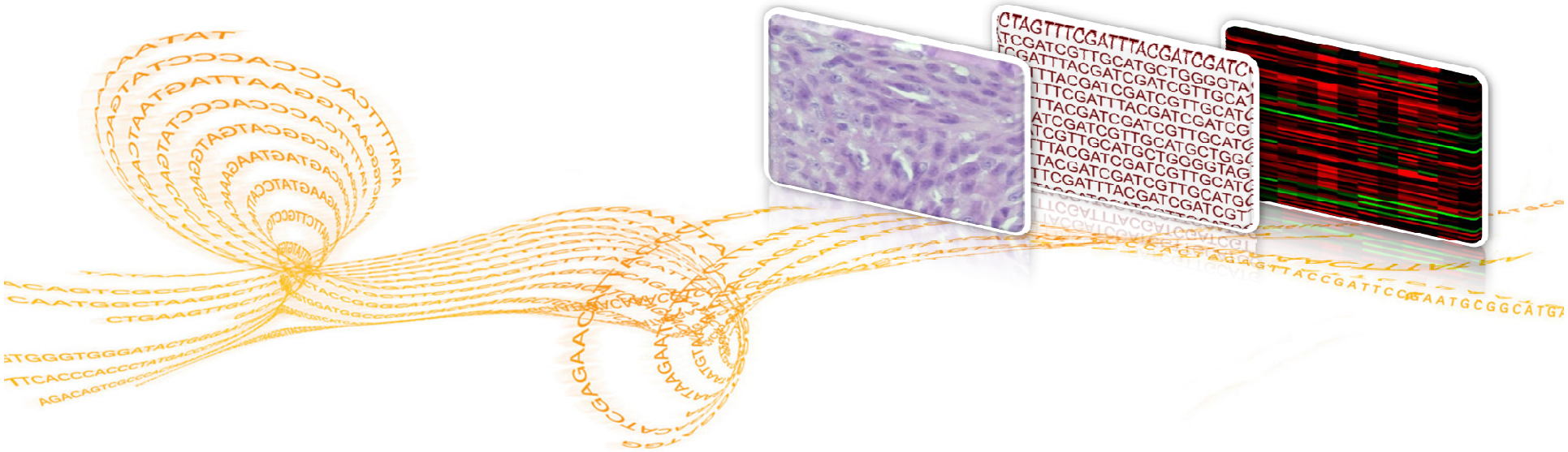


Illumina's next generation sequencing technology

Presented by field applications scientist
Pernille Albertus
Denmark/Norway

illumina®

Illumina

- headquarter in San Diego, California
- 1800+ employees globally
- develop and sell innovative technologies for studying genetic variation and function enabling rapid advances in disease research, drug development, and the development of molecular tests in the clinic
- founded in 1998 (GoldenGate genotyping)
- acquired Solexa in 2006 (Sequencing By Synthesis)

Illumina Sequencers

Next Generation Sequencing made accessible.	Two proven technologies. One powerful platform.	Most widely adopted NGS platform.	Redefining the trajectory of sequencing.

GA_{IIe}	HiScanSQ	GA_{IIx}	HiSeq2000

illumina®

Illumina Array Platforms

Low- to mid-plex molecular testing.	Dedicated array instrument.	Sequencing-compatible array instrument.	Two proven technologies. One powerful platform.

BeadXpress	iScan	HiScan	HiScanSQ

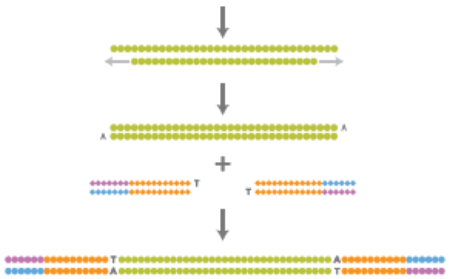
illumina®

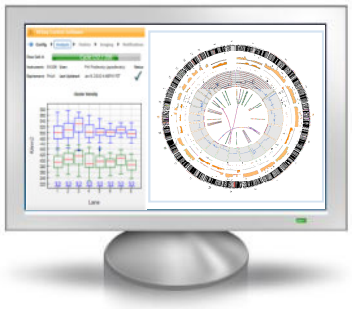
Sequencing by synthesis chemistry

illumina®

Workflow

SAMPLE PREP	cBot CLUSTER GENERATION	Genome Analyzer SEQUENCING	DATA PROCESSING & ANALYSIS





illumina

The flow cell - a core component

EVERYTHING EXCEPT SAMPLE PREPARATION IS COMPLETED ON THE FLOW CELL

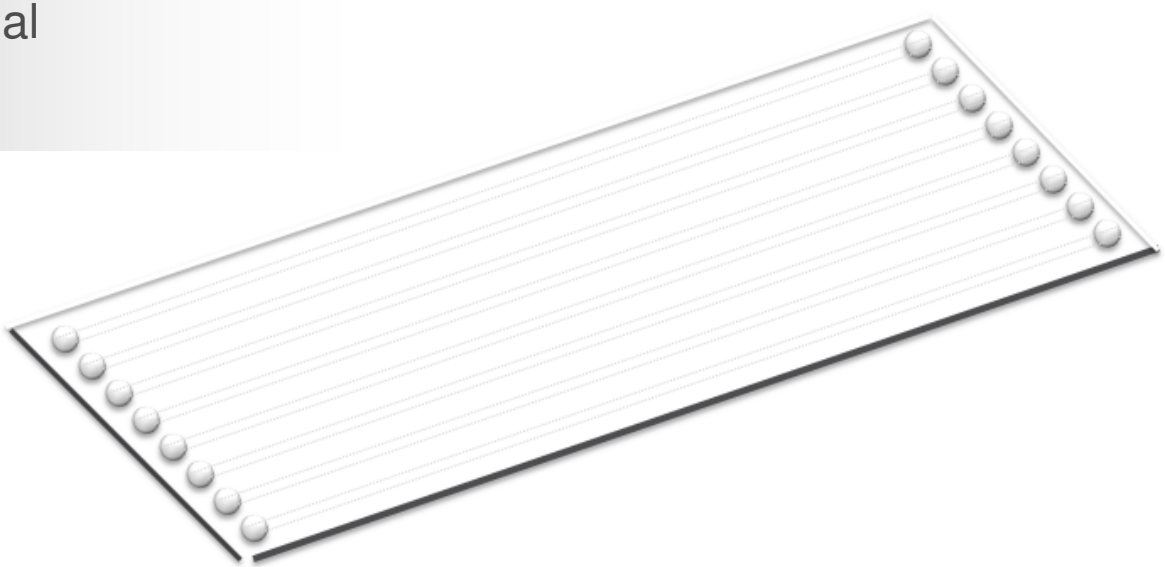
template annealing (1 - 96 samples)

template amplification

sequencing primer hybridization

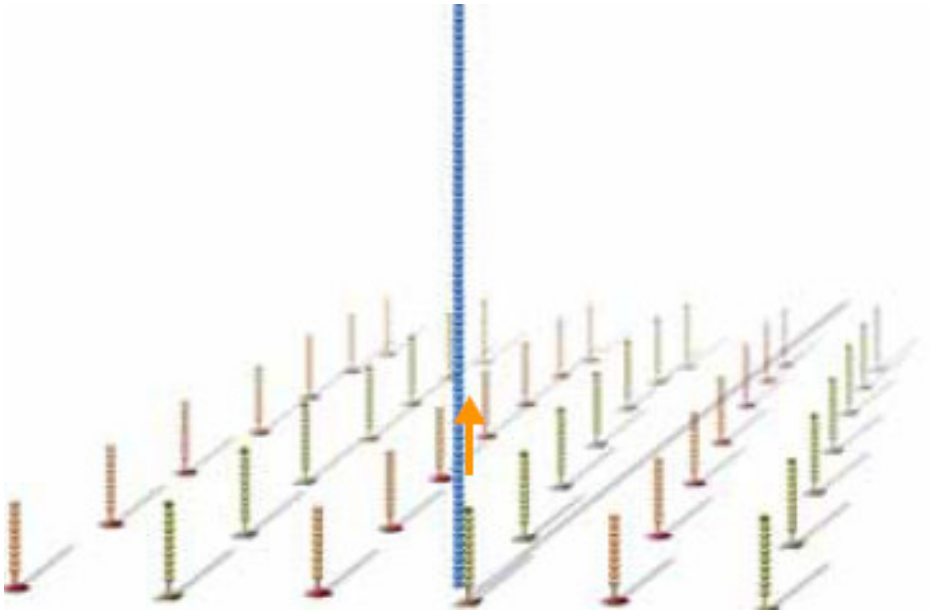
Sequencing-by-synthesis reaction

generation of fluorescent signal



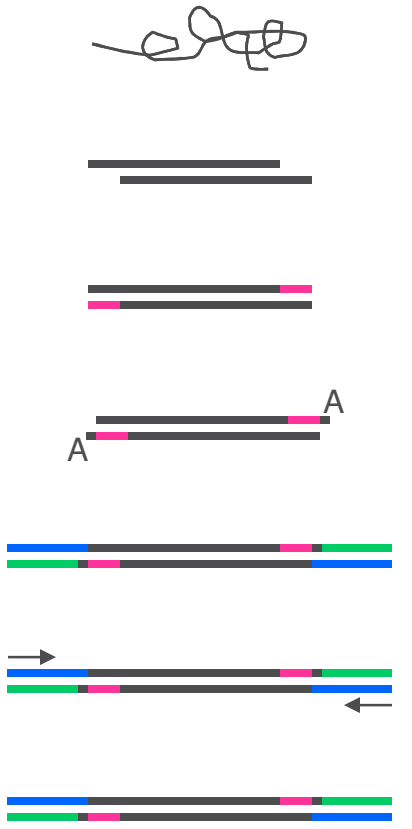
illumina

The flow cell surface is coated with oligos



illumina

Preparation of template



template DNA

⇩

fragment

⇩

repair ends

⇩

add A overhang

⇩

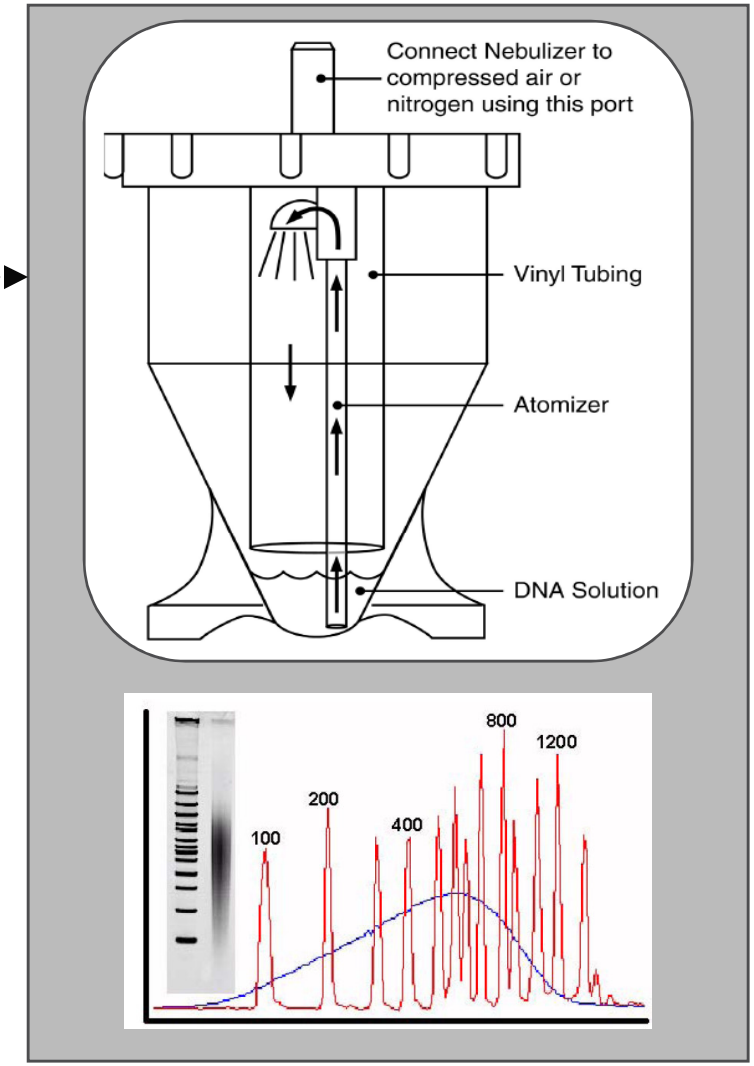
ligate adaptors & purify on gel

⇩

enrich

⇩

genomic library & library QC



illumina

The flow cell is mounted on the cBot

AUTOMATICALLY

loads library into the lanes of the flow cell

amplifies templates

anneals sequencing primer to templates

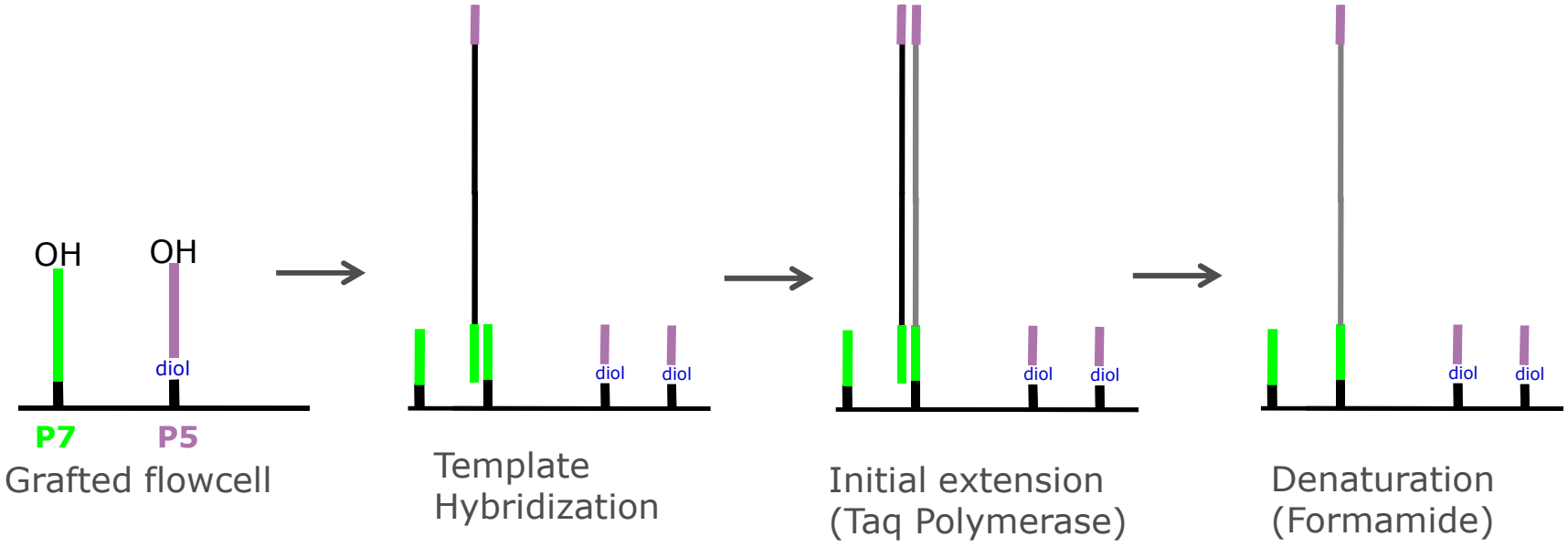
FEATURES

intervention-free clonal amplification in 4 hours

simple touch screen operation

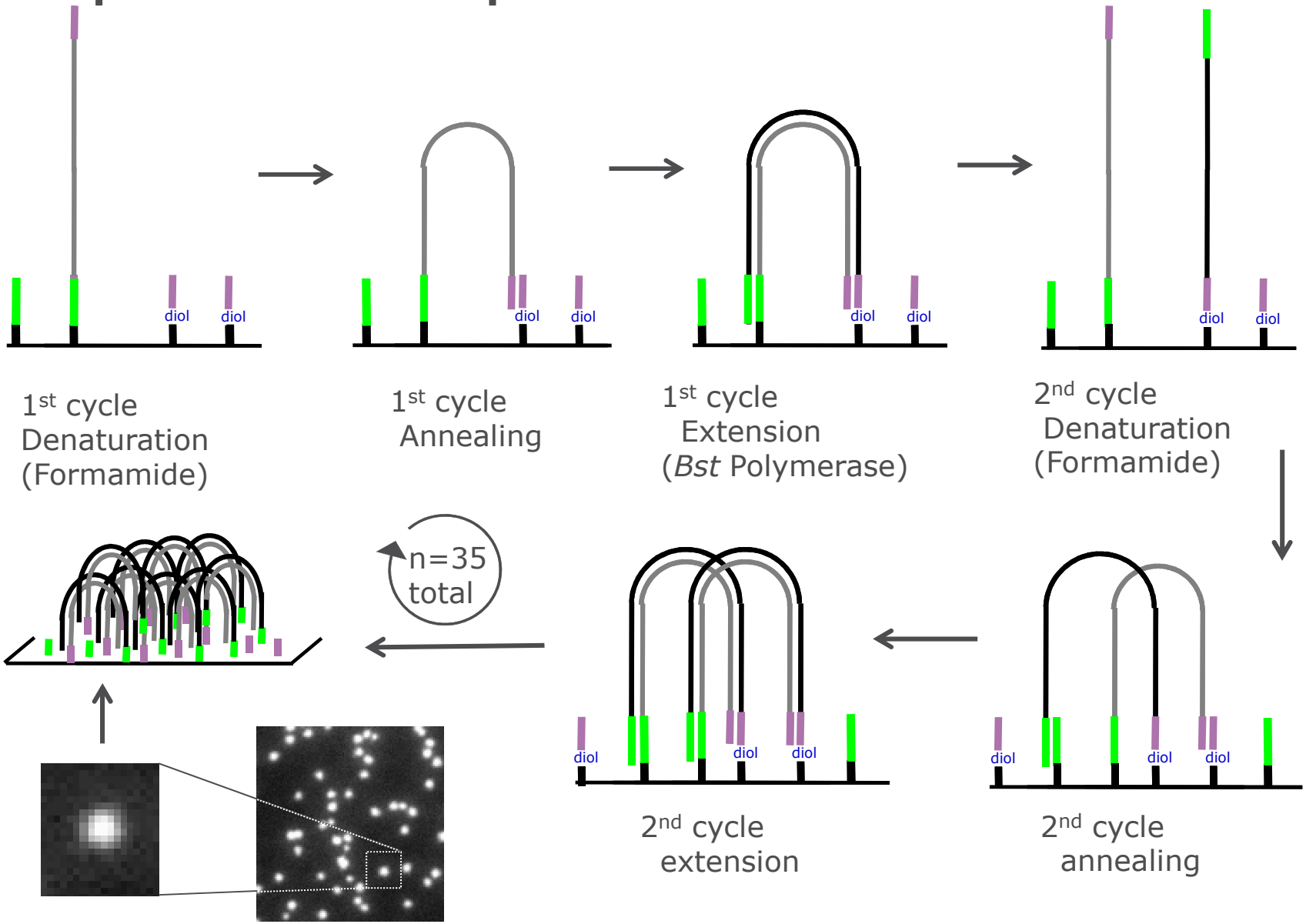
illumina

Hybridization of template



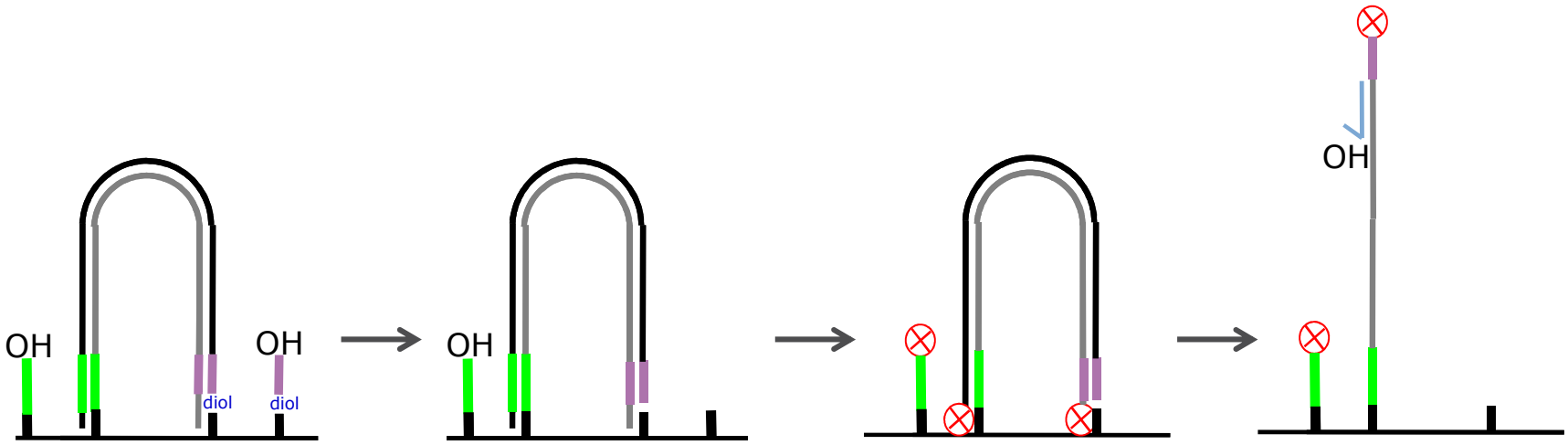
illumina

Amplification of template



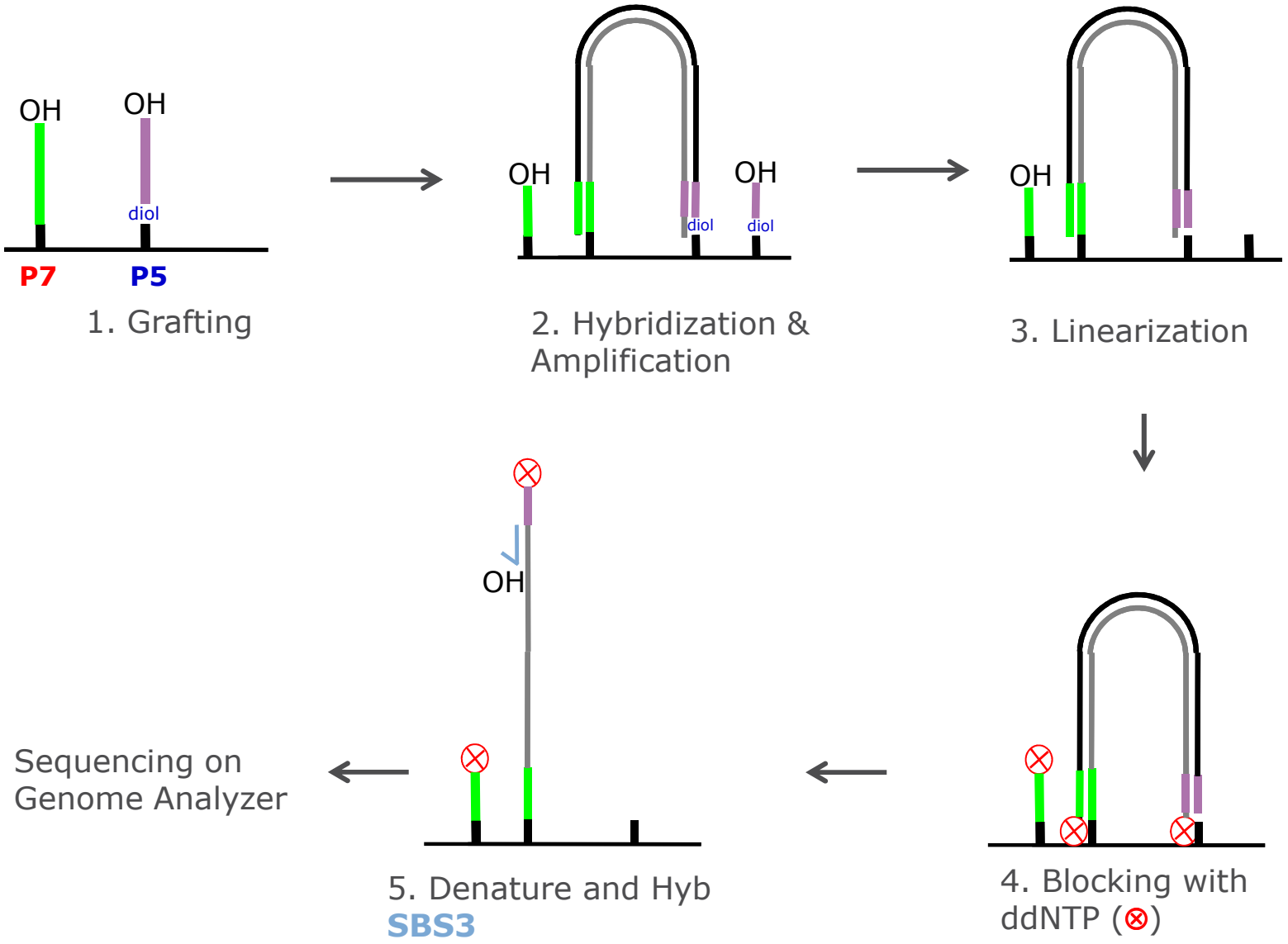
illumina

Annealing of sequencing primer to template



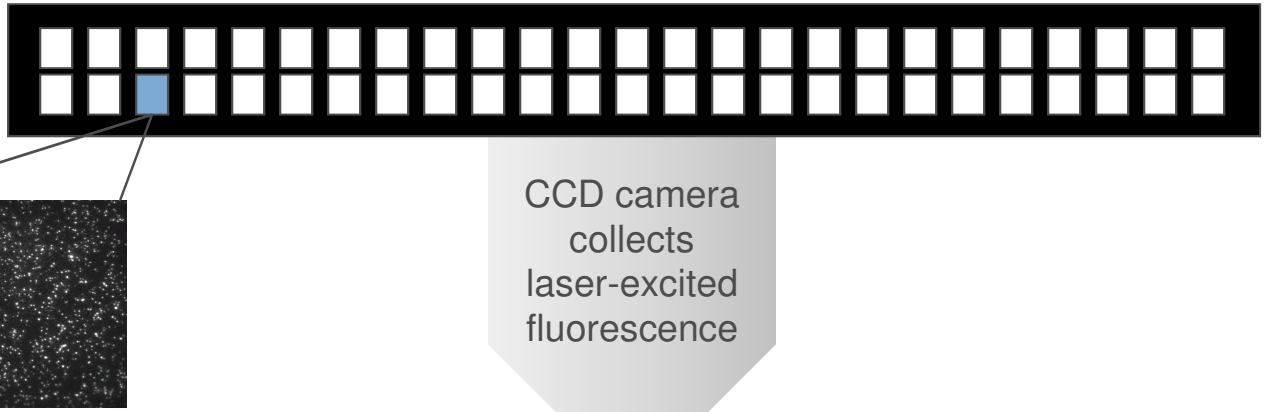
illumina

Summary - "cluster generation"



illumina

The flow cell is mounted on the sequencer



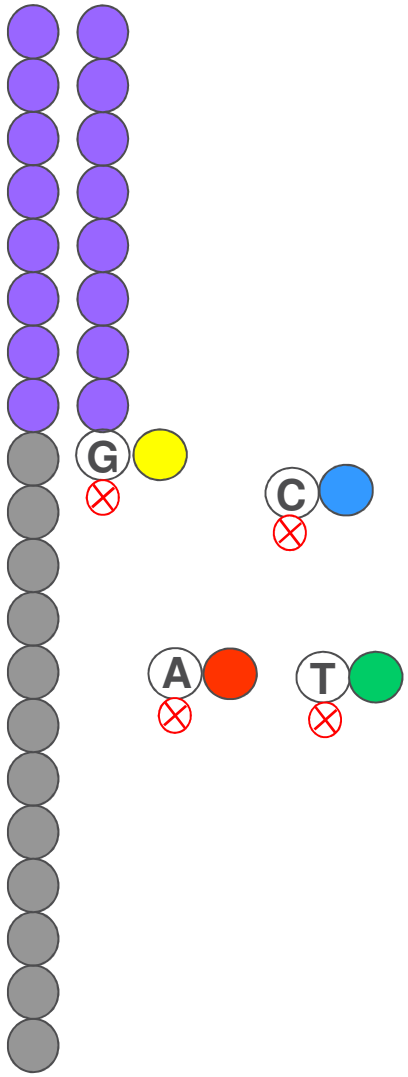
CCD camera collects laser-excited fluorescence

sequencing reagents pass through the 8 lanes inside the flow cell

sequencing reaction is temperature controlled

illumina

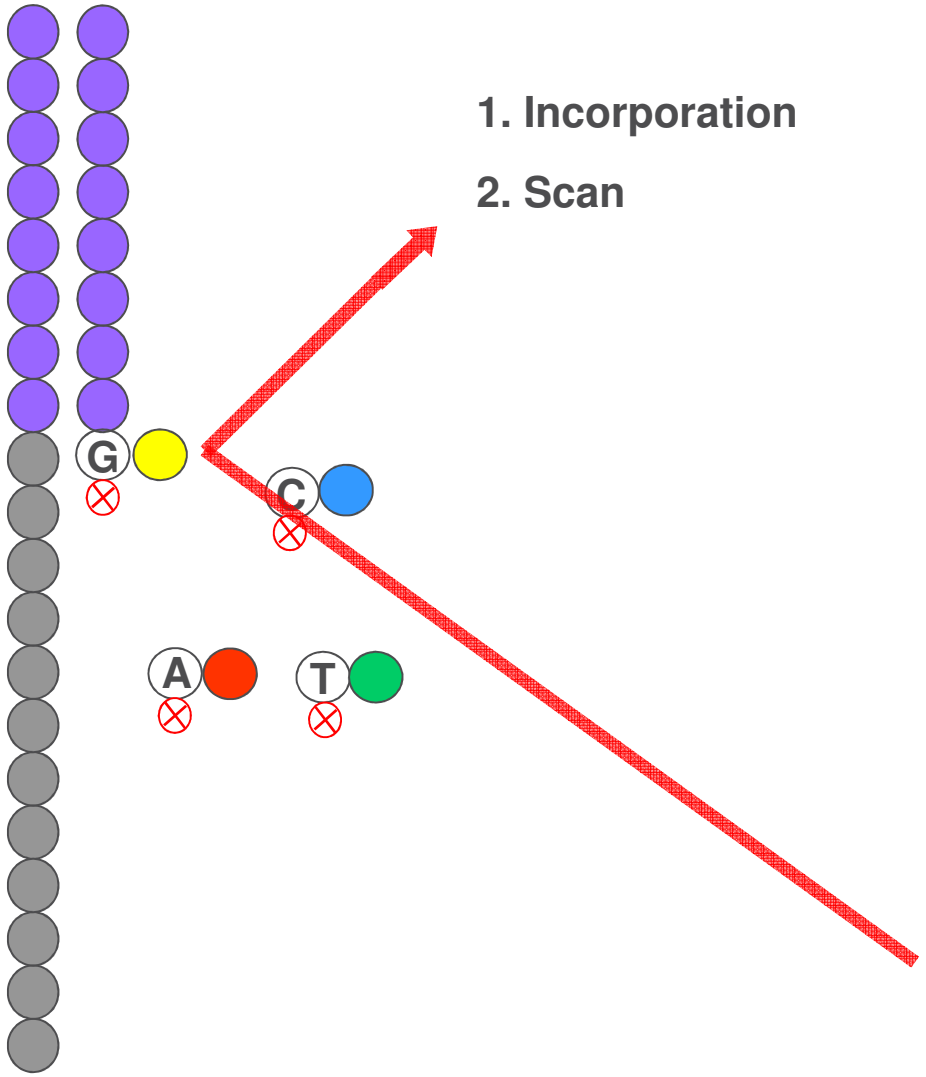
Incorporation



1. Incorporation

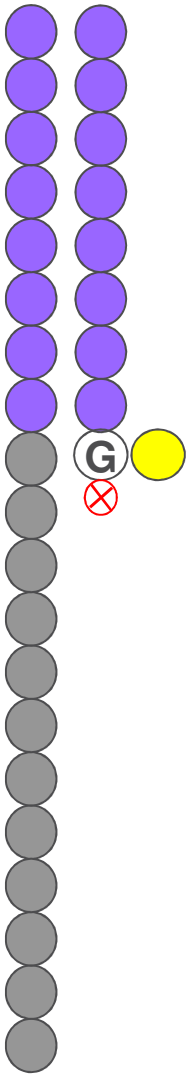
illumina

Scanning



illumina

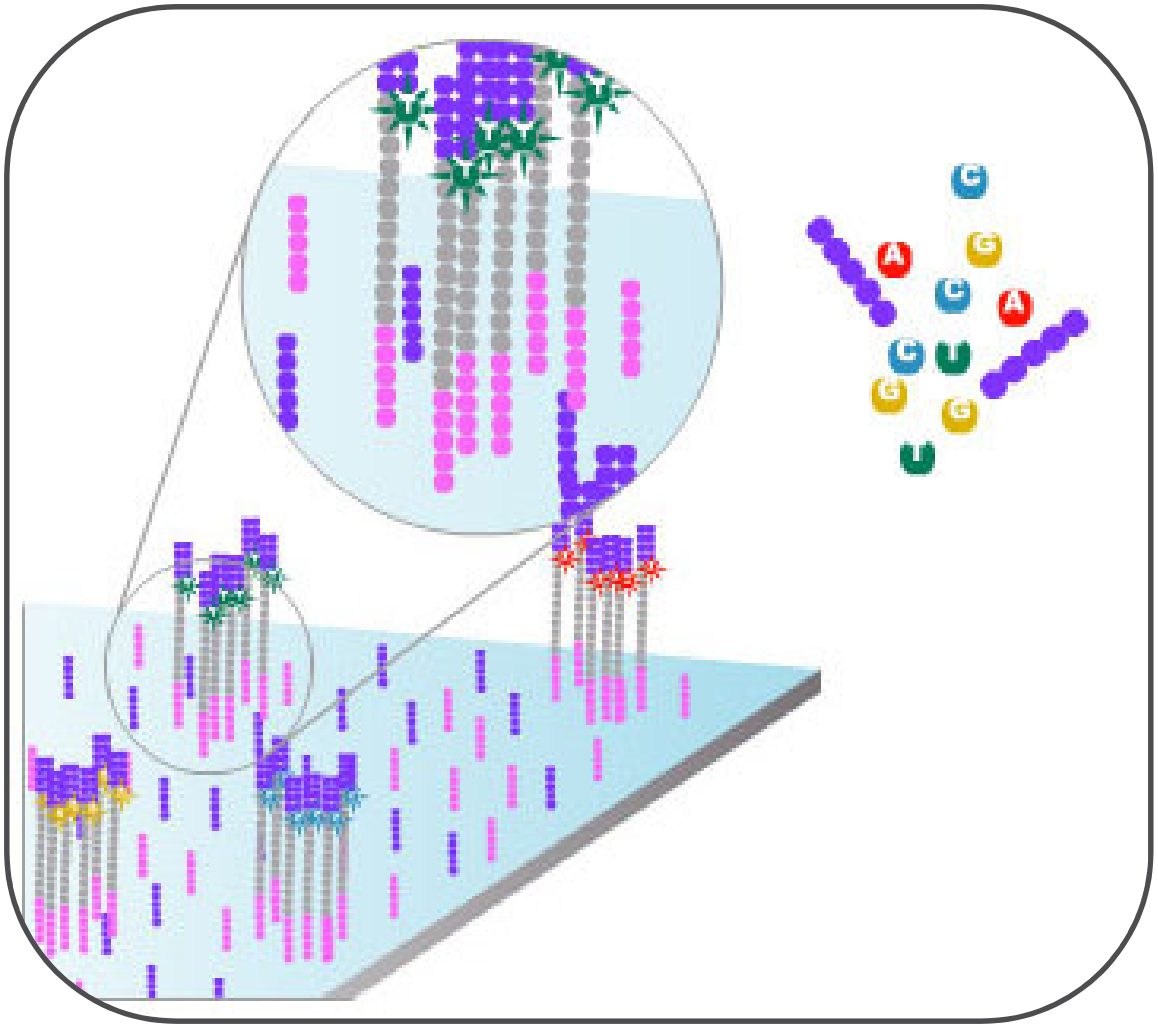
Cleavage



1. Incorporation
2. Scan
3. Cleavage

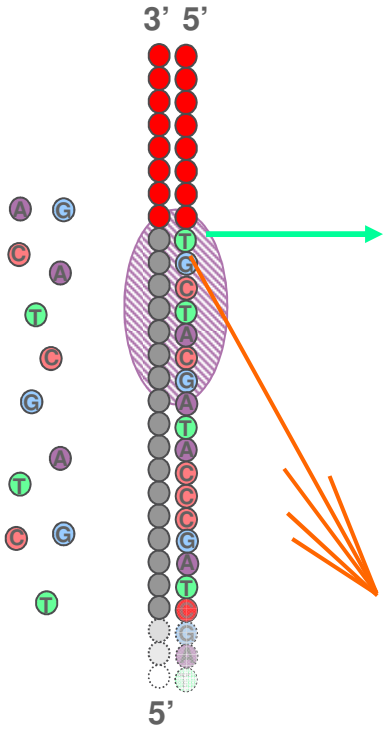
illumina

Millions of clusters are sequenced in parallel



illumina

A picture is taken every time a new base is added



Sequencing

36bp – 100bp

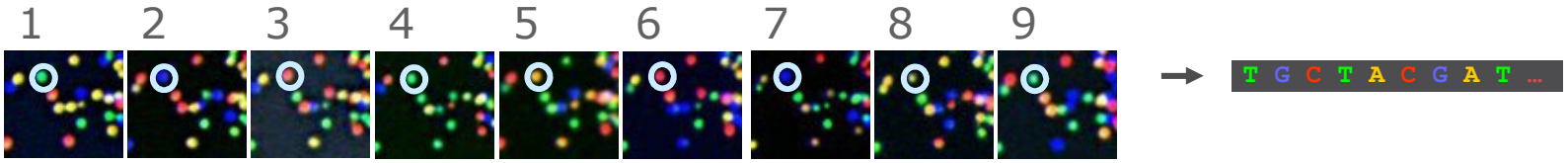
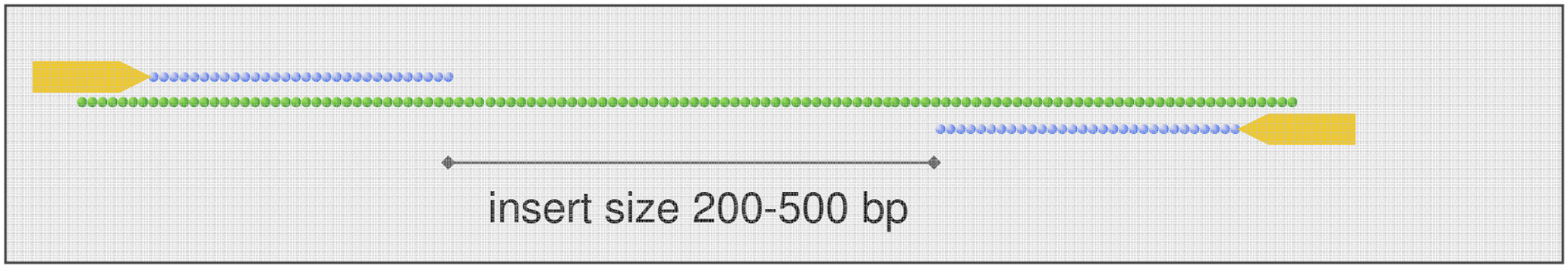


Image acquisition

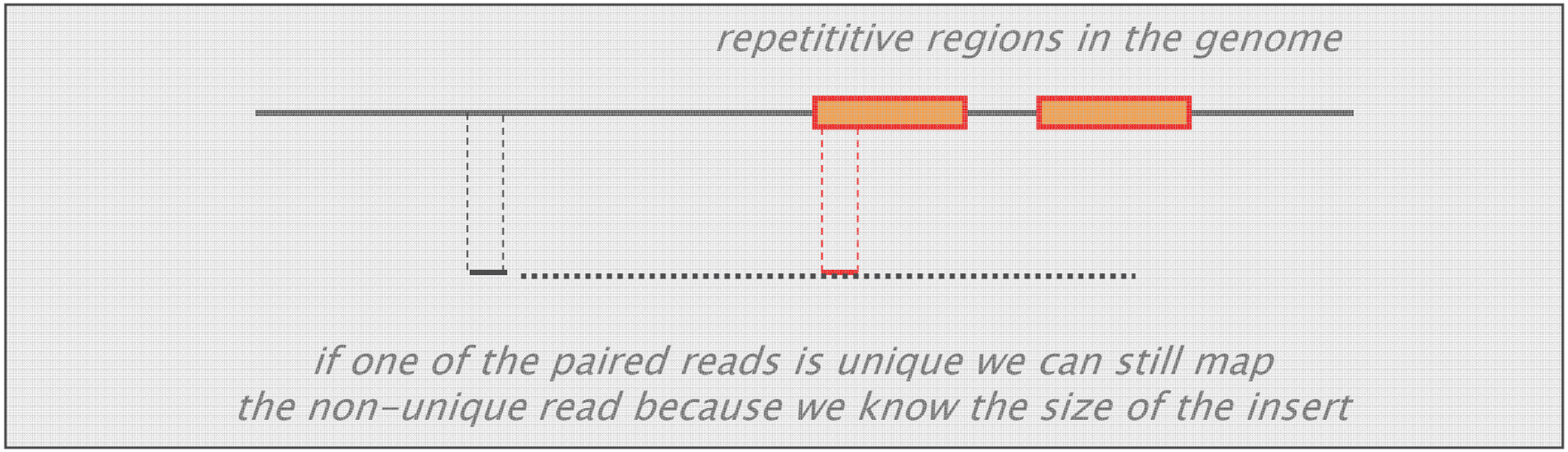
Base calling

illumina

"Paired-end" sequencing - a core concept

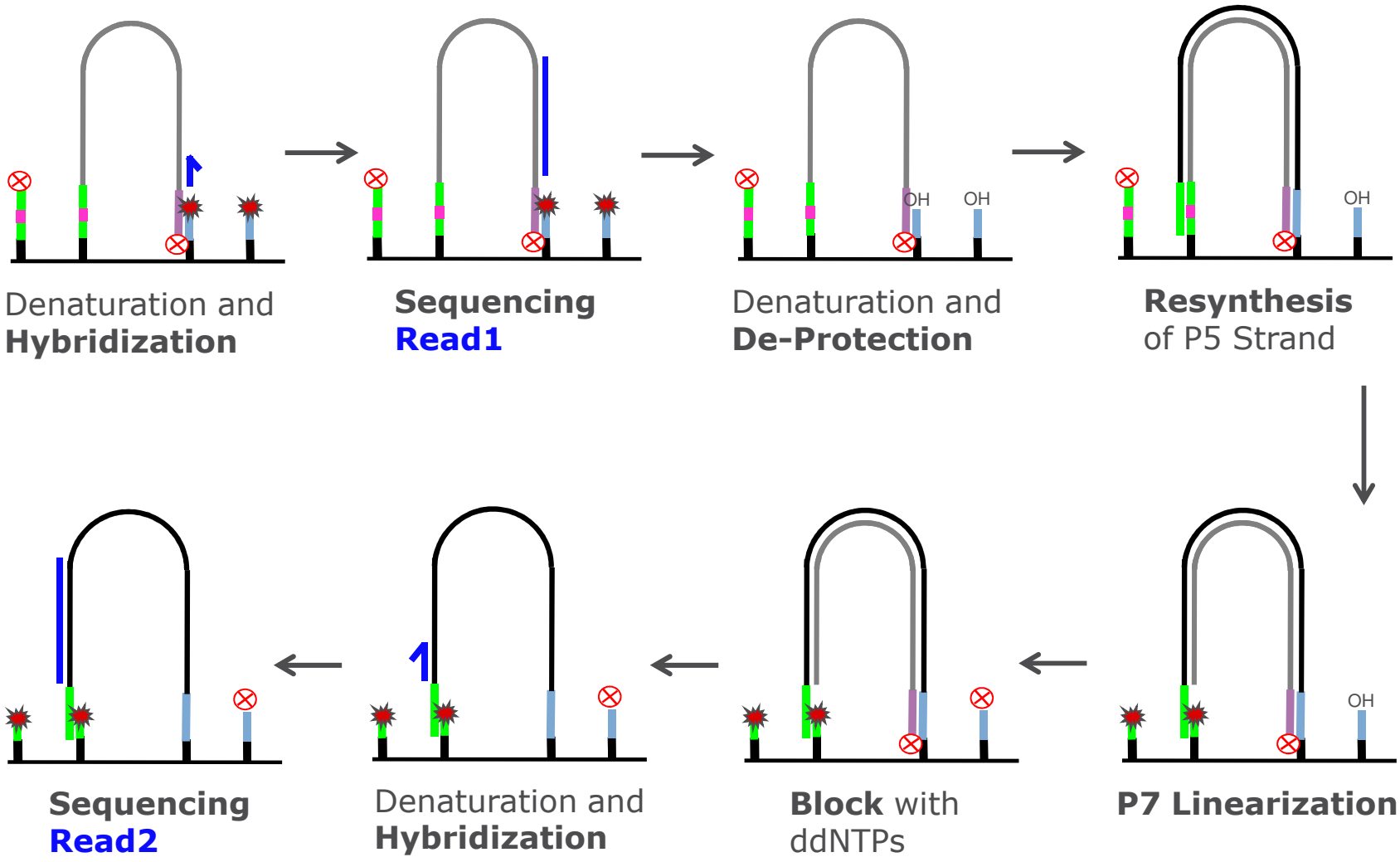


- allows unique mapping of more data
- combined with single reads and mate pair complex structural changes can be discovered



illumina

Hybridization of second sequencing primer is done *in-situ* on the sequencer



illumina

Instrument specifications and throughput

illumina®

Illumina Sequencer for Everyone!

Next Generation Sequencing made accessible	Unique combination of sequencing & arrays	Most widely adopted NGS platform	Changing the trajectory of sequencing
			
GA_{IIe}	iScanSQ	GA_{IIx}	HiSeq 2000

illumina®

Genome Analyzer IIx

Most widely adopted NGS platform

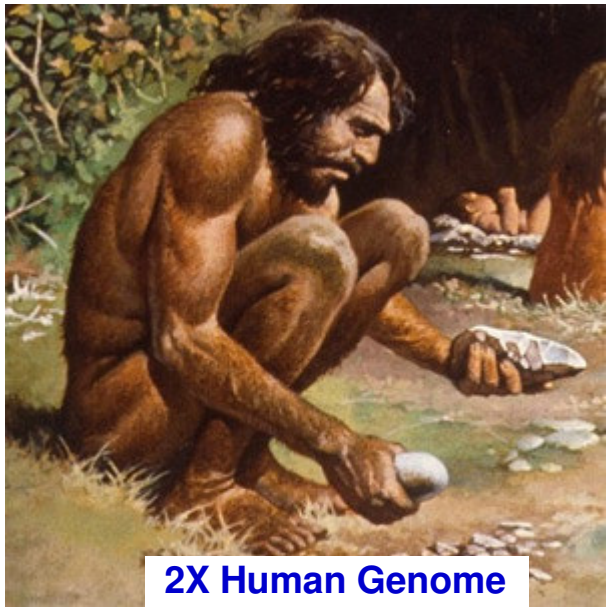
illumina

Genome Analyzer$_{IIx}$ Performance Specifications

Performance Parameters
50 Gb of high quality data / run
5 Gb / day
500 M reads per paired-end run
2 x 100 bp supported read length
Raw Accuracy: ≥ 98% (2 x 100) ≥ 99% (2 x 50)
Run Time: 2 x 100 bp in 9.5 days 2 x 50 bp in 5 days 1 x 35 bp in 2 days
Consensus accuracy 99.999%
12 to 96 multiplex sequencing/channel

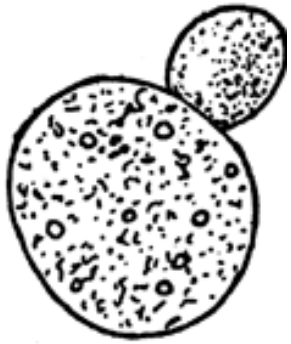
illumina

How much can you do with just *one lane* of GA data?



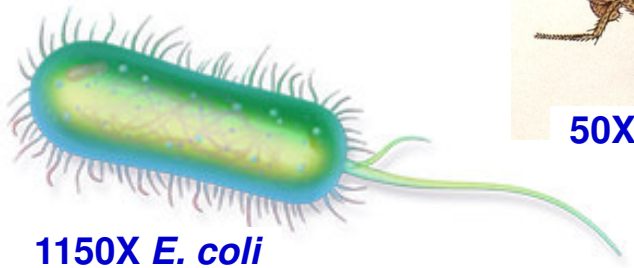
2X Human Genome

50X Arabidopsis



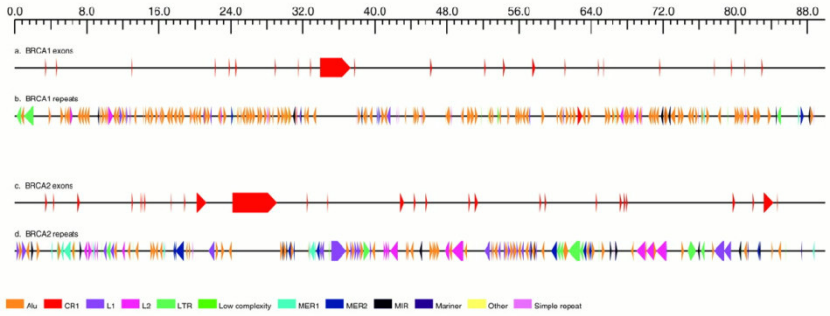
500X Yeast Genome

50X Drosophila



1150X *E. coli*

3000X BRCA1+BRCA2, 12 samples per lane



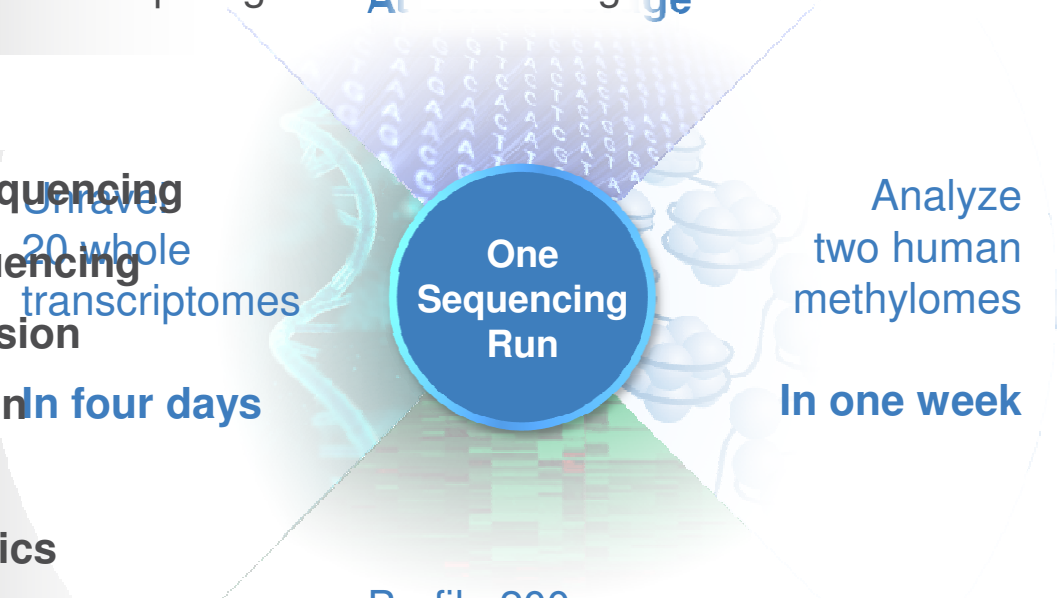
illumina

What if, in one sequencing run you could…

SIMULTANEOUSLY

Run multiple applications requiring different read lengths

- **Whole genome sequencing**
- **Targeted resequencing**
- **Gene expression**
- **Methylation**
- **De novo**
- **Metagenomics**
- **ChIP-seq**
- **Whole transcriptome**



Sequence one cancer & one normal genome

Unravel 20 whole transcriptomes

In four days

Analyze two human methylomes

In one week

Profile 200 gene expression samples

In less than two days

illumina

HiSeq 2000

OUTPUT

Initially capable of up to 200 Gb per run

DATA RATE

~25 Gb/day

7-8 days for 2 x 100 bp

NUMBER OF READS

One billion single-end reads*

Two billion paired-end reads*

*Based on one billion clusters passing filter

illumina

HiSeq 2000

Comparison with the Genome Analyzer

	HiSeq 2000 (at launch)	GA*IIx* (at 50G)	GA*IIx* (at 95G)
Gb per run	150-200	50	95
Gb per day	20-25	5	7
Cluster density in KClusters/mm²**	260-350	490	620
Read length	2 x100	2 x100	2 x150
Available surface area (mm²)*	2880	510	510

*GA*IIx* with single surface, single FC, HiSeq 2000 with dual surface, dual FC
**Clusters passing filter

illumina

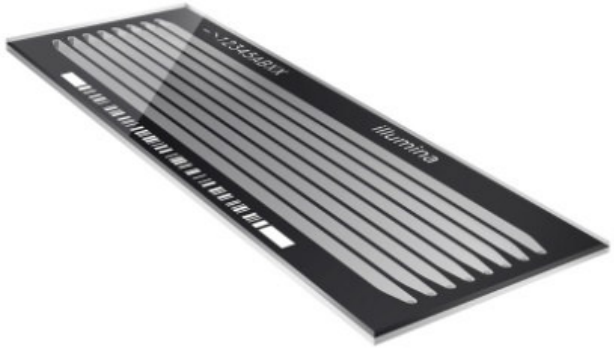
HiSeq 2000

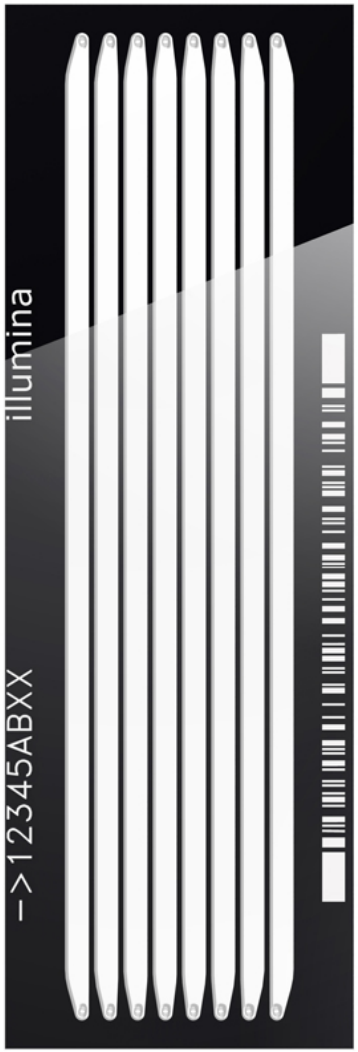
New flow cell design

LARGER, DUAL-SURFACE ENABLED

>5x increase in imaging area

Retains 8 lane format





illumina

HiSeq 2000 dual flow cell design

TWO INDEPENDENT FLOW CELLS

Simultaneously run applications that require different read lengths

Run in single or dual flow cell mode

SIMPLE FLOW CELL LOADING

Flow cells held by vacuum

No oil needed

LED switch ensures correct connection

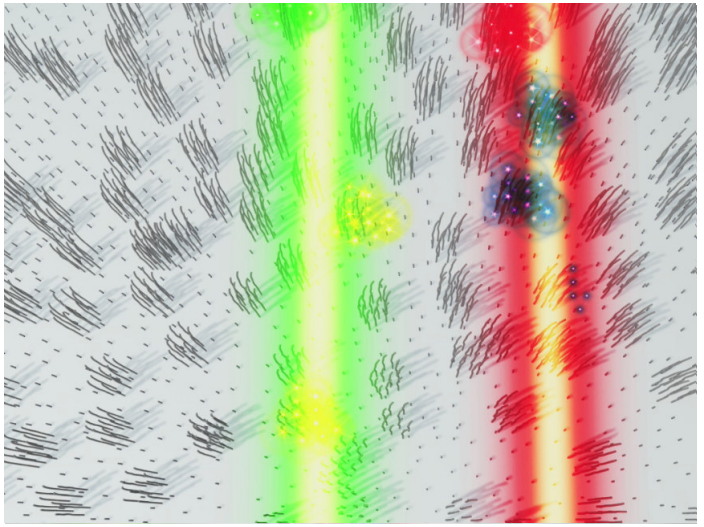
illumina

Dual surface imaging

Cutting-edge imaging technology

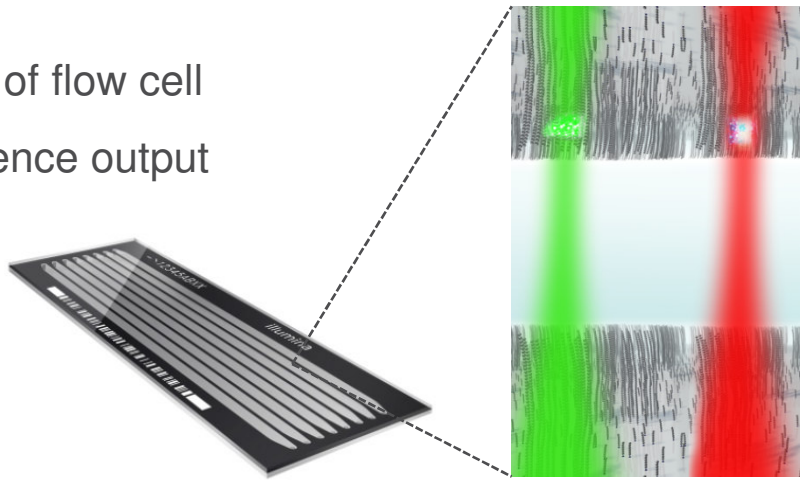
TDI line-scanning technology with four CCDs for imaging

Fastest scanning and imaging method



Images clusters grown on both surfaces of flow cell

Huge gain in number of reads and sequence output

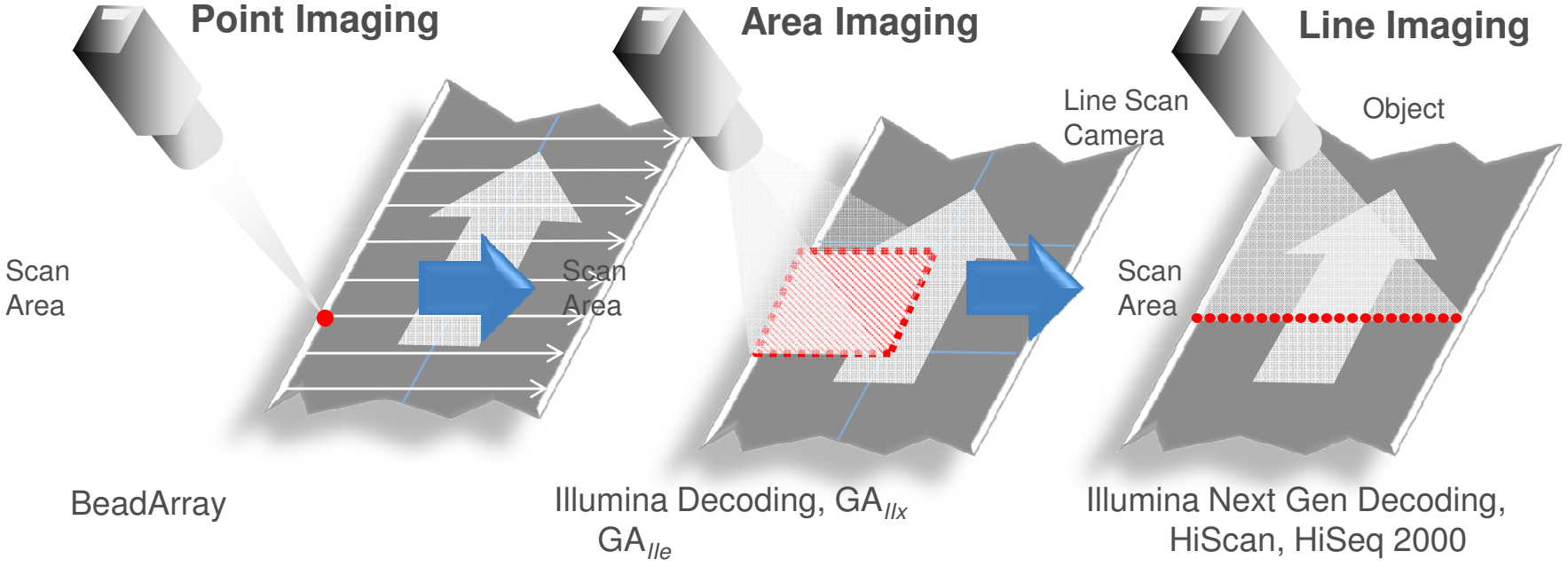


illumina®

The power of line scanning

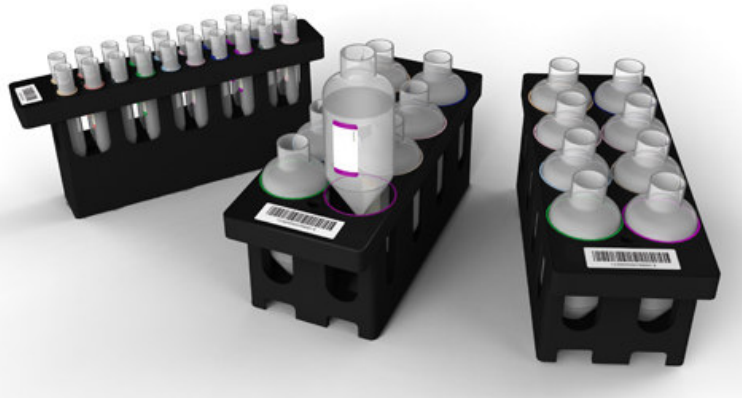
Maximizing data rate

	Point Imaging	Area Imaging	Line Imaging
Stage & filter movement delays	+	–	+
Data transfer delays	+	–	+
Practical data acquisition limit	–	+	+
Data quality/background rejection	+	–	+



illumina

HiSeq 2000
Plug-and-play reagents



PRE-CONFIGURED SEQUENCING REAGENTS

Only two minutes hands-on time

Up to 200 cycles per flow cell

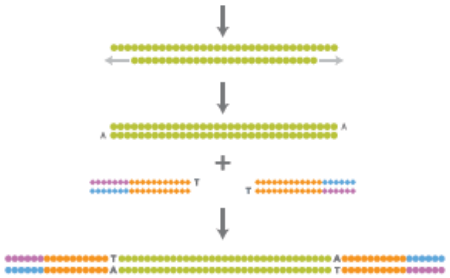
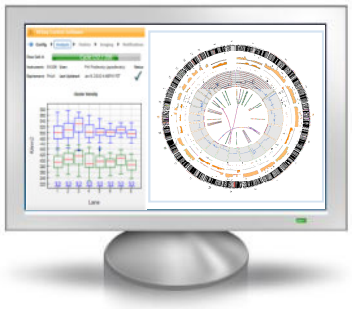
Bar-coded for tracking

Temperature-controlled compartment

Integrated paired-end fluidics

illumina

Workflow

SIMPLIFIED SAMPLE PREP	cBot CLUSTER GENERATION	Genome Analyzer SEQUENCING	DATA PROCESSING & ANALYSIS
			

Data management and analysis

illumina®

Instrument computer specifications

INSTRUMENT CONTROL COMPUTER (HISEQ)

Base Unit: 2x Intel Xeon X5560 2.8 GHz CPU

Memory: 48 GB RAM

Hard Drive: 4x 1.0 TB 7200 RPM SATA

Operating System: Windows Vista

DATA ANALYSIS COMPUTER

HP ProLiant DL580 G5 Rack Server (any 64-bit Unix)

Red Hat Linux

Four quad-core 2.93GHz 64-bit Intel Xeon processors

32 GB fault-tolerant RAM

Data analysis flow

GENERATING SEQUENCING IMAGES

PERFORMING IMAGE ANALYSIS

cluster positions / intensities / noise

BASE CALLING

cluster sequence

quality calibration

filtering results

INSTRUMENT PC

PRIMARY ANALYSIS

SCS

DEMULTIPLEXING

ALIGNING TO REFERENCE GENOME

DETECTING VARIANTS AND COUNTING

expression levels of exons, genes, splice variants

LINUX SERVER

SECONDARY ANALYSIS

CASAVA

VIEWING RESULTS

build consensus sequence

call SNPs

detect indels

count RNA reads

ANY PC

GENOMESTUDIO

illumina

qseq.txt file

- Tab-delimited: easy to parse, easy to import into databases

Instrument	Run ID	Lane	Tile	X-coord	Y-coord	Index #	Read #	Sequence	ASCII Character Q-score	PF (0,1)
IL10	2732	4	25	1782	1896	0	1	ACGCTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGAATTGGCAC	\bbKbba_aaaaab`baba_aaab^b_`ab_bWabW__`aaaZ_aTTY__a_]]	1
IL10	2732	4	25	1782	1250	0	1	TGTCTACAGTAGAGTCAATAGCAAGGCCACGACGCAATGGAGAAAGACGGAGAG	a\Z]aXaZQU_Ya^Y`Z_X`_WVXRWTNW\G[_BBBBBBBBBBBBBBBBBBB	1
IL10	2732	4	25	1782	1226	0	1	CGAGAGCGGTCAGTAGCAATCCAAACTTTGTTACTCGTCAGAAAATCGAAATCA	a_X\bbba[_b_b_ab``bbaba_ba`\]`_a`aaa\Va\`L`b^```SPI]a_	1
IL10	2732	4	25	1782	1757	0	1	AACGTACCTTCAAGAAGTCCTTTACCAGCTTTAGCCATAGCACCAGAAACAAAA	abbb`\bbb````ba`aab`baa^aa^`a`^`aaa``^`a`_]__a^^__Q^^	1
IL10	2732	4	25	1782	1583	0	1	CAACAGGAGCAGGAAAGCGAGGGTATCCTACAAAGTCCAGCGTACCATAAACGC	abb\]__^`b```b``abb`bab^```^`b]a]Z]]_^``]aaX`[``_`]]a`W	1
IL10	2732	4	25	1782	1223	0	1	AAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAA	a^``abb]]a`_]Yaa``aba[PY_`a]WYOR]QVLT]H`[LT[[]RGY^_`BB	1
IL10	2732	4	25	1782	999	0	1	CCAATCTGACCAGCAAGGAAGCCAAGATGGGAAAGGTCATGCGGCATACGCTCG	]baabbbbbaaa]bb`Wab[``a]X]b_bb\VZZ]baY`^`aa`aa`aX`aaR``	1
IL10	2732	4	25	1782	162	0	1	ATGGAACAACTCACTAAAAACCAAGCTGTCGCTACTTCCCAAGAAGCTGTTCAG	aabb_bbaaa`aZabbabbba]_aa`]`Z]`__``]\\``X`_[]YX\VJXVSQ	1
IL10	2732	4	25	1782	854	0	1	CGCCAATATGAGAAGAGCCATACCGCTGATTCTGCGTTTGCTGATGAACTAAGT	[bab[bbabY^`^QQWaab[aaa`aa_`^aaaa[a^___aa_`_a_R__aX[BB	1
IL10	2732	4	25	1782	1814	0	1	TGAACAATCCGTACGTTTCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGC	`bYZ[[^[a^XP\a`V^a``L`a[_XXPUTDMV\]]ZQ_ZKU__]]]]YBBBB	1
IL10	2732	4	25	1782	1816	0	1	TAGGGTCAACGCTACCTGTAGGAAGTGTCCGCATAAAGTGCACCGCATGGAAAT	aabab^a`aaaa\a`a_aUa`aaV]Z_U_`[__^aY_`Ua_^`_``___^_BB	1
IL10	2732	4	25	1782	1951	0	1	TGACGGGATGAACATAATAAGCAATGACGGCAGCAATAAACTCAACAGGAGCAG	ab__ba_]`a^_aaaa^aYS[]`[^`_^`]Y\``^XU^W^\\Y_[^^ZVVNYWS[	1
IL10	2732	4	25	1782	631	0	1	TCAGAAAATCGAAATCATCTTCGGTTAAATCCAAAACGGCAGAAGCCTGAATGA	aaab^[`_aabV]_`a`aa`a`aa]]LMSZT\UY]BBBBBBBBBBBBBBBBBBB	1
IL10	2732	4	25	1782	1982	0	1	TCTTTAGCTCCTAGACCTTTAGCAGCAAGGTCCATATCTGACTTTTTGTTAACG	a`aa`^``^aa_a_^]]]^__[^^YUQ^XZWZUSWZX^VLVXWYZXXXVPVRQ	1
IL10	2732	4	25	1782	157	0	1	TGCCGTTTCTGATAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGC	aaaa``baaaa]`]__aa_aaaZ\a_Y]W^[_[Z^___\SWU\ZY^^WV^Z	1
IL10	2732	4	25	1782	456	0	1	ATCAGCTTTACCGTCTTTCCAGAAATTGTTCCAAGTATCGGCAACAGCTTTATC	^aabbaabbaaab_`a`baab]`[_ab]Za`Y^aaXaaZ[WLL_^__\]Ua]_\	1
IL10	2732	4	25	1782	375	0	1	GAAAGAGTAGAAATGCCACAAGCCTCAATAGCAGGTTTAAGAGCCTCGATACGC	^a_]a`aaaa`]abbbaa`a_``aaa`abaa`aaa_aa``a`\]aaRa[`a`\^	1
IL10	2732	4	25	1782	401	0	1	GCAAGCTCTTTAACTGGCGGCGATTGCGTACCCGACGACCAAACTCAGGGACAA	a_`a]aaI`Z`]a\BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB	1
IL10	2732	4	25	1782	915	0	1	GTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAGCCGCTTAAAGCTACCA	aaaa`]a__a__a_``_^`]W`Y[__^_VFVZY\[W\WZ`ZRXLYUUV[BBBB	1
IL10	2732	4	25	1782	979	0	1	TTTATGTTTGGTGCTATTGCTGGCGGAGTTGTTTCTTTTTCTGCTTGTGGGGGC	ab`Z`a^_a`^Va]]P]a_aa`BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB	1
IL10	2732	4	25	1782	774	0	1	GCACGCTCCCAAGCATTAAGCTCAGGAAATGCAGCAGCAAGATAATCACGAGTA	aa^_aba^aa]]aa_`][_``_][Q]XR__YU]_]__]Q`U^`X^UXY]YBBB	1
IL10	2732	4	25	1782	1616	0	1	CTTGGTCAACCCCTCAGCGGCAAAAATTAAAATTTTTACCGCTTCGGCGTTATA	abbabbbZabbbabaW^b_aa\V^Vaa\WUSYa]aaa^Y^U`^_`_^\^WYVOY	1
IL10	2732	4	25	1782	1354	0	1	CTGAATAGCAAAGCCTCTACGCGATTTCATAGTGGAGGCCTCCAGCAATCTTGA	abaaabab_`aaababab`bbbaabab``b_a`aa__`aaa``aaaTL\[aaa`	1

illumina

Base calling quality score

- A quality score is a prediction of the probability of an error in base calling
  - produced by a model that uses quality predictors as inputs and produces Q-values as outputs
- Q=-10log10 (probability that the base is wrong)
  - Q40: 1 error in 10.000 base calls
  - Q30: 1 error in 1.000 base calls
  - Q20: 1 error in 100 base calls
- The Phred score is a method for assigning quality scores to sequencing data, using numerial predictors of base quality
- Q score are represented as ASCII characters
  - from ASCI to phred = ASCII value + 64
- Why not use the capillary sequencing standard Phred algorithm/predictors ?
  - Phred depends crucially on the quality predictors and their statistical distributions
  - good predictors for SBS data are much different than good predictors for capillary sequencing data

Alignment and alignment scoring

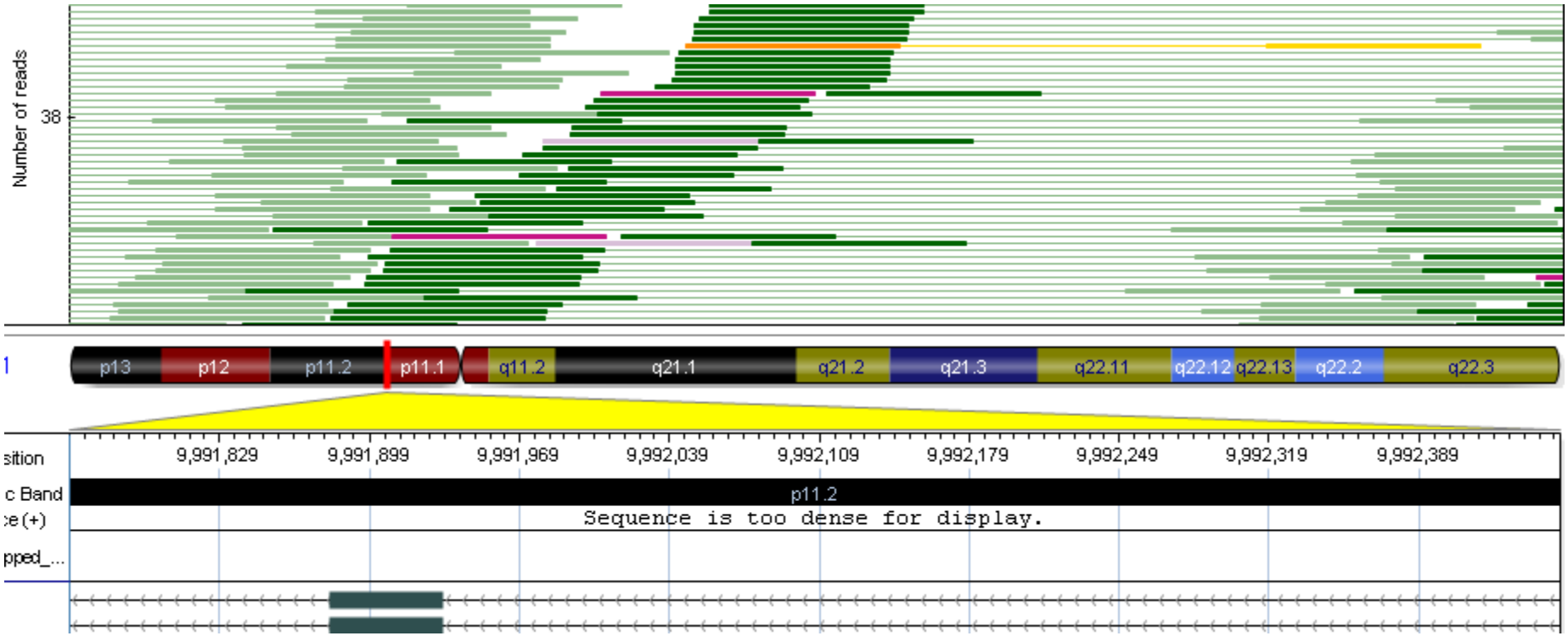
- ELAND v2
- reference genome is squashed
- multiseed, gapped alignment allows for detection of indels (<20 bp)
- each candidate position gets a probability
  - Base quality scores and mismatches are used in this calculation
  - Alignment score is expressed on the Phred scale (log odds ratio)

Data quality is assessed by checking a set of metrics and plots

Lane Info		Tile Mean +/- SD for Lane							
Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters	% Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	384339	374821 +/- 2479	317111 +/- 2469	500 +/- 15	80.94 +/- 1.82	84.60 +/- 0.52	84.61 +/- 0.25	326.08 +/- 8.17	1.24 +/- 0.46
2	359280	335506 +/- 4367	296436 +/- 3392	495 +/- 13	80.28 +/- 1.27	88.36 +/- 0.28	99.10 +/- 0.11	520.47 +/- 3.21	0.68 +/- 0.10
3	363356	351148 +/- 12389	299799 +/- 6395	476 +/- 17	78.23 +/- 1.37	85.42 +/- 1.34	98.47 +/- 0.00	460.72 +/- 11.11	1.28 +/- 0.36
4	384594	374144 +/- 21316	317322 +/- 12469	515 +/- 17	80.03 +/- 2.38	84.90 +/- 1.75	84.47 +/- 0.27	322.65 +/- 4.62	1.20 +/- 0.16
5	382373	377654 +/- 10799	315489 +/- 8581	494 +/- 21	79.81 +/- 1.09	83.54 +/- 0.51	84.51 +/- 0.28	322.09 +/- 4.96	1.22 +/- 0.19
6	358367	357009 +/- 11548	295682 +/- 6826	451 +/- 27	78.09 +/- 1.85	82.85 +/- 1.09	98.46 +/- 0.28	461.38 +/- 11.17	1.27 +/- 0.41
7	380393	375250 +/- 13209	313856 +/- 7465	494 +/- 18	79.75 +/- 1.93	83.67 +/- 1.14	84.69 +/- 0.10	324.06 +/- 5.23	1.19 +/- 0.19
8	374541	372954 +/- 3179	309027 +/- 3476	535 +/- 83	78.60 +/- 2.66	82.86 +/- 0.53	84.65 +/- 0.00	321.03 +/- 8.67	1.24 +/- 0.30

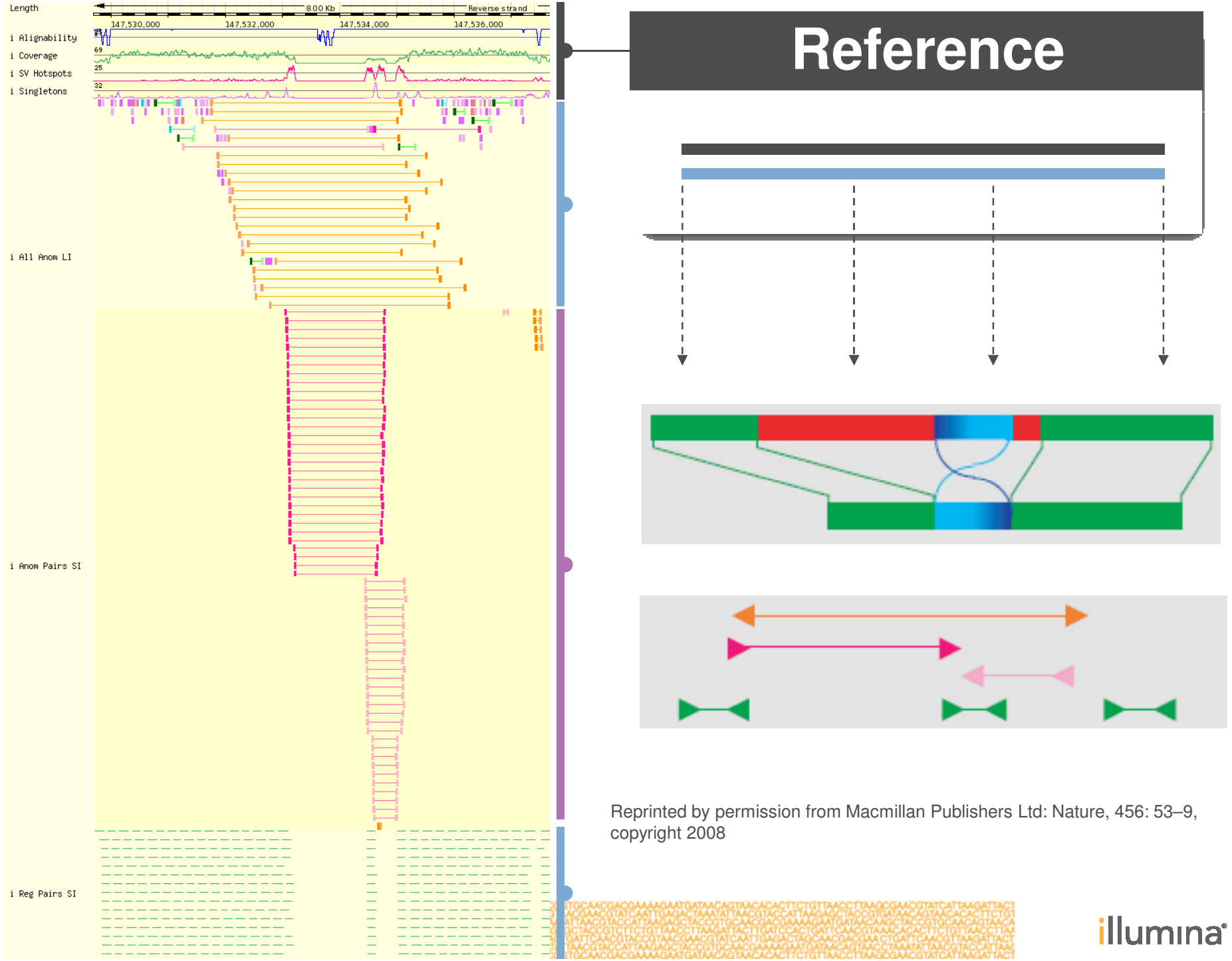
GenomeStudio

visualization of paired-end reads



From the TPTE gene on Chromosome 21

illumina



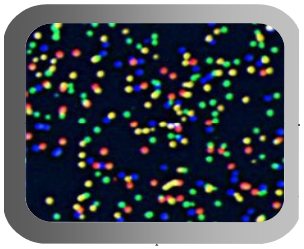
Reprinted by permission from Macmillan Publishers Ltd: Nature, 456: 53–9, copyright 2008

illumina

Applications

illumina®

Broadest range of customer demonstrated applications



Human Genome Resequencing

Bacterial Sequencing | Targeted Resequencing

De novo Sequencing | ChIP-Seq | Small RNA Discovery

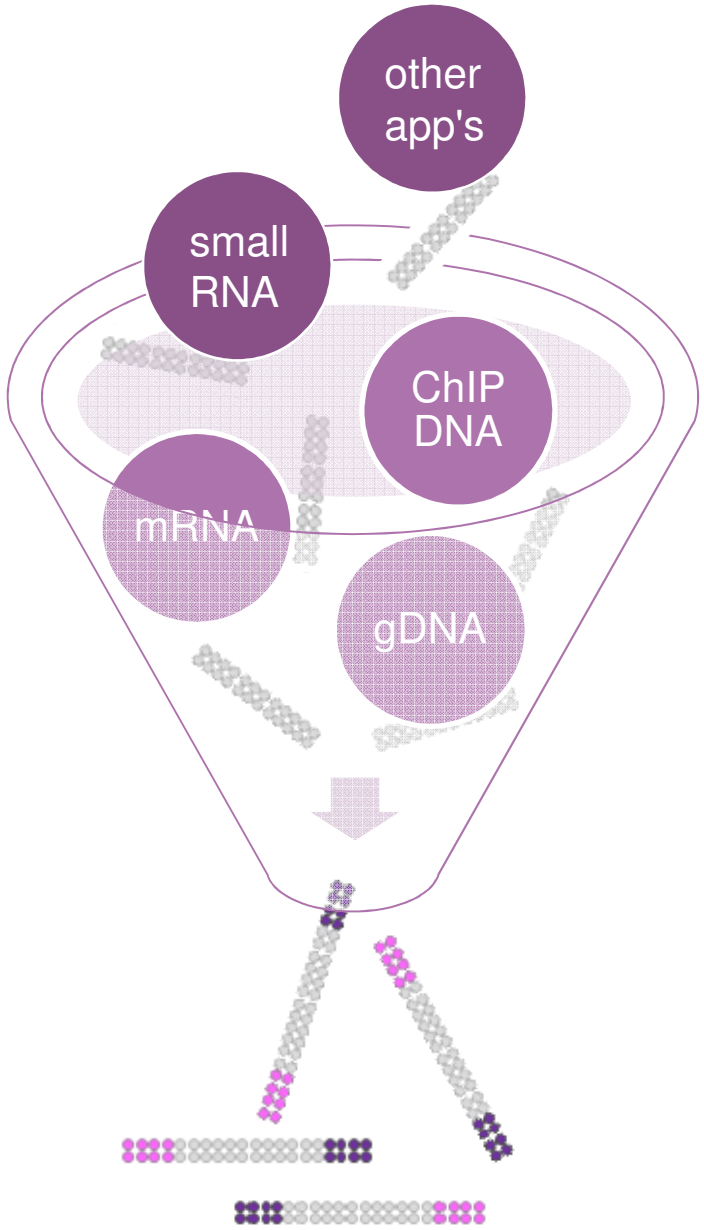
Tag profiling | mRNA-Seq | DNase I Hypersensitivity | Methylation

Molecular Cytogenetics | AFLP | ChIA-PET | CNV | Nucleosome Mapping

"The Genome Analyzer is enabling our clients to do things that used to be impossible, experiments that they only dreamed of doing, but can do now at a reasonable cost. The Genome Analyzer has completely changed our business."

- Laurent Farinelli, Ph.D., Fasteris

illumina



illumina

A comprehensive catalogue of somatic mutations from a human cancer genome

Pleasance et al - Nature 2010

"..provides insights into the forces that shape a cancer genome."

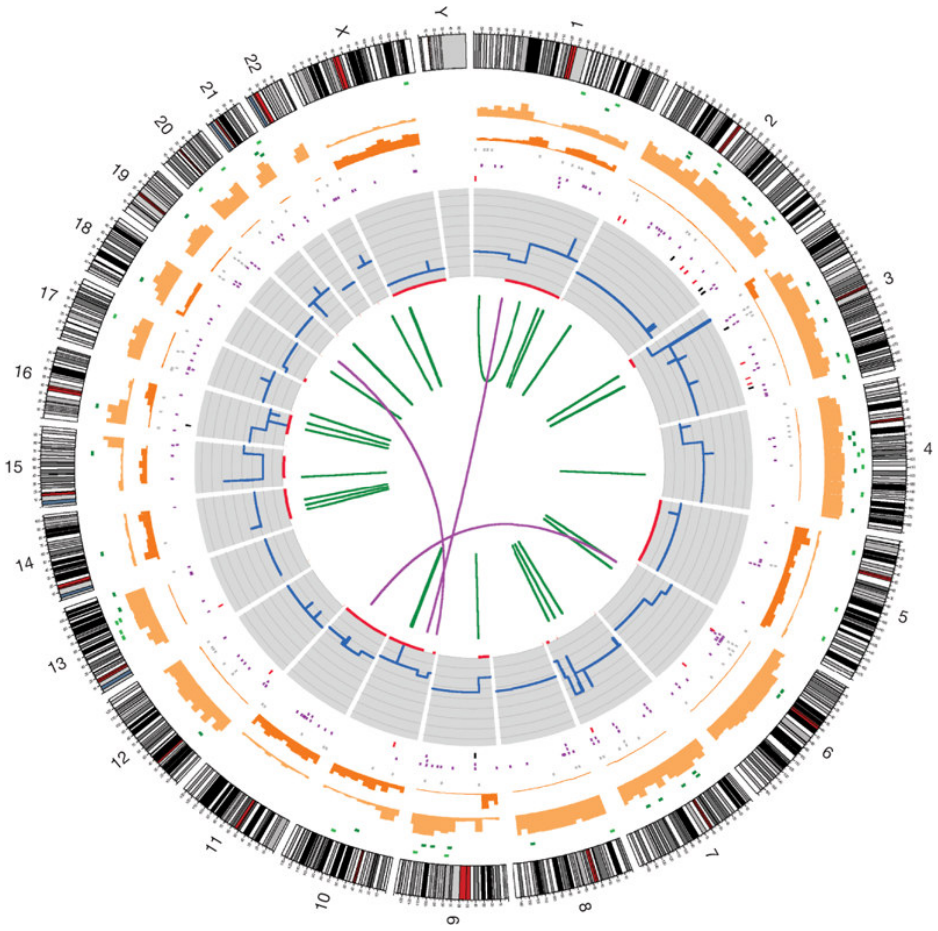
"..reveal traces of the DNA damage, repir, mutation and selection processes that were operative years before the cancer became symptomatic"

- Method
  - combined 2x75bp PE reads and 2x50bp mate pair libraries (2/3/4 kb)
  - COLO-829 cancer cell line from a metastasis of a malignant melanoma and COLO-829BL lymphoblastoid line from same patient
  - obtained > 40x average haploid genome coverage from COLO-829 and 32-fold from COLO-829BL

nature

illumina

The catalogue of somatic mutations in COLO 829



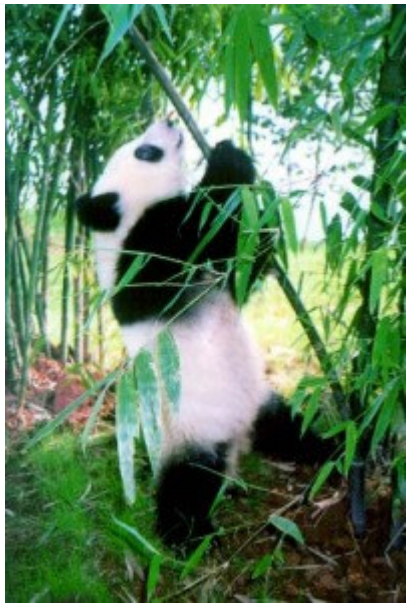
Results

- 33,345 single base substitutions
  - 292 coding
- 1018 small indels
  - 14 coding
- 37 structural rearrangements
  - 34 intrachromosomal
  - 3 interchromosomal
  - 19 breakpoints in genes
- 198 changes in copy number

ED Pleasance *et al.* Nature. 2010 Jan 14;463(7278):191-6

nature

illumina



The sequence and *de novo* assembly of the giant panda genome

Ruiqiang Li *et al*, Nature 2010 Jan 21;463(7279):311-7

The Giant Panda lives in bamboo forests high in the mountains of Western China. It eats 12 - 38 kg bamboo per day.

1,600 individuals remained in the wild in 2004.

- Method
  - insert sizes of 150 bp, 500 bp, 2 kb, 5 kb and 10 kb
  - generated 176 gigabases of usable sequence (equal to 73x coverage of the whole genome)
  - average read length of 52 bp
  - assembled short reads using "SOAPdenovo"
- Results
  - genome size 2.40 gigabases
  - dietary preferences seem to be related to gut microbiome; genetically speaking the Panda is carnivorous

illumina

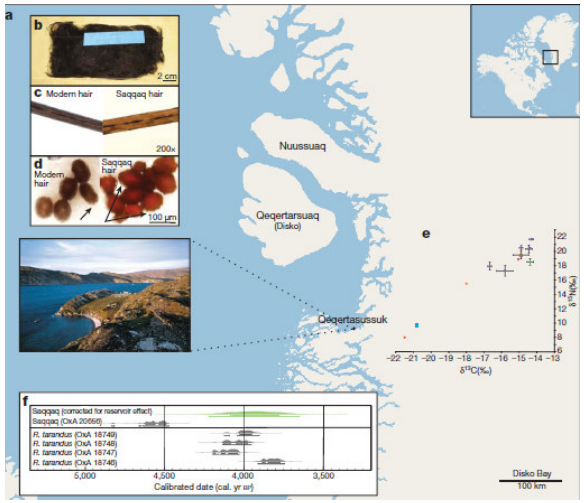
Ancient DNA sequencing

- DNA isolated from 4000 year old permafrost-preserved hair
- 20x coverage
- provides evidence for a migration from Siberia into the New World some 5,500 years ago, independent of that giving rise to the modern Native Americans and Inuit

Ancient human genome sequence of an extinct Palaeo-Eskimo

Morten Rasmussen[1,2]*, Yingrui Li[2,3]*, Stinus Lindgreen[1,4]*, Jakob Skou Pedersen[4], Anders Albrechtsen[4], Ida Moltke[4], Mait Metspalu[5], Ene Metspalu[5], Toomas Kivisild[5,6], Ramneek Gupta[7], Marcelo Bertalan[7], Kasper Nielsen[7], M. Thomas P. Gilbert[1,2], Yong Wang[8], Maanasa Raghavan[1,9], Paula F. Campos[1], Hanne Munkholm Kamp[1,4], Andrew S. Wilson[10], Andrew Gledhill[10], Silvana Tridico[11,12], Michael Bunce[12], Eline D. Lorenzen[1], Jonas Binladen[1], Xiaosen Guo[2,3], Jing Zhao[2,3], Xiuqing Zhang[2,3], Hao Zhang[2,3], Zhuo Li[2,3], Minfeng Chen[2,3], Ludovic Orlando[13], Karsten Kristiansen[2,3,4], Mads Bak[14], Niels Tommerup[14], Christian Bendixen[15], Tracey L. Pierre[16], Bjarne Grønnow[17], Morten Meldgaard[18], Claus Andreasen[19], Sardana A. Fedorova[5,20], Ludmila P. Osipova[21], Thomas F. G. Higham[9], Christopher Bronk Ramsey[10], Thomas v. O. Hansen[22], Finn C. Nielsen[22], Michael H. Crawford[23], Søren Brunak[7,24], Thomas Sicheritz-Pontén[7], Richard Villems[5], Rasmus Nielsen[4,8], Anders Krogh[2,4], Jun Wang[2,3,4] & Eske Willerslev[1,2]

We report here the genome sequence of an ancient human. Obtained from ~4,000-year-old permafrost-preserved hair, the genome represents a male individual from the first known culture to settle in Greenland. Sequenced to an average depth of 20×, we recover 79% of the diploid genome, an amount close to the practical limit of current sequencing technologies. We identify 353,151 high-confidence single-nucleotide polymorphisms (SNPs), of which 6.8% have not been reported



illumina

The impact of scale in sequencing

Human Genome finished here

Gb / run

Year

10^4 scale in throughput; 10^7 scale in parallelisation in 5 years

illumina