An Introduction to Illumina's Next Gen Sequencing

Craig Fishman
Field Application Scientist
cfishman@illumina.com

illumina

Agenda-May 24, 2010

- "Illumina"- the company…
- Where we are today…
- "Solexa Sequencing chemistry"
  - Sample Prep
  - Cluster generation
  - Sequencing By Synthesis (SBS)
- Data Analysis
- Q & A…

illumina

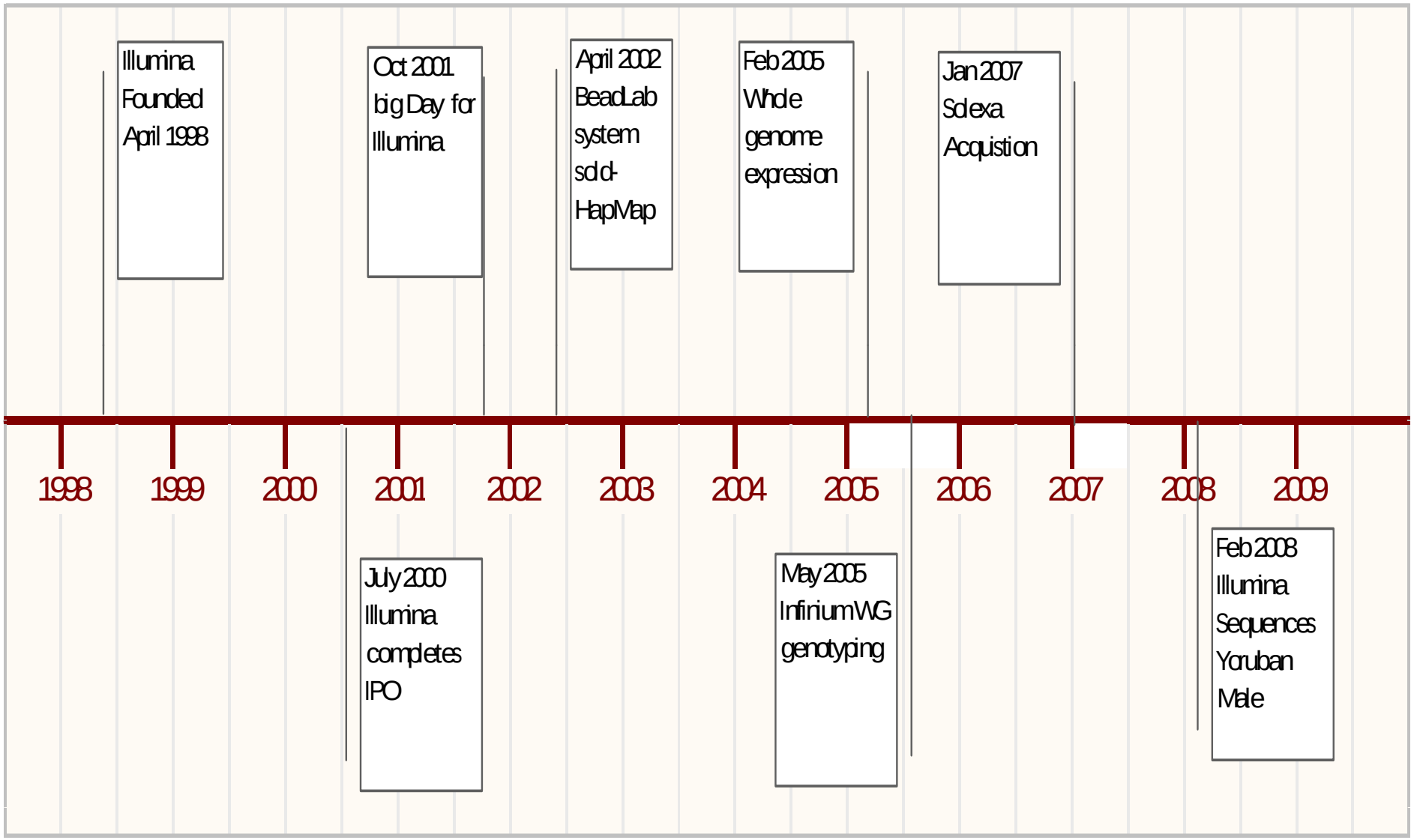
Corporate Summary

Founded: May 1998	IPO in 2000
Headquarters:	San Diego, Hayward, Cambridge
Solexa Acquisition:	Jan 2007
Employees:	>2000 worldwide
Ticker Symbol:	ILMN

Illumina Headquarters, San Diego, CA

illumina

Historical Timeline (in brief)



Typical Genome Center in the Year 2006

- Large Sequencing Factories
- 100 to 150 Capillary Sequencers
- 5 to 10 Colony Picking Robots
- Dozens of PCR Machines
- Several Liquid Handling Robots
- Thousands of 384 well plates
- Dozens of lab personnel
- Multi-million dollar budgets

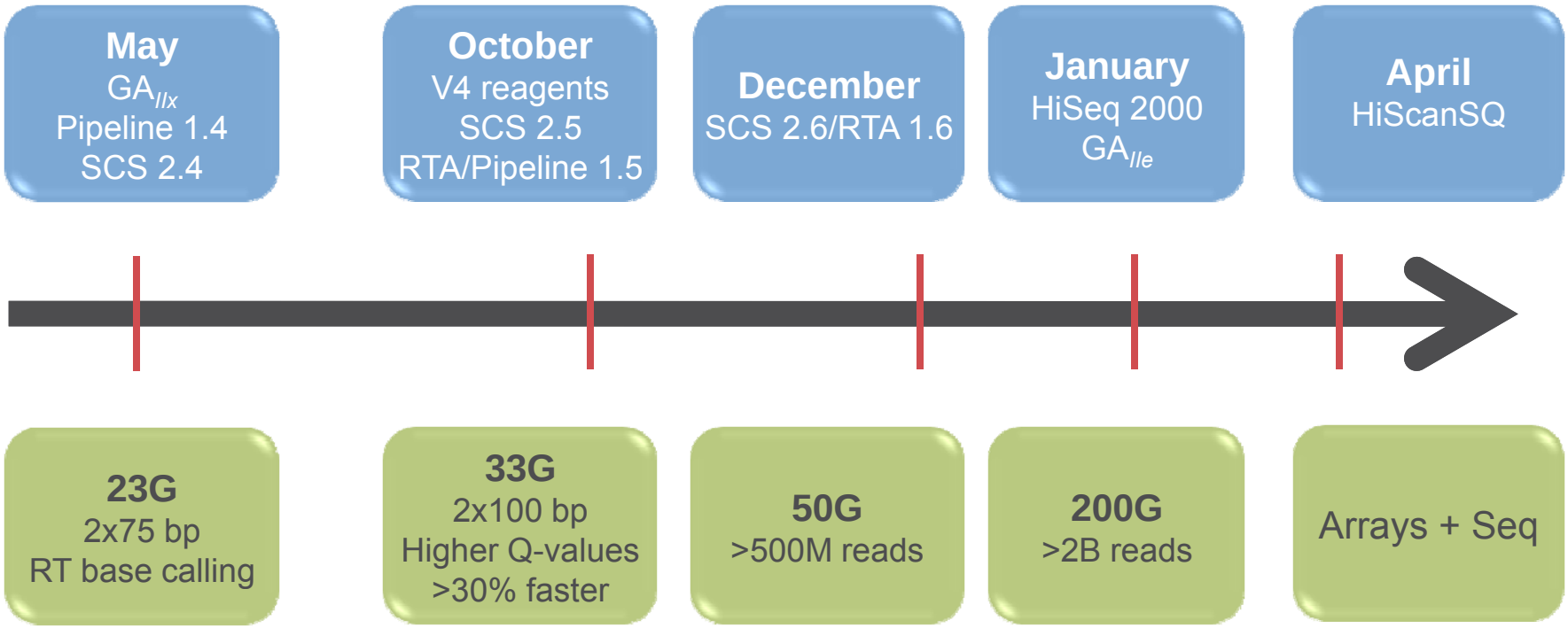
Broad, WashU, Sanger,
Baylor, BGI, Venter, JGI, etc.

→ Output of
3 to 5 Gb per month

Key Developments Over the Last 12 Months

Substantial enhancements on the Genome Analyzer

- >4x increase in output
- 2x increase in read length
- >2x the tags per run
- Further streamlined workflow



illumina

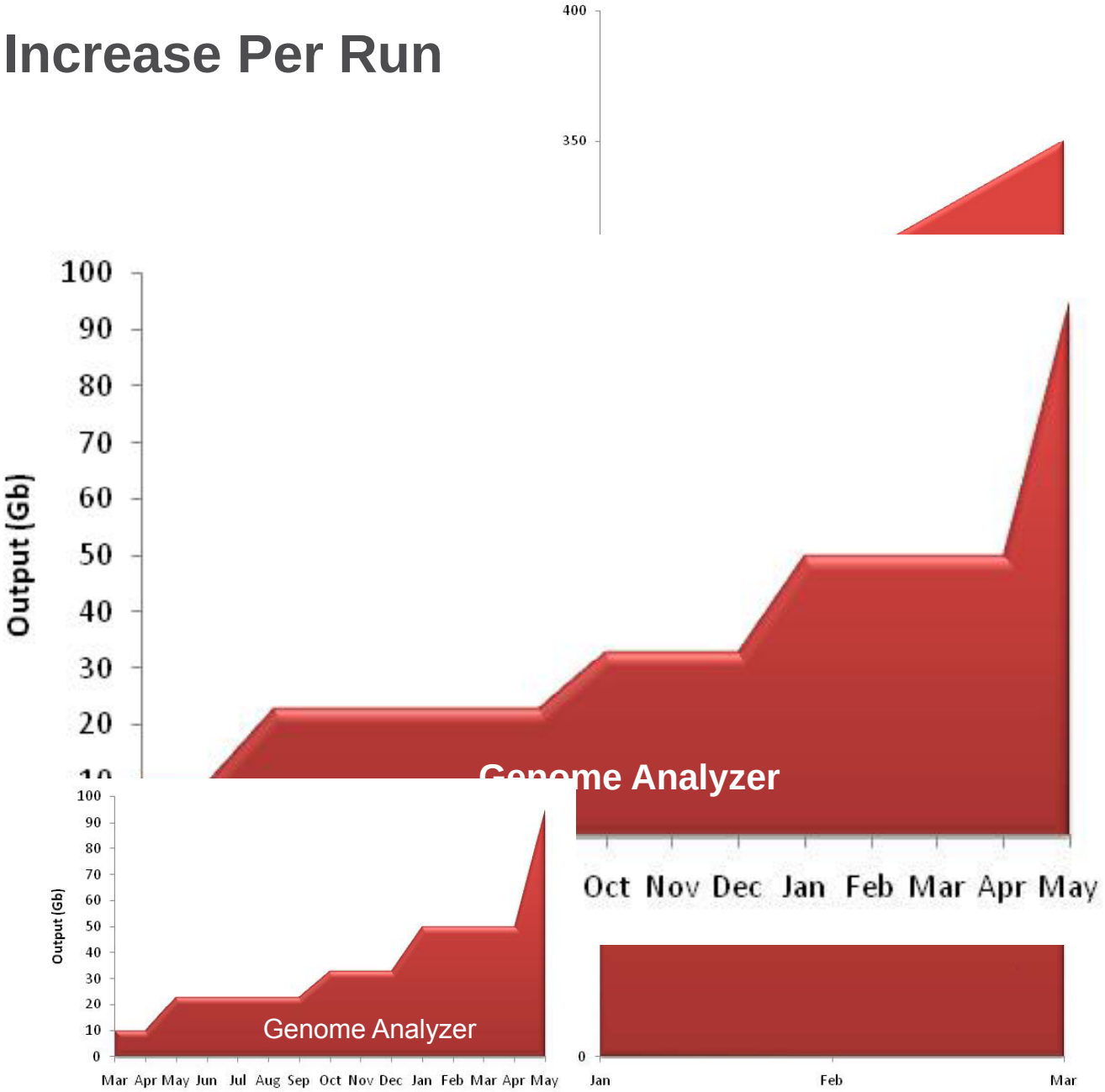
An Illumina Sequencer for Everyone

Next Generation Sequencing made accessible. 40GB/run	Two proven technologies. One powerful platform. 50GB/run	Most widely adopted NGS platform. 95GB/run	Redefining the trajectory of sequencing. 200GB/run

GA_{IIe}	HiScanSQ	GA_{IIx}	HiSeq2000

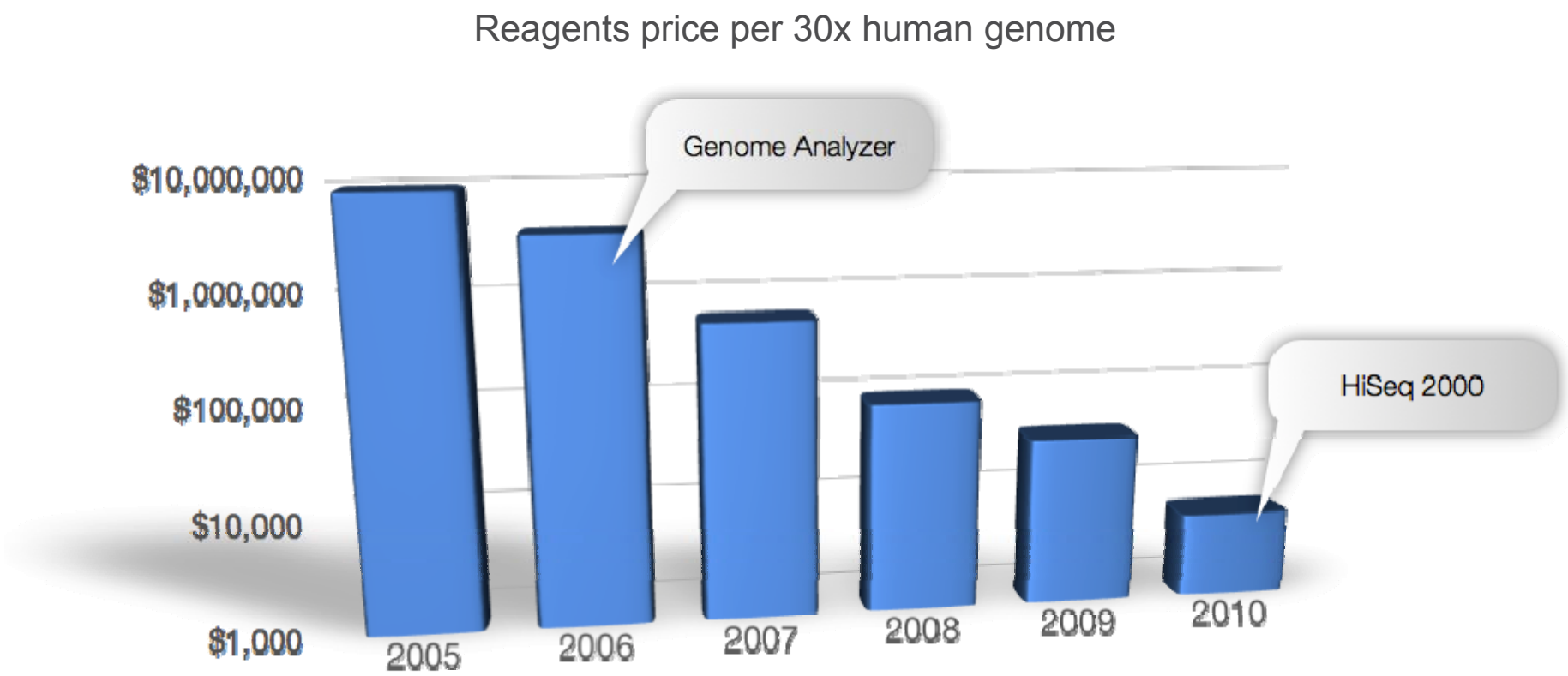
illumina®

Output Increase Per Run



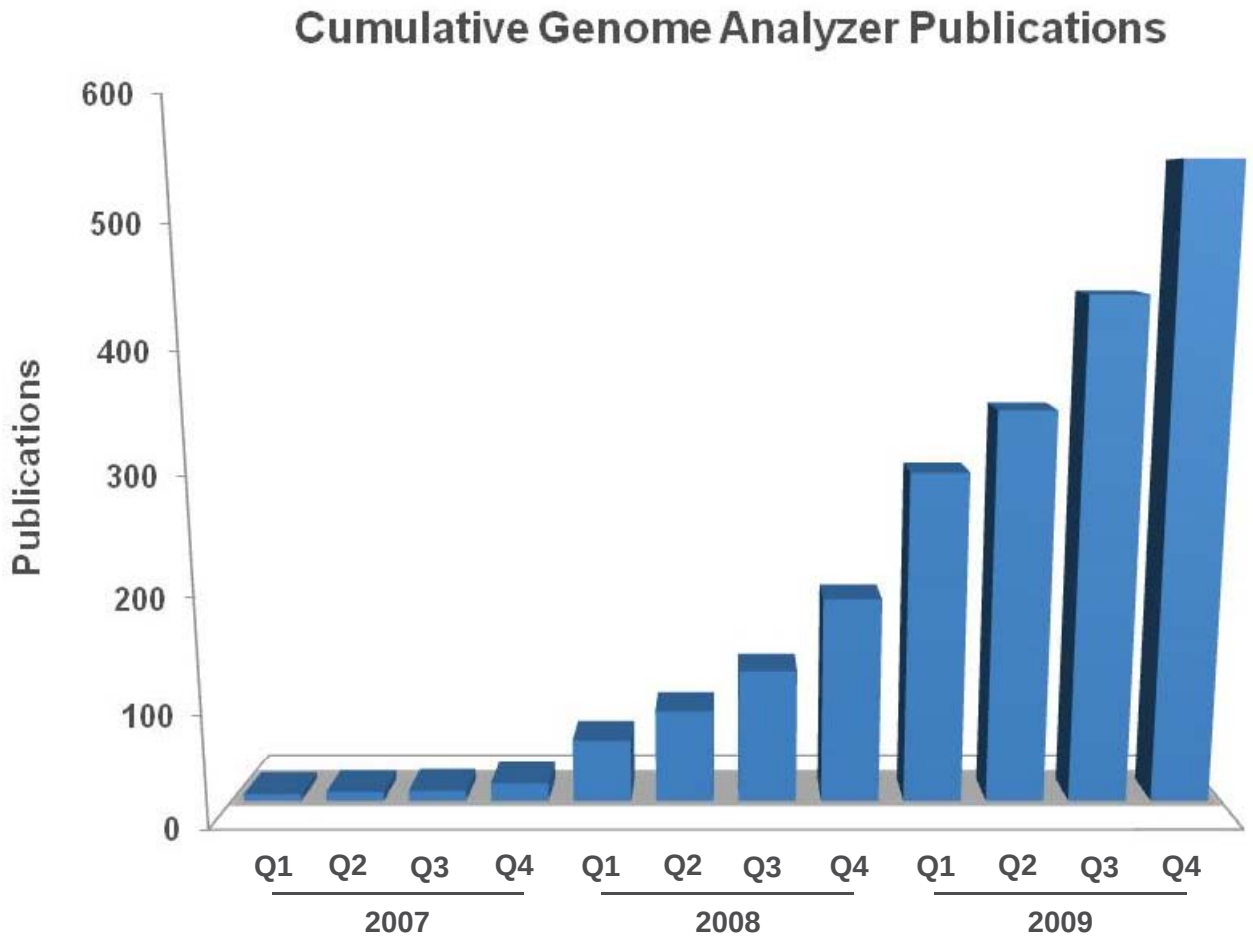
illumina

Driving down the cost of human genome sequencing



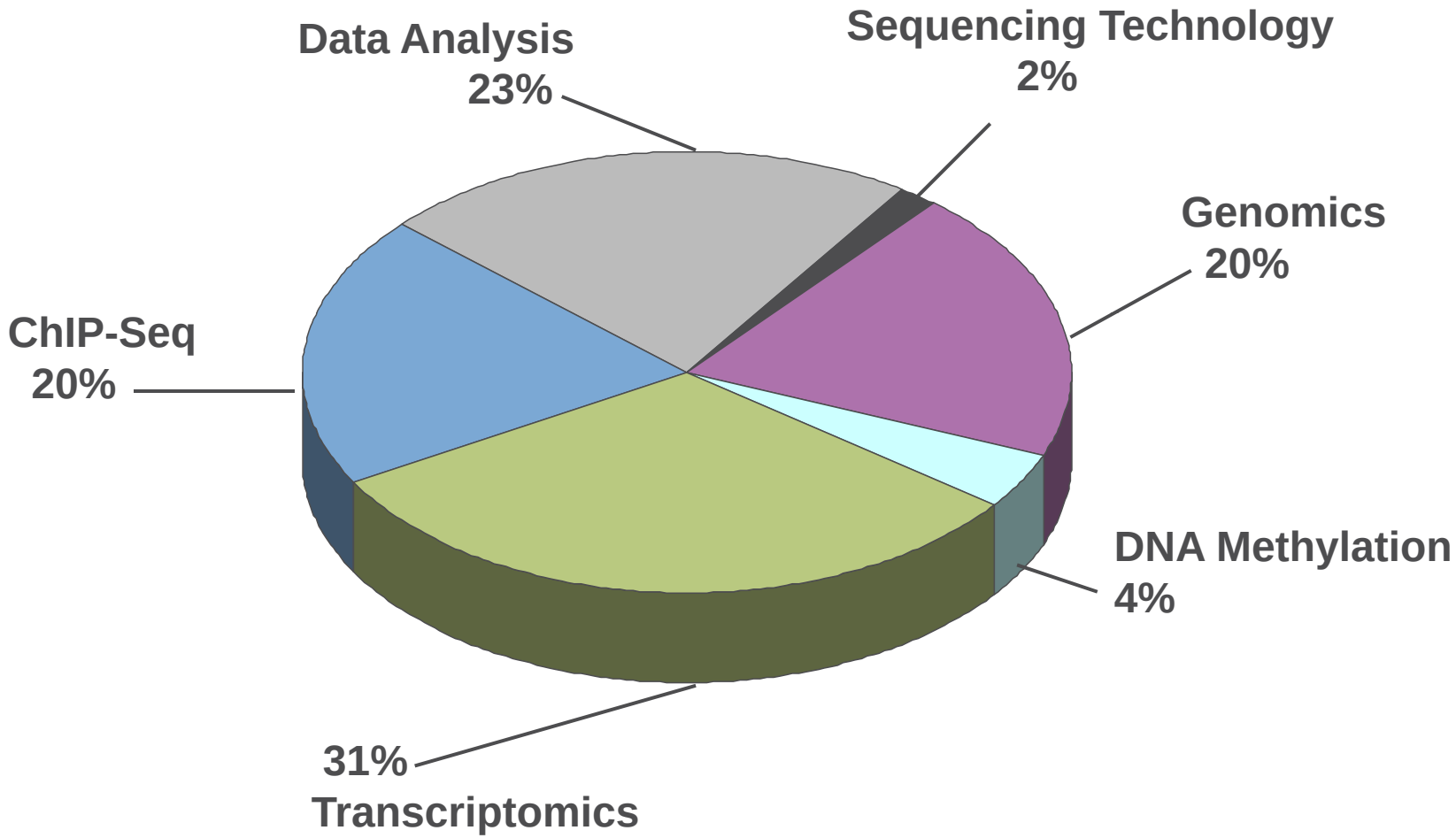
illumina

Publications



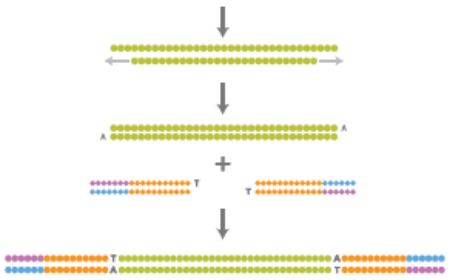
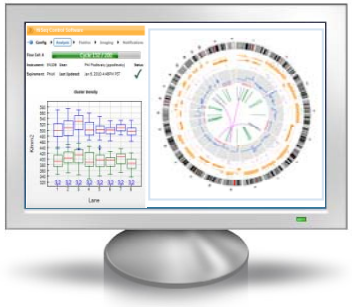
illumina

Publications on Genome Analyzer



illumina

Simplest Sequencing Workflow

SIMPLIFIED SAMPLE PREP	cBot CLUSTER GENERATION	HiSeq 2000 SEQUENCING	DATA PROCESSING & ANALYSIS
			
Parallel sample processing	Automated cluster generation	Automated sequencing	Simple, efficient data analysis

HiSeq 2000

Redefining the trajectory of sequencing

HIGHEST OUTPUT

Initially capable of up to 200 Gb per run

FASTEST DATA RATE

~25 Gb/day

7-8 days for 2 x 100 bp

HIGHEST NUMBER OF READS

One billion single-end reads*

Two billion paired-end reads*

*Based on one billion clusters passing filter

illumina

Dual Surface Imaging

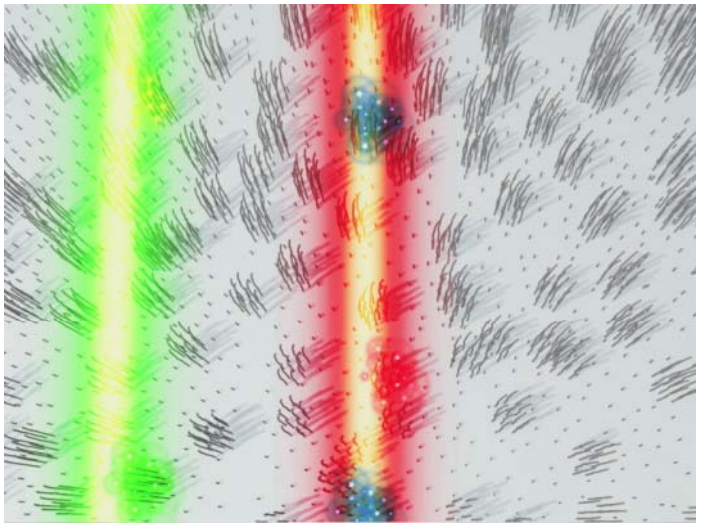
Cutting-edge imaging technology

TDI line-scanning technology, four CCDs

Fastest scanning and imaging method

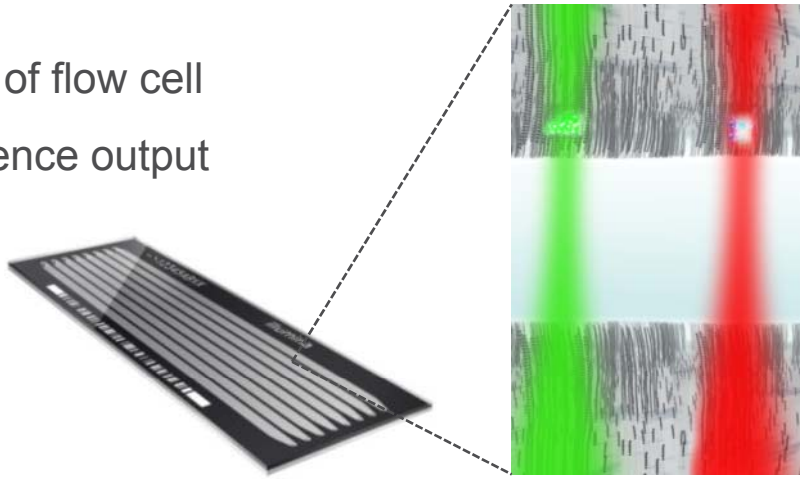
Eliminates Mode Scrambler

Automated real-time auto-focus



Images clusters grown on both surfaces of flow cell

Huge gain in number of reads and sequence output

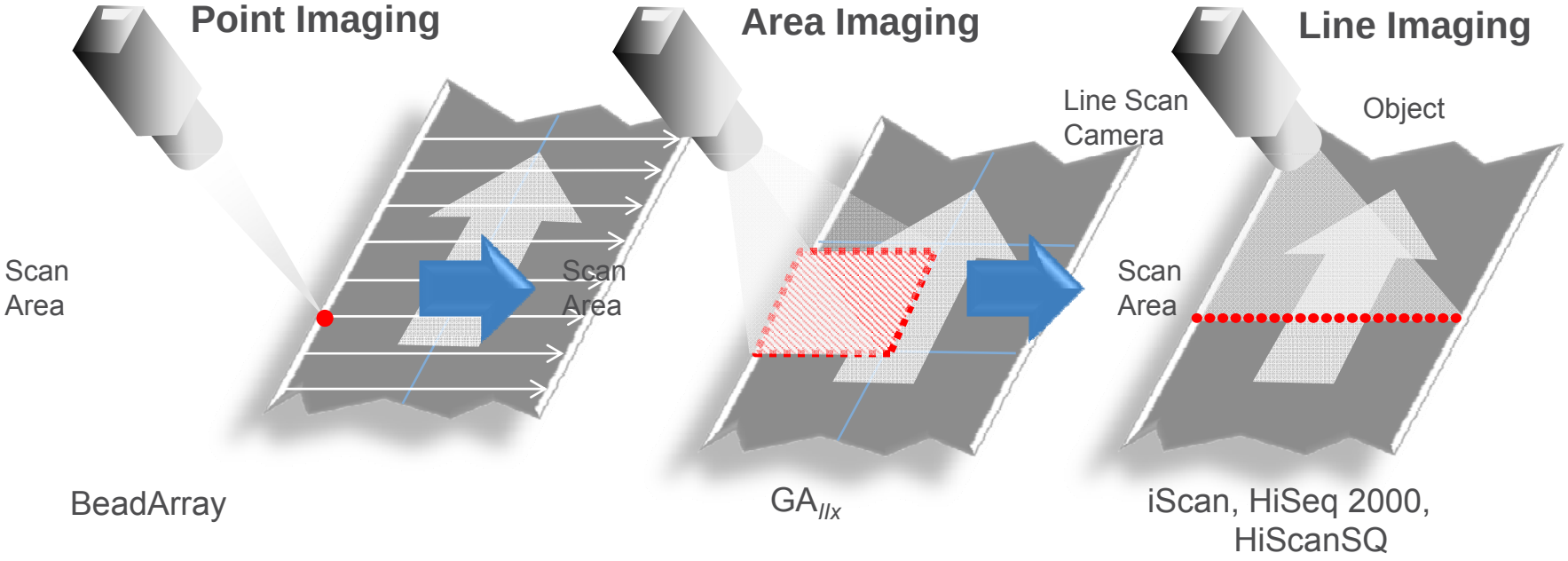


illumina

The Power of Line Scanning

Maximizing data rate

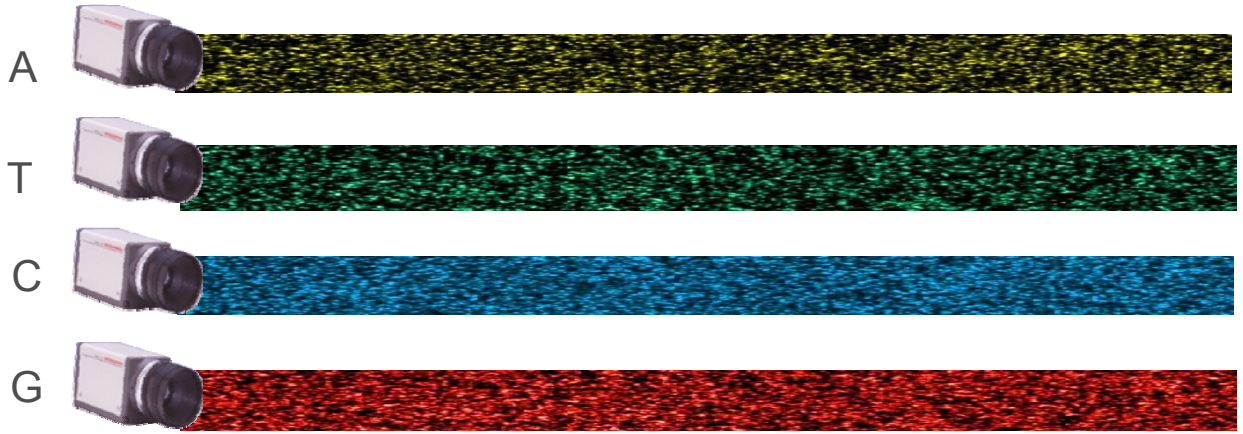
	Point Imaging	Area Imaging	Line Imaging
Stage & filter movement delays	+	–	+
Data transfer delays	+	–	+
Practical data acquisition limit	–	+	+
Data quality/background rejection	+	–	+



illumina

Four CCD Image Capture
Maximizes Throughput

- Cameras capture data almost 100% of time
- Dual flow cells - alternation of imaging and chemistry
- Simultaneous vs. sequential imaging



illumina

HiSeq 2000 Dual Flow Cell Design

Instrument scalability and experimental flexibility

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

HiSeq 2000

Plug-and-play reagents

PRE-CONFIGURED SEQUENCING REAGENTS

Only two minutes hands-on time

Up to 200 cycles per flow cell

Color & bar-coded for tracking

Temperature-controlled compartment

Integrated paired-end fluidics

illumina®

HiSeq Control Software

Touch screen user interface

SIMPLE OPERATION

Step-by-step run set up

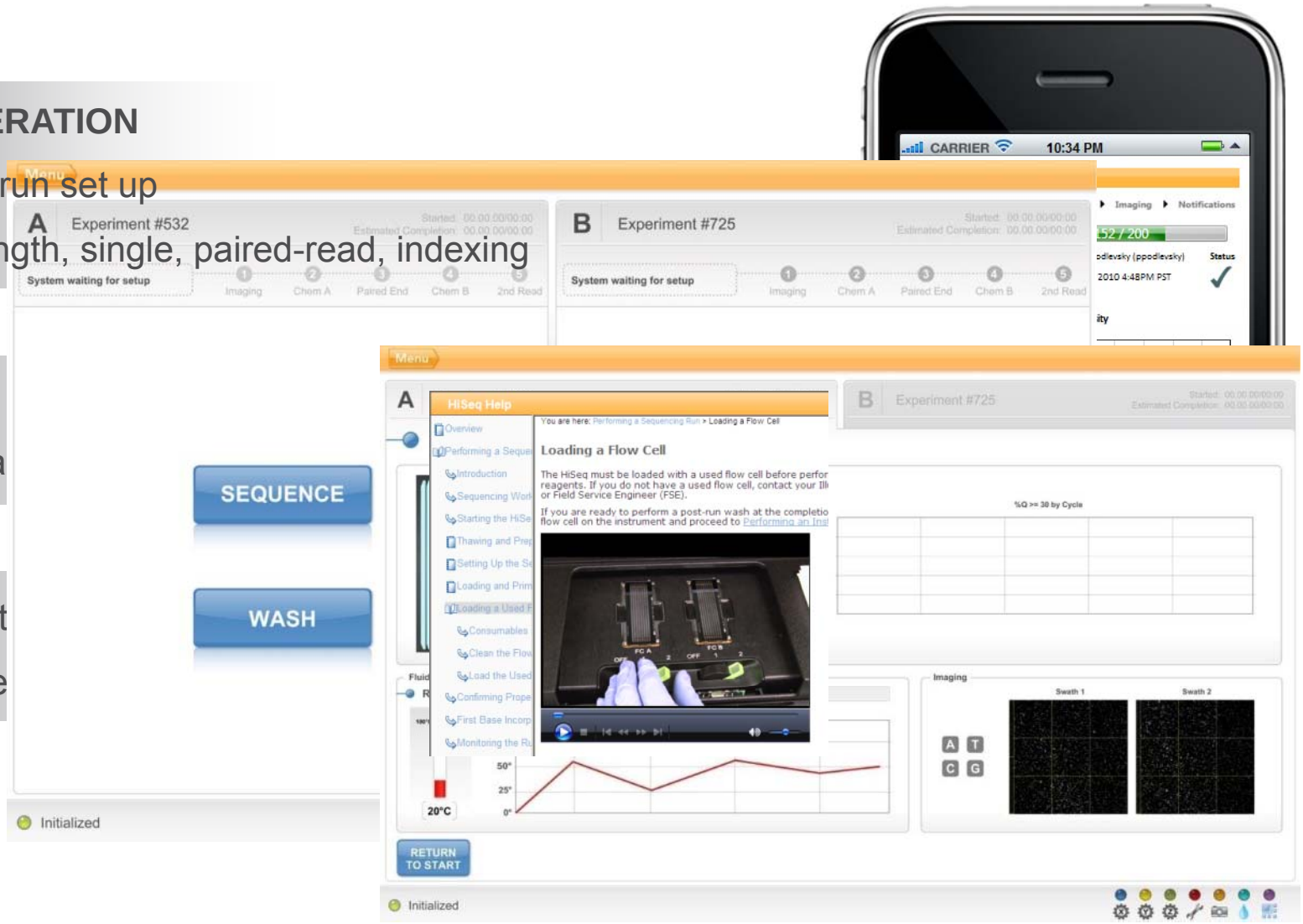
Enter read length, single, paired-read, indexing

Real time run

Reagents usa

Remote monit

Multimedia he



illumina

Genome Analyzer$_{IIx}$ Roadmap

World's most adopted and proven next-gen sequencer

COMMITTED ON PATH TO 95GB

Internal runs have generated >95Gb

RTA 1.8 - detects more clusters

Version 5 reagents - 2 x 150 bp reads

illumina

95G Data Examples

Internal runs

Run	Ave error rate	Yield
A	2.02%	109Gb
B	1.74%	93.7Gb
C	1.21%	94.5Gb

Customer data

Run	Ave error rate	Yield
D*	1.63%	94.8Gb
E**	1.4%	97.8 Gb

**Part of early Access programme; extrapolated from 4 lanes*
***Broad Institute, courtesy of Sheila Fisher*

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

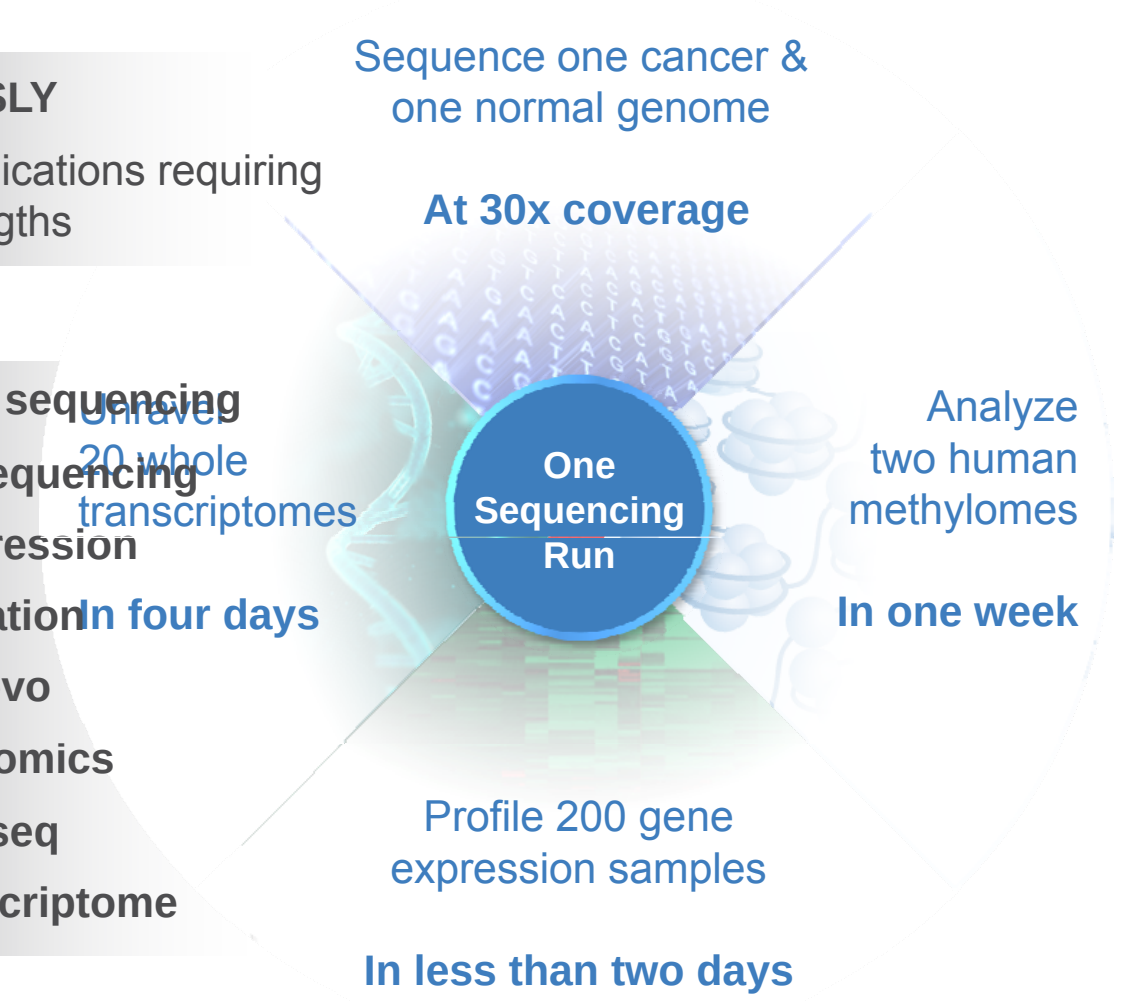
illumina

In one sequencing run you can…

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

At 30x coverage

Unravel 20 whole transcriptomes

In four days

Analyze two human methylomes

In one week

Profile 200 gene expression samples

In less than two days

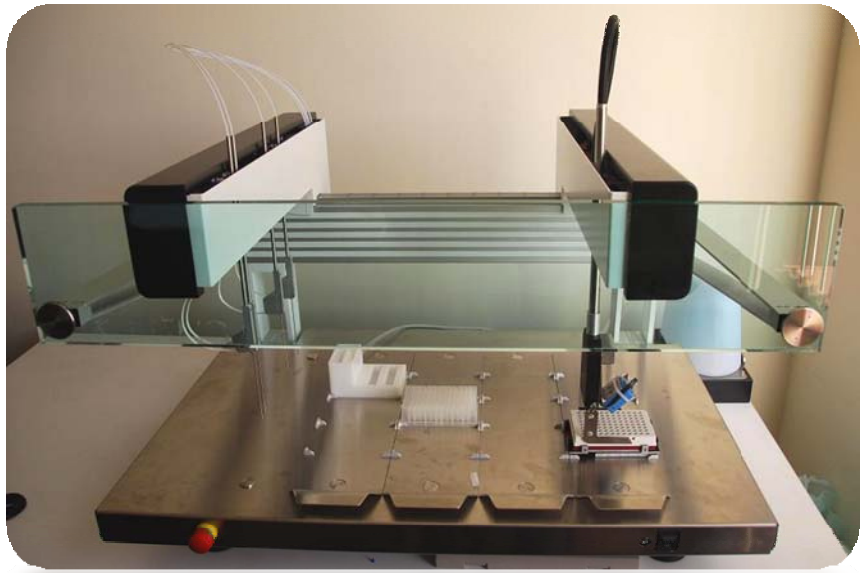
illumina

New Sequencing Sample Prep Automation*

FOR GENOMIC DNA & mRNA-SEQ

Increases scalability

Reduces pipetting errors and variability



PRE-CONFIGURED PROTOCOLS

Accelerates adoption and transition times

No extensive development and verification needed

Serviced and supported by Illumina

*Expected availability End 2010

illumina®

cBot

New automated cluster generation system

EASE OF USE

Pre-configured, plug-and-play reagents

Simple touch screen operation

Barcode scanner for reagent and sample tracking

Real-time run monitoring and self-diagnostics

AUTOMATION

Intervention-free clonal amplification in 4 hours

QUALITY

10-20% increase in cluster density

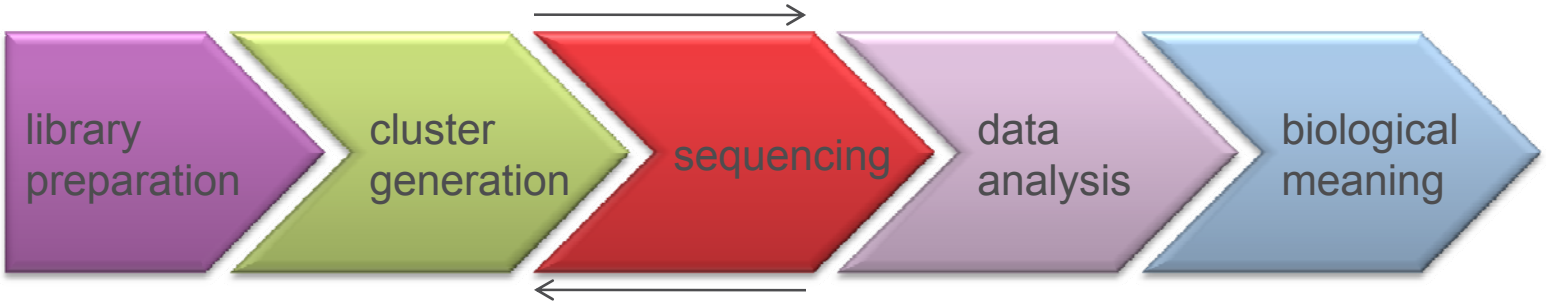
Less variability vs. cluster station

illumina

Sequencing By Synthesis (SBS)

illumina

Sample preparation



illumina

1. Fragment template

	Nebulization	Covaris	Hydro Shear
Technology	Mechanical shearing	Adaptive Focused Acoustics	Hydrodynamic Shearing
Advantages	Low Cost Ease of Use	Low Sample Loss Tight Size Distrib. @ >1Kb	Enables Mate Pair analysis
Limitations	Large sample Loss Broad Size Distribution	Expensive, $40,000 (S2) Broad Size Distrib. @ <1Kb	PreClearing of samples required. Assembly calibration required; assemblies can clog Low Throughput
Minimum Length of Starting Material	2 Kb	600bp	3 Kb
Maximum Product Fragment Length	5 Kb	4 Kb	20 Kb
Automatable	NO	YES (E210)	NO
Applications	ReSeq, PE	ReSeq, PE, Targeted ReSeq	Mate Pair

2. Repair Ends

- Converts overhangs into blunt ends and phosphorylates 5’ end
- Reagents:
  - dNTP, T4 DNA pol, Klenow
  - Kinase/ATP (T4 PNK)
- Simple enzymatic reaction
- 30 minutes incubation

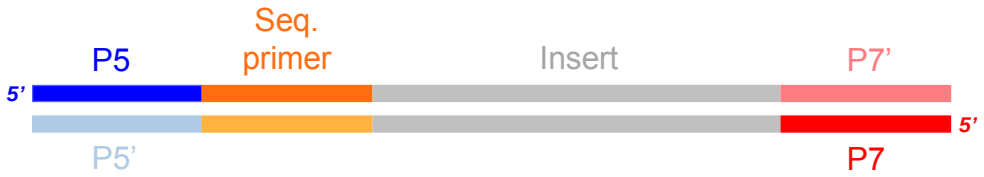
3. Adenylate Ends

- Adds ‘A’ base to the 3' end of the blunt, phosphorylated DNA fragments
- Prevents formation of adapters dimers and concatemers
- Reagents:
  - 1 mM dATP, Klenow exo (3' to 5' exo minus)

4. Link adapters

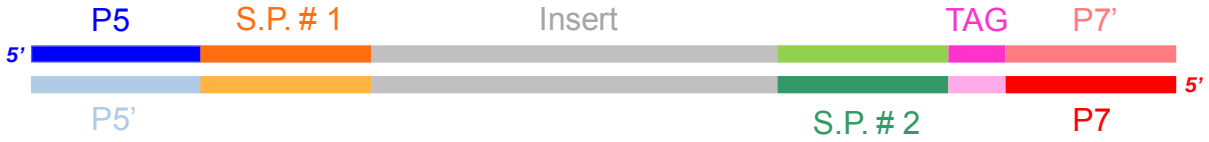
Same general template architecture regardless of application (gDNA, RNA, miRNA, ChIP-seq, exon pull-down, etc.)

Single Read



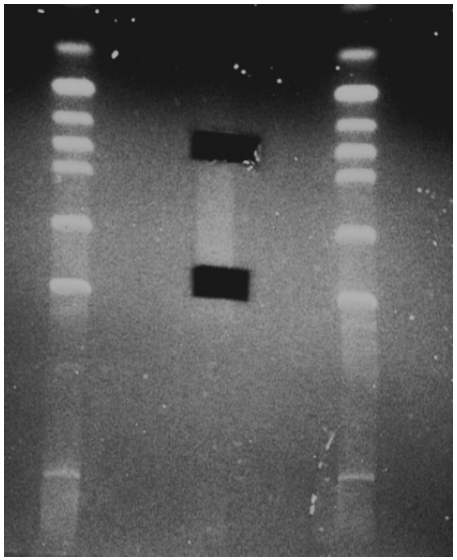
Paired Read

Indexed Paired Read



illumina

5. Gel purify



Size selection

- remove excess adapter and adapter dimers
- tighten the range of fragment sizes
- cut out a ~2 mm wide gel slice

Excision range

- 200 bp insert size target (+/- 10%) for 2×75 bp or shorter
- in practice, this translates to a 2 mm gel slice at ~300 bp to account for the length of the adapter sequences flanking the inserts
- 300 bp or greater insert size target for 2×100 bp or longer (unless you intentionally want to sequence overlapping read pairs)

6. Template enrichment /Final Adapter ligation

- Recommended number of cycles

Starting DNA (µg)	Volume of ligation product to PCR (µl)	Number of Cycle
5	1	10
1	5	10
0.5	10	10/12
0.1	10	12
0.05	all*	12

7. QC library

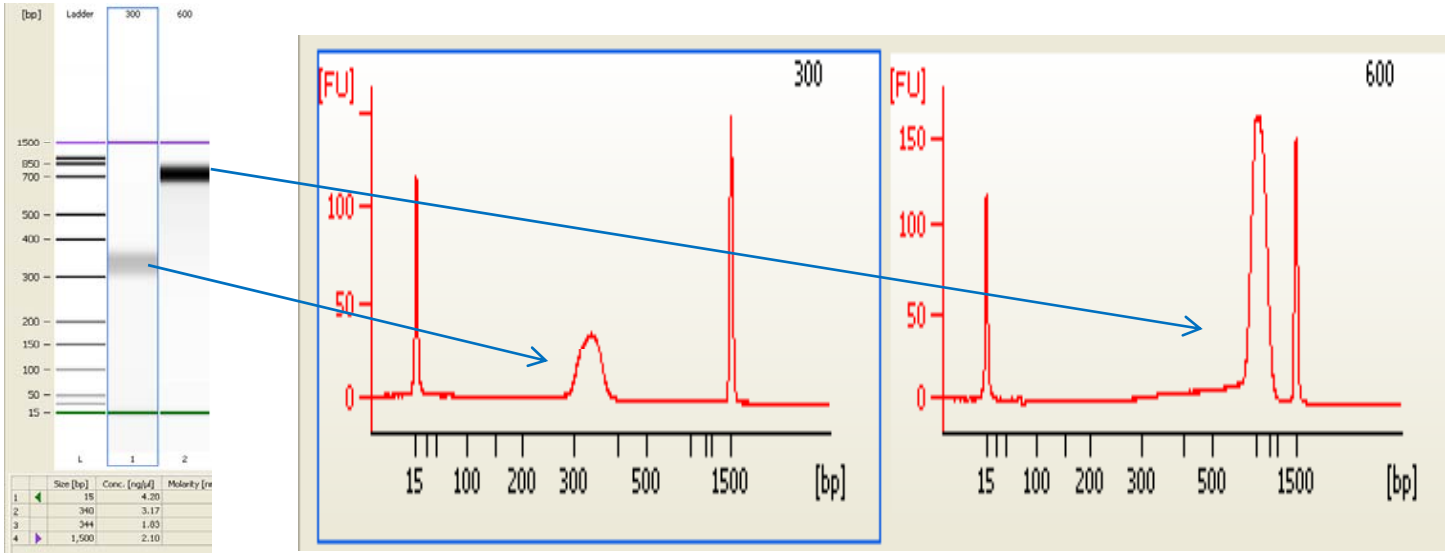
Quantify your completed library using

- NanoDrop (ng/µl)
- Agilent Bioanalyzer (size confirmation)

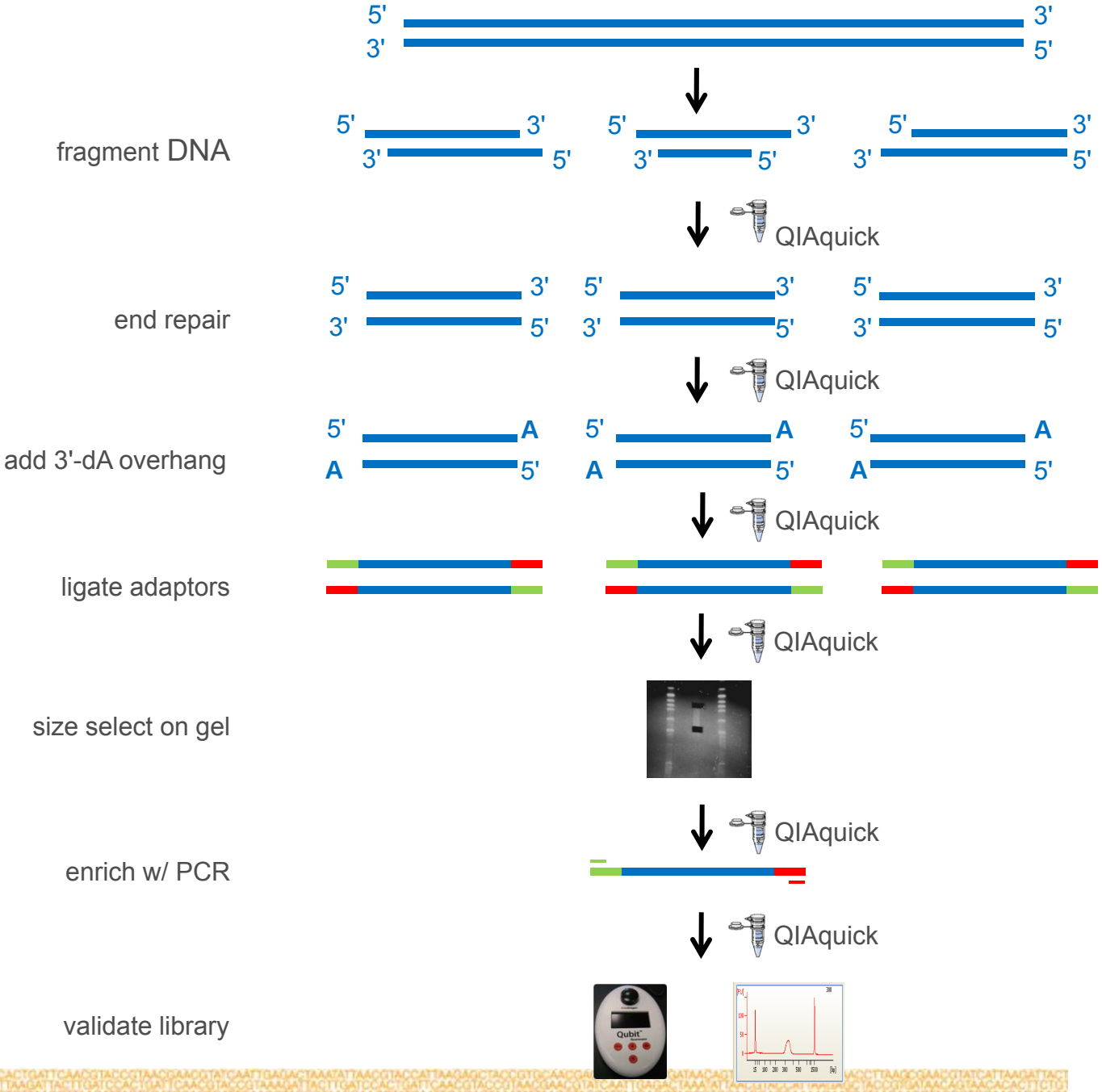
Bioanalyzer traces:

lane 1: 300bp library

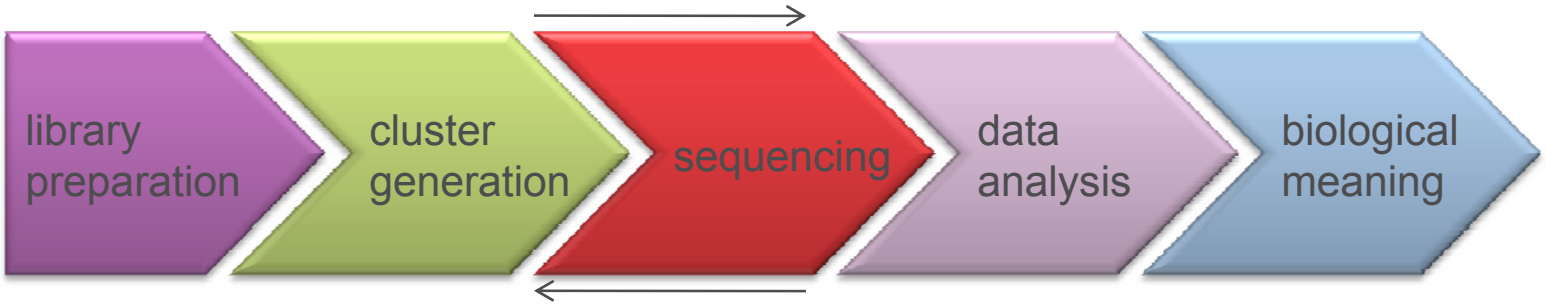
lane 2: 600bp library



illumina


fragment DNA
QIAquick
end repair
QIAquick
add 3'-dA overhang
QIAquick
ligate adaptors
QIAquick
size select on gel
QIAquick
enrich w/ PCR
QIAquick
validate library
concentration
size

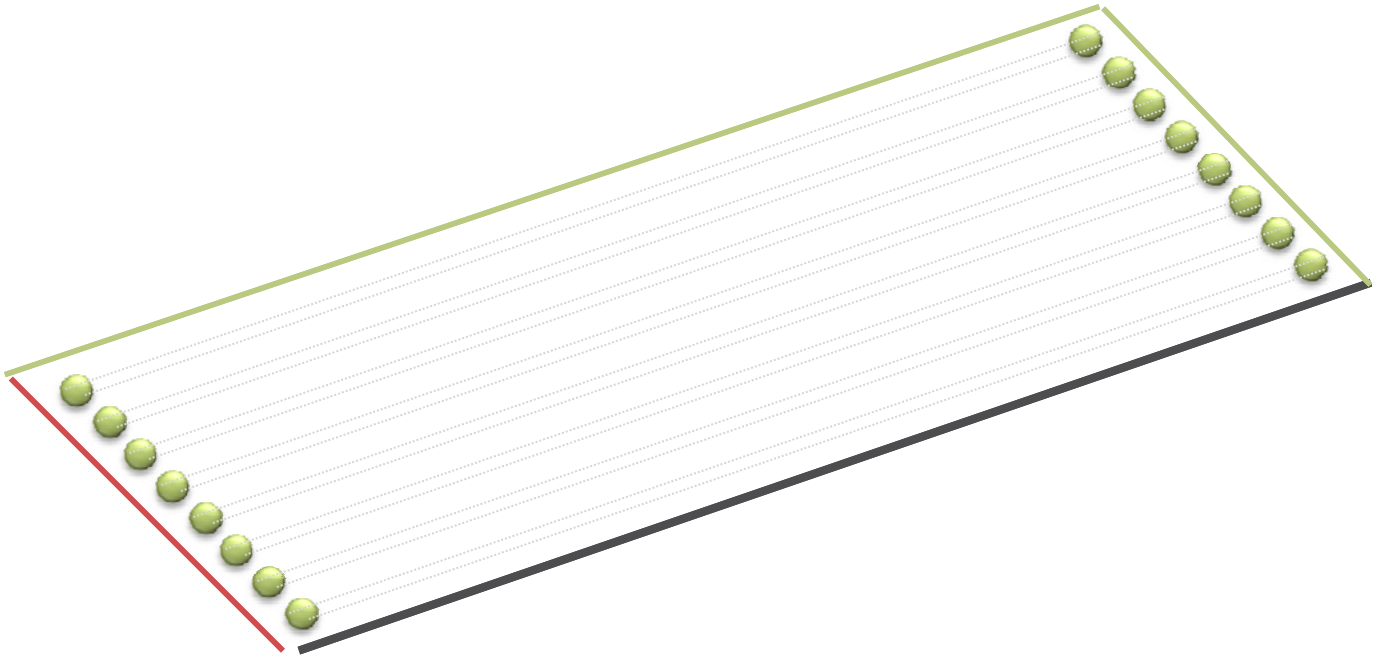
Cluster generation



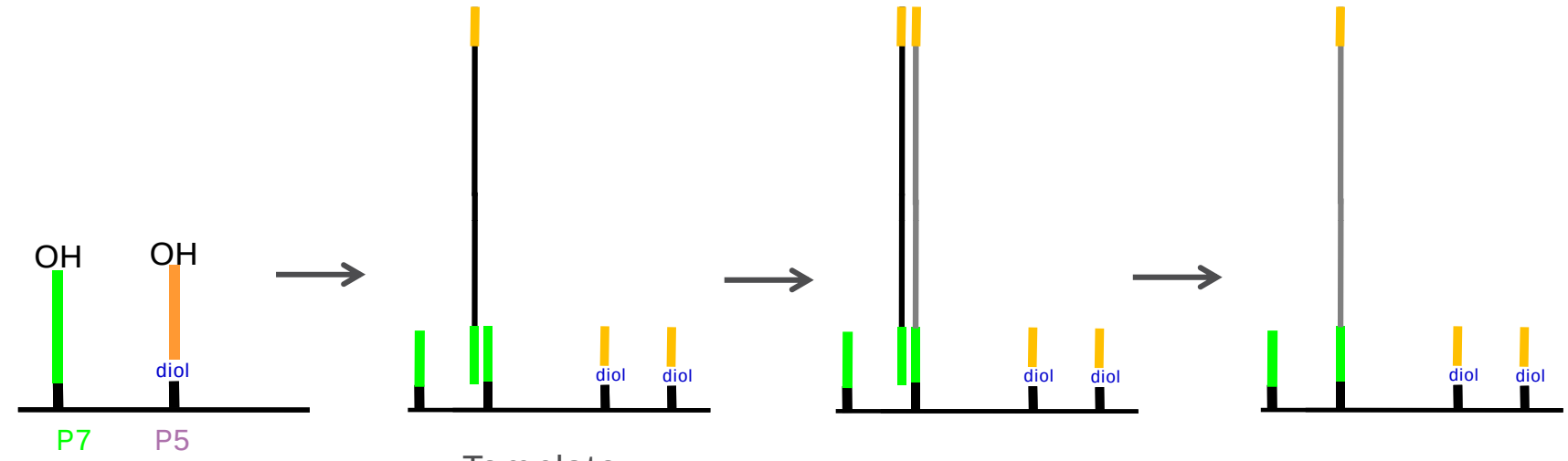
illumina

The flow cell

- Everything except sample preparation is completed on the flow cell-Self Contained
- Cluster Generation
- Sequencing by Synthesis Reaction
- Up to 96 samples
- Closed system

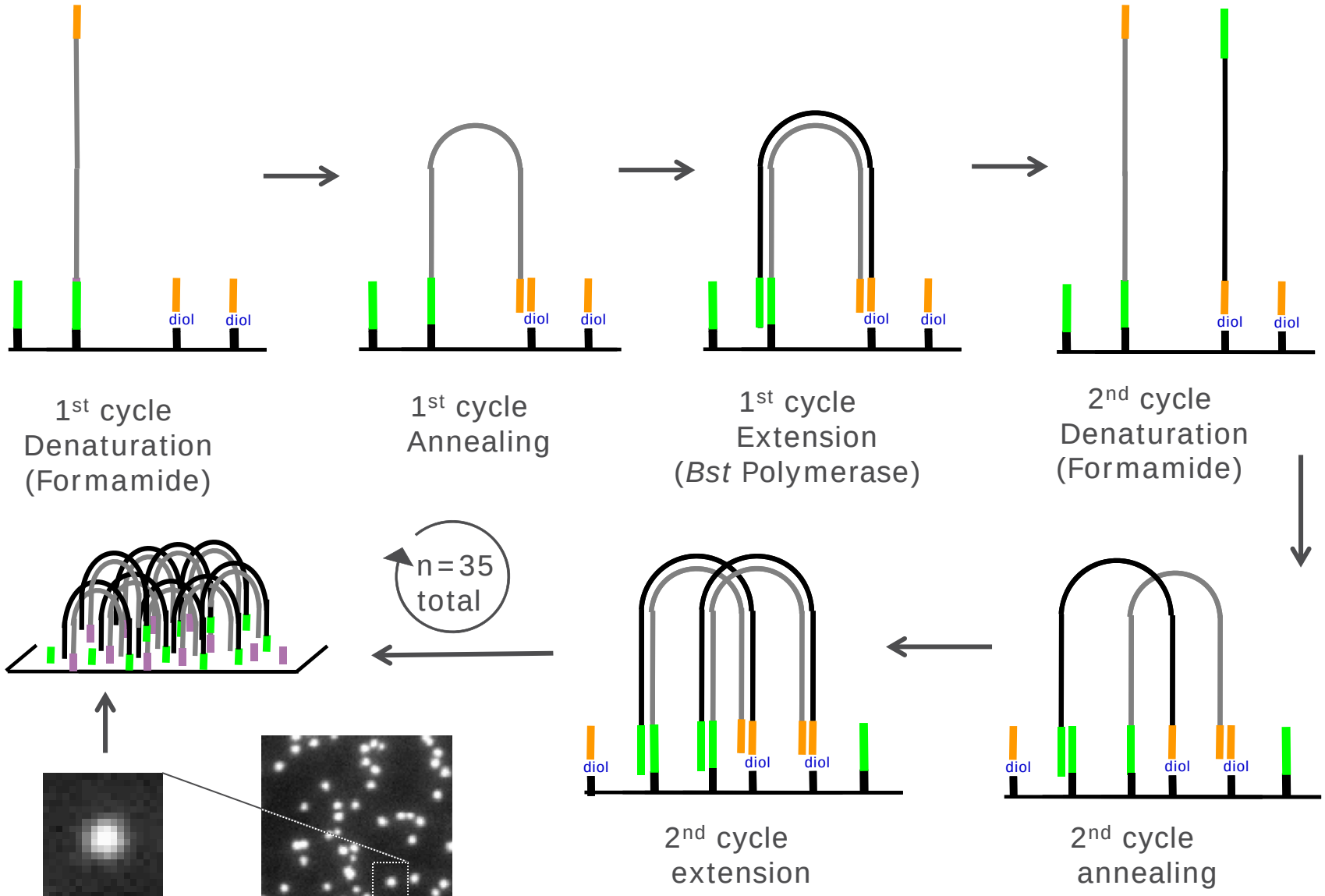


Cluster generation – hybridization and amplification



illumina

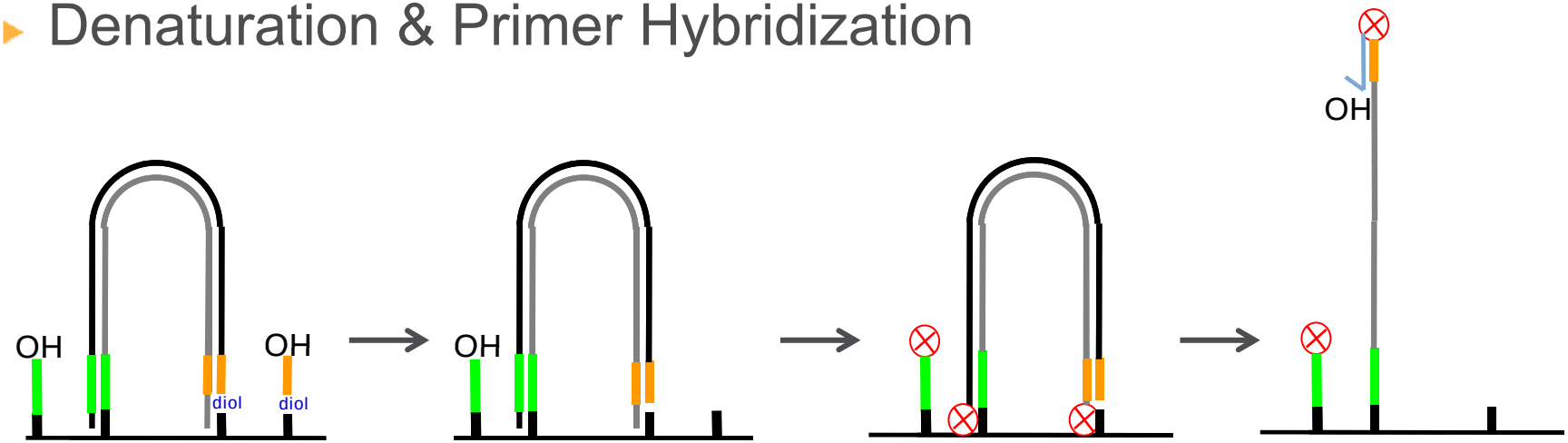
Cluster generation – hybridization and amplification



illumina

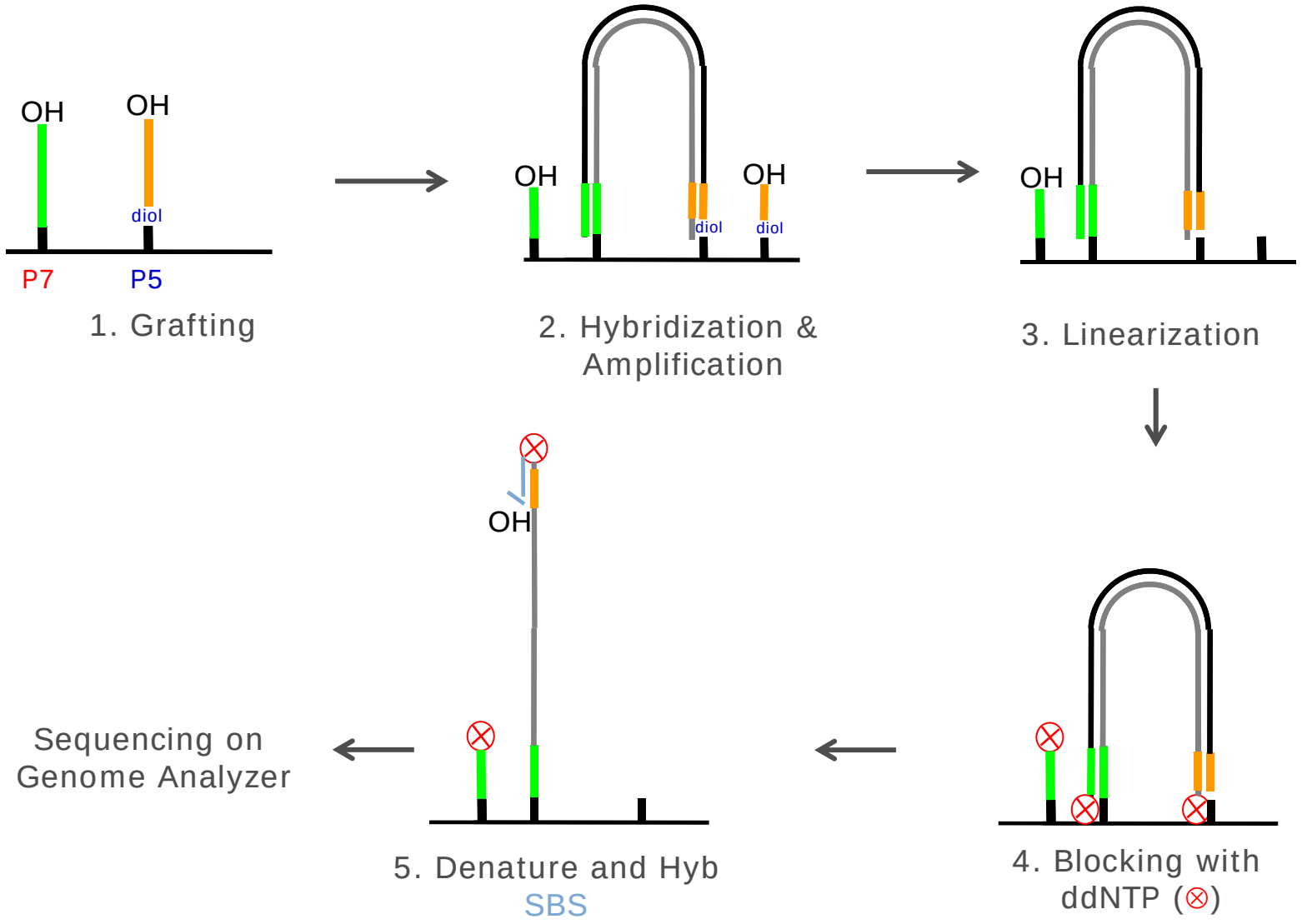
Cluster generation for single read runs

- Linearization & Blocking
- Denaturation & Primer Hybridization



illumina

Summary of cluster generation for single reads

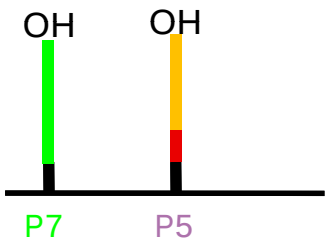


illumina

How does this differ for paired end sequencing?

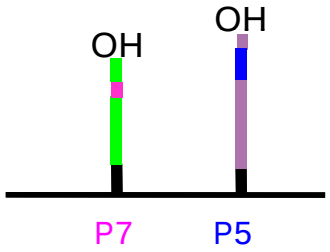
- *The answer is in the flow cell oligos*

SINGLE READ



Periodate Linearization

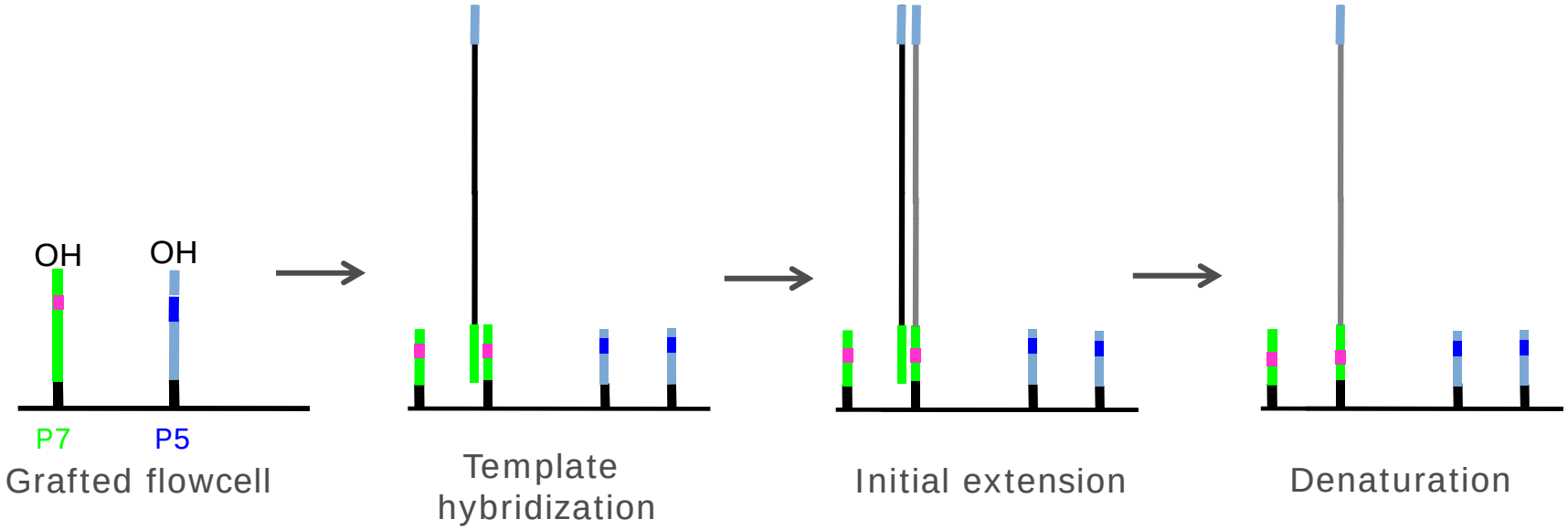
PAIRED END



Linearization 1 Enzyme
Linearization 2 Enzyme

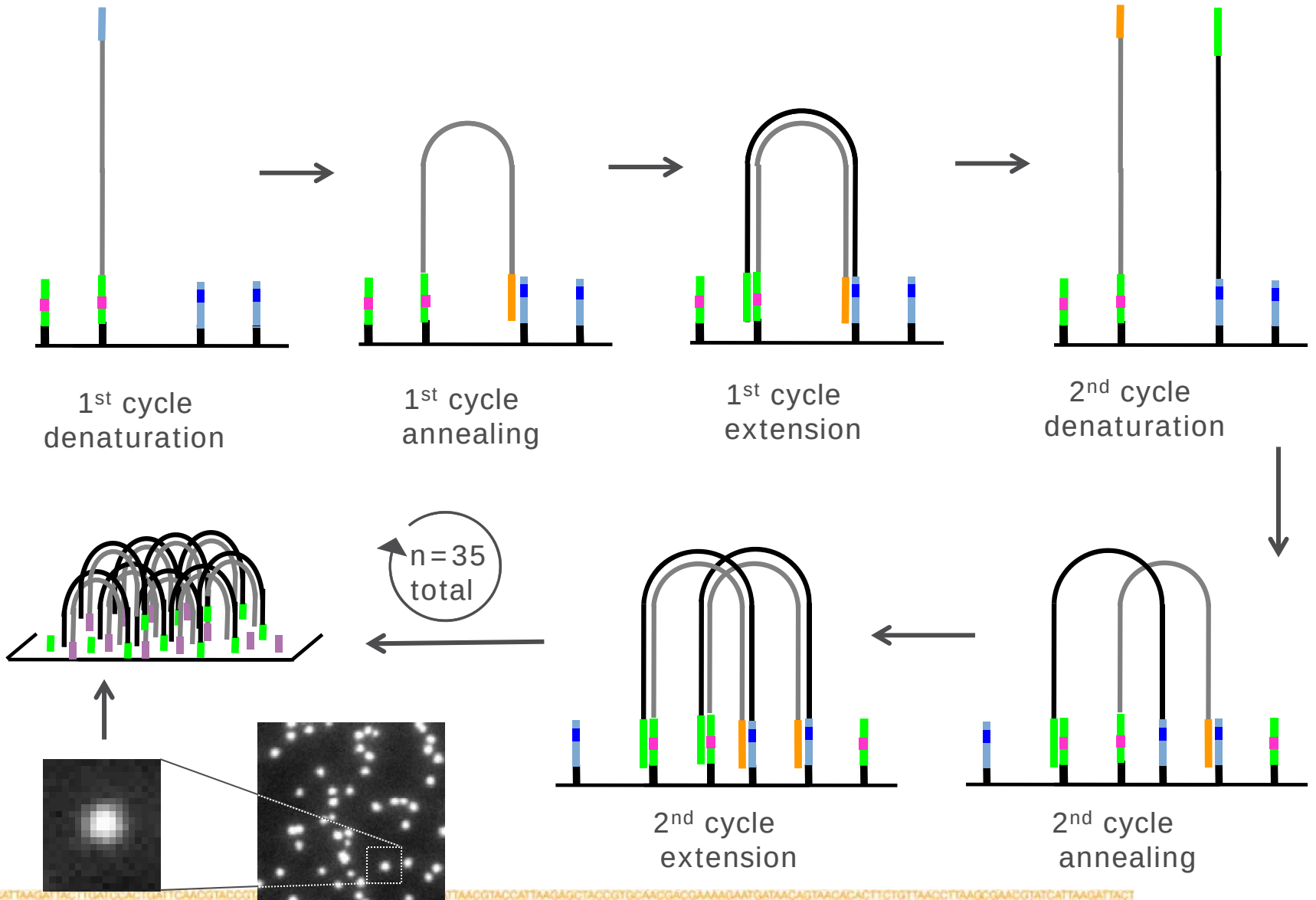
illumina

Paired end cluster generation



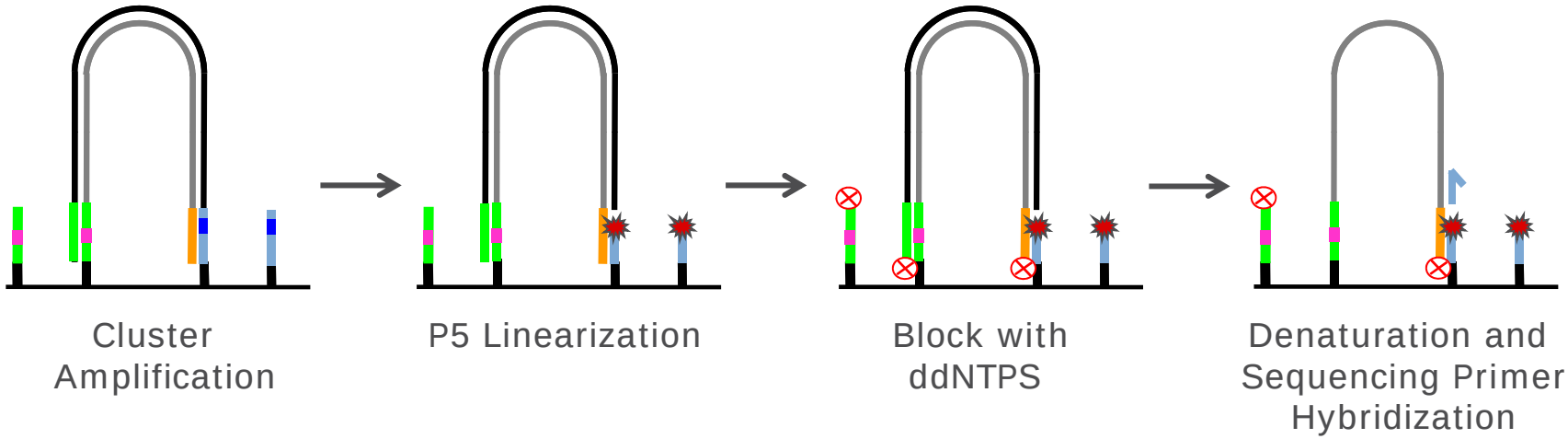
illumina

Paired end cluster generation



illumina

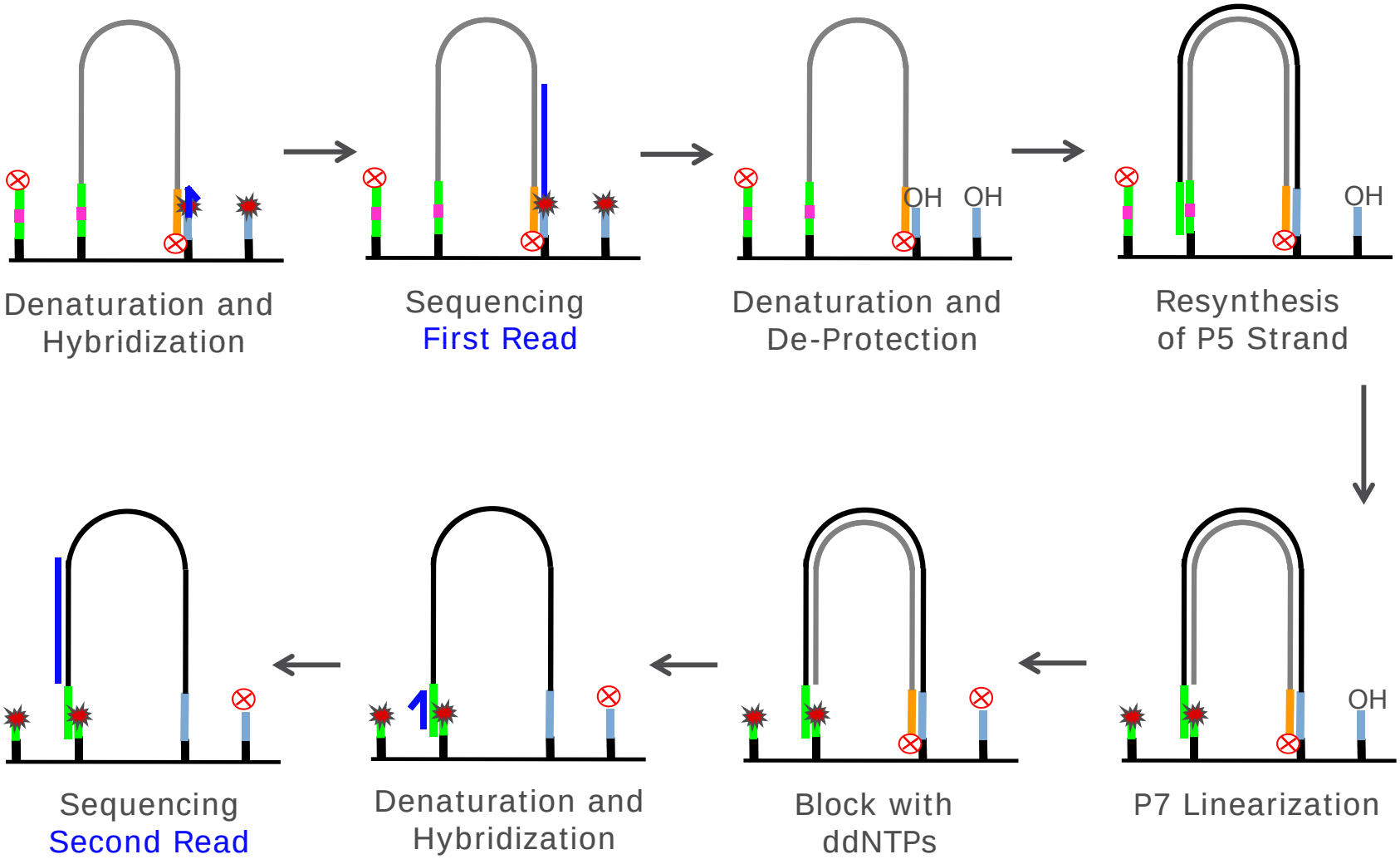
Paired end linearization



So far, it's all the same as Single Read except we use an ENZYME (Linearization Enzyme 1) to perform the linearization instead of a CHEMICAL (periodate)

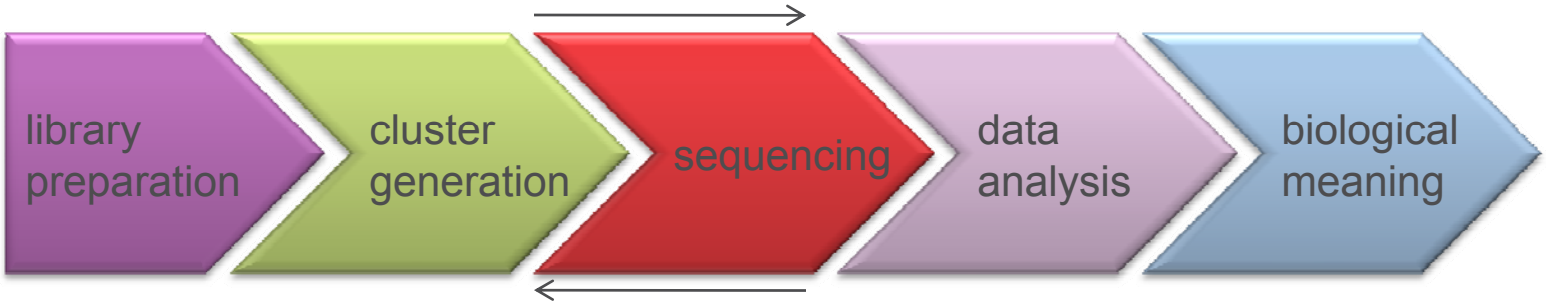
illumina

Paired end sequencing – read 1 and read 2



illumina

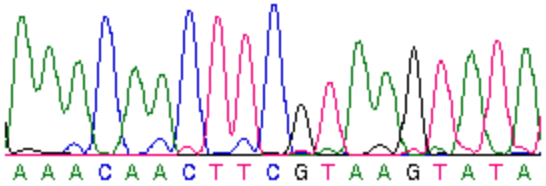
Sequencing



illumina

NexGen Sequencing vs Sanger

Sanger sequencing



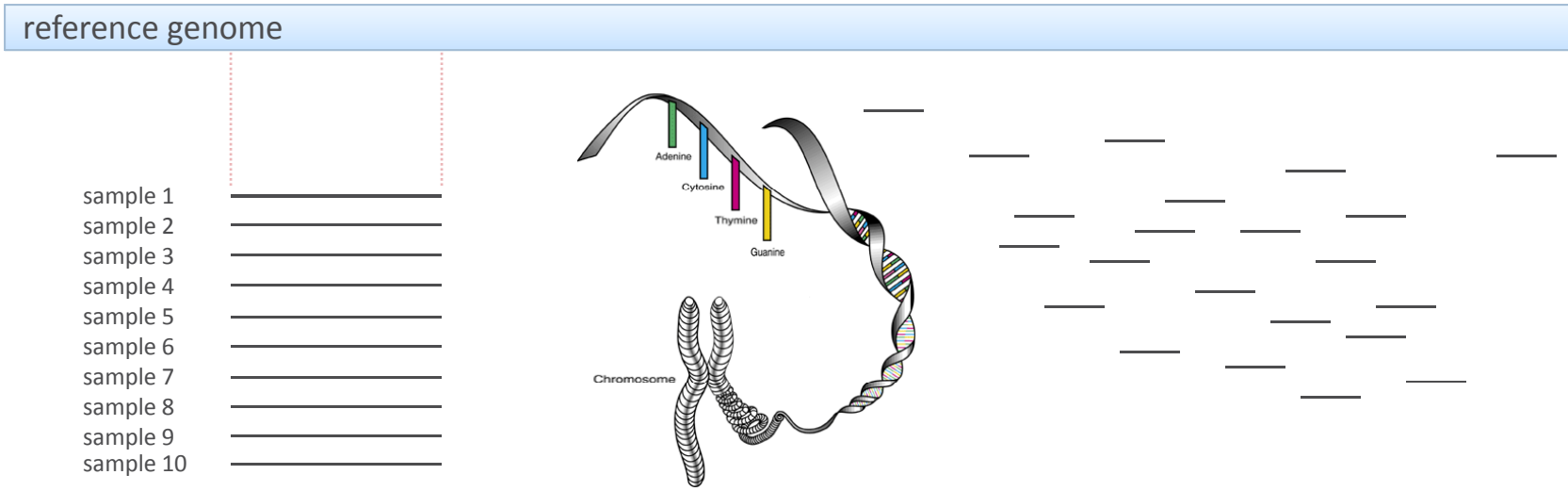
.....

one region or PCR product
(read length 350 - 850 bp)

Illumina sequencing

1 cycle = 1 bp
36 cycles = 36 bp

whole genome
(read length 18 - 100 bp)

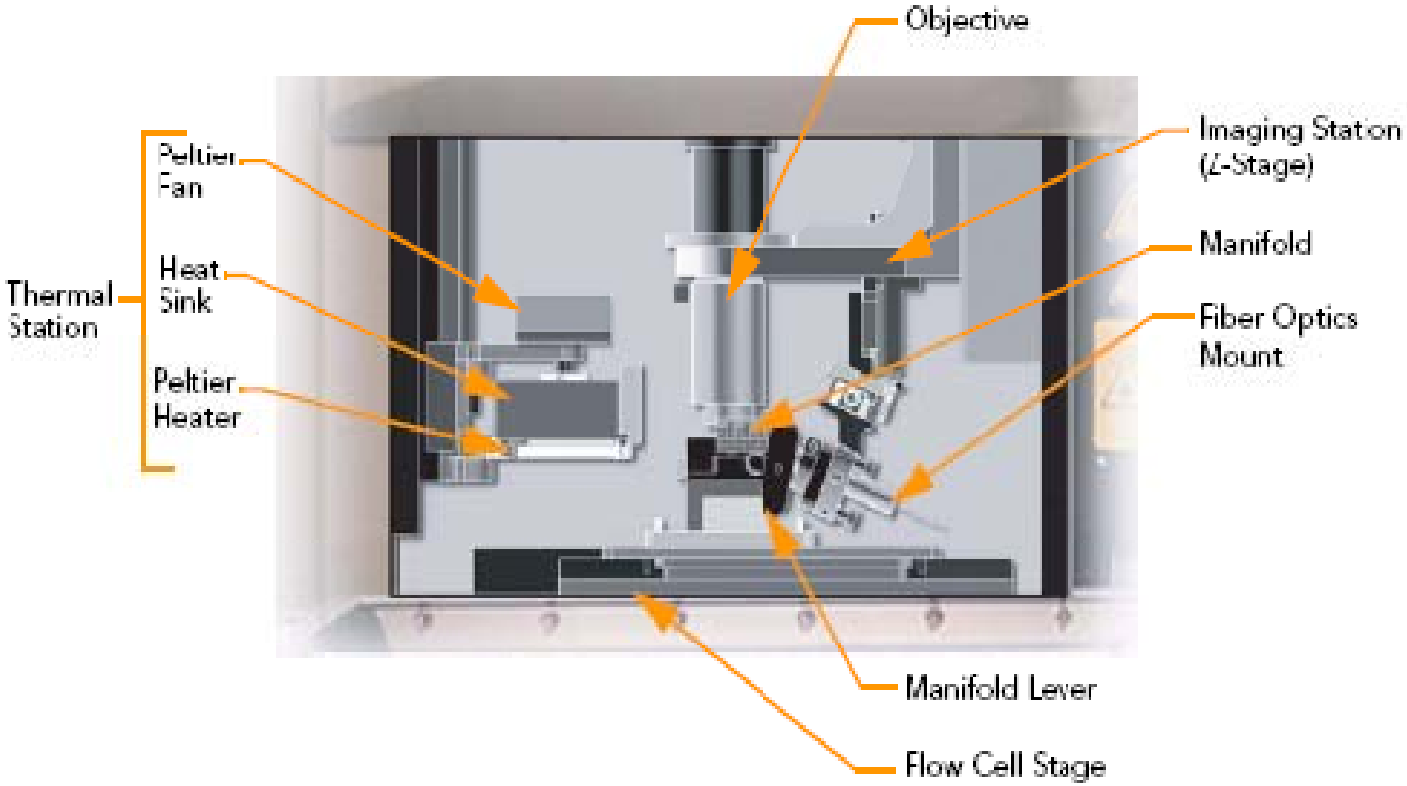


illumina

Genome Analyzer IIx

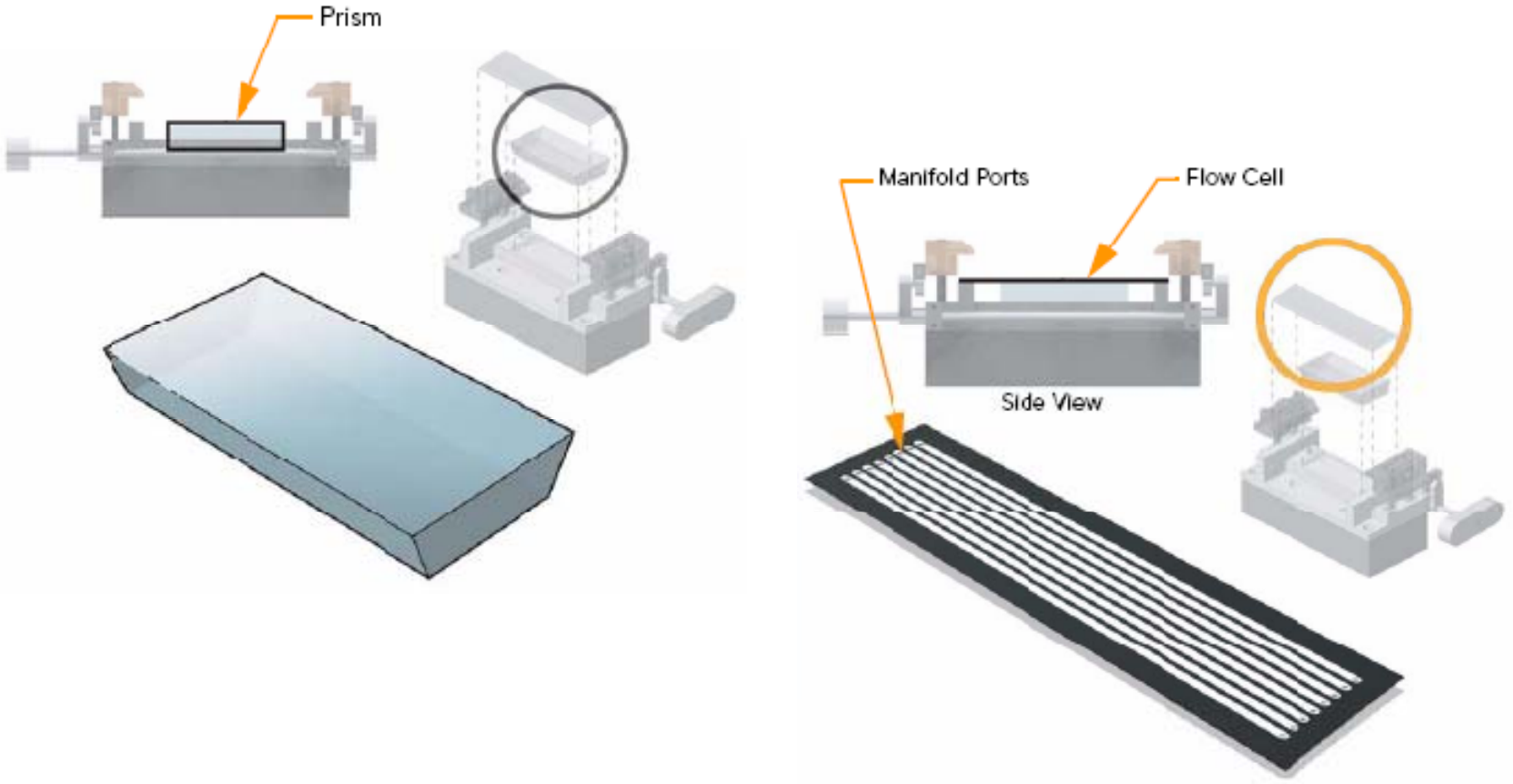
illumina

Genome Analyzer IIx - Imaging compartment



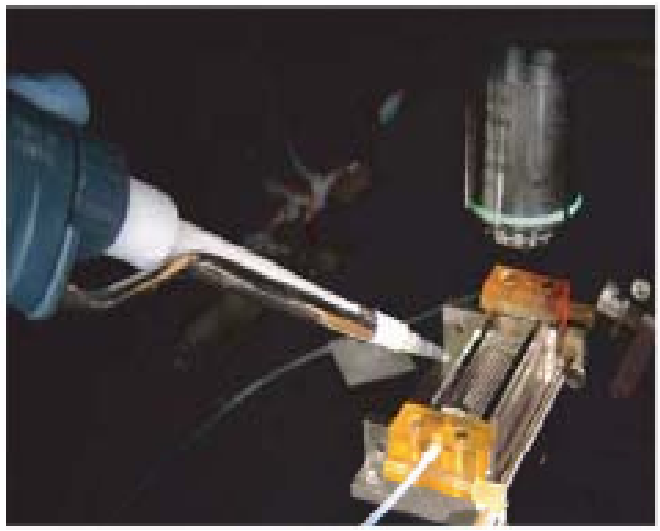
illumina

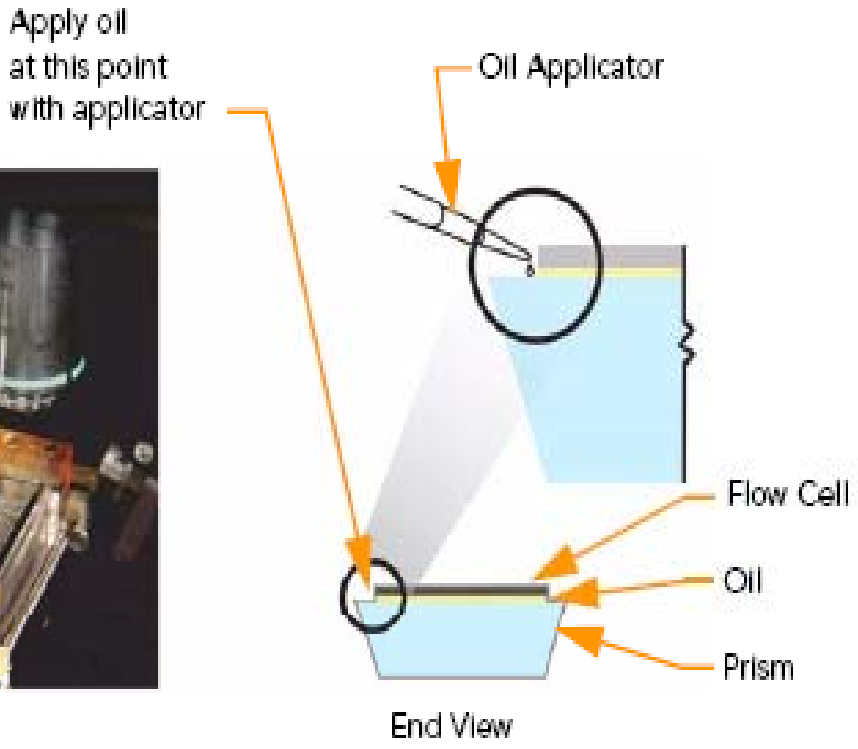
Imaging Compartment



illumina

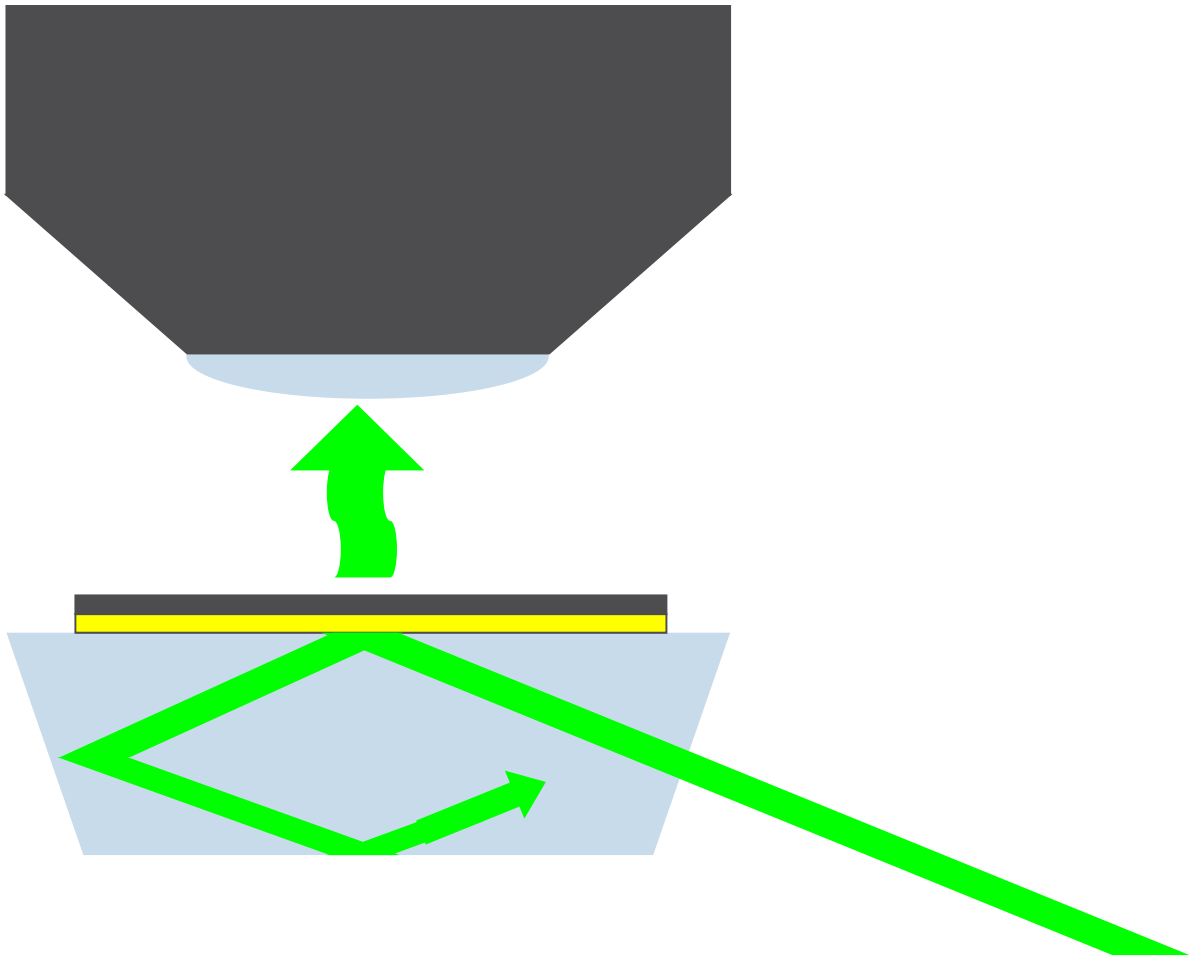
GA IIx - Imaging Compartment



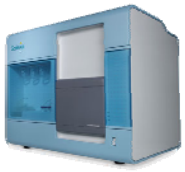


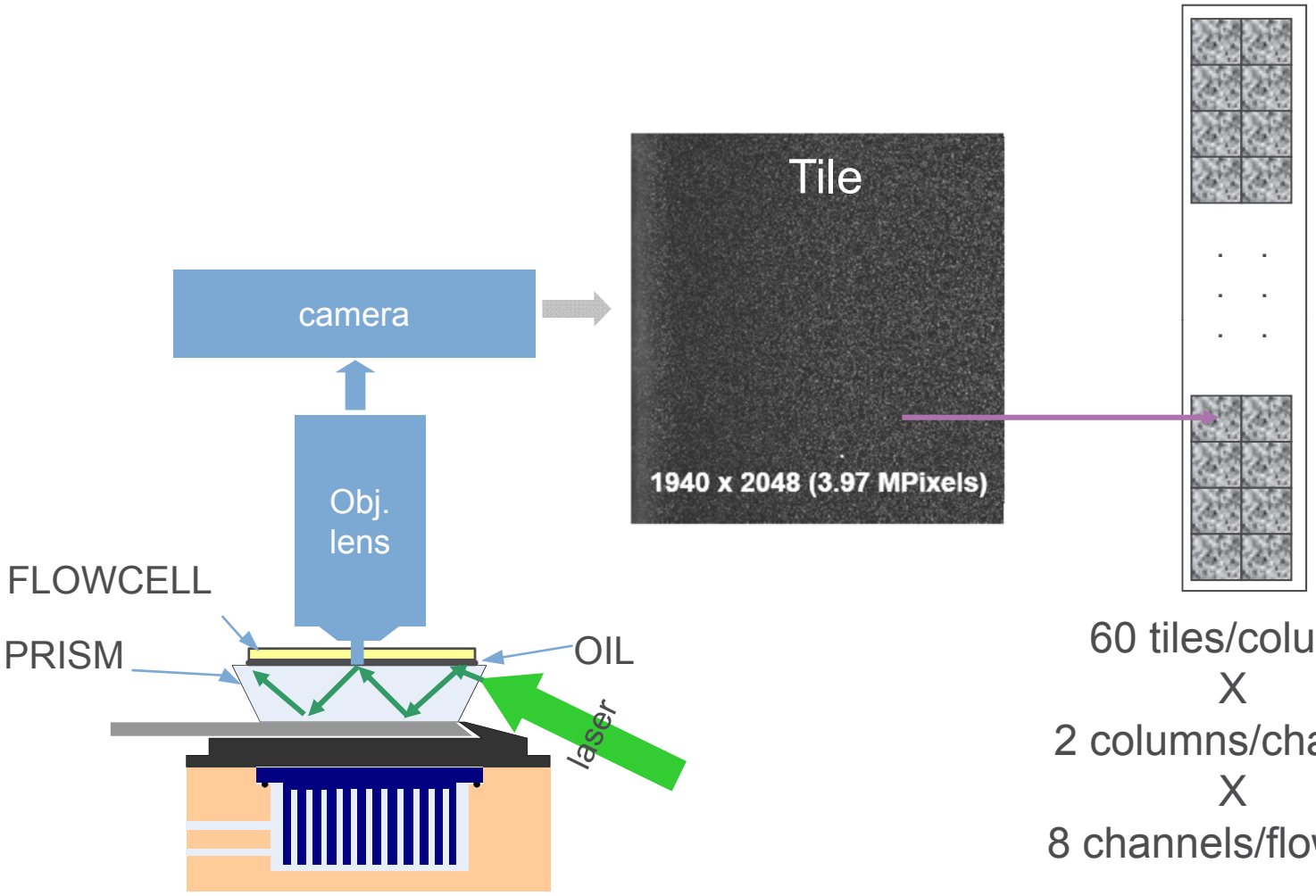
illumina

Imaging Compartment



GA IIx - Imaging Set Up

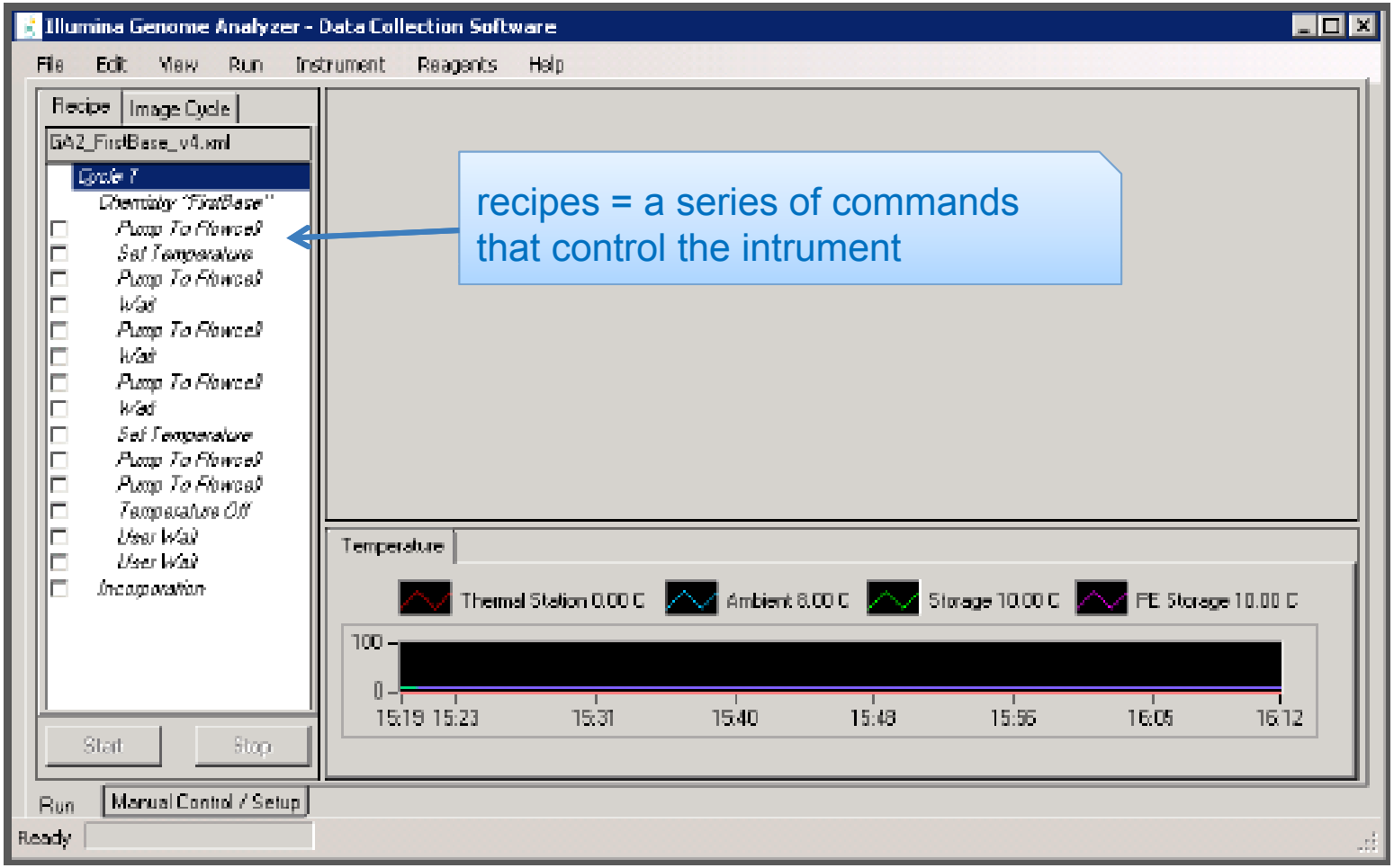




60 tiles/column
X
2 columns/channel
X
8 channels/flow cell

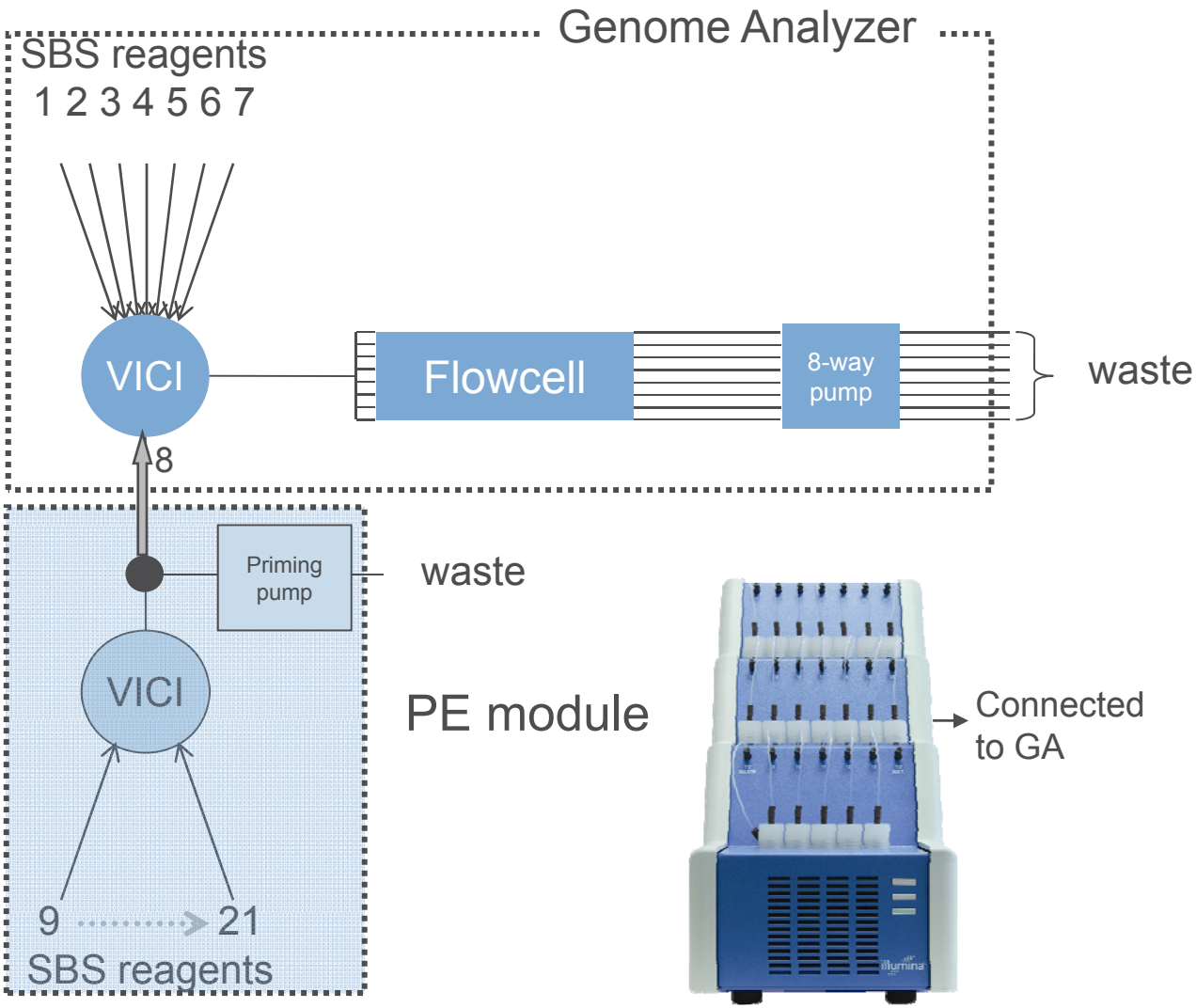
illumina

Genome Analyzer IIx
SCS - the data collection software



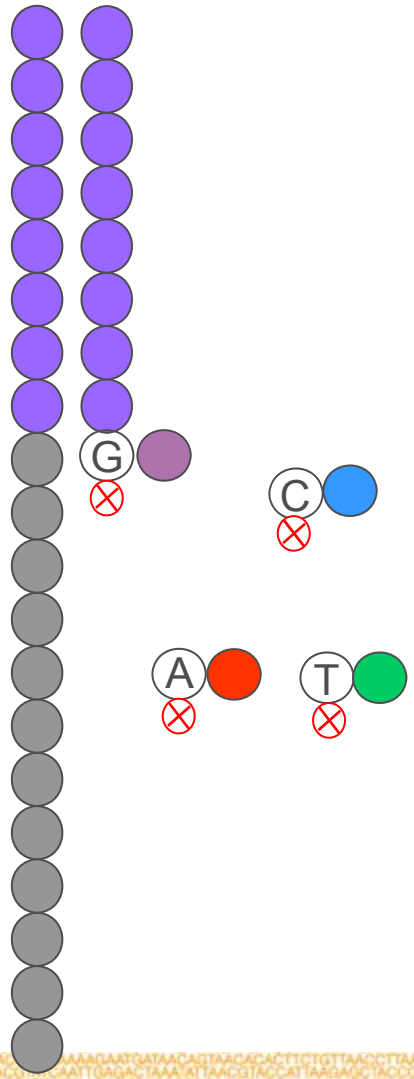
illumina

GAIIx - Paired end sequencing



illumina

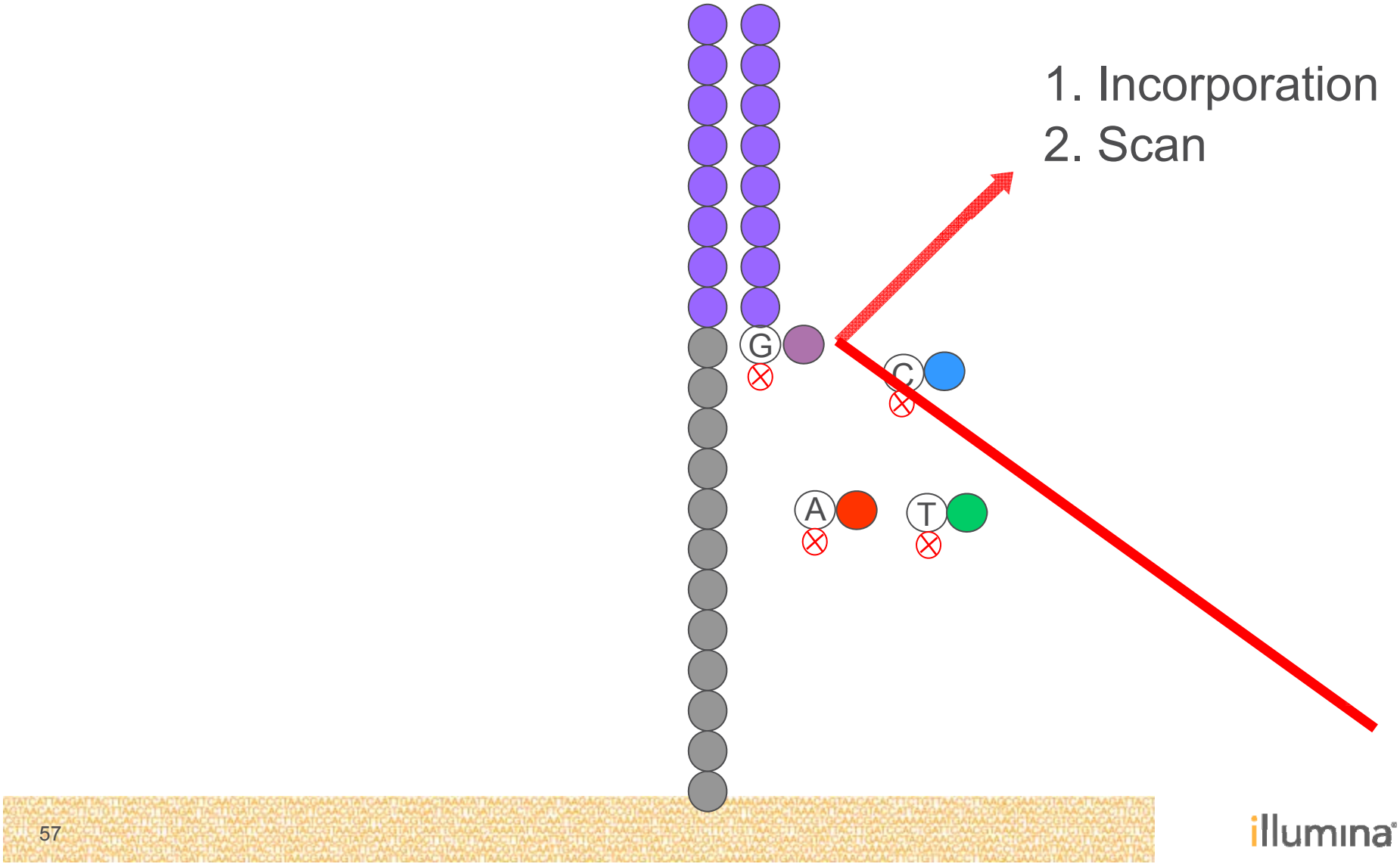
Sequencing by synthesis



1. Incorporation

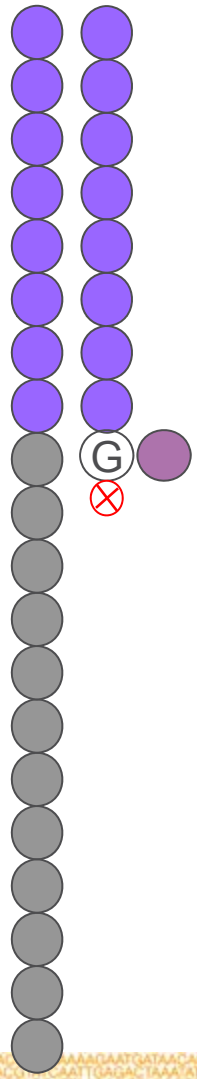
illumina

Sequencing by synthesis



illumina

Sequencing by synthesis



1. Incorporation
2. Scan
3. Cleavage

illumina

Genome Analyzer$_{IIx}$ Performance Specifications

Performance Parameters
27.5 – 33 Gb of high quality data / run
2.7 – 3.3 Gb / day
276-338M 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

illumina

The ENCODE Project

- a project with a vision to identify functional genomic elements

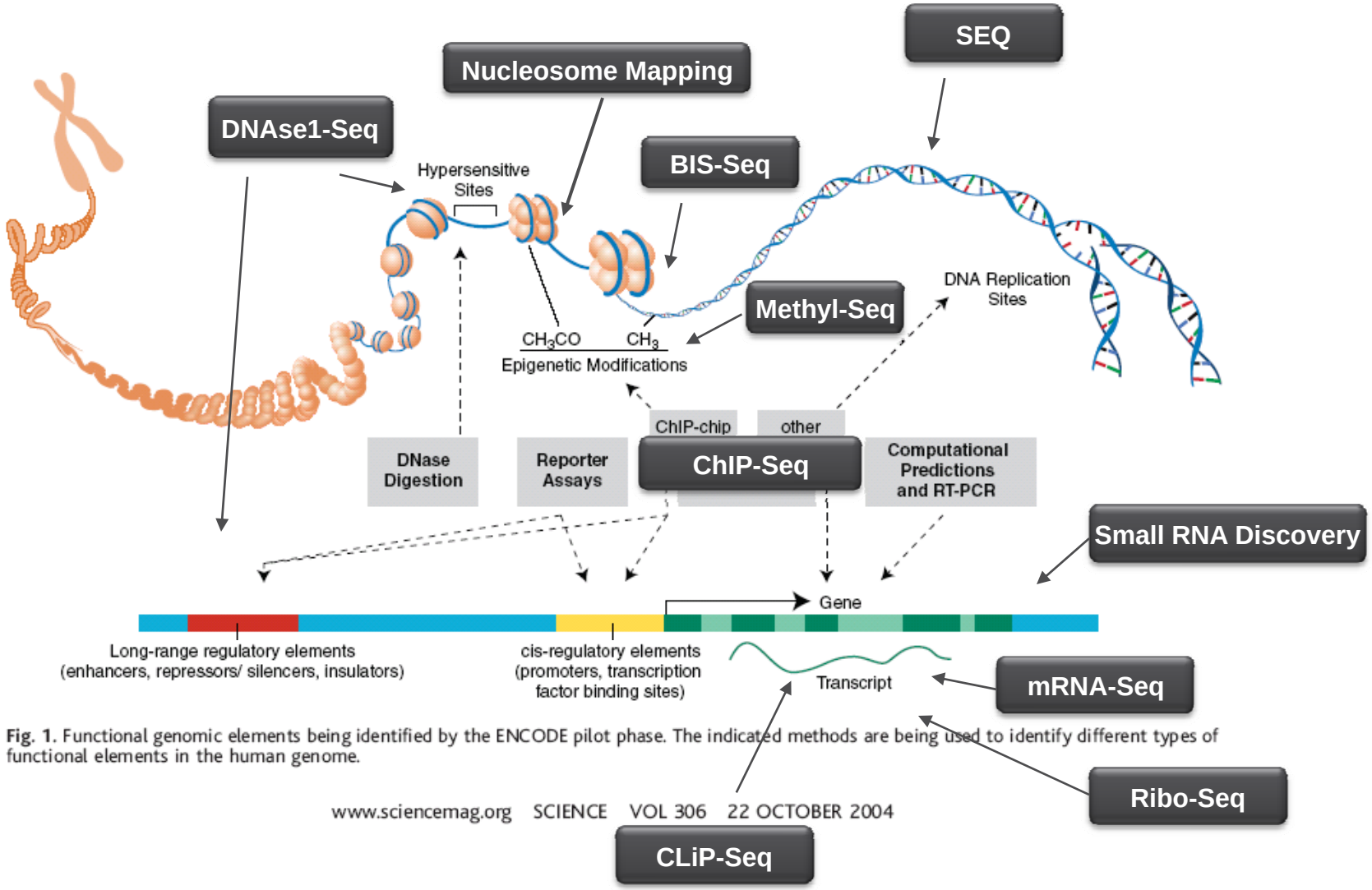


Fig. 1. Functional genomic elements being identified by the ENCODE pilot phase. The indicated methods are being used to identify different types of functional elements in the human genome.

www.sciencemag.org SCIENCE VOL 306 22 OCTOBER 2004

illumina

Analysis

illumina

Data Analysis Overview

Instrument Control (SCS)

Real Time Analysis (RTA)

Image Analysis → Base Caller → QC and Stats

Align

Allele call

Indels

Transcripts

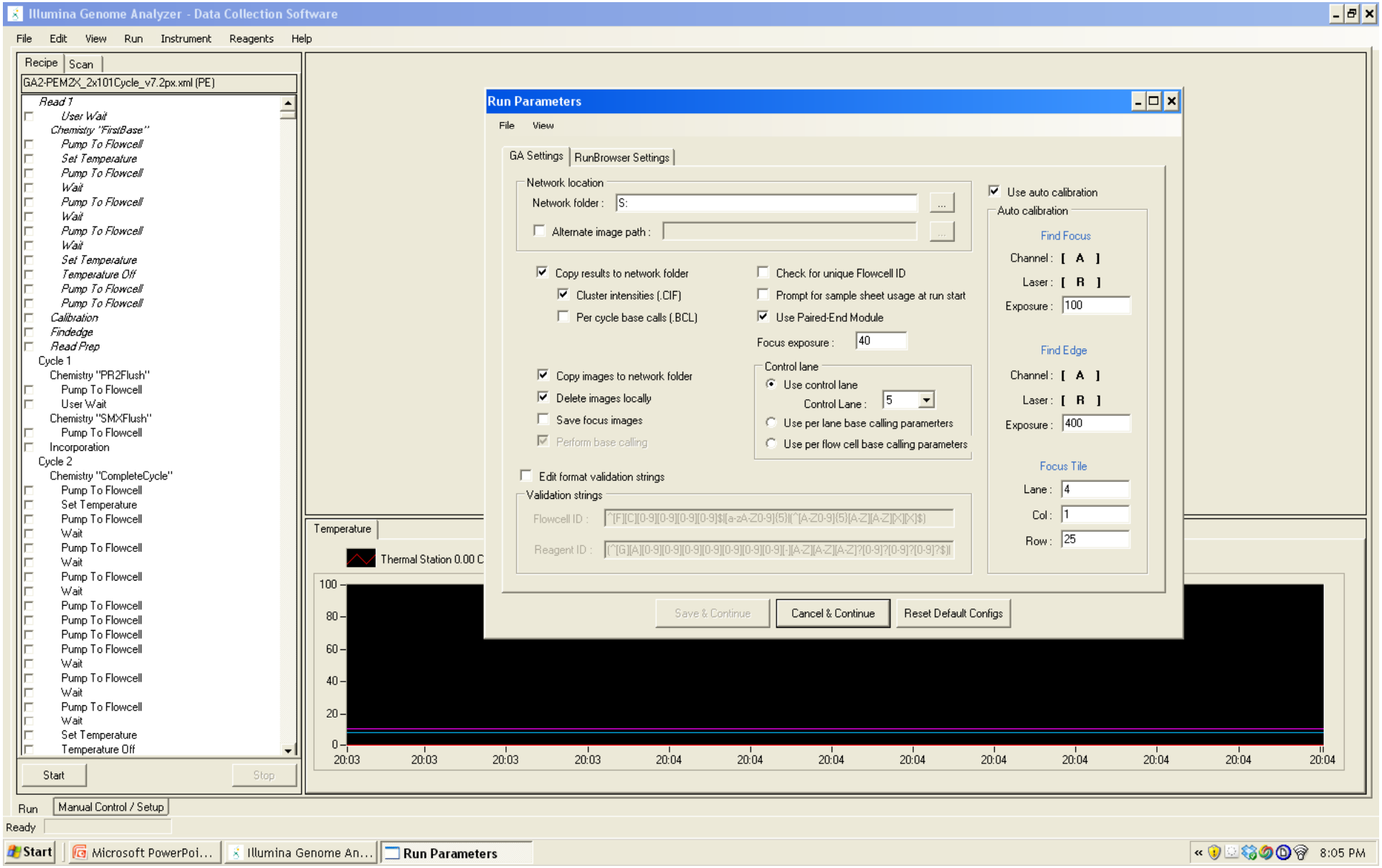
CASAVA

Raw Images → Processed data

Software

- SCS 2.6 (**S**equencing **C**ontrol **S**tudio) → G*A PC*
- RTA 1.6 (Integrated **R**eal **T**ime **A**nalysis) → *GA PC*
- OLB 1.6 (**O**ff-Line **B**asecaller) → *Server*
- CASAVA 1.6 (**C**onsensus **A**ssessment of **S**equence **A**nd **Va**riation) → *Server*
  - *GERALD* (**Ge**neration of **R**ecursive **A**nalyses **L**inked by **D**ependency)
    - ELAND v2 (**E**fficient **L**arge-Scale **A**lignment of **N**ucleotide **D**atabases)
    - PhageAlign

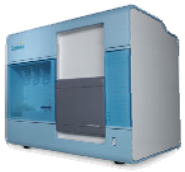
Sequencing Control Software v2.6

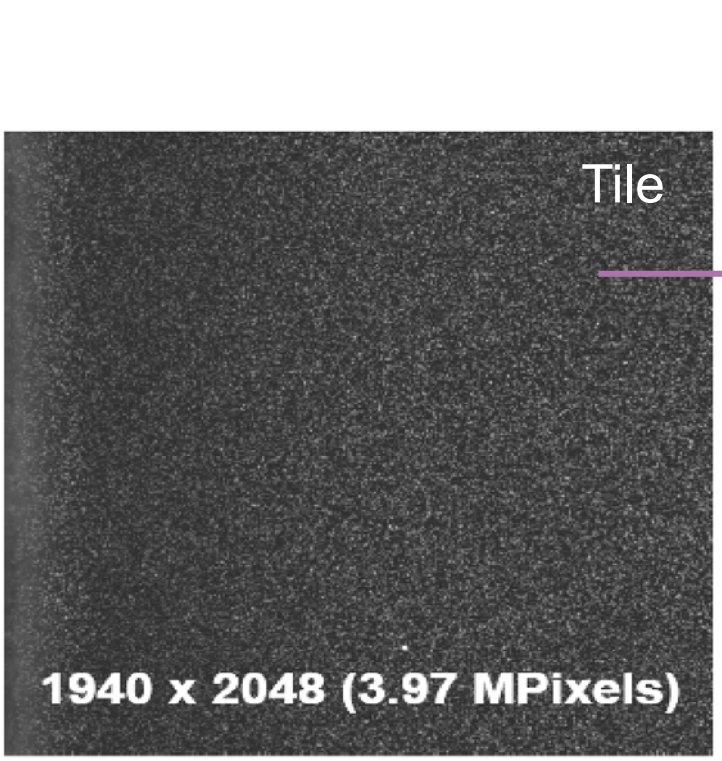


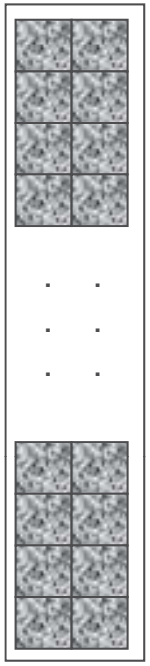
Real Time Analysis-RTA

- About RTA
  - State Machine
  - Output of every process is a file that acts as the trigger for the next step
  - Multi-threaded
  - Works in the background
- Input:
  - Image files (.tiff)
- Output:
  - qSeq.txt files
- Provides Real-Time metrics through graphs
- Transfers data from GA computer to server

Real-Time Analysis Input







60 tiles/column
X
2 columns/channel
X
8 channels/flow cell

How many images?

Real Time Analysis (RTA)

- Firecrest:
  - Image Analysis/Template generation or cluster positioning
  - Registration
    - Aligning template cluster positions with image
  - Intensity Extraction
  - Color Matrix (Cross Talk adjustment)
  - Phasing/Prephasing
- Bustard:
  - Basecalling
  - qSeq file generation

Real Time Analysis-RTA

	Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
MIN	332K	433K	549K	622K	332K	472K	540K	620K
MAX	464K	608K	664K	725K	442K	593K	679K	722K
P50	375K	524K	593K	665K	374K	515K	604K	660K

illumina

Real Time Analysis-RTA Output

- Qseq.txt
- Tab-delimited: easy to parse, easy to import into databases
- Split files per read on a read pair / multiple read run

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	]baabbbbaaa]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

OLB 1.6

- “Offline Basecaller” (OLB)
- OLB is simply the old “Pipeline” software, but without the Gerald module
- OLB contains image analysis (i.e., “Firecrest”) and basecalling (“Bustard”)
- OLB is Linux-based
- All OLB algorithms identical to RTA 1.6 algorithms (equivalence directly tested vs. RTA)
- OLB will be released as a “tarball” (compress Linux .tar file)
- OLB Users Guide a separate downloadable .pdf file via iCom
- DVD’s no longer provided (not for PL, OLB or SCS)

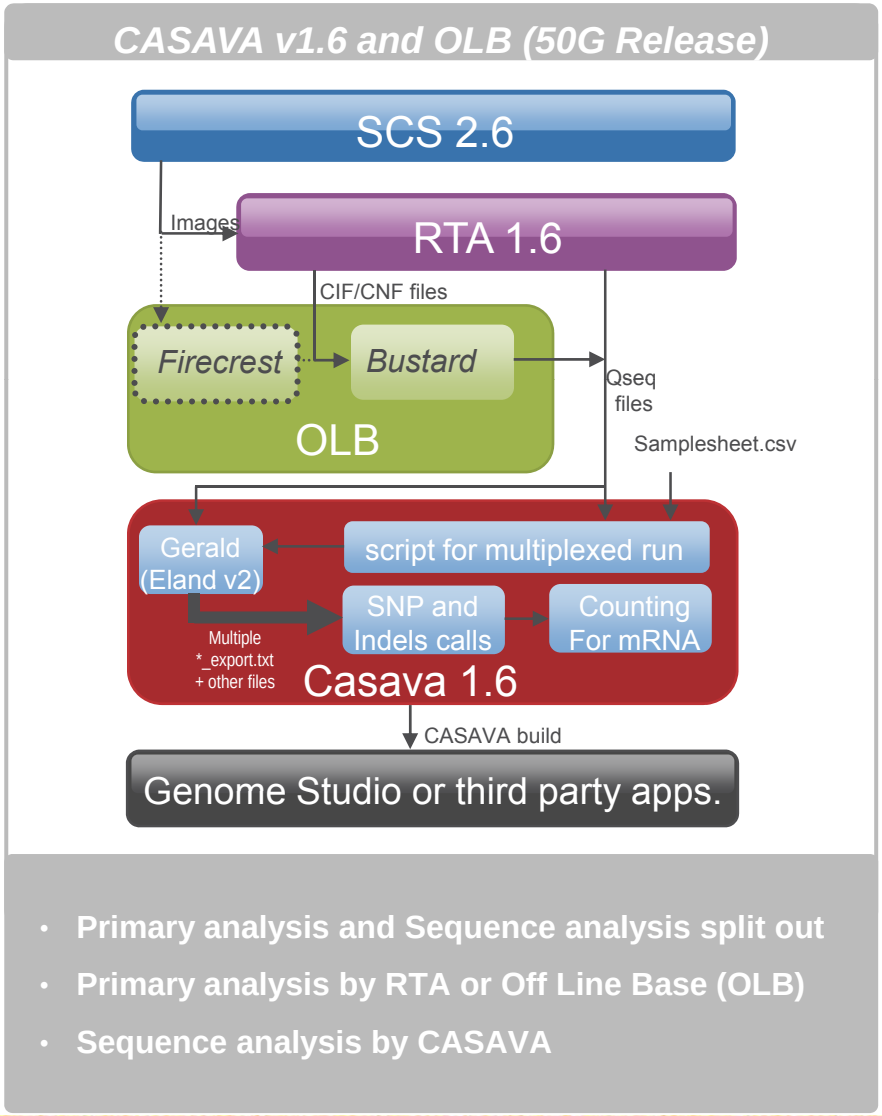
CASAVA 1.6 High Level Features Summary

- GERALD and ELANDv2 Aligner
  - Supports gapped and multiseed alignment.
  - Use of Custom quality calibration tables is supported
  - Workflow simplification with the introduction of static make files
- Demultiplexing
  - Support for indexed runs (RTA via Demultiplex.pl and multiplexedGERALD.pl
  - Sample Sheet - simple .csv file implementation
- Updated Allele Caller
  - Significantly faster (approximately 20-fold)
  - Correctly handles reads with gapped alignments and 100bp reads
- Updated SNP Caller
  - SNP caller updated to increase heterozygous SNP sensitivity.

CASAVA 1.6 High Level Features Summary

- GROUPER – Indel Detector
  - Uses singleton/shadow read pairs to detect indels.
- RNA SEQ improvements
  - ReadBases (counting by number of bases) is now the default counting method
  - Normalization of RNA counts for exons and genes now based on RPKM (reads per kb per million aligned reads)
- Performance Improvements
  - Lower memory footprint and bandwidth
  - Removed chromosome name parsing
  - Chromosomes will be designated by the fasta reference filename

Overall Analysis Workflow



- **Primary analysis and Sequence analysis split out**
- **Primary analysis by RTA or Off Line Base (OLB)**
- **Sequence analysis by CASAVA**

illumina

Genome Studio

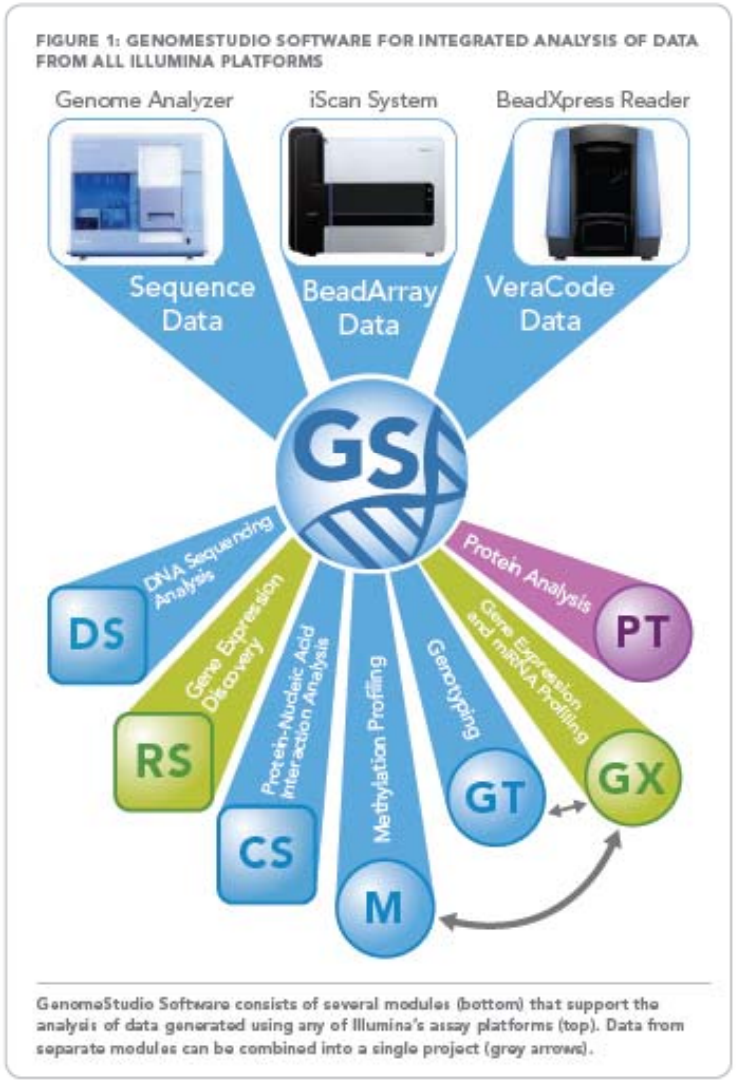


FIGURE 1: GENOMESTUDIO SOFTWARE FOR INTEGRATED ANALYSIS OF DATA FROM ALL ILLUMINA PLATFORMS

GenomeStudio Software consists of several modules (bottom) that support the analysis of data generated using any of Illumina's assay platforms (top). Data from separate modules can be combined into a single project (grey arrows).

DNA Sequencing Module (DS)

- Software tools can be analyzed to discover and confirm SNPs and chromosomal breakpoint regions. Visualization tools display consensus reads in the reassembled genome and indicate SNPs with colored letters

ChIP Sequencing Module (CS)

- Create global binding site maps of DNA-associated proteins. Differential binding levels between experimental groups can be identified by comparing sequences, regions, and peaks in table or chromosome views.

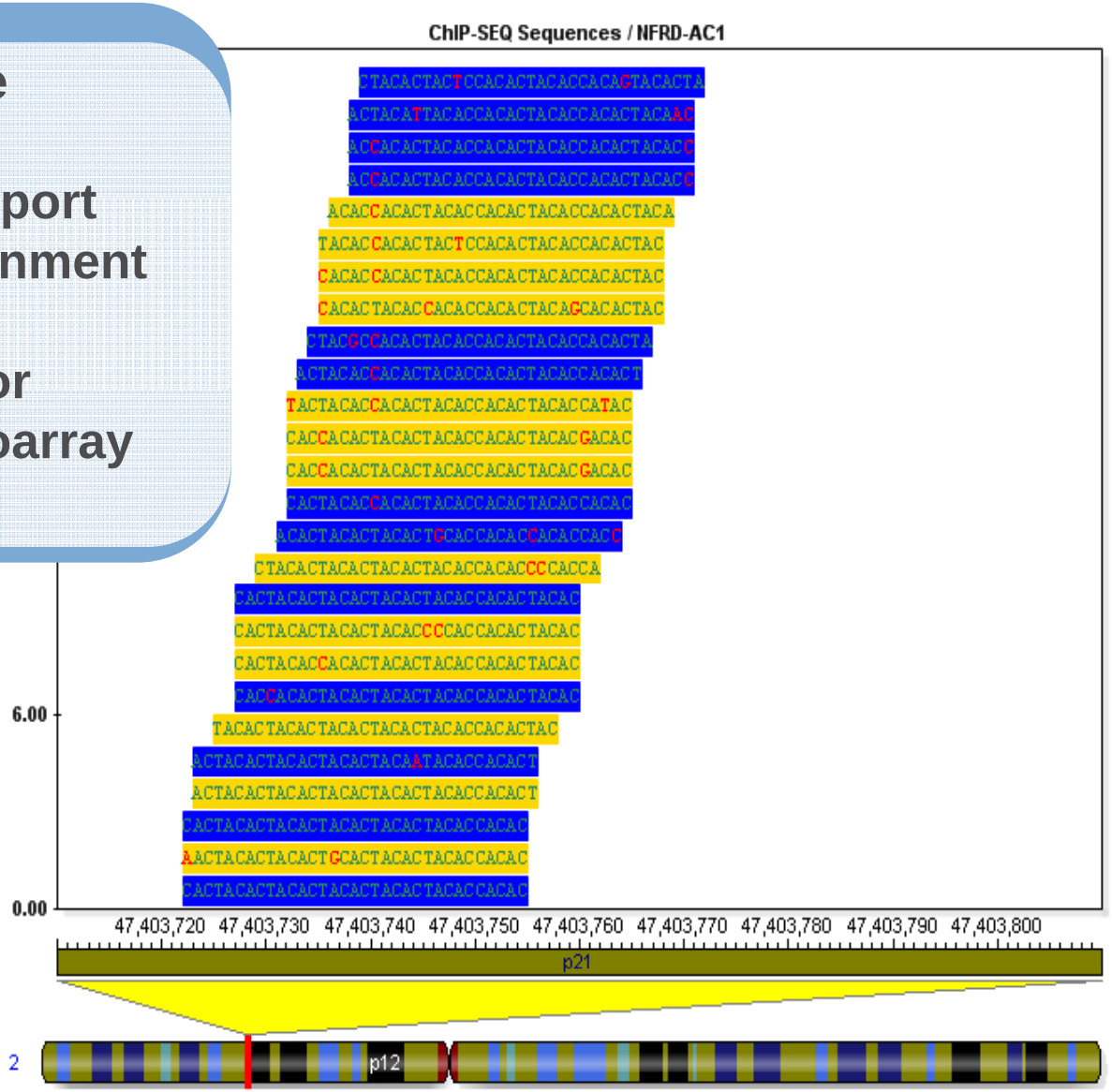
RNA Sequencing Module (RS)

- This module performs digital gene expression profiling by aggregating data from the Genome Analyzer Pipeline Software to count the abundance of reads falling within specific exons, genes, and splice junctions. The data are then graphically displayed as tables or plots with GenomeStudio Software

illumina

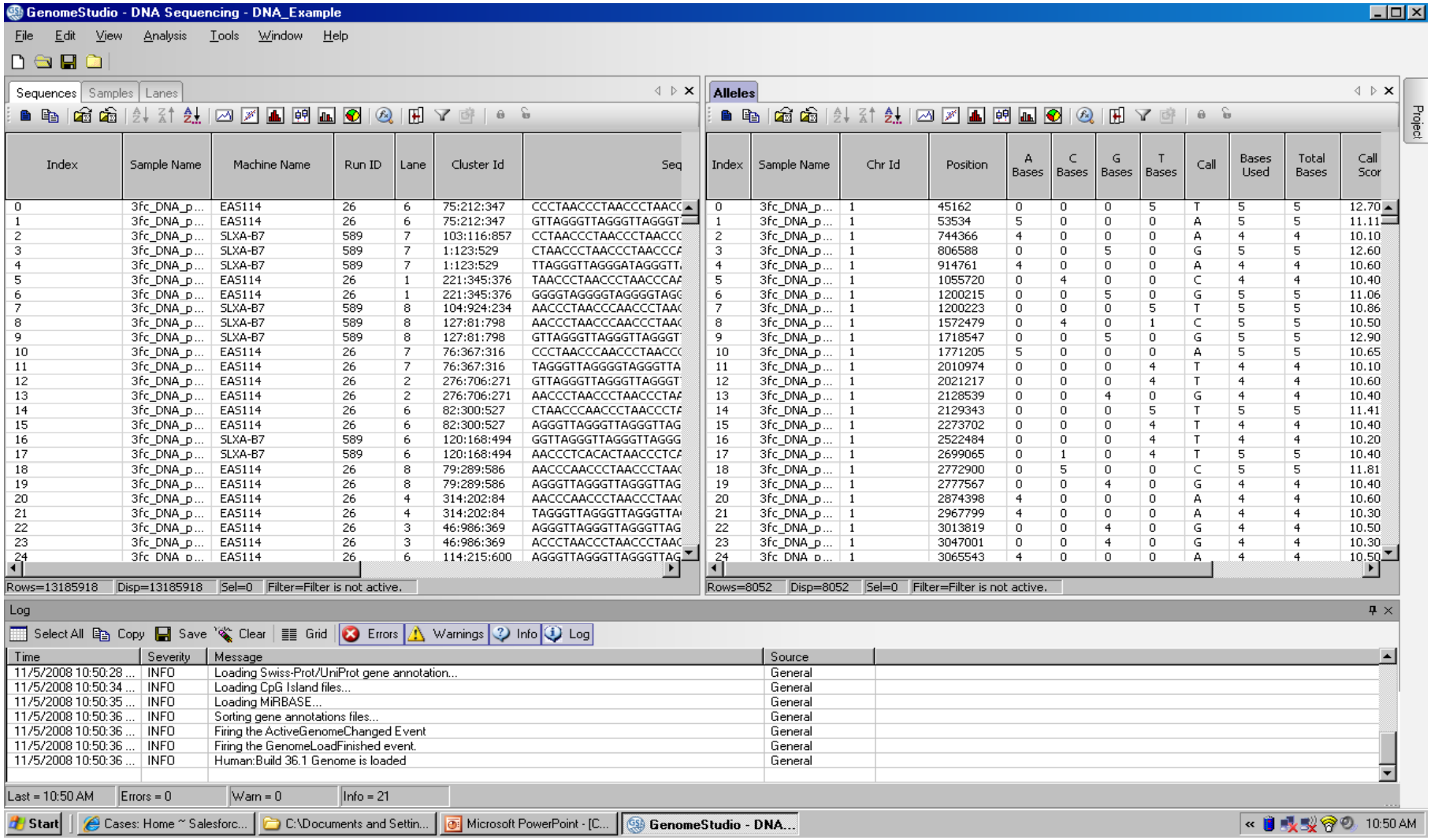
Genome Studio Viewer

- An integrated genome browser
- Fast and easy data import
- Standard stacked alignment plots
- A common platform for sequencing and microarray applications



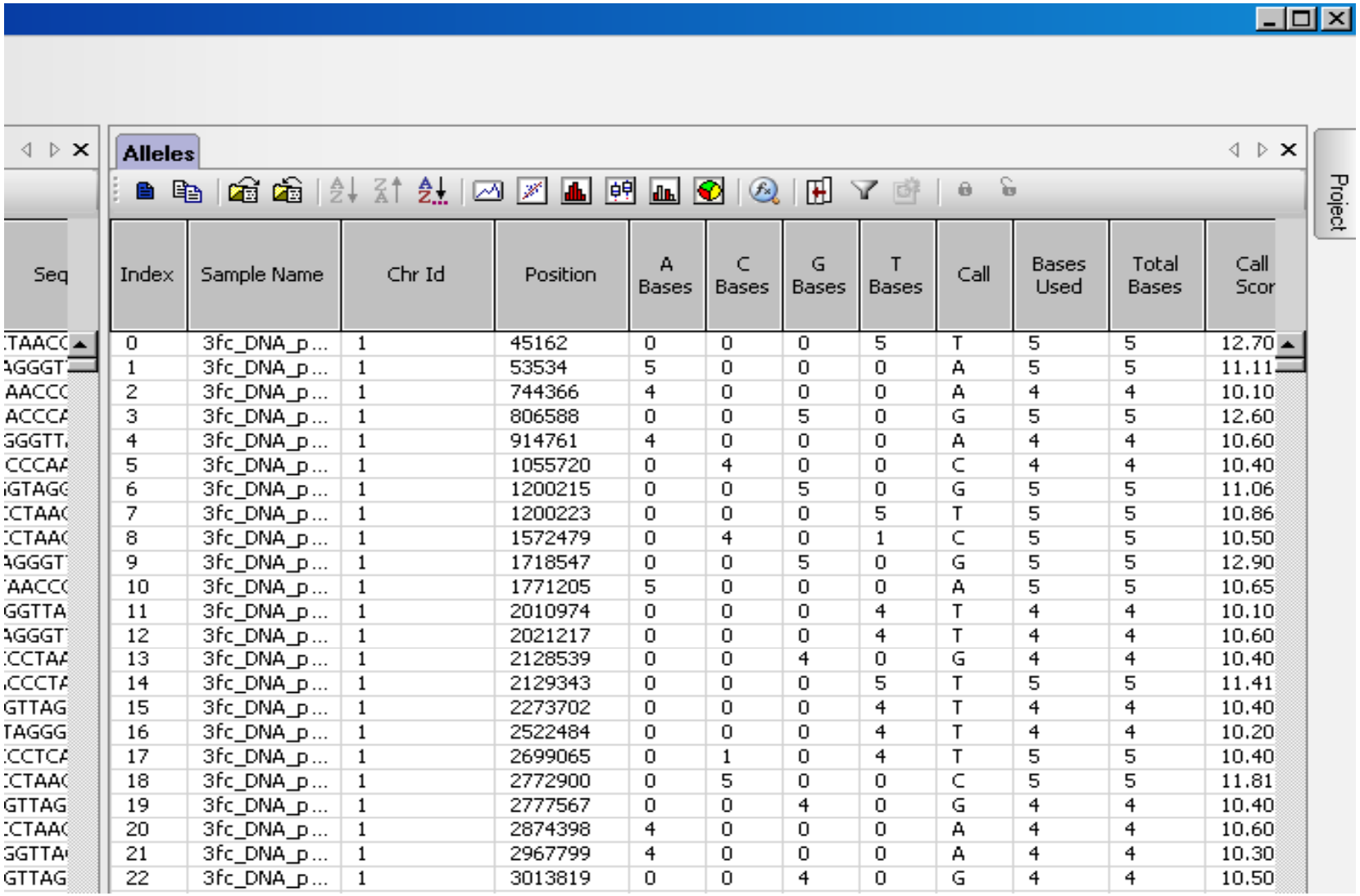
illumina

Genome Studio Module



illumina

Genome Studio Module



Seq	Index	Sample Name	Chr Id	Position	A Bases	C Bases	G Bases	T Bases	Call	Bases Used	Total Bases	Call Scor
TAACC	0	3fc_DNA_p...	1	45162	0	0	0	5	T	5	5	12.70
AGGGT	1	3fc_DNA_p...	1	53534	5	0	0	0	A	5	5	11.11
AACCC	2	3fc_DNA_p...	1	744366	4	0	0	0	A	4	4	10.10
ACCCA	3	3fc_DNA_p...	1	806588	0	0	5	0	G	5	5	12.60
GGGTT	4	3fc_DNA_p...	1	914761	4	0	0	0	A	4	4	10.60
CCCAA	5	3fc_DNA_p...	1	1055720	0	4	0	0	C	4	4	10.40
GTAGG	6	3fc_DNA_p...	1	1200215	0	0	5	0	G	5	5	11.06
CTAAC	7	3fc_DNA_p...	1	1200223	0	0	0	5	T	5	5	10.86
CTAAC	8	3fc_DNA_p...	1	1572479	0	4	0	1	C	5	5	10.50
AGGGT	9	3fc_DNA_p...	1	1718547	0	0	5	0	G	5	5	12.90
AACCC	10	3fc_DNA_p...	1	1771205	5	0	0	0	A	5	5	10.65
GGTTA	11	3fc_DNA_p...	1	2010974	0	0	0	4	T	4	4	10.10
AGGGT	12	3fc_DNA_p...	1	2021217	0	0	0	4	T	4	4	10.60
CCTAA	13	3fc_DNA_p...	1	2128539	0	0	4	0	G	4	4	10.40
CCCTA	14	3fc_DNA_p...	1	2129343	0	0	0	5	T	5	5	11.41
GTTAG	15	3fc_DNA_p...	1	2273702	0	0	0	4	T	4	4	10.40
TAGGG	16	3fc_DNA_p...	1	2522484	0	0	0	4	T	4	4	10.20
CCTCA	17	3fc_DNA_p...	1	2699065	0	1	0	4	T	5	5	10.40
CTAAC	18	3fc_DNA_p...	1	2772900	0	5	0	0	C	5	5	11.81
GTTAG	19	3fc_DNA_p...	1	2777567	0	0	4	0	G	4	4	10.40
CTAAC	20	3fc_DNA_p...	1	2874398	4	0	0	0	A	4	4	10.60
GGTTA	21	3fc_DNA_p...	1	2967799	4	0	0	0	A	4	4	10.30
GTTAG	22	3fc_DNA_p...	1	3013819	0	0	4	0	G	4	4	10.50

illumina

Illumina evaluated and third Party Tools

third-party genome analyzer data analysis tools

The evolution of next-generation sequencing technology will continue to require collabor researchers, thought leaders, and industry. Illumina encourages the open exchange of in and ideas, and we are committed to rapidly delivering solutions that will maximize your success.

Below you will find tools developed by Illumina sequencing customers that you can use t your Genome Analyzer data. These software applications* are designed to be used for a sequencing applications such as resequencing, ChIP-Seq, and Digital Gene Expression. C customers have made these tools freely available to you.

Genome Alignment Browsers

- **Gbrowse - Genomic Browsing**
  Generic Model Organism Database Project
  http://www.gmod.org/wiki/index.php/Gbrowse
- **UCSC Browser - Genome browsing and comprehensive annotation**
  Generic Model Organism Database Project
  http://genome.ucsc.edu/goldenPath/help/customTrack.html
- **Staden Tools (GAP4, TGAP) - Alignment and Visualization for Small Data S**
  James Bonifield (initially developed by Rodger Staden), Wellcome Trust Sanger In
  http://sourceforge.net/projects/staden/

Alignment and Polymorphism Detection

- **BFAST – Blat-like Fast Accurate Search Tool**
  Nils Homer, Stanley F. Nelson and Barry Merriman, University of California, Los An
  http://genome.ucla.edu/bfast
- **MAQ – Mapping and Assembly with Quality**
  Heng Li, Sanger Centre
  http://maq.sourceforge.net/maq-man.shtml
- **Bowtie - An ultrafast memory-efficient short read aligner**
  Ben Langmead and Cole Trapnell, Center for Bioinformatics and Computational Bi University of Maryland
  http://bowtie-bio.sourceforge.net/

Genomic Assembly

- **Velvet – De novo assembly of short reads**
  Daniel Zerbino and Ewan Birney, EMBL-EBI
  http://www.ebi.ac.uk/~zerbino/velvet/
- **SSAKE – Assembly of short reads**
  Rene Warren, et al, British Columbia Cancer Agency
  http://bioinformatics.oxfordjournals.org/cgi/content/full/23/4/500
- **Euler – Genomic Assembly**
  Pavel Pevzner and Mark Chaisson, University of California, San Diego
  http://nbcr.sdsc.edu/euler/

ChIP Sequencing

- **ChIP-Seq Peak Finder**
  Barbara Wold, Cal Tech and Rick Meyers, Stanford University
  http://woldlab.caltech.edu/html/software/

Digital Gene Expression

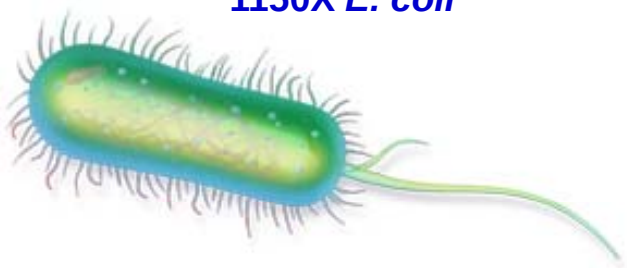
- **Comparative Count Display**

illumina

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

2x Human Genome

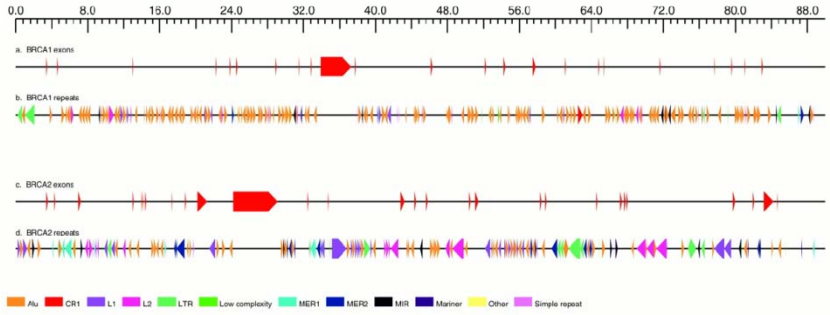
1150X *E. coli*

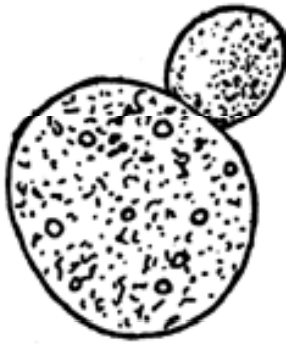


50X Arabidopsis

50X Drosophila

3000X BRCA1+BRCA2, 12 samples per lane





500X Yeast Genome

illumina

Illumina Next-Gen Sequencing

Sequence on a scale never before possible

ENABLES DEEP SEQUENCING OF MANY LARGE, COMPLEX GENOMES AND TRANSCRIPTOMES

FUEL MAJOR STUDIES INTO CANCER AND OTHER COMPLEX GENETIC DISEASES

DRIVE TRANSLATIONAL MEDICINE

MAKE ROUTINE WHOLE GENOME SEQUENCING POSSIBLE

Human
Crop
Livestock
Consumer Genomics

illumina®