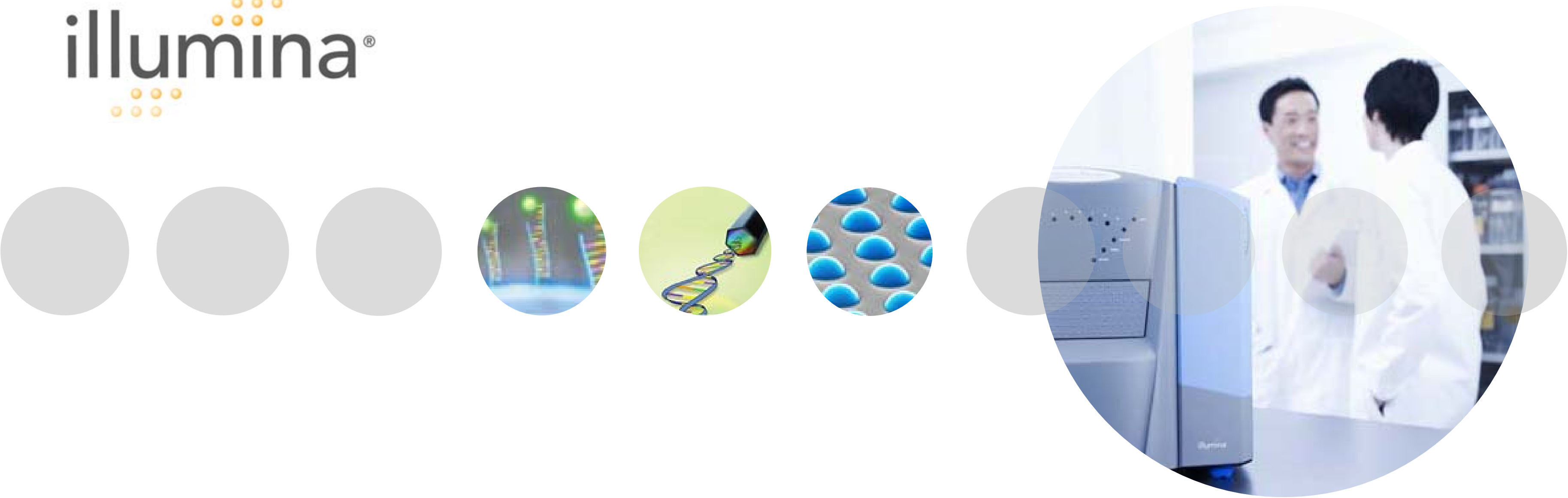


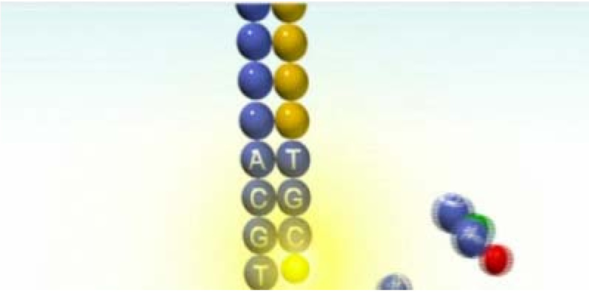
Innovations in Genomic Analysis:

Downstream analysis of Illumina Sequencing Data

Marco Cappelletti
Product Marketing Manager
Europe

Agenda

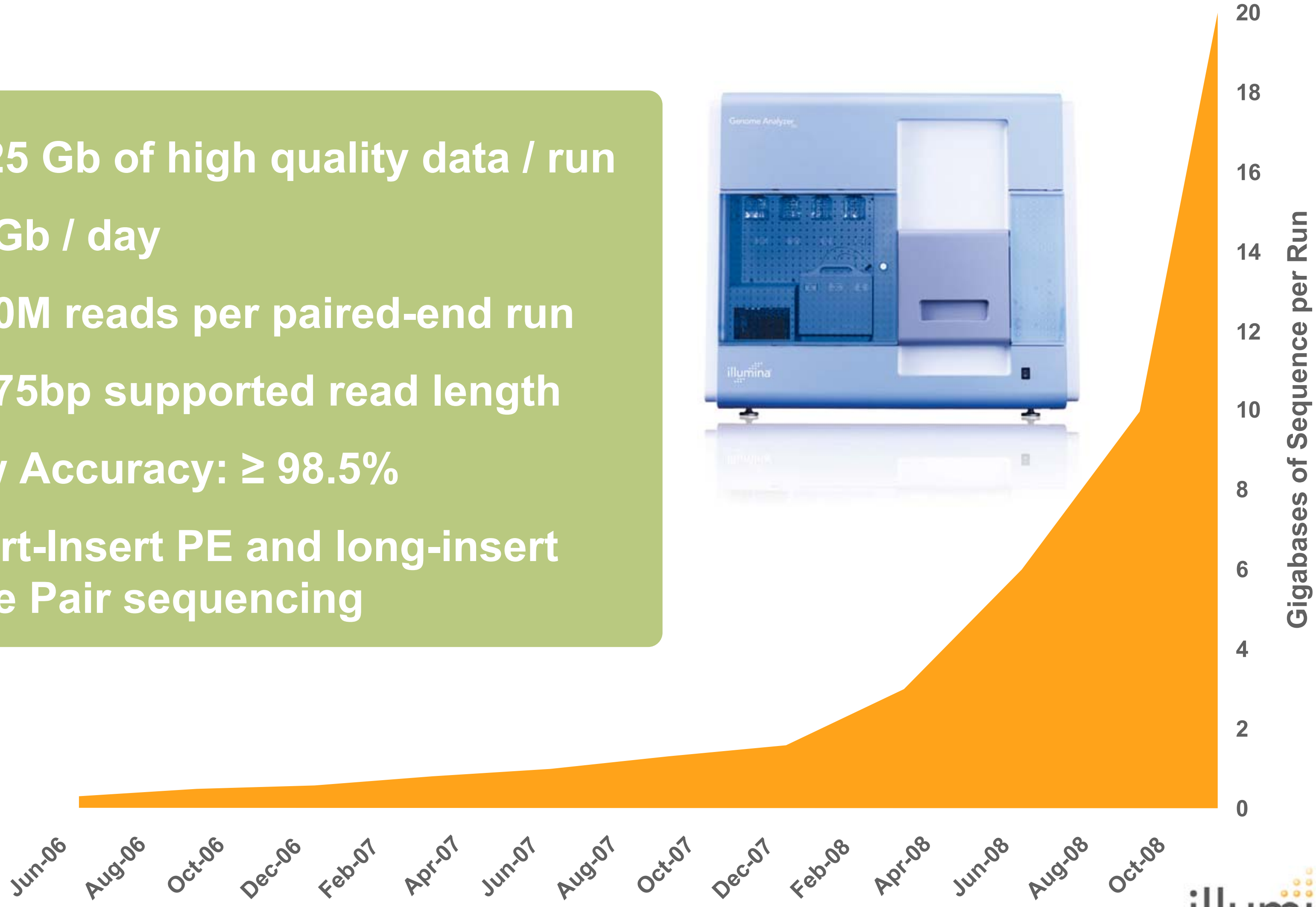
- Genome Analyzer IIx Sequencing Technology
- Applications overview
- GA*II*x SW Improvements
- Downstream Data Analysis



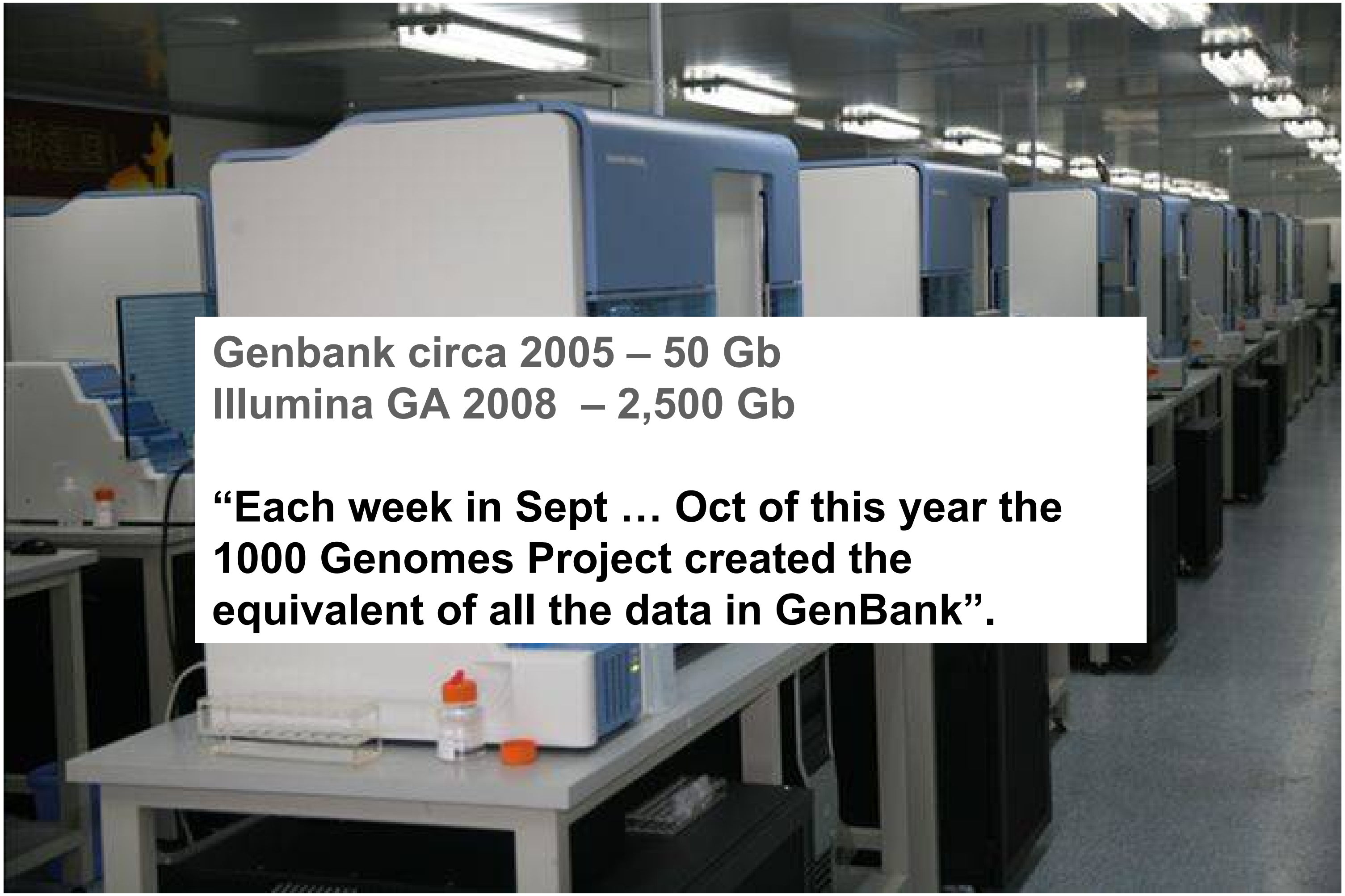
The Genome Analyzer$_{\text{IIx}}$ and Software Advancements

65% Increase in data output

- 20-25 Gb of high quality data / run
- 2.5 Gb / day
- >300M reads per paired-end run
- 2 x 75bp supported read length
- Raw Accuracy: ≥ 98.5%
- Short-Insert PE and long-insert Mate Pair sequencing



Illumina Genome Analyzer: A paradigm shift



Genbank circa 2005 – 50 Gb
Illumina GA 2008 – 2,500 Gb

“Each week in Sept … Oct of this year the 1000 Genomes Project created the equivalent of all the data in GenBank”.

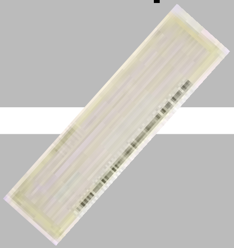
Simplest Sequencing Process - 5 to 10 days WF

1 *Library prep (~ 6 hrs)*

Fragment DNA
↓
Repair ends / Add A overhang
↓
Ligate adapters
↓
Select ligated DNA

2 *Automated Cluster Generation (~ 5 hrs)*

Up to 96 samples



Hybridize to flow cell
↓
Extend hybridized oligos
↓
Perform bridge amplification

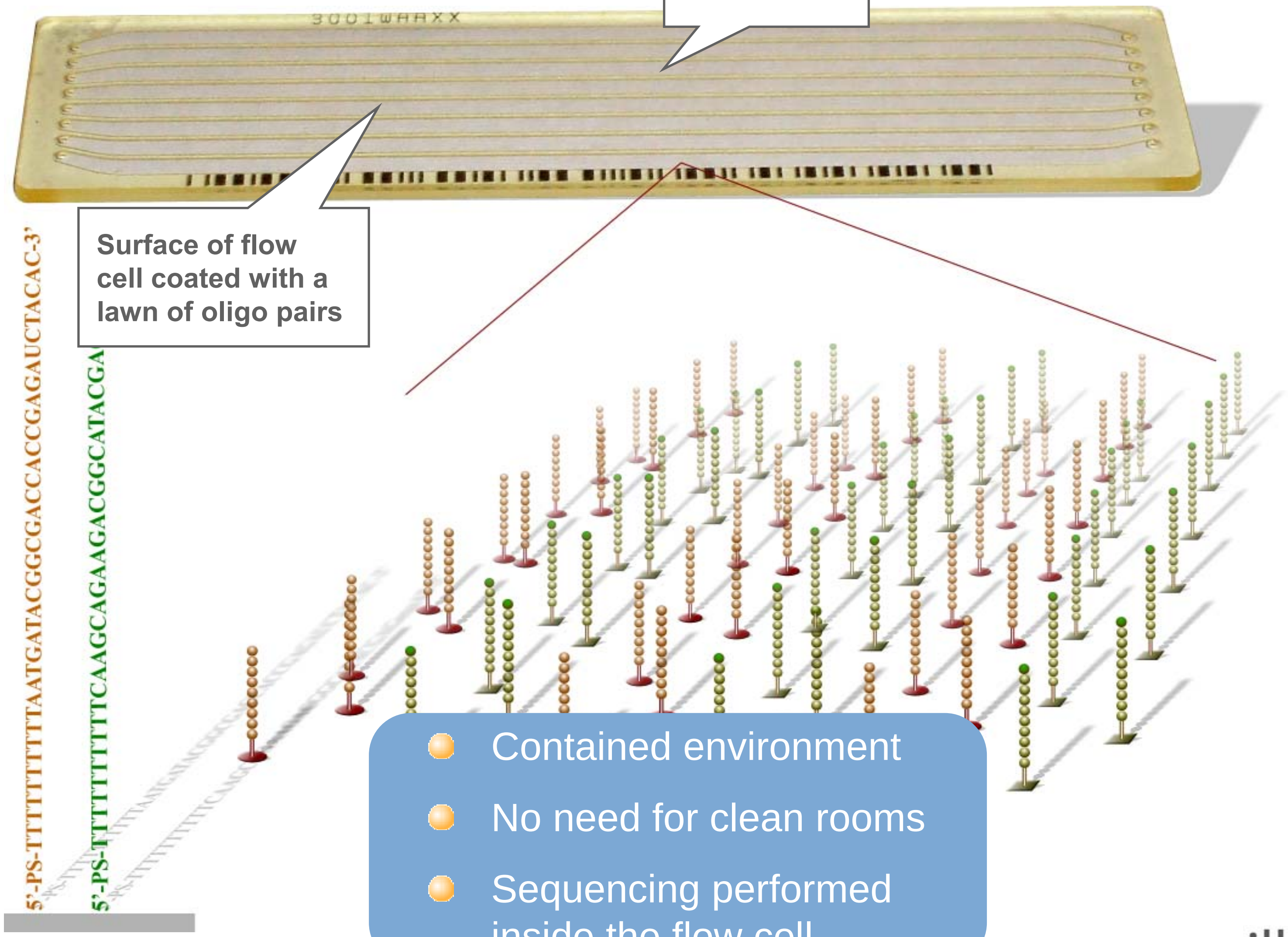
3 *Sequencing (~ 4-9 days*)*

Up to 96 samples

Perform sequencing on forward strand
↓
Re-generate reverse strand
↓
Perform sequencing on reverse strand

illumina

Flow cell

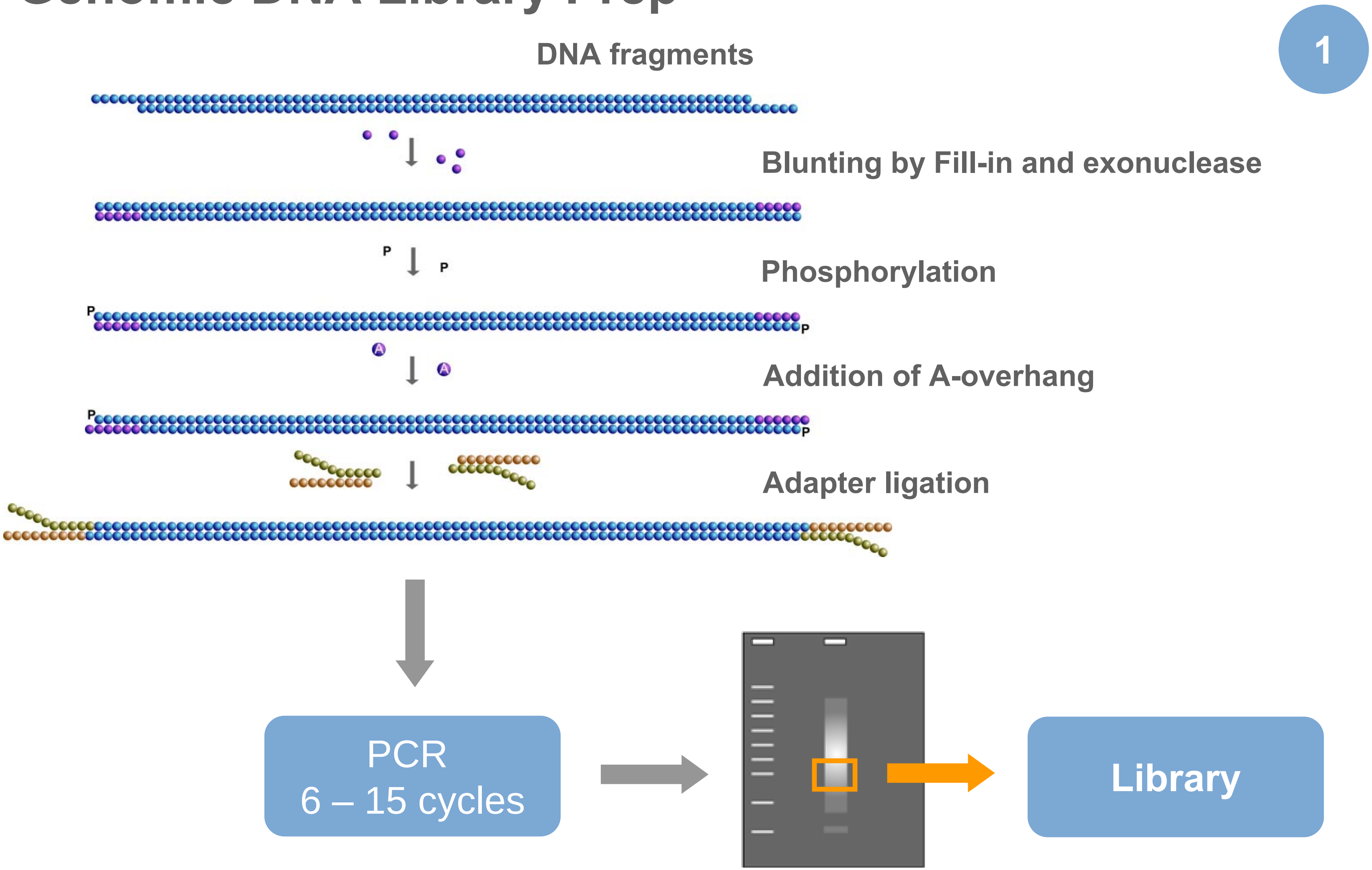

3001WAAXX
5'-PS-TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC-3'
5'-PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGA

8 channels

Surface of flow cell coated with a lawn of oligo pairs

- Contained environment
- No need for clean rooms
- Sequencing performed inside the flow cell

Genomic DNA Library Prep



1

DNA fragments

Blunting by Fill-in and exonuclease

Phosphorylation

Addition of A-overhang

Adapter ligation

PCR
6 – 15 cycles

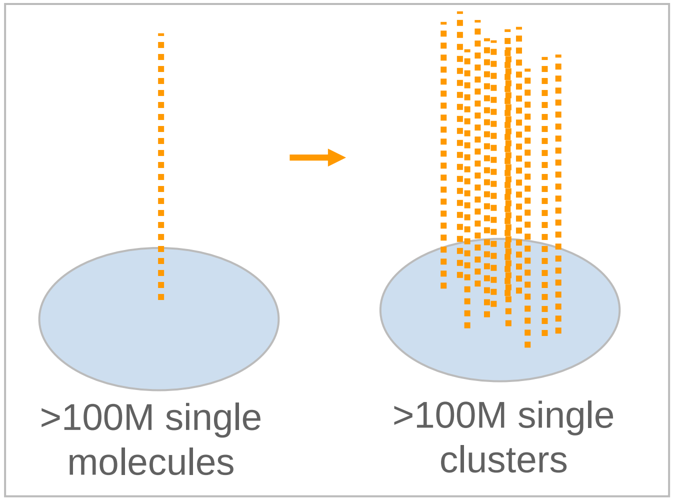
Library

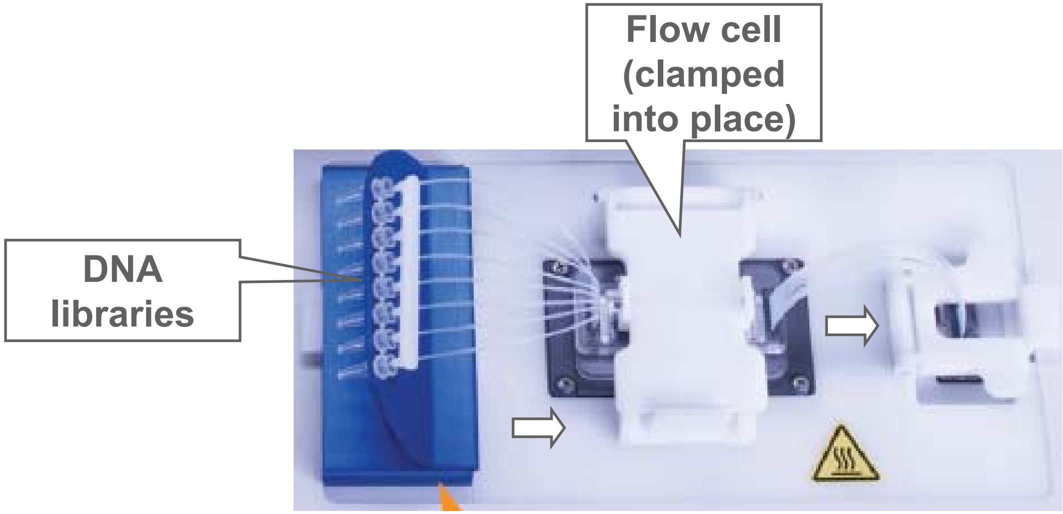
Cluster Generation

Cluster station

2

- Aspirates DNA samples into flow cell
- Automated amplified clonal clusters

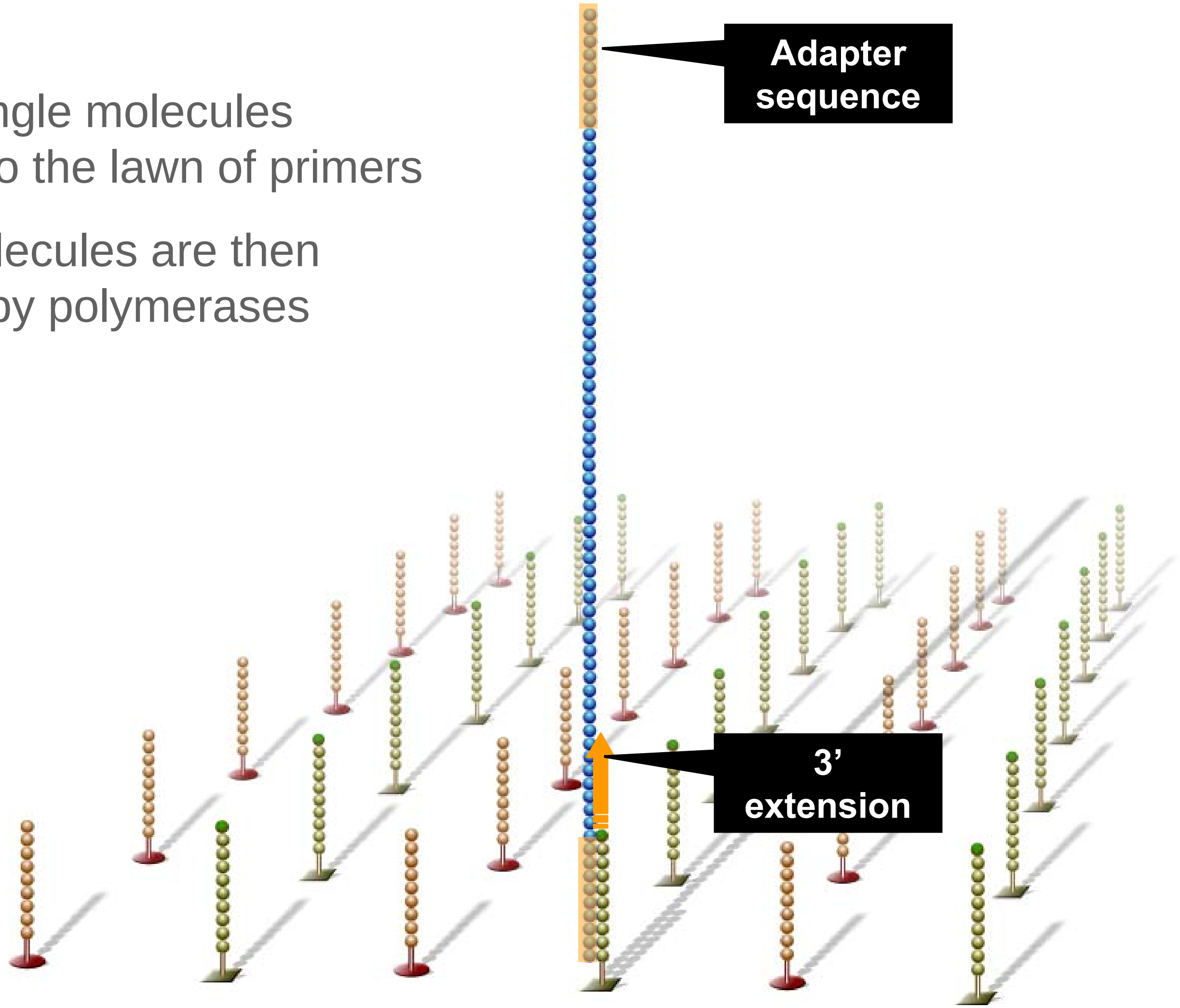




Cluster Generation

Hybridize Fragment & Extend

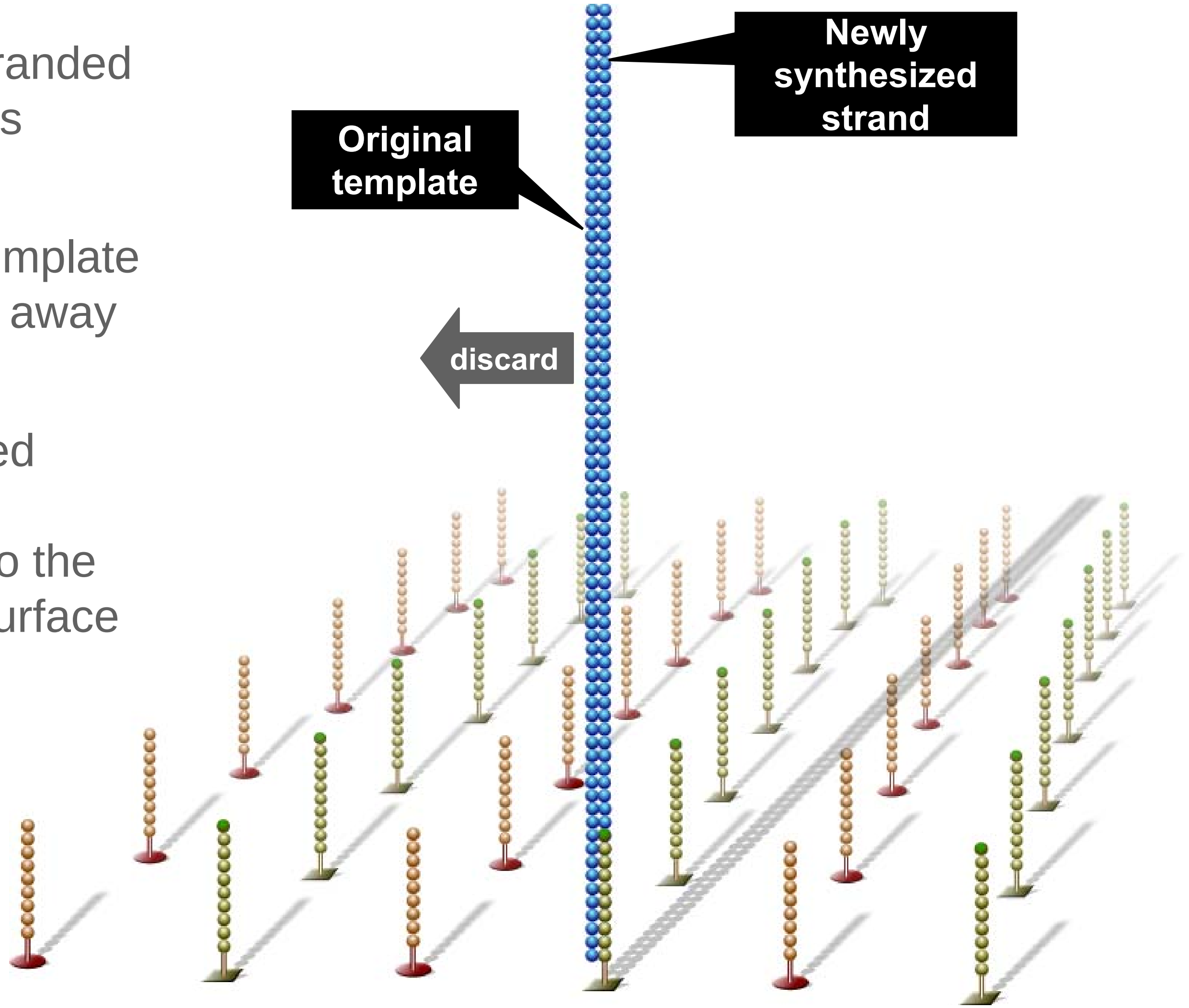
- >150 M single molecules hybridize to the lawn of primers
- Bound molecules are then extended by polymerases



Cluster Generation

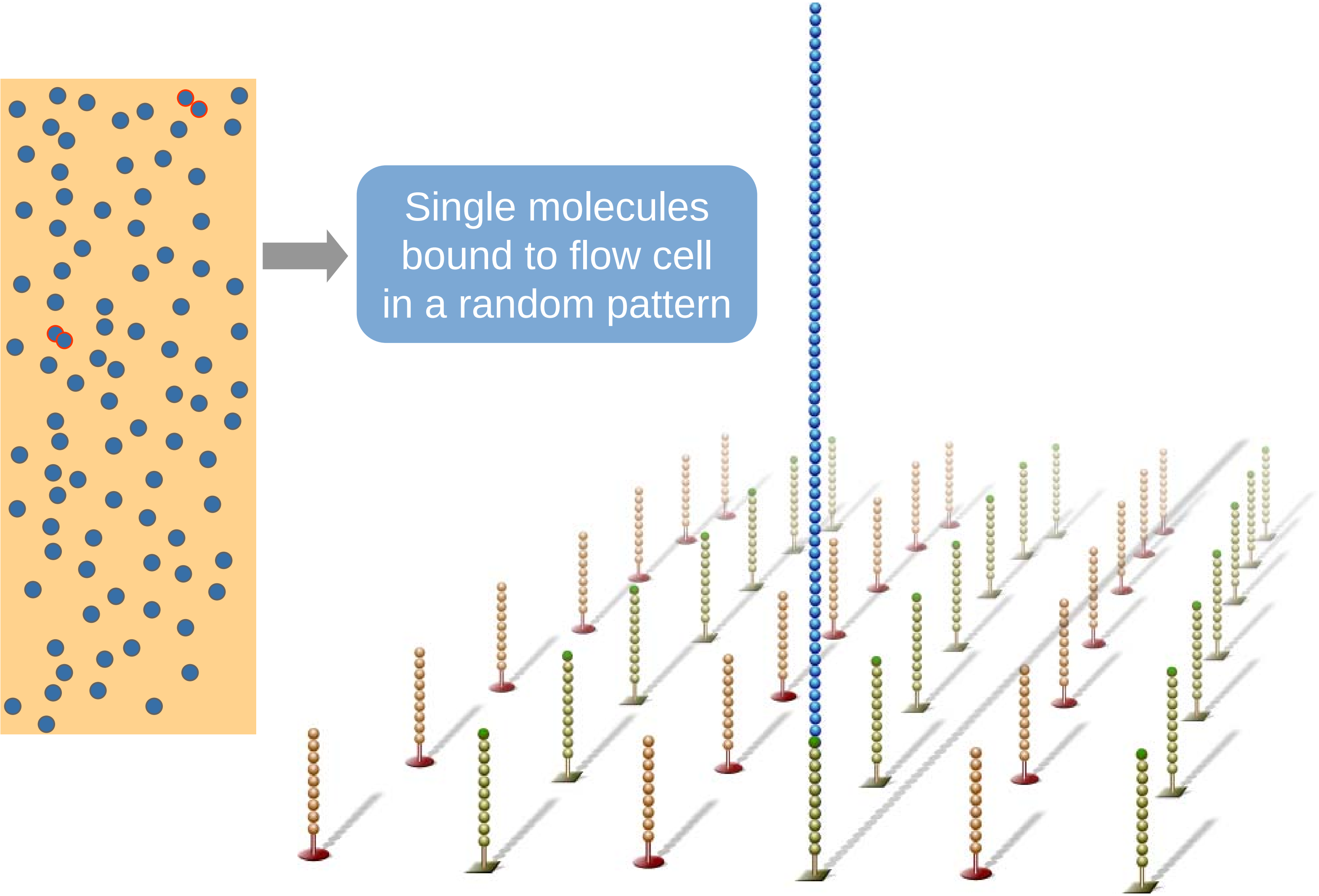
Denature Double-stranded DNA

- Double-stranded molecule is denatured
- Original template is washed away
- Newly synthesized covalently attached to the flow cell surface



Cluster Generation

Covalently-Bound Spatially Separated Single Molecules

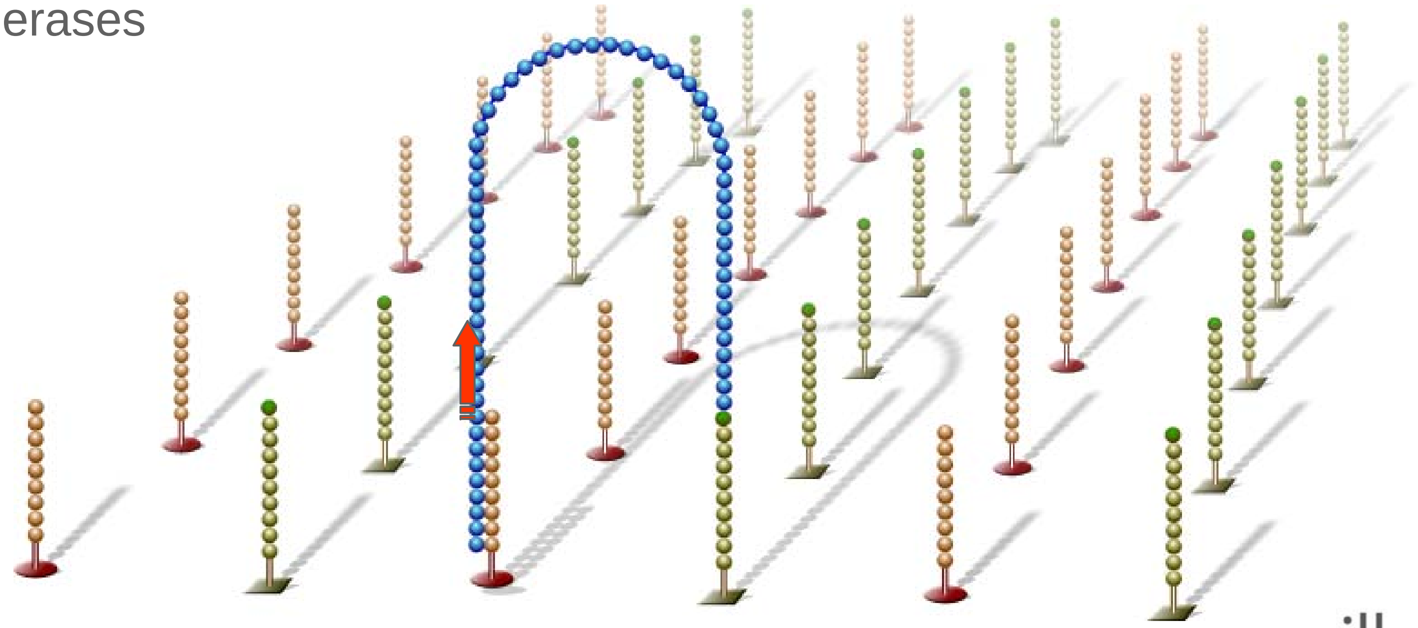


Single molecules bound to flow cell in a random pattern

Cluster Generation

Bridge Amplification

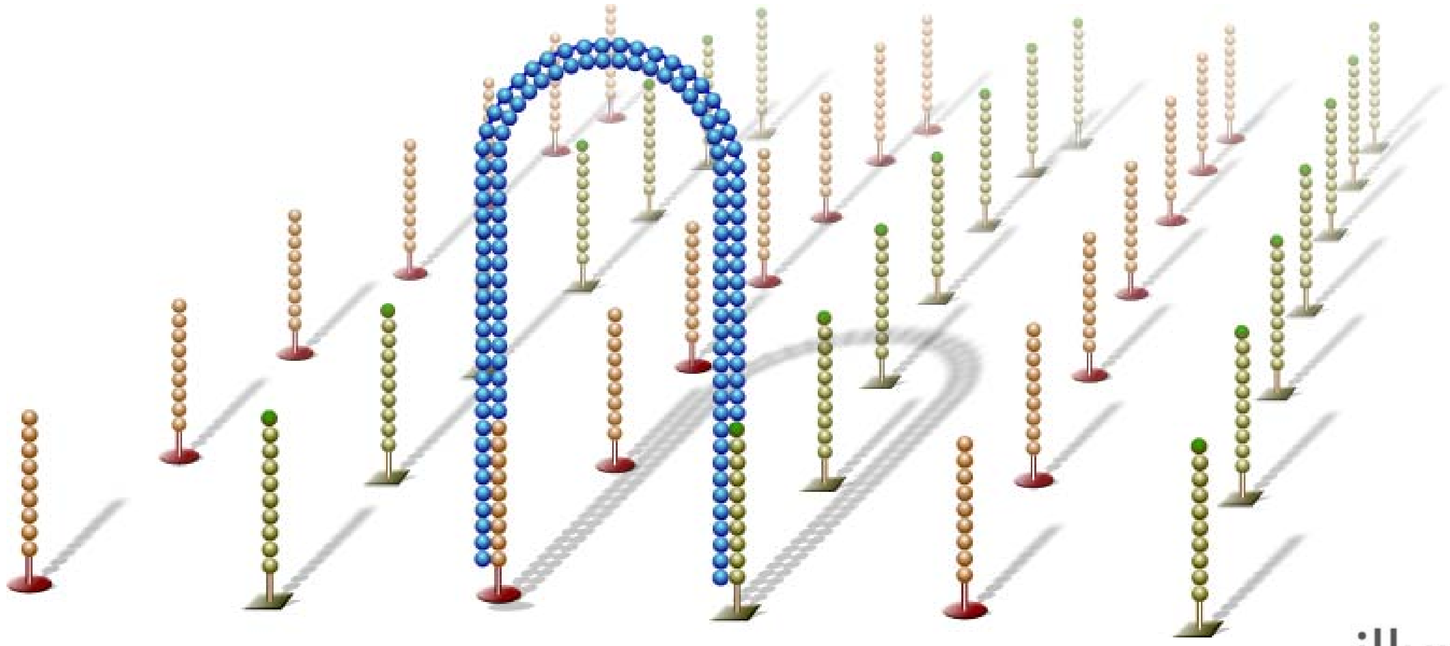
- Single-strand flips over to hybridize to adjacent primers to form a bridge
- Hybridized primer is extended by polymerases



Cluster Generation

Bridge Amplification

- Double-stranded bridge is formed

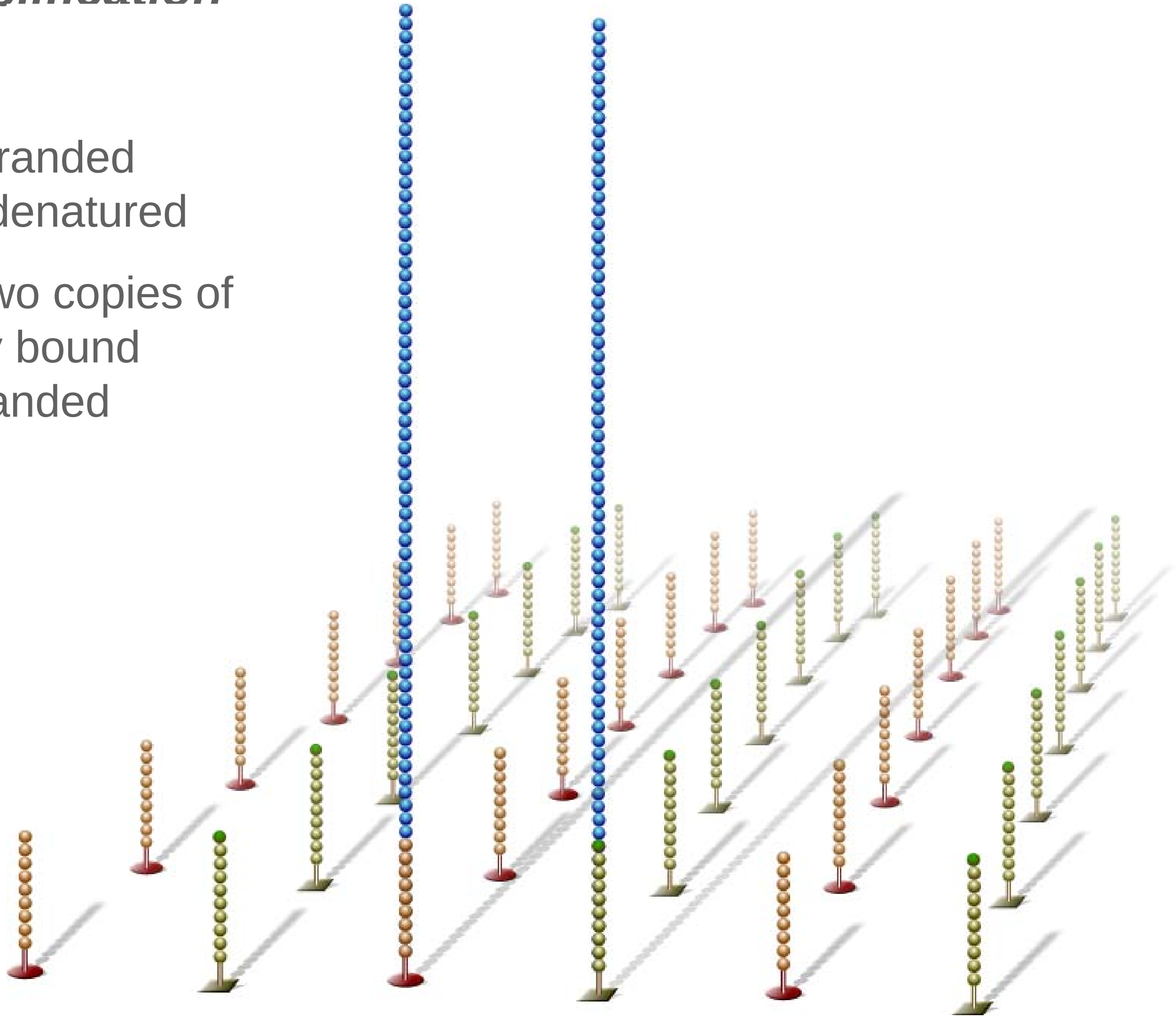


illumina

Cluster Generation

Bridge Amplification

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

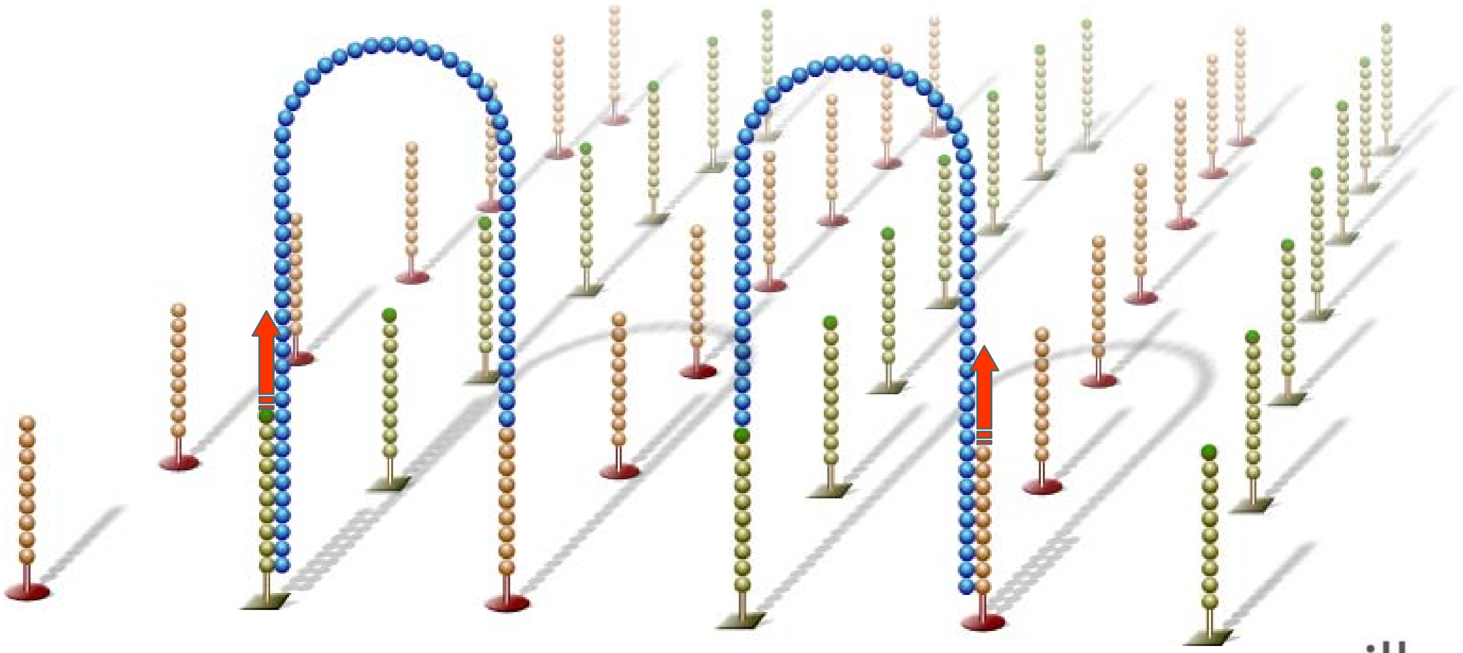


illumina

Cluster Generation

Bridge Amplification

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

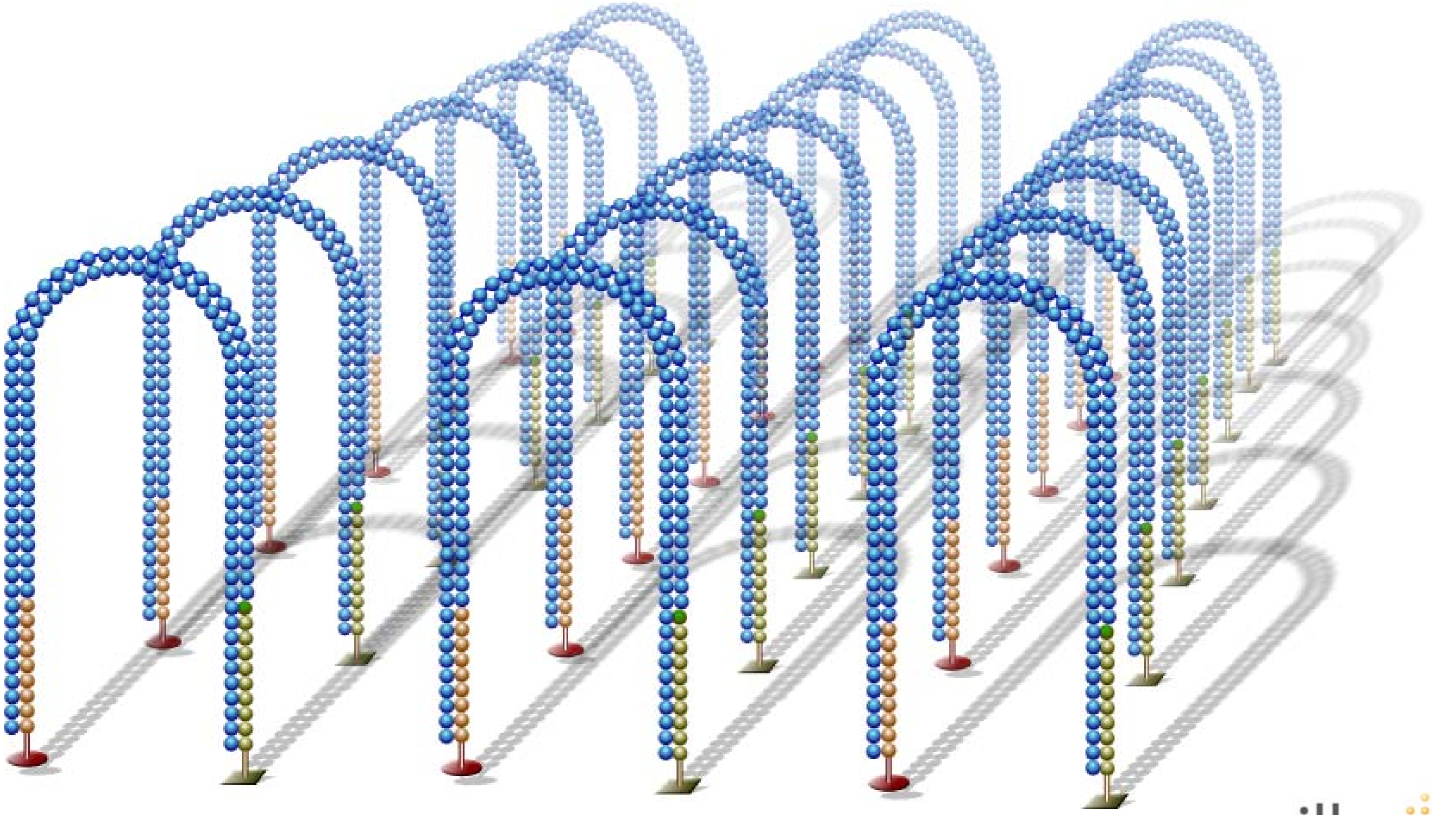


illumina

Cluster Generation

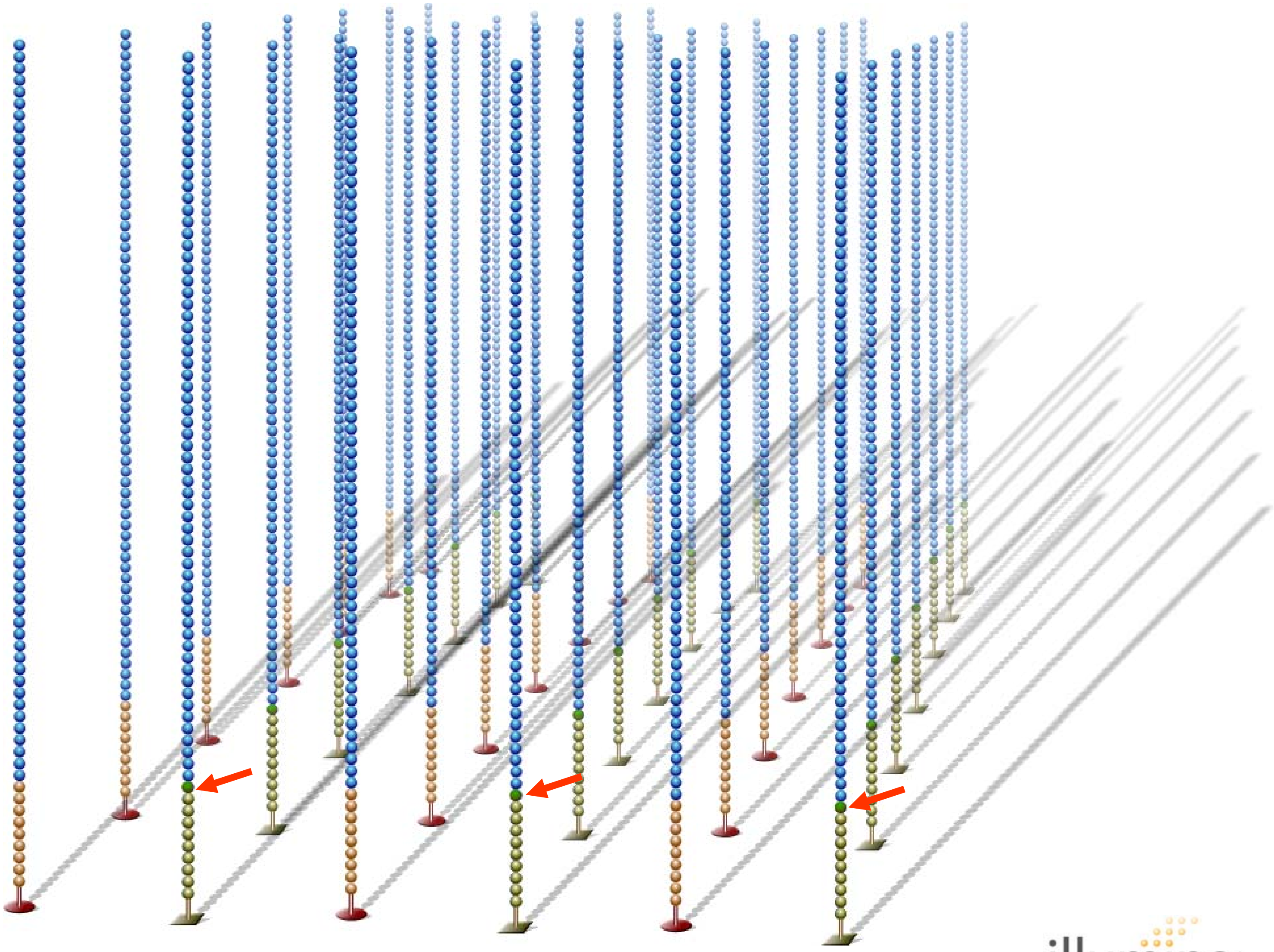
Bridge Amplification

- Bridge amplification cycle repeated until multiple bridges are formed



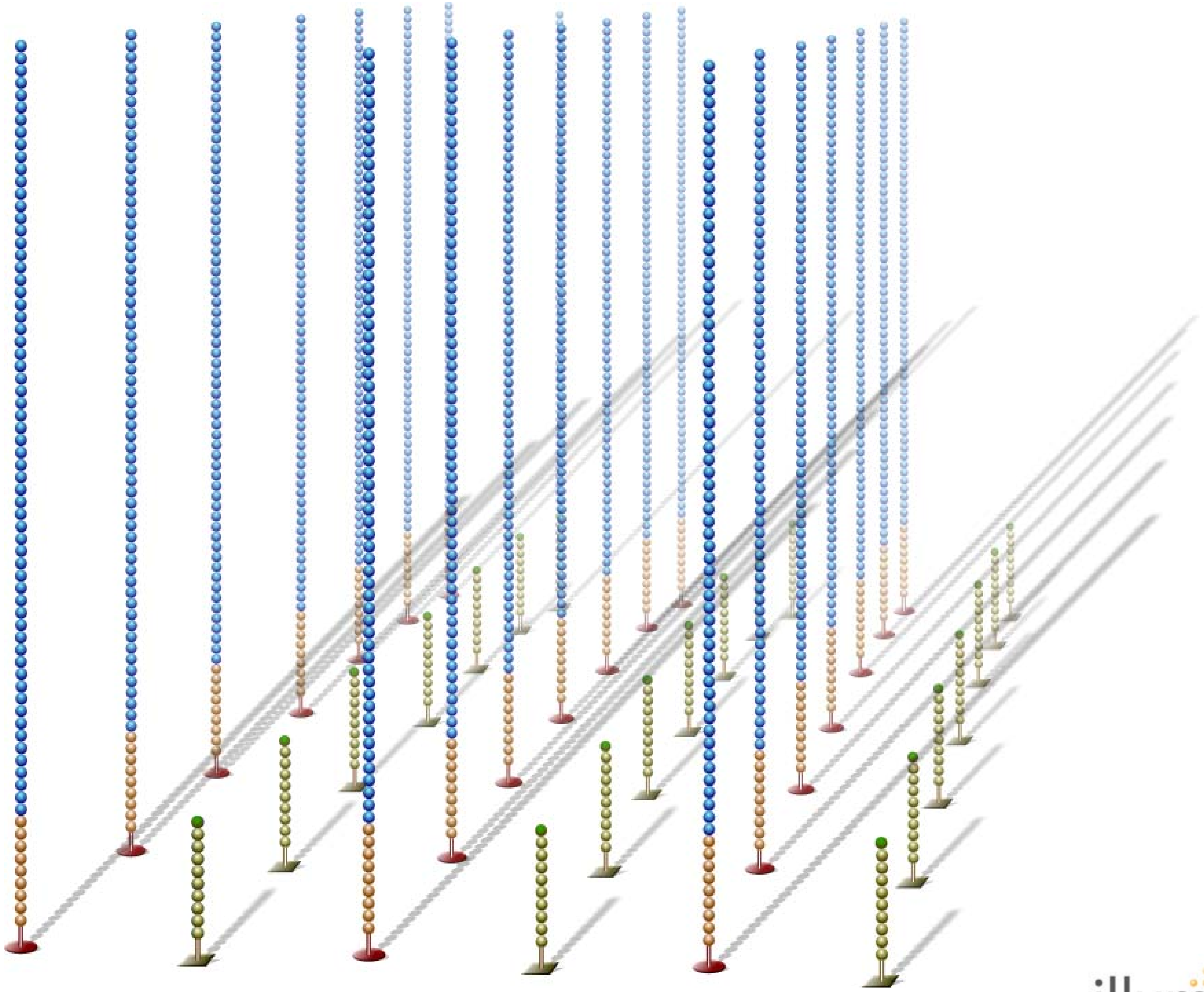
Cluster Generation

- dsDNA bridges denatured
- Reverse strands cleaved and washed away



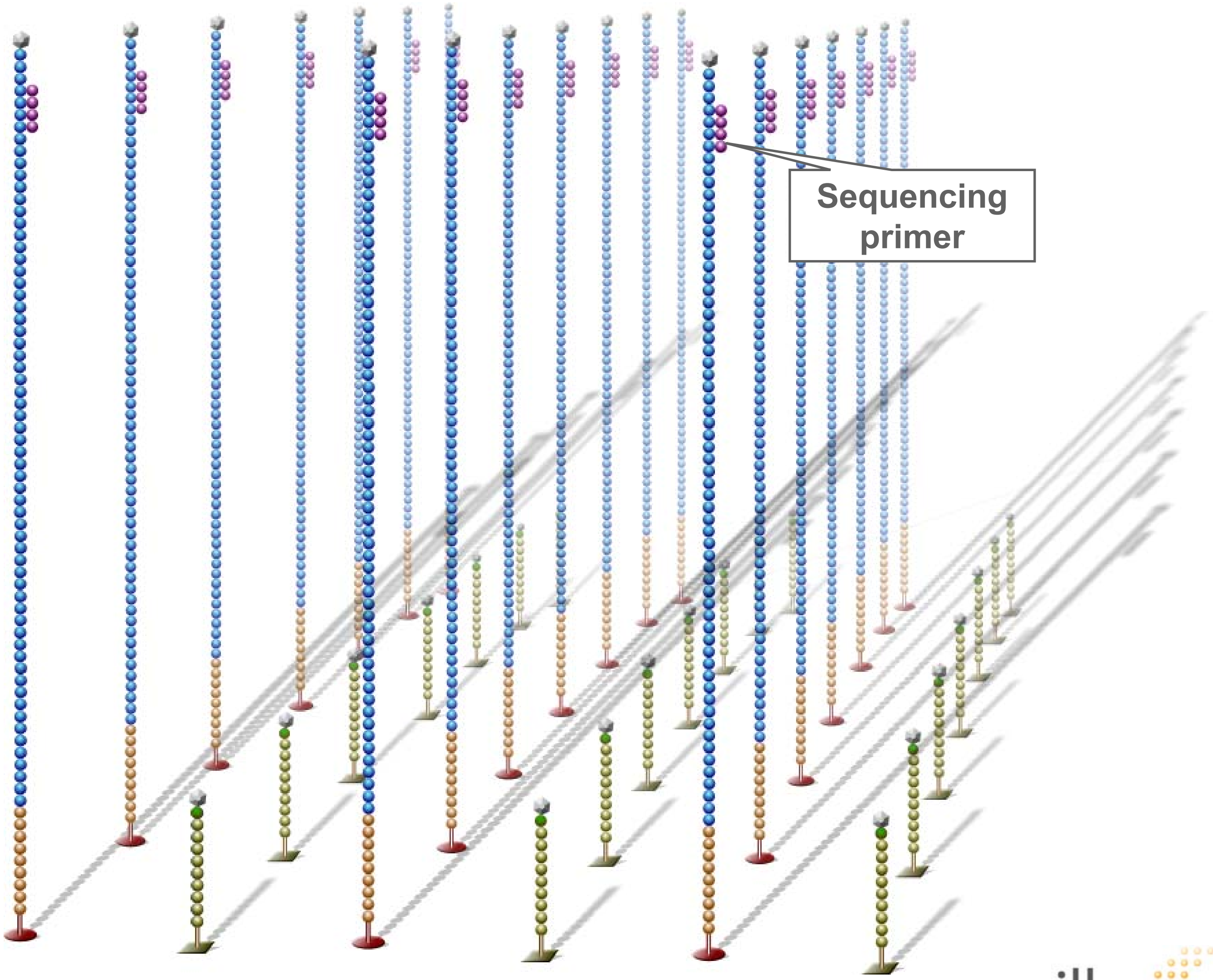
Cluster Generation

- Leaving a cluster with forward strands only



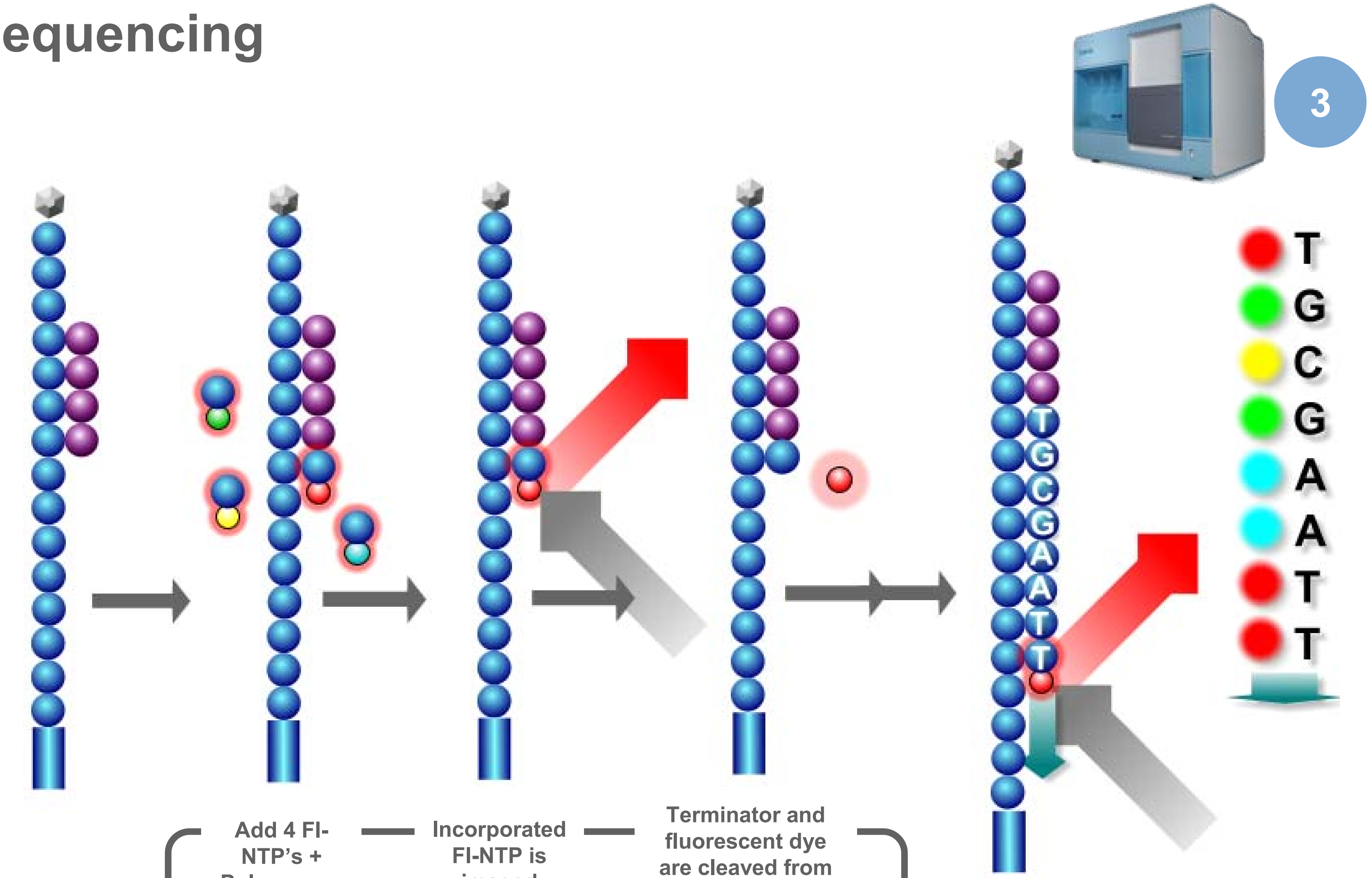
Genome Analyzer Sequencing reaction

- Sequencing primer is hybridized to adapter sequence



Sequencing primer

Sequencing



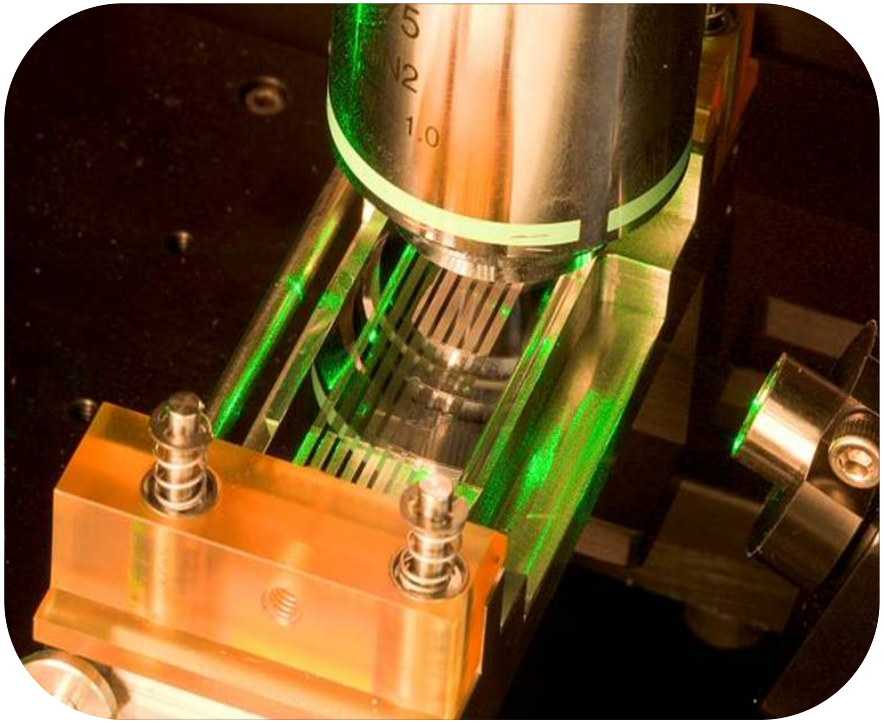
Add 4 Fl-NTP's + Polymerase

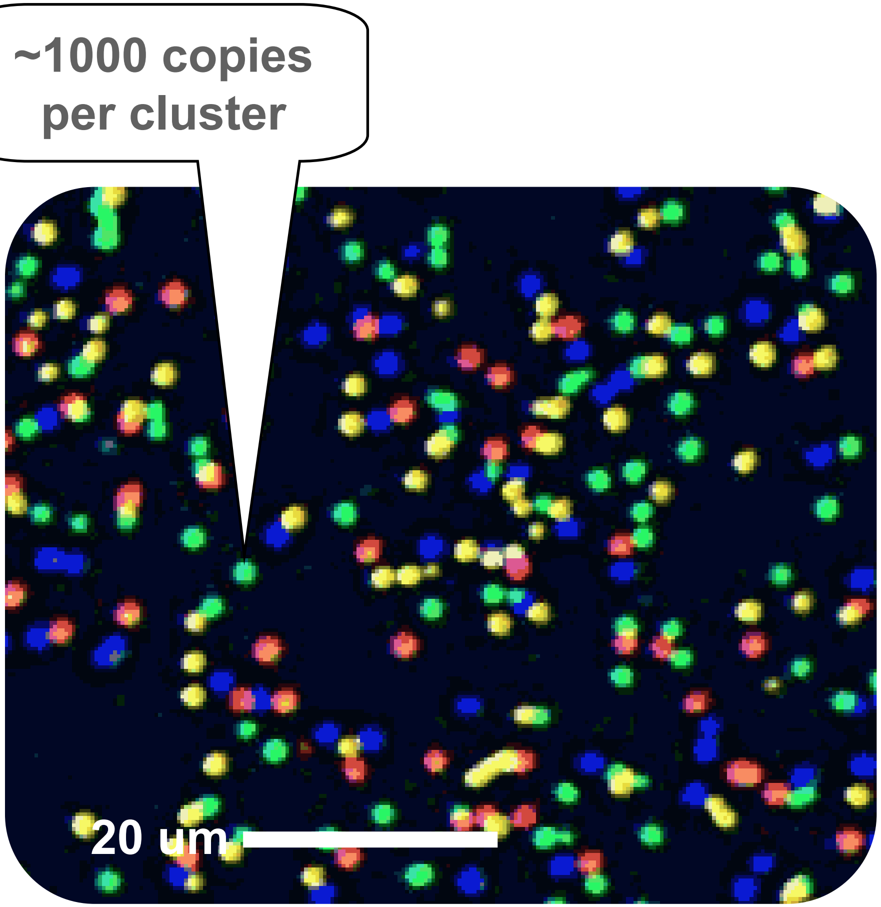
Incorporated Fl-NTP is imaged

Terminator and fluorescent dye are cleaved from the Fl-NTP

X 36 - 75

Sequencing





>150 Million Clusters Per Flow Cell

Broadest range of applications

Optimized, streamlined and easy-to-use reagent solutions

Sample Prep

Whole genome
- Resequencing
- De-novo
- Targeted
- Metagenomics

Transcriptome
- RNA-Seq
- DGE
- Small RNA
- miRNA

Regulation
- Methylation
- ChIP-Seq

Automated Cluster Generation

Sequencing

Epigenomics

Transcriptomics

de novo sequencing

Metagenomics

Genomics

…go where the biology takes you

Structural Variation Analysis

Epigenomics

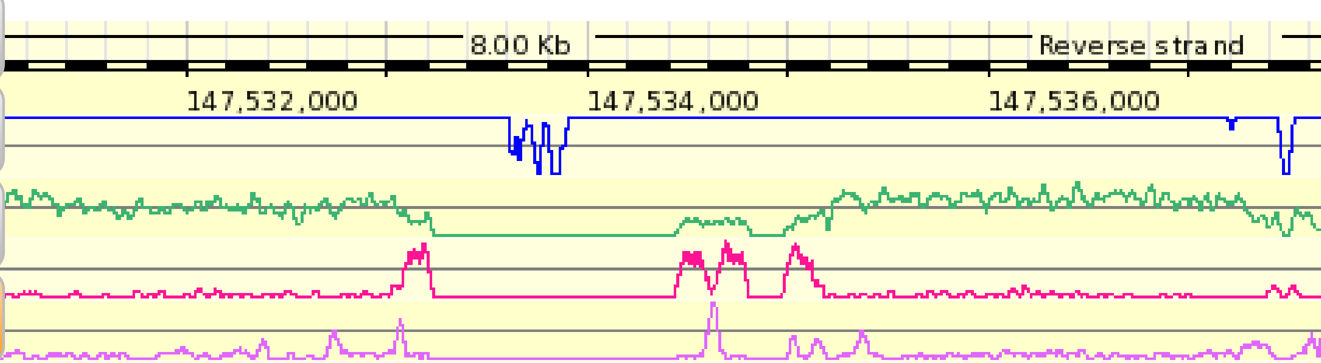
Transcriptomics

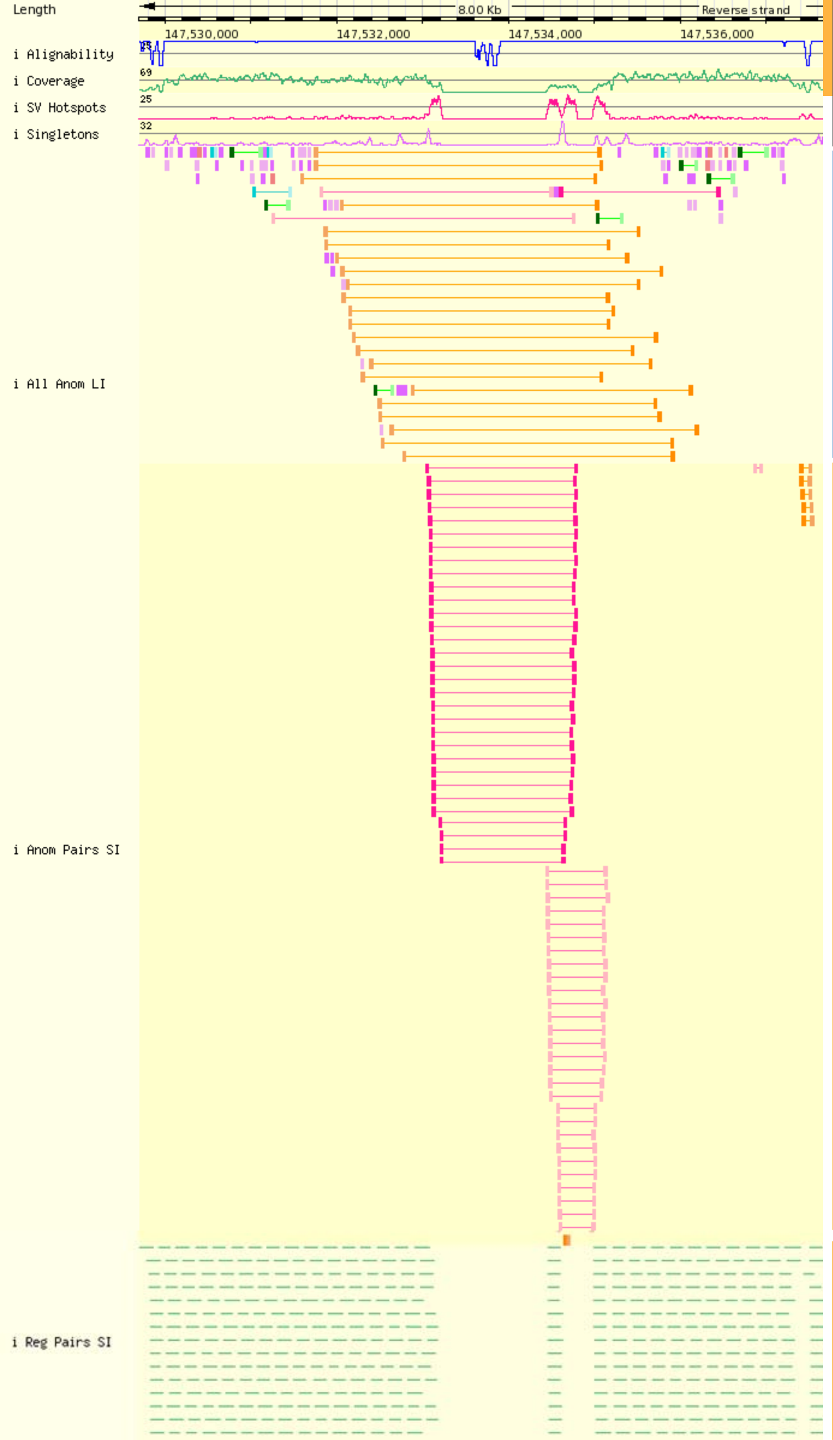
de novo sequencing

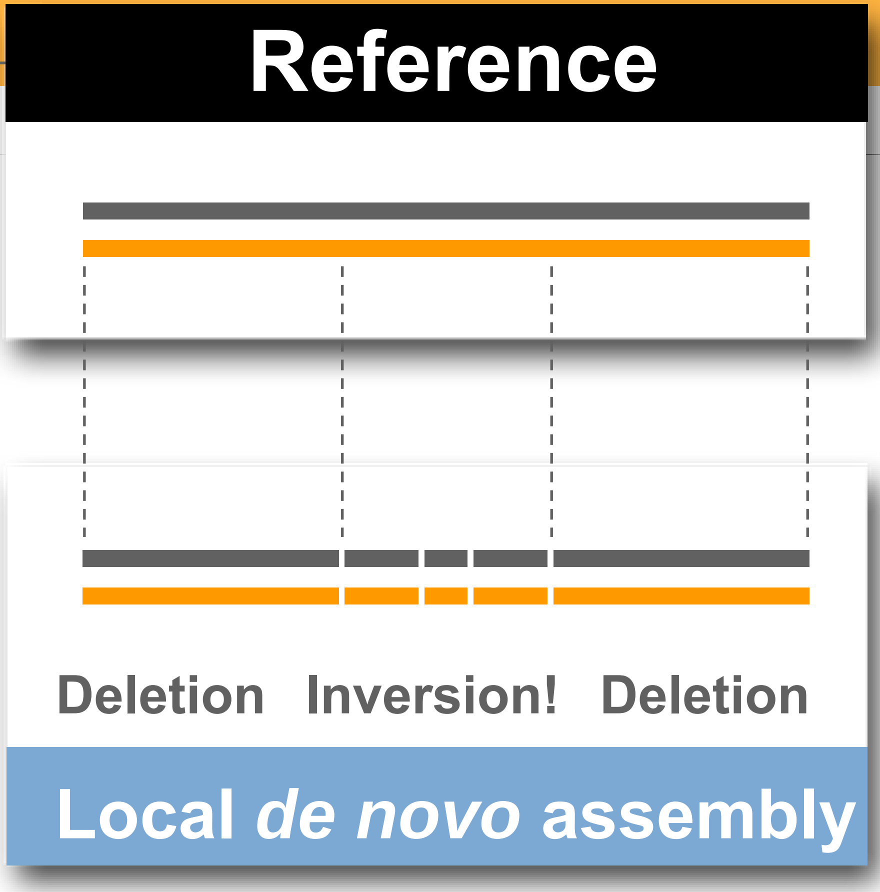
Metagenomics

Genomics

…to go where the biology takes you







Long Inserts	**Deletion?**
Short Inserts	**Inversions?**
Middle	**Normal?**

Epigenomics

Transcriptomics

de novo sequencing

Metagenomics

Genomics

Nature Publications

Author	Number of Reads	Average Length
Shi, *et al.*	389K	97 bp
Dinsdale, *et al.*	14.6M	105 bp
Mou, *et al.*	307K	96 bp
Warnecke, *et al.*	300K	100 bp
Turnbaugh, *et al.*	1.7M	93 bp

Nature Publications

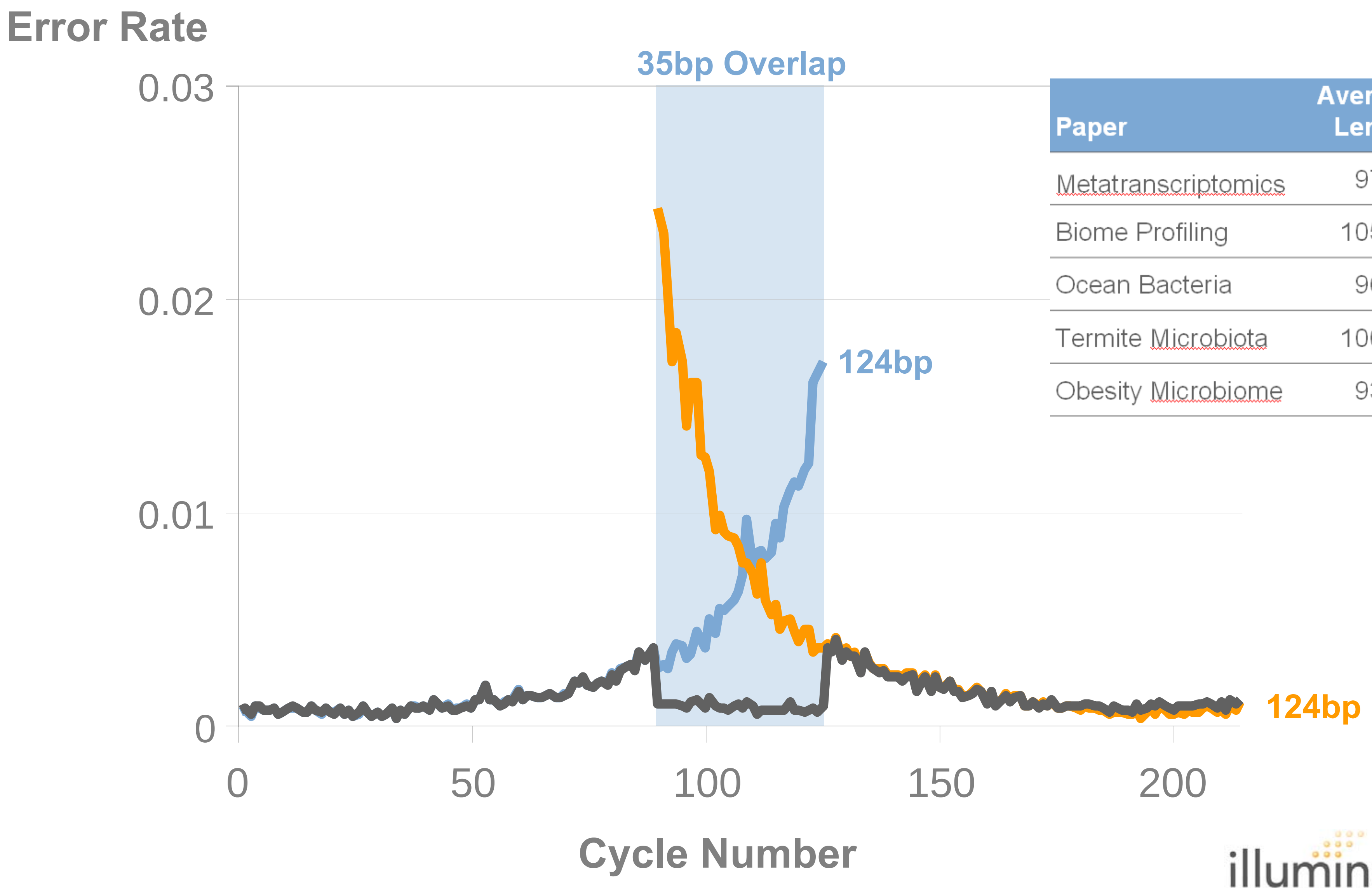
Paper	Number of Reads	Average Length
Metatranscriptomics	389K	97 bp
Biome Profiling	14.6M	105 bp
Ocean Bacteria	307K	96 bp
Termite Microbiota	300K	100 bp
Obesity Microbiome	1.7M	93 bp

Nature Publications

Reads per run: 150-300 Million	Number of Reads	% of a 150M read run
	389K	0.3%
	14.6M	9.7%
	307K	0.2%
	300K	0.2%
	1.7M	1.1%

illumina

The Value of Long, Overlapping Reads

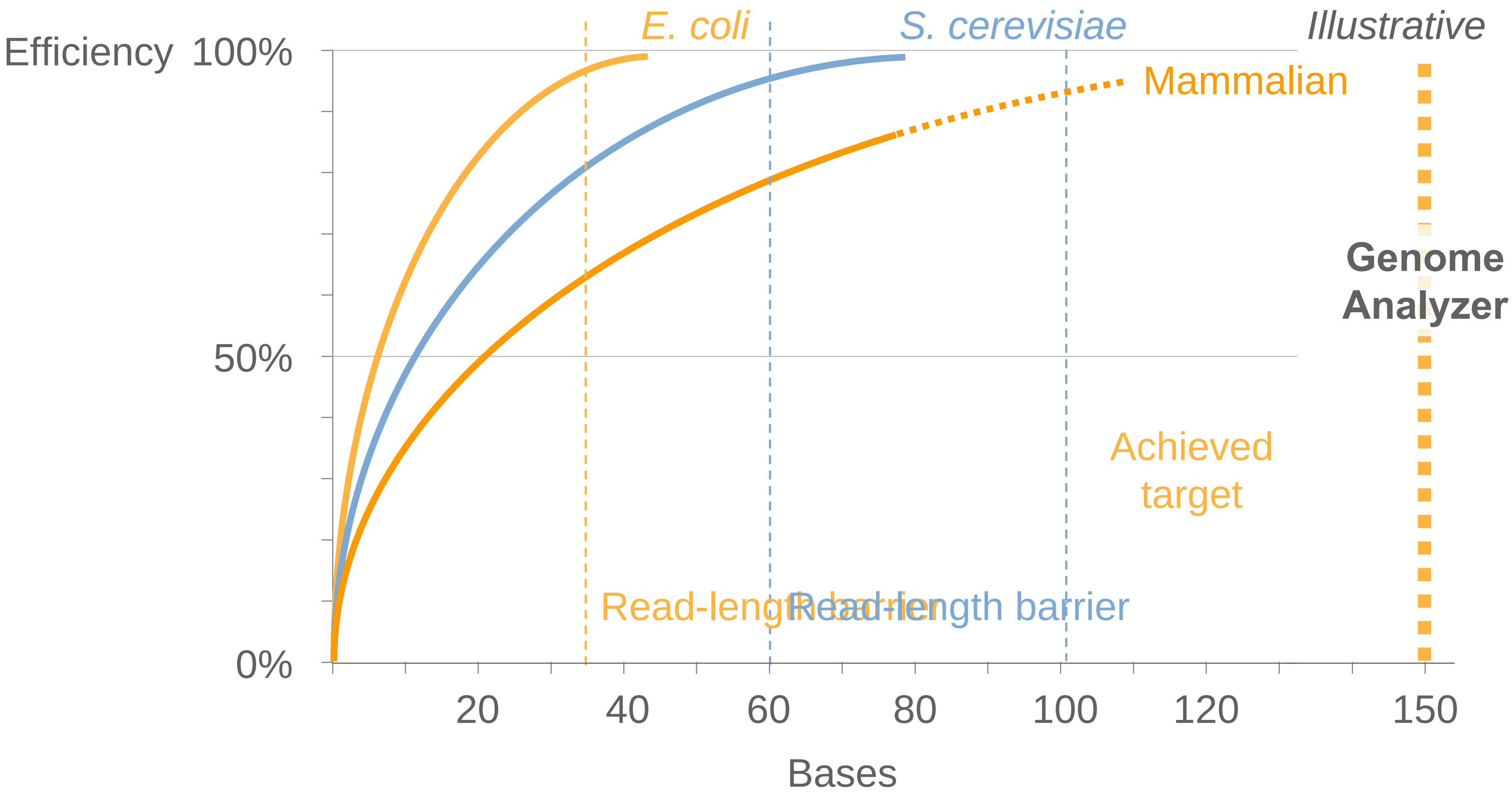


Paper	Average Length
Metatranscriptomics	97 bp
Biome Profiling	105 bp
Ocean Bacteria	96 bp
Termite Microbiota	100 bp
Obesity Microbiome	93 bp

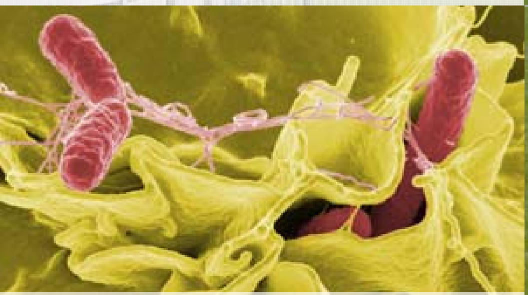
- Epigenomics
- Transcriptomics
- *de novo* sequencing
- Metagenomics
- Genomics

> When the read length exceeds a certain threshold, the ***read length barrier***, the efficiency reaches nearly 100%, so that the read length indeed does not matter.

Chaisson et al., De novo *Fragment Assembly with Short Mate-Paired Reads: Does the Read Length Matter?, Genome Research,* October 22, 2008


Efficiency
100%
50%
0%
E. coli
S. cerevisiae
Illustrative
Mammalian
Genome Analyzer
Achieved target
Read-length barrier
Read-length barrier
20
40
60
80
100
120
150
Bases

October, 2008



Salmonella seftenberg

	Illumina
N50 contig size:	139,353
Largest contig:	395,600
Average contig:	63,969
Total bases of contigs:	4.80Mb
Coverage of genome:	99.8%

Beijing Genome Institute

January 9, 2009

- 3GB genome
- Paired 75-base reads
- >95% gene regions
- N50 contig: 300Kb

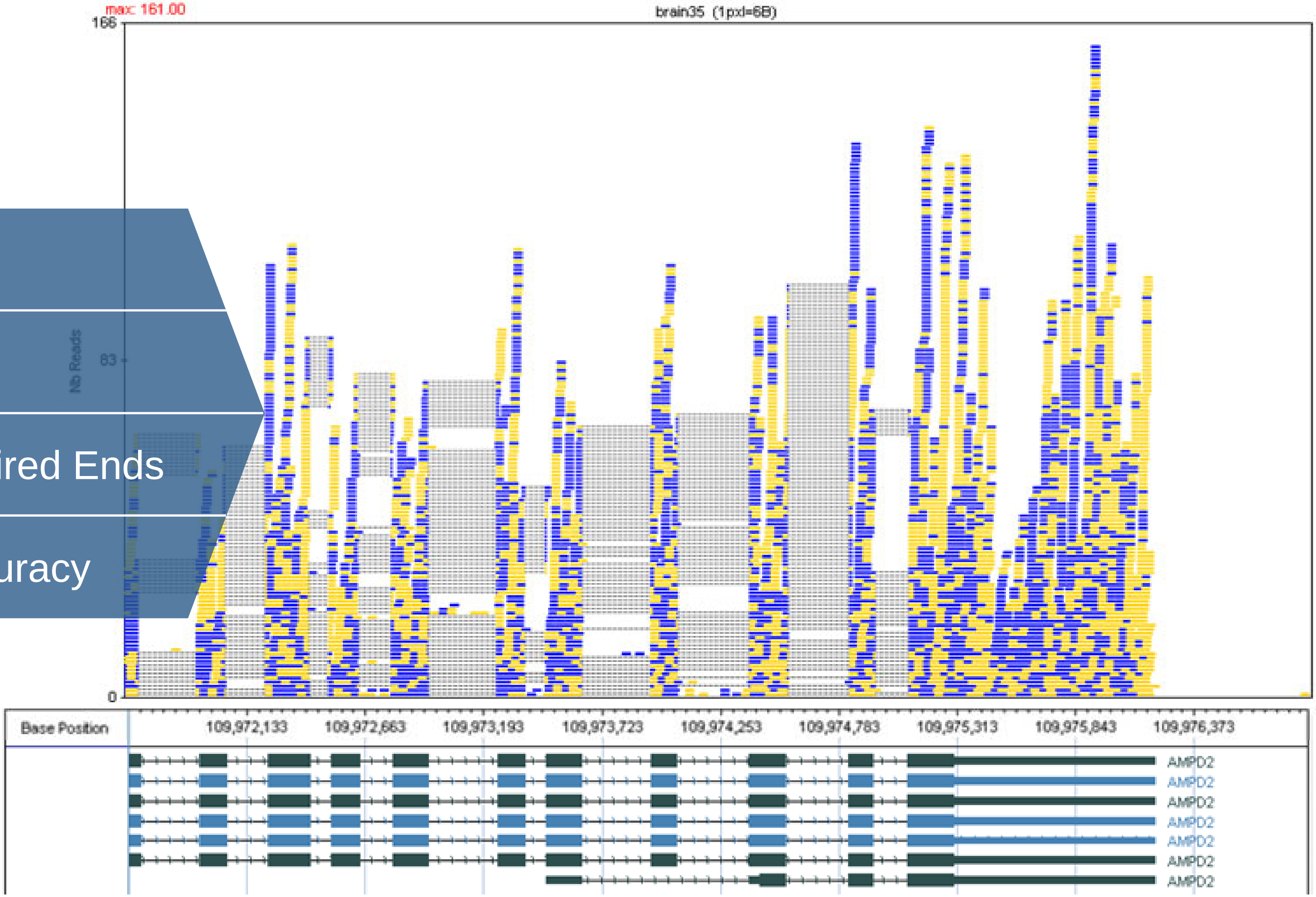
- Epigenomics
- Transcriptomics
- *de novo* sequencing
- Metagenomics
- Genomics

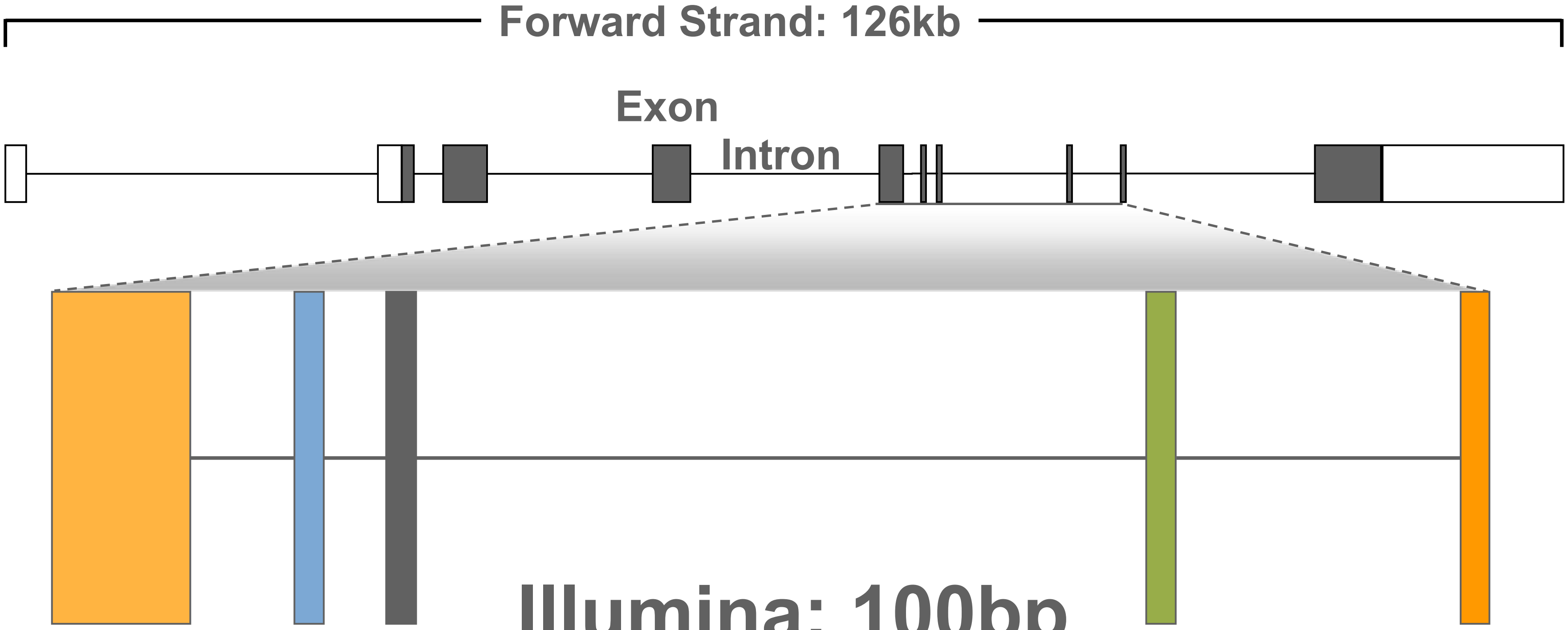
Capabilities

Length of read

Short-insert Paired Ends

Raw Read Accuracy





Forward Strand: 126kb

Exon

Intron

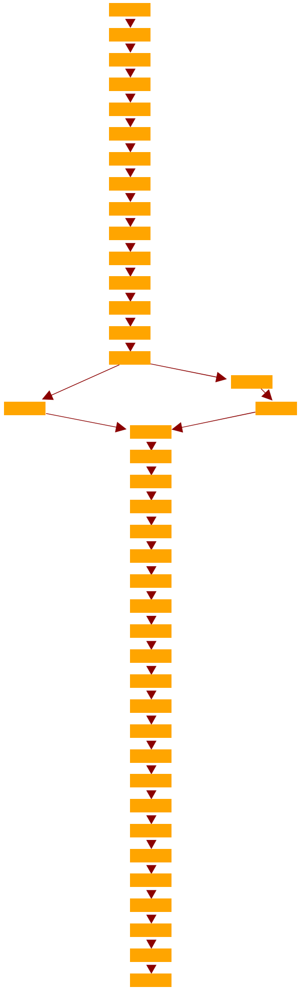
Illumina: 100bp

GAGTGAAGCTCCTGGAGGAACTCAGATGGAGAAGTATCCAGTATGCATCTCGGGGAGAGAGACATTCAGCCTATAATGAATGGAAAAAGGCCCTCTTCAA

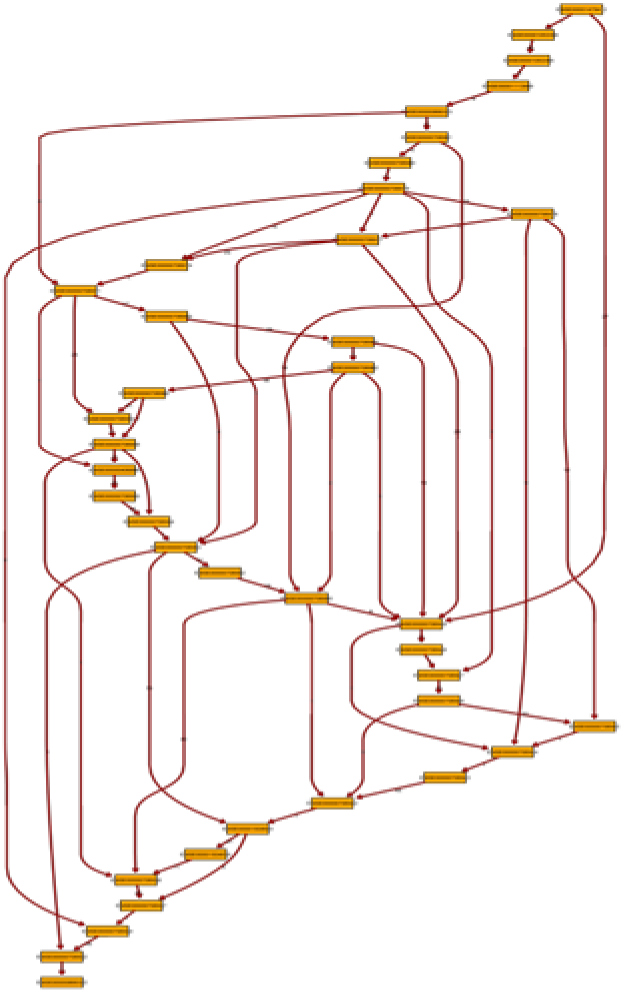
AGAACAAAGCACAAGAGTGAAGCTCCTGGAGGAACTCAGATGGAGAAGTATCCAGTATGCATCTCGGGGAGAGAGACATTCAGCCTATAATGAATGGAAAAAGGCCCTCTTCAAGCCTC

Source: Illumina R&D

Ensembl



Illumina GA



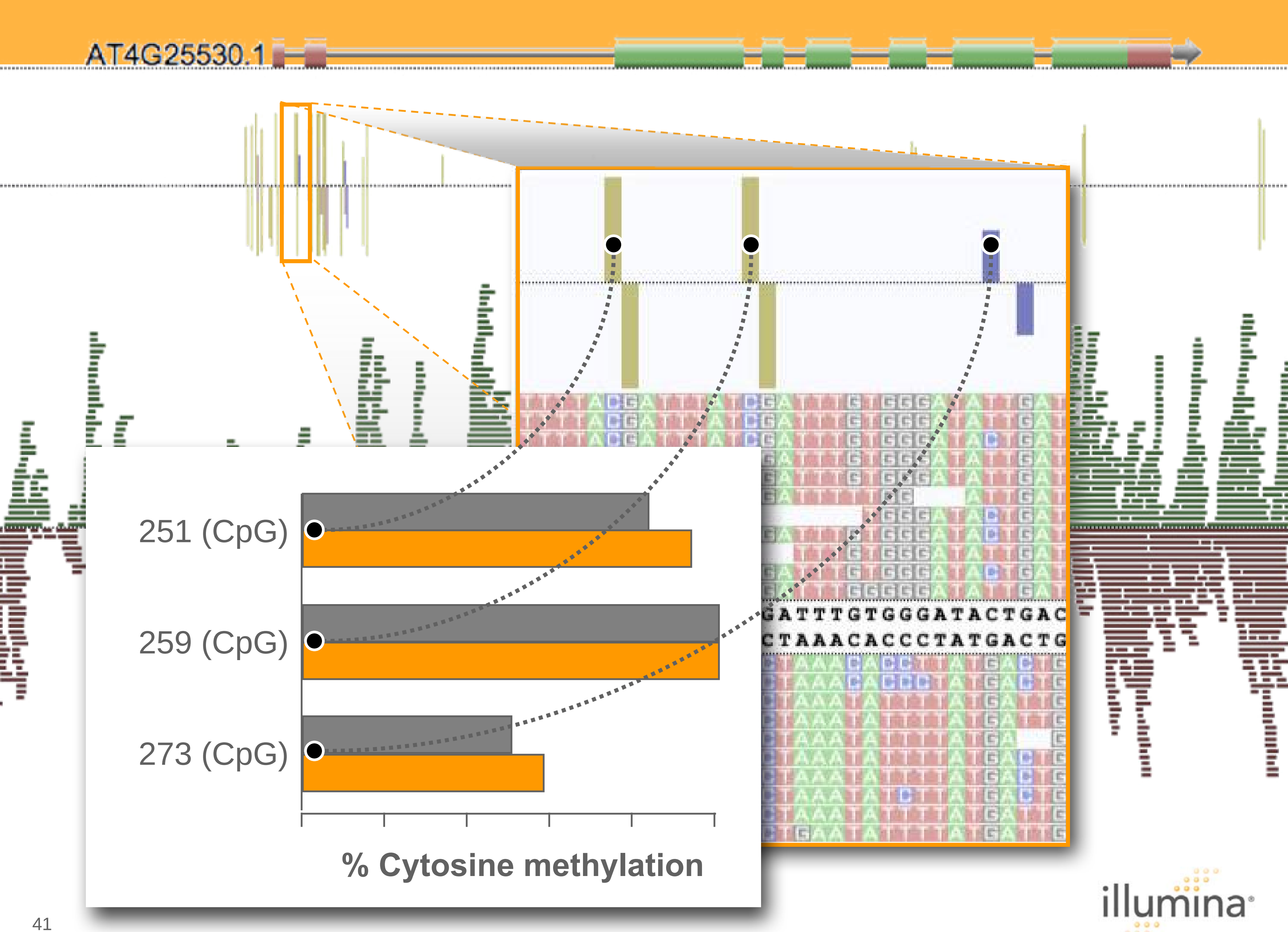
Epigenomics

Transcriptomics

***de novo* sequencing**

Metagenomics

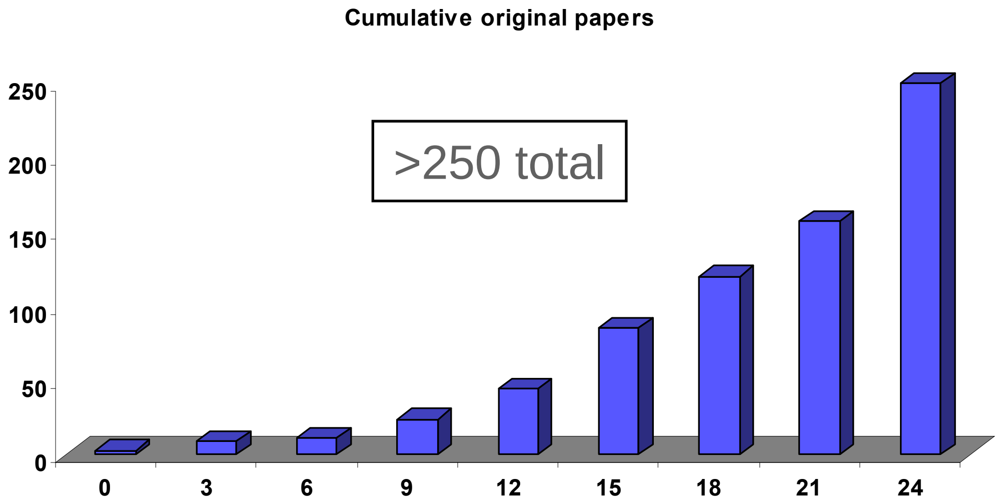
Genomics



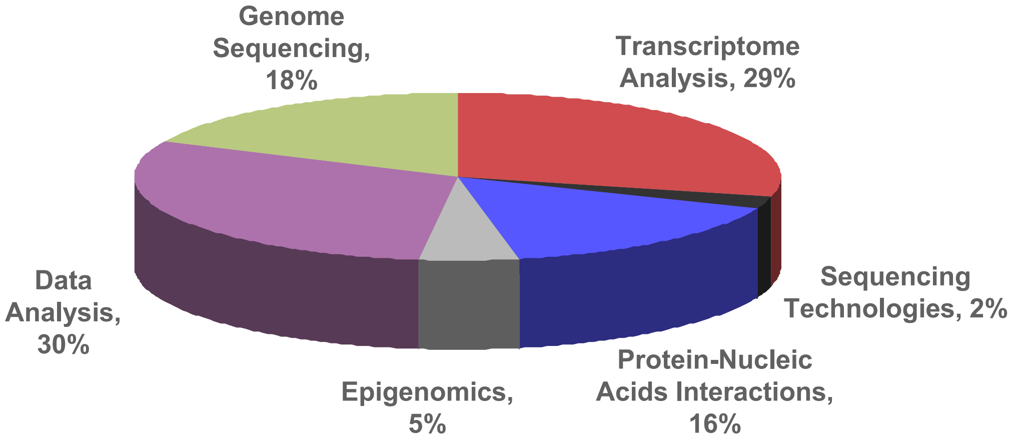
Lister, R., et al., (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 133(3), 523-536.

GA Applications Published 24 months from launch

nature	42
Science	14



>250 total



Genome Analyzer$_{IIx}$ Software Advancements

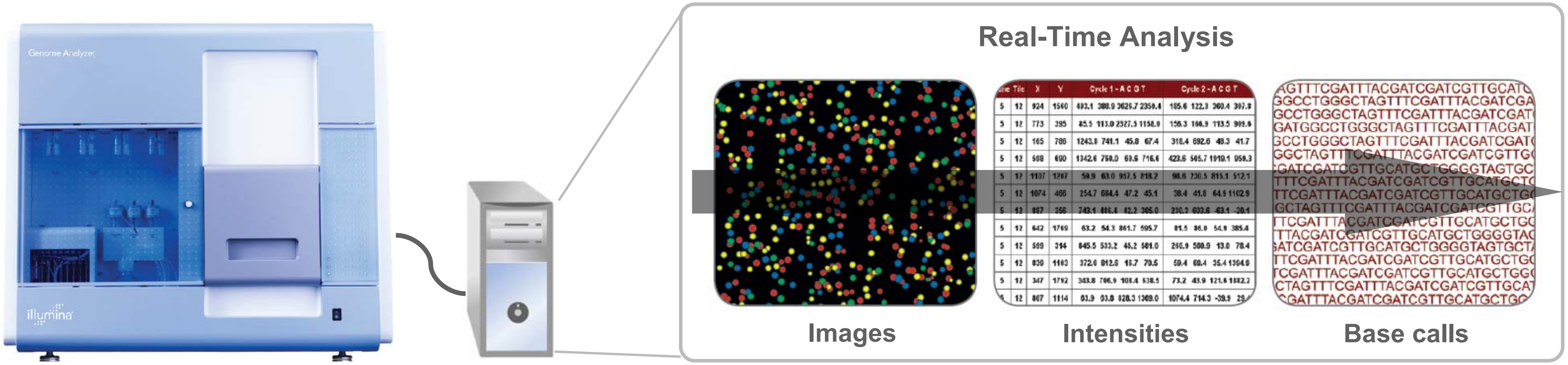
Increased Output, Simplified Computing

Increased output with reduced computing infrastructure

More gigabases of data for fewer gigabytes of computing power!

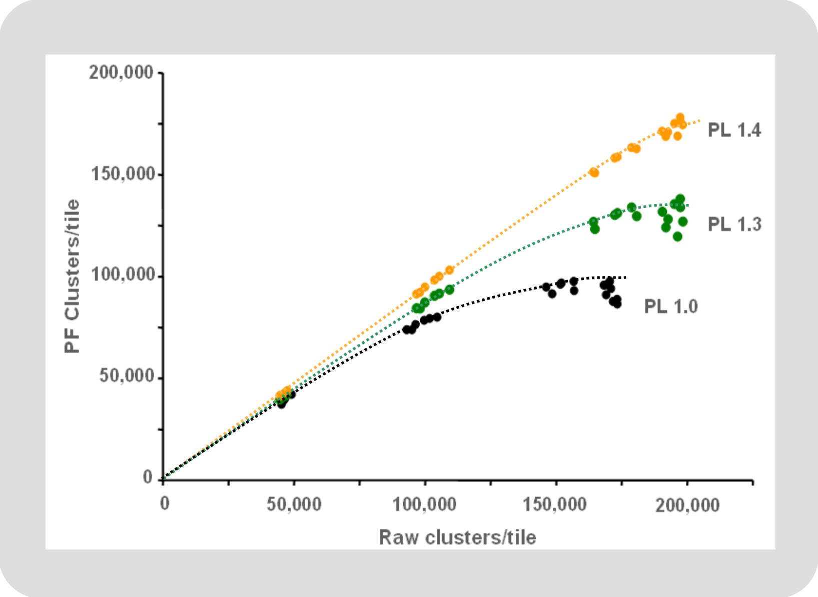
Sequencing Control Software v2.4

- Includes new Real Time Analysis (RTA) feature
- Image extraction and real time base calling on instrument computer
- Shorter time to results
  - Performed simultaneously with sequencing
  - Eliminates need to transfer images and intensities across network
  - Base calls and quality scores within hours of end of run

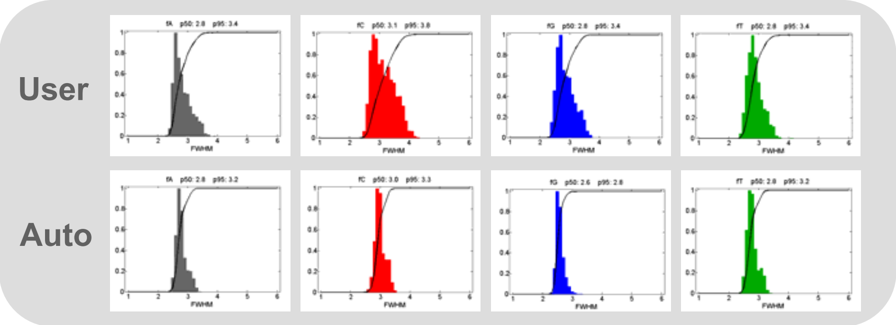


New Software Delivers Up to 40% More Data Per Run

- Pipeline 1.4 – Enhanced analysis algorithm
  - Increases yield, improves accuracy
  - Improved cluster delineation
  - More clusters pass filter
  - Lower error rates



- SCS 2.4 – Integrated autofocus
  - Easier to use, removes user error
  - Less variability in focus quality



What is *Real Time Analysis*?

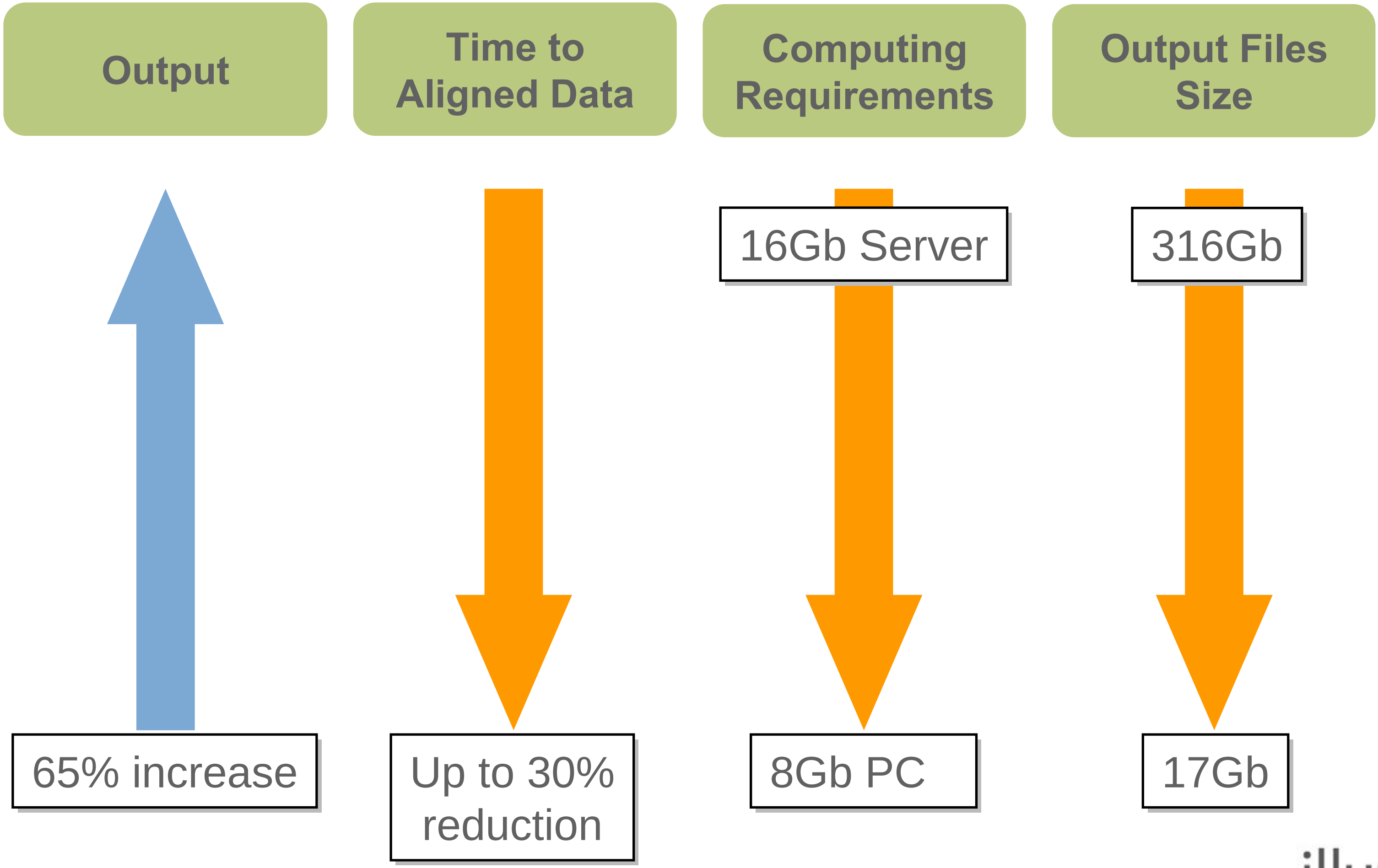
- The RTA module analyzes data as it leaves the Genome Analyzer
  - **Produces base calls, including Phred-like quality scores**
  - **Generates reports, to assess run and library quality**
  - Performs image analysis, generation of cluster intensities

- RTA simplifies the data management process
  - **Eliminates the need to transfer images from computer to computer**
  - Includes optional mechanisms for complete or selective archiving of images
  - Includes optional mechanisms for archiving of intensities

- RTA improves the system performance
  - **Minimizes time to results – base calls and qualities generated within hours of the end of the run**
  - Removes dependencies on network availability
  - Minimizes the time spent analyzing data after the run

Reversing the Trend

Simplified computing, Smaller storage needs, Faster analysis



Output

65% increase

Time to Aligned Data

Up to 30% reduction

Computing Requirements

16Gb Server

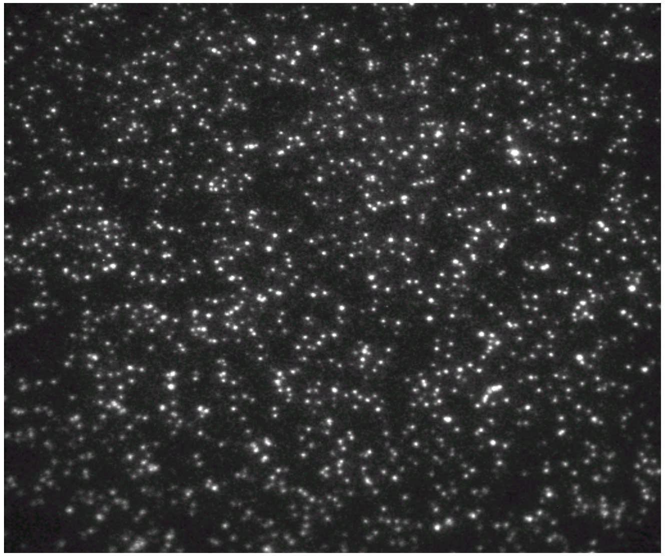
8Gb PC

Output Files Size

316Gb

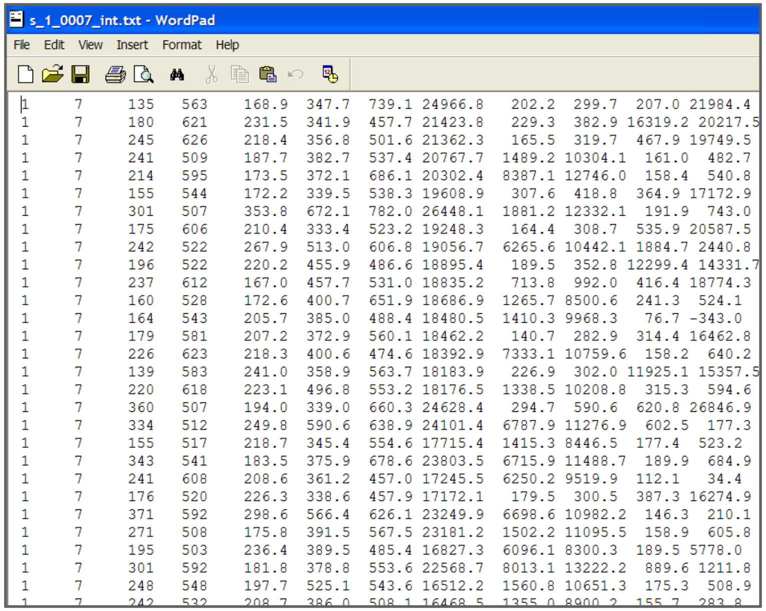
17Gb

FireCrest module: Start From the Spots…



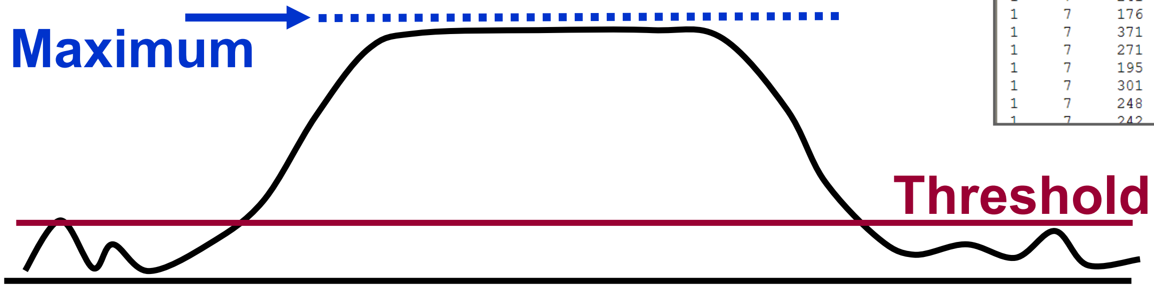
tiff image files

- Clusters identification and assigns intensities to them
- The output is a simple text file



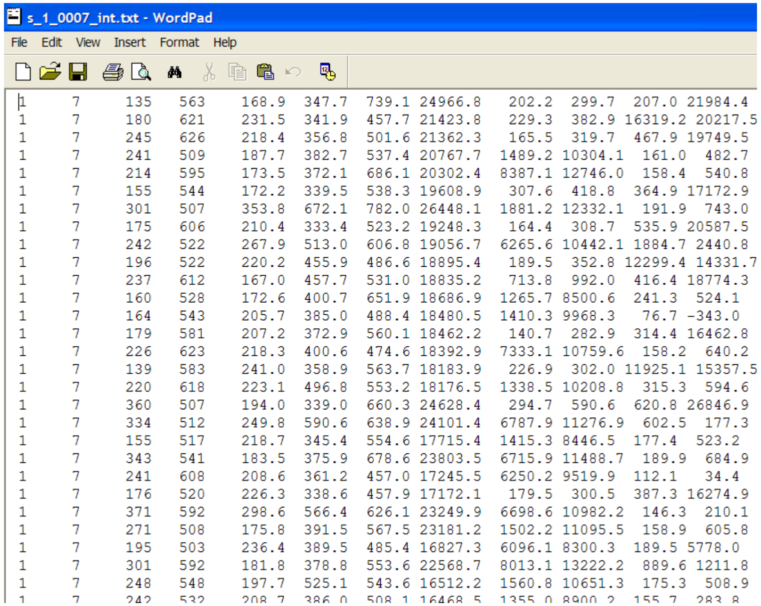
```
1       7       135     563     168.9   347.7   739.1 24966.8    202.2   299.7   207.0 21984.4
1       7       180     621     231.5   341.9   457.7 21423.8    229.3   382.9 16319.2 20217.5
1       7       245     626     218.4   356.8   501.6 21362.3    165.5   319.7   467.9 19749.5
1       7       241     509     187.7   382.7   537.4 20767.7   1489.2 10304.1   161.0   482.7
1       7       214     595     173.5   372.1   686.1 20302.4   8387.1 12746.0   158.4   540.8
1       7       155     544     172.2   339.5   538.3 19608.9    307.6   418.8   364.9 17172.9
1       7       301     507     353.8   672.1   782.0 26448.1   1881.2 12332.1   191.9   743.0
1       7       175     606     210.4   333.4   523.2 19248.3    164.4   308.7   535.9 20587.5
1       7       242     522     267.9   513.0   606.8 19056.7   6265.6 10442.1  1884.7  2440.8
1       7       196     522     220.2   455.9   486.6 18895.4    189.5   352.8 12299.4 14331.7
1       7       237     612     167.0   457.7   531.0 18835.2    713.8   992.0   416.4 18774.3
1       7       160     528     172.6   400.7   651.9 18686.9   1265.7  8500.6   241.3   524.1
1       7       164     543     205.7   385.0   488.4 18480.5   1410.3  9968.3    76.7  -343.0
1       7       179     581     207.2   372.9   560.1 18462.2    140.7   282.9   314.4 16462.8
1       7       226     623     218.3   400.6   474.6 18392.9   7333.1 10759.6   158.2   640.2
1       7       139     583     241.0   358.9   563.7 18183.9    226.9   302.0 11925.1 15357.5
1       7       220     618     223.1   496.8   553.2 18176.5   1338.5 10208.8   315.3   594.6
1       7       360     507     194.0   339.0   660.3 24628.4    294.7   590.6   620.8 26846.9
1       7       334     512     249.8   590.6   638.9 24101.4   6787.9 11276.9   602.5   177.3
1       7       155     517     218.7   345.4   554.6 17715.4   1415.3  8446.5   177.4   523.2
1       7       343     541     183.5   375.9   678.6 23803.5   6715.9 11488.7   189.9   684.9
1       7       241     608     208.6   361.2   457.0 17245.5   6250.2  9519.9   112.1    34.4
1       7       176     520     226.3   338.6   457.9 17172.1    179.5   300.5   387.3 16274.9
1       7       371     592     298.6   566.4   626.1 23249.9   6698.6 10982.2   146.3   210.1
1       7       271     508     175.8   391.5   567.5 23181.2   1502.2 11095.5   158.9   605.8
1       7       195     503     236.4   389.5   485.4 16827.3   6096.1  8300.3   189.5  5778.0
1       7       301     592     181.8   378.8   553.6 22568.7   8013.1 13222.2   889.6  1211.8
1       7       248     548     197.7   525.1   543.6 16512.2   1560.8 10651.3   175.3   508.9
1       7       242     532     208.7   386.0   508.1 16468.5   1355.0  8900.2   155.7   283.8
```

intensity files

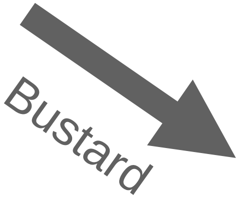


From Intensities to Reads

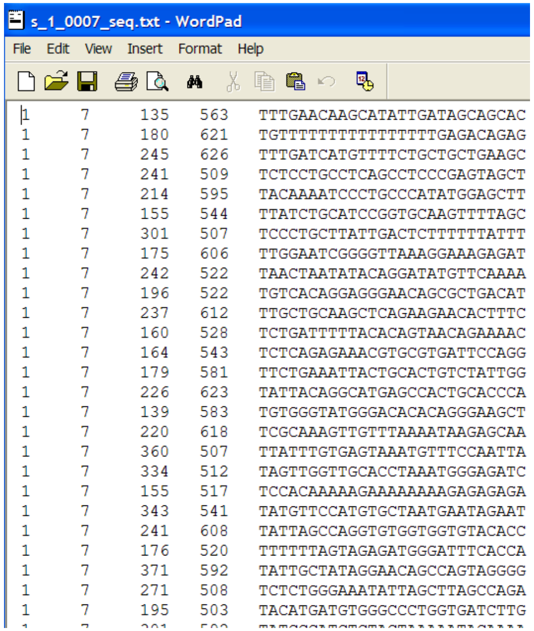
Intensity Files



```
1     7     135   563    168.9  347.7  739.1 24966.8   202.2   299.7   207.0 21984.4
1     7     180   621    231.5  341.9  457.7 21423.8   229.3   382.9 16319.2 20217.5
1     7     245   626    218.4  356.8  501.6 21362.3   165.5   319.7   467.9 19749.5
1     7     241   509    187.7  382.7  537.4 20767.7  1489.2 10304.1   161.0   482.7
1     7     214   595    173.5  372.1  686.1 20302.4  8387.1 12746.0   158.4   540.8
1     7     155   544    172.2  339.5  538.3 19608.9   307.6   418.8   364.9 17172.9
1     7     301   507    353.8  672.1  782.0 26448.1  1881.2 12332.1   191.9   743.0
1     7     175   606    210.4  333.4  523.2 19248.3   164.4   308.7   535.9 20587.5
1     7     242   522    267.9  513.0  606.8 19056.7  6265.6 10442.1  1884.7  2440.8
1     7     196   522    220.2  455.9  486.6 18895.4   189.5   352.8 12299.4 14331.7
1     7     237   612    167.0  457.7  531.0 18835.2   713.8   992.0   416.4 18774.3
1     7     160   528    172.6  400.7  651.9 18686.9  1265.7  8500.6   241.3   524.1
1     7     164   543    205.7  385.0  488.4 18480.5  1410.3  9968.3    76.7  -343.0
1     7     179   581    207.2  372.9  560.1 18462.2   140.7   282.9   314.4 16462.8
1     7     226   623    218.3  400.6  474.6 18392.9  7333.1 10759.6   158.2   640.2
1     7     139   583    241.0  358.9  563.7 18183.9   226.9   302.0 11925.1 15357.5
1     7     220   618    223.1  496.8  553.2 18176.5  1338.5 10208.8   315.3   594.6
1     7     360   507    194.0  339.0  660.3 24628.4   294.7   590.6   620.8 26846.9
1     7     334   512    249.8  590.6  638.9 24101.4  6787.9 11276.9   602.5   177.3
1     7     155   517    218.7  345.4  554.6 17715.4  1415.3  8446.5   177.4   523.2
1     7     343   541    183.5  375.9  678.6 23803.5  6715.9 11488.7   189.9   684.9
1     7     241   608    208.6  361.2  457.0 17245.5  6250.2  9519.9   112.1    34.4
1     7     176   520    226.3  338.6  457.9 17172.1   179.5   300.5   387.3 16274.9
1     7     371   592    298.6  566.4  626.1 23249.9  6698.6 10982.2   146.3   210.1
1     7     271   508    175.8  391.5  567.5 23181.2  1502.2 11095.5   158.9   605.8
1     7     195   503    236.4  389.5  485.4 16827.3  6096.1  8300.3   189.5  5778.0
1     7     301   592    181.8  378.8  553.6 22568.7  8013.1 13222.2   889.6  1211.8
1     7     248   548    197.7  525.1  543.6 16512.2  1560.8 10651.3   175.3   508.9
1     7     242   532    208.7  386.0  508.1 16468.5  1355.0  8900.2   155.7   283.8
```



Sequence Files

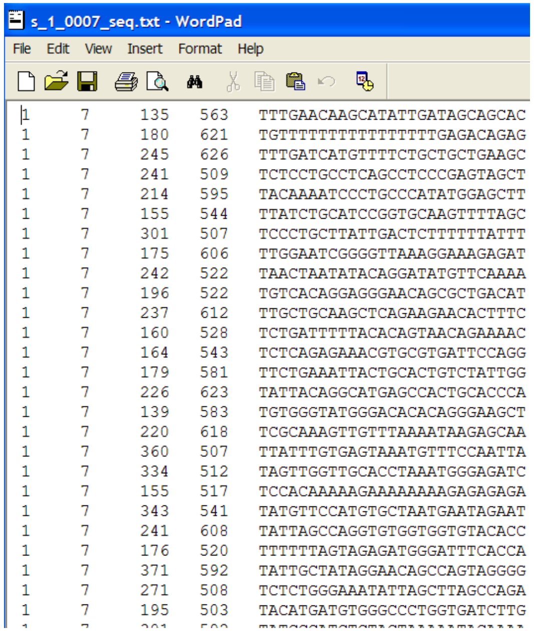


```
1      7      135    563    TTTGAACAAGCATATTGATAGCAGCAC
1      7      180    621    TGTTTTTTTTTTTTTTTGAGACAGAG
1      7      245    626    TTTGATCATGTTTTCTGCTGCTGAAGC
1      7      241    509    TCTCCTGCCTCAGCCTCCCGAGTAGCT
1      7      214    595    TACAAAATCCCTGCCCATATGGAGCTT
1      7      155    544    TTATCTGCATCCGGTGCAAGTTTTAGC
1      7      301    507    TCCCTGCTTATTGACTCTTTTTTATTT
1      7      175    606    TTGGAATCGGGGTTAAAGGAAAGAGAT
1      7      242    522    TAACTAATATACAGGATATGTTCAAAA
1      7      196    522    TGTCACAGGAGGGAACAGCGCTGACAT
1      7      237    612    TTGCTGCAAGCTCAGAAGAACACTTTC
1      7      160    528    TCTGATTTTTACACAGTAACAGAAAAC
1      7      164    543    TCTCAGAGAAACGTGCGTGATTCCAGG
1      7      179    581    TTCTGAAATTACTGCACTGTCTATTGG
1      7      226    623    TATTACAGGCATGAGCCACTGCACCCA
1      7      139    583    TGTGGGTATGGGACACACAGGGAAGCT
1      7      220    618    TCGCAAAGTTGTTTAAAATAAGAGCAA
1      7      360    507    TTATTTGTGAGTAAATGTTTCCAATTA
1      7      334    512    TAGTTGGTTGCACCTAAATGGGAGATC
1      7      155    517    TCCACAAAAAGAAAAAAAAGAGAGAGA
1      7      343    541    TATGTTCCATGTGCTAATGAATAGAAT
1      7      241    608    TATTAGCCAGGTGTGGTGGTGTACACC
1      7      176    520    TTTTTTAGTAGAGATGGGATTTCACCA
1      7      371    592    TATTGCTATAGGAACAGCCAGTAGGGG
1      7      271    508    TCTCTGGGAAATATTAGCTTAGCCAGA
1      7      195    503    TACATGATGTGGGCCCTGGTGATCTTG
1      7      301    502
```

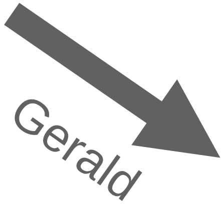
- Transforms intensities into base-calls

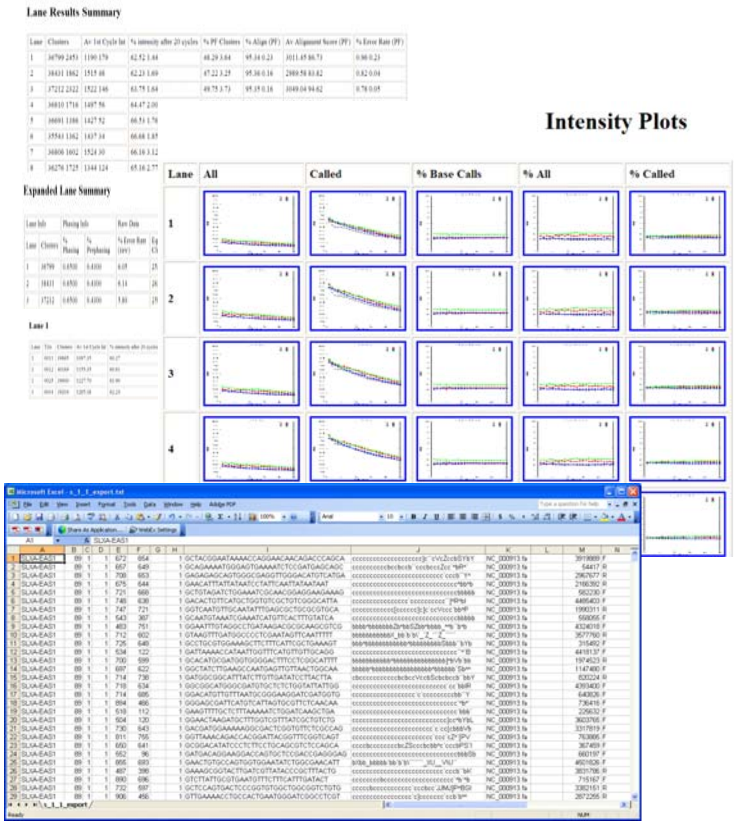
From Reads to Aligned Sequences

sequence files



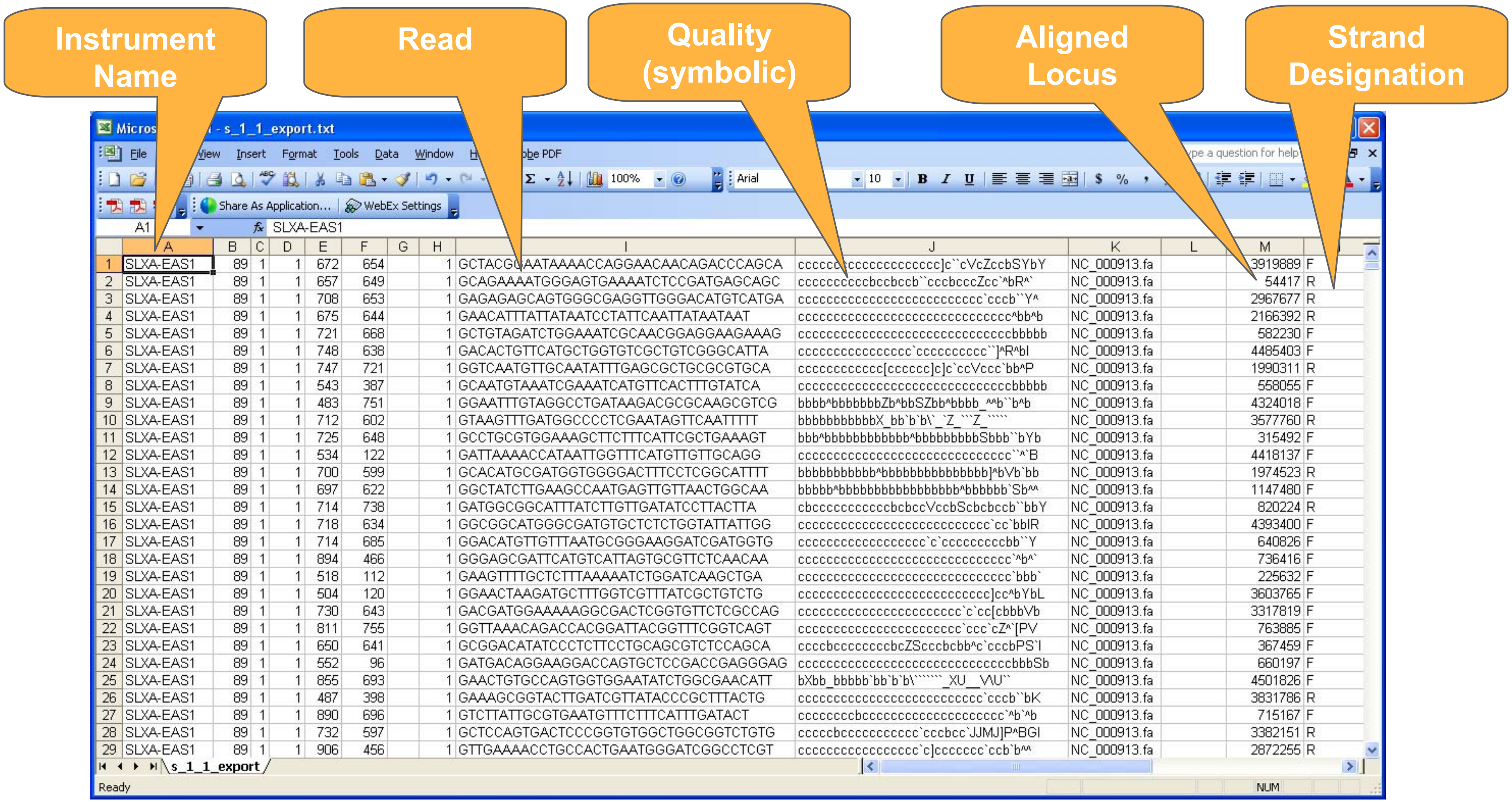
```
1    7    135    563    TTTGAACAAGCATATTGATAGCAGCAC
1    7    180    621    TGTTTTTTTTTTTTTTTTGAGACAGAG
1    7    245    626    TTTGATCATGTTTTCTGCTGCTGAAGC
1    7    241    509    TCTCCTGCCTCAGCCTCCCGAGTAGCT
1    7    214    595    TACAAAATCCCTGCCCATATGGAGCTT
1    7    155    544    TTATCTGCATCCGGTGCAAGTTTTAGC
1    7    301    507    TCCCTGCTTATTGACTCTTTTTTATTT
1    7    175    606    TTGGAATCGGGGTTAAAGGAAAGAGAT
1    7    242    522    TAACTAATATACAGGATATGTTCAAAA
1    7    196    522    TGTCACAGGAGGGAACAGCGCTGACAT
1    7    237    612    TTGCTGCAAGCTCAGAAGAACACTTTC
1    7    160    528    TCTGATTTTTACACAGTAACAGAAAAC
1    7    164    543    TCTCAGAGAAACGTGCGTGATTCCAGG
1    7    179    581    TTCTGAAATTACTGCACTGTCTATTGG
1    7    226    623    TATTACAGGCATGAGCCACTGCACCCA
1    7    139    583    TGTGGGTATGGGACACACAGGGAAGCT
1    7    220    618    TCGCAAAGTTGTTTAAAATAAGAGCAA
1    7    360    507    TTATTTGTGAGTAAATGTTTCCAATTA
1    7    334    512    TAGTTGGTTGCACCTAAATGGGAGATC
1    7    155    517    TCCACAAAAAGAAAAAAAAGAGAGAGA
1    7    343    541    TATGTTCCATGTGCTAATGAATAGAAT
1    7    241    608    TATTAGCCAGGTGTGGTGGTGTACACC
1    7    176    520    TTTTTTAGTAGAGATGGGATTTCACCA
1    7    371    592    TATTGCTATAGGAACAGCCAGTAGGGG
1    7    271    508    TCTCTGGGAAATATTAGCTTAGCCAGA
1    7    195    503    TACATGATGTGGGCCCTGGTGATCTTG
```





Export.txt

_export.txt – Flexible Output for Myriad Applications



Instrument Name

Read

Quality (symbolic)

Aligned Locus

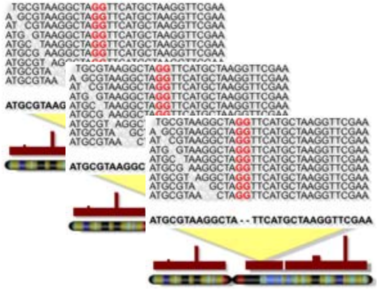
Strand Designation

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	SLXA-EAS1	89	1	1	672	654		1	GCTACGGAATAAAACCAGGAACAACAGACCCAGCA	cccccccccccccccccccc]c``cVcZccbSYbY	NC_000913.fa		3919889	F
2	SLXA-EAS1	89	1	1	657	649		1	GCAGAAAATGGGAGTGAAAATCTCCGATGAGCAGC	ccccccccccbccbccb``cccbcccZcc`^bR^`	NC_000913.fa		54417	R
3	SLXA-EAS1	89	1	1	708	653		1	GAGAGAGCAGTGGGCGAGGTTGGGACATGTCATGA	cccccccccccccccccccccccccc`cccb``Y^	NC_000913.fa		2967677	R
4	SLXA-EAS1	89	1	1	675	644		1	GAACATTTATTATAATCCTATTCAATTATAATAAT	cccccccccccccccccccccccccccccc^bb^b	NC_000913.fa		2166392	R
5	SLXA-EAS1	89	1	1	721	668		1	GCTGTAGATCTGGAAATCGCAACGGAGGAAGAAAG	ccccccccccccccccccccccccccccccbbbbb	NC_000913.fa		582230	F
6	SLXA-EAS1	89	1	1	748	638		1	GACACTGTTCATGCTGGTGTCGCTGTCGGGCATTA	cccccccccccccccc`cccccccccc``]^R^bl	NC_000913.fa		4485403	F
7	SLXA-EAS1	89	1	1	747	721		1	GGTCAATGTTGCAATATTTGAGCGCTGCGCGTGCA	cccccccccccc[cccccc]c]c`ccVccc`bb^P	NC_000913.fa		1990311	R
8	SLXA-EAS1	89	1	1	543	387		1	GCAATGTAAATCGAAATCATGTTCACTTTGTATCA	ccccccccccccccccccccccccccccccbbbbb	NC_000913.fa		558055	F
9	SLXA-EAS1	89	1	1	483	751		1	GGAATTTGTAGGCCTGATAAGACGCGCAAGCGTCG	bbbb^bbbbbbbZb^bbSZbb^bbbb_^^b``b^b	NC_000913.fa		4324018	F
10	SLXA-EAS1	89	1	1	712	602		1	GTAAGTTTGATGGCCCCTCGAATAGTTCAATTTTT	bbbbbbbbbbbX_bb`b`b\`_`Z_```Z_`````	NC_000913.fa		3577760	R
11	SLXA-EAS1	89	1	1	725	648		1	GCCTGCGTGGAAAGCTTCTTTCATTCGCTGAAAGT	bbb^bbbbbbbbbbbbb^bbbbbbbbSbbb``bYb	NC_000913.fa		315492	F
12	SLXA-EAS1	89	1	1	534	122		1	GATTAAAACCATAATTGGTTTCATGTTGTTGCAGG	cccccccccccccccccccccccccccccc``^`B	NC_000913.fa		4418137	F
13	SLXA-EAS1	89	1	1	700	599		1	GCACATGCGATGGTGGGGACTTTCCTCGGCATTTT	bbbbbbbbbbb^bbbbbbbbbbbbbbb]^bVb`bb	NC_000913.fa		1974523	R
14	SLXA-EAS1	89	1	1	697	622		1	GGCTATCTTGAAGCCAATGAGTTGTTAACTGGCAA	bbbbb^bbbbbbbbbbbbbbbbb^bbbbbb`Sb^^	NC_000913.fa		1147480	F
15	SLXA-EAS1	89	1	1	714	738		1	GATGGCGGCATTTATCTTGTTGATATCCTTACTTA	cbcccccccccccbcbccVccbScbcbccb``bbY	NC_000913.fa		820224	R
16	SLXA-EAS1	89	1	1	718	634		1	GGCGGCATGGGCGATGTGCTCTCTGGTATTATTGG	ccccccccccccccccccccccccccc`cc`bblR	NC_000913.fa		4393400	F
17	SLXA-EAS1	89	1	1	714	685		1	GGACATGTTGTTTAATGCGGGAAGGATCGATGGTG	ccccccccccccccccc`c`ccccccccccbb``Y	NC_000913.fa		640826	F
18	SLXA-EAS1	89	1	1	894	466		1	GGGAGCGATTCATGTCATTAGTGCGTTCTCAACAA	cccccccccccccccccccccccccccccc`^b^`	NC_000913.fa		736416	F
19	SLXA-EAS1	89	1	1	518	112		1	GAAGTTTTGCTCTTTAAAAATCTGGATCAAGCTGA	cccccccccccccccccccccccccccccc`bbb`	NC_000913.fa		225632	F
20	SLXA-EAS1	89	1	1	504	120		1	GGAACTAAGATGCTTTGGTCGTTTATCGCTGTCTG	ccccccccccccccccccccccccccc]cc^bYbL	NC_000913.fa		3603765	F
21	SLXA-EAS1	89	1	1	730	643		1	GACGATGGAAAAAGGCGACTCGGTGTTCTCGCCAG	ccccccccccccccccccccccc`c`cc[cbbbVb	NC_000913.fa		3317819	F
22	SLXA-EAS1	89	1	1	811	755		1	GGTTAAACAGACCACGGATTACGGTTTCGGTCAGT	ccccccccccccccccccccccc`ccc`cZ^`[PV	NC_000913.fa		763885	F
23	SLXA-EAS1	89	1	1	650	641		1	GCGGACATATCCCTCTTCCTGCAGCGTCTCCAGCA	ccccbccccccccbcZScccbcbb^c`cccbPS`l	NC_000913.fa		367459	F
24	SLXA-EAS1	89	1	1	552	96		1	GATGACAGGAAGGACCAGTGCTCCGACCGAGGGAG	ccccccccccccccccccccccccccccccbbbSb	NC_000913.fa		660197	F
25	SLXA-EAS1	89	1	1	855	693		1	GAACTGTGCCAGTGGTGGAATATCTGGCGAACATT	bXbb_bbbbb`bb`b`b\``````_XU__V\U``	NC_000913.fa		4501826	F
26	SLXA-EAS1	89	1	1	487	398		1	GAAAGCGGTACTTGATCGTTATACCCGCTTTACTG	cccccccccccccccccccccccccc`cccb``bK	NC_000913.fa		3831786	R
27	SLXA-EAS1	89	1	1	890	696		1	GTCTTATTGCGTGAATGTTTCTTTCATTTGATACT	ccccccccbcccccccccccccccccccc`^b`^b	NC_000913.fa		715167	F
28	SLXA-EAS1	89	1	1	732	597		1	GCTCCAGTGACTCCCGGTGTGGCTGGCGGTCTGTG	cccccbccccccccccc`cccbcc`JJMJ]P^BGl	NC_000913.fa		3382151	R
29	SLXA-EAS1	89	1	1	906	456		1	GTTGAAAACCTGCCACTGAATGGGATCGGCCTCGT	ccccccccccccccccc`c]ccccccc`ccb`b^^	NC_000913.fa		2872255	R

CASAVA

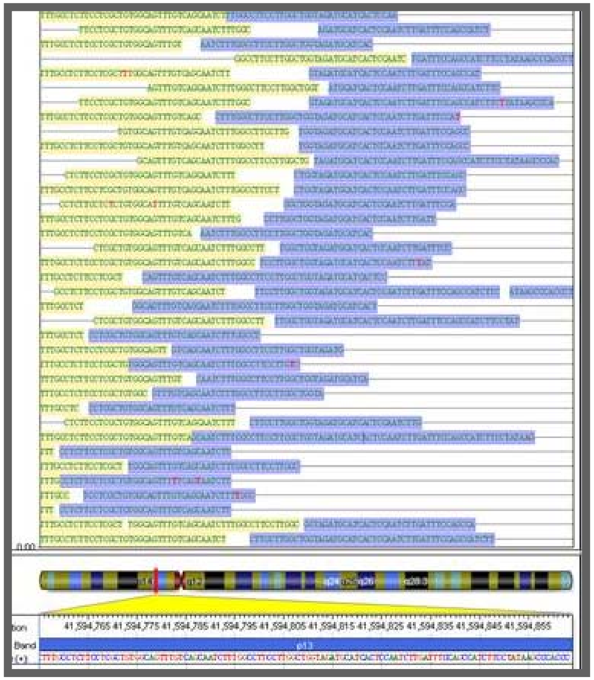
Consensus Assessment of Sequence and Variation

Aligned Reads

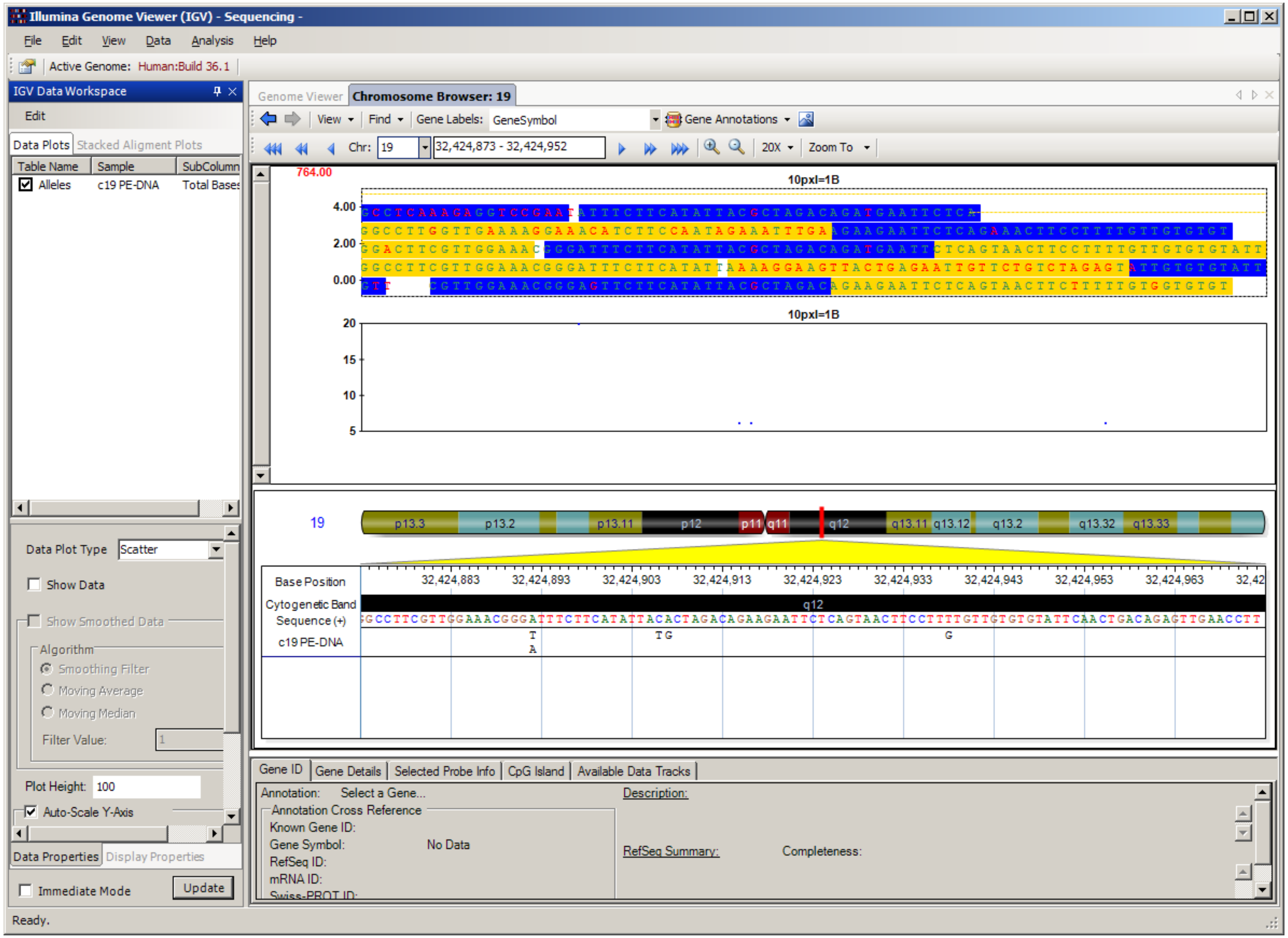

TGCGTAAGGCTAGGTTCATGCTAAGGTTCGAA
ATGCGTAAGGCTA - - TTCATGCTAAGGTTCGAA

Consensus assembly

BINS
SNPs
Counts



Consensus Sequence



Illumina Genome Viewer (IGV) - Sequencing -

Active Genome: Human:Build 36.1

Chromosome Browser: 19

Gene Labels: GeneSymbol

Chr: 19 32,424,873 - 32,424,952 20X

Table Name	Sample	SubColumn
☑ Alleles	c19 PE-DNA	Total Bases

Data Plot Type: Scatter

□ Show Data

□ Show Smoothed Data

Algorithm

● Smoothing Filter

○ Moving Average

○ Moving Median

Filter Value: 1

Plot Height: 100

☑ Auto-Scale Y-Axis

□ Immediate Mode

Base Position: 32,424,883 32,424,893 32,424,903 32,424,913 32,424,923 32,424,933 32,424,943 32,424,953 32,424,963 32,42

Cytogenetic Band: q12

Sequence (+): GGCCTTCGTTGGAAACGGGATTTCTTCATATTACACTAGACAGAAGAATTCTCAGTAACTTCCTTTTGTTGTGTGTATTCAACTGACAGAGTTGAACCTT

c19 PE-DNA: T A TG G

Annotation: Select a Gene...

Annotation Cross Reference

Known Gene ID:

Gene Symbol: No Data

RefSeq ID:

mRNA ID:

Swiss-PROT ID:

Description:

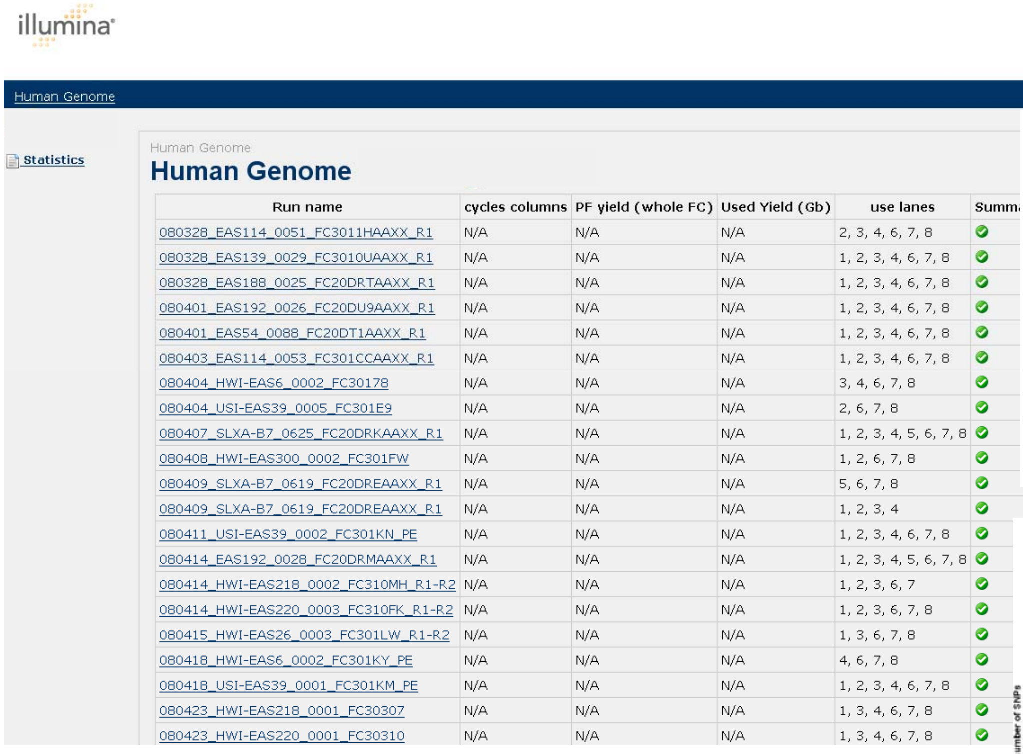
RefSeq Summary:

Completeness:

Ready.

Output Interpretation - html

- Project directory/html/Home.html



illumina

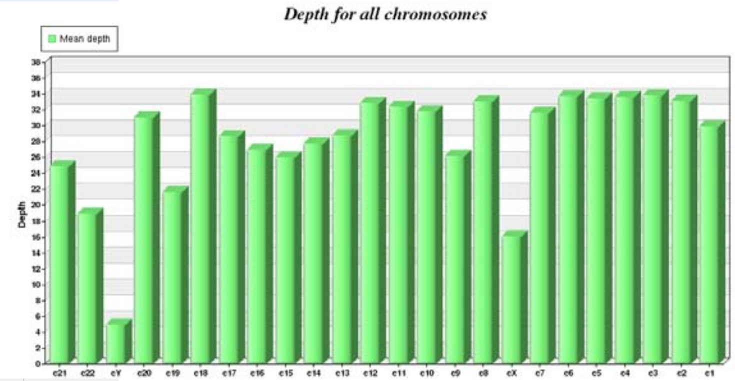
Human Genome

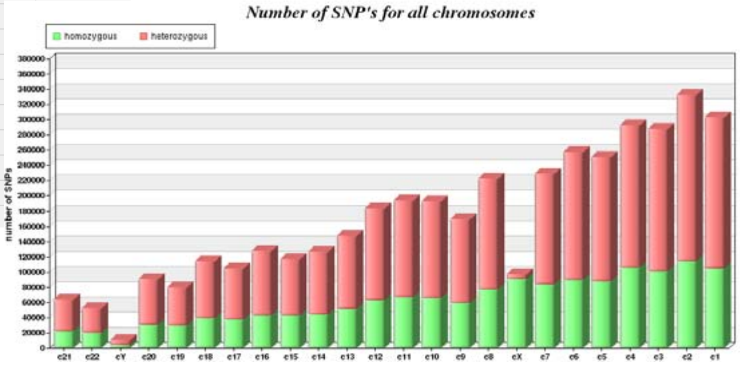
Statistics

Human Genome

Human Genome

Run name	cycles columns	PF yield (whole FC)	Used Yield (Gb)	use lanes	Summa
080328_EAS114_0051_FC3011HAAXX_R1	N/A	N/A	N/A	2, 3, 4, 6, 7, 8	✓
080328_EAS139_0029_FC3010UAAXX_R1	N/A	N/A	N/A	1, 2, 3, 4, 6, 7, 8	✓
080328_EAS188_0025_FC20DRTAAXX_R1	N/A	N/A	N/A	1, 2, 3, 4, 6, 7, 8	✓
080401_EAS192_0026_FC20DU9AAXX_R1	N/A	N/A	N/A	1, 2, 3, 4, 6, 7, 8	✓
080401_EAS54_0088_FC20DT1AAXX_R1	N/A	N/A	N/A	1, 2, 3, 4, 6, 7, 8	✓
080403_EAS114_0053_FC301CCAAXX_R1	N/A	N/A	N/A	1, 2, 3, 4, 6, 7, 8	✓
080404_HWI-EAS6_0002_FC30178	N/A	N/A	N/A	3, 4, 6, 7, 8	✓
080404_USI-EAS39_0005_FC301E9	N/A	N/A	N/A	2, 6, 7, 8	✓
080407_SLXA-B7_0625_FC20DRKAAXX_R1	N/A	N/A	N/A	1, 2, 3, 4, 5, 6, 7, 8	✓
080408_HWI-EAS300_0002_FC301FW	N/A	N/A	N/A	1, 2, 6, 7, 8	✓
080409_SLXA-B7_0619_FC20DREAAXX_R1	N/A	N/A	N/A	5, 6, 7, 8	✓
080409_SLXA-B7_0619_FC20DREAAXX_R1	N/A	N/A	N/A	1, 2, 3, 4	✓
080411_USI-EAS39_0002_FC301KN_PE	N/A	N/A	N/A	1, 2, 3, 4, 6, 7, 8	✓
080414_EAS192_0028_FC20DRMAAXX_R1	N/A	N/A	N/A	1, 2, 3, 4, 5, 6, 7, 8	✓
080414_HWI-EAS218_0002_FC310MH_R1-R2	N/A	N/A	N/A	1, 2, 3, 6, 7	✓
080414_HWI-EAS220_0003_FC310FK_R1-R2	N/A	N/A	N/A	1, 2, 3, 6, 7, 8	✓
080415_HWI-EAS26_0003_FC301LW_R1-R2	N/A	N/A	N/A	1, 3, 6, 7, 8	✓
080418_HWI-EAS6_0002_FC301KY_PE	N/A	N/A	N/A	4, 6, 7, 8	✓
080418_USI-EAS39_0001_FC301KM_PE	N/A	N/A	N/A	1, 2, 3, 4, 6, 7, 8	✓
080423_HWI-EAS218_0001_FC30307	N/A	N/A	N/A	1, 3, 4, 6, 7, 8	✓
080423_HWI-EAS220_0001_FC30310	N/A	N/A	N/A	1, 3, 4, 6, 7, 8	✓



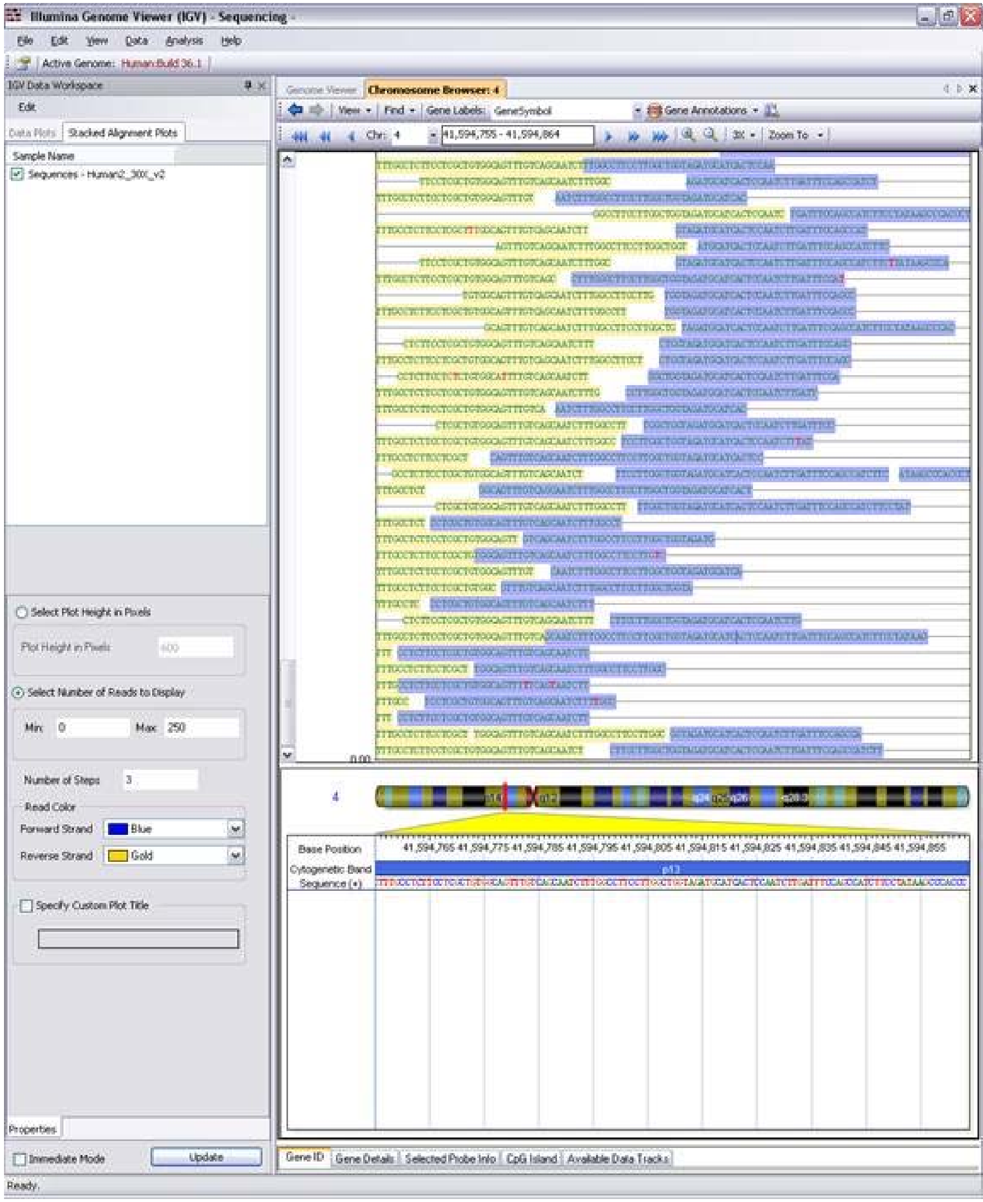


Overview of CASAVA Outputs for GenomeStudio™

- Output utilized for DNA experiments
  - Sorted.txt files binned by chromosome and subdivided into 10 megabase bins
  - SNP.txt (1 per chromosome)
  - Run_summary.xml
  - Run.conf file
  - Project.conf file
- Output utilized for RNA experiments
  - Sorted.txt files binned by chromosome and subdivided into 10 megabase bins
  - SNP.txt (1 per chromosome)
  - Exon counts file (1 per chromosome)
  - Gene counts file (1 per chromosome)
  - Splice Junction counts file (1 per chromosome)
  - Run_summary.xml
  - Run.conf file
  - Project.conf file

DNA Sequencing Module

- Direct import of data from Pipeline/CASAVA
- Visualization of SNPs
- Browsing of coverage and consensus reads
- Export of SNP tables



Active Genome: Human Build 36.1

Gene Labels: GeneSymbol

Chr: 4 41,594,755 - 41,594,864

☑ Sequences - Human2_30X_v2

○ Select Plot Height in Pixels

Plot Height in Pixels 600

● Select Number of Reads to Display

Min: 0 Max: 250

Number of Steps 3

Read Color

Forward Strand 🟦 Blue

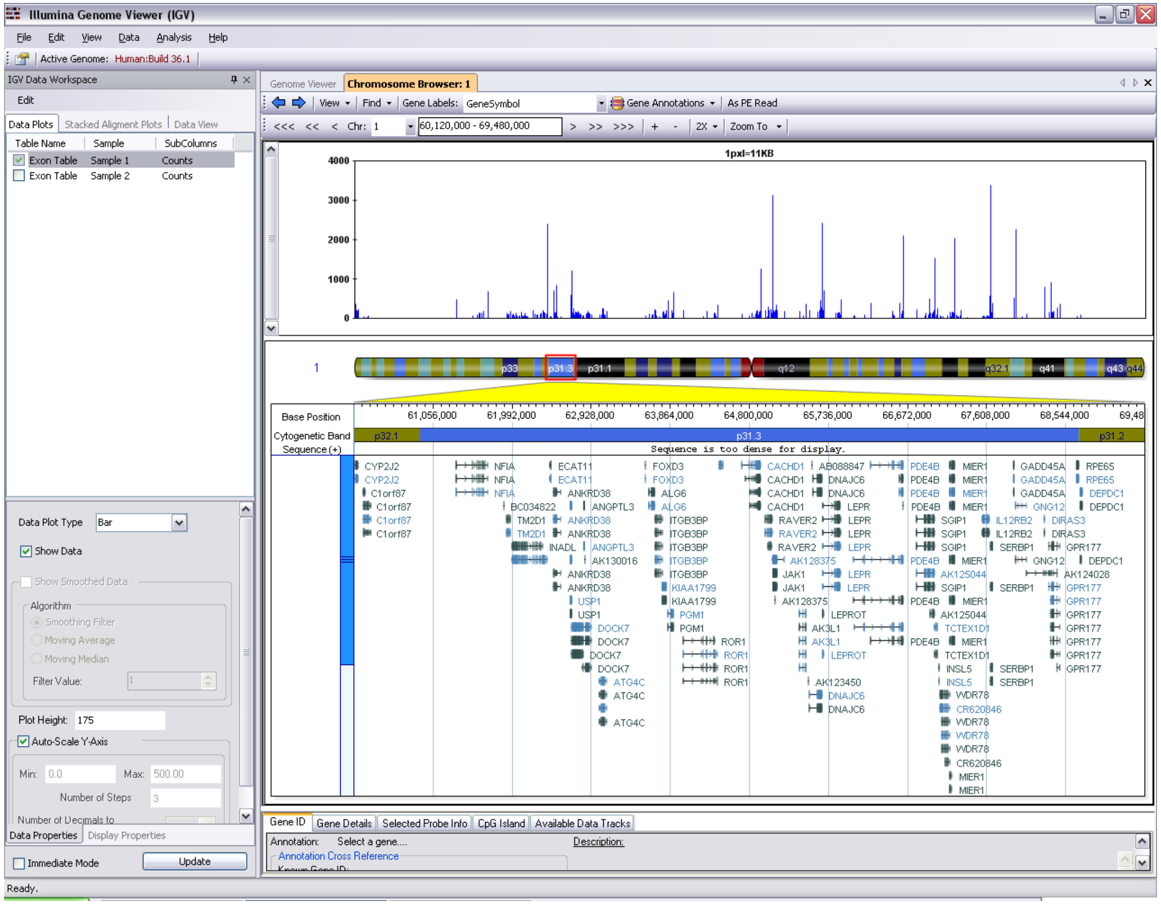
Reverse Strand 🟨 Gold

□ Specify Custom Plot Title

□ Immediate Mode

Exon Counts from RNA Data

Browser view of exon counts



Active Genome: Human:Build 36.1

Table Name	Sample	SubColumns
☑ Exon Table	Sample 1	Counts
□ Exon Table	Sample 2	Counts

Data Plot Type: Bar

☑ Show Data

□ Show Smoothed Data

Algorithm

● Smoothing Filter

○ Moving Average

○ Moving Median

Filter Value: 1

Plot Height: 175

☑ Auto-Scale Y-Axis

Min: 0.0 Max: 500.00

Number of Steps 3

Number of Decimals to

□ Immediate Mode

Chr: 1 60,120,000 - 69,480,000

Annotation: Select a gene....

Description:

Annotation Cross Reference

Known Gene ID:

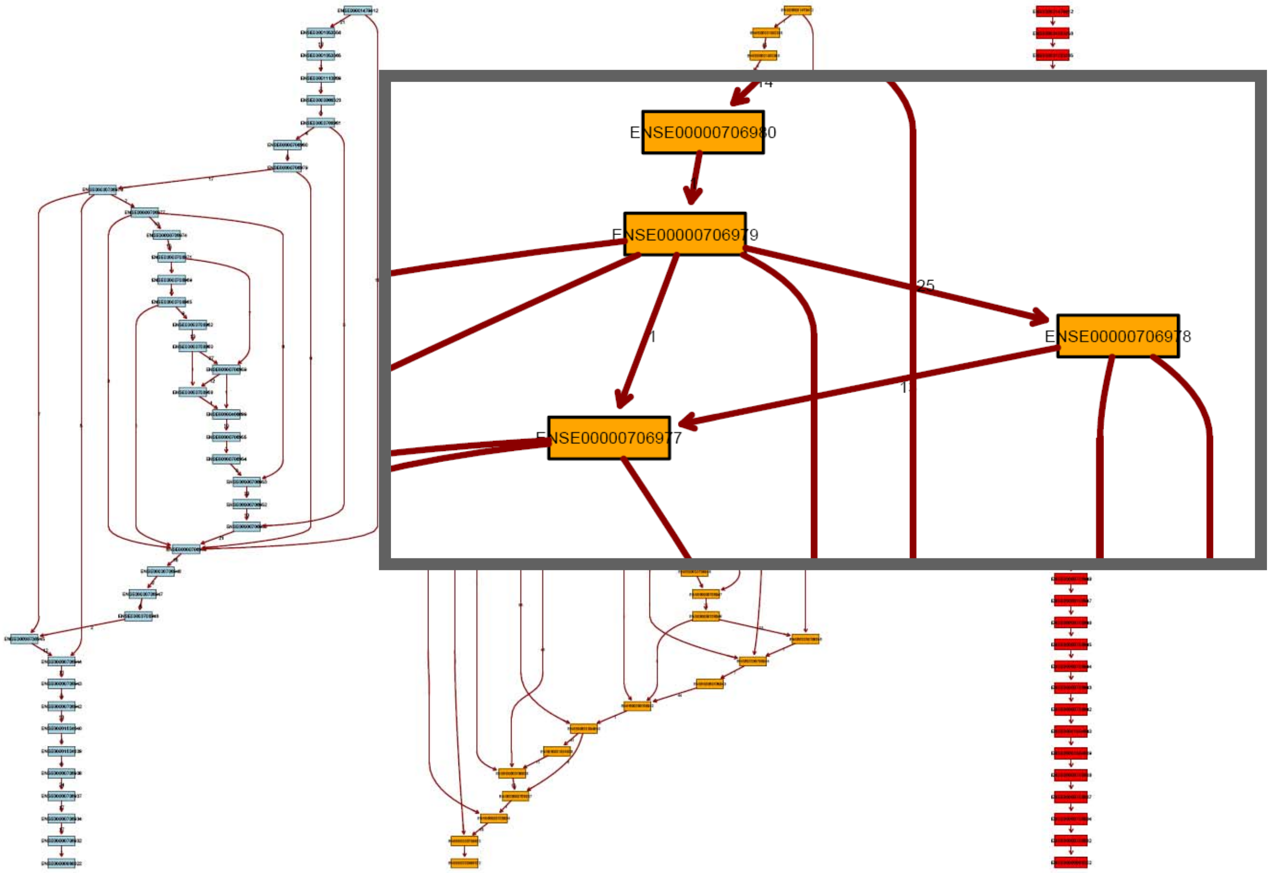
Ready.

A New Look at Alternative Splicing

Brain

UHR

Ensembl



Leveraging the GA Informatics Community

De Novo 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
  - Group: Rene Warren, et al; British Columbia
  - http://bioinformatics.oxfordjournals.org/cgi/content/full/23/4/500
- Euler SR – Genomic Assembly
  - Group: Pavel Pevzner, Mark Chaisson; UC San Diego
  - http://nbcr.sdsc.edu/euler/

Rapidly Expanding Choice of Open Source Tools

Genomic Alignment Browsers

- Gbrowse – Genomic Browsing
  - Generic Model Organism Database Project
  - http://www.gmod.org/wlk/index.pho/Gbrowse
- UCSC Browser – Genome browsing and comprehensive annotation
  - Generic Model Organism Database Project
  - http://www.genome.ucsc.edu/goldenPath/help/customTrack.html
- Anno-J – Genome Annotation and Visualization
  - Computational Systems Biology Center of Excellence
  - http://www.annoj.org/csb_index.shtml

Leveraging the GA Informatics Community

Alignment and Polymorphism Detection

- MAQ – Mapping and Assembly with Quality
  - Heng Li, Sanger Centre
  - http://maq.sourceforge.net/maq-man.shtml
- SOAP – Short Oligonucleotide Alignment Program
  - Ruiqiang Li, Beijing Genomics Institute
  - http://soap.genomics.org.cn/
- Consed – Alignement and Polymorphism Detection
  - Green Lab, U. Washington (commercial offering)
  - http://bozeman.mbt.washington.edu/consed/consed.html

Rapidly Expanding Choice of Open Source Tools

ChIP Sequencing

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

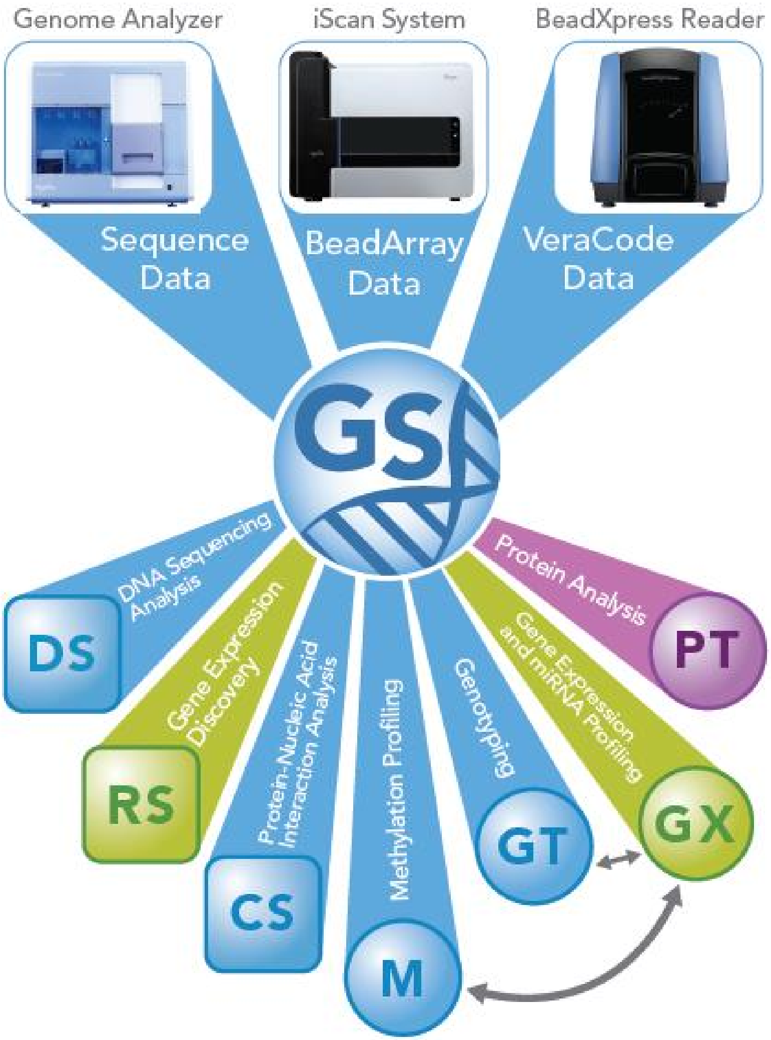
Digital Gene Expression

- Comparative Count Display
  - Alex Lash, NIH
  - ftp://ftp.ncbi.nlm.nih.gov/pub/sage/obsolete/bin/ccd/
- SAGE DGED Tool
  - Cancer Genome Anatomy Project
  - http://cgap.nci.nih.gov/SAGE/SDGED_Wizard?METHOD=SS10,LS108ORG=Hs

Genome Studio: Data Analysis Platform for Many Applications

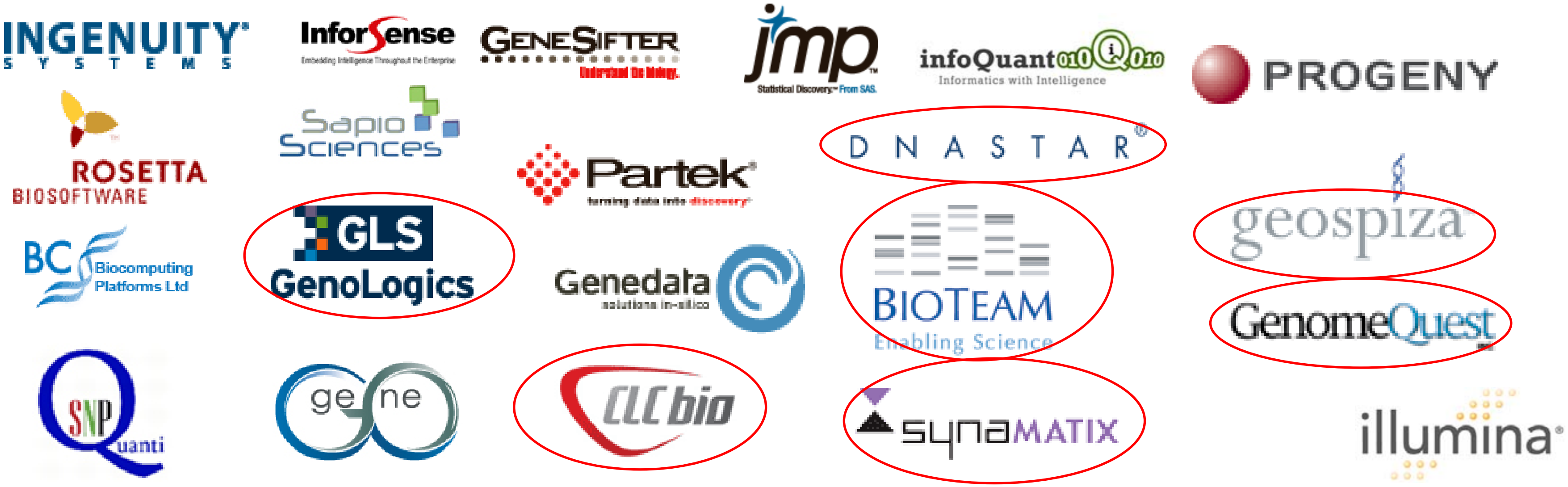
A single workbench with a growing number of modules

- Sequencing
- Genotyping (GT)
- Gene Expression (GX)
- Regulation (M)
- ChiP Sequencing (CS)
- We are changing the name to reflect a new platform →
  - microarray + sequencing



Connecting with the larger informatics universe

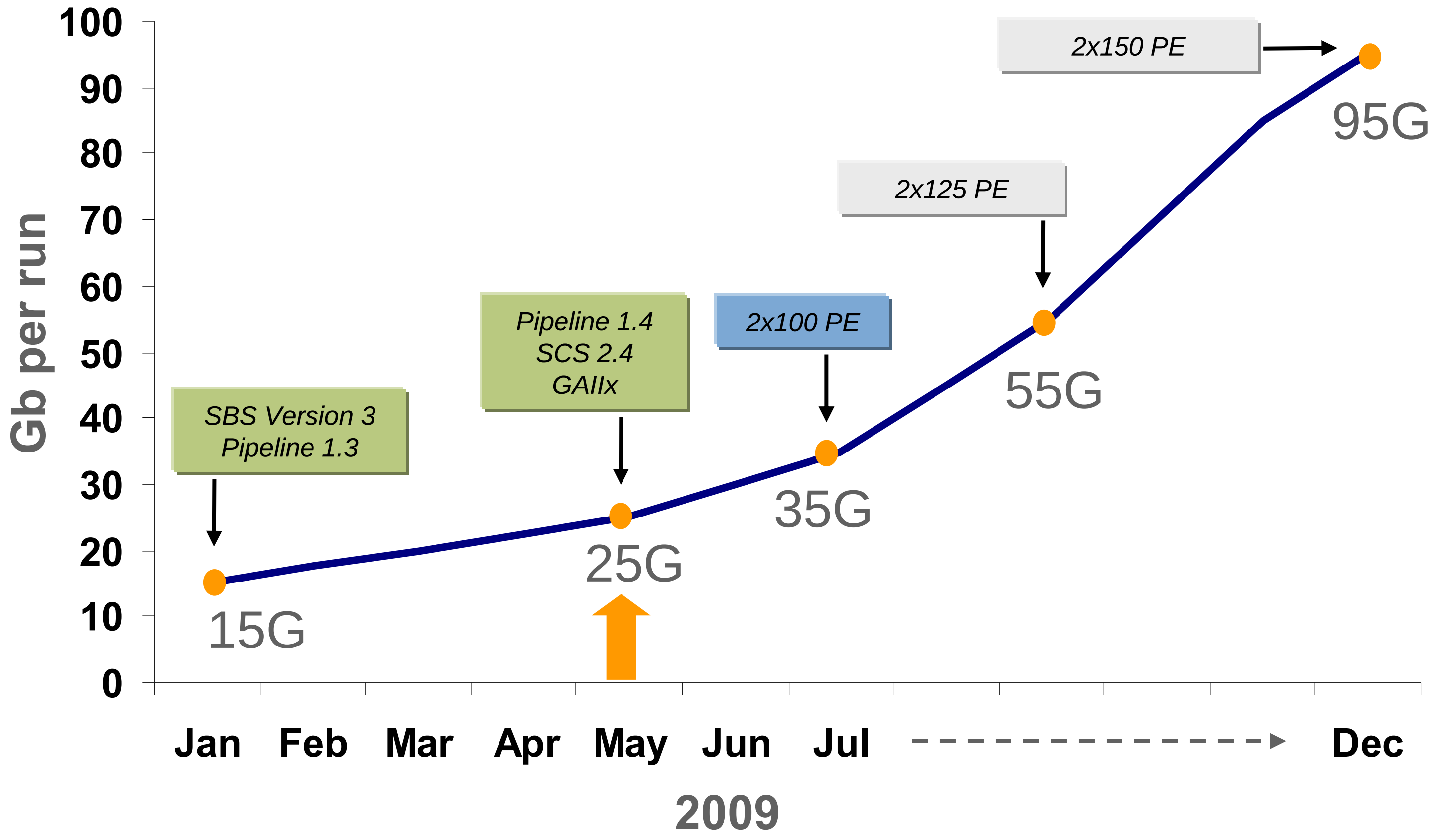
- Illumina has a 3rd party partnership program – illumina•Connect – launched last year designed to "increase software/hardware ecosystem connecting Illumina to researchers in genomics, genetics and sequencing communities"
  - ~30 vendors and academic partners in the program
  - http://www.illumina.com/pagesnrn.ilmn?ID=229
  - Integrated tools connecting GenomeStudio to several 3rd party apps for microarray and sequencing data analysis



Delivering on Roadmap Milestones

15x increase in 2008

4-5x increase in 2009



Sequencing Enabled iScan

What is needed?
How does it work?
What can it do?

iScan Sequencing Module technical Update

- Add on to enable sequencing on the iScan
  - Uses lasers and optics of iScan for imaging of flow cell
  - Fluidics module holding sequencing reagents, pumps and reagent chiller
- Sequencing specs
  - Throughput: ~ 0.5 GB per day, up to 5 GB per run
  - Data density: ~ 32 M clusters per run, 60M Reads
  - Flexible read length: from short single reads to 2x75 bb reads
  - 8 lanes, 1-12 samples per lane
- Applications supported
  - smallRNA, CHiP-Seq & mRNA seq
  - Targeted re-sequencing & re-sequencing of small genomes
  - High Density GWAS, Medium and HD Custom GTP (Array Applications)
  - WG Gene Expression Profiling, Human, Mouse and Rat (Array Applications)

Thank You

Marco Cappelletti

mcappelletti@illumina.com