Introduction to transcriptome analysis using High-Throughput Sequencing technologies

D. Puthier

2012

Nature Methods - **5**, 585 - 587 (2008)
doi:10.1038/nmeth0708-585

The beginning of the end for microarrays?

Jay Shendure

Jay Shendure is in the Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. shendure@u.washington.edu

**Two complementary app
successfully tackled the
once revealing unpreced**

Published online 15 October 2008 | *Nature* **455**, 847 (2008) |
doi:10.1038/455847a

News

The death of microarrays?

High-throughput gene sequencing seems to be stealing a march on microarrays. Heidi Ledford looks at a genome technology facing intense competition.

Heidi Ledford

Stem cell transcriptome profiling via massive-scale mRNA sequencing pp613 - 619
Nicole Cloonan, Alistair R R Forrest, Gabriel Kolle, Brooke B A Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, Alan J Robertson, Andrew C Perkins, Stephen J Bruce, Clarence C Lee, Swati S Ranade, Heather E Peckham, Jonathan M Manning, Kevin J McKernan & Sean M Grimmond

Published online: 30 May 2008 | doi:10.1038/nmeth.1223

Application of next-generation sequencing using the ABI SOLiD technology to mammalian transcriptome analysis enabled a survey of the content, the complexity and the developmental dynamics of the embryonic stem cell transcriptome in the mouse. Also in this issue, Mortazavi *et al.* report Illumina technology–based RNA-Seq analysis of the mouse transcriptome in three different tissues.

Mapping and quantifying mammalian transcriptomes by RNA-Seq pp621 - 628
Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer & Barbara Wold

Published online: 30 May 2008 | doi:10.1038/nmeth.1226

The mouse transcriptome in three tissue types has been analyzed using Illumina next-generation sequencing technology. This quantitative RNA-Seq methodology has been used for expression analysis and splice isoform discovery and to confirm or extend reference gene models. Also in this issue, another paper reports application of the ABI SOLiD technology to sequence the transcriptome in mouse embryonic stem cells.

Nature Methods - **5**, 585 - 587 (2008)
doi:10.1038/nmeth0708-585

The beginning of the end for microarrays?

Jay Shendure

Jay Shendure is in the Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. shendure@u.washington.edu

**Two complementary app
successfully tackled the
once revealing unpreced**

Published online 15 October 2008 | *Nature* **455**, 847 (2008) |
doi:10.1038/455847a

News

The death of microarrays?

High-throughput gene sequencing seems to be stealing a march on microarrays. Heidi Ledford looks at a genome technology facing intense competition.

Heidi Ledford

Stem cell transcriptome profiling via massive-scale mRNA sequencing pp613 - 619
Nicole Cloonan, Alistair R R Forrest, Gabriel Kolle, Brooke B A Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, Alan J Robertson, Andrew C Perkins, Stephen J Bruce, Clarence C Lee, Swati S Ranade, Heather E Peckham, Jonathan M Manning, Kevin J McKernan & Sean M Grimmond

Published online: 30 May 2008 | doi:10.1038/nmeth.1223

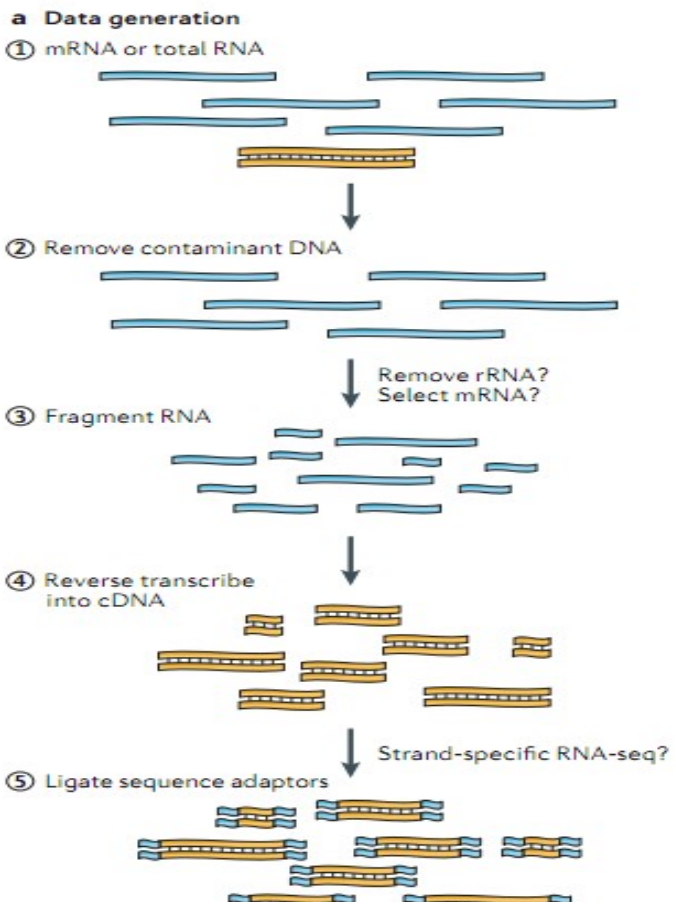
Application of next-generation sequencing using the ABI SOLiD technology to mammalian transcriptome analysis enabled a survey of the content, the complexity and the developmental dynamics of the embryonic stem cell transcriptome in the mouse. Also in this issue, Mortazavi *et al.* report Illumina technology–based RNA-Seq analysis of the mouse transcriptome in three different tissues.

Mapping and quantifying mammalian transcriptomes by RNA-Seq pp621 - 628
Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer & Barbara Wold

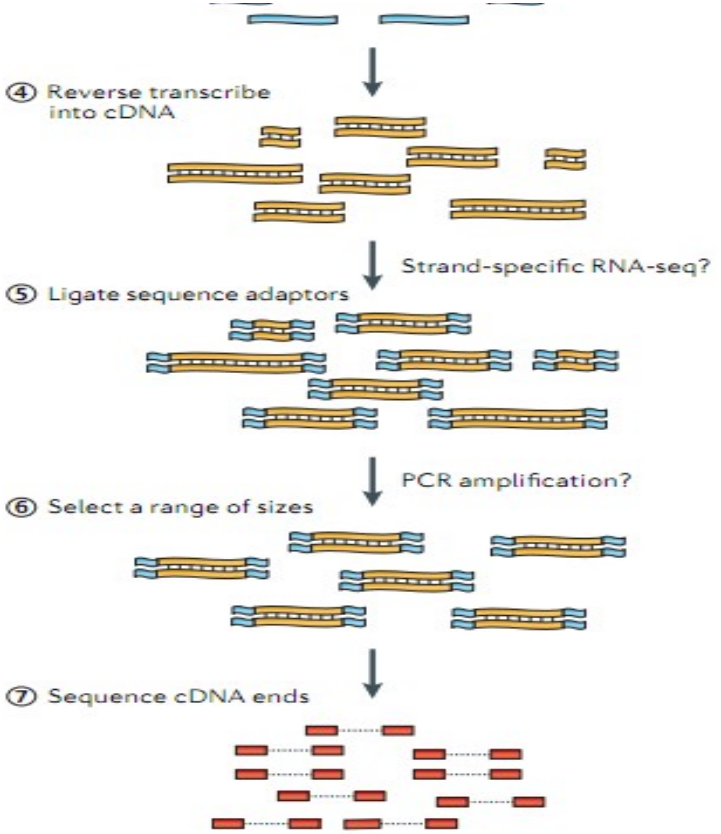
Published online: 30 May 2008 | doi:10.1038/nmeth.1226

The mouse transcriptome in three tissue types has been analyzed using Illumina next-generation sequencing technology. This quantitative RNA-Seq methodology has been used for expression analysis and splice isoform discovery and to confirm or extend reference gene models. Also in this issue, another paper reports application of the ABI SOLiD technology to sequence the transcriptome in mouse embryonic stem cells.

A typical RNA-Seq experiment



- Library construction



Nature Reviews Genetics **12**, 671-682 (October 2011) | doi:10.1038/nrg3068

ARTICLE SERIES: Study designs

Next-generation transcriptome assembly

Jeffrey A. Martin[1] & Zhong Wang[1] About the authors

Protocol variations

- Fragmentation methods
  - RNA: nebulization, magnesium-catalyzed hydrolysis, enzymatic clivage (RNAse III)
  - cDNA: sonication, Dnase I treatment
- Depletion of highly abundant transcripts
  - Ribosomal RNA (rRNA)
    - Positive selection of mRNA . Poly(A) selection.
    - Negative selection. (RiboMinus™)
      - Select also pre-messenger
- Strand specificity
  - Most RNA sequencing is not strand-specific
- Single-end or Paired-end sequencing

http://www.bioconductor.org/help/course-materials/2009/EMBLJune09/Talks/RNAseq-Paul.pdf

Single and paired-end sequencing

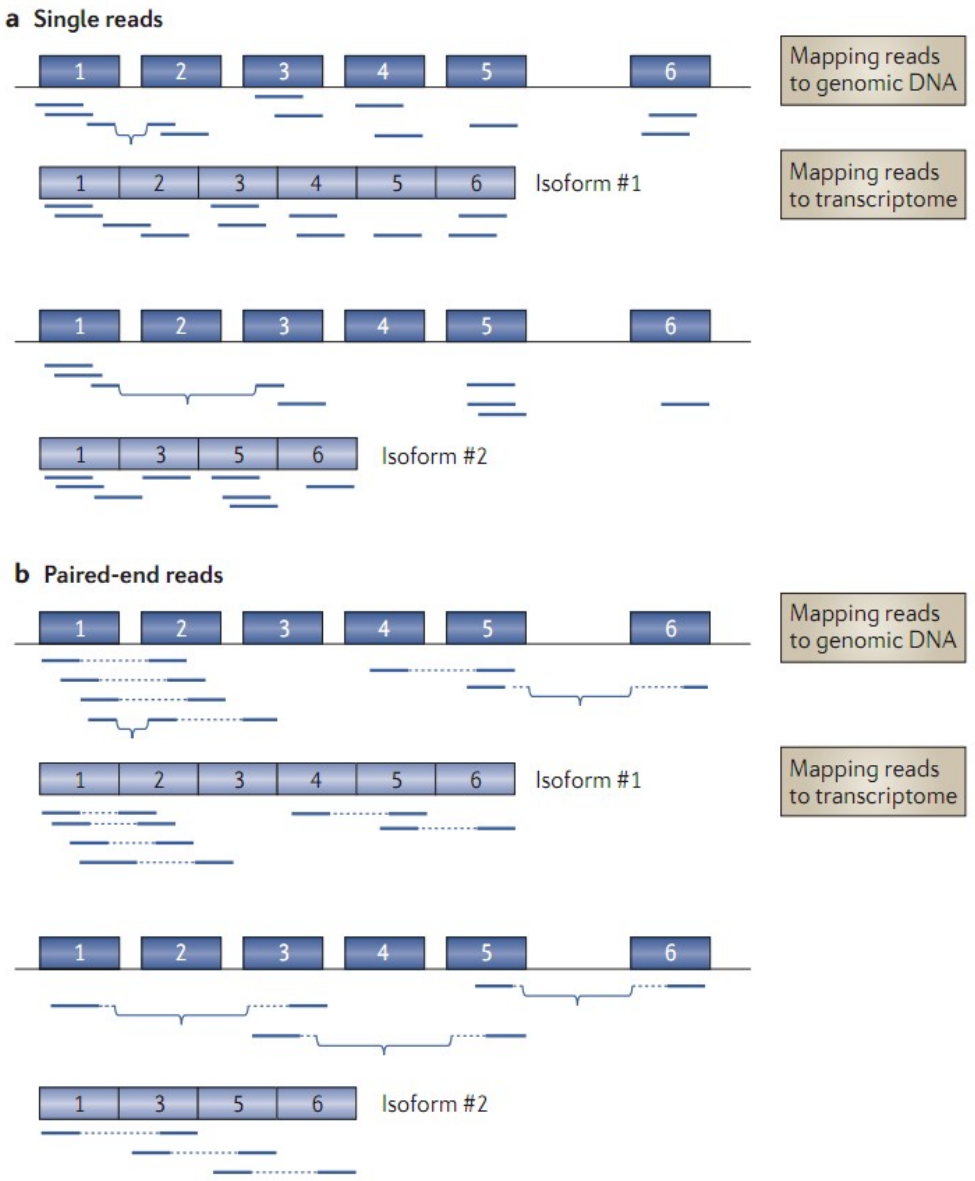
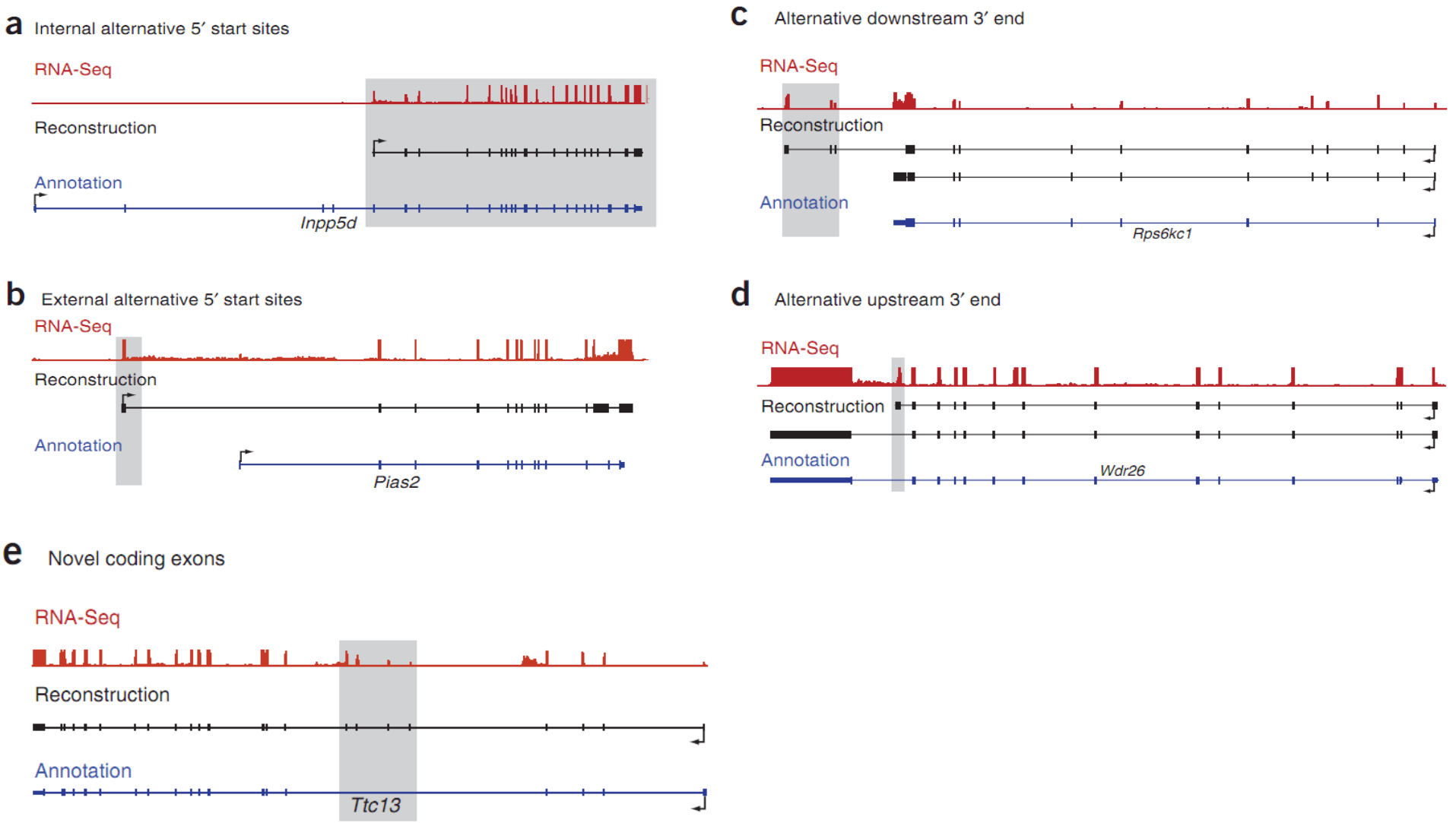


Figure 1 | **RNA-seq for detection of alternative splicing events. a** | Sequence reads are mapped to genomic DNA or to a transcriptome reference to detect alternative isoforms of an RNA transcript. Mapping is based simply on read counts to each exon and reads that span the exonic boundaries. One infers the absence of the genomic exon in the transcript by virtue of no reads mapping to the genomic location. **b** | Paired sequence reads provide additional information about exonic splicing events, as demonstrated by matching the first read in one exon and placing the second read in the downstream exon, creating a map of the transcript structure.

Microarrays versus RNA-seq

- RNA-seq
  - Counting
  - Absolute abundance of transcripts
  - All transcripts are present and can be analyzed
    - mRNA / ncRNA (snoRNA, linc/lncRNA, eRNA,miRNA,...)
  - Several types of analyses
    - Gene discovery
    - Gene structure (new transcript models)
    - Differential expression
      - Gene/isoforms
    - Allele specific gene expression
    - Detection of fusions and other structural variations
    - ...

Microarrays versus RNA-seq



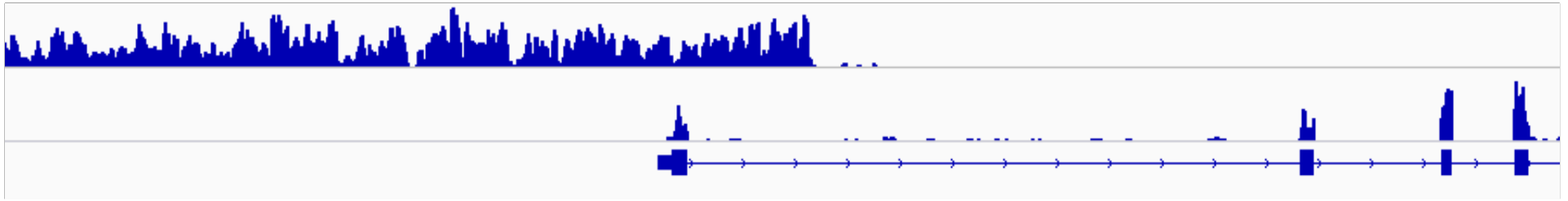
Nat Biotechnol. 2010 May;28(5):503-10. Epub 2010 May 2.

Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A.

Microarrays versus RNA-seq

- Strand-specific analysis



Microarrays versus RNA-seq

- Microarrays
  - Indirect record of expression level (complementary probes)
  - Relative abundance
  - Cross-hybridization
  - Content limited (can only show you what you're already looking for)

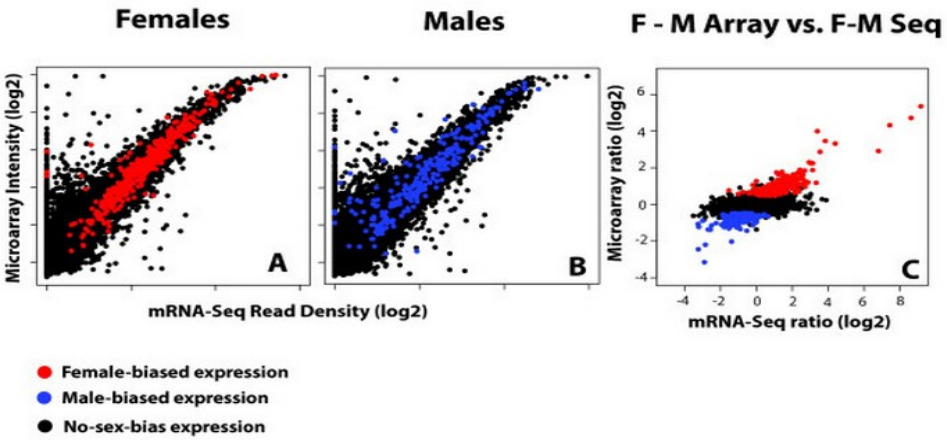
Some RNA-Seq drawbacks

- Current disadvantages
  - More expensive than standard expression arrays
  - More time consuming than any microarray technology
  - Some (lots of) data analysis issues
    - Mapping reads to splice junctions
    - Computing accurate transcript models
    - Contribution of high-abundance RNAs (eg ribosomal) could dilute the remaining transcript population; sequencing depth is important

http://www.bioconductor.org/help/course-materials/2009/EMBLJune09/Talks/RNAseq-Paul.pdf

Do arrays and RNA-Seq tell a consistent story?

- Do arrays and RNA-Seq tell a consistent story?
  - ”The relationship is not quite linear ... but the vast majority of the expression values are similar between the methods. Scatter increases at low expression ... as background correction methods for arrays are complicated when signal levels approach noise levels. Similarly, RNA-Seq is a sampling method and stochastic events become a source of error in the quantification of rare transcripts ”
  - ”Given the substantial agreement between the two methods, the array data in the literature should be durable”



Review

Highly accessed Open Access

Microarrays, deep sequencing and the true measure of the transcriptome

John H Malone and Brian Oliver
Laboratory of Cellular and Developmental Biology, National Institute of Digestive, Diabetes, and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
author email corresponding author email

BMC Biology 2011, **9**:34 doi:10.1186/1741-7007-9-34

Comparison of array and RNA-Seq data for measuring differential gene expression in the heads of male and female D. pseudoobscura

Sequencing technologies

- RNA-Seq has been performed using SOLiD, Illumina and 454 platform
  - However, RNA-Seq needs high dynamic range
    - SOLiD and Illumina sequencers
    - Need PCR amplification
      - Low coverage of GC rich transcrits
  - Future ?
    - Helicos
      - No PCR amplification
      - Direct RNA sequencing seem possible
      - High error rates
    - Pacific Biosciences

Sequencing technologies

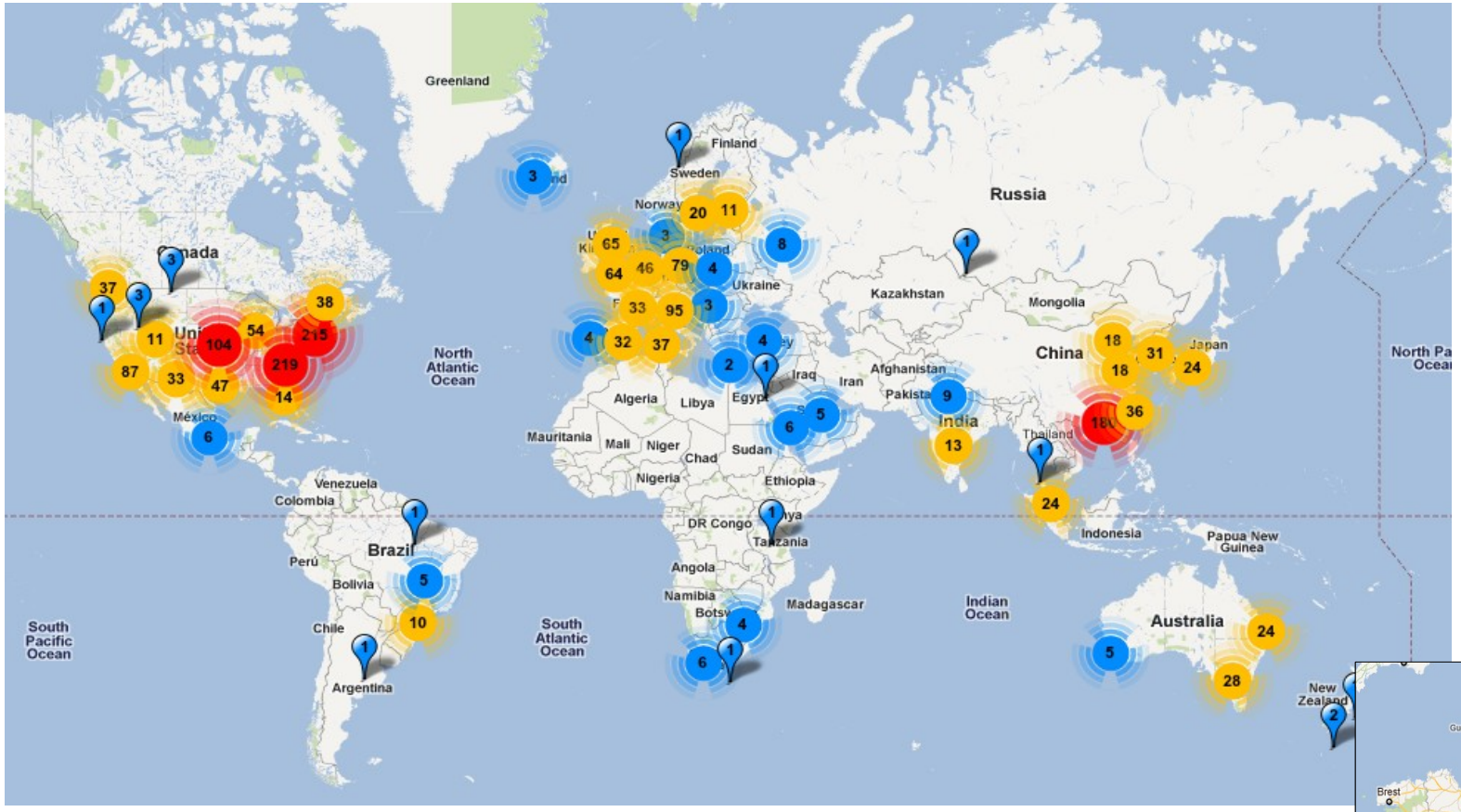
Platform	Library/template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

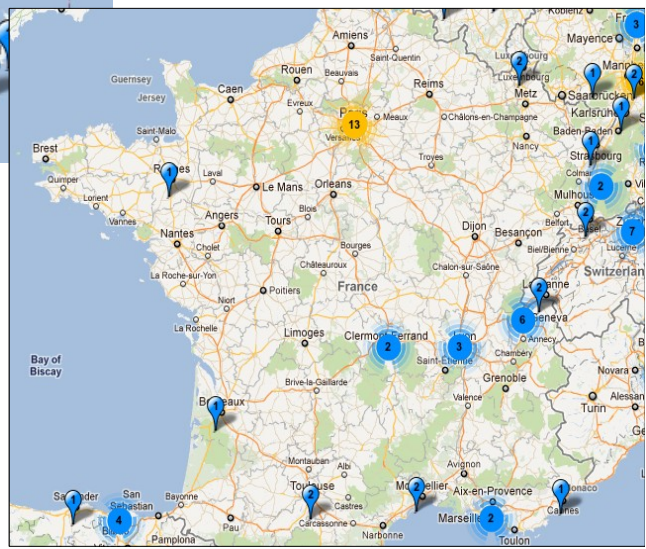
*Average read-lengths.

‡Fragment run.

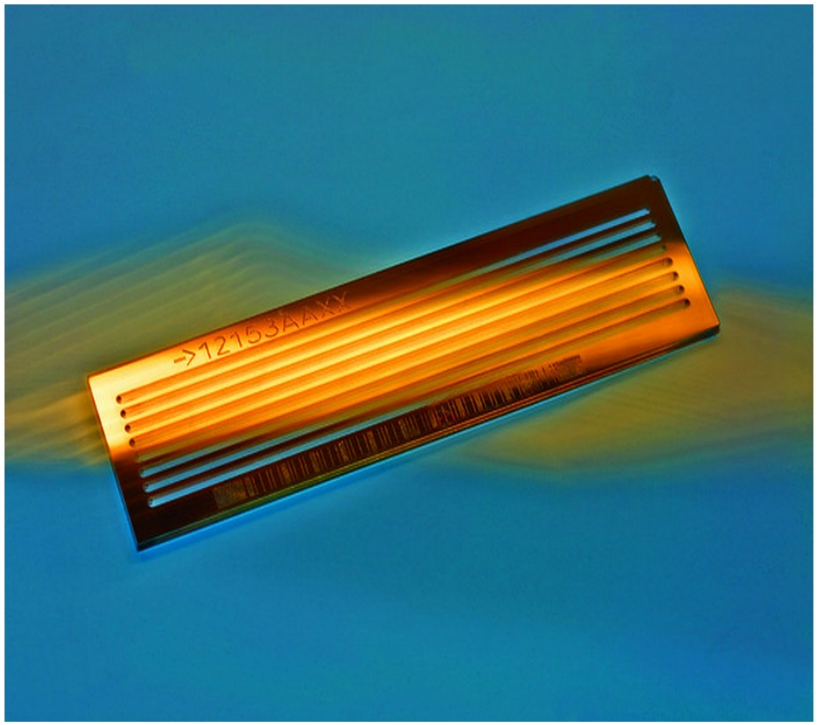
§Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

World Map of High-throughput Sequencers

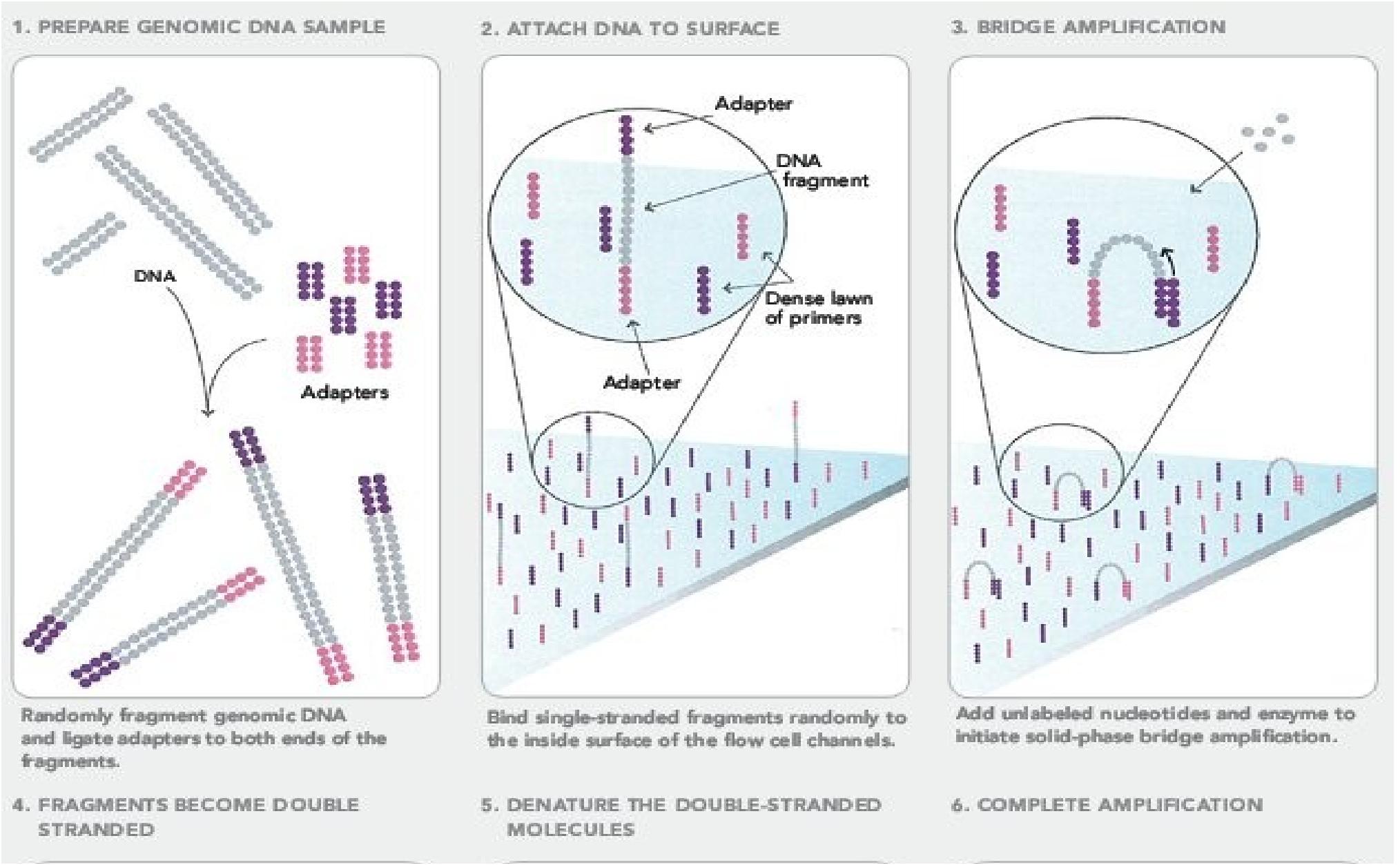




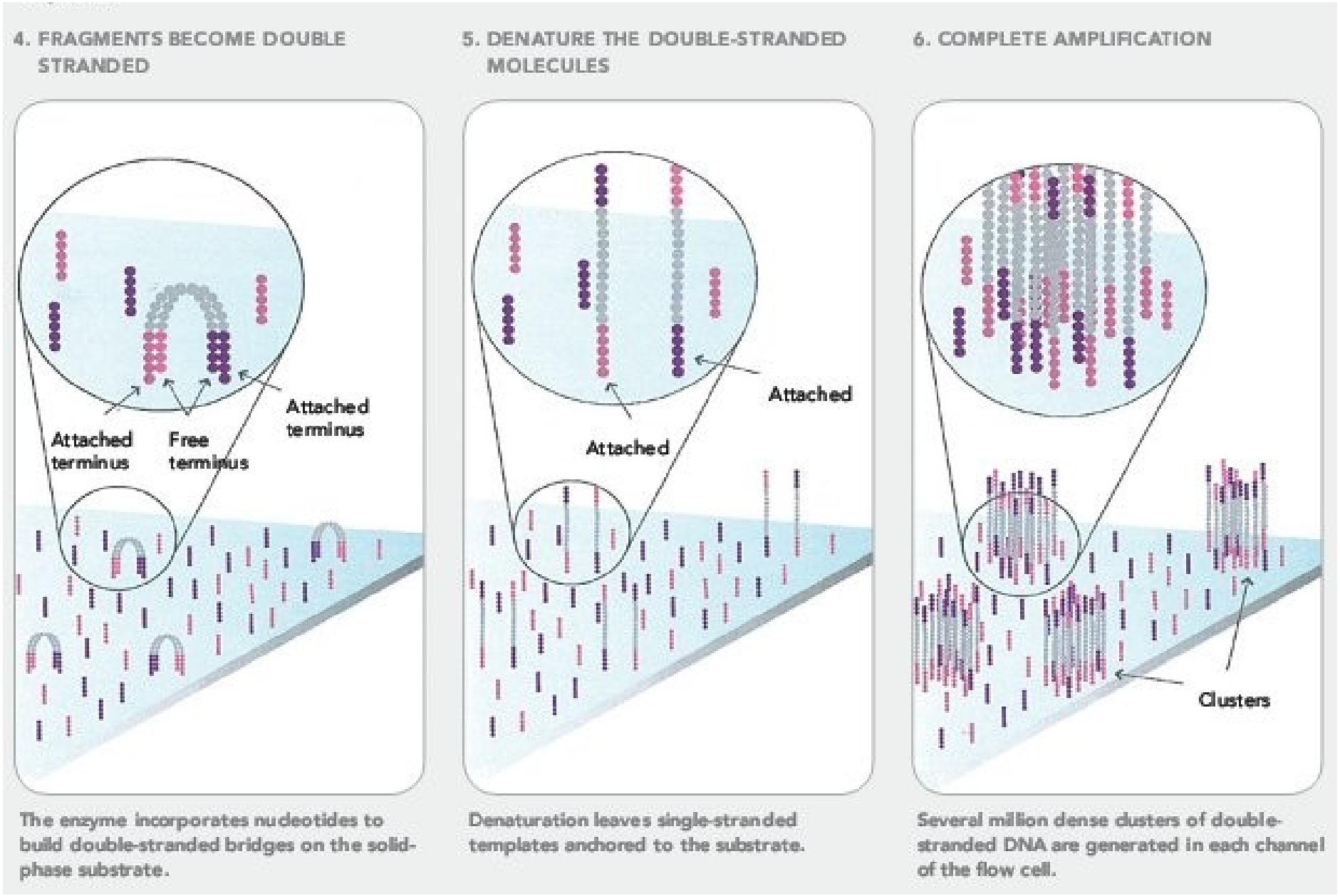
Illumina



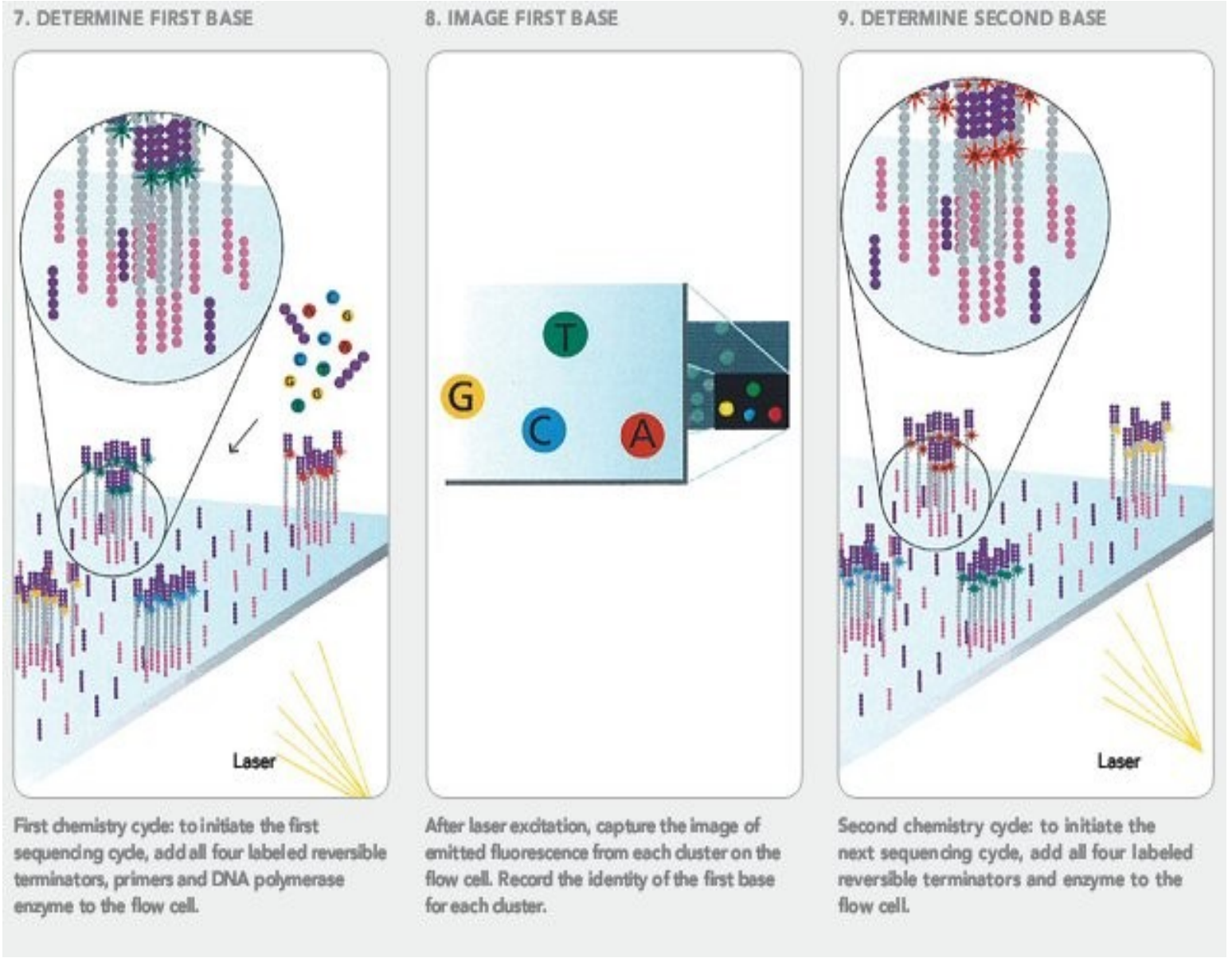
Illumina Genome Analyzer (I)



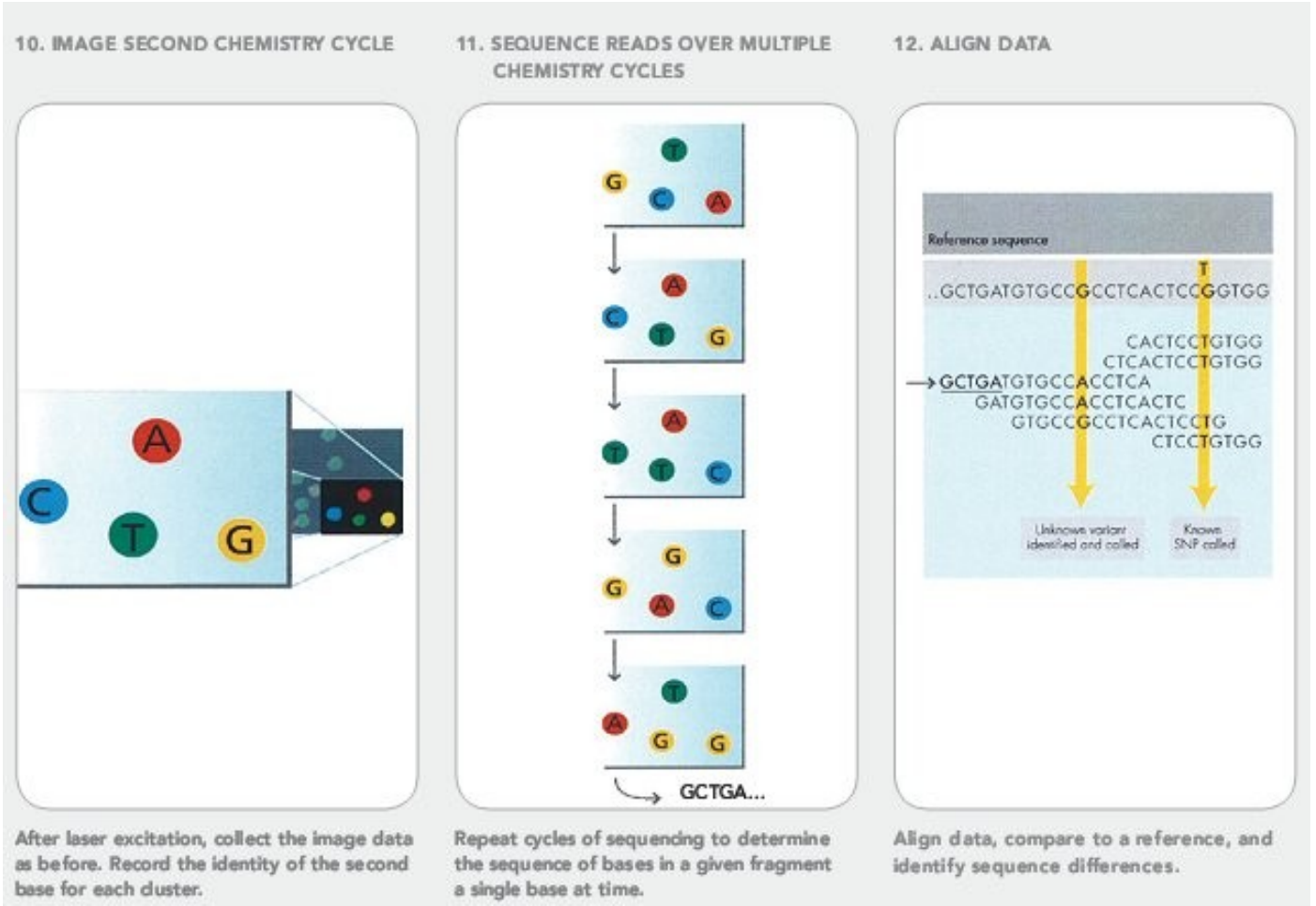
Illumina Genome Analyzer (II)



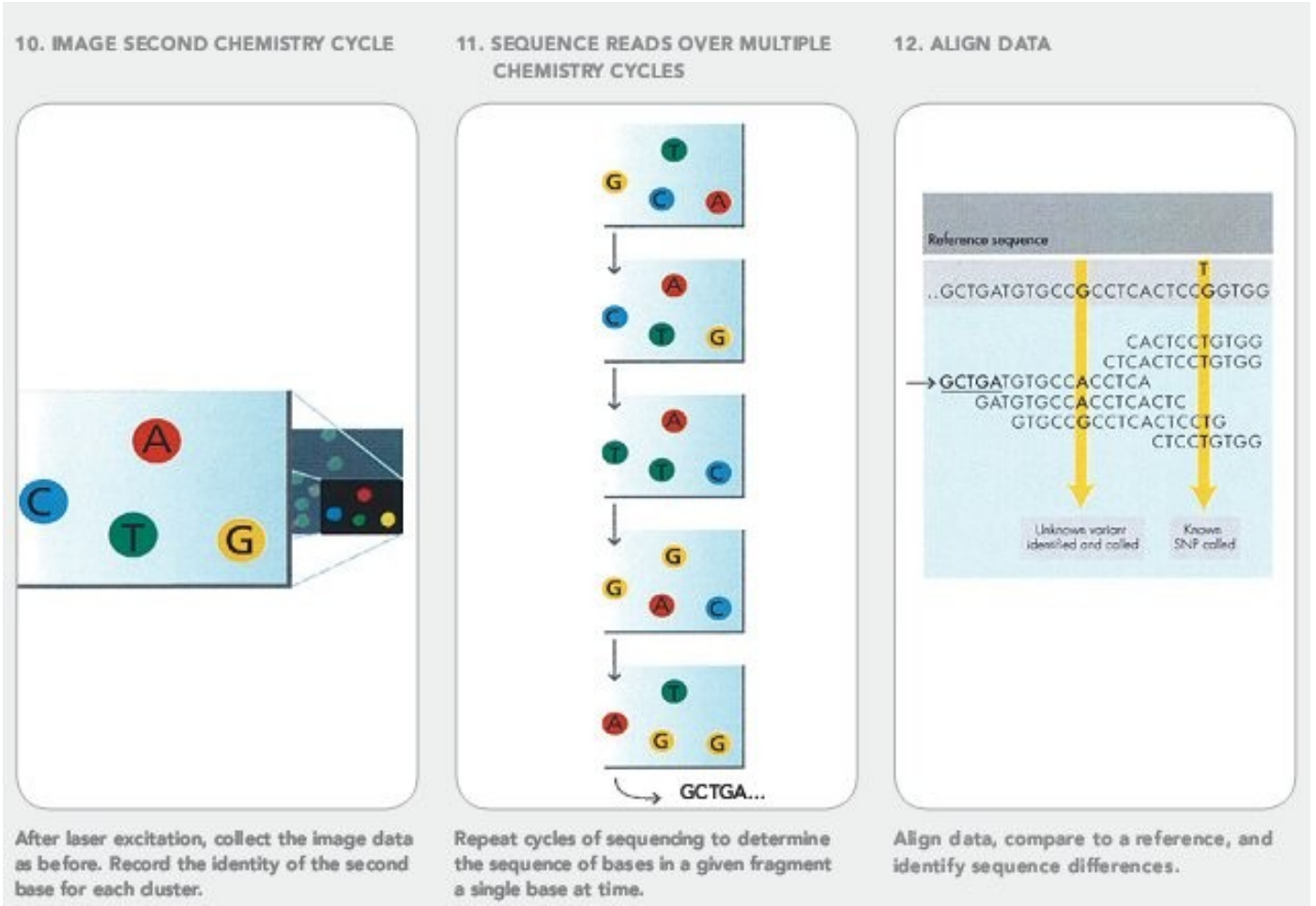
Illumina Genome Analyzer (III)



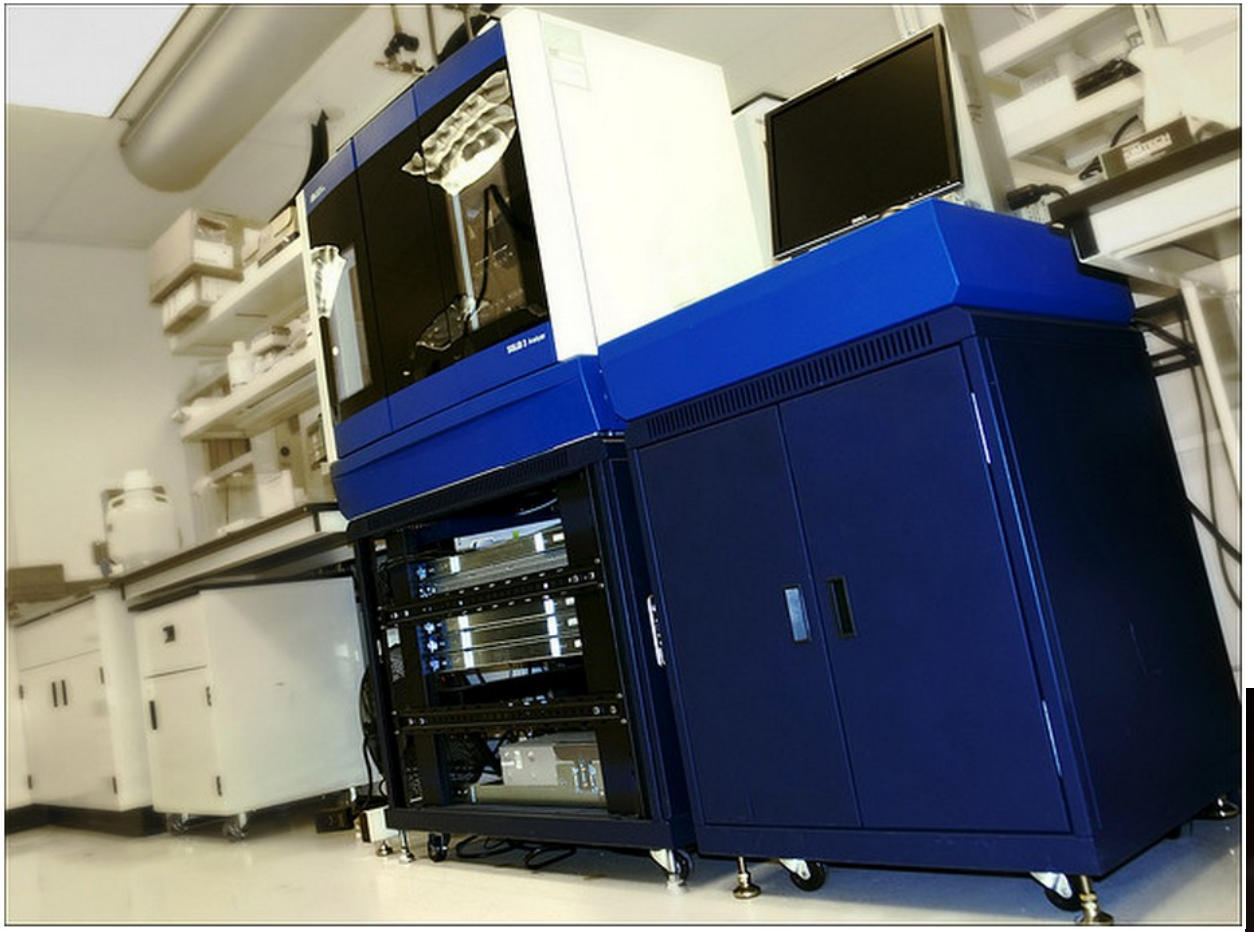
Illumina Genome Analyzer

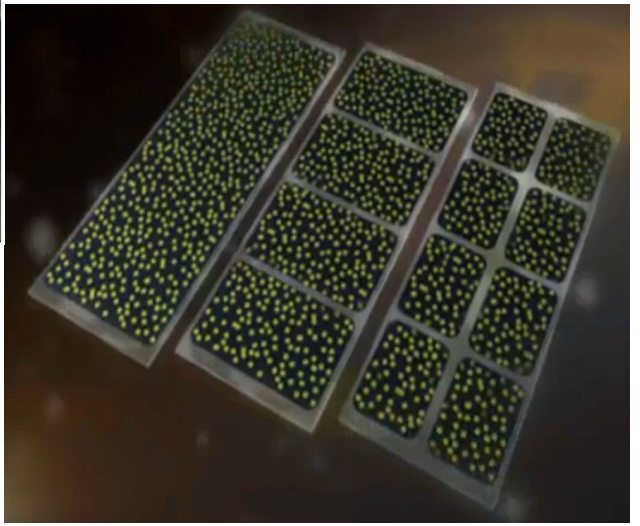


Illumina Genome Analyzer



ABI Next Gen Sequencing: SOLiD





http://www.flickr.com/photos/ricardipus/4452215186/

ABI Next Gen Sequencing: SOLiD

- Use emPCR on magnetic beads
- Sequencing by ligation using fluorescent probes

Sequencing by ligation (SOLiD)

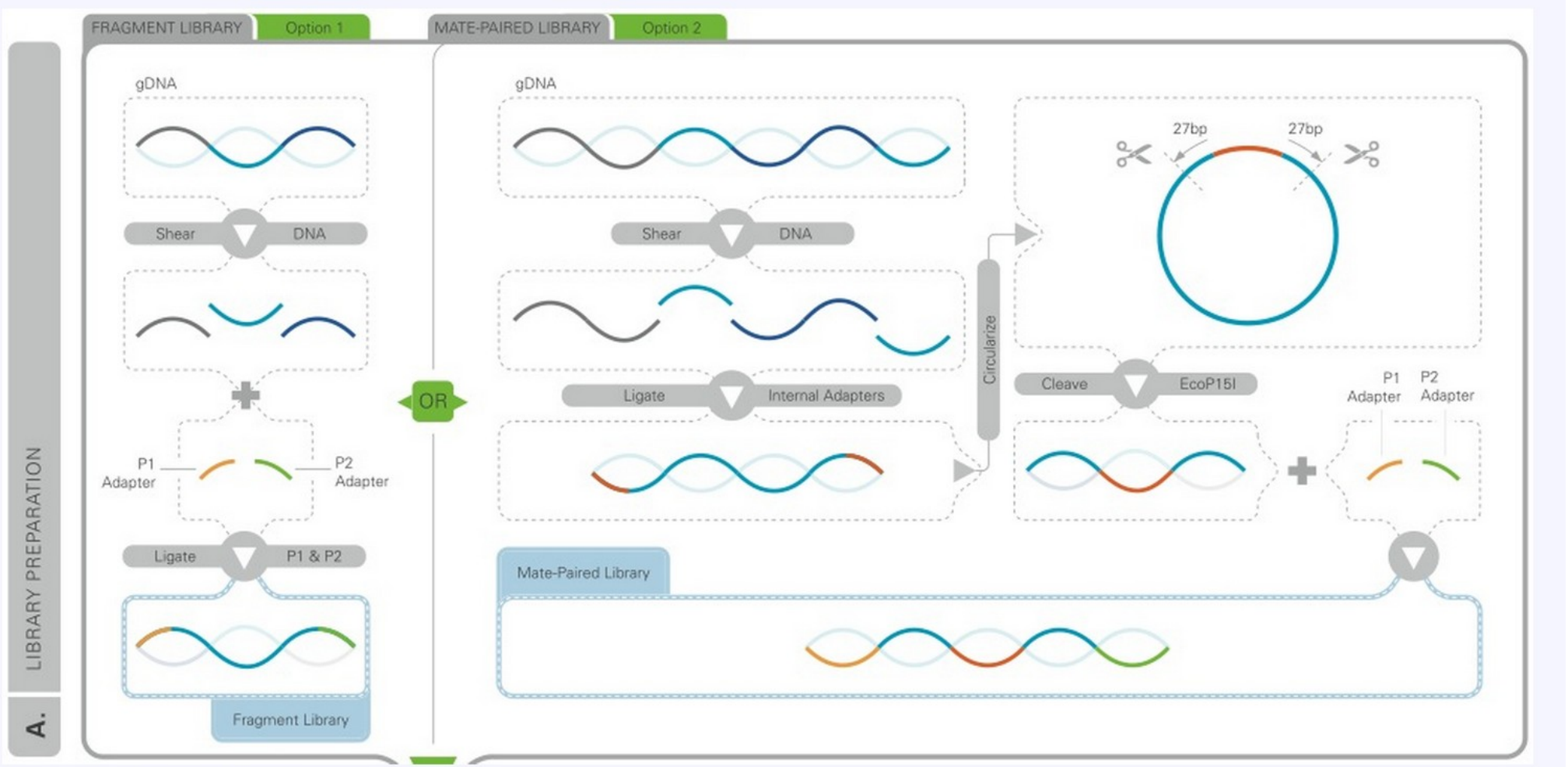


Figure 1. Library generation schematic.

Sequencing by ligation (SOLiD)

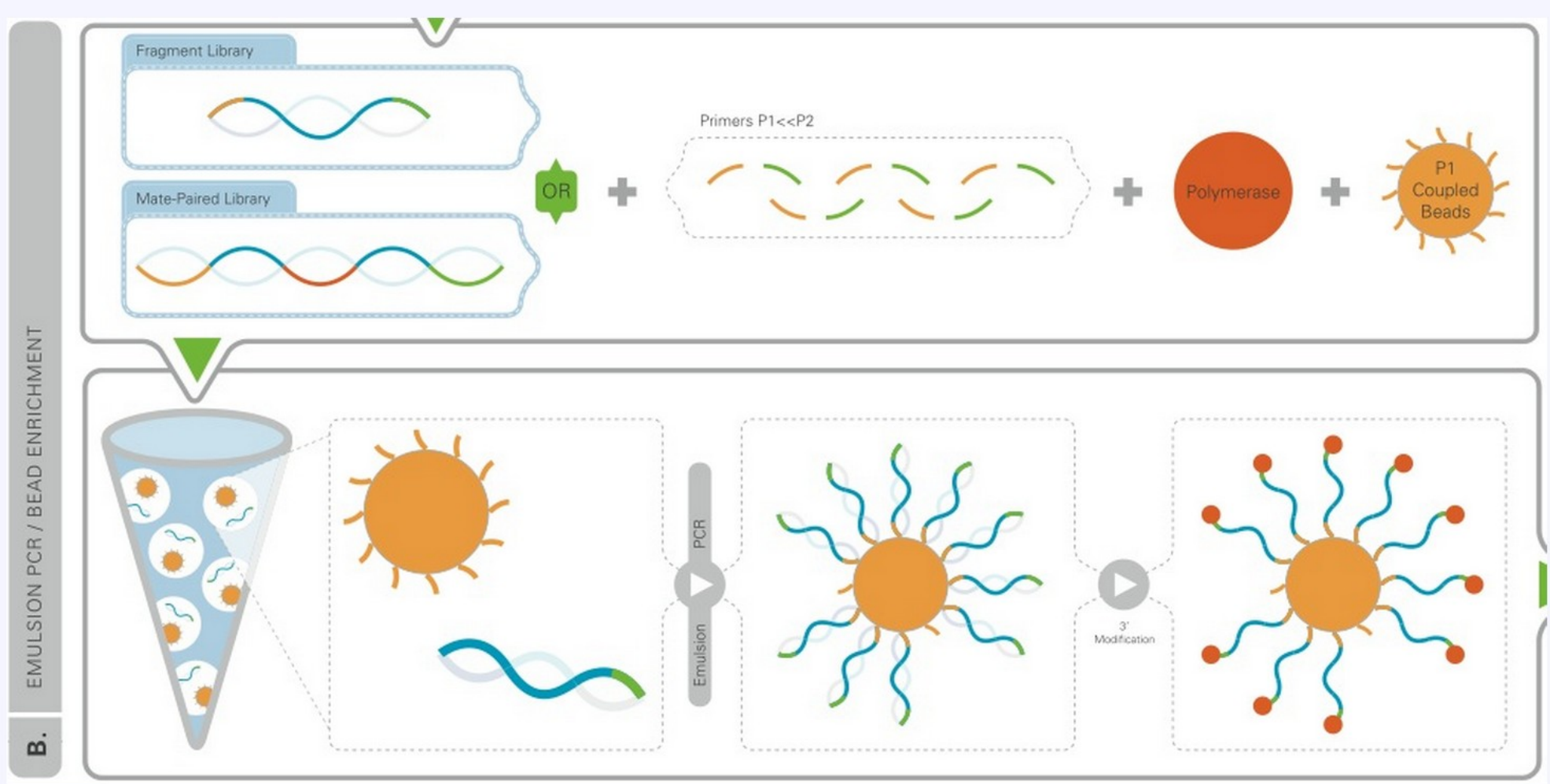


Figure 2. Clonal bead library generation via emulsion PCR.

Sequencing by ligation (SOLiD)

Each bead is then attached to the surface of a flow cell via 3′ modifications to the DNA strands.

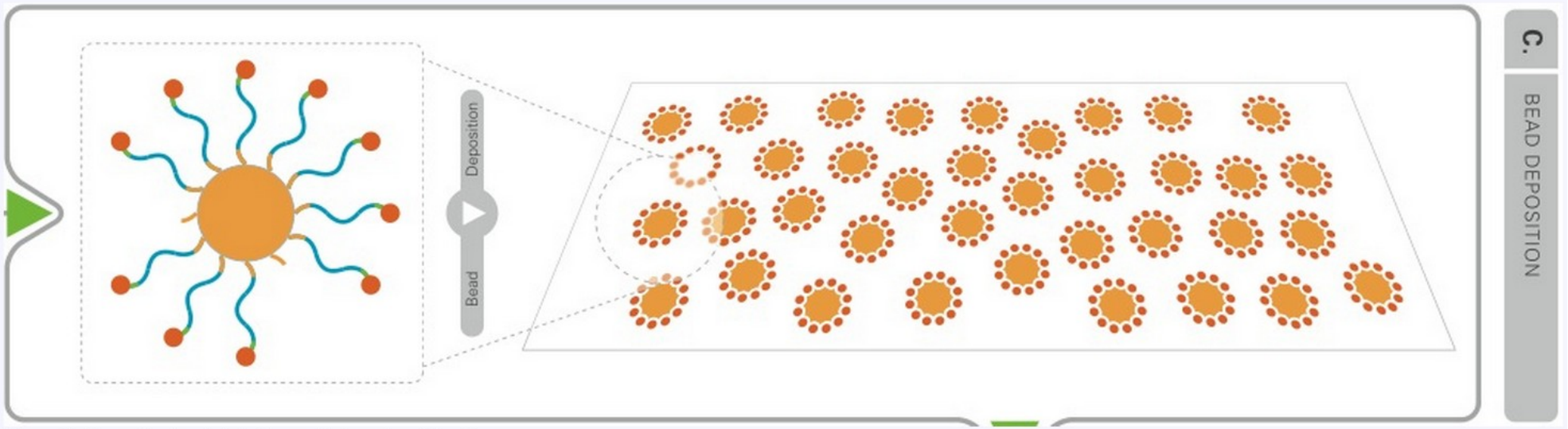
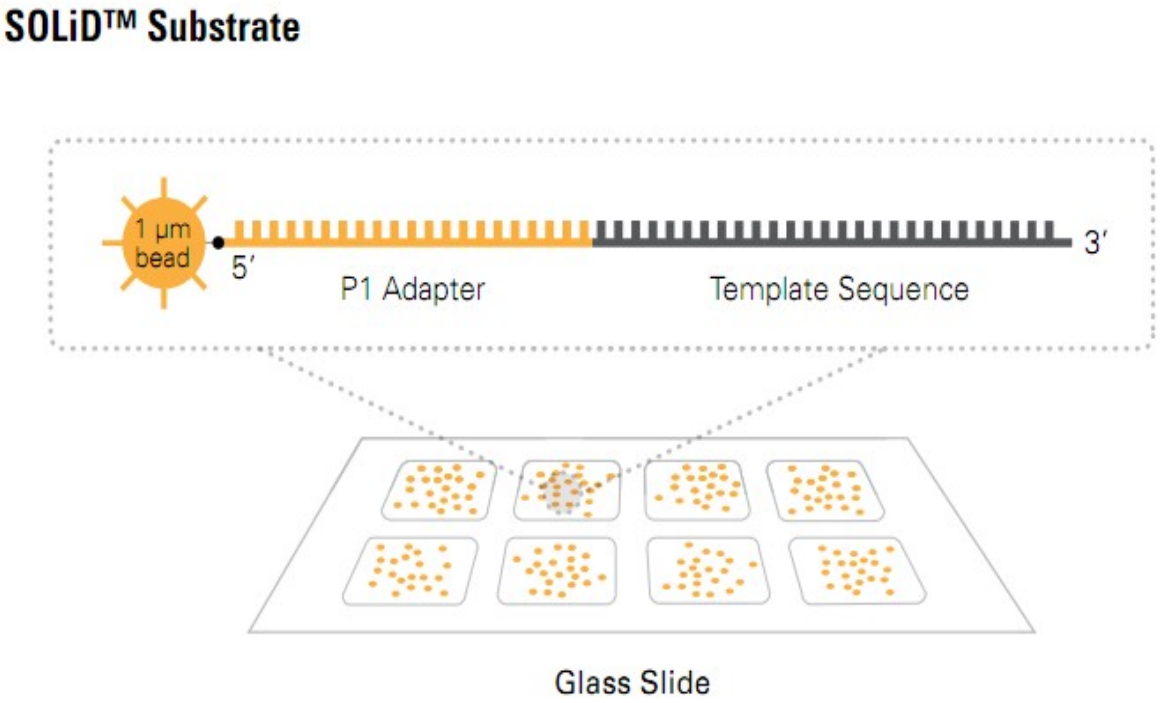
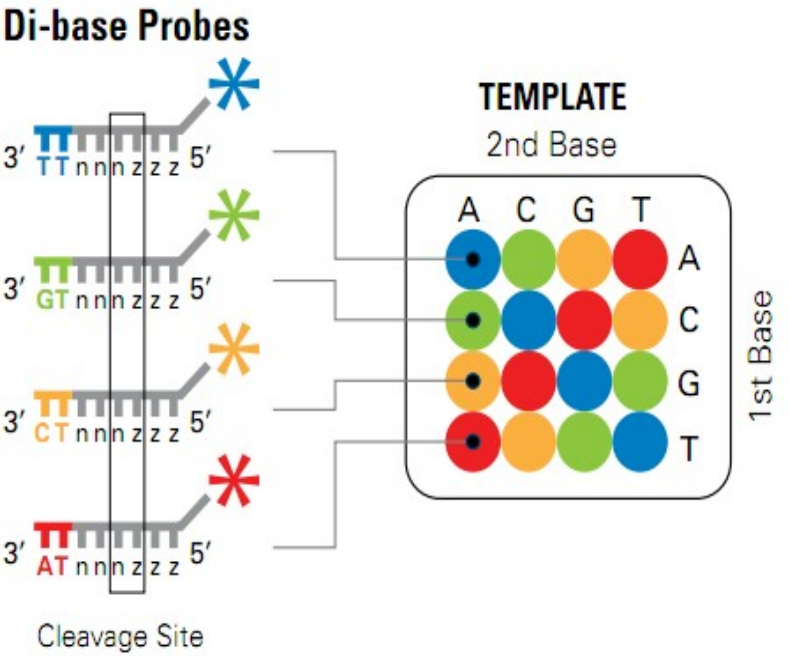


Figure 3. Depositing beads into flow cell via end modifications.

Sequencing by ligation (SOLiD)





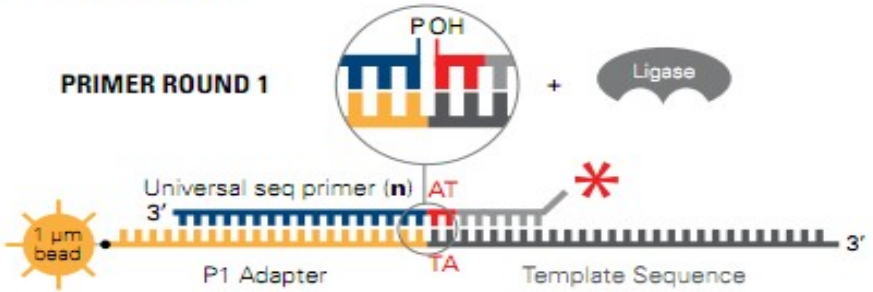
We have a flow cell (basically a microscope slide that can be serially exposed to any liquids desired) whose surface is coated with thousands of beads each containing a single genomic DNA species, with unique adapters on either end. Each microbead can be considered a separate sequencing reaction which is monitored simultaneously via sequential digital imaging. Up to this point all next-gen sequencing technologies are very similar, this is where ABI/SOLiD diverges dramatically (see figure 4)

code	0	1	2	3
dye	FAM	Cy3	TXR	Cy5
	AA	AC	AG	AT
	CC	CA	GA	TA
	GG	GT	CT	CG
	TT	TG	TC	GC

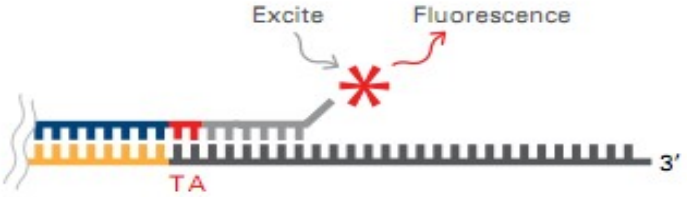
Figure 1: SOLiD™ System's 2 base Coding Scheme. The column under code *i* lists the corresponding dye and the di-bases (adjacent nucleotides) encoded by color *i*. For example, GT is labeled with Cy3 and coded as "1".

Sequencing by ligation (SOLiD)

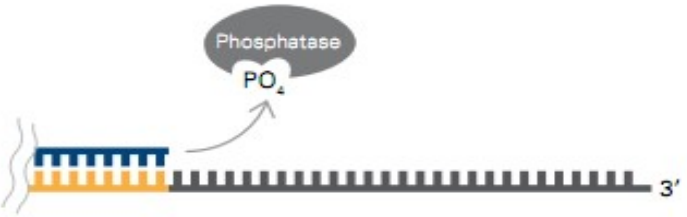
1. Prime and Ligate



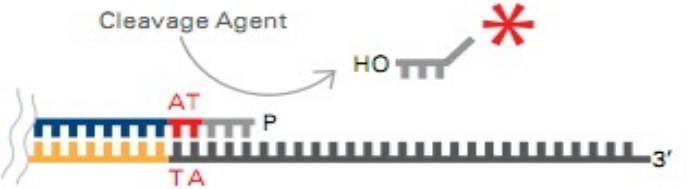
2. Image



3. Cap Unextended Strands



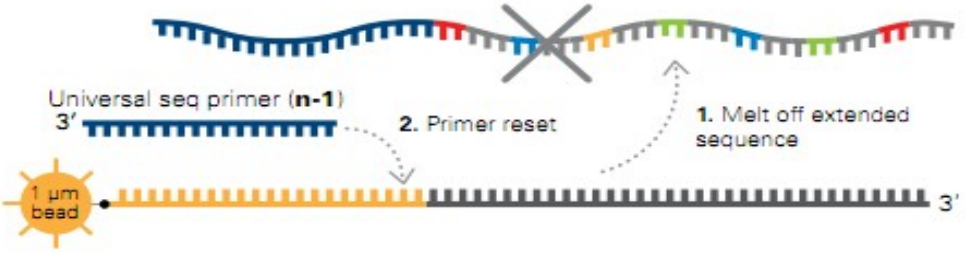
4. Cleave off Fluor



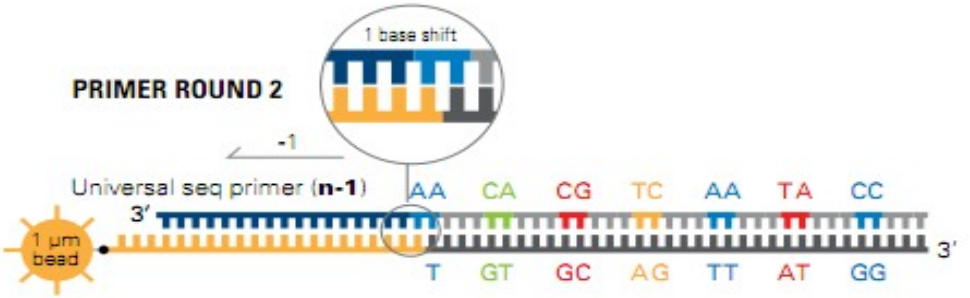
8. Repeat Reset with , n-2, n-3, n-4 primers

5. Repeat steps 1-4 to Extend Sequence

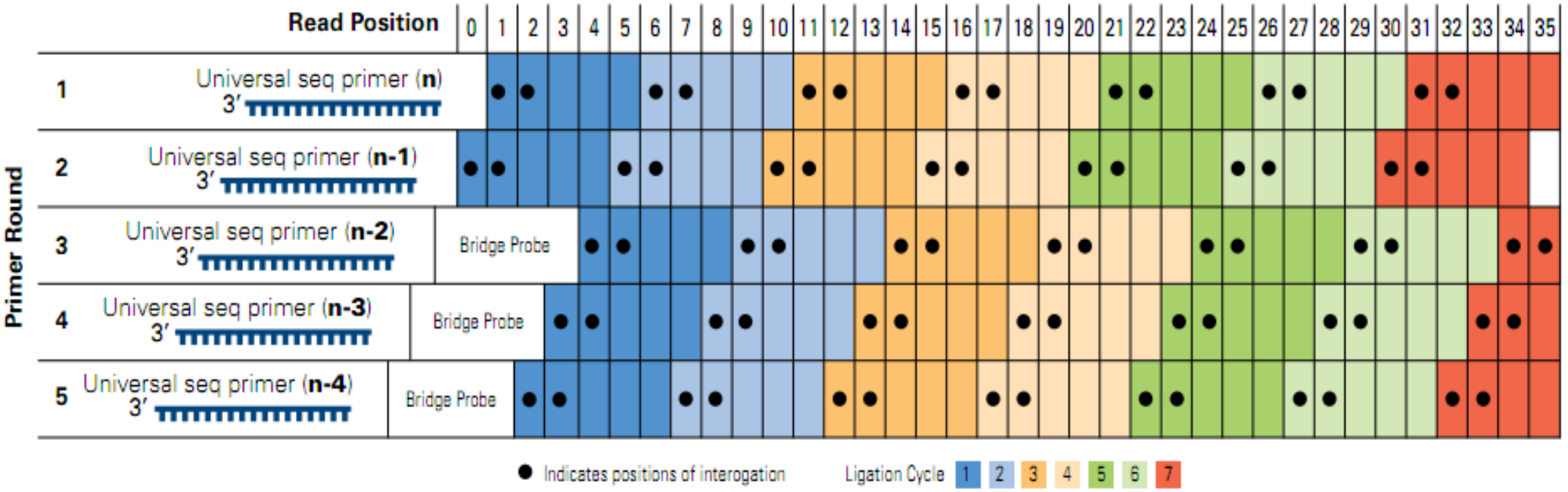
6. Primer Reset



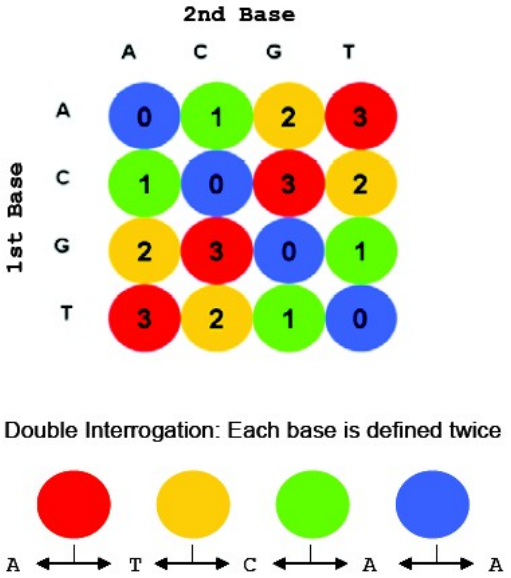
7. Repeat steps 1-5 with new primer

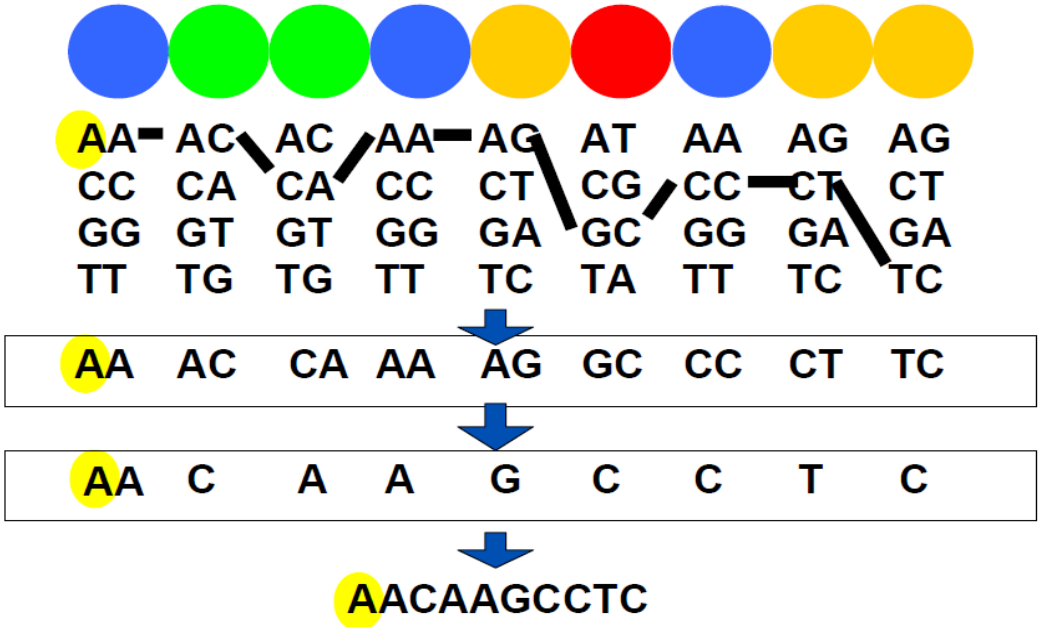


Sequencing by ligation (SOLiD)



Sequencing by ligation (SOLiD)

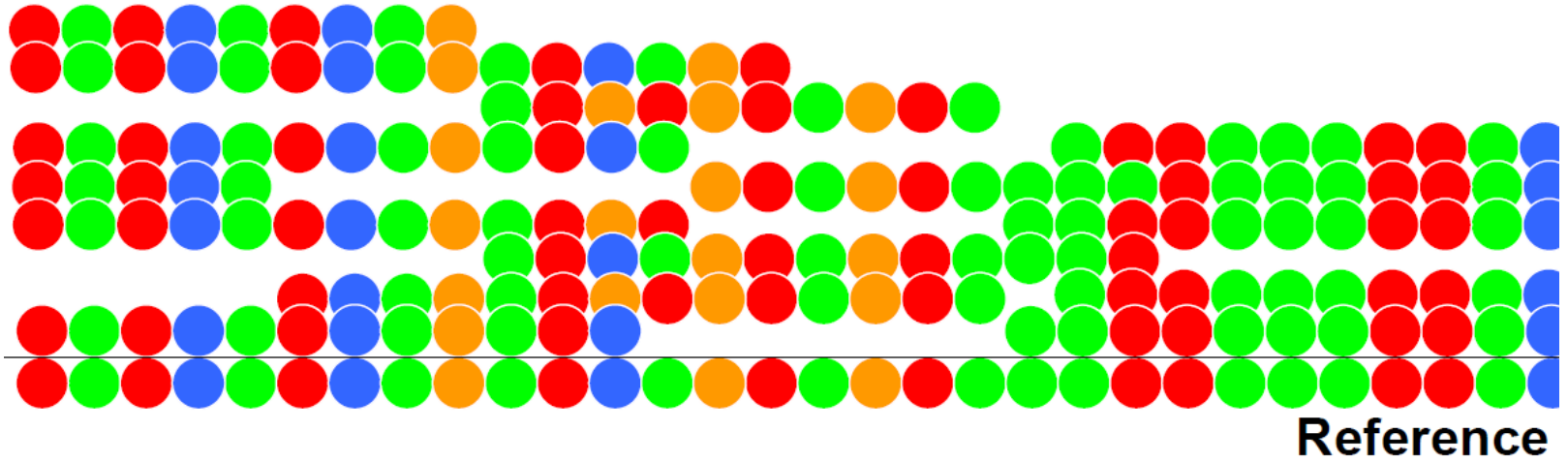


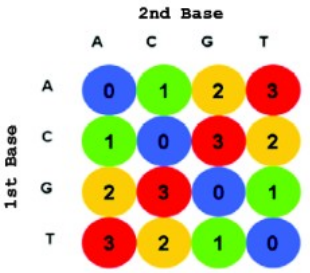


- According to colour code table four different dinucleotides may correspond to the same colour.
  - This ambiguity is resolved using primer n-1 round
    - The first ligation of this primer round analyses dinucleotide with one known nucleotide

Sequencing by ligation (SOLiD)

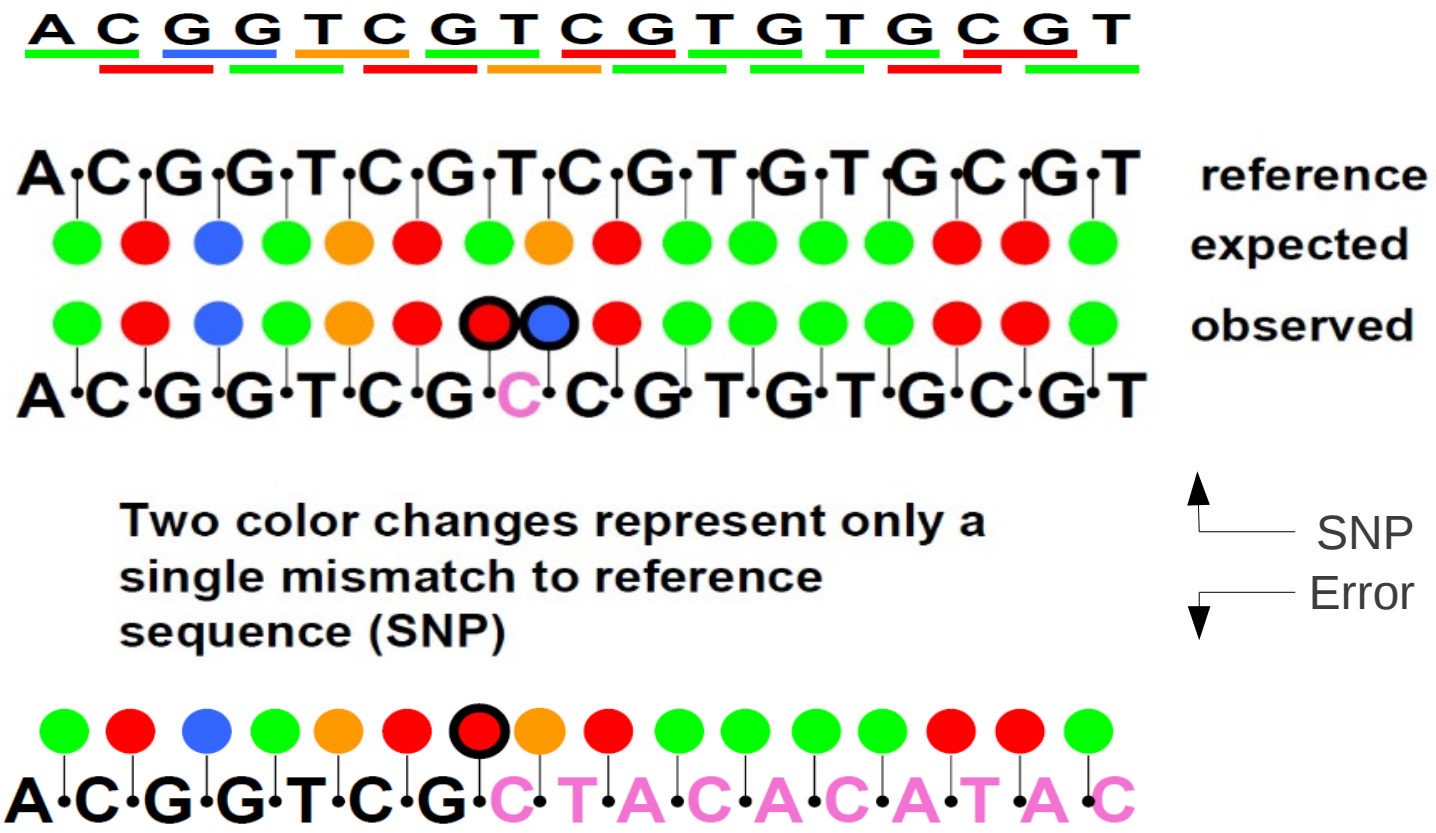
- Alignements are performed in color space
  - Alignments of nucleotide transitions

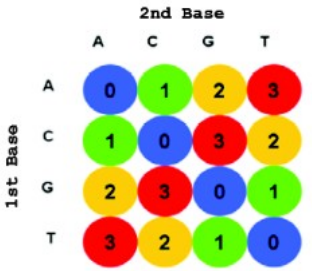




Advantages of 2 base pair encoding

- Double base interrogation eases the discrimination between system errors and true polymorphisms





Fastq file format

- Header
- Sequence
- + (optional header)
- Quality (default Sanger-style)

```
@QSEQ32.249996 HWUSI-EAS1691:3:1:17036:13000#0/1 PF=0 length=36
GGGGGTCATCATCATTTGATCTGGGAAAGGCTACTG
+
=.+5:<<<<>AA?0A>;A*A################
@QSEQ32.249997 HWUSI-EAS1691:3:1:17257:12994#0/1 PF=1 length=36
TGTACAACAACAACCTGAATGGCATACTGGTTGCTG
+
DDDD<BDBDB??BB*DD:D################
```

Solid output

- Read sequence in color (csfasta)

```
>1831_573_1004_F3
T000301333122121113000110213101322222
>1831_573_1567_F3
T033303222303221121310102211021221 13
```

- Quality scores (qual)

```
>1831_573_1004_F3
  4 29 34 34 32 32 24 24 20 17 10 34 29 20 34 13 30 34 22 24 11 28 19
17 34       17 24 17 25 34 7 24 14 12 22
>1831_573_1567_F3
  8 26 31 31 16 22 30 31 28 29 22 30 30 31 32 23 30 28 28 31 19 32 30
32 19   8   32 10 13 6 32 10 6 16 11
```

Solid output in fastq format

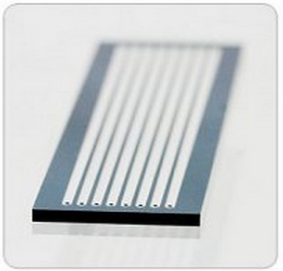
```
@1831_573_1004
T000301333122121113000110213101322222
+1831_573_1004
%>CCAA9952+C>5C.?C79,=42C292:C(9/-7
@1831_573_1004
T033303222303221121310102211021221l3
+1831_573_1004
);@@17?@=>7??@A8?==@4A?A4)A+.'A+'1,
```

Illumina sequence identifiers

- Sequences from the Illumina software use a systematic identifier:

```
@SRR038538.sra.2 HWI-EAS434:4:1:1:1701 length=36
NAATCGGAAATTTTATTTGTTCAGTACACCAAATAG
+SRR038538.sra.2 HWI-EAS434:4:1:1:1701 length=36
!0<<;:::<<<<<<<<<<<<<<;;;<<<<<<<<<;76
```

HWI-EAS434	Unique instrument name
4	Flowcell lane
1	Tile number within the flow cell
1	'x'-coordinate of the cluster within the tile
1701	'y'-coordinate
#0	Index number for a multiplexed sample (opt.)
/1	/1 or /2 for paired-end and maite-pair sequencing (opt.)



Sanger quality score

- Sanger quality score (Phred quality score): Measure the quality of each base call
  - Based on p, the probality of error (the probability that the corresponding base call is incorrect)
  - Qsanger= -10*log10(p)
  - p = 0.01 <=> Qsanger 20
- Quality score are in ASCII 33
- Note that SRA has adopted Sanger quality score although original fastq files may use different quality score (see: http://en.wikipedia.org/wiki/FASTQ_format)

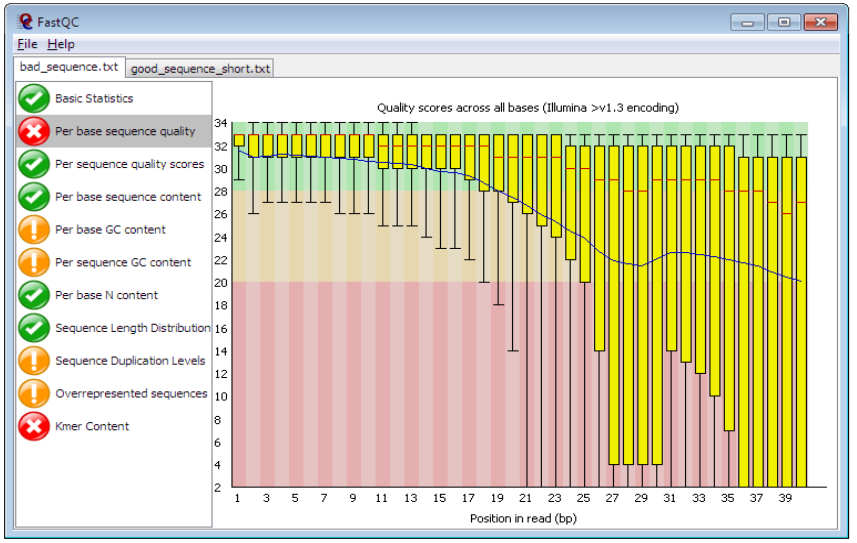
ASCII 33

- Storing PHRED scores as single characters gave a simple and space efficient encoding:
- Character ”!” means a quality of 0
- Range 0-40

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t

Quality control for high throughput sequence data

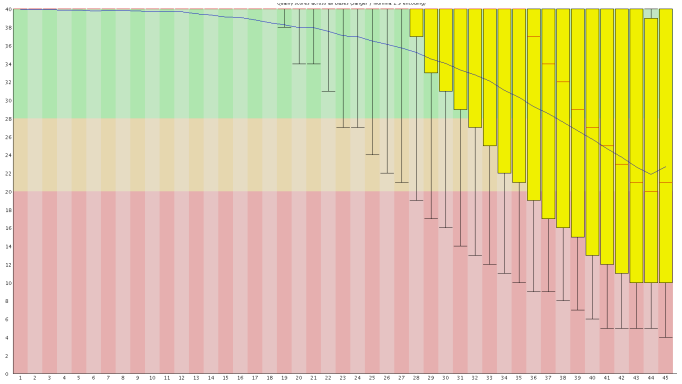
- FastQC
  - GUI / command line
  - http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc



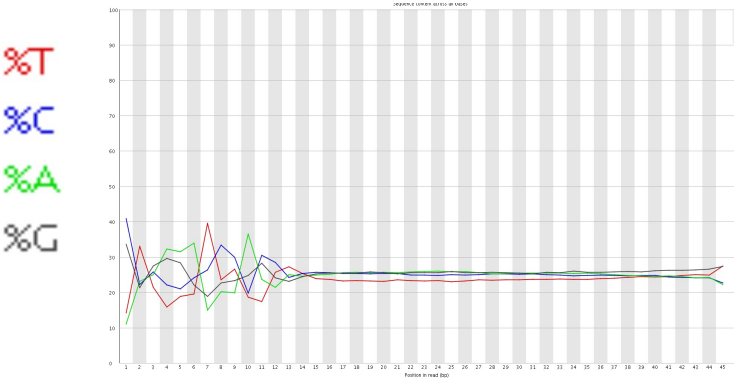
  - ShortRead
    - Bioconductor package

Quality control with FastQC

Quality

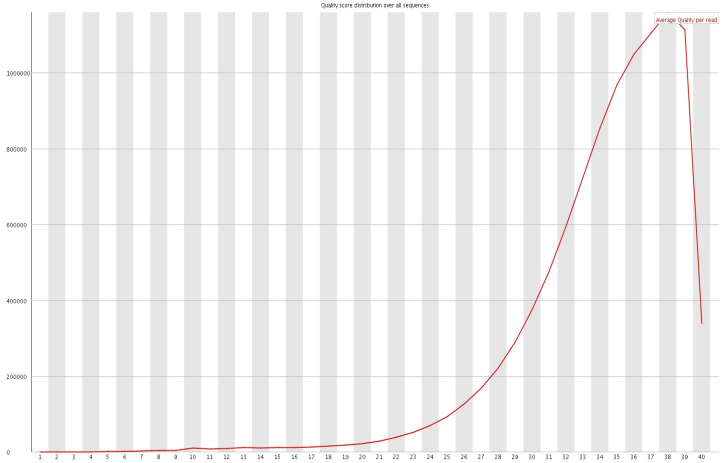


Position in read



Position in read

Nb Reads



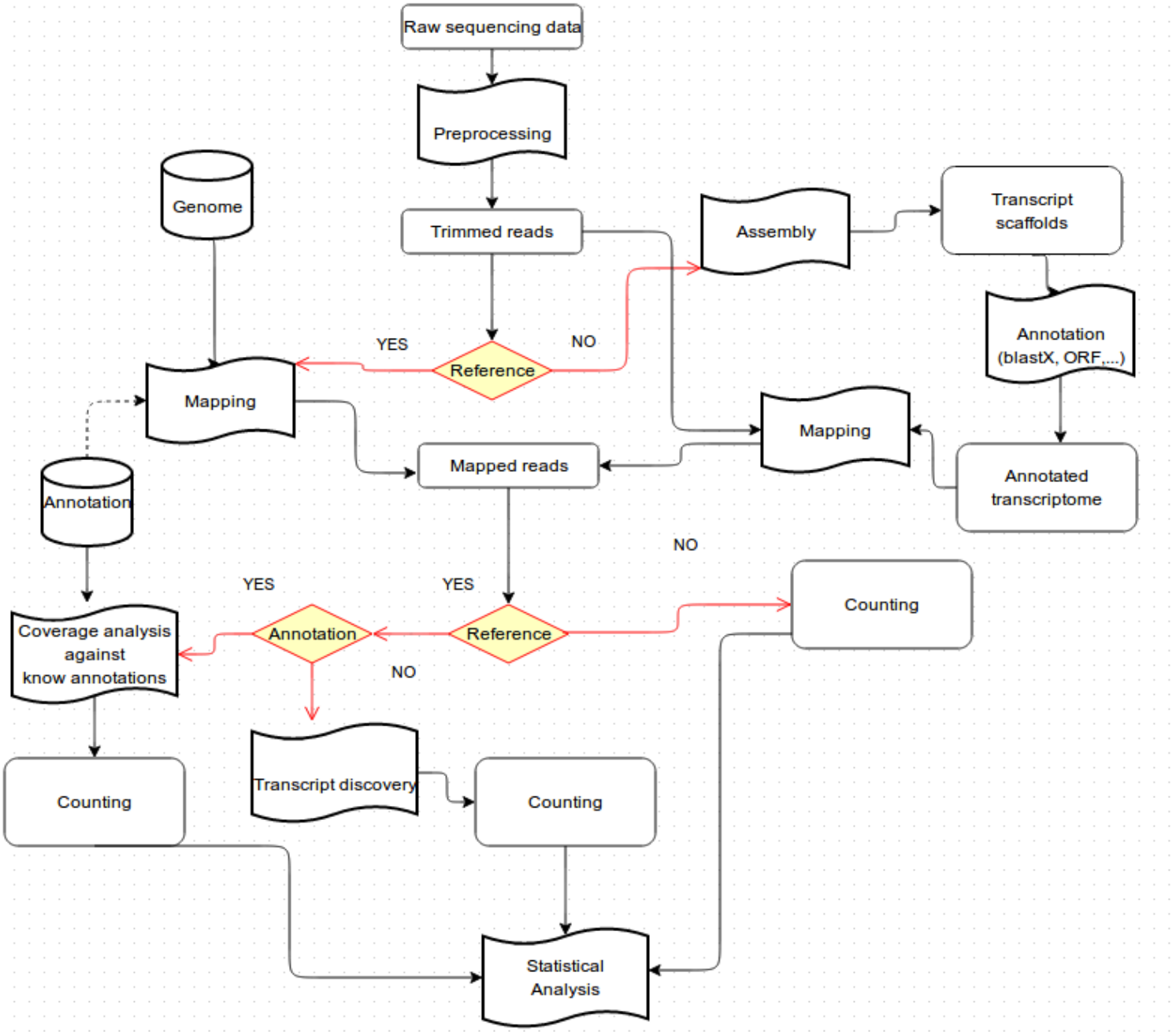
Mean Phred Score

Look also at over-represented sequences

Trimming

- Essential step (at least when using bowtie)
  - Almost mandatory when using tophat
- FASTX-Toolkit
- Sickle
  - Window-based trimming (unpublished)
- ShortRead
  - Bioconductor package
- csfasta_quality_filter.pl
  - SOLiD
    - Mean quality
    - Continuous run of bad colors at the end of the read

Reference mapping and de novo assembly



Mapping reads to genome: general softwares

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

[a]Work well for Sanger and 454 reads, allowing gaps and clipping.

[b]Paired end mapping.

[c]Make use of base quality in alignment.dBWA trims the primer base and the first color for a color read.

[e]Long-read alignment implemented in the BWA-SW module. fMAQ only does gapped alignment for Illumina paired-end reads.

[g]Free executable for non-profit projects only.

Brief Bioinform. 2010 Sep;11(5):473-83. Epub 2010 May 11.

A survey of sequence alignment algorithms for next-generation sequencing.

Li H, Homer N.

Broad Institute, Cambridge, MA 02142, USA. hengli@broadinstitute.org

Storing alignment: SAM Format

- Store information related to alignement
  - Read ID
  - CIGAR String
  - Bitwise FLAG
    - read paired
    - read mapped in proper pair
    - read unmapped, ...
  - Alignment position
  - Mapping quality
  - ...

The extended CIGAR string

- Exemple flags:
  - M alignment match (can be a sequence match or mismatch)
  - I insertion to the reference
  - D deletion from the reference
  - http://samtools.sourceforge.net/SAM1.pdf

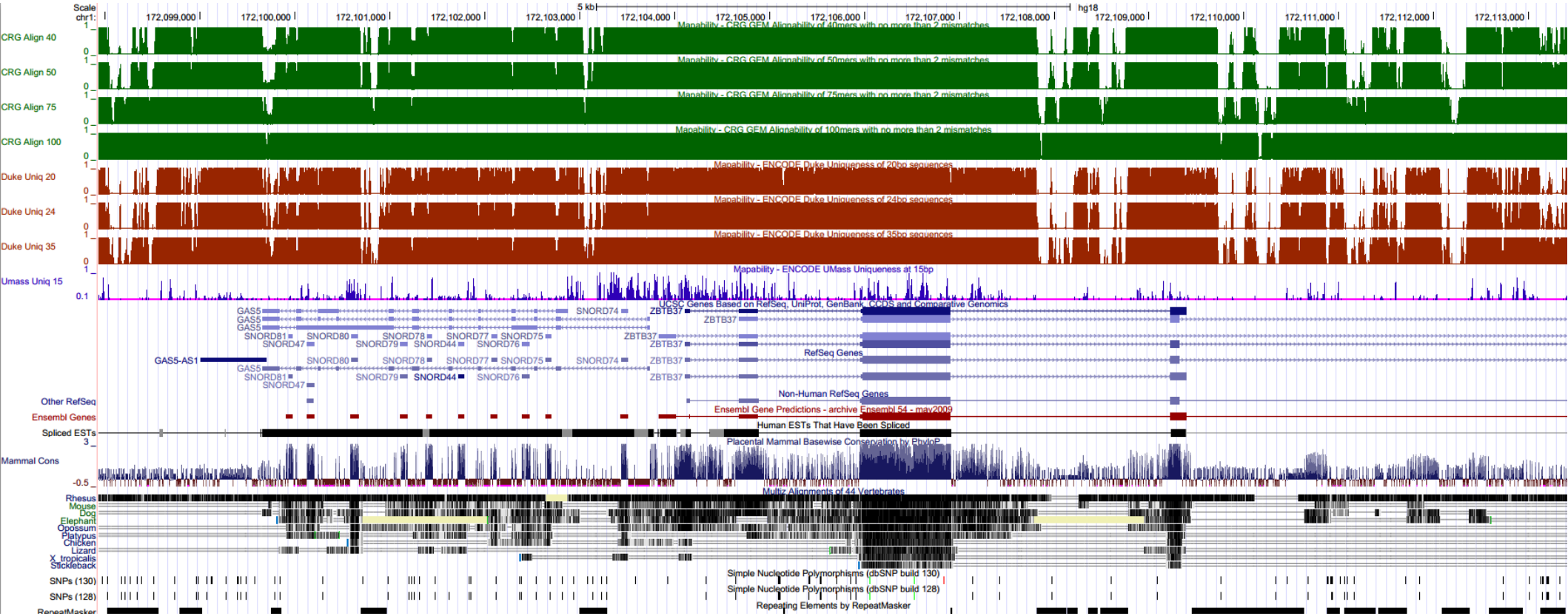
```
ATTCAGATGCAGTA
ATTCA--TGCAGTA

 5M2D7M
```

Mapping reads

- Main Issues:
  - Number of allowed mismatches
    - Depend on sequence size (sometimes heterogeneous length)
    - Depend of the aligner
  - Number of multi-hits
    - Issue with short reads
  - PCR duplicates
    - Accepted with RNA-Seq
    - Warning with ChIP-Seq (library complexity)
  - Mates expected distance (mate/paired-sequencing)
- RNA-Seq specific
  - Considering exon junctions (RNA-Seq)

Mappability

- Sequence uniqueness of the reference
- These tracks display the level of sequence uniqueness of the reference NCBI36/hg18 genome assembly. They were generated using different window sizes, and high signal will be found in areas where the sequence is unique.



Mapping read spanning exons

- One limit of bowtie
  - mapping reads spanning exons
- Solution: splice-aware short-read aligners

Assembler	*De novo*?	Parallelism	Support for paired-end reads?	Support for stranded reads?	Support for multiple insert sizes?	Outputs transcript counts?	Software availability	Refs
G-Mo.R-Se	No	None	No	No	No	No	http://www.genoscope.cns.fr/externe/gmorse/	17
Cufflinks	No	MP	Yes	Yes	Yes	Yes	http://cufflinks.cbcb.umd.edu/	20
Scripture	No	None	Yes	Yes	Yes	Yes	http://www.broadinstitute.org/software/scripture/	16
ERANGE	No	None	Yes	Yes	Yes	Yes	http://woldlab.caltech.edu/rnaseq	50
Multiple-k	Yes	None	Yes	Yes	Yes	No	http://www.surget-groba.ch/downloads/	19
Rnnotator	Yes	MP	Yes	Yes	Yes	Yes	Contact David Gilbert (DEGilbert@lbl.gov)	15
Trans-ABySS	Yes	MPI	Yes	No	Yes	Yes	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	18
Oases	Yes	MP	Yes	Yes	Yes	No	http://www.ebi.ac.uk/~zerbino/oases/	-
Trinity	Yes	MP	Yes	Yes	No	Yes	http://trinityrnaseq.sourceforge.net/	59

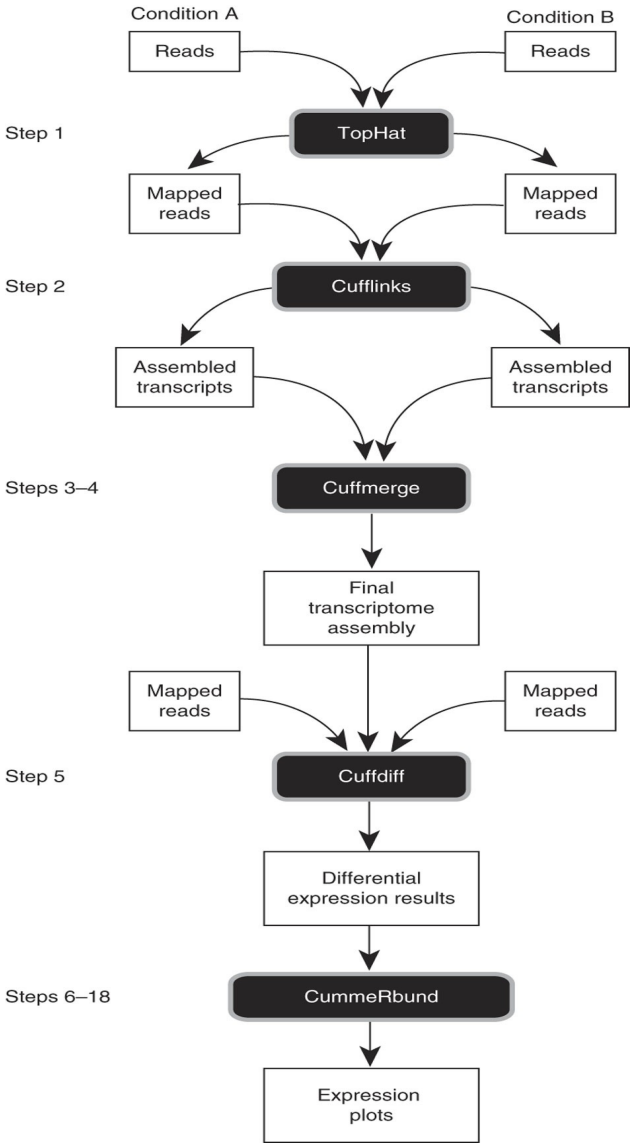
MP, multiple processor support (assembler takes advantage of many cores from a single computer); MPI, message-passing interface support (assembler runs in parallel on multiple computers within a cluster).

Nat Rev Genet. 2011 Sep 7;12(10):671-82. doi: 10.1038/nrg3068.

Next-generation transcriptome assembly.

Martin JA, Wang Z.

The Tuxedo pipeline



Nat Protoc. 2012 Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016.

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L.

TopHat pipeline

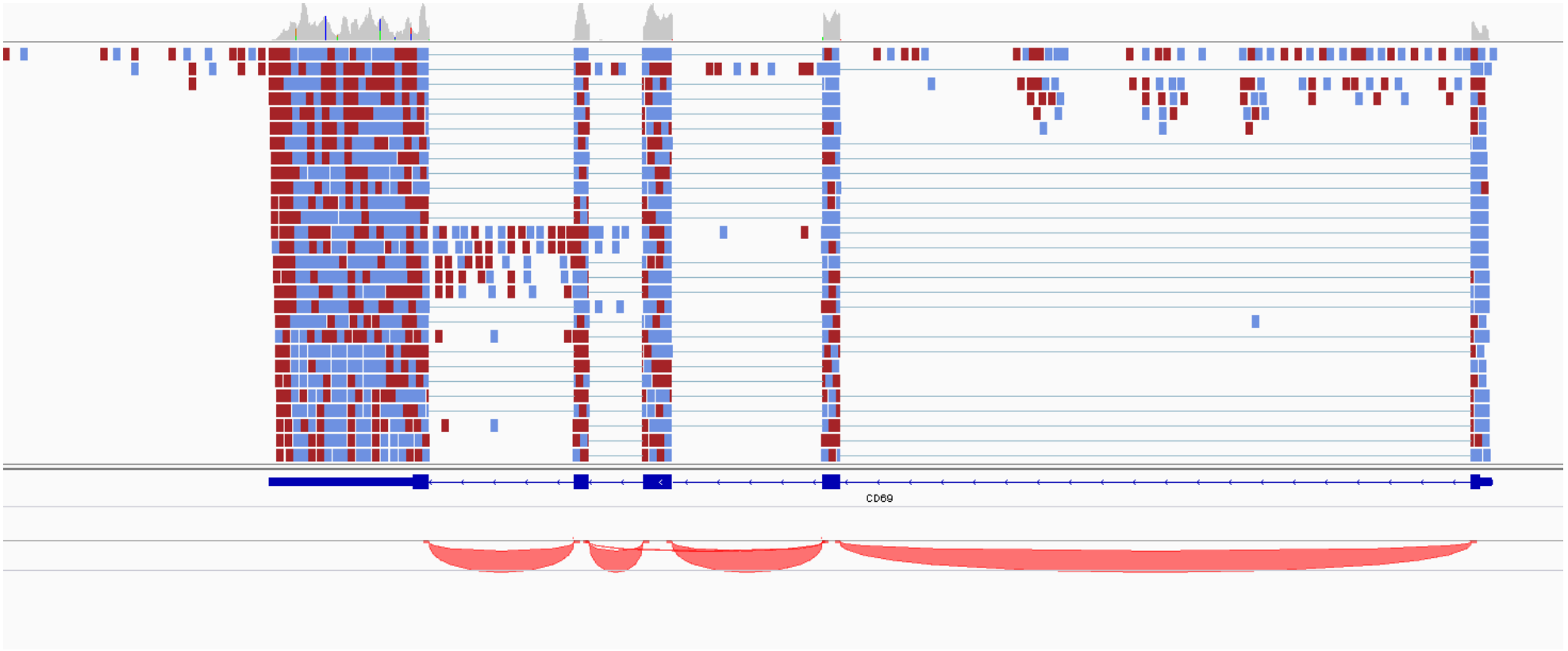
- RNA-Seq reads are mapped against the whole reference genome (bowtie).
- TopHat allows Bowtie to report more than one alignment for a read (default=10), and suppresses all alignments for reads that have more than this number
- Reads that do not map are set aside (initially unmapped reads, or IUM reads)
- TopHat then assembles the mapped reads using the assembly module in Maq. An initial consensus of mapped regions is computed.
- The ends of exons in the pseudoconsensus will initially be covered by few reads (most reads covering the ends of exons will also span splice junctions)
  - Tophat a small amount of flanking sequence of each island (default=45 bp).

Bioinformatics. 2009 May 1;25(9):1105-11. Epub 2009 Mar 16.

TopHat: discovering splice junctions with RNA-Seq.

Trapnell C, Pachter L, Salzberg SL.

TopHat example results



- Output : BAM file (compressed version of SAM)

TopHat pipeline

- Weakly expressed genes should be poorly covered
  - Exons may have gaps
- To map reads to splice junctions, TopHat first enumerates all canonical donor and acceptor sites within the island sequences (as well as their reverse complements)

- Next, tophat considers all pairings of these sites that could form canonical (GT–AG) introns between neighboring (but not necessarily adjacent) islands.
  - By default, TopHat examines potential introns longer than 70 bp and shorter than 20 000 bp (more than 93% of mouse introns in the UCSC known gene set fall within this range)
- Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions.
- Read are mapped onto these junction library

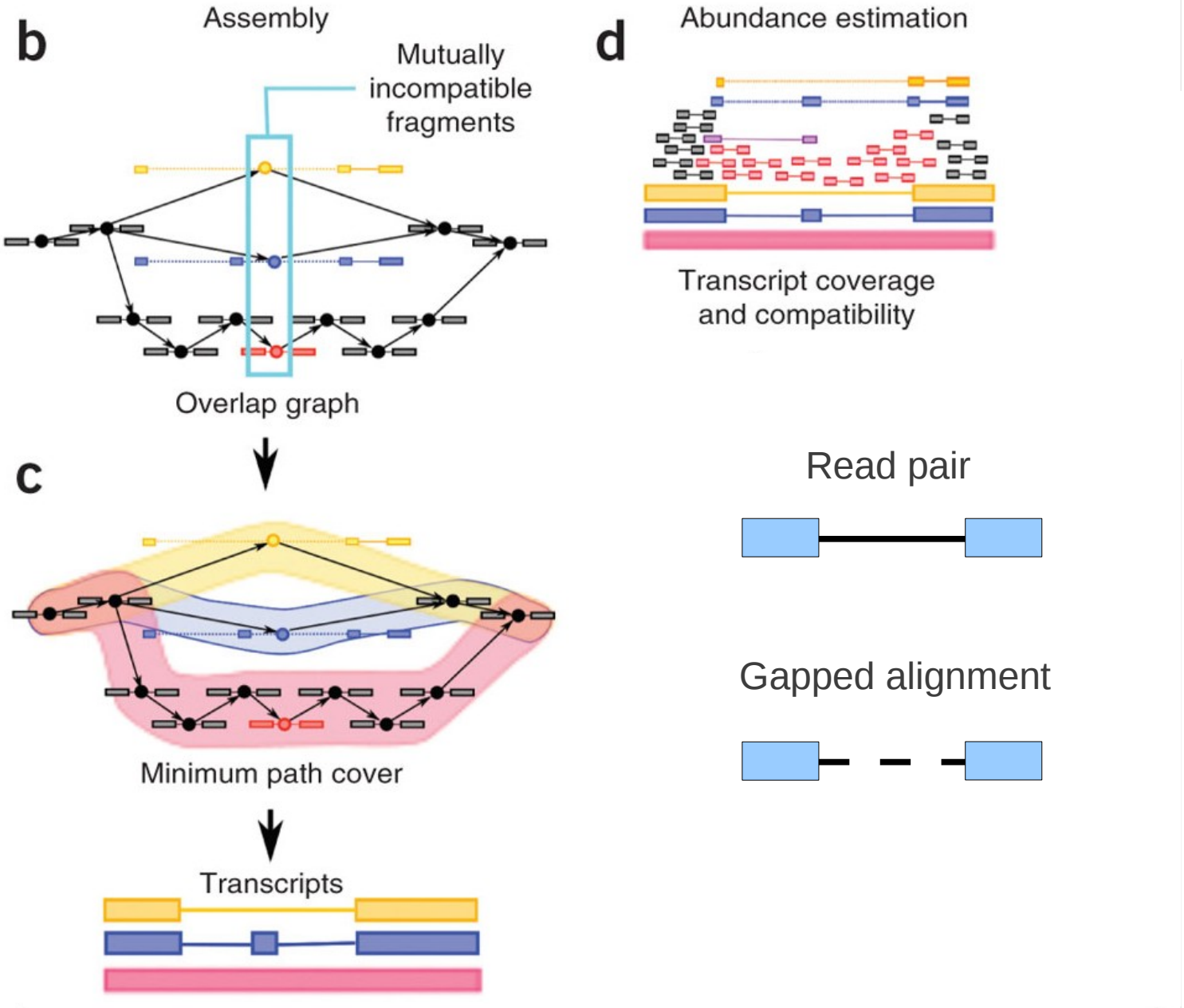
Compressing and indexing files

- Needed before visualization in Genome Browser

```
[u@m] samtools view output.bam # output SAM format
[u@m] samtools sort output.bam output.sorted
[u@m] samtools index output.sorted.bam
```

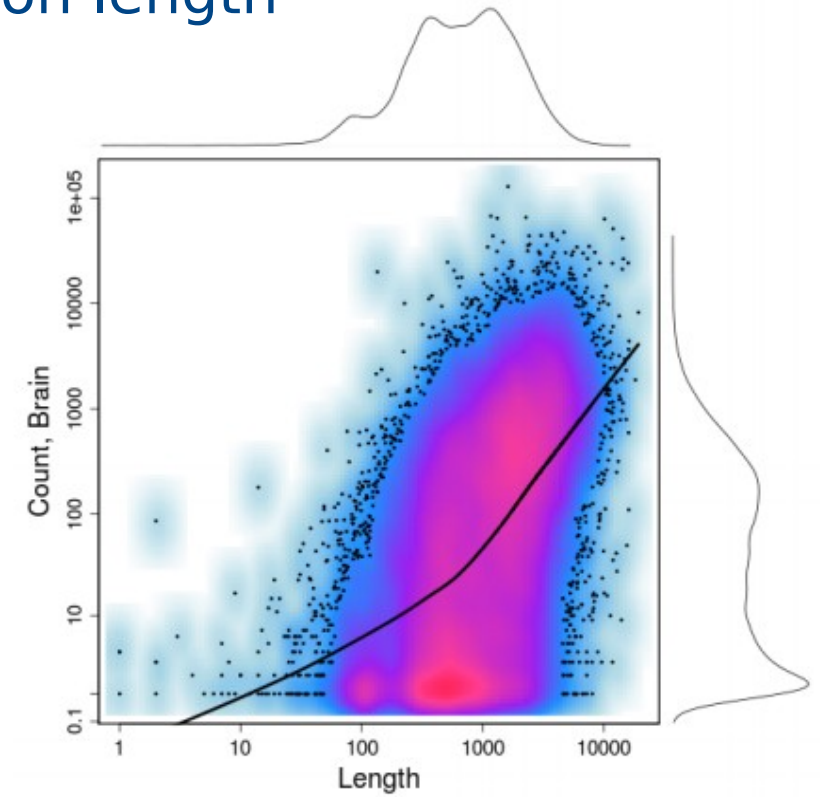
- Or use Galaxy or IGVtools

Cufflinks: transcript assembly and quantification



Summarization issues

- Positive association between gene counts and length
  - suggests higher expression among longer genes or non-linear dependence of gene counts on length
  - Need to scale by gene length



Count vs. length

BMC Bioinformatics. 2010 Feb 18;11:94.

Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.

Bullard JH, Purdom E, Hansen KD, Dudoit S.

RPKM/FPKM normalization

- RPKM: Tag count is normalized for transcript length and total read number in the measurement (RPKM, Reads Per Kilobase of exon model per Million mapped reads)
  - 2kb transcript with 3000 alignments in a sample of 10 millions of mappable reads
  - RPKM = 3000/(2 * 10) = 150
- FPKM, Fragments Per Kilobase of exon model per Million mapped reads (paired-end sequencing)

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber[1], Manfred G Grabherr[1], Mitchell Guttman[1,2] & Cole Trapnell[1,3]

Accurate quantification of transcriptome from RNA-Seq data by effective length normalization

Soohyun Lee[1], Chae Hwa Seo[1], Byungho Lim[2], Jin Ok Yang[1], Jeongsu Oh[1], Minjin Kim[2], Sooncheol Lee[2], Byungwook Lee[1], Changwon Kang[2] and Sanghyuk Lee[1,3,*]

Limits of RPKM/FPKM

- If a large number of genes are highly expressed in, one experimental condition, the expression values of remaining genes will be decreased.
  - Can force the differential expression analysis to be skewed towards one experimental condition.

Other normalization methods

- **Several methods proposed**

- Total count (TC): Gene counts are divided by the **total number of mapped reads** (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset.
  - First proposed
- Median (Med): Also similar to TC, the total counts are replaced by the **median counts different from 0** in the computation of the normalization factors.
  - Warning : if lots of weakly expressed values
- Upper Quartile (UQ): the total counts are replaced by the **upper quartile** of counts different from 0 in the computation of the normalization factors.
  - Very similar in principle to TC (but really more powerful).
- Trimmed Mean of M-values (TMM): This normalization method is implemented in the **edgeR Bioconductor** package (version 2.4.0). Scaling is based on a subset of M values.
  - TMM seems to provide a robust scaling factor.
- Quantiles (Q): First proposed in the context of microarray data, this normalization method consists in **matching distributions** of gene counts across lanes.
  - Use with caution when comparing distantly related tissues.
- **Reads Per Kilobase per Million mapped reads** (RPKM): This approach was initially introduced to facilitate comparisons between genes within a sample.
  - Not sufficient (need to be combined with inter-sample normalization method)

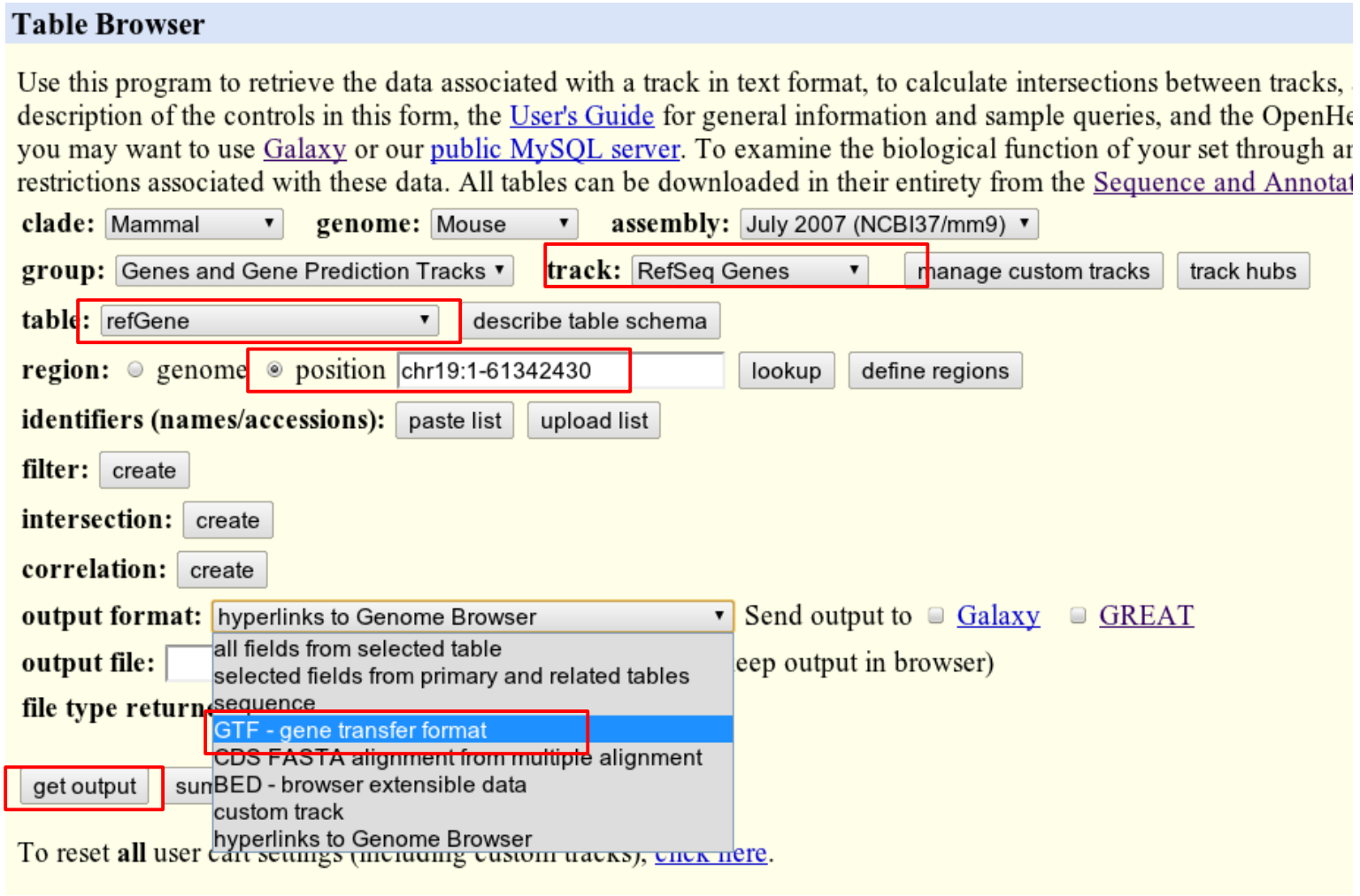
Brief Bioinform. 2012 Sep 17. [Epub ahead of print]

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; on behalf of The French StatOmique Consortium.

Getting a GTF/GFF file

- Ask UCSC

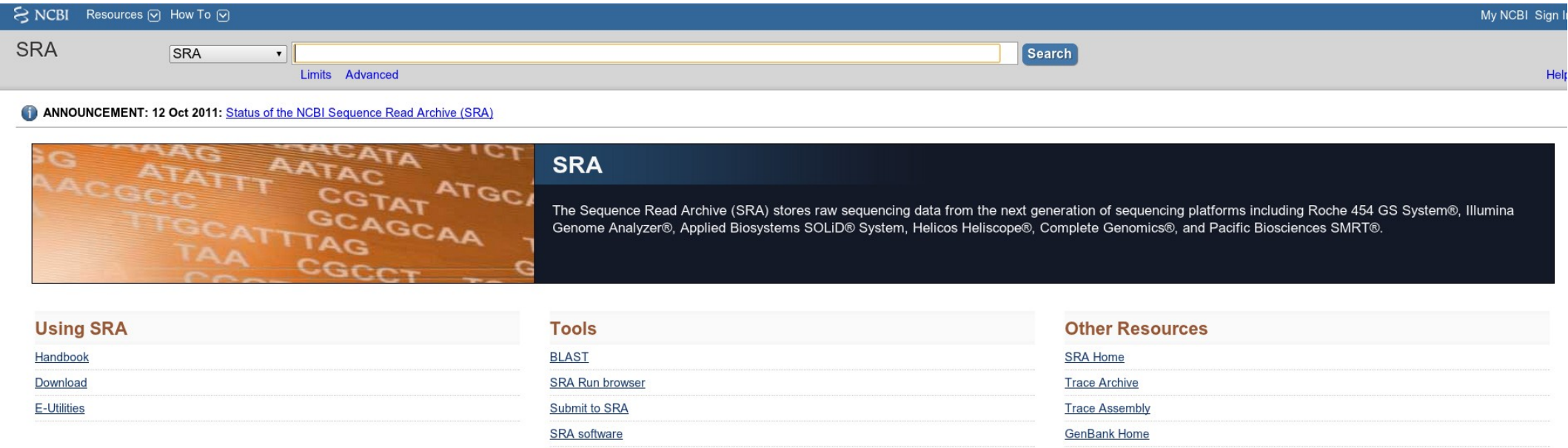


Differential expression analysis

- Cufflinks includes a program, "Cuffdiff"
  - Find significant changes in
    - transcript expression
    - Gene expression
  - From the command line, run cuffdiff as follows:

```
[u@m] cuffdiff MyTranscrit.gtf tophat_result_1.bam tophat_result_2.bam
 -o cuffdiff
```

Sequence read Archive (SRA)



- The SRA archives high-throughput sequencing data that are associated with:
- RNA-Seq, ChIP-Seq, and epigenomic data that are submitted to GEO

SRA growth

Display Settings: Abstract

Send to:

Nucleic Acids Res. 2011 Oct 18. [Epub ahead of print]

The sequence read archive: explosive growth of sequencing data.

Kodama Y, Shumway M, Leinonen R; on behalf of the International Nucleotide Sequence Database Collaboration.

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Abstract

New generation sequencing platforms are producing data with significantly higher throughput and lower cost. A portion of this capacity is devoted to individual and community scientific projects. As these projects reach publication, raw sequencing datasets are submitted into the primary next-generation sequence data archive, the Sequence Read Archive (SRA). Archiving experimental data is the key to the progress of reproducible science. The SRA was established as a public repository for next-generation sequence data as a part of the International Nucleotide Sequence Database Collaboration (INSDC). INSDC is composed of the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). The SRA is accessible at www.ncbi.nlm.nih.gov/sra from NCBI, at www.ebi.ac.uk/ena from EBI and at trace.ddbj.nig.ac.jp from DDBJ. In this article, we present the content and structure of the SRA and report on updated metadata structures, submission file formats and supported sequencing platforms. We also briefly outline our various responses to the challenge of explosive data growth.

PMID: 22009675 [PubMed - as supplied by publisher] **Free full text**

In 2011 the SRA surpassed 100 Terabases of open-access genetic sequence reads from next generation sequencing technologies. The Illumina™ platform comprises 84% of sequenced bases, with SOLiD™ and Roche/454™ platforms accounting for 12% and 2%, respectively. The most active SRA submitters in terms of submitted bases are the Broad Institute, the Wellcome Trust Sanger Institute and Baylor College of Medicine with 31, 13 and 11%, respectively. The largest individual global project generating next-generation sequence is the 1000 Genomes project which has contributed nearly one third of all bases. The most sequenced organisms are *Homo sapiens* with 61%, human metagenome with 6% and *Mus musculus* with 5% share of all bases. The common

SRP000698 study

SRP000698 Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing.

Study Type:	Transcriptome Analysis
Submission:	SRA008367 by Barts and The London on 2009-04-23 09:50:34
Abstract:	Many disease associated variants identified by genome-wide association (GWA) studies are expected to regulate gene expression. Allele specific expression (ASE) quantifies transcription from both haplotypes using individuals heterozygous at tested SNPs. We performed deep human transcriptome-wide resequencing (RNA-seq) for ASE analysis and expression quantitative trait locus (eQTL) discovery. We resequenced double poly(A) selected RNA from primary CD4+ T-cells (n=4 individuals, both activated and untreated conditions) and developed tools for paired end RNA-seq alignment and ASE analysis. We generated an average of 20 million uniquely mapping 45 base reads per sample. We obtained sufficient read depth to test 1,371 unique transcripts for ASE. Multiple biases inflate the false discovery rate which we estimate to be approximately 50% for random SNPs. However, after controlling for these biases and considering the subset of SNPs that pass HapMap QC, 4.6% of heterozygous SNP-sample pairs show evidence of imbalance (p < 0.001). We validated four findings by both bacterial cloning and Sanger sequencing assays. We also found convincing evidence for allelic imbalance at multiple reporter exonic SNPs in CD6 for two samples heterozygous at the multiple sclerosis associated variant rs17824933, linking GWA findings with variation in gene expression. Finally, we show in CD4+ T-cells from a further individual that high throughput sequencing of genomic DNA and RNA-seq following enrichment for targeted gene sequences by sequence capture methods offers an unbiased means to increase the read depth for transcripts of interest, and therefore a method to investigate the regulatory role of many disease associated genetic variants
Description:	n/a
Center Project:	CD4 T cell RNA-Seq
NCBI Link:	Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing.

Show Entrez docsums for all experiments

Download reads for entire study as sra or sra-lite

What is "sra" and "sra-lite" formats?

(use Aspera plugin for fast download)

Experiments

Show RUNs for each experiment

⇑Accession	Spots	Bases
Total: 15	194.1M	17.1G
SRX011543	13.6M	1.2G
SRX011544	7.6M	679.8M
SRX011545	23.5M	2.1G
SRX011546	27.8M	2.5G
SRX011547	13.3M	1.2G
SRX011548	10.7M	963.2M
SRX011549	10.0M	902.4M
SRX011550	9.7M	870.3M
SRX011551	20.2M	3.1G
SRX011552	9.4M	1.4G
SRX011553	4.9M	218.8M
SRX011554	6.9M	305.6M
SRX011555	7.6M	336.3M
SRX011556	21.3M	950.2M
SRX011557	7.7M	337.5M

SRA Concepts

- Data submitted to SRA is organized using a metadata model consisting of six objects:

- Study – A set of experiments with an overall goal and literature references.
- Experiment – An experiment is a consistent set of laboratory operations on input material with an expected result.
- Sample – An experiment targets one or more samples. Results are expressed in terms of individual samples or bundles of samples as defined by the experiment.
- Run – Results are called runs. Runs comprise the data gathered for a sample or sample bundle and refer to a defining experiment.

Getting fastq files using SRA toolkit

- *.sra to fastq conversion
- Fastq-dump

- `fastq-dump -A SRRxxxx.sra`

- Note: use –split-files argument for paired-end library

SRX011549 & SRX011550 experiments

- SRX011550: human naïve T-cells
- SRX011549: human activated CD4 T-Cells

Display Settings: Full Send to:

"Illumina sequencing of Human CD4 T cells RNA-Seq paired-end library "

Accession: SRX011549
Experiment design: not provided
Submission: SRA008367 by Barts and The London
Study summary: Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. (SRP000698) • Study • All experiments (more...)
Sample: CD4 T cell Transcriptome Activated for 3 hours with CD3 Cd28 antibodies (SRS005034) (more...)
Library: RNA-Seq Individual 4 Activated PE (more...)
Platform: Illumina (more...)
Processing:
Base calls: Base Space, Illumina primary analysis
Quality score: Illumina primary analysis, 80x1
Spot descriptor:

1 forward 46 reverse

Paired-end (2x45)

Total: 2 runs, 10M spots, 902.4M bases
Download reads for this experiment in sra or sra-lite formats

#	Run	# of Spots	# of Bases
1.	SRR027888	5,402,892	486.3M
2.	SRR027890	4,623,729	416.1M

Runs

ID: 11836

Getting fastq files using SRA toolkit

- *.sra to fastq conversion
- Fastq-dump

- `fastq-dump -A SRRxxxx.sra`

- Note: use –split-files argument for paired-end library

Merci