THE UNIVERSITY OF IOWA

Galaxy Workshop 1-8-13

Intros:

- Tom Bair
  - thomas-bair@uiowa.edu
- Ann Black-Ziegelbein
  - annblack@eng.uiowa.edu
- Srinivas Maddhi
  - srinivas-maddhi@uiowa.edu

What is galaxy good for

```
~/cufflinks-1.2.1.Linux_x86_64/cuffdiff ~/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf /Sample_tube1_081611ns/tophat_out/accepted_hits.bam,/Sample_tube2_101011ns/tophat_out/accepted
e3_102611ns/tophat_out/accepted_hits.bam,./Sample_tube4_120111ns/tophat_out/accepted_hits.bam ./Sample_tube5_012011sm/tophat_out/accepted_hits.bam,./Sample_tube6_012111c_sm/tophat_out/acc
e_tube7_030111sm/tophat_out/accepted_hits.bam,./Sample_tube8_032211a_sm/tophat_out/accepted_hits.bam
~/cufflinks-1.2.1.Linux_x86_64/cuffdiff ~/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf ./Sample_tube1_081611ns/tophat_out/accepted_hits.bam,./Sample_tube2_101011ns/tophat_out/accept
ube3_102611ns/tophat_out/accepted_hits.bam,./Sample_tube4_120111ns/tophat_out/accepted_hits.bam ./Sample_tube5_012011sm/tophat_out/accepted_hits.bam,./Sample_tube6_012111c_sm/tophat_out/a
ple_tube7_030111sm/tophat_out/accepted_hits.bam,./Sample_tube8_032211a_sm/tophat_out/accepted_hits.bam
~/cufflinks-2.0.2.Linux_x86_64/cuffdiff ~/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf ./Sample_tube1_081611ns/tophat_out/accepted_hits.bam,./Sample_tube2_101011ns/tophat_out/accept
ube3_102611ns/tophat_out/accepted_hits.bam,./Sample_tube4_120111ns/tophat_out/accepted_hits.bam ./Sample_tube5_012011sm/tophat_out/accepted_hits.bam,./Sample_tube6_012111c_sm/tophat_out/a
ple_tube7_030111sm/tophat_out/accepted_hits.bam,./Sample_tube8_032211a_sm/tophat_out/accepted_hits.bam
~/cufflinks-2.0.2.Linux_x86_64/cuffdiff ~/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf ./Sample_tube1_081611ns/tophat_out/accepted_hits.bam,./Sample_tube2_101011ns/tophat_out/accept
ube3_102611ns/tophat_out/accepted_hits.bam,./Sample_tube4_120111ns/tophat_out/accepted_hits.bam ./Sample_tube5_012011sm/tophat_out/accepted_hits.bam,./Sample_tube6_012111c_sm/tophat_out/a
ple_tube7_030111sm/tophat_out/accepted_hits.bam,./Sample_tube8_032211a_sm/tophat_out/accepted_hits.bam
```

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:

5: Galaxy5-[brain_2...fastqsanger

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Built-ins were indexed using default options

Select a reference genome:

e_coli Full

If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:

Single-end

TopHat settings to use:

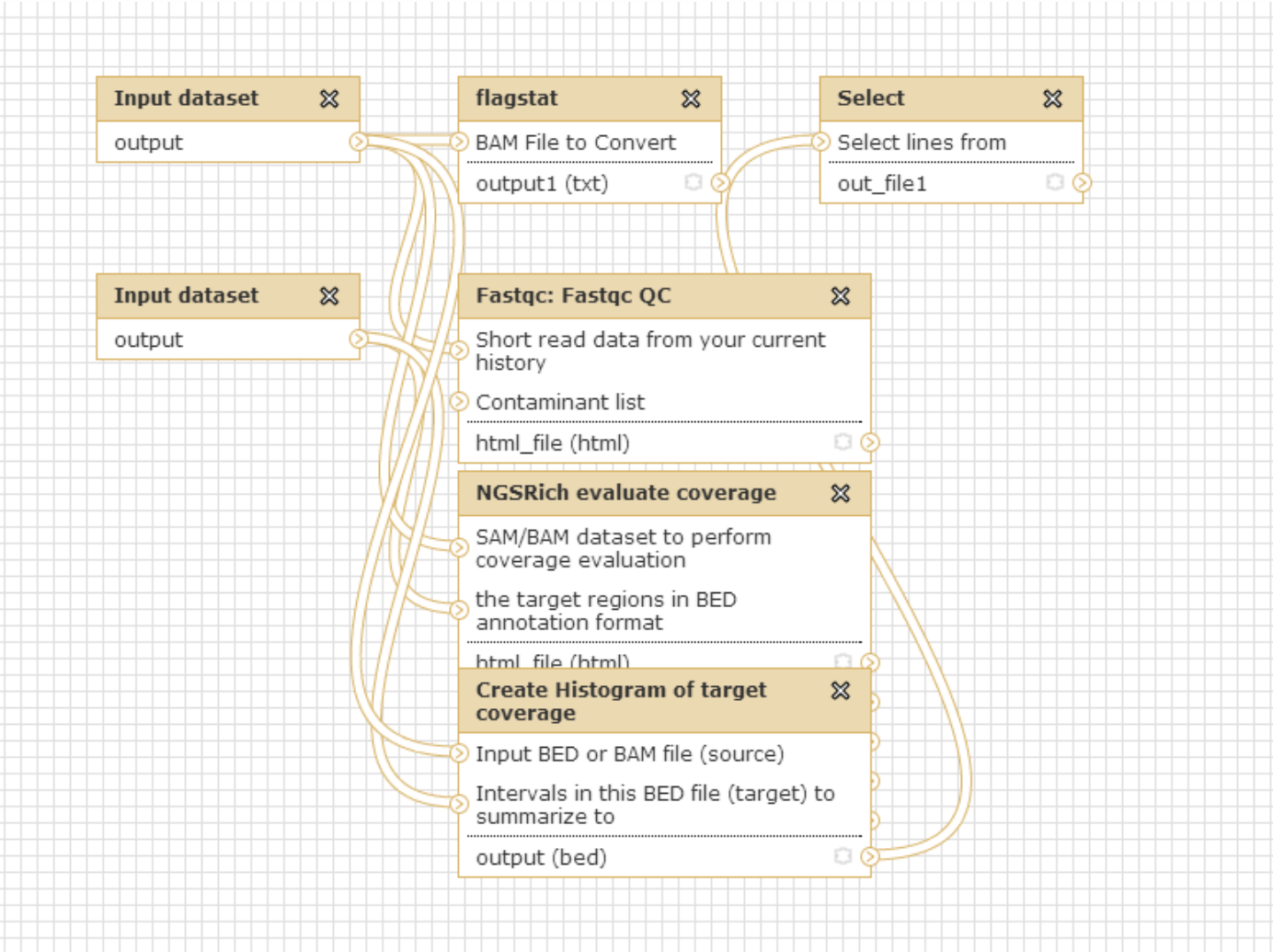
Use Defaults

You can use the default settings or set custom values for any of Tophat's parameters.

Execute

Access to resources

Documentation



What doesn't Galaxy do?

- Galaxy does not know the best or even correct way to do an analysis
- Very good and providing a "wrapper" for command line tools
  - If no command line tool Galaxy won't work very well until one is developed
- Hard tasks are still hard
  - If the application requires massive memory or computational resources it still will be difficult
  - Group resources rather than individual
    - Much more CPU/Memory/Storage
      - But you are in contention with others

Where is this magical “galaxy” thing

- https://main.g2.bx.psu.edu/
  - Galaxy “Main”
    - Very useful but somewhat limited on storage resources (250G of storage)
- https://galaxy.hpc.uiowa.edu/
  - IIHG Galaxy
    - Fewer restrictions on storage space
    - Currently not as full featured as galaxy main
    - Runs on local hardware
- https://localhost , https://your_cloud _provider
  - Open source and well supported ways to run anywhere

FastQ? Bed? Sam? Bam? Understanding data formats

- Most files are just plain old TEXT files
  - With rules/specifications that dictate file structure (just like grammar)
  - Can view, edit and save with your favorite text editor
    - Like notepad!

File Format	Data Purpose
.fastq	Format of the raw short read data from the sequencer
.sam	Sequence Alignment/Map format that represents aligned reads
.bam	compressed/binary format of SAM (Binary Alignment Map)
.bed	Interval file, tab delimited, depicting regions of the genome
.vcf	Variant Call Format that have lines that represent information about structural genetic variants

- For more information …
  - UCSC Browser: http://genome.ucsc.edu/FAQ/FAQformat
  - IGV: http://www.broadinstitute.org/igv/FileFormats

FASTQ Format

Illumina sequence identifiers

Sequences from the Illumina software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

Versions of the Illumina pipeline since 1.4 appear to use **#NNNNNN** instead of **#0** for the multiplex ID, where **NNNNNN** is the sequence of the multiplex tag.

- **FASTQ format**
  - **Stores sequences and Phred Quality scores in a concise format**
  - **Encodes Illumina Sequencer Identification Information**

```
@HWI-ST821_0101:1:1101:1981:2081#ATCACG/2
CACAAACCTCTCCCTATTCCTCCTTTCCCTCTATCCTTCTCAGCCCCTAGTATTCTTTTTACTTCTATGAGATCAACTTTTTTTTTTTGTAACACATGGT
+HWI-ST821_0101:1:1101:1981:2081#ATCACG/2
abaeeecegceggbgihhiiiaggggbbgggfhiiiffghhhheggggbiig^affegggeegffbgfb[d\Z_b\VHMMVZ`^BBBBBBBBBBBBBBB
@HWI-ST821_0101:1:1101:1770:2136#ATCACG/2
```

- **FASTQ variants**

The range of quality scores in a FASTQ file will depend on the technology and the base caller used.

- If the quality scores contain characters in the range ASCII 33 - 58 -> can only be Sanger
- If FastQ file is known to be from an Illumina/Solexa platform **AND** the quality scores contain characters in the range ASCII 59 - 63 -> can only be Solexa/Illumina 1.0
- If ASCII characters 64 or 65 are used in quality scores -> cannot be Illumina 1.5+

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Thank you Wikipedia!

SAM format

- Flexible format to store alignment information from any alignment program
- Index-able and stream-able to allow efficient retrieval and lower memory requirements

```
HWI-ST1122:213:C1DWVACXX:8:1101:4101:1904    16    chr15  66795583    42    33M    *    0    0
CATCAGAACATCCGAGAAAATCATGTGGTTCAG    IEJIJIJJJJIJJJJJJJJJJHHHHHFFFDD=1
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

BAM format

- Binary format
  - Smaller but not human readable
- Easy to convert between SAM and BAM
- BAM much more space and computationally efficient

BED Format

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2
itemRgb="On"
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```

Quiz:

- What file format is this?

track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512

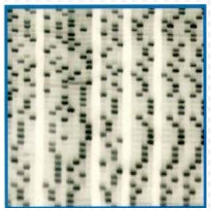
- How do you convert between fastq and SAM?

- What file format is this?

ï¿½ï¿½ï¿½ï¿½)ï¿½,UfÚ,ï¿½:?~ï¿½ï¿½×¯Hï¿½8ï¿½9ï¿½ï¿½lNï¿½ï¿½S<ï¿½

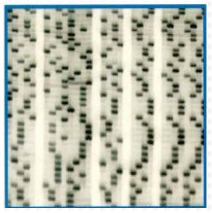
Evolution of DNA Sequencers

1st Generation DNA Sequencers
Sanger dideoxy-based

Pre-1992 "old fashioned way"	1992-1999 ABI 373/377	1999 ABI 3700	2003 ABI 3730XL
			
S35 ddNTPs **Gels** **Manual loading** **Manual base calling**	**Fluorescent ddNTPs*** **Gels** **Manual loading** **Automated base calling***	**Fluorescent ddNTPs** **Capillaries*** **Robotic loading*** **Automated base calling** **Breaks down frequently**	**Fluorescent ddNTPs** **Capillaries** **Robotic loading** **Automated base calling** **Reliable***

Evolution of DNA Sequencers

1st Generation DNA Sequencers
Sanger dideoxy-based



2005
Roche GS20
Aka "the 454"

Early 2007
Solexa 1G

Late 2007
ABI SOLiD

2nd (Next) Generation Sequencers
"Single Molecule Detection After Amplification"

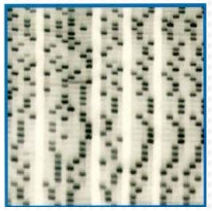
"Long-Read" Instrument
300 base reads
400K reads/run
70 billion bases/run

"Short-Read" Instrument
36 Base reads
40M reads/run
1 billion bases/run

"Short-Read" Instrument
35 Base reads
40M reads/run
1 billion bases/run

Evolution of DNA Sequencers

1st Generation DNA Sequencers
Sanger dideoxy-based




2nd (Next) Generation Sequencers
"Single Molecule Detection After Amplification"



2008
Helicos HeliScope

2010
PacBio RS

2012
Oxford Nanopore
GridION/MinION

3rd (Next-Next) Generation Sequencers
"Single Molecule Detection"



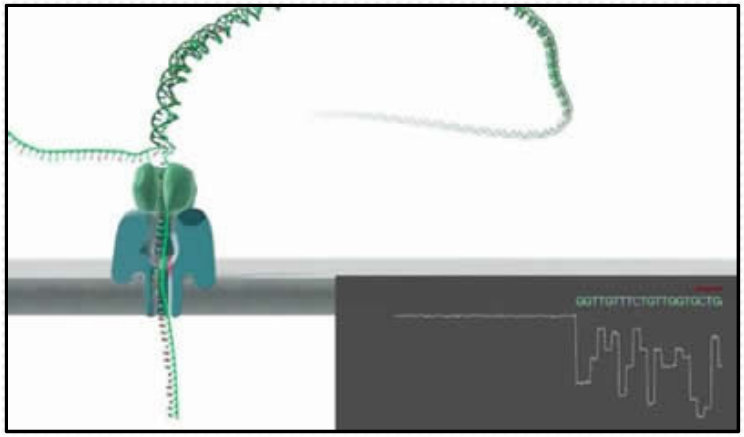
tSMS
True Single-Molecule Sequencing
Short-read system
35 base average read
~25 Gb/run
Up to 4,800 samples/run

SMRT Sequencing
Single-molecule Real-time
Long-Read System
3 kb average read, up to 10 kb

Nanopore Sensing
Scalable for long- or short-read

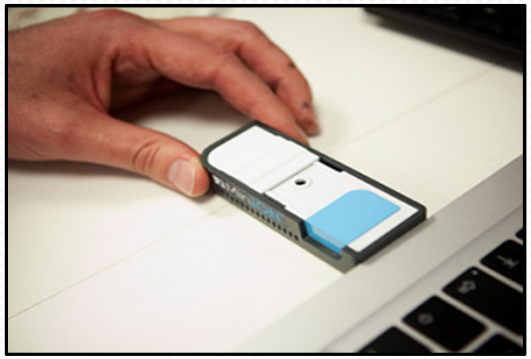
Oxford Nanopore

An electric current is applied across a pore, and when analytes bind at a site engineered within the pore, a signature change in the current specific to the compound passing through is generated. Hence, the order of the individual base units of DNA – cytosine, adenine, guanine and thymine – can be recorded electronically as a strand of DNA passes though the pore.



GridION

MinION



Scalable Read Lengths
>10,000 nt
>40 – 100 GB-GridION
>1 GB-MinION

Cost Comparisons

Type	Mbp/run	Run time	$/Mbp	$/Machine
Sanger	0.08	2 hr	100	$250,000
NextGen	600	1 wk	10	$750,000
NextNextGen	900	6 hr	1	$1,000 OR LESS!!!

Evolution of NGS Platforms

2nd (Next) Generation Sequencers
"Single Molecule Detection After Amplification"

454

Solexa 1G

ABI SOLiD

Roche GS-FLX

Illumina GAIIx

ABI SOLiDv4

GS-FLX Titanium

Illumina HiSeq2000

ABI 5500xl

ABI Ion Torrent

Junior

Roche GS-FLX+

Illumina MiSeq

Illumina HiSeq2500

PGM

Proton

Pros and Cons of NGS

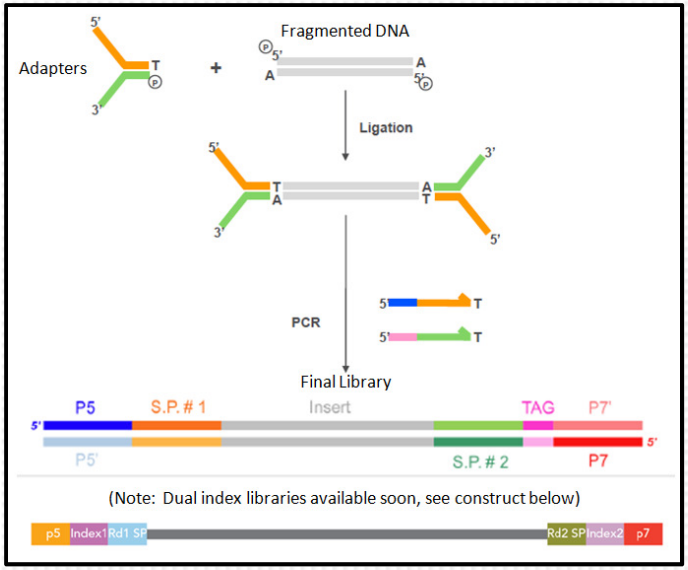
Pros

- No sub-cloning, no need for bacterial host
  - Less cloning bias
  - Easier to create more libraries
- Large number of individual sequence reads generated
  - Permits quantification by counting # of reads
  - Enhanced dynamic range
  - Permit detection of rare variants
- Readily adaptable to a variety of applications
  - Whole genomes (DNA-Seq)
  - Whole transcriptomes (RNA-Seq)
  - Epigenomes (Methyl-Seq)
  - Small RNA (miRNA-Seq)
- Landscape is changing rapidly
  - Dramatic decrease in cost and speed of sequencing data
  - Permitted experiments that were previously cost prohibitive
  - Revolutionized the way we do sequencing

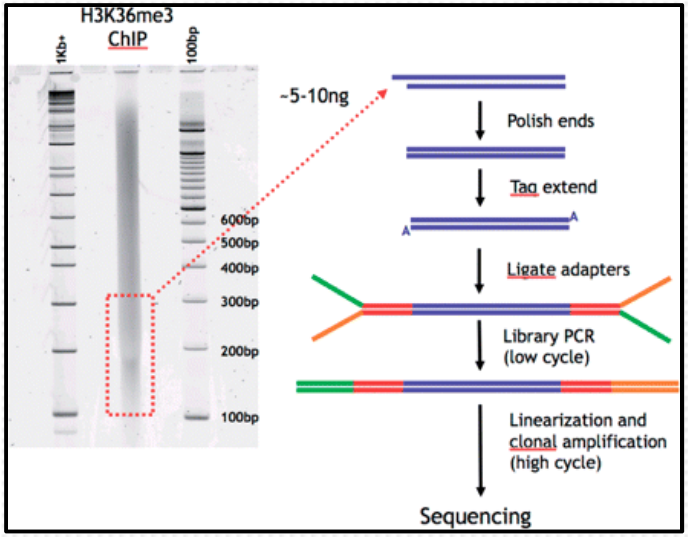
Cons

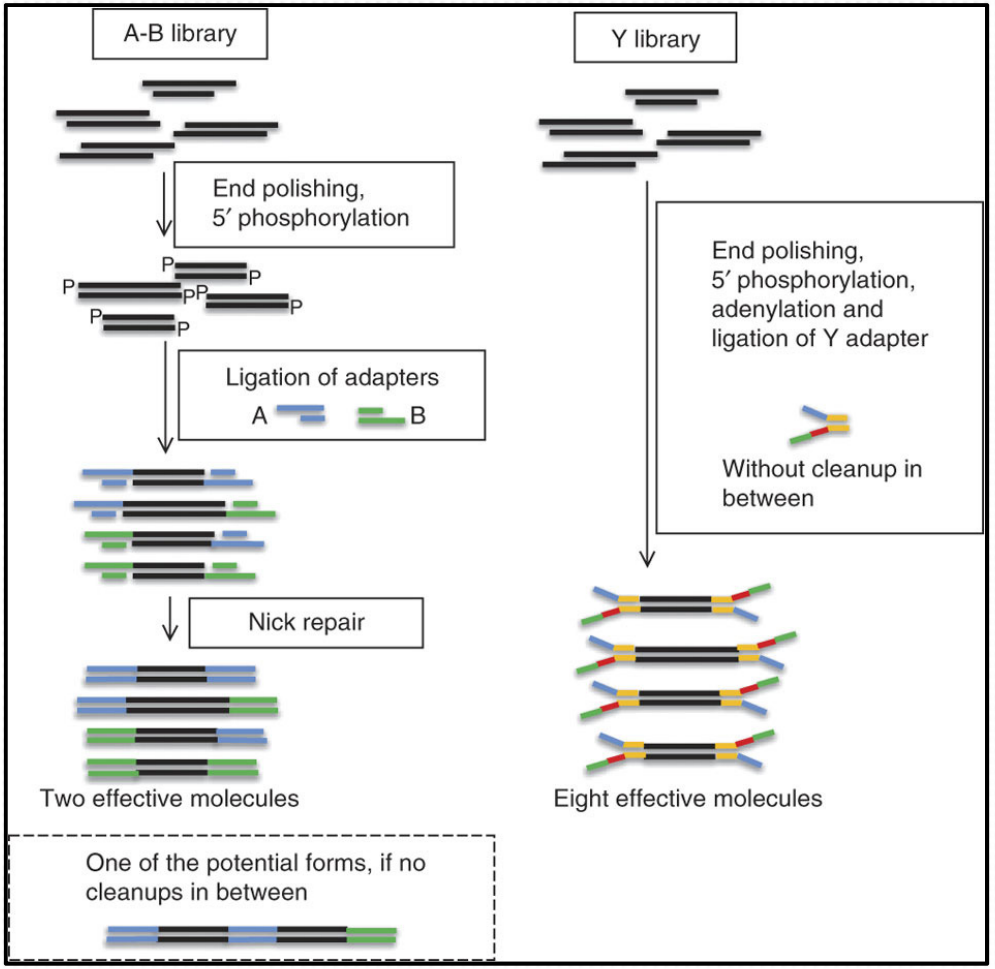
- Landscape is changing rapidly
  - Platforms
  - Analysis software
  - Sequencing chemistries
  - Difficult to have continuity within projects
- Vendors overstate performance
  - Fail to meet performance expectation deadlines
  - Fragile instruments

Library Prep Overview



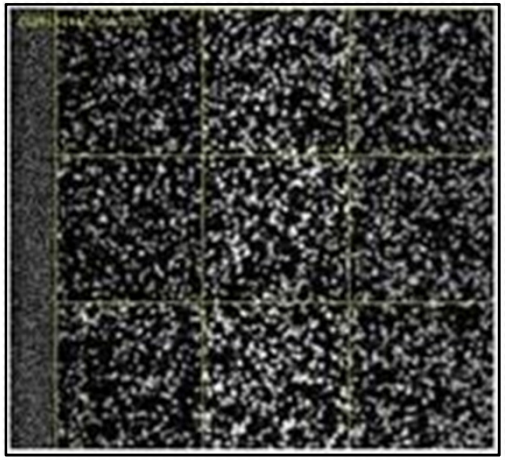
Illumina Library Prep





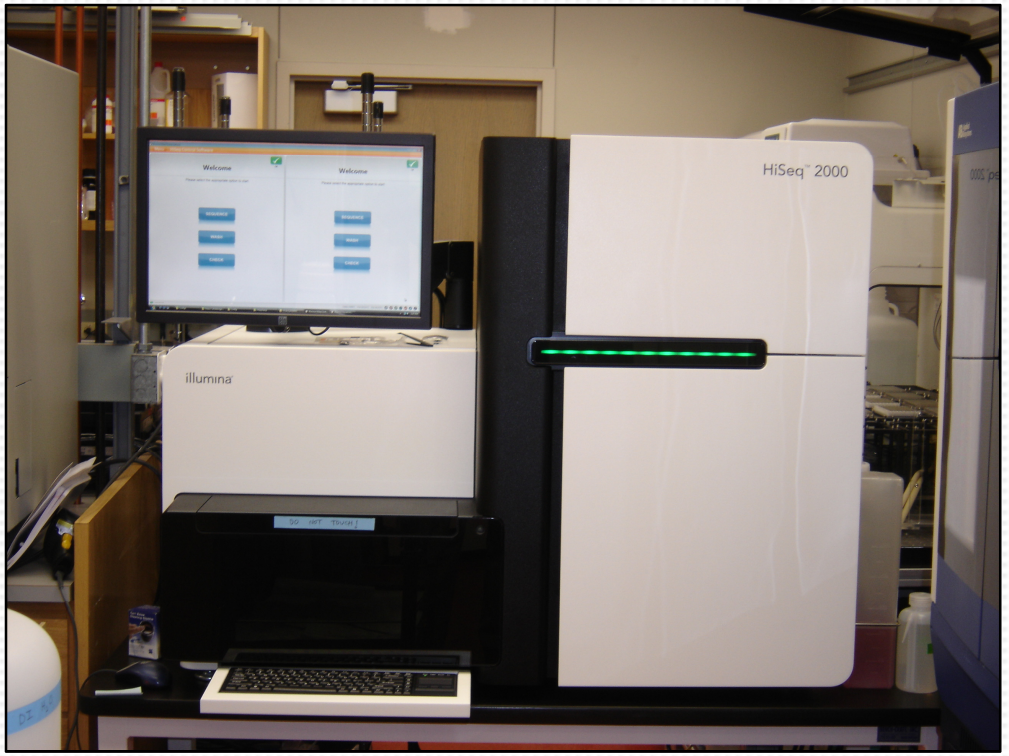
GS-FLX Library Prep

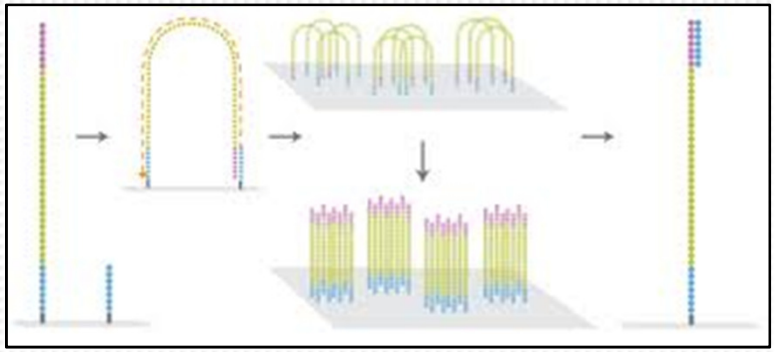
Illumina HiSEQ 2000



Sequence Clusters

Flow Cell





Cluster Generation

- Applications
  - Resequencing
  - ChipSeq
  - miRNA identification & quantification
  - Gene expression
  - Methylation profiling
  - Target capture

Illumina HiSEQ 2000 (short-read system)

2 Flowcell System

- **8 lanes/flowcell**
- **Read Length: 35-100 bp**
- **No. of Reads/Lane: >150M**
- **Accuracy**
  - >85% bases higher than Q30 at 2 x 50 nt
  - >80% bases higher than Q30 at 2 x 100 nt

- **Total: Up to 3 Billion SE or 6 Billion PE reads**
- **Run Times: days to weeks**
  - 1 x 35 nt = ~1.5 days
  - 2 x 100 nt = ~8 - 11 days

Illumina MiSeq v2

Read Length	Total Time	MiSeq v2 Output
1 × 36	~4 hrs	540-610 Mb
2 × 25	~5.5 hrs	750-850 Mb
2 × 100	~16.5 hrs	3.0-3.4 Gb
2 × 150	~24 hrs	4.5-5.1 Gb
2 × 250	~39 hrs	7.5-8.5 Gb

Longer and More Reads – a new 500-cycle reagent kit supports 2 x 250 nt runs

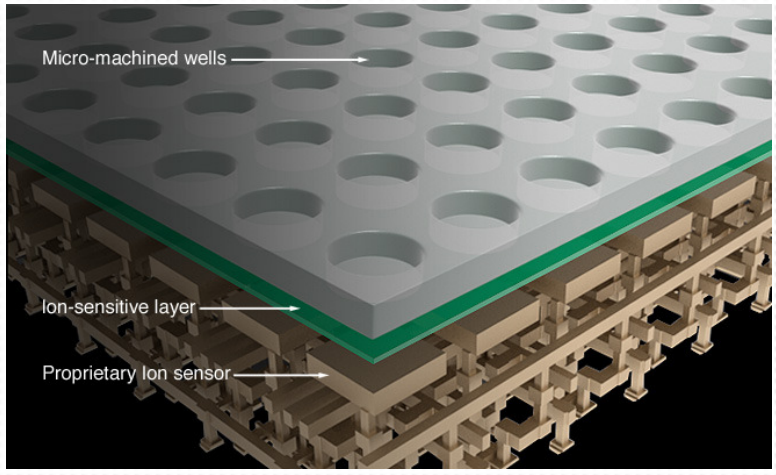
Ion Torrent Personal Genome Machine

Pairs semiconductor technology with sequencing chemistry

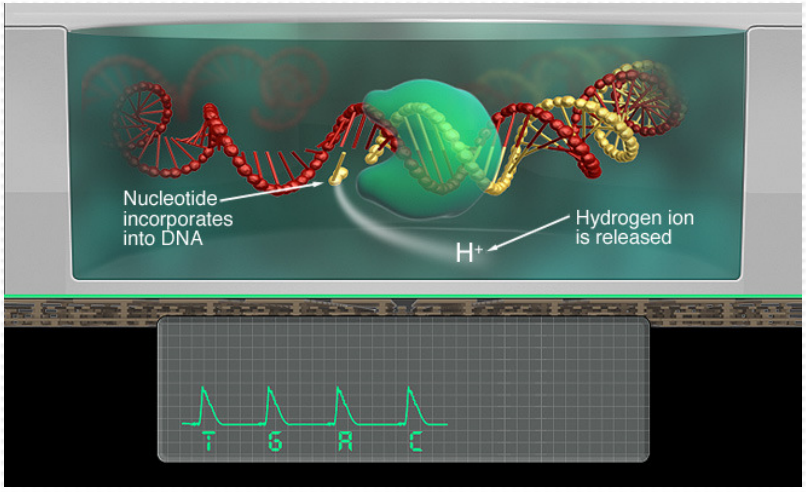
High-density array of micro-machined wells to perform reactions in a massively parallel way and beneath the wells is an ion-sensitive layer and beneath that a proprietary ion sensor.

When a nucleotide is incorporated into a strand of DNA by polymerase, a hydrogen ion is released as a byproduct.

PGM Sequencer

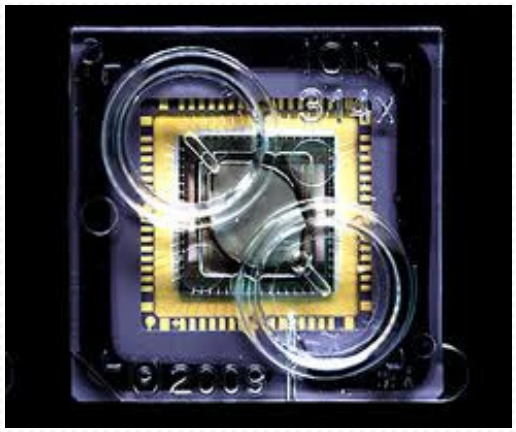


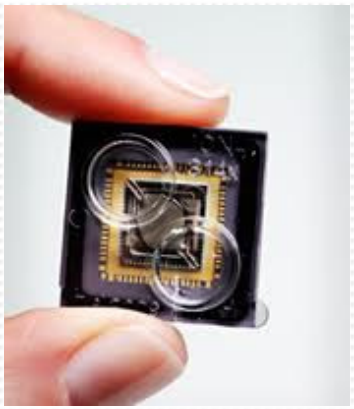
Cross-Section of Flow Cell



Addition of a Base Gives Signal

Ion Torrent Personal Genome Machine





Applications:

- Small Genome Sequencing
- Targeted (Amplicon) Re-sequencing
- Target Capture
- Copy Number
- Chip-Seq
- RNA-Seq
- Bar Codes, Paired End Reads

314 Chip give ~1 million reads/run
316 & 318 Chips give ~2 million reads/run

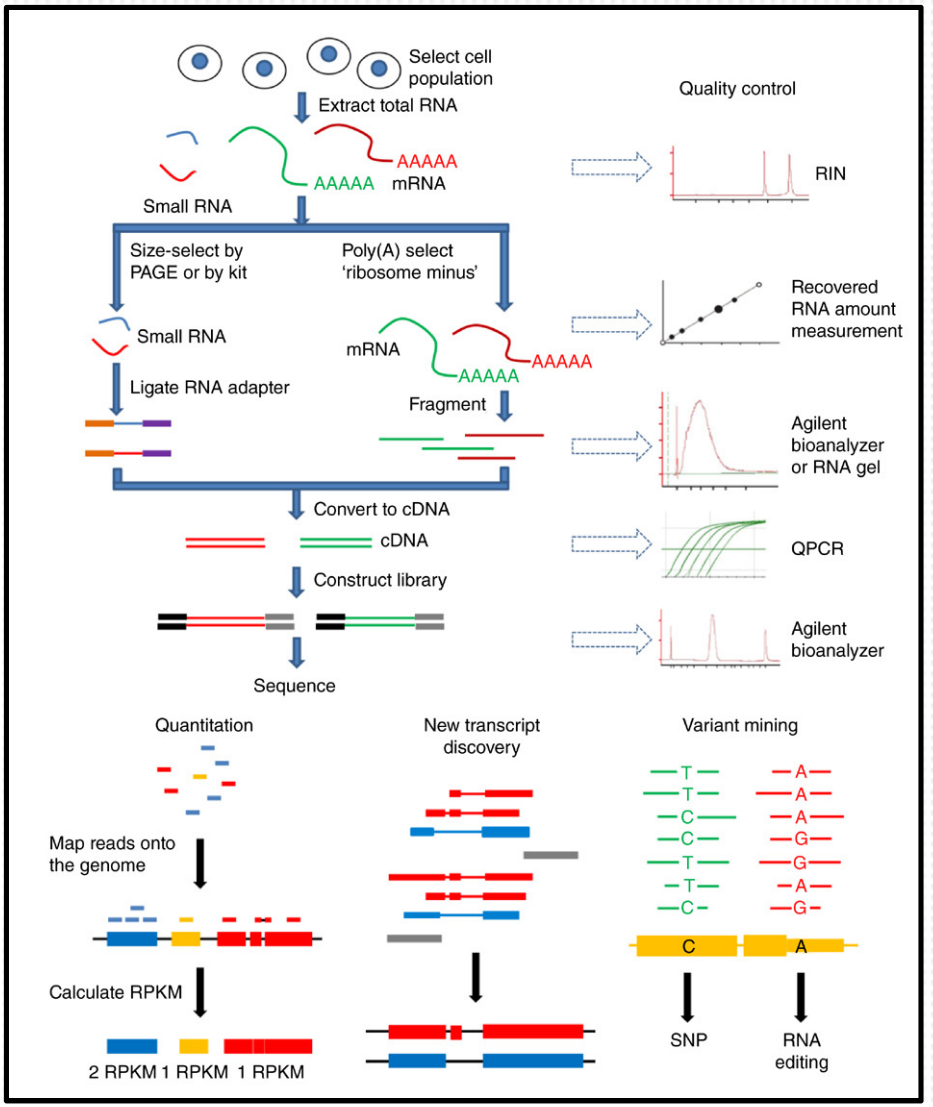
Ion Semiconductor Sequencing Chip	Output	Read Length		Total Sequencing Time
		2011	2012	
314	> 10Mb			
316	> 100Mb	> 200bp	> 400bp	< 2 hours
318	> 1Gb			
Accuracy:	>99.99% consensus accuracy and >99.5% raw accuracy.			

RNA-Seq (Whole Transcriptome Shotgun Sequencing)

Uses massively parallel DNA sequencing to sequence cDNA to acquire information about a sample's RNA content.

Depending on depth of coverage, RNA-Seq can be used to measure:

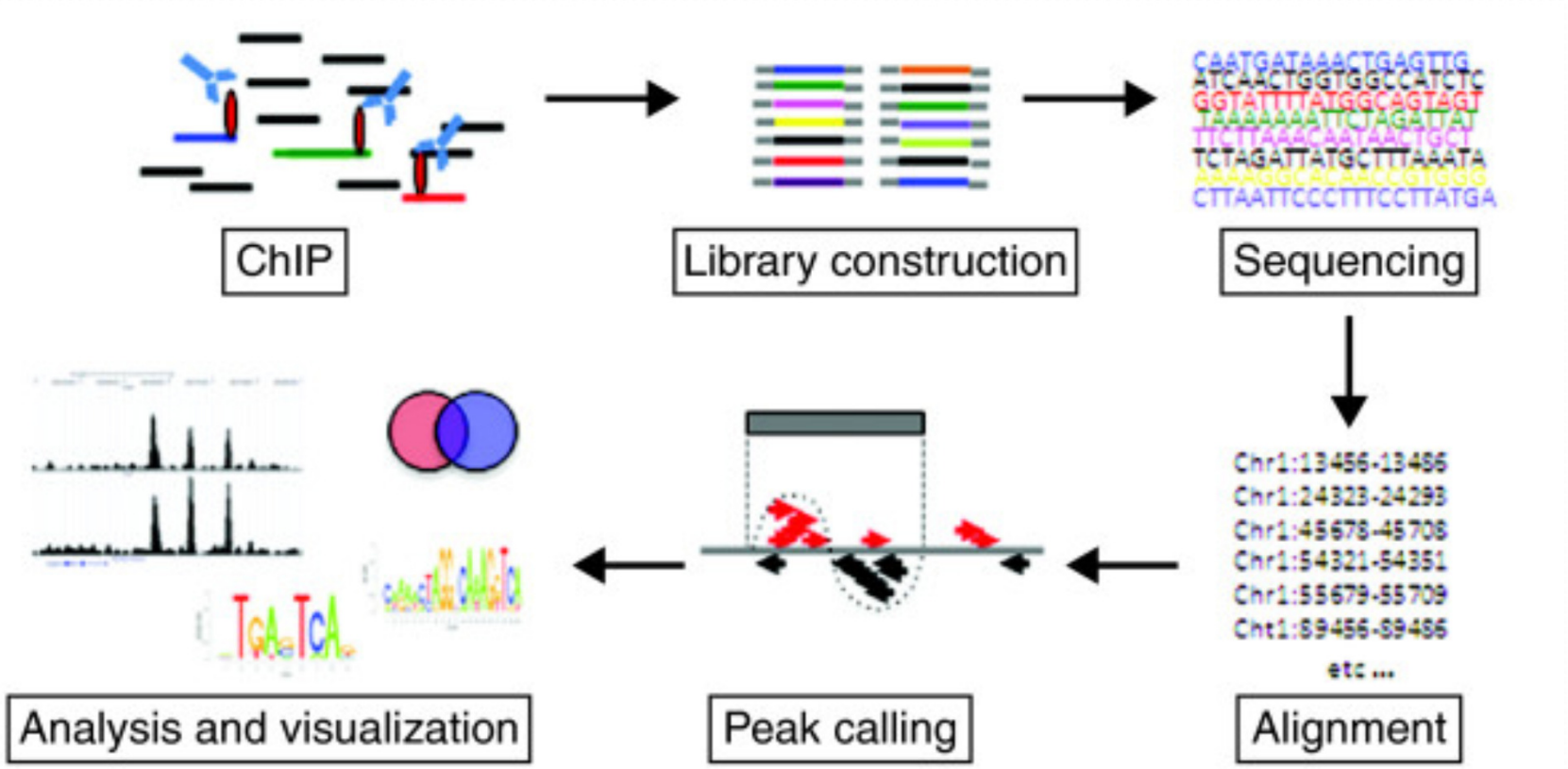
- Differential expression of genes
  - Gene alleles
  - Differently spliced transcripts
- Non-coding RNAs
- Post-transcriptional mutations/editing
- Gene fusions



ChIP-Seq

Used to analyze protein interactions with DNA

Combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins



Sequence Target Capture

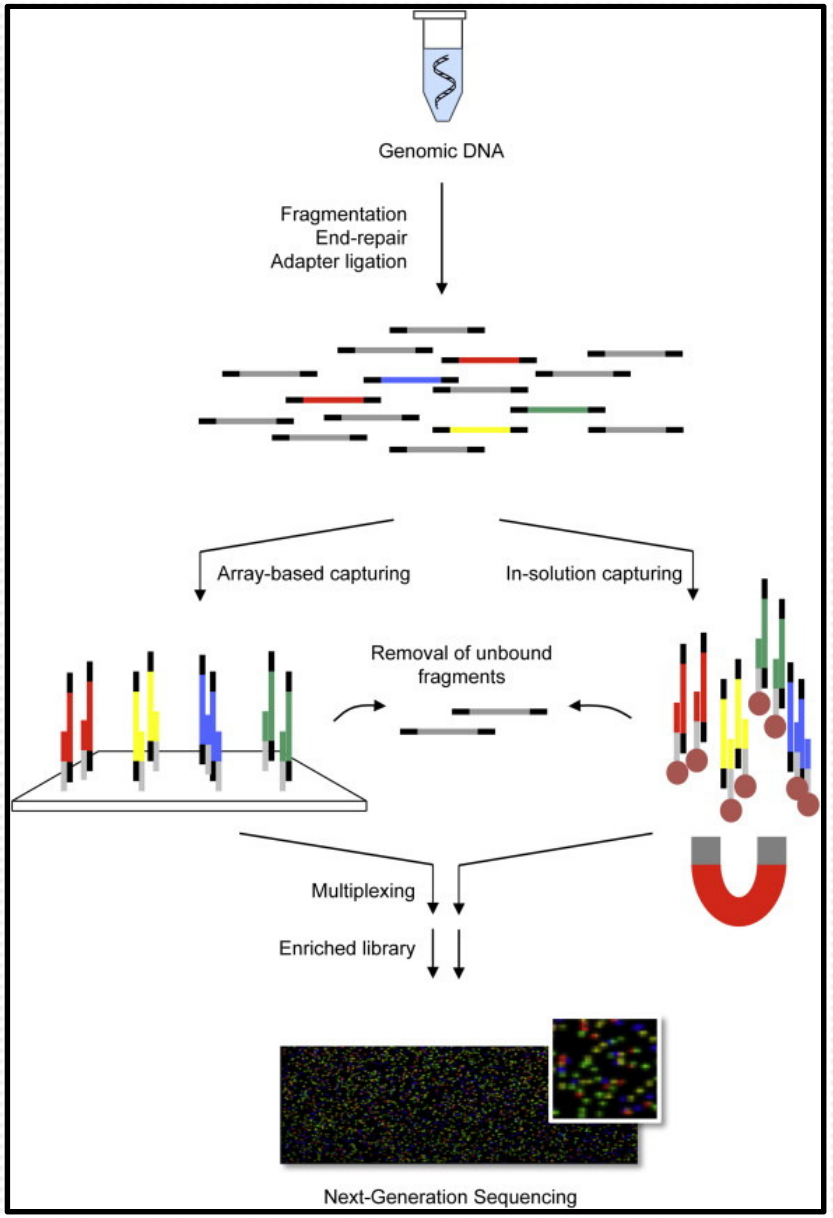
A process for the enrichment of selected genomic regions from the full complexity an entire genome

<u>Why Target</u>?

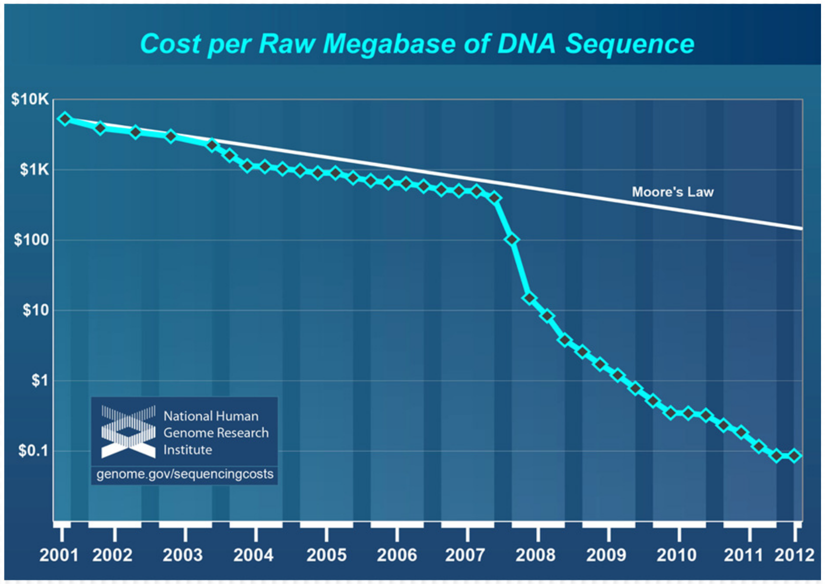
- **Can focus on only regions or genes of interest**
- **It still costs too much to sequence the entire genome**
- **Can process more samples per sequencing run**

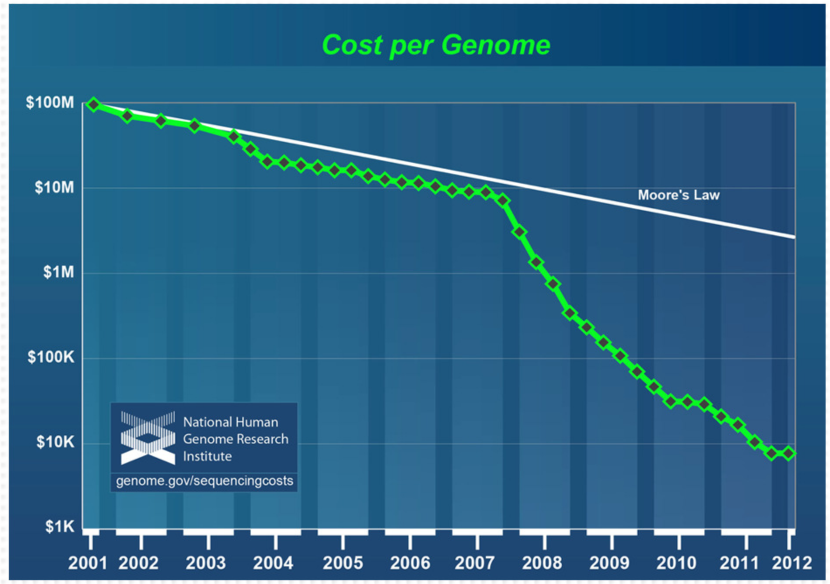
<u>Potential Issues</u>

- **Poor or no capture of some targets**
- **Sequencing whole genome may soon be cost effective**
- **For Exome products, not all genes are targeted**



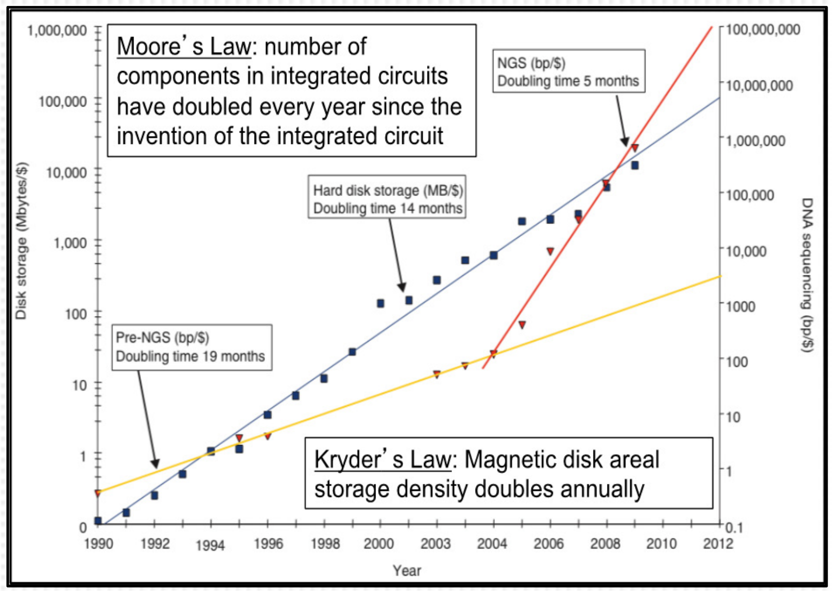
Paradigm Shift in DNA Sequencing Costs





Date	Cost per Mb	Cost per Genome	Date	Cost per Mb	Cost per Genome
Sep-01	$5,292.39	$95,263,072	Jul-07	$495.96	$8,927,342
Mar-02	$3,898.64	$70,175,437	Oct-07	$397.09	$7,147,571
Sep-02	$3,413.80	$61,448,422	Jan-08	$102.13	$3,063,820
Mar-03	$2,986.20	$53,751,684	Apr-08	$15.03	$1,352,982
Oct-03	$2,230.98	$40,157,554	Jul-08	$8.36	$752,080
Jan-04	$1,598.91	$28,780,376	Oct-08	$3.81	$342,502
Apr-04	$1,135.70	$20,442,576	Jan-09	$2.59	$232,735
Jul-04	$1,107.46	$19,934,346	Apr-09	$1.72	$154,714
Oct-04	$1,028.85	$18,519,312	Jul-09	$1.20	$108,065
Jan-05	$974.16	$17,534,970	Oct-09	$0.78	$70,333
Apr-05	$897.76	$16,159,699	Jan-10	$0.52	$46,774
Jul-05	$898.90	$16,180,224	Apr-10	$0.35	$31,512
Oct-05	$766.73	$13,801,124	Jul-10	$0.35	$31,125
Jan-06	$699.20	$12,585,659	Oct-10	$0.32	$29,092
Apr-06	$651.81	$11,732,535	Jan-11	$0.23	$20,963
Jul-06	$636.41	$11,455,315	Apr-11	$0.19	$16,712
Oct-06	$581.92	$10,474,556	Jul-11	$0.12	$10,497
Jan-07	$522.71	$9,408,739	Oct-11	$0.09	$7,743
Apr-07	$502.61	$9,047,003	Jan-12	$0.09	$7,666

National Human Genome Research Institute
genome.gov/sequencingcosts



http://www.genome.gov/sequencingcosts/

Lincoln Stein, *Genome Biology 2010, 11:207*

Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

"I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients.
… patients included a child with irritable bowel disease, a child with severe combined immunodeficiency, two siblings affected with Miller syndrome, and several with cancers of different types.

…each presenter emphasized the rapidity with which these data can now be generated using next-generation sequencing instruments, they also listed the large number of people involved in the analysis of these datasets. The required expertise to 'solve' each case included molecular and computational biologists, geneticists, pathologists and physicians with exquisite knowledge of the disease and of treatment modalities, research nurses, genetic counselors, and IT and systems support specialists, among others.

…although the idea of clinical whole-genome sequencing for diagnosis is exciting and potentially life-changing …one does wonder how … such a 'dream team' of specialists would be assembled for each case. **In other words, even if the cost and speed of generating sequencing data continue their precipitous decreases, the cost of 'team' analysis seems unlikely to immediately follow suit.**"

Group Discussion

- What NGS projects have you done?
- How did you analyze your data?
- Data storage?
- What NGS projects do you hope to work on?