An Introduction to Second-Generation Sequencing

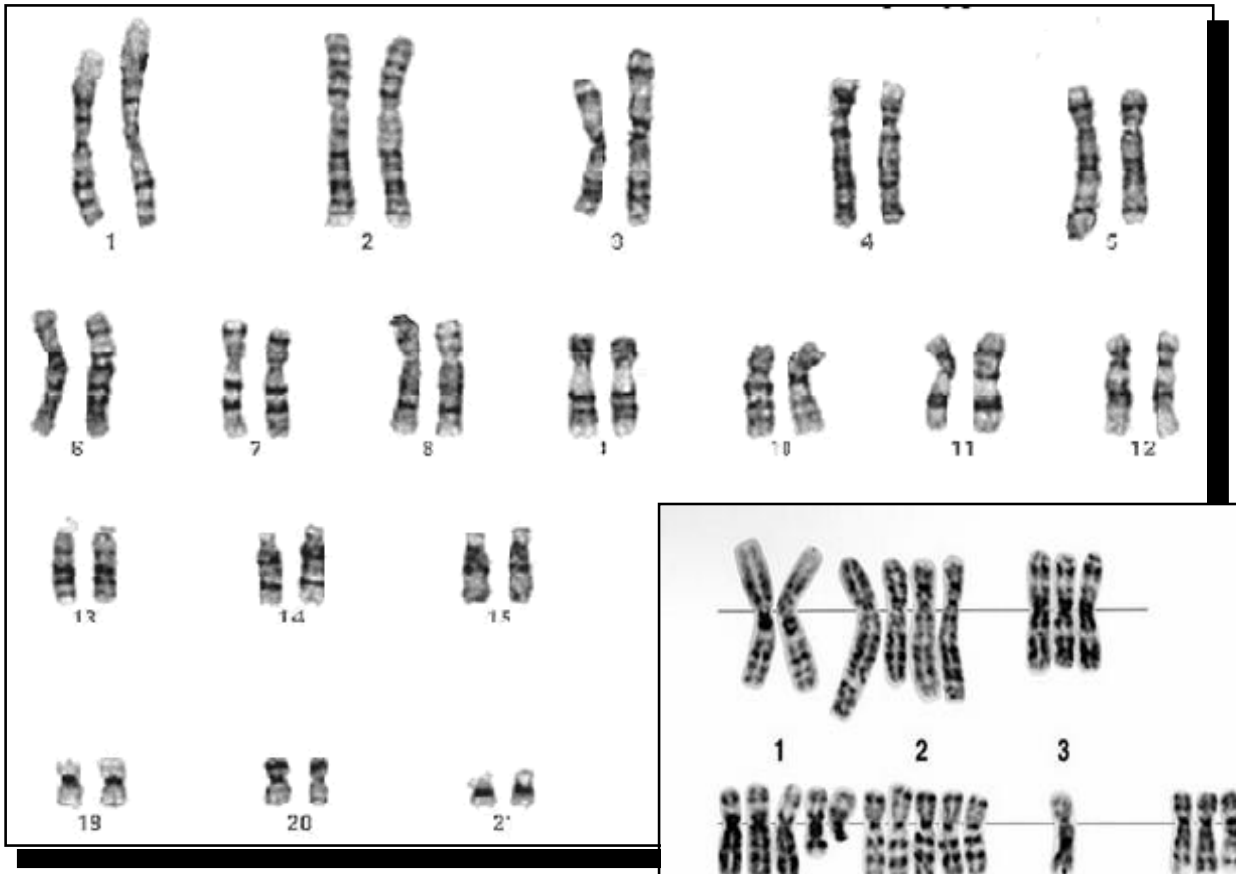
Using Helix and Biowulf

Sean Davis, M.D., Ph.D.

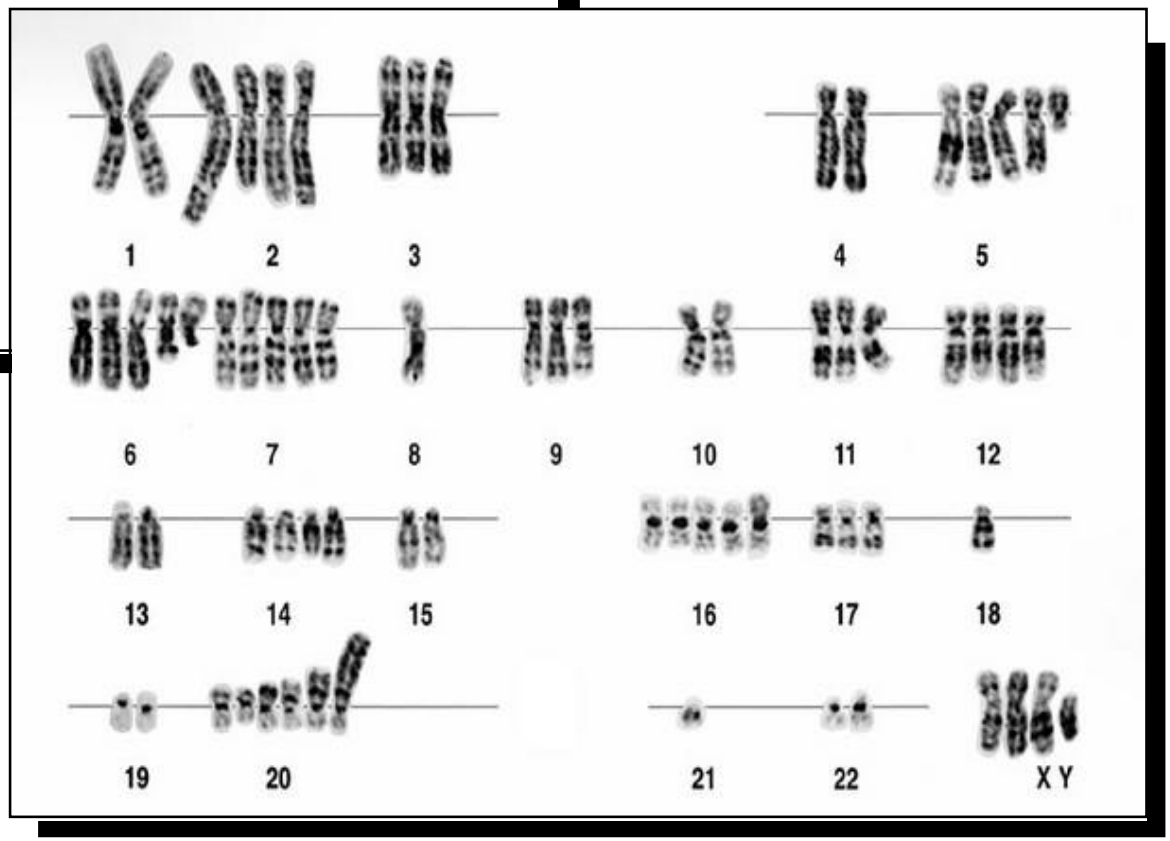
Genetics Branch, Center for Cancer Research

National Cancer Institute

National Institutes of Health

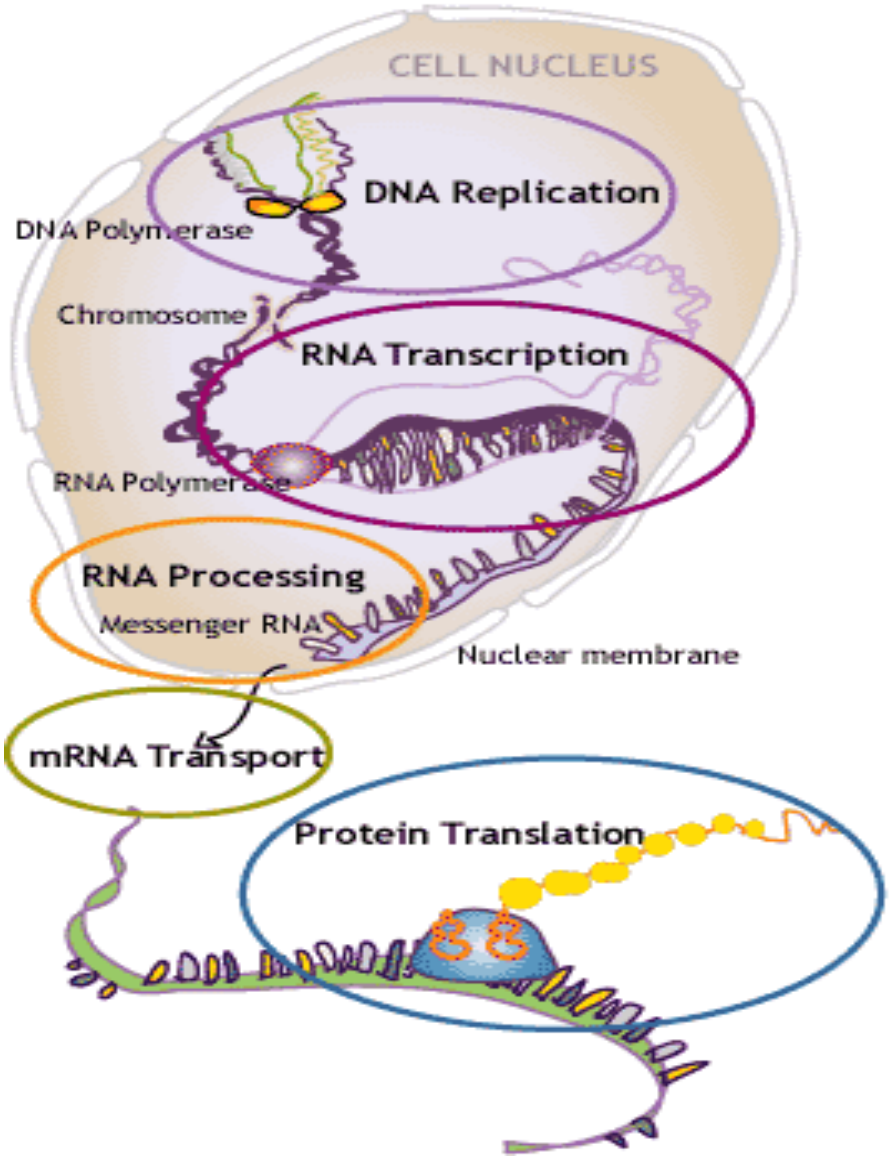


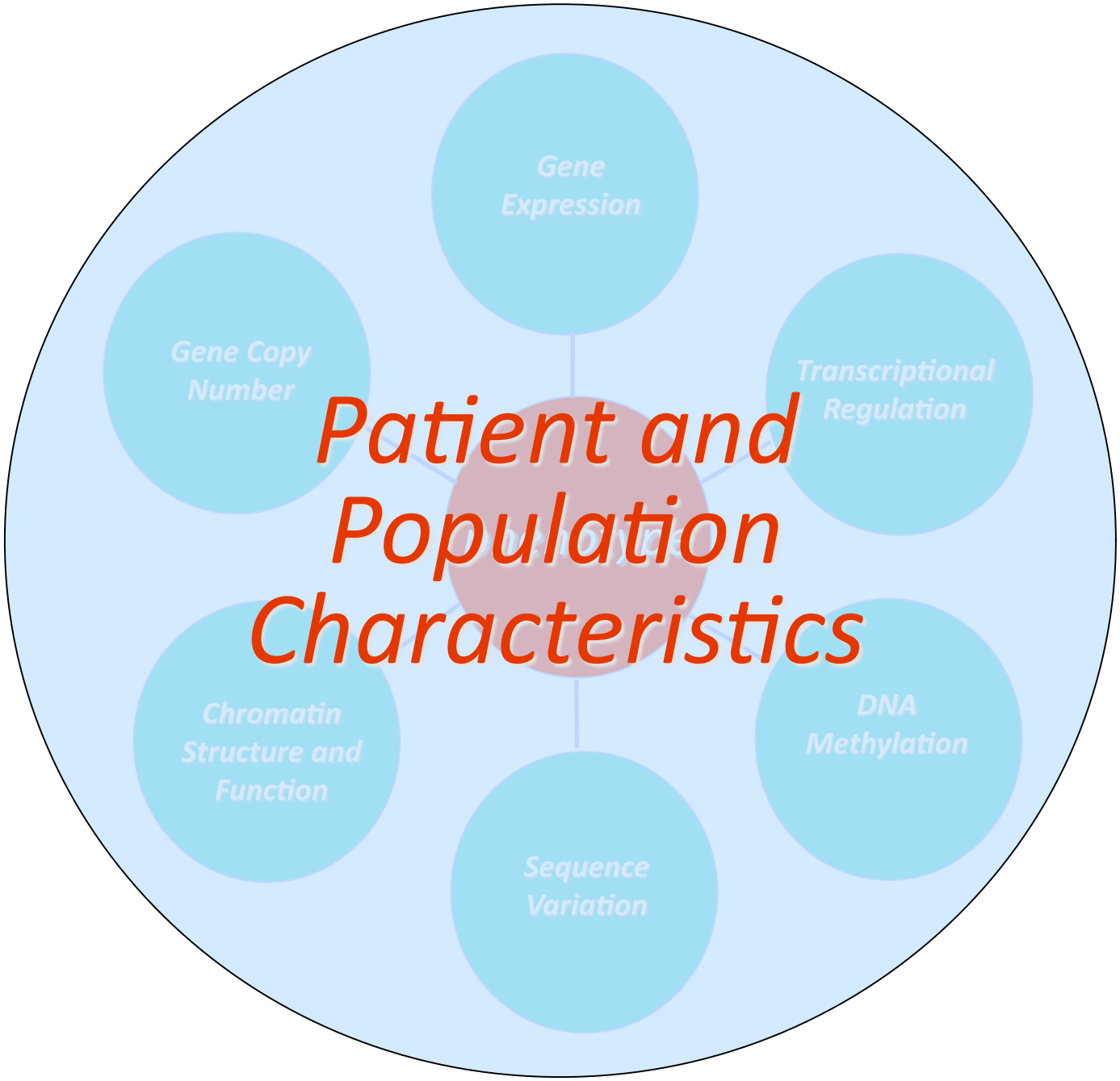
Normal Karyotype



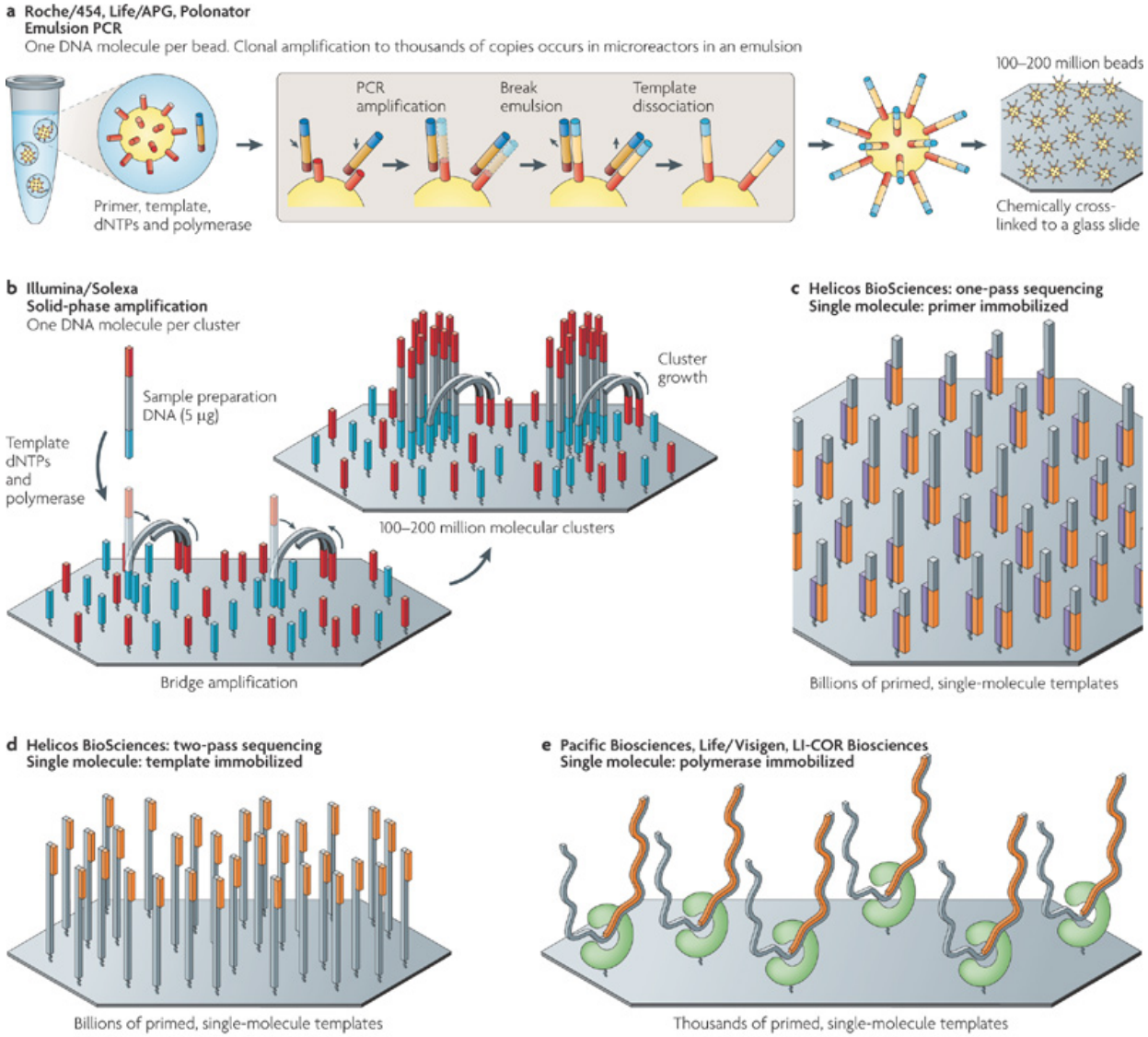
Tumor Karyotype

The Central Dogma





Patient and Population Characteristics


a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a glass slide
b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification
c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized
Billions of primed, single-molecule templates
d Helicos BioSciences: two-pass sequencing
Single molecule: template immobilized
Billions of primed, single-molecule templates
e Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized
Thousands of primed, single-molecule templates

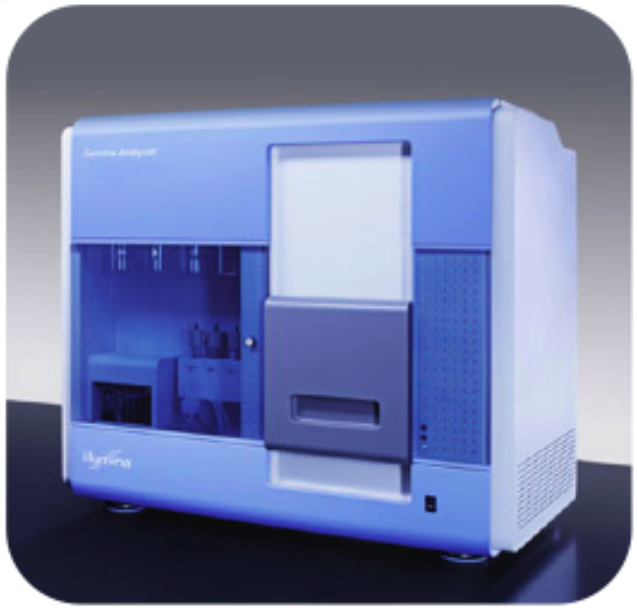
Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent 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.

Illumina GA2

Setting the Standard in "Now Generation" Sequencing

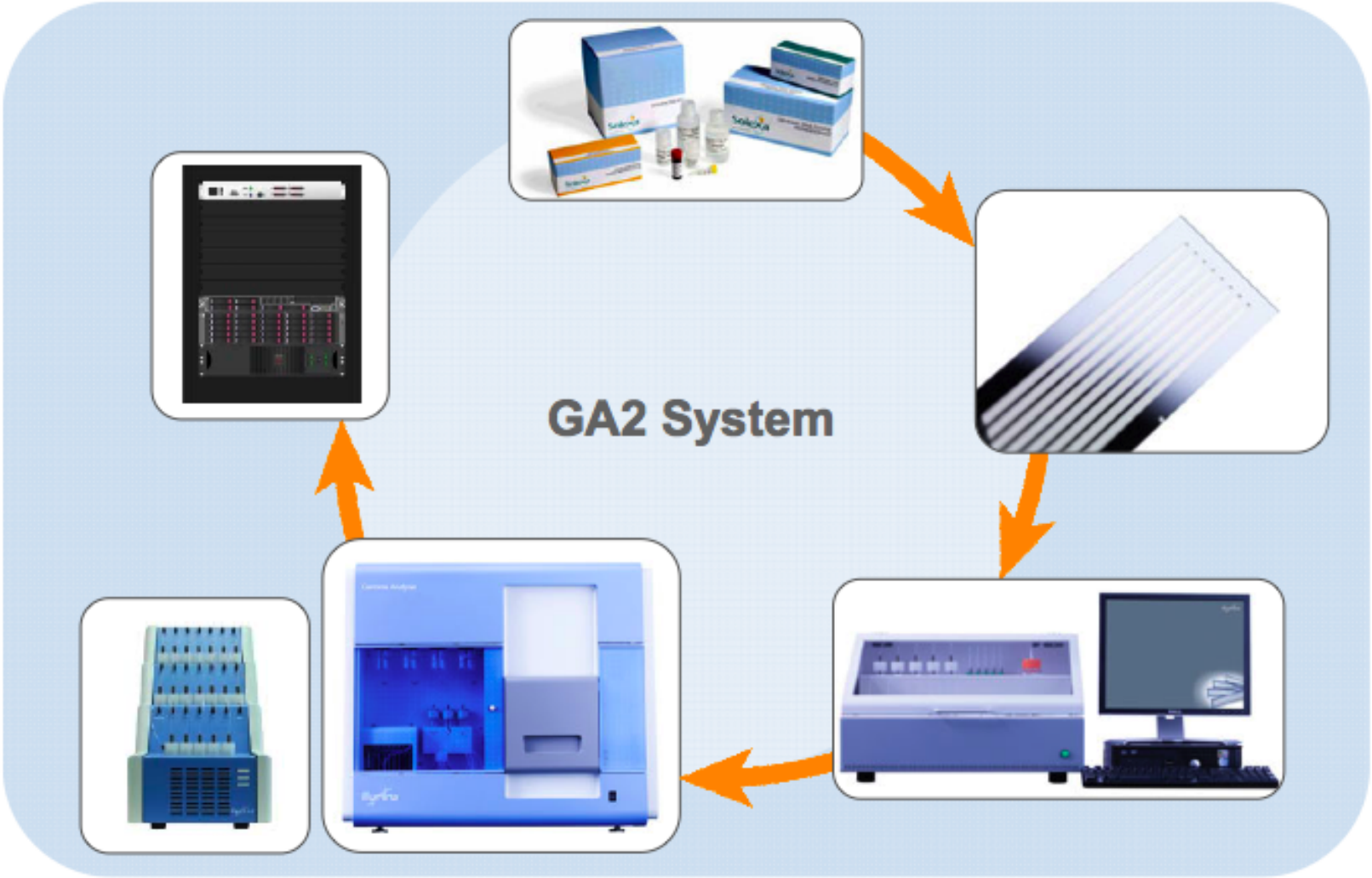
Industry Leading Ease of Use…



- Highest accuracy
- Highest percentage of perfect reads
- 1.5Gb+ single read in 2.5 days
- 3.0Gb+ paired read in 5 days
- 600MB per day today (minimum performance specification)
  - Supports future expansion
- Lowest operating cost
- >250 instruments installed

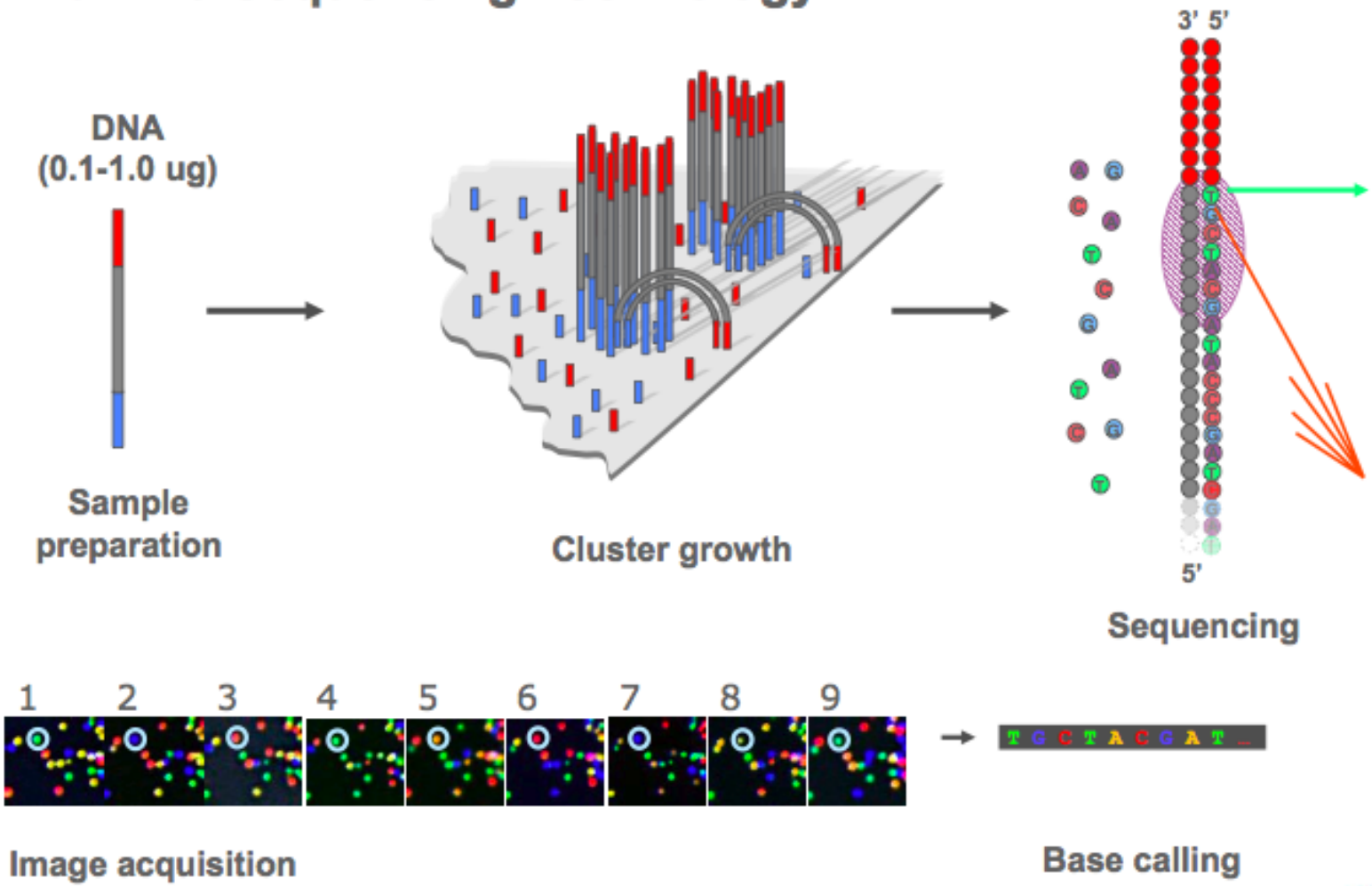
…with Applications Flexibility

illumina

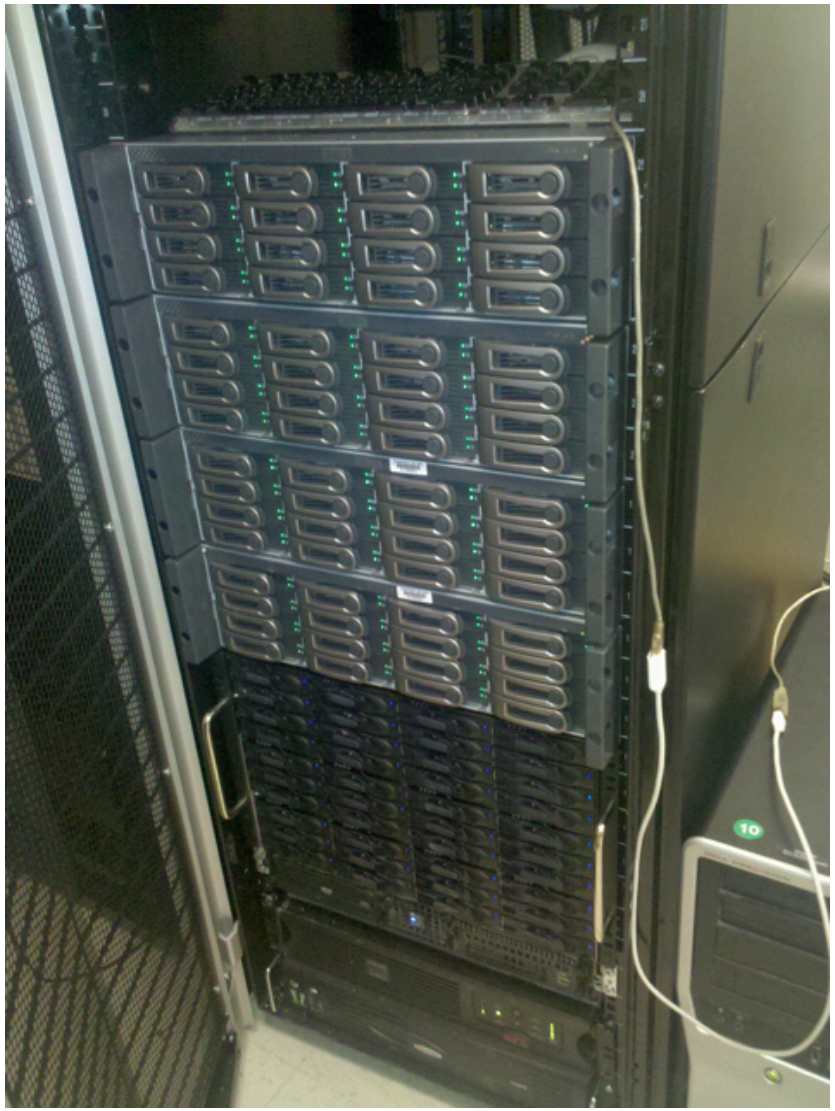


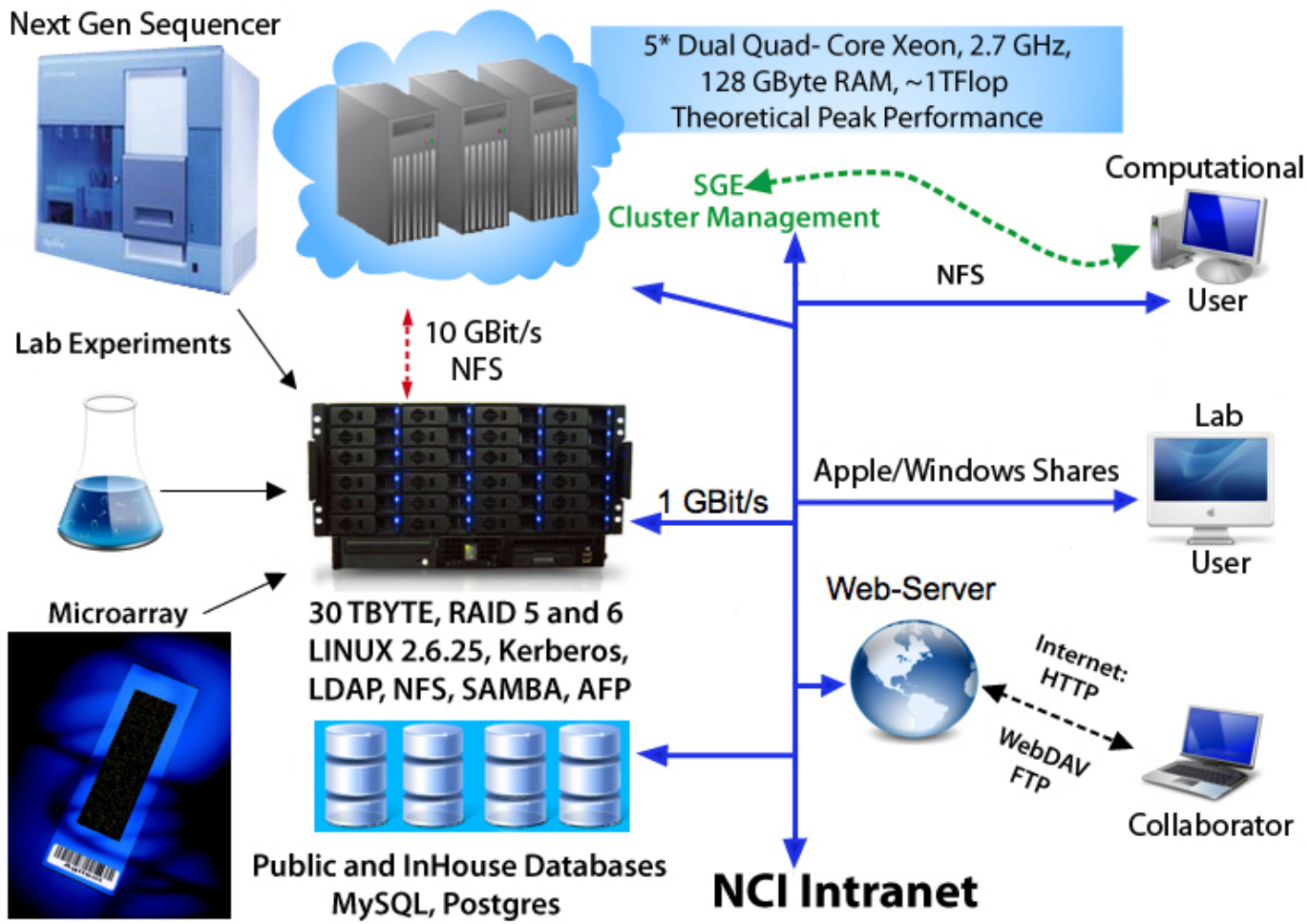
illumina

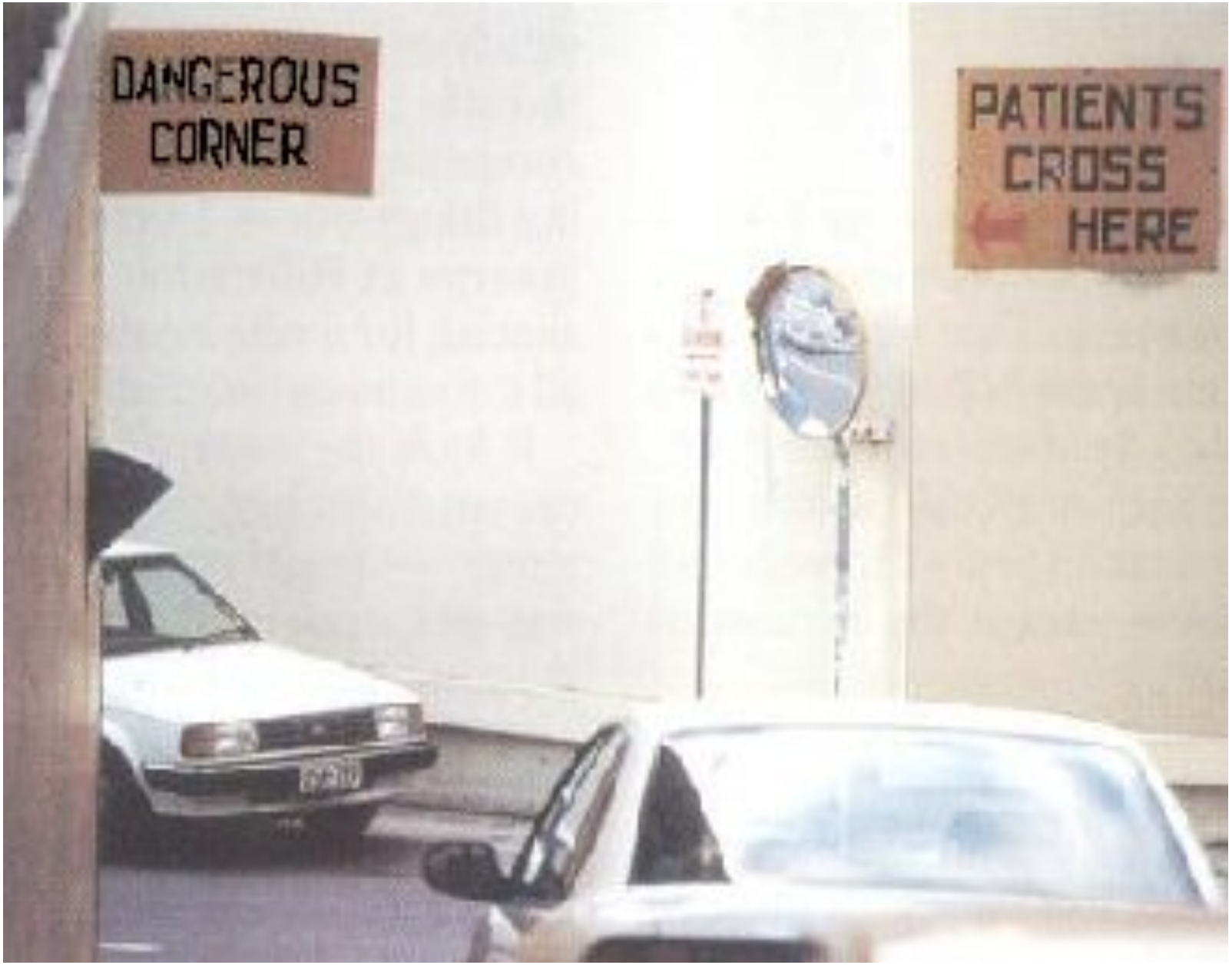
Illumina Sequencing Technology



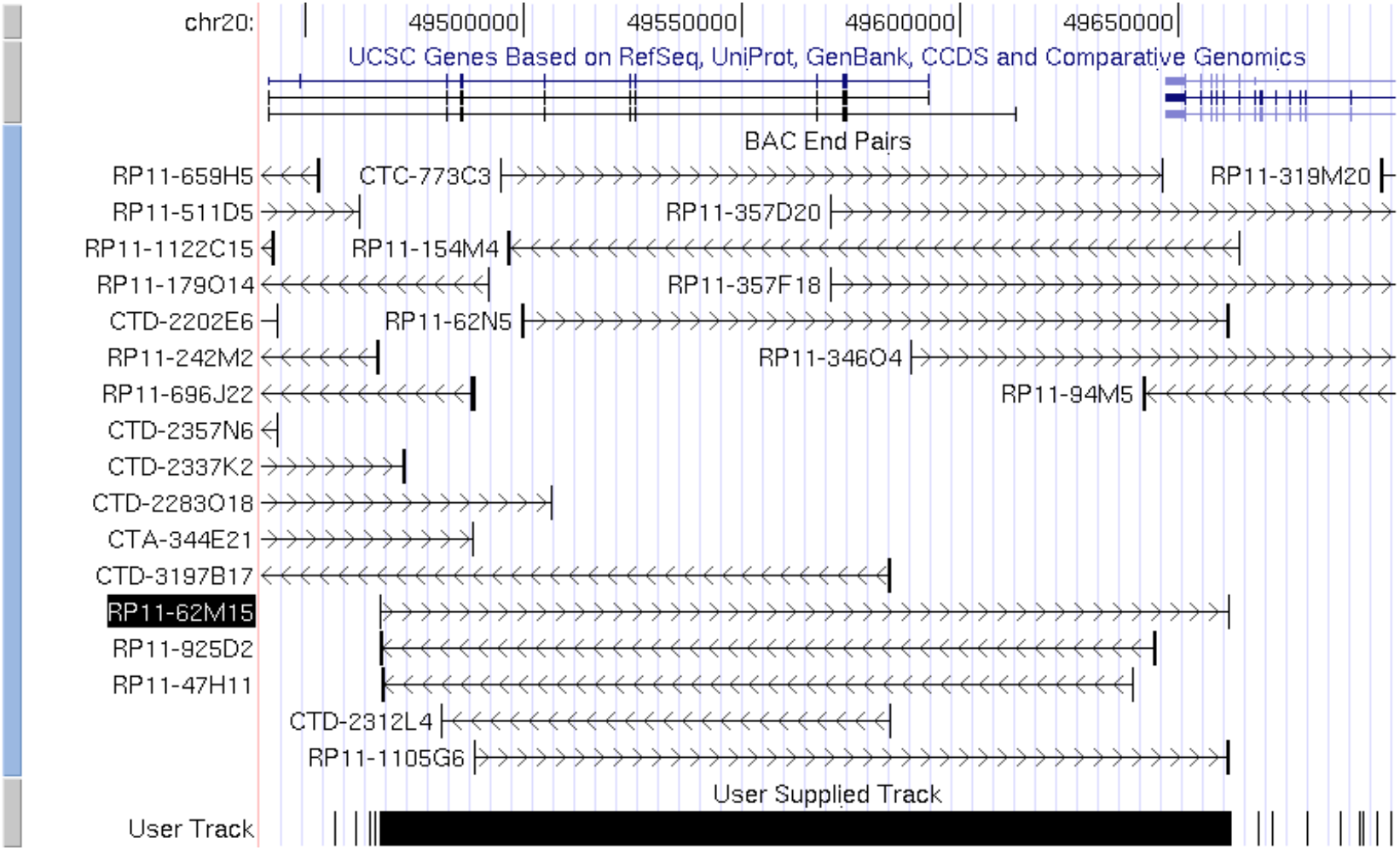
illumina



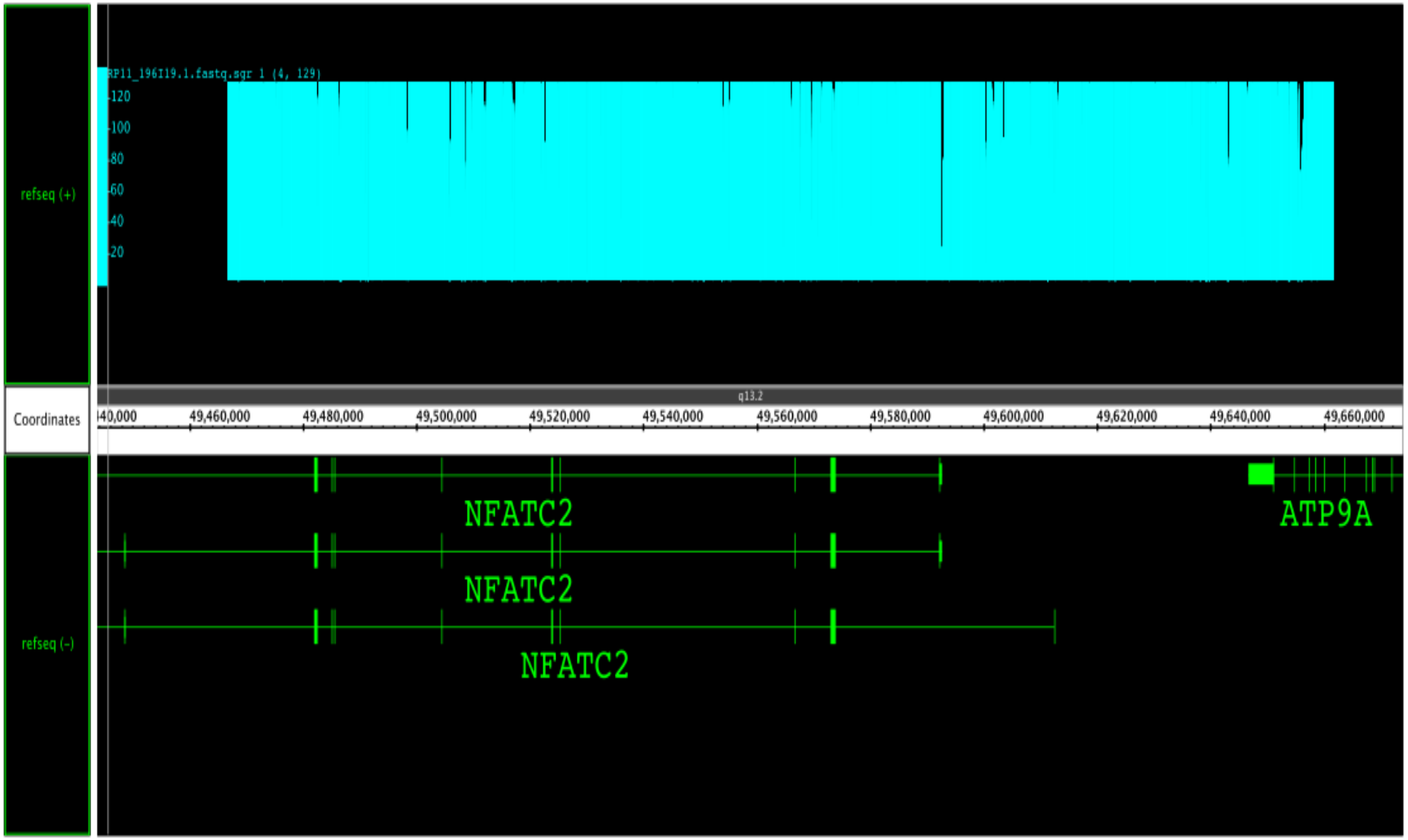

Next Gen Sequencer
5* Dual Quad- Core Xeon, 2.7 GHz,
128 GByte RAM, ~1TFlop
Theoretical Peak Performance
SGE
Cluster Management
Computational
NFS
User
Lab Experiments
10 GBit/s
NFS
Lab
Apple/Windows Shares
1 GBit/s
User
Microarray
30 TBYTE, RAID 5 and 6
LINUX 2.6.25, Kerberos,
LDAP, NFS, SAMBA, AFP
Web-Server
Internet:
HTTP
WebDAV
FTP
Collaborator
Public and InHouse Databases
MySQL, Postgres
NCI Intranet


DANGEROUS
CORNER
PATIENTS
CROSS
HERE

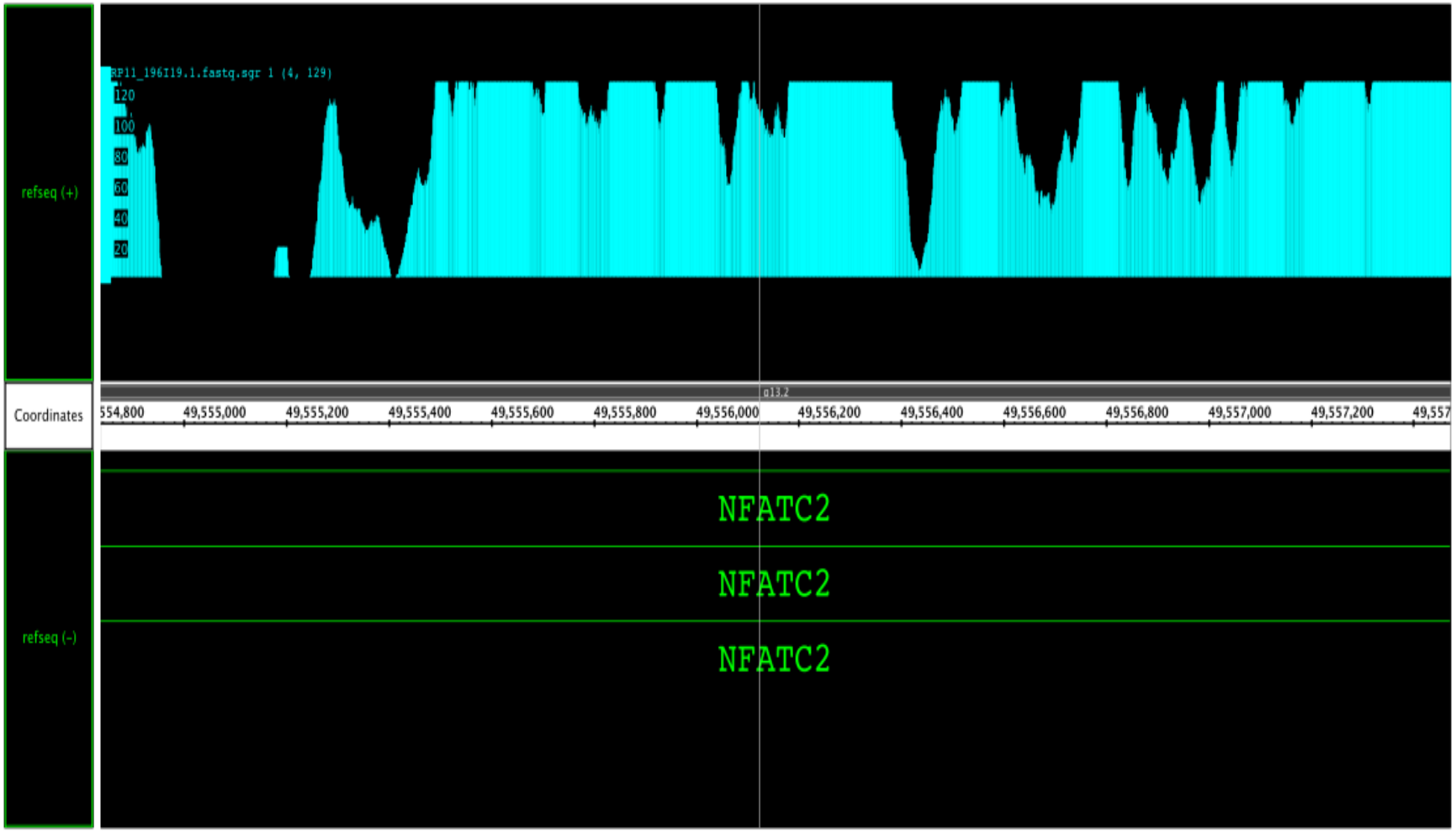
Overview of BAC in the Genome



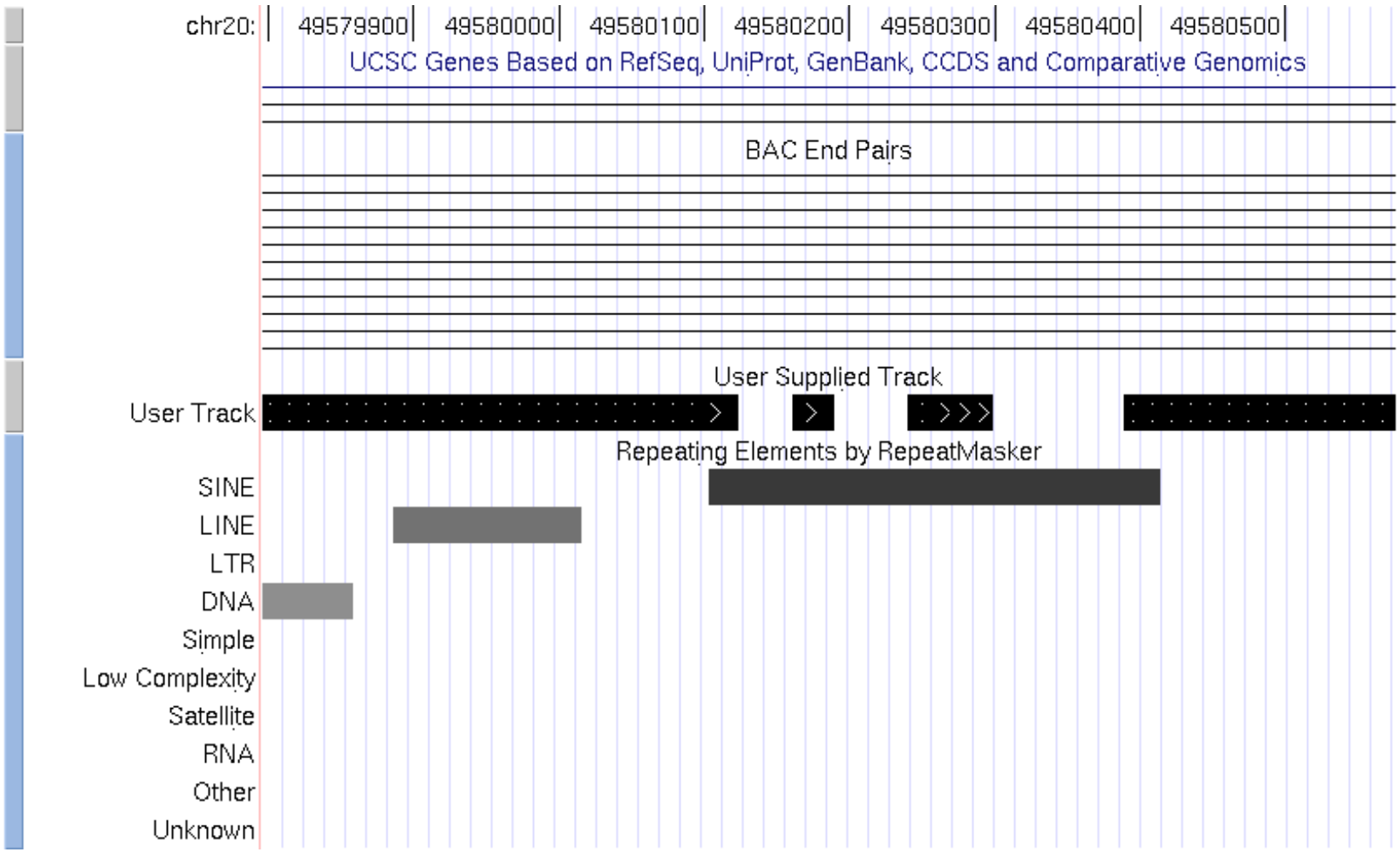
Sequencing a BAC



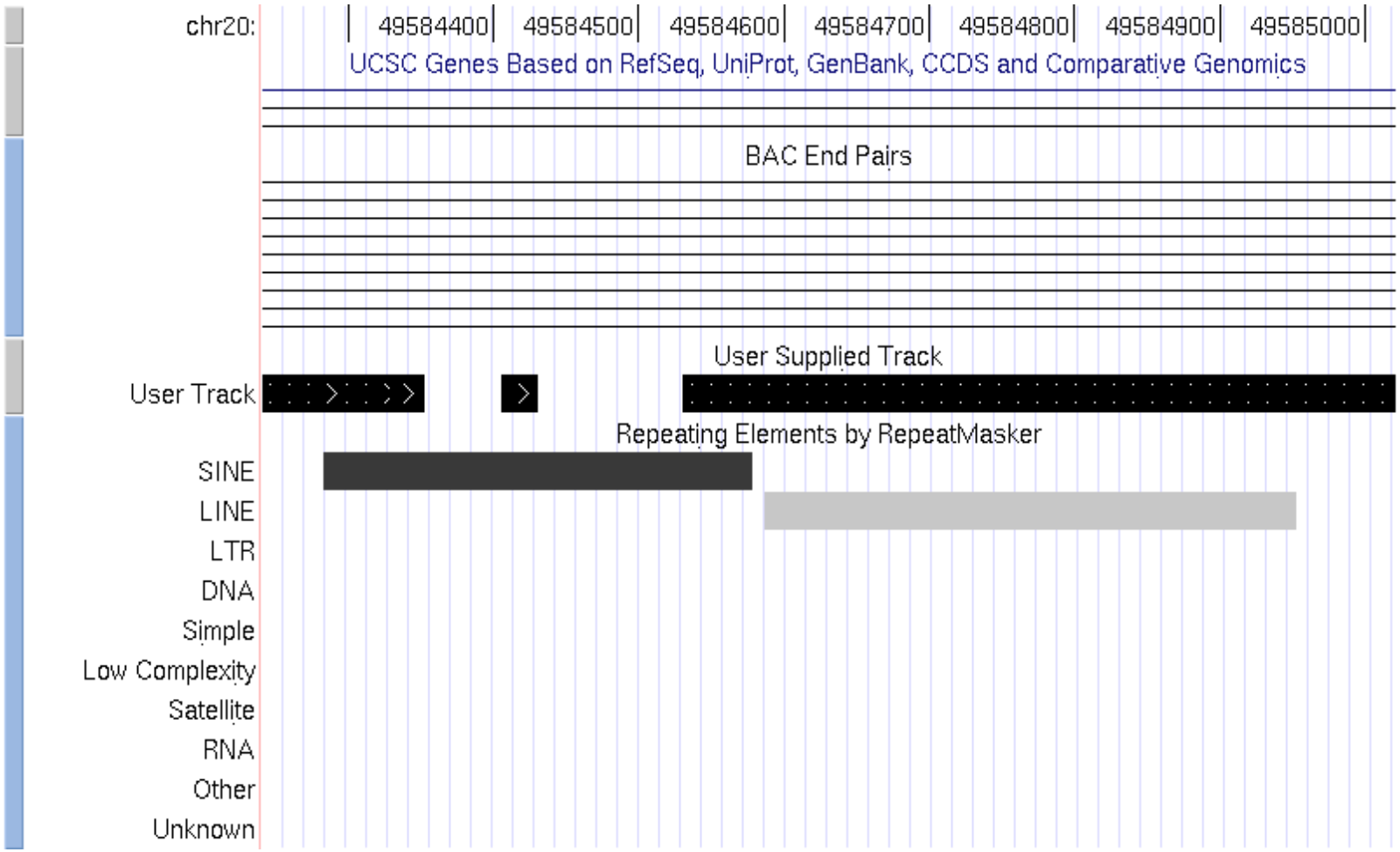
Sequence Coverage



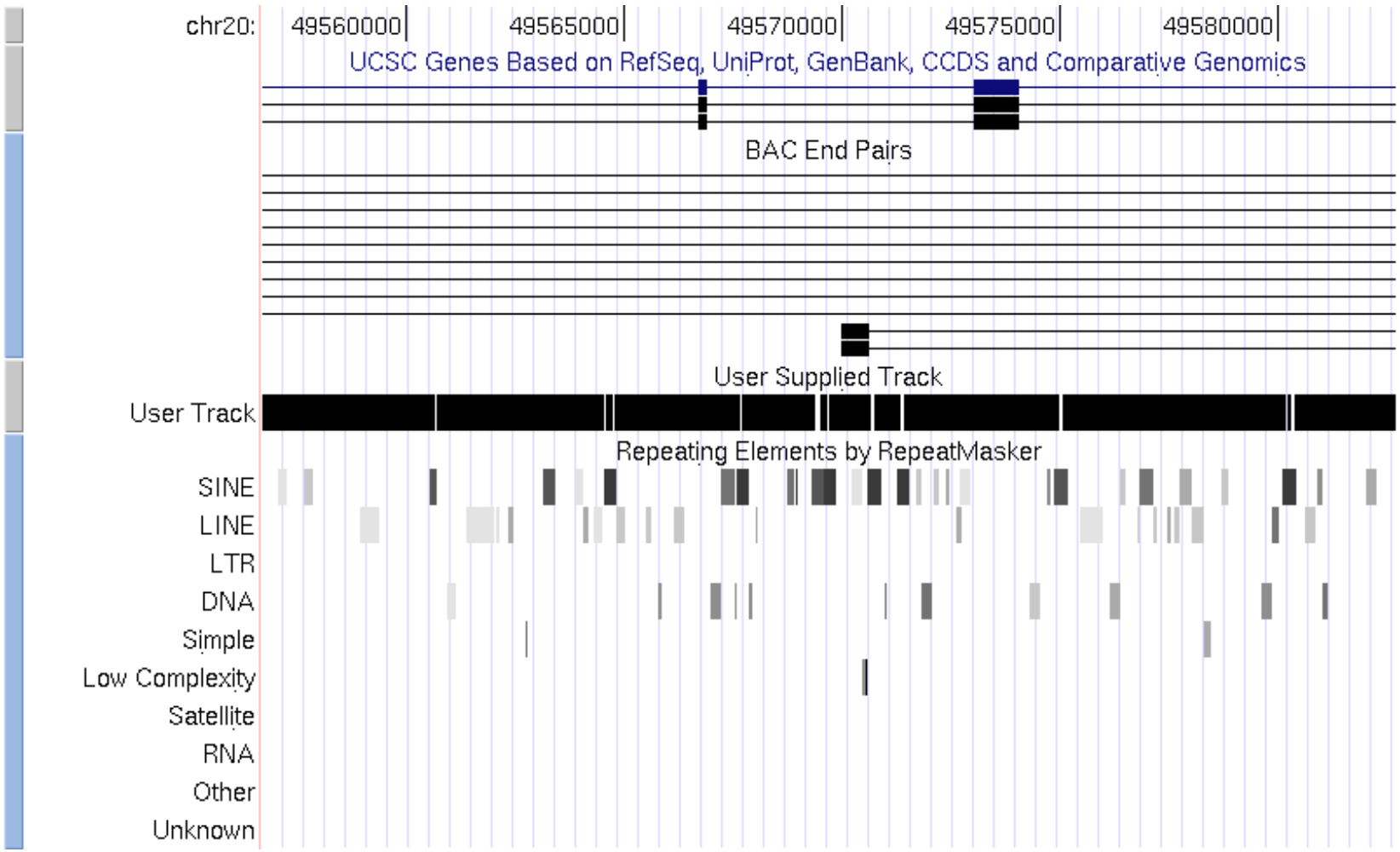
Repeats

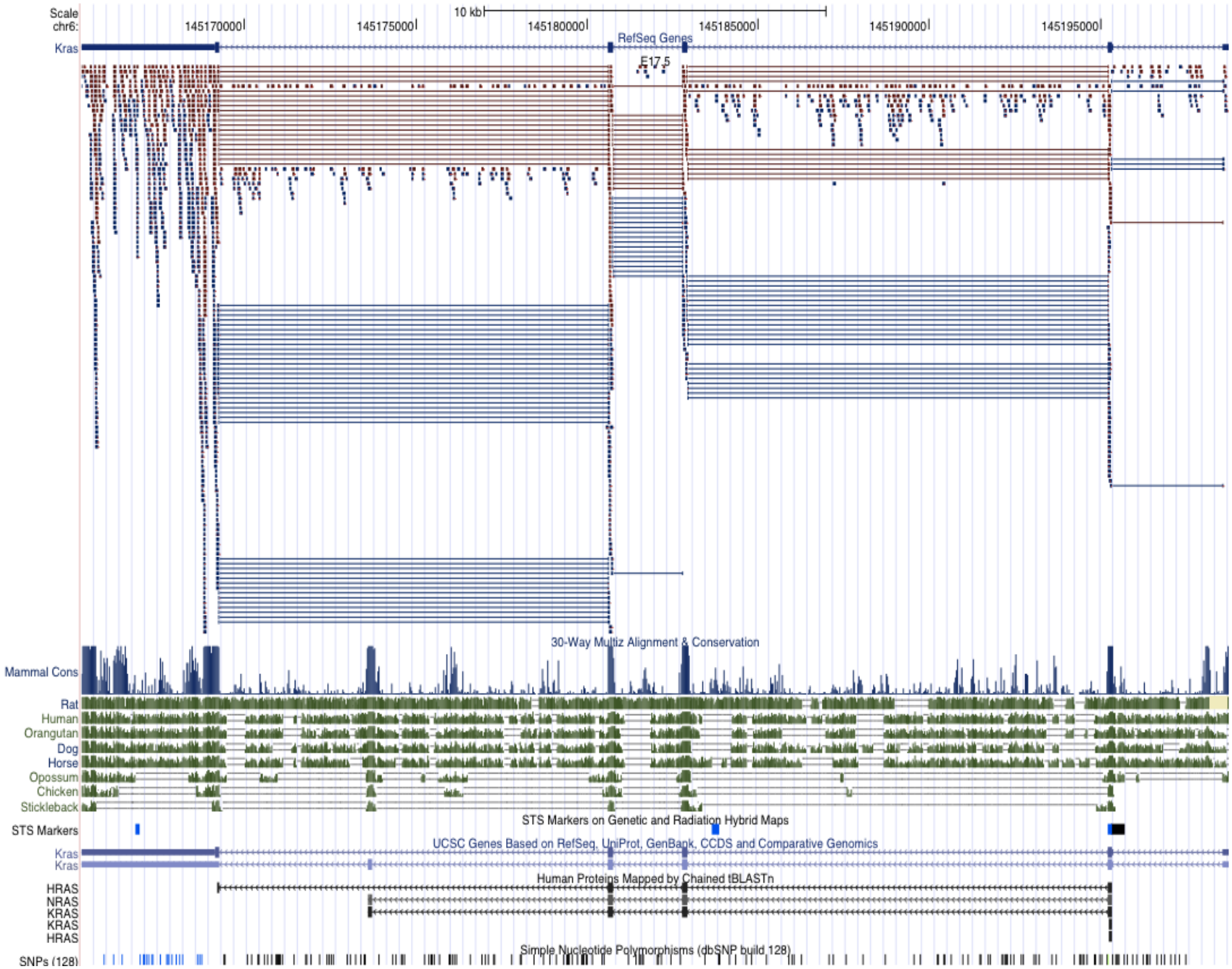

chr20:
49579900
49580000
49580100
49580200
49580300
49580400
49580500
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BAC End Pairs
User Supplied Track
User Track
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Repeats

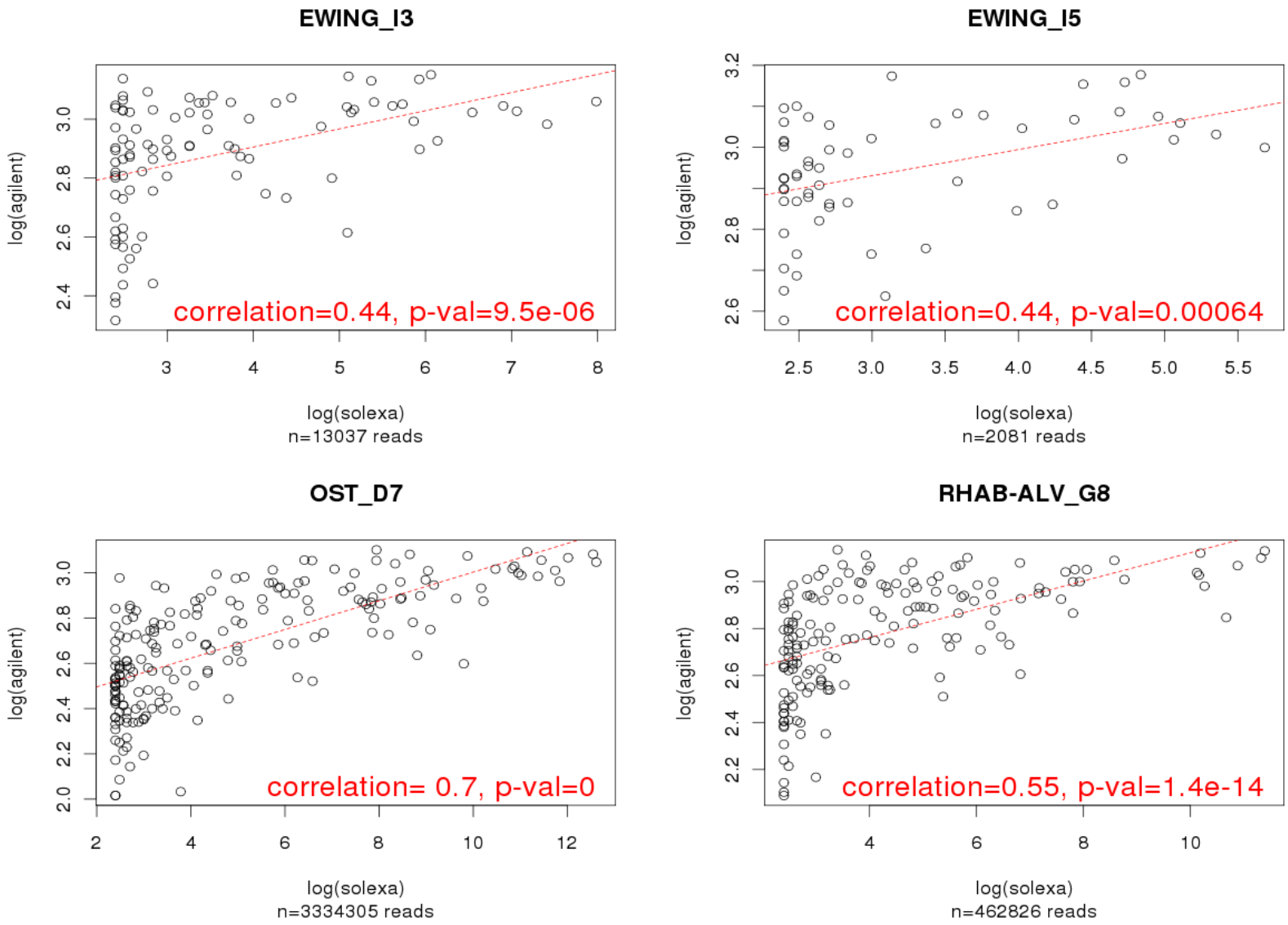

chr20:
49584400
49584500
49584600
49584700
49584800
49584900
49585000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BAC End Pairs
User Supplied Track
User Track
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Repeats are not created equal




Scale
chr6:
10 kb
145170000
145175000
145180000
145185000
145190000
145195000
Kras
RefSeq Genes
E17.5
30-Way Multiz Alignment & Conservation
Mammal Cons
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback
STS Markers
STS Markers on Genetic and Radiation Hybrid Maps
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Kras
Kras
Human Proteins Mapped by Chained tBLASTn
HRAS
NRAS
KRAS
KRAS
HRAS
Simple Nucleotide Polymorphisms (dbSNP build 128)
SNPs (128)

Sarcoma miRNA

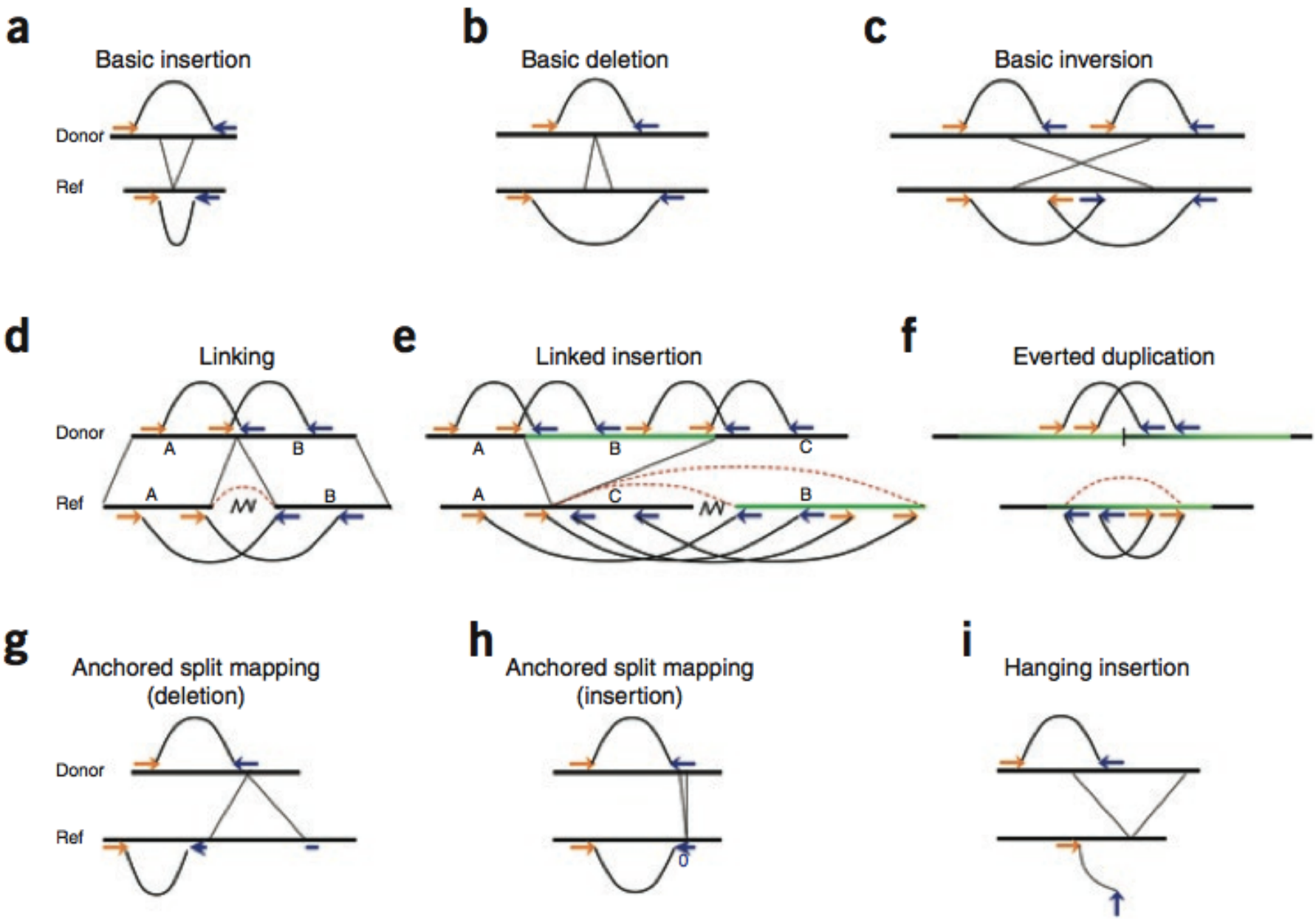


Computational methods for discovering structural variation with next-generation sequencing

Paul Medvedev[1], Monica Stanciu[1] & Michael Brudno[1,2]

In the last several years, a number of studies have described large-scale structural variation in several genomes. Traditionally, such methods have used whole-genome array comparative genome hybridization or single-nucleotide polymorphism arrays to detect large regions subject to copy-number variation. Later techniques have been based on paired-end mapping of Sanger sequencing data, providing better resolution and accuracy. With the advent of next-generation sequencing, a new generation of methods is being developed to tackle the challenges of short reads, while taking advantage of the high coverage the new sequencing technologies provide. In this survey, we describe these methods, including their strengths and their limitations, and future research directions.

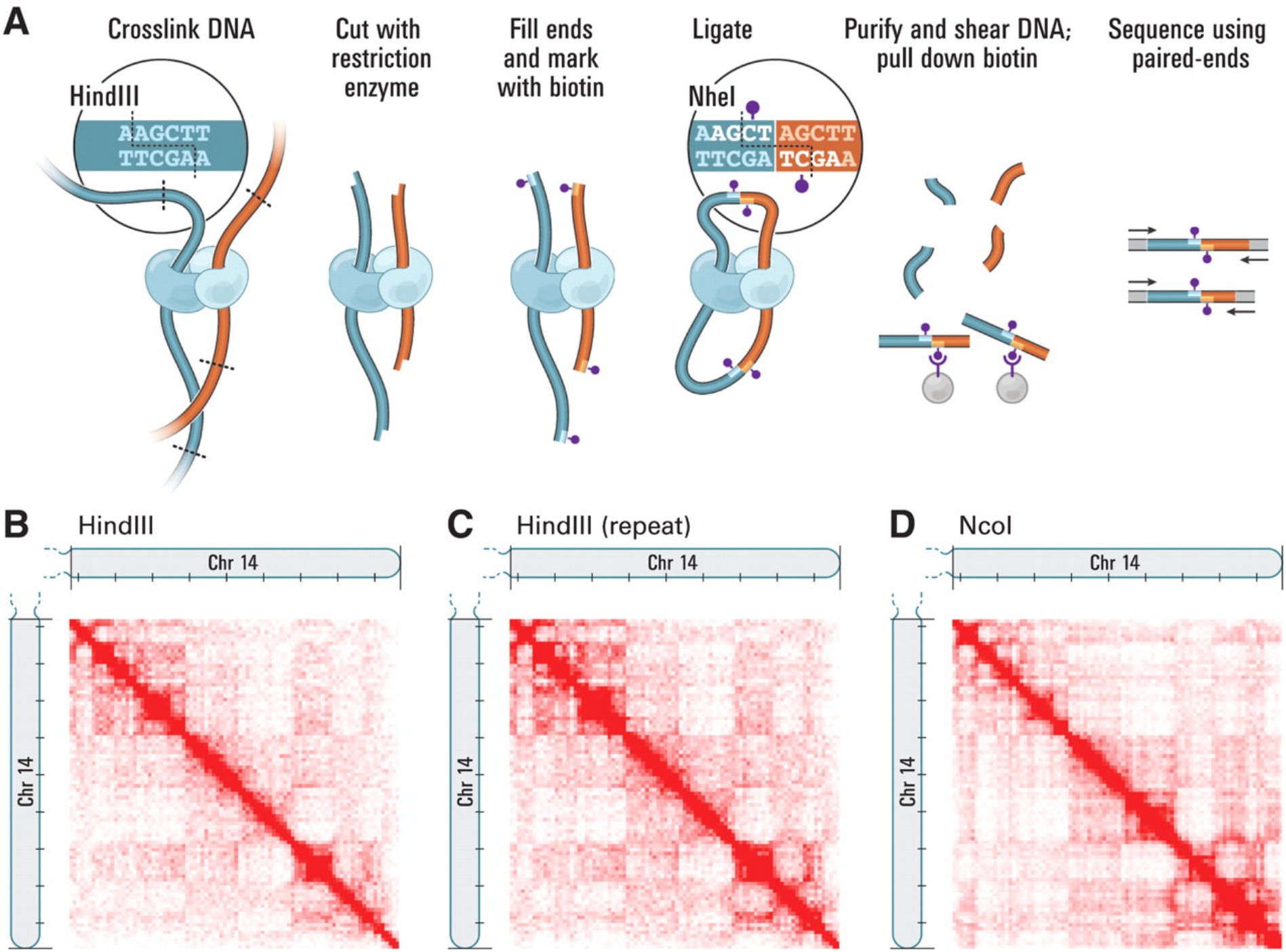
Medvedev et al., Nature 2009


a
Basic insertion
Donor
Ref
b
Basic deletion
c
Basic inversion
d
Linking
Donor
A
B
Ref
A
B
e
Linked insertion
A
B
C
A
C
B
f
Everted duplication
g
Anchored split mapping
(deletion)
Donor
Ref
h
Anchored split mapping
(insertion)
0
i
Hanging insertion

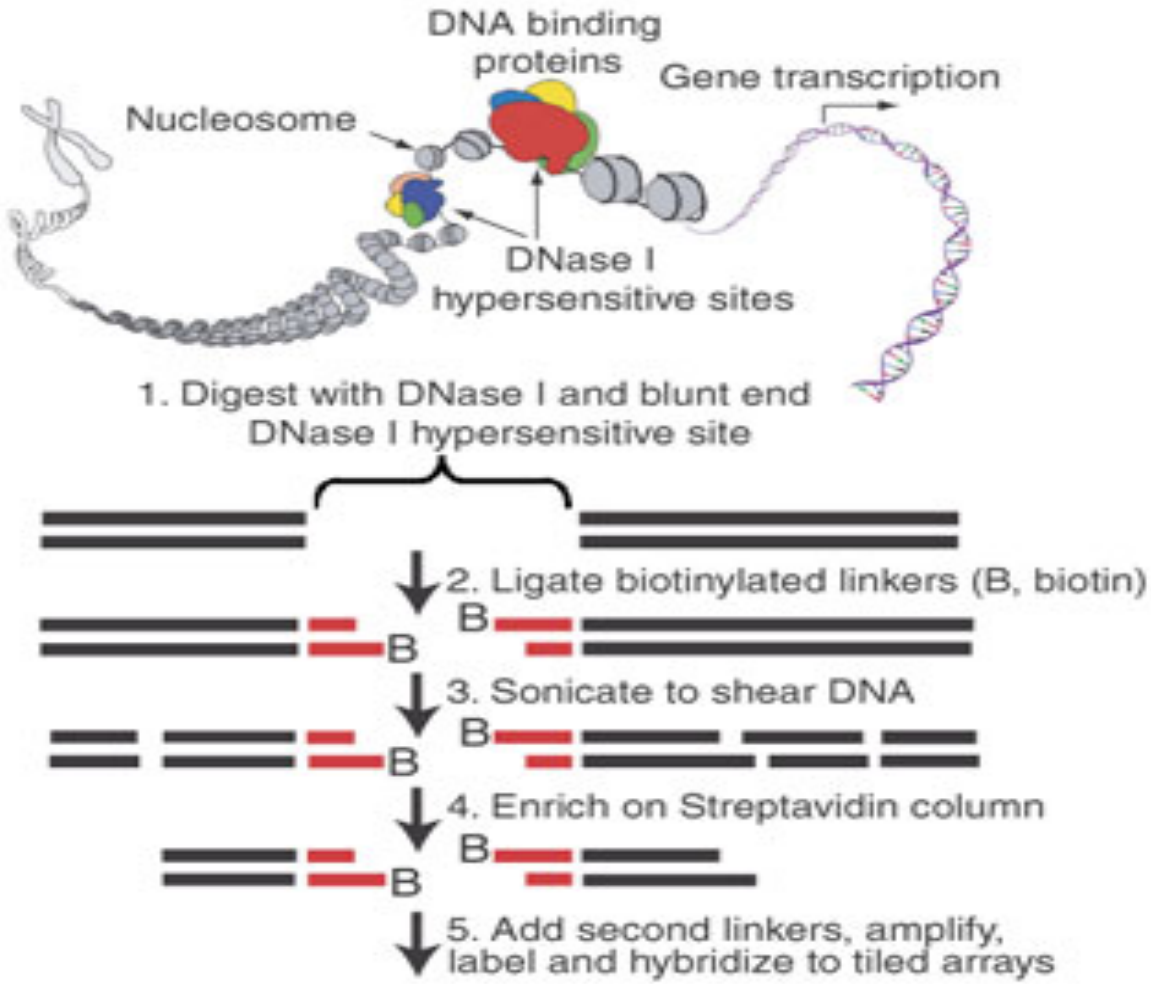
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. We constructed spatial proximity maps of the human genome with Hi-C at a resolution of 1 megabase. These maps confirm the presence of chromosome territories and the spatial proximity of small, gene-rich chromosomes. We identified an additional level of genome organization that is characterized by the spatial segregation of open and closed chromatin to form two genome-wide compartments. At the megabase scale, the chromatin conformation is consistent with a fractal globule, a knot-free, polymer conformation that enables maximally dense packing while preserving the ability to easily fold and unfold any genomic locus. The fractal globule is distinct from the more commonly used globular equilibrium model. Our results demonstrate the power of Hi-C to map the dynamic conformations of whole genomes.

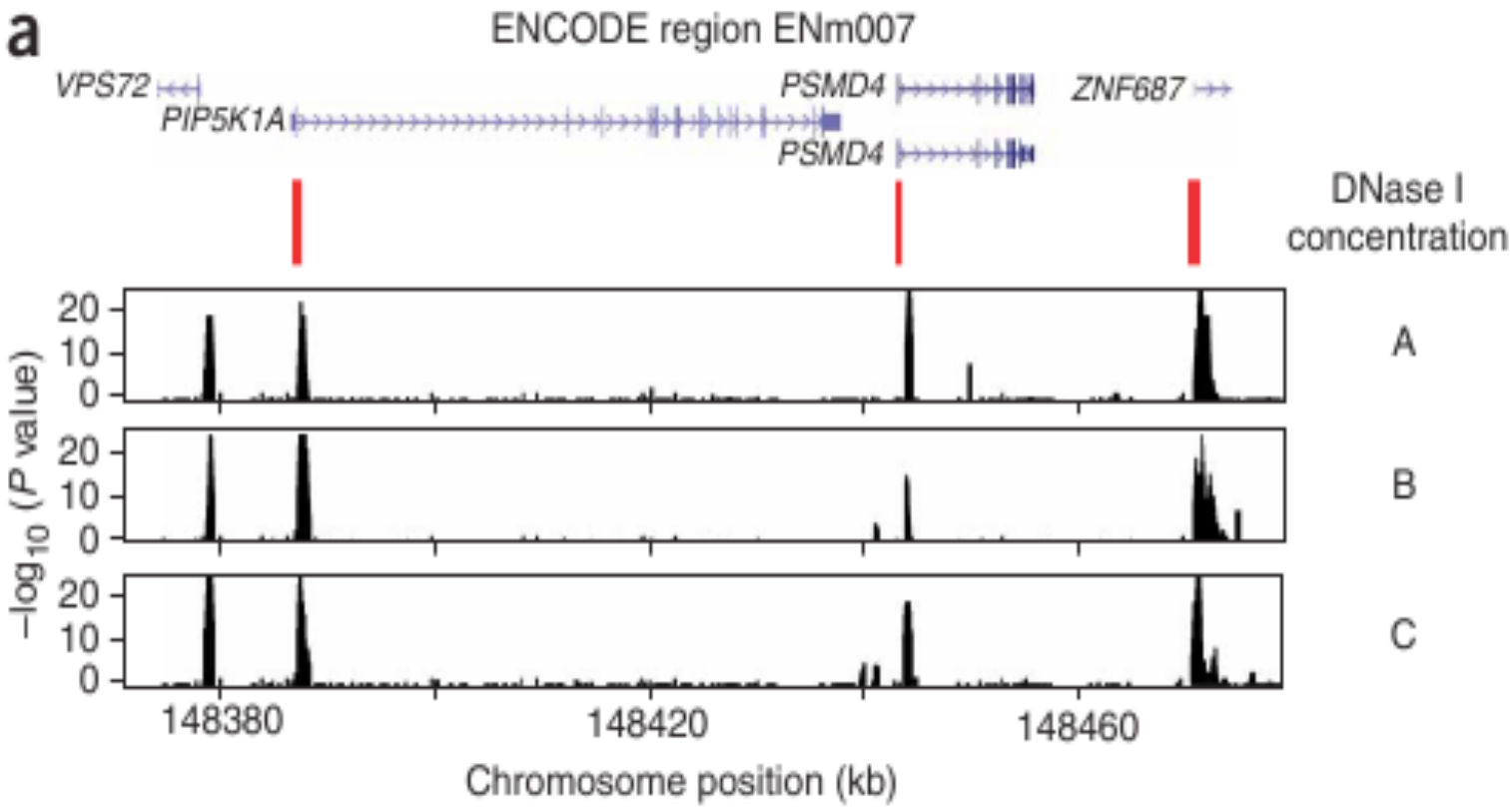

A
Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends
B
HindIII
Chr 14
Chr 14
C
HindIII (repeat)
Chr 14
Chr 14
D
NcoI
Chr 14
Chr 14

DNAse-chip Method

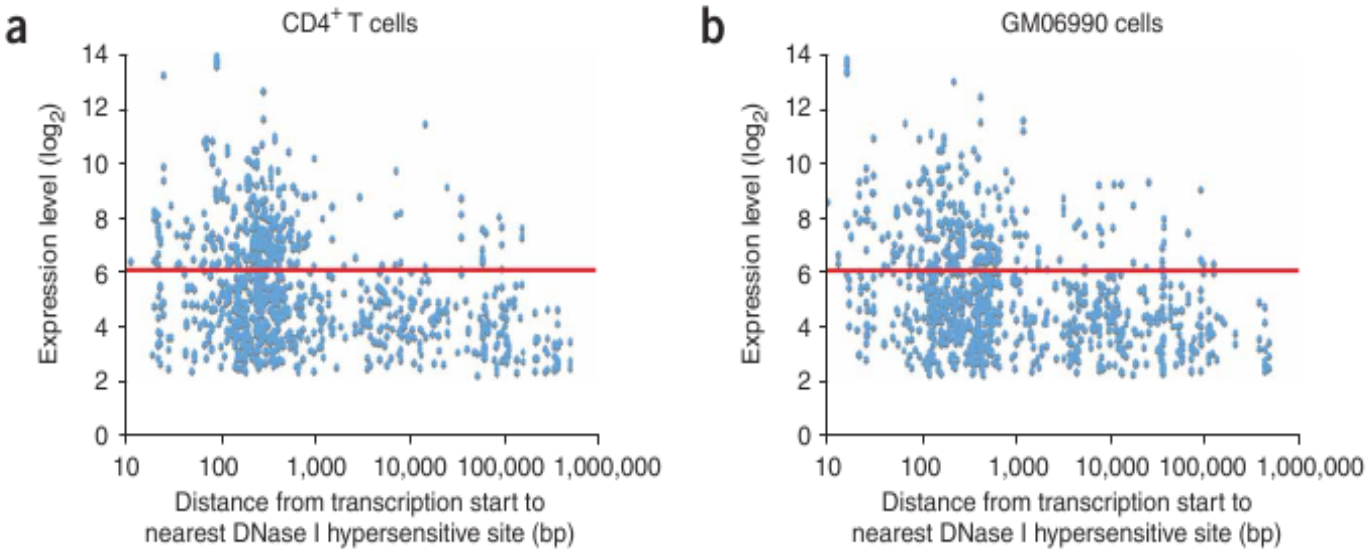


Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G., and Collins, F.S. *Nat Methods*, 2006

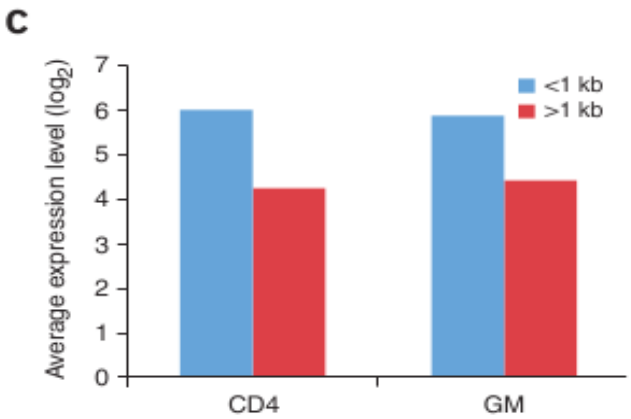
DNAse Sites Relative to Genes



DNAse HS Sites and Gene Expression



- DNAse HS sites near transcription start sites are associated with actively transcribed genes.



You must explain your problem clearly

sedavis@helix:/data/ngs/rna-seq/fastq

Default | sedavis@helix:/data/ngs/r...

```
@SRR065491.4 HWUSI-EAS627_1:1:1:0:16 length=75
NNNNNNNNNNNNNNNNCGGGNTAGGCCATCCCACAGGCACCCCCGCCGCNCNNNNNCACGCCCCCACTGCCCACA
+SRR065491.4 HWUSI-EAS627_1:1:1:0:16 length=75
!!!!!!!!!!!!!!!!!####!##############################!#!!!!##################
@SRR065491.5 HWUSI-EAS627_1:1:1:0:57 length=75
NNNNNNNNNNNNNNNNAAACNTGCATGCAATGTGAGCCGTGTGGCAATCNANNNNNGGGGCATAGCCGGCGCTTA
+SRR065491.5 HWUSI-EAS627_1:1:1:0:57 length=75
!!!!!!!!!!!!!!!!!####!##############################!#!!!!##################
@SRR065491.6 HWUSI-EAS627_1:1:1:0:173 length=75
NNNNNNNNNNNNNNNNTTCTNTGCCAAAATTAAACAGTACAACACAAACNTNNNNNCTATAATCTTCATCTATGA
+SRR065491.6 HWUSI-EAS627_1:1:1:0:173 length=75
!!!!!!!!!!!!!!!!!####!##############################!#!!!!##################
@SRR065491.7 HWUSI-EAS627_1:1:1:0:176 length=75
NNNNNNNNNNNNNNNNCTAGNGTATCTGGCTGGGTCCCAAATTTCTTCCNCNNNNNGGCCCGGTGGGCCAGCCCC
+SRR065491.7 HWUSI-EAS627_1:1:1:0:176 length=75
!!!!!!!!!!!!!!!!!####!##############################!#!!!!##################
@SRR065491.8 HWUSI-EAS627_1:1:1:0:268 length=75
NNNNNNNNNNNNNNNNCTACNTGCATGCTGTGTGACTTTAGGCCAGTCANTNNNNNCTTGATTCTCTGACCCTCA
+SRR065491.8 HWUSI-EAS627_1:1:1:0:268 length=75
!!!!!!!!!!!!!!!!!####!##############################!#!!!!##################
@SRR065491.9 HWUSI-EAS627_1:1:1:0:373 length=75
NNNNNNNNNNNNNNNNAAAANAGTAGCTACAATAAGGATATTTCAACCTNANNNNNAGAAGTGATAAAACATCAA
--More--(0%)
```

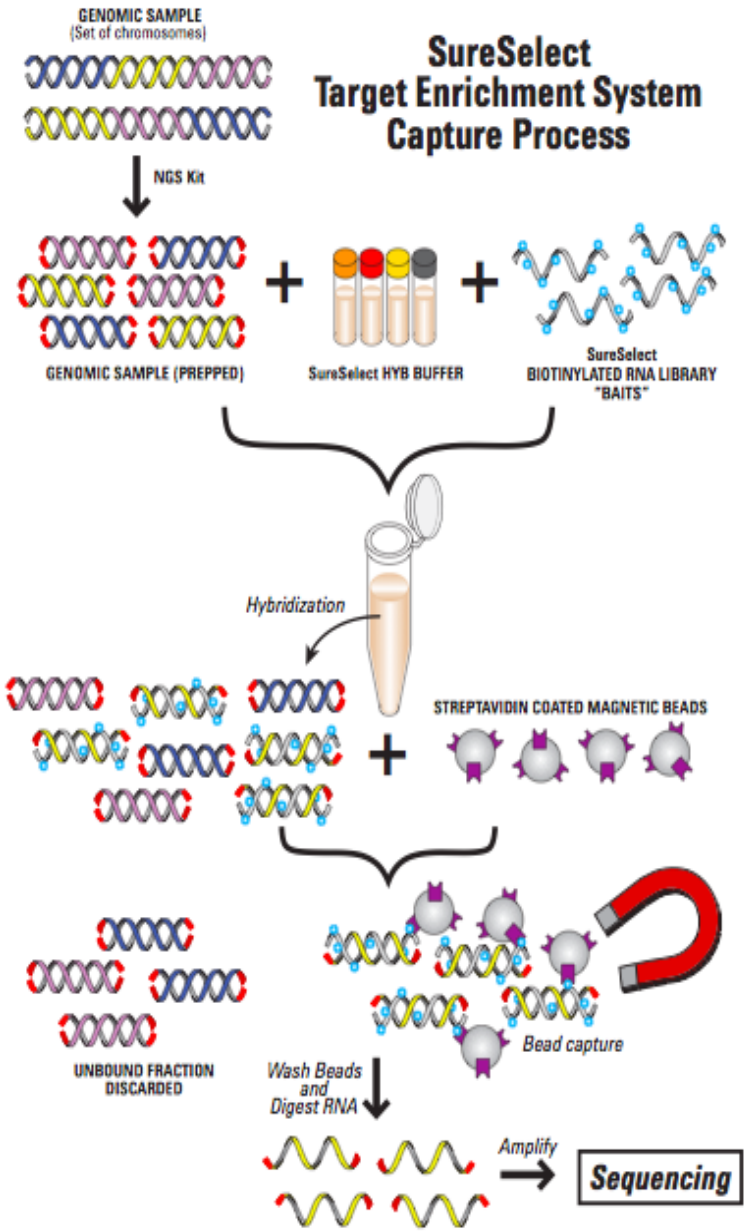
$$Q = -10 \log_{10} p$$

SAM Format

A general, standard format for representing alignments of sequences to reference sequences

Genomic Sequencing

Targeting the Exome



- Long oligos synthesized on arrays (DNA)
- RNA baits synthesized from DNA oligo template
- RNA baits hybridized to DNA sequencing library
- Targets captured using beads and biotin-labeled baits
- RNA bait degraded, leaving sequencing library enriched for target regions

Data Flow

- FASTQ files generated by Illumina GAIIx pipeline
- Aligned to reference genome (hg18, excluding _random, unmapped, and hap) using Novoalign
  - SAM/BAM used extensively
- Follow Broad Institute GATK pipeline for exome capture
- Use picard java library for quality assessment
- Processed BAM files available via local http for browsing

Data Pipeline....

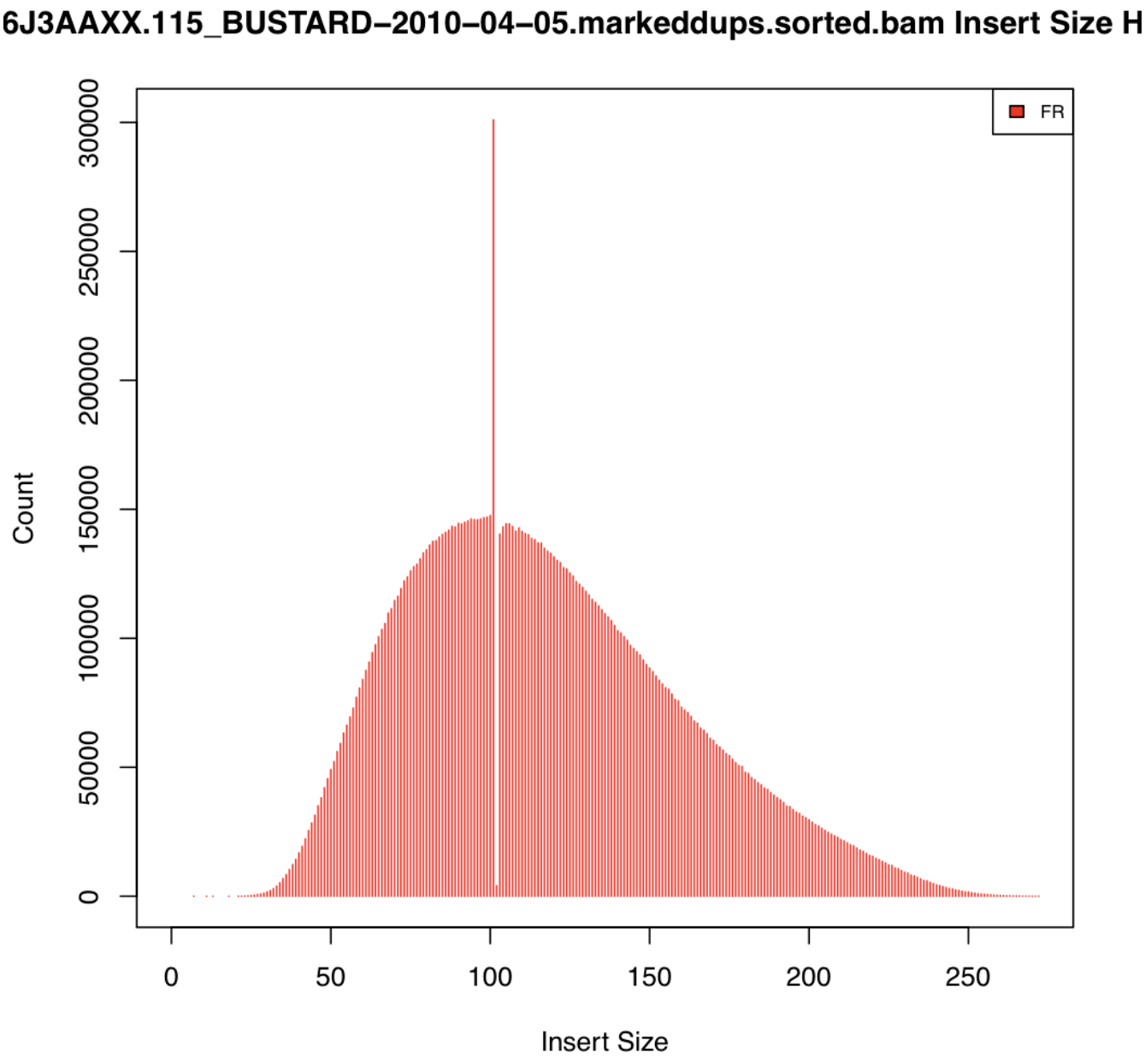
- Samtools import
- Samtools sort
- Picard MarkDuplicates
- GATK Indel Realignment
- GATK Quality Recalibration
- Picard QC metrics

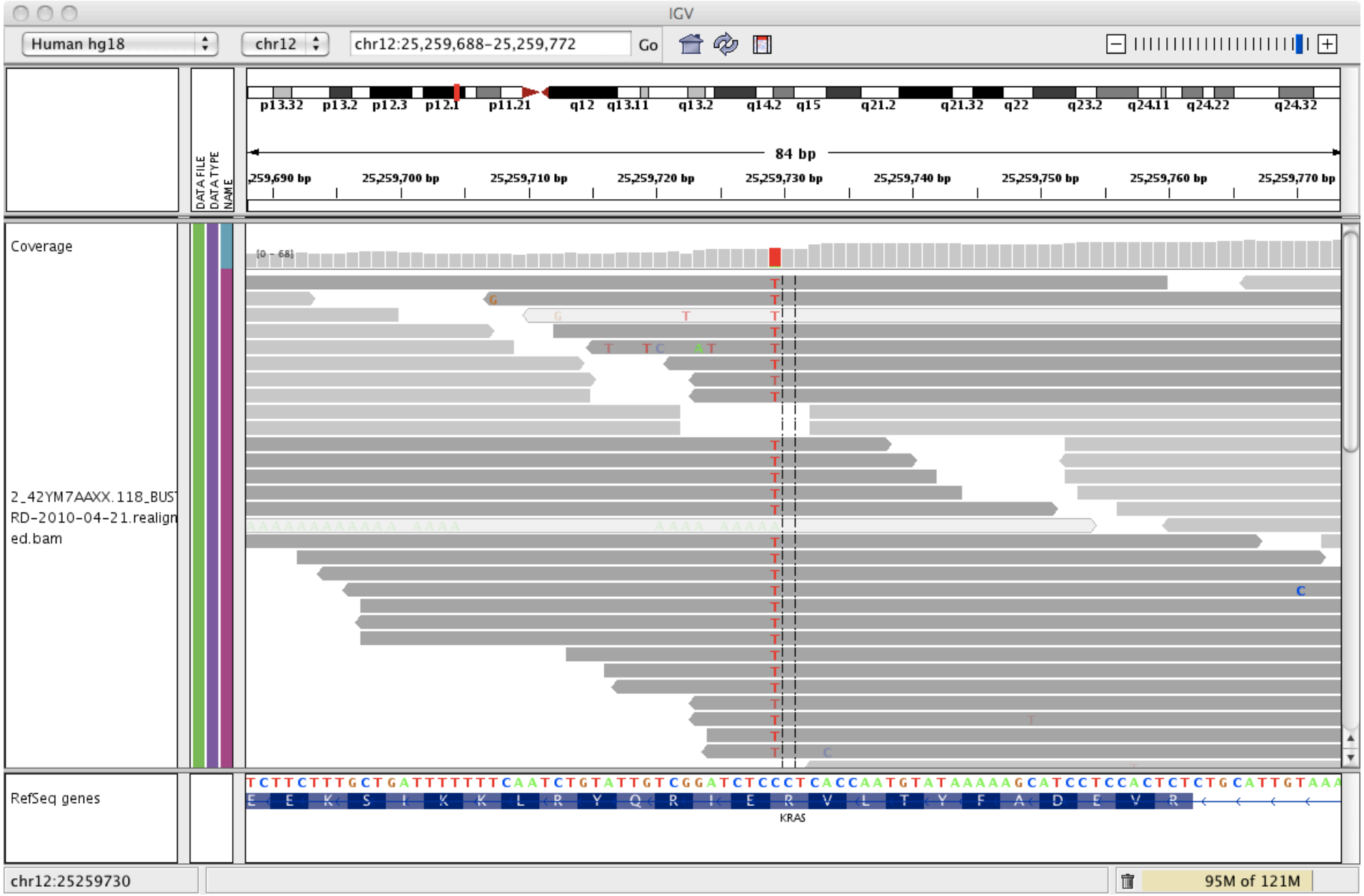
Realignment around Indels

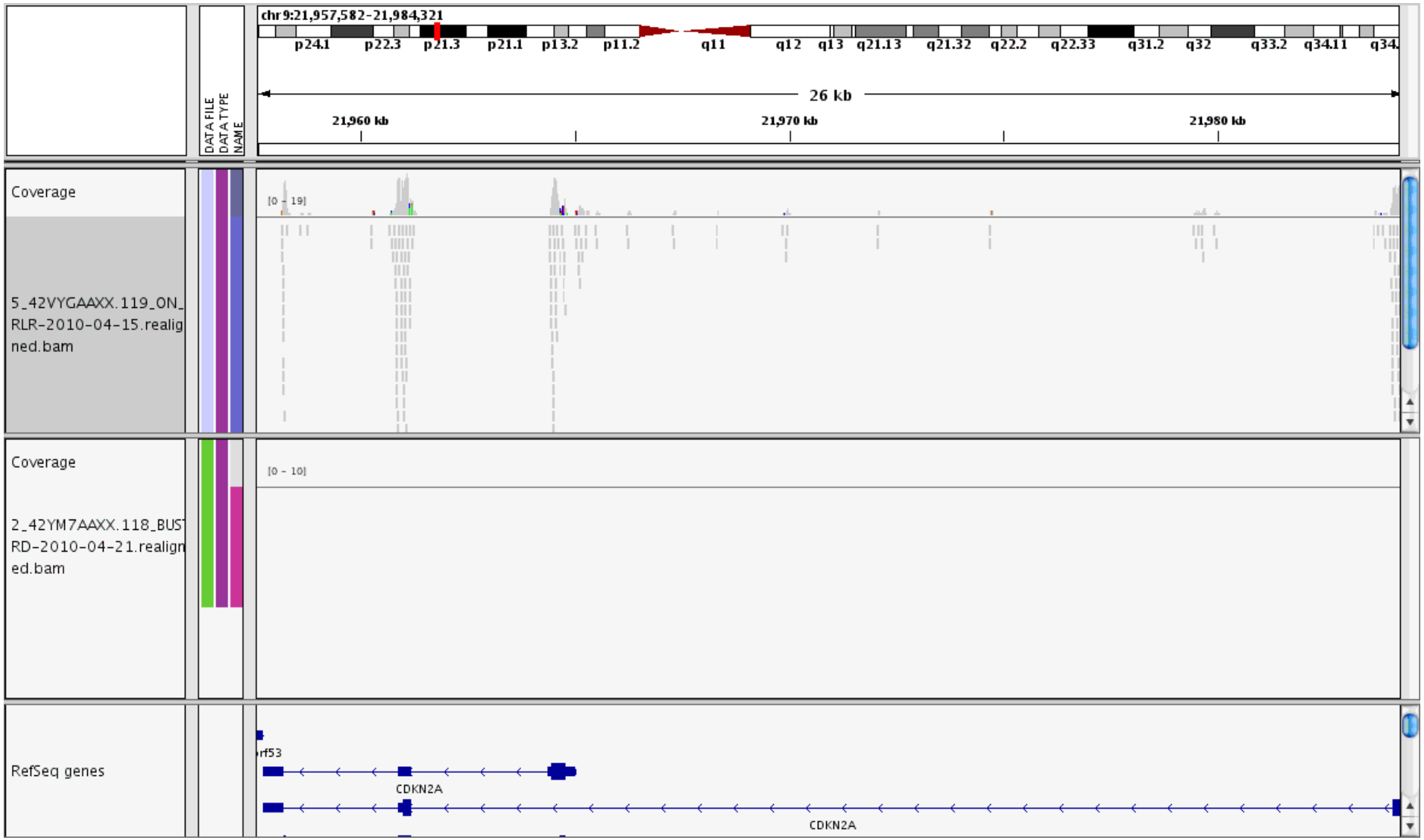
- The problem
  - Aligners align each read independently
  - Potentially leads to increased error rates around indels
- A potential solution
  - Locally realign reads in regions that might harbor an indel
  - Goal is to align reads overlying indels more accurately, reducing errors in each read and, in turn, reducing SNV call error rates

Quality Recalibration

- Since most SNV callers will rely on quality scores to estimate error probabilities, having the best possible estimates for error rates is important
- Reported error rates from the Illumina sequencer generally reflect technical parameters of the base call process, but not other systematic biases
- Quality recalibration can include covariates to account for systematic biases
    - Cycle count, dinucleotide context, original quality, and sample/library variables

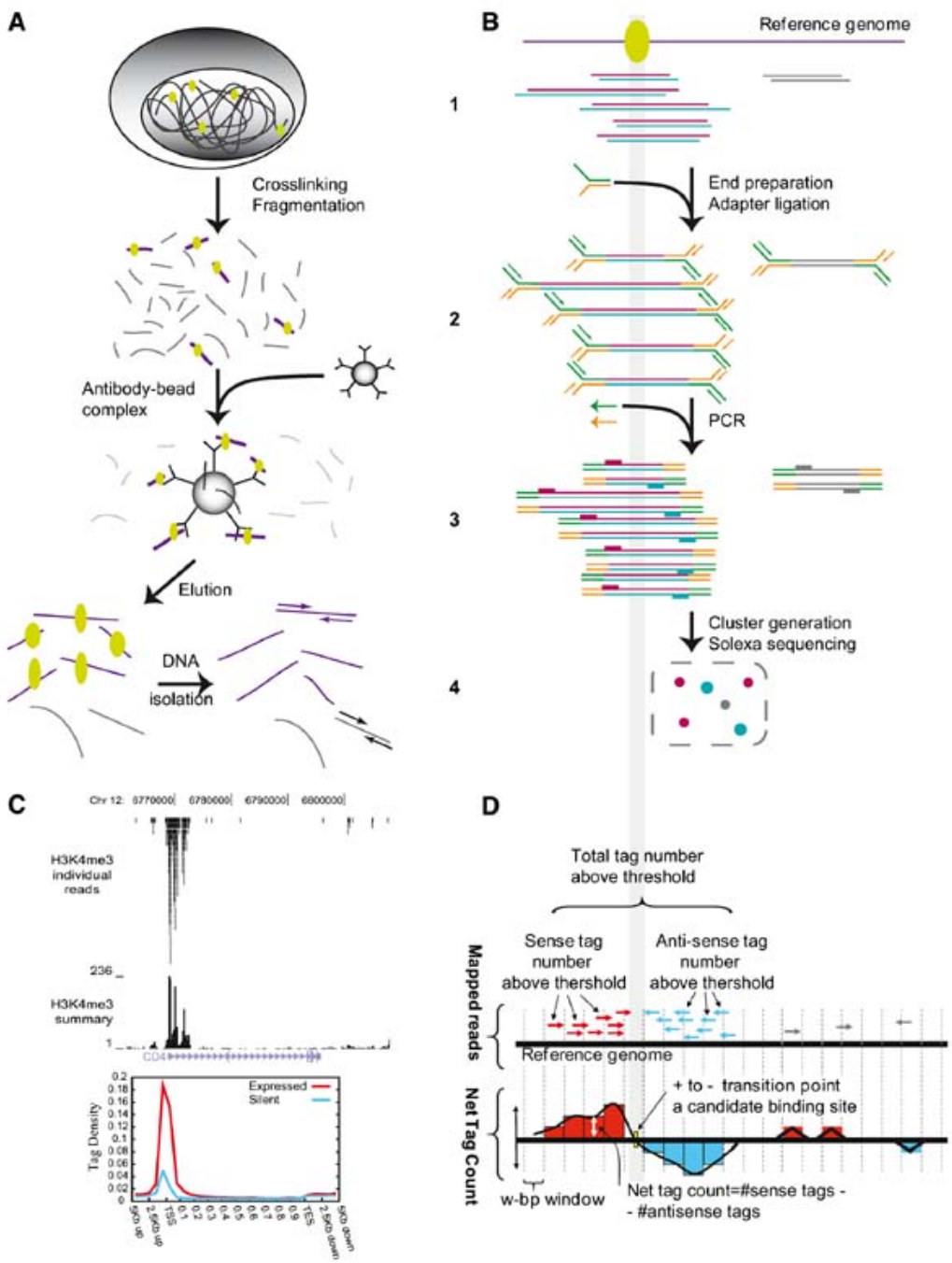

6J3AAXX.115_BUSTARD−2010−04−05.markeddups.sorted.bam Insert Size H
FR
Count
0
50000
100000
150000
200000
250000
300000
0
50
100
150
200
250
Insert Size


IGV
Human hg18
chr12
chr12:25,259,688-25,259,772
Go
p13.32 p13.2 p12.3 p12.1 p11.21 q12 q13.11 q13.2 q14.2 q15 q21.2 q21.32 q22 q23.2 q24.11 q24.22 q24.32
84 bp
25,259,690 bp 25,259,700 bp 25,259,710 bp 25,259,720 bp 25,259,730 bp 25,259,740 bp 25,259,750 bp 25,259,760 bp 25,259,770 bp
DATA FILE
DATA TYPE
NAME
Coverage
[0 - 68]
2_42YM7AAXX.118_BUSTRD-2010-04-21.realigned.bam
RefSeq genes
KRAS
chr12:25259730
95M of 121M


chr 9:21,957,582-21,984,321
p24.1
p22.3
p21.3
p21.1
p13.2
p11.2
q11
q12
q13
q21.13
q21.32
q22.2
q22.33
q31.2
q32
q33.2
q34.11
26 kb
21,960 kb
21,970 kb
21,980 kb
DATA FILE
DATA TYPE
NAME
Coverage
[0 - 19]
5_42VYGAAXX.119_ON_RLR-2010-04-15.realigned.bam
Coverage
[0 - 10]
2_42YM7AAXX.118_BUSTRD-2010-04-21.realigned.bam
RefSeq genes
CDKN2A
CDKN2A

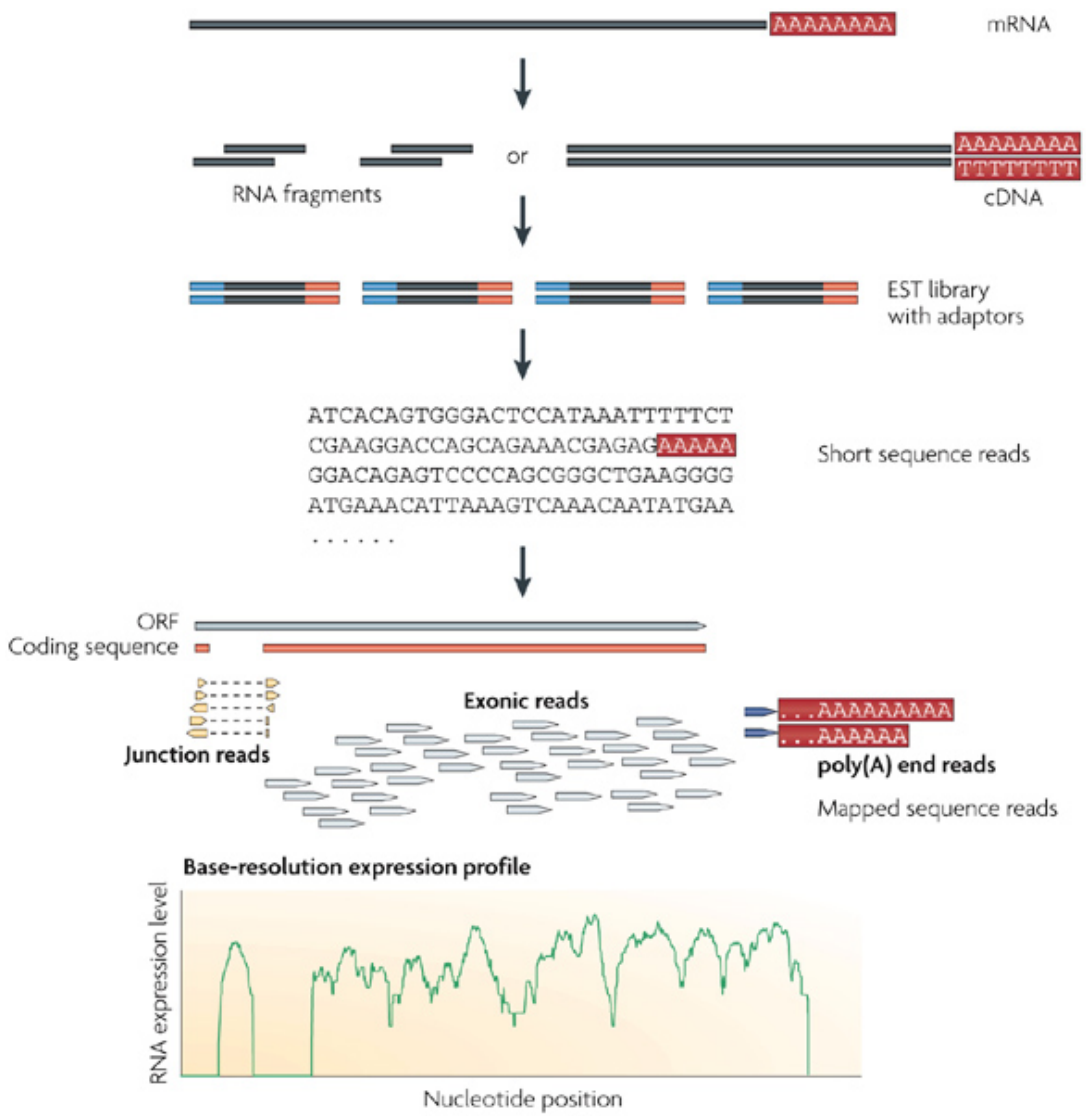
Variant Calling and Evaluation

ChIP-Seq



Barski A, Zhao K. Genomic location analysis by ChIP-Seq. J Cell Biochem. 2009

RNA-Seq



Nature Reviews | Genetics

RNA-Seq Blog

rna-seqblog.com

RNA-SEQ BLOG

TRANSCRIPTOME ANALYSIS: SEQUENCING AND PROFILING

HOME CONTACT

Search for:

Search

Categories

Select Category

RNA-Seq Data Analysis Tools

ArrayExpressHTS - is an R based pipeline for pre-processing, expression estimation and data quality assessment of high throughput sequencing transcriptional profiling (RNA-seq) datasets.

iAssembler – a standalone package to assemble ESTs generated using Sanger and/or Roche-454 pyrosequencing technologies into contigs.

Trinity RNA-Seq Assembly – software solutions targeted to the reconstruction of full-length transcripts and alternatively spliced isoforms from Illumina RNA-Seq data.

MAR 23

RNA-Seq revealing transcriptomes of commercially important marine species

Filed Under Publications | Leave a Comment

Transcriptome characterization of the South African abalone Haliotis midae using sequencing-by-synthesis

Worldwide, the genus Haliotis is represented by 56 extant species and several of these are commercially cultured. Among the six abalone species found in South Africa, Haliotis midae is the only aquacultured species. Despite its economic importance, genomic sequence resources for H. midae, and for abalone in general, are still scarce. Next generation sequencing technologies provide a fast and efficient tool to generate large sequence collections that can be used to characterize the transcriptome and identify expressed genes associated with economically important traits like growth and disease resistance.[1]

Novel and Conserved Micrornas in Dalian Purple Urchin (Strongylocentrotus Nudus) Identified by Next Generation Sequencing

The purple urchin, Strongylocentrotus nudus, is one of the most important marine economic animals that widely distributed in the cold seas along the coasts of eastern pacific area. To date, only 45 microRNAs have been identified in a related species, Strongylocentrotus purpuratus, and there is no

You can do it yourself.

Learn More

Avadis NGS

by Strand Scientific Intelligence

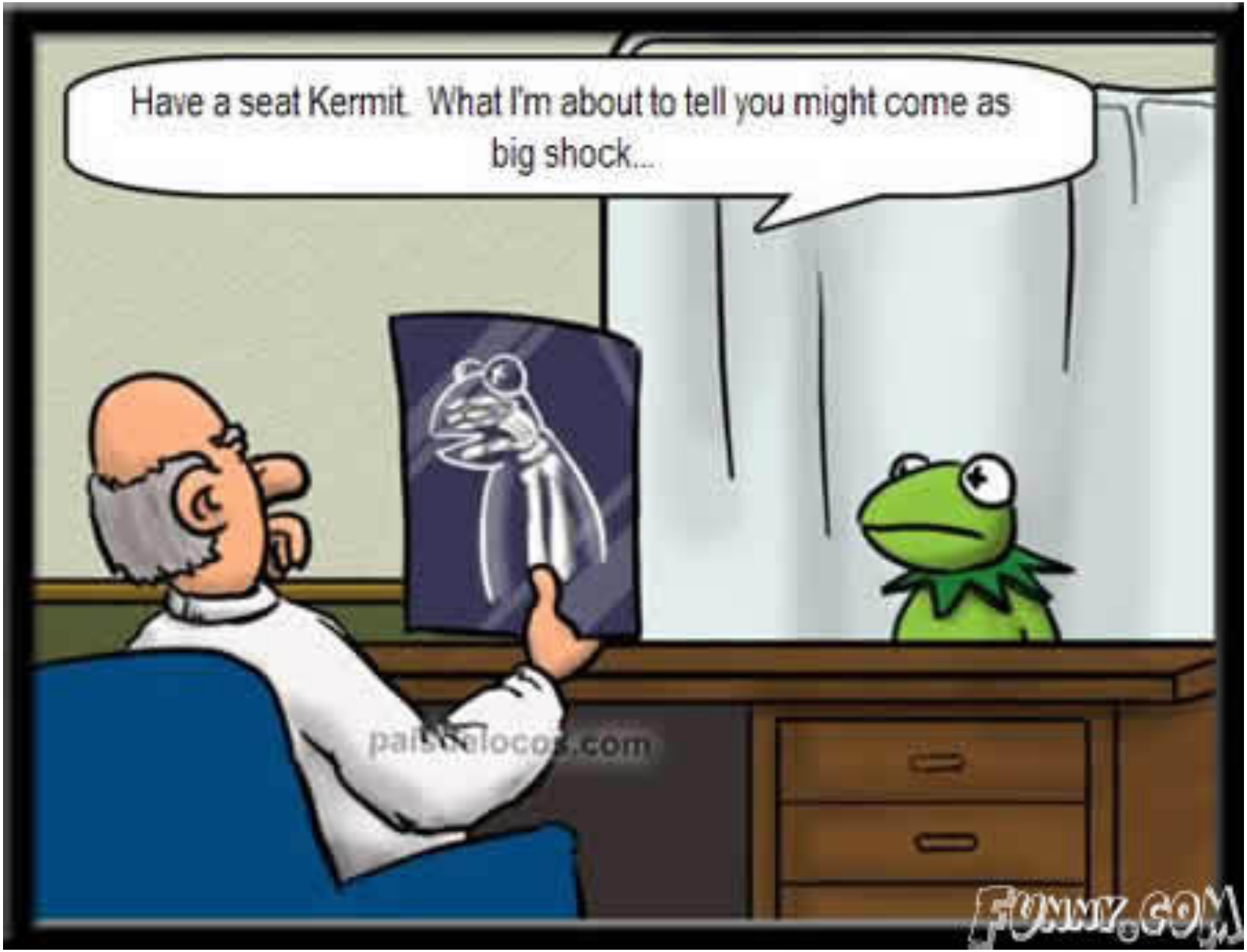
Subscribe via RSS

Subscribe by Email

Polls

Which technology will ultimately rise above all others as the choice method for high-throughput sequencing?

- Pyrosequencing
- DNA Nanoball Sequencing
- Reversible Dye- Terminator Sequencing

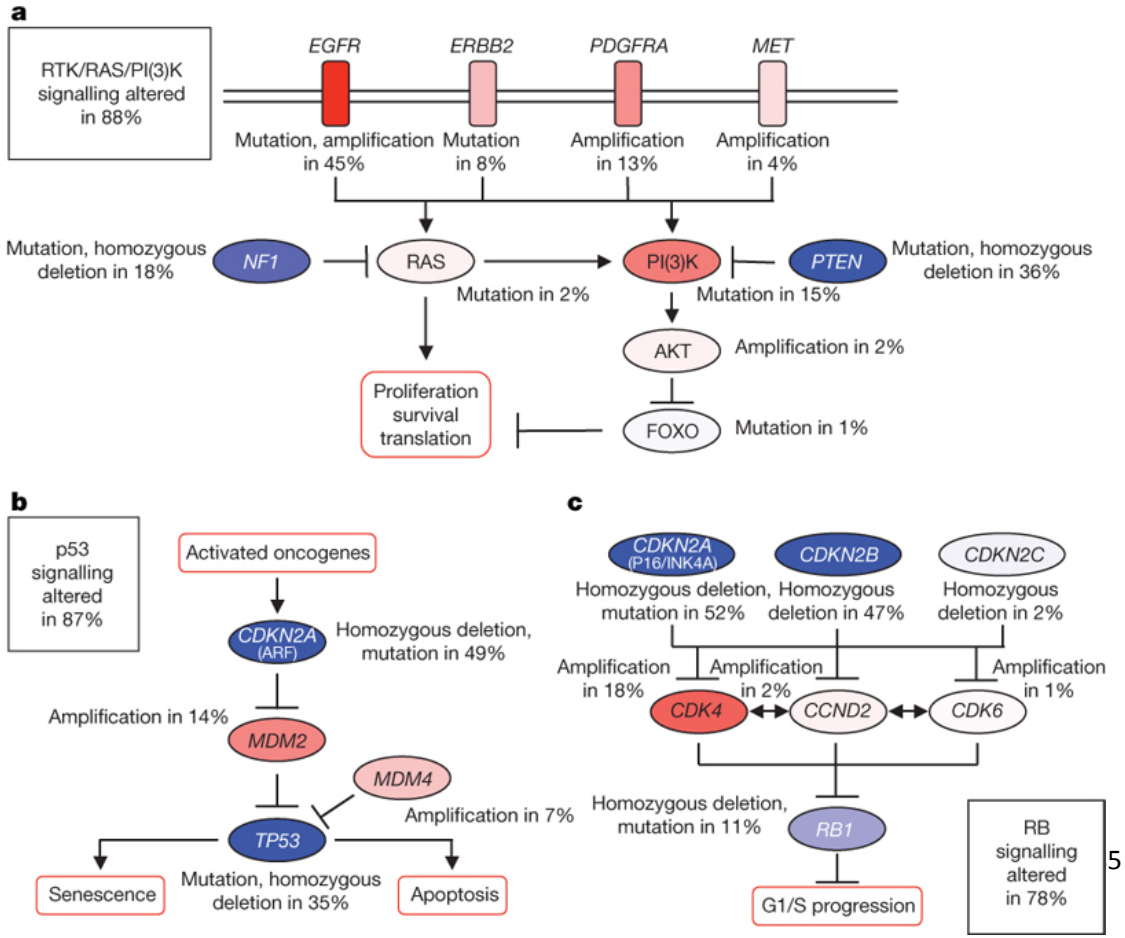

Have a seat Kermit. What I'm about to tell you might come as big shock...
FUNNY.COM

Data Integration and Interpretation

Public Data

- 1000 Genomes, HapMap, Encyclopedia of DNA Elements (ENCODE), The Cancer Genome Atlas (TCGA)
- NCBI Gene Expression Omnibus (GEO)
- Sequence Read Archive (SRA)
- Hundreds of datasets not submitted publicly or housed only in proprietary databases

Frequent genetic alterations in three critical signalling pathways.



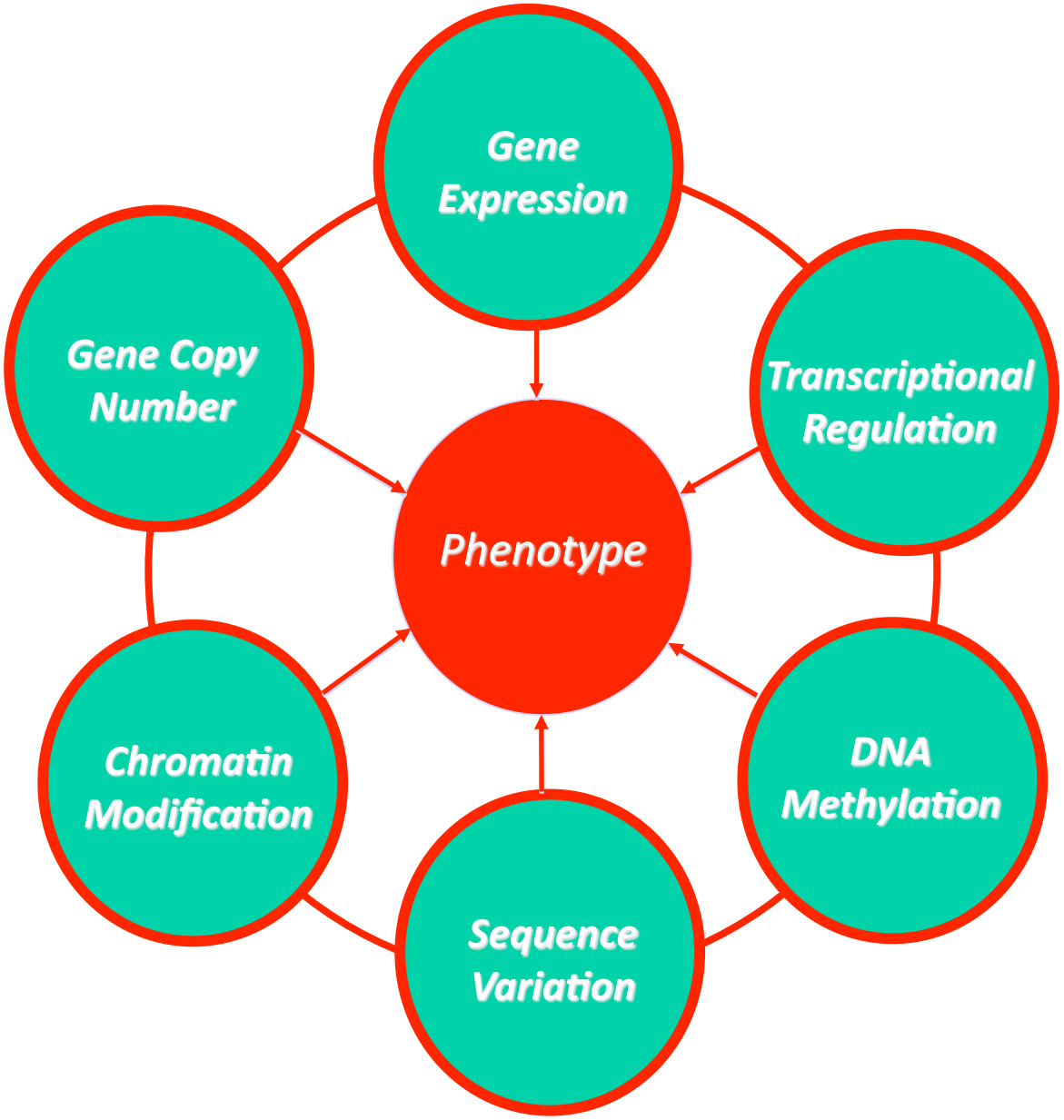
nature

Data Deluge

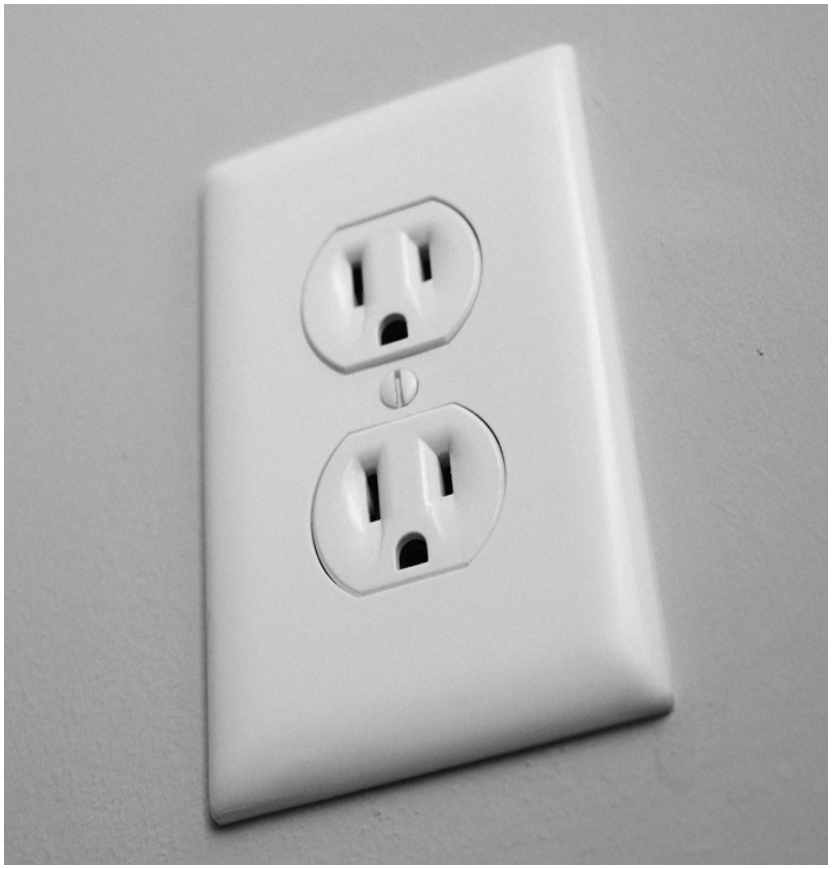
- Lots of data, little information
- Data integration is an enormous challenge that cannot be easily generalized
- Study design and hypothesis-driven research still has a place in a post-genomic world
- Functional validation of genomic data is non-trivial and remains low-throughput, generally
- Data must be correlated with existing biological knowledge!

Reproducible Research

- In an era of biologic discovery driven by large datasets and computational analyses:
  - Tracking research is problematic with different versions of databases, data updates, software updates, etc.
  - Describing methods succinctly but fully may best be done in computer code and not manuscript text
  - Making data available for validation may be problematic due to size and legal/ethical constraints
  - What constitutes the “raw data”?


Gene Expression
Gene Copy Number
Transcriptional Regulation
Phenotype
Chromatin Modification
DNA Methylation
Sequence Variation

The Last Mile



Tutorial Location:

http://watson.nci.nih.gov/~sdavis/

```
qsub –I –l nodes=1:o2200
echo $SHELL
# IF NOT bash, type:
bash
source /data/ngs/bashrc
```

```
#Mounting helixdrive
#Windows:
\\helixdrive.nih.gov\ngs

#Mac Finder → Connect to server
smb://helixdrive.nih.gov/ngs
```