生物信息学：导论与方法
Bioinformatics: Introduction and Methods

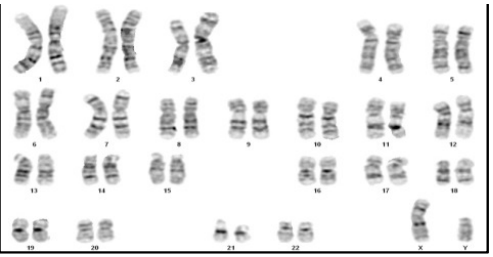
生物信息学：导论与方法

Bioinformatics: Introduction and Methods

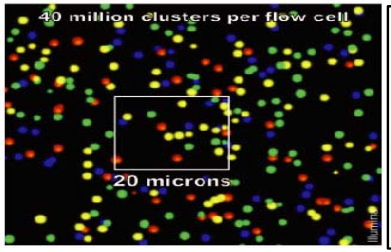
北京大学生物信息学中心 高歌、魏丽萍

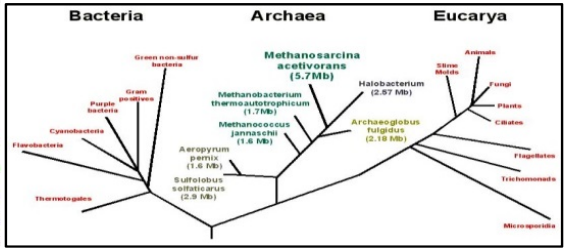
Ge Gao & Liping Wei

Center for Bioinformatics, Peking University



TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



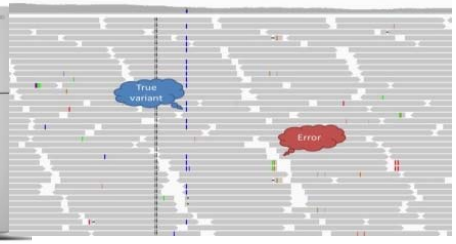


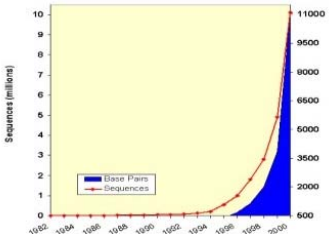
Next Generation Sequencing (NGS): Reads Mapping

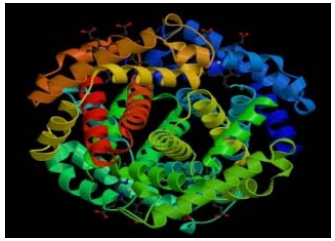
北京大学生物信息学中心 高歌

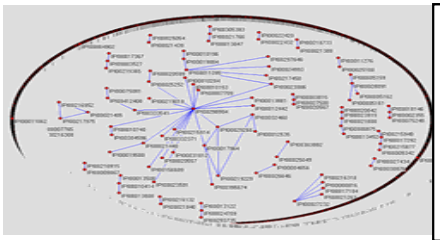
Ge Gao, Ph.D.

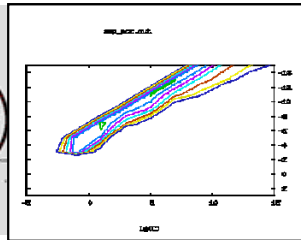
Center for Bioinformatics, Peking University

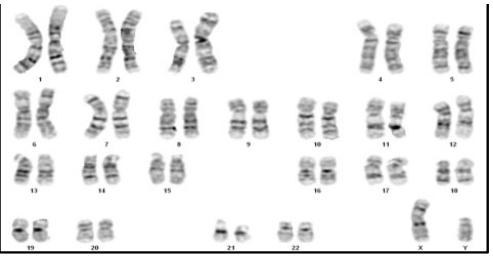




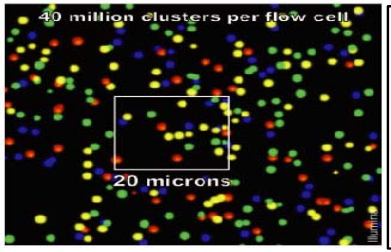


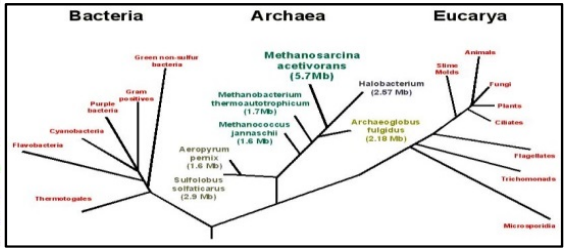






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



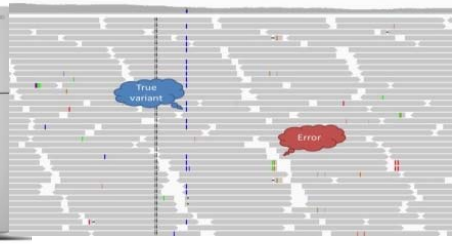


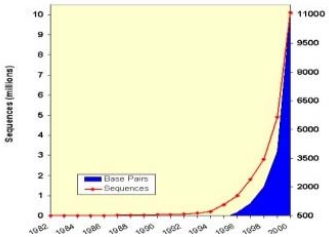
Unit 1:
From Sequencing to NGS

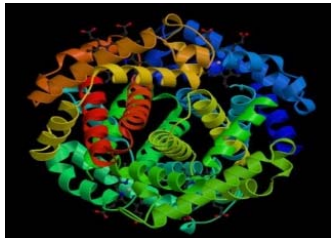
北京大学生物信息学中心 高歌

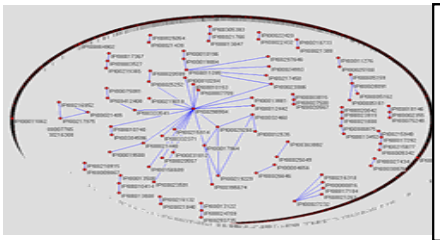
Ge Gao, Ph.D.

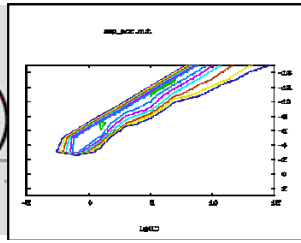
Center for Bioinformatics, Peking University











So simple, yet so mysterious...

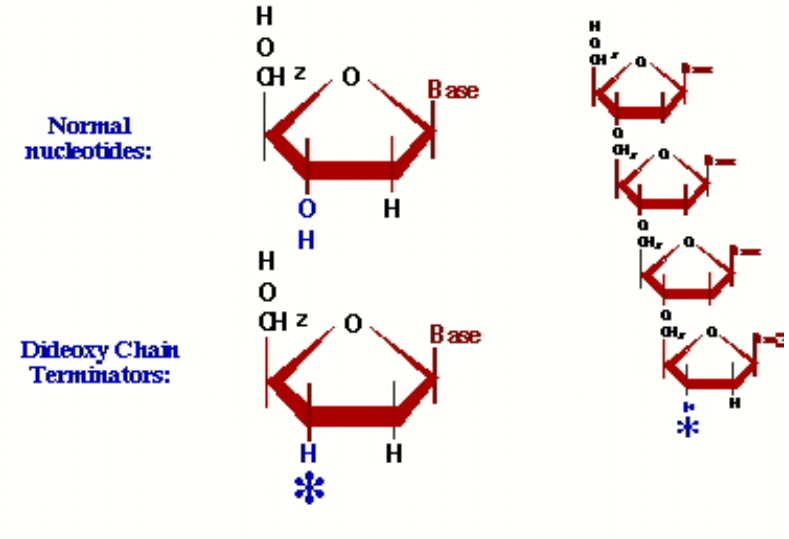
Human genome has **3.1 billion** base pairs.

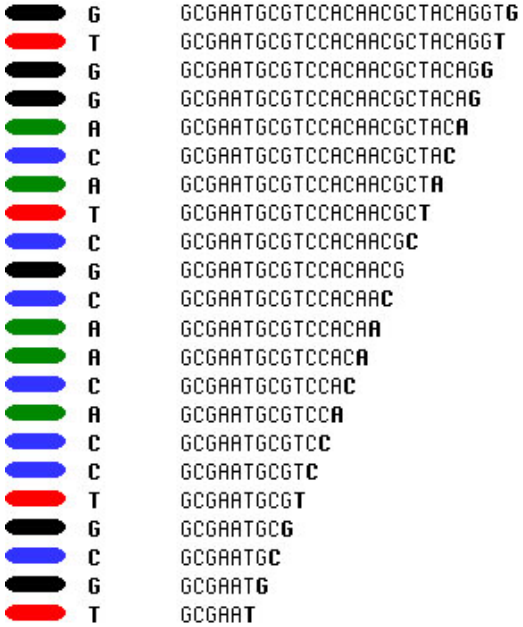
~2.9% of the bases encode **genes.**

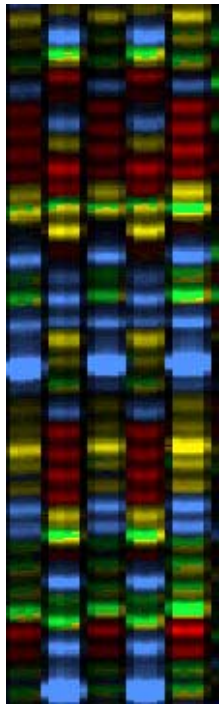
~97% of the genome were previously called “junk”.

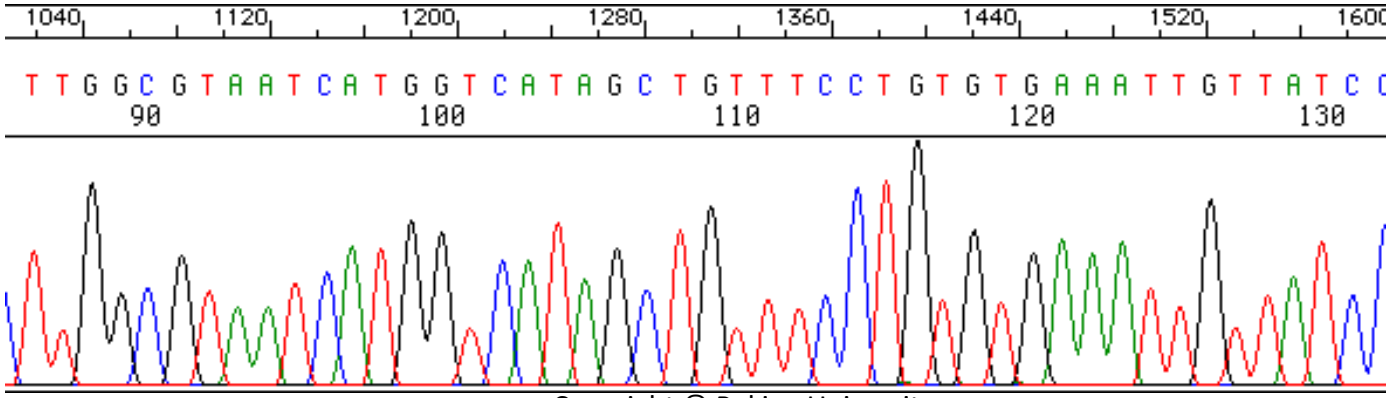
But we now found that they contain the **regulatory elements** that encode instructions on **when, where, which, and how much** proteins to make.

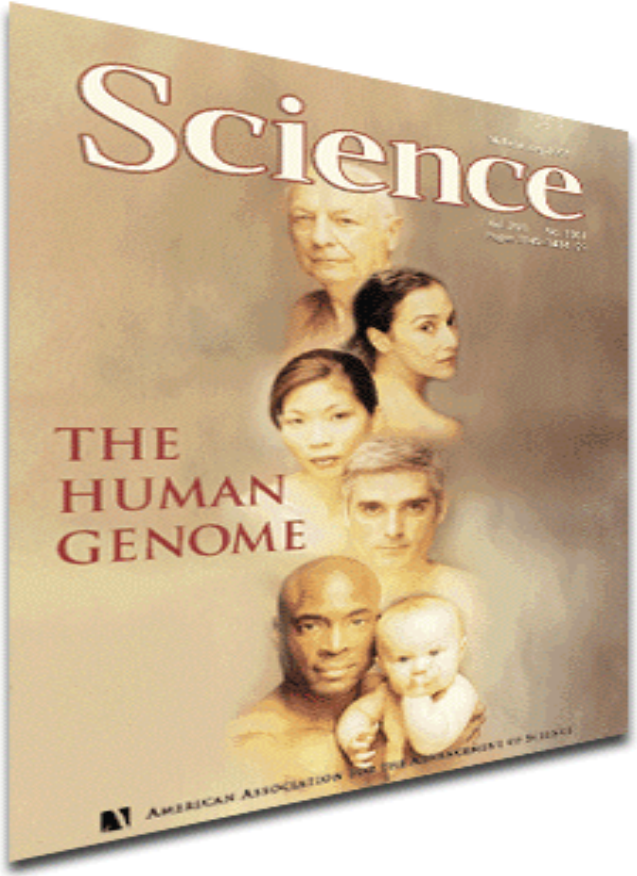
Chain Termination Sequencing





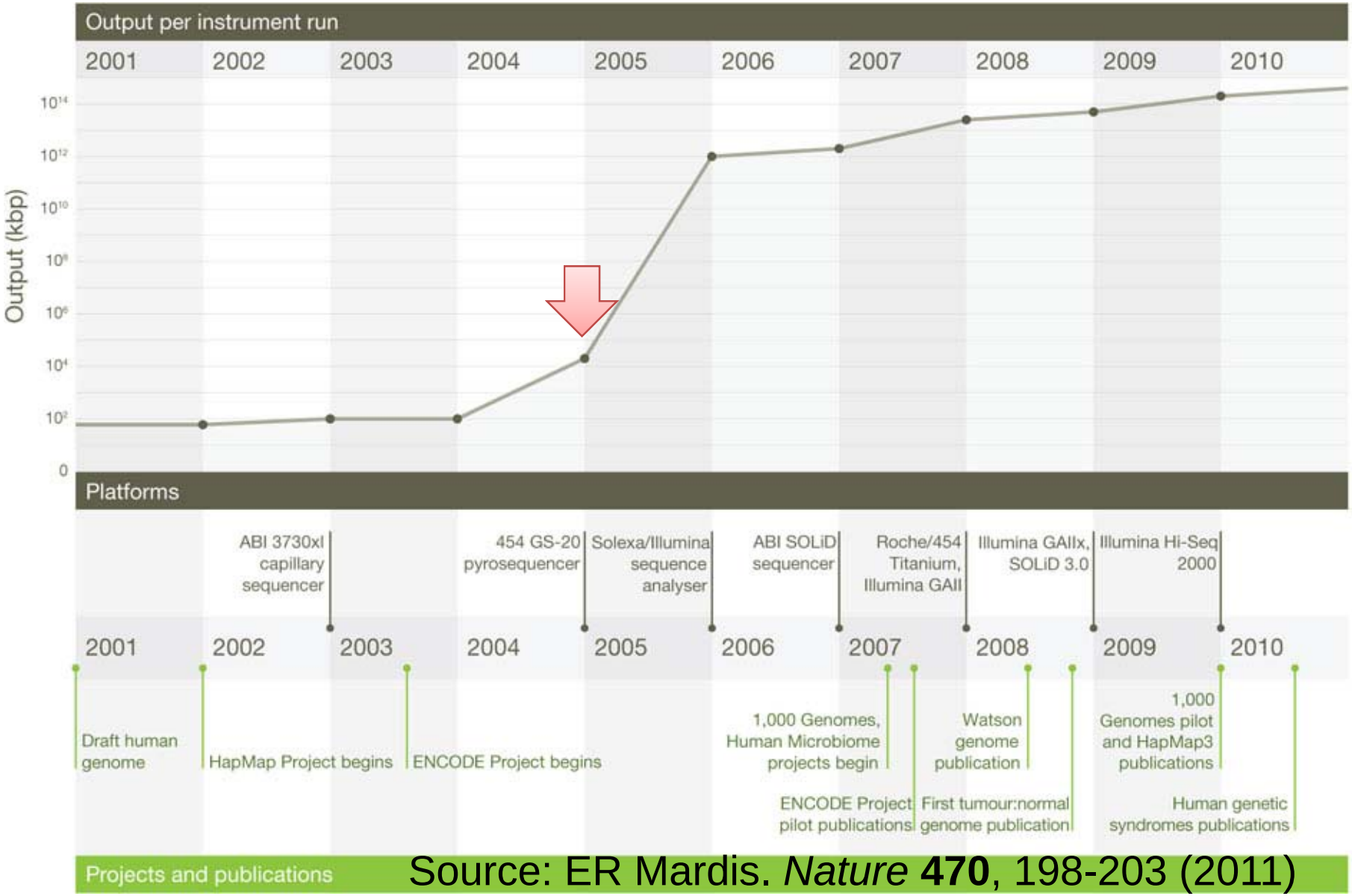






"90% of the three billion base pairs comprising the genome have been read and recorded. The completed work delivers surprises. Perhaps the biggest is that the human genome, estimated at the beginning of the project to contain 80,000 to 100,000 coding genes, appears to possess fewer than 25,000. "

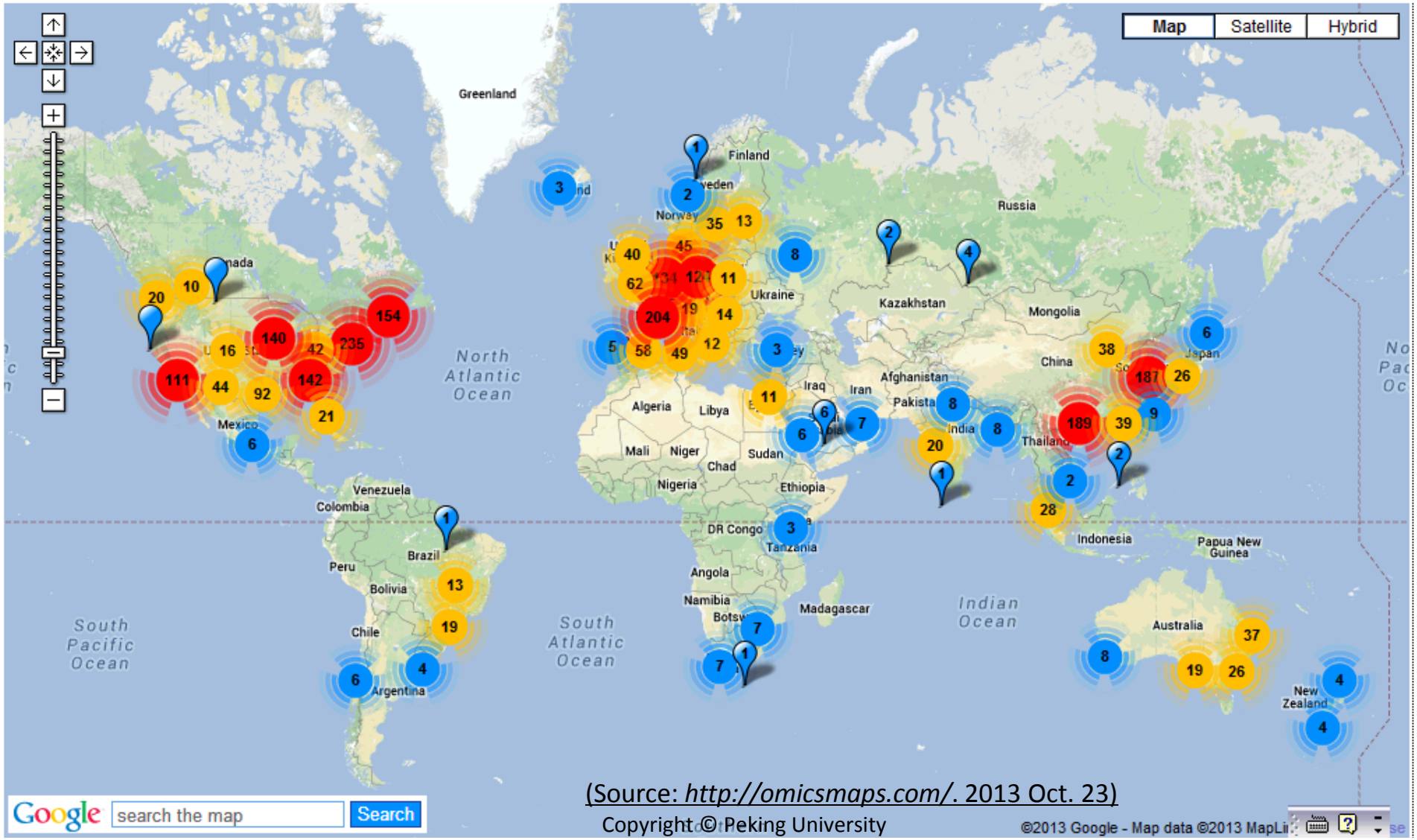
(Source: http://www.lifesciencesfoundation.org/events-The_Book_of_Life.html)



Source: ER Mardis. *Nature* **470**, 198-203 (2011)

Next Generation Genomics: World Map of High-throughput Sequencers

Show all platforms ☐ 454 ☐ HiSeq ☐ Illumina GA2 ☐ Ion Torrent ☐ MiSeq ☐ PacBio ☐ Polonator ☐ Proton ☐ SOLiD ☐ Service Provider

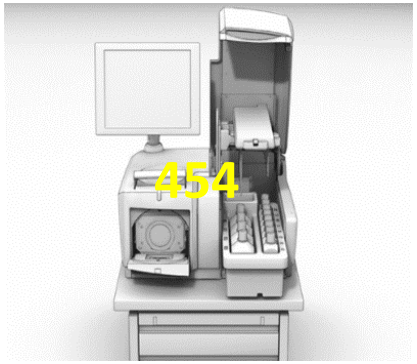


(Source: *http://omicsmaps.com/*. 2013 Oct. 23)

Next Generation Sequencing/Deep Sequencing ***Sanger Sequencing***

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

(Source: *J Biomed Biotechnol*. 2012: 251364.)

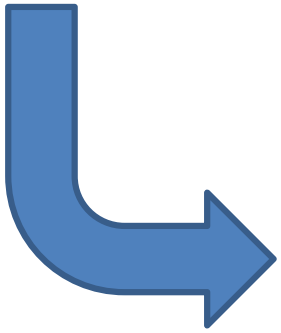


Read: A short DNA fragment which is *read out* by sequencer.

- DNA sequence (symbols)
- Quality information

In FASTQ format

```
@test_fastq
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
+
!''*((((***+))%%%++)(%%%%).1***-+*''
```



```
Seq ID: test_fastq
Sequence: GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
Quality:  !''*((((***+))%%%++)(%%%%).1***-+*''
```

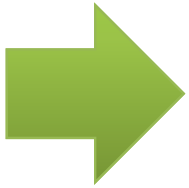
Quality: Given p = the probability of a base calling is *wrong*, its Quality Score can be written as

$$Q = -10 * log_{10}(p)$$

p	Q
0.1	10
0.01	20
0.001	30
0.0001	40

```
0          10         20         30         40
|          |          |          |          |
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|          |          |          |          |
0          10         20         30         40
```

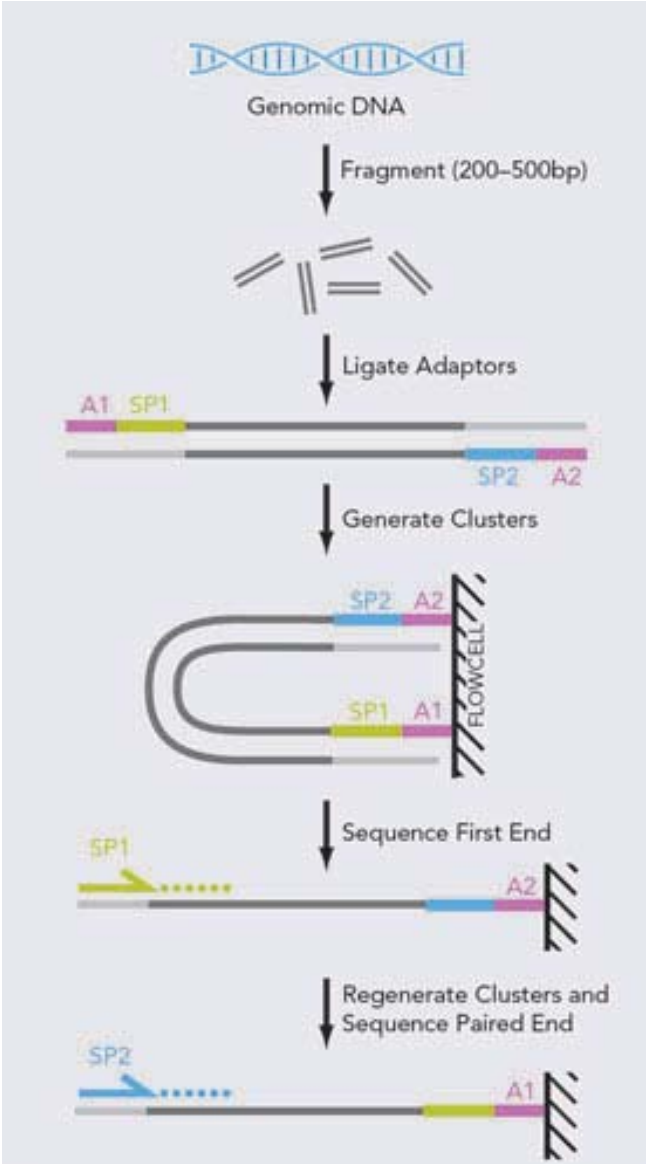
```
@test_fastq
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
+
!''*((((***+))%%%++)(%%%%).1***-+*''
```



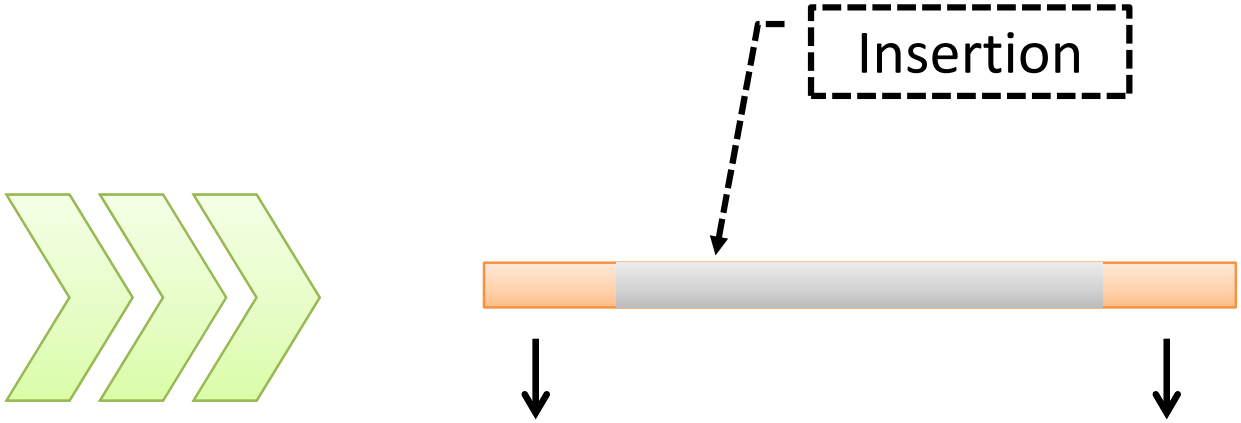
Seq	Quality Symbol	Quality Score	p
G	!	0	1.00
A	'	6	0.25
T	'	6	0.25
T	*	9	0.13
T	(	7	0.20
G	(	7	0.20
G	(	7	0.20
G	(	7	0.20
G	*	9	0.13
T	*	9	0.13
T	*	9	0.13
C	+	10	0.10
A	)	8	0.16
A	)	8	0.16
A	%	4	0.40
G	%	4	0.40
C	%	4	0.40
A	+	10	0.10
G	+	10	0.10
T	)	8	0.16
A	(	7	0.20
T	%	4	0.40
C	%	4	0.40
G	%	4	0.40
A	%	4	0.40
T	)	8	0.16
C	.	13	0.05
A	1	16	0.03
A	*	9	0.13
A	*	9	0.13
T	*	9	0.13
A	-	12	0.06
G	+	10	0.10
T	*	9	0.13
A	'	6	0.25
A	'	6	0.25

```
0         10        20        30        40
|         |         |         |         |
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|         |         |         |         |
0         10        20        30        40
```

Paired-End Reads

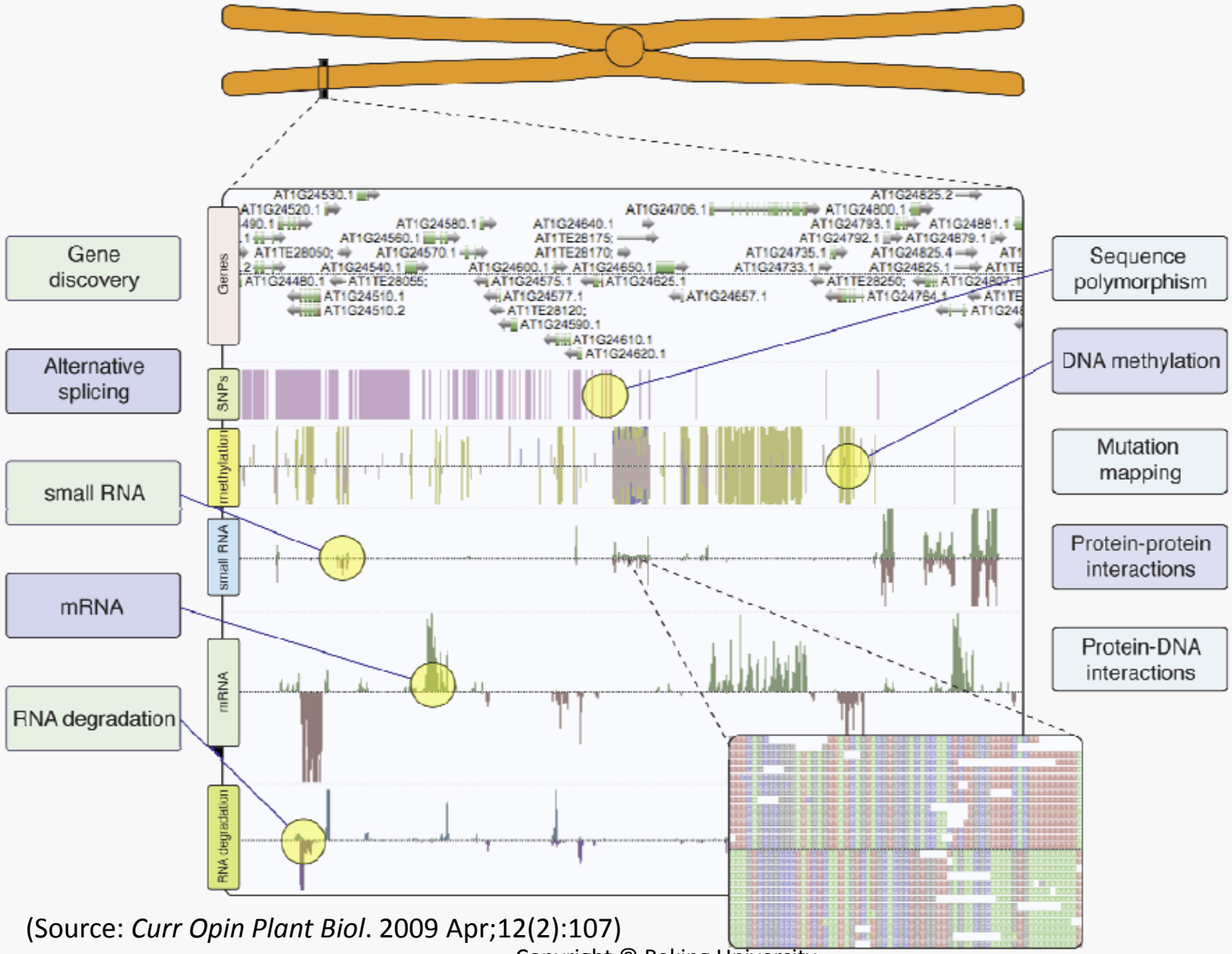


http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn

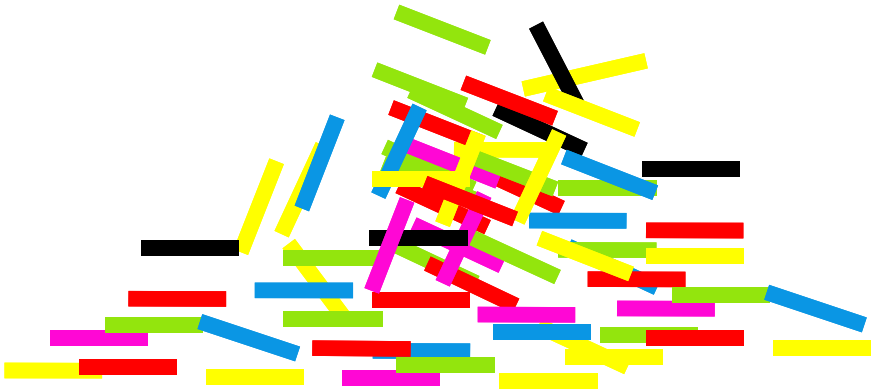



```
@test_fastq/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
+
!''*((((***+))%%%++)(%%%%).1***-+*''
```

```
@test_fastq/2
ACATACTATTACTCATTACTCCTCATANNNNTNCNN
+
BBB1',9,66<B>9<74<=BB@4=93'!!!!)!'!9
```



(Source: *Curr Opin Plant Biol*. 2009 Apr;12(2):107)



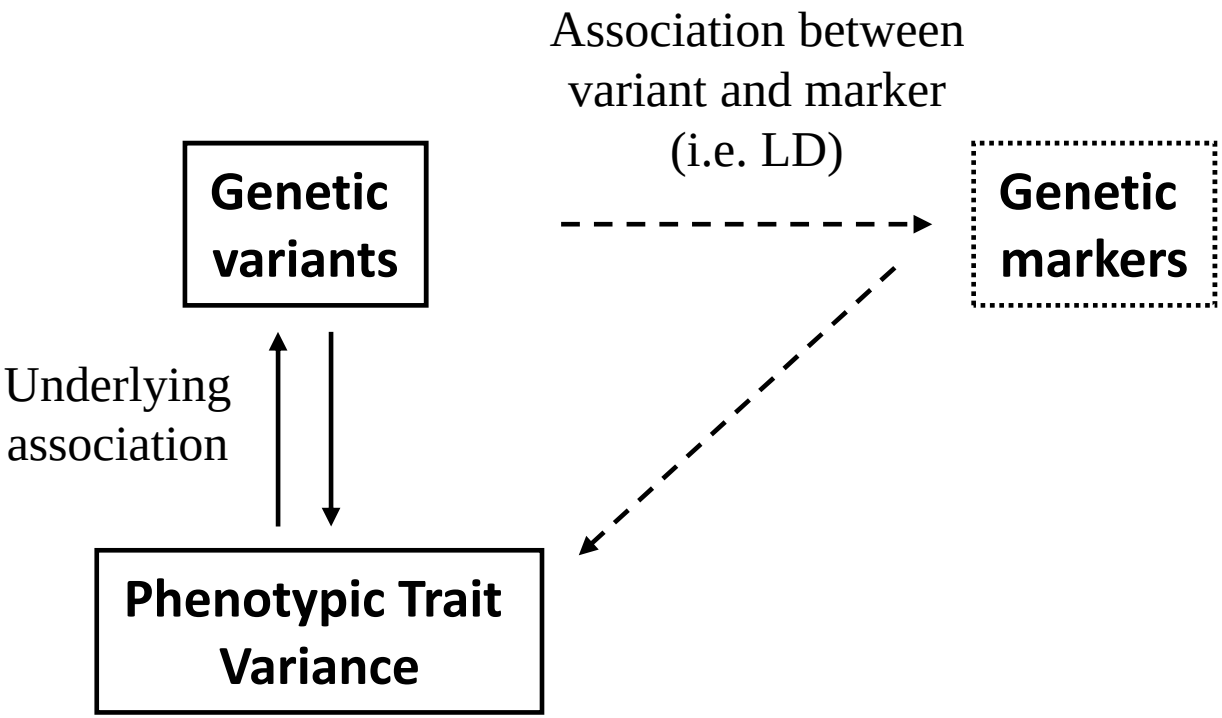
mapping

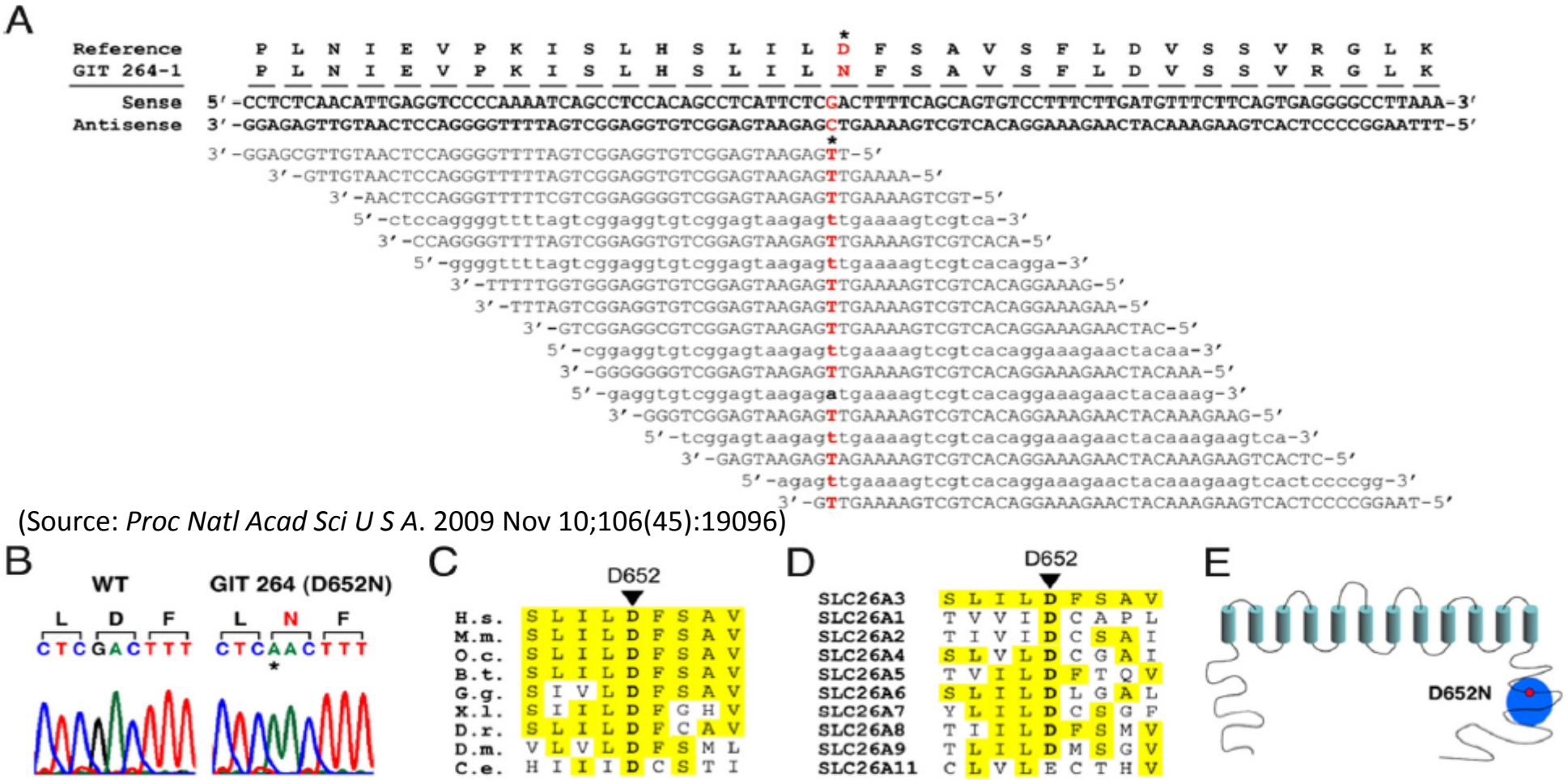
Reference Genome

```
                                                  GGTATAC...
...CCATAG        TAT  CGCCC     CGGAAATTT  CGGTATAC
...CCAT      CTATAT  CG        TCGGAAATT   CGGTATAC
...CCAT  GGCTATAT  CGC  CTATCGGAAA     GCGGTATA
...CCA  AGGCTATAT  CGC CCTATCGGA     TTGCGGTA    C...
...CCA  AGGCTATAT    GCCCTATCG      TTTGCGGT     C...
...CC   AGGCTATAT    GCCCTATCG  AAATTTGC     ATAC...
...CC  TAGGCTATA    CGCCCTA      AAATTTGC  GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

Genetic variants

<u>Association Study</u>: for the given phenotypic trait, "functional variants" could be identified by comparing allele frequencies at hundreds of thousands of polymorphic sites, *i.e* allele A is associated with phenotypic trait P if (and only if) people who have P also have A more (or less) often than would be predicted from individual frequencies of A and P in the assessed population.

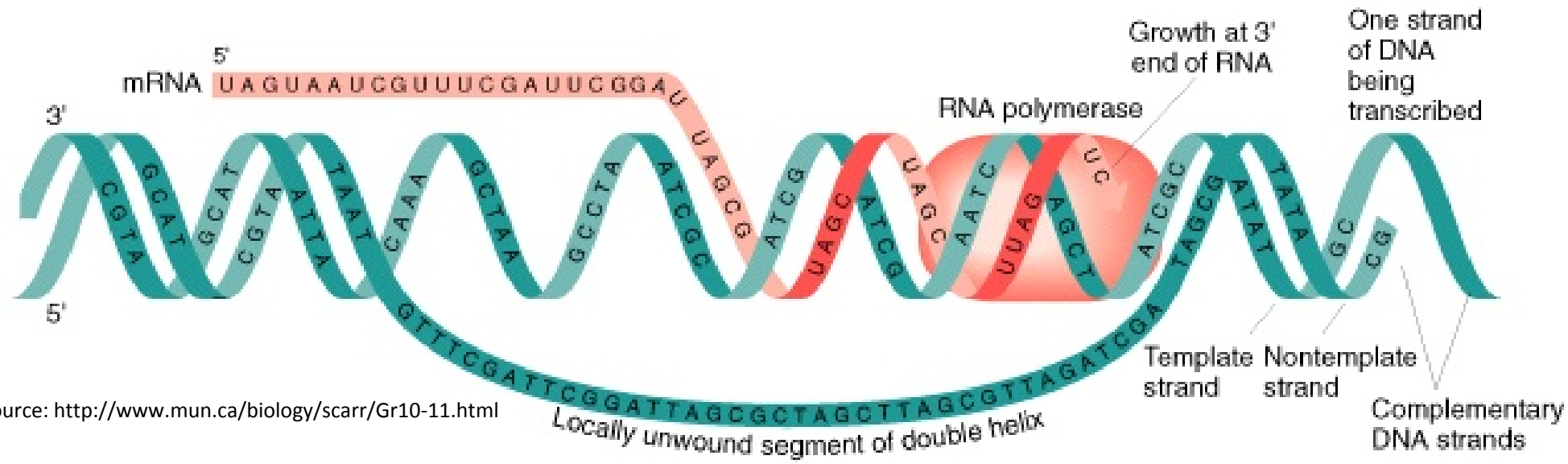




(Source: *Proc Natl Acad Sci U S A*. 2009 Nov 10;106(45):19096)

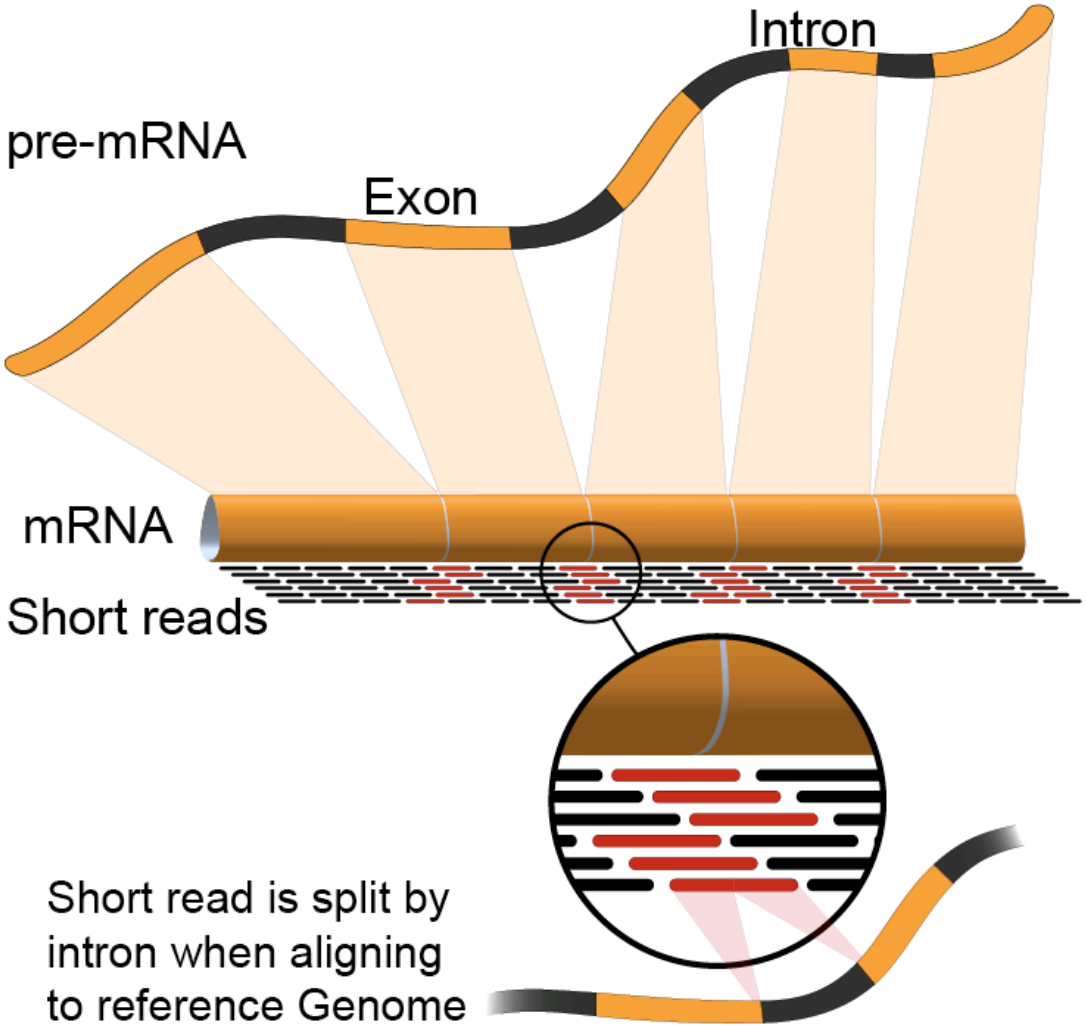
Choi *et al.*used whole-exome sequencing to discover the cause of disease in an individual with an unclear diagnosis. They identified a missense mutations in positions that were highly conserved from invertebrates to humans, in a gene known to cause congenital chloride-losing diarrhoea, consistent with the patient's symptoms.

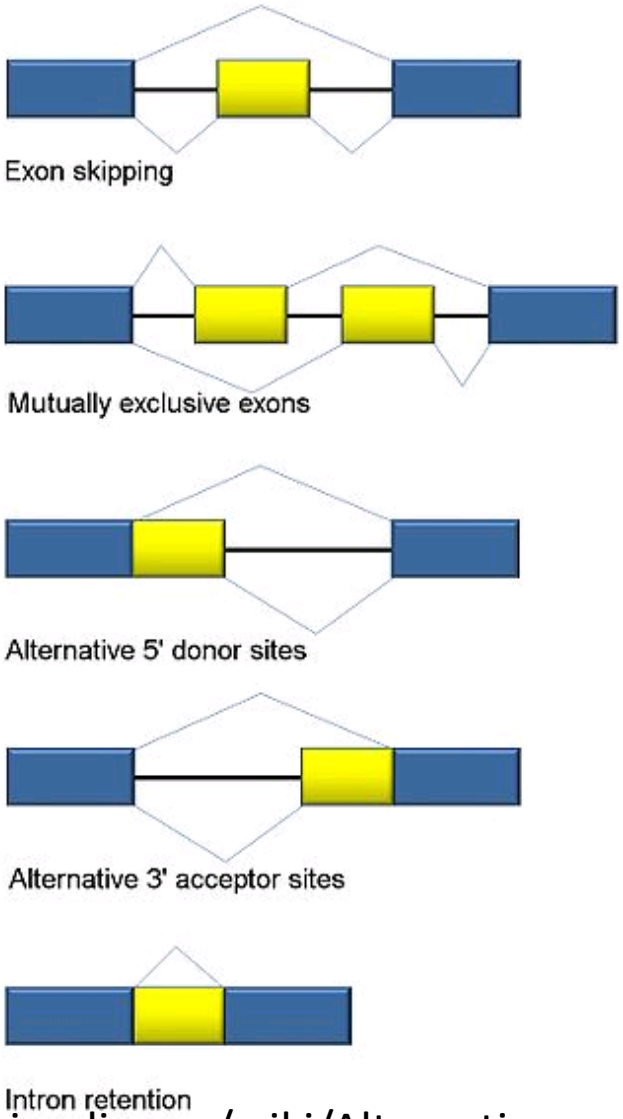
RNA-Seq: Explore the transcriptome



Source: http://www.mun.ca/biology/scarr/Gr10-11.html

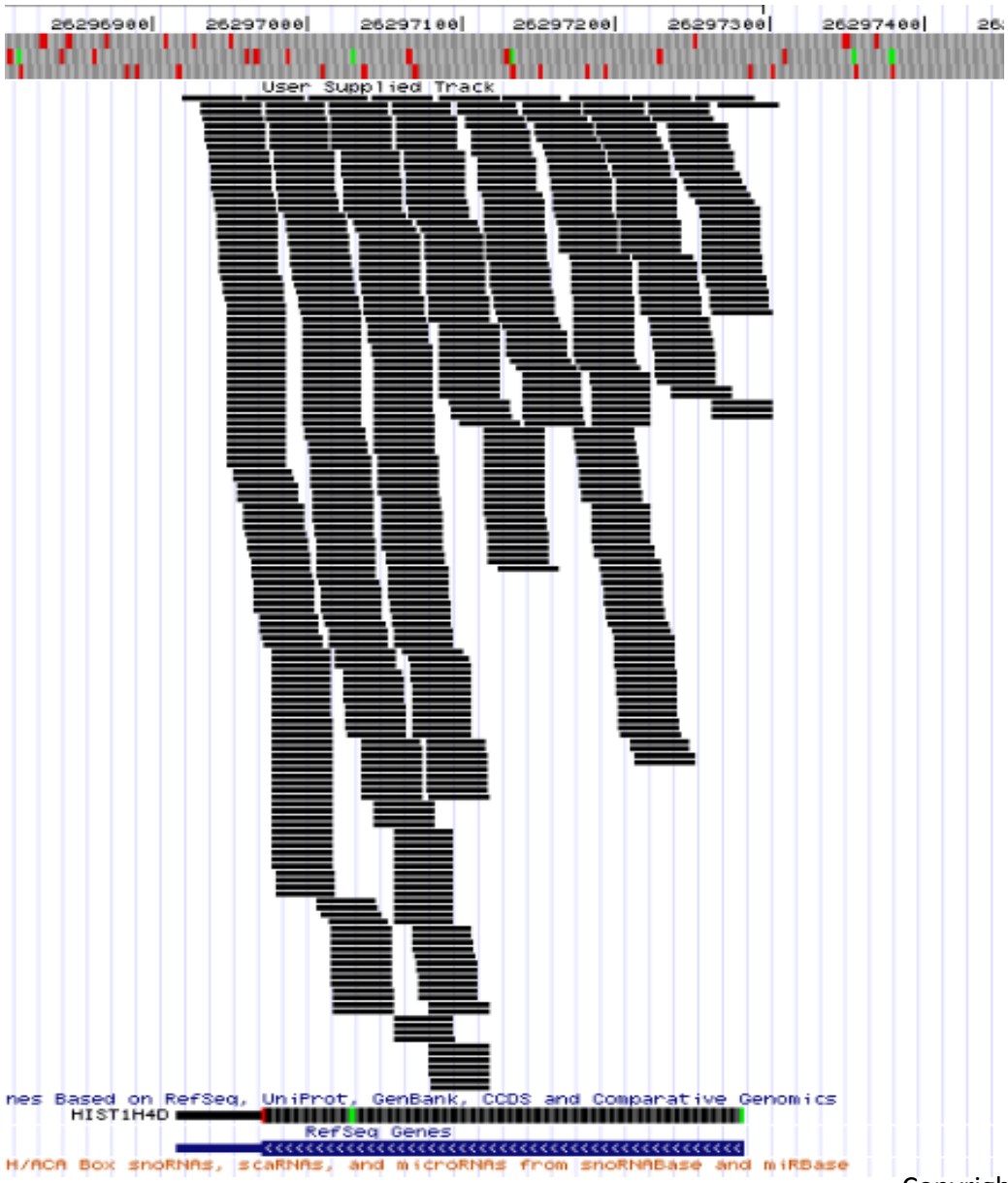
"*A transcriptome is a collection of all the transcripts present in a given cell.*" (NHGRI factsheet, NIH, US)





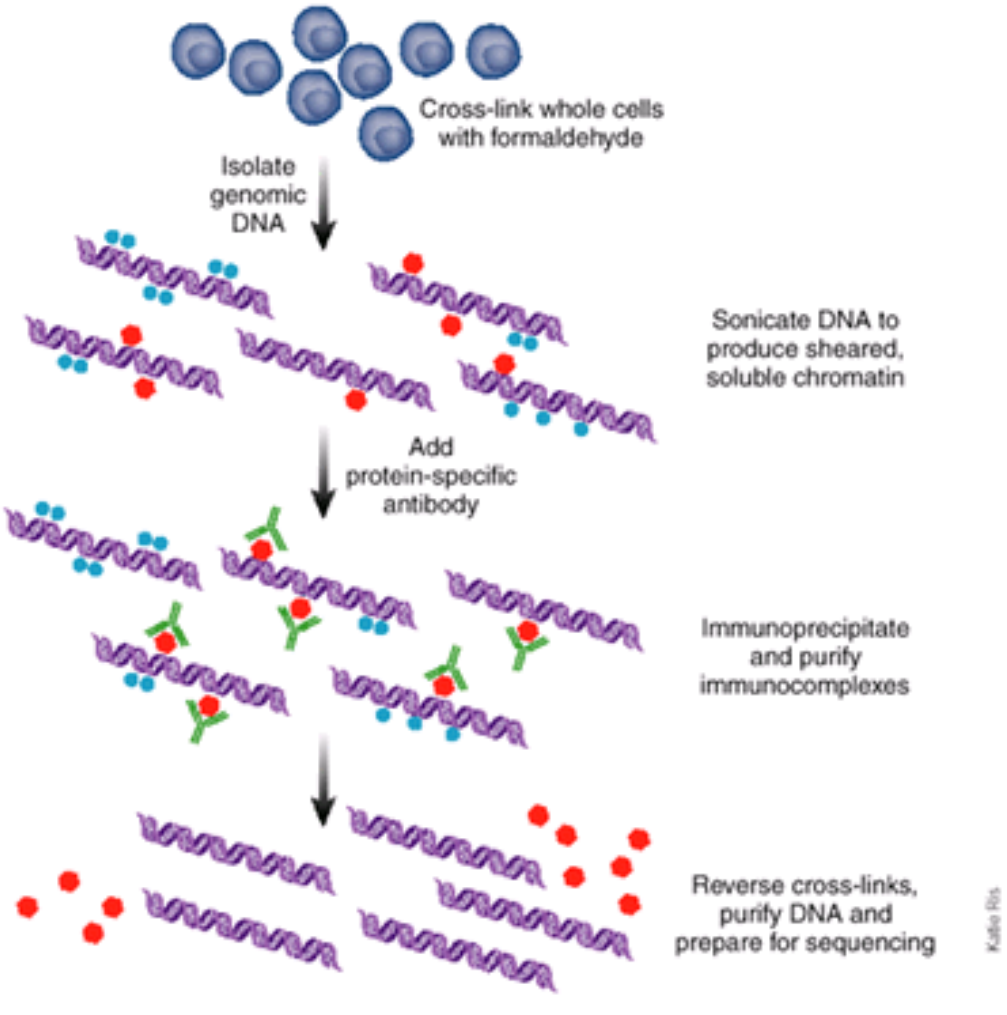
http://en.wikipedia.org/wiki/RNA-Seq

http://en.wikipedia.org/wiki/Alternative_splicing



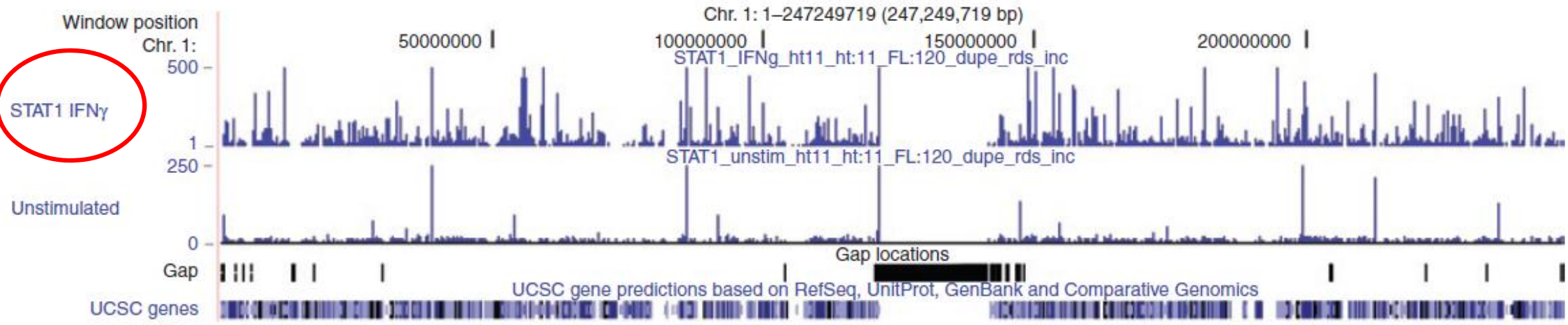
	B	C	D	E	F
1	gene	nsc1	nsc1 SE	nsc2	nsc2 SE
2	brain protein	18.9574	3.79952	21.5848	3.02241
3	Cluster Incl AW1	110.513	7.84625	114.894	7.95669
4	Cluster Incl AI8	235.873	35.6748	210.349	27.612
5	Cluster Incl AV3	47.4605	3.94976	29.6941	3.6586
6	Cluster Incl AV1	28.4527	3.74512	15.2986	3.62097
7	Cluster Incl AV1	80.302	6.45368	107.23	8.09591
8	Cluster Incl AV3	40.8113	5.13418	54.0835	3.18591
9	Cluster Incl AI1	53.1437	3.63392	58.635	5.50994

Chromatin ImmunoPrecipitation Sequencing (ChIP-Seq):

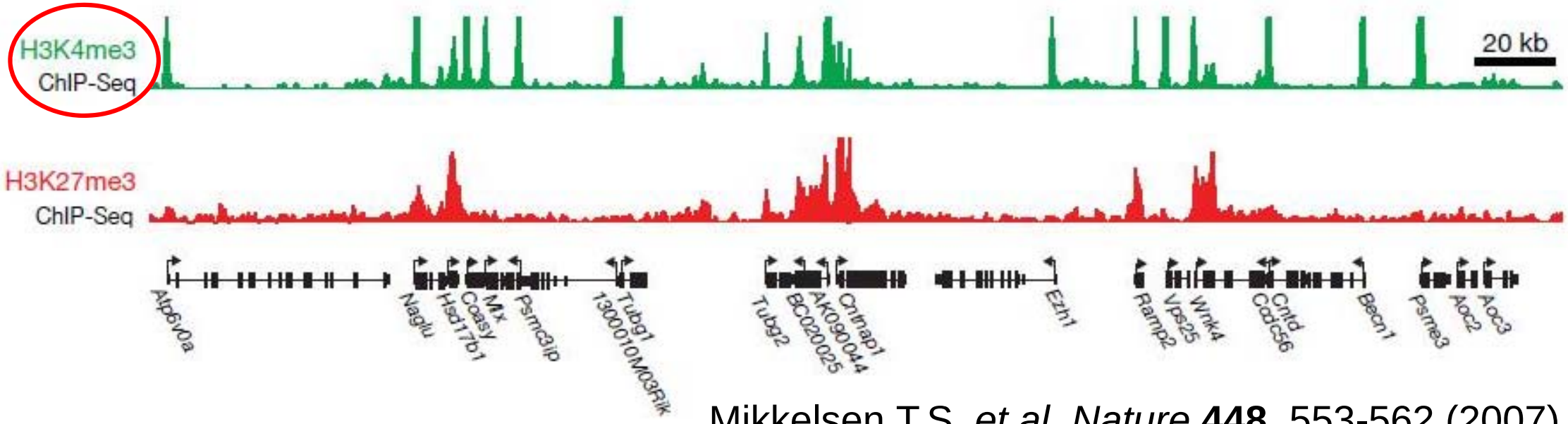


Source: *Nature Methods* 4:613

Profile Protein-DNA interaction



Robertson, G. *et al*. *Nat. Methods* **4**, 651-657 (2007)



Mikkelsen,T.S. *et al*. *Nature* **448**, 553-562 (2007)

生物信息学：导论与方法
Bioinformatics: Introduction and Methods