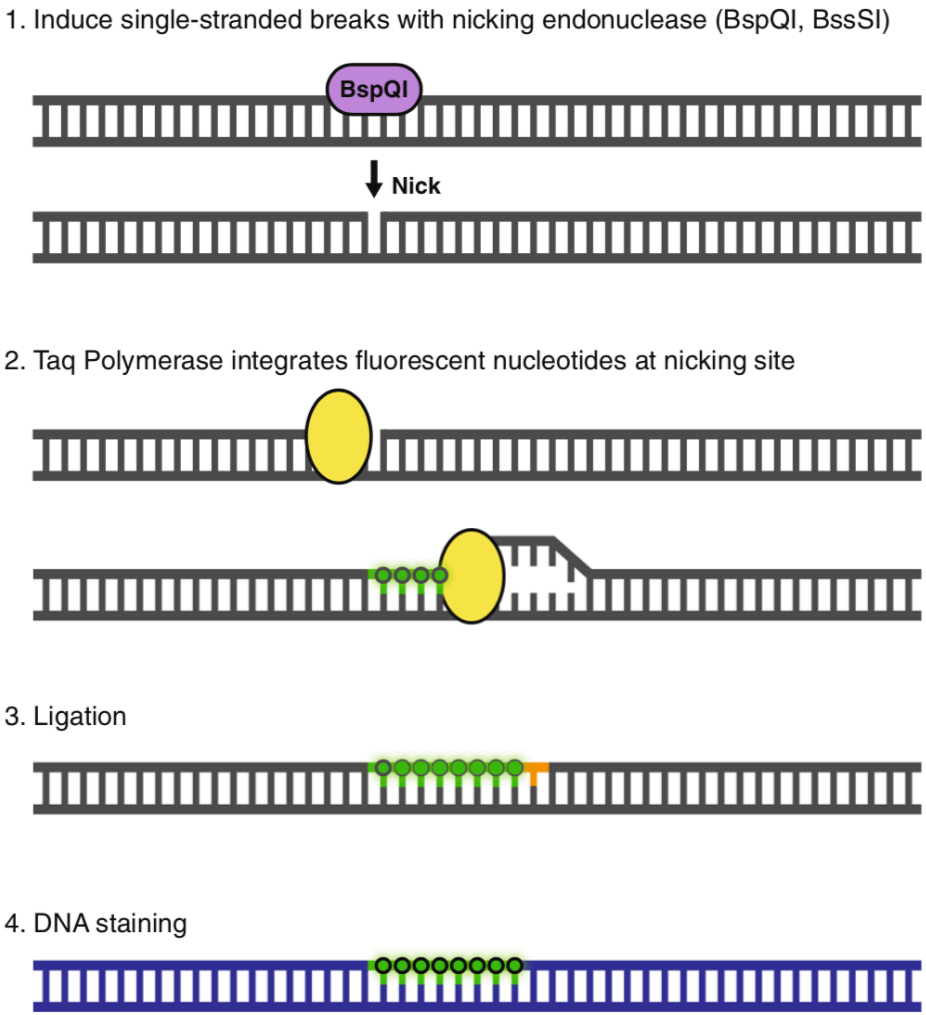
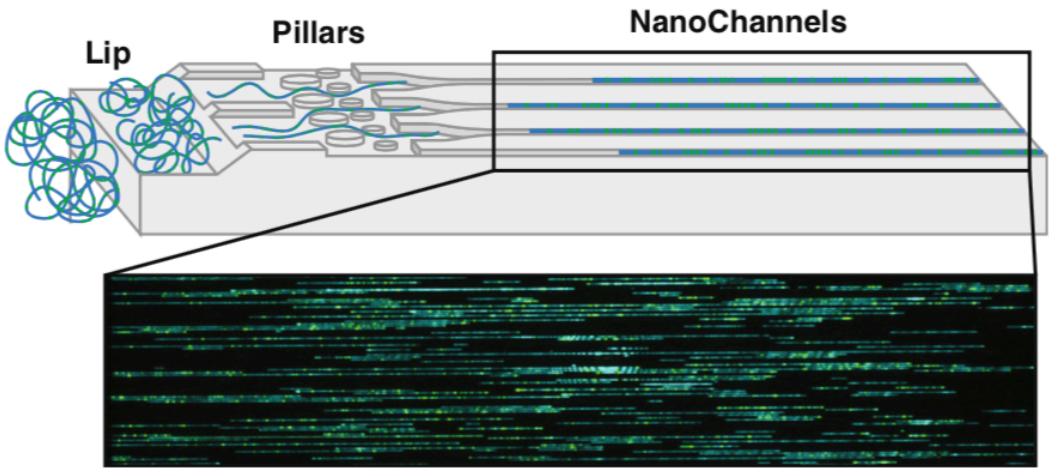
Optical mapping: bionano DNA labeling



The DNA labeling workflow is divided into four consecutive steps. First, the high molecular weight DNA is nicked with an endonuclease of choice that introduces single strand nicks throughout the genome. Second, Taq polymerase recognizes these sites and replaces several nucleotides with fluorescently tagged nucleotides added to the solution. Third, the two ends of the DNA are ligated together using DNA ligase. Fourth, the DNA backbone is stained with DNA Stain.

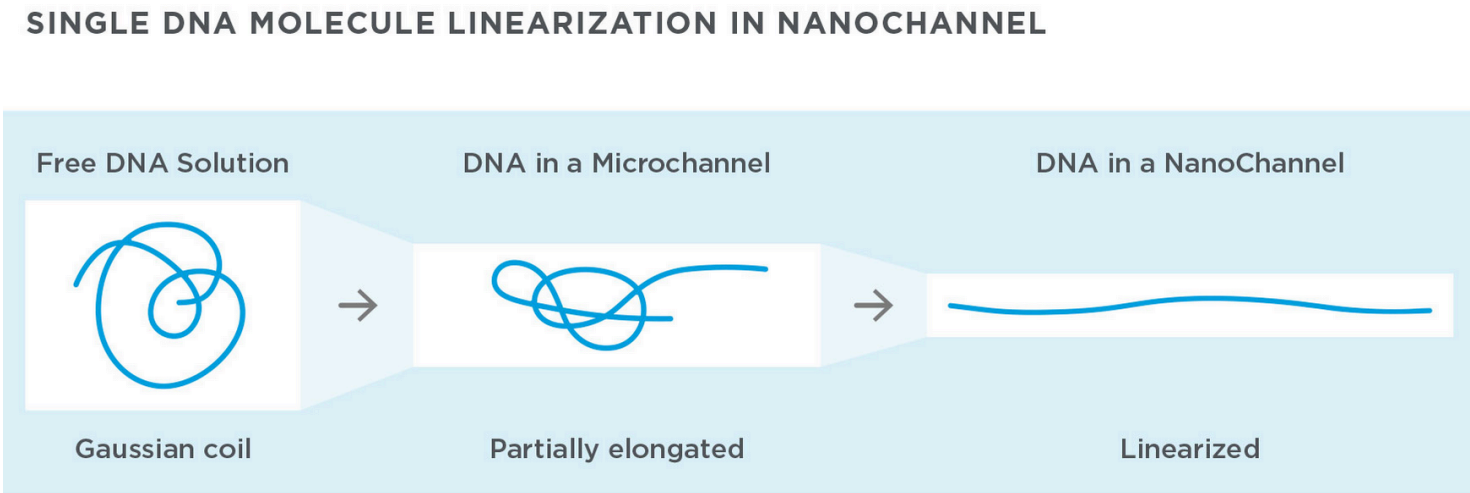
Barseghyan et al. Genome Medicine (2017)

Optical mapping: bionano DNA loading



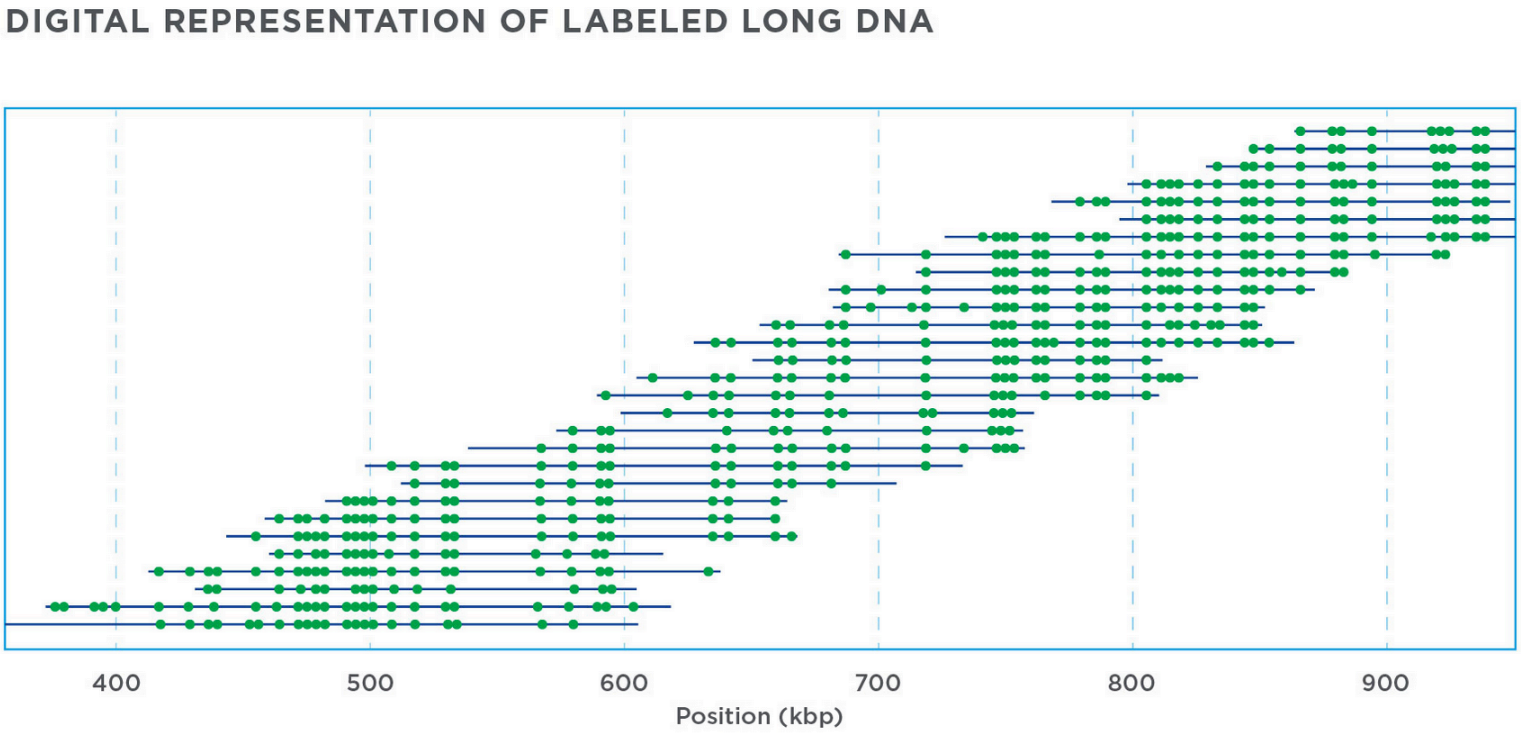
Barseghyan et al. Genome Medicine (2017)

The labeled dsDNA is loaded into chip flowcells. The applied voltage concentrates the coiled DNA at the lip (left). Later, DNA is pushed through pillars (middle) to uncoil/straighten, then into nanochannels (right). DNA is stopped and imaged in the nanochannels. Blue=staining of DNA backbone, green=fluorescently labeled nicked sites



https://bionanogenomics.com/

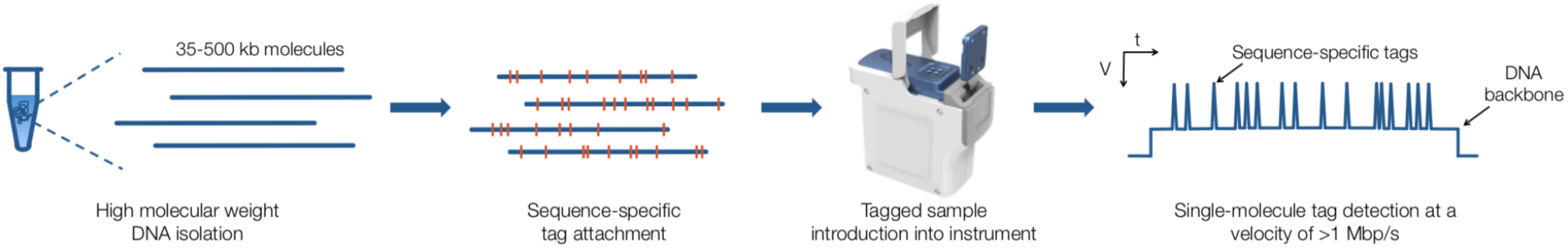
Optical mapping: bionano DNA visualization



Once raw image data of labeled long DNA molecules is captured, it is converted into digital representations of the motif-specific label pattern. A proprietary software then assembles the data *de novo* to recreate a whole genome map assembly.

https://bionanogenomics.com/

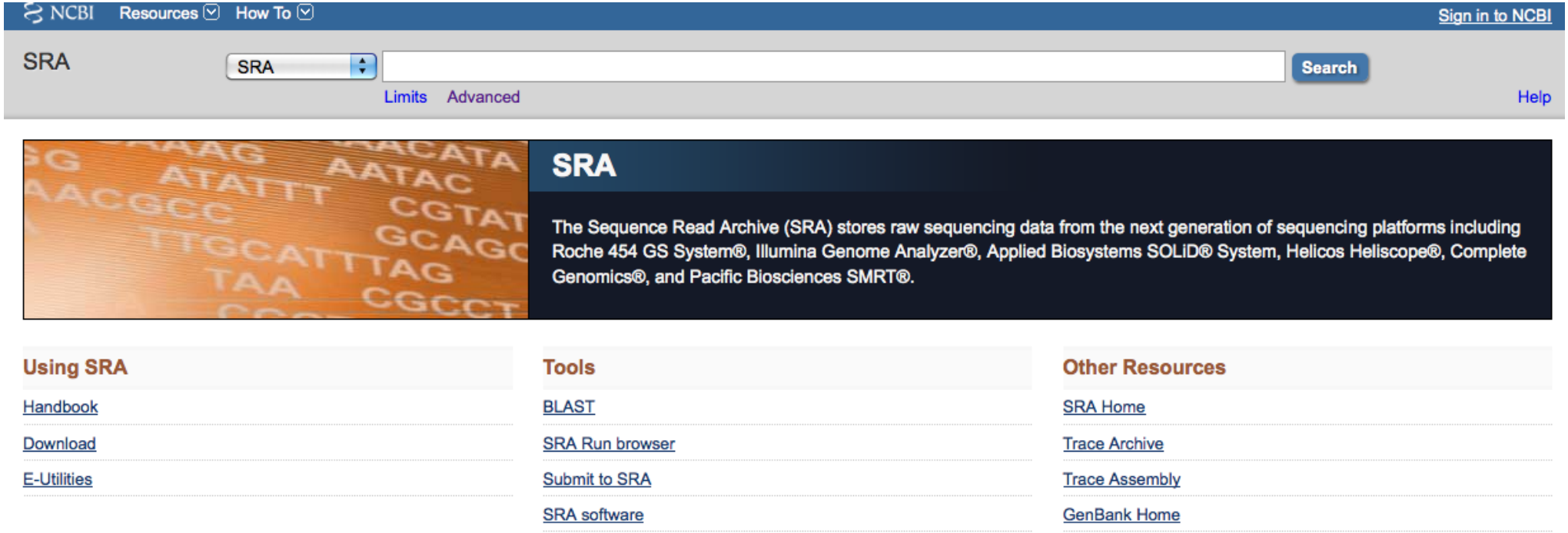
Electronic mapping: Nabsys



In order to construct whole genome electronic maps, high molecular weight genomic DNA is isolated from the cells or tissue of choice. The high per-molecule information content of Nabsys single-molecule reads allows for a solution-phase DNA isolation procedure, producing DNA in the 35-500 kb range, obviating the need for time consuming gel plug isolation protocols. Following purification, the DNA is tagged in a sequence-specific manner through an enzymatic nicking reaction. As single molecules pass through the detector, the presence of the DNA backbone and attached tags are sensed as changes in the resistance of the detector. The resulting data indicate the time between tag sites on each single-molecule DNA backbone. The temporal events are then converted to distance-based events where the distance between tags (termed an “interval”) is reported in base-pairs.

http://nabsys.com/Home.aspx

SRA database



Display Settings: Full

Send to:

WGS of Sa_JKD6272

Accession: SRX031534
Experiment design: n/a
Submission: SRA026511 by University of Melbourne
Study summary: Sequencing Australian cMRSA (**SRP004474**) • Study • All experiments (more...)
Sample: (SRS121467) (more...)
Library: Sa_JKD6272 (more...)
Platform: Illumina (more...)
Processing:
Base calls: Base Space, Solexa primary analysis
Quality score: Solexa primary analysis, 80x1
Spot descriptor:

Total: 1 run, 11M spots, 789.7M bases
Download reads for this experiment in sra (795.3M) *or sra-lite* (795.3M) *formats*

#	Run	# of Spots	# of Bases
1.	SRR073304	10,967,442	789.7M

ID: 38437

Related information
BioSample
PubMed

Recent activity
Turn Off Clear
SRP004474 (3) SRA
SRP007764 (1) SRA
exome (38813) SRA
SRP007744 (1) SRA
SRP (0) SRA
See more...

Alias: Sa_JKD6272
Instrument model: Illumina Genome Analyzer II
Date of run:
Run center: University of Melbourne
Statistics:
Number of spots: 10967442
Number of reads: 21934884

Design:
Platform: Illumina
Sample: WGS of cMRSA strain Sa_JKD6272
Library:
Name: Sa_JKD6272
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED (NOMINAL_LENGTH=250, NOMINAL_SDEV=50)
Construction Protocol:

Download (hide):

Object		.lite.sra				.sra			
Run	SRR073304	834.0M	HTTP	FTP	Aspera	834.0M	HTTP	FTP	Aspera
Experiment	SRX031534	834.0M	HTTP	FTP	Aspera	834.0M	HTTP	FTP	Aspera
Study	SRP004474	1.8G	HTTP	FTP	Aspera	1.8G	HTTP	FTP	Aspera

Filter: Find Filtered Download What does it do?
What can the filter be applied to?

< 1 1 1096745 >

View: ☑ reads (customize)

1. SRR073304.1
name: HWI-EASXXX:1:1:7:1420, member: default
x: 7, y: 1420

2. SRR073304.2
name: HWI-EASXXX:1:1:7:1227, member: default
x: 7, y: 1227

3. SRR073304.3
name: HWI-EASXXX:1:1:7:857, member: default
x: 7, y: 857

4. SRR073304.4
name: HWI-EASXXX:1:1:8:1518, member: default

Reads (separated)

>gnl|SRA|SRR073304.1.1 HWI-EASXXX:1:1:7:1420 *F (Biological)*
GTNCAATTCGCCGTAATCGTGCTGGGTTTGATGACG

One channel quality score

```
33 27  0 30 33 30 33 34 34 34 34 34 29 31 33 26 31 29 25 32
30 33 33 29 15 15 31 33 30  7 25 31 15 31 29  2
```

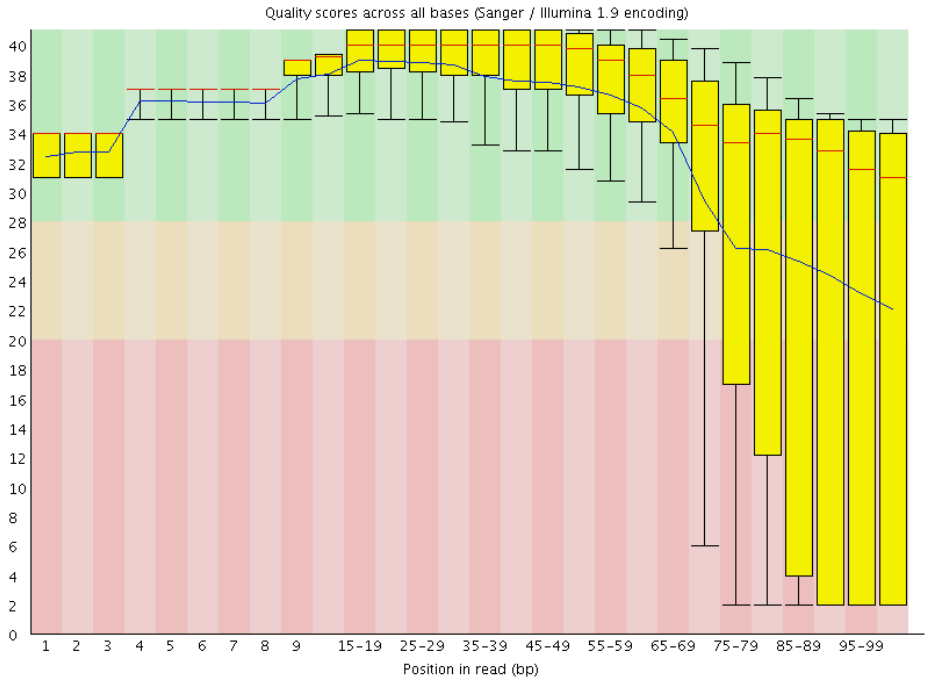
>gnl|SRA|SRR073304.1.2 HWI-EASXXX:1:1:7:1420 *R (Biological)*
NCTTTATTACCAAATAGATCAATTGCTAATTTTTTA

One channel quality score

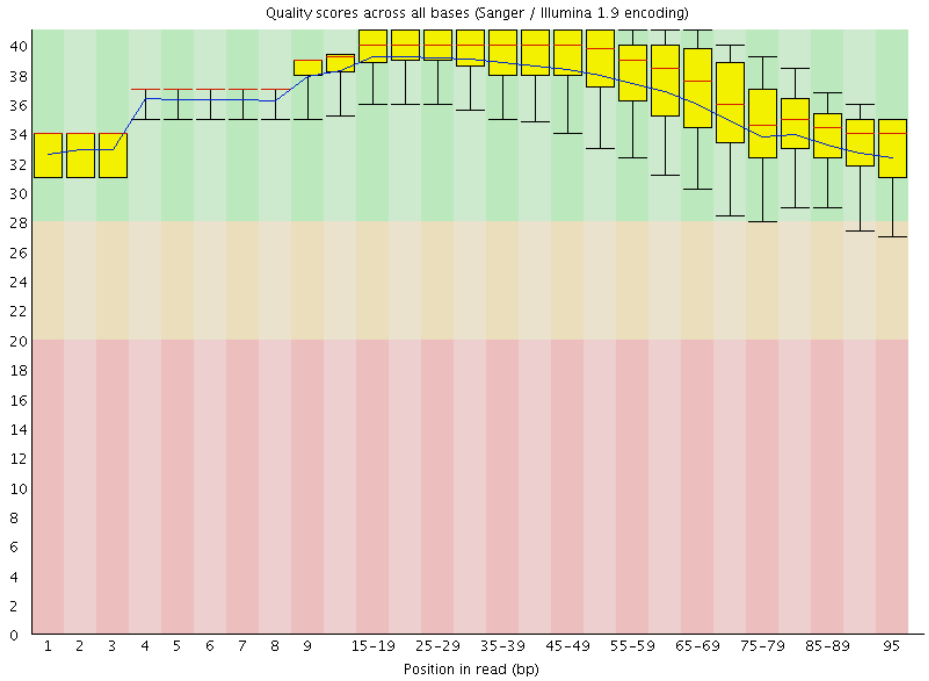
```
 0 13 23 27 27 22 24 24 26 23 25 17 24 25 24 25 14 18 25 22
12 17 24 25 19 25 21 17 18 25 17  6 18 27 22 20
```

Quality check

Per base sequence quality



After sequencing

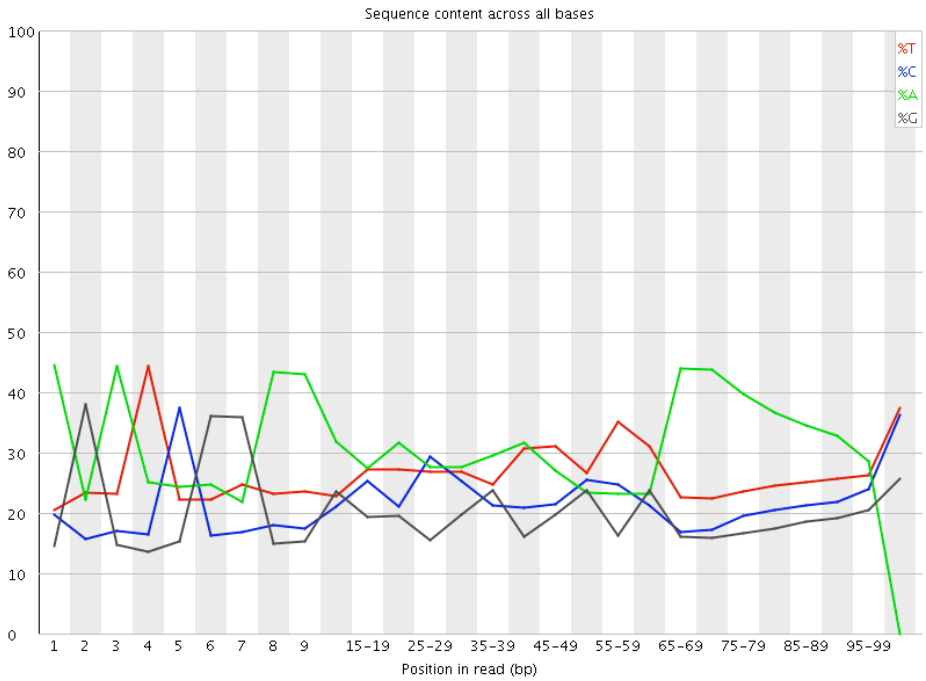


After adaptor trimming and removal of low quality regions

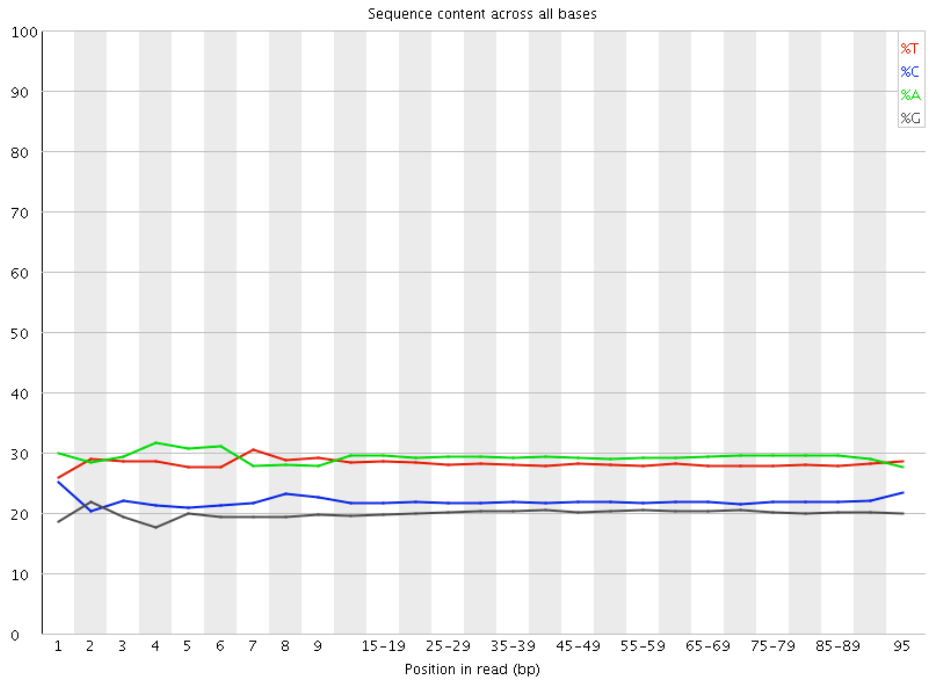
Generated by FASTQC software
http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Quality check

Per base sequence content



After sequencing



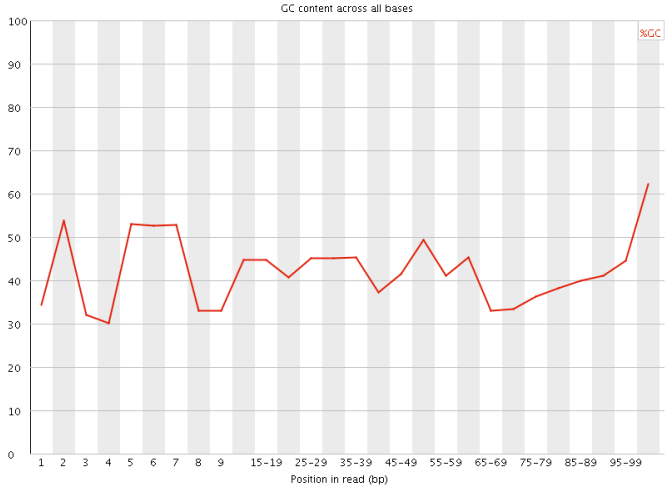
After adaptor trimming and removal of low quality regions

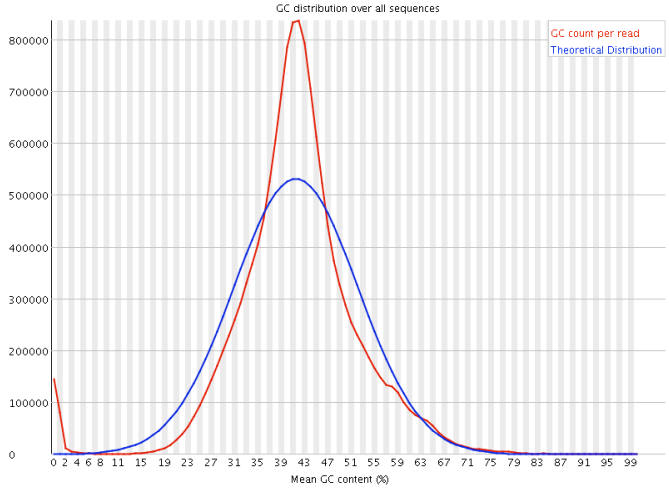
Generated by FASTQC software
http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Quality check

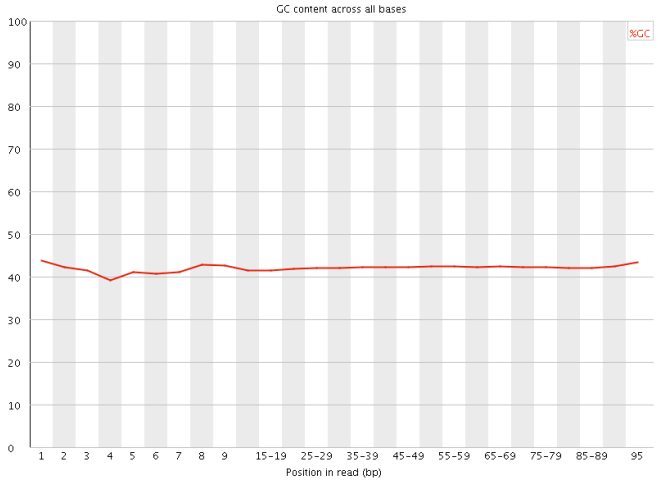
GC content

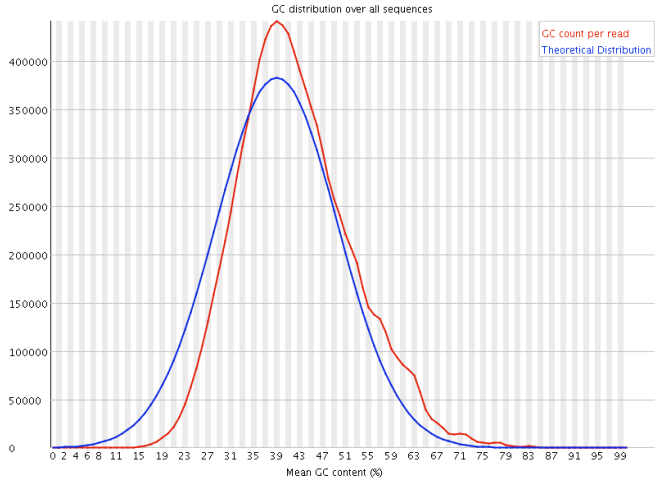
After sequencing





After adaptor trimming and removal of low quality regions

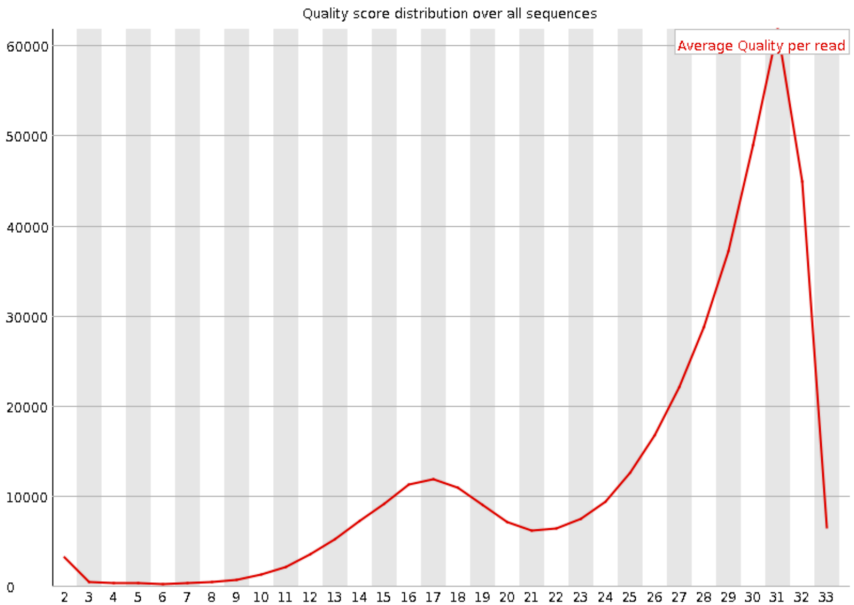




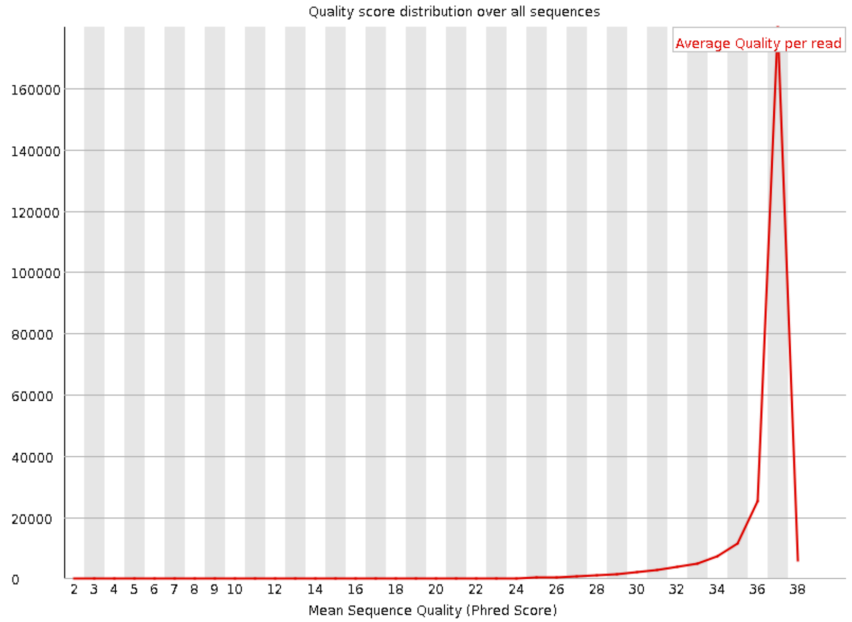
Generated by FASTQC software
http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Quality check

Per sequence quality score



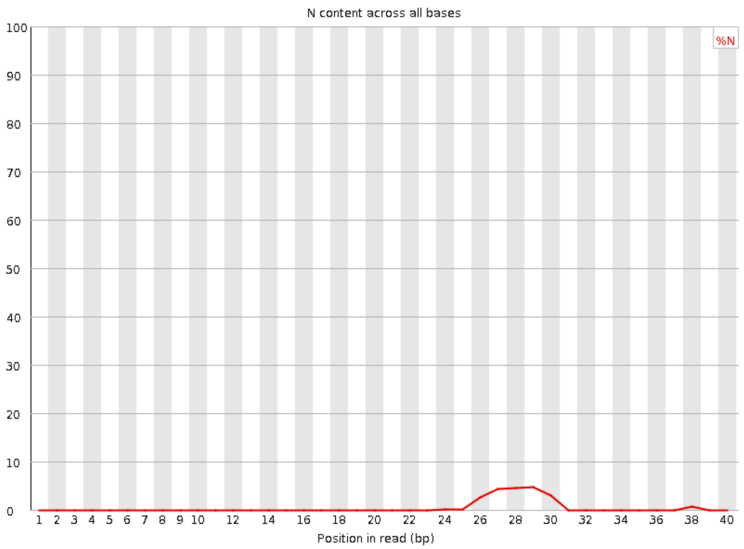
Bad sequencing



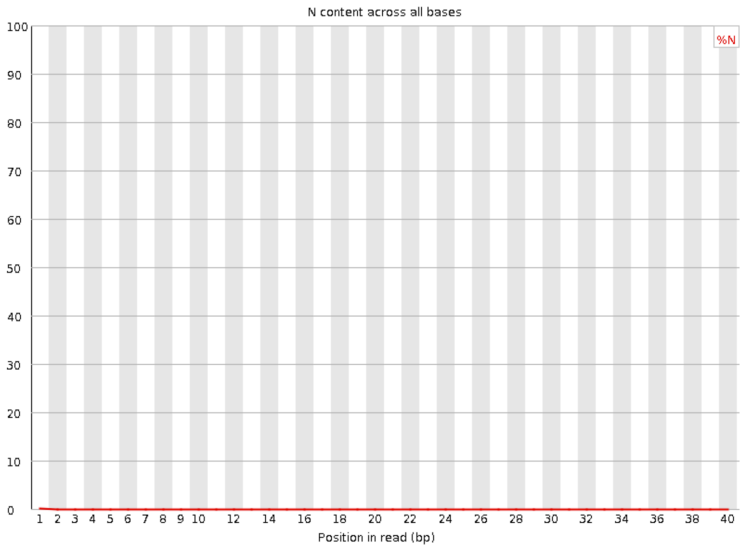
Good sequencing

Quality check

Per base N content



Bad sequencing



Good sequencing

Quality check

Over-represented sequences

Overrepresented sequences

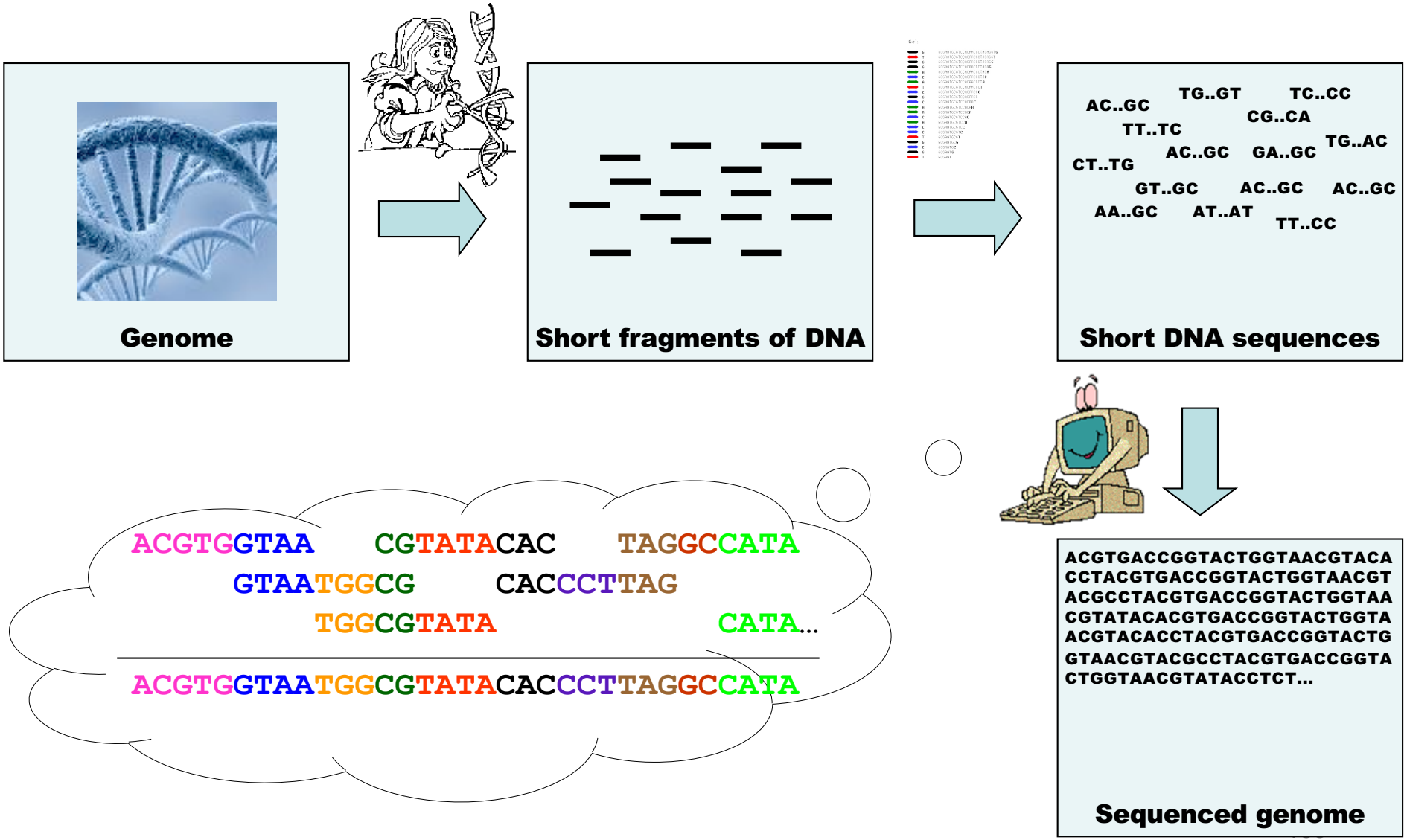
Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit

Applications of NGS platforms

- **DNA sequencing**
  - genome re-sequencing (SNPs, CNV, GWAS)
  - *de novo* sequencing
  - identification of genome structural variants (cancer genome)
  - 3D chromatin interactions
  - Epigenomics (chromatin state and genome methylation)
  - Metagenomics (taxonomic analysis of environmental samples)

- **RNA sequencing**
  - Qualitative and quantitative analysis of the Transcriptome
  - Identification and characterization of miRNAs and other ncRNAs
  - RNA editing
  - Metatrancriptomics (functional analysis of environmental samples)

Genome re-sequencing



Whole genome sequencing

Vol 452|17 April 2008 nature

NEWS & VIEWS

HUMAN GENETICS

Dr Watson's base pairs

Maynard V. Olson

The application of new technology to sequence the genome of an individual yields few biological insights. Nonetheless, the feat heralds an era of 'personal genomics' based on cheap sequencing.

This issue of *Nature* contains a paper that is, in a curious way, a sequel to one published 55 years ago — the description by James Watson and Francis Crick[1] of the double-helical structure of DNA. At the information-carrying core of this beautiful structure, with its far-reaching implications for biology and medicine, are the base pairs that Watson discovered by fitting together cardboard cut-outs of the bases adenine, thymine, guanine and cytosine. Now, on page 872, Wheeler *et al.*[2] describe the use of massively parallel DNA sequencing to determine the order of the base pairs in Watson's own genome. This achievement is a technical *tour de force* that points towards routine use of whole-genome sequencing as a research tool in human genetics. Given the choice of James Watson as an identified research subject, the paper is also a conspicuous effort to publicize the arrival of the era of personal genomics and the willingness of a famous geneticist to put his genome sequence in the public domain.

Technically, the paper's interest stems from its reliance on a DNA-sequencing platform that differs greatly from the one used during the first great era of genome sequencing, which culminated in the Human Genome Project (HGP). In the HGP platform, each kilobase-pair fragment of genomic DNA was captured as a bacterial 'clone' using recombinant-DNA techniques and processed in its own microlitre-scale well in a microtitre plate. Following a series of biochemical steps, each sample was analysed electrophoretically in a dedicated, metre-long glass capillary. To achieve the required redundancy in sequence coverage, centres that looked more like manufacturing plants than laboratories. The data-production costs alone were hundreds of millions of dollars.

James Watson decoded.

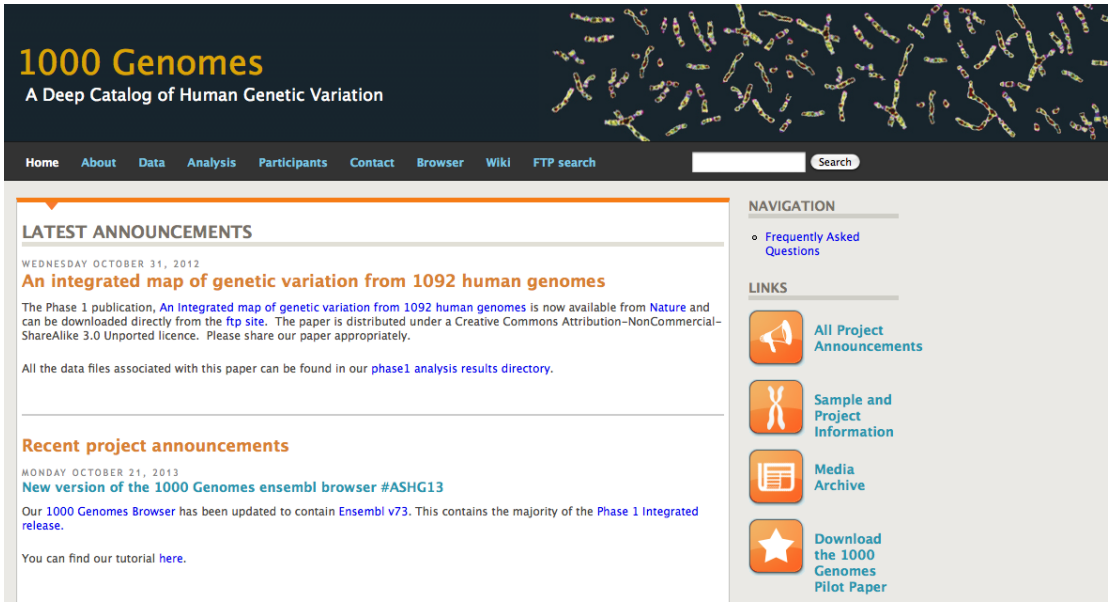
PHOTO BETTMANN/CORBIS

Wheeler *et al.*[2] used one of several new DNA-sequencing platforms that can achieve much the same result at perhaps 1% of the cost[3,4]. Note, efficiency of the new methods lies in massive parallelization of the biochemical and measurement steps. The instruments used by Wheeler *et al.* are marketed by 454 Life Sciences, a component of Roche Diagnostics, which joined forces with the Human Genome Sequencing Center at Baylor College of Medicine in Houston, Texas, to sequence Watson's genome.

The 454 instruments achieve massive parallelization in two different ways[5]. In an initial step, single DNA molecules are attached to synthetic beads and then amplified enzymatically. During amplification, the beads are trapped in tiny water droplets within a water–oil emulsion; hence, more than 100,000 samples can be processed in parallel in a single test tube. In a later step, during which optical measurements are used to collect the actual sequencing data, each bead is confined to a picolitre-scale well etched into the end of a glass fibre within a fibre-optic bundle. Although costs have not yet dropped to the much-ballyhooed target of US$1,000 per genome[6], they are now low enough to make the era of personal genomics a reality rather than a distant dream.

What can we expect to learn from the sequences of individual genomes? The main lesson from the analyses by Wheeler *et al.* is that it will be extremely difficult to extract medically, or even biologically, reliable inferences from individual sequences. Consider the challenge of interpreting Watson's single nucleotide polymorphisms (SNPs — simple substitutions of one base for another at a particular site in the genome). Wheeler *et al.* report about 3,300,000 SNPs in Watson's genome relative to the HGP reference

http://www.internationalgenome.org/



Home > The 100,000 Genomes Project

https://www.genomicsengland.co.uk/the-100000-genomes-project/

The 100,000 Genomes Project

The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

The project will also enable new medical research. Combining genomic sequence data with medical records is a ground-breaking resource. Researchers will study how best to use genomics in healthcare and how best to interpret the data to help patients. The causes, diagnosis and treatment of disease will also be investigated. We also aim to kick-start a UK genomics industry. This is currently the largest national sequencing project of its kind in the world.

UNIVERSITY OF ALBERTA

1000 Plants

Search this site

- HOME
- CONTACT INFO
- GREEN PLANTS
- MEDIA
- SUB-PROJECTS
  - AGRICULTURE
  - ANGIOSPERMS
  - BIOCHEMISTRY
  - EXTREMOPHYTES
  - GREEN ALGAE
  - MEDICINES
  - NON-FLOWERING
- SITEMAP

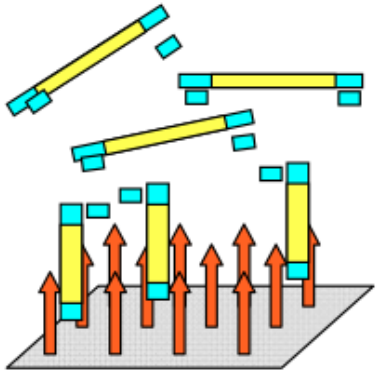
Home

The 1000 plants (oneKP or 1KP) initiative is an international multi-disciplinary consortium that has generated large-scale gene sequencing data for over 1000 species of plants. Major supporters include Alberta Ministry of Innovation and Advanced Education, Musea Ventures (Somekh Family Foundation), Beijing Genomics Institute in Shenzhen (BGI-Shenzhen), China National GeneBank (CNGB), iPlant Tree-of-Life (iPToL) Grand Challenge, Compute Canada (Westgrid), Alberta Innovates Technology Futures (AITF-iCORE Strategic Chair). The sample selection was originally based on a series of overlapping sub-projects with scientific objectives that could be addressed by sequencing multiple plant species (links on left). As more collaborators joined 1KP, however, the objectives evolved and are now exemplified by the diverse collection of papers described by the links below.

https://sites.google.com/a/ualberta.ca/onekp/

Capturing

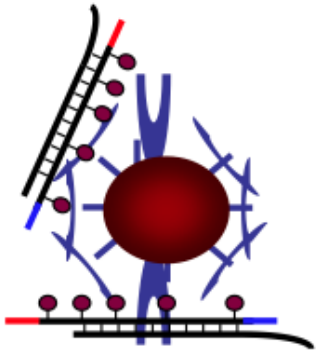
- Array hybridization



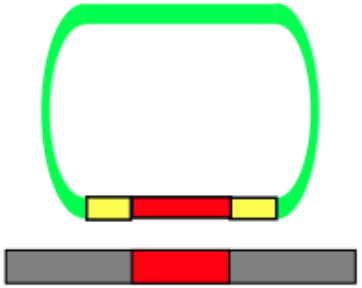
Agilent Technologies

NimbleGen SYSTEMS, INC.

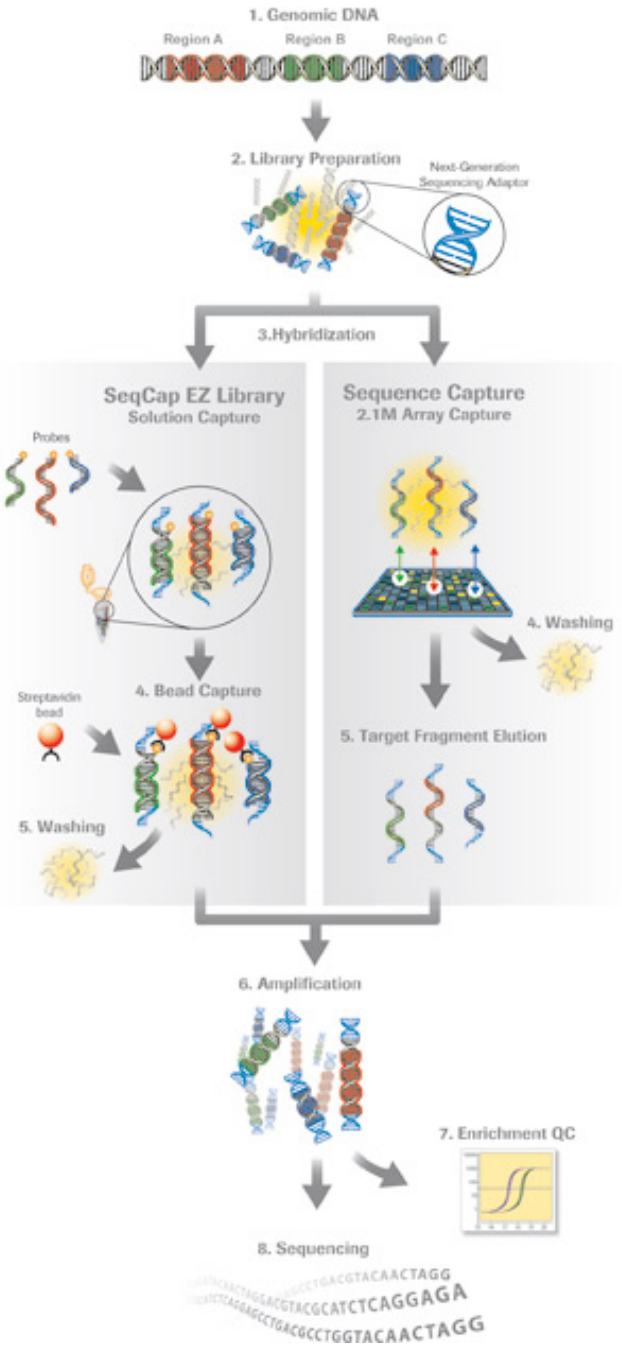
- In solution hybridization

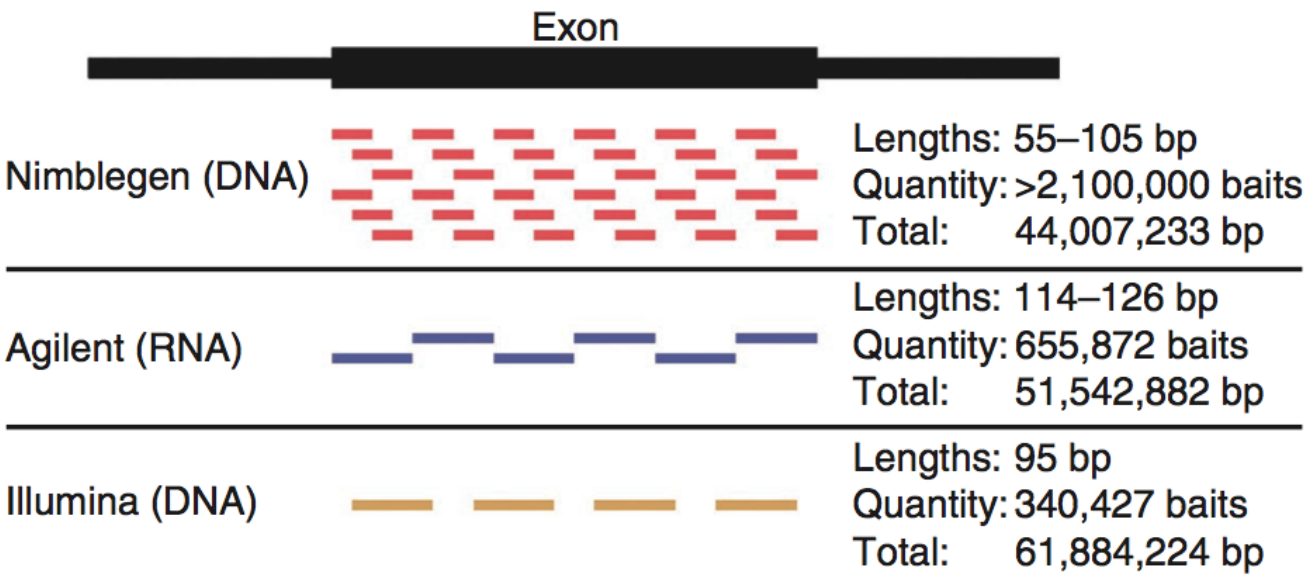


- Molecular inversion probes

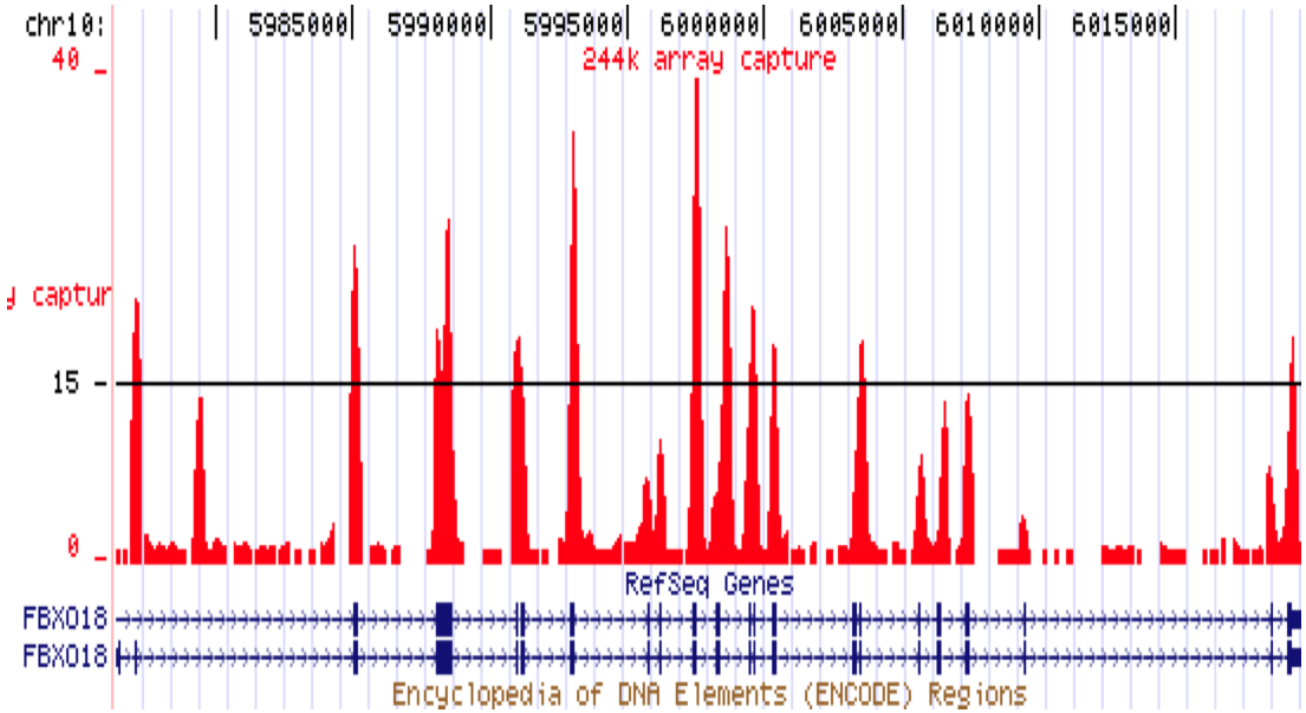


Exome sequencing

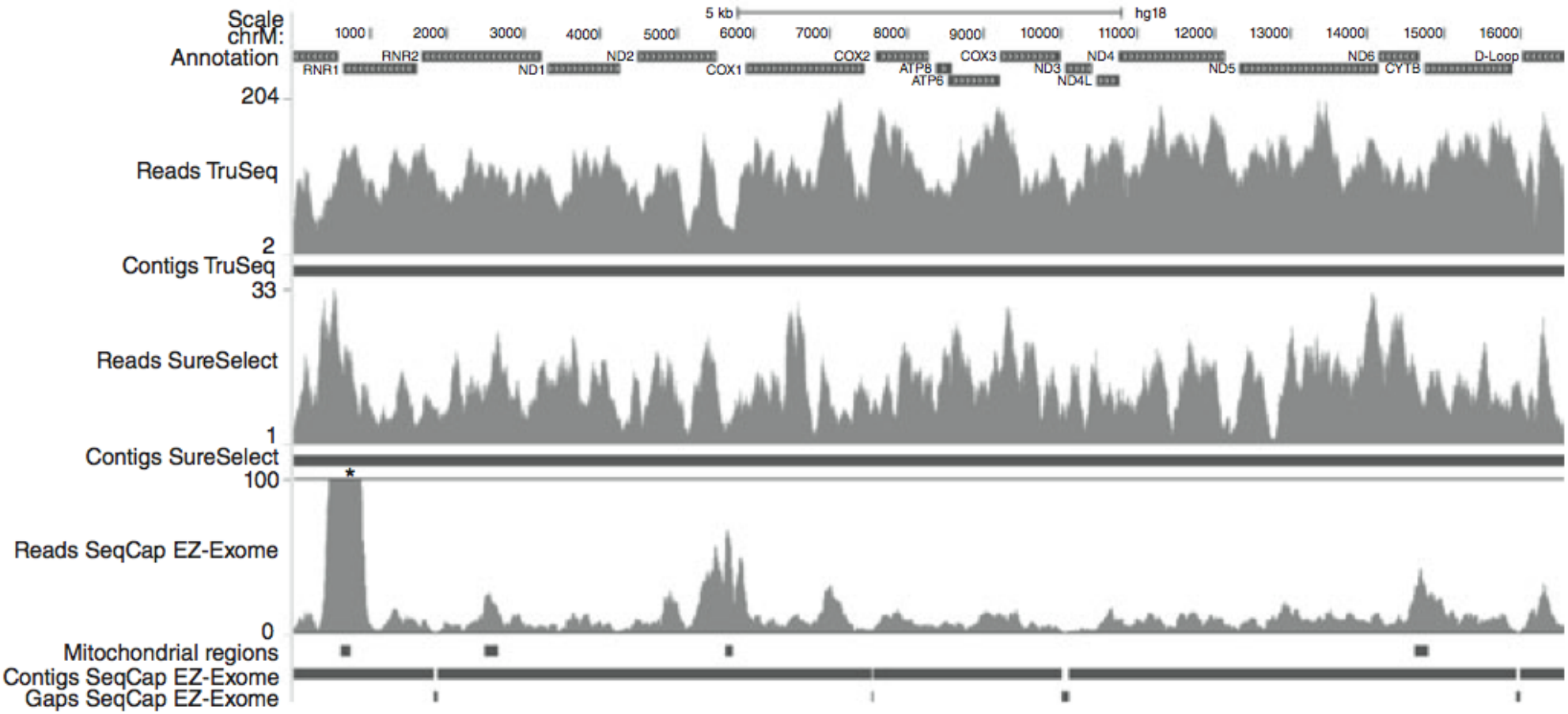




Exome sequencing



Exome sequencing



Picardi and Pesole 2013 Nature Methods

Structural Variations

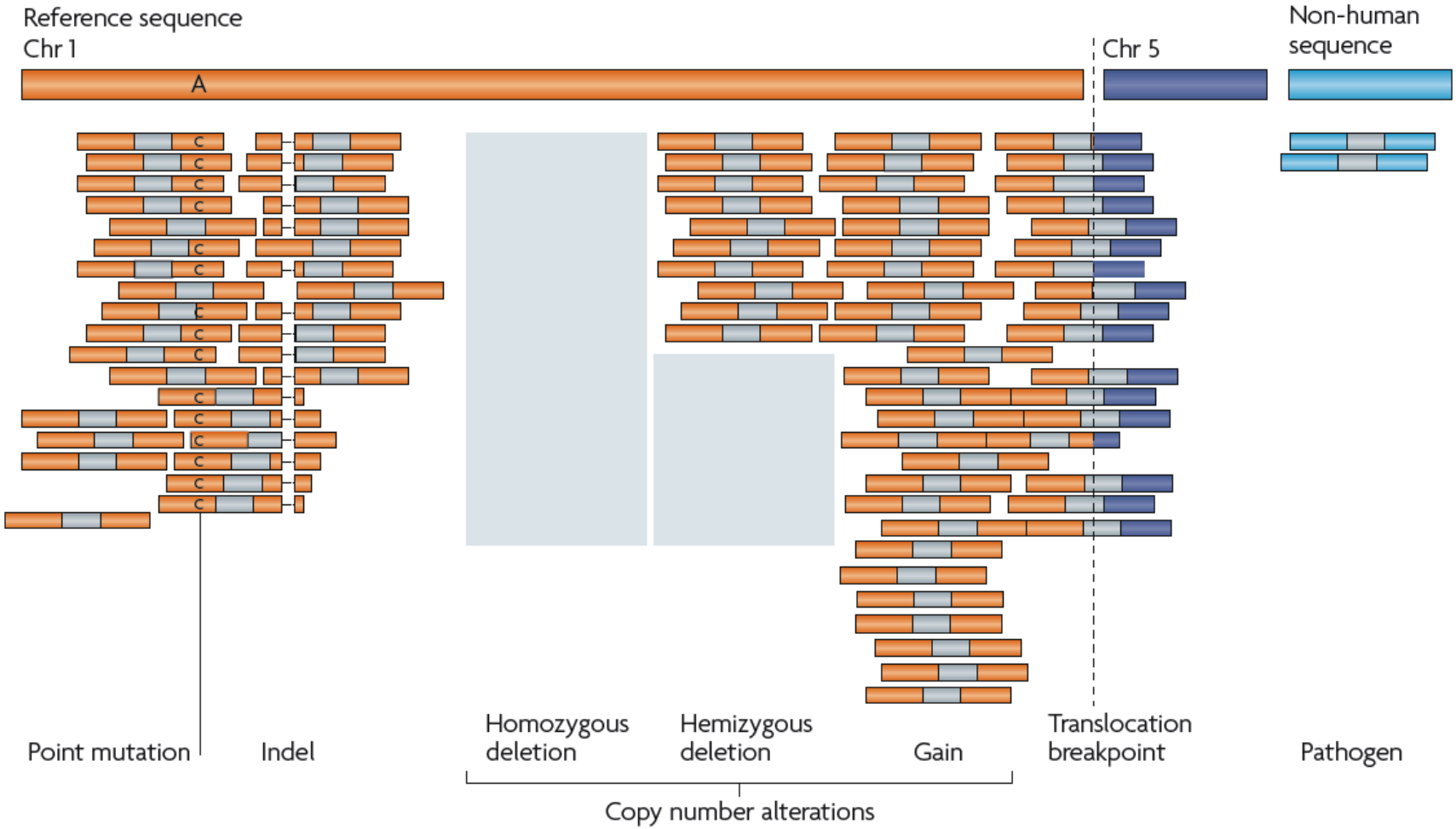
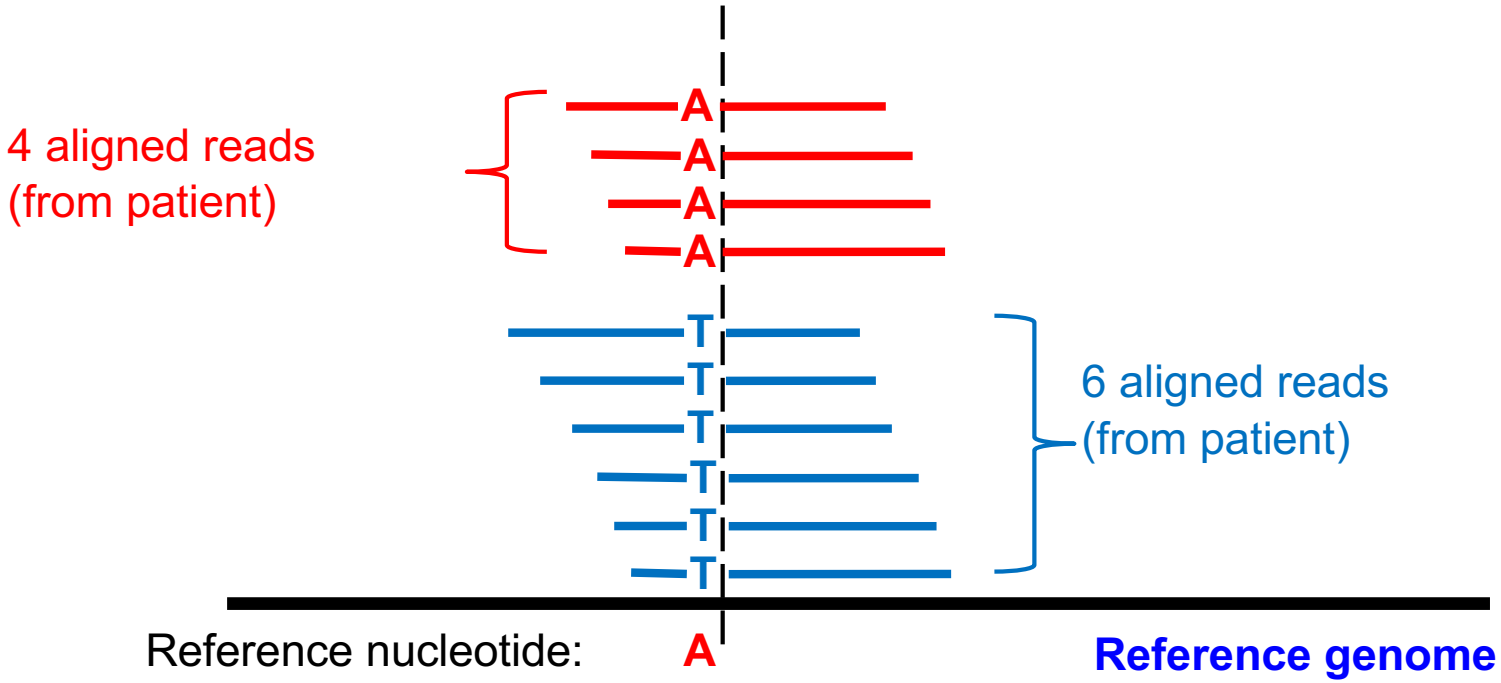


Figure 3 | **Types of genome alterations that can be detected by second-generation sequencing.** Sequenced

SNP detection



Total number of reads covering indicated position: 10

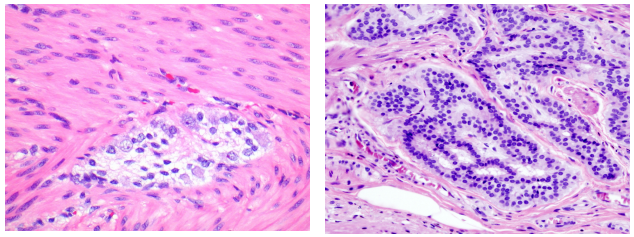
Frequency of reads supporting variant: 6/10 = 60%

Heterozygous

RNA-Seq

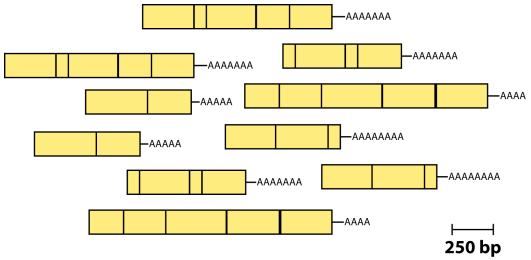
RNA-Seq refers to experimental procedures that generate sequence reads derived from the entire RNA molecule. It can be used to build a complete map of the transcriptome across all cell types, perturbations and states.

Samples of interest

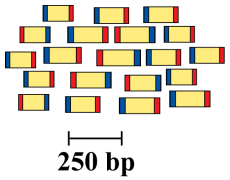


Condition 1 (normal colon) Condition 2 (colon tumor)

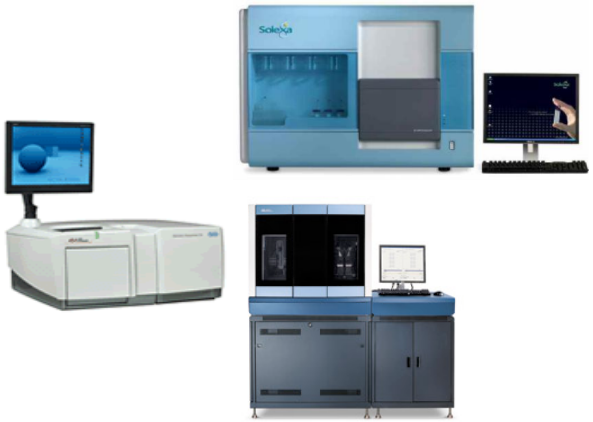
Isolate RNAs

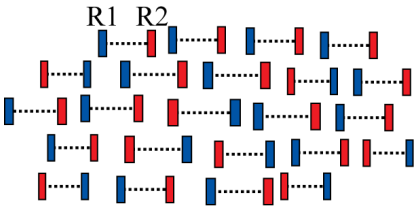


Generate cDNA, fragment, size select, add linkers



Sequence

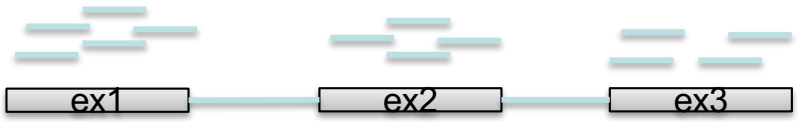




Map to genome, transcriptome, and predicted exon junctions

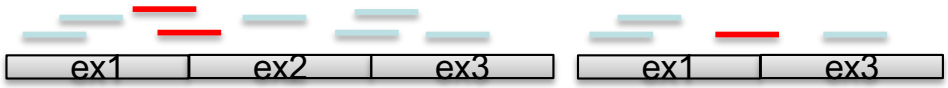
RNA-Seq: Applications

- ✓ Gene/transcript expression

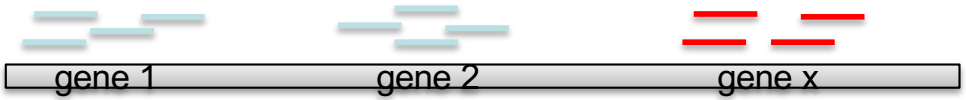


- ✓ Isoform reconstruction

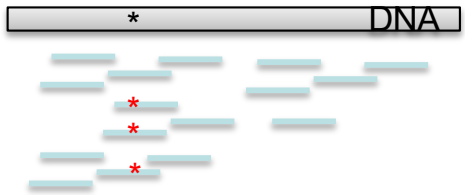
- ✓ Alternative splicing detection



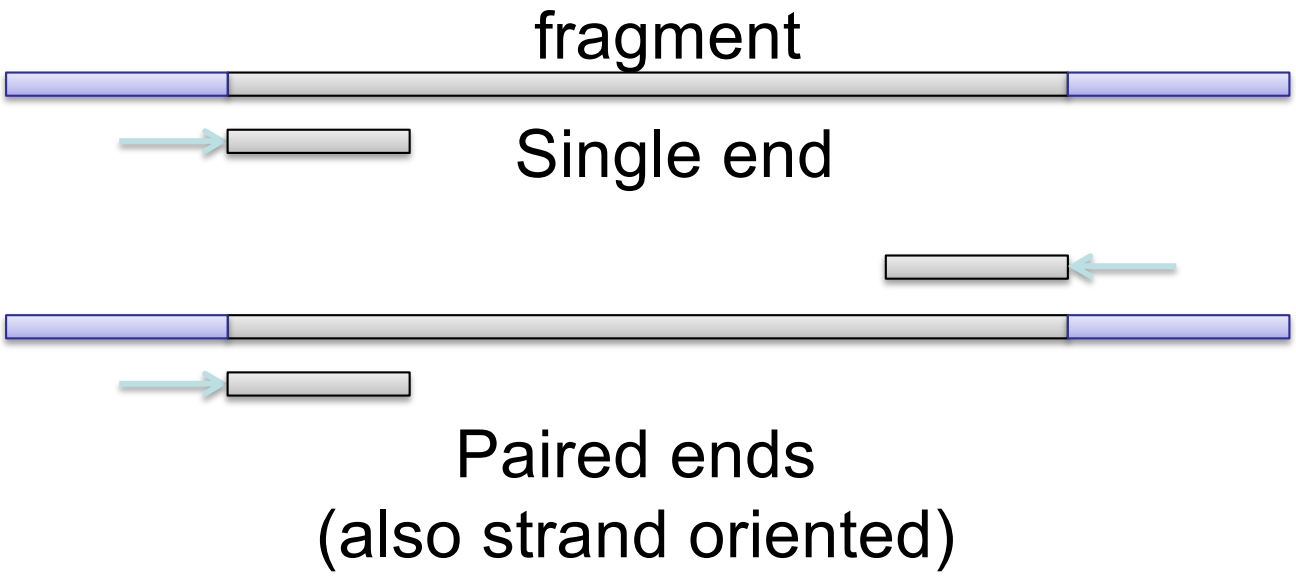
- ✓ Gene discovery



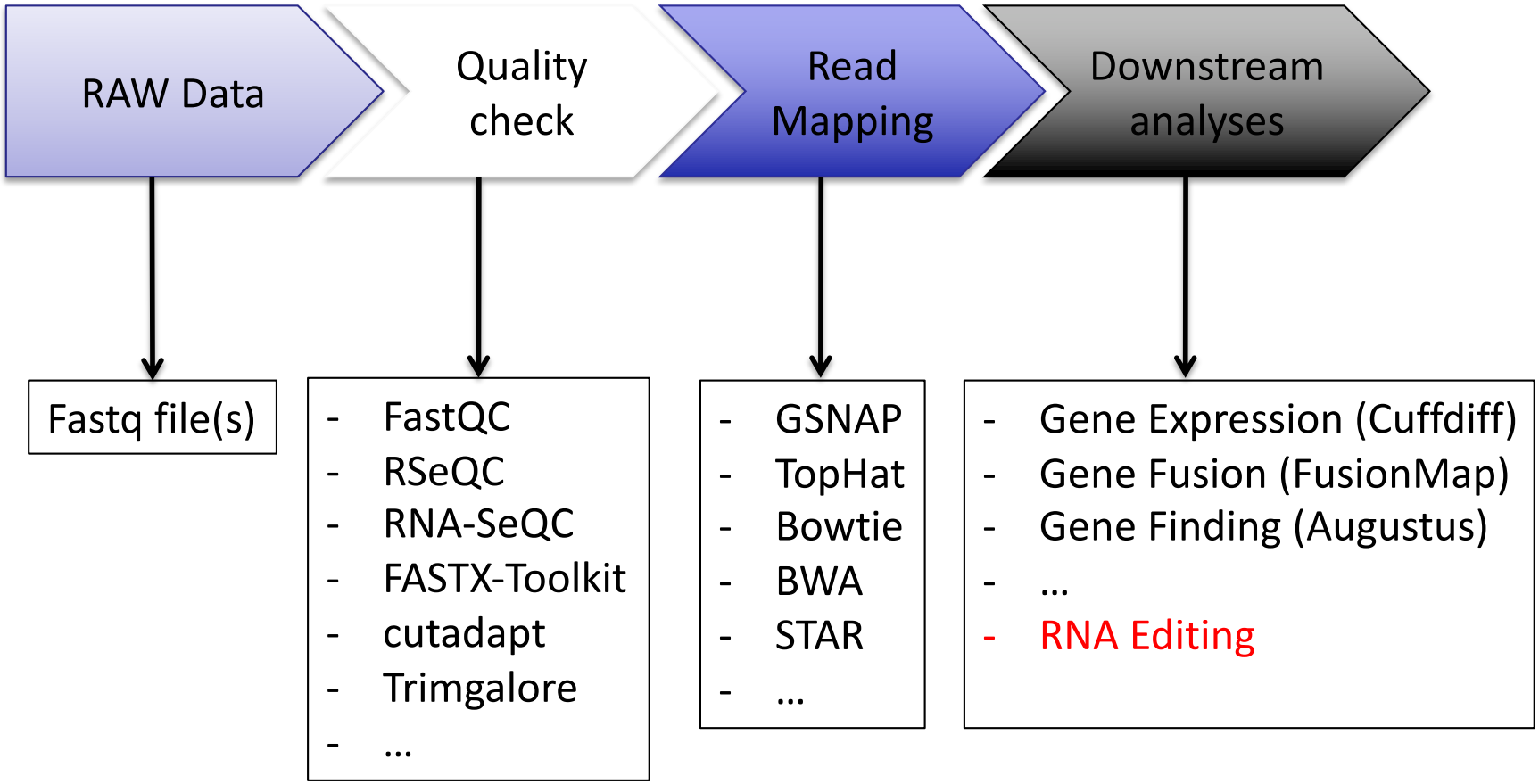
- ✓ RNA editing identification



RNA-Seq: read types



RNA-Seq analysis workflow

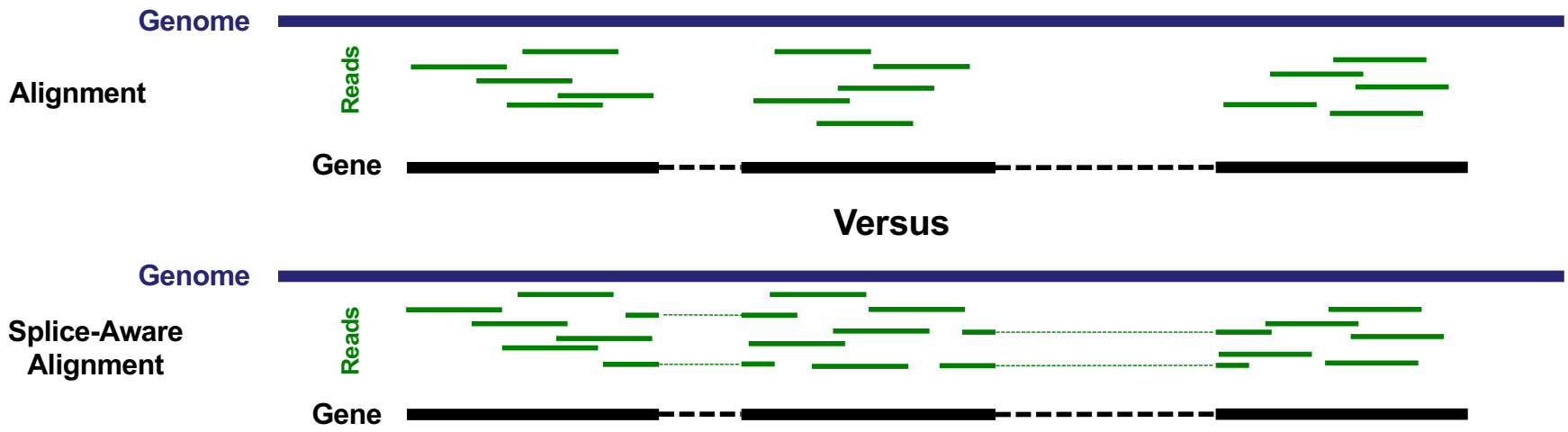


RNA-Seq: read mapping

We need to align the sequence data to our genome of interest

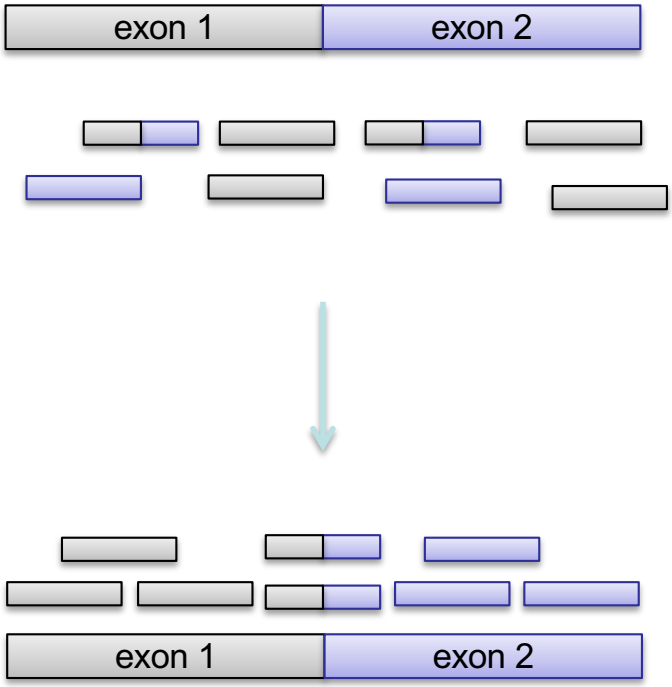
- ✧ In aligning RNA-Seq data to the genome always pick a slice-aware aligner:

TopHat2, MapSplice, SOAPSplice, Passion, SpliceMap, RUM, ABMapper, CRAC, GSNAP, HMMSplicer, Olego, BLAT



RNA-Seq: read mapping

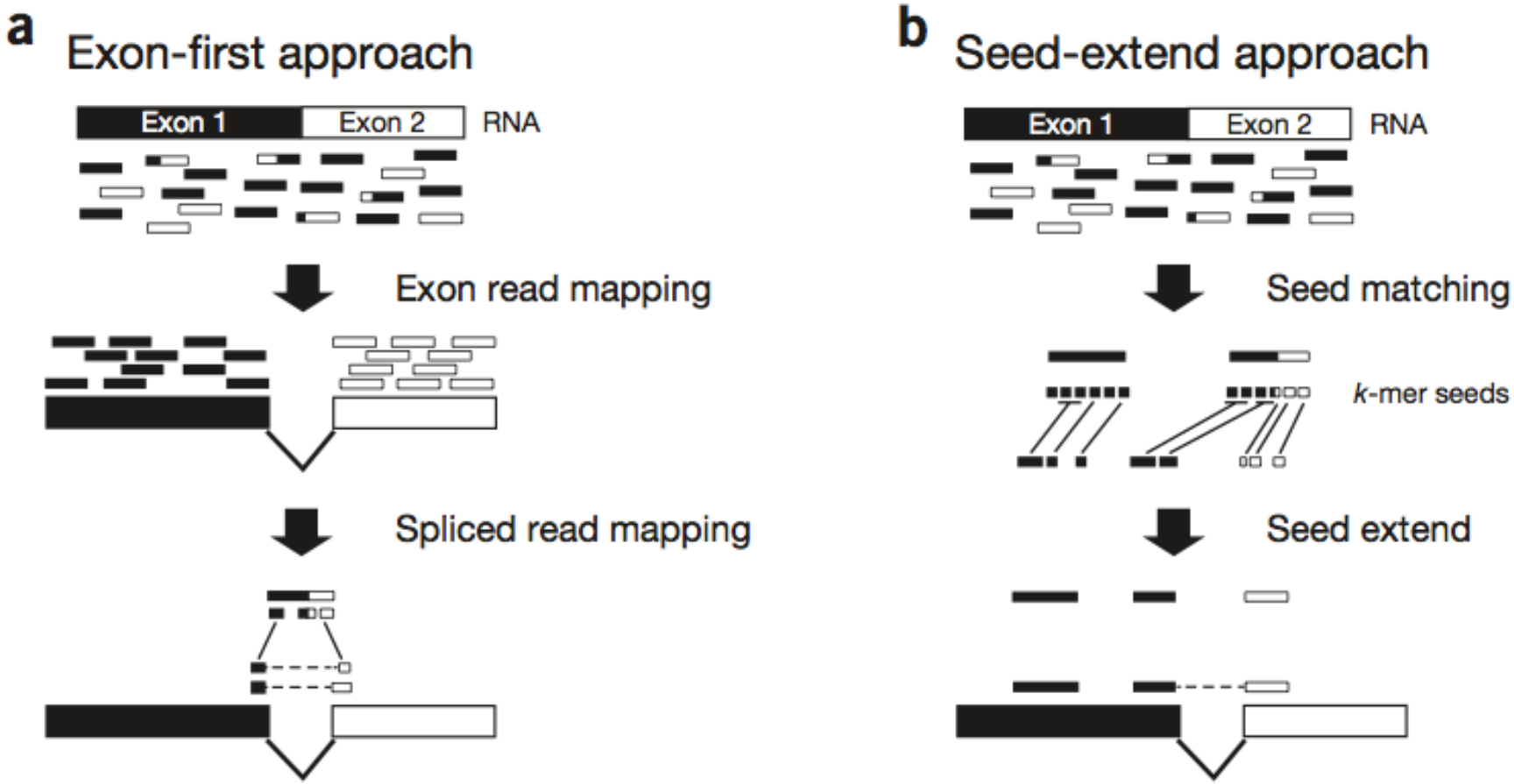
Against transcriptome



Tools as BWA, Bowtie, MAQ, SOAP, GSNAP … and others can be used.

RNA-Seq: read mapping

Against whole genome

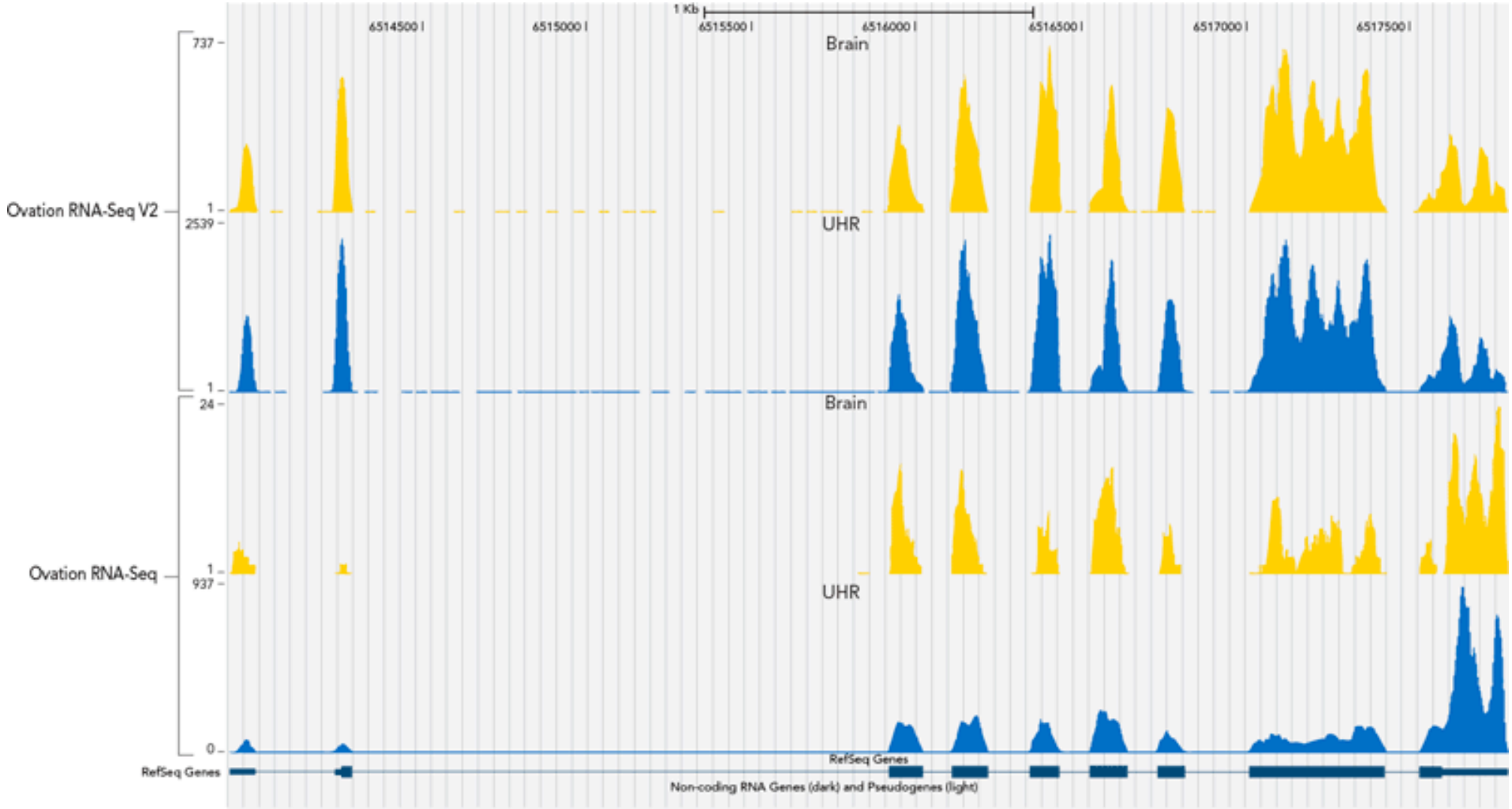


Tophat, MapSplice …

GSNAP

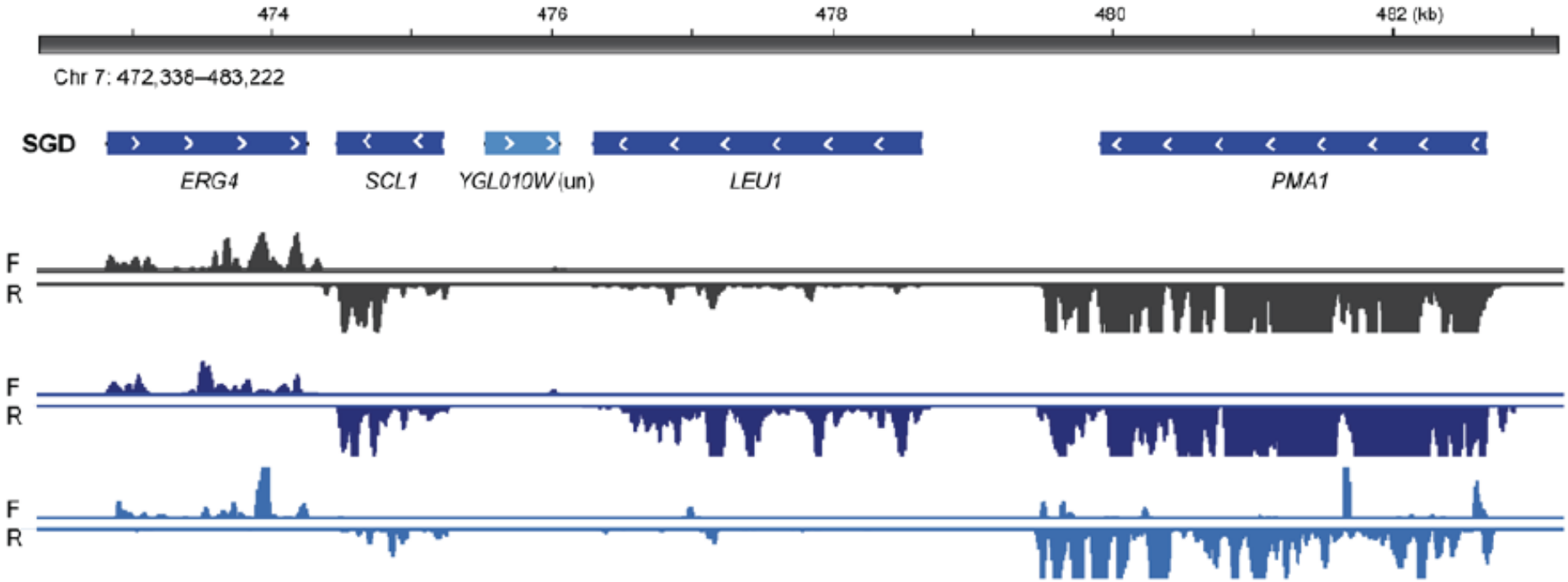
Garber at al. (2011) Nature Methods

RNA-Seq: visualization



RNA-Seq: visualization

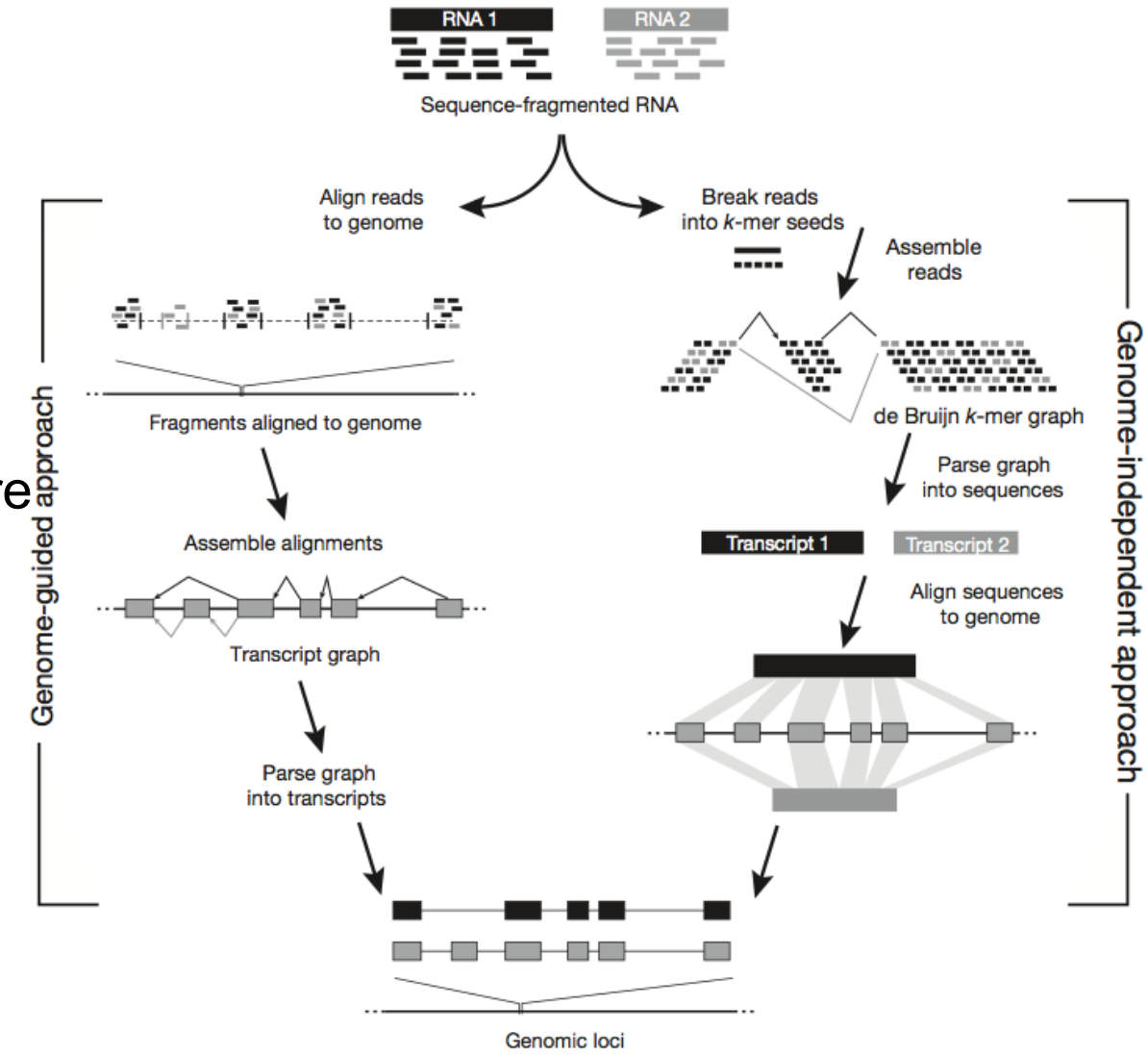
strand oriented reads



RNA-Seq: transcriptome reconstruction

Transcript assembly

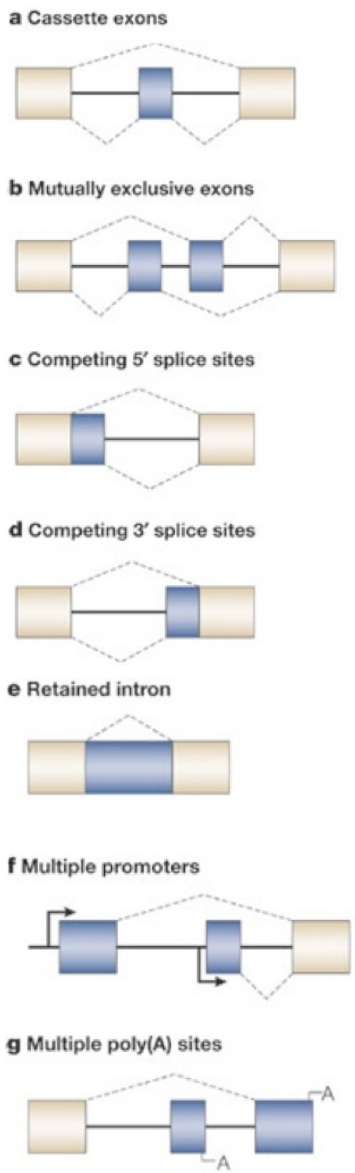
Cufflinks, Scripture



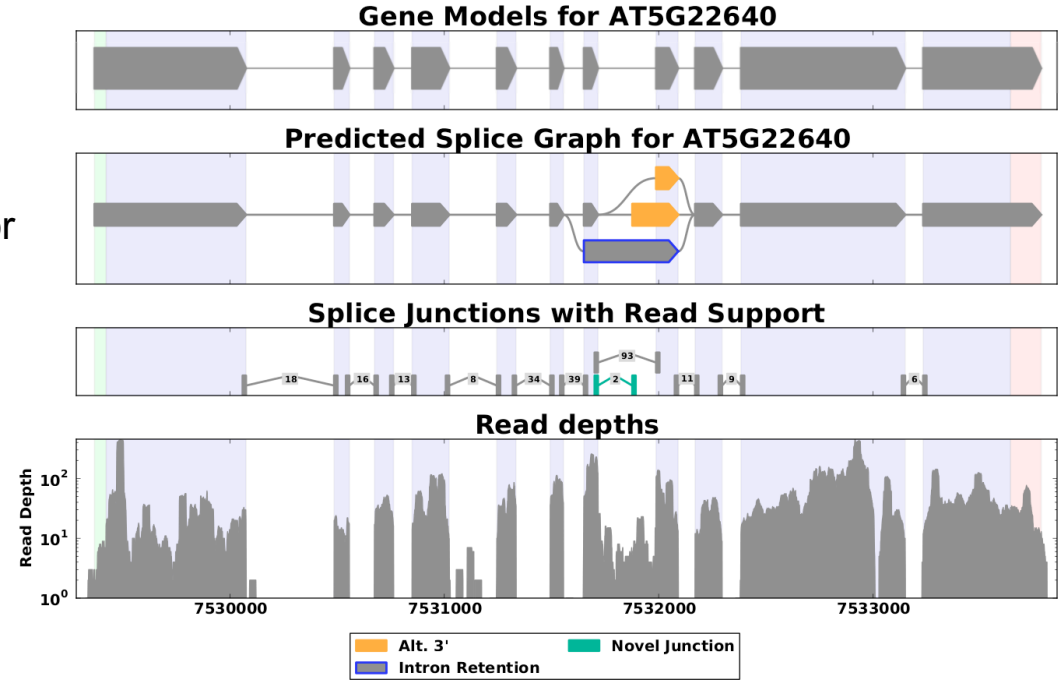
Abyss, Trinity

Garber at al. (2011) Nature Methods

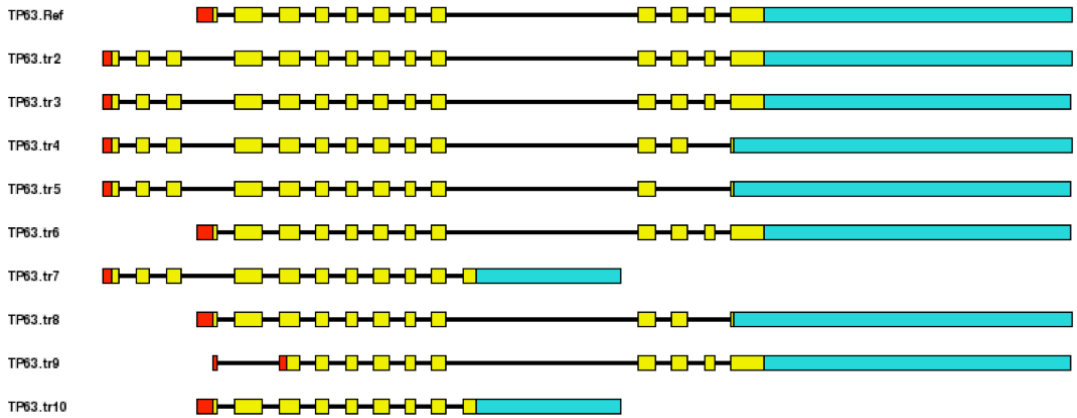
RNA-Seq: alternative splicing



Detection by known or novel splice junctions

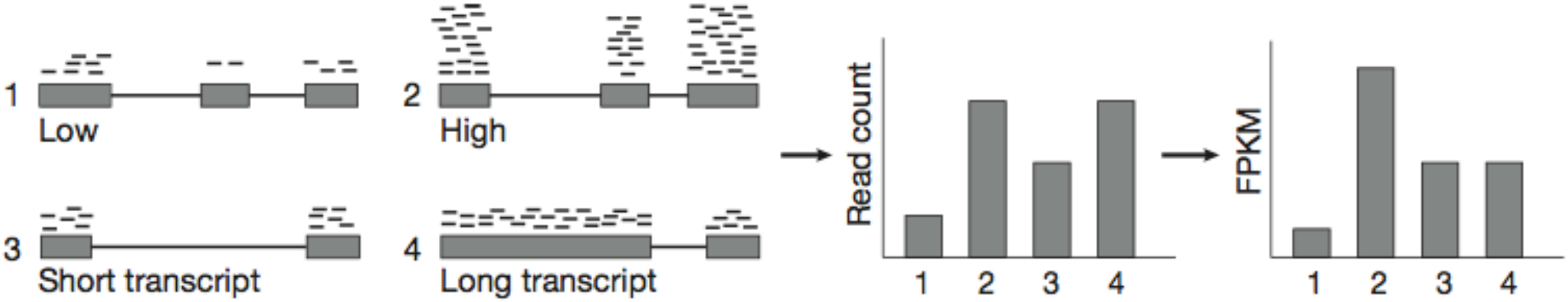


Detection by isoform comparison



Matlin et al. (2005) Nature Reviews Molecular Cell Biology

RNA-Seq: gene/transcript expression



Garber at al. (2011) Nature Methods

When using RNA-seq to estimate gene expression, read counts need to be properly normalized to extract meaningful expression estimates

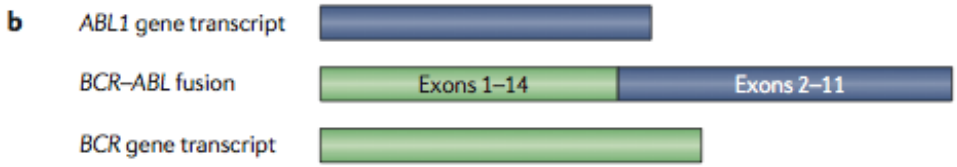
- RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample;
- The variability in the number of reads produced for each run causes fluctuations in the number of fragments mapped across samples;

To account for these issues, the reads per kilobase of transcript per million mapped reads (RPKM) metric normalizes a transcript's read count by both its length and the total number of mapped reads in the sample.

$$RPKM(FPKM) = 10^9 \times \frac{C}{NL}$$

C= the number of reads mapped onto the gene's exons
N= total number of reads in the experiment
L= the sum of the exons in base pairs.

RNA-Seq: gene fusions



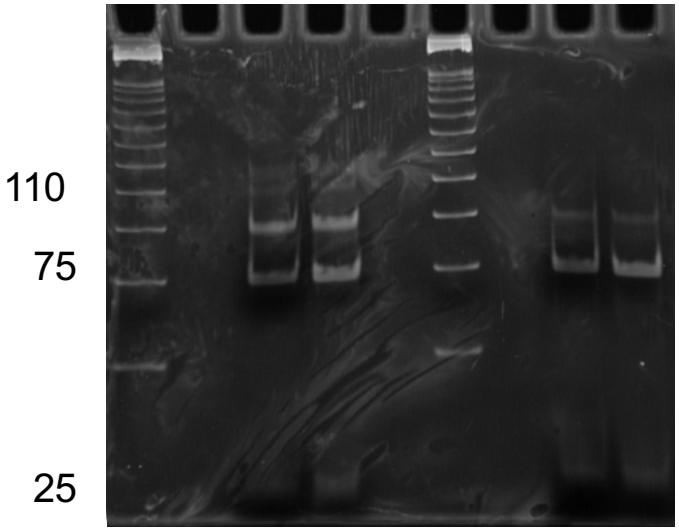
c

```
GTCATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAATCTGTACTGCACCCTGGAGGTGGATTCCTTTGGGTATTTT      BCR

AGGCATGGGGGTCCACACTGCAATGTTTTTGTGGAACATGAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCAG      ABL1

GTCATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCAG      BCR–ABL fusion transcript


GTCAT-GTCCACTCAGC-ACTGGATT-AAGCAGAGTTCAAAAGC
 TCAT-GTCCACTCAGCCACTGGATTTAA-CAGAGTTCAAAAGC
       TCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCC
        CCACTCAGCTACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGC
             CAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGC
             CAGCCA-TGGATTTAAGC-GAGTTCAAAAGCCCTTCAG
              AGCCACTGGATTTAAGCAGAGTTCAAAAG
               GCCACTGGATTTAAGCAGAGTTCAAAAGCCCT
                 CACTGGATTTAAGCAGAGTTCAAAAGCCCTT
                   CTGGATTTAAGCAGAGTTCAAAAGC--TTCAGCGGC-AGTAG
                    TG-ATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGC
                       ATTTAAGCAGAGTTCAAAAGCCCTTCAGCG-CCAGTAGCA
                         TTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAG
                            AGCAGAGT-CAAAAGCCCTTC-GCGGCCAGTAGCATCTGACTTTGA-C
                            AT-A-AGTTCAAAAGCC-TTCAGCGGCCA-TAGCATCTG
                              CAGAGTTCAAAAGCCCTTCAGCGGCCAG
                              CAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTG
                                 AGTTCAAAAGCCCTTCAGCG-CCAGT-GCATCT
                                  GTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACT
                                    TCAAA-GCCCT-C-GCGGCCAGTAGCATCTGAC
                                    TCAAA-GCCCTTCAGCGGCCAGTAGCATCTGACTTTGAG
                                     CAAA-GCCCTTCAG-GGCCAGTAGCATCTGACTTTGAGCCTCAG
                                      AAAAGCCCTTCAG-GGCCAGTAGCATCTGACTTTGAG
```

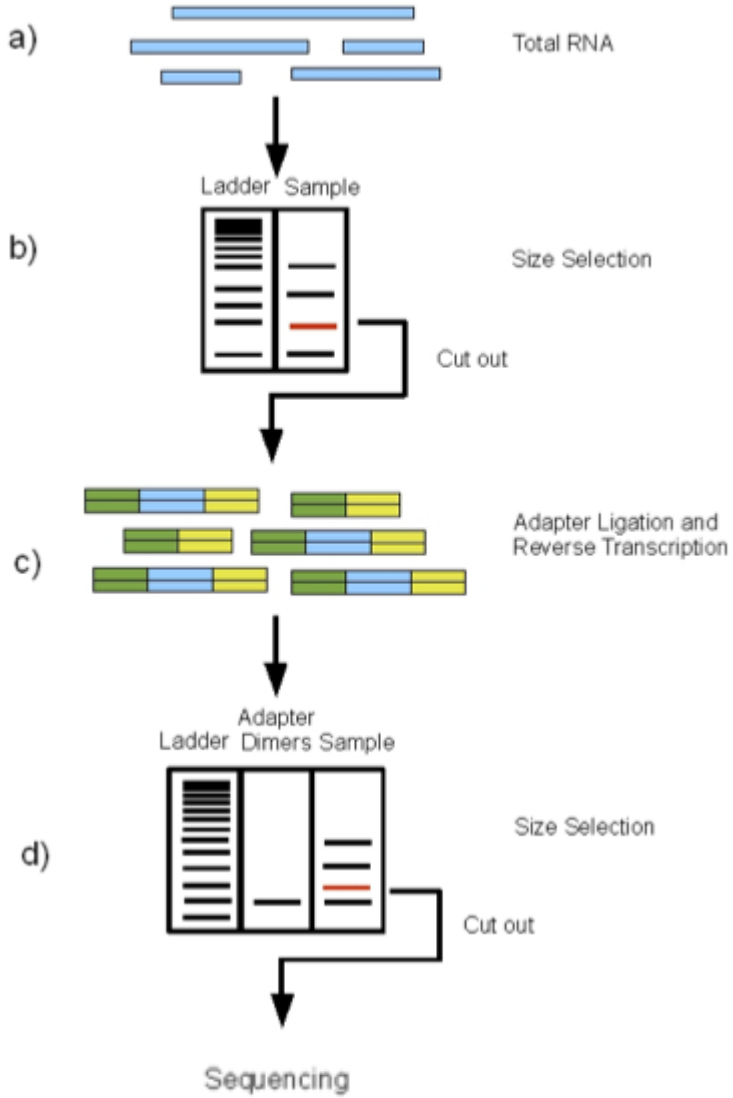
RNA-Seq: small RNAs

sequencing

smallRNA separation: PAGE



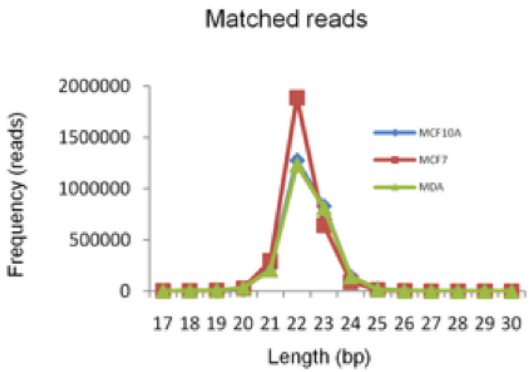
→small RNA < 35bp



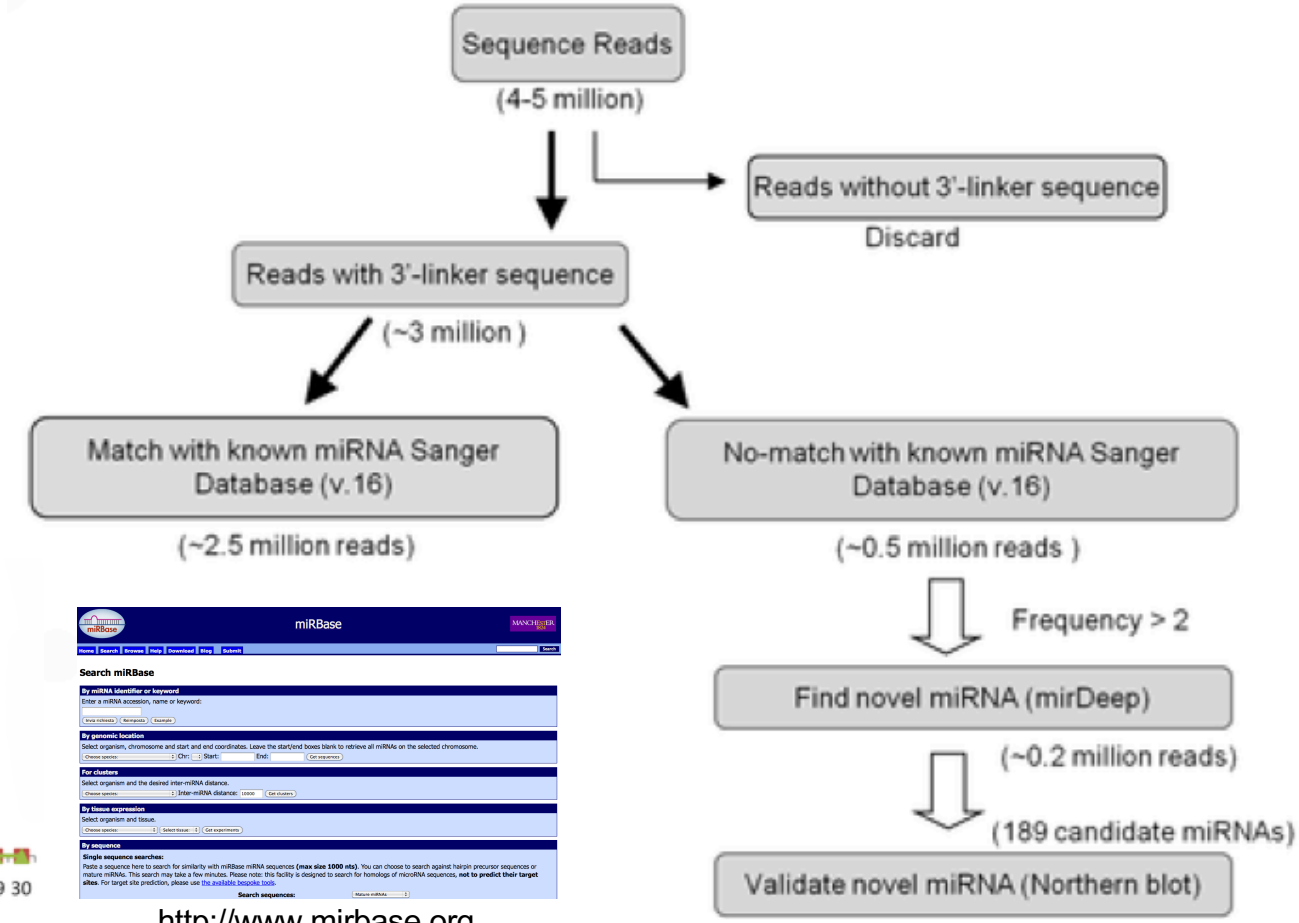
Motameny et al. (2010) Genes

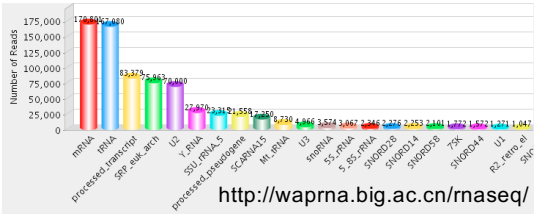
RNA-Seq: small RNAs

analysis



http://www.mirbase.org



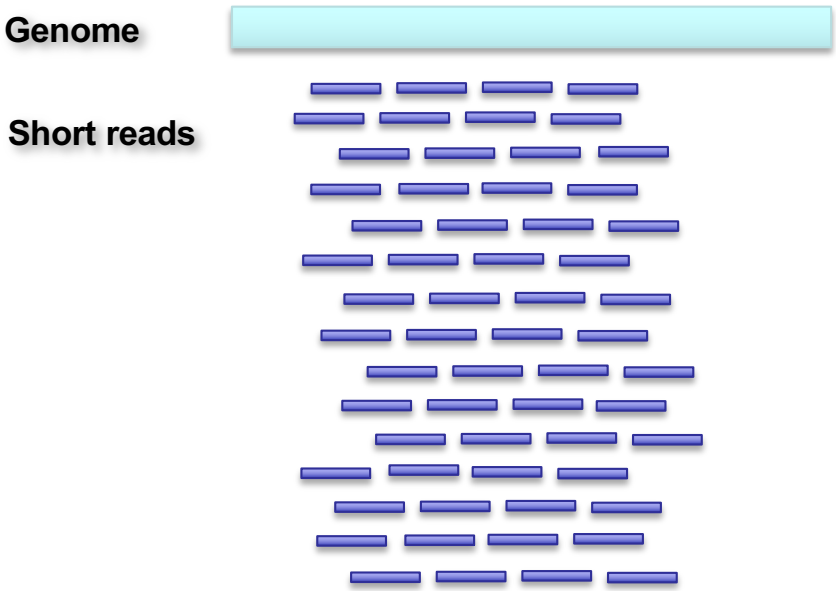


http://waprna.big.ac.cn/rnaseq/

Ryu et al. (2011) PlosOne

RNA-Seq: RNA editing

Massive RNA sequencing can facilitate the study of entire transcriptomes as well as post-transcriptional events occurring herein as alternative splicing and RNA editing.



Using NGS, each genomic position can be supported by a large number of sequences and this can greatly improve the detection of RNA editing substitutions.

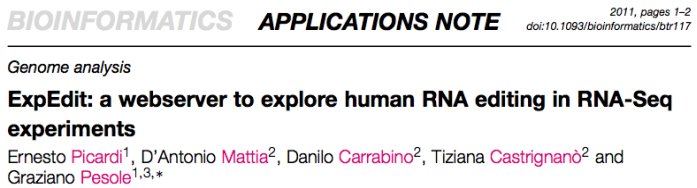
RNA-Seq: RNA editing detection

We can employ NGS data (RNA-Seq, genome resequencing and exome sequencing) to study RNA editing at different levels:

✓ genome/exome Vs RNA-Seq to identify new events (REDItools);

```
Exome    r1   GGGTGCCTTTATGCAGCAAGGATGCGATATT
         r2   GGGTGTCTTTATGCAGCAAGGATGCGATACTTCGC
         r3   GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
         r4   GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
         r5   GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
                ..............A.....................
gDNA     TGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCC
                ..............G.....................
RNA-Seq  r1   GGGTGCCTTTATGCGGCAAGGATGCGATATT
         r2   GGGTGTCTTTATGCAGCAAGGATGCGATACTTCGC
         r3   GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
         r4   GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
         r5   GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
```

✓ RNA-Seq to explore the presence of known A-to-I conversions;


BIOINFORMATICS APPLICATIONS NOTE

2011, pages 1–2
doi:10.1093/bioinformatics/btr117

Genome analysis

ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments

Ernesto Picardi[1], D'Antonio Mattia[2], Danilo Carrabino[2], Tiziana Castrignanò[2] and Graziano Pesole[1,3,*]

16 of 865 rows match filter(s)

Location	Position	Reference base	Strand	Gene	Region	Nucleotide				Coverage	Editing extent	Source
						As	Cs	Gs	Ts			
All	= AND =	All	All		CDS	= AND =	= AND =	= AND =	= AND =	≥ 10 AND =	> 0 AND =	D
chr1	6081149	A	+	KCNAB2	CDS	0	0	31	0	31	1.000	D
chr4	158477325	A	+	GRIA2	CDS	0	0	11	0	11	1.000	D
chr5	150619602	A	+	GM2A	CDS	0	0	16	0	16	1.000	D
chr5	150619632	A	+	GM2A	CDS	0	0	12	0	12	1.000	D
chr11	62214851	A	-	BSCL2	CDS	0	0	20	0	20	1.000	D
chr16	57102927	A	+	NDRG4	CDS	0	0	93	0	93	1.000	D
chr11	77468301	A	-	NDUFC2	CDS	9	0	17	0	26	0.654	D

http://www.caspur.it/ExpEdit/

✓ RNA-Seq to detect *de novo* new editing candidates;


OPEN ACCESS Freely available online PLOS | ONE

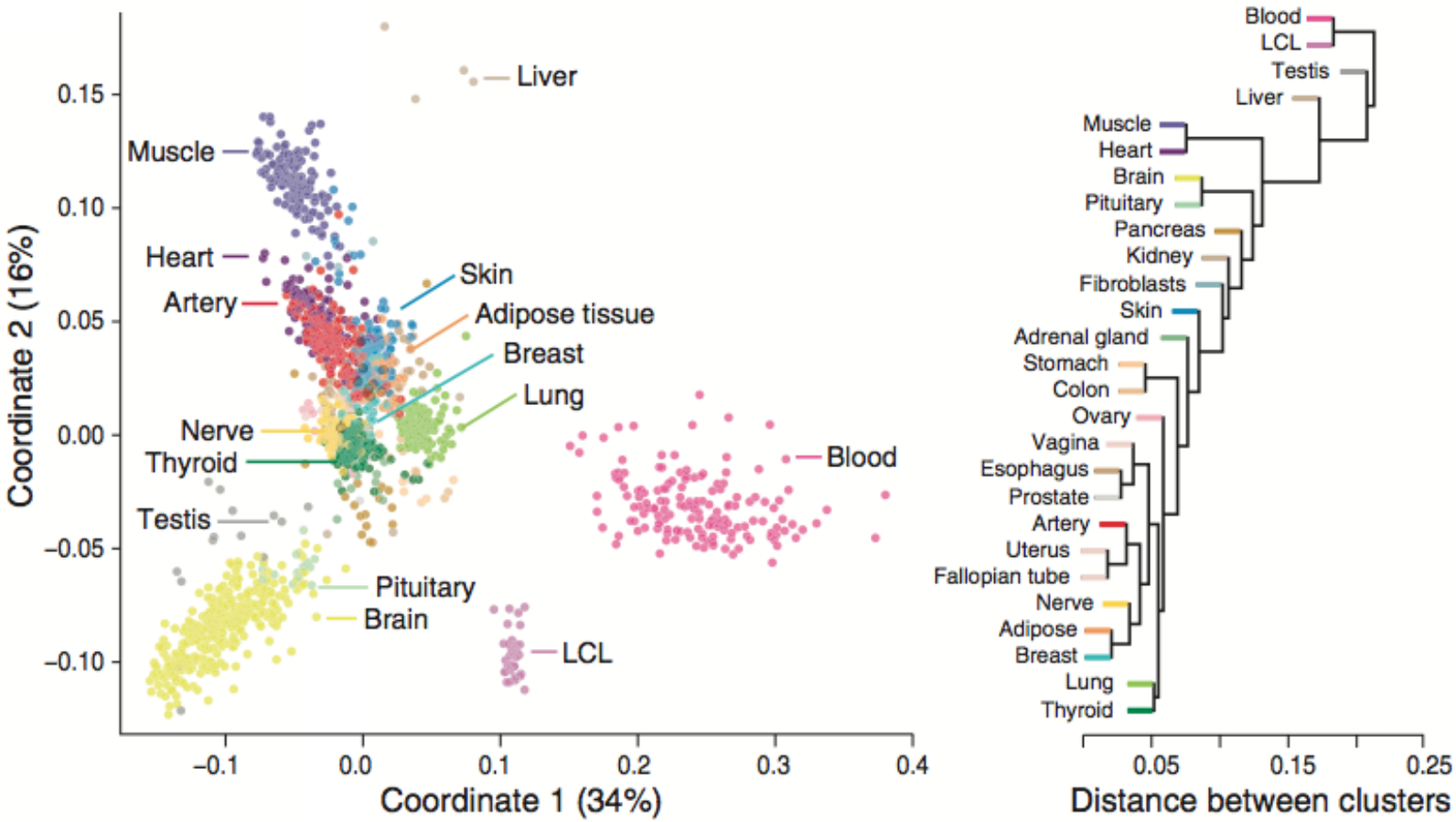
A Novel Computational Strategy to Identify A-to-I RNA Editing Sites by RNA-Seq Data: *De Novo* Detection in Human Spinal Cord Tissue

Ernesto Picardi[1,2], Angela Gallo[3], Federica Galeano[3], Sara Tomaselli[3], Graziano Pesole[1,2*]

1 Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università di Bari, Bari, Italy, 2 Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy, 3 RNA Editing Laboratory, Oncohaematology Department, Ospedale Pediatrico "Bambino Gesù", IRCCS, Rome, Italy

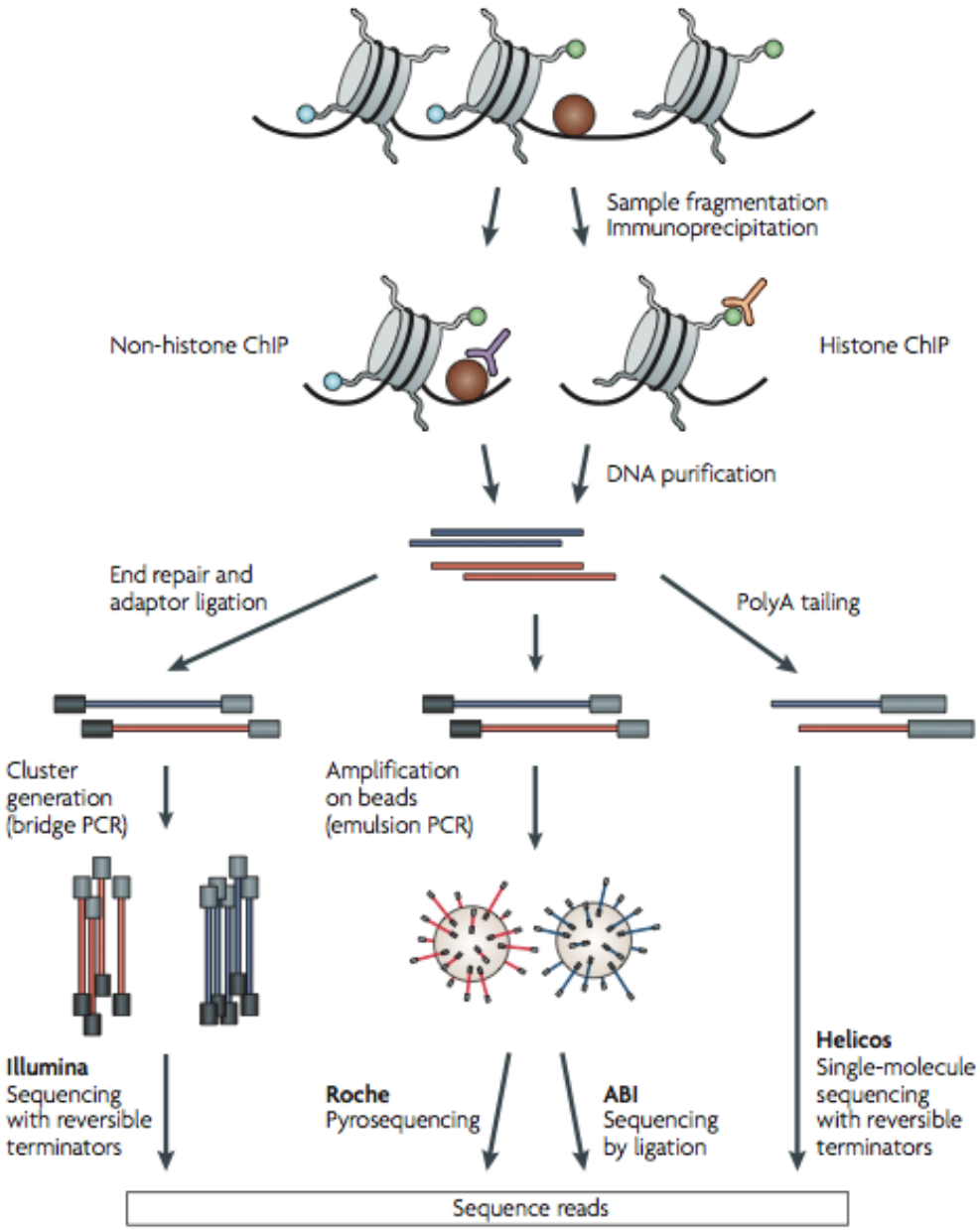
```
gDNA     AGCTGGCCAGATACATTAAGACCAGTGCTCACTATGAAG
                ................G.....................
RNA-Seq  r1   GCTGGCCAGATACATTGAGACCAGTGCTCAC
         r2   GCTGGCCAGATACATTAAGACCAGTGCTCAC
         r3    CTGGCCAGATACATTGAGACCAGTGCTCACTATGAAG
         r4    CTGGCCAGATACATTGAGACCAGTGCTCACTATG
         r5    CTGGCCAGATACATTAAGACCAGTGCTCACTATGAAG
         r6    CTGGCCAGATACATTAAGACCAGTGCTCACTATGAAG
         r7    CTGGCCAGATACATTGGGACCAGTGCTCACTATGAAG
         r8    CTGGCCAGATACATTGAGACCAGTGCTCACT
         r9    CTGGCCAGATACATTGAGACCAGTGCTCACTATGAAG
```

Human Transcriptome profile

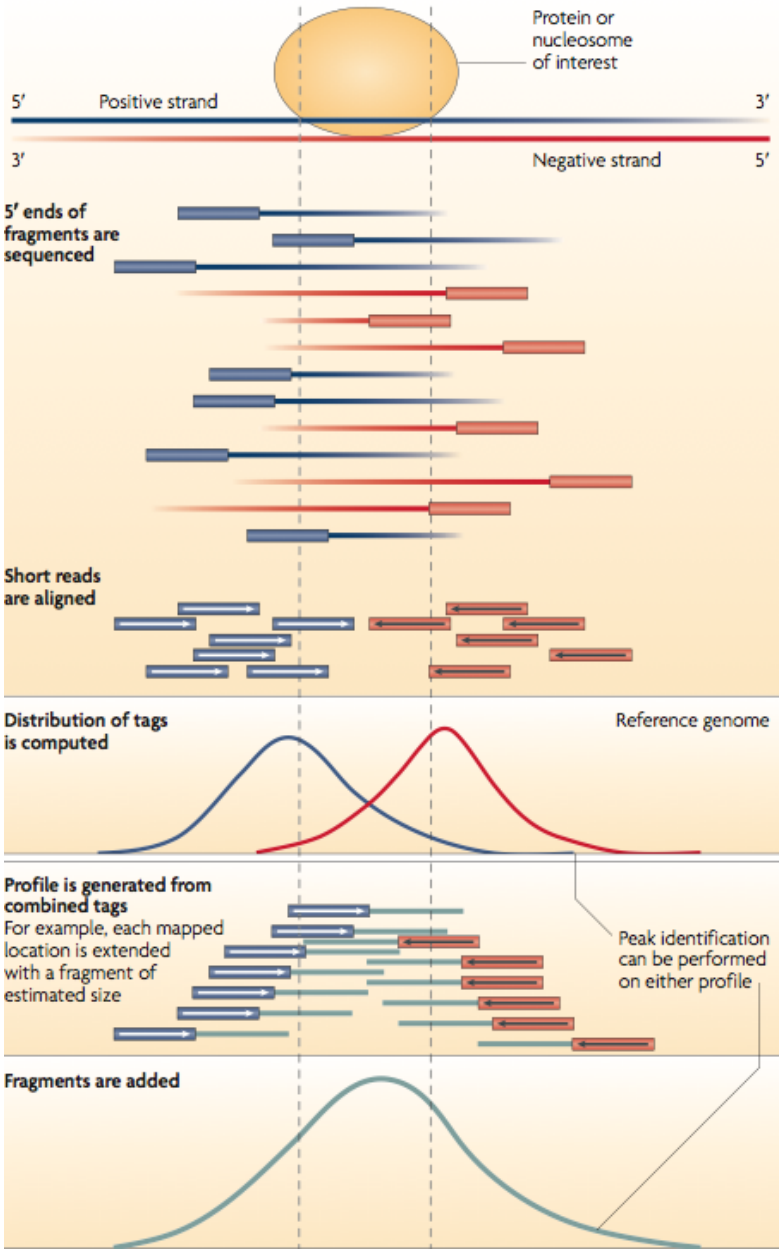


Melè et al. 2015 Science

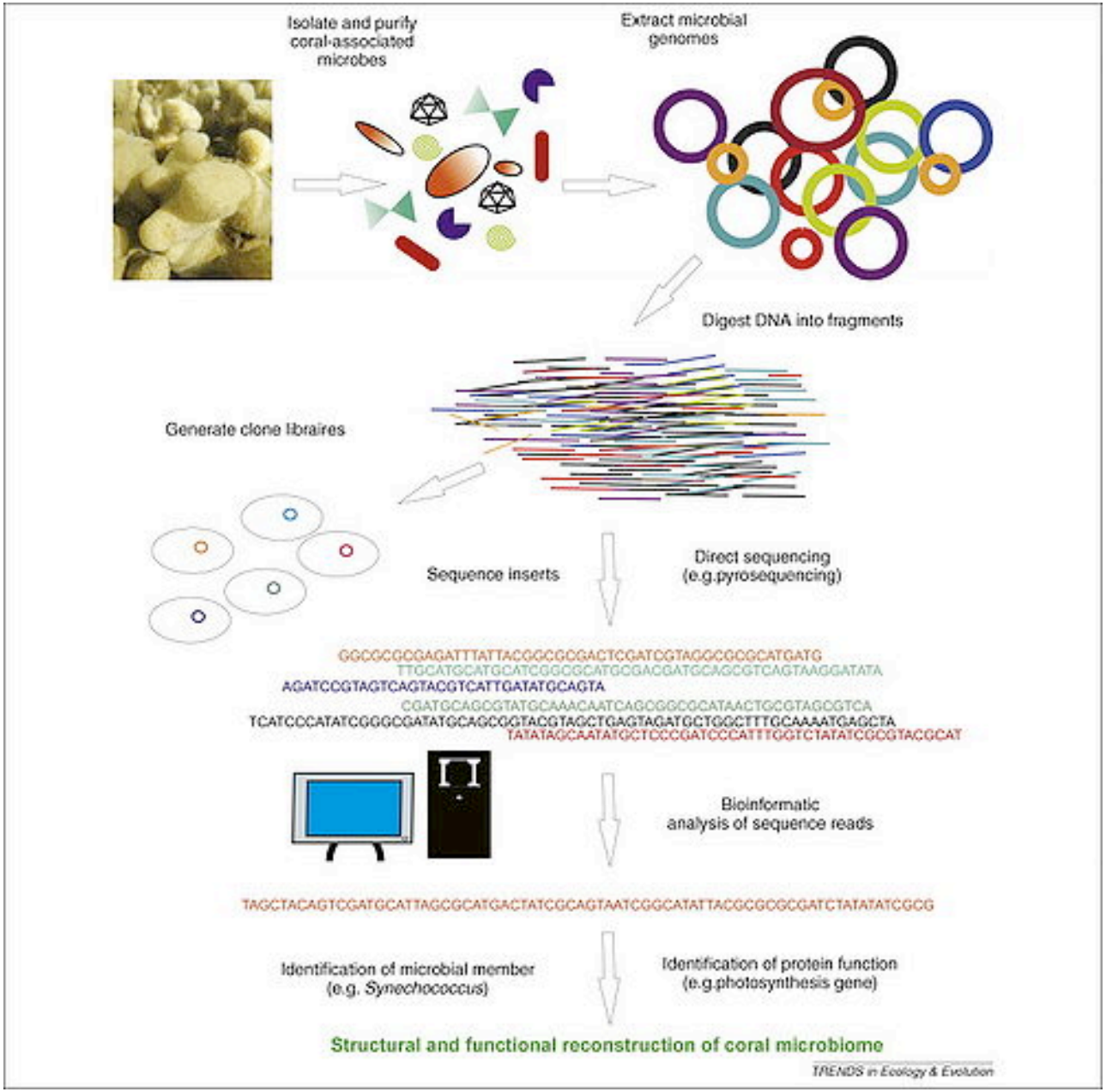
ChIP-Seq


Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and
adaptor ligation
PolyA tailing
Cluster
generation
(bridge PCR)
Amplification
on beads
(emulsion PCR)
Illumina
Sequencing
with reversible
terminators
Roche
Pyrosequencing
ABI
Sequencing
by ligation
Helicos
Single-molecule
sequencing
with reversible
terminators
Sequence reads

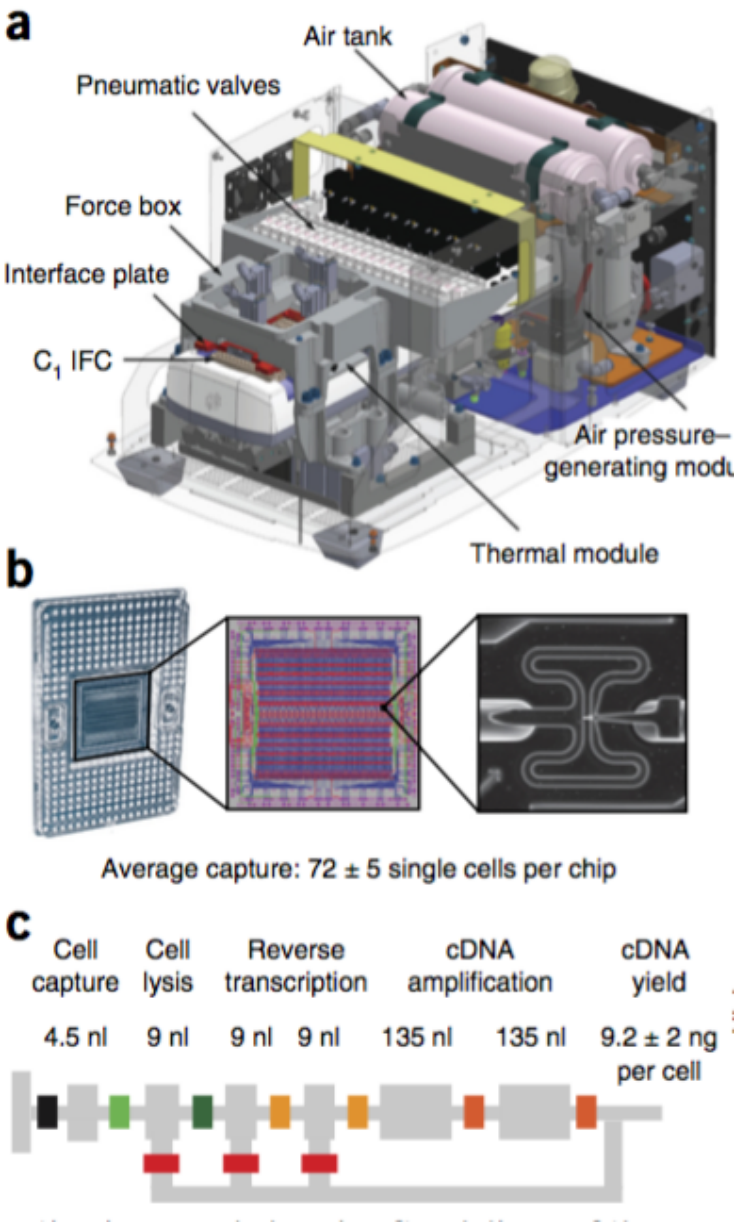
ChIP-Seq


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Metagenomics



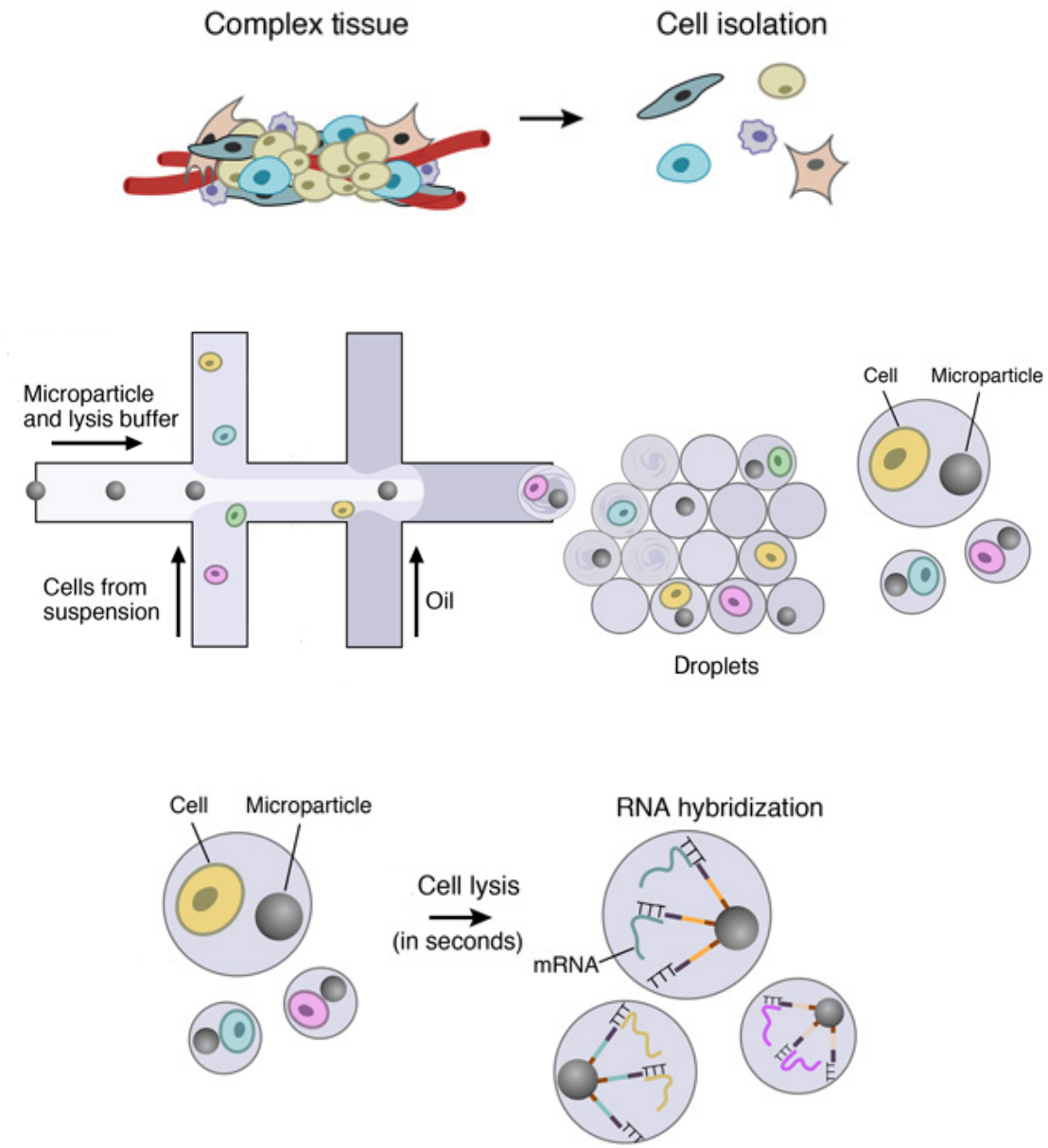
Single-cell sequencing – C1



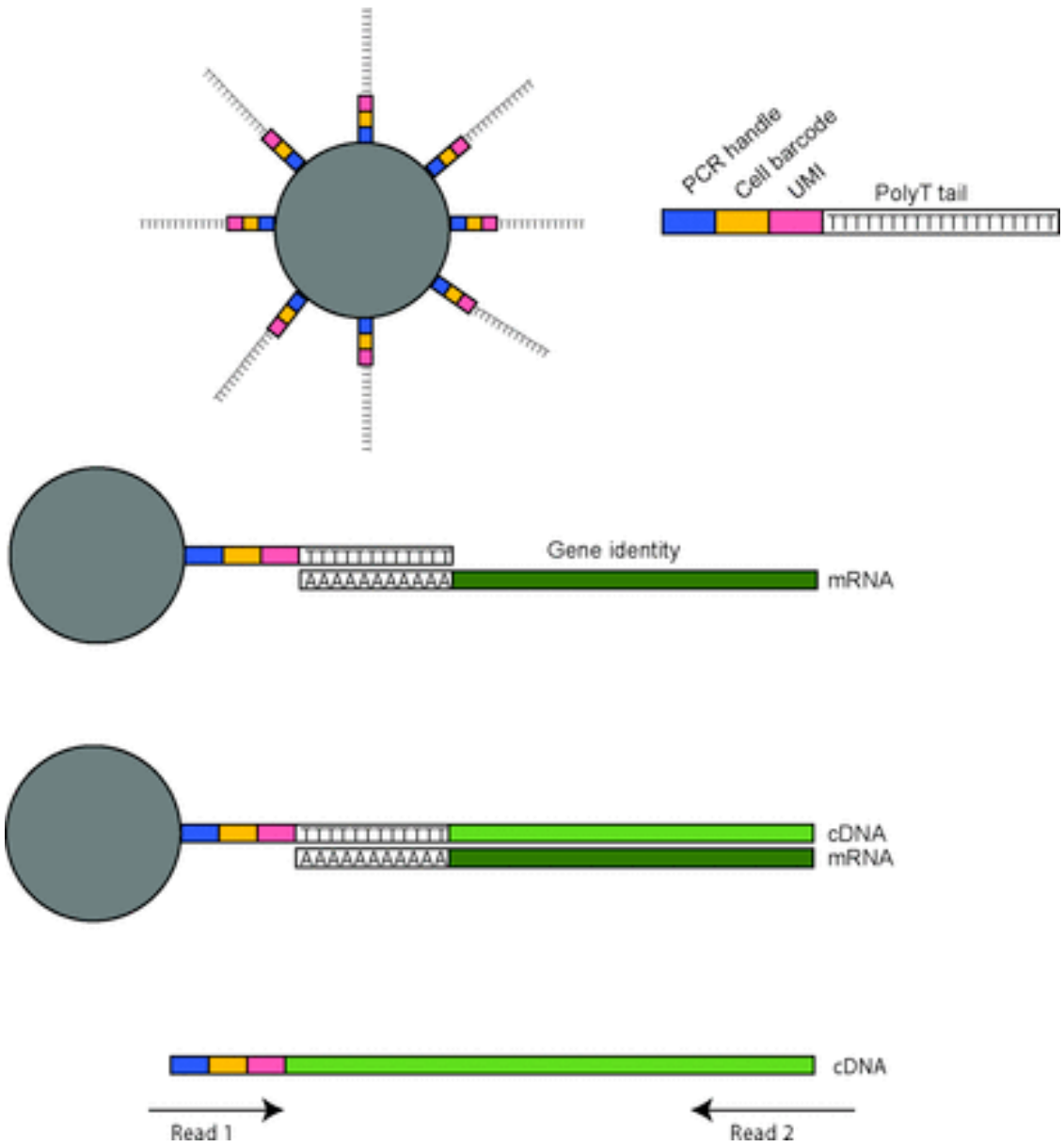
Pollen et al. 2014 Nature Biotech.

Capturing single cells and quantifying mRNA levels using the C1 Single-Cell Auto Prep System. (a) Key functional components of the C1 System are labeled, including the pneumatic components necessary for control of the microfluidic integrated fluidic circuit (IFC) and the thermal components necessary for preparatory chemistry. (b) Left, complete IFC with carrier; reagents and cells are loaded into dedicated carrier wells, and reaction products are exported to other dedicated carrier wells.(c) Schematic for a C1 reaction line, with the reaction line colored light gray and the isolation valves shown in varied colors.

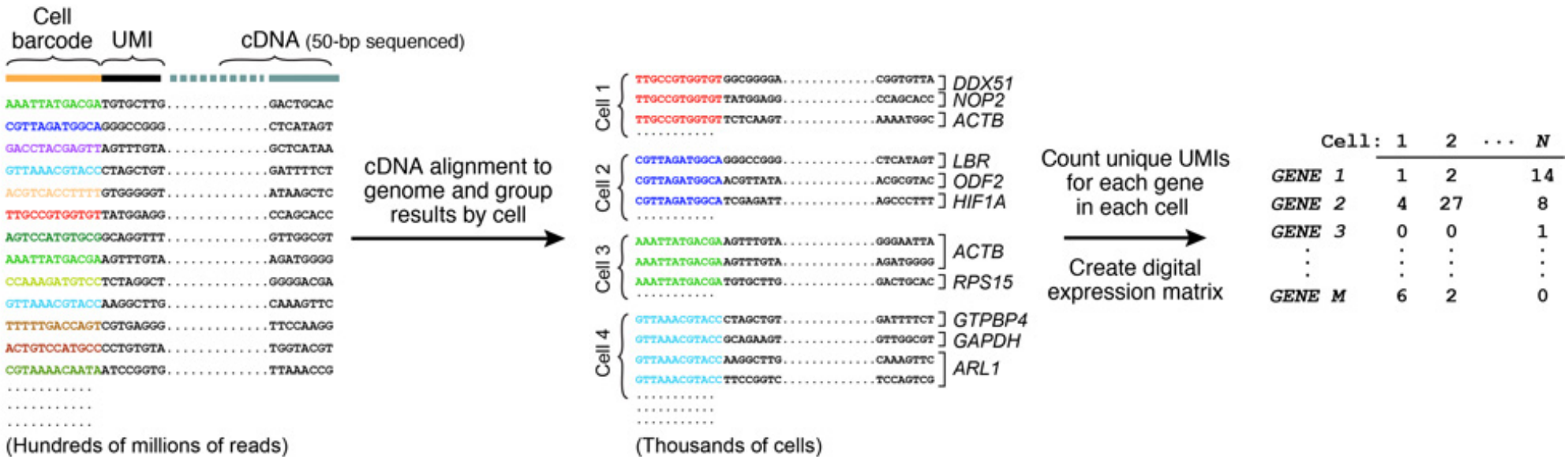
Droplet-Sequencing (Drop-Seq)



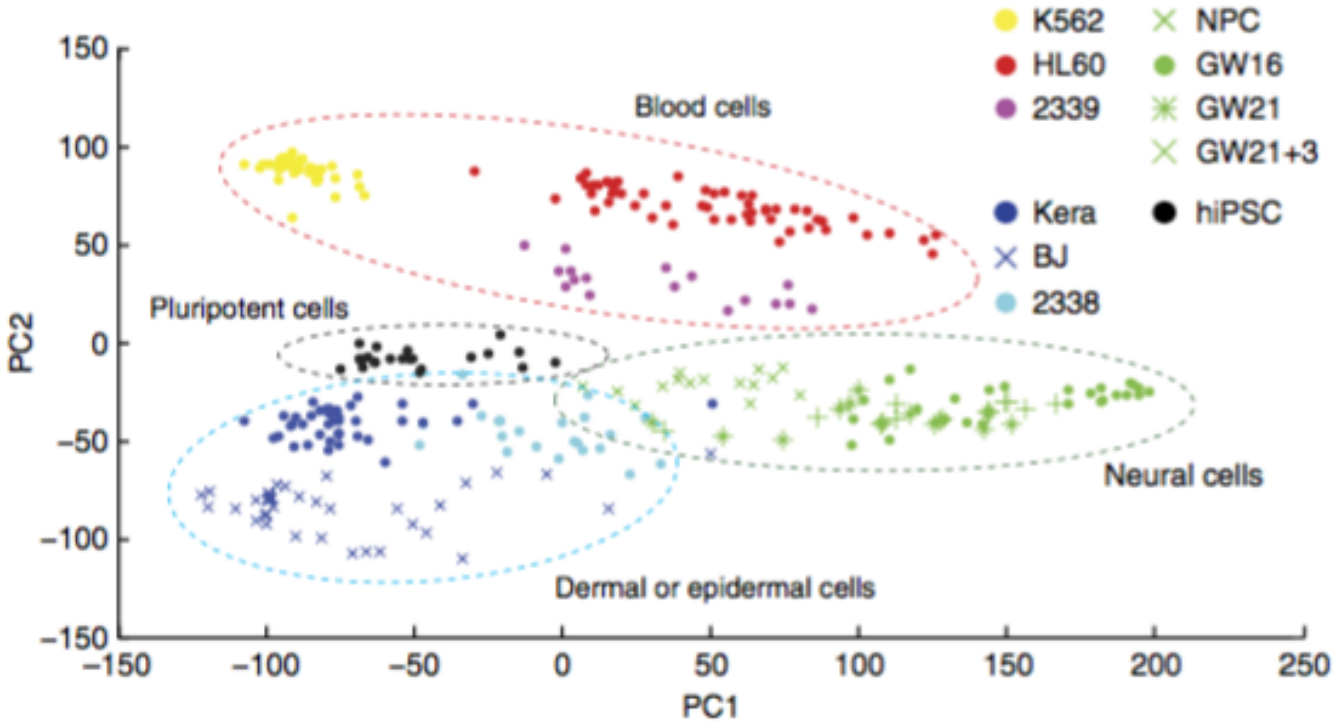
Droplet-Sequencing (Drop-Seq)



Droplet-Sequencing (Drop-Seq)



Single-cell transcriptome



Distinct groups of cells corresponding to pluripotent, blood, skin and neural cells can be identified by PCA. NPC, neural progenitor cell.

Pollen et al. 2014 Nature Biotech.

Table 3 Bioinformatics tools for short-read sequencing

Program	Categories	Author(s)	Reference	URL
Cross_match	Alignment	Phil Green, Brent Ewing and David Gordon		http://www.phrap.org/phredphrapconsed.html
ELAND	Alignment	Anthony J. Cox		http://www.illumina.com/
Exonerate	Alignment	Guy S. Slater and Ewan Birney	72	http://www.ebi.ac.uk/~guy/exonerate
MAQ	Alignment and variant detection	Heng Li	37	http://maq.sourceforge.net
Mosaik	Alignment	Michael Strömberg and Gabor Marth		http://bioinformatics.bc.edu/marthlab/Mosaik
RMAP	Alignment	Andrew Smith, Zhenyu Xuan and Michael Zhang	73	http://rulai.cshl.edu/rmap
SHRiMP	Alignment	Michael Brudno and Stephen Rumble		http://compbio.cs.toronto.edu/shrimp
SOAP	Alignment	Ruiqiang Li *et al.*	35	http://soap.genomics.org.cn
SSAHA2	Alignment	Zemin Ning *et al.*	36	http://www.sanger.ac.uk/Software/analysis/SSAHA2
SXOligoSearch	Alignment	Synamatix		http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
ALLPATHS	Assembly	Jonathan Butler *et al.*	38	
Edena	Assembly	David Hernandez *et al.*	74	http://www.genomic.ch/edena
Euler-SR	Assembly	Mark Chaisson and Pavel Pevzner	75	
SHARCGS	Assembly	Juliane Dohm *et al.*	76	http://sharcgs.molgen.mpg.de
SHRAP	Assembly	Andreas Sundquist *et al.*	39	
SSAKE	Assembly	René Warren *et al.*	40	http://www.bcgsc.ca/platform/bioinfo/software/ssake
VCAKE	Assembly	William Jeck	77	http://sourceforge.net/projects/vcake
Velvet	Assembly	Daniel Zerbino and Ewan Birney	41	http://www.ebi.ac.uk/%7Ezerbino/velvet
PyroBayes	Base caller	Aaron Quinlan *et al.*	34	http://bioinformatics.bc.edu/marthlab/PyroBayes
PbShort	Variant detection	Gabor Marth		http://bioinformatics.bc.edu/marthlab/PbShort
ssahaSNP	Variant detection	Zemin Ning *et al.*		http://www.sanger.ac.uk/Software/analysis/ssahaSNP

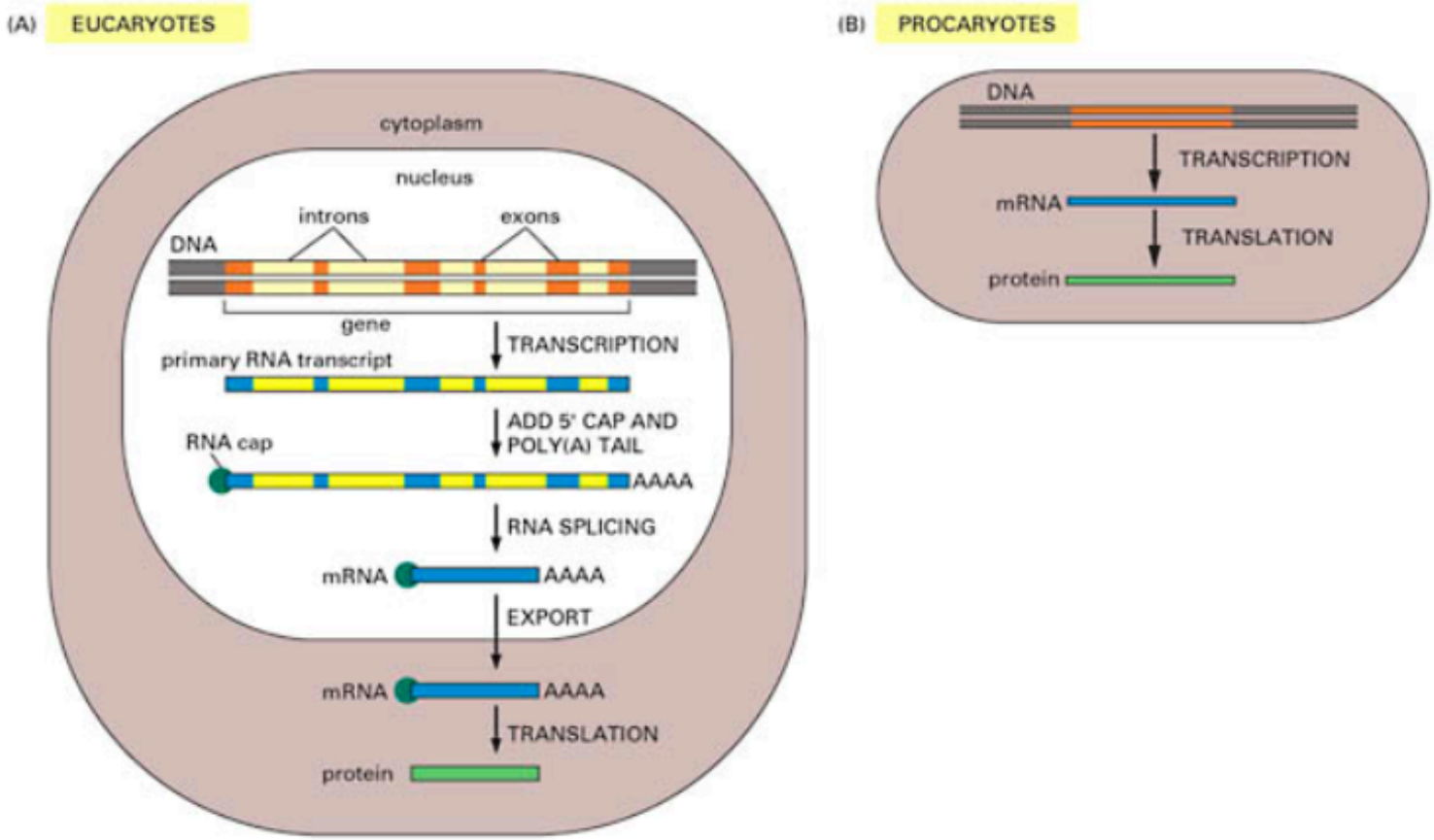
Incomplete list compiled from sources, including http://seqanswers.com/forums/showthread.php?t=43 and http://www.sanger.ac.uk/Users/lh3/seq-nt.html.

What is Computational Gene Finding?

Given an uncharacterized DNA sequence, find out:

- Which region codes for a protein?
- Which DNA strand is used to encode the gene?
- Which reading frame is used in that strand?
- Where does the gene starts and ends?
- Where are the exon-intron boundaries (in eukaryotes)?
- (optionally) Where are the regulatory sequences for that gene?

Gene Structure



Prokaryotic Vs. Eukaryotic Gene Finding

Prokaryotes:

- small genomes 0.5 – $10 \cdot 10^6$ bp
- high coding density (>90%)
- no introns

⬇

– Gene identification relatively easy, with success rate ~ 99%

Problems:

- overlapping ORFs
- short genes
- finding TSS and promoters

Eukaryotes:

- large genomes 10^7 – 10^{10} bp
- low coding density (<50%)
- intron/exon structure

⬇

– Gene identification a complex problem, gene level accuracy ~50%

Problems:

- many

Gene Finding: Different Approaches

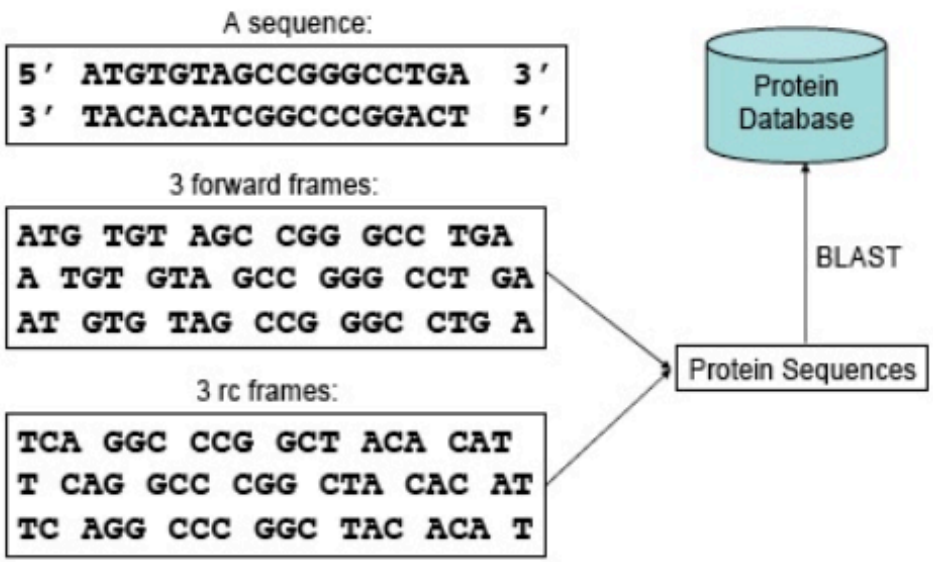
- Similarity-based methods (extrinsic) - use similarity to annotated sequences:
  - proteins
  - cDNAs
  - ESTs
- Comparative genomics - Aligning genomic sequences from different species
- *Ab initio* gene-finding (intrinsic)
- Integrated approaches

Similarity-based methods

- Based on sequence conservation due to functional constraints
- Use local alignment tools (Smith-Waterman algo, BLAST, FASTA) to search protein, cDNA, and EST databases
- Will not identify genes that code for proteins not already in databases (cannot identify new genes)
- Limits of the regions of similarity not well defined

Similarity-based methods

Similarity search for genes



Summary for Extrinsic Approaches

Strengths:

- Rely on accumulated pre-existing biological data, thus should produce biologically relevant predictions

Weaknesses:

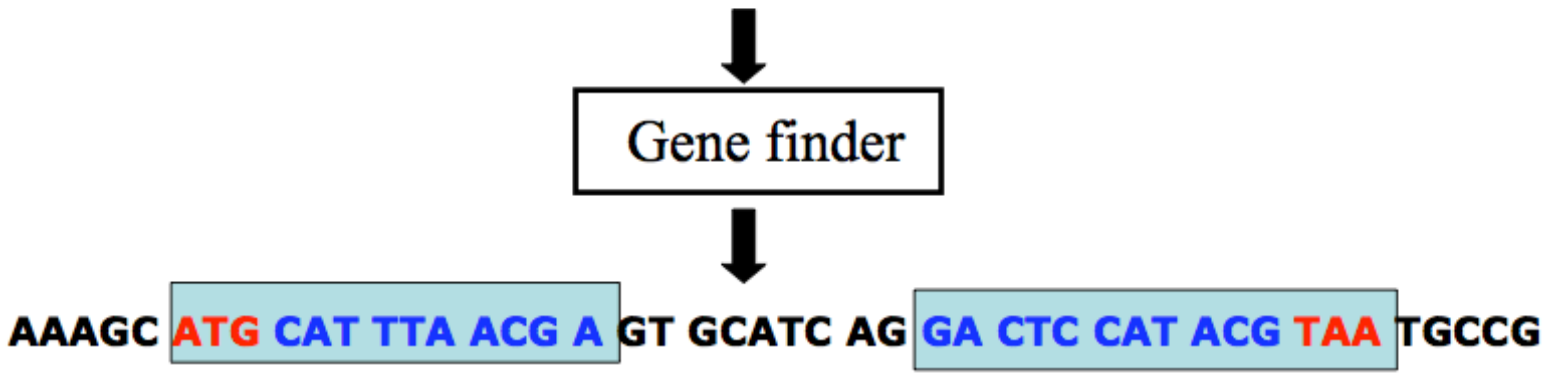
- Limited to pre-existing biological data
- Errors in databases
- Difficult to find limits of similarity

Ab initio Gene Finding, Part 1

Input: A DNA string over the alphabet *{A,C,G,T}*

Output: An annotation of the string showing for every nucleotide whether it is coding or non-coding

AAAGCATGCATTTAACGAGTGCATCAGGACTCCATACGTAATGCCG



Coding Statistics, Part 1

- Unequal usage of codons in the coding regions is a universal feature of the genomes
  - uneven usage of amino acids in existing proteins
  - uneven usage of synonymous codons (correlates with the abundance of corresponding tRNAs)
- We can use this feature to differentiate between coding and non-coding regions of the genome
- Coding statistics - a function that for a given DNA sequence computes a likelihood that the sequence is coding for a protein

Coding Statistics, Part 2

- Many different ones
  - codon usage
  - hexamer usage
  - GC content
  - compositional bias between codon positions
  - nucleotide periodicity
  - …

An Example of Coding Statistics, Part 1

The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

An Example of Coding Statistics, Part 2

- Let $F(c)$ be the frequency (probability) of codon c in the genes of the species under consideration

- Given the sequence of codons $C=c_1c_2\ldots c_m$ and assuming independence between adjacent codons:

$$P(C)=F(c_1)F(c_2)\ldots F(c_m)$$

is probability of finding C, knowing that C codes for protein

Example: S=AGGACC $\quad c_1$=AGG $\quad c_2$= ACC

$$P(S) = F(\text{AGG})\cdot F(\text{ACC}) = 0.022 \cdot 0.038 = 0.000836$$

An Example of Coding Statistics, Part 3

- Let $F_0(c)$ be the frequency of codon c in a non-coding sequence.

$$P_0\,(C)=F_0(c_1)F_0(c_2)\ldots F_0(c_m)$$

is the probability of finding C, knowing that C is non-coding

- Assuming the random model of non-coding DNA, $F_0(c)$ = 1/64= 0.0156 for all codons

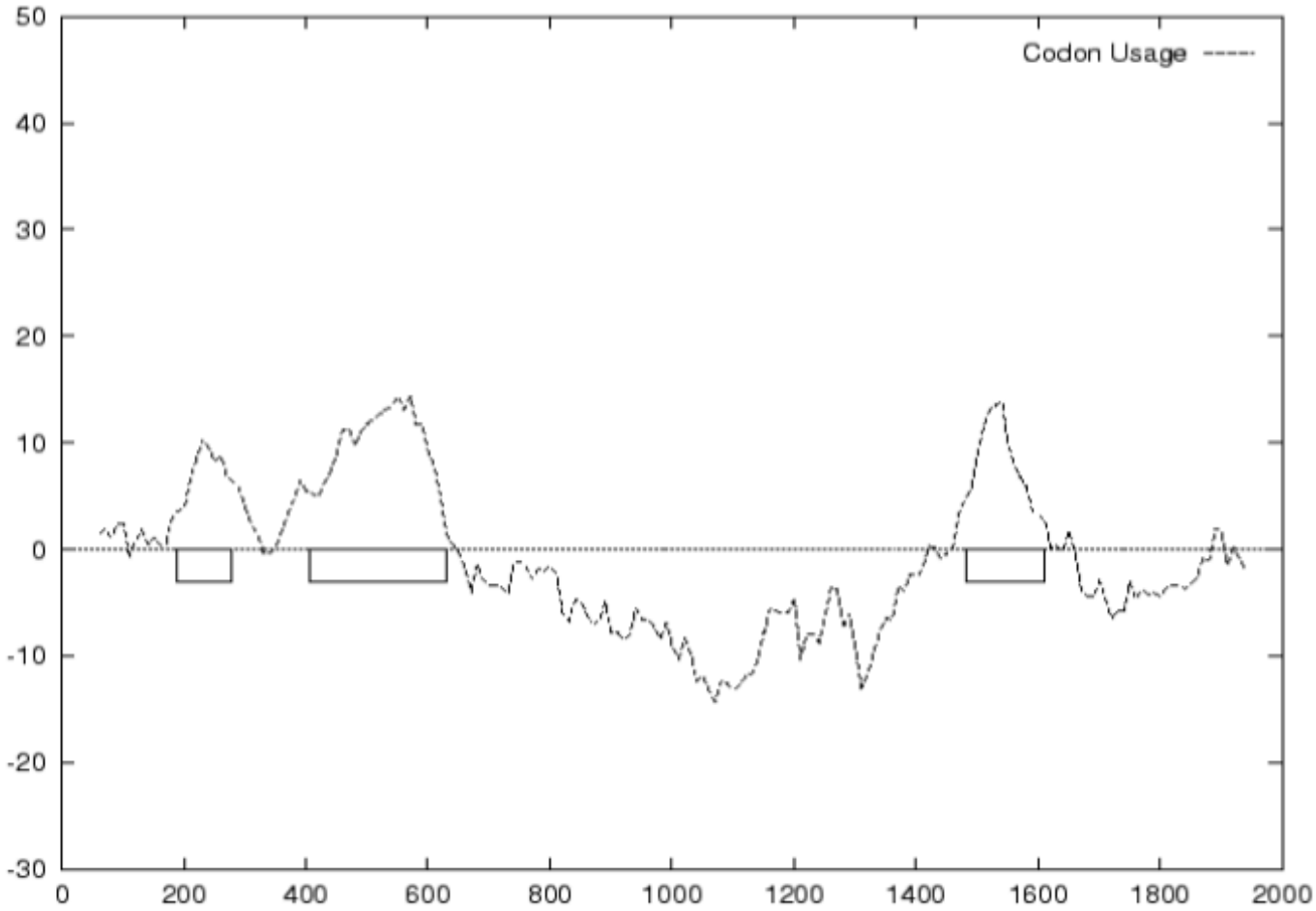
$$P_0\,(S) = 0.0156 \cdot 0.0156 = 0.000244$$

- The log-likelihood (LP) ratio for S is:

$$LP(S) = log(0.000836/0.000244) = log(3.43) = 0.53$$

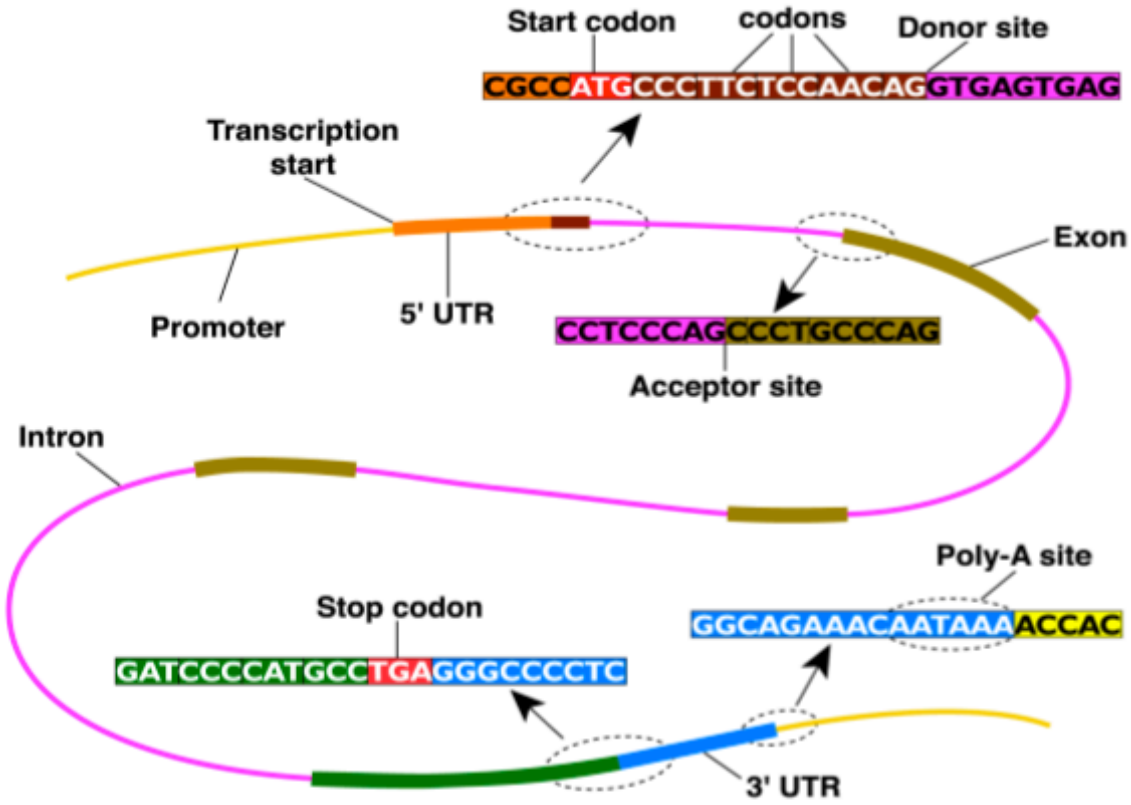
$LP(S) > 0$ ➡ S is coding

Coding Profile of ß-globin gene



Signal Sensors, Part 1

- Signal – a string of DNA recognized by the cellular machinery



Signal Sensors, Part 2

- Various pattern recognition method are used for identification of these signals:
  - consensus sequences
  - weight matrices
  - weight arrays
  - decision trees
  - Hidden Markov Models (HMMs)
  - neural networks
  - …

Prokaryotic Vs. Eukaryotic Gene Signals

Prokaryotes:

- Start codon: ATG
- Stop codon: TAA, TGA, TAG
- Promoters: TATAAT, -10 upstream
- Codon bias

Signal recognition for TATAAT

Positional weight matrix: %

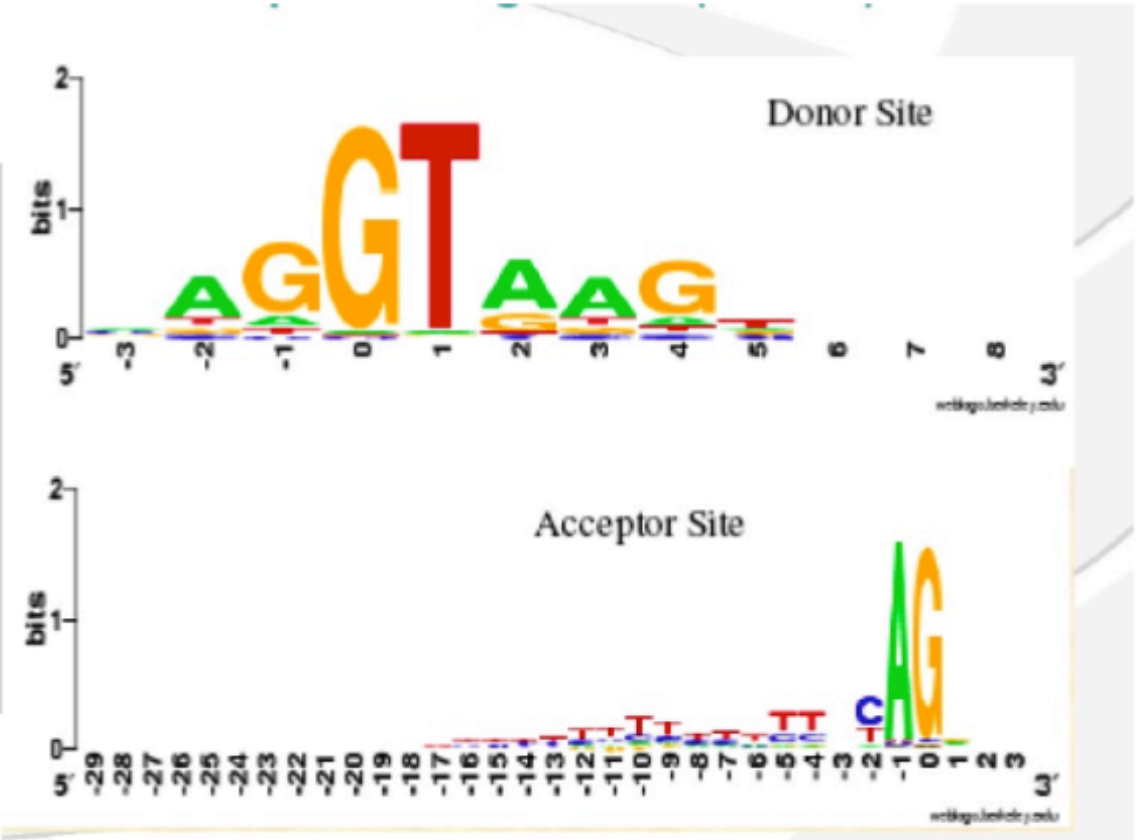
position	1	2	3	4	5	6
A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

Eukaryotes:

- Core promoter (CpG-rich, TATA-box, CAAT-box, ..)
- Start codon: ATG
- Stop codon: TAA, TGA, TAG
- Donor site: GT
- Acceptor site: AG
- Branching site
- Poly-A tail

Eukaryotic Gene Signals

- Start codon context (ryyAUGr)
- Splice signals
- polyA-signal
- Regulatory elements (CpG island, TATA-, CAAT-box,..)



PWM from 15412 Human introns

	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
A:	0,29	0,28	0,32	0,63	0,09	0,00	0,00	0,54	0,69	0,08	0,17	0,27	0,21	0,20	0,20
C:	0,26	0,28	0,38	0,12	0,03	0,00	0,01	0,03	0,08	0,06	0,17	0,21	0,28	0,29	0,26
G:	0,21	0,24	0,19	0,12	0,81	1,00	0,00	0,40	0,13	0,80	0,21	0,33	0,26	0,26	0,28
T:	0,25	0,20	0,11	0,14	0,07	0,00	0,99	0,03	0,11	0,06	0,45	0,20	0,26	0,25	0,26

GeneID ... in action

First step: splice sites, start and stop codons are predicted and scored along the sequence using Position Weight Matrices (PWMs)

Start codon

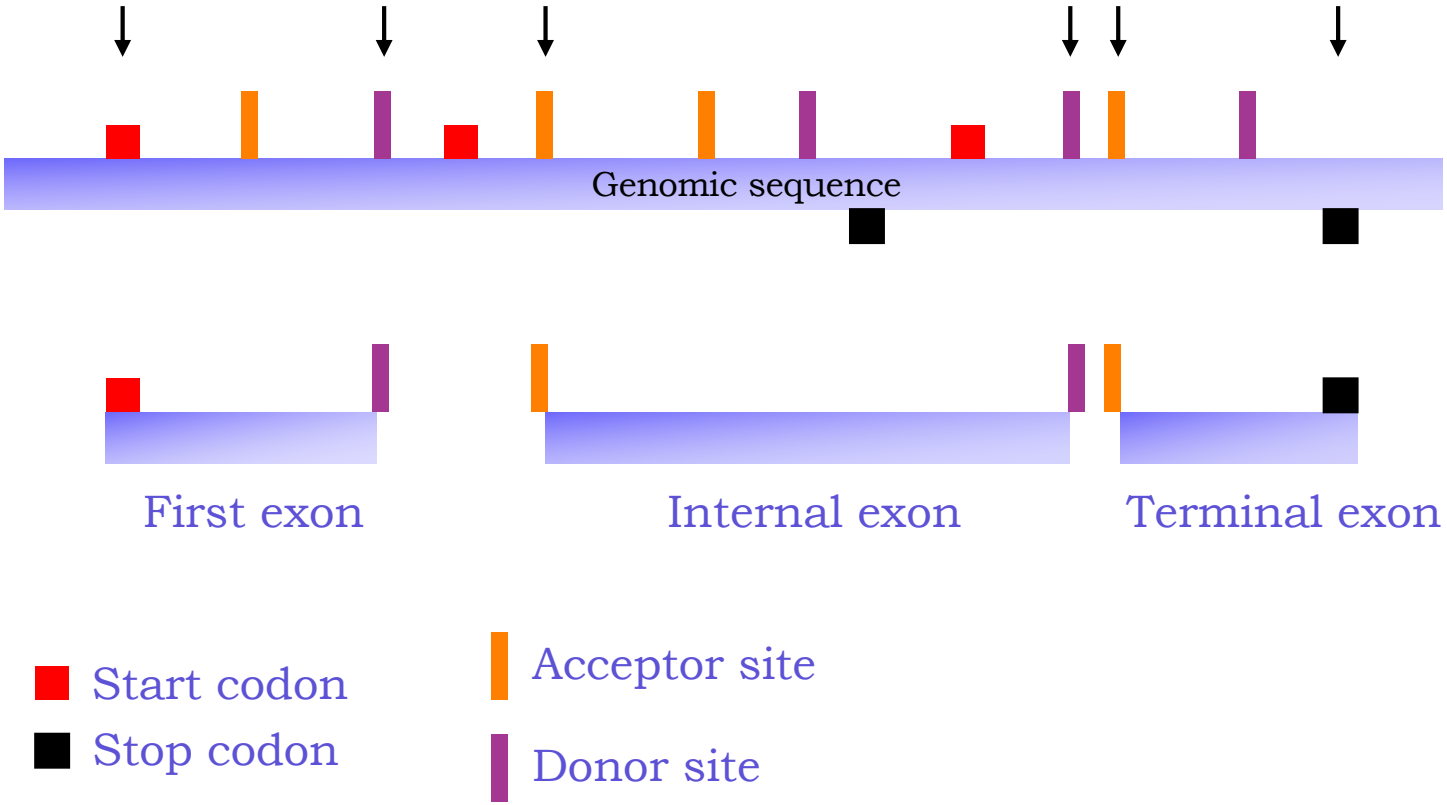
Stop codon

Acceptor site

Donor site

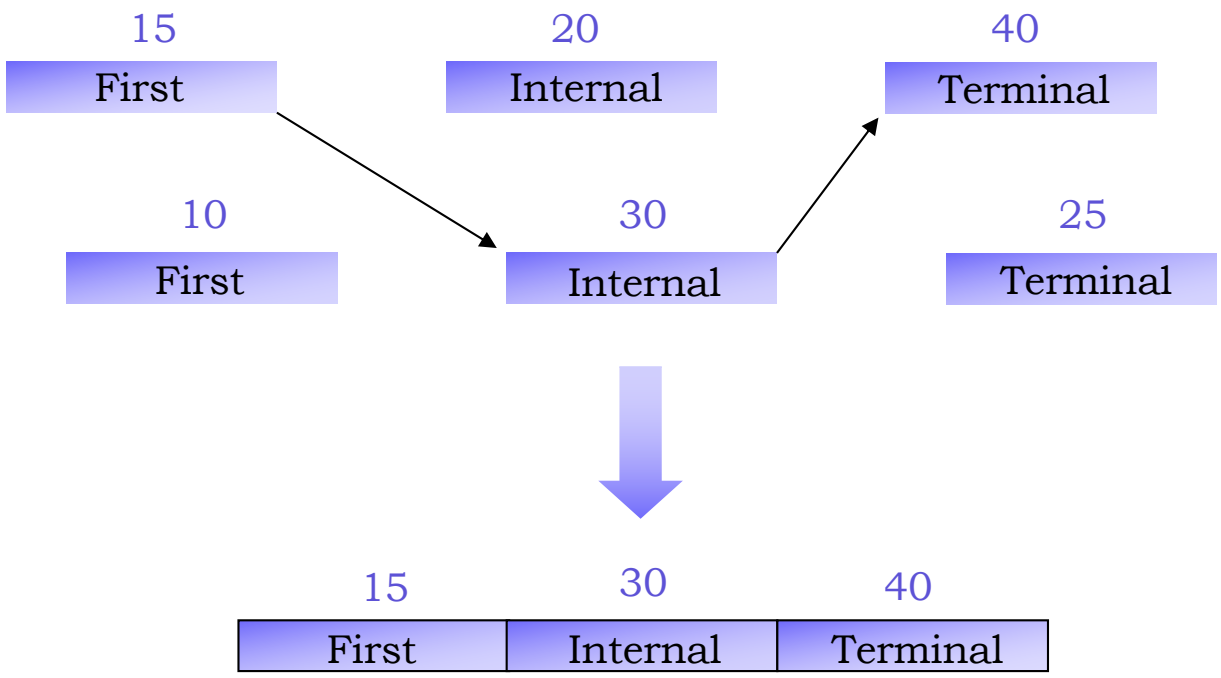
GeneID … in action

Second step: Exons are scored as the sum of the scores of the defining sites, plus the the log-likelihood ratio of a Markov Model for coding DNA



GeneID … in action

Last step: from the set of predicted exons, the gene structure is assembled maximizing the sum of the scores of the assembled exons



Predicted gene with score (15+30+40)

geneid 1.2 Web Server 2005

Paste your FASTA sequence here

... or search a FASTA file to process

Sfoglia...

Paste your GFF evidences here (Field separator: tab)

... or search a GFF file containing evidences to process

Sfoglia...

☐ **Do you want a graphical representation of the predictions ?**
(it might be time consuming depending on the size of the sequence)
Maximum sequence size for plots: 100,000 bps

http://genome.crg.es/geneid.html

Submit | Reset form

Ab initio Gene Prediction Tools

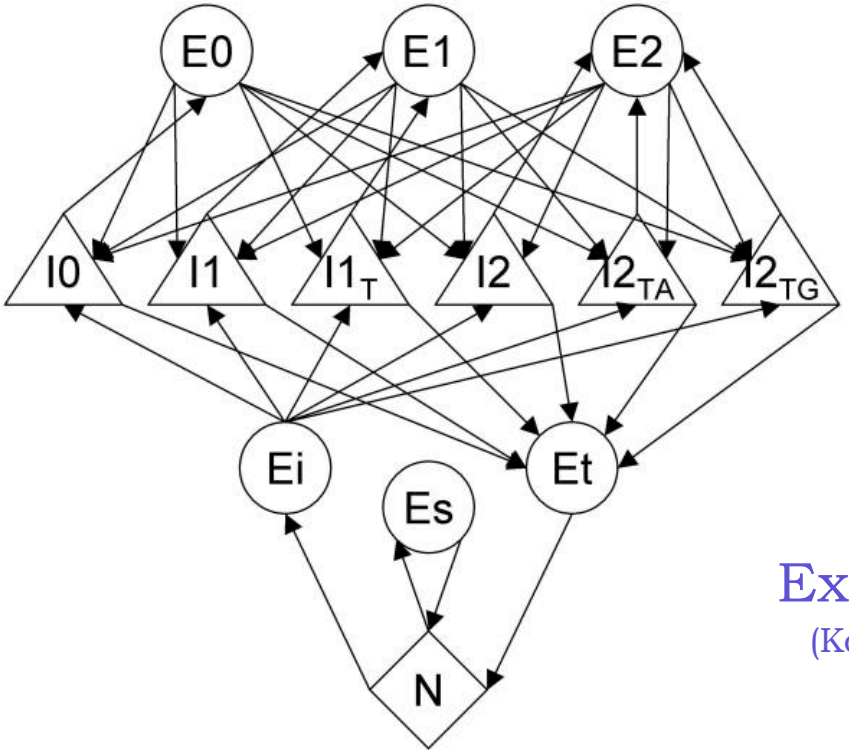
Ab initio gene finders remain the simplest and cost-effective technology for translating a genome to a set of exon-intron structures and the proteins they encode.

A lot of such *ab initio* gene predictions are freely available on world wide web:

- **Genscan** (http://genes.mit.edu/)
- **FGENESH** (http://www.softberry.com/berry.phtml?topic=gfind)
- **GeneMark.hmm** (http://opal.biology.gatech.edu/GeneMark)
- **GlimmerHMM** (http://www.genomics.jhu.edu/GlimmerHMM)
- **SNAP** (http://homepage.mac.com/iankorf)
- **Genie** (http://www.fruitfly.org/~martinr/doc/genie.html)

Ab initio
Gene Prediction Tools

A great part of these *ab initio* gene finders is based on HMM and thus on very complex probabilistic models.



LEGEND:
N: intergenic,
Es: single-exon gene
Ei: initial exon
Et terminal exon
E0–E2: exons in phase 0–2
I0–I2: introns in phases 0–2

Example…from SNAP
(Korf, 2004 *BMC Bioinformatics*)

The GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

For information about Genscan, click here

Server update, November, 2009: We've been recently upgrading the GENSCAN webserver hardware, which resulted in some problems in the output of GENSCAN. We apologize for the inconvenience. These output errors were resolved.

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page).

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (upper or lower case, spaces/numbers ignored): Sfoglia...

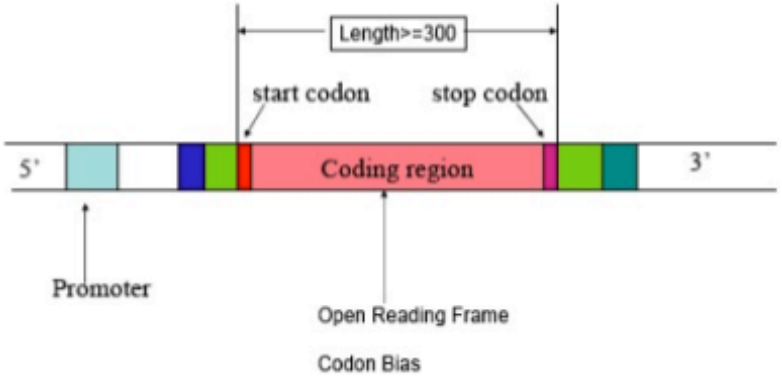
Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

http://genes.mit.edu/GENSCAN.html

Eukaryotic Gene Finding: GenScan

```
GENSCAN 1.0      Date run:  7-Nov-105     Time: 06:43:01

Sequence 11 : 9164 bp : 37.79% C+G : Isochore 1 ( 0 - 43 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:


Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.04 PlyA -     26     21    6                                1.05
 1.03 Term -    262    134  129  1  0  116   43   119 0.963   7.40
 1.02 Intr -   1335   1113  223  2  1  100   96   217 0.999  20.91
 1.01 Init -   1557   1466   92  0  2  103   77   133 0.990  13.71
 1.00 Prom -   1644   1605   40                               -4.75

 2.04 PlyA -   3125   3120    6                                1.05
 2.03 Term -   7628   7500  129  2  0  101   50    83 0.373   3.00
 2.02 Intr -   8749   8527  223  0  1   83   96   181 0.999  15.61
 2.01 Init -   8969   8878   92  2  2  103   77   142 0.997  14.61
 2.00 Prom -   9056   9017   40                               -2.75
```

Prokaryotic Gene Finding



Open Reading Frame

A sequence:

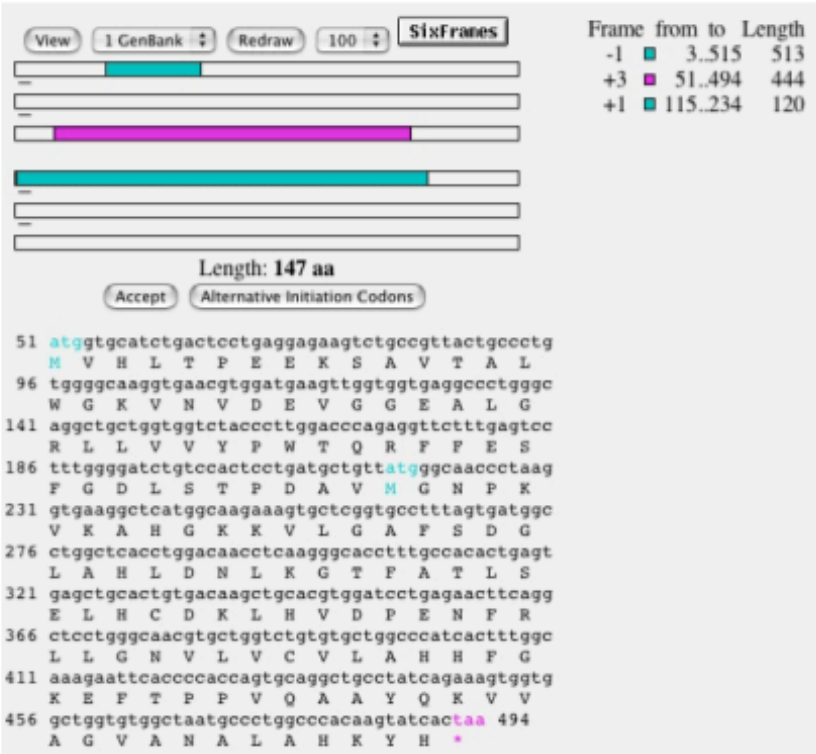
```
5' ATGTGTAGCCGGGCCTGA  3'
3' TACACATCGGCCCGGACT  5'
```

3 forward frames:

```
ATG TGT AGC CGG GCC TGA
A TGT GTA GCC GGG CCT GA
AT GTG TAG CCG GGC CTG A
```

3 rc frames:

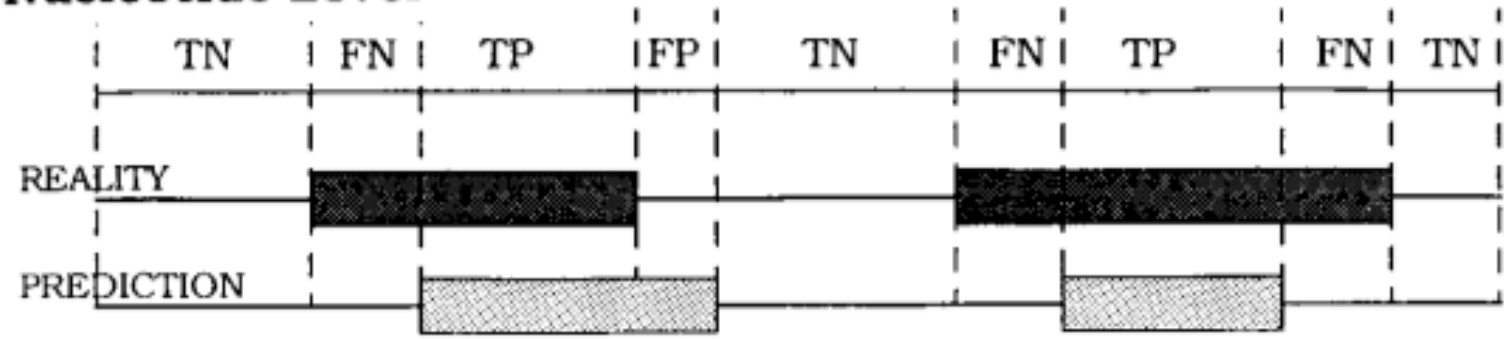
```
TCA GGC CCG GCT ACA CAT
T CAG GCC CGG CTA CAC AT
TC AGG CCC GGC TAC ACA T
```



A compilation of widespread *ab initio* and evidence-based gene prediction programs

Program	Web Page	Evidence
GENSCAN	http://genes.mit.edu/GENSCAN.html	No
GENEID	http://www1.imim.es/geneid.html	No
SNAP	http://homepage.mac.com/iankorf/	No
GlimmerHMM	http://www.genomics.jhu.edu/GlimmerHMM/	No
GeneMark	http://exon.gatech.edu/GeneMark/eukhmm.cgi	No
AUGUSTUS	http://augustus.gobics.de/	ESTs, cDNAs and proteins
SGP2	http://genome.imim.es/software/sgp2/sgp2.html	TBLASTX hits
GENOMESCAN	http://genes.mit.edu/genomescan.html	BLASTX hits
TWINSCAN	http://mblab.wustl.edu/nscan/submit/	BLASTN hits and ESTs
GENOMINER	http://pentagramma.caspur.it/GenoMinerNew/	Complete Genomes
ENSEMBL	http://www.ensembl.org/	ESTs, cDNAs and proteins
N-SCAN	http://mblab.wustl.edu/nscan/submit/	ESTs, complete genomes
EXOGEAN	http://www.biologie.ens.fr/dyogen/spip.php?rubrique4&lang=en	ESTs, cDNAs and proteins
GENEWISE	http://www.ebi.ac.uk/Wise2/index.html	Proteins
ASPIC	http://t.caspur.it/ASPIC/	ESTs and cDNAs
Eugène	http://www.inra.fr/mia/T/EuGene/	ESTs, cDNAs and proteins
GAZE	http://www.sanger.ac.uk/Software/analysis/GAZE/	All available + ab initio
JIGSAW	http://www.cbcb.umd.edu/software/jigsaw/	All available + ab initio

Nucleotide Level



PREDICTION \ REALITY	coding	no coding	
coding	TP	FP	TP+FP
no coding	FN	TN	FN+TN
	TP+FN	TF+TN	

$$Sn = \frac{TP}{TP + FN}$$

Sensitivity

$$Sp = \frac{TP}{TP + FP}$$

Specificity

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

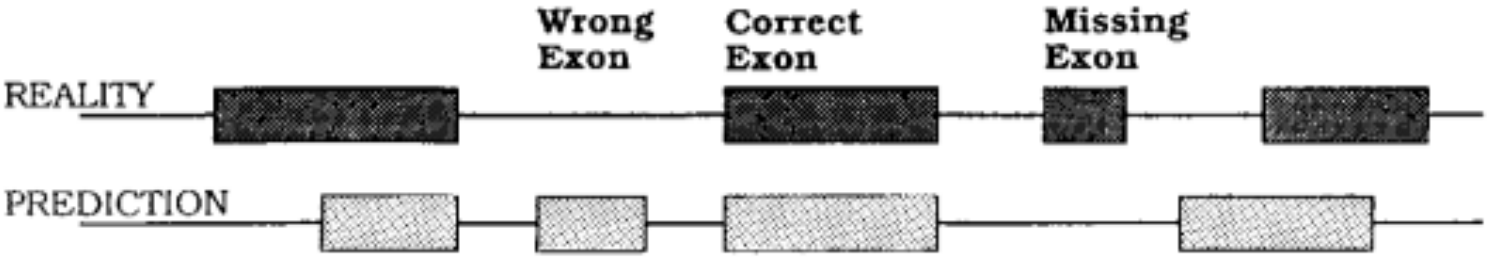
Correlation Coefficient

$$ACP = \frac{1}{4}\left[\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right]$$

$$AC = (ACP - 0.5) \times 2$$

Approximate Correlation

Exon Level



$$Sn = \frac{\text{number of Correct Exons}}{\text{number of Actual Exons}}$$ **Sensitivity**

$$Sn = \frac{\text{number of Correct Exons}}{\text{number of Predicted Exons}}$$ **Specificity**

$$ME = \frac{\text{number of Missing Exons}}{\text{number of Actual Exons}}$$ **(Sensitivity)**

$$WE = \frac{\text{number of Wrong Exons}}{\text{number of Predicted Exons}}$$ **(Specificity)**

Efficiency of gene finding programs