ABiMS

11/06/2014

RNA Seq analysis

Plateforme ABiMS

Galaxy

UPMC SORBONNE UNIVERSITÉS

CNRS

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

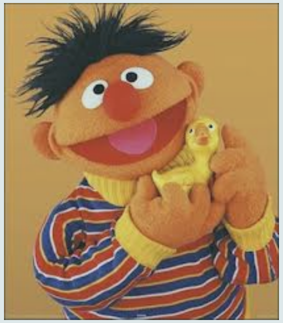
IBiSA INFRASTRUCTURES BIOLOGIE SANTÉ ET AGRONOMIE

EMBRC EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE

OCEANOMICS

IDEALG seaweed for the future

Biogenouest


Misharl Monsoor

Pierre Pericard

Alexandre Cormier

Gildas Le Corguillé

Erwan Corre

ABiMS – UMR 8227

Thanks - Sources

Brian Haas and Trinity team :
“Leveraging RNA-Seq for Genome-free Transcriptome Studies” 2012
“Transcriptome assembly with Trinity: How it works” 2011

genotoul bioinfo

PEPI

- RNA-Seq de novo assembly training session Sigenea 2013. C. Cabau . C. Klopp
- PEPI IBIS -Partage de pratique et d’experience en informatique INRA: Ingénierie Bio Informatique et Statistique pour les données haut-débit (IBIS)

EMBL 40 YEARS

Simon Anders Research Scientist at Huber Group

RNA-Seq Blog
Transcriptome Sequencing Research & Industry News

http://www.rna-seqblog.com/

nescent academy
Hands-on instruction for your evolutionary toolbox

Keyshawn Goldsby

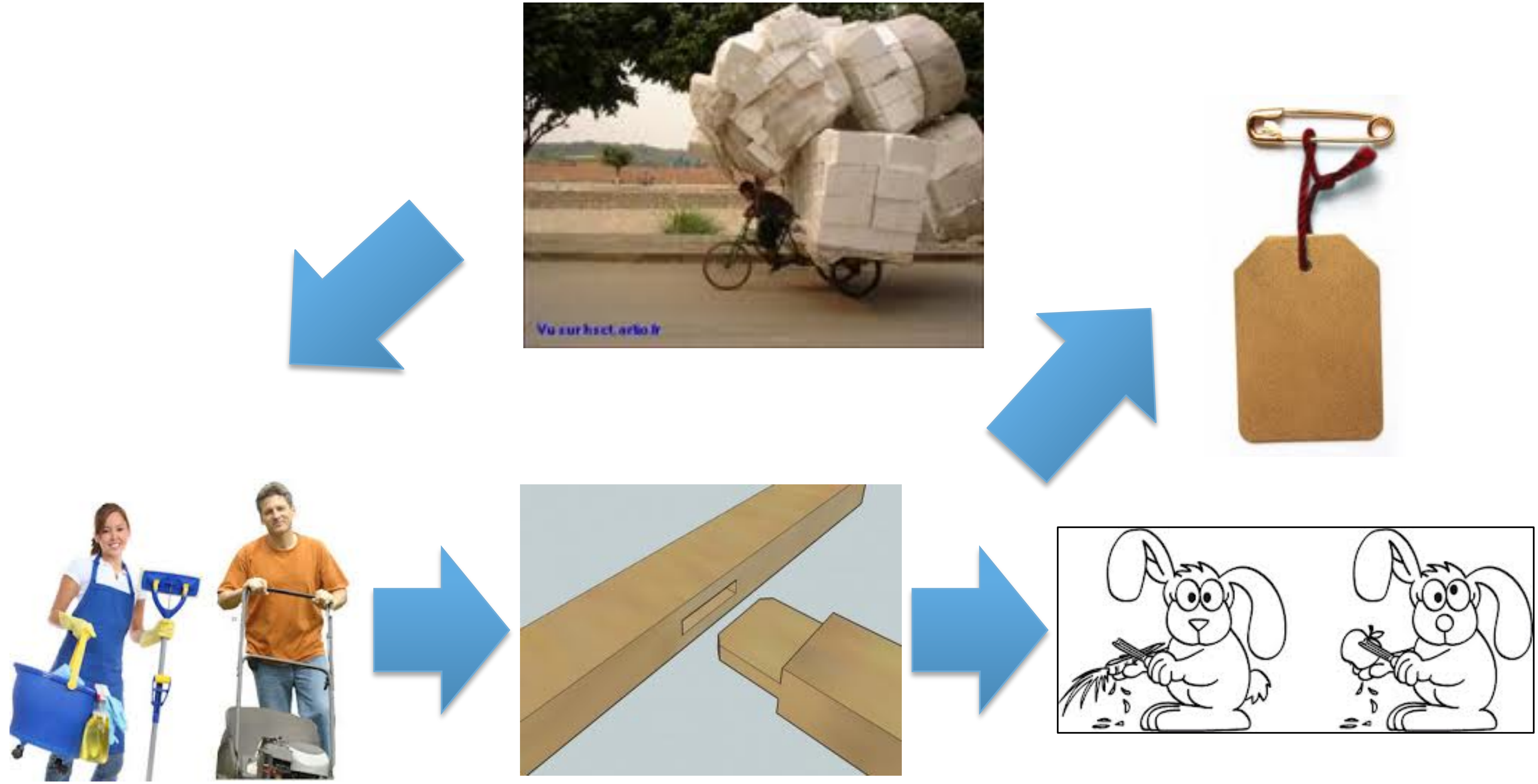
ABiMS

ABiMS collaborations projects

WIKIPÉDIA
L’Encyclopédie libre

Google

RNA Seq analysis

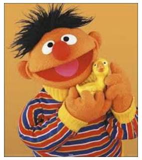


Courses Organisation

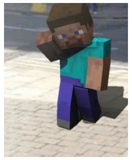
Wednesday

- Introduction
  - Transcriptome definitions and variability
  - RNAseq vs Micro-array
  - How deep is enough
  - Library construction bias
  - Sequencing terminology

- Data cleaning
- RNAseq analysis
  - Assembly algorithm
  - RNASeq assembly



- De novo assembly TP
- Reference assembly TP



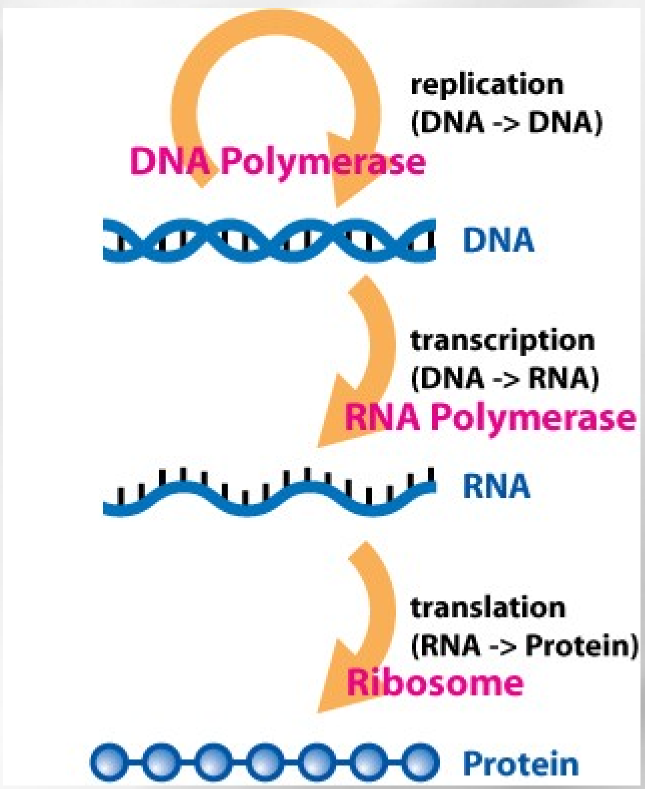
Thursday

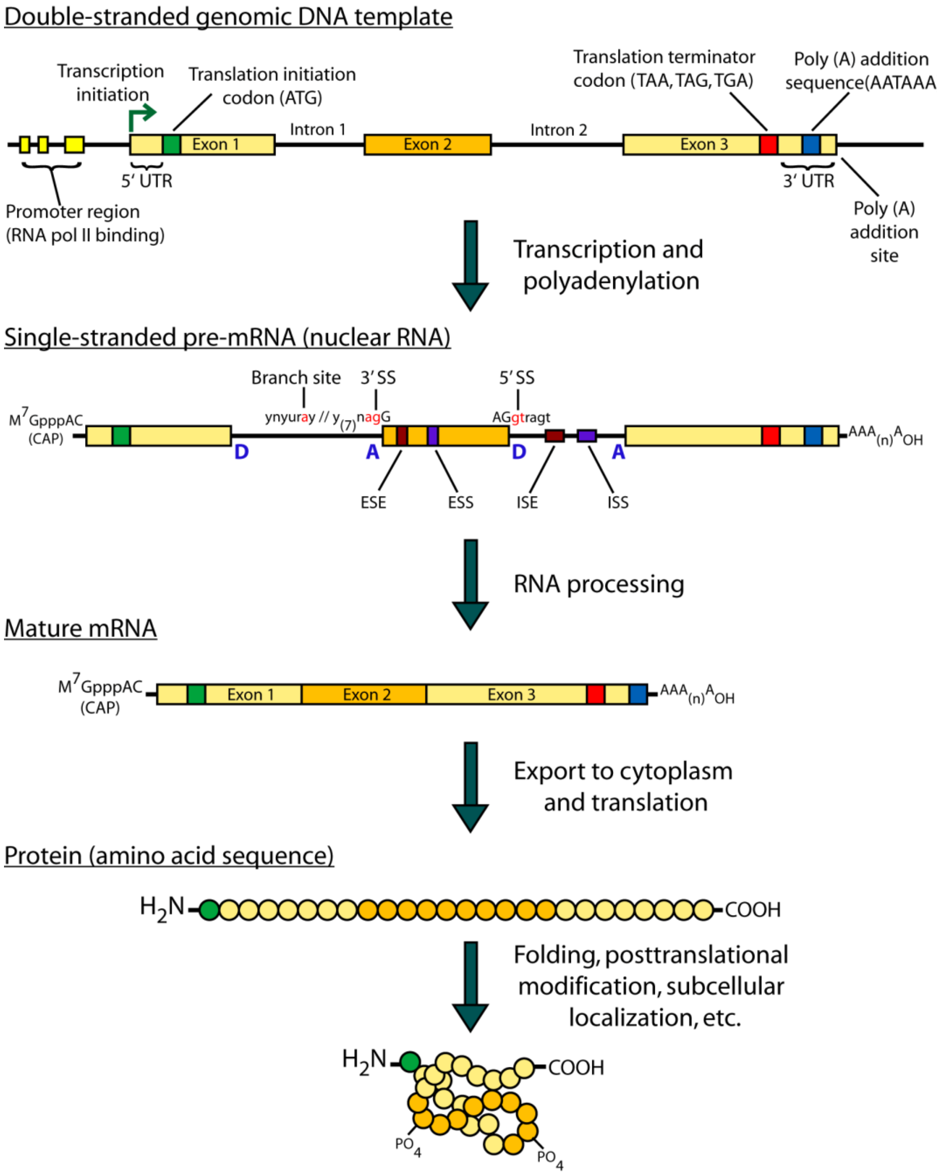
- Expression analysis
- Transcriptome annotation

Introduction

TRANSCRIPTOME DEFINITIONS AND VARIABILITY

Gene expression




Double-stranded genomic DNA template
Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Intron 1
Intron 2
Exon 1
Exon 2
Exon 3
5′ UTR
3′ UTR
Promoter region (RNA pol II binding)
Poly (A) addition site
Transcription and polyadenylation
Single-stranded pre-mRNA (nuclear RNA)
Branch site
3′ SS
5′ SS
M^7GpppAC (CAP)
ynyuray // $y_{(7)}$nagG
AGgtragt
D
A
D
A
ESE
ESS
ISE
ISS
$AAA_{(n)}A_{OH}$
RNA processing
Mature mRNA
M^7GpppAC (CAP)
Exon 1
Exon 2
Exon 3
$AAA_{(n)}A_{OH}$
Export to cytoplasm and translation
Protein (amino acid sequence)
H_2N
COOH
Folding, posttranslational modification, subcellular localization, etc.
H_2N
COOH
PO_4
PO_4

Transcription

Definition :

Transcription is the process of creating a complementary RNA copy of a sequence of DNA. Transcription is the first step leading to gene expression.

Transcription product

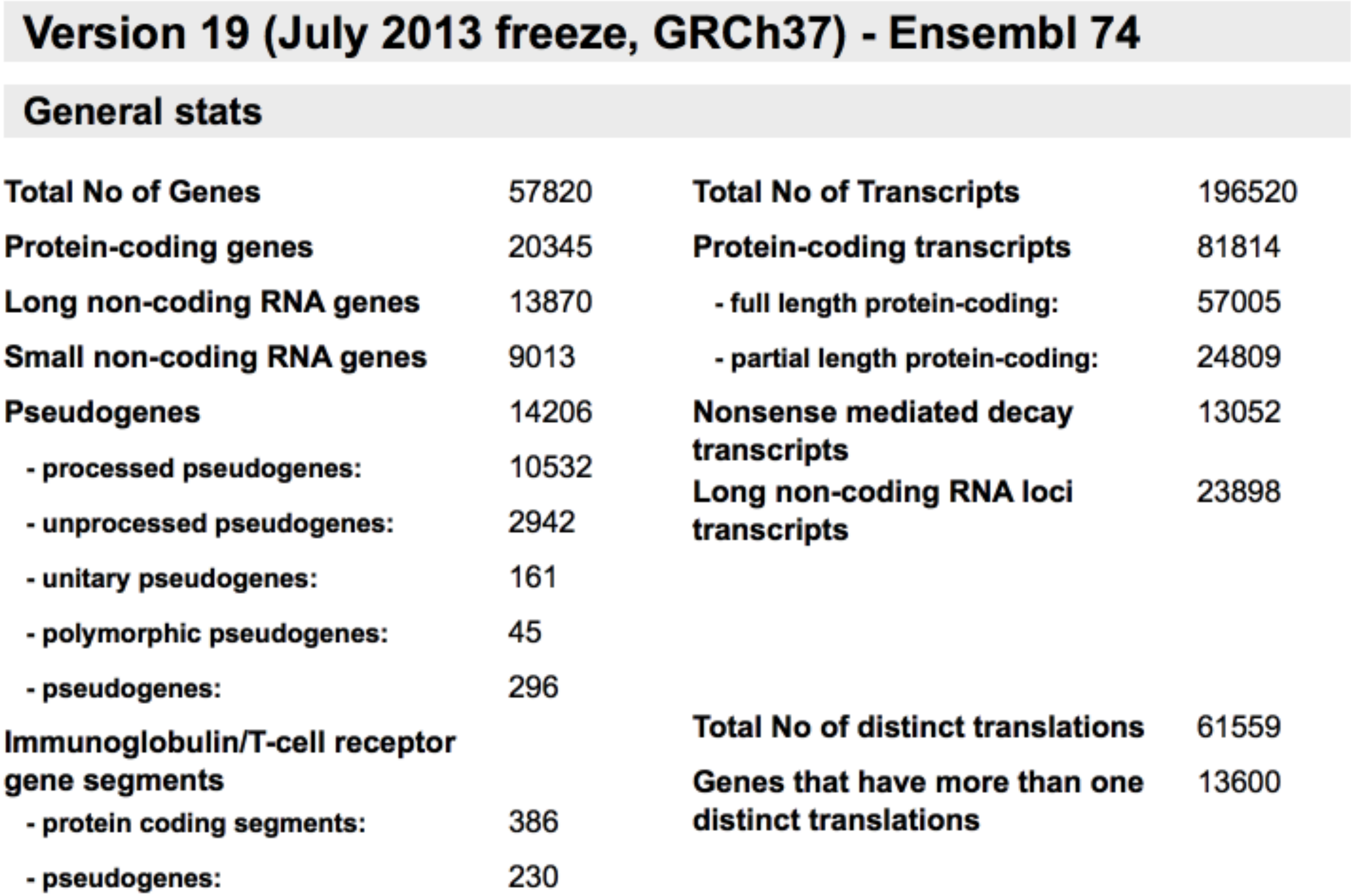
Protein coding gene: transcribed in mRNA

ncRNA : highly abundant and functionally important RNA (up 95%)

- tRNA,
- rRNA,
- Regulatory RNA
    - snoRNAs (rRNA maturation)
    - microRNAs (post-transcriptional regulators)
    - siRNAs (mRNA degradation)
    - piRNAs (block the activity of the mobile elements)
    - LincRNA (regulators of diverse cellular processes)
    - VlincRNA...

The GENCODE Project:

Encyclopædia of genes and gene variants



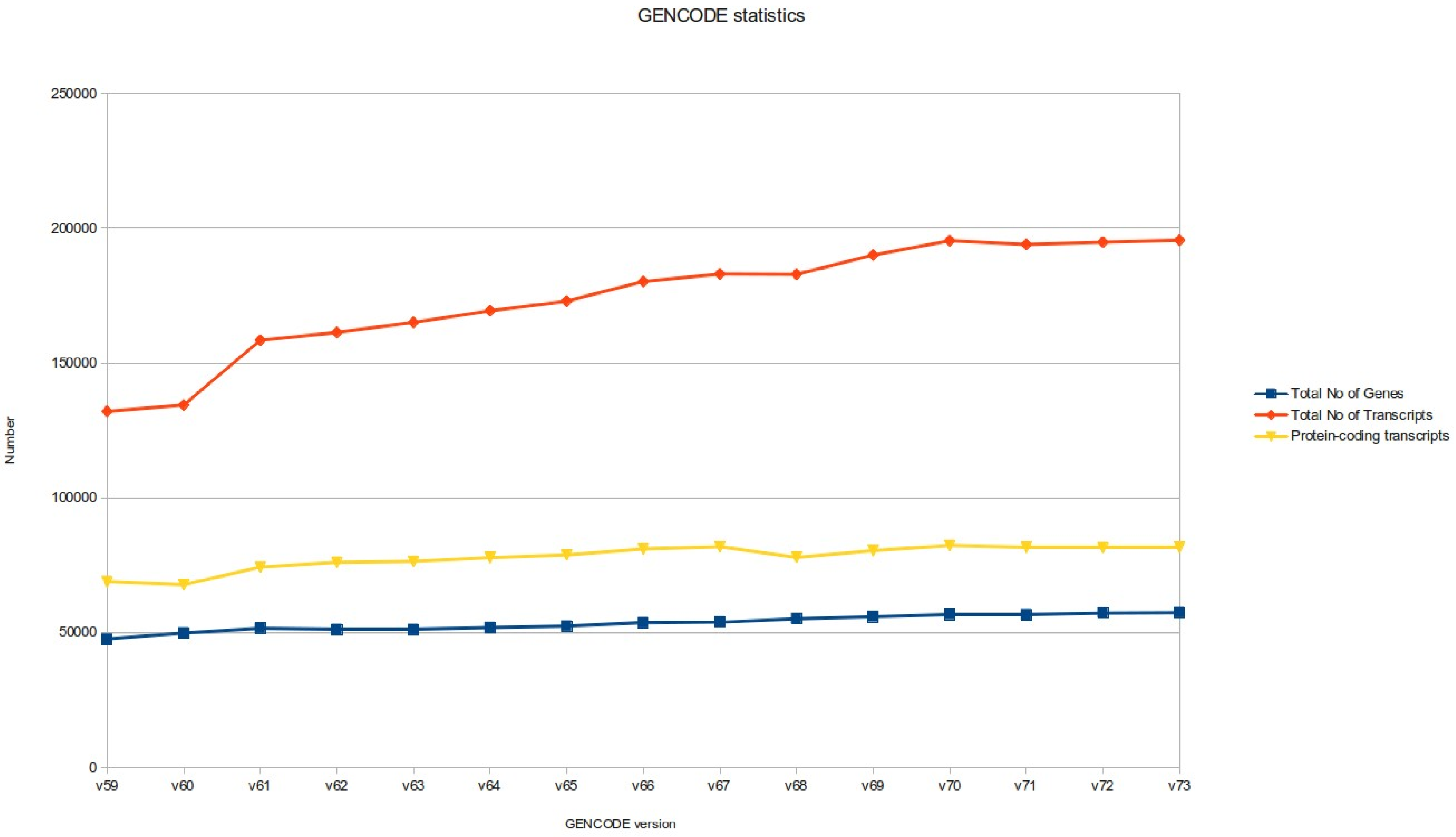
Version 19 (July 2013 freeze, GRCh37) - Ensembl 74

General stats

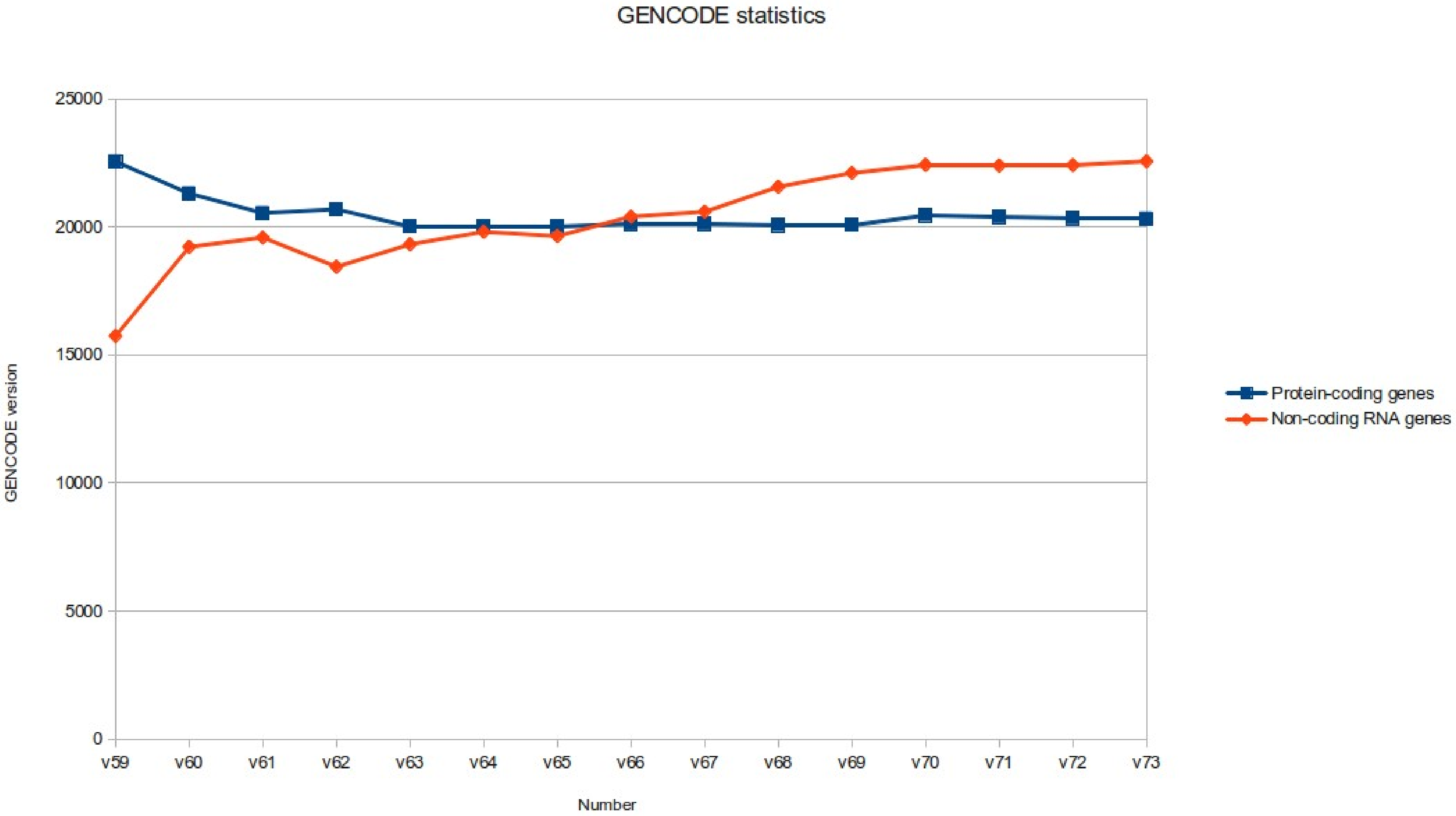
Total No of Genes	57820
Protein-coding genes	20345
Long non-coding RNA genes	13870
Small non-coding RNA genes	9013
Pseudogenes	14206
- processed pseudogenes:	10532
- unprocessed pseudogenes:	2942
- unitary pseudogenes:	161
- polymorphic pseudogenes:	45
- pseudogenes:	296
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments:	386
- pseudogenes:	230

Total No of Transcripts	196520
Protein-coding transcripts	81814
- full length protein-coding:	57005
- partial length protein-coding:	24809
Nonsense mediated decay transcripts	13052
Long non-coding RNA loci transcripts	23898
Total No of distinct translations	61559
Genes that have more than one distinct translations	13600

Non proteins-coding transcripts



Non - coding genes



Sources of variability

Biological elements which tend to blur the signal

- Repeats
- Gene families
- Pseudogenes
- Alternative splicing
- Intron retention
- (Cis-)natural anti-sens transcript
- Fusion genes

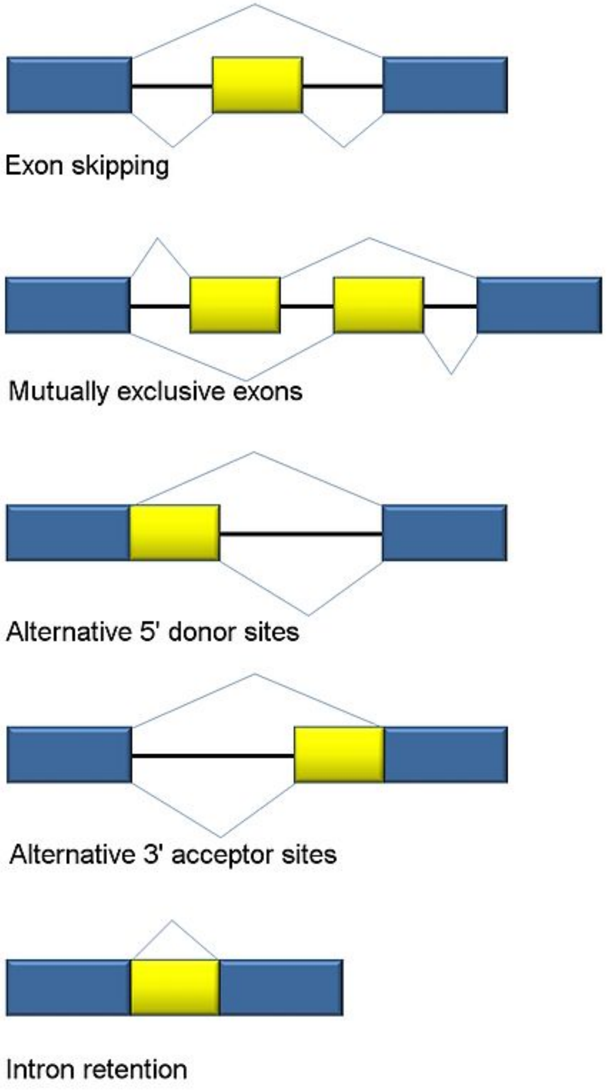
Elements removing or masking the signal

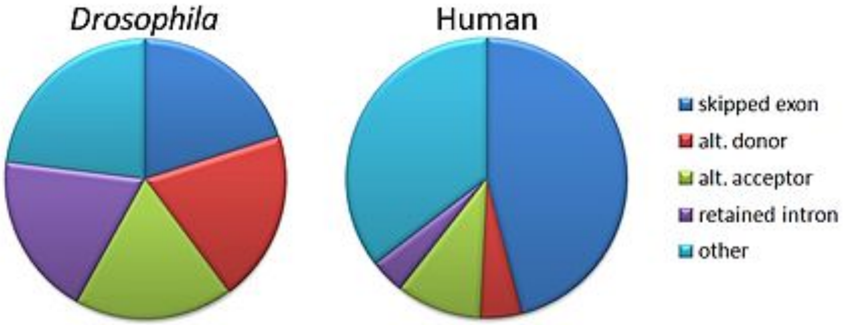
- Transcript decay
- Sequencing protocol biases
- Sequencing depth

Other elements :

- PolyAtails
- Adapters
- Contamination

Alternative Splicing and circular mrna



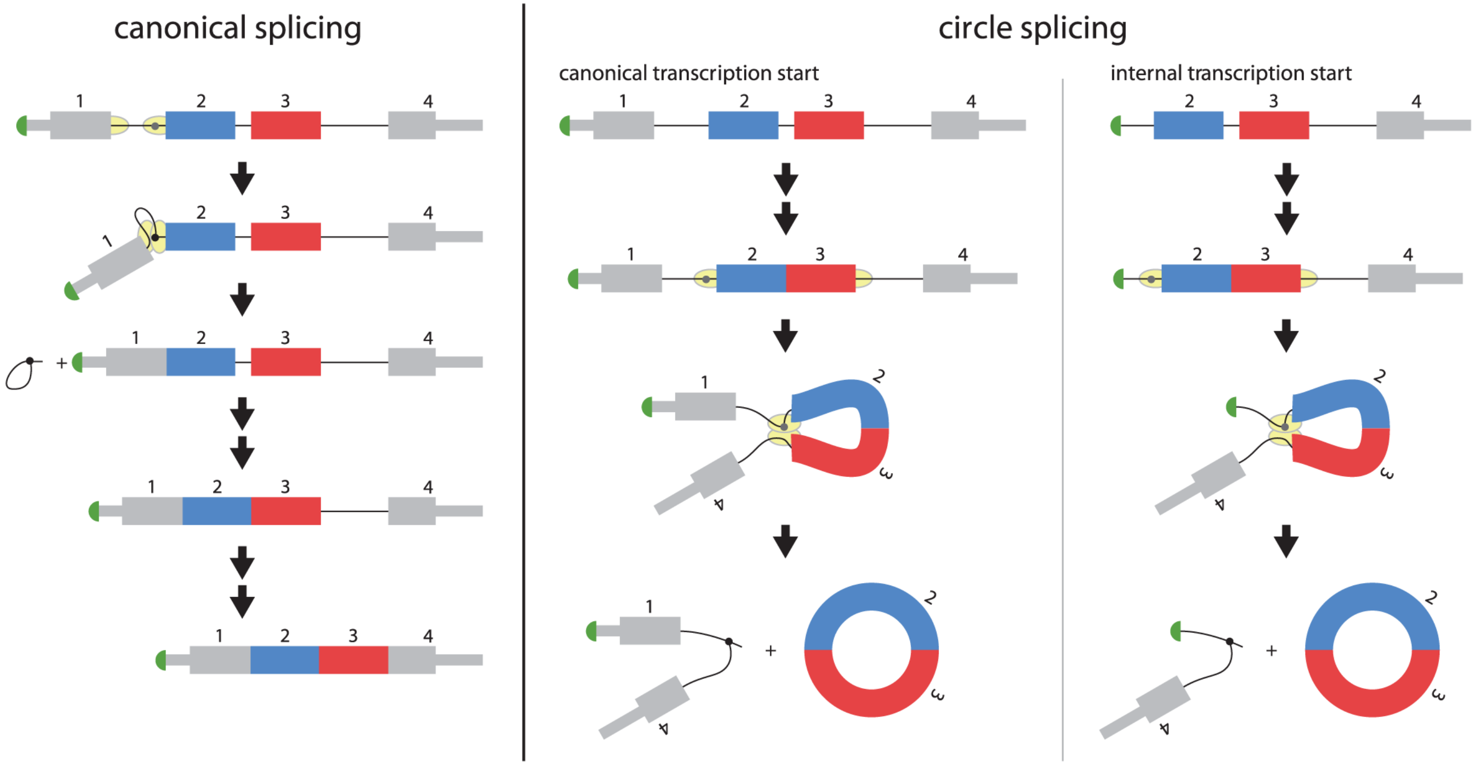


« In humans, for example, there is evidence for alternative splicing of more than **95% of genes** [1], with an average of more than **five isoforms per gene.**
Somewhat surprisingly, alternatively spliced isoforms from a single gene can also have very different, even antagonistic, functions [2] »

[1] Pan et al. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 40: 1413-1415.

[2] Boise et al. (1993) Bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 74: 597-608.

Circular mrna



Salzman, J. et al. (2012). Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types PLoS ONE, 7 (2) DOI: 10.1371/journal.pone.0030733

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., et al. (2014). A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biology, 15(2), R34.

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19: 141-157.

Nitsche, A., Doose, G., Tafer, H., Robinson, M., Saha, N. R., Gerdol, M., et al. (2013). Atypical RNAs in the coelacanth transcriptome. Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution.

Natural anti-sense transcripts

Natural anti-sense transcripts (NATs) are a group of RNAs encoded within a cell that have transcript complementarity to other RNA transcripts.

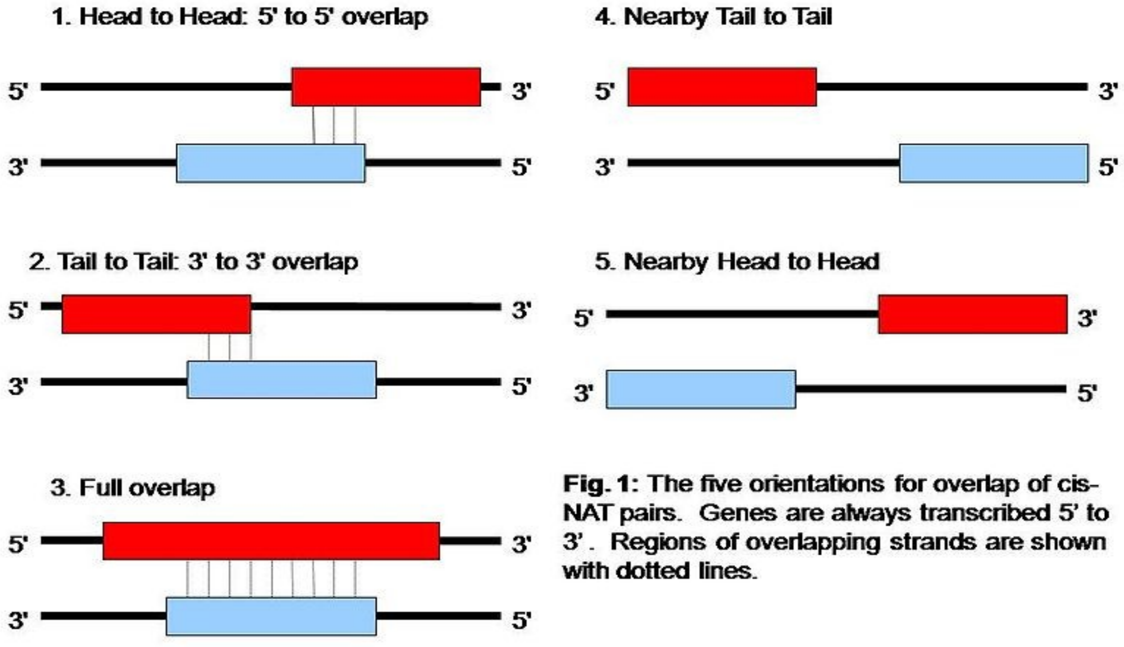


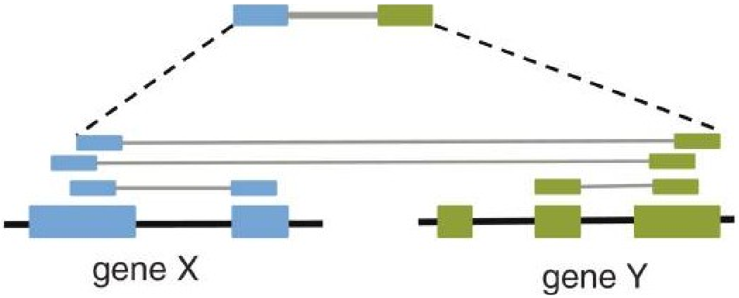
Fig. 1: The five orientations for overlap of cis-NAT pairs. Genes are always transcribed 5' to 3'. Regions of overlapping strands are shown with dotted lines.

http://en.wikipedia.org/wiki/Cis-natural_antisense_transcript

Fusion gene

A fusion gene is a hybrid gene formed from two previously separate genes. It can occur as the result of a translocation, interstitial deletion, or chromosomal inversion. Often, fusion genes are oncogenes.

They often come from trans-splicing : Trans-splicing is a special form of RNA processing in eukaryotes where exons from two different primary RNA transcripts are joined end to end and ligated.



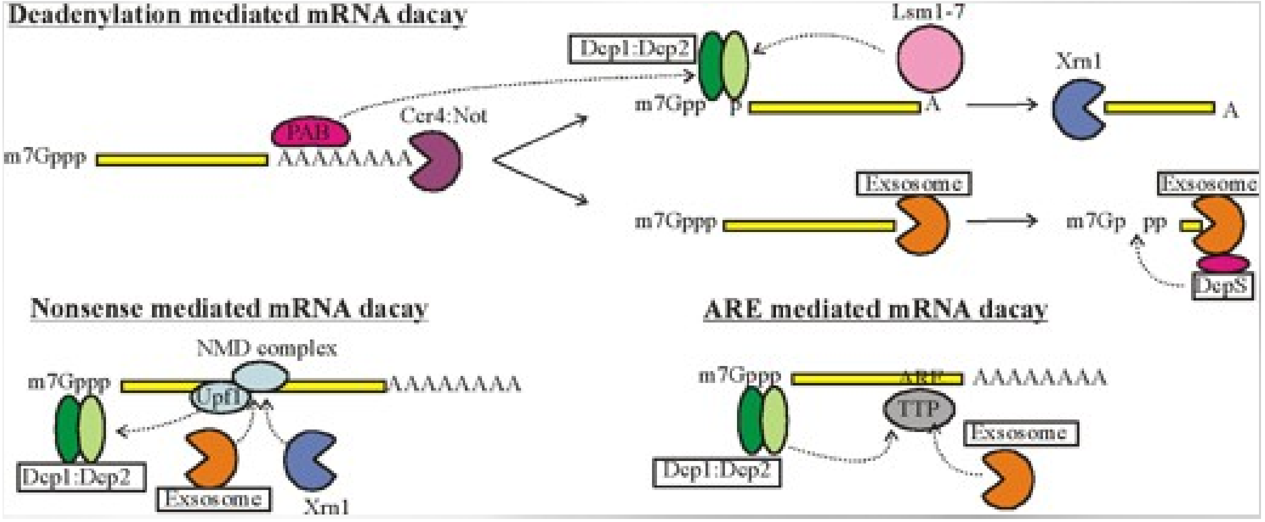
Distinguishing gene fusions from noise in RNA-Seq data is extremely difficult :

Panagopoulos I, Thorsen J, Gorunova L, Micci F, Heim S. (2014) **Sequential combination of karyotyping and RNA-sequencing in the search for cancer-specific fusion genes.** Int J Biochem Cell Biol.

Transcript decay

After export to the cytoplasm, mRNA is protected from degradation by a 5’ cap structure and a 3’ poly adenine tail.

In the deadenylation dependent mRNA decay pathway, the polyA tail is gradually shortened by exonucleases. This ultimately attracts the degradation machinery that rapidly degrades the mRNA in both in the 5’ to 3’ direction and in the 3’ to 5’ direction.



http://www.eb.tuebingen.mpg.de/research/research-groups/remco-sprangers/mrna-degradation.html

Introduction

RNASEQ ANALYSIS

How RNA-seq works

Methodology:

– RNA is isolated from cells,

– Fragmented at random positions,

– Copied into complementary DNA,

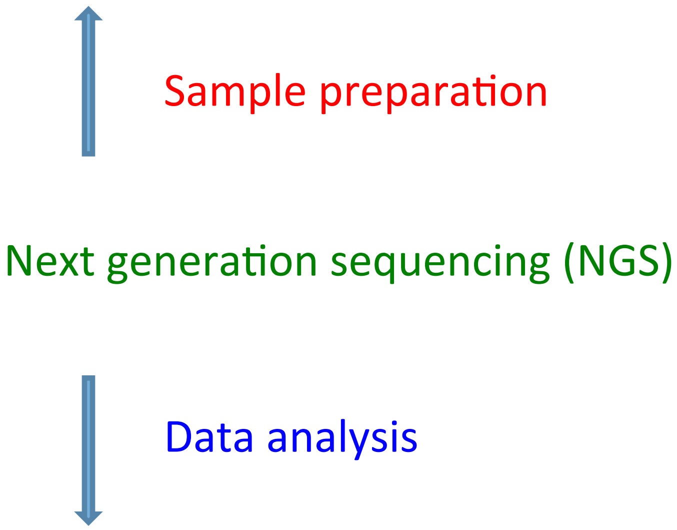
– Selection of fragments with a certain size range,

– Amplification using PCR,

– Sequencing,

– Reads are aligned to a reference genome or de novo assembled,

– The number of sequencing reads mapped to each gene in the reference is tabulated.



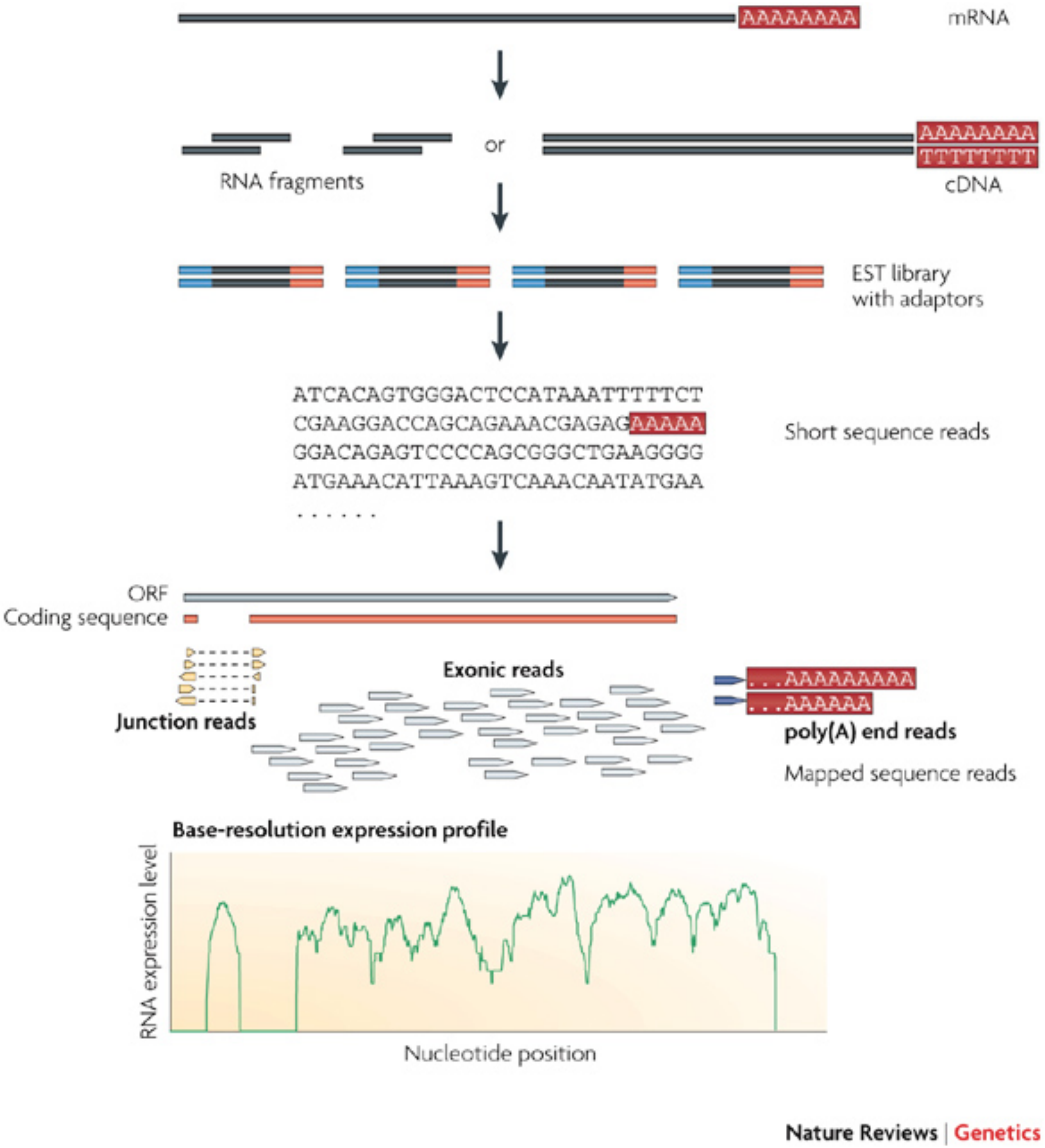
Sample preparation

Next generation sequencing (NGS)

Data analysis

Figure from Wang et. al, **RNA-Seq: a revolutionary tool for transcriptomics,** Nat. Rev. Genetics 10, 57-63, 2009**.**

How RNA-seq works



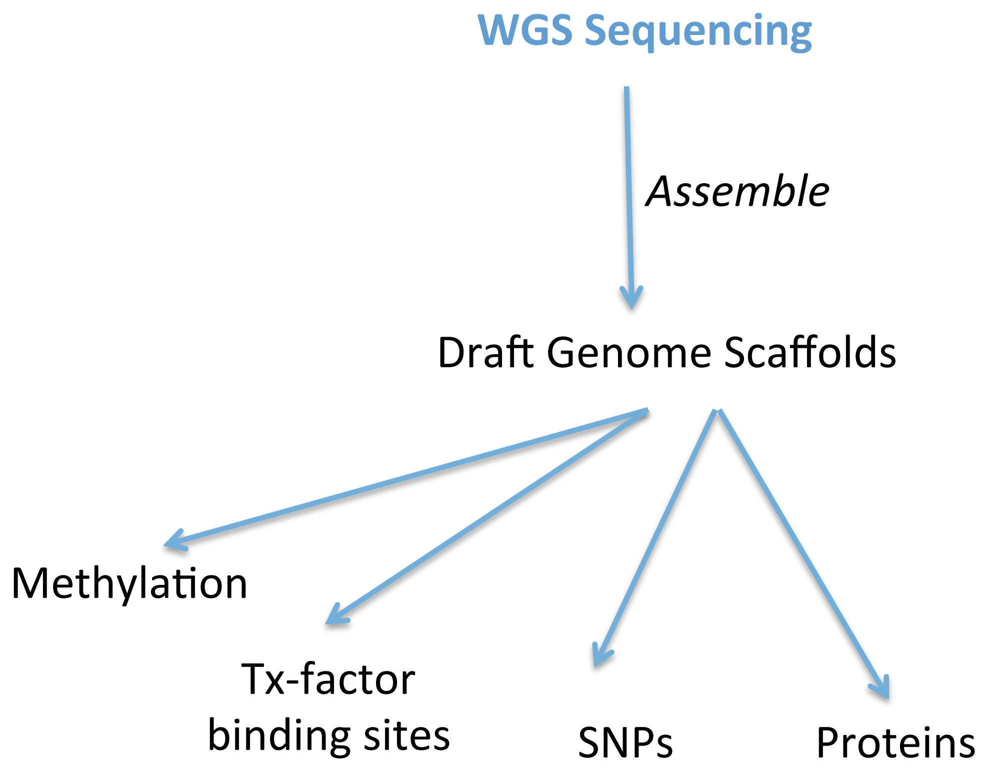
Sample preparation

Next generation sequencing (NGS)

Data analysis

Figure from Wang et. al, **RNA-Seq: a revolutionary tool for transcriptomics,** Nat. Rev. Genetics 10, 57-63, 2009).

A Paradigm for Genomic Research



WGS Sequencing

Assemble

Draft Genome Scaffolds

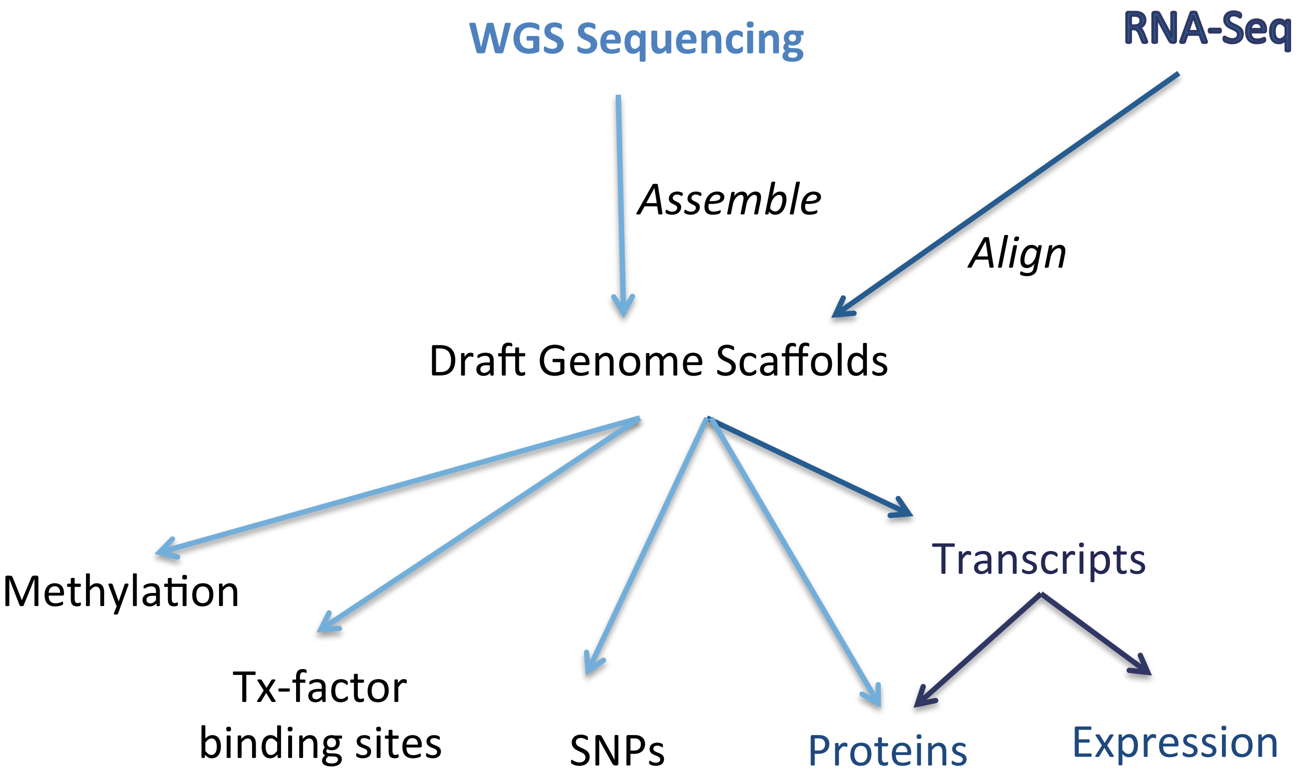
Methylation

Tx-factor
binding sites

SNPs

Proteins

A Paradigm for Genomic Research



WGS Sequencing

RNA-Seq

Assemble

Align

Draft Genome Scaffolds

Methylation

Tx-factor binding sites

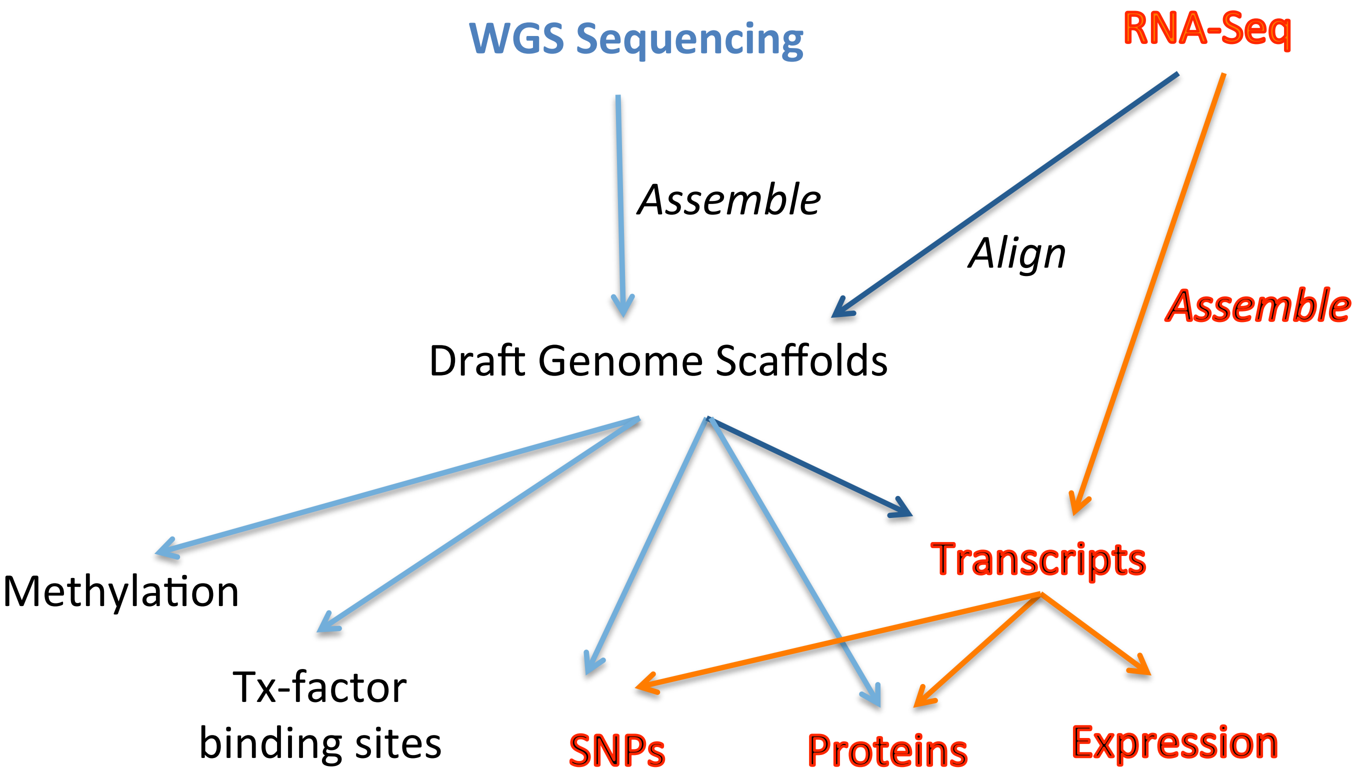
SNPs

Proteins

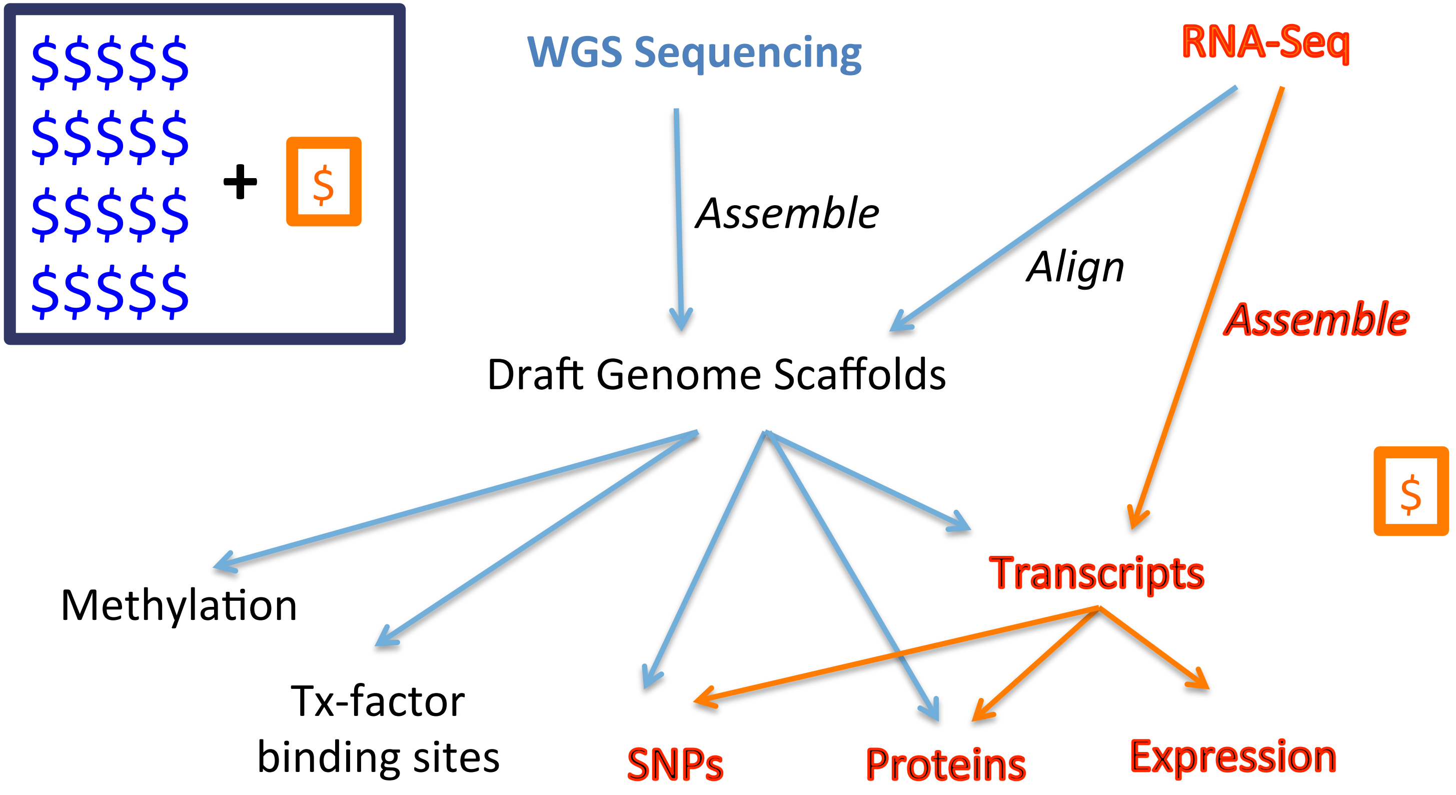
Transcripts

Expression

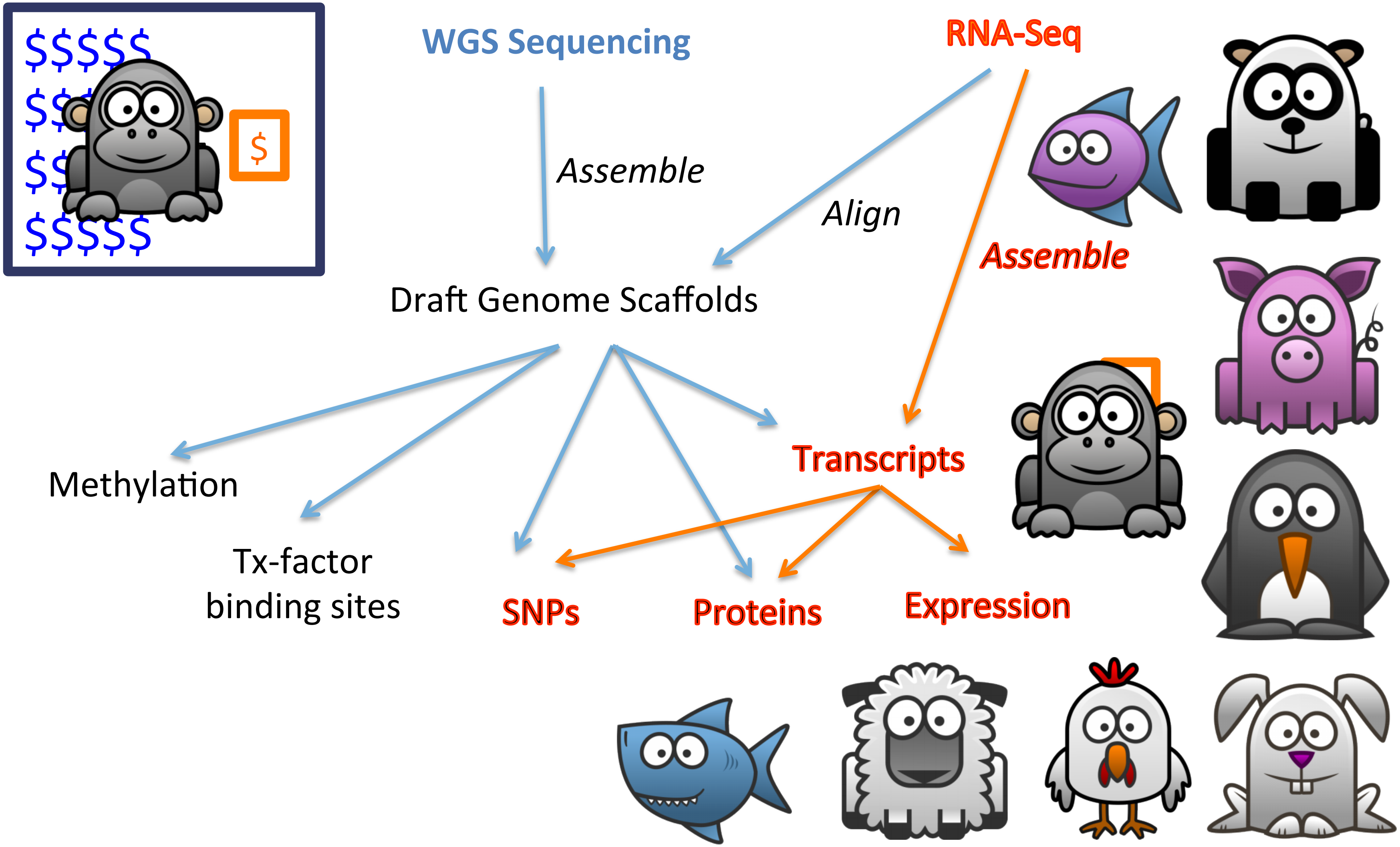
A *Maturing* Paradigm for Transcriptome Research



A Maturing Paradigm for Transcriptome Research



A **Maturing** Paradigm for Transcriptome Research



WGS Sequencing

Assemble

RNA-Seq

Align

Assemble

Draft Genome Scaffolds

Methylation

Tx-factor binding sites

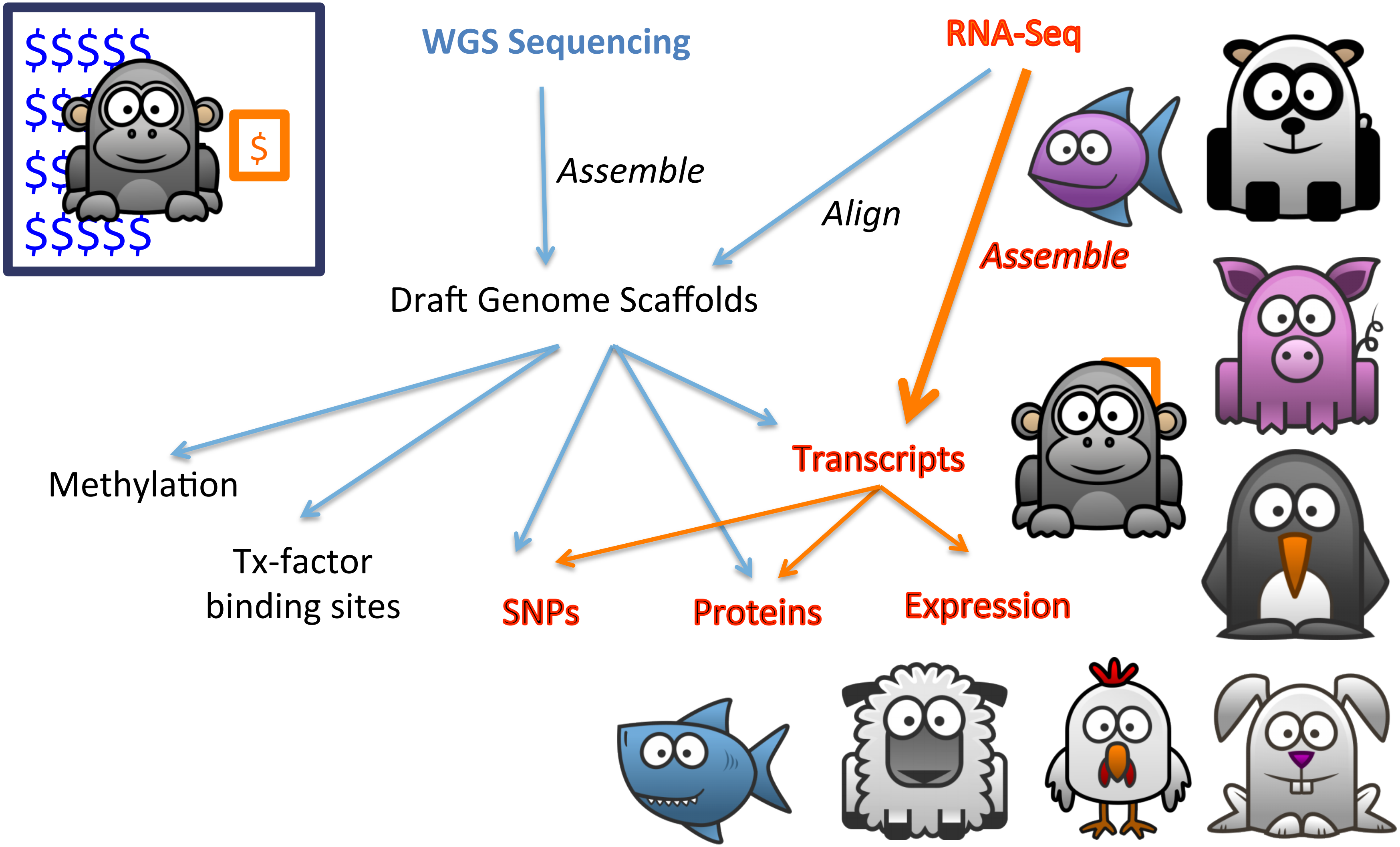
SNPs

Proteins

Transcripts

Expression

A **Maturing** Paradigm for Transcriptome Research



WGS Sequencing

Assemble

Draft Genome Scaffolds

RNA-Seq

Align

Assemble

Transcripts

Methylation

Tx-factor binding sites

SNPs

Proteins

Expression

mRNA-Seq Technology Applications (1)

Differential gene expression analysis

- Healthy vs. Diseased
- Time course experiments
- Different genotypes

Transcriptional profiling

- Tissue-specific expression

Novel gene identification

- Transcriptome assembly

mRNA-Seq Technology Applications (2)

Identification of splice variants :

- analysis of exon borders,
- patterns of alternative splicing and the study of protein isoforms.

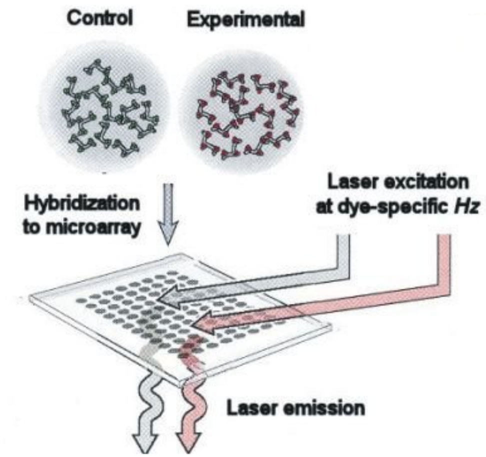
SNP finding

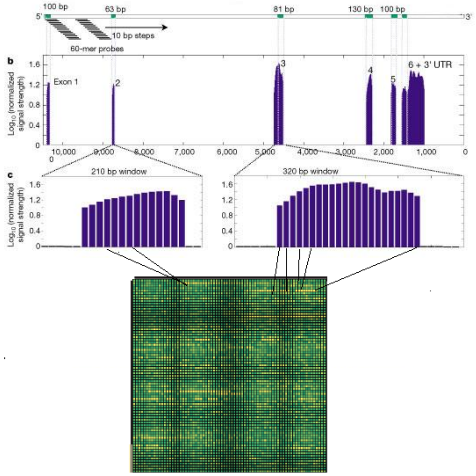
RNA editing

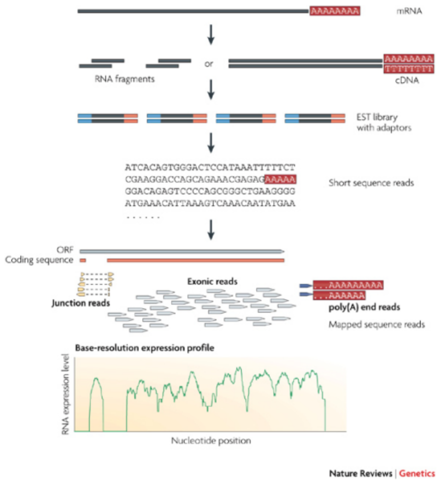
Discovery of "small" RNA ("small RNAs" snRNA, snoRNA, siRNA, miRNA, piRNA ("Piwi-interacting RNAs"), ...) of small size (20-30 nucleotides) and prediction of their secondary structures

The evolution of transcriptomics

Hybridization-based


Control
Experimental
Hybridization to microarray
Laser excitation at dye-specific *Hz*
Laser emission





Nature Reviews | Genetics

1995 P. Brown, et. al. Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

RNA-seq is still a technology under active development

RNAseq vs microarray

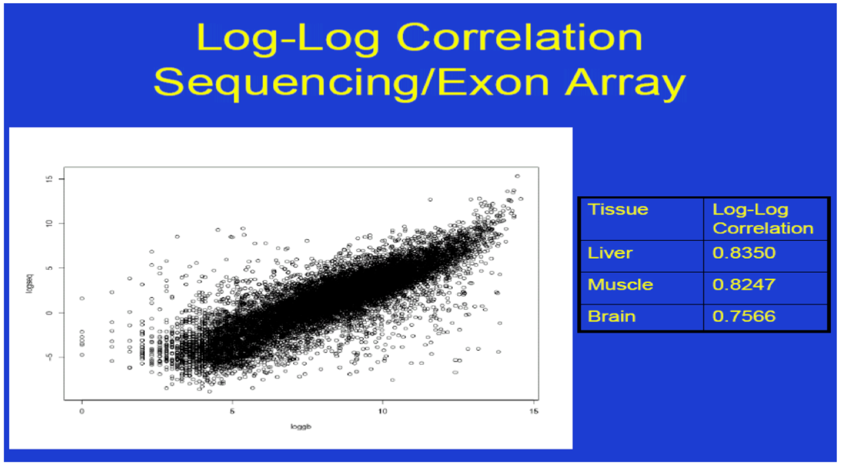
RNAseq vs microarray

- **Microarrays** always have a fixed number of fluorescent probes and therefore have a constant amount of data per run (probe can saturate or fall to background level, however)

- **RNA seq** : Digital data in the form of aligned read-counts but the total amount of sequence can vary significantly both between runs and between genes within a given run

RNAseq vs microarray

- RNAseq sensitivity 10 to 100 order of magnitude higher than microarrays. it allows a very wide dynamic range : detection of rare transcripts
- RNAseq allows **de novo** gene expression analysis
- Good correlation RNAseq vs. Micro-array

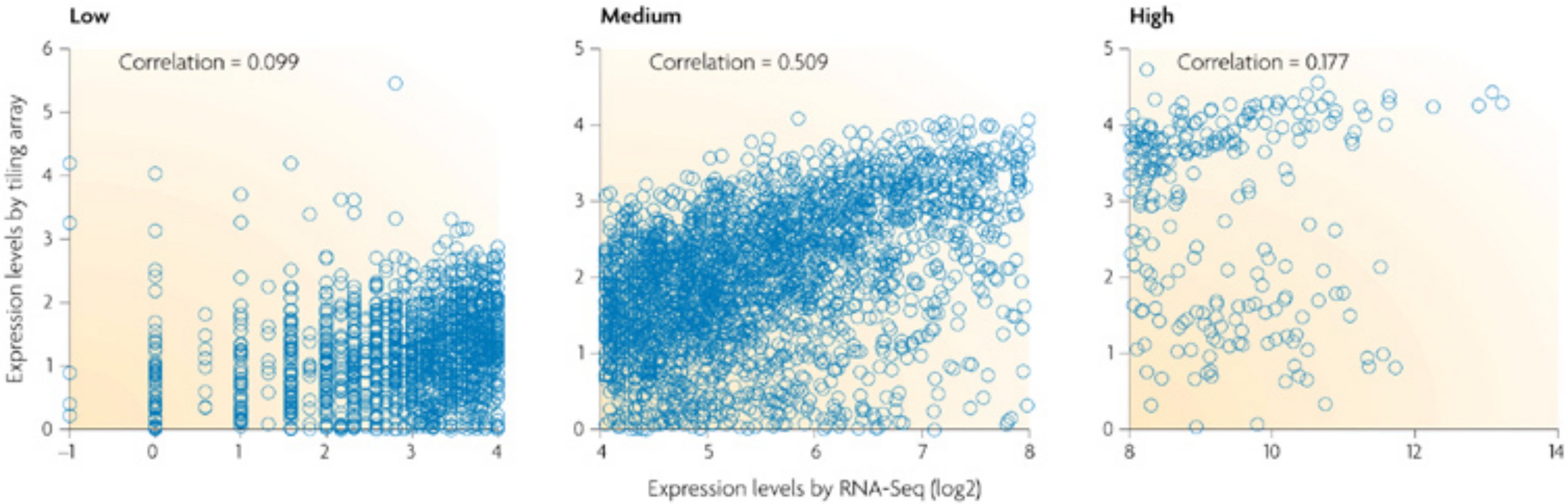


Log-Log Correlation
Sequencing/Exon Array

Tissue	Log-Log Correlation
Liver	0.8350
Muscle	0.8247
Brain	0.7566

RNAseq vs microarray

RNA-seq and microarray agree fairly well only for genes with medium levels of expression

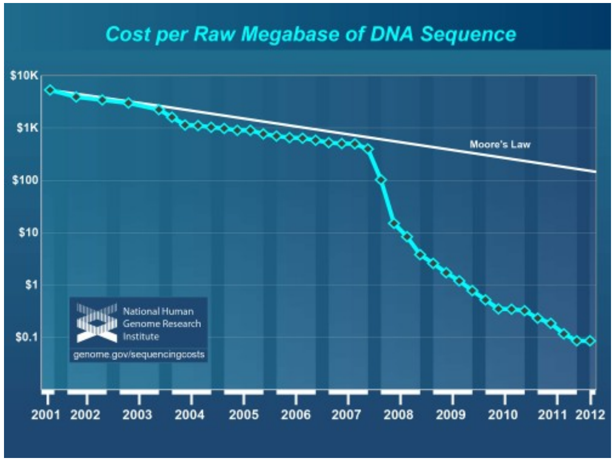


Nature Reviews | Genetics

Saccharomyces cerevisiae cells grown in nutrient-rich media. Correlation is very low for genes with either low or high expression levels.

RNAseq vs microarray

- RNAseq cost decrease



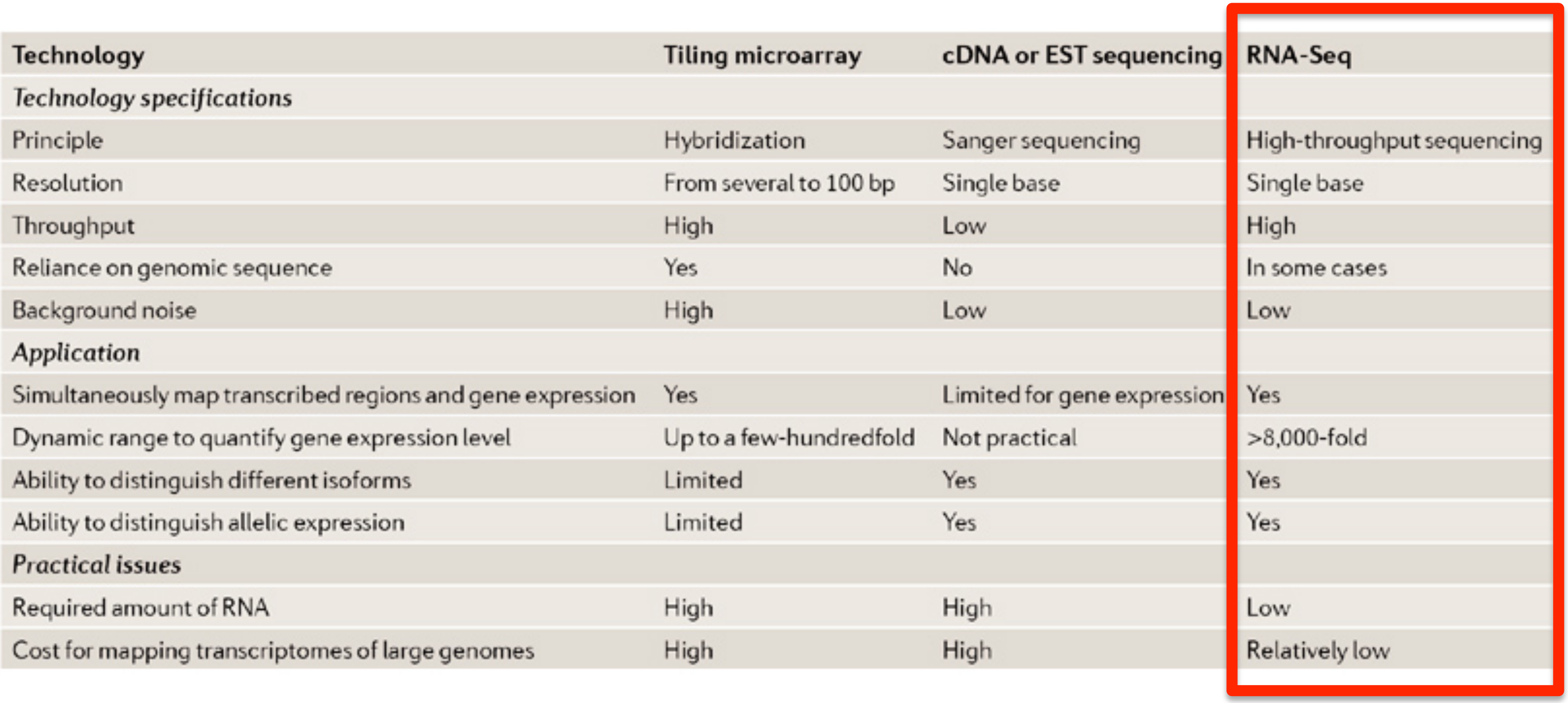
For example, using Rapid Run mode on the HiSeq 2500, one can run as many as 24 samples at 25 million* paired-end 75 basepair reads) (50 million total reads) per sample in less than 24 hours and in a single run.

(75 bp in 16h and 150 bp in 40h).

(*)Depth and format at which published studies have reported the detection of novel features such as gene fusions and alternative transcripts

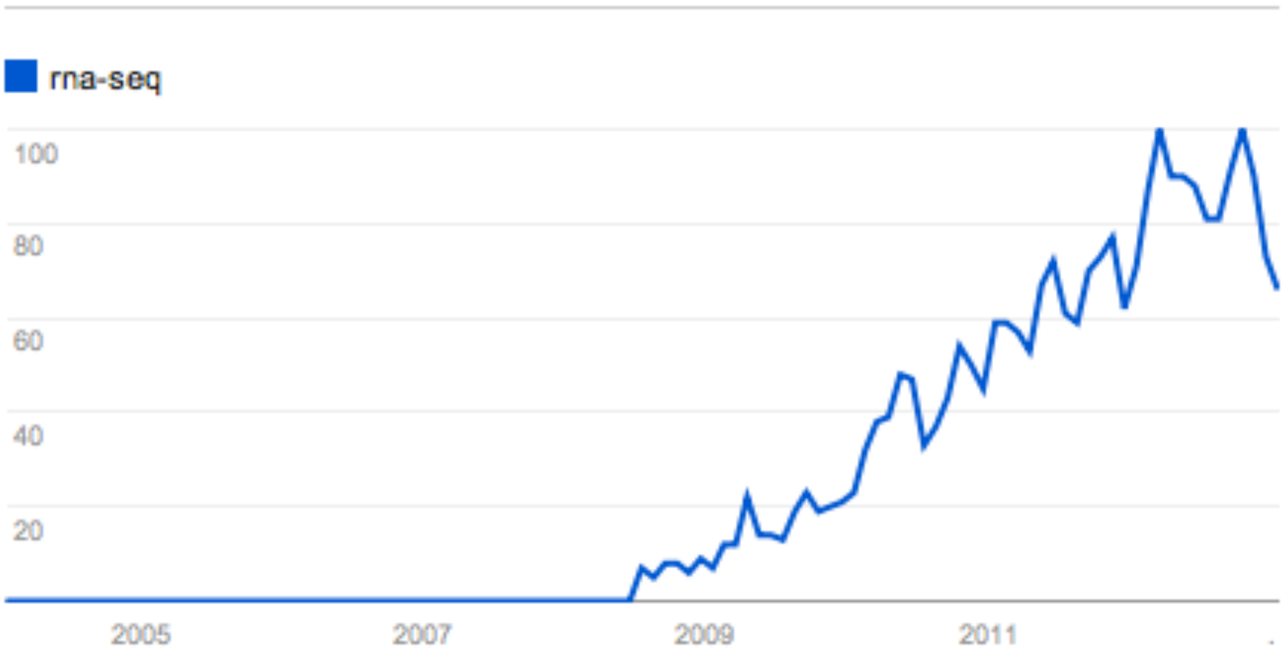
Transcriptomics methods

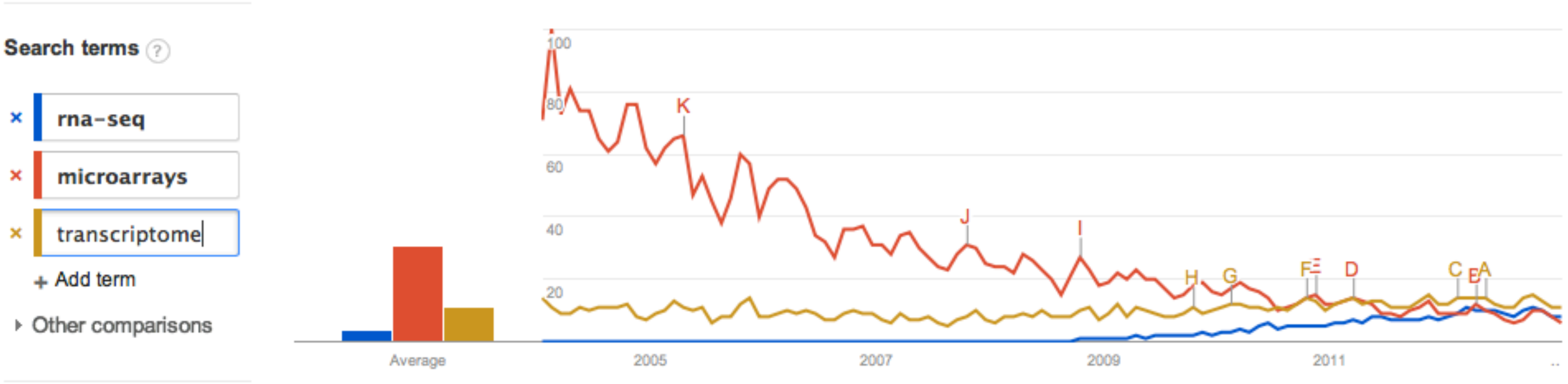
Advantages of RNA-Seq compared with other transcriptomics methods



Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

The evolution of transcriptomics





Search terms

× rna-seq

× microarrays

× transcriptome

The evolution of transcriptomics

RNA-Seq is Hot!

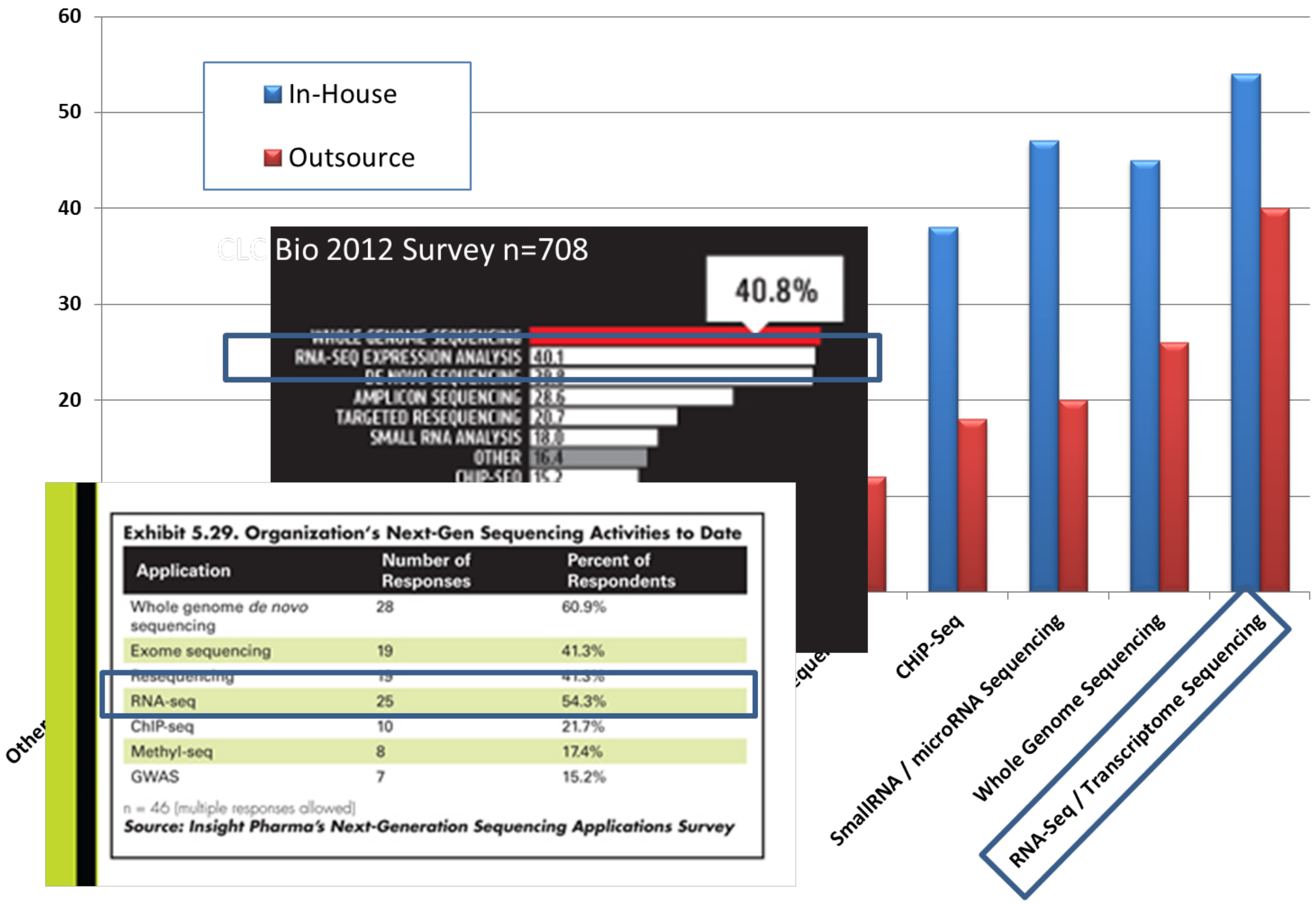


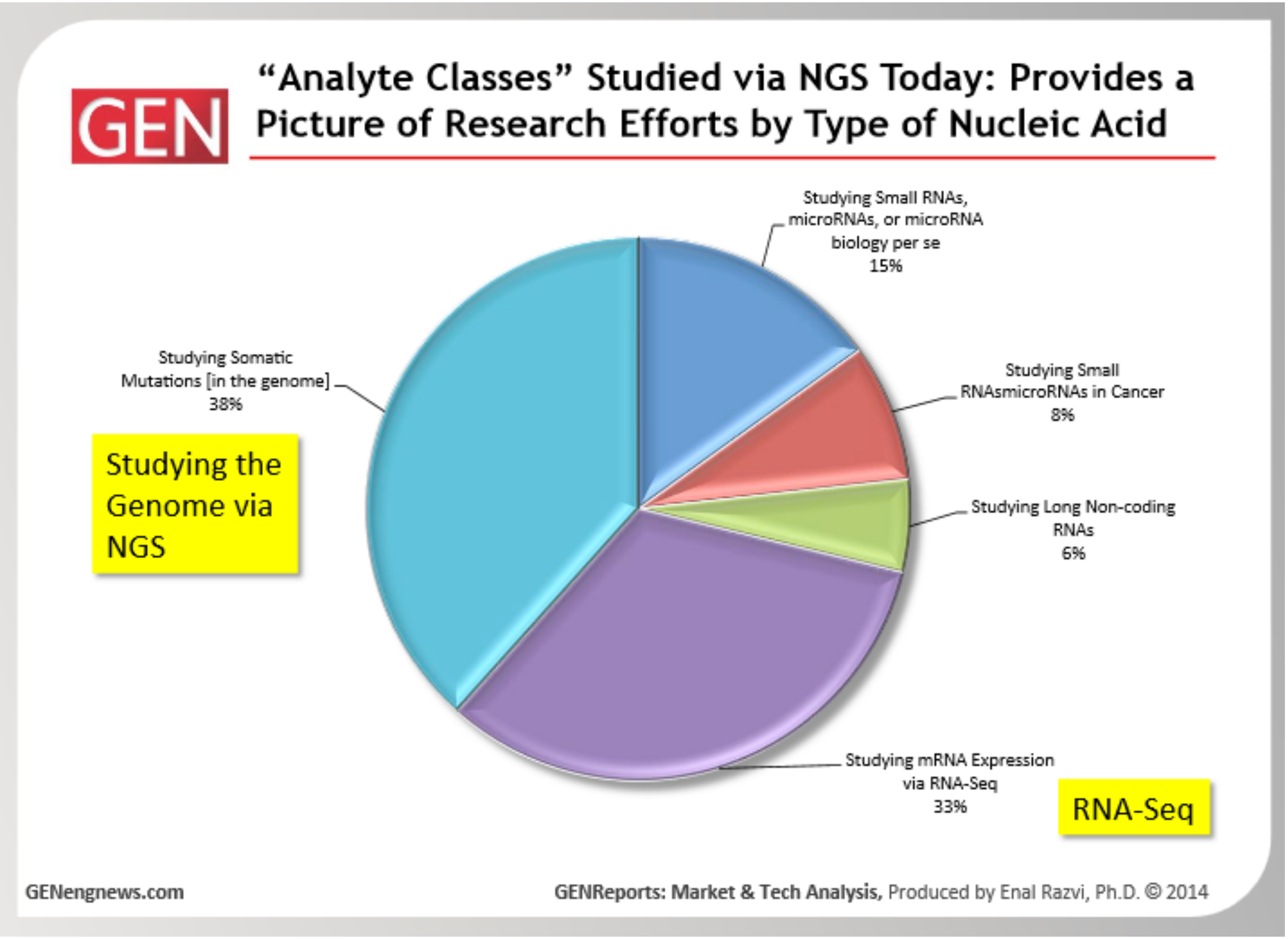
Exhibit 5.29. Organization's Next-Gen Sequencing Activities to Date

Application	Number of Responses	Percent of Respondents
Whole genome *de novo* sequencing	28	60.9%
Exome sequencing	19	41.3%
Resequencing	19	41.3%
RNA-seq	25	54.3%
ChIP-seq	10	21.7%
Methyl-seq	8	17.4%
GWAS	7	15.2%

n = 46 (multiple responses allowed)

Source: Insight Pharma's Next-Generation Sequencing Applications Survey

The evolution of transcriptomics



GEN “Analyte Classes” Studied via NGS Today: Provides a Picture of Research Efforts by Type of Nucleic Acid

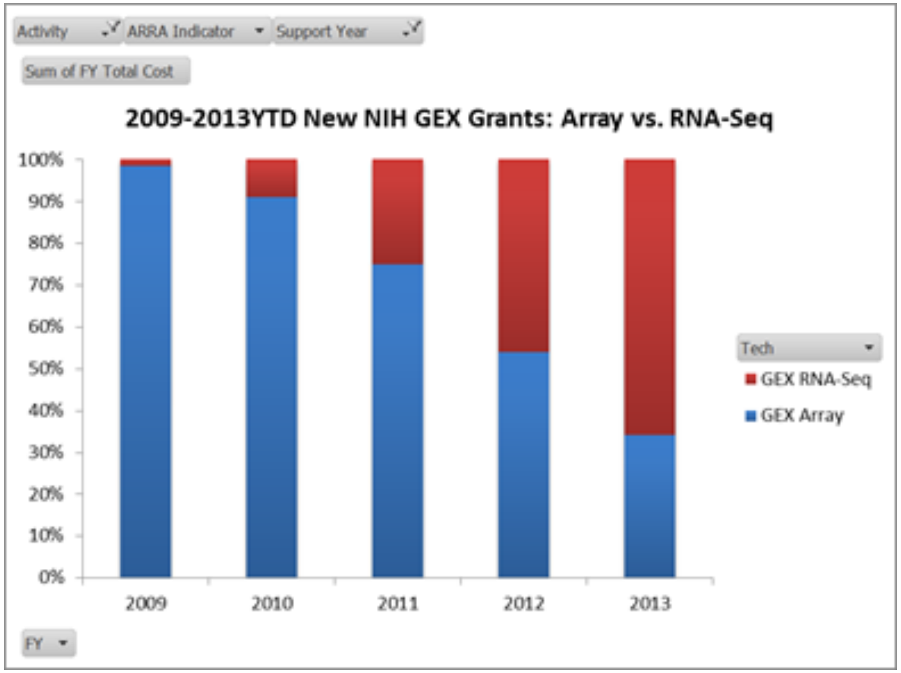
Studying the Genome via NGS

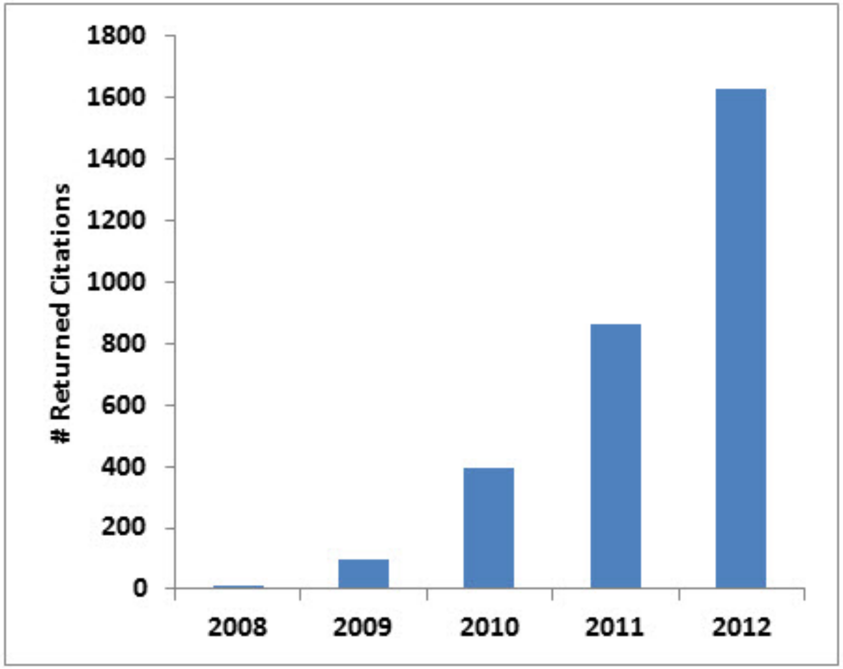
RNA-Seq

GENengnews.com

GENReports: Market & Tech Analysis, Produced by Enal Razvi, Ph.D. © 2014

The evolution of transcriptomics





Introduction

EXPERIMENT DESIGN

Experiment design

What is RNAseq experiment design ?

- Answer to a clear biological question
- Take in account the indentified variation factor , the material and money constrains
- Plan the bioinformatic and biostatistics analysis
- Follow the R. A. Fisher (1935) principles adapted to RNA-seq :

- Repeats
- Shuffeling repartion
- Bloc constitution

- Biological repetitions
- Sequencing depth
- Multiplexing

Experiment design

Sources of variance

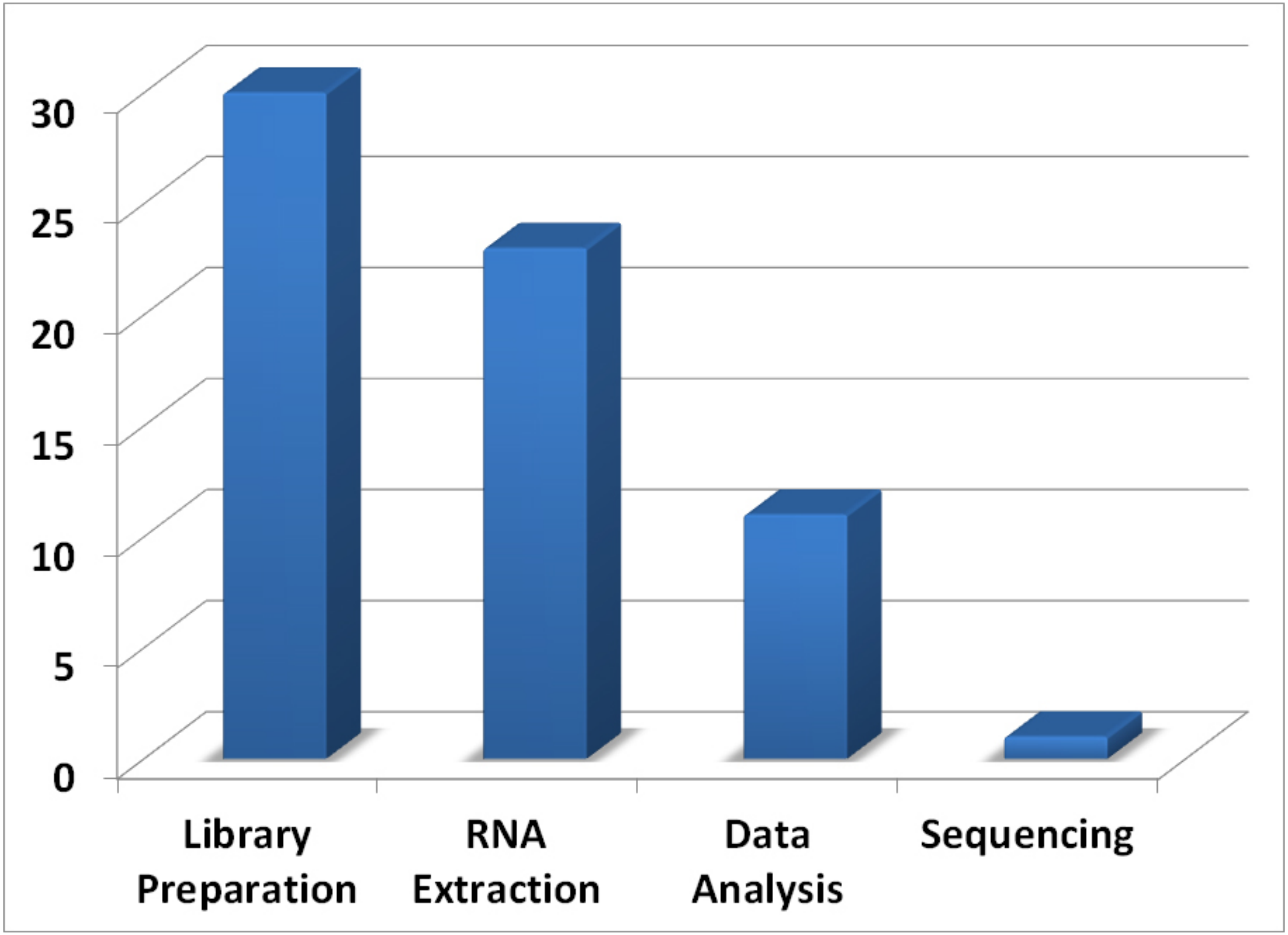
- **Sampling (fragment) variance:** Even though NGS is capable of producing millions of sequence reads, these represent only a small fraction of the nucleic acid that is actually present in the library. But also **subject sampling** (for a larger population) and **RNA sampling** (from different cells or tissues)

- **Technical variance:**
  - RNA extracted : Quality and Quantity
  - Library preparation (fragmentation, enrichment, purification, amplification, GC %, fragment orientation)
  - NGS sequencing procedures (multiplexing- sequencing kit)

- **Biological variance:** The nascent variance that is present within a treatment or control group.

- **Variance effect :** line < run < Library preparation << biological variance

Poll : Which step of an RNA-Seq experiment is the greatest source of technical variability?



Source RNASEQ Blog june 2014

Experiment design

1) how deep does one need to sequence?

2) how many biological replicates are necessary to observe a significant change in expression?

Read coverage

- Coverage = (Total Sequence)/Transcriptome Size

- Transcriptome = ~500K transcripts
  - Contamination
  - Mitochondrial, etc…
- Average Transcript length = 1000bp
- **Transcriptome size = 500K x 1kb = 500Mb**
- Total Sequence = 30M reads x 100bp x 2 = 6Gb
- Coverage = 6Gb/500Mb =12X

How deep is enought

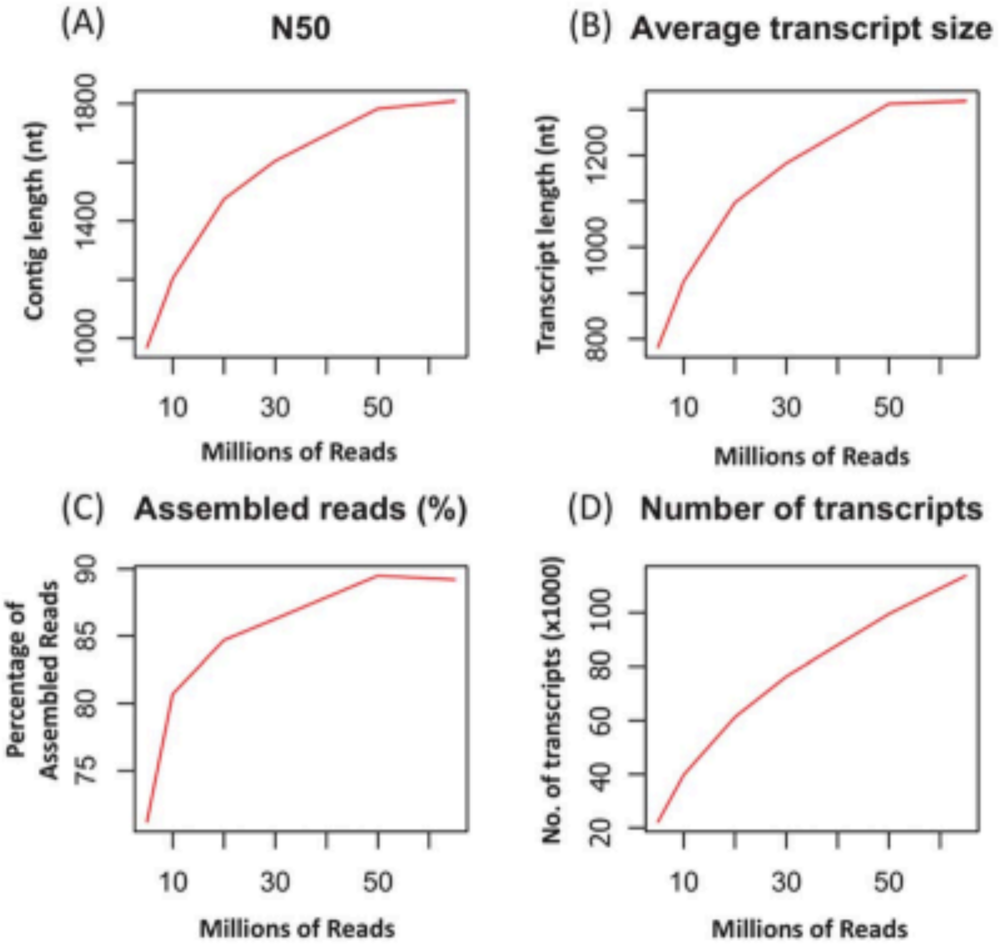


Fig. 6 Effect of sequencing depth on a transcriptome assembly. Four Paired-End assemblies using 5, 10, 20, 30, 50 and 65 million reads were generated using Oases.[37] The N50 contig size (A), average transcript size (B), percentage of reads used in the assembly (C), and number of transcripts (D) *versus* number of reads used in the assembly are shown.

Góngora-Castillo, E., & Buell, C. R. (2013). Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Natural Product Reports. doi: 10.1039/c3np20099j

How deep is enought

Human

Majority of expressed genes and AS events can be detected with **modest sequencing depths (~100 M filtered reads**), the estimated gene expression levels and exon/intron inclusion levels were less accurate

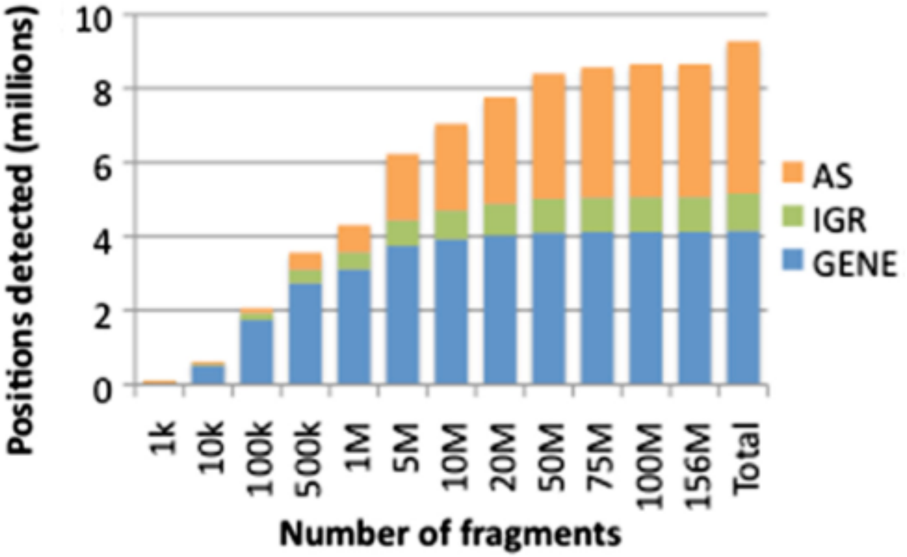
- To detect expressed genes and AS events, ~100 to 150 million (M) filtered reads were needed.
- For a DE analysis and detect 80% of events, ~300 M filtered reads were needed
- For detecting Differential AS and detect 80% of events, at least 400 M filtered reads were necessary

Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose.
Yichuan Liu et al., 2013.

How deep is enought ?

Bacteria

E. Coli : 5000 genes
intergenic (IGR)
antisense to ORFs or ncRNAs (AS)



« A sequencing depth of **5-10 million** non- rRNA fragments enables profiling of the vast majority of transcriptional activity in diverse species grown under di- verse culture conditions. »

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., & Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC genomics, 13, 734. doi:10.1186/1471-2164-13-734

How deep is enought ?

Depends on the purpose of the experiment and the nature of the samples (ENCODE).

- 100M of reads is sufficient to detect 90% of the transcripts and 81% of the genes of the human transcriptome. (Tung et al. 2011)
- 20M reads (75bp) is sufficient to detect transcripts expressed at a medium or low level in the chicken. (Wang et al. 2011)
- 10 M of reads allow 90% of transcripts (human, zebrafish) to be covered by an average of 10 reads. (Hart et al. 2013)

Sample size

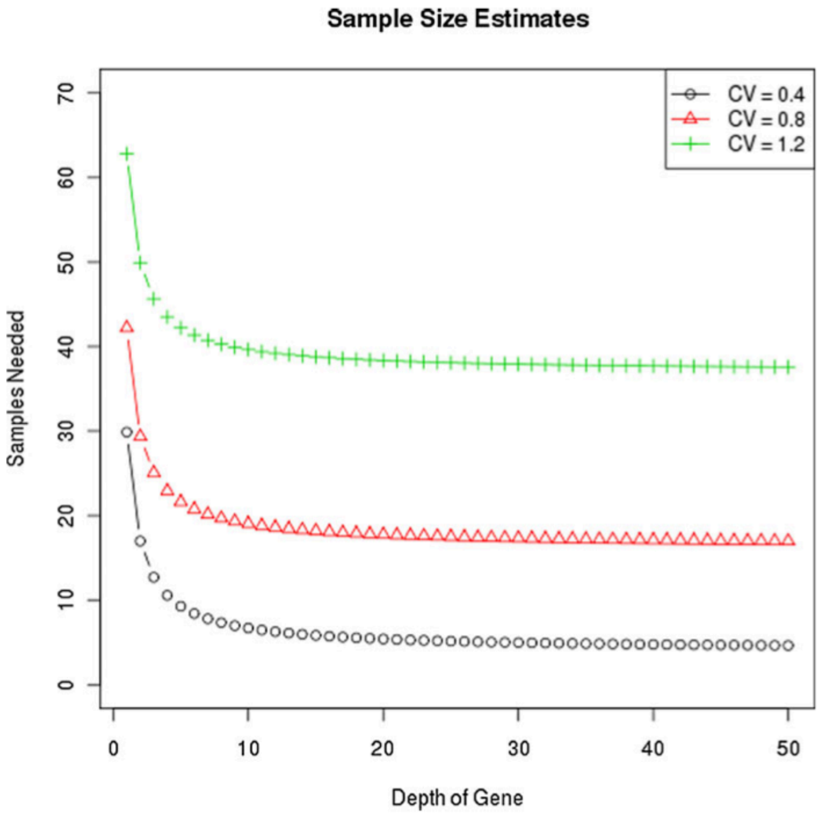


FIG. 3. Sample size estimates for identifying a two-fold change vary by CV, not coverage. The *y-axis* is the sample size needed to detect a two-fold difference in expression with 80% power, and 5% type 1 error, given at alpha = 0.01 for three different biological CV's and sequencing depths.

Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A., & Kocher, J.-P. (2013). Calculating Sample Size Estimates for RNA Sequencing Data. Journal of Computational Biology. doi:10.1089/cmb.2012.0283

Sample size

Why increase the number of biological replicates?

- Generalizing the results to the population

- Estimate more accurately the variation of each transcript individually (Hart et al. 2013)

- Improve the detection of differential transcripts and rate control false positives: TRUE from 3 (Sonenson et al, 2013, Robles et al 2012.)

Sample size vs depth

It’s up to you! (Haas et al., 2012, Liu Y. et al 2013)

- **Detection of differential transcripts:**
  - (+) biological replicates
- **Construction / transcriptome annotation:**
  - (+) depth & (+) conditions
- **Search variants:**
  - (+) biological replicates & (+) depth

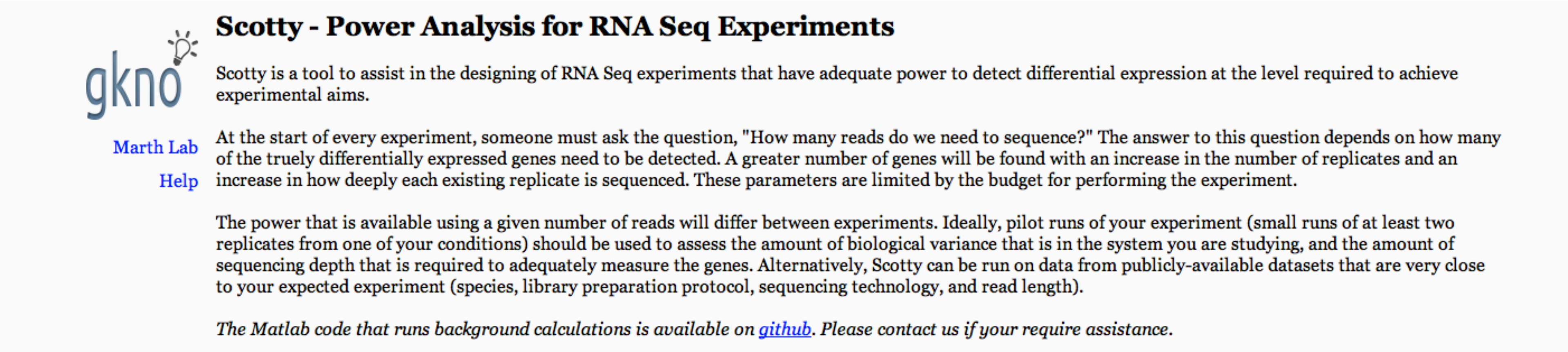
Deciphering Sample Heterogeneity

- “RNA prepared from heterogeneous tissue samples might contain only a fraction of the total cell subpopulation of interest. Consequently, the expression signal of any gene detected directly from a complex sample is a convolution of expressions of all present cell types”

Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD. PLoS One. 2011; 6(11):e27156. Epub 2011 Nov 16.

- DeConRNAseq : R package

http://euler.bc.edu/marthlab/scotty/scotty.php



Scotty - Power Analysis for RNA Seq Experiments

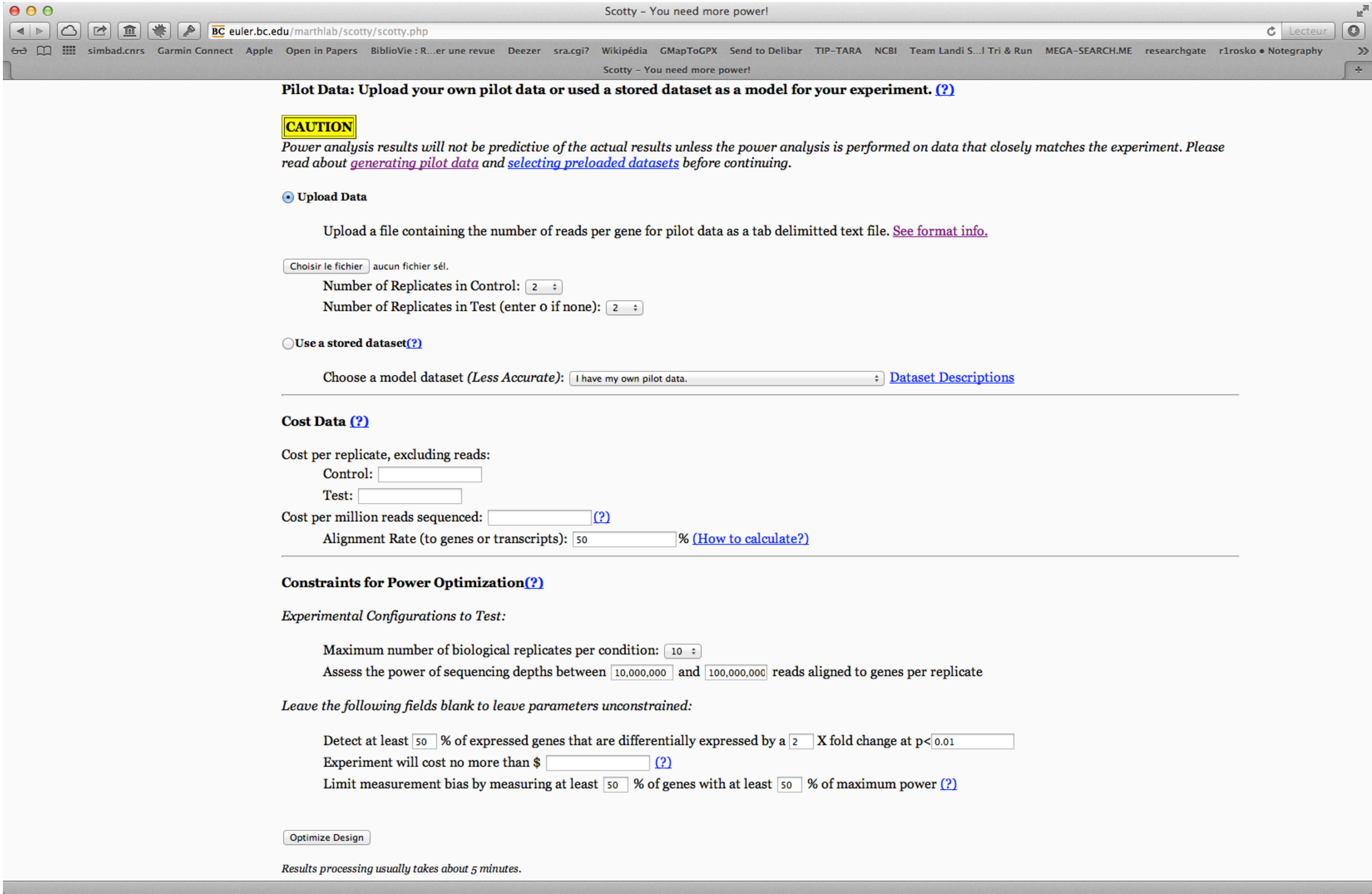
Scotty is a tool to assist in the designing of RNA Seq experiments that have adequate power to detect differential expression at the level required to achieve experimental aims.

At the start of every experiment, someone must ask the question, "How many reads do we need to sequence?" The answer to this question depends on how many of the truely differentially expressed genes need to be detected. A greater number of genes will be found with an increase in the number of replicates and an increase in how deeply each existing replicate is sequenced. These parameters are limited by the budget for performing the experiment.

The power that is available using a given number of reads will differ between experiments. Ideally, pilot runs of your experiment (small runs of at least two replicates from one of your conditions) should be used to assess the amount of biological variance that is in the system you are studying, and the amount of sequencing depth that is required to adequately measure the genes. Alternatively, Scotty can be run on data from publicly-available datasets that are very close to your expected experiment (species, library preparation protocol, sequencing technology, and read length).

The Matlab code that runs background calculations is available on github. Please contact us if your require assistance.

http://euler.bc.edu/marthlab/scotty/scotty.php


Scotty – You need more power!
euler.bc.edu/marthlab/scotty/scotty.php
Lecteur
simbad.cnrs Garmin Connect Apple Open in Papers BiblioVie : R...er une revue Deezer sra.cgi? Wikipédia GMapToGPX Send to Delibar TIP-TARA NCBI Team Landi S...l Tri & Run MEGA-SEARCH.ME researchgate r1rosko • Notegraphy
Scotty – You need more power!

Pilot Data: Upload your own pilot data or used a stored dataset as a model for your experiment. (?)

CAUTION

Power analysis results will not be predictive of the actual results unless the power analysis is performed on data that closely matches the experiment. Please read about generating pilot data and selecting preloaded datasets before continuing.

● **Upload Data**

Upload a file containing the number of reads per gene for pilot data as a tab delimitted text file. See format info.

Choisir le fichier aucun fichier sél.

Number of Replicates in Control: 2

Number of Replicates in Test (enter 0 if none): 2

○ **Use a stored dataset**(?)

Choose a model dataset *(Less Accurate)*: I have my own pilot data. Dataset Descriptions

Cost Data (?)

Cost per replicate, excluding reads:

Control:

Test:

Cost per million reads sequenced: (?)

Alignment Rate (to genes or transcripts): 50 % (How to calculate?)

Constraints for Power Optimization(?)

Experimental Configurations to Test:

Maximum number of biological replicates per condition: 10

Assess the power of sequencing depths between 10,000,000 and 100,000,000 reads aligned to genes per replicate

Leave the following fields blank to leave parameters unconstrained:

Detect at least 50 % of expressed genes that are differentially expressed by a 2 X fold change at p< 0.01

Experiment will cost no more than $ (?)

Limit measurement bias by measuring at least 50 % of genes with at least 50 % of maximum power (?)

Optimize Design

Results processing usually takes about 5 minutes.

Experiment design : PEPI IBIS recommandations

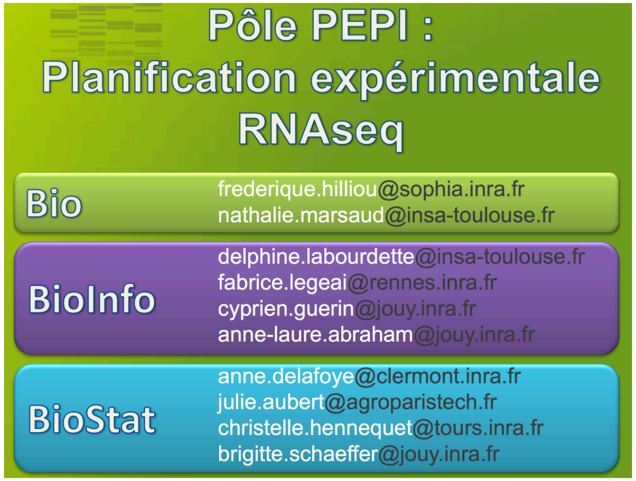
- Clarify the biological question:

  RNA-seq can answer a lot of questions, but all questions will not reply with a single RNA-seq experience level.

- Biologist / Bioinformatician / Statistician Trio required at start construction of the project and in discussions

- Make biological replicates!
- Think multiplexing! € ➘
- Repeats/ Depth depends of the biological question

- Your budget should include :
  - Extraction of biological data,
  - Sequencing data storage,
  - Bioinformatics analyzes, statistical analyzes



Pôle PEPI :
Planification expérimentale RNAseq

Bio	frederique.hilliou@sophia.inra.fr nathalie.marsaud@insa-toulouse.fr
BioInfo	delphine.labourdette@insa-toulouse.fr fabrice.legeai@rennes.inra.fr cyprien.guerin@jouy.inra.fr anne-laure.abraham@jouy.inra.fr
BioStat	anne.delafoye@clermont.inra.fr julie.aubert@agroparistech.fr christelle.hennequet@tours.inra.fr brigitte.schaeffer@jouy.inra.fr

Experiment design : ENCODE recommandations

- RNA-Seq is not a mature technology.
- Experiments should be performed with **two or more biological replicates**, unless there is a compelling reason why this is impractical or wasteful
- A typical **R2** (Pearson) correlation of gene expression (RPKM) between two biological replicates, for RNAs that are detected in both samples using RPKM or read counts, should be between **0.92 to 0.98**. Experiments with biological correlations that fall below 0.9 should be either be repeated or explained.
- Between **30M and 100M reads** per sample depending on the study.
- **NB.** Guidelines for the information to publish with the data.

http://encodeproject.org/ENCODE/dataStandards.html

A statistical answer : Conclusions

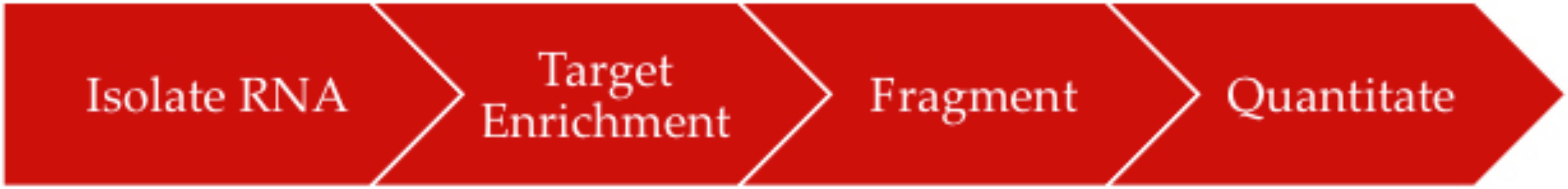
This work quantitatively explores comparisons between contemporary analysis tools and experimental design choices for the detection of differential expression using RNA-Seq. …With regard to testing of various experimental designs, this work strongly suggests **that greater power is gained through the use of biological replicates relative to library (technical) replicates and sequencing depth**. Strikingly, **sequencing depth could be reduced as low as 15% without substantial impacts on false positive or true positive rates.**

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics, 13, 484.

Introduction

RNA EXTRACTION AND LIBRARY CONSTRUCTION BIAIS

RNA extraction



1 2 3 4

- (1) Isolation and purification of RNA typically involves disrupting cells in the presence of detergents and chaotropic agents.

- (1) After homogenization, RNA can be recovered and purified from the total cell lysate using either liquid-liquid partitioning or solid-phase extraction.

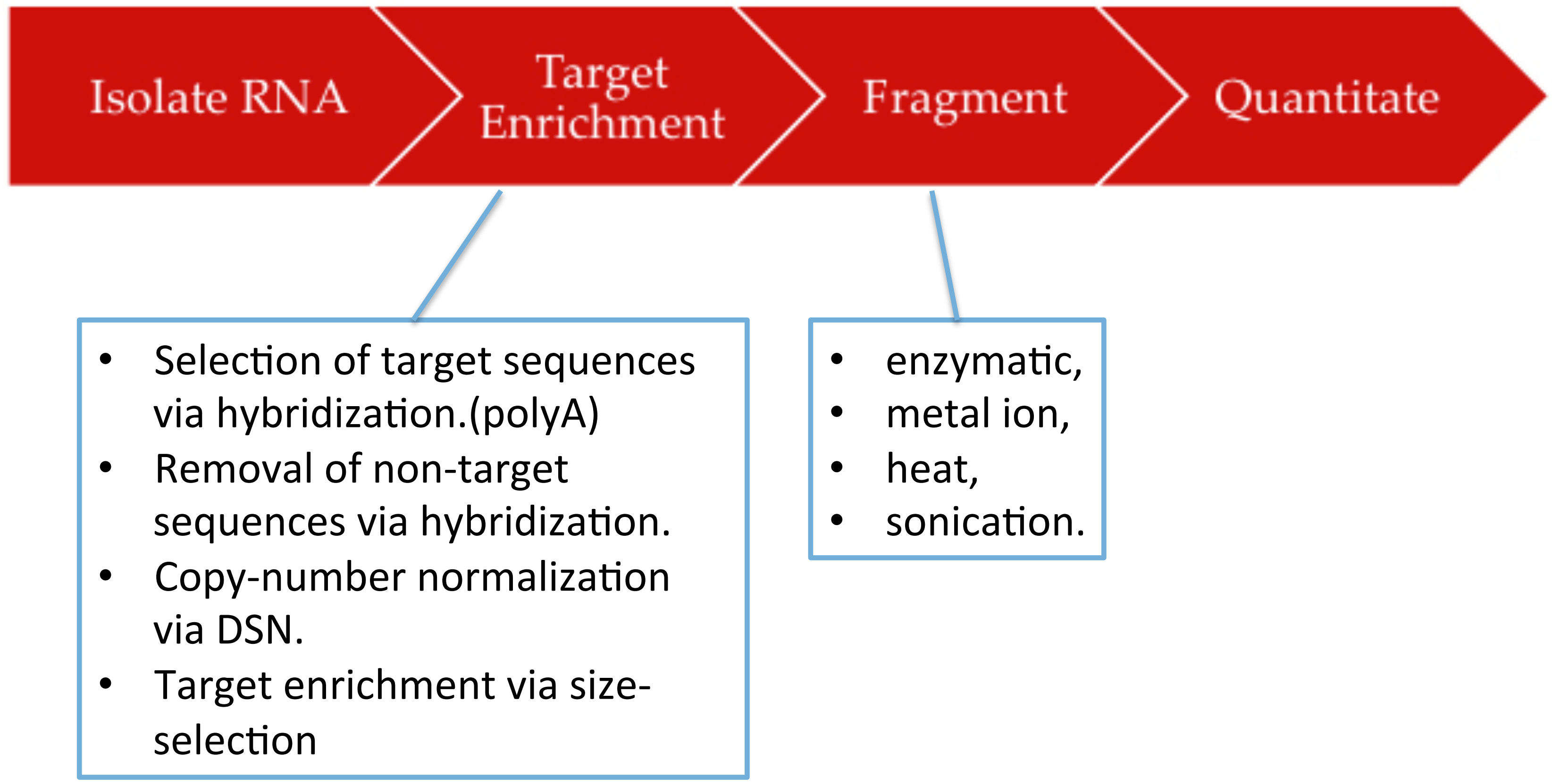
- (2) Typically the total RNA is then enriched for messenger RNA (mRNA). This can be done by either directly selecting mRNA or by selectively removing ribosomal RNA (rRNA).

- (3) To make the RNA suitable for RNA-seq it is typically fragmented

- (4) And then the quality and fragmentation are assessed.

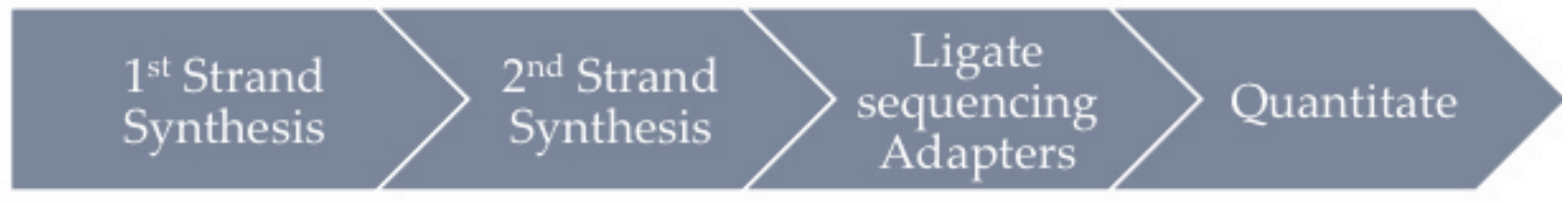
http://rnaseq.uoregon.edu

RNA extraction



- Selection of target sequences via hybridization.(polyA)
- Removal of non-target sequences via hybridization.
- Copy-number normalization via DSN.
- Target enrichment via size-selection

- enzymatic,
- metal ion,
- heat,
- sonication.

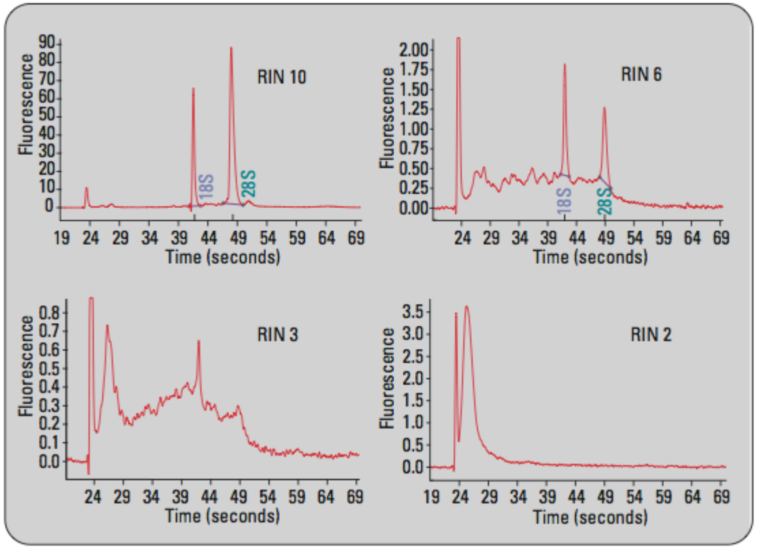


http://rnaseq.uoregon.edu

RIN : RNA Integrity Number

The integrity of RNA is a major concern for gene expression studies :

The RIN algorithm is applied to electrophoretic RNA measurements and based on a combination of different features that contribute information about the RNA integrity to provide a more robust universal measure.



- Values over 8 are good enough for transcriptome analysis (euk).
- Values over 9 for bacterial RNA
- For small RNAseq prefer values above 8.5 to ensure that you are fishing just the physiological RNAs and not degradation products

Degraded RNA samples :
- Over representation of 3'-end fragments of transcripts (poly A targetted)
- Highly fragmented transcriptome -> hundreds of thousands transcripts

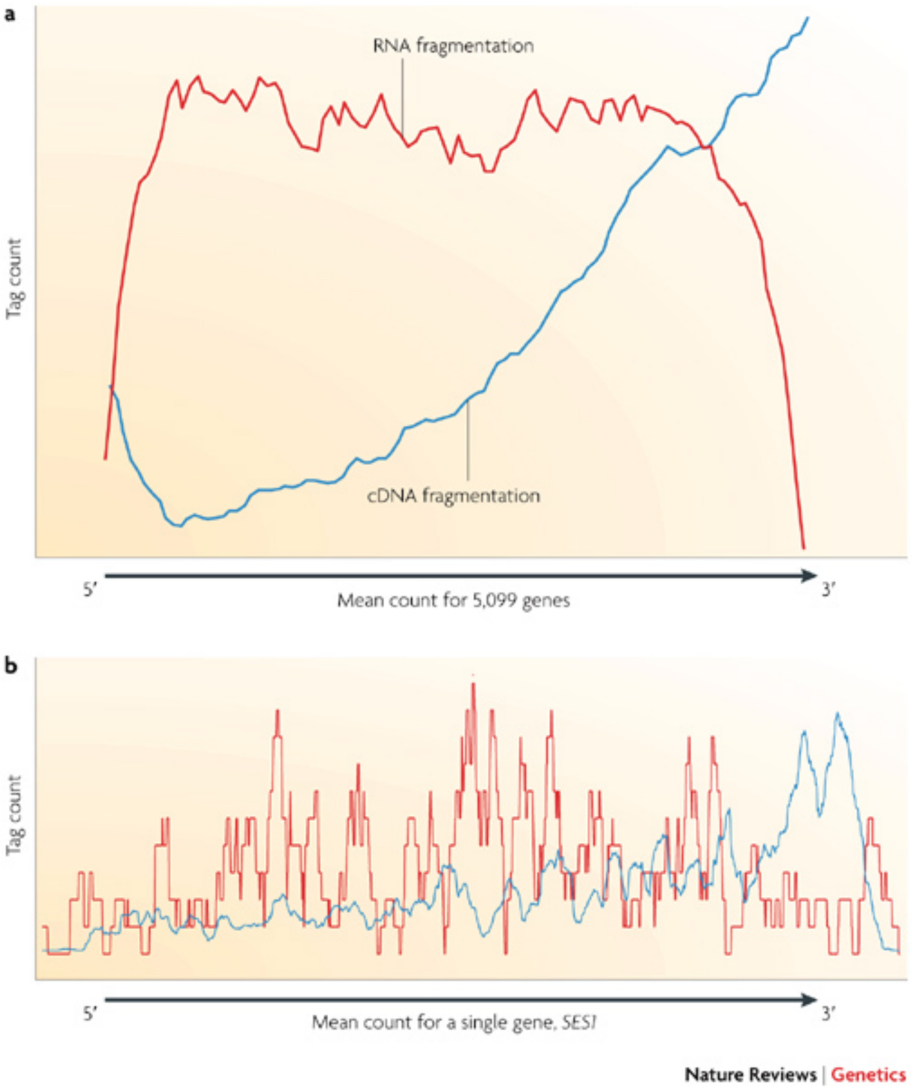
Library construction : fragmentation

- Unlike small RNAs (microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs) and many others), which can be directly sequenced after adaptor ligation, larger RNA molecules must be fragmented into smaller pieces (200–500 bp) to be compatible with most deep-sequencing technologies.

- Common fragmentation methods include RNA fragmentation (RNA hydrolysis or nebulization) or/and cDNA fragmentation (DNase I treatment or sonication).

Each of these methods creates a different bias in the outcome.

Library construction : fragmentation



- Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript.
- RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends.

A specific yeast gene, SES1 (seryl-tRNA synthetase)

Library construction

PCR artefacts

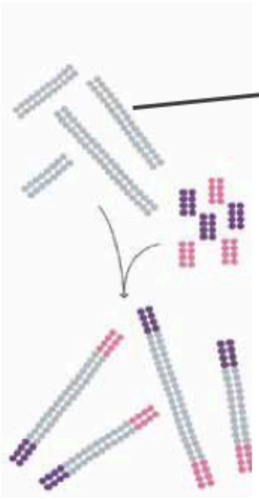
- Many shorts reads that are identical to each other can be obtained from cDNA libraries that have been amplified. These could be a genuine reflection of abundant RNA species, or they could be PCR artefacts.
- Use replicates

Whether or not to prepare strand-specific libraries

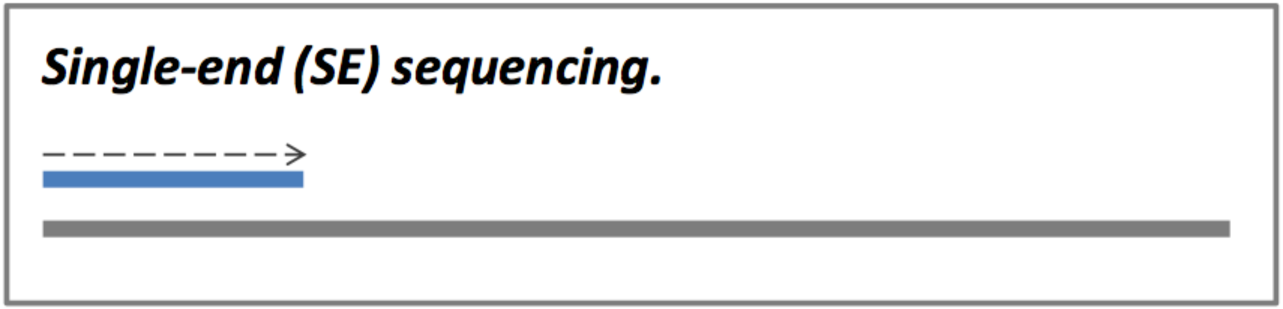
- Strand-specific libraries are valuable for transcriptome annotation, especially for regions with overlapping transcription from opposite direction
- strand-specific libraries are currently laborious to produce because they require many steps or direct RNA–RNA ligation, which is inefficient

Sequencing terminology

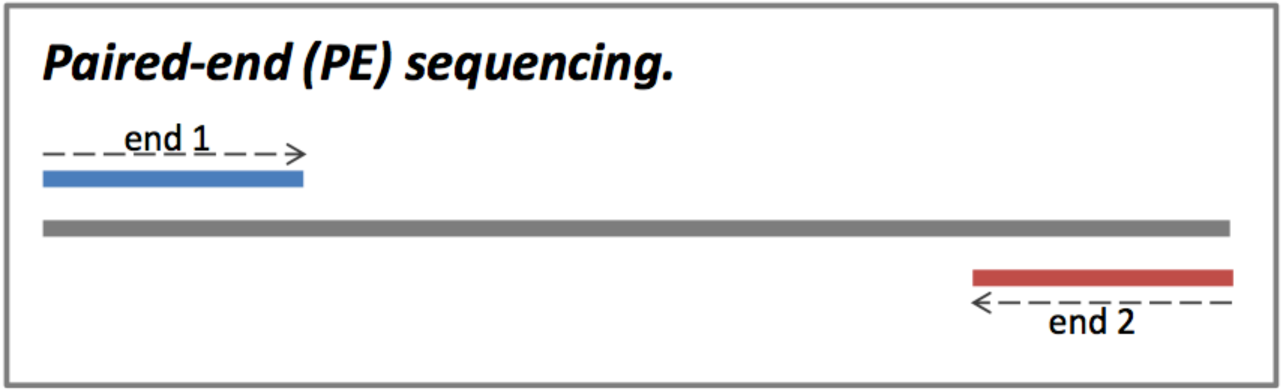
Prepare sequence fragments. Ligate adapters.



DNA/RNA fragment of known length



Single-end (SE) sequencing.



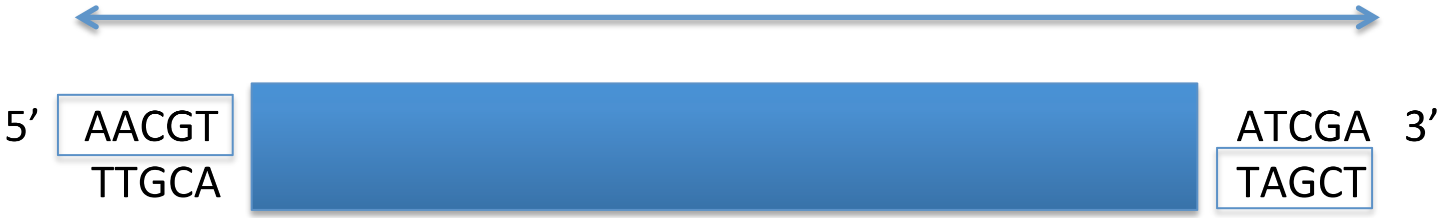
Paired-end (PE) sequencing.

Sequencing terminology

Shotgun fragments

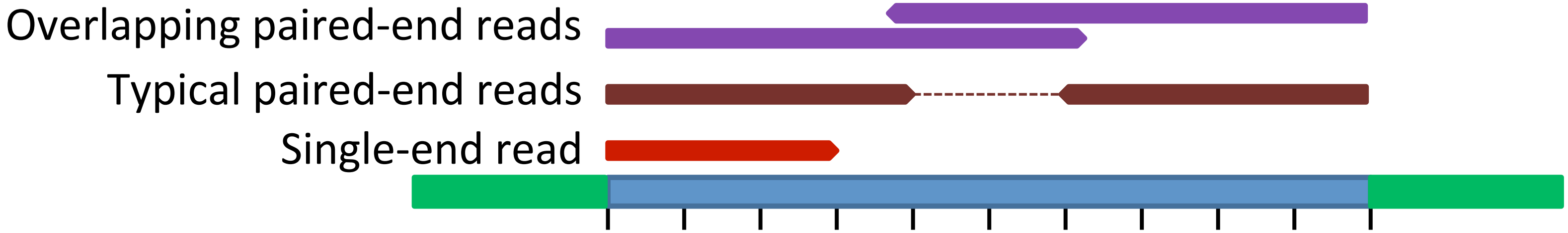
Fragments vs. Reads

Insert size



5’ AACGT
TTGCA

ATCGA 3’
TAGCT

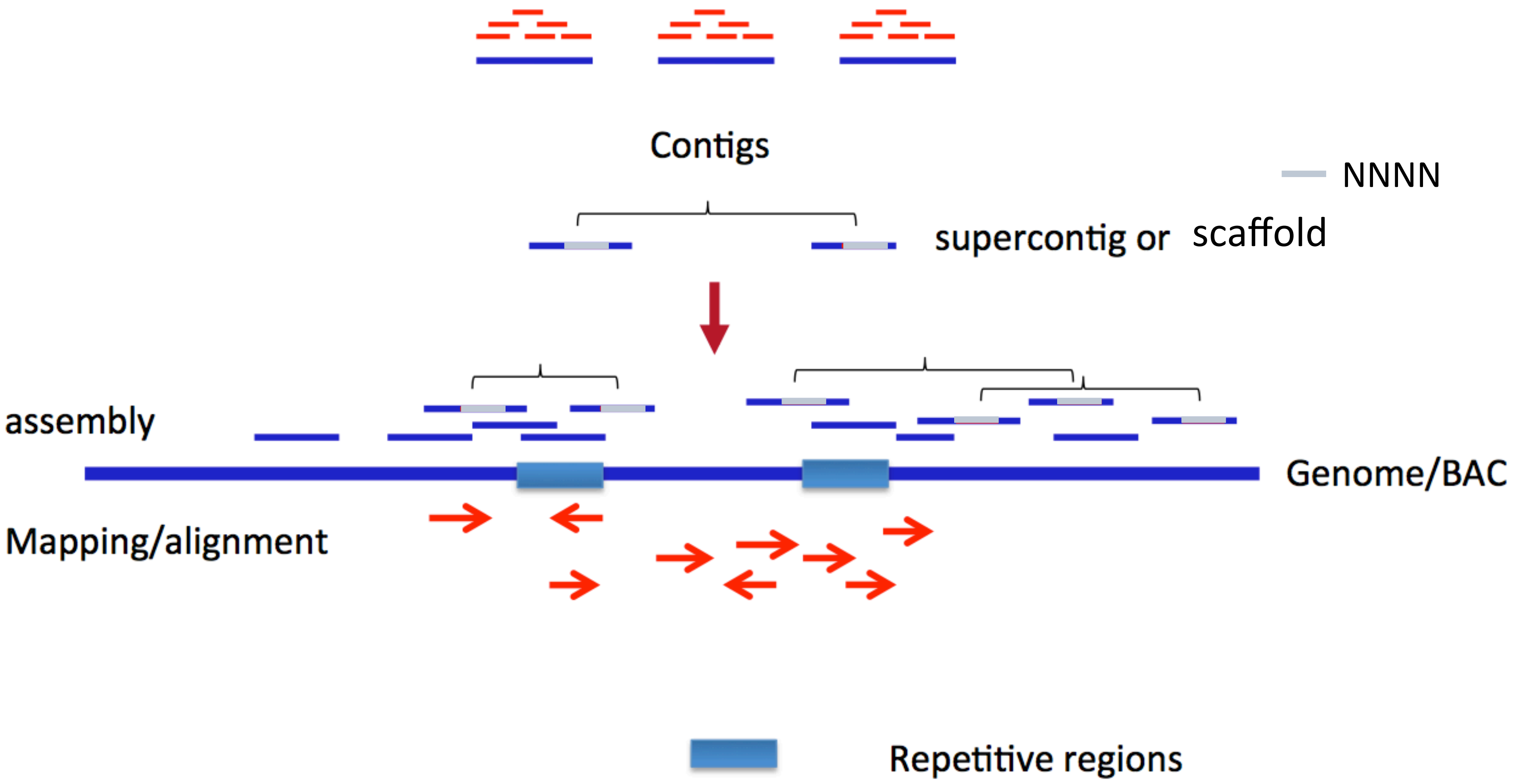


Overlapping paired-end reads

Typical paired-end reads

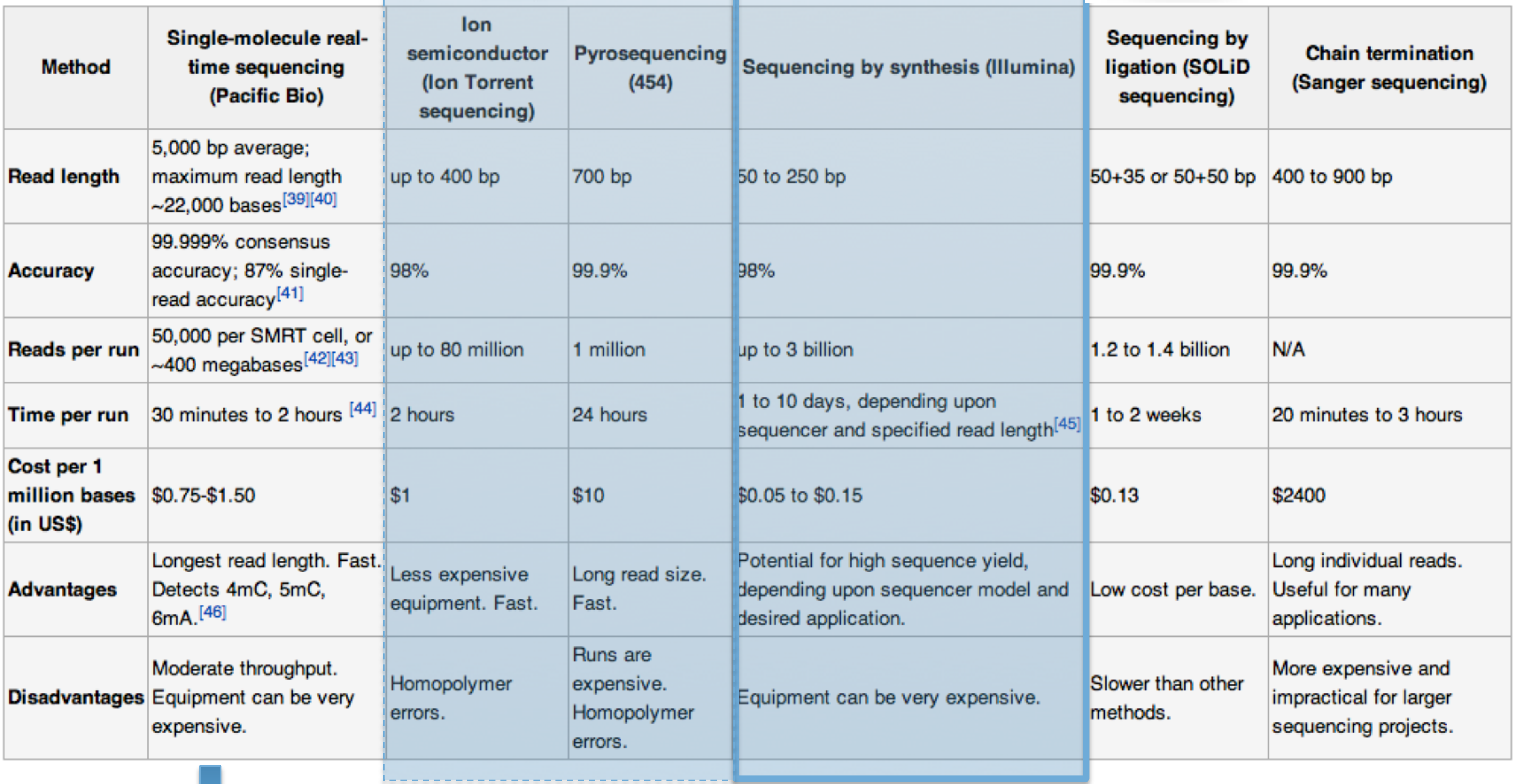
Single-end read

Sequencing terminology



CNRS UPMC
Station Biologique
Roscoff

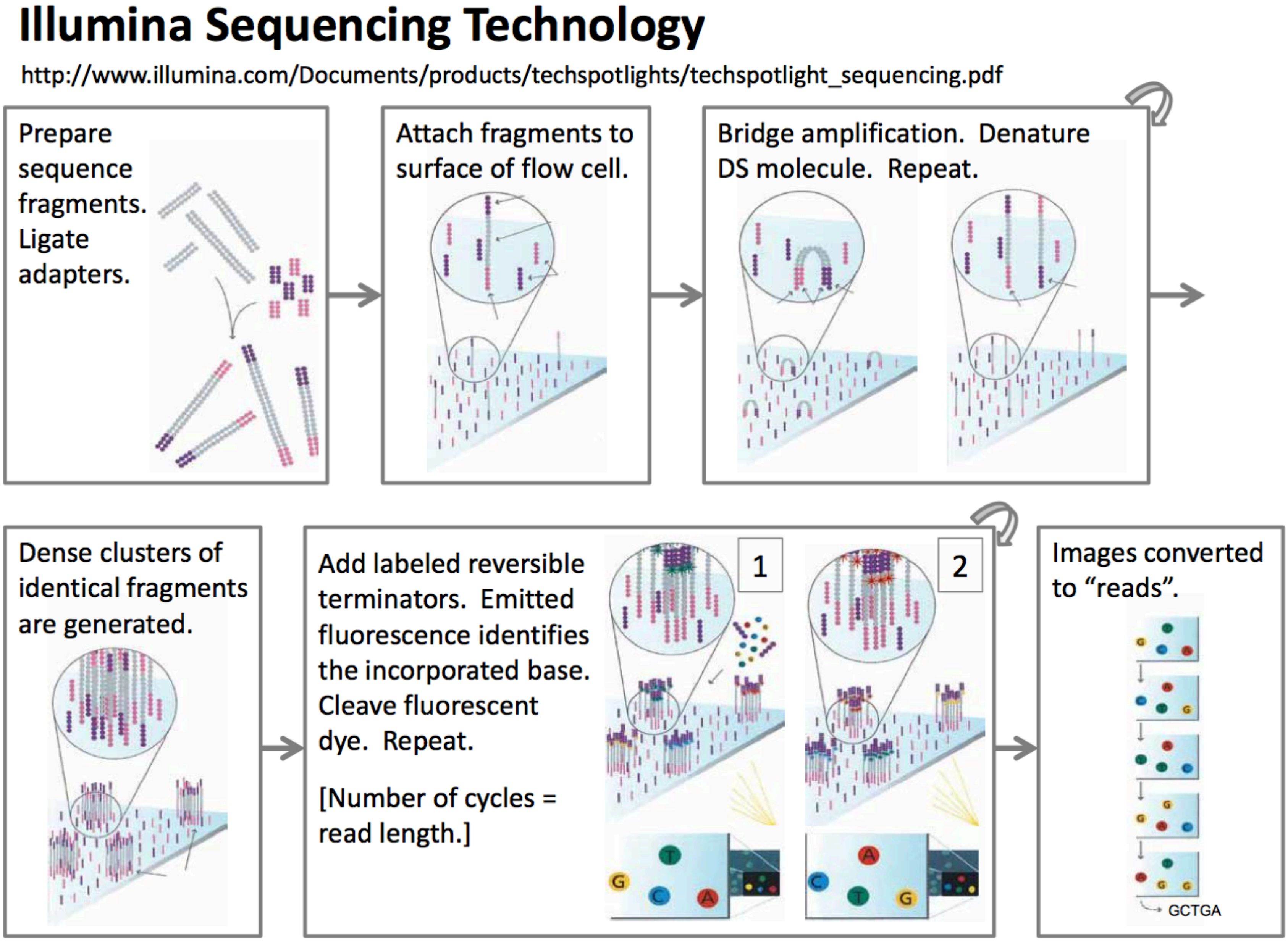
NGS for RNAseq



Method	Single-molecule real-time sequencing (Pacific Bio)	Ion semiconductor (Ion Torrent sequencing)	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	5,000 bp average; maximum read length ~22,000 bases[39][40]	up to 400 bp	700 bp	50 to 250 bp	50+35 or 50+50 bp	400 to 900 bp
Accuracy	99.999% consensus accuracy; 87% single-read accuracy[41]	98%	99.9%	98%	99.9%	99.9%
Reads per run	50,000 per SMRT cell, or ~400 megabases[42][43]	up to 80 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours [44]	2 hours	24 hours	1 to 10 days, depending upon sequencer and specified read length[45]	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bases (in US$)	$0.75-$1.50	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantages	Longest read length. Fast. Detects 4mC, 5mC, 6mA.[46]	Less expensive equipment. Fast.	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model and desired application.	Low cost per base.	Long individual reads. Useful for many applications.
Disadvantages	Moderate throughput. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive.	Slower than other methods.	More expensive and impractical for larger sequencing projects.

Iso-Seq Method: Full-length transcript sequencing (2014)

Illumina sequencing technology



Illumina Sequencing Technology

http://www.illumina.com/Documents/products/techspotlights/techspotlight_sequencing.pdf

Prepare sequence fragments. Ligate adapters.

Attach fragments to surface of flow cell.

Bridge amplification. Denature DS molecule. Repeat.

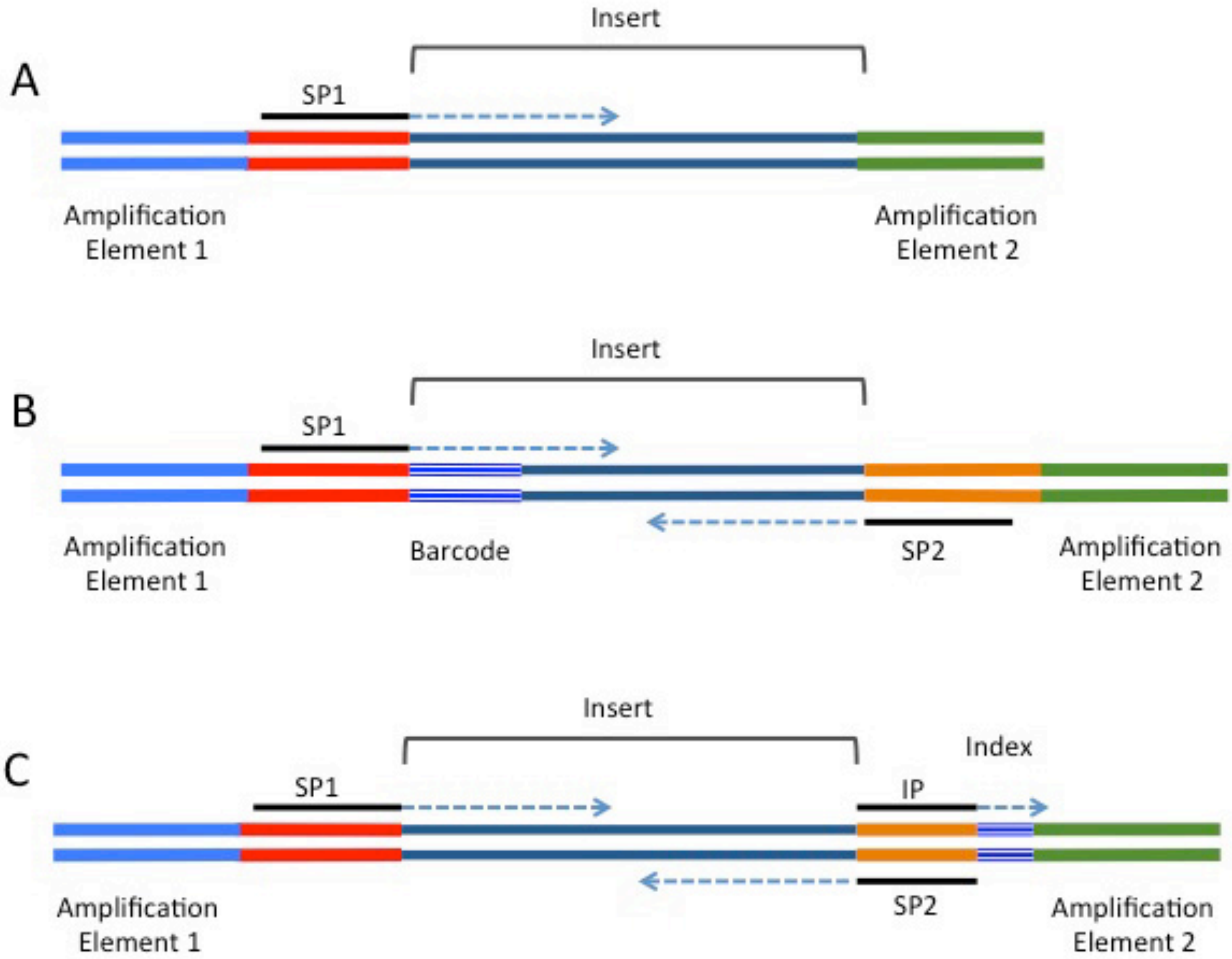
Dense clusters of identical fragments are generated.

Add labeled reversible terminators. Emitted fluorescence identifies the incorporated base. Cleave fluorescent dye. Repeat.

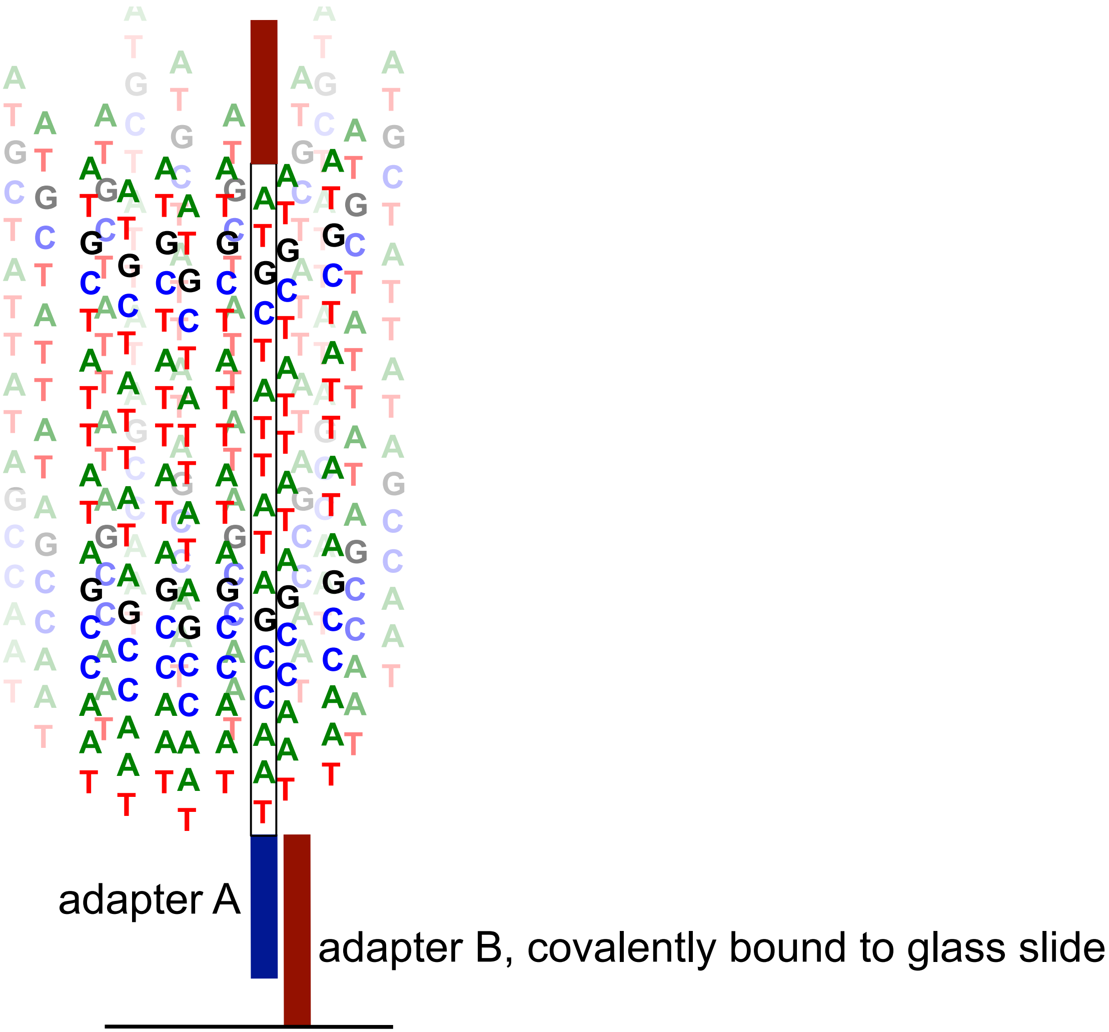
[Number of cycles = read length.]

Images converted to “reads”.

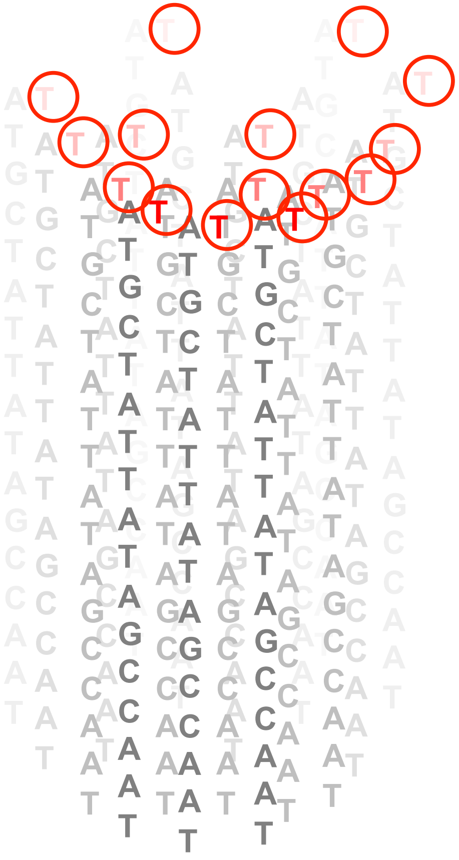
Illumina sequencing technology



Illumina cluster generation



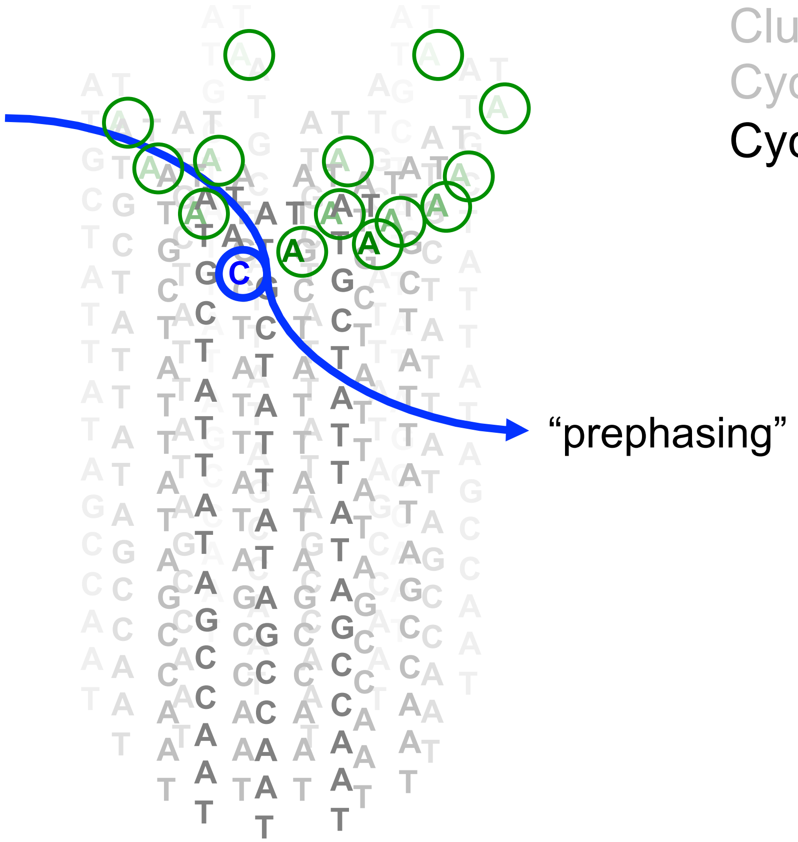
Cycle 1 strong signal



Cluster generation

Cycle 1 *read as*: T T

Cycle 2 : errors

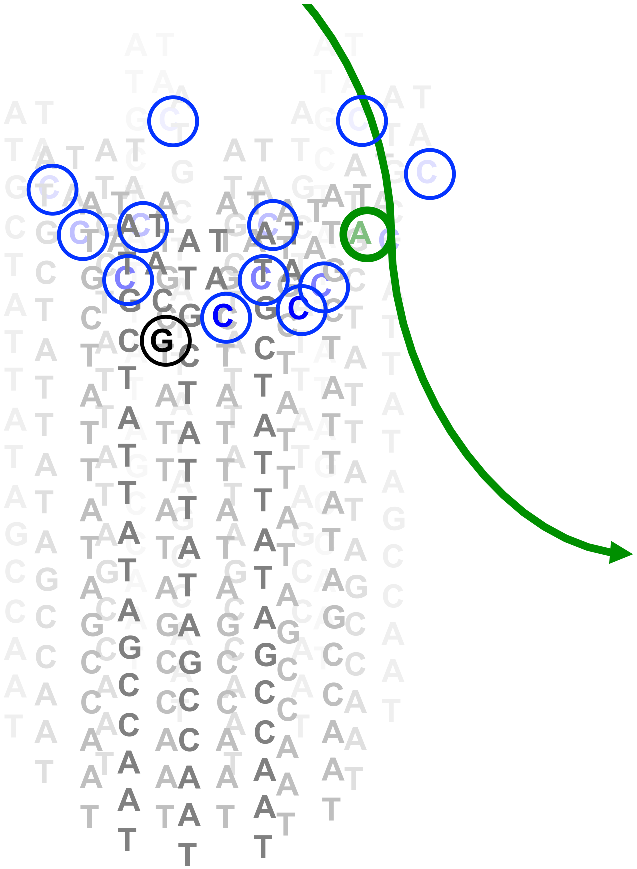


Cluster generation

Cycle 1 *read as*: T T

Cycle 2 *read as*: A A

Cycle 3 : errors

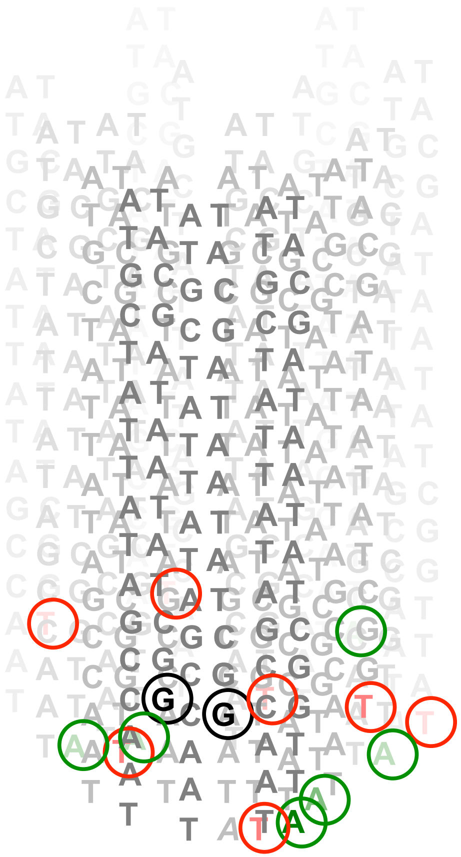


“postphasing”

Cluster generation

Cycle 1	*read as*:	T	T
Cycle 2	*read as*:	A	A
Cycle 3	*read as*:	C	C

Later cycles with more errors



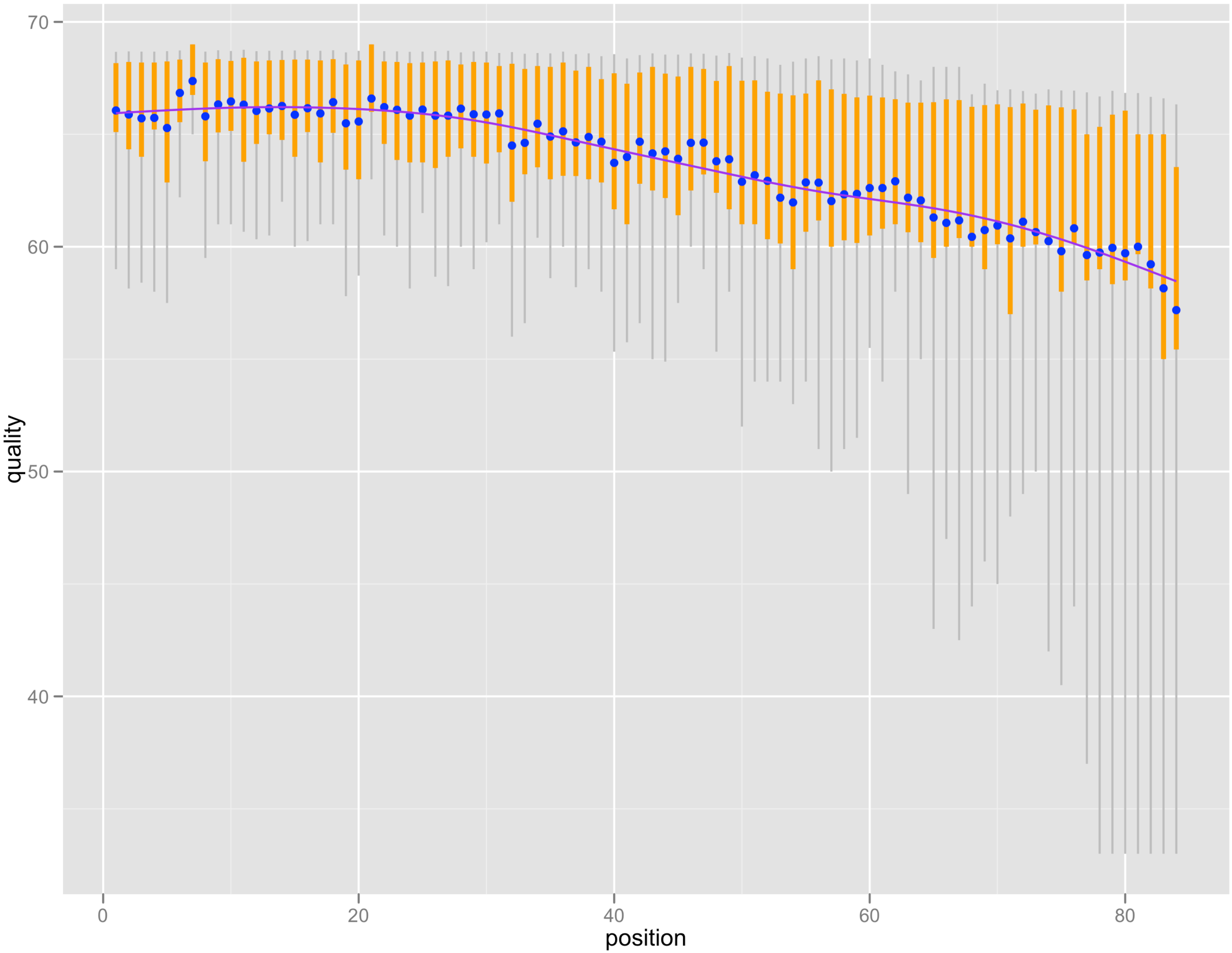
Cluster generation

Cycle 1	*read as*:	T	T
Cycle 2	*read as*:	A	A
Cycle 3	*read as*:	C	C
Cycle 4	*read as*:	G	G
Cycle 5	*read as*:	A	A
Cycle 6	*read as*:	T	T
Cycle 7	*read as*:	A	A
Cycle 8	*read as*:	A	A
Cycle 9	*read as*:	T	T
Cycle 10	*read as*:	A	A
Cycle 11	*read as*:	T	?
Cycle 12	*read as*:	C	?
Cycle 13	*read as*:	G	?
Cycle 14	*read as*:	G	?
Cycle 15	*read as*:	T	?
Cycle 16	*read as*:	T	?

CNRS UPMC Station Biologique Roscoff

Abims

Basecall qualities



Data Cleaning