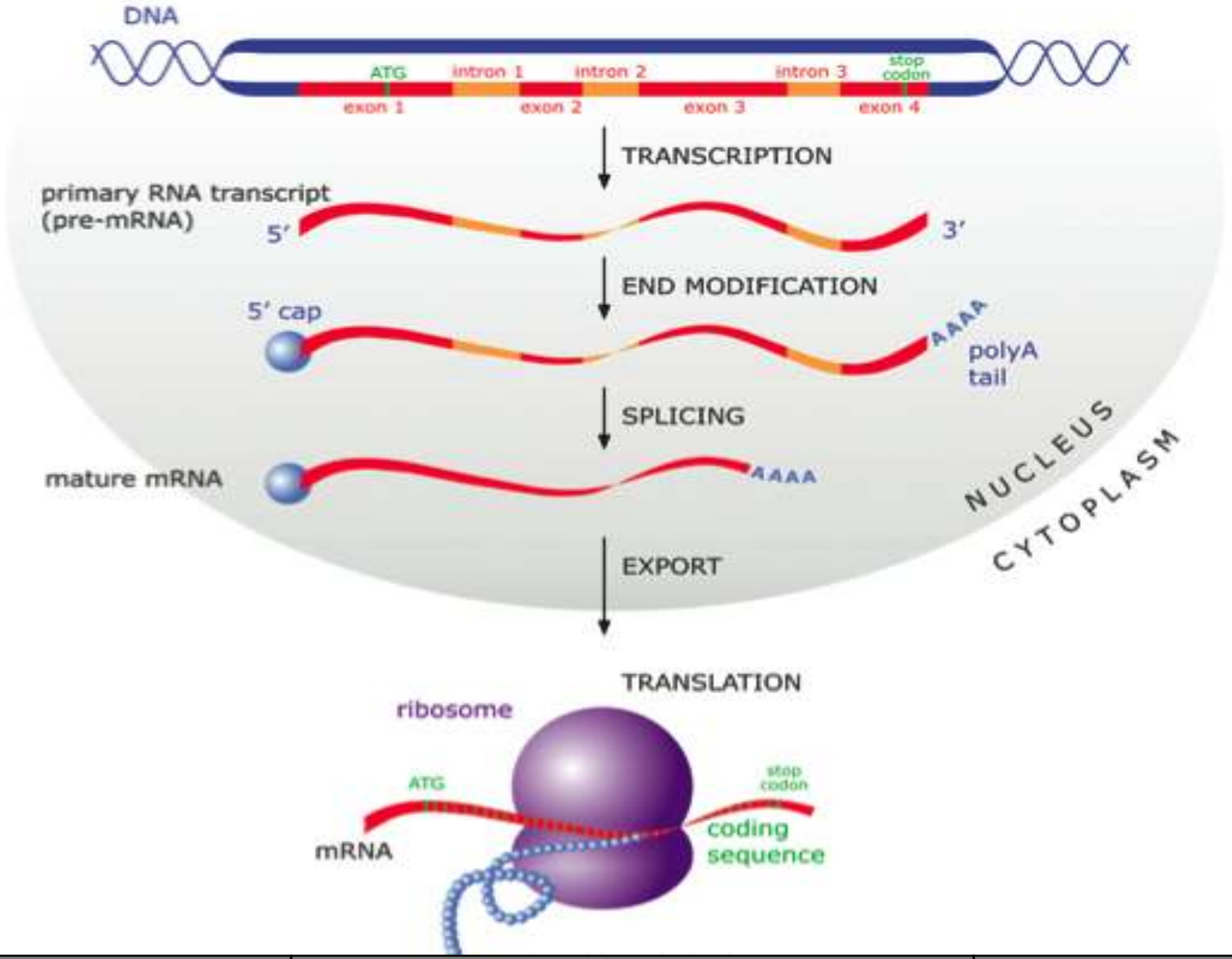
Worshop on Comparative Genomis, Cesky Krumlov, 2011

Transcriptome characterization with RNASeq

Roderic Guigó, roderic.guigo@crg.cat
Center for Genomic Regulation, Barcelona

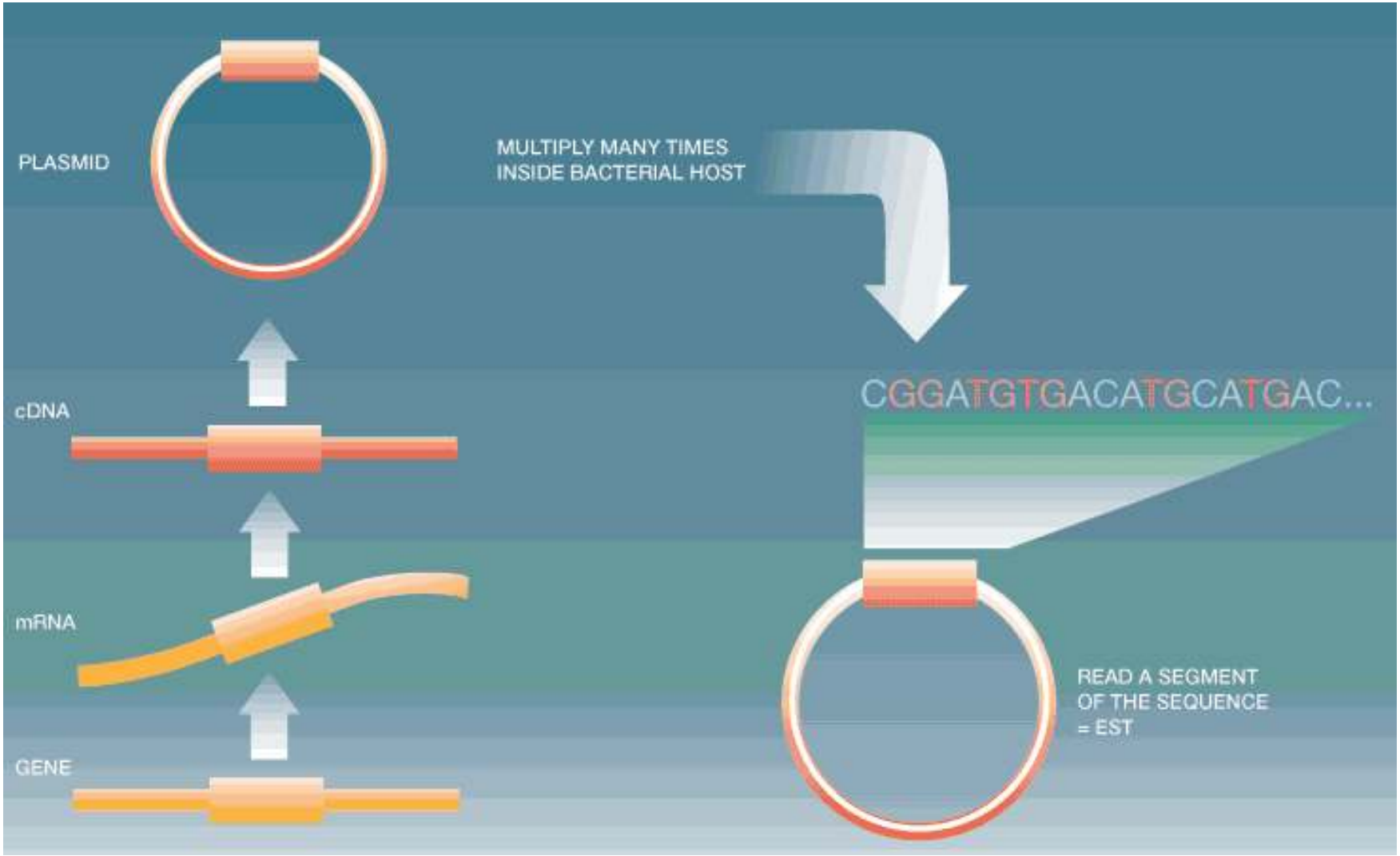
Central Dogma of Molecular Biology



RNA (as the first phenotype)

- Transcription to RNA and subsequent processing is the first step in the unfolding of the intructions encoded in the DNA.
- There is a one-to-one correspondence between the RNA content of the cell and the cellular phenotype (which is obviolusly not true for the DNA)
- The phenotypic effects of the variations in the DNA are ultimately mediated by variations in the RNA content of cell.

cDNA cloning and sequencing

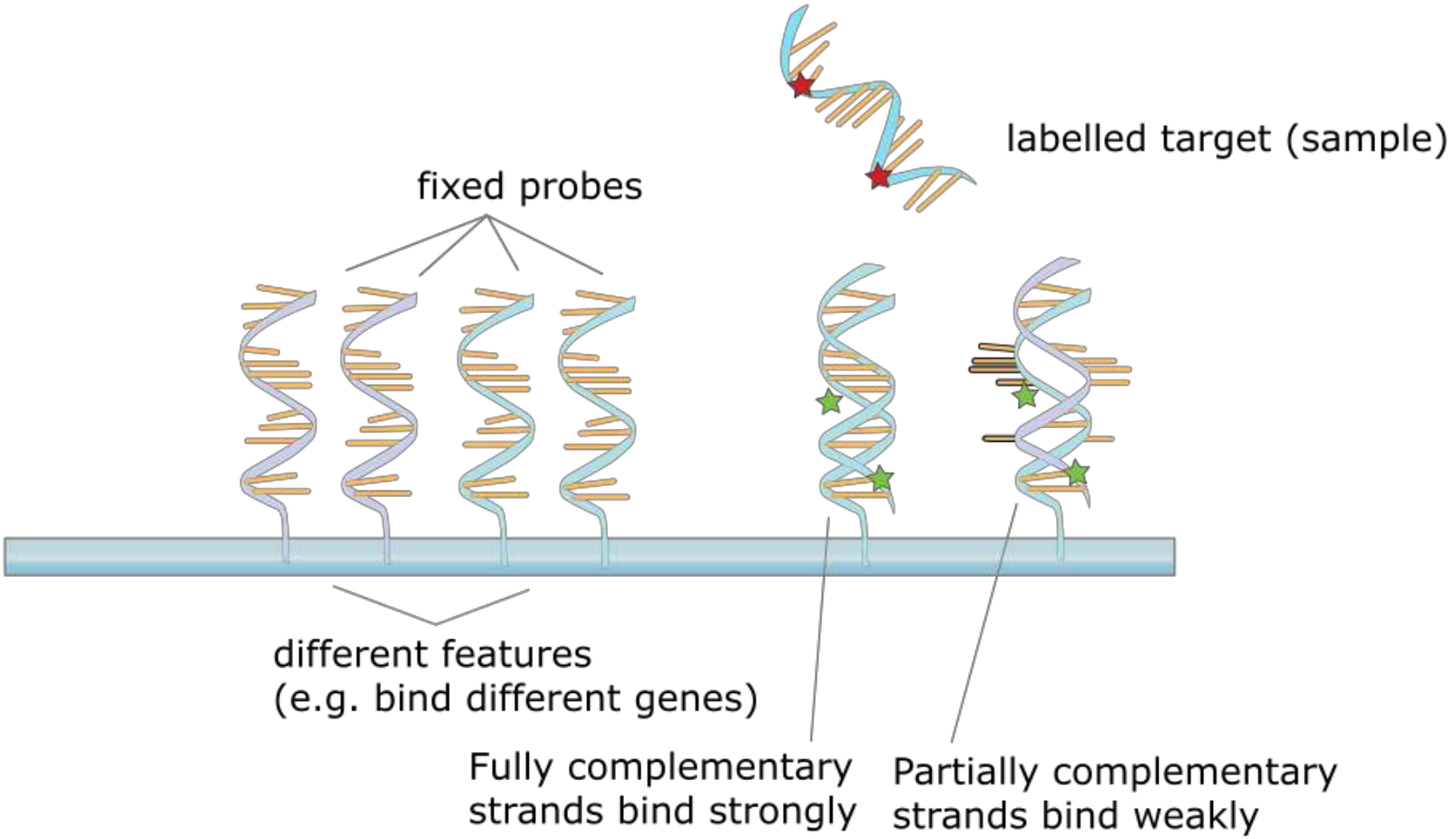


IBM Systems Journal, Inman et al (2001) , Deep Computing for life Sciences

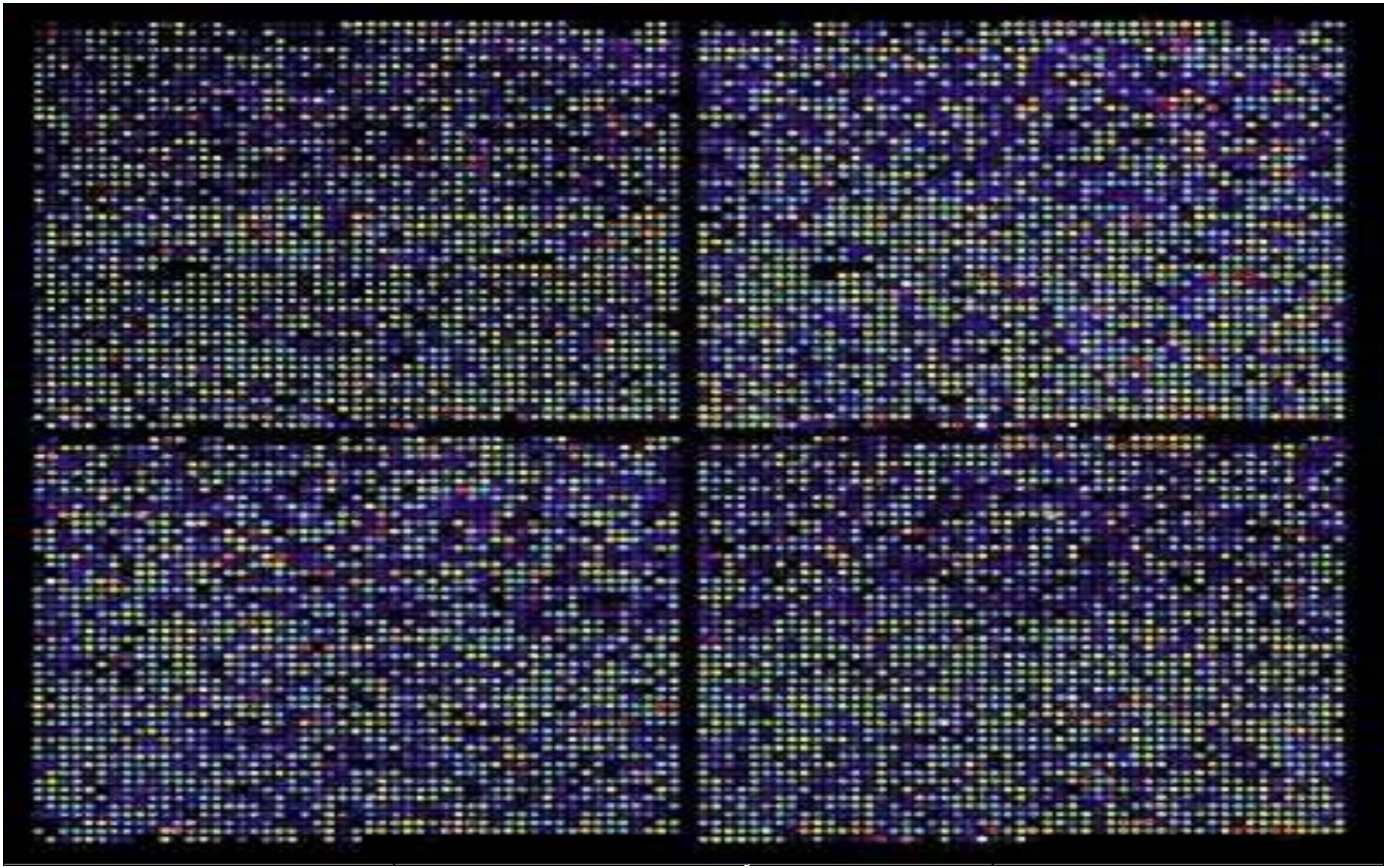
high dynamic range in RNA abundance

- RNA species vary greatly in abundance. Typically (Carninci et al., 2000)
  - 5-10 **highly expressed** species. thousands of copies ~ 20% of the mRNA mass
  - 500-2000 **intermediate** species. hundreds of copies 40%-60% of the mRNA mass
  - 10,000-20,000 **rare** messages. a few copies <20%-40% of the mRNA mass
- Random clone selection is ineffective in recovering low abundance transcripts.
- Normalization strategies are required, but the destroy transript abundances

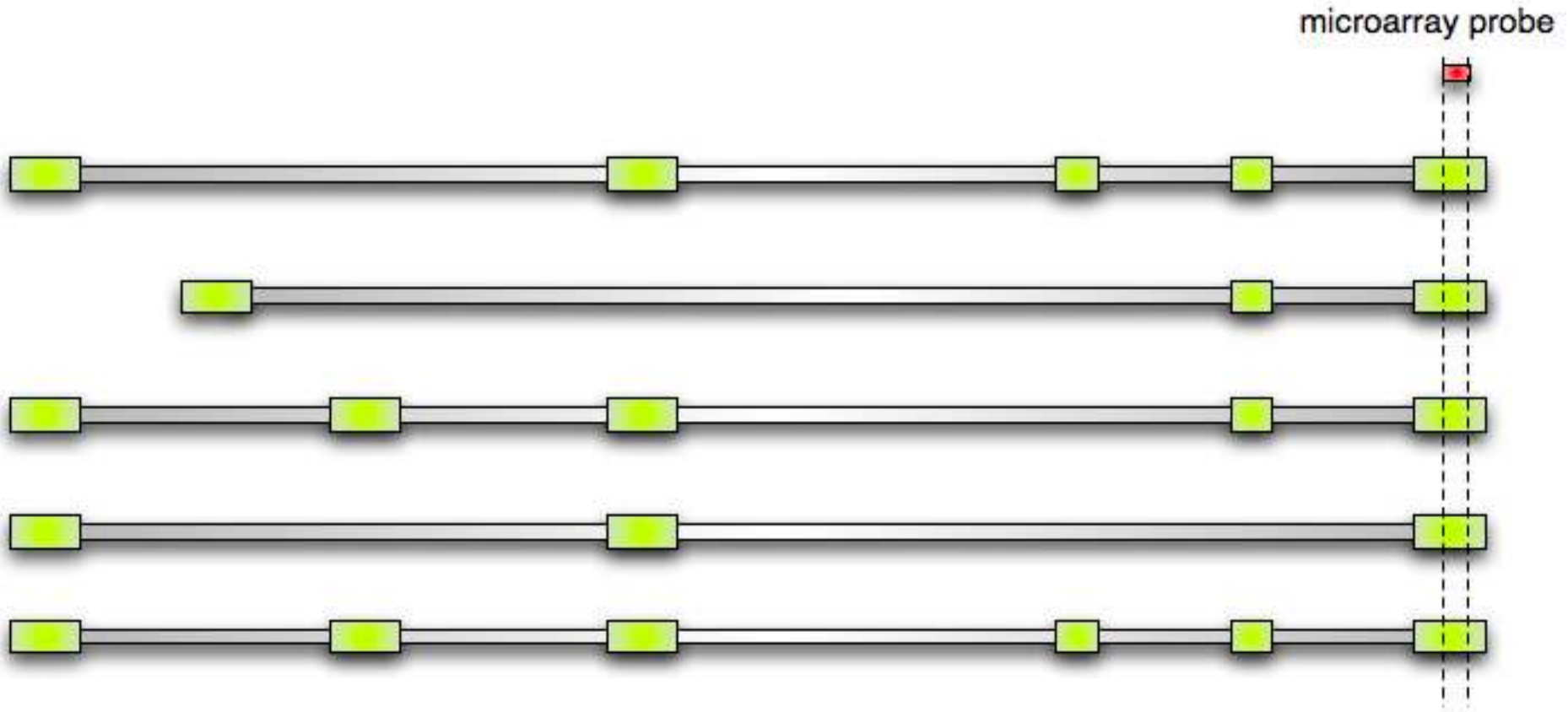
DNA microarrays

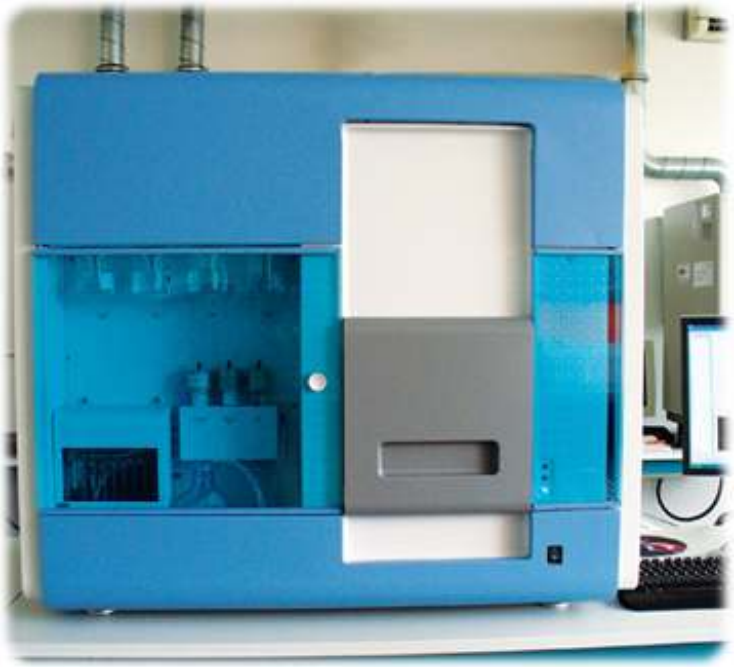


DNA microarrays



Alternative splicing





Next generation sequencing technologies.

- They MAY provide enough sequence depth, so that the (unbiased) sequencing of the RNA complement of the cell MAY become feasible.

Further Evolution of Large-scale Genome Sequencing

- 2000: Human genome working drafts
- Data unit of approximately 10x coverage of human
  - 10 years and cost about $3 billion

- 2008: Major genome centers can sequence the same number of base pairs **every 4 days**
  - 1000 Genome project launched
  - World-wide capacity dramatically increasing

- 2009: **Every 4 hours ($25,000)**
- 2010: **Every 14 minutes ($5,000)**
  - Illumina HiSeq2000 machine produces 200 gigabases per 8 day run (BGI have 128)

Many diverse technologies

- Illumina
- 454 Life Sciences
- ABI Solid
- Helicos Biosciences
- Pacific Biosciences
- Knome
- Etc.

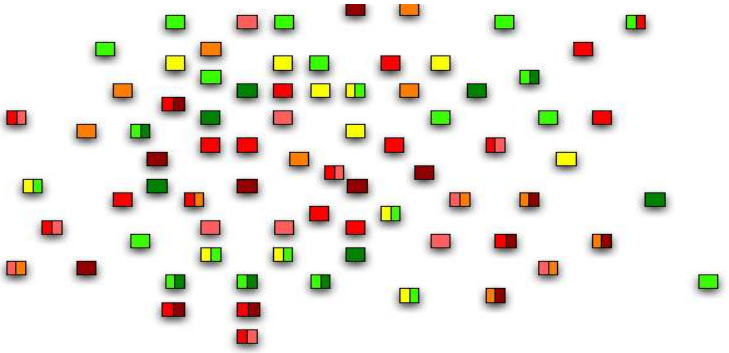
RNASeq experiment

- RNA population of interest
- Copied to cDNA
- Fragment cDNA sequences
- Select Fragment size (insert length)
- Sequence one end (single-end) or both ends (paired-end) of the fragment
- Length of Sequences short (< 100bp)

RNASeq experiment

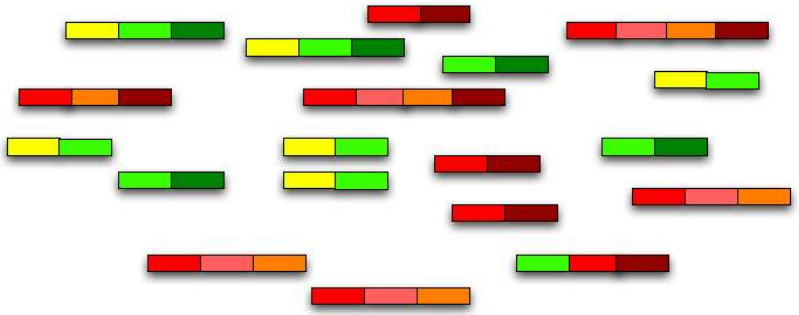
- RNA population of interest
- Copied to cDNA
- Fragment cDNA sequences
- Select Fragment size (insert length)
- Sequence one end (single-end) or both ends (paired-end) of the fragment
- **Length of Sequences short (< 100bp)**

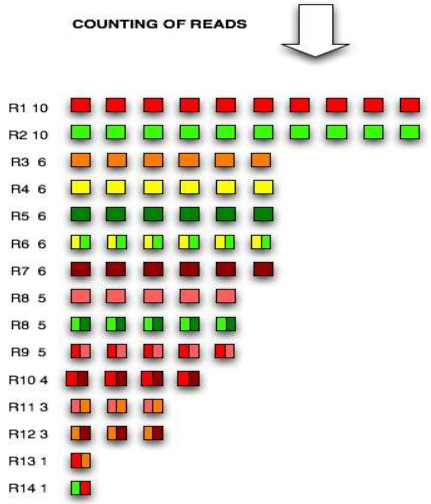
RNASeq experiment

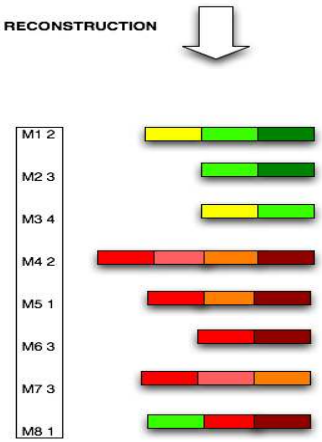


sequencing

Transcripts in the cell







Transcript reconstruction

RNASeq: long polyA+ RNA from the nucleous of K562 cells (Gingeras Lab. ENCODE project)

- Three lanes
- 234,227,244 reads
- 2x75bp reads
- 20 Gb of sequence

RNASeq steps

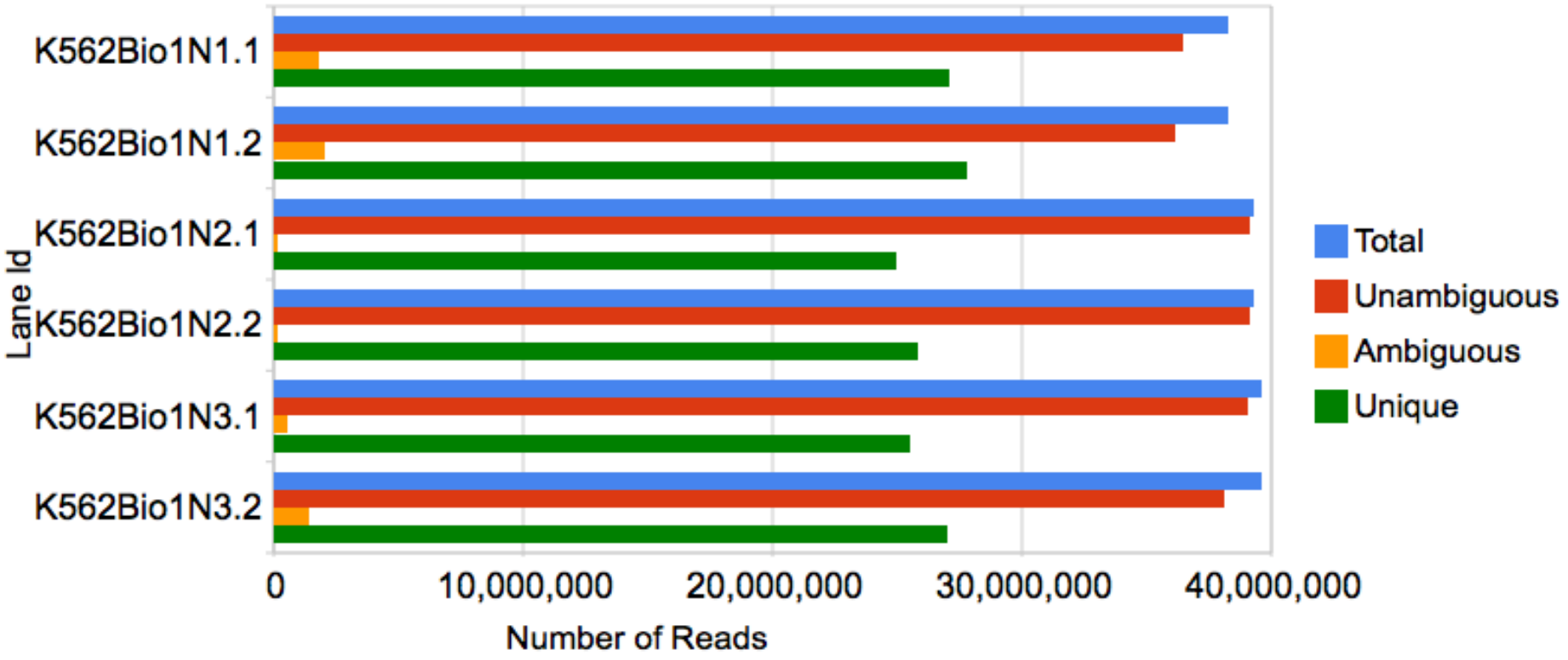
- Analysis of read quality
- Mapping of reads
- Transcript quantitation
- Transcript discovery

FASTAQ File

```
@Homo_sapiens_1_2_1125_702|p1
CAGACTTAAAAGGCCATTCCCTACTTCAAAGTTACATTCAACACCACTATGATCCCCTTCCGTTATTAGCAACTTT
+Homo_sapiens_1_2_1125_702|p1
abaabbbabbaa`aa`bbaaaaababbaa_]abababaaaaaababababaaaaa^_aaa``\_aaaa\F[^^_```
@Homo_sapiens_1_2_1125_961|p1
GCNCGTTTGAGTTTTGATCTTAAAGGACTTGTACCTTCCTCATCCGTGACTCAATGTCAGCACGACGATTCGGCAT
+Homo_sapiens_1_2_1125_961|p1
VbBb_aabJXJ[bba[]b`b`bb`SNWaaaaaaab_a]b``b`bba_a]_aa_a_]`a_Q`W`J_]`^aaaIU\NU
```

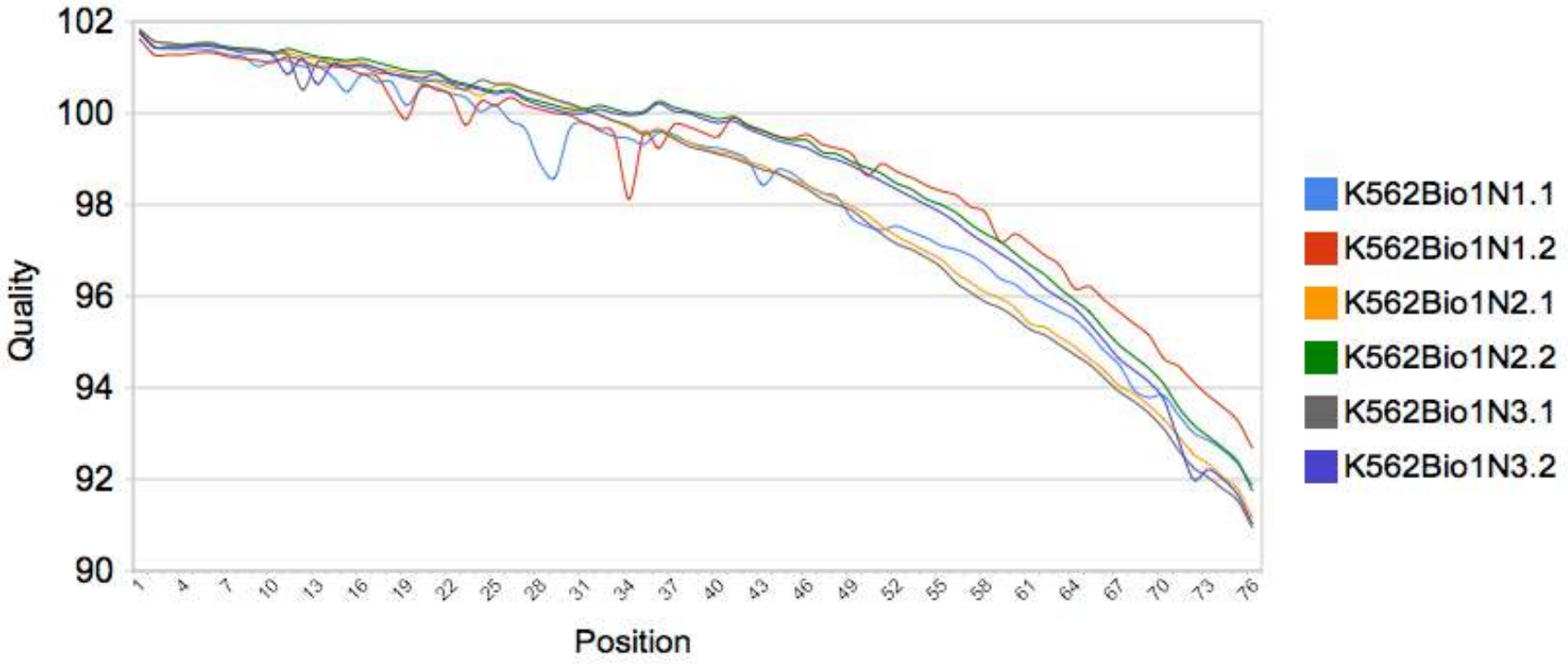
Illumina Quality scores range from 104 to 64

Read Quality: ambigous bases



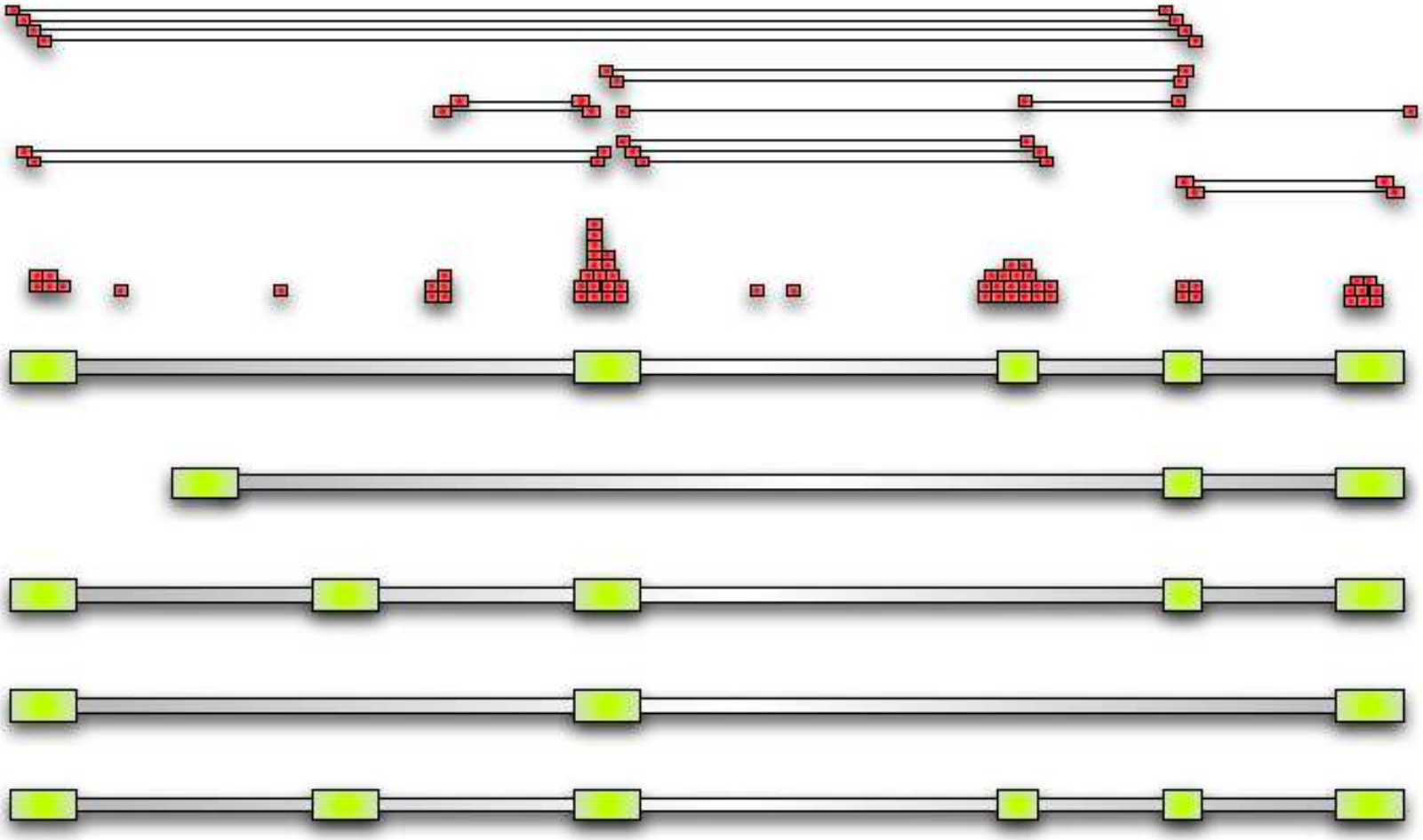
Total unambiguos reads: 228,028,503

Read Quality: Quality score per position



MAPPING OF READS

mapping the reads to the genome & transcriptome



Short-Read Sequence Alignment [edit]

Name	Description	Multi-threaded	License	Link
BFAST	Explicit time and accuracy tradeoff with a prior accuracy estimation, supported by indexing the reference sequences. Optimally compresses indexes. Can handle billions of short reads. Can handle insertions, deletions, SNPs, and color errors (can map ABI SOLiD color space reads). Performs a full Smith Waterman alignment.	Yes (POSIX)	GPL	link
BLASTN	BLAST's nucleotide alignment program, slow and not accurate for short reads, and uses a sequence database (EST, sanger sequence) rather than a reference genome.			link
BLAT	Made by Jim Kent. Can handle one mismatch in initial alignment step.	Yes (client/server).	Free for academic and non-commercial use.	link
Bowtie	Uses a Burrows-Wheeler transform to create a permanent, reusable index of the genome; 1.3 GB memory footprint for human genome. Aligns more than 25 million Illumina reads in 1 CPU hour. Supports Maq-like and SOAP-like alignment policies.	Yes (POSIX)	Artistic License	link
CASHX	Quantify and manage large quantities of short-read sequence data. CASHX pipeline contains a set of tools that can be used together or as independent modules on their own. This algorithm is very accurate for perfect hits to a reference genome.	No	Free for academic and non-commercial use.	link
CUDA-EC	Short-read alignment error correction using GPUs.	Yes (GPU enabled)		CUDA-EC
ELAND	Implemented by Illumina. Includes ungapped alignment with a finite read length.			
GMAP and GSNAP	Robust, fast, short-read alignment. GMAP: singleton reads; GSNAP: paired reads. Useful for digital gene expression, SNP and indel genotyping. Developed by Tom Wu at Genentech. Implemented by NCGR in Alpheus.	Yes	Free for academic and non-commercial use.	[4]
LAST				link
MAQ	Ungapped alignment that takes into account quality scores for each base.		GPL	link
MOM	MOM or maximum oligonucleotide mapping is a query matching tool that captures a maximal length match within the short read.	Yes		[5]
MOSAIK	Fast gapped aligner and reference-guided assembler. Aligns reads using a banded Smith-Waterman algorithm seeded by results from a k-mer hashing scheme. Supports reads ranging in size from very short to very long.	Yes		link
Novoalign	Gapped alignment of single end and paired end Illumina GA I & II reads and reads from the new Helicos Heliscope Genome Analyzer. High sensitivity and specificity, using base qualities at all steps in the alignment. Includes adapter trimming, base quality calibration, Bi-Seq alignment, and option to report multiple alignments per read.	Multi-threading and MPI versions available with paid license.	Single threaded version free for academic and non-commercial use.	Novocraft
PALMapper	PALMapper, efficiently computes both spliced and unspliced alignments at high accuracy. Relying on a machine learning strategy combined with a fast mapping based on a banded Smith-Waterman-like algorithm it aligns around 7 million reads per hour on a single CPU. It refines the originally proposed QPALMA approach.	Yes	GPL	[6]
PerM	Indexes the genome with periodic seeds to quickly find alignments with full sensitivity up to four mismatches. It can map Illumina and SOLiD reads. Unlike most mapping programs, speed increases for longer read lengths.	Yes	GPL	link
QPalma	Is able to take advantage of quality scores, intron lengths and computation splice site predictions to perform and performs an unbiased alignment. Can be trained to the specifics of a RNA-seq experiment and genome. Useful for splice site/intron discovery and for gene model building. (See PALMapper for a faster version).	Yes (client/server)	GPLv2	link
RazerS	No read length limit. Hamming or edit distance mapping with configurable error rates. Configurable and predictable sensitivity (runtime/sensitivity tradeoff). Supports paired-end read mapping.		LGPL	link
RMAP	Read lengths can range from 20bp to at most 64bp. Uses the "exclusion principle" to allow for mismatches and look-up reads in an index.			link
SeqMap	Up to 5 mixed substitutions and insertions/deletions. Various tuning options and input/output formats.		Free for academic and non-commercial use.	link
Shrec	Short read error correction with a Suffix trie data structure.	Yes (Java)		link
SHRiMP	Indexes the reads instead of the reference genome. Uses masks to generate possible keys. Can map ABI SOLiD color space reads.		BSD derivative	link
SLIDER	Slider is an application for the Illumina Sequence Analyzer output that uses the "probability" files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences.			link
SLIM Search	Extremely fast, tolerant to high indel and substitution counts. Includes full read alignment.	Yes	Commercial/High-use customers can join beta programme (Q2 '09)	link
SOAP	Robust with a small (1-3) number of gaps and mismatches. Speed improvement over BLAT, uses a 12 letter hash table.Now SOAP2 is much faster than the first version.	YES(MULTITHREAD)	GPL; SOAP2 source is currently unavailable	link
SOCS	For ABI SOLiD technologies. Significant increase in time to map reads with mismatches (or color errors). Uses an iterative version of the Rabin-Karp string search algorithm.	Yes		link
SSAHA and SSAHA2	Fast for a small number of variants.		Free for academic and non-commercial use.	link
Taipan	de-novo Assembler for Illumina reads		Free for academic and non-commercial use.	link
ZOOM	100% sensitivity for a reads between 15 - 240bp with practical mismatches. Very fast. Support insertions and deletions. Works with Illumina & SOLiD instruments, not 454.	Yes (GUI) No (CLI).	Commercial	link

Short read sequence alignment programs

- More succesful approaches use a data structure known as suffix array
  - Invented by Manber and Myers, 1991

- GEM (Paolo Ribeca, http://gemlibrary.sourceforge.net)

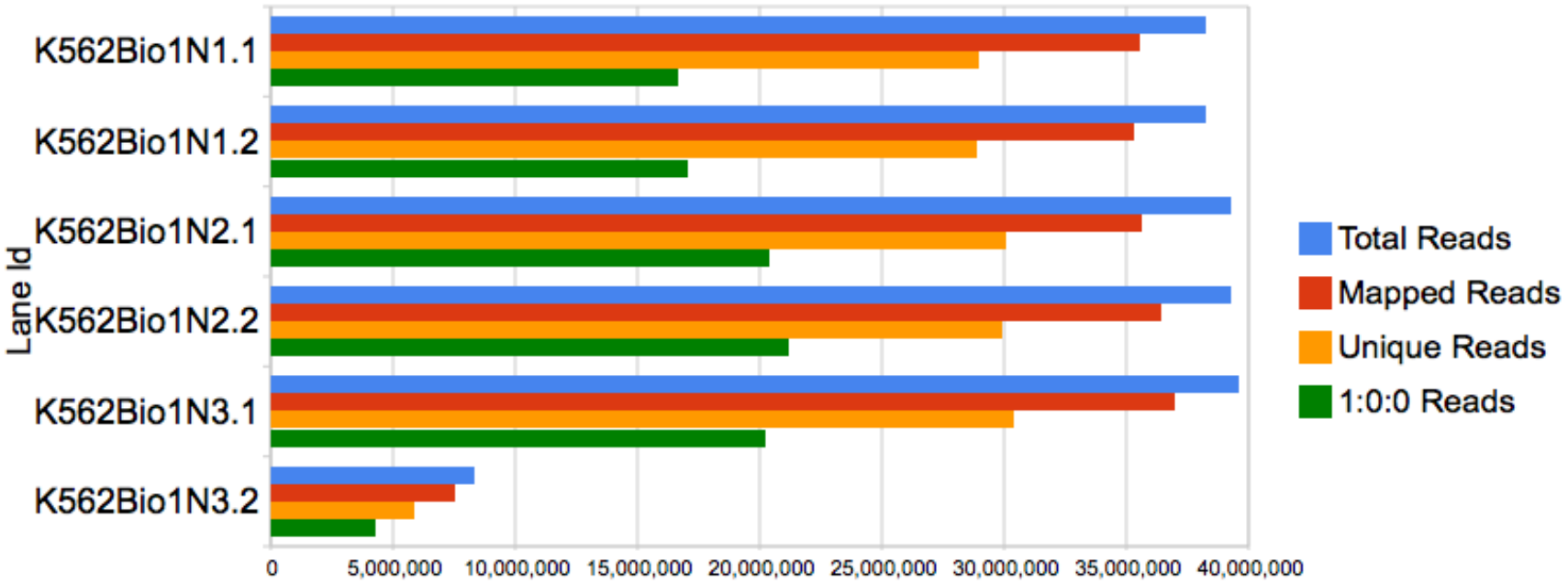
- Need for a new generation of mapping tools

Read mapping

```
Homo_sapiens_1_2_1125_702|p1
CAGACTTAAAAGGCCATTCCCTACTTCAAAGTTACATTCAACACCACTATGATCCCCTTCCGTTATTAGCAACTTT
abaabbbabbaa`aa`bbaaaaababbaa_]abababaaaaaababababaaaaaa^_aaa``\_aaaa\F[^^_```
1:0:0chr9:F114983623@0/0
Homo_sapiens_1_2_1125_961|p1
GCCCGTTTGAGTTTTGATCTTAAAGGACTTGTACCTTCCTCATCCGTGACTCAATGTCAGCACGACGATTCGGCAT
Vbab_aabJXJ[bba[]b`b`bb`SNWaaaaaaab_a]b``b`bba_a]_aa_a_]`a_Q`W`J_]`^aaaIU\NU
1:0:0chrX:R1720981@0/0


BILLIEHOLIDAY_0008:1:11:13870:15798#0/1
AAAACTAATAACTTAAAACTGCCACACNCAAAAAAGAAAACCAAAGTGGTCCACAAAACATTCTCCTTTCCTTCTG
hhhhhhgghhhhhhdhhhhhhhhhhffBfffffffhhhhhhhghhhdhhhhhhfhhhhhhhfhehhhhhhhhhggh
0:4:4:3
chr6:F74227323G28@2/1,chr7:F22550021G28@2/1,chr9:R135896396G28@2/1,chr13:R96271784A
   28@2/1,chr5:F14651837T27G28@40/2,chr12:F8233706G28<+2>43@2/1,chr12:F19608973G26G
   28@40/2,chr10:R129322971T27G28@40/2
```

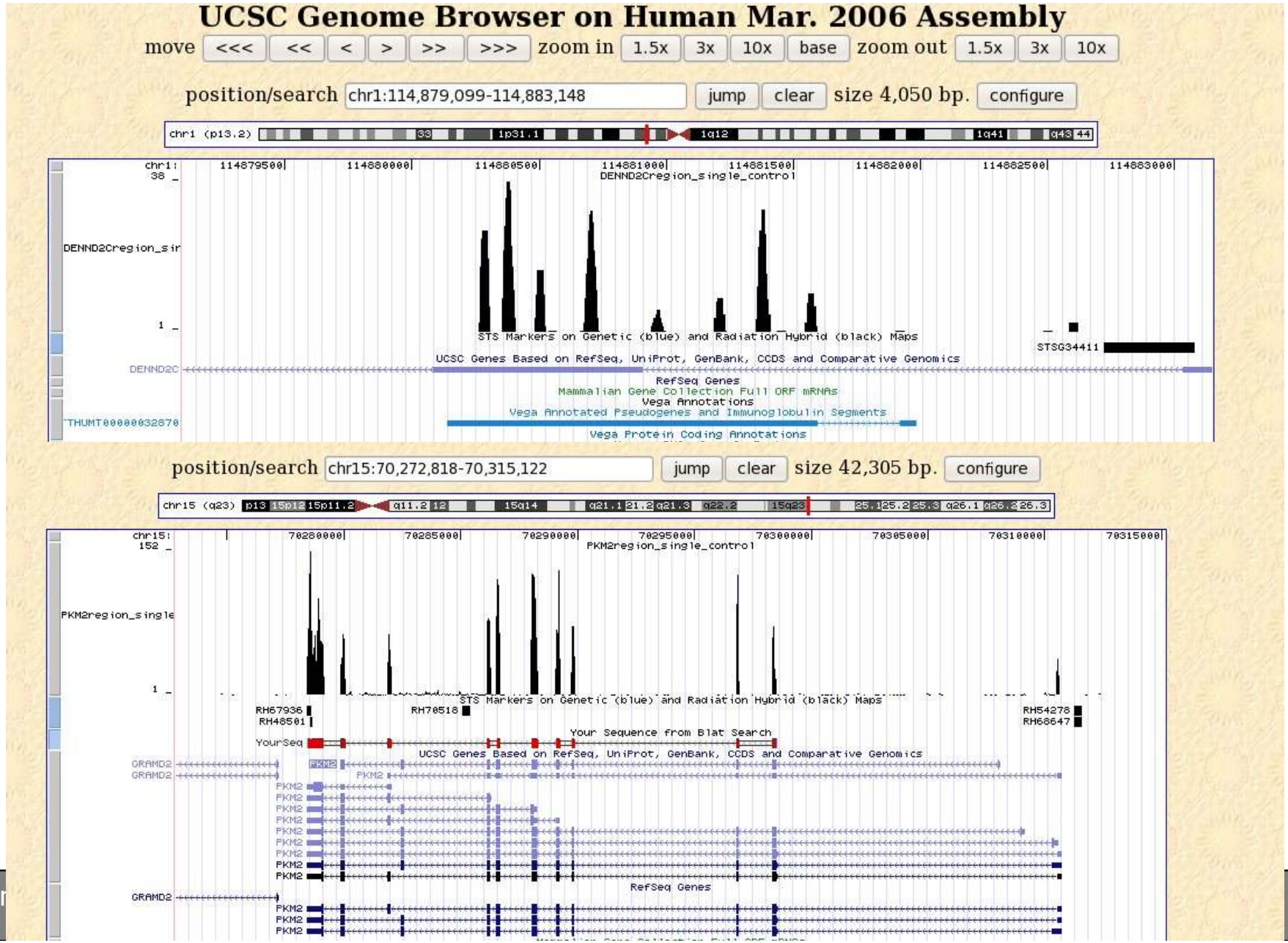
Read mapping



MAPPING SUMMARY

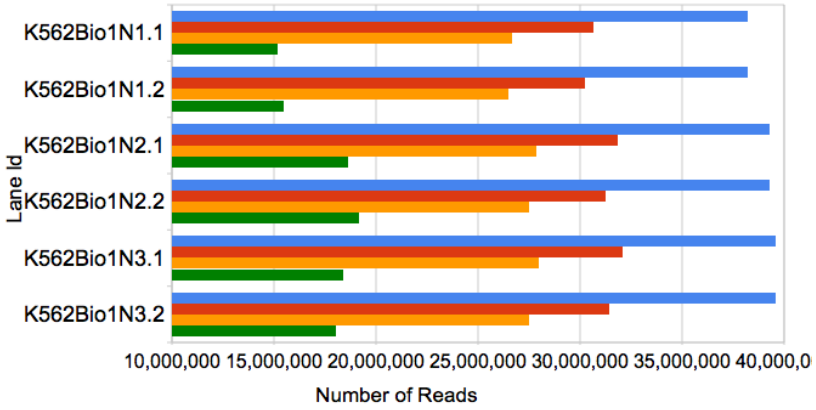
	Total	Percent
Uniquely Mapped Reads	154,185,171	76.0%
Multi-Mapped Reads	33,227,498	16.4%
Unmapped Reads	15,589,960	7.7%

Mapping of RNAseq reads is not trivial

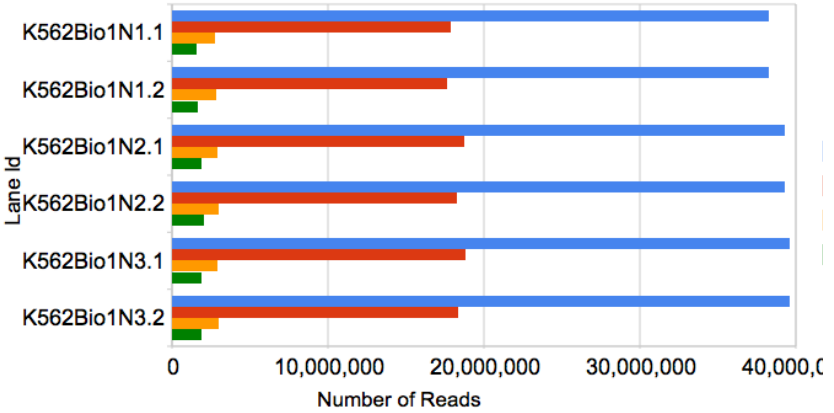


A more complex mapping strategy

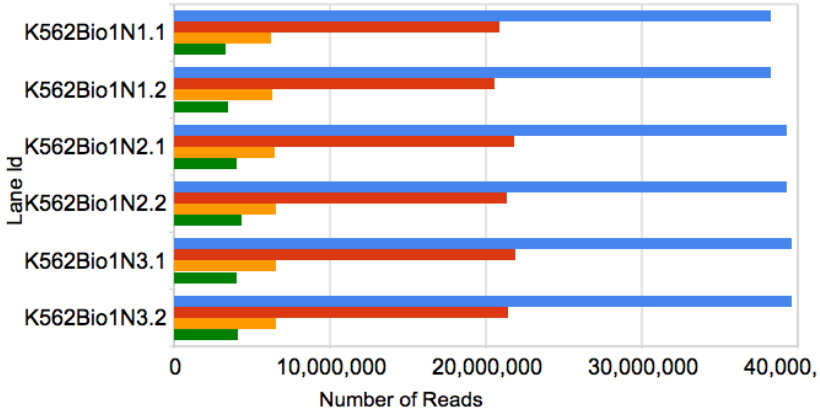
Genome mapping



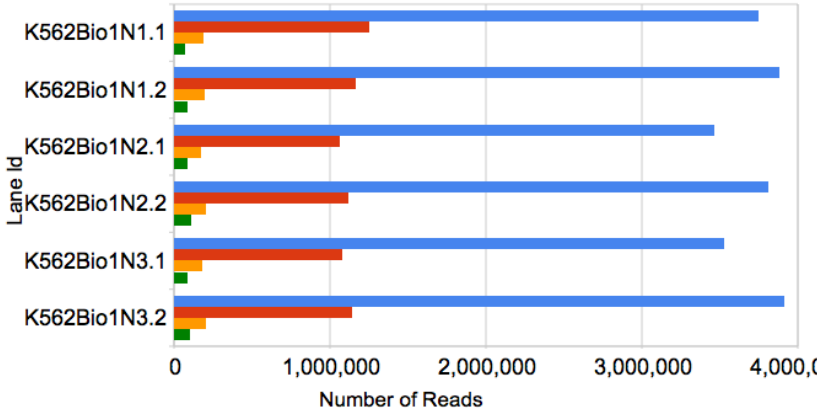
Junction mapping



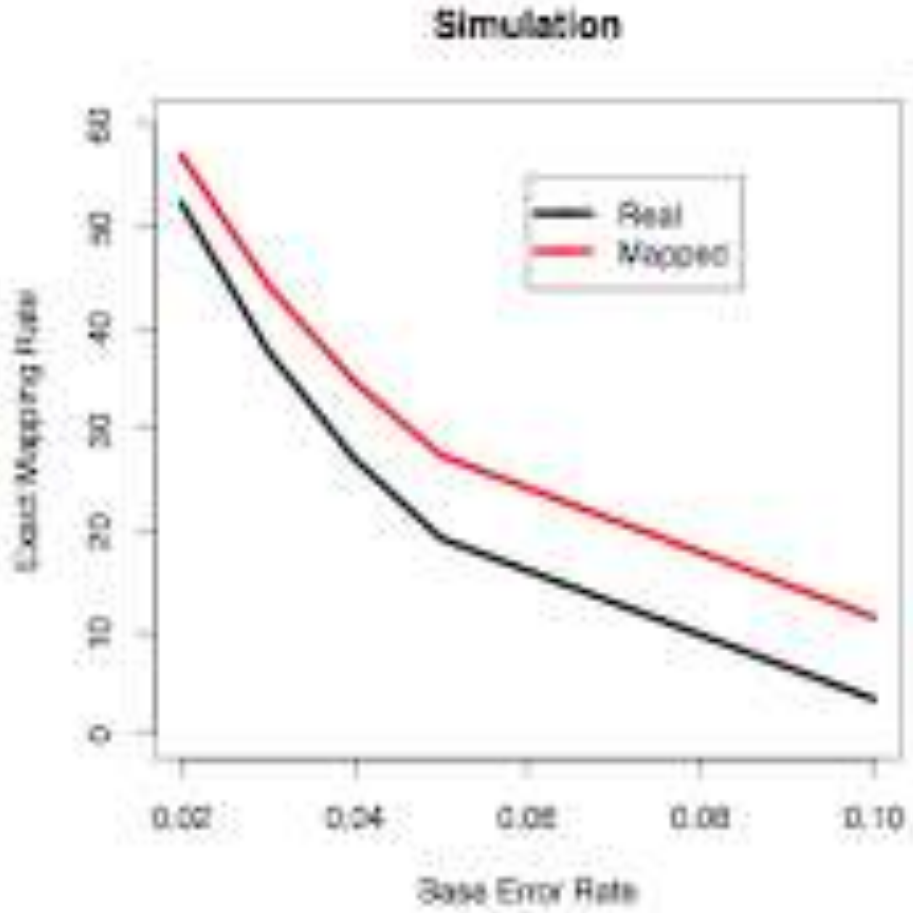
Transcriptome mapping



Split mapping



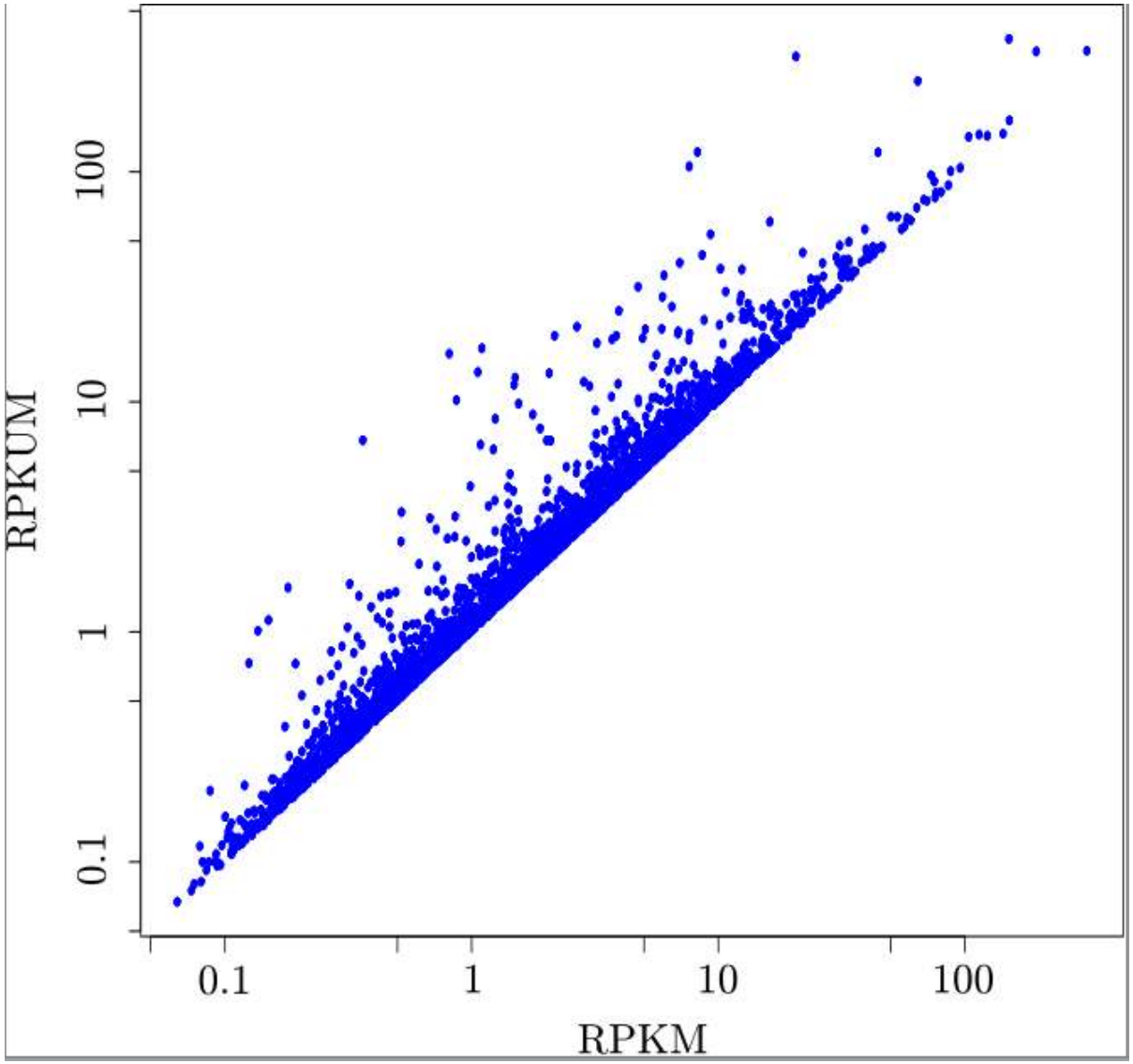
Mapping of RNASeq reads is not trivial



- Extract 32nt reads from genome.
- Apply *x*% error rate.
- Record number of reads without mismatches and exactly mapped reads.
- 5-15% of exactly mapping locations are incorrect (based on this SIMULATION).

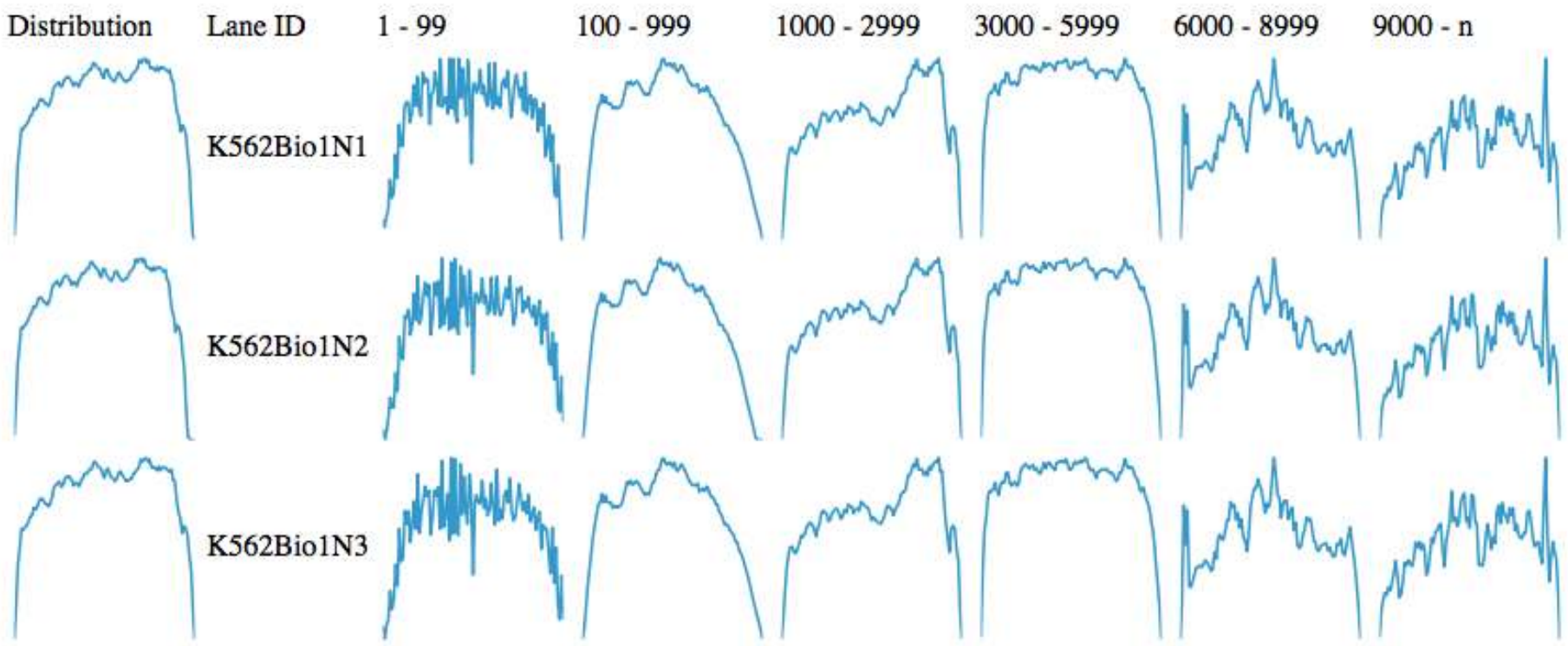
Timo Lassman, Riken

Mapping issues

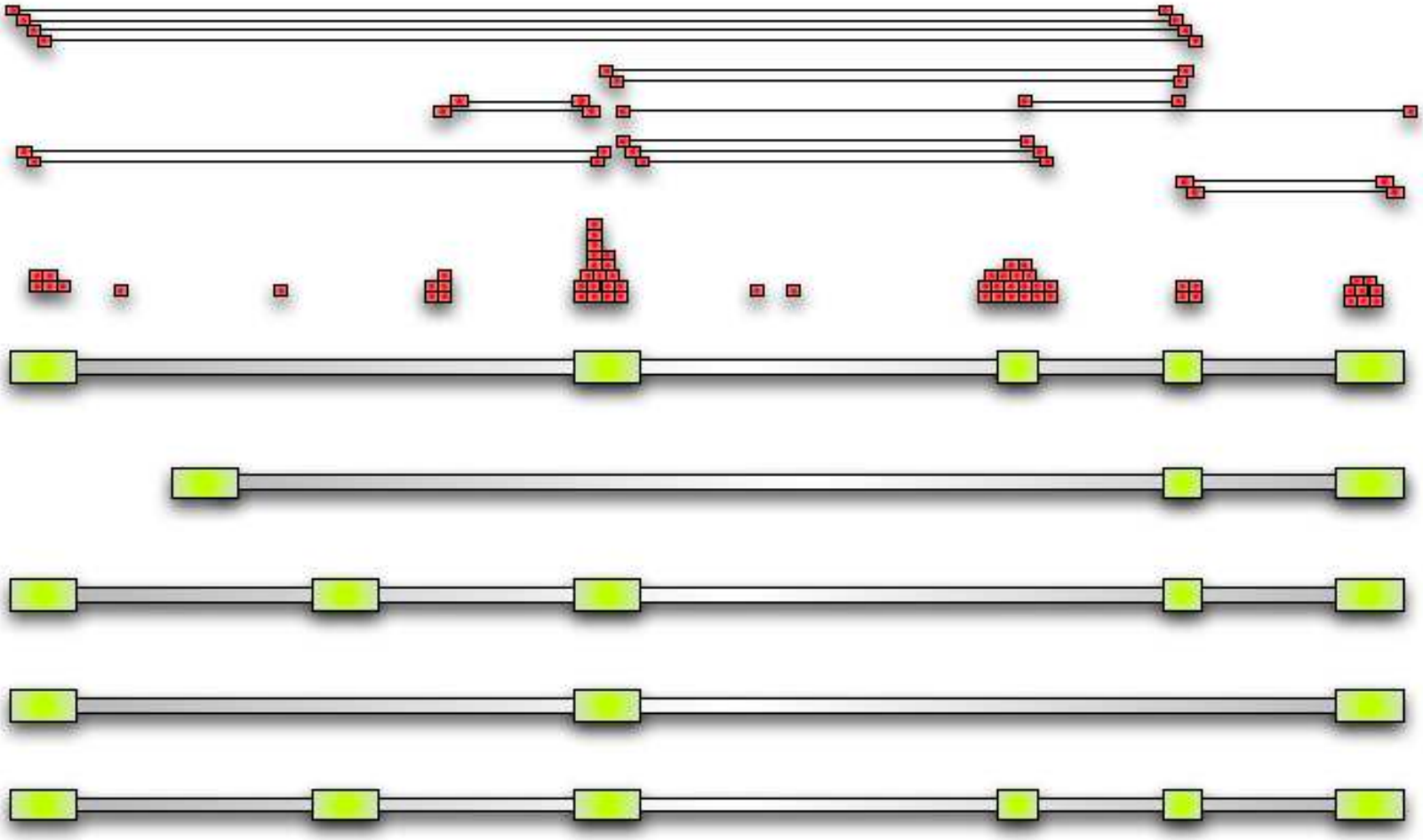


Paolo Ribeca

Read distribution along transcripts

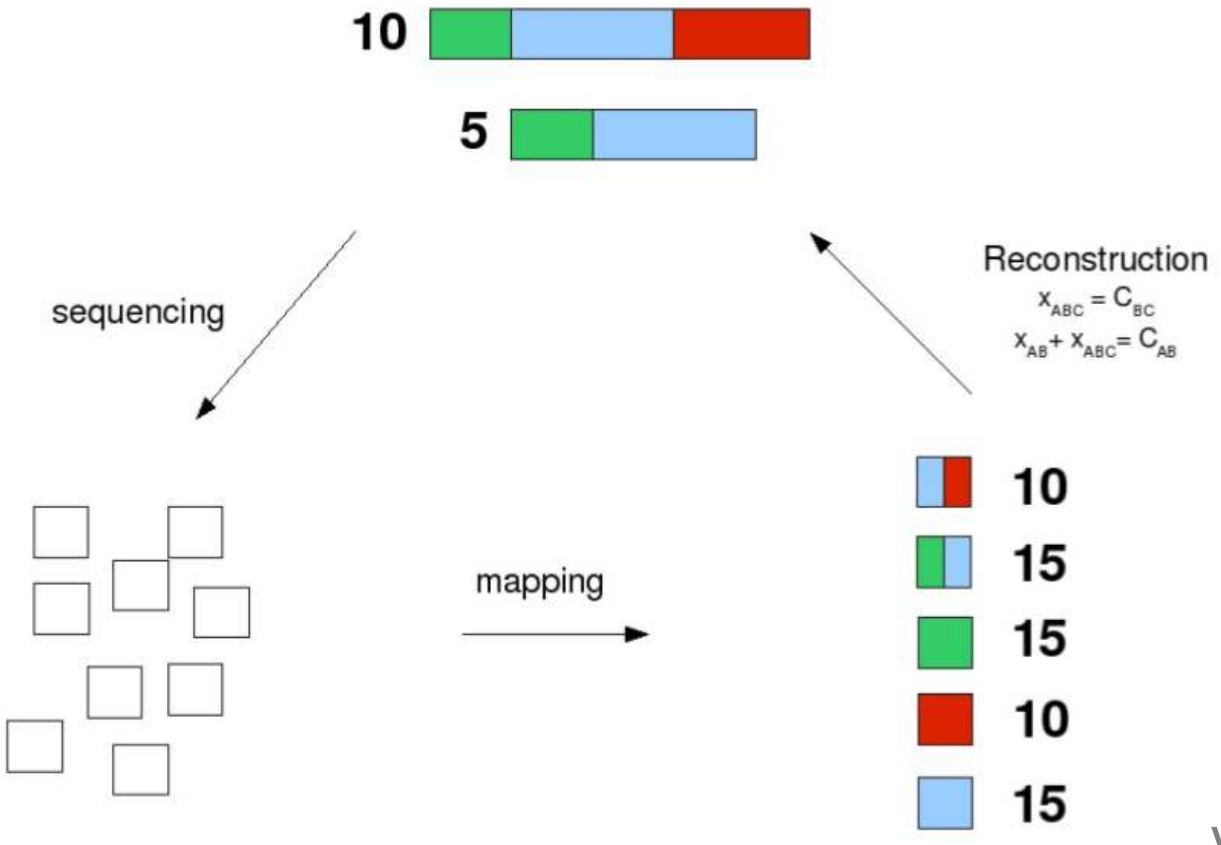


mapping the reads to the genome & transcriptome



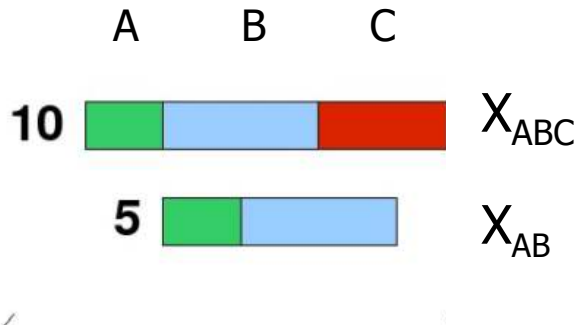
TRANSCRIPT QUANTITATION

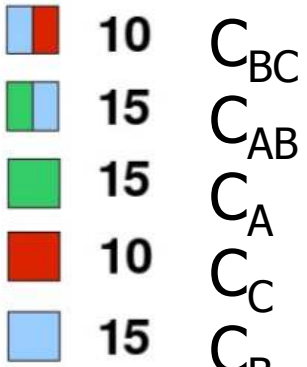
Reconstructing transcript abundances from short sequence reads



Micha Sammet
Vincent Lacroix
Paolo Ribeca

the problem can be posed as a system of linear equations

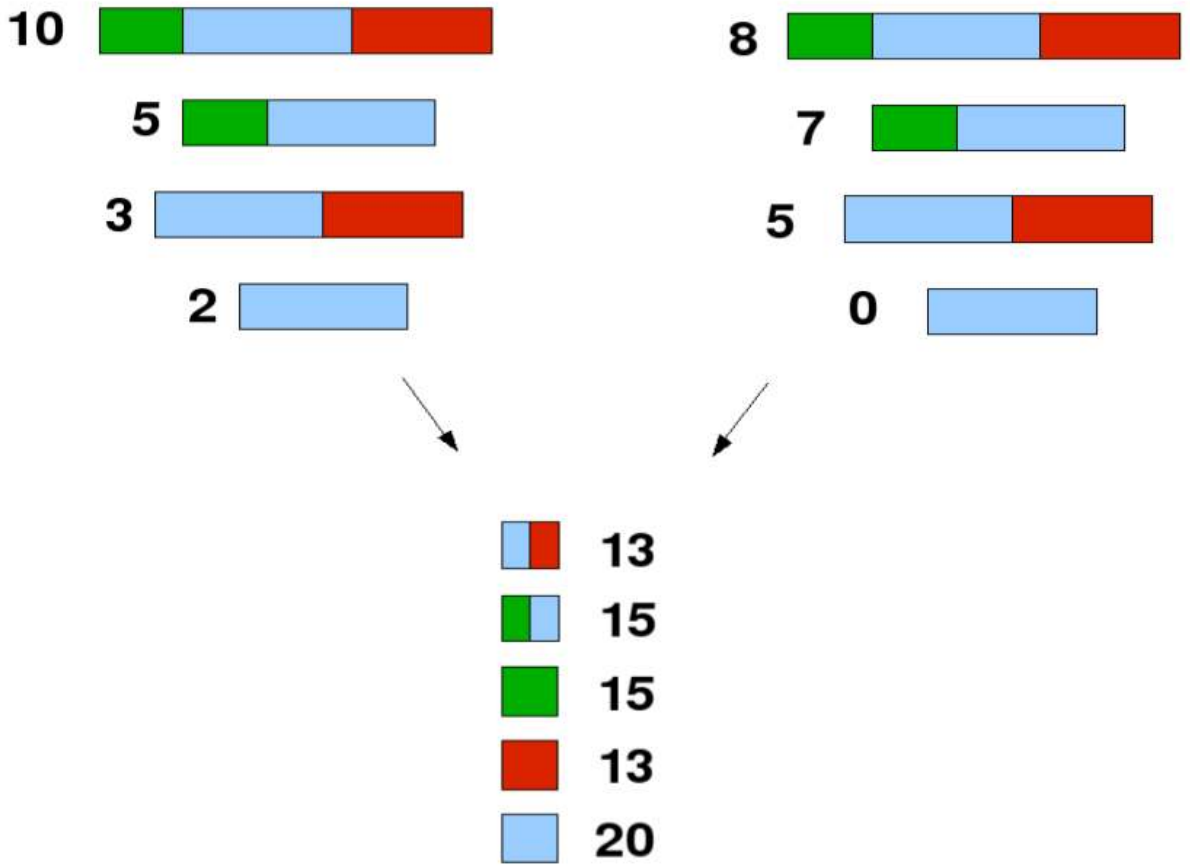




$$C_{BC} = X_{ABC}$$
$$C_{AB} = X_{ABC} + X_{AB}$$
$$C_A = X_{ABC} + X_{AB}$$
$$C_C = X_{ABC}$$
$$C_B = X_{ABC} + X_{AB}$$

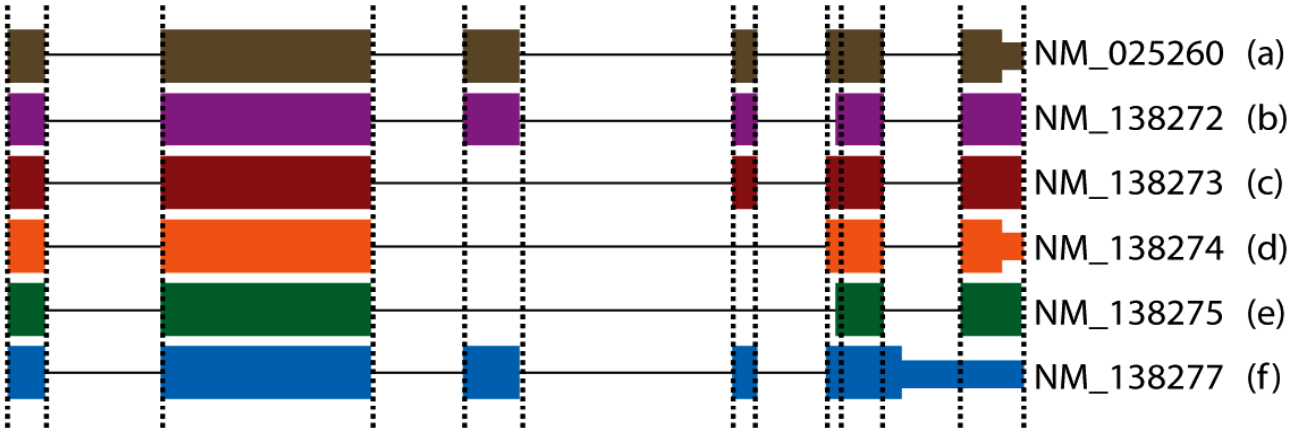
V. Lacroix, M. Sammeth, R. Guigo, A. Bergeron. Exact transcriptome reconstruction from short sequence reads. WABI (2008)

The system of equations is often underdetermined: different transcriptomes result in exactly the same set of sequenced reads

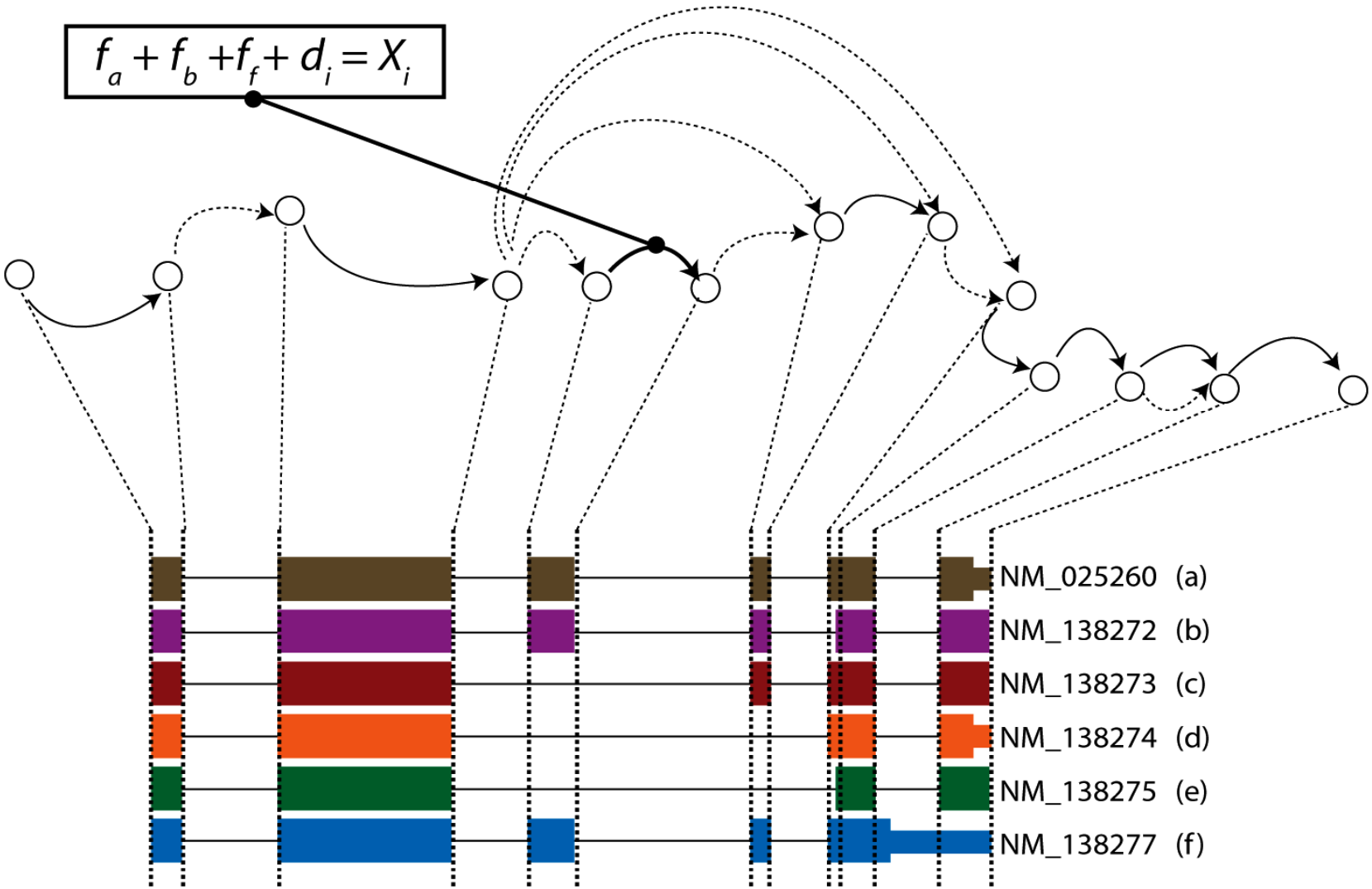


The Flux Capacitor.

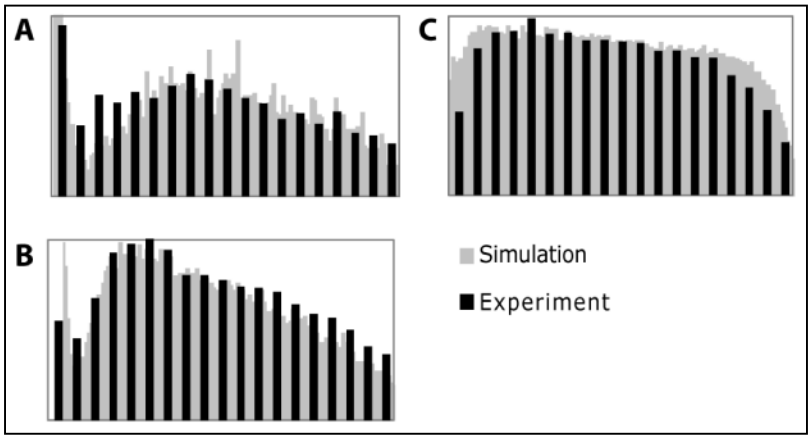
Micha Sammeth, http://flux.sammeth.net

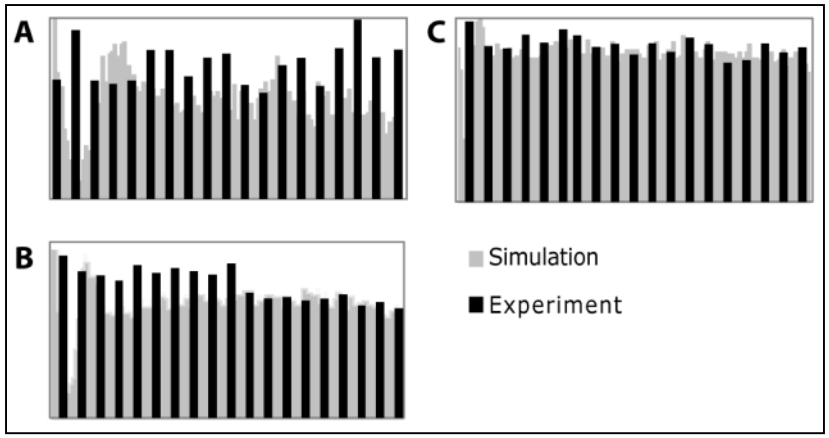


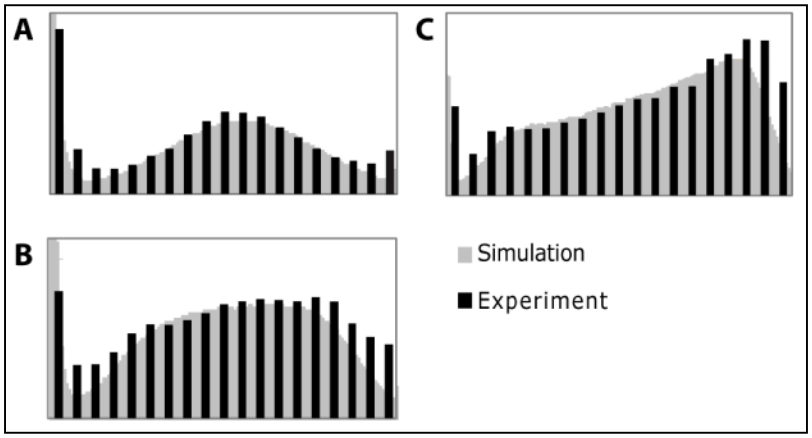
The Flux Capacitor

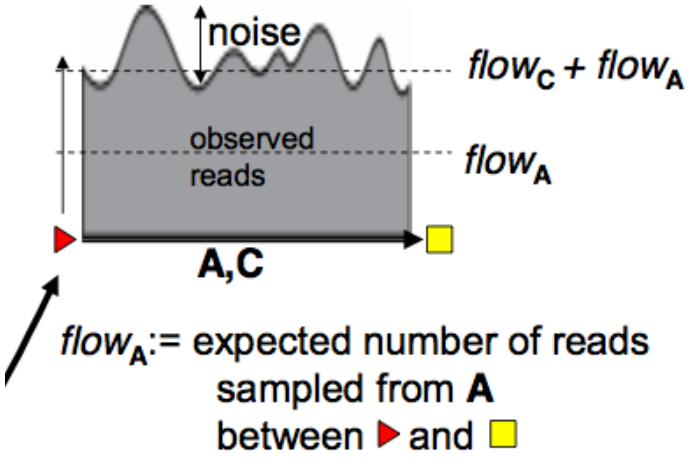


Read density bias along transcripts









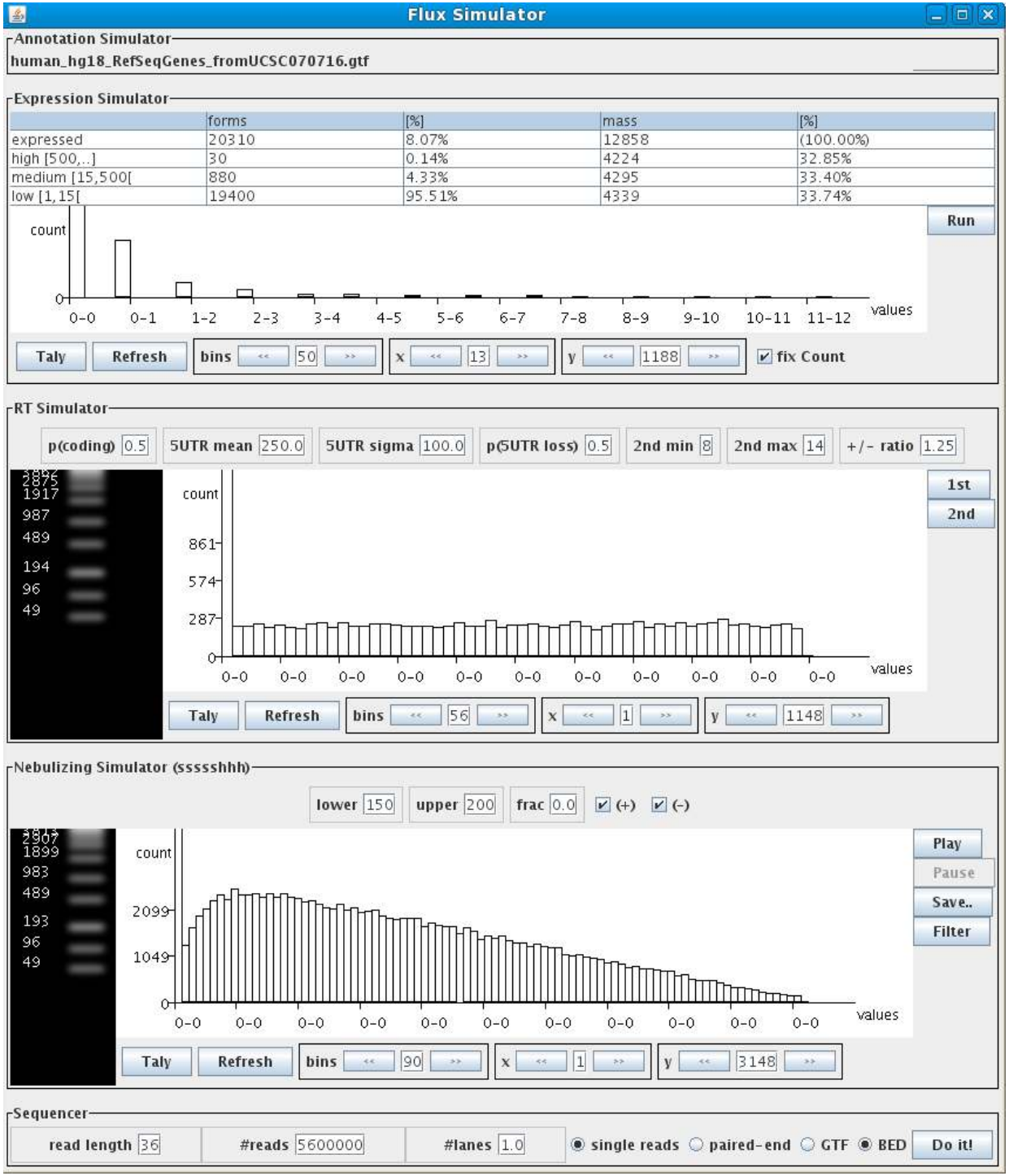
RNAseq simulator

First, given a transcript annotation, and a model of transcript abundance

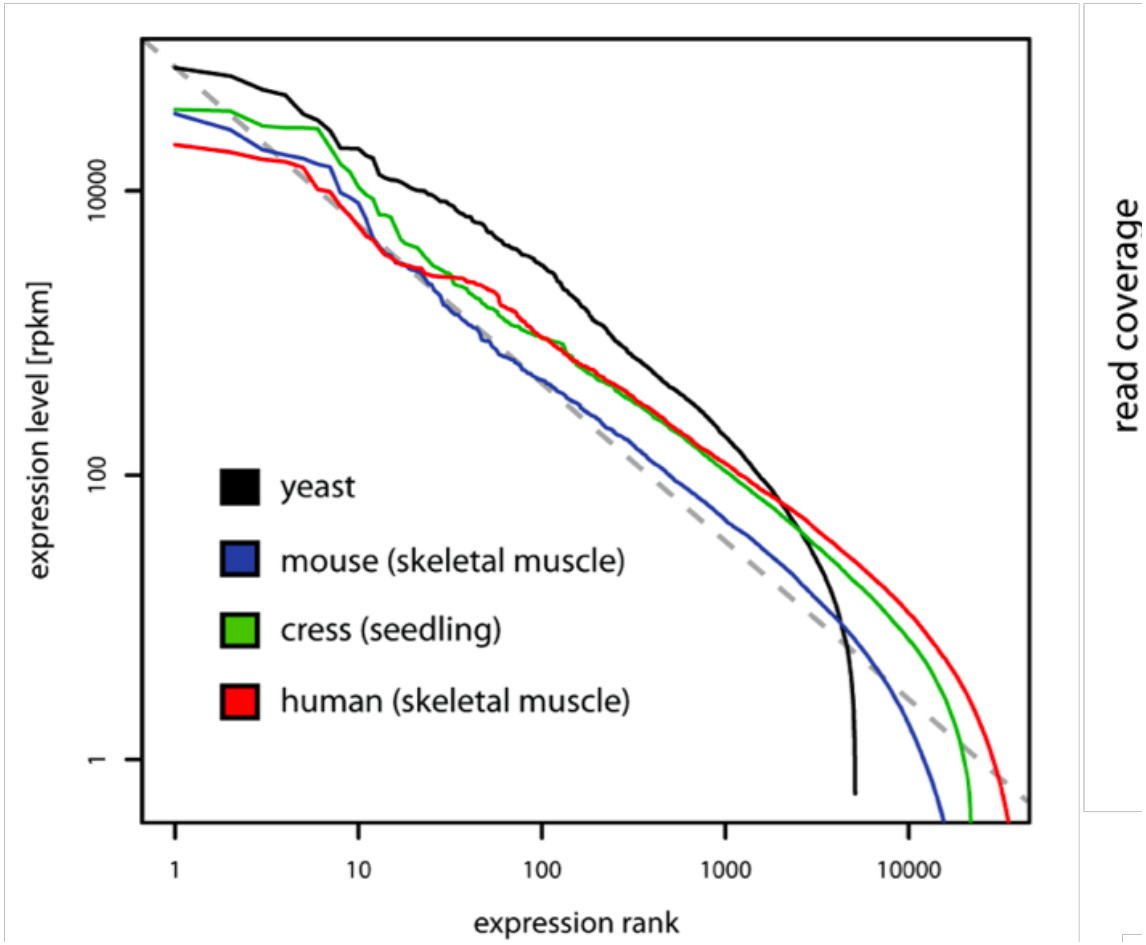
Then, the simulator produces a transcriptome (a set of transcripts with abundances attached)

Second, it simulates the different steps of the sequencing experiment (RT, nebulization, etc)

To **produce** a set of sequence reads as realistic as possible



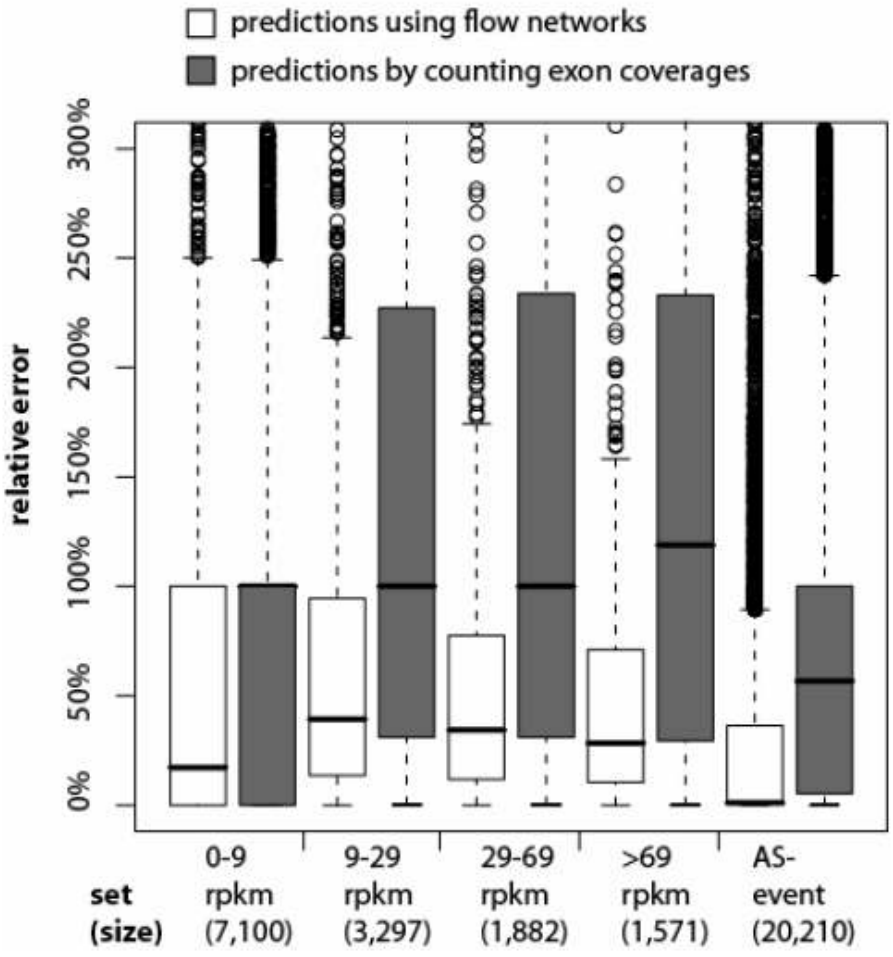
RNAseq simulator; simulating the transcriptome



RPKM (Mortazavi et al.)

- Reads Per Kilobase per Milion mapped reads.
- Example
  - Experiment generated 3,000,000 mapped reads
  - 200 reads map in a region (exon) 500 bp long
  - The RPKM of the region (exon) is
    - (200 x 2 = 400 reads per kilobase) / 3 = **133.33**

Quantitative assessment



relative error :=

$$\frac{|\, pred_cov - ref_cov \,|}{ref_cov}$$

resp. 100% for false positives (i.e., *ref_cov* = 0).

pred_cov = predicted coverage

ref_cov = reference coverage

rpkm = reads per kilobase per million mapped reads (coverage measure)

welcome trust sanger institute

Search

Information | Projects | Other Services

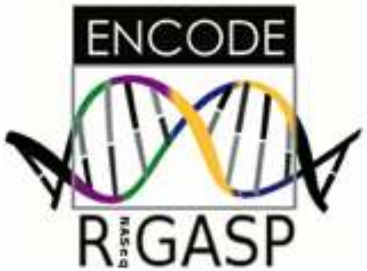
Functional Genomics
Human (HGP)
Pathogens
Blast

ENCODE
Home
Data Management
Participants
News
RGASP
Pilot Phase

Website Search
People Search
Library Services
Site Map
Feedback / Help

Hide Navigation

RGASP - The RNAseq Genome Annotation Assessment Project



Introduction

Following the successful format of the EGASP workshop in 2005, the RNAseq genome annotation assessment project is being launched to assess the current progress of automatic gene building using RNAseq as its primary dataset.

The main aim of the workshop is to **assess the status of computational methods to map human RNAseq data, assemble them into transcripts and quantify the abundance of that transcript in particular datasets**.

Transcript predictions will be evaluated against the GENCODE annotation produced as part of the ENCODE project. Special attention/assessment will be given to newly manually annotated chromosomes not previously publicly released before the workshop release. Promising transcript predictions not covered by Gencode annotation will be validated by experimental methods

The project is open to all researchers, but places are limited.
We will also include modEncode drosophila data and would like to encourage participants working on model organisms to participate.
More details will be posted on this website.

Dates

Registration deadline:	03.07.2009
Submission of predictions:	21.09.2009 (Extended)
Workshop days:	10. & 11.11.2009

Overall correlation of transcript-level human RNA-seq quantitation with nanostring

Ali Mortazavi, RGASP

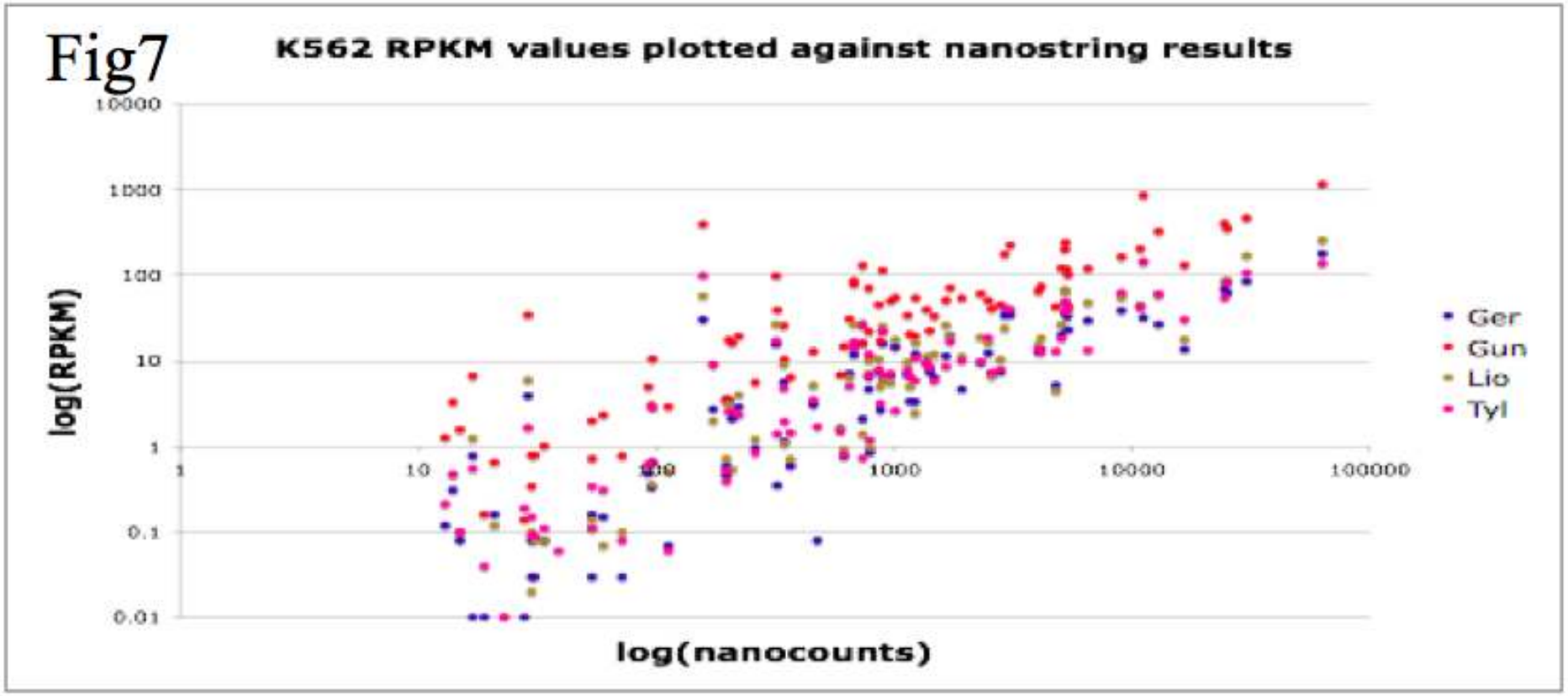
All probes (141)

R^2 with nanostring	Gerstein	Raetsch	Pachter	Guigo	Solyvyev
HepG2	0.49	0.55	0.47	0.66	0.29
GM12878	0.51	0.57	0.48	0.51	-
K562	0.59	0.65	0.55	0.44	-

Only within design T_m probes (105)

sample	Gerstein	Raetsch	Pachter	Guigo	Solyvyev
HepG2 (R^2 pearson)	0.65	0.65	0.63	0.78	0.31
GM12878 (R^2 pearson)	0.56	0.63	0.55	0.55	-
K562 (R^2 pearson)	**0.88**	**0.73**	**0.81**	**0.59**	**-**
K562 (rho spearman)	**0.87**	**0.85**	**0.86**	**0.86**	

Overall correlation of transcript-level human RNA-seq quantitation with nanostring



Overall correlation of transcript-level worm L3 RNA-seq quantitation with nanostring

All probes (157)

R^2 with nanostring	Raetsch	Stanke	Guigo	Solyvyev
Caltech L3	0.49	0.55	0.59	0.27

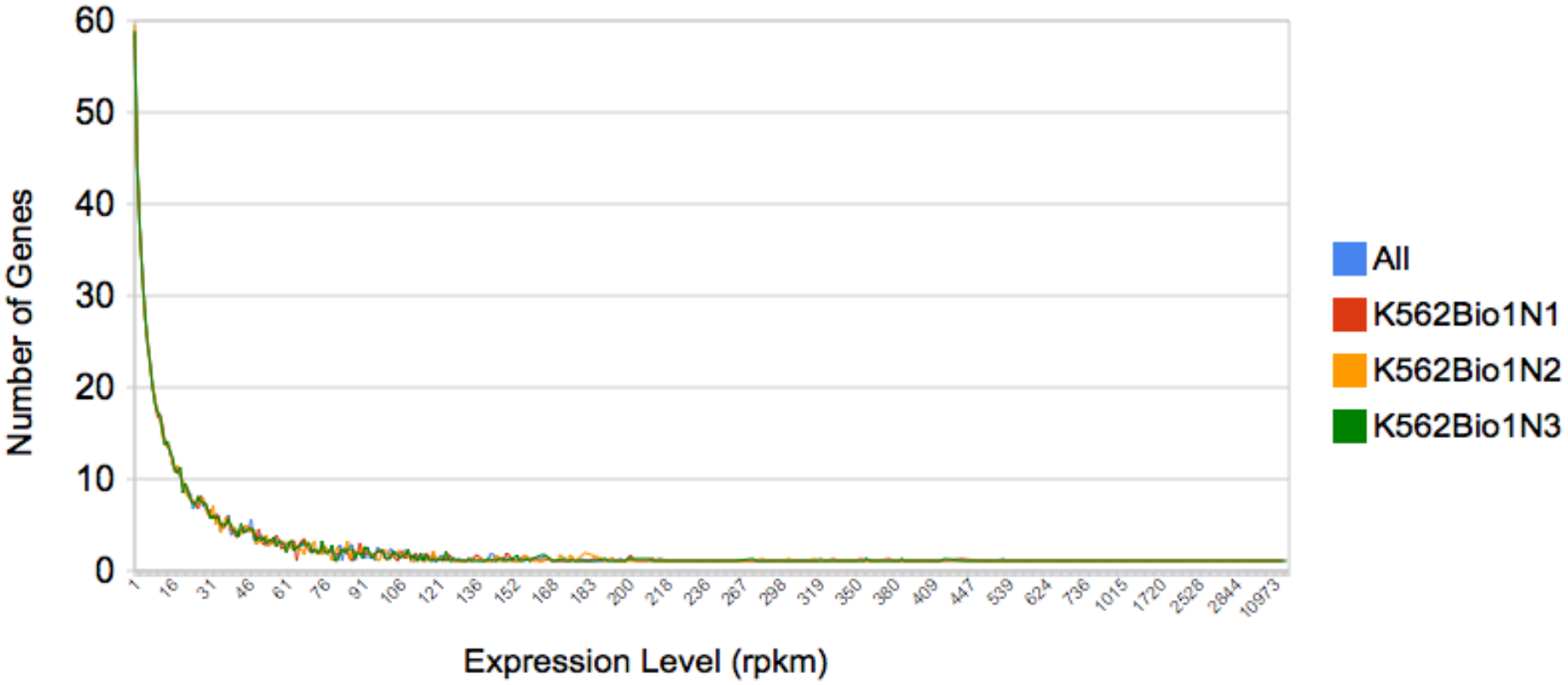
Only within design T_m probes (145)

Caltech C.e. L3	Raetsch	Stanke	Guigo	Solyvyev
Pearson R^2	0.49	0.55	0.59	0.27
Spearman rho	0.65	0.69	0.65	0.77

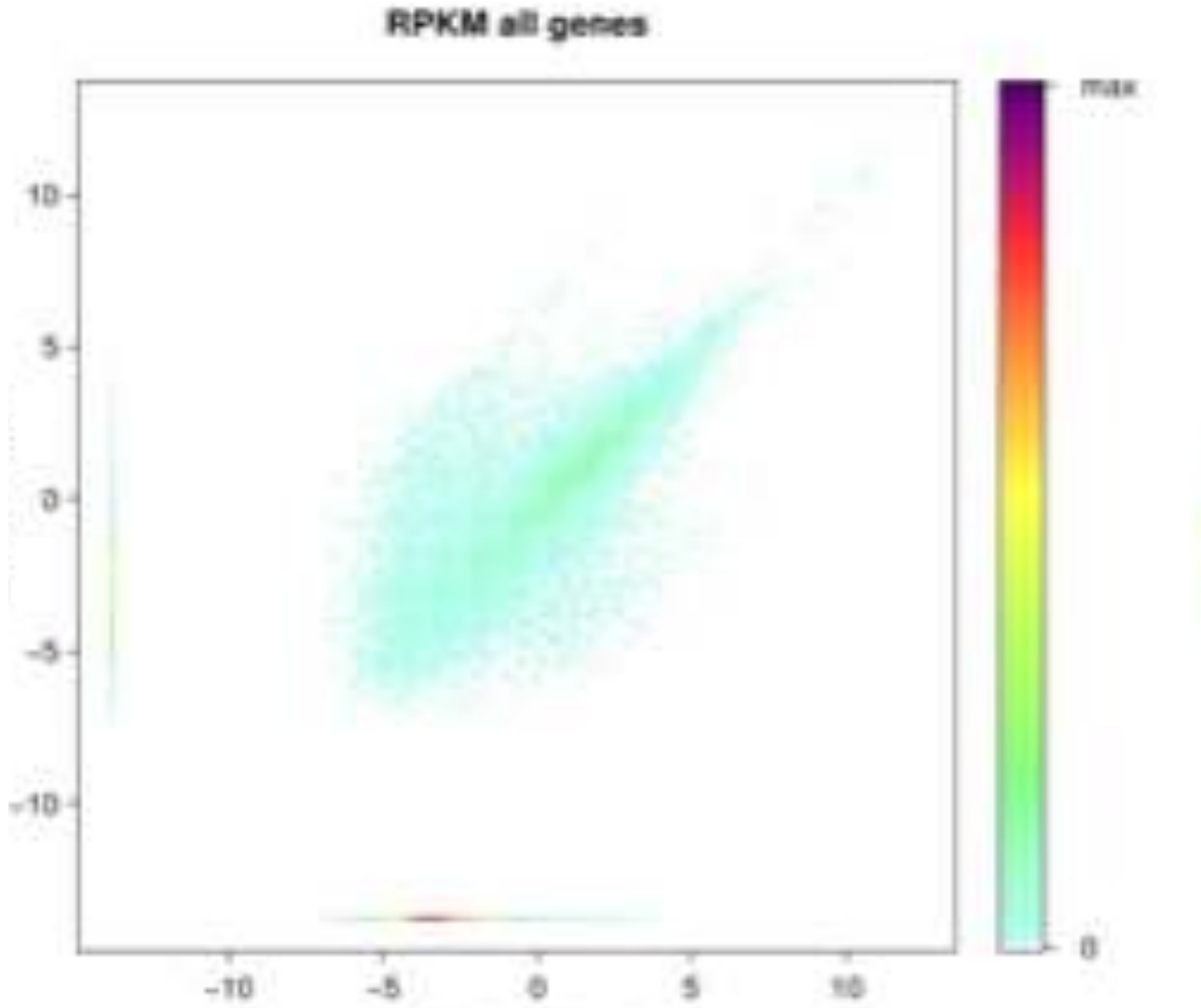
Gene expression

Gene Id	Length	Strand (+/-)	Locus	# Exons	# Transcripts	Expression Value Total	Expression Value Lane 1	Expression Value Lane 2	Expression Value Lane 3
ENSG00000210082	1,558	+	M	1	1	50,059	16,931	16,174	16,954
ENSG00000198804	1,542	+	M	1	1	30,479	10,973	8,518	10,988
ENSG00000198712	684	+	M	1	1	15,130	5,268	4,593	5,269
ENSG00000239846	366	-	1	1	1	10,808	3,607	3,596	3,605
ENSG00000244516	402	-	8	1	1	10,098	3,340	3,274	3,484
ENSG00000198886	1,378	+	M	1	1	8,384	2,816	2,724	2,844
ENSG00000198938	781	+	M	1	1	7,865	2,773	2,304	2,788
ENSG00000199916	264	-	9	1	1	7,741	2,558	2,628	2,555
ENSG00000222855	616	+	21	1	1	7,713	2,629	2,591	2,493
ENSG00000240831	165	+	12	1	1	7,532	2,544	2,528	2,460
ENSG00000211459	954	+	M	1	1	5,801	2,020	1,753	2,028
ENSG00000212907	297	+	M	1	1	5,582	1,911	1,817	1,854
ENSG00000202198	331	+	6	1	1	5,170	1,999	1,179	1,992
ENSG00000239907	510	+	X	1	1	4,572	1,479	1,555	1,538
ENSG00000232673	135	+	1	1	1	4,237	1,402	1,432	1,403
ENSG00000198727	1,135	+	M	1	1	3,916	1,307	1,303	1,306
ENSG00000223073	777	+	M	1	1	3,044	1,004	1,018	1,022
ENSG00000198786	1,812	+	M	1	1	2,583	856	848	879
ENSG00000226958	432	-	X	1	1	2,275	720	783	772
ENSG00000243172	293	+	11	1	1	2,274	736	774	764

Gene expression



Differential gene expression



Condition 2

Condition 1

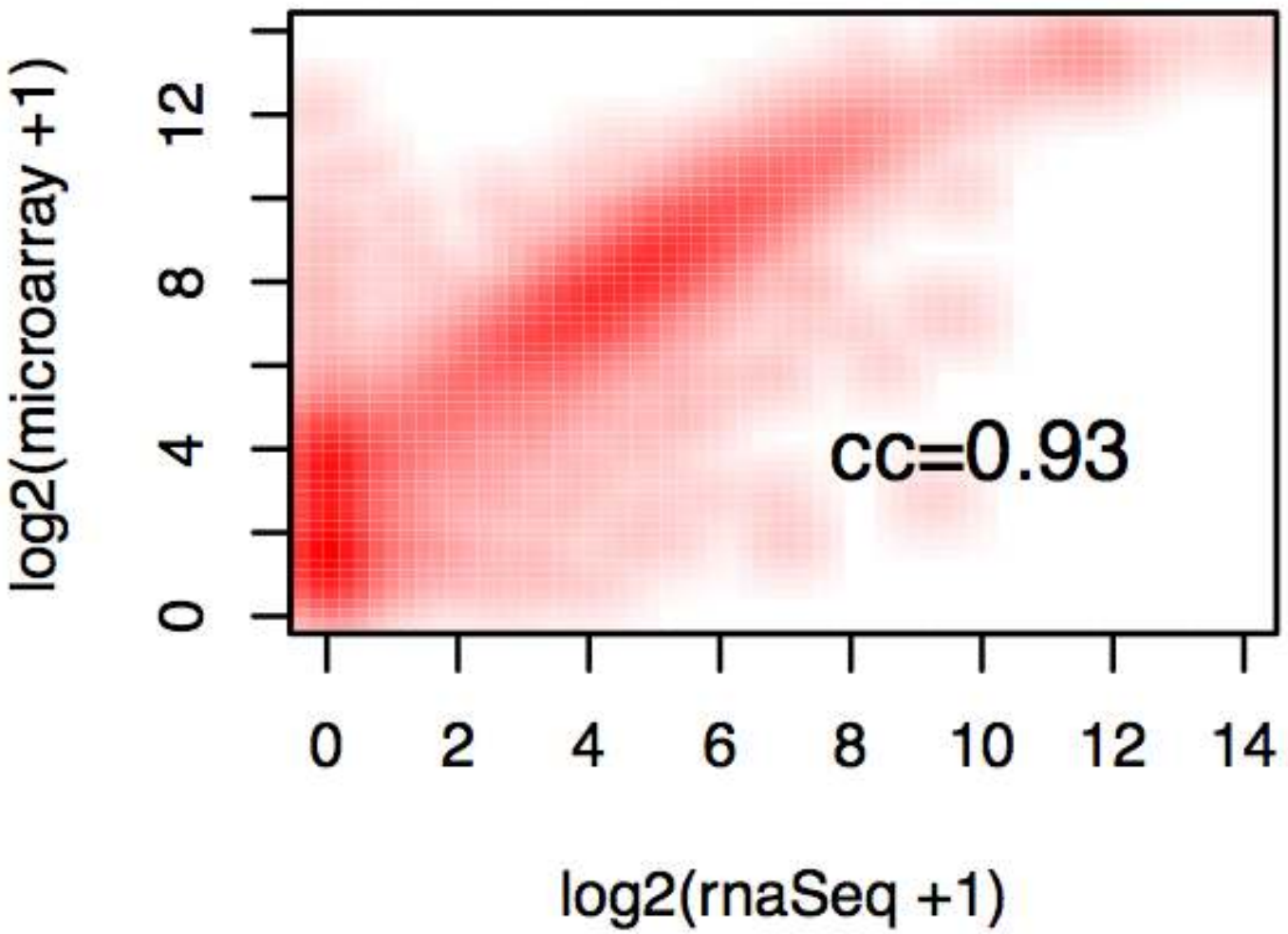
Angelika Merkel

RNASeq vs microarrays

Fly wing
imaginal discs

80M reads
50x2

M Corominas
H Tilgner

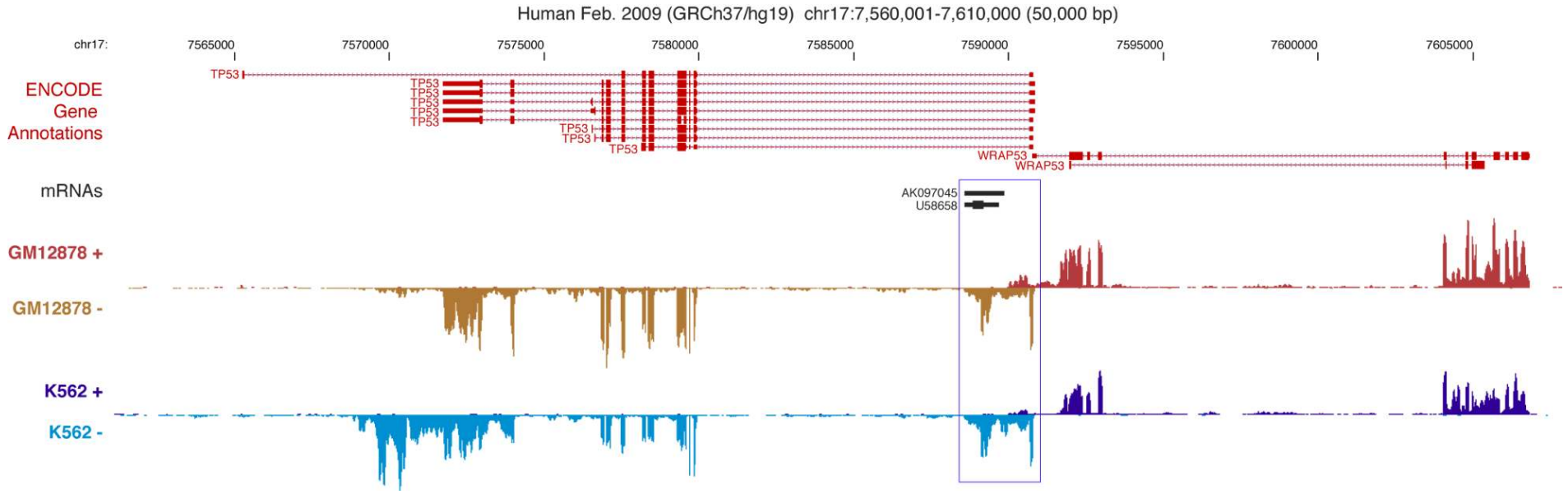


Transcript expression

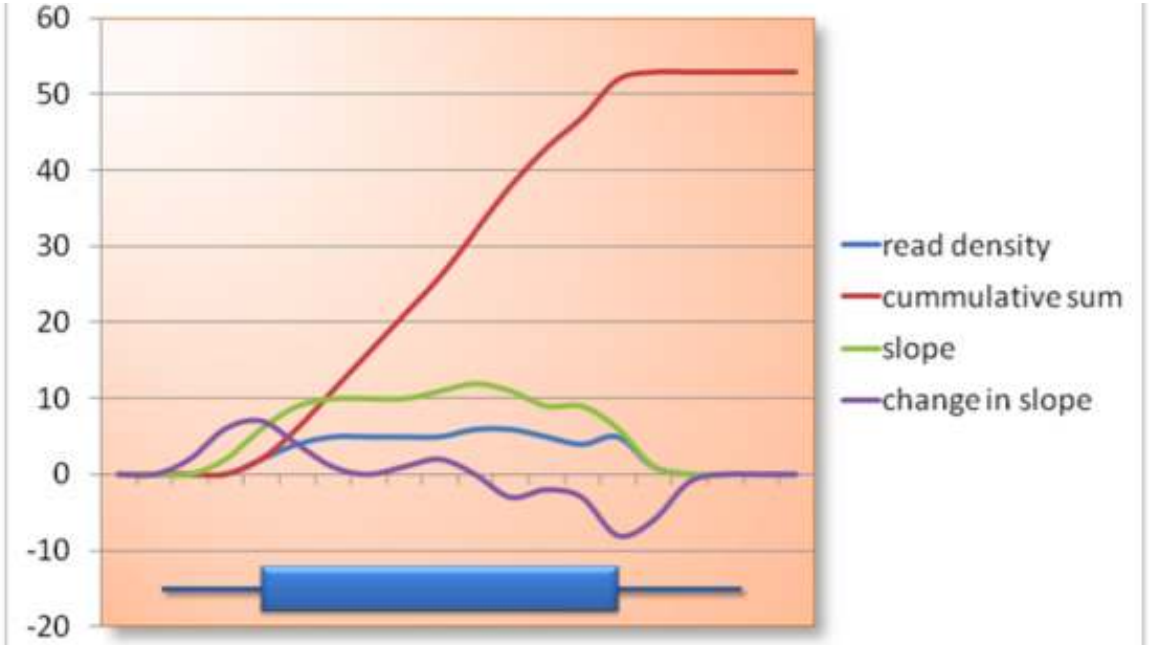
Gene Id	Length	Strand (+/-)	Locus	# Exons	Expression Value Total	Expression Value Lane 1	Expression Value Lane 2	Expression Value Lane 3
ENST00000387347	1,558	+	M	1	740,572	245,922	246,997	247,653
ENST00000361624	1,542	+	M	1	581,997	193,103	194,930	193,964
ENST00000410923	616	+	21	1	251,187	82,466	85,545	83,176
ENST00000361739	684	+	M	1	240,665	79,991	80,209	80,465
ENST00000495862	861	-	X	2	204,293	0	117,916	86,377
ENST00000491213	510	+	X	1	178,827	57,194	60,738	60,895
ENST00000414273	1,543	+	1	1	170,239	56,171	57,224	56,844
ENST00000493956	402	-	8	1	167,007	52,934	56,579	57,494
ENST00000363046	264	-	9	1	158,774	52,384	53,999	52,391
ENST00000365328	331	+	6	1	145,766	47,484	49,269	49,013
ENST00000362079	781	+	M	1	137,151	45,215	46,014	45,922
ENST00000389680	954	+	M	1	130,699	43,246	43,891	43,562
ENST00000361381	1,378	+	M	1	123,634	40,770	41,517	41,347
ENST00000363673	317	+	16	1	107,694	34,655	36,537	36,502
ENST00000339647	3,384	-	12	2	102,525	33,825	34,221	34,479
ENST00000463737	332	+	11	1	96,302	30,476	32,947	32,879
ENST00000419932	432	-	X	1	93,870	29,423	32,167	32,280
ENST00000436459	1,744	+	9	1	78,337	27,699	25,311	25,327
ENST00000427426	682	+	1	1	70,127	22,695	23,733	23,699
ENST00000361335	297	+	M	1	69,381	23,142	23,047	23,192

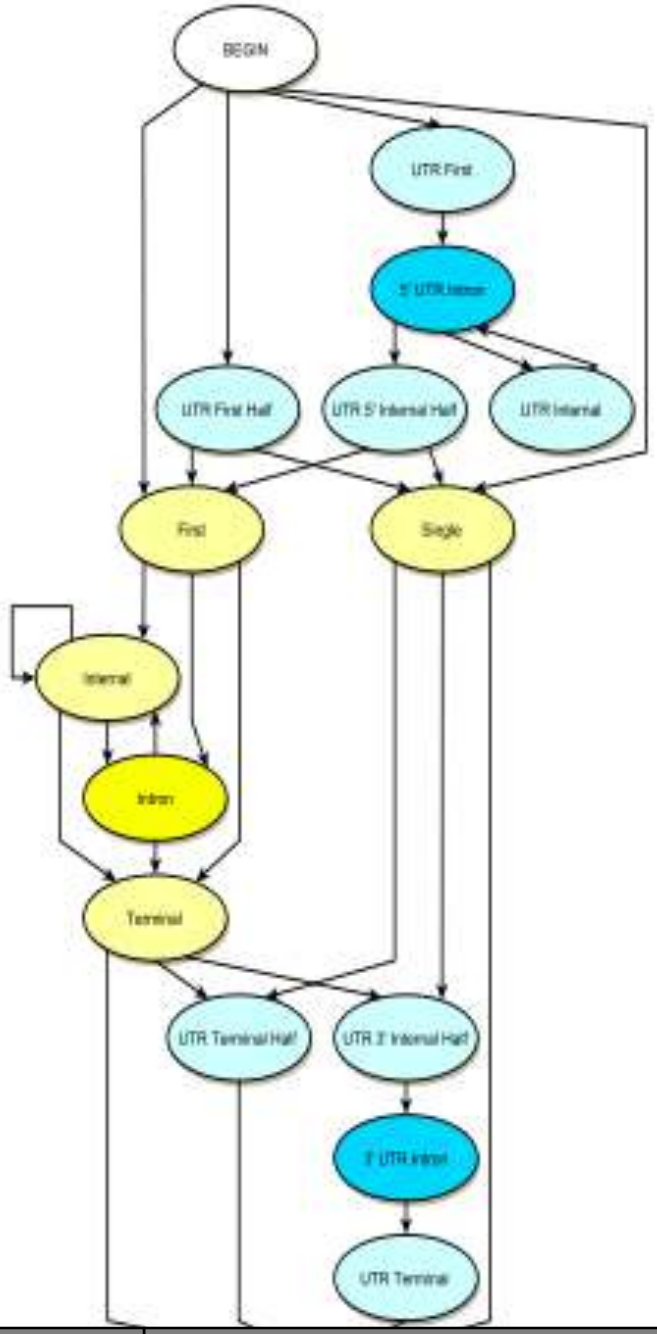
TRANSCRIPT DISCOVERY

Evidence for additional transcription



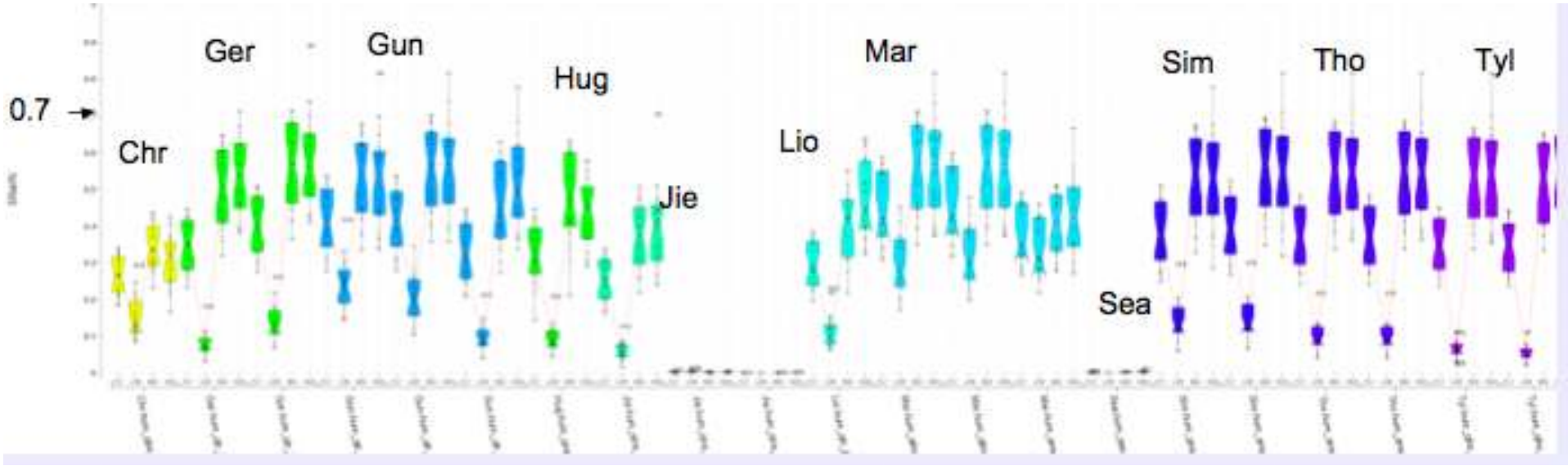
NextGeneId,
Tyler Alioto





RGASP

Intron and exon accuracy of
RNASeq based gene modelers againts GENCODE



Summary

- RNASeq is a powerful tool for transcriptome characterization
  - It may become standard within the near future
- Still, many issues need to be addressed at the level of experimental protocols and bioinformatic/quantitative analysis
- Technologies are in flux, and they may become soon obsolete.
- For large-scale expression analysis, microarrays may still be a cost-effective alternatives.

acknwoledgements

- Paolo Ribeca (CNAG), GEM
- Micha Sammeth (CNAG), FluxCapacitor
- Tyler Alioto (CNAG), NextGeneId

Maik Roder
David Gonzalez

ENCODE Project

Projects • ENCODE

Ging001N

Overview | Read Summary | Mapping Summary | Expression Summary | Splicing Summary

ABOUT

Long RNA poly-A+ RNAseq data from K562 Nucleus. K562 is an immortalized cell line produced from a female patient with chronic myelogenous leukemia (CML). It is a widely used model for cell biology, biochemistry, and erythropoiesis. It grows well, is transfectable, and represents the mesoderm linage. Cells will be obtained from the America Type Culture Collection (ATCC) [atcc.org] (ATCC Number CCL-243). This is the first of Two Biological replicates the second being Ging002N. It contains Trhess lanes of paired end data

Species: **Homo sapiens**

READ SUMMARY

Overview of the reads that were mapped in this experiment.

- Unique reads shows the number of unique sequences in the set, reflecting the redundancy of the data.
- Unambiguous reads are those reads for which every base was resolved by the sequencer.
- Ambiguous reads are those reads for which at least one nt was not resolved by the sequencer (assigned an N).

	Total	Percent
Unique Reads	158,210,524	67.5%
Unambiguous Reads	228,028,503	97.4%
Ambiguous Reads	6,198,741	2.6%

MAPPING SUMMARY

Overview of the results after mapping.

- Uniquely mapped reads are those reads for which we can find a unique alignment that is better (with a lower number of mismatches) than any of the other possible alignments.
- Multi mapped reads shows the number of reads for which we find 2 or more equally good possible alignments.
- Unmapped reads are those reads for which we were unable to find an alignment within the parameters used in the mapping.

	Total	Percent
Uniquely Mapped Reads	154,185,171	76.0%
Multi-Mapped Reads	33,227,498	16.4%
Unmapped Reads		