L3.4 Transcriptomes (Beyond RNA-seq)

AGENDA

1. Some conclusion from studies done using the approaches discussed In the last couple of days:
    a- New concept of gene
    b- RNA catalogue

2. Accessing transcriptomic data

3. Other techniques beyond RNA seq

Textbook

Perspective

Functional transcriptomics in the post-ENCODE era

Jonathan M. Mudge,[1] Adam Frankish, and Jennifer Harrow

Department of Informatics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom

The last decade has seen tremendous effort committed to the annotation of the human genome sequence, most notably perhaps in the form of the ENCODE project. One of the major findings of ENCODE, and other genome analysis projects, is that the human transcriptome is far larger and more complex than previously thought. This complexity manifests, for example, as alternative splicing within protein-coding genes, as well as in the discovery of thousands of long noncoding RNAs. It is also possible that significant numbers of human transcripts have not yet been described by annotation projects, while existing transcript models are frequently incomplete. The question as to what proportion of this complexity is truly functional remains open, however, and this ambiguity presents a serious challenge to genome scientists. In this article, we will discuss the current state of human transcriptome annotation, drawing on our experience gained in generating the GENCODE gene annotation set. We highlight the gaps in our knowledge of transcript functionality that remain, and consider the potential computational and experimental strategies that can be used to help close them. We propose that an understanding of the true overlap between transcriptional complexity and functionality will not be gained in the short term. However, significant steps toward obtaining this knowledge can now be taken by using an integrated strategy, combining all of the experimental resources at our disposal.

23:1961–1973 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

TRANSCRIPTIONAL COMPLEXITY

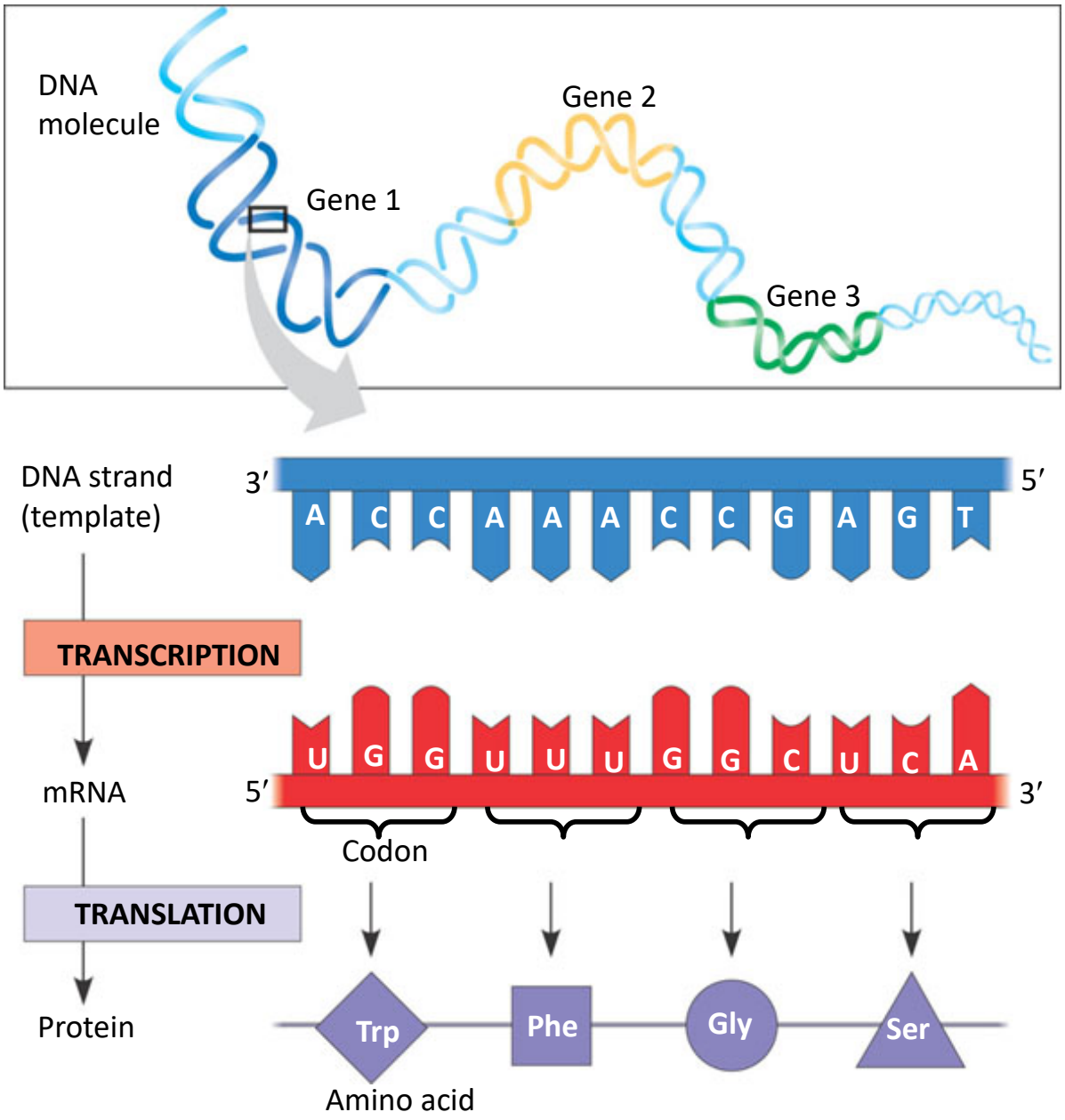
"…Over the 20th Century, the biological definition of the gene evolved from "the site of a hereditable trait" to "the genomic region from where the mRNA that encodes a protein is transcribed," i.e., the "central dogma" of molecular biology (Fig. 1A; Crick 1970). In the 21st Century, however, our view of transcription is becoming more complicated….

….a locus may generate multiple transcripts due to alternative splicing (AS) and read-through transcription, while the discovery of long noncoding RNAs (lncRNAs) suggests that most human transcripts may not encode proteins). In fact, the bulk of the genome appears to be "pervasively" transcribed (The ENCODE Project Consortium 2012), although the functional relevance of this process remains a source of debate (Ball 2013; Doolittle 2013; Graur et al. 2013).

We use the term "transcriptional complexity" to refer to these phenomena collectively…."

From Mudge et al., 2013

From Gene to Phenotype



Transcriptomics Aim: identify the full set of transcripts

→ Identification of distict transcripts biotypes

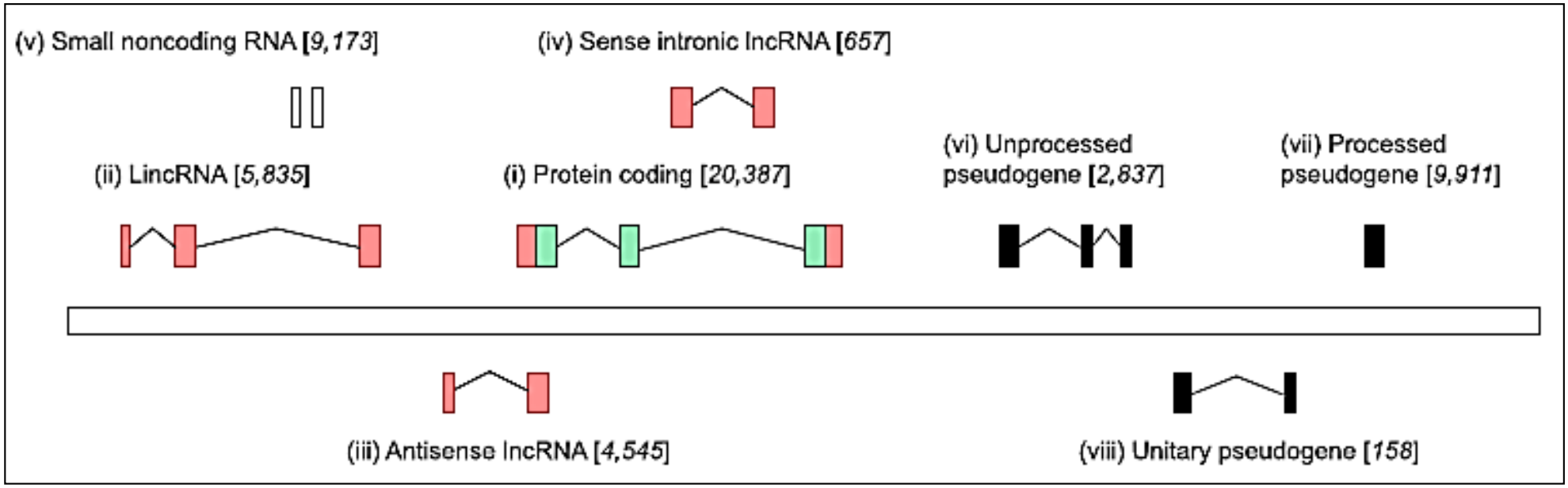


Figure 2. A summary of locus biotypes in GENCODE.

From Mudge et al., 2013

Table 1 | **Summary statistics of human and mouse genomes and gene sets**

Characteristic	Human (GRCh38)	Mouse (GRCm38)
Genome size (nt)	3,088,269,832	2,725,521,370
Unplaced scaffolds (nt)	11,464,317	5,334,105
Number of chromosomes	22 + X + Y	19 + X + Y
Chain alignments (nt)	2,735,135,097	2,465,275,732
Number of genes	58,037	48,709
Number of transcripts	198,093	118,925
Number of protein-coding genes	19,950	22,018
Number of protein-coding 1 to 1 orthologues	15,893	
Number of protein-coding transcripts	80,087	52,382
Number of lncRNA genes	15,767	9,989
Number of lncRNA orthologues	851 (REF. 22), 1,100 (REF. 35), 1,587 (REF. 36), 2,720 (REF. 34)	
Number of lncRNA transcripts	27,692	13,904
Number of pseudogenes	14,650	10,096
Number of small RNAs	7,258	6,110
Number of miRNAs	2,588	1,915
Number of snRNAs	1,900	1,383
Number of snoRNAs	944	1,508
Number of tRNAs	631	471

Annotation counts were retrieved from the GENCODE website (v25 for human and vM11 for mouse). The number of microRNAs (miRNAs) was obtained from miRBase v21 (REF. 48). The number of transfer RNAs (tRNAs) was obtained from GtRNAdb[51]. The number of protein-coding orthologues was taken from Ensembl Compara[23] (v86), and the numbers of orthologous long non-coding RNAs (lncRNAs) were obtained from different sources[22,34–36]. snoRNAs, small nucleolar RNAs; snRNAs, small nuclear RNAs.

Breschi et al., Nature Review Genetics, 2017

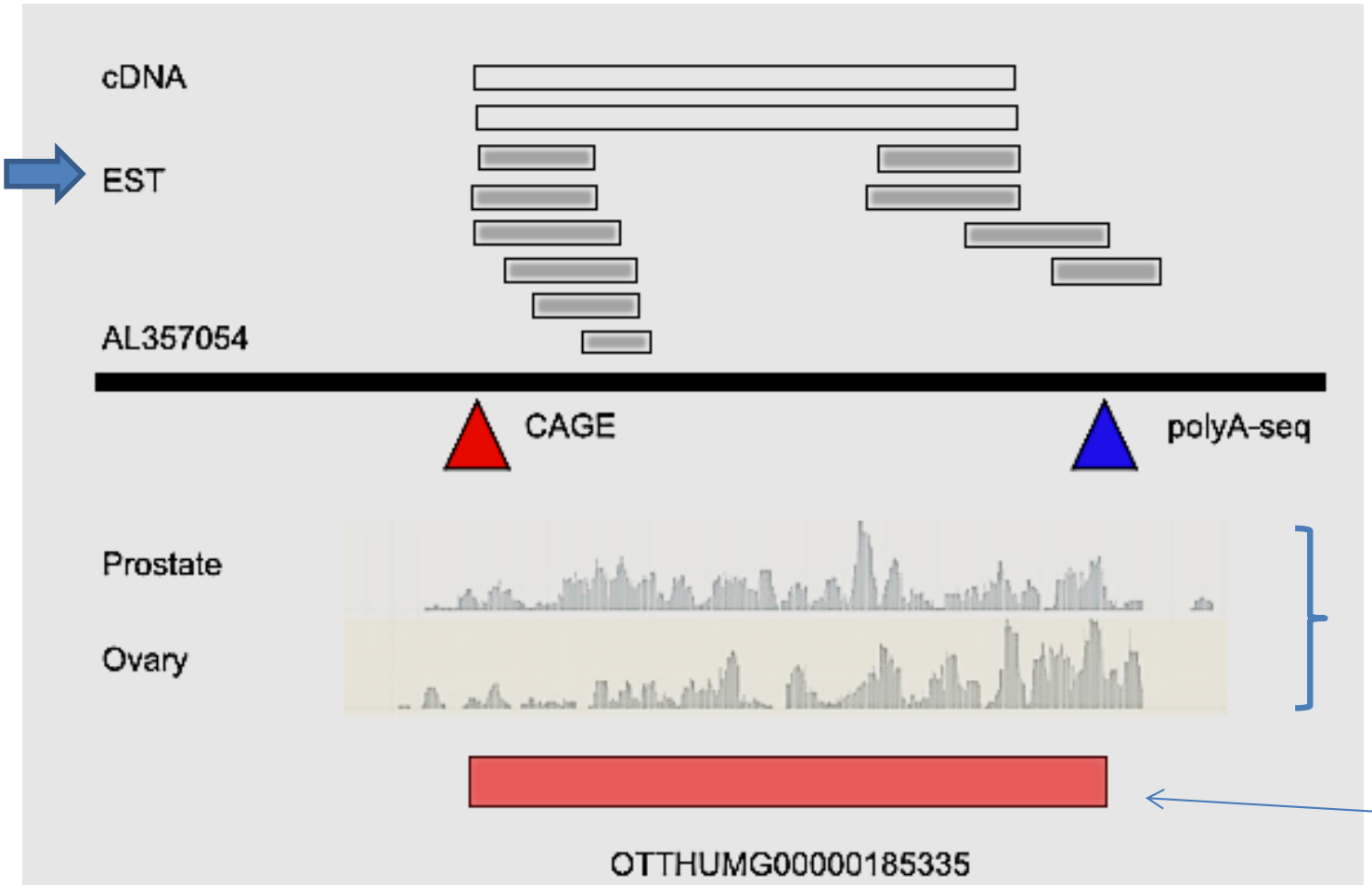
ARE WE DONE YET?

"… we must recognize that transcriptomes can differ significantly between the cells of distinct tissues and developmental stages, in terms of both the transcripts produced and their levels of expression …… splicing abnormalities are commonly observed in cancer cells and immortalized cell lines…. Finally, it is becoming apparent that AS patterns can show notable polymorphism ….

While there may be no single human transcriptome, it will often make practical sense to work with a "consensus" transcriptome that combines all known transcripts into one gene set. Even then, the question of missing transcripts is difficult to answer…."

Need to use combination of approaches to move forward
(see example next slide)

https://www.ncbi.nlm.nih.gov/nuccore/



AX747750, AK092822

DA816101, DA923061, DA427401, BG541155, AI831721, AW135930, CB052137, DB332016, CB052136, AA004346

EST

BAC AL357054

RNA-Seq

Newly defined lncRNA

Figure 4. The annotation of a novel lncRNA locus in GENCODE

https://www.ncbi.nlm.nih.gov/pubmed/19277571
https://www.ncbi.nlm.nih.gov/nucest/

Mudge, 2013

«Classical» versus modern view of a gene : what is a «gene»?

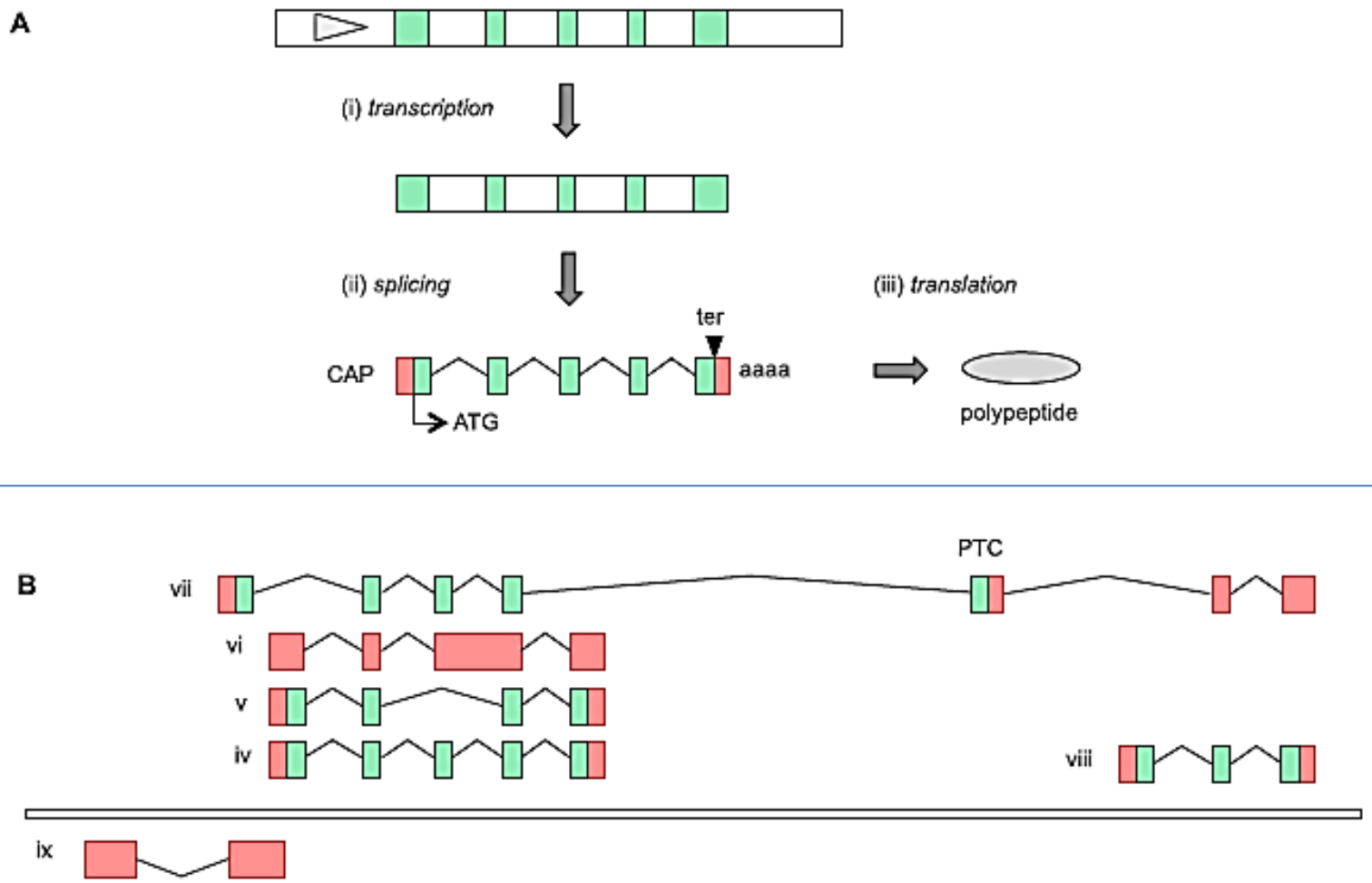
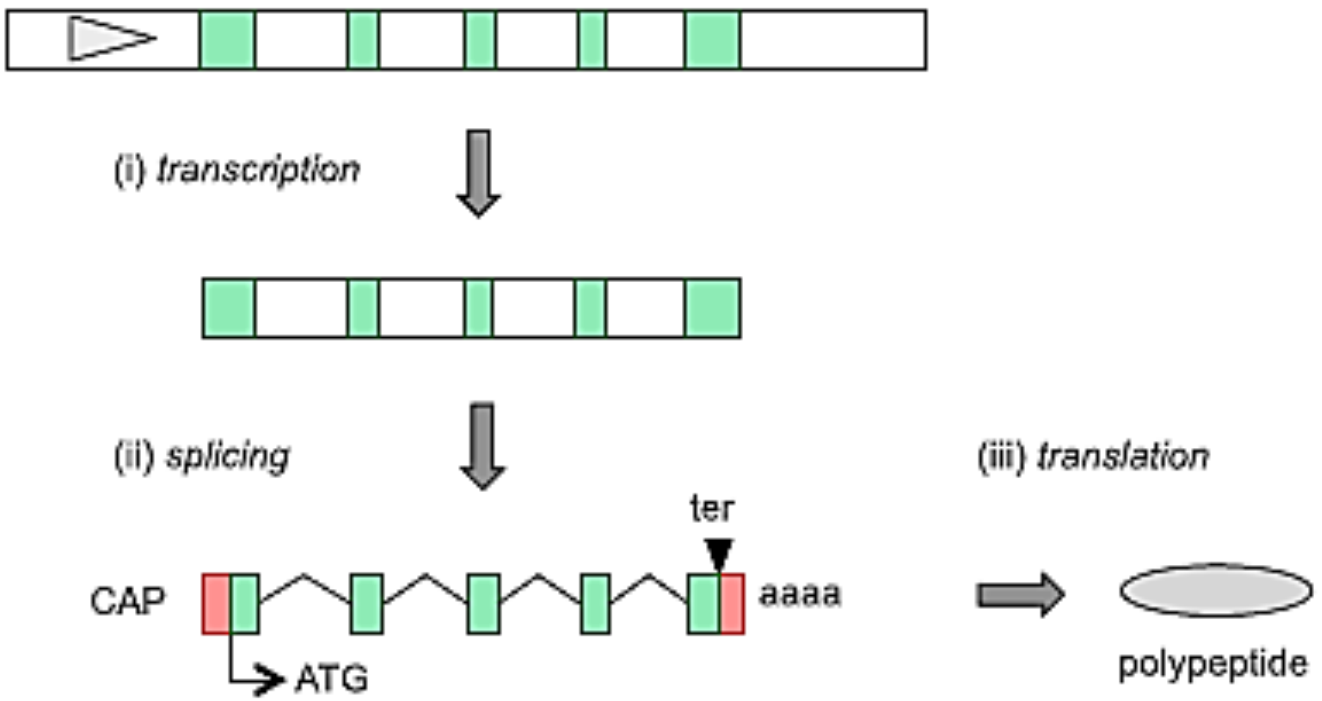


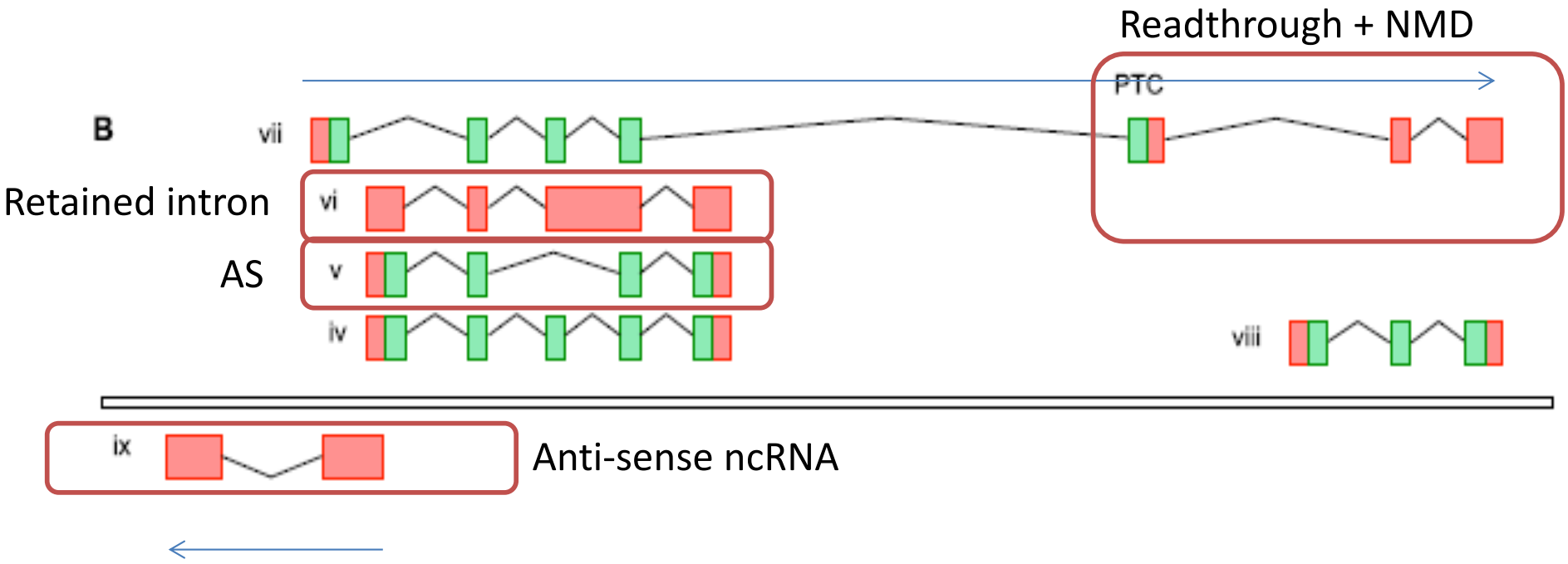
Figure 1. The evolving dogma of gene transcription.

From Mudge et al., 2013



(A) The **historical "central dogma"** of molecular biology. By this model, (i) transcription generates the primary transcript (exons in green, introns in white), with the initial interaction between the RNA polymerase complex and the genome being mediated by a promoter region (gray triangle). (ii) The introns of the primary transcript are removed by the spliceosome, and a mature mRNA is generated by 5' end capping (CAP) and polyadenylation (aaaa) (coding region [CDS] shown in green, untranslated 5' and 3' UTRs in red). (iii) The mRNA is translated into a polypeptide by the ribosome complex, with translation proceeding from the initiation codon (ATG) and ending at the termination codon (ter).

From Mudge et al., 2013

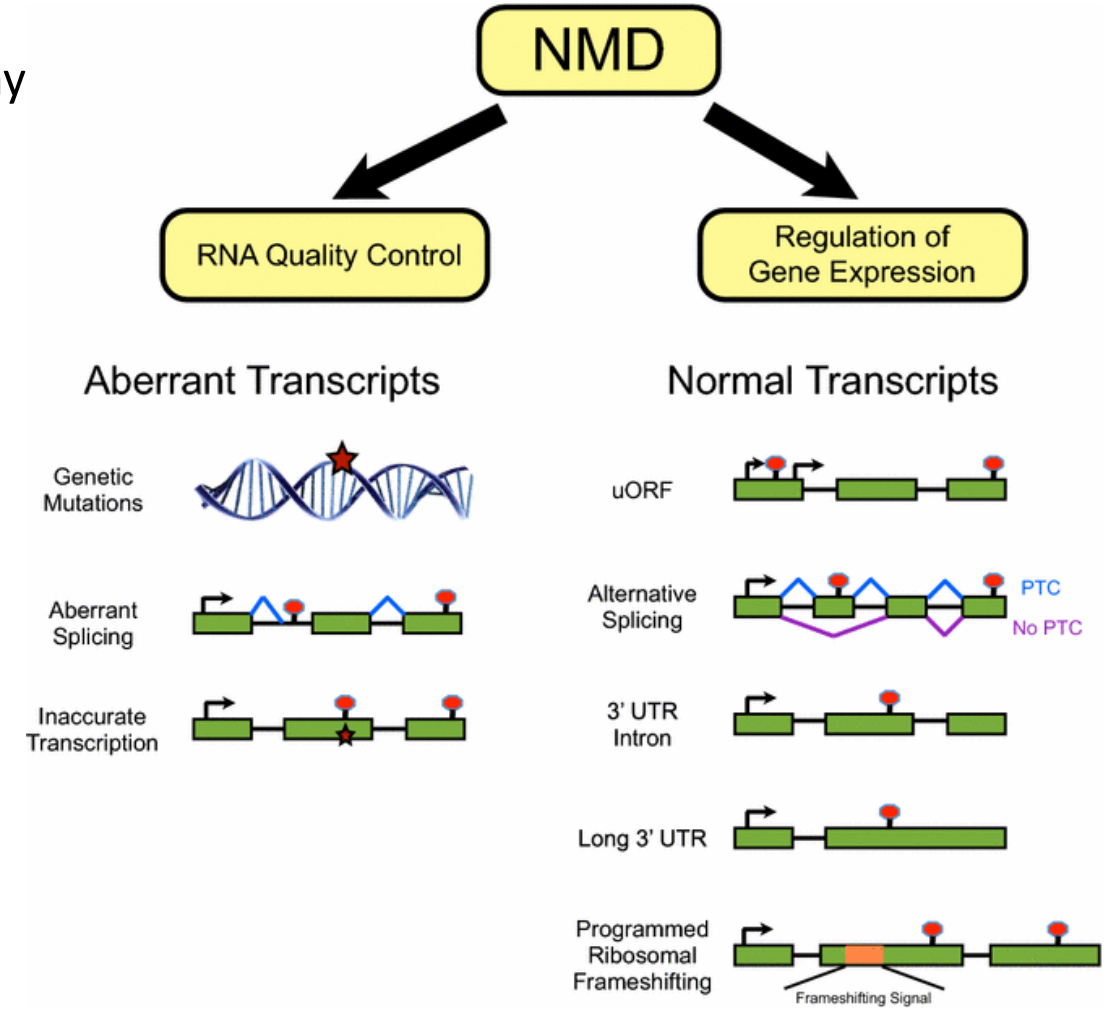


(B) An **updated model** reflecting a modern view of transcriptional complexity. Here, the same gene (iv) undergoes alternative splicing (AS), for example an exon skipping event that does not change the frame of the CDS (v); this event thus has the potential to generate an alternative protein isoform. However, products of AS cannot be assumed to be functional; this gene has generated a retained intron transcript (vi), perhaps due to the failure of the spliceosome to remove this intron. Further complexity comes from a read-through transcription event (vii), whereby a transcript is generated that also includes exons from a neighboring protein-coding locus (viii). In this example, the read-through transcript has an alternative first exon compared with the upstream gene that contains a potential alternative ATG codon, although the presence of a subsequent premature termination codon (PTC) prior to two splice junctions indicates that this transcript is likely subjected to the nonsense mediated decay (NMD) degradation pathway. Finally, model ix is a transcript that is antisense to the upstream gene; both loci are potentially generated under the control of a bidirectional promoter.

From Mudge et al., 2013

How much of the transcriptome is functional?

i.e. Non sense-mediated decay



Long transcripts

- Protein coding
- lncRNA antisense to protein-coding or other noncoding genes
- lncRNA “divergent” from a protein-coding or noncoding gene
- LincRNA = long intergenic RNA
- eRNA = RNA transcribed form enhancers in both directions (Chapter 4)
- circRNA = circular RNAs
- Other (pseudogene transcripts, RNA from transposable elements, etc)

How is the classification «noncoding» attributed ?

- ORF search in all the possible frames
- Short ORFs evaluated on «codon usage»
- Proteomic database interrogated
- Association with ribosomes (poly-ribosome purification and RNA-seq)

Annotated small RNAs

Currently, a total of 7,053 small RNAs are annotated by GENCODE, 85% of which correspond to four major classes: small nuclear (sn)RNAs, small nucleolar (sno)RNAs, micro (mi)RNAs and transfer (t)RNAs (Table 2a). Overall we find 28% of all annotated small RNAs

small RNA classes were enriched in those compartments where they are known to perform their functions: miRNAs and tRNAs in the cytosol, and snoRNAs in the nucleus. Interestingly, snRNAs were equally abundant in both the nucleus and the cytosol. When specifically interrogating the subnuclear compartments of the K562 cell line, however, snRNAs seem to be present in very high abundance in the chromatin-associated RNA fraction (Supplementary Fig. 14b, c). This striking enrichment is consistent with splicing being predominantly co-transcriptional[16,25].

Unannotated short RNAs

We detected two types of unannotated short RNAs. The first type corresponds to subfragments of annotated small RNAs. Because we

The second and largest source of unannotated short RNAs corresponds to novel short RNAs (Table 2b) that map outside of annotated ones. Almost 90% of these are only observed in one cell line and are present at low copy numbers. Nearly 40% of these unannotated short RNAs are associated with promoter and terminator regions of annotated genes (promoter-associated short RNAs (PASRs) and termini-associated short RNAs (TASRs)), and their position relative to TSSs and transcription termination sites is similar to previous results[4].

Djebali et al., 2012

Stable short noncoding RNAs

snoRNAs	Small nucleolar RNAs	Small ncRNAs that guide chemical modifications of other non-coding RNAs
siRNAs	Small interfering RNAs	Double-stranded RNA molecules, 20–25 nucleotides in length, that act in various silencing pathways
miRNAs	microRNAs	Single-stranded RNA molecules of 21–24 nucleotides in length, which regulate gene expression
snRNA	Small nuclear RNAs	Nuclear small non-coding RNAs involved in various functions including splicing
piRNA	Piwi interacting RNAs	26–31 nt long RNAs involved in transcriptional gene silencing, including retrotransposons

tRNA tranfert RNA

Short name of RNA classes	Full name of RNA classes	Notes
PALRs	Promoter-associated long RNAs	Hundreds nt long RNAs spanning regions on proximal promoters to the first exon
PASRs	Promoter-associated short RNAs	20–70 nt long RNAs spanning regions around core promoters
TASRs	Termini-associated short RNAs	20–70 nt long RNAs spanning regions around transcription termination sites
PROMPTs	Promoter upstream transcripts	Unstable transcripts mapping 0.5–2 kb upstream the transcription starting sites
TSSa-RNAs	Transcription start sites antisense RNAs	RNAs, generally short and non-coding, generated from bidirectional activity of mammalian RNA Polymerase II
NRO-RNAs	Nuclear run-on assay derived RNAs	Short RNA detected by nuclear run-on assays, mapping 20 to 50 downstream to transcriptions starting sites of mRNAs
RE RNAs	Retrotransposon-derived RNAs	A heterogeneous class of RNAs which starting sites overlap retrotransposon elements
tiRNAs	Tiny transcription initiation RNAs	RNAs about 18 nt long, positioned about 20 bp after the transcription starting sites of highly expressed mRNAs

Unstable small RNA accompanying gene transcription

Exercise: visualize the complexity of a specific locus using online databases

Example: human p16 locus (Chr 9p21)

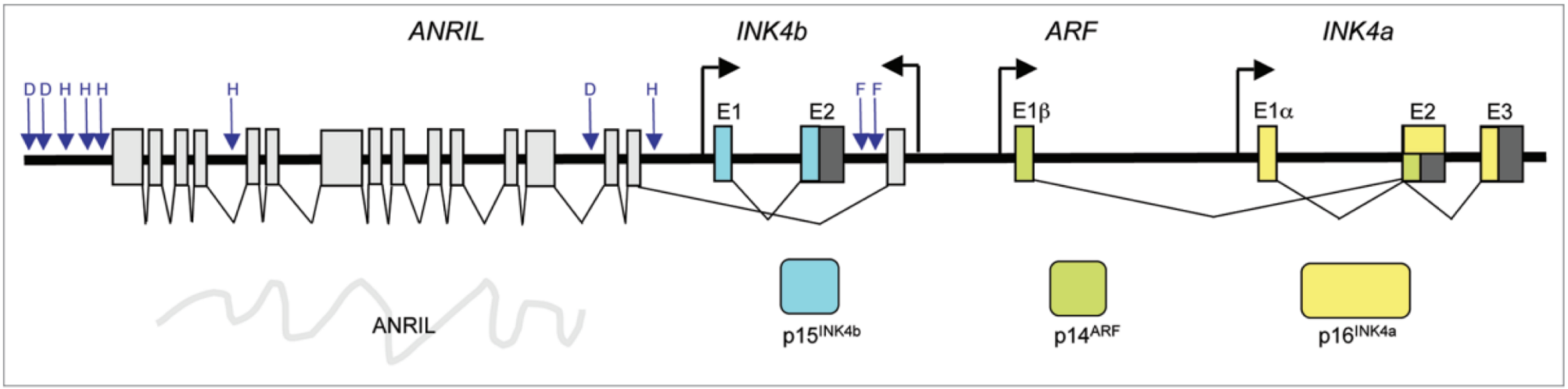
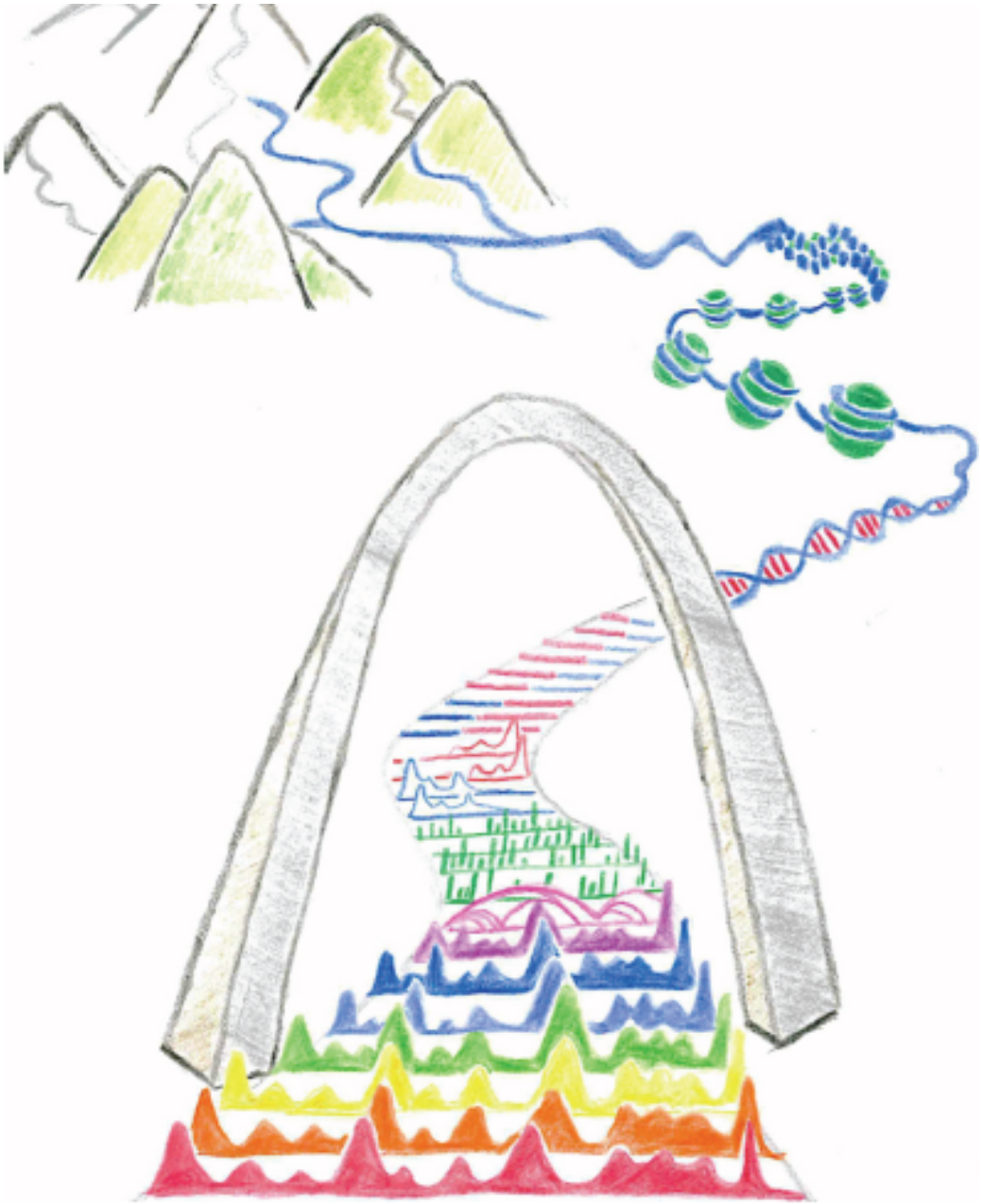


Figure 1. Organization of the *INK4b-ARF-INK4a* locus and disease-associated SNPs. Genetic structure of the human *INK4b-ARF-INK4a* locus. The coding exons are shown in colors and non-coding exons are shown in light gray for ANRIL and dark gray for the other genes of the locus. The approximate position of single nucleotide polymorphisms (SNPs) associated with disease states is indicated by blue arrows. SNPs associated with type 2 diabetes mellitus (D), vascular heart disease (H) and frailty (F) are indicated. Map is not drawn to scale and positions are approximate.

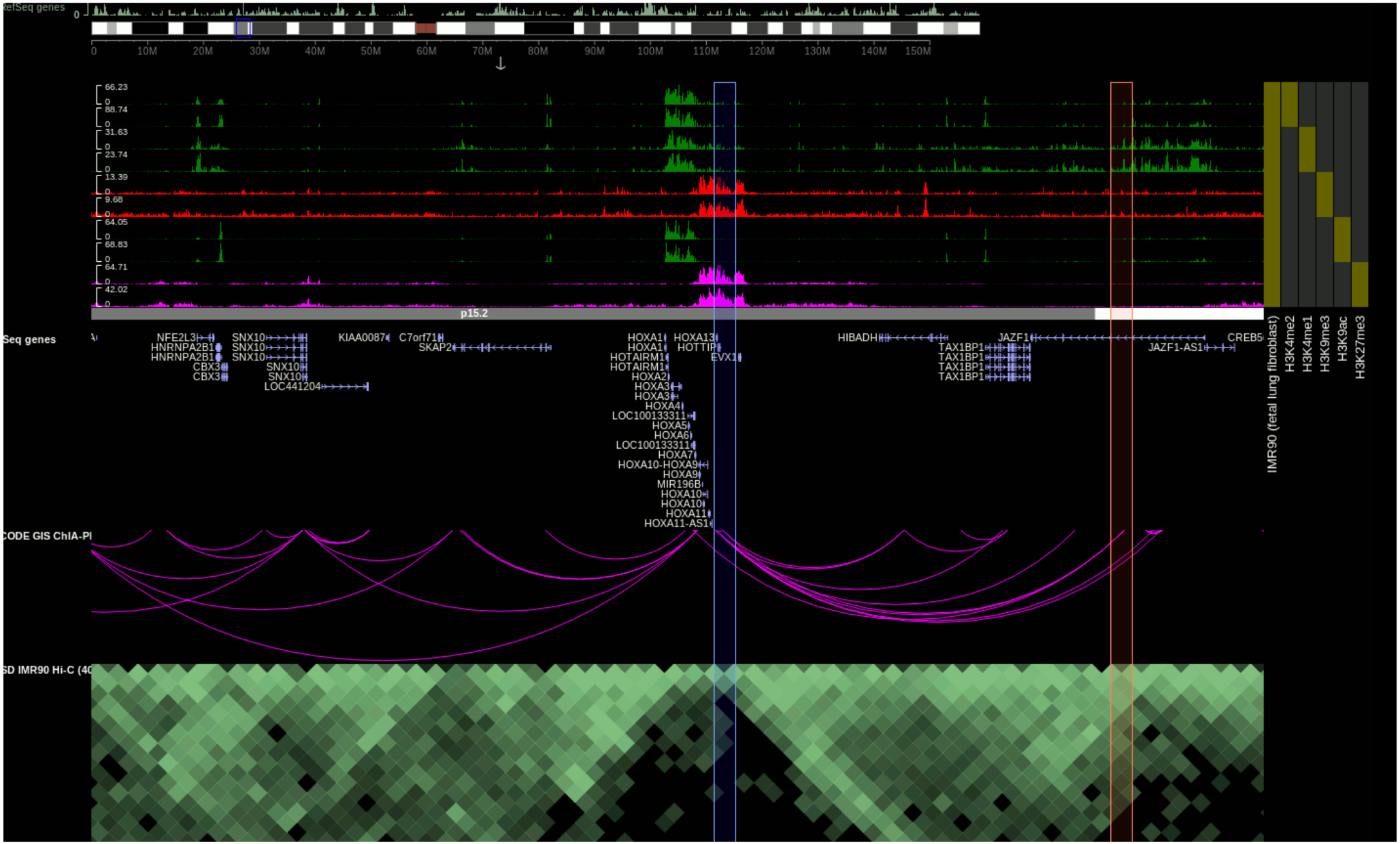


WASHU EPIGENOME BROWSER

2018

epigenomegateway.wustl.edu

Browser Navigation Instructions from this Workshop are included as “Extra” PDF material


RefSeq genes
0
0
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
140M
150M
66.23
88.74
31.63
23.74
13.39
9.68
64.05
68.83
64.71
42.02
p15.2
Seq genes
NFE2L3
HNRNPA2B1
HNRNPA2B1
CBX3
CBX3
SNX10
SNX10
SNX10
SNX10
SNX10
LOC441204
KIAA0087
C7orf71
SKAP2
HOXA1
HOXA1
HOTAIRM1
HOTAIRM1
HOXA2
HOXA3
HOXA3
HOXA4
LOC100133311
HOXA5
HOXA6
LOC100133311
HOXA7
HOXA10-HOXA9
HOXA9
MIR196B
HOXA10
HOXA10
HOXA11
HOXA11-AS1
HOXA13
HOTTIP
EVX1
HIBADH
JAZF1
TAX1BP1
TAX1BP1
TAX1BP1
TAX1BP1
JAZF1-AS1
CREB5
IMR90 (fetal lung fibroblast)
H3K4me2
H3K4me1
H3K9me3
H3K9ac
H3K27me3
CODE GIS ChIA-PI
SD IMR90 Hi-C (40

BREAK

Other NGS-based approaches

- RNA-seq variations (direct RNA sequencing, single cell sequencing

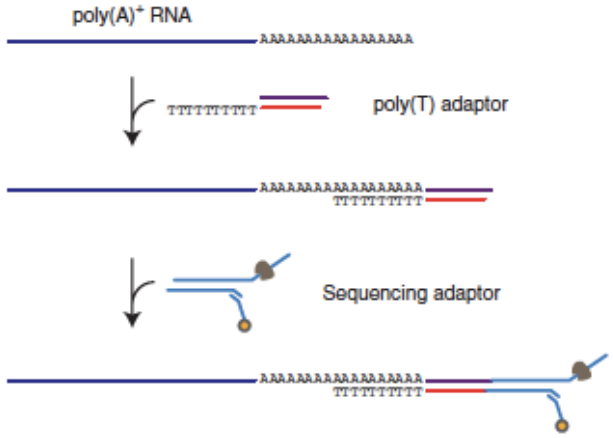
- Alternative NGS approaches to sequence/measure transcripts

Nanopore RNA-seq

Highly parallel direct RNA sequencing on an array of nanopores

Daniel R Garalde[1], Elizabeth A Snell[1], Daniel Jachimowicz[1], Botond Sipos[1], Joseph H Lloyd[1], Mark Bruce[1], Nadia Pantic[1], Tigist Admassu[1], Phillip James[1], Anthony Warland[1], Michael Jordan[1], Jonah Ciccone[1], Sabrina Serra[1], Jemma Keenan[1], Samuel Martin[1], Luke McNeill[1], E Jayne Wallace[1], Lakmal Jayasinghe[1], Chris Wright[1], Javier Blasco[1], Stephen Young[1], Denise Brocklebank[1], Sissel Juul[2], James Clarke[1], Andrew J Heron[1] & Daniel J Turner[1]

Sequencing the RNA in a biological sample can unlock a wealth of information, including the identity of bacteria and viruses, the nuances of alternative splicing or the transcriptional state of organisms. However, current methods have limitations due to short read lengths and reverse transcription or amplification biases. Here we demonstrate nanopore direct RNA-seq, a highly parallel, real-time, single-molecule method that circumvents reverse transcription or amplification steps. This method yields full-length, strand-specific RNA sequences and enables the direct detection of nucleotide analogs in RNA.



Direct RNA sequencing of yeast transcripts

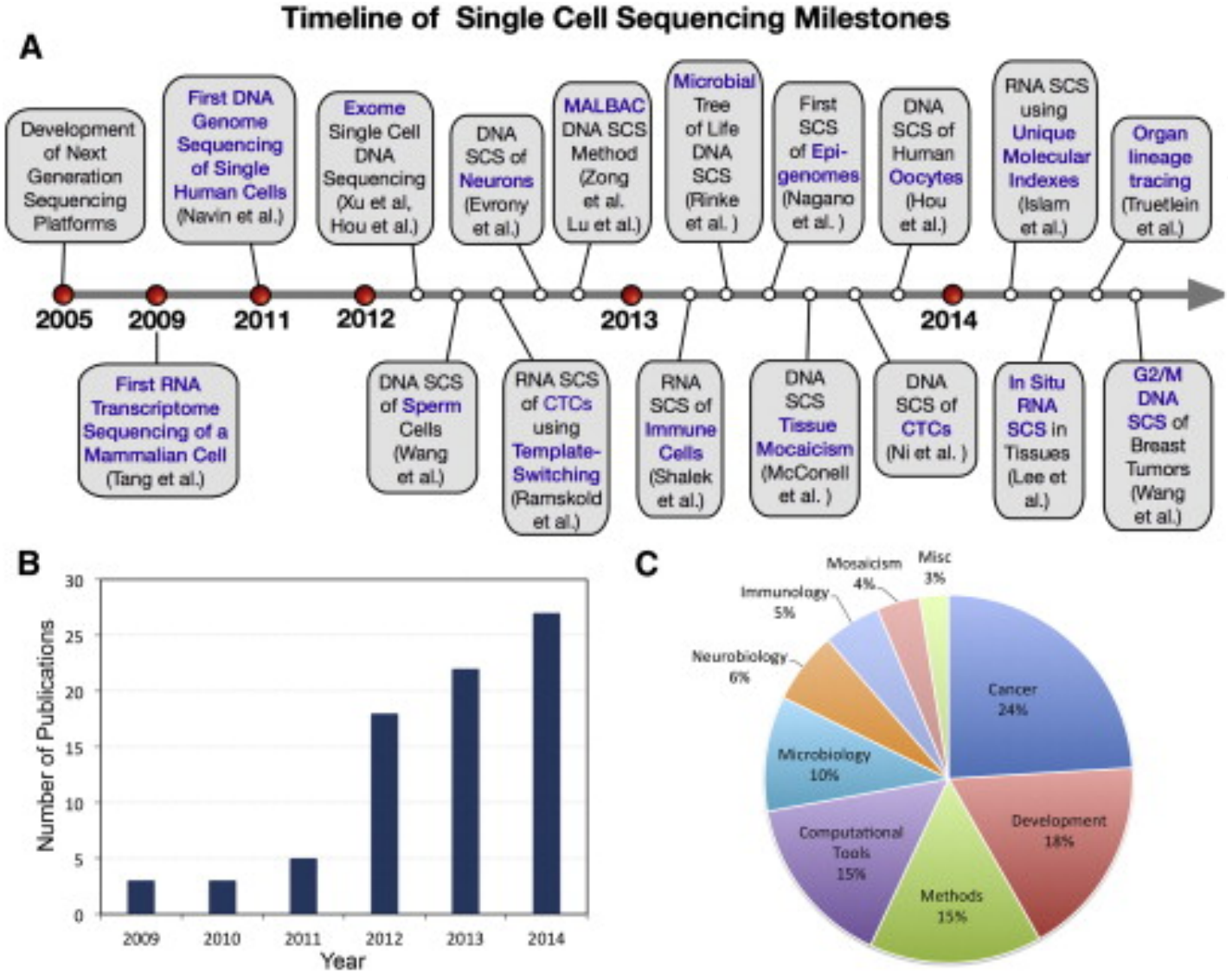
“ To assess the performance of the direct RNA-seq method, we sequenced a direct RNA library from yeast poly(A)+ RNA on a MinION MkIb with R9.4 flowcells. Using the MinKNOW instrument software, we recorded the nanopore current as each strand of RNA translocated through a nanopore.
As an adaptor-ligated yeast RNA enters the pore, the adaptor oligo is detected first, followed by the poly(A) tail, then the body of the transcript. The nanopore current returns to a high open-pore level as the transcript exits the pore on the opposite side of the membrane. We used Albacore 1.2.1 (Oxford Nanopore Technologies Ltd.) to call the bases, and we aligned the resulting reads to the Saccharomyces cerevisiae transcriptome.”

“We also sequenced the same yeast mRNA sample on a MinION cDNA run and an Illumina 100-nucleotide paired-end run…. We calculated read-count correlations between the three data sets as described, and we obtained good agreement. The direct RNA and cDNA nanopore data sets gave the highest correlation (Spearman’s rho = 0.89), and both nanopore data sets gave similar correlation values to those of the Illumina data set (Spearman’s rho = 0.81 for direct RNA and 0.79 for cDNA; Fig. 2a). Five technical replicates of different direct RNA yeast libraries correlated very well (Spearman’s rho = 0.94–0.96; n = 6,713 transcripts), showing that the library prep and sequencing is reproducible”

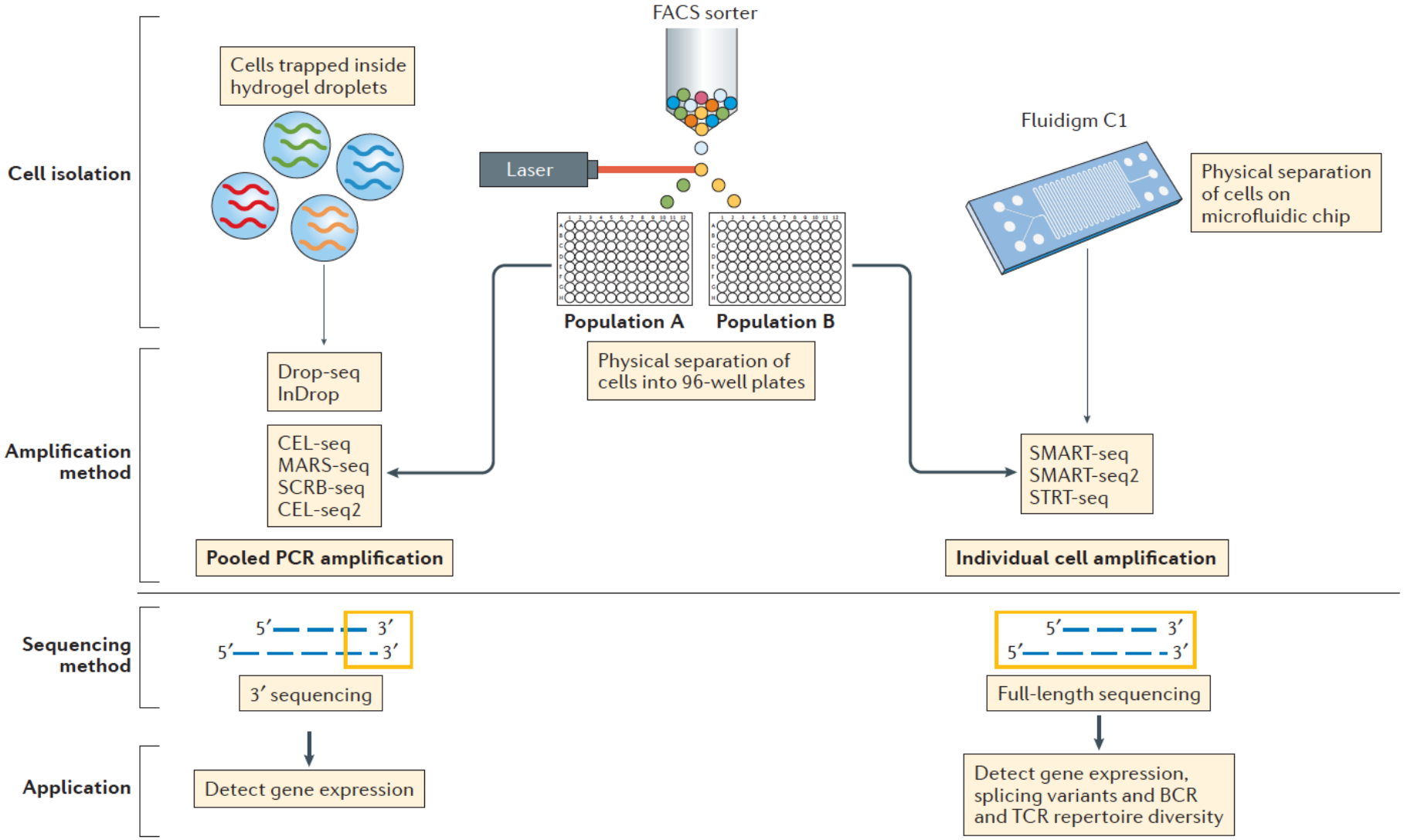
2.2 scRNA-seq

scRNA-seq

- A **new** technology, first publication by (Tang et al., 2009)
- Did not gain widespread popularity until ~2014 when new protocols and lower sequencing costs made it more accessible
- Measures the **distribution of expression levels** for each gene across a population of cells
- Allows to study new biological questions in which **cell-specific changes in transcriptome are important**, e.g. cell type identification, heterogeneity of cell responses, stochasticity of gene expression, inference of gene regulatory networks across the cells.



Overview of scRNA-seq technologies. Single-cell RNA sequencing (scRNA-seq) technologies use many different methods for cell isolation and transcript amplification. Some technologies capture cells using microfluidic devices that trap cells inside hydrogel droplets, other technologies rely on methods (such as fluorescence-activated cell sorting (FACS) into 96-well plates and the microfluidic chips used by Fluidigm C1) that physically separate one cell from another in wells.



Papalexi and Satija, Nature Reviews Imm 2018

OPTIONS for CELL ISOLATION and CAPTURE

Well-based platforms: cells are isolated using for example pipette or laser capture and placed in micro fluidic wells. One advantage of well-based methods is that they can be combined with fluorescent activated cell sorting (FACS), making it possible to select cells based on surface markers. Another advantage is that one can take pictures of the cells. Main drawback: these platforms are low-throughput.

Micro fluidic platforms (i.e Fluidigm's C1): more integrated system for capturing cells and for carrying out the reactions necessary for the library preparations. Thus, higher throughput than microwell based platforms. Typically, only around 10% of cells are captured and thus they are not appropriate if one is dealing with rare cell-types or very small amounts of input. Chips are relatively expensive.

Droplet based methods: each individual cell is encapsulated inside a nanoliter droplet together with a bead. The bead is loaded with the enzymes required to construct the library. Each bead contains a unique barcode which is attached to all of the reads originating from that cell. Thus, droplets can be pooled, sequenced together and the reads can subsequently be assigned to the cell of origin based on the barcodes. Droplet platforms typically have the highest throughput.

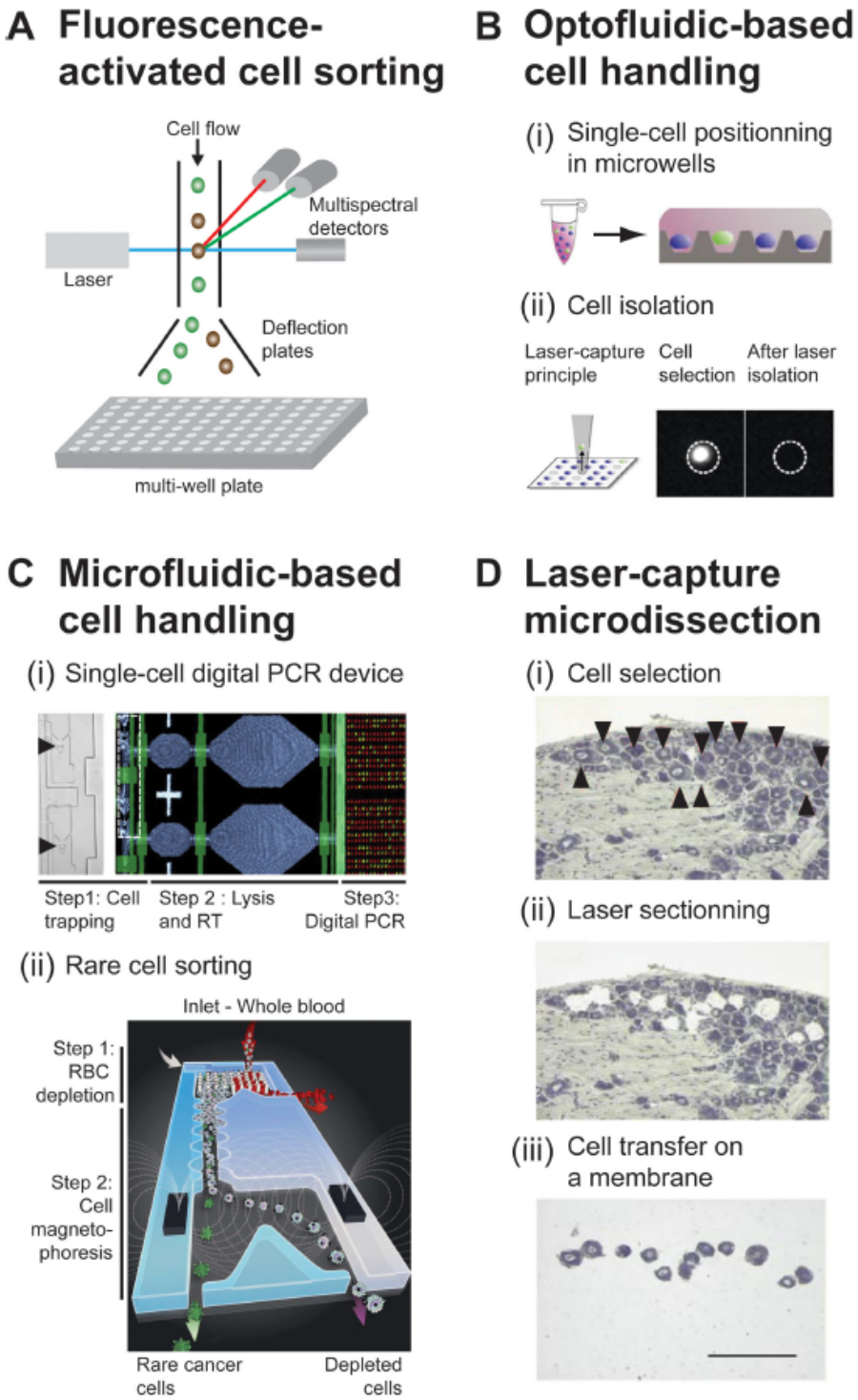


Figure 1. Methods to isolate single cells.

(A) Principle of fluorescence-activated cell sorting. A stream of droplets, each containing a single cell, passes through an excitation laser beam and the fluorescence signal emerging from each cell is analyzed by a multispectral detector. If the emitted fluorescence signal gates in the preset window, the respective cell will be sorted into a multi-well plate.

(B) Optofluidic-based cell handling. (i) A cell suspension is arrayed in a plate, each well containing a single cell. (ii) A laser, also called an optical tweezer, is used to manipulate each individual cell.

(C) Two examples of emerging microfluidics-based cell sorting technologies. (i) A microfluidic system integrates all steps from single-cell trapping to gene expression analysis. (ii)Representation of a microfluidic system that integrates all steps necessary for sorting extremely rare cancer cells from whole blood

(D) Laser capture microdissection. (i) Cells of interest are identified in a stained section of rat cervical dorsal root ganglia (indicated by black arrowheads), (ii) cut with a UV laser and (iii) transferred onto a membrane

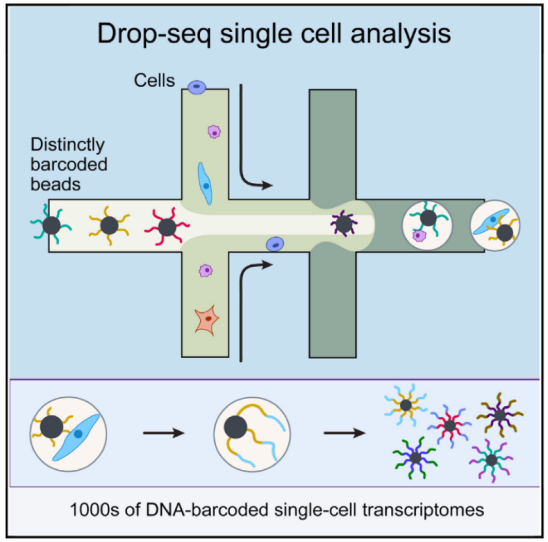
Saliba et al., NAR 2014

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Evan Z. Macosko,[1,2,3,*] Anindita Basu,[4,5] Rahul Satija,[4,6,7] James Nemesh,[1,2,3] Karthik Shekhar,[4] Melissa Goldman,[1,2] Itay Tirosh,[4] Allison R. Bialas,[8] Nolan Kamitaki,[1,2,3] Emily M. Martersteck,[9] John J. Trombetta,[4] David A. Weitz,[5,10] Joshua R. Sanes,[9] Alex K. Shalek,[4,11,12] Aviv Regev,[4,13,14] and Steven A. McCarroll[1,2,3,*]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[2]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

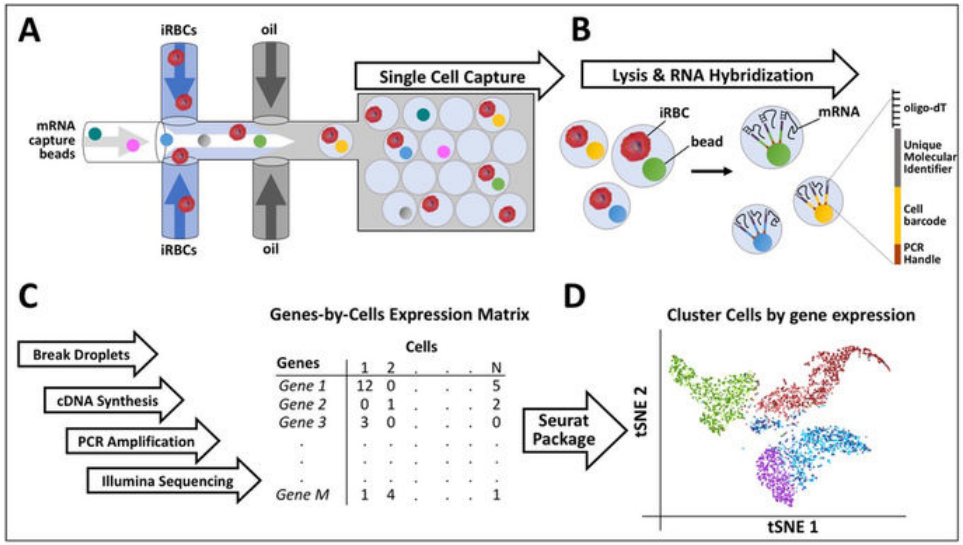


Highlights

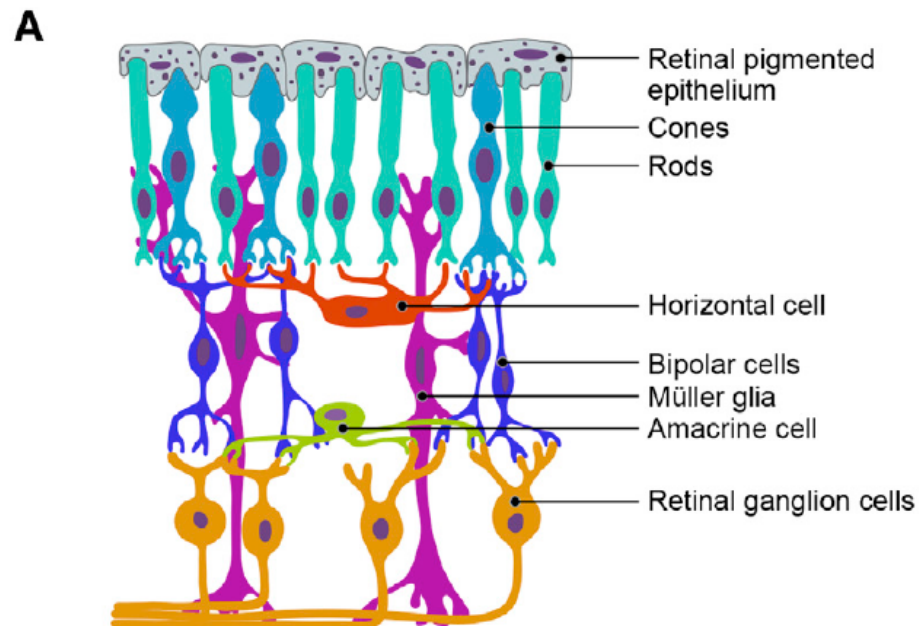
- Drop-seq enables highly parallel analysis of individual cells by RNA-seq
- Drop-seq encapsulates cells in nanoliter droplets together with DNA-barcoded beads
- Systematic evaluation of Drop-seq library quality using species mixing experiments
- Drop-seq analysis of 44,808 cells identifies 39 cell populations in the retina

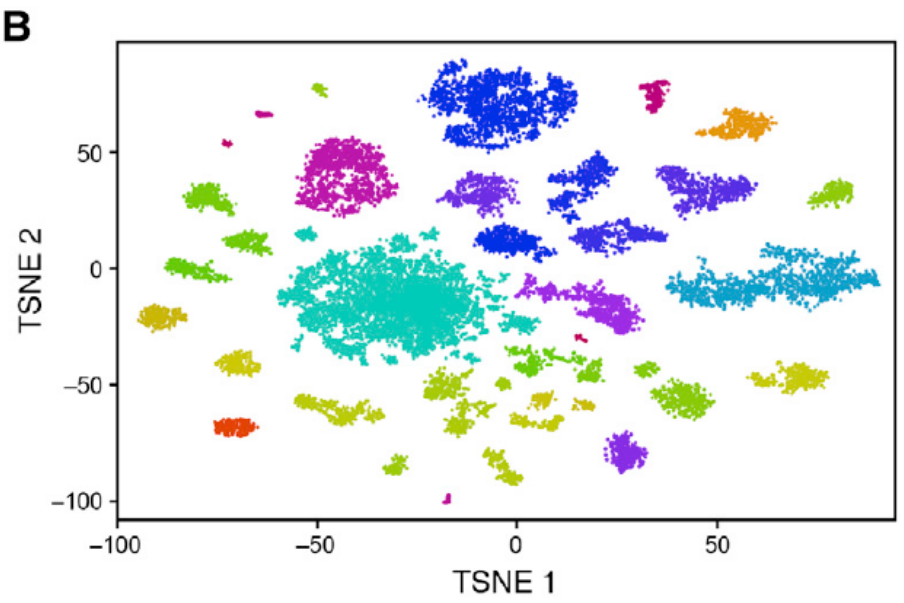
Mcoscko et al., Cell, 2015

DROP-seq
Protocol Schematic



Poran et al., Nature, 2017





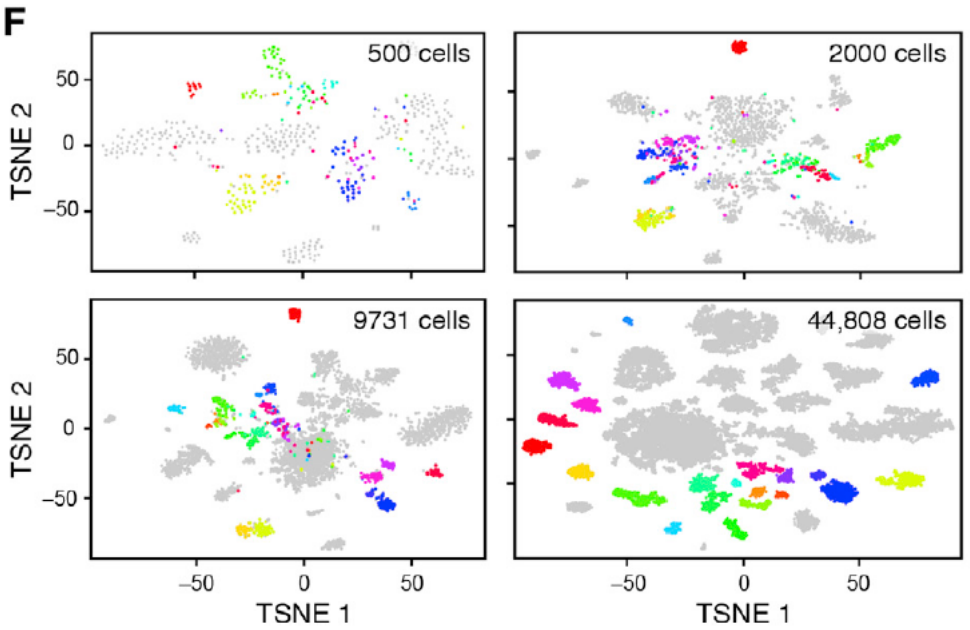


Figure 5. **Reconstruction of Retinal Cell Types from 44,808 Single-Cell Transcription Profiles**

(A) Schematic representation of major cell classes in the retina.

(B) Clustering of 44,808 Drop-seq single-cell expression profiles into 39 retinal cell populations. The plot shows a two-dimensional representation (tSNE) of global gene expression relationships among 44,808 cells; clusters are colored by cell class, according to Figure 5A.

(C) Trajectory of amacrine clustering as a function of number of cells analyzed. Analyses of smaller numbers of cells incompletely distinguished these subpopulations from one another.

Macoscko et al., Cell, 2015

GRO-seq

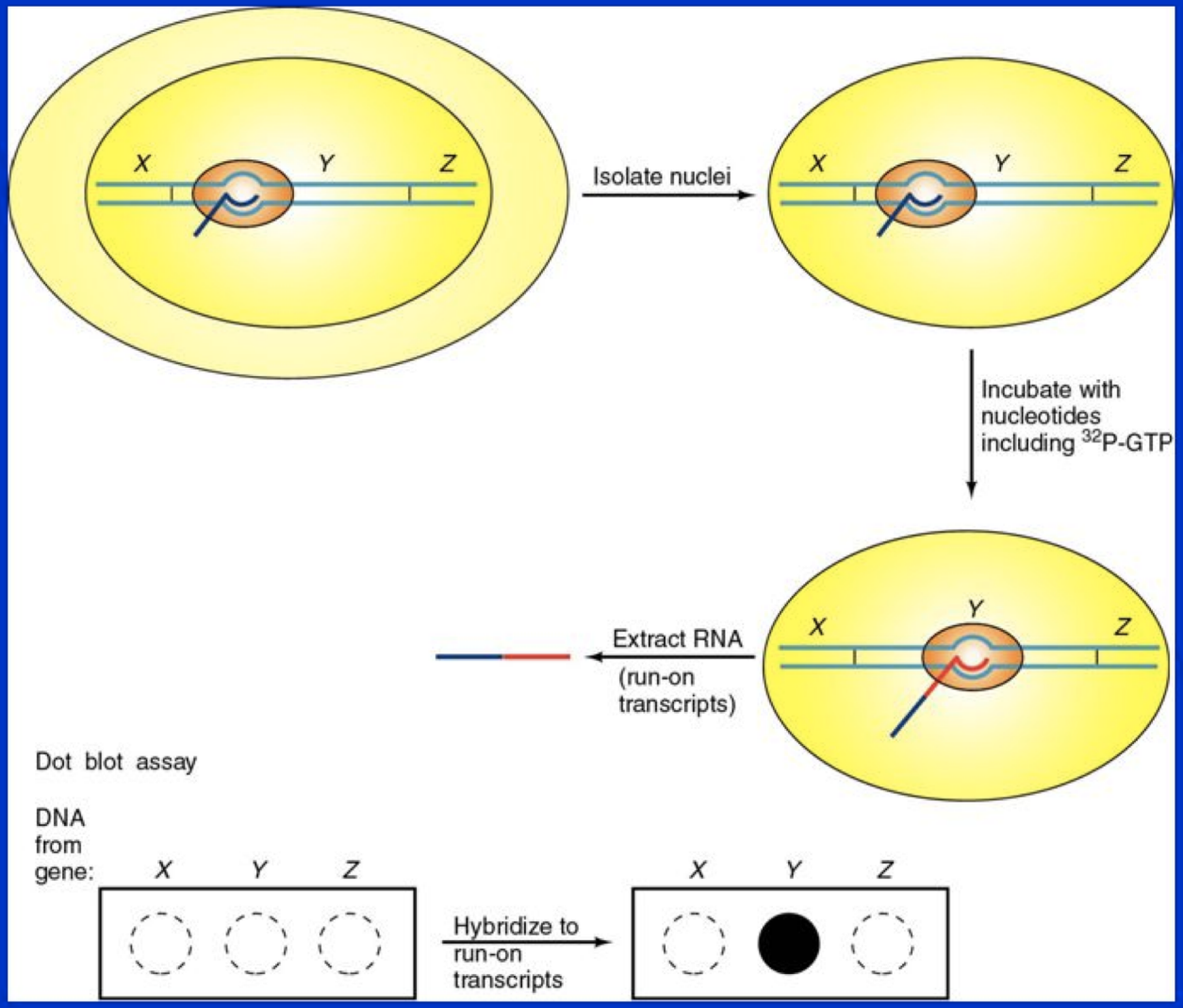
Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters

Leighton J. Core,* Joshua J. Waterfall,* John T. Lis†

RNA polymerases are highly regulated molecular machines. We present a method (global run-on sequencing, GRO-seq) that maps the position, amount, and orientation of transcriptionally engaged RNA polymerases genome-wide. In this method, nuclear run-on RNA molecules are subjected to large-scale parallel sequencing and mapped to the genome. We show that peaks of promoter-proximal polymerase reside on ~30% of human genes, transcription extends beyond pre-messenger RNA 3′ cleavage, and antisense transcription is prevalent. Additionally, most promoters have an engaged polymerase upstream and in an orientation opposite to the annotated gene. This divergent polymerase is associated with active genes but does not elongate effectively beyond the promoter. These results imply that the interplay between polymerases and regulators over broad promoter regions dictates the orientation and efficiency of productive transcription.

SCIENCE VOL 322 19 DECEMBER 2008

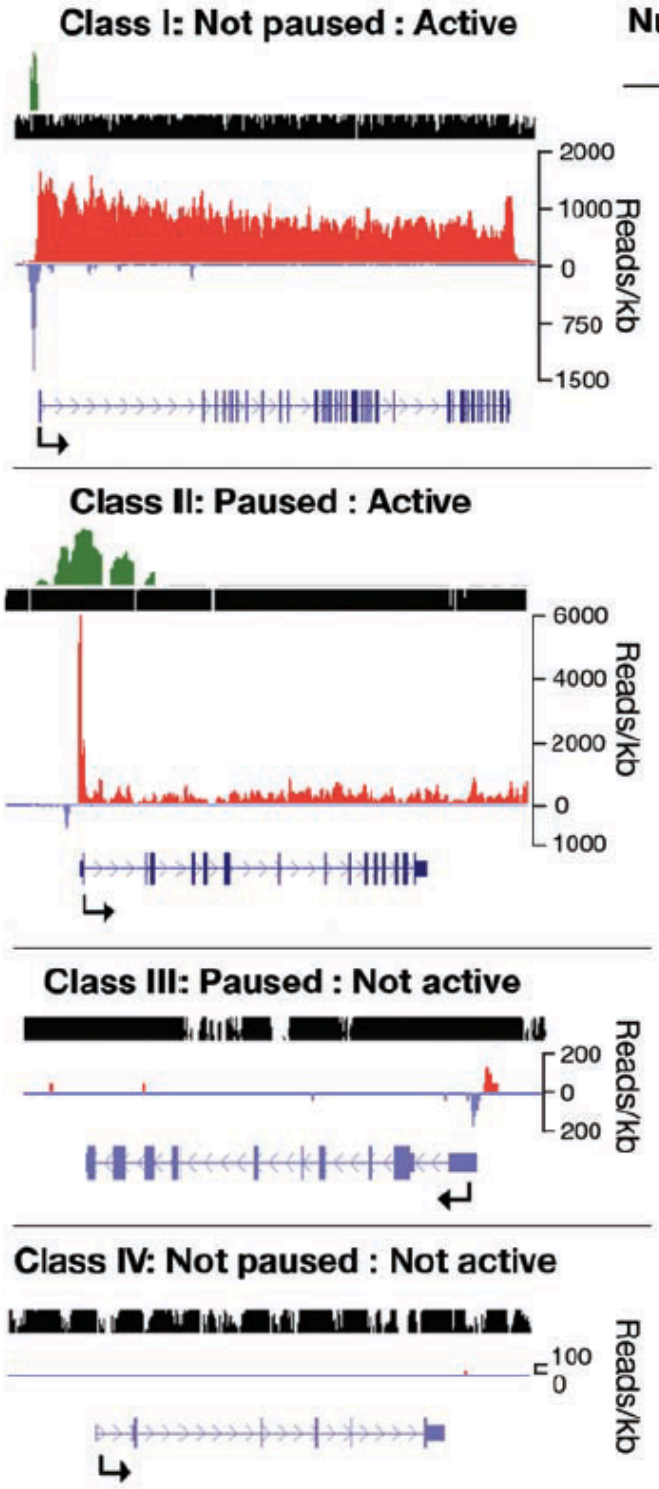
Nuclear Run-on Transcription Assay



Measure transcripts as they are produced rather than transcripts accumulated in the cell

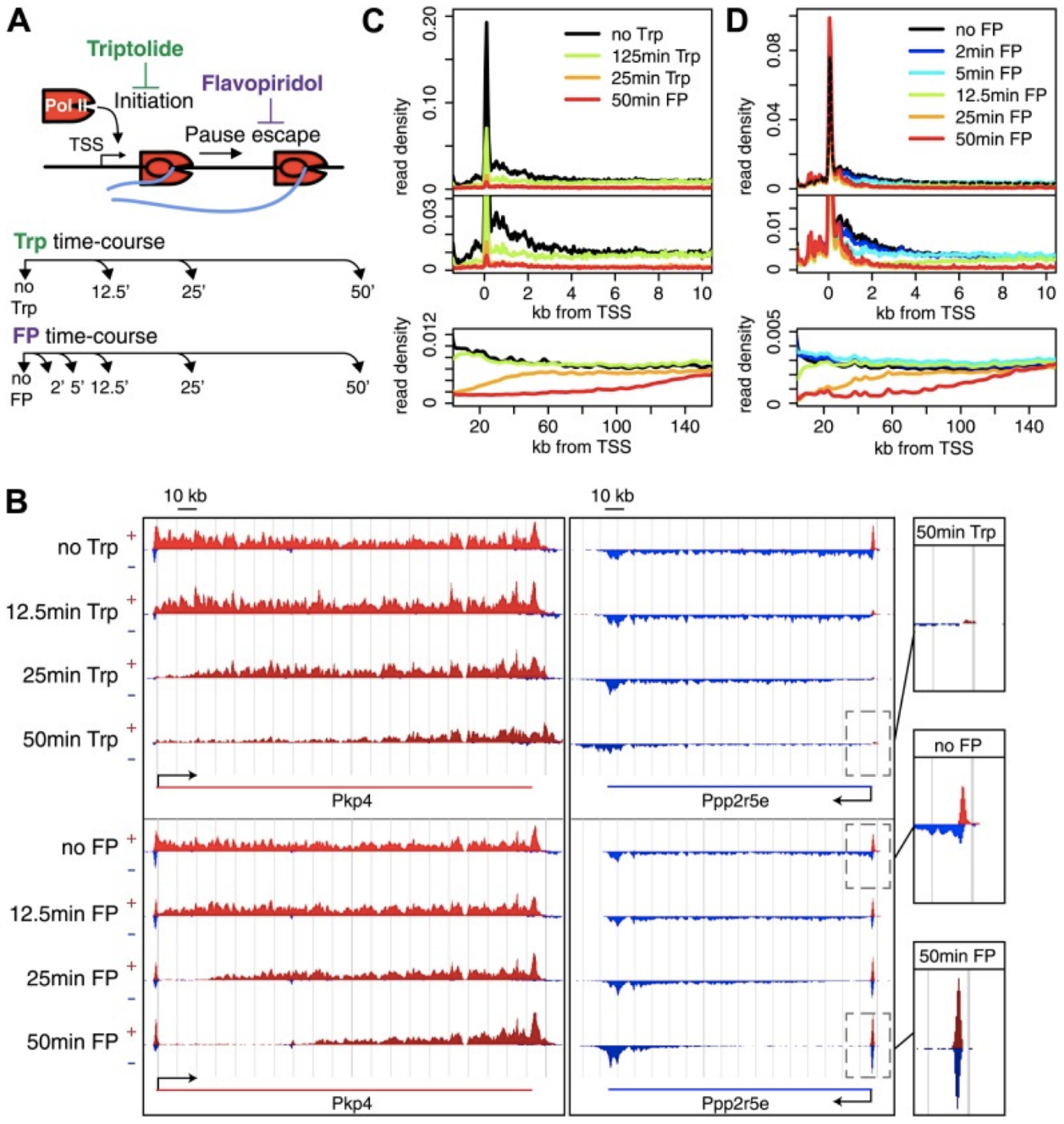
Transcription versus mRNA stabilization

Fig. 3. Comparison of pausing with gene activity. Four classes of genes are found when comparing genes with a paused polymerase and transcription activity either by microarray or GRO-seq density in the downstream portions of genes. An example of each class is shown, with tracks shown in the UCSC genome browser as in Fig. 1. The gene names, pausing index, and *P* value, from top to bottom, respectively, are as follows: *TRIO*, 1.1, 0.62; *FUS*, 41, 2.8×10^{-43}; *IZUMO1*, 410, 7.6×10^{-3}; and *GALP* (which has no reads and therefore no pausing index). The number of genes represented in each class is shown to the right.



Number of genes by method:	
Array	GRO-seq
3867	9850
4197	7032
1224	25
8012	5687

Leighton et al., Science 2008



Tag density plots
(all transcripts)

Genome browser
tracks
(single gene)

Jonkers et al., eLIFE 2014

OTHER VARIATIONS

PRO-seq

NET-seq

PRO-seq: Precision Run-On Sequencing (Lis Lab)

PRO-Seq maps RNA POL2 active sites with base-pair resolution. This approach is similar to GRO-Seq, but the protocol is slightly modified to achieve single-base resolution.

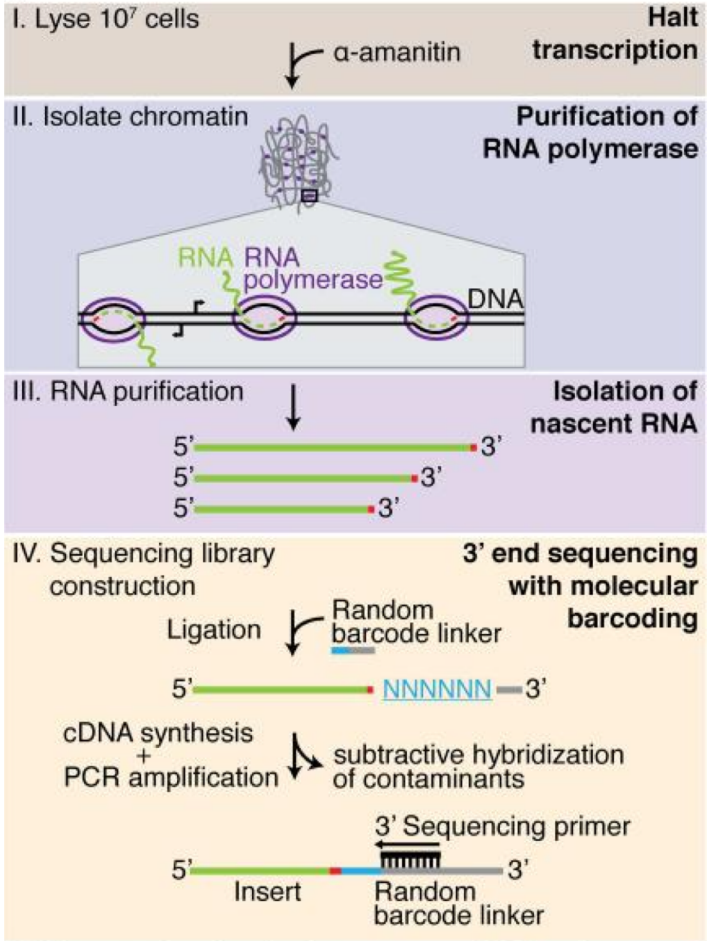
PRO-Cap is a variation of the same protocol that allows mapping of RNAPII initiation sites.

NET-seq: Native Elongating Transcript Sequencing (Churchman, Weissman Labs)

NET-seq is based on 3'-end sequencing of nascent RNA isolated with engaged RNA PolII from the chromatin fraction. Single nucleotide resolution for POL2 mapping

START-seq (Adelman Lab)

High-throughput sequencing of nascent 25–65 nucleotide (nt) capped RNA species associated with RNAP



Cell *161*, 526–540, April 23, 2015 ©2015

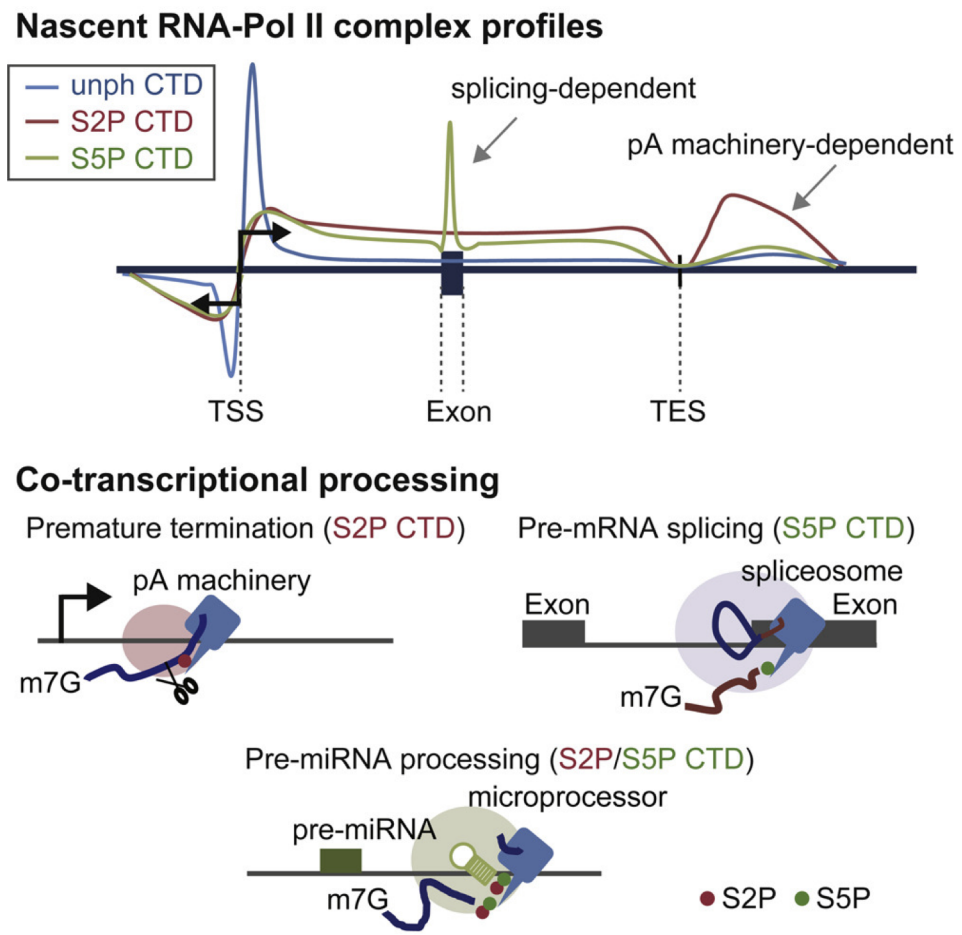
Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing

Takayuki Nojima,[1,4] Tomás Gomes,[2,4] Ana Rita Fialho Grosso,[2] Hiroshi Kimura,[3] Michael J. Dye,[1] Somdutta Dhir,[1] Maria Carmo-Fonseca,[2,*] and Nicholas J. Proudfoot[1,*]

[1]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

[2]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal

[3]Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 226-8501 Yokohama, Japan



SUMMARY

Transcription is a highly dynamic process. Consequently, we have developed native elongating transcript sequencing technology for mammalian chromatin (mNET-seq), which generates single-nucleotide resolution, nascent transcription profiles. Nascent RNA was detected in the active site of RNA polymerase II (Pol II) along with associated RNA processing intermediates. In particular, we detected 5′splice site cleavage by the spliceosome, showing that cleaved upstream exon transcripts are associated with Pol II CTD phosphorylated on the serine 5 position (S5P), which is accumulated over downstream exons. Also, depletion of termination factors substantially reduces Pol II pausing at gene ends, leading to termination defects. Notably, termination factors play an additional promoter role by restricting non-productive RNA synthesis in a Pol II CTD S2P-specific manner. Our results suggest that CTD phosphorylation patterns established for yeast transcription are significantly different in mammals. Taken together, mNET-seq provides dynamic and detailed snapshots of the complex events underlying transcription in mammals.

BUONA PASQUA!!!