Non-coding RNAs in the human ENCODE project: perspectives for "non-model" organisms.

IGDR : Institute of Genetics and Development of Rennes

UMR6290 - CNRS - Université de Rennes
Canine Genetics Group

Thomas DERRIEN

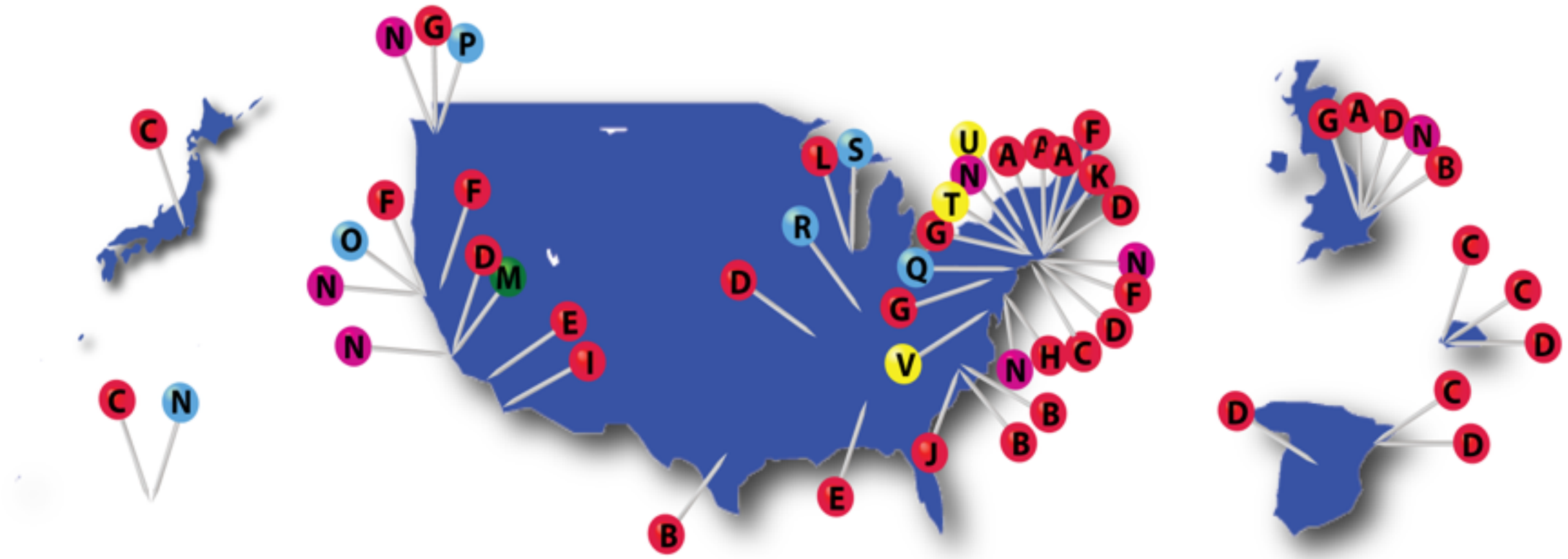
The ENCODE project and consortium

- ENCODE = **Enc**yclopedia **o**f **D**NA **E**lements.
- Aim: identify all functional elements present in the human genome.
- Launched by the National Human Genome Research Institute (NHGRI, USA) as a public research consortium in September 2003, after the Human Genome Project.

- 2/3 components:
  - Pilot phase: focusing on 1% of the genome (ENCODE regions).
  - Technology development phase: on 100% of the genome.
  - Mouse ENCODE project (End 2014)
- Involve many investigators:
  - with diverse backgrounds and expertise,
  - mainly from USA, but also from Europe and Asia.

A comparative encyclopedia of DNA elements in the mouse genome

ENCODE institution location



Production Groups

- **A** **Broad Institute**, Harvard University, MIT, Cambridge University.
- **B** **Duke University**, University of Texas, Austin, UNC-Chapel Hill, EBI.
- **C** **CSHL**, University of Geneva, Centre for Genomic Regulation, RIKEN, University of Lausanne, Genome Institute of Singapore.
- **D** **Sanger Institute** Washington University, Yale University, Centre for Genomic Regulation, UCSC, MIT, University of Lausanne, CNIO.
- **E** **HudsonAlpha**, Caltech.
- **F** **Stanford University**, Yale University, UC Davis, Harvard University.
- **G** **University of Washington**, U Mass Med School, EBI, Princeton University.
- **H** **Penn State**.
- **I** **UC - San Diego**.
- **J** **UNC - Chapel Hill**.
- **K** **Boston University**.
- **L** **University of Chicago**.

Data Coordination Center

- **M** **UCSC**

Data Analysis Center

- **N** **EBI**, UC Berkeley, Yale University Penn State, UCSC, University of Washington, U Mass Med School.

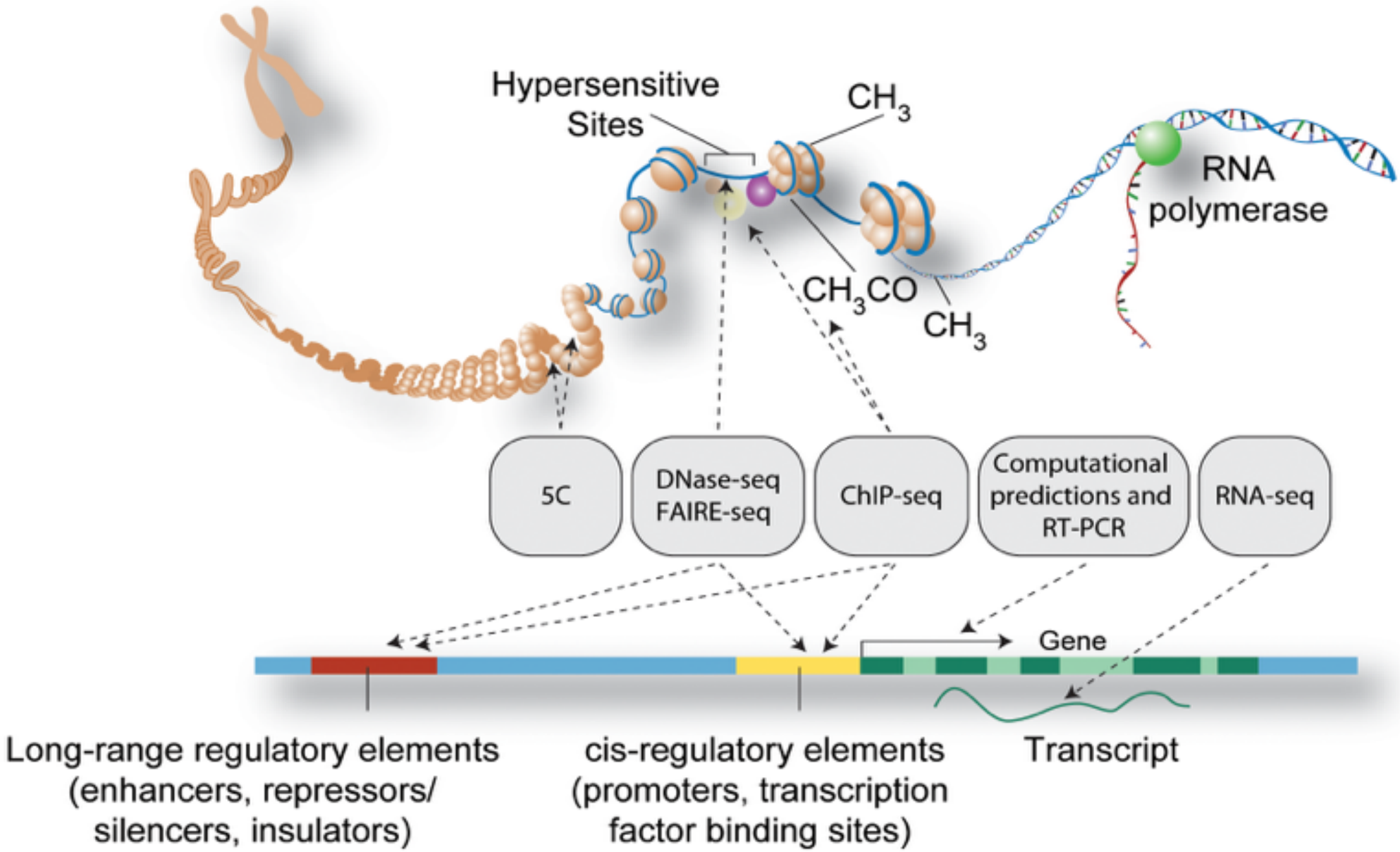
Technology Groups

- **N** **Genome Institute of Singapore**.
- **O** **Stanford University**.
- **P** **University of Washington**.
- **Q** **Albert Einstein College of Med**.
- **R** **Indiana University**.
- **S** **University of Chicago**.

Other Groups

- **T** **U Mass Med School**
- **U** **University at Albany**
- **V** **NHGRI**

Whole genome ENCODE overview



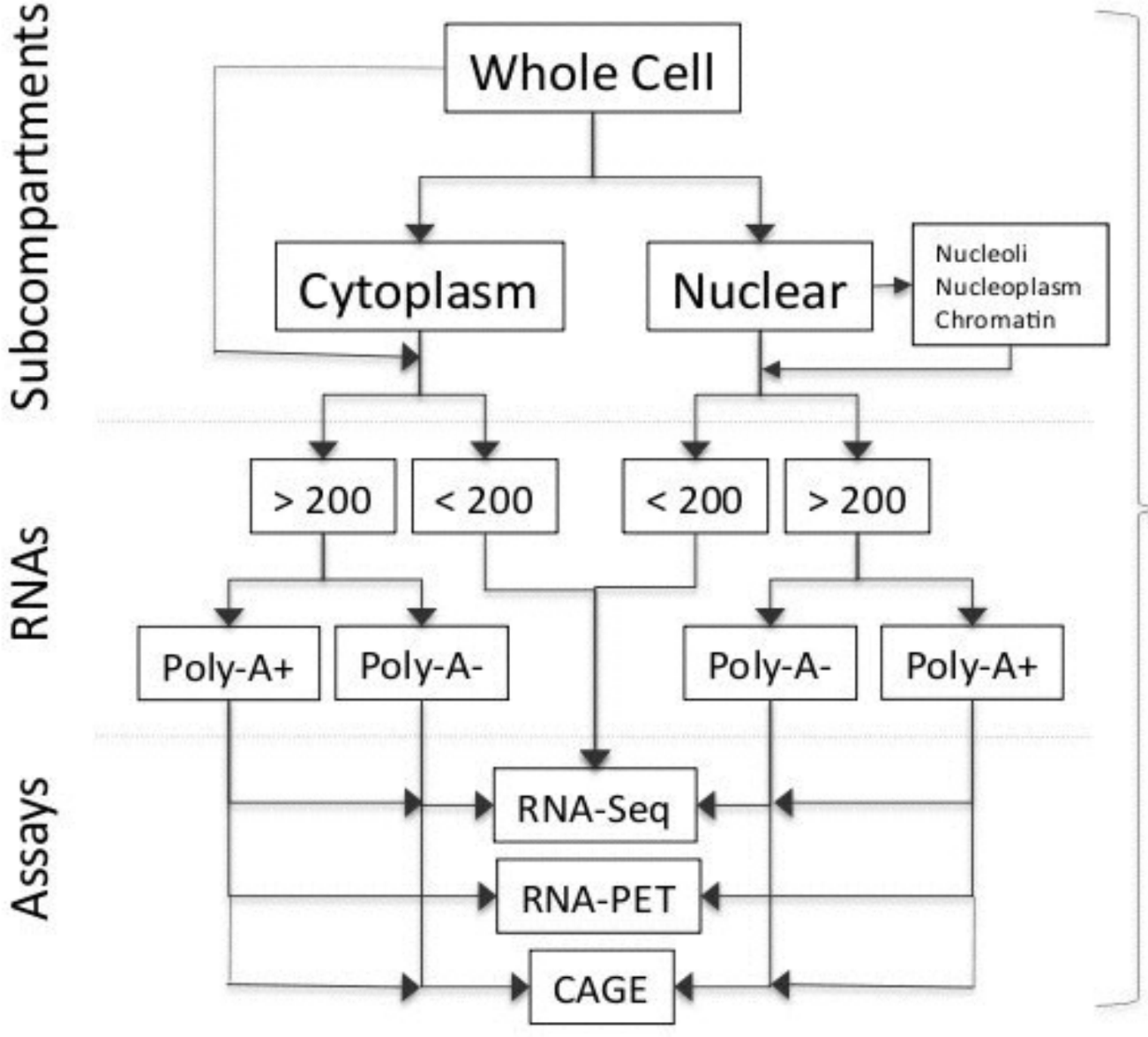
ENCODE project consortium, PLoS Biology, 2011

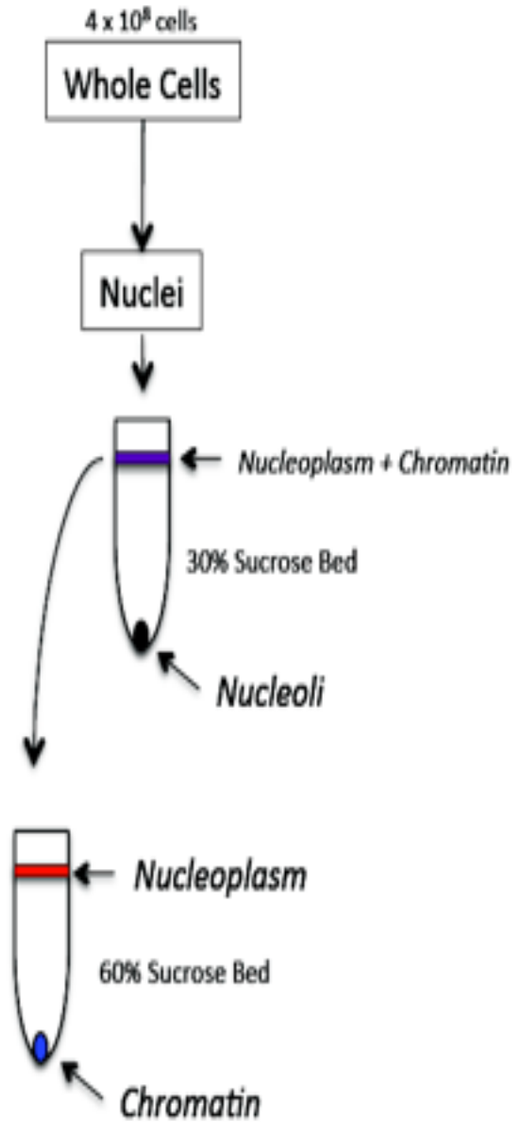
Whole genome ENCODE production

Research Group	Institution	Research Goals
Bradley Bernstein	Broad Institute of MIT and Harvard	Map histone modifications using chromatin immunoprecipitation followed by high-throughput sequencing.
Gregory Crawford	Duke University	Identify and characterize regions of open chromatin using DNaseI hypersensitivity assays, formaldehyde-assisted isolation of regulatory elements, and chromatin immunoprecipitation.
Morgan Giddings	University of North Carolina, Chapel Hill	Produce large-scale proteomic data sets on ENCODE cell lines using mass spectrometry.
Thomas Gingeras	Affymetrix, Inc.	Identify protein-coding and non-protein coding RNA transcripts using microarrays, high-throughput sequencing, sequence paired-end tags, and sequenced cap analysis of gene expression tags.
Timothy Hubbard	Wellcome Trust Sanger Institute	Annotate gene features using computational methods, manual annotation, and targeted experiments.
Richard Myers	HudsonAlpha Institute for Biotechnology	Identify transcription factor binding sites using chromatin immunoprecipitation followed by high-throughput DNA sequencing; Pilot effort to determine the methylation status of CpG-rich regions.
Michael Snyder	Stanford University	Identify transcription factor binding sites using chromatin immunoprecipitation followed by high-throughput DNA sequencing.
John Stamatoyannopoulos	University of Washington, Seattle	Map and functionally classify DNaseI hypersensitive sites by digital DNaseI and histone modification mapping using high-throughput sequencing.
Thomas Tullius	Boston University	Develop high-throughput methods for collecting hydroxyl radical cleavage data; locate structural features in human genome that are under selective evolutionary pressure, but for which the exact nucleotide sequence is not under selection.
Kevin White	University of Chicago	Epitope tag transcription factors for chromatin immunoprecipitation using BAC recombineering.

Landscape of transcription in human cells

The ENCODE RNA assays





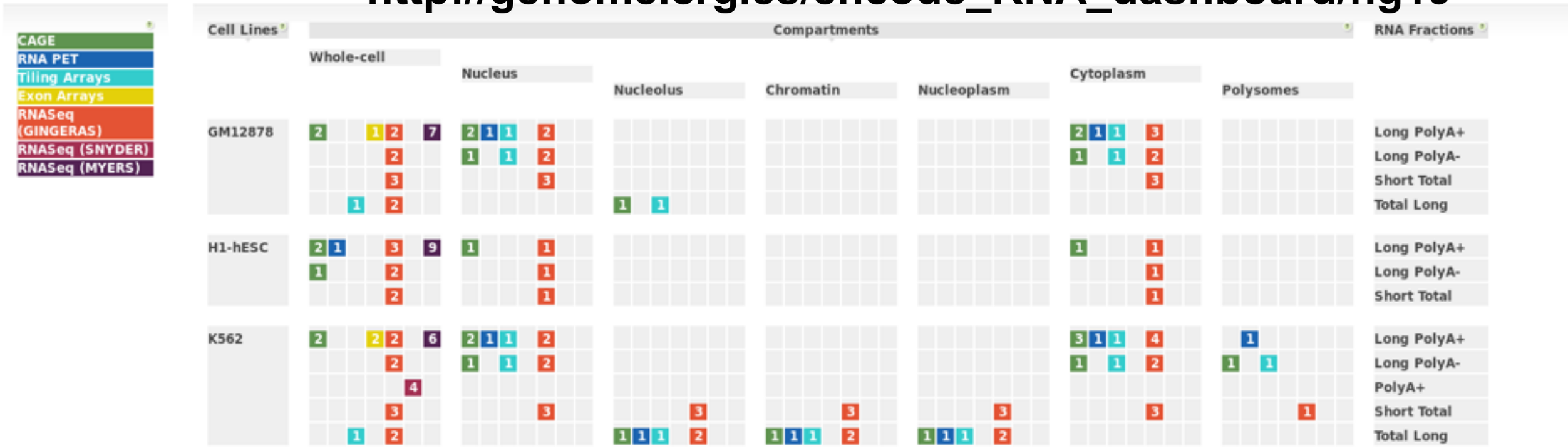
Allows to assess reproducibility

Djebali, Davis et al., Nature, 2012

Data distribution: the RNA dashboard

(Julien Lagarde - CRG)

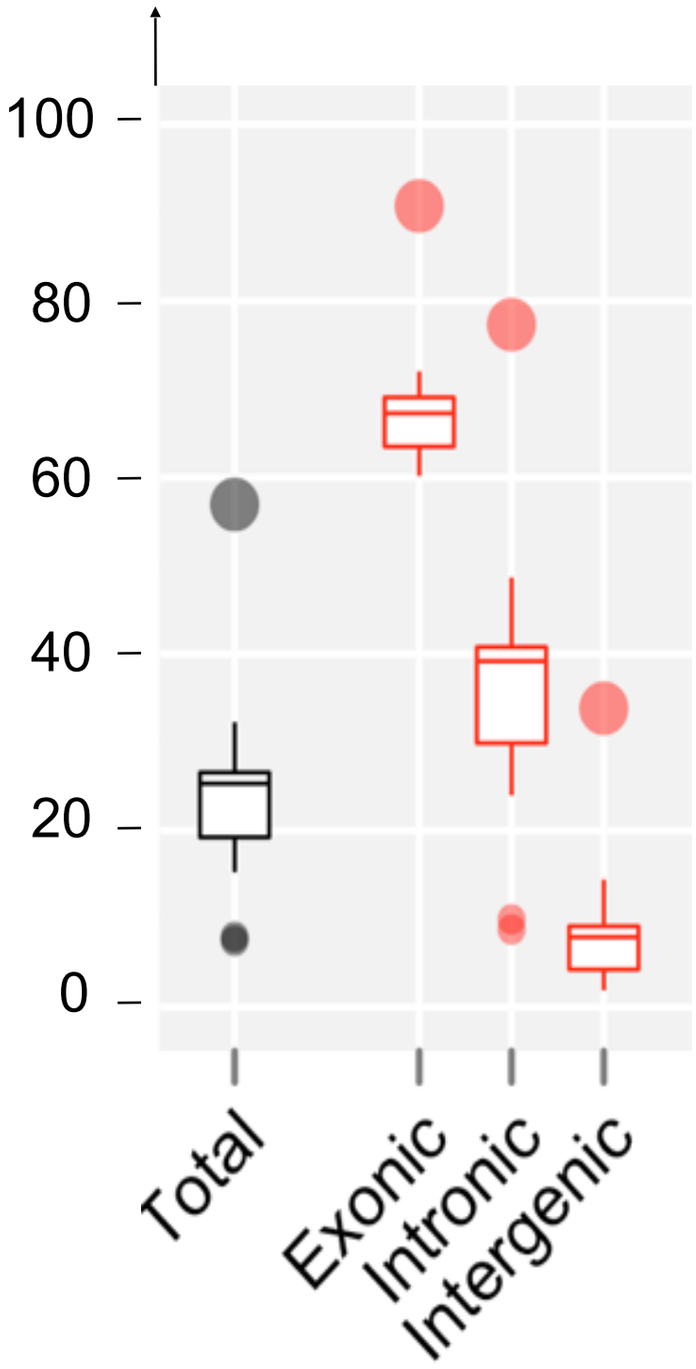
http://genome.crg.es/encode_RNA_dashboard/hg19



- The number inside each coloured box represents the **number of experiments** that have been performed for the corresponding metadata (cell line/ cell compartment/ RNA fraction).

- Clicking on a box expands it and provides the user with links to files of both **raw and processed data** available for the corresponding experiments.

ENCODE RNA-seq coverage of the human genome



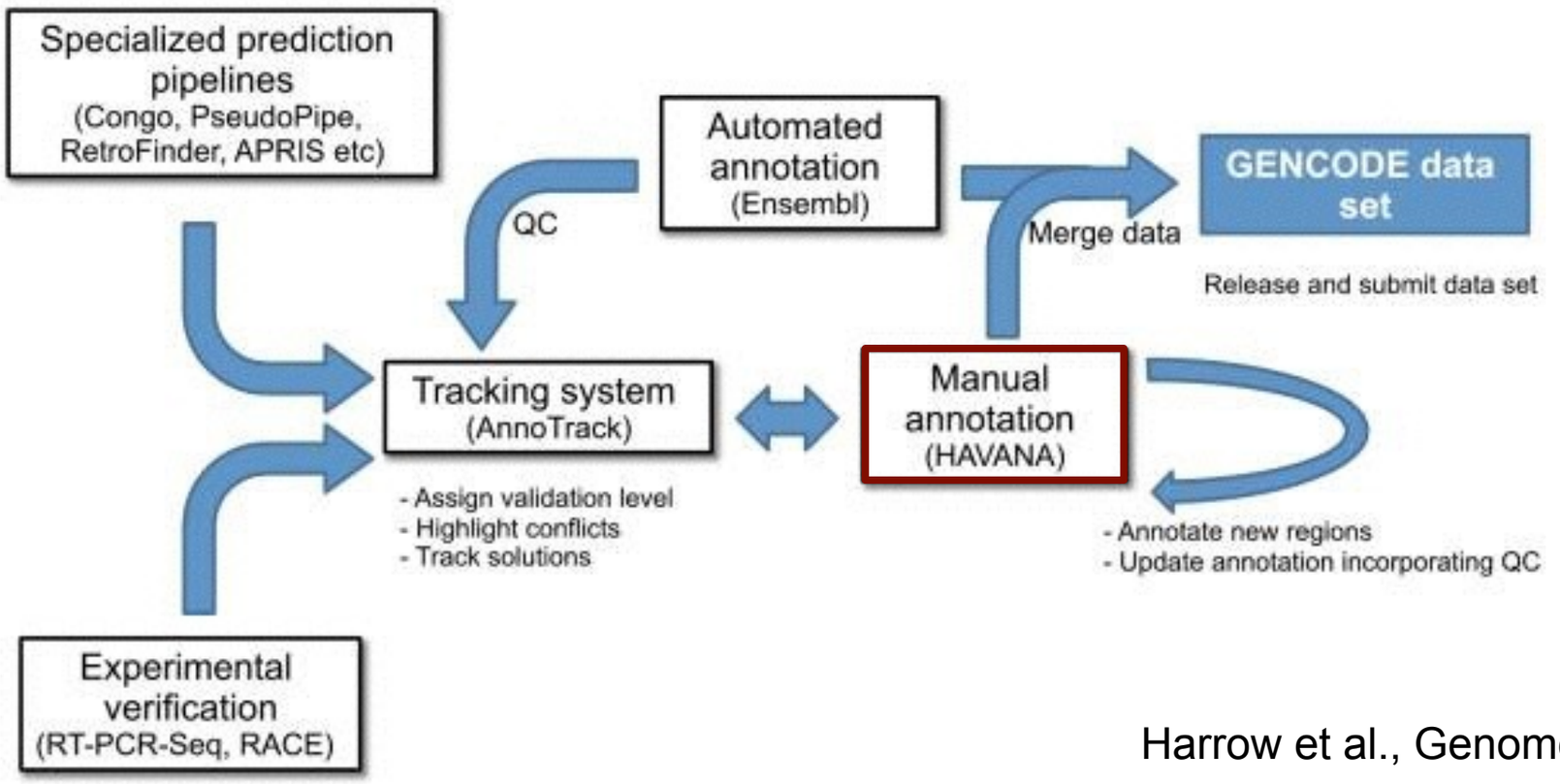
Red: Proportion of nucleotides (nt) in genomic domains covered by RNA-seq contigs;

Cumulatively (● or ●), is covered by RNA-seq:
57% of the genome,
91% of the exonic nt,
77% of the intronic nt,
34% of the intergenic nt.

Djebali, Davis et al., Nature, 2012

The Gencode Reference annotation

Gencode as the reference gene annotation



Harrow et al., Genome Research, 2012

Many different biotypes for transcripts and genes: 4 broad types:
- protein coding (mRNA),
- long non-coding (lncRNA),
- small non-coding (sRNA),
- pseudogenes.

Several objects annotated:
- gene
- transcript
- exon
- CDS
- UTR

3 confidence levels for transcripts and genes:
- level 1: validated,
- level 2: manually annotated,
- level 3: automatically annotated.

http://www.gencodegenes.org/

- Project
- Phase 2 GENCODE Goals
- Data
- Statistics - Human
- Statistics - Mouse
- Participants
- Publications
- lncRNA microarray
- RGASP 1/2
- RGASP 3
- Blog
- GENCODE workshops
- Contact us

The GENCODE Project:

Encyclopædia of genes and gene variants

Current GENCODE version

The current version in **Human** is **Gencode 19**, released on the 10/12/2013.
For more information about the human releases please see the **README.txt** file.

The current version in **Mouse** is **Gencode M2**, released on the 10/12/2013.
For more information about the mouse releases please see the **README.txt** file.

** NEW ** Two publications now out on our RNASeq genome annotation assessment project (RGASP):

- **Assessment of transcript reconstruction methods for RNA-seq.**
  Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consortium, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, Bertone P, Bohnert R, Bucher P, Cloonan N, Derrien T, Djebali S, Du J, Dudoit S, Engström PG, Gerstein M, Gingeras TR, Gonzalez D, Grimmond SM, Guigó R, Habegger L, Harrow J, Hubbard TJ, Iseli C, Jean G, Kahles A, Kokocinski F, Lagarde J, Leng J, Lefebvre G, Lewis S, Mortazavi A, Niermann P, Rätsch G, Reymond A, Ribeca P, Richard H, Rougemont J, Rozowsky J, Sammeth M, Sboner A, Schulz MH, Searle SM, Solorzano ND, Solovyev V, Stanke M, Steijger T, Stevenson BJ, Stockinger H, Valsesia A, Weese D, White S, Wold BJ, Wu J, Wu TD, Zeller G, Zerbino D, Zhang MQ, Hubbard TJ, Guigó R, Harrow J and Bertone P
  Nature methods 2013;**10**;12;1177-84
  PUBMED: **24185837**; PMC: **3851240**; DOI: **10.1038/nmeth.2714**
- **Systematic evaluation of spliced alignment programs for RNA-seq data.**
  Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, RGASP Consortium, Alioto T, Behr J, Bertone P, Bohnert R, Campagna D, Davis CA, Dobin A, Engström PG, Gingeras TR, Goldman N, Grant GR, Guigó R, Harrow J, Hubbard TJ, Jean G, Kahles A, Kosarev P, Li S, Liu J, Mason CE, Molodtsov V, Ning Z, Ponstingl H, Prins JF, Rätsch G, Ribeca P, Seledtsov I, Sipos B, Solovyev V, Steijger T, Valle G, Vitulo N, Wang K, Wu TD, Zeller G, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R and Bertone P
  Nature methods 2013;**10**;12;1185-91
  PUBMED: **24185836**; DOI: **10.1038/nmeth.2722**

Introduction

Gencode statistics

Version 21 (June 2014 freeze, GRCh38) - Ensembl 77

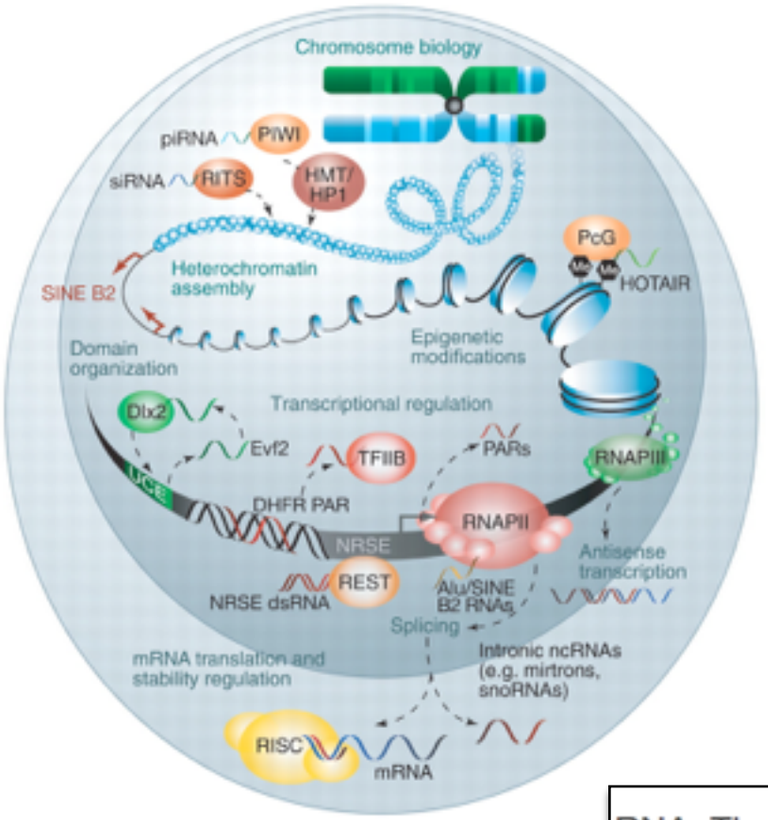
General stats

Total No of Genes	60155	**Total No of Transcripts**	196327
Protein-coding genes	19881	**Protein-coding transcripts**	79377
Long non-coding RNA genes	15877	**- full length protein-coding:**	54420
Small non-coding RNA genes	9534	**- partial length protein-coding:**	24957
Pseudogenes	14467	**Nonsense mediated decay transcripts**	13222
- processed pseudogenes:	10753	**Long non-coding RNA loci transcripts**	26414
- unprocessed pseudogenes:	3230		
- unitary pseudogenes:	170		
- polymorphic pseudogenes:	59		
- pseudogenes:	29		
Immunoglobulin/T-cell receptor gene segments		**Total No of distinct translations**	59512
- protein coding segments:	395	**Genes that have more than one distinct translations**	13526
- pseudogenes:	226		

The Gencode v7 catalog of human long noncoding RNA: analysis of their gene structure, evolution and expression.

Why lncRNAs?

- ~60% of the human genome is transcribed (only 2% correspond to mRNAs)
- **Back to the future**: The cell as an RNA machinery



(from Amaral P, et al., 2008)

RNA: The genome's rising stars

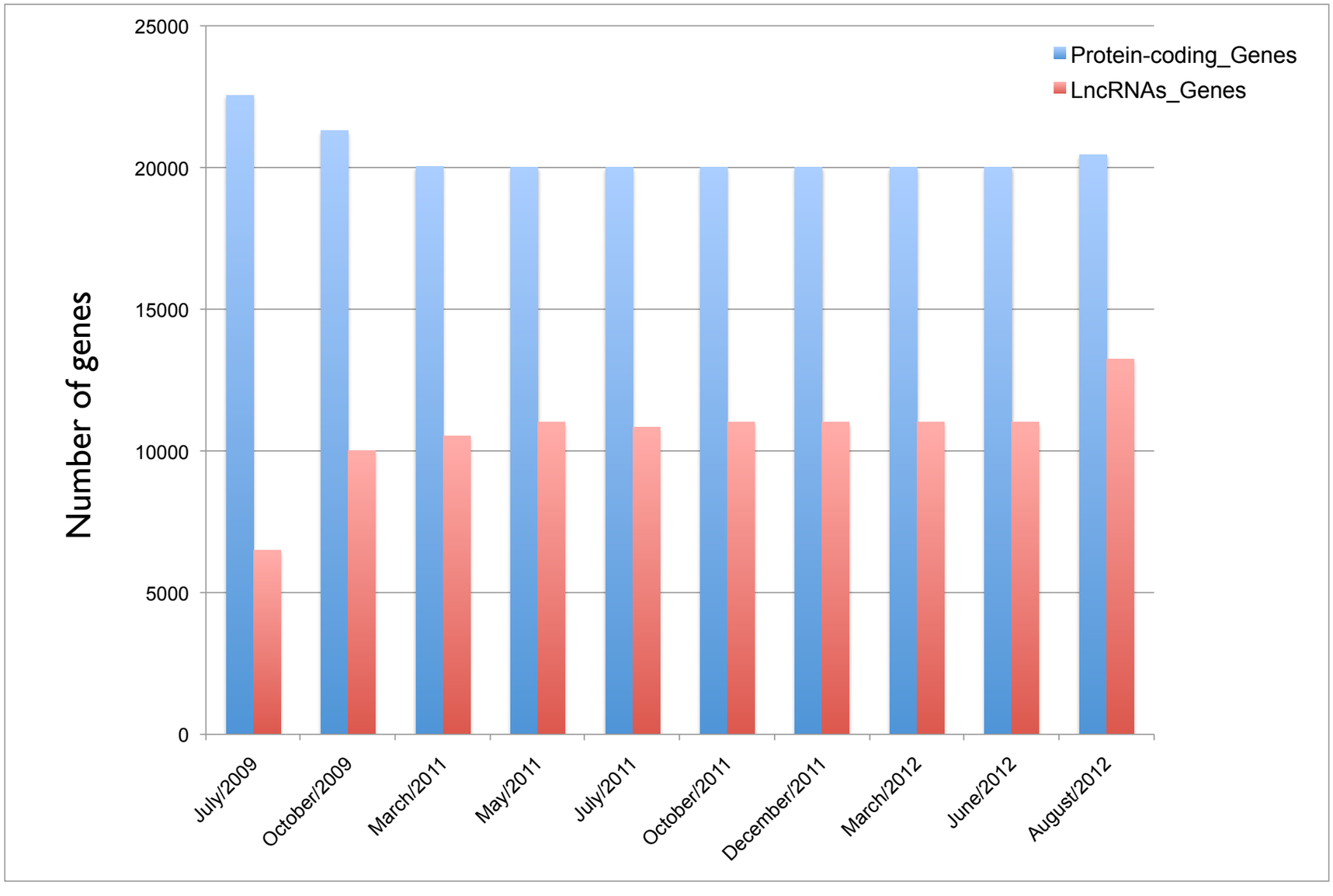
Amy Maxmen

Nature 496, 127-129 (2013) doi:10.1038/nj7443-127a

Type	functions
mRNAs	many..
…	…
miRNAs	Regulation of gene expression
siRNAs	RNA interference pathway
snoRNAs	Chemical modification of rRNA, tRNAs and small RNAs
piRNAs	transposon defense - regulate euchromatin formation
snRNA	splicing, regulation of TFs, telomere stability...
long ncRNAs	**Various**

What is known about lncRNAs

- Definition : Transcripts without coding potential , >200 nt, spliced, polyA+/- (Derrien et al., 2012)
- Annotation in human : e.g **GENCODE** reference annotation (Harrow et al., 2012, 1000 genomes project)



- "Famous" lncRNAs: *XIST*, *H19*, *HOTAIR*... (Guttman et al., Duret et al., Navarro et al., Ponting et al.,)
- Known functions: regulation of mRNAs expression, X chromosome inactivation, imprinting...

LncRNAs Functions

doi:10.1038/nature10398

lincRNAs act in the circuitry controlling pluripotency and differentiation

The EMBO Journal (2012) 31, 515–516 | © 2012 European Molecular Biology Organization | All Rights Reserved 0261-4189/12
www.embojournal.org

RNA driving the epigenetic bus

Cell

Transcription of Two Long Noncoding RNAs Mediates Mating-Type Control of Gametogenesis in Budding Yeast

Nature. 2012 Nov 15;491(7424):454-7. doi: 10.1038/nature11508. Epub 2012 Oct 14.

Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.

Epigenetic Regulation by Long Noncoding RNAs
Jeannie T. Lee
Science **338**, 1435 (2012);
DOI: 10.1126/science.1231776

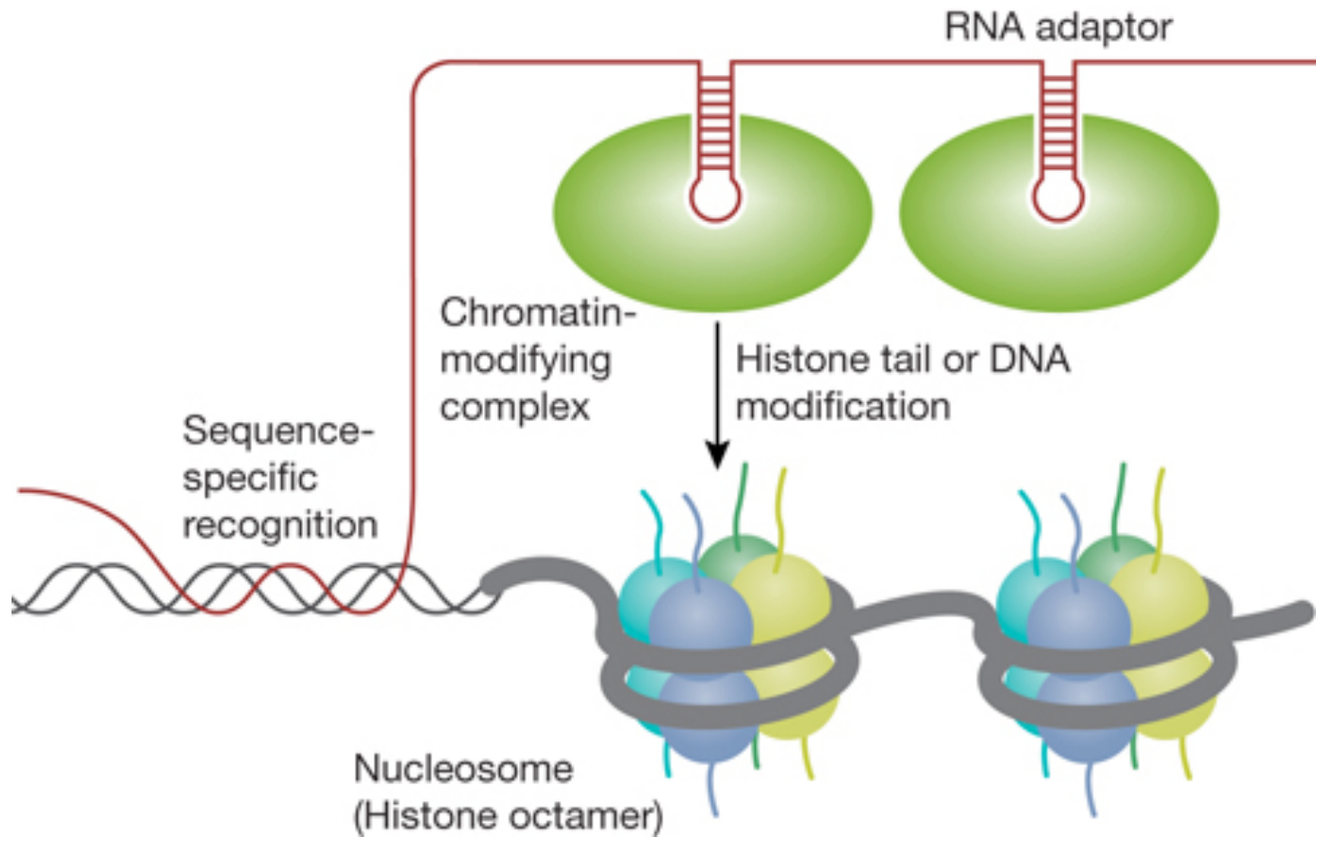
The Emergence of lncRNAs in Cancer Biology

John R. Prensner and Arul M. Chinnaiyan

The functional role of long non-coding RNA in human carcinomas

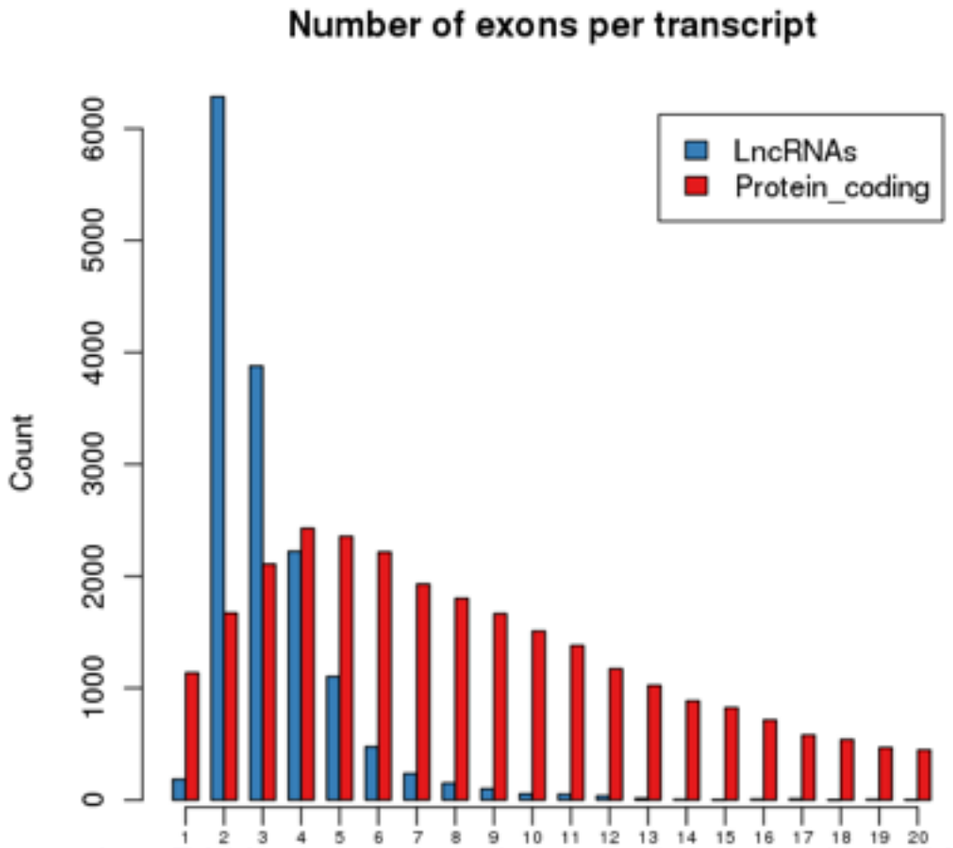
LncRNAs Functions

- Can enhance or repress transcription of targeted mRNA(s)
- Can act in *cis* or in *trans*
- sponge for miRNAs
- Serve as "flexible scaffolds"
- Examples:
  - **XIST** : binds PRC2 (DNMT3A) => DNA hypermethylation => silencing X chromosome
  - **HOTTIP** : binds MLL1 => H3K4me3 => activation of HOXA genes

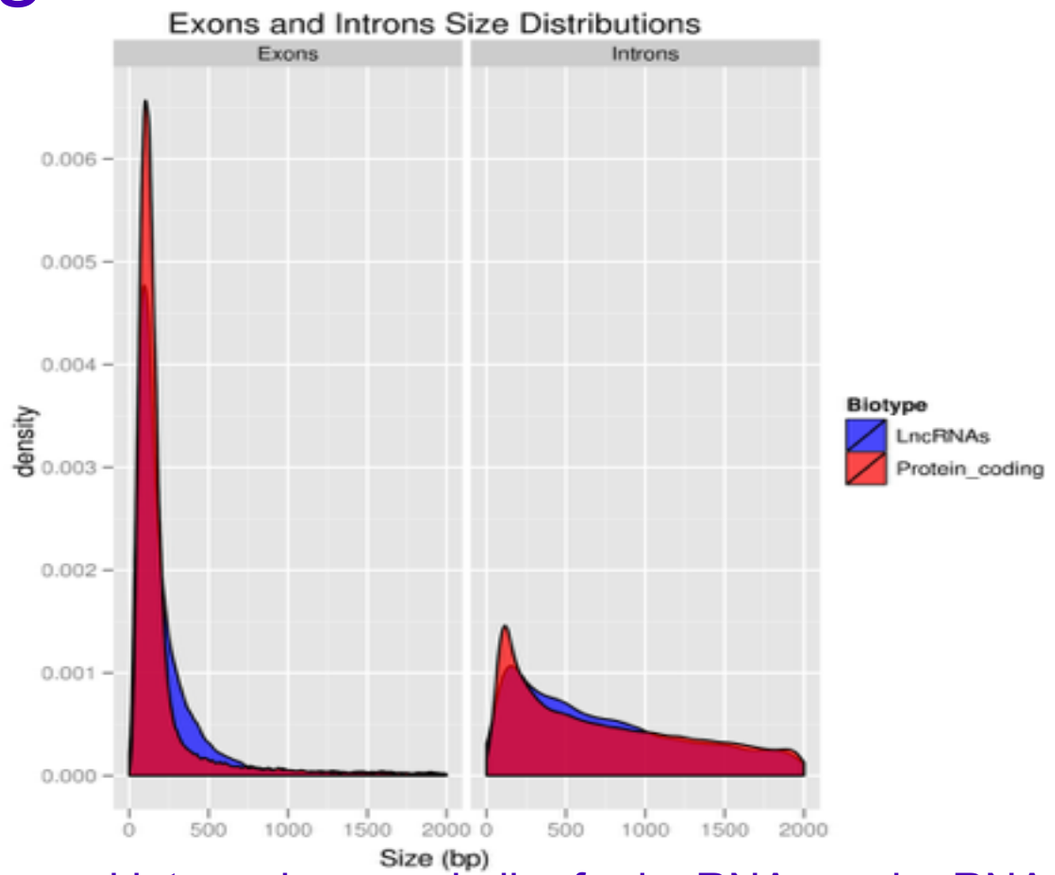


(from Mattick JS, et al., 2010)

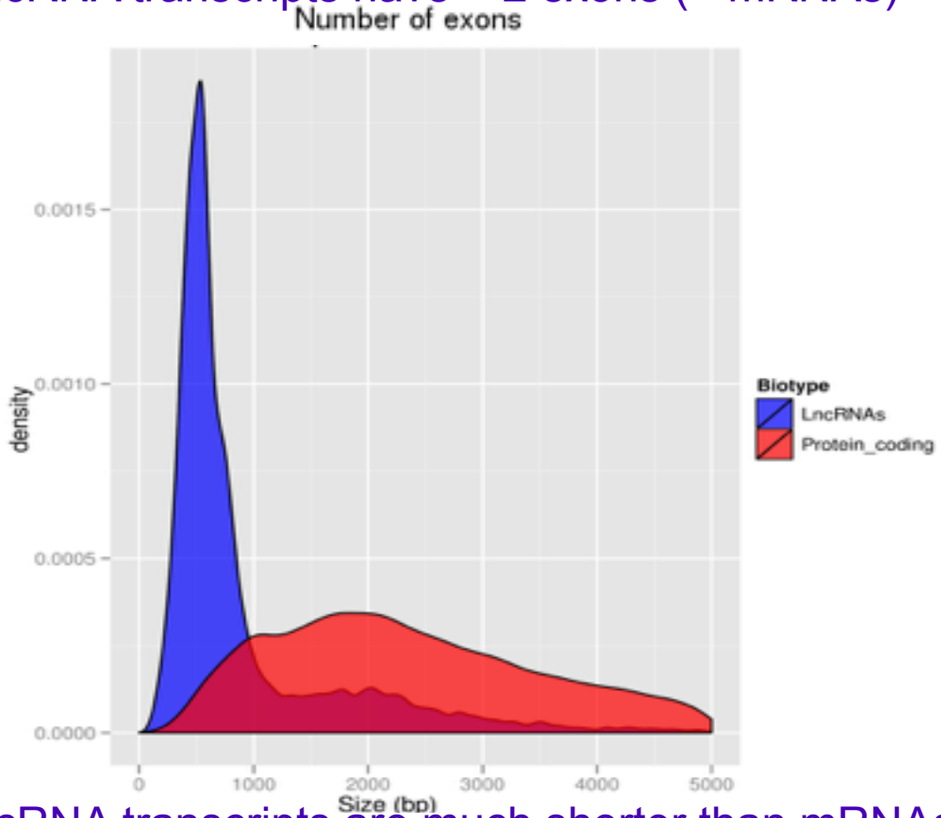
Features of lncRNA gene structure



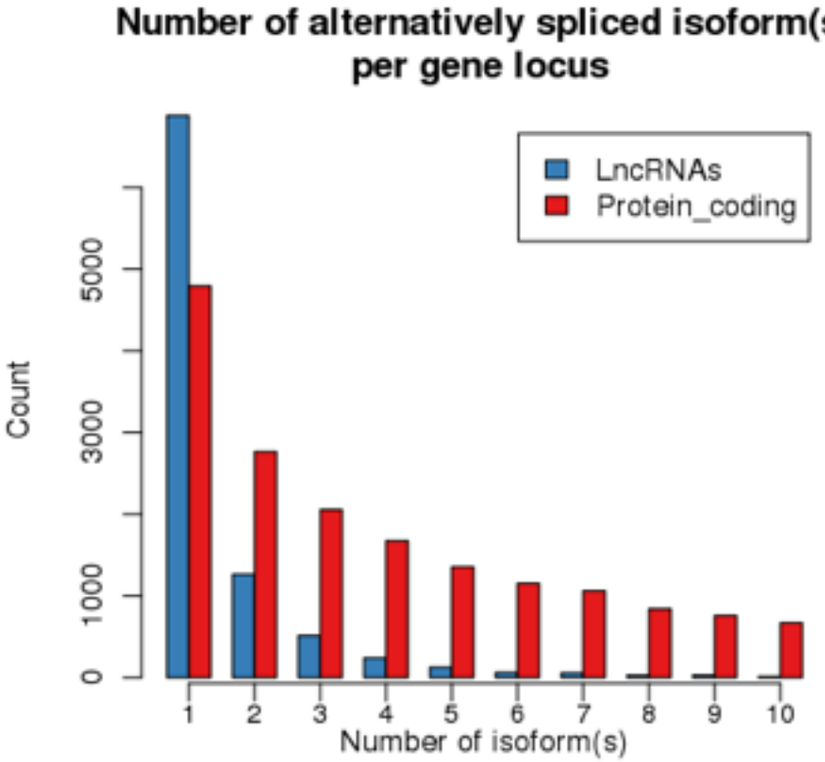
LncRNA transcripts have ~ 2 exons (< mRNAs)



Exon and intron size are similar for lncRNAs and mRNAs



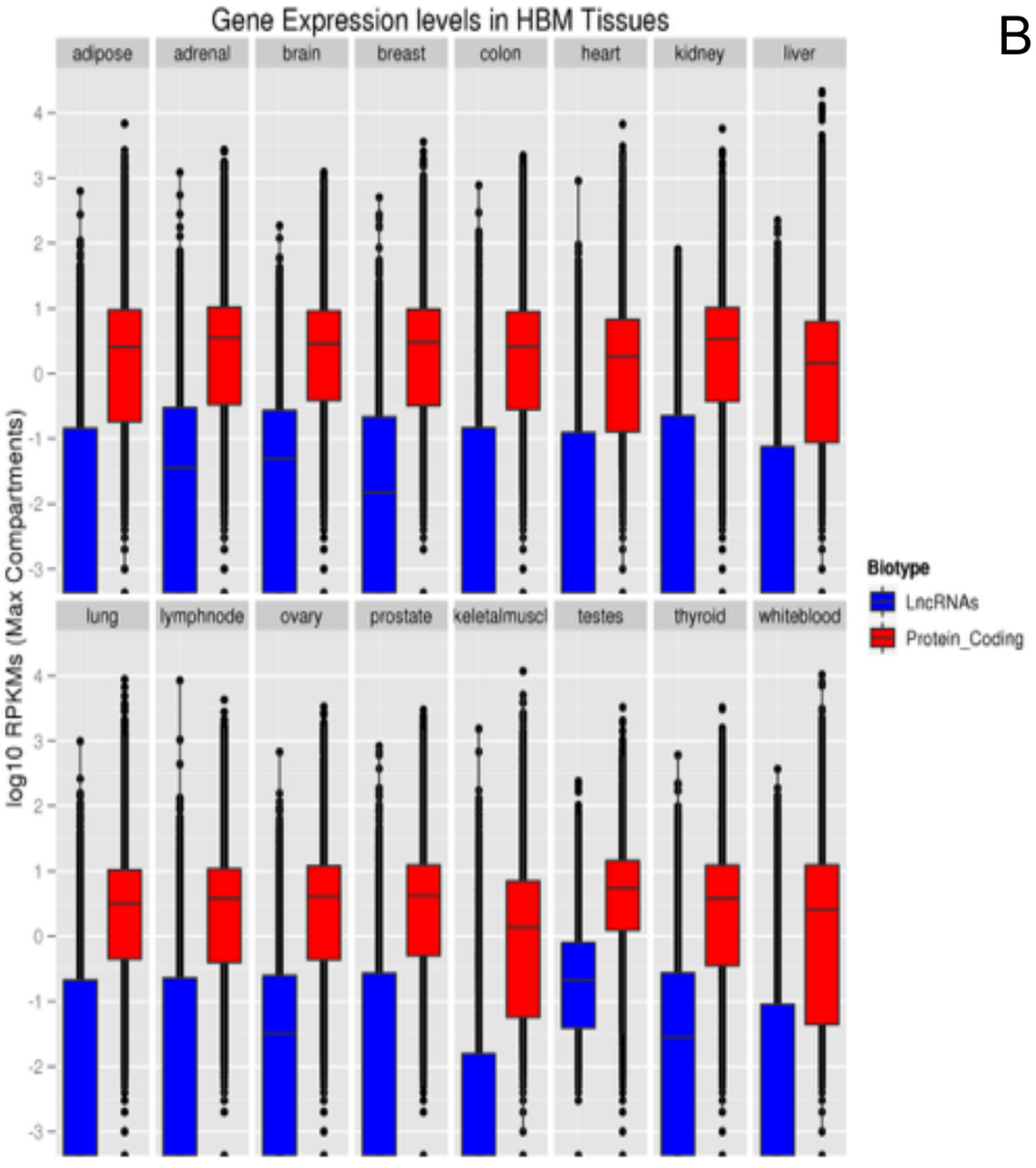
LncRNA transcripts are much shorter than mRNAs



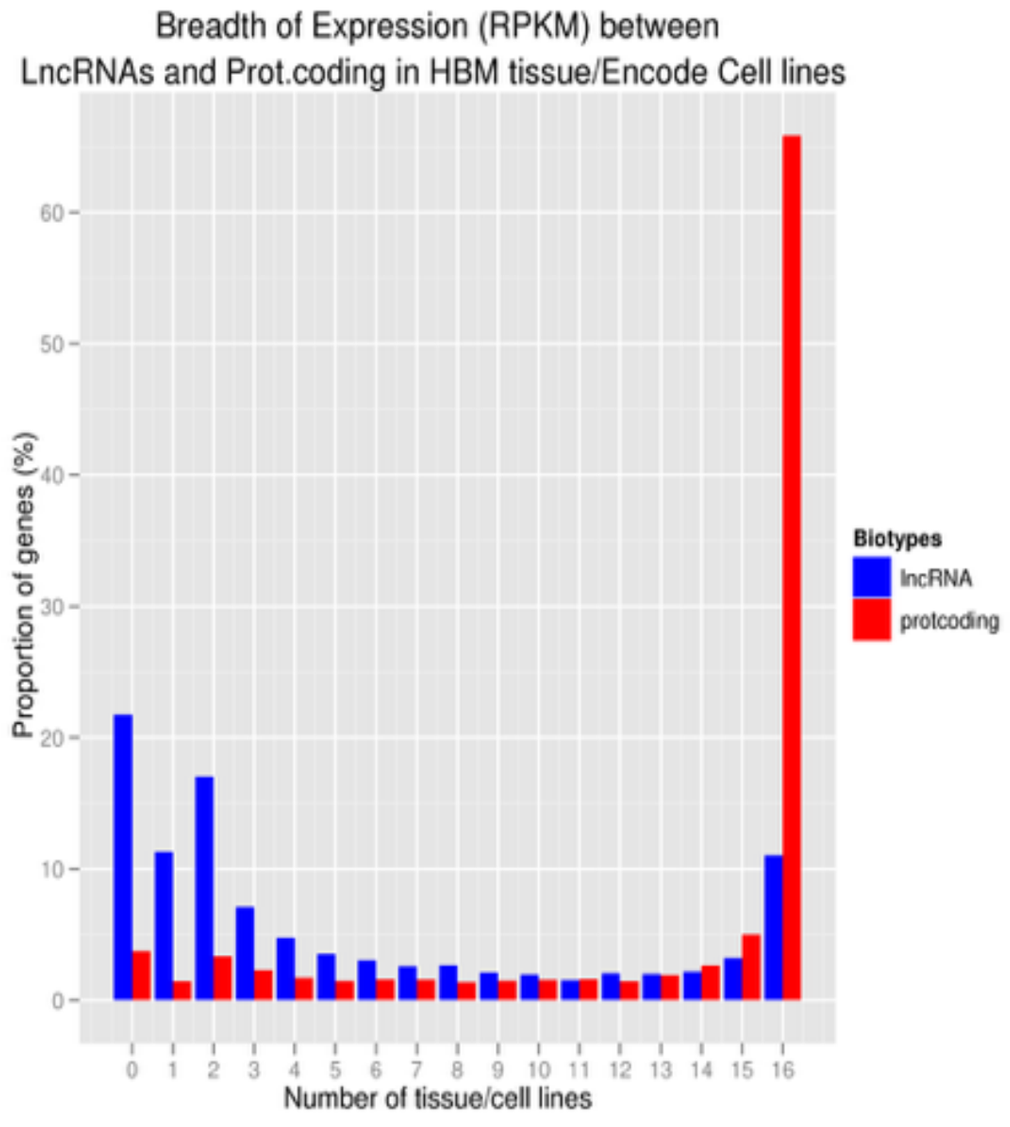
LncRNA genes have less isoforms than PCG genes

Characteristics of lncRNA expression in human cell types

A.



B.



<u>LncRNAs are less expressed than mRNAs in terms of:</u>

- expression levels (A),
- number of cell types in which they are found (B).

Derrien, Johnson et al., Genome research, 2012

ENCODE main messages

Whole genome ENCODE main messages

- The vast majority (**80.4%**) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.

- Nearly **60%** of the genome appears to be transcribed.

- Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.

- The ENCODE data (raw and processed) are available through dedicated websites (DCC) to the scientific community.

MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.

EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

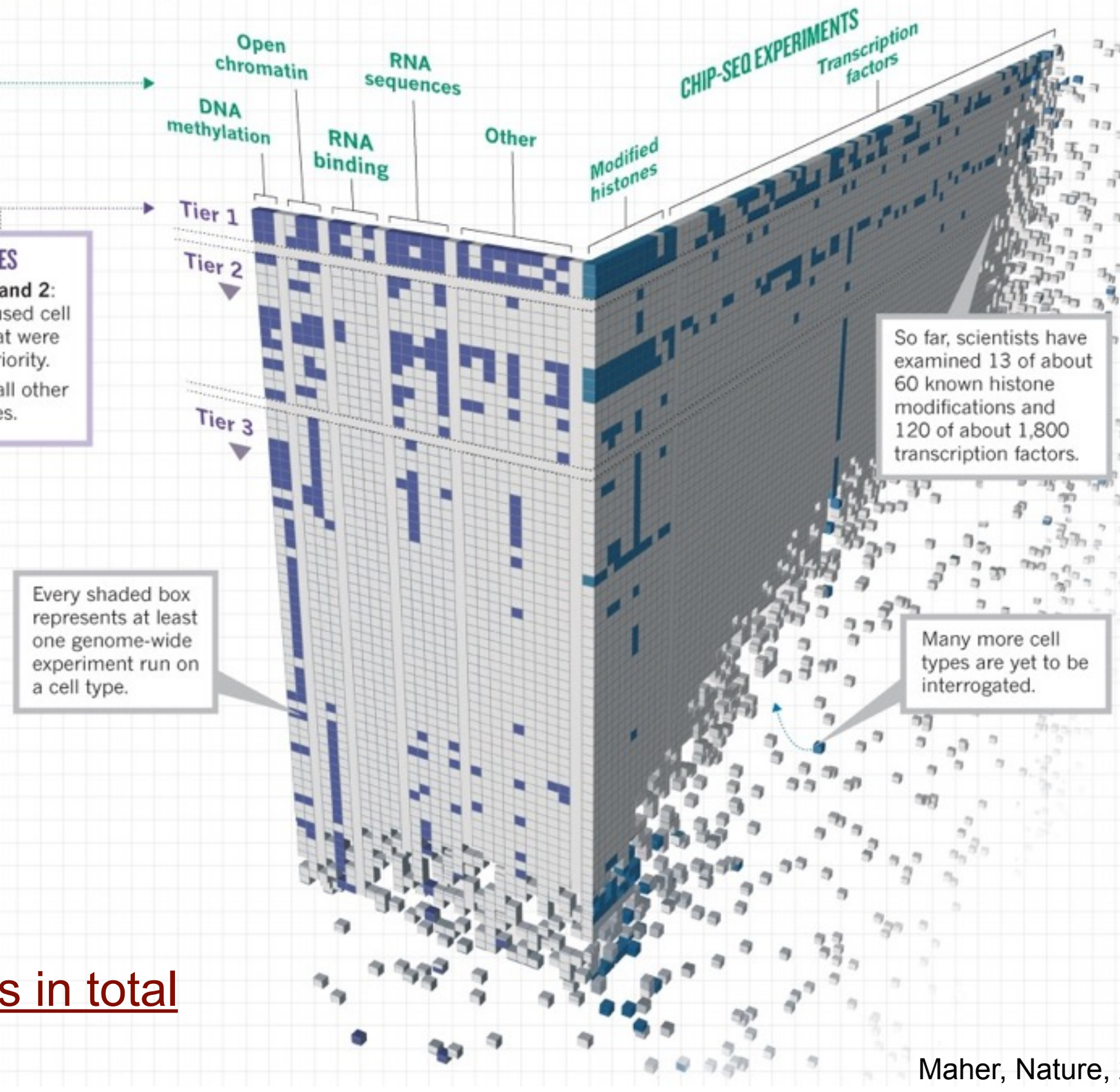
Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.



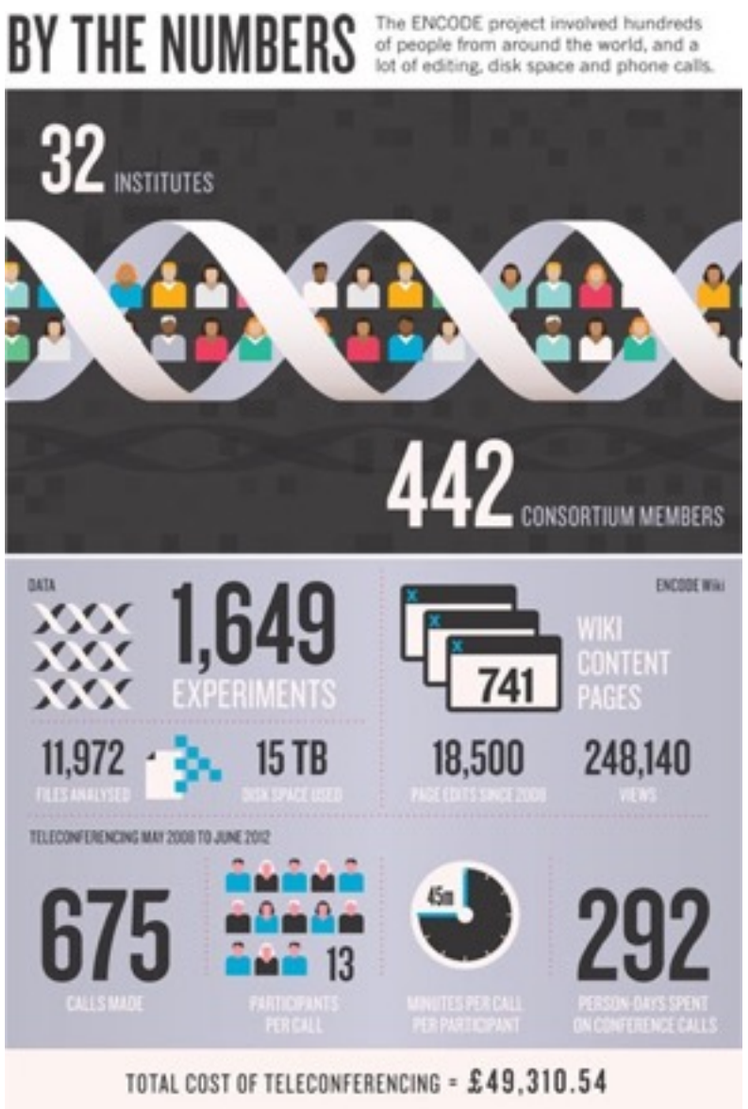
1,649 datasets in total

Maher, Nature, 2012

Whole genome ENCODE in numbers

- 442 consortium members in 32 institutes: coordination needed:
  - One analysis (AWG) call every week,
  - One transcriptome call every week (coordinated by CRG),
  - One DCC call every week,
  - One consortium call every month,
  - One PI call every month,
  - 2 meetings per year.
- Working for the community more than for one-self
  - Discussing ideas, be open for collaborations...

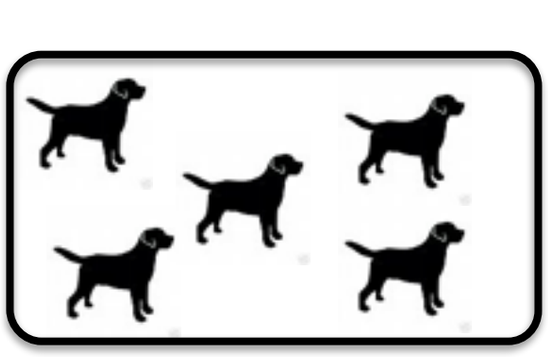
Birney, Nature, 2012

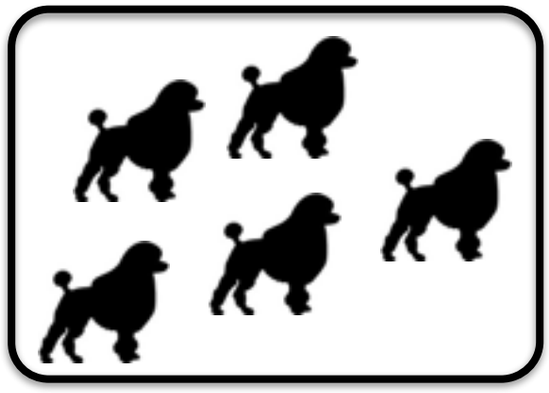


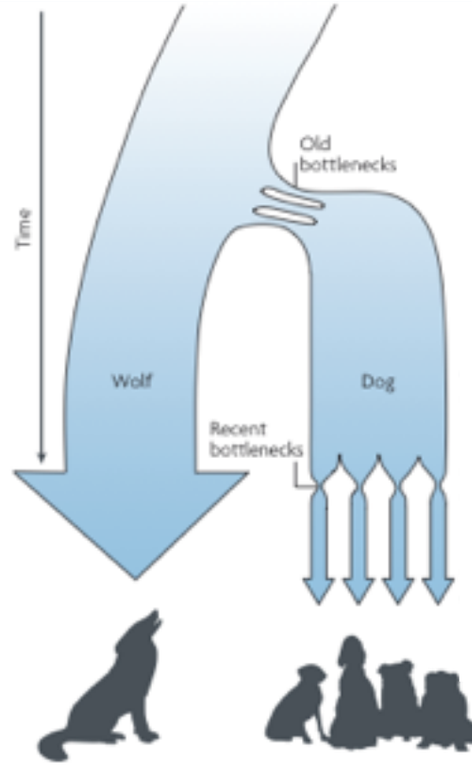
➡ ENCODE-like project in "non-model" organisms
(example in dogs)

Why dogs?

- Unique evolutionary history => unique population structure







- High heterogeneity bw breeds vs. High homogeneity within a breed
- One breed = One genetic isolate

➡ Most of the traits are governed by a few variants with high phenotypical effects

➡ **Dog model facilitates the identification of Genotype/Phenotype relationship**

Dog and disease/cancers

- Unique history => high prevalence of diseases/cancers

➡ Cancers in dogs :

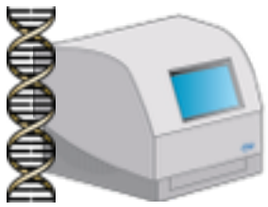
- Homologous to human cancers
- Breed-specific (high frequency within a breed ≈ 20%)
- Spontaneous cancers (and not induced like in mouse)
- Dogs share the same environment as humans
- Receive a high level of health care

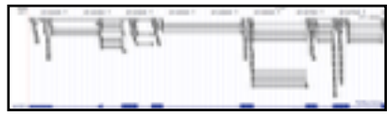
➡ **Dog is a good model for studying diseases/cancers**

(Dog genome sequenced: 4th mammals (K. Lindblad-Toh, et al., Nature, 2005)**)**

A typical project in the dog genetics team (Cancers…)

Vision for a complete ("ideal") workflow





Storing samples and characterization (BRC)

High-Throughput Sequencing technologies

Bioinformatics analyses
- resources
- dedicated programs

Results - Functional Validation

BRC : CaniDNA canidna.univ-rennes1.fr/

Sample Storing and characterization (BRC)

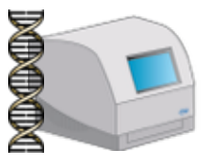
Vet network | Vet schools | Dog Owners

Sample characterization/ phenotyping

- histological
- types of cancers...

Biobank
CaniDNA

~20,000 DNA
>3,000 RNA



Sample Storing and characterization (BRC)

High-Throughput Sequencing technologies

BRC- Biobank
CaniDNA

Genotyping GWAS

DNASeq
- Exome
- Capture...

RNASeq
illumina

Genomic Locus associated with the disease

Variants (SNP, INDELs) associated with the disease

Annotation of coding and non-coding genome

RNASeq samples available in dog

UNIVERSITÉ DE RENNES 1

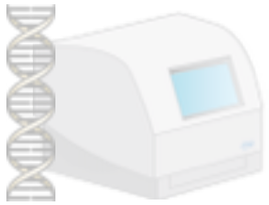
- 34 samples
- 28 from dogs at GIGA (Liège)
- 6 from dogs at CNG (Evry)
- Unstranded

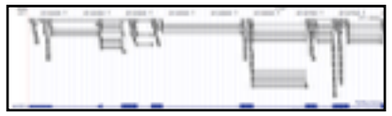
NCBI Site map | All databases | PubMed

Sequence Read Archive

- 24 samples
- 18 from dogs
- 6 from wolves
- Stranded and Not stranded

- 58 RNAseq
- 33 Dogs
- 10 Breeds
- 17 Tissues





Sample Storing and characterization (BRC)

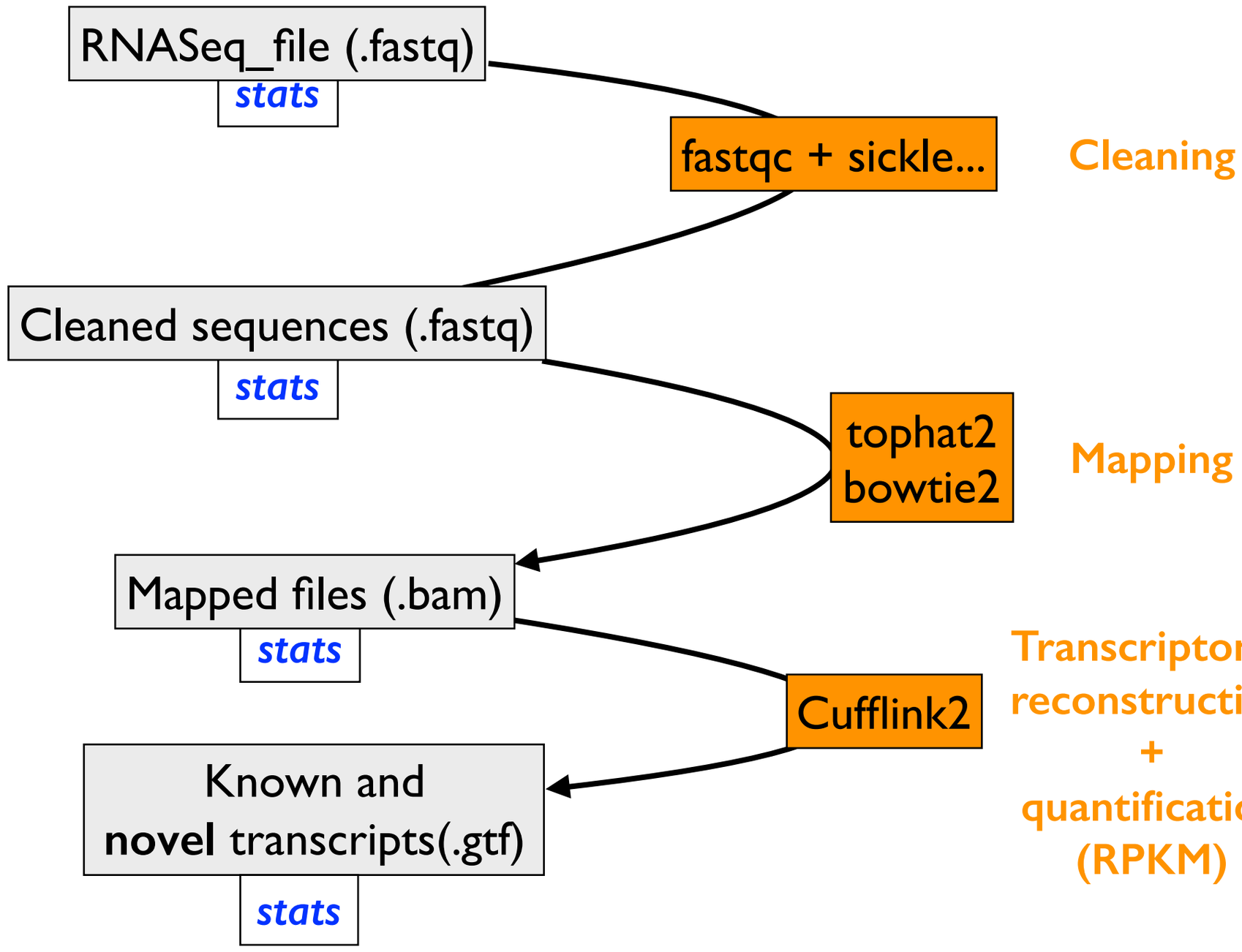
High-Throughput Sequencing technologies

Bioinformatics analyses
- resources
- dedicated programs

Pipeline for dog RNASeq analysis

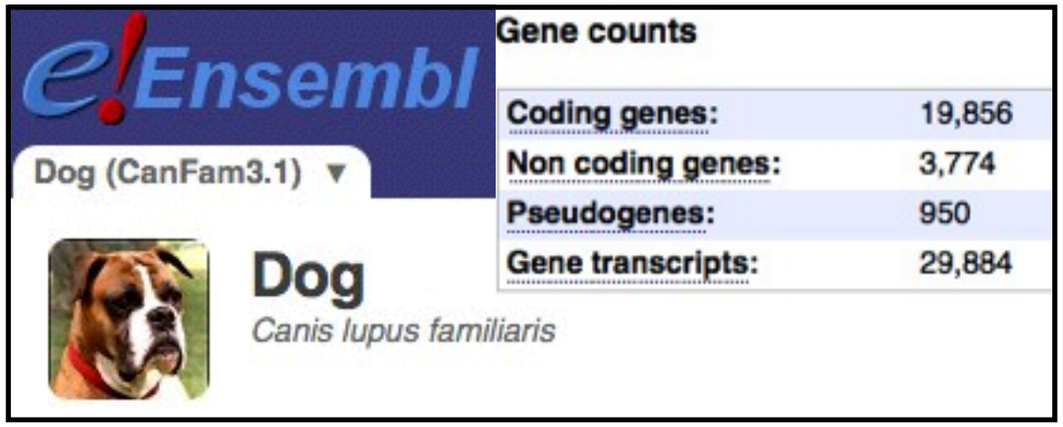
Dog Reference genome : canFam3
Dog Reference annotation : Ensembl (v75)

Christophe Hitte



Example of Brain (cortex) RNASeq

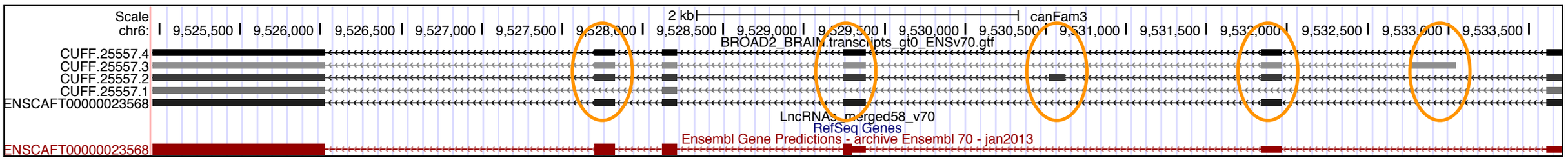
Current dog annotation



One RNASeq Experiment

	BRAIN RNASeq
-#Genes:	29,878
-#tcpts:	44,831

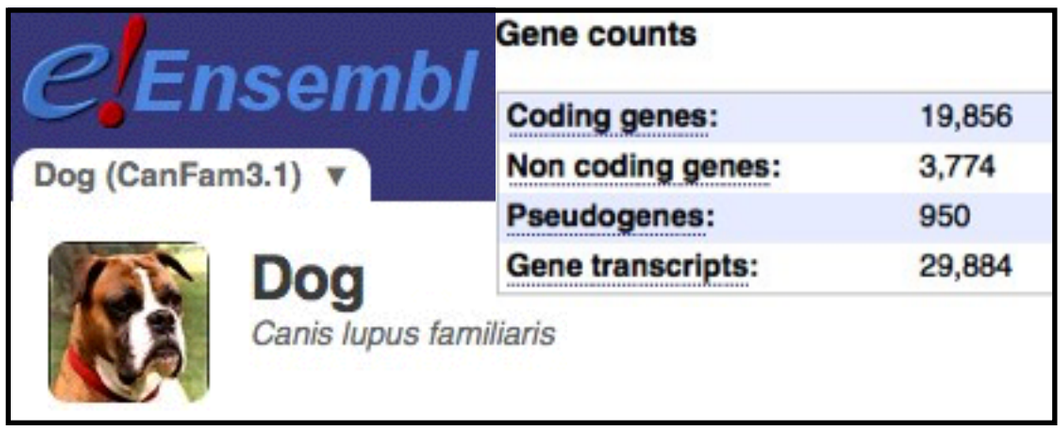
ZNF3-201



=> RNASeq allows to annotate new isoforms w.r.t to current reference annotations

Example of Brain (cortex) RNASeq

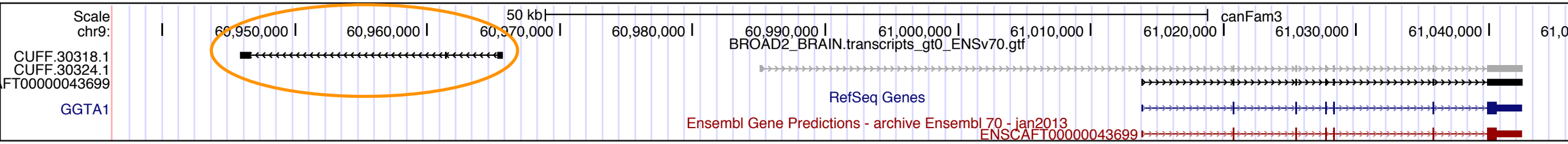
Current dog annotation



One RNASeq Experiment

	BRAIN RNASeq
-#Genes:	29,878
-#tcpts:	44,831

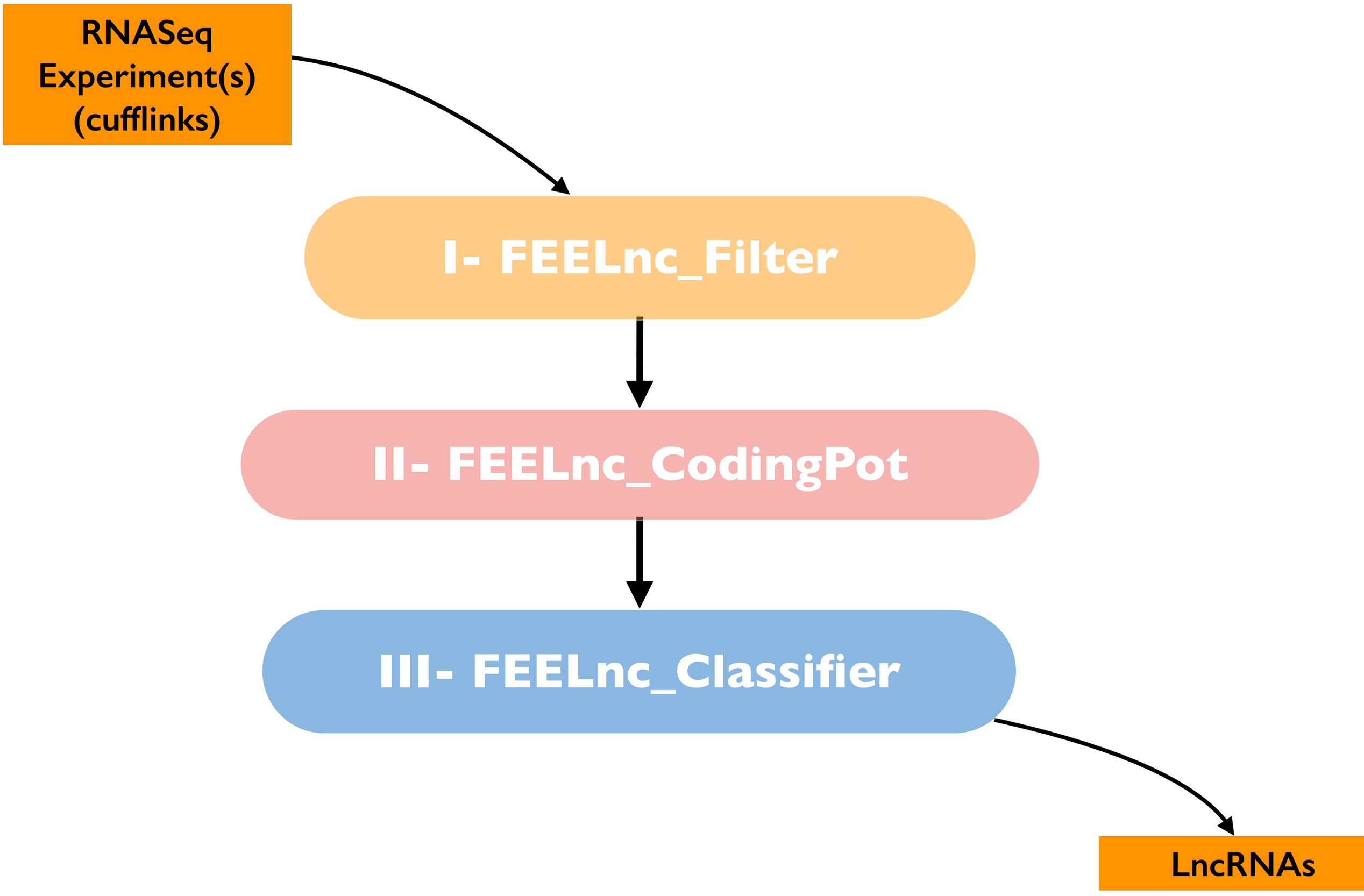
New transcript



=> RNASeq allows to annotate new (expressed) transcripts

=> Are these lncRNAs?

FEELnc : Fast and Effective Extraction of LncRNAs



FEELnc : Filters

Merged RNASeq samples (cuffmerge)

Known and novel transcripts

I- FEELnc_Filters

```
* Mandatory arguments:
    -i,--infile=file.gtf
    -a,--mRNAfile=file.gtf

* Filtering arguments:
    -s,--size=200
    -b,--biotype
    -l,--linconly
    --monoex=-1|0|1
    --biex=25

* Overlapping specification:
    -f,--minfrac_over=0
    -p,--proc=4
```

Candidate lncRNAs

-- biotype : only remove tx overlapping mRNAs ?

-- linconly : only keep intergenic tx

-- monoex : keep Antisense monoexonic tx?

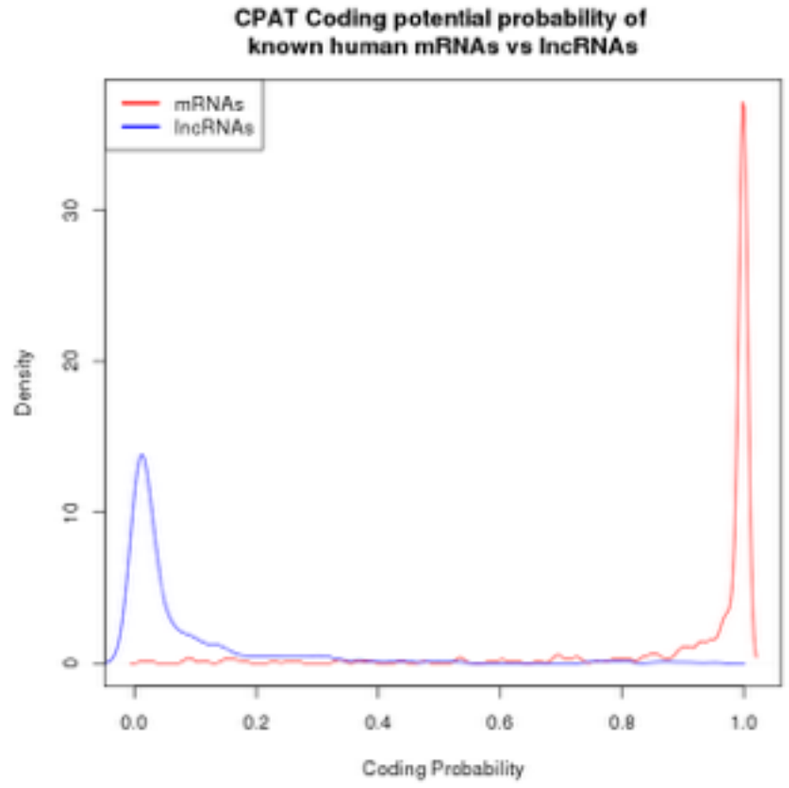
FEELnc : Coding potential

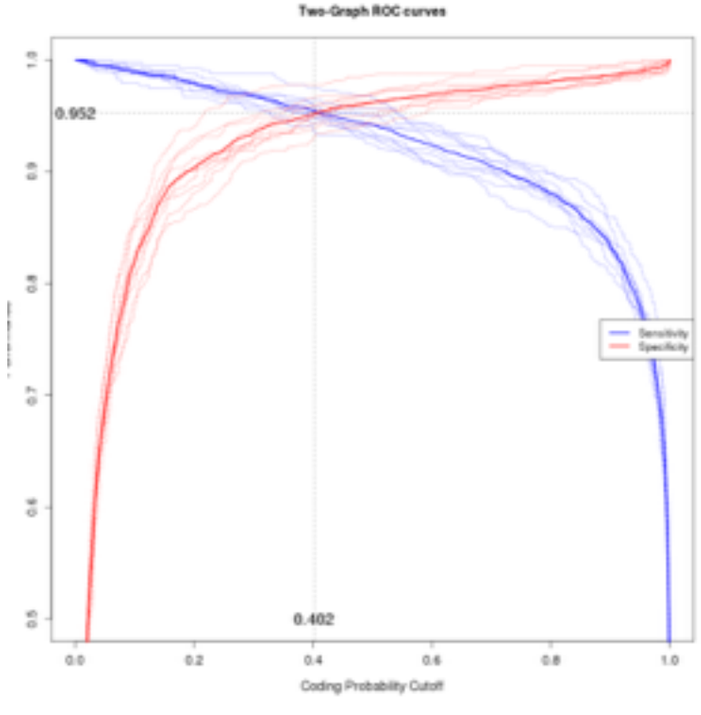
Candidate lncRNAs

II- FEELnc_CodingPot.

CPAT : Coding Potential Assessment Tool (Wang et al.)

- Alignment-free tool => fast
- logistic regression model based on 4 features
- Coding potential probability





New set of lncRNAs
(also new mRNAs)

FEELnc : Classifier

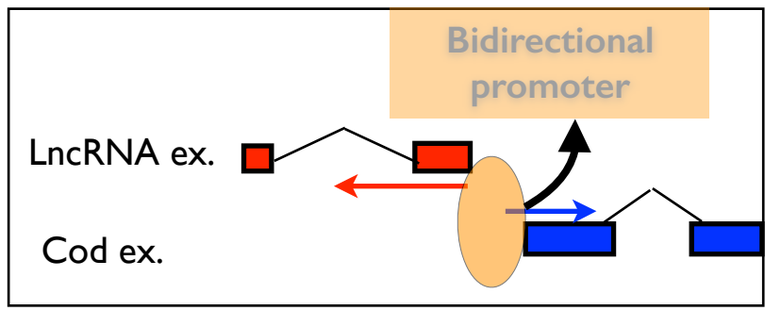
- Classifying lncRNAs genomic context wrt to mRNAs could help predict functionality

Set of lncRNAs

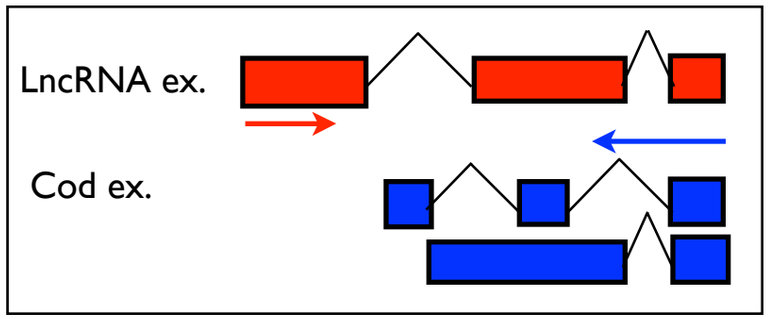
III- FEELnc_Classifier

Intergenic (lincRNA)	Genic (mRNA overlap)
Divergent	Exon (AS)
Convergent	Intron (S/AS)
Same Orient.	Contain (S/AS)

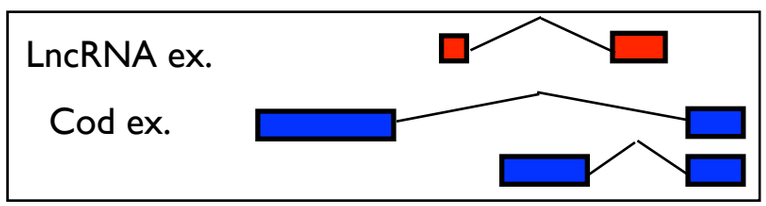
Schematic overlapping scenario



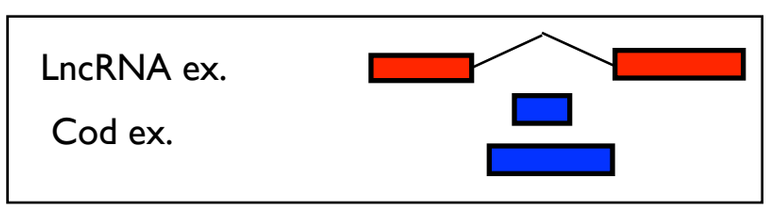
Divergent



Exonic AS



Intronic



Contain

FEELnc : In dog and chicken (S. Lagarrigue)

~60 RNASeq

RNASeq Experiment(s)

I- FEELnc_Filter

6 RNASeq
(adipose and liver tissues)

II- FEELnc_CodingPot

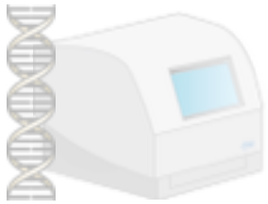
III- FEELnc_Classifier

Classified LncRNAs

lncRNA catalogue
-#tcpts: 18,050
-#genes: 9,810

lncRNA catalogue
-#tcpts: ~2,000
-#genes: 1,750

Vision for a complete (ideal) workflow



Sample Storing and characterization (BRC)

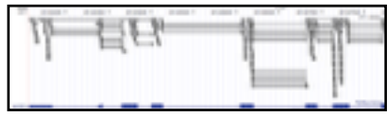
High-Throughput Sequencing technologies

Bioinformatics analyses
- resources
- dedicated programs

Results - Functional Validation

Conclusion

Conclusion (some critical points…)



Sample Storing and characterization (BRC)

High-Throughput Sequencing technologies

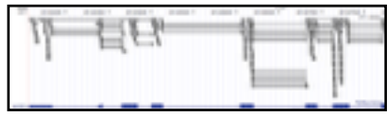
Bioinformatics analyses
- resources
- dedicated programs

Results - Functional Validation

- high level resources
- phenotyping

Which technology?

Conclusion (some critical points…)



Sample Storing and characterization (BRC)

- High resources
- phenotyping+++

High-Throughput Sequencing technologies

Which technology?

Bioinformatics analyses
- resources
- dedicated programs

- Quality of the reference genome/ annotation
- Bioinformatic platform needed
- biostatistics

Results - Functional Validation

- Which cell lines? (time consuming)
- Biochemical activity does not always mean function…

ACKNOWLEDGEMENTS

- IGDR. CNRS-UMR6290, Rennes

Christophe Hitte
Mathieu Bahin
Benoit Hédan
Amaury Vaysse
Jocelyn Plassais
Edouard Cadieu
Catherine ANDRÉ
Laetitia Lagoutte
Anne-Sophie Guillory
Clotilde de brito
Melanie Rault
Ronan Ulvé
Morgane Bunel

- Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine. University Liège

Benoit HENNUY
Wouter COPPIETERS

- BROAD Institute - Boston/Uppsala University

Jennifer MEADOW
Kerstin LINDBLAD-TOH

- Center for Genomic Regulation -Barcelona-

Sarah Djebali
Rory JOHNSON
Giovanni BUSSOTTI
Cédric NOTREDAME
Roderic GUIGÓ

- AgroCampus ouest Rennes

Sandrine Lagarrigue
Frederic Lecerf

- GABI - Jouy en Josas

Andrea Rau

- Biogenouest - INRIA - Genscale Team

Fabrice Legeai, Claire Lemaître, Pierre Peterlongo, Guillaume Rizk, D. Lavenier, **Olivier Collin**

LUPA