Genomic Characterization of Human Long Noncoding RNAs

Julien Lagarde

PhD defense

January 17th, 2020

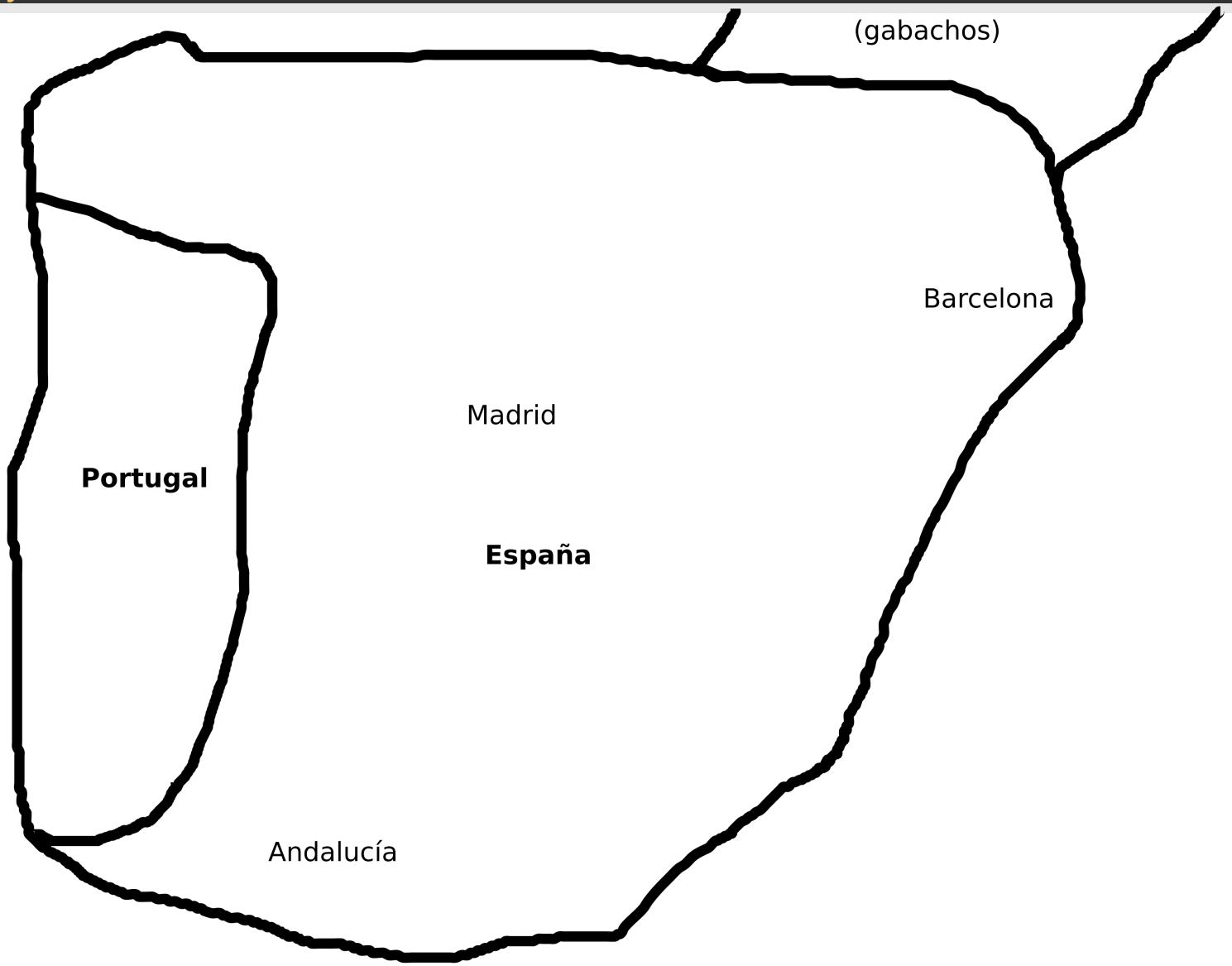
UNIVERSITAT DE BARCELONA

The current state of long noncoding RNA (lncRNA) annotation

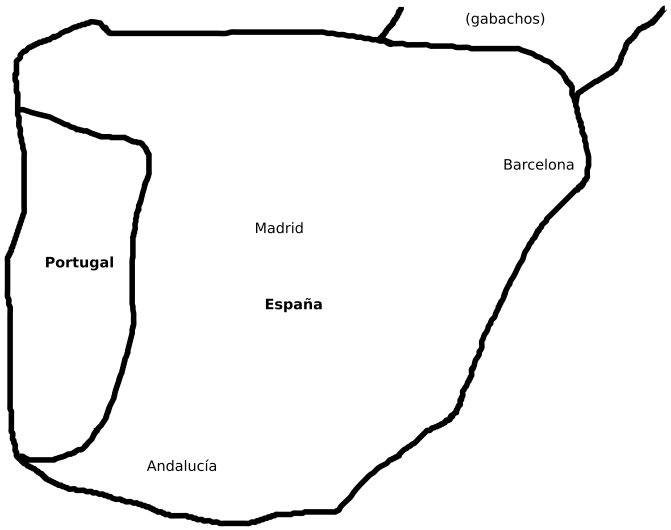
The goal

The current state of long noncoding RNA (lncRNA) annotation

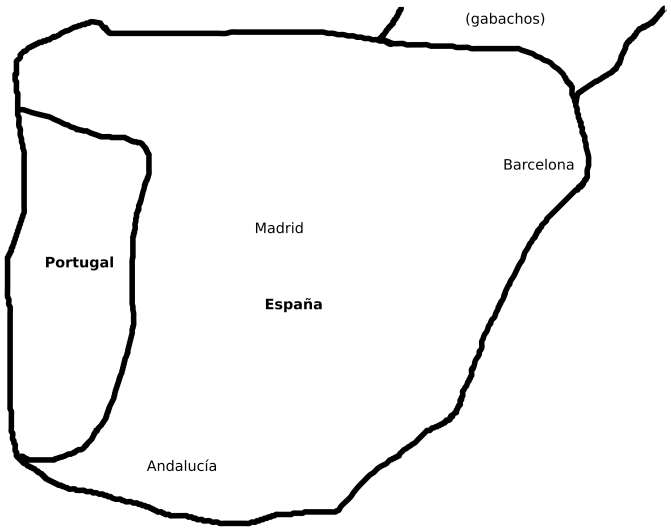
The reality



What this work is about: improving lncRNA annotation

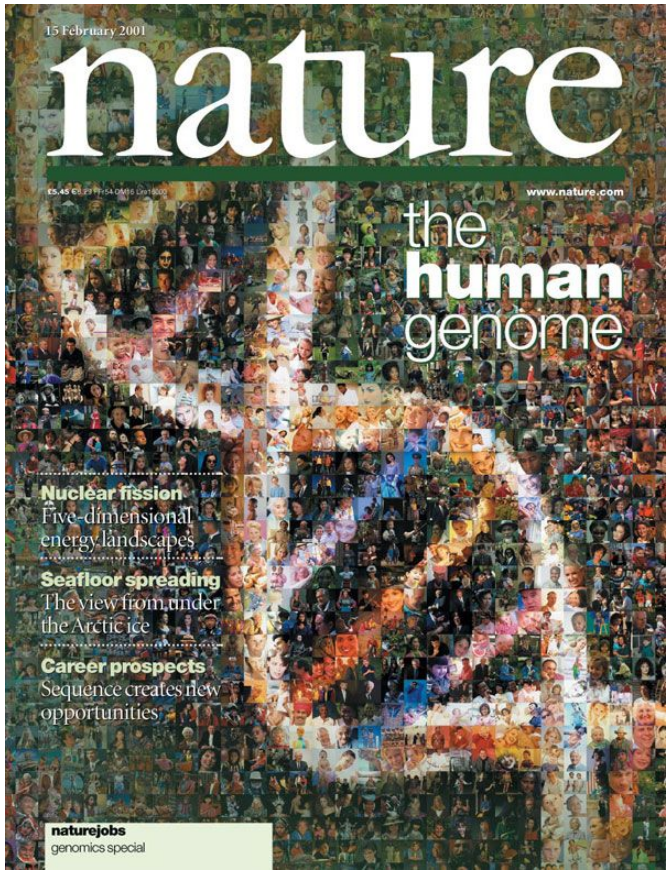


What this work is about: improving lncRNA annotation



Sequencing the Human Genome

Venter *et al. Science*, 2001



Lander *et al. Nature*, 2001

The complete sequence of the human genome

.....GGCTTGTAGTCACCTCGGAATTGTTAGGCGACGAAAAAAACTACTCACCCCAGGTGTCCCTGCGTCAGAGGCGGGTTTC
TCTCCTGTCTCCCTTTGCTCGCCCCTAAATCCTTCCTCGGACTCCAACTCCCGCCGTGCCTTGCGCTGAGATCCCTGCGGCAAAG
AACCGGGCTGTGTCCAAAGTGTTCTCTGGAAGTTGTAGTTCCTGTATTGGTGAGGCAAGGAGGAGGCGGAGTGACTCGGCGGCCA
TTAGCTGTGTGTAGTTGCCCGGGACTAGGAGCTTAAGTGAAGAGGTACGCCTTGTTCGGTGGAAATCAGCCGTAGCCATGAGTTT
CTGCCGGGGCTAGCCCTAGAGTACGGAGCAGGCGGACTTTTCGGTTCCCCGCCCCGCCAGGTGGCGGGGCCTACTAGGCCTCCGG
GCATCCCCGGTCTCAAGTAGGCCTCATCTGCCGGCAAGGGCGCCCGAAACGCGGGAGGCGCCATGTCGCTGGTTGCTTACGCCAG
CAGCGATGAGAGCGAGCCGGATGAGGCTGAGCCCGAGCCGGAGGAAGAGGAGGCGGTGGCTCCTACATCTGGGCCCGCTTTAGGG
GGCTTGTTCGCTTCTCTCCCTGCGCCCAAGGGTCCGGCCTTGCTGCCTCCGCCCCCTCAGATGCTGGCGCCAGCCTTTCCCCCGC
CGCTGTTGCTTCCCCCACCCACCGGAGACCCCAGGCTTCAGCCTCCTCCCCCCTTGCCCTTCGGCCTGGGAGGCTTCCCCCCACC
TCCAGGCGTGAGCCCGGCTGAAGCGGCGGGAGTTGGGGAGGGACTGGGATTGGGGTTGCCCTCGCCCCGAGGCCCTGGCCTCAAT
CTGCCCCCTCCAATTGGCGGTGCCGGTCCCCCGCTGGGGCTTCCCAAGCCAAAGAAGAGGAAAGAGCCCGTGAAGATCGCGGCGC
CGGAGTTGCATAAGGGAGATGTGAGTATCCGGGGAAGGCACCCCCAAACTGTCCATCGGGTTTGAGTCGGCTGTAGGAGGGACAT
GTGGAGATTTCAGAACTCCTTCCTACAAACGCATCCGTGGATTCAAGGCCACTGGAATCCTCTTATAATGCGATTAATTGAGCTG
AAGTTGGTCAACTTGGGGGAGTTTGGGTGTTTGCCCATGTGGGTTTCCTGAA.....

The complete (bare) sequence of the human genome

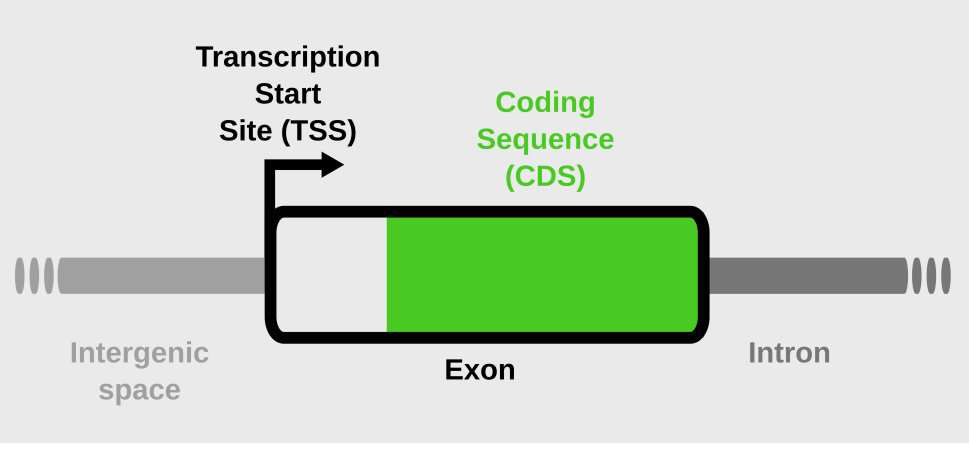
Annotating the human genome

The complete sequence of the human genome

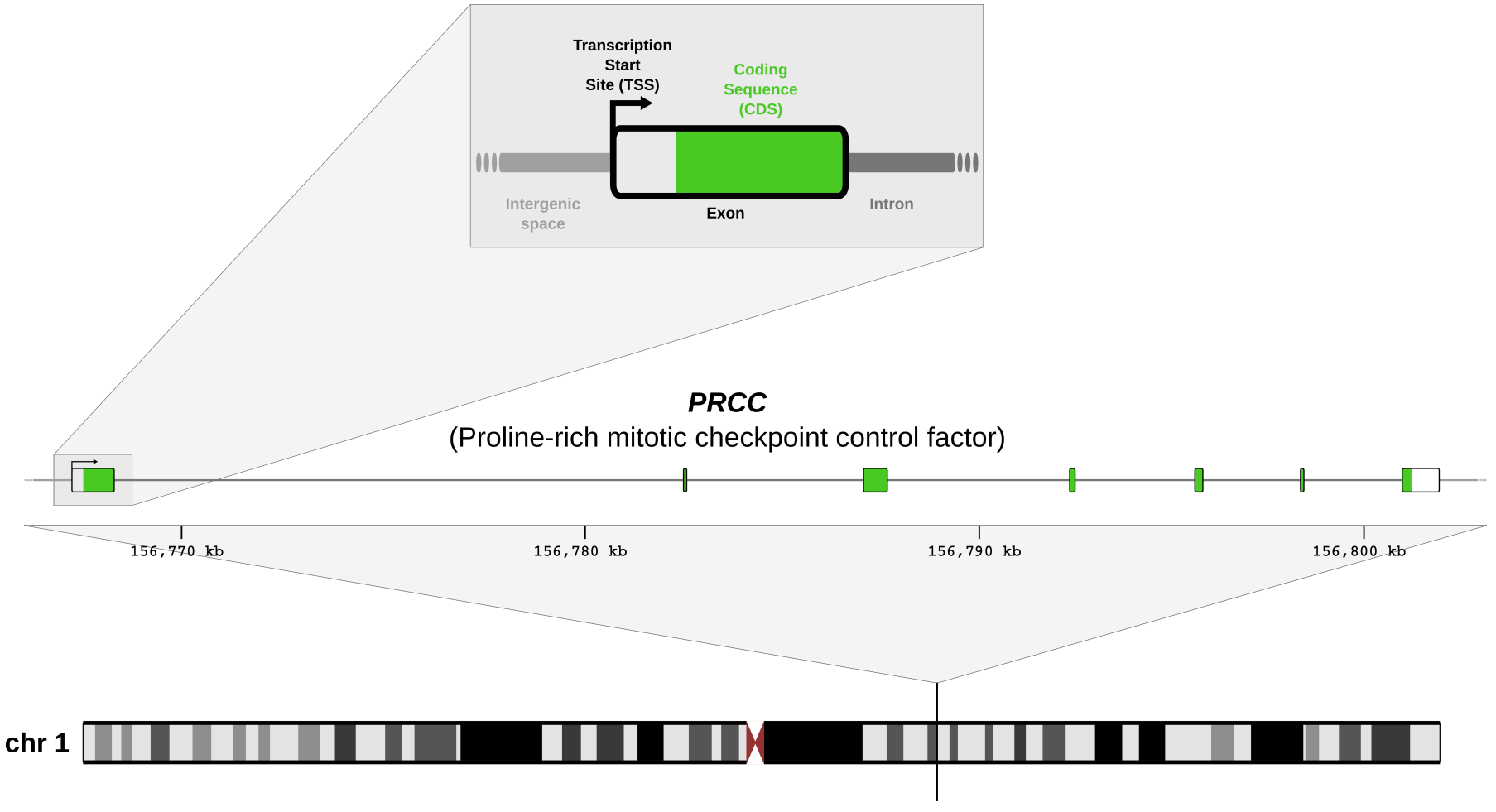
.....GGCTTGTAGTCACCTCGGAATTGTTAGGCGACGAAAAAAACTACTCACCCCAGGTGTCCCTGCGTCAGAGGCGGGTTTC
TCTCCTGTCTCCCTTTGCTCGCCCCTAAATCCTTCCTCGGACTCCAACTCCCGCCGTGCCTTGCGCTGAGATCCCTGCGGCAAAG
AACCGGGCTGTGTCCAAAGTGTTCTCTGGAAGTTGTAGTTCCTGTATTGGTGAGGCAAGGAGGAGGCGGAGTGACTCGGCGGCCA
TTAGCTGTGTGTAGTTGCCCGGGACTAGGAGCTTAAGTGAAGAGGTACGCCTTGTTCGGTGGAAATCAGCCGTAGCCATGAGTTT
CTGCCGGGGCTAGCCCTAGAGTACGGAGCAGGCGGACTTTTCGGTTCCCCGCCCCGCCAGGTGGCGGGGCCTACTAGGCCTCCGG
GCATCCCCGGTCTCAAGTAGGCCTCATCTGCCGGCAAGGGCGCCCGAAACGCGGGAGGCGCCATGTCGCTGGTTGCTTACGCCAG
CAGCGATGAGAGCGAGCCGGATGAGGCTGAGCCCGAGCCGGAGGAAGAGGAGGCGGTGGCTCCTACATCTGGGCCCGCTTTAGGG
GGCTTGTTCGCTTCTCTCCCTGCGCCCAAGGGTCCGGCCTTGCTGCCTCCGCCCCCTCAGATGCTGGCGCCAGCCTTTCCCCCGC
CGCTGTTGCTTCCCCCACCCACCGGAGACCCCAGGCTTCAGCCTCCTCCCCCCTTGCCCTTCGGCCTGGGAGGCTTCCCCCCACC
TCCAGGCGTGAGCCCGGCTGAAGCGGCGGGAGTTGGGGAGGGACTGGGATTGGGGTTGCCCTCGCCCCGAGGCCCTGGCCTCAAT
CTGCCCCCTCCAATTGGCGGTGCCGGTCCCCCGCTGGGGCTTCCCAAGCCAAAGAAGAGGAAAGAGCCCGTGAAGATCGCGGCGC
CGGAGTTGCATAAGGGAGATGTGAGTATCCGGGGAAGGCACCCCCAAACTGTCCATCGGGTTTGAGTCGGCTGTAGGAGGGACAT
GTGGAGATTTCAGAACTCCTTCCTACAAACGCATCCGTGGATTCAAGGCCACTGGAATCCTCTTATAATGCGATTAATTGAGCTG
AAGTTGGTCAACTTGGGGGAGTTTGGGTGTTTGCCCATGTGGGTTTCCTGAA.....

The annotated human genome

.....GGCTTGTAGTCACCTCGGAATTGTTAGGCGACGAAAAAAACTACTCACCCCAGGTGTCCCTGCGTCAGAGGCGGGTTTC
TCTCCTGTCTCCCTTTGCTCGCCCCTAAATCCTTCCTCGGACTCCAACTCCCGCCGTGCCTTGCGCTGAGATCCCTGCGGCAAAG
AACCGGGCTGTGTCCAAAGTGTTCTCTGGAAGTTGTAGTTCCTGTATTGGTGAGGCAAGGAGGAGGCGGAGTGACTCGGCGGCCA
TTAGCTGTGTGTAGTTGCCCGGGACTAGGAGCTTAAGTGAAGAGGTACGCCTTGTTCGGTGGAAATCAGCCGTAGCCATGAGTTT
CTGCCGGGGCTAGCCCTAGAGTACGGAGCAGGCGGACTTTTCGGTTCCCCGCCCCGCCAGGTGGCGGGGCCTACTAGGCCTCCGG
GCATCCCCGGTCTCAAGTAGGCCTCATCTGCCGGCAAGGGCGCCCGAAACGCGGGAGGCGCCATGTCGCTGGTTGCTTACGCCAG
CAGCGATGAGAGCGAGCCGGATGAGGCTGAGCCCGAGCCGGAGGAAGAGGAGGCGGTGGCTCCTACATCTGGGCCCGCTTTAGGG
GGCTTGTTCGCTTCTCTCCCTGCGCCCAAGGGTCCGGCCTTGCTGCCTCCGCCCCCTCAGATGCTGGCGCCAGCCTTTCCCCCGC
CGCTGTTGCTTCCCCCACCCACCGGAGACCCCAGGCTTCAGCCTCCTCCCCCCTTGCCCTTCGGCCTGGGAGGCTTCCCCCCACC
TCCAGGCGTGAGCCCGGCTGAAGCGGCGGGAGTTGGGGAGGGACTGGGATTGGGGTTGCCCTCGCCCCGAGGCCCTGGCCTCAAT
CTGCCCCCTCCAATTGGCGGTGCCGGTCCCCCGCTGGGGCTTCCCAAGCCAAAGAAGAGGAAAGAGCCCGTGAAGATCGCGGCGC
CGGAGTTGCATAAGGGAGATGTGAGTATCCGGGGAAGGCACCCCCAAACTGTCCATCGGGTTTGAGTCGGCTGTAGGAGGGACAT
GTGGAGATTTCAGAACTCCTTCCTACAAACGCATCCGTGGATTCAAGGCCACTGGAATCCTCTTATAATGCGATTAATTGAGCTG
AAGTTGGTCAACTTGGGGGAGTTTGGGTGTTTGCCCATGTGGGTTTCCTGAA.....



The annotated human genome

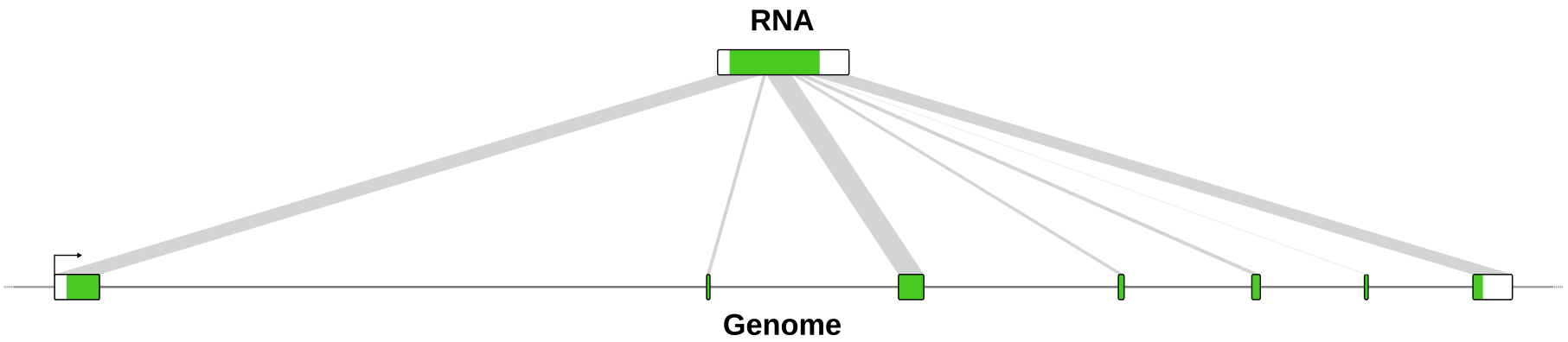


How to annotate genes

The gold standard: evidence-based, manually-curated gene annotation

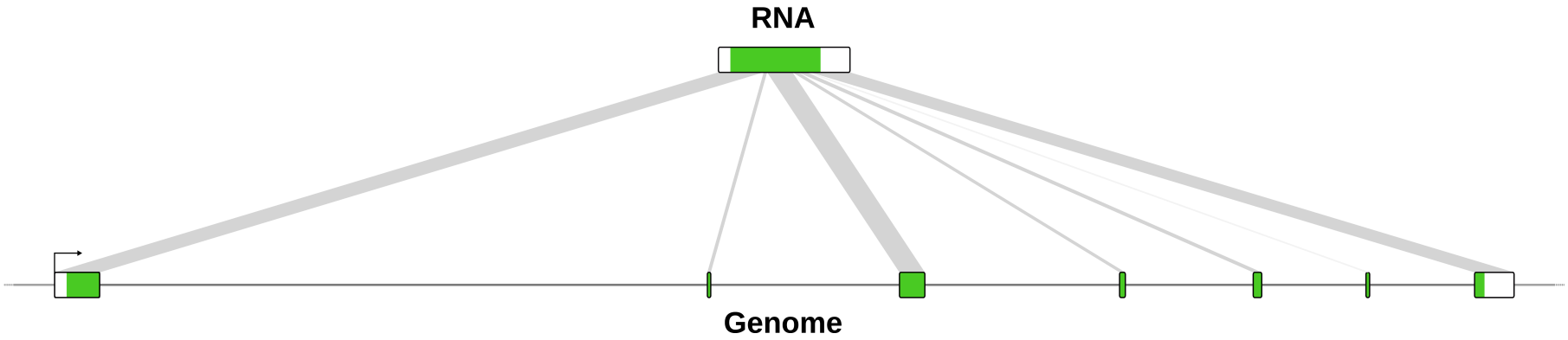
How to annotate genes

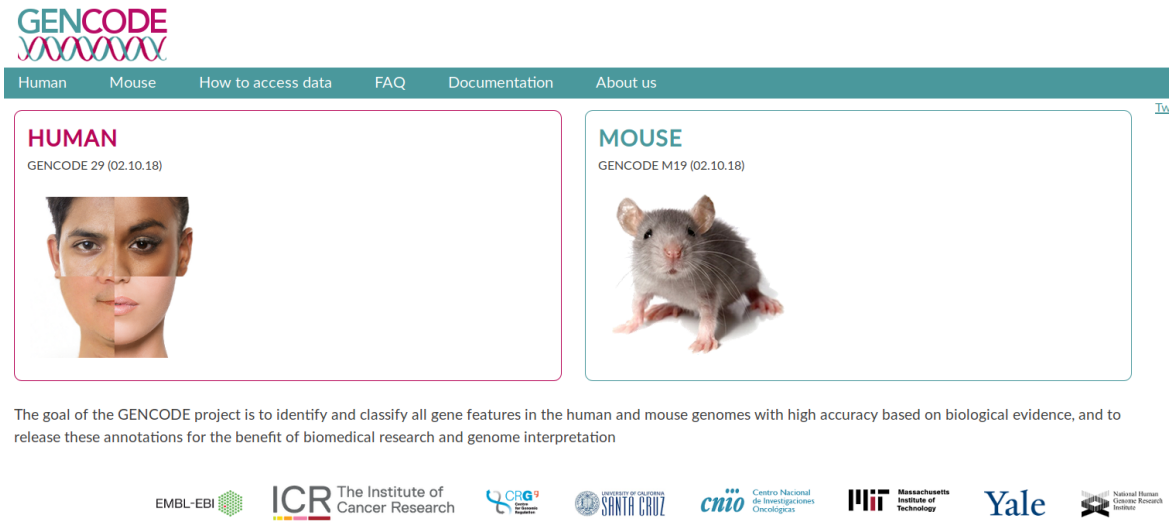
The gold standard: evidence-based, manually-curated gene annotation



How to annotate genes

The gold standard: evidence-based, manually-curated gene annotation





https://www.gencodegenes.org/

Sequencing RNA (cDNA)

Comparison of DNA sequencing platforms (*ca.* 2016)

Performance	poor					good

	1st-gen.
Technology	**Sanger**
Available	1977-
Read length (bases)	~800
Yield (# reads)	(low)
Base accuracy (%)	99.999

Goodwin, McPherson, and McCombie *Nature Reviews. Genetics*, 2016

Sequencing RNA (cDNA)

Comparison of DNA sequencing platforms (*ca.* 2016)

Performance	poor					good

	1st-gen.	2nd-gen. (SGS)
Technology	**Sanger**	**Illumina**
Available	1977-	2006-
Read length (bases)	~800	150
Yield (# reads)	(low)	>200M
Base accuracy (%)	99.999	99.9

Goodwin, McPherson, and McCombie *Nature Reviews. Genetics*, 2016

Sequencing RNA (cDNA)

Comparison of DNA sequencing platforms (*ca.* 2016)

Performance	poor					good

	1st-gen.	2nd-gen. (SGS)	3rd-gen. (TGS)	
Technology	**Sanger**	**Illumina**	**PacBio (RSII)**	**ONT**
Available	1977-	2006-	2009-	2014-
Read length (bases)	~800	150	<20k	<200k
Yield (# reads)	(low)	>200M	55k	100k
Base accuracy (%)	99.999	99.9	99	88

Goodwin, McPherson, and McCombie *Nature Reviews. Genetics*, 2016

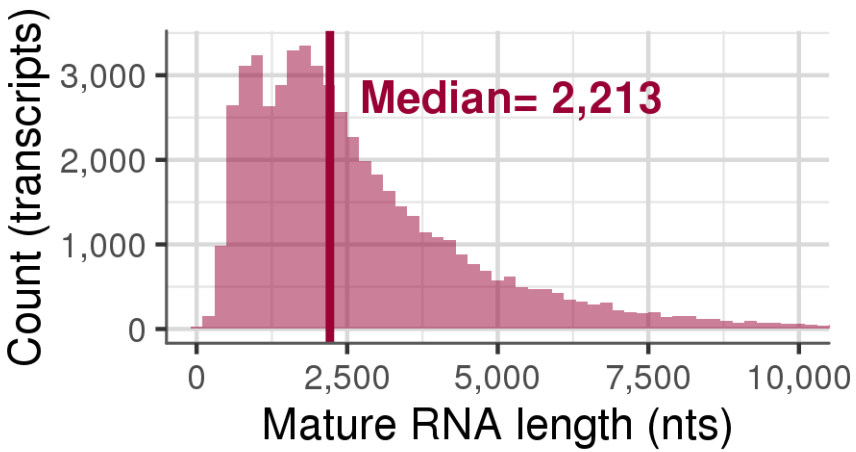
Sequencing RNA (cDNA)

Comparison of DNA sequencing platforms (*ca.* 2016)

Performance: poor → good

	1st-gen.	2nd-gen. (SGS)	3rd-gen. (TGS)	
Technology	**Sanger**	**Illumina**	**PacBio (RSII)**	**ONT**
Available	1977-	2006-	2009-	2014-
Read length (bases)	~800	150	<20k	<200k
Yield (# reads)	(low)	>200M	55k	100k
Base accuracy (%)	99.999	99.9	99	88

Goodwin, McPherson, and McCombie *Nature Reviews. Genetics*, 2016



(GENCODE v21)

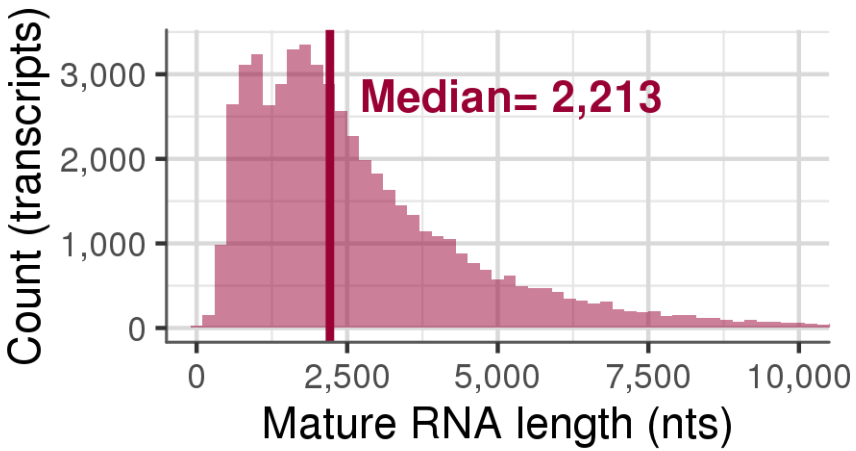
Sequencing RNA (cDNA)

Comparison of DNA sequencing platforms (*ca.* 2016)

Performance	poor	good

	1st-gen.	2nd-gen. (SGS)	3rd-gen. (TGS)	
Technology	**Sanger**	**Illumina**	**PacBio (RSII)**	**ONT**
Available	1977-	2006-	2009-	2014-
Read length (bases)	~800	150	<20k	<200k
Yield (# reads)	(low)	>200M	55k	100k
Base accuracy (%)	99.999	99.9	99	88

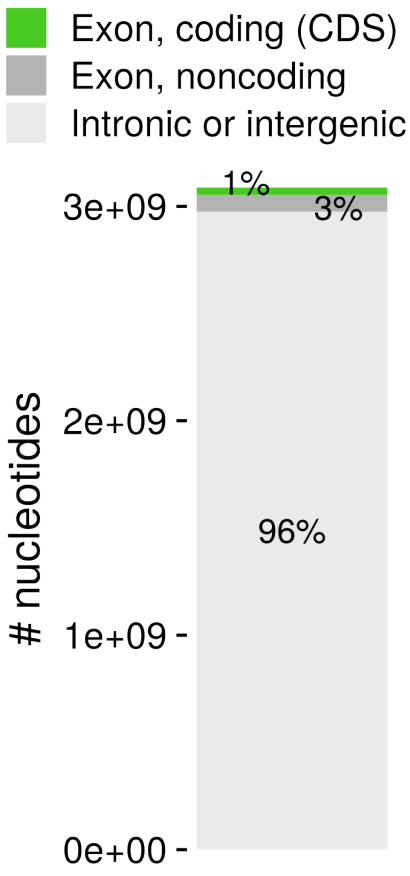
Goodwin, McPherson, and McCombie *Nature Reviews. Genetics*, 2016



(GENCODE v21)

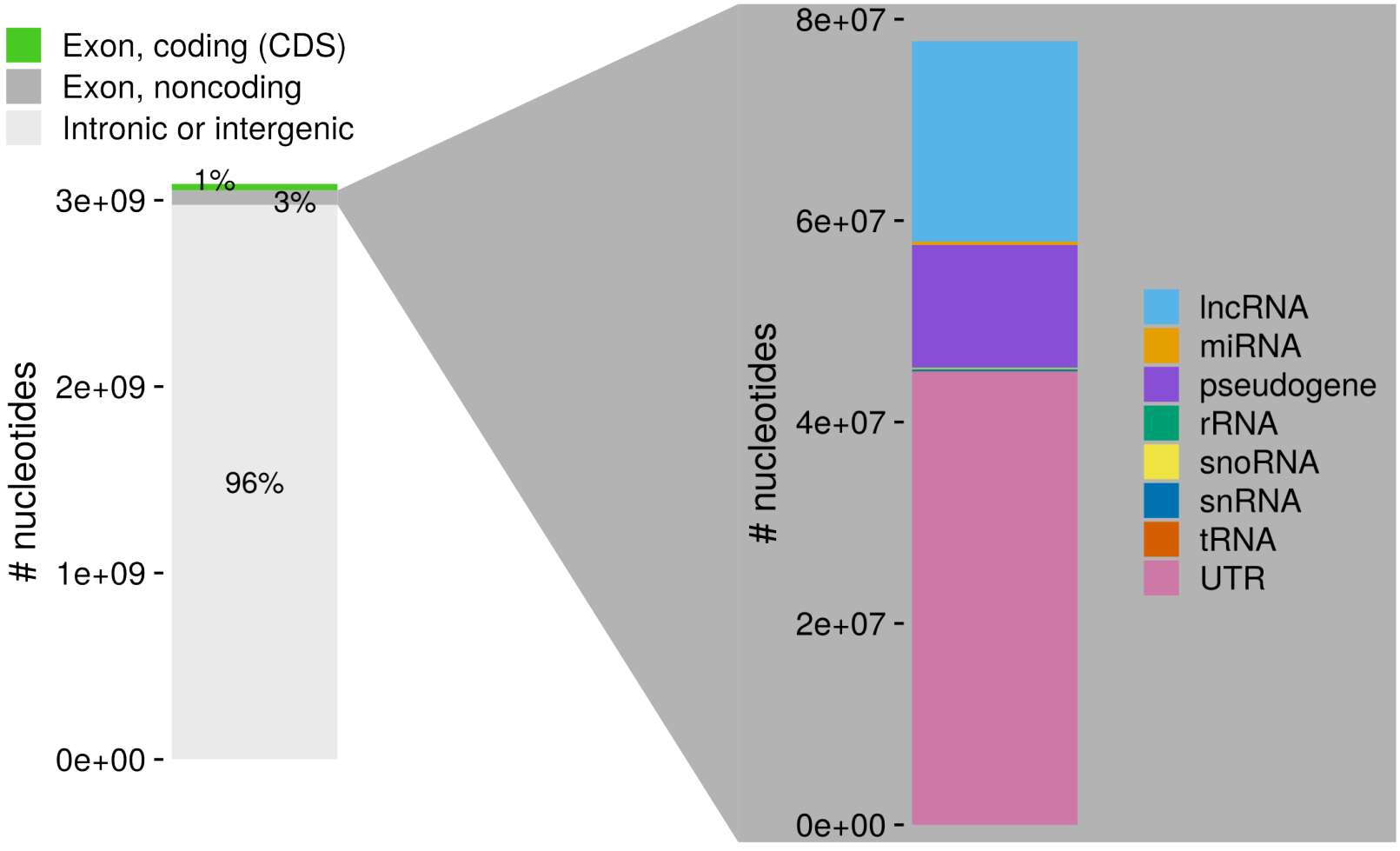
High-throughput, **end-to-end** transcript sequencing is now possible thanks to **TGS**

How complete is the human genome's annotation?



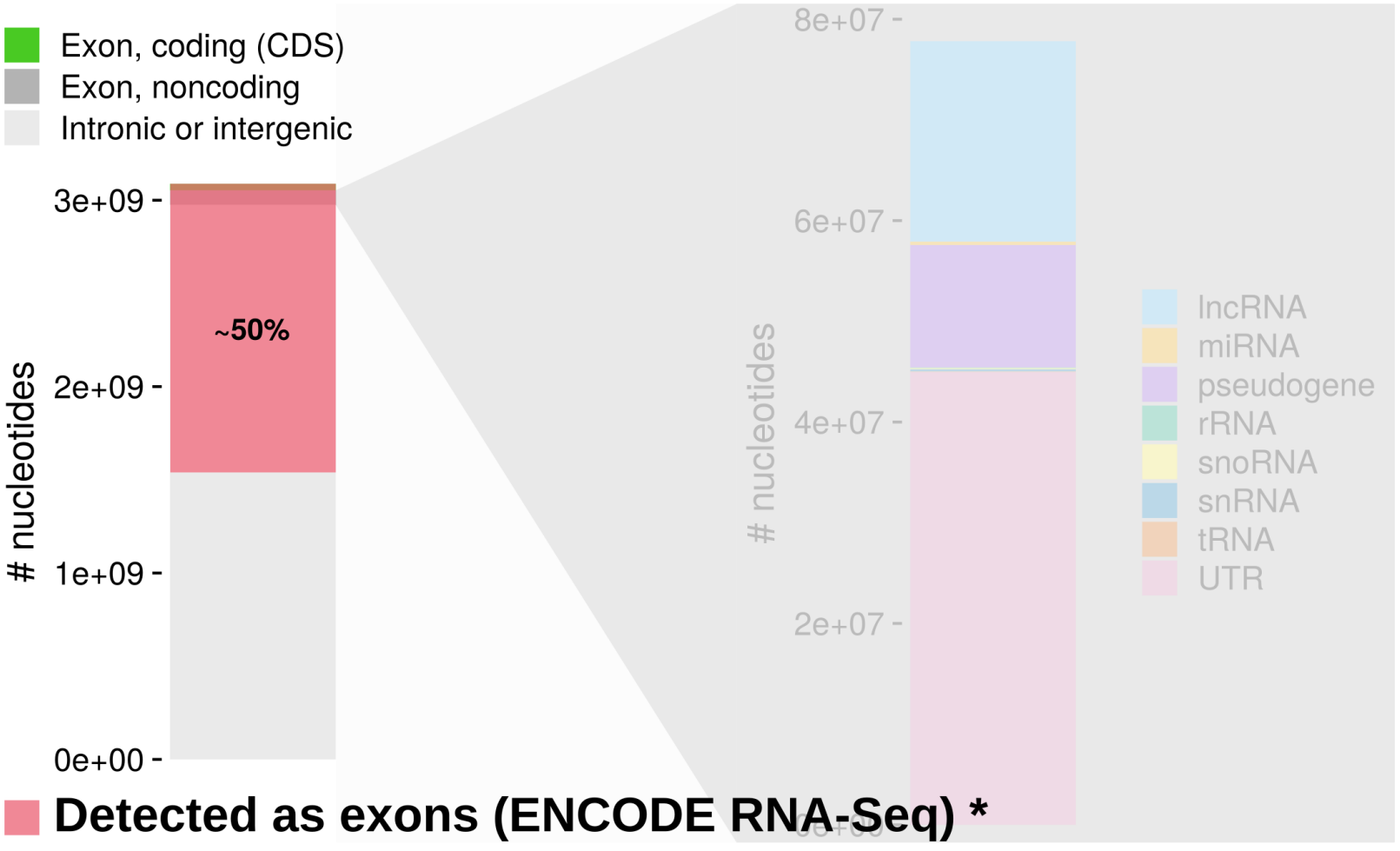
(GENCODE v21)

How complete is the human genome's annotation?



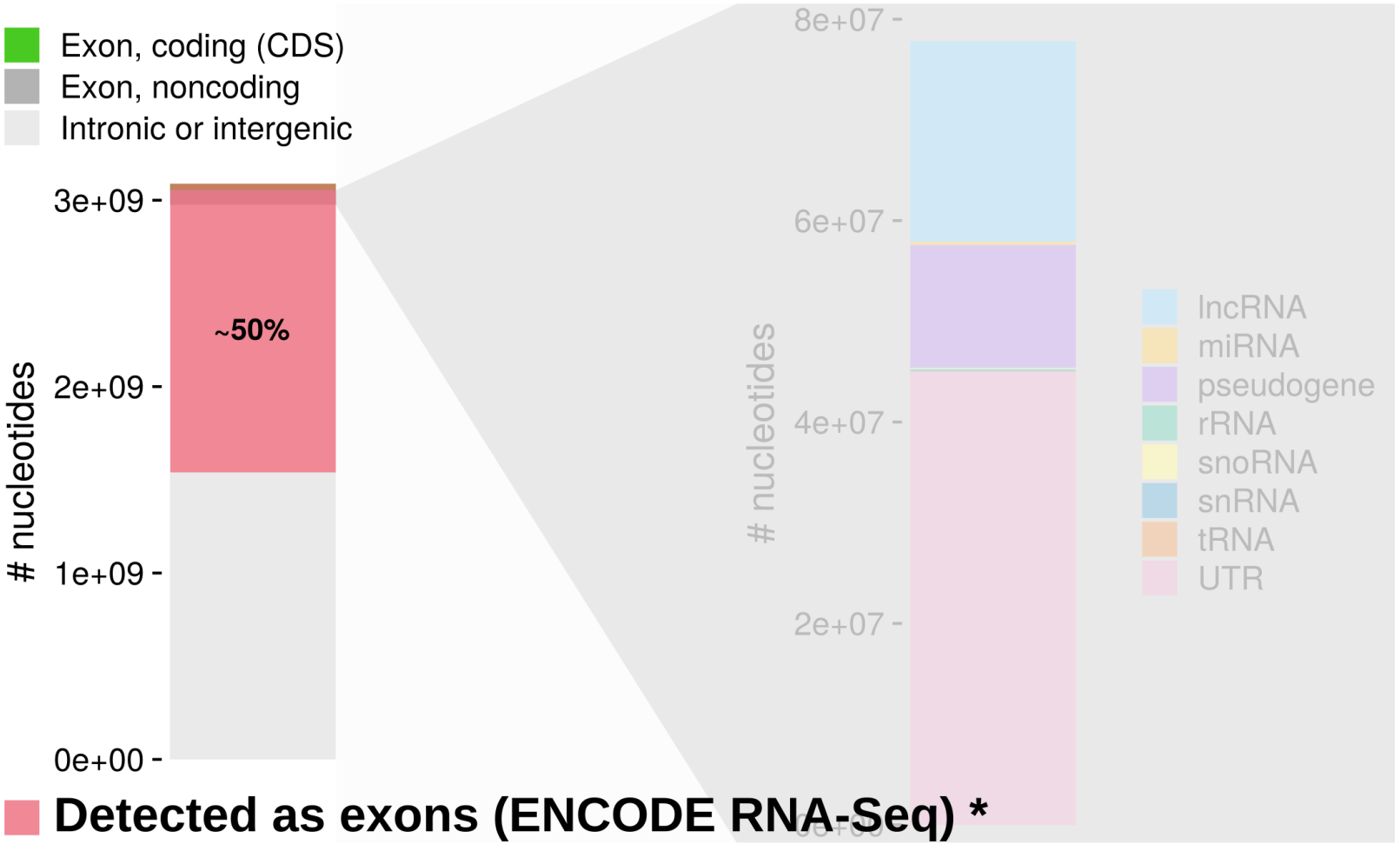
(GENCODE v21)

How complete is the human genome's annotation?



(GENCODE v21)
* Djebali *et al.*
Nature, 2012

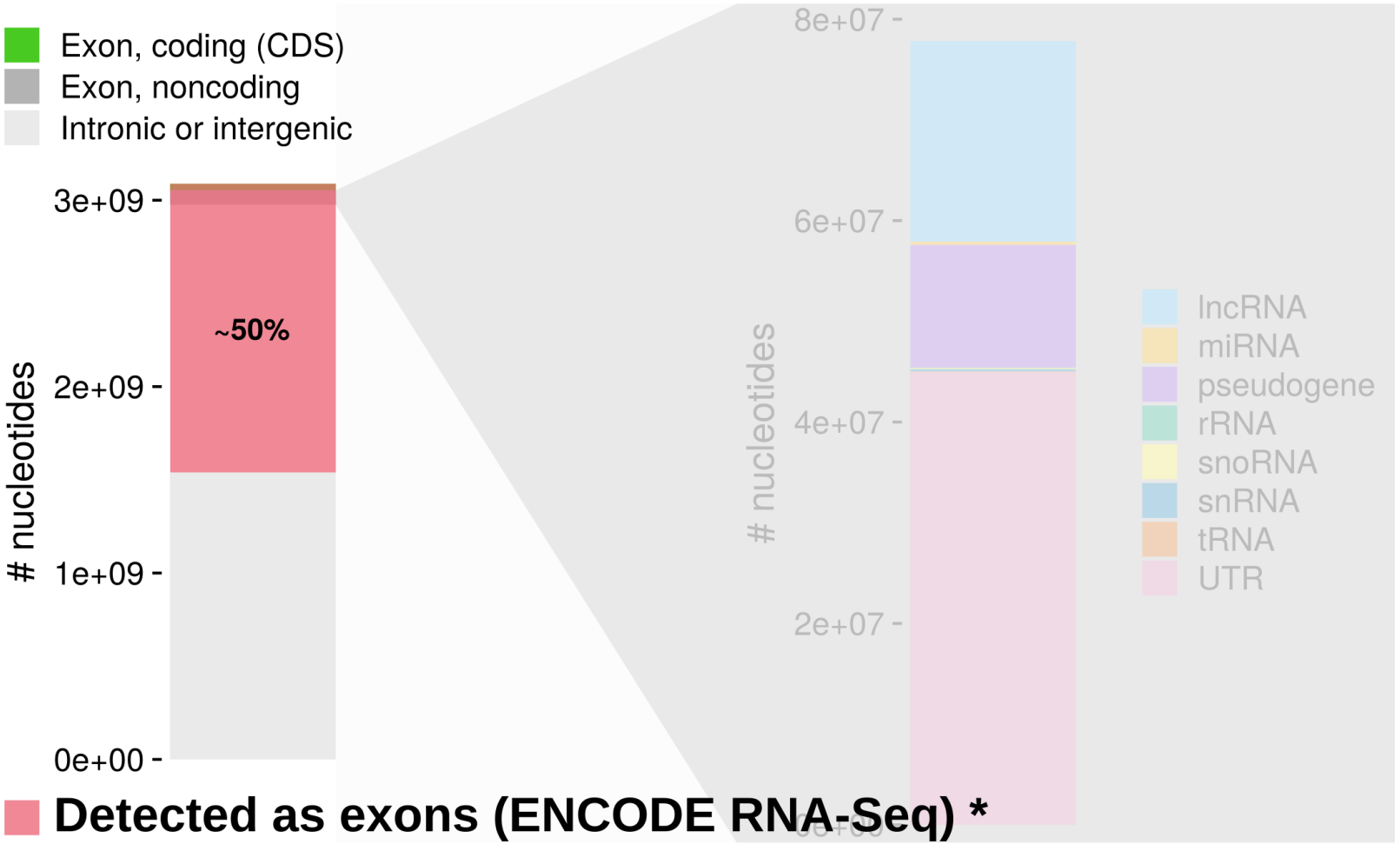
How complete is the human genome's annotation?



(GENCODE v21)
* Djebali *et al.* *Nature*, 2012

- The human genome is overwhelmingly **noncoding**

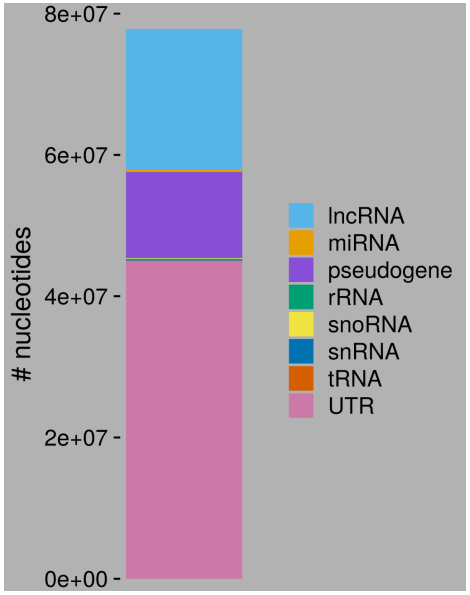
How complete is the human genome's annotation?



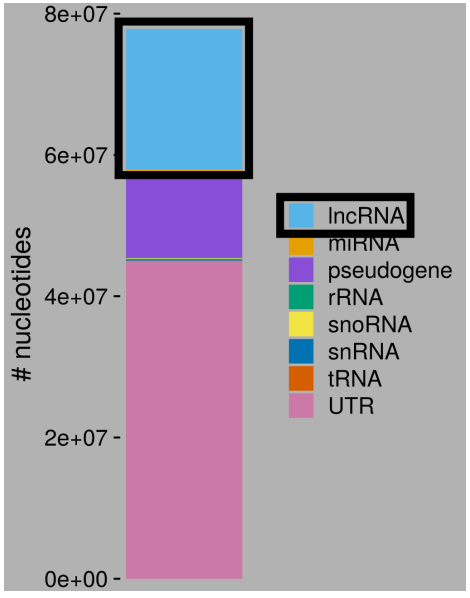
(GENCODE v21)
* Djebali *et al.* *Nature*, 2012

- The human genome is overwhelmingly **noncoding**
- Only a small fraction of the **noncoding "exome"** is annotated as such in **reference gene catalogs**

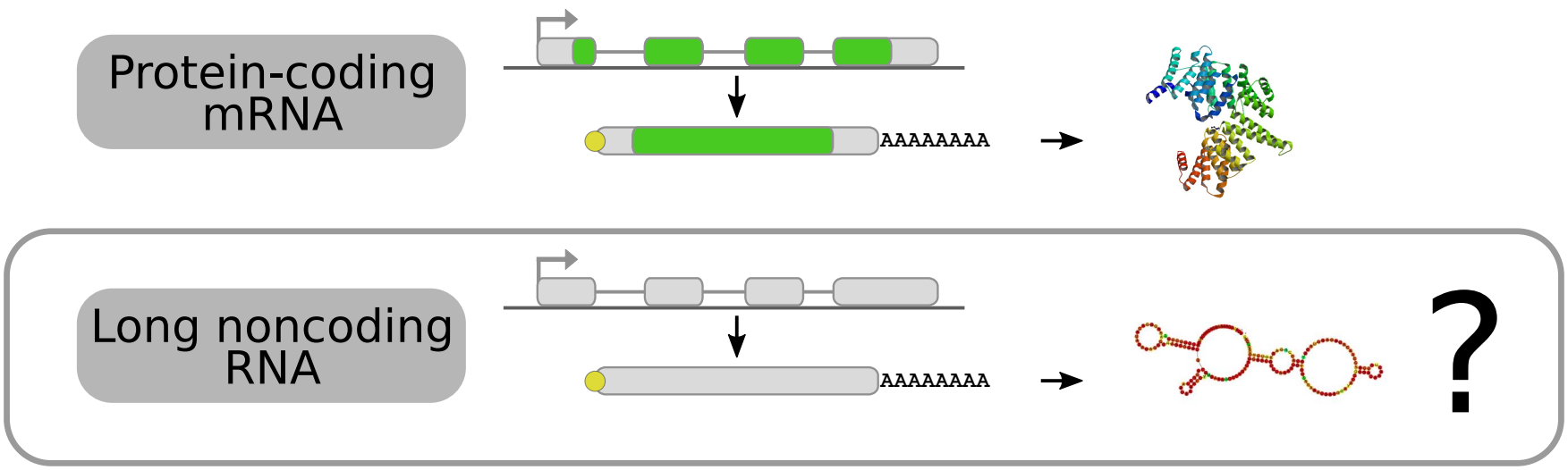
Long (intervening) noncoding RNAs



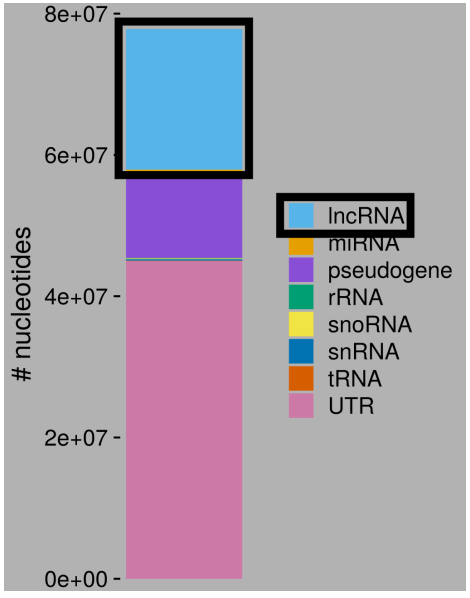
Long (intervening) noncoding RNAs



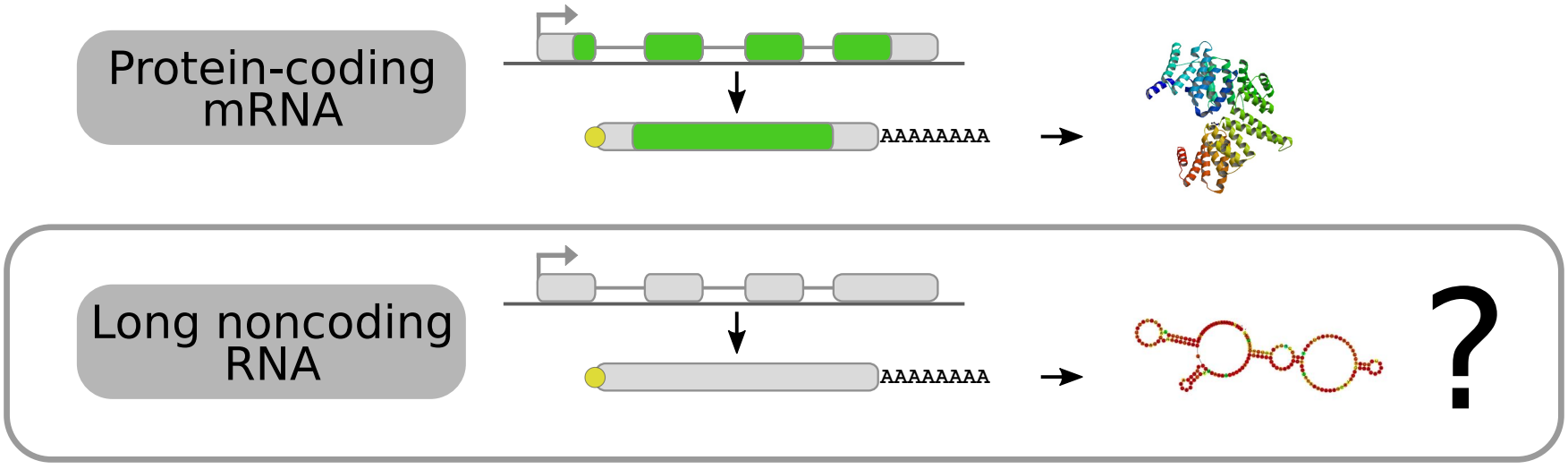
- **Long** (>200nts) and **non-coding**



Long (intervening) noncoding RNAs



- **Long** (>200nts) and **non-coding**
- Most are **poorly conserved** at the sequence level



LncRNAs perform various functions in the cell

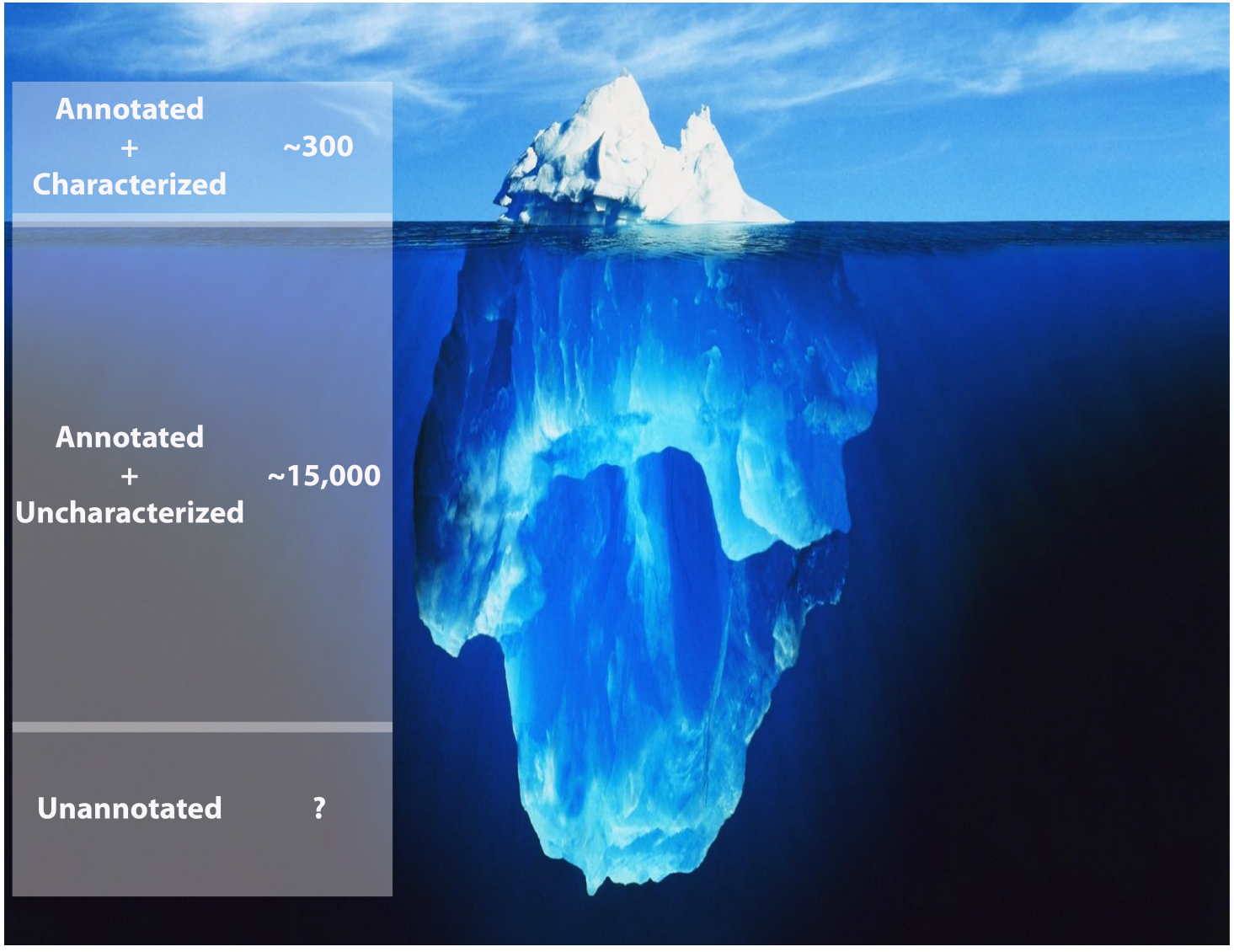
Some examples of functionally characterized lncRNAs:

- ***H19***: control of cell growth and proliferation - Brannan *et al. Molecular and cellular biology*, 1990
- ***Xist***: chromosome X inactivation - Brown *et al. Nature*, 1991
- ***Air***: *Ig2fr* gene silencing in *cis* - Sleutels, Zwart, and Barlow *Nature*, 2002
- ***MALAT1***: nuclear organization - Ji *et al. Oncogene*, 2003
- ***HOTAIR***: chromatin scaffolding - Rinn *et al. Cell*, 2007
- ***lincRNA-p21***: global gene repression - Huarte *et al. Cell*, 2010
- ***Upperhand***: *Hand2* gene regulation (heart development) - Anderson *et al. Nature*, 2016
- ***...***

(Low sequence conservation)

Functional characterization of lincRNAs

A Genetic *Terra Incognita*



(Human)
(lincRNAdb – GENCODE v21)

LincRNAs and mRNAs show distinct genomic properties

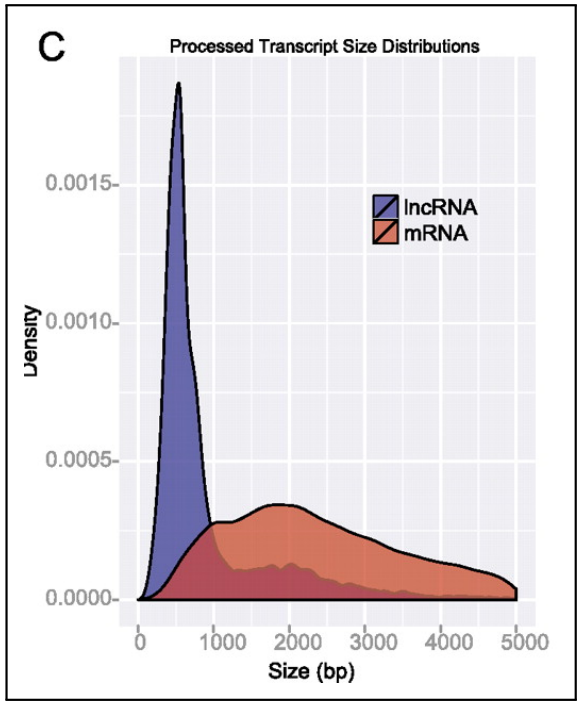
(GENCODE v7, Derrien *et al. Genome Research*, 2012)

LncRNA 5' ends are depleted in markers of transcription initiation (**15% less CAGE coverage**)

LincRNAs and mRNAs show distinct genomic properties

(GENCODE v7, Derrien *et al. Genome Research*, 2012)

LncRNA 5' ends are depleted in markers of transcription initiation (**15% less CAGE coverage**)

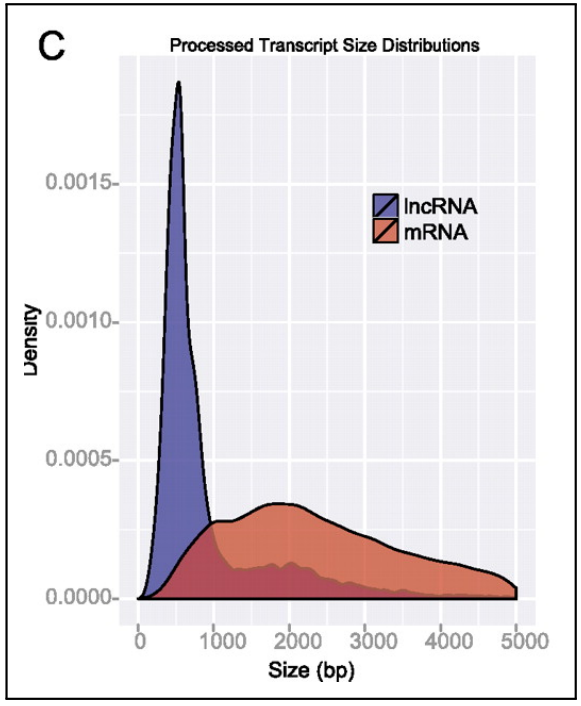


LncRNAs are shorter than mRNAs

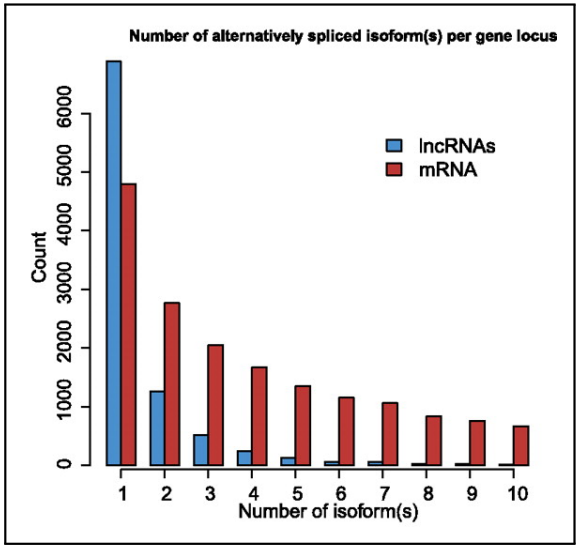
LincRNAs and mRNAs show distinct genomic properties

(GENCODE v7, Derrien *et al. Genome Research*, 2012)

LncRNA 5' ends are depleted in markers of transcription initiation (**15% less CAGE coverage**)



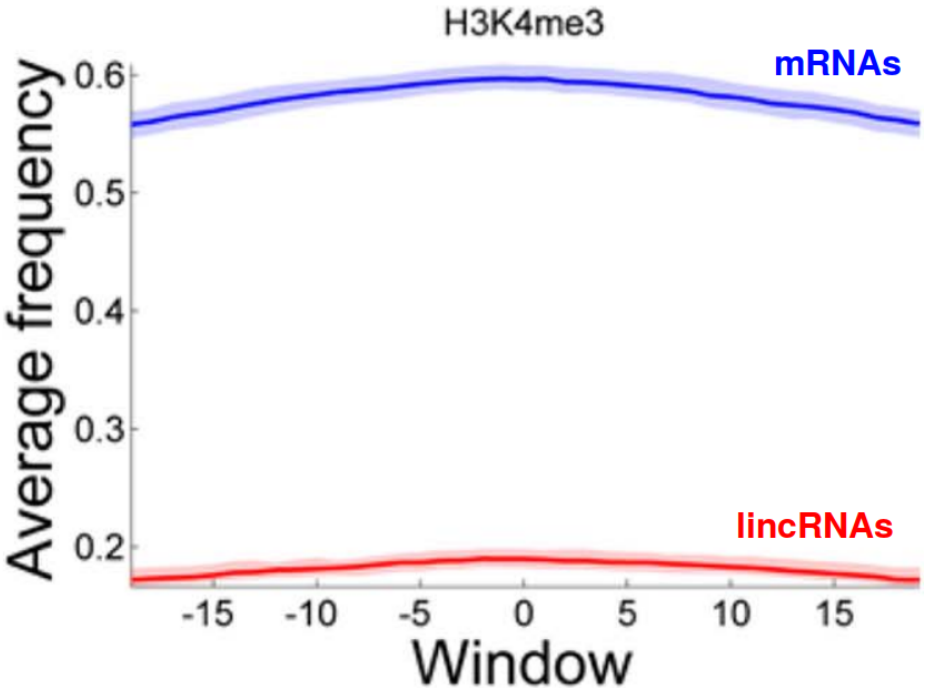
LncRNAs are shorter than mRNAs



LncRNAs are less alternatively spliced than mRNAs

LincRNA promoter environment

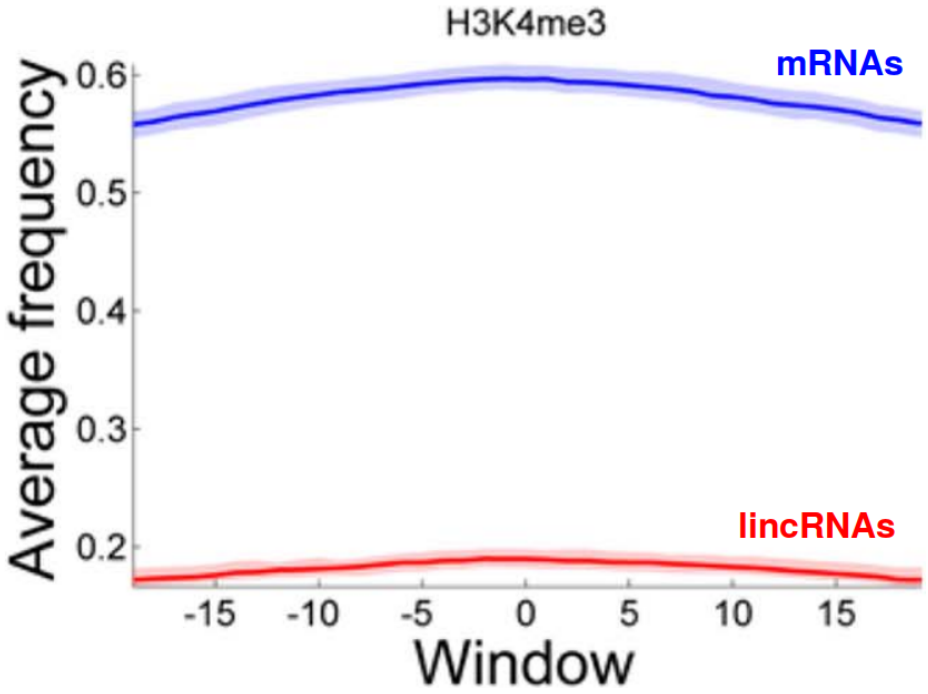
"Promoter Analysis Reveals **Globally Differential Regulation** of Human Long Non-Coding RNA and Protein-Coding Genes"
(Alam *et al. PLoS ONE*, 2014)



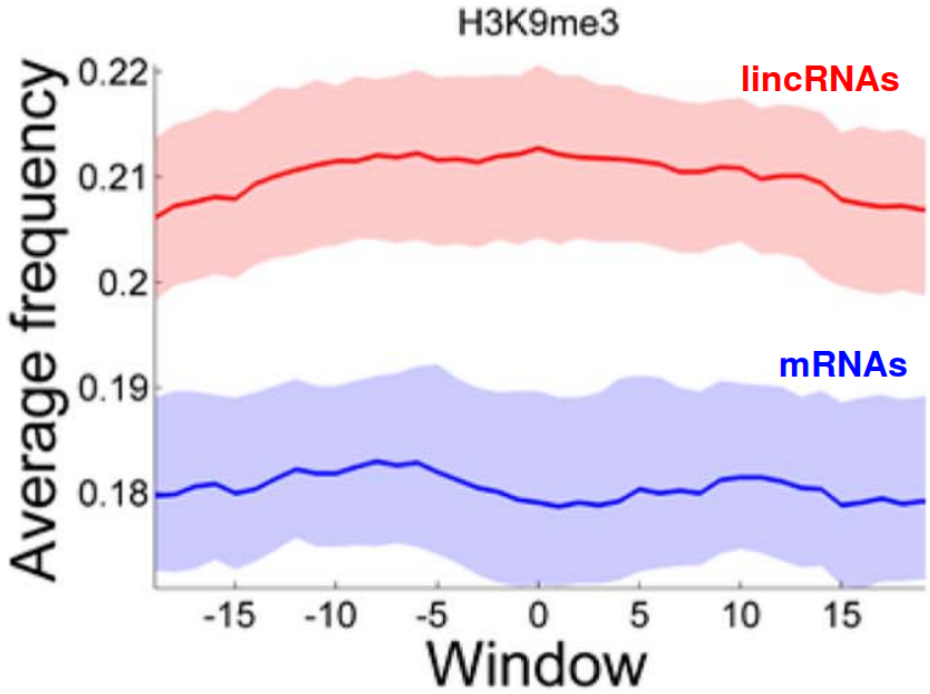
LincRNA promoters are **depleted in active chromatin marks**

LincRNA promoter environment

"Promoter Analysis Reveals **Globally Differential Regulation** of Human Long Non-Coding RNA and Protein-Coding Genes"
(Alam *et al. PLoS ONE*, 2014)



LincRNA promoters are **depleted in active chromatin marks**



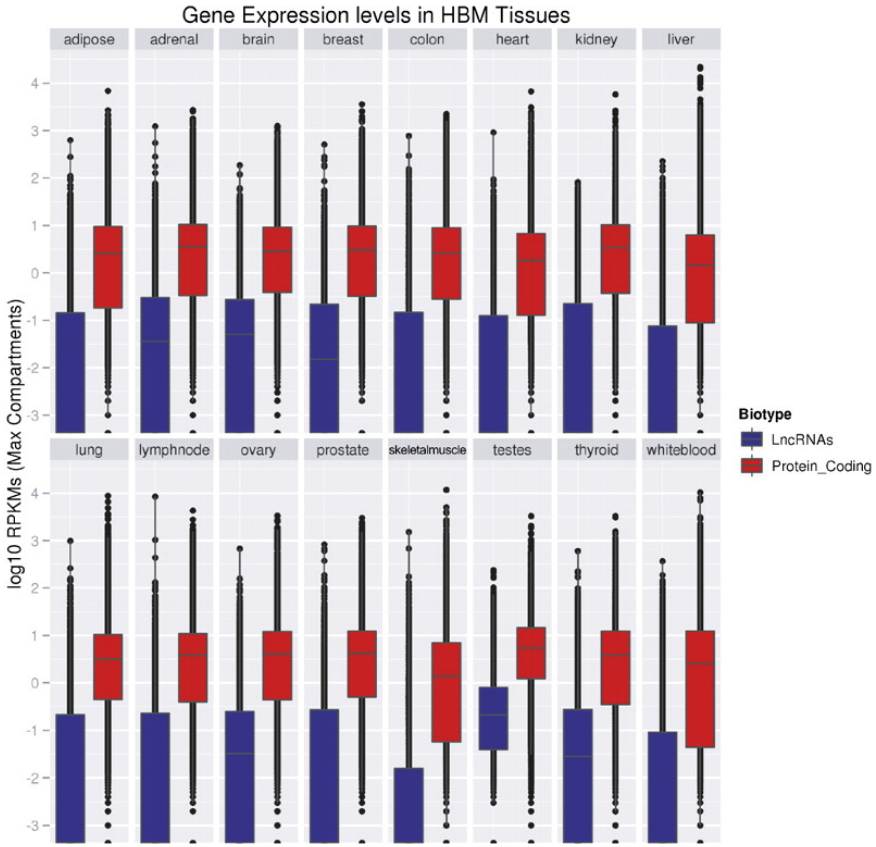
LincRNA promoters are **enriched in repressive chromatin marks**

... or are they, really?

... or are they, really?

LincRNAs have low expression levels

(GENCODE v7, Derrien *et al. Genome Research*, 2012)



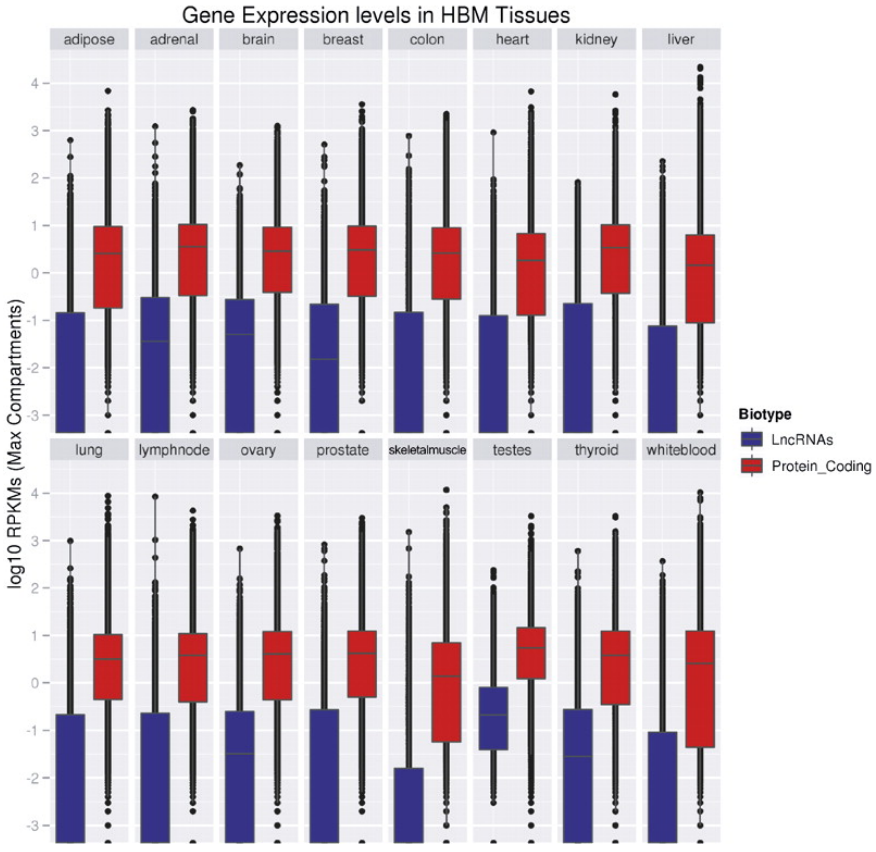
In a typical (polyA+) RNA-Seq experiment, lincRNAs account for **<3% of the reads**

(Bakel *et al. PLoS biology*, 2010)

... or are they, really?

LincRNAs have low expression levels

(GENCODE v7, Derrien *et al. Genome Research*, 2012)



Low expression

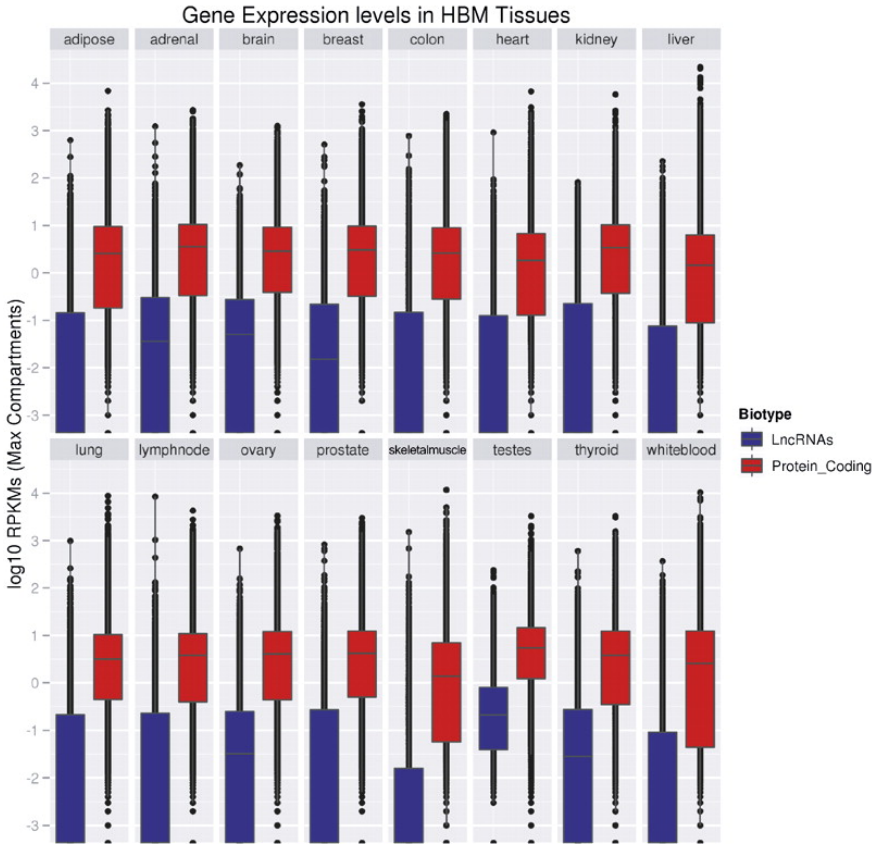
In a typical (polyA+) RNA-Seq experiment, lincRNAs account for **<3% of the reads**

(Bakel *et al. PLoS biology*, 2010)

... or are they, really?

LincRNAs have low expression levels

(GENCODE v7, Derrien *et al. Genome Research*, 2012)



Low expression

⇓

Under-representation of intact lncRNA transcripts in cDNA libraries

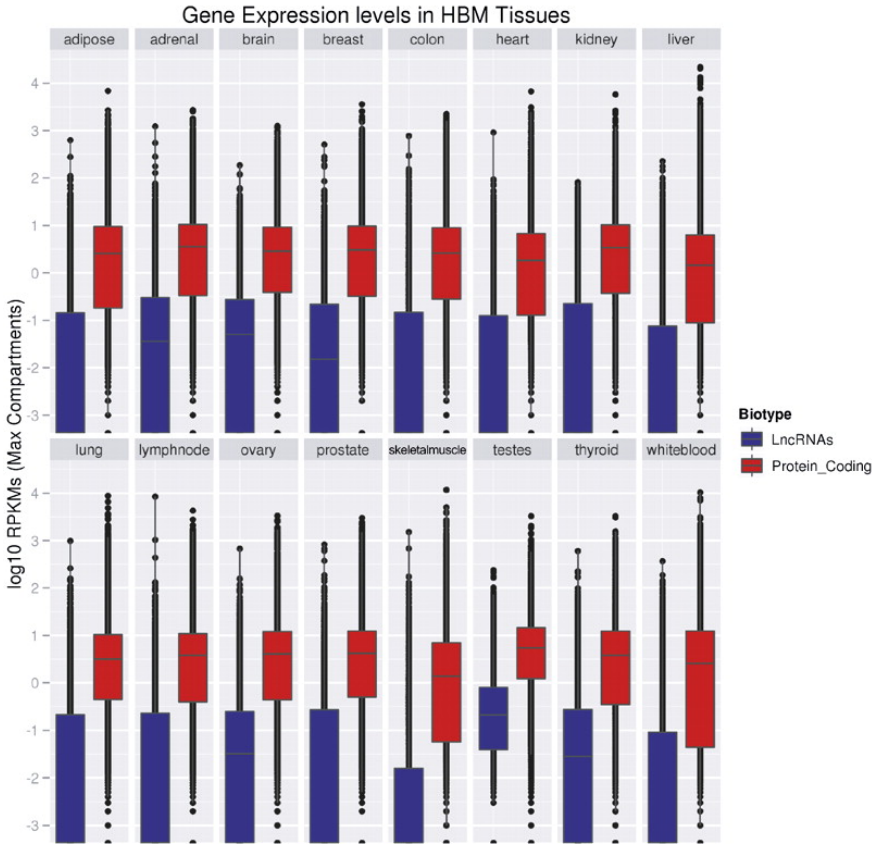
In a typical (polyA+) RNA-Seq experiment, lincRNAs account for **<3% of the reads**

(Bakel *et al. PLoS biology*, 2010)

... or are they, really?

LincRNAs have low expression levels

(GENCODE v7, Derrien *et al. Genome Research*, 2012)



Low expression

⇓

Under-representation of intact lncRNA transcripts in cDNA libraries

⇓

Annotation artifacts

(truncated lncRNA transcript models)

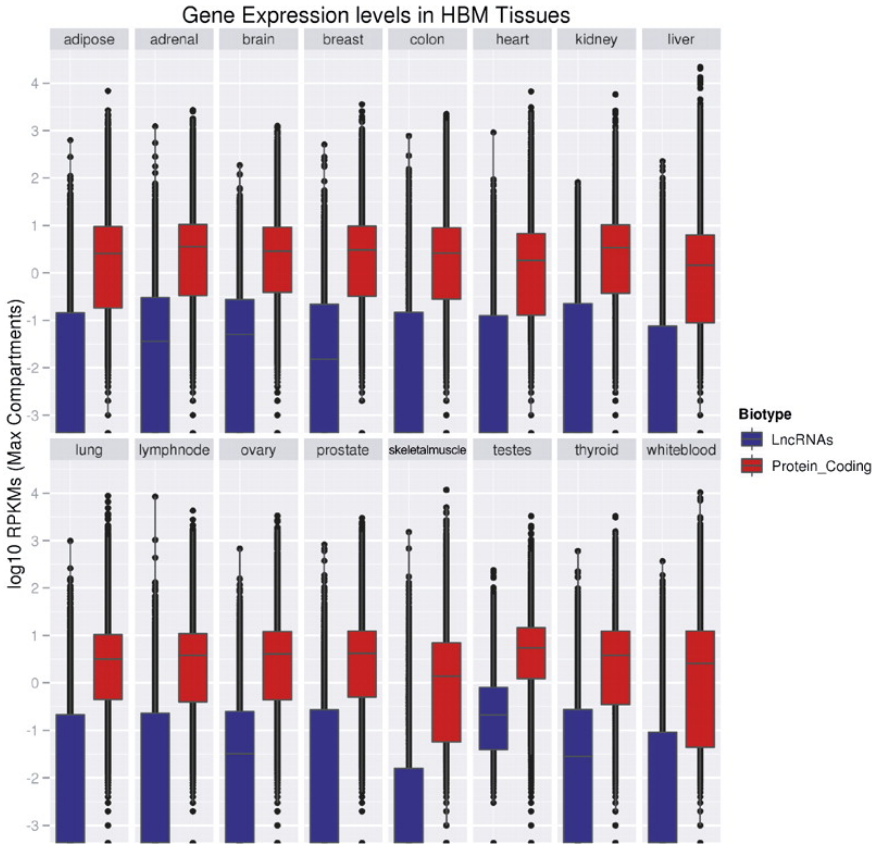
In a typical (polyA+) RNA-Seq experiment, lincRNAs account for **<3% of the reads**

(Bakel *et al. PLoS biology*, 2010)

... or are they, really?

LincRNAs have low expression levels

(GENCODE v7, Derrien *et al. Genome Research*, 2012)



In a typical (polyA+) RNA-Seq experiment, lincRNAs account for **<3% of the reads**

(Bakel *et al. PLoS biology*, 2010)

Low expression

⇓

Under-representation of intact lncRNA transcripts in cDNA libraries

⇓

Annotation artifacts

(truncated lncRNA transcript models)

⇒ **Need for targeted sequencing techniques**

Other lncRNA catalogs

Catalog	# lincRNA loci	Short-read assembly?	Ref.
NONCODE	96,000	**Yes (partially)**	*Fang et al.*, 2017
MiTranscriptome	63,000	**Yes**	Iyer *et al.*, 2015
FANTOM CAT	28,000	**Yes**	Hon *et al.*, 2017
RefSeq	15,000	**Yes (partially)**	O'Leary *et al.*, 2016
GENCODE	15,000	**No**	Frankish *et al.*, 2019
BigTranscriptome	14,000	**Yes**	You *et al.*, 2017

Other lncRNA catalogs

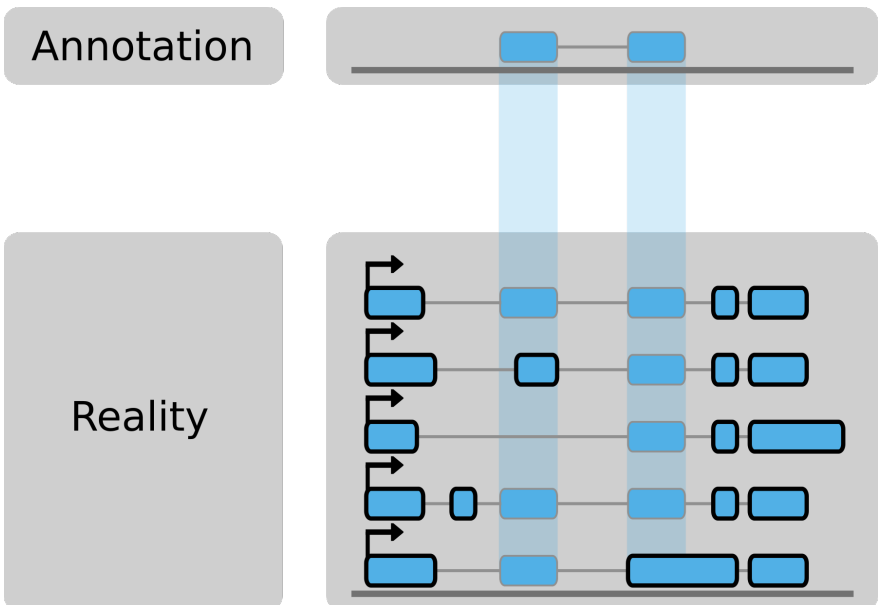
Catalog	# lincRNA loci	Short-read assembly?	Ref.
NONCODE	96,000	**Yes (partially)**	*Fang et al.*, 2017
MiTranscriptome	63,000	**Yes**	Iyer *et al.*, 2015
FANTOM CAT	28,000	**Yes**	Hon *et al.*, 2017
RefSeq	15,000	**Yes (partially)**	O'Leary *et al.*, 2016
GENCODE	15,000	**No**	Frankish *et al.*, 2019
BigTranscriptome	14,000	**Yes**	You *et al.*, 2017

Short-read-based assembly leads to **inaccurate transcript models**, especially:

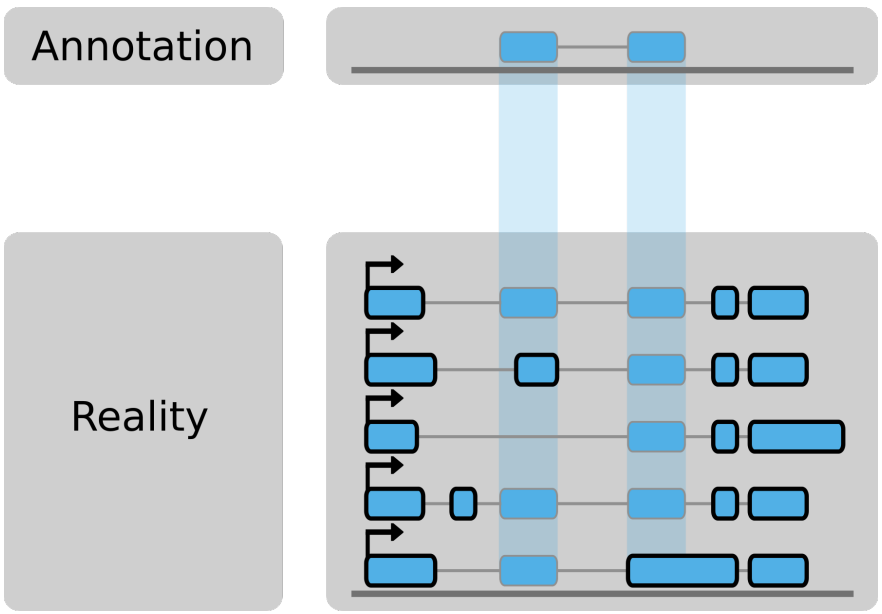
- Wrong 5'/3' boundaries
- Wrong lncRNA models

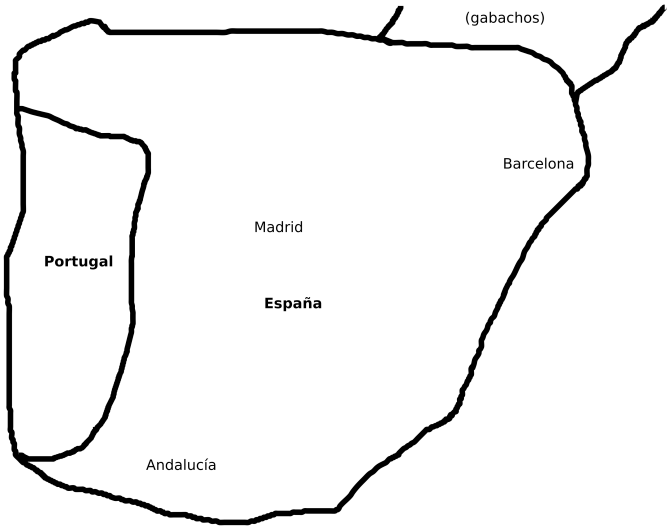
(Steijger *et al. Nature Methods*, 2013)

Accurate gene annotations: a foundation for lncRNA functional characterization

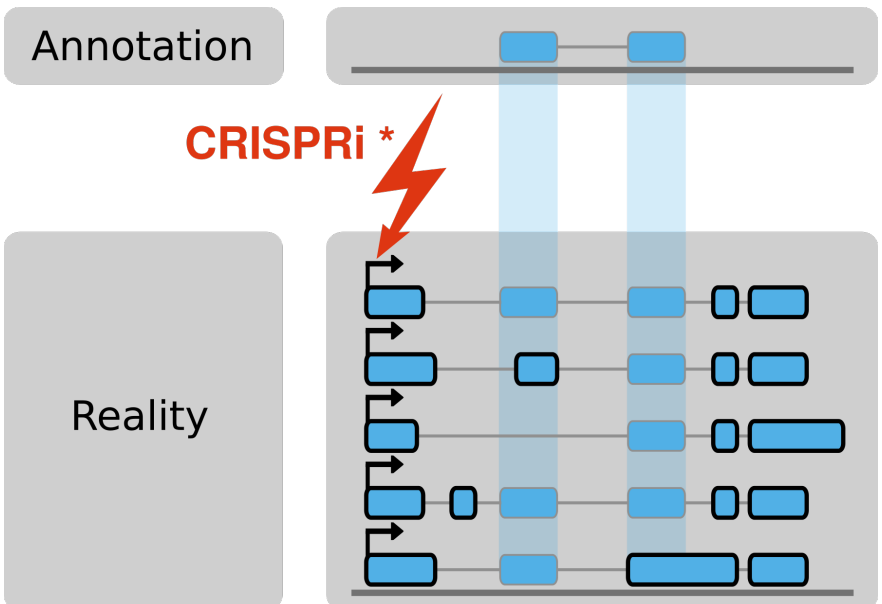


Accurate gene annotations: a foundation for lncRNA functional characterization

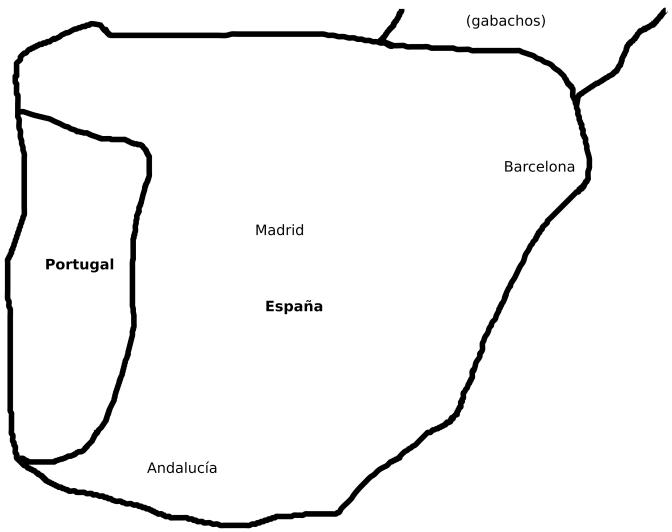




Accurate gene annotations: a foundation for lncRNA functional characterization



(* *e.g.* Liu *et al. Science*, 2017)



Recapitulating...

- The **function** of most lncRNAs is **unknown**

Recapitulating...

- The **function** of most lncRNAs is **unknown**
- **Accurate annotation** of lncRNAs is crucial to understand their **biological roles**

Recapitulating...

- The **function** of most lncRNAs is **unknown**
- **Accurate annotation** of lncRNAs is crucial to understand their **biological roles**
- Currently, reference lncRNA catalogs are **incomplete/inaccurate**

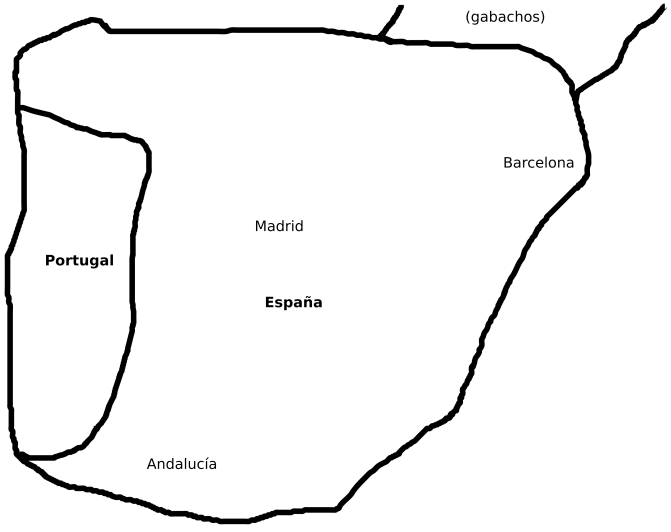
Recapitulating...

- The **function** of most lncRNAs is **unknown**
- **Accurate annotation** of lncRNAs is crucial to understand their **biological roles**
- Currently, reference lncRNA catalogs are **incomplete/inaccurate**
- LncRNAs are expressed at **low levels**

Recapitulating...

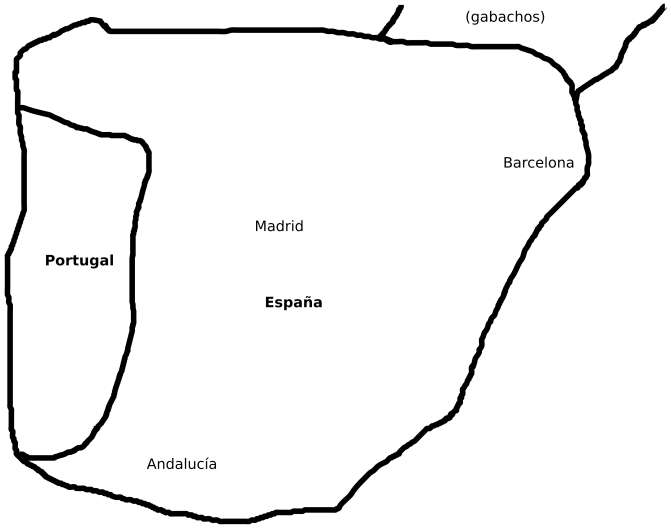
- The **function** of most lncRNAs is **unknown**
- **Accurate annotation** of lncRNAs is crucial to understand their **biological roles**
- Currently, reference lncRNA catalogs are **incomplete/inaccurate**
- LncRNAs are expressed at **low levels**
- Long-read (TGS) sequencing (**PacBio**) promises to revolutionize gene annotation:
  - Full-length, high-quality transcript sequencing
  - High-throughput
  - Lower depth than short-read SGS (Illumina)

Objectives



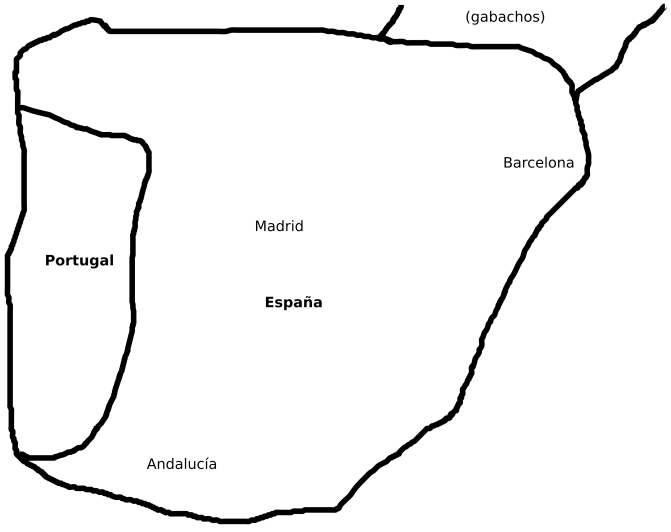
- Towards a more **complete, high-quality lncRNA map** in the human genome:

Objectives



- Towards a more **complete, high-quality lncRNA map** in the human genome:
  - Using **targeted** long-read sequencing (**PacBio**)
  - In a **high-throughput** manner, with minimal manual intervention

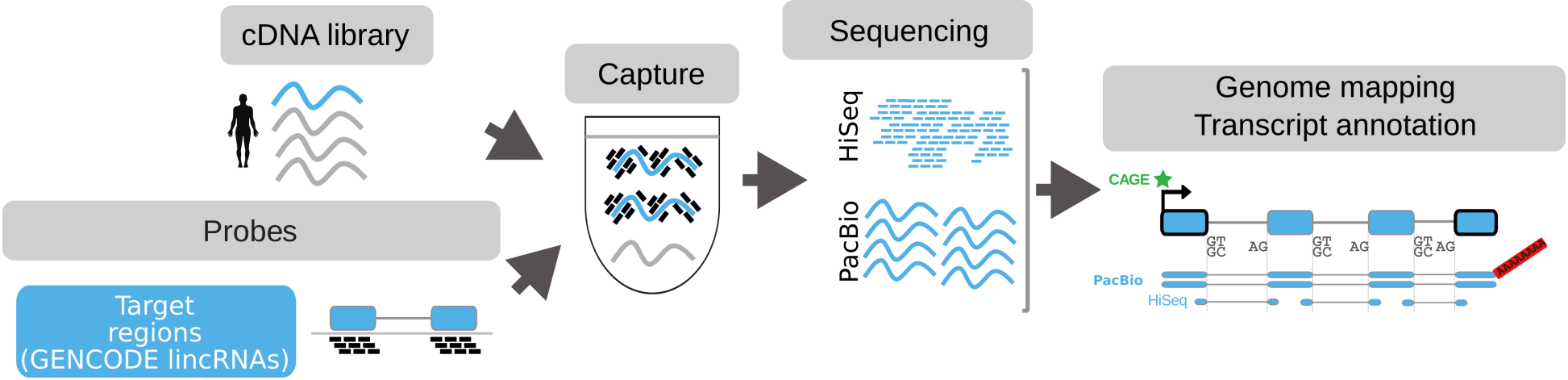
Objectives



- Towards a more **complete, high-quality lncRNA map** in the human genome:
  - Using **targeted** long-read sequencing (**PacBio**)
  - In a **high-throughput** manner, with minimal manual intervention
- Using this enhanced map, re-evaluate some of the **genomic properties of lncRNAs**

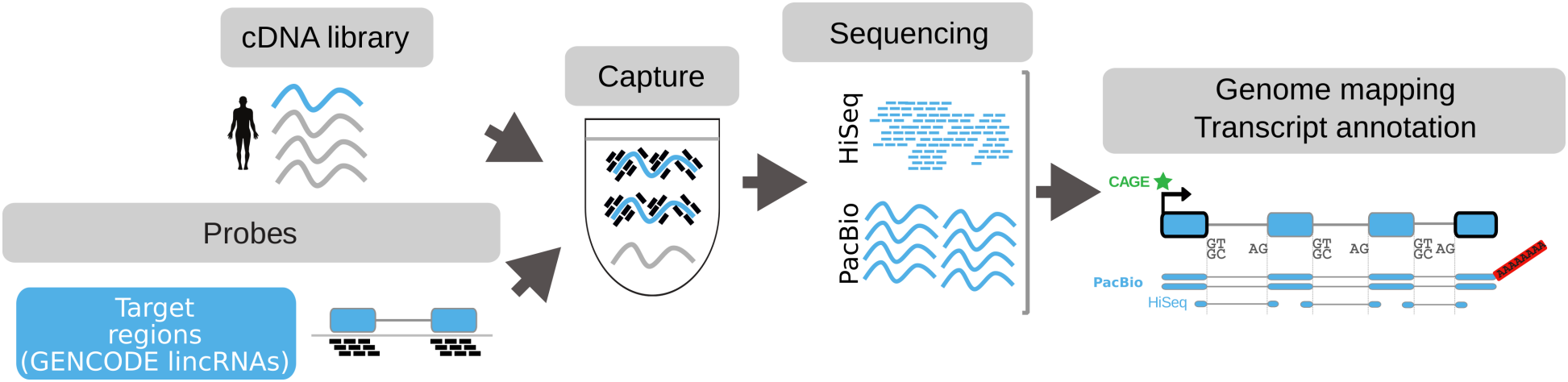
The CLS method: Capture Long-read Sequencing

High-throughput lncRNA annotation



The CLS method: Capture Long-read Sequencing

High-throughput lncRNA annotation



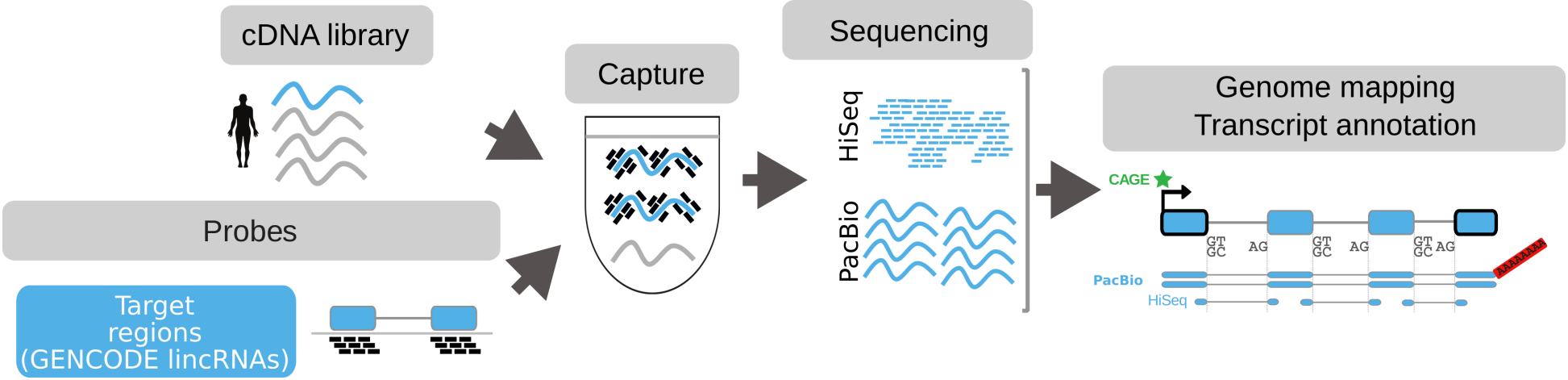
CLS targets:

- ~6,000 human lincRNA loci
- ~10 Mb (~50% of total annotated)

Lagarde *et al. Nature Genetics*, 2017

The CLS method: Capture Long-read Sequencing

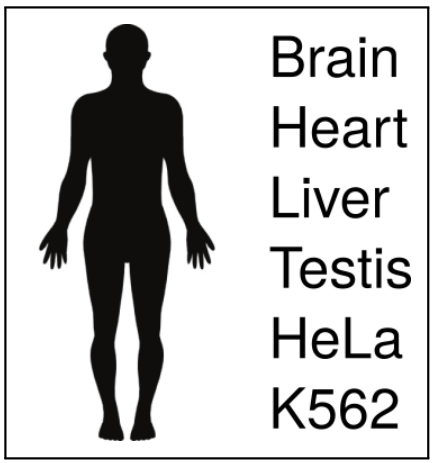
High-throughput lncRNA annotation



CLS targets:

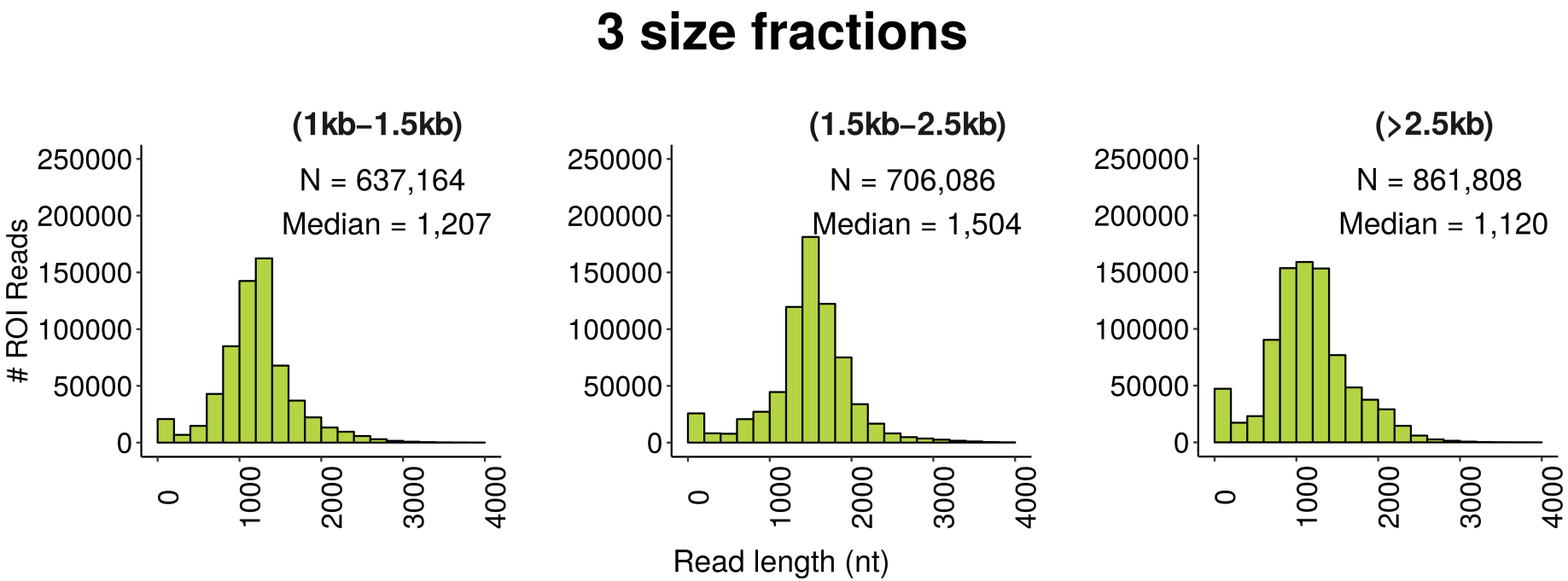
- ~6,000 human lincRNA loci
- ~10 Mb (~50% of total annotated)

Lagarde *et al. Nature Genetics*, 2017



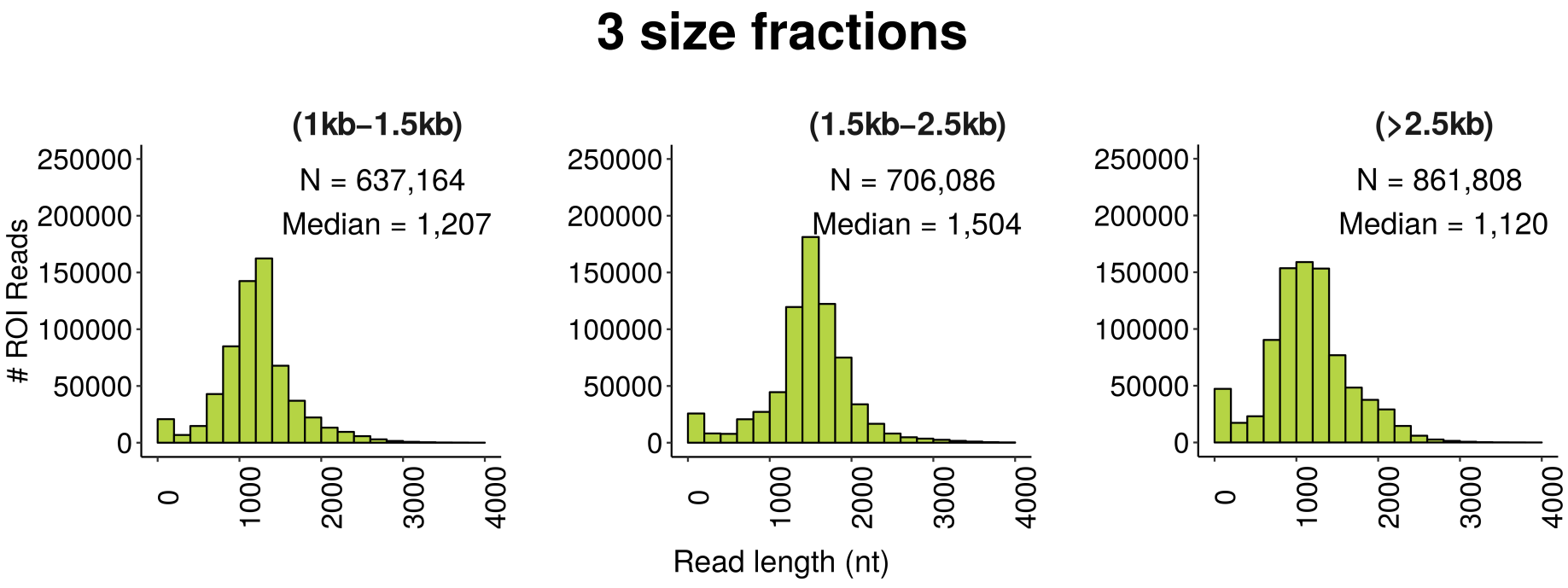
CLS sequencing output

PacBio RSII:



CLS sequencing output

PacBio RSII:



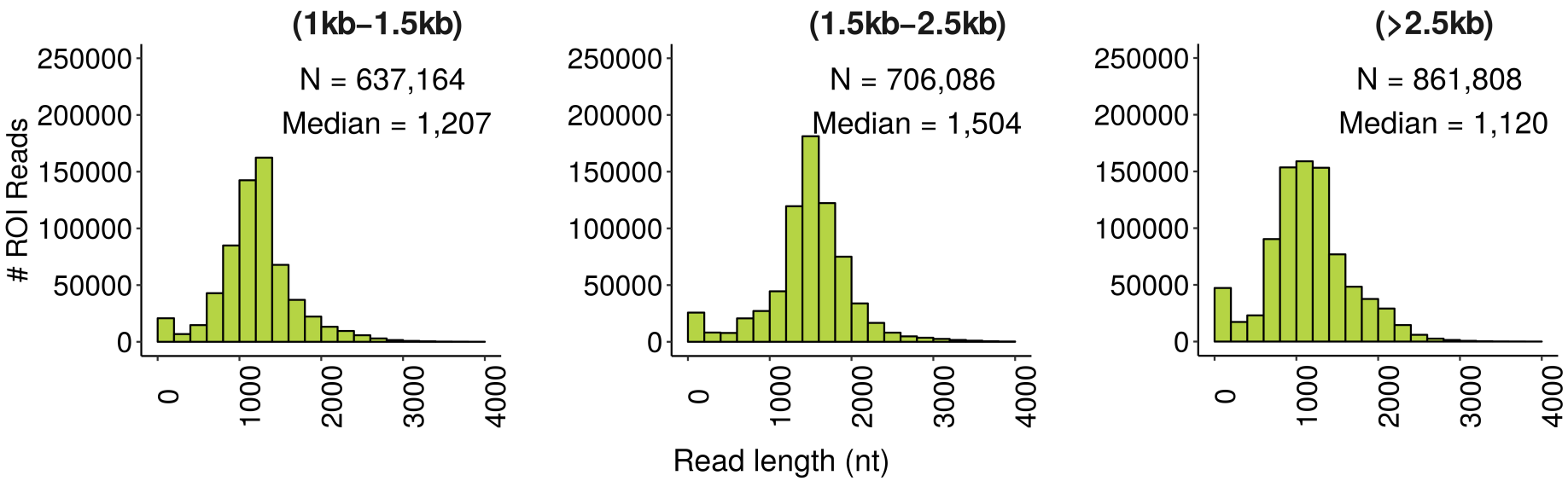
Total:

- ~2.2 million PacBio reads
- ~1.3kb median read length

CLS sequencing output

PacBio RSII:

3 size fractions



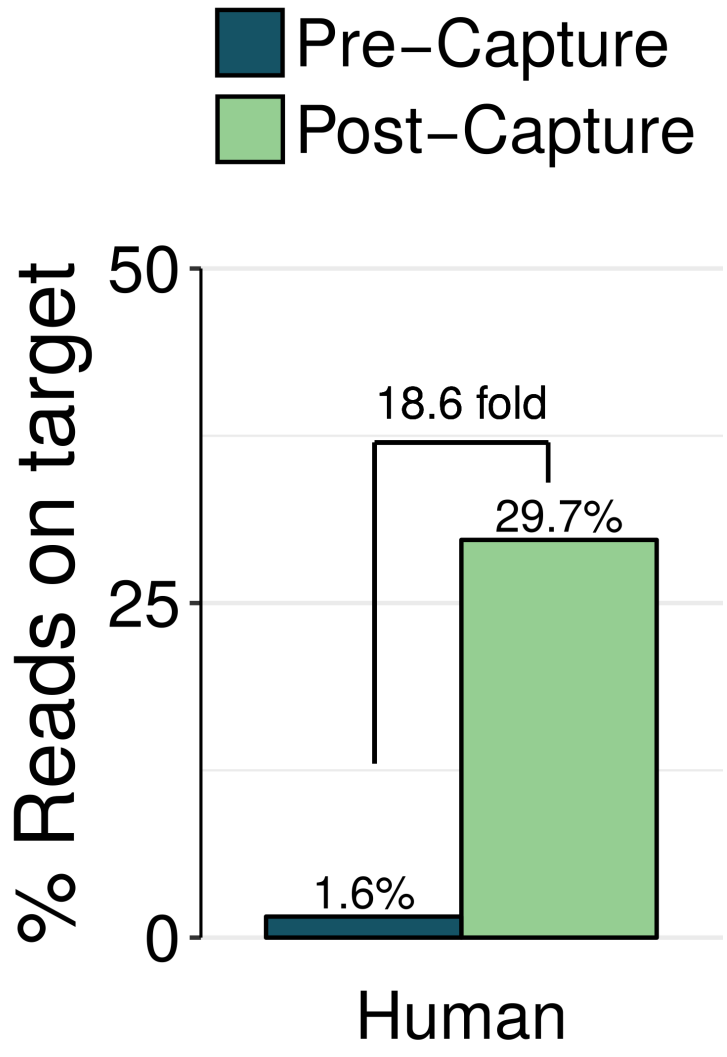
Total:

- ~2.2 million PacBio reads
- ~1.3kb median read length

+ **HiSeq** 2500: 212 million PE stranded reads

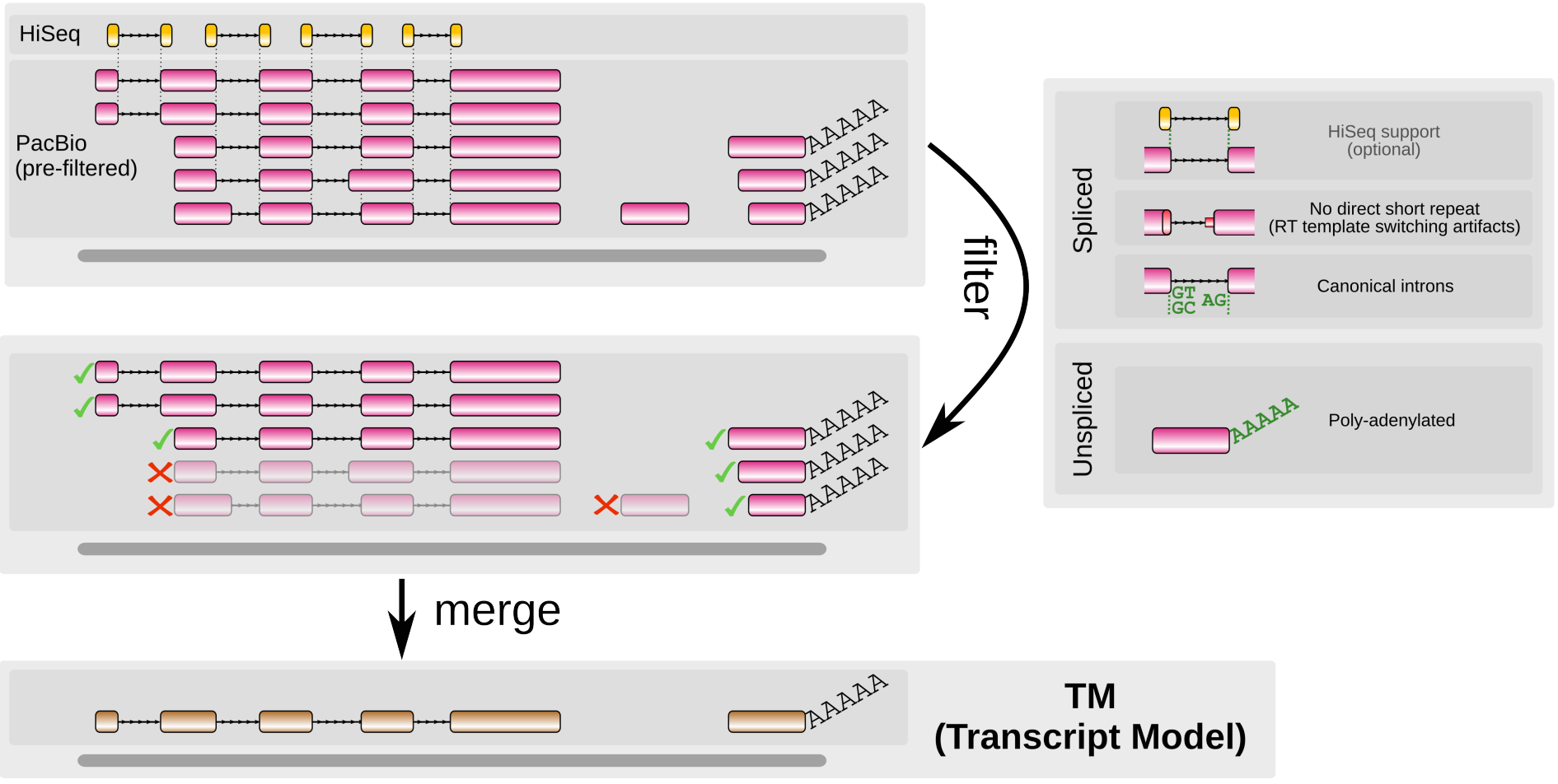
Lagarde *et al. Nature Genetics*, 2017

CLS libraries are highly enriched for targets

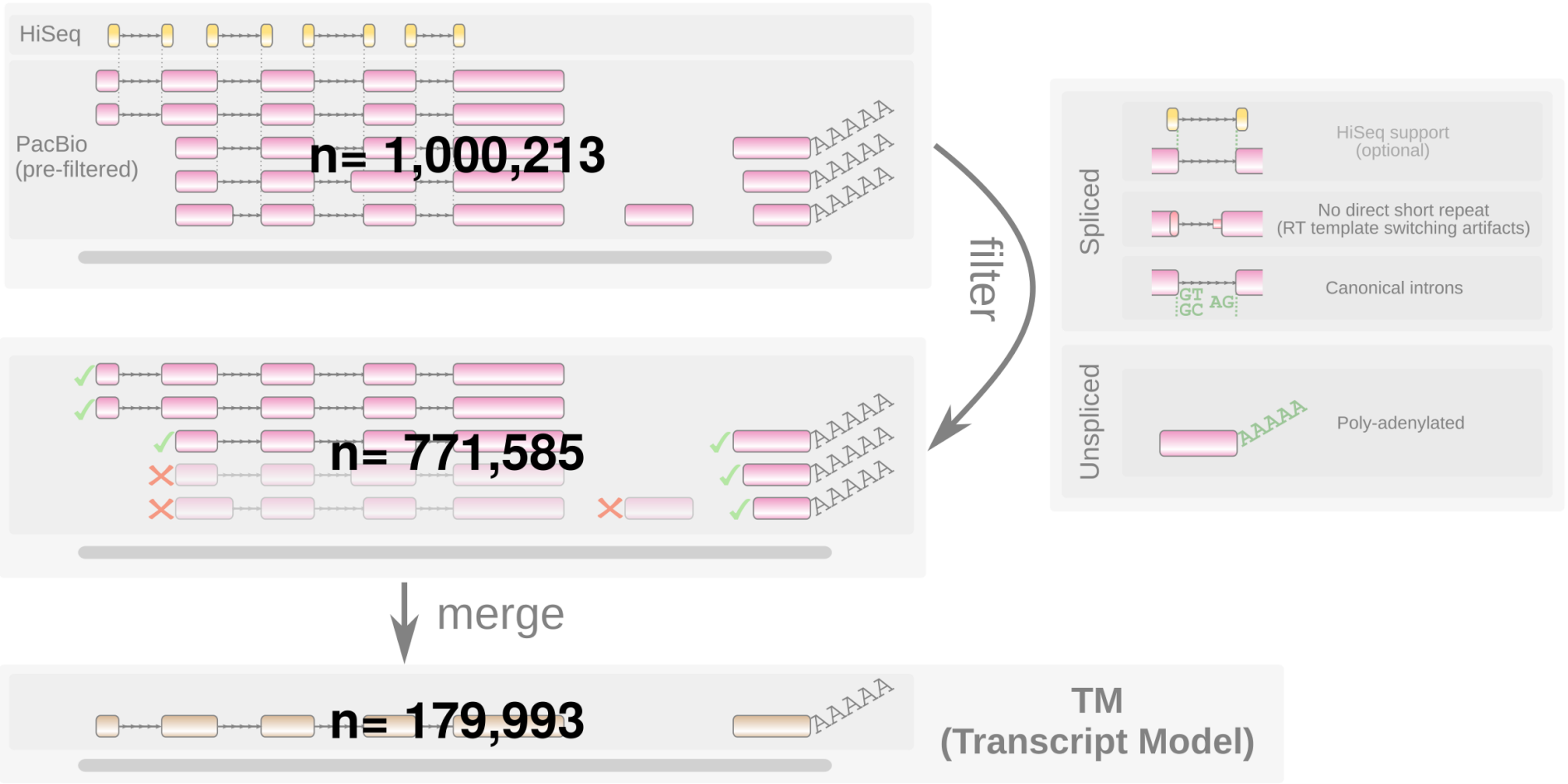


Lagarde *et al. Nature Genetics*, 2017

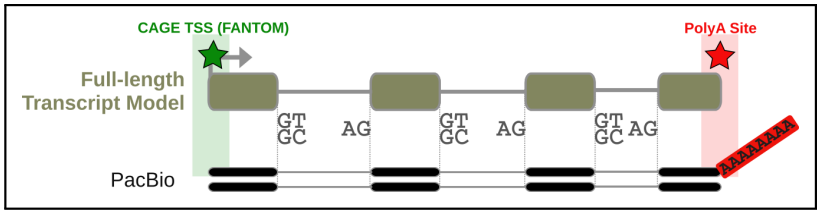
Read quality filtering and merging



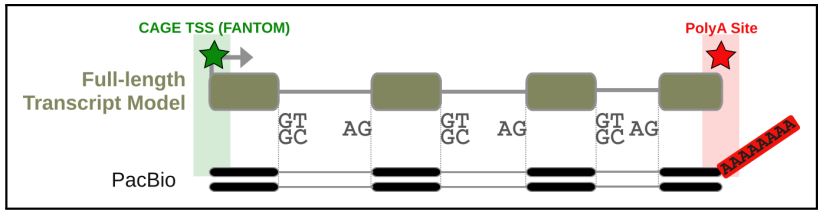
Read quality filtering and merging

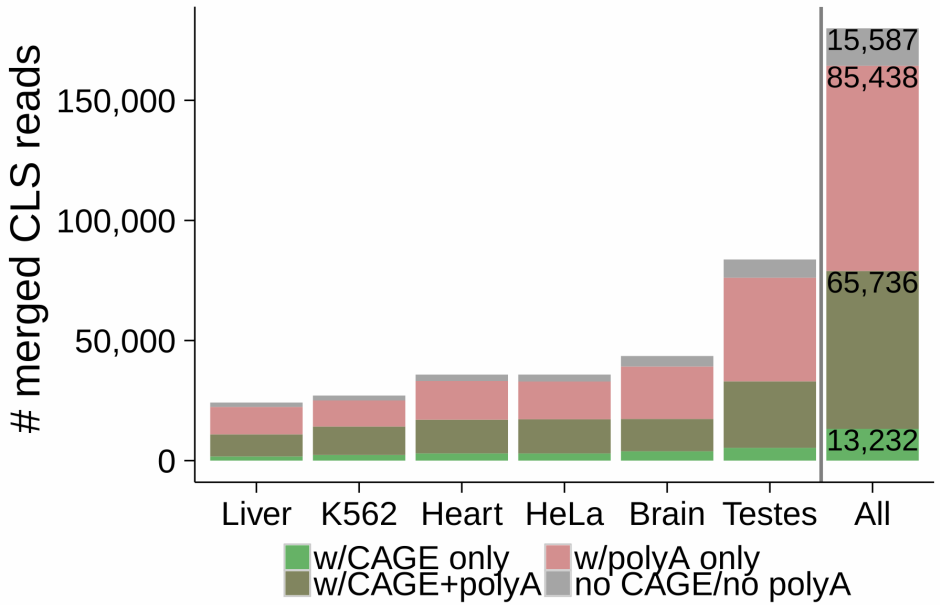


A full-length transcript catalog



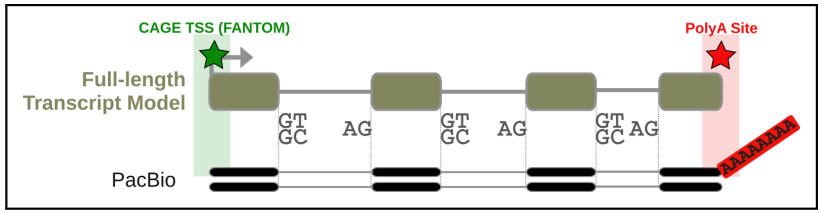
A full-length transcript catalog

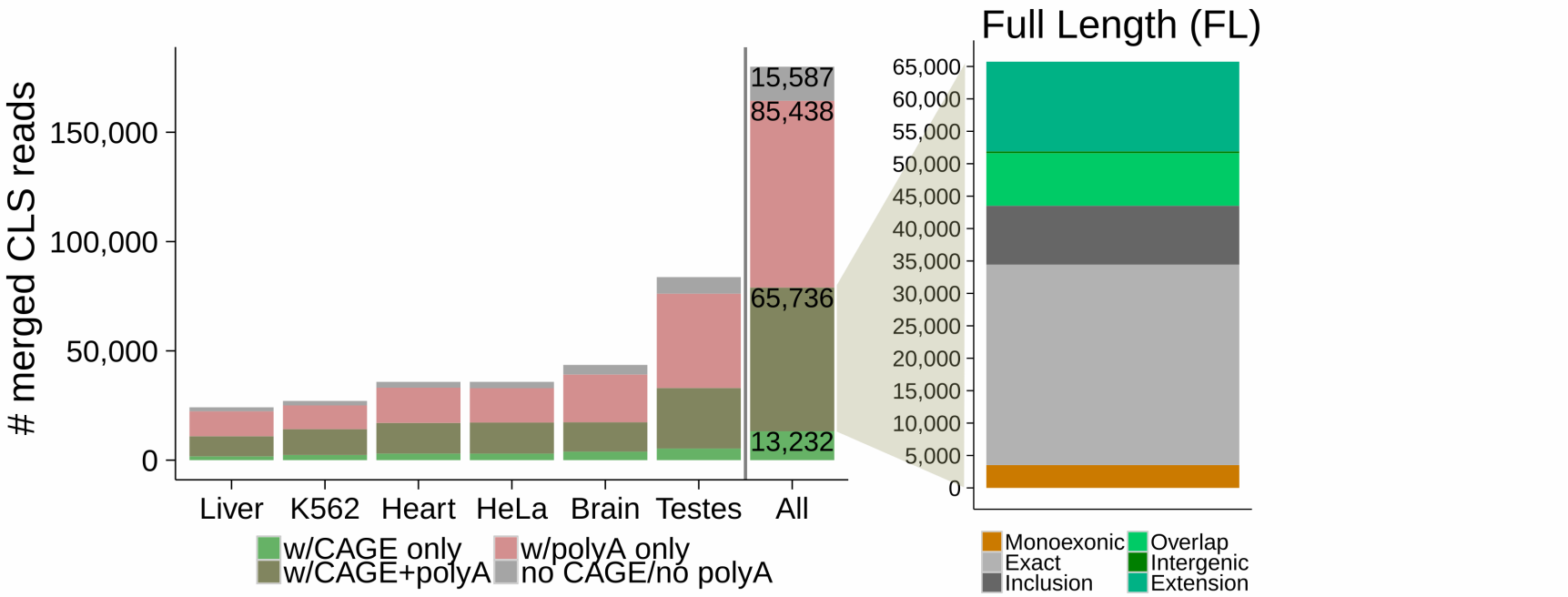




Lagarde *et al. Nature Genetics*, 2017

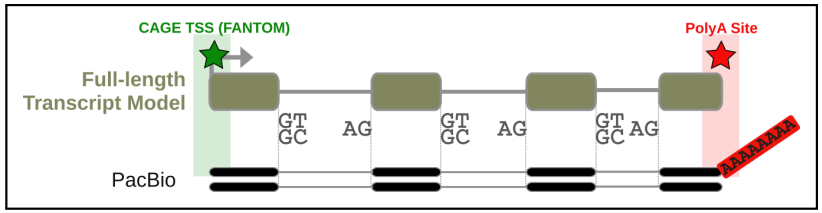
A full-length transcript catalog

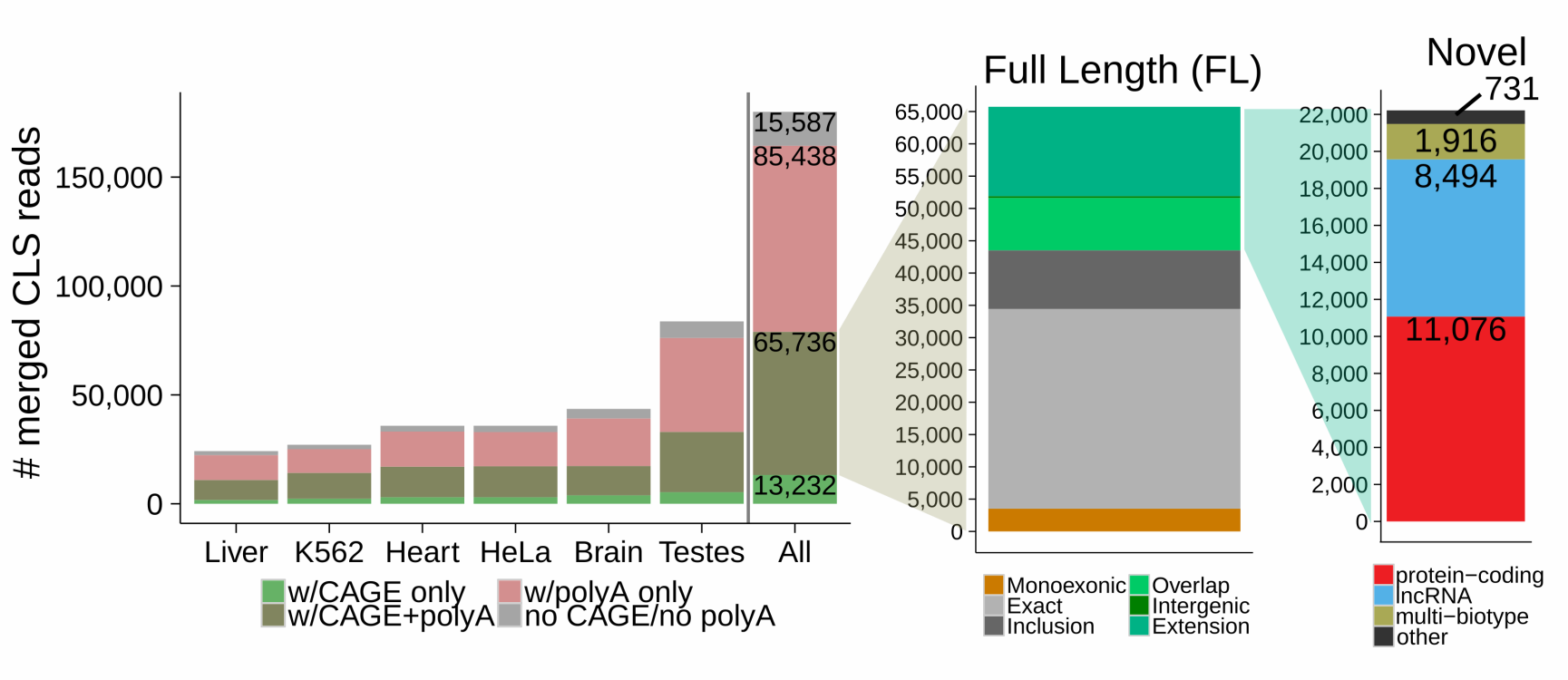




Lagarde *et al. Nature Genetics*, 2017

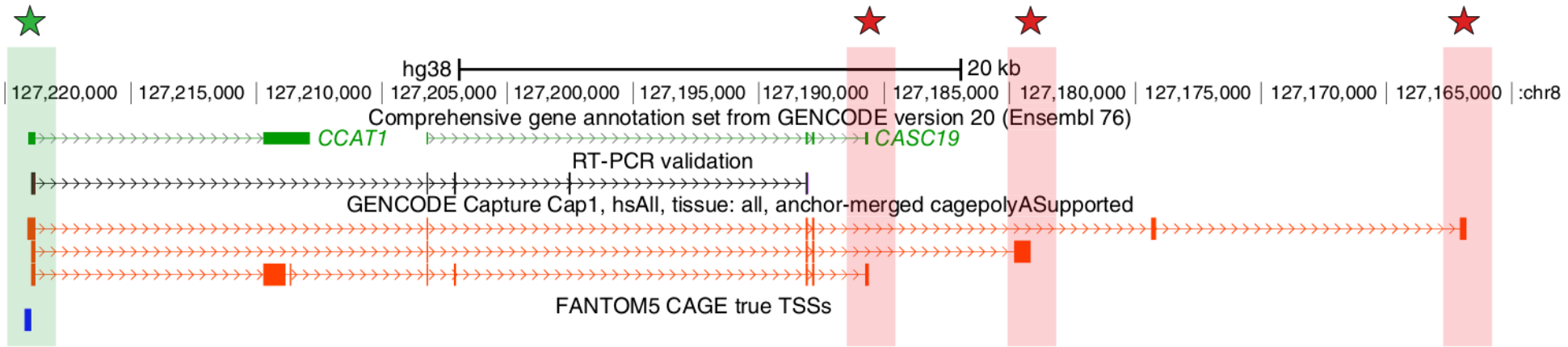
A full-length transcript catalog





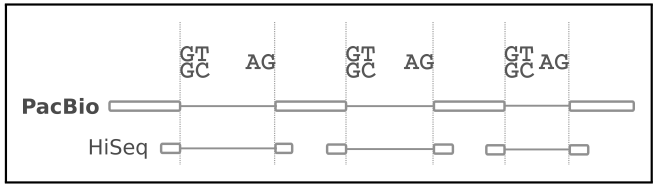
Lagarde *et al. Nature Genetics*, 2017

Novel full-length transcript structures in the *CCAT1* lincRNA locus

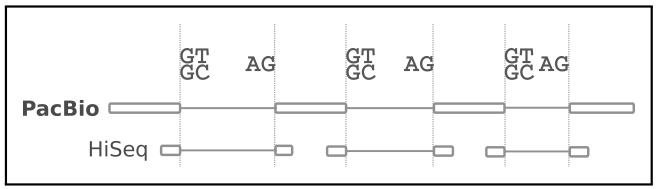


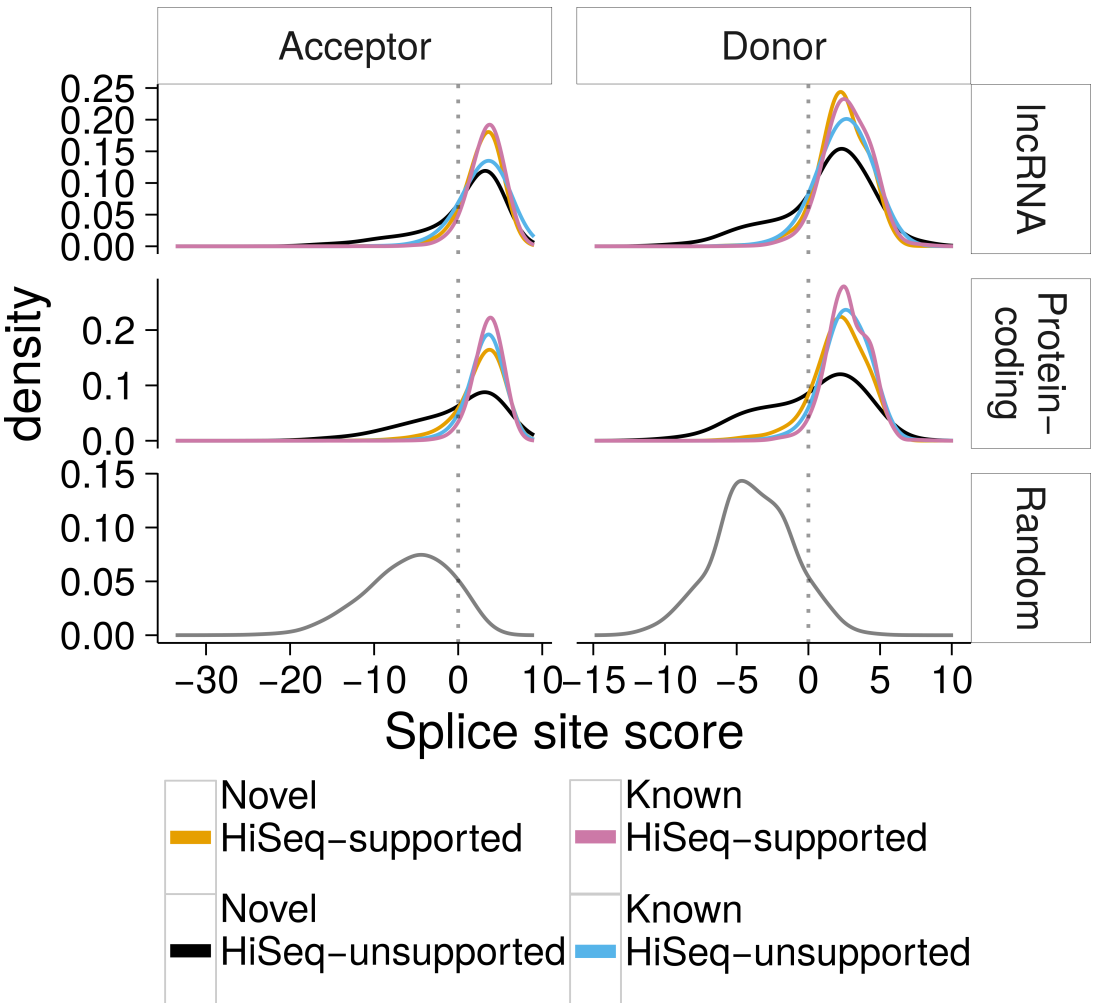
Lagarde *et al. Nature Genetics*, 2017

CLS splice junctions match the quality of GENCODE-annotated ones

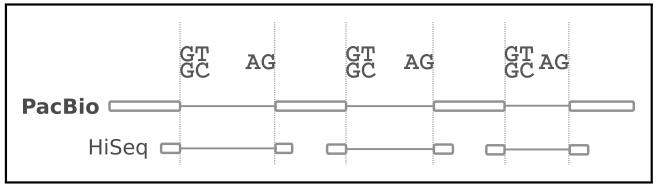


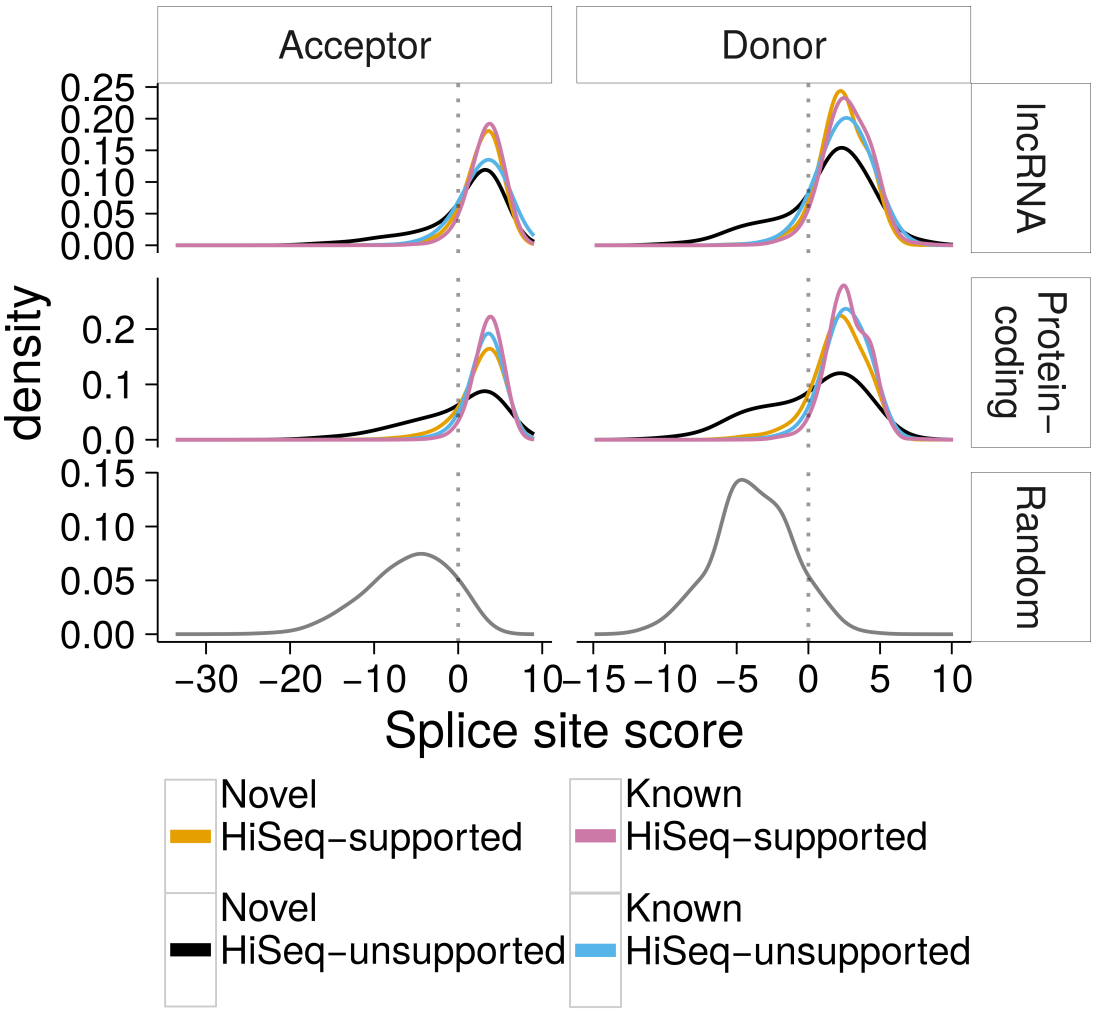
CLS splice junctions match the quality of GENCODE-annotated ones





CLS splice junctions match the quality of GENCODE-annotated ones

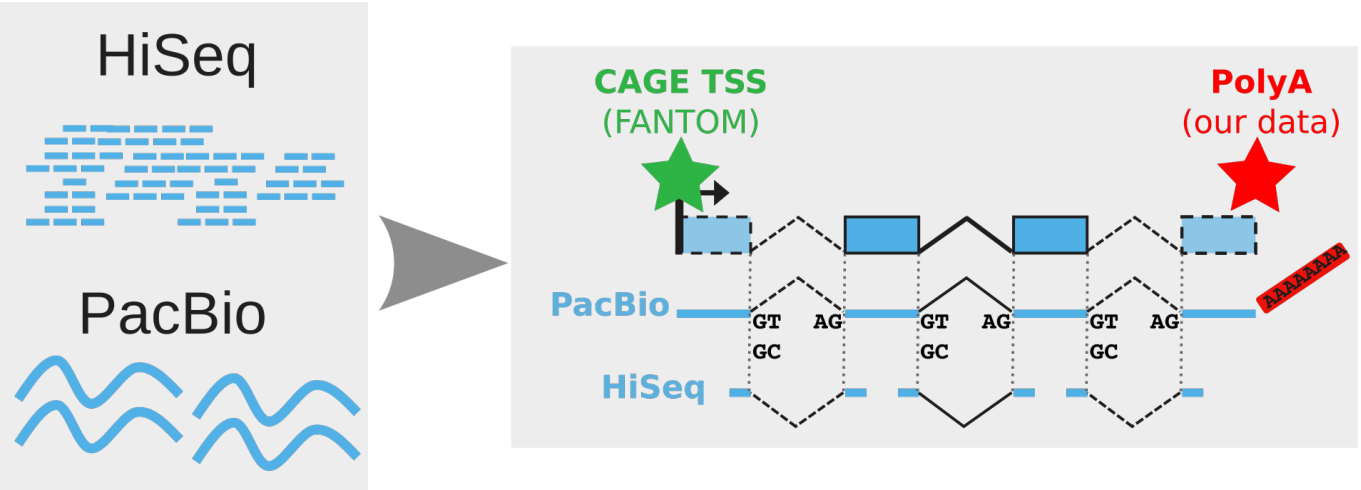




86% of CLS TMs are fully HiSeq-supported

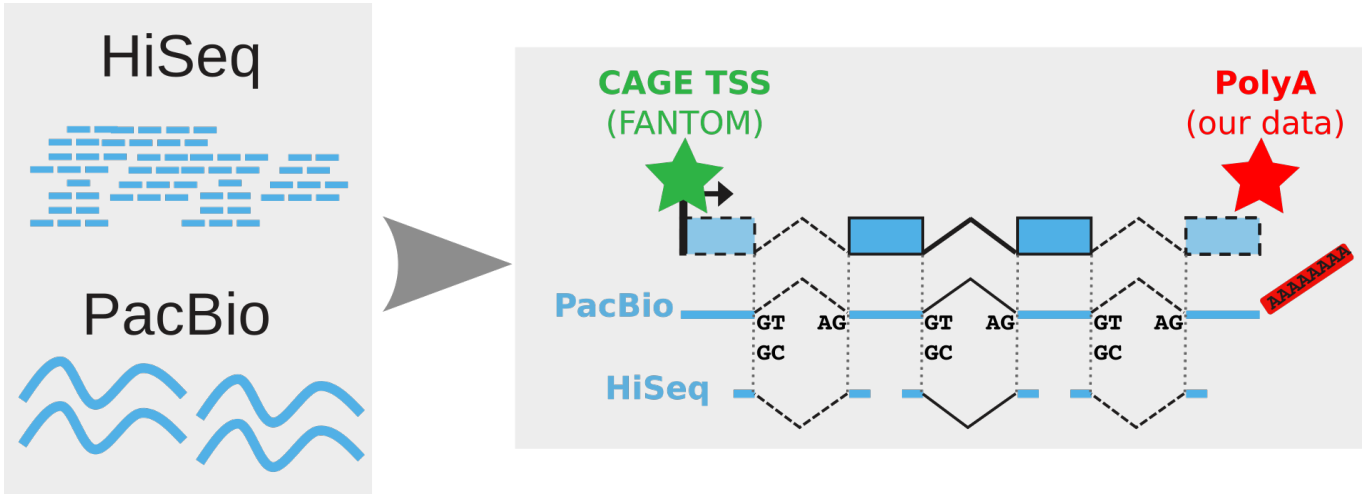
Lagarde *et al. Nature Genetics*, 2017

Short- *vs* long-read sequencing



CLS

Short- *vs* long-read sequencing



CLS

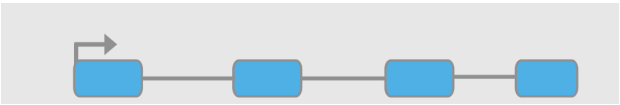
vs

Transcript reconstruction software *

(* Pertea *et al. Nature Biotechnology*, 2015)

CLS discovers a wealth of novel lincRNA transcript structures...

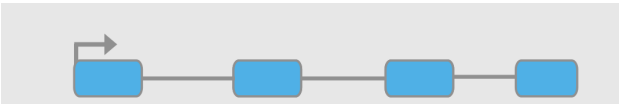
... and vastly outperforms StringTie



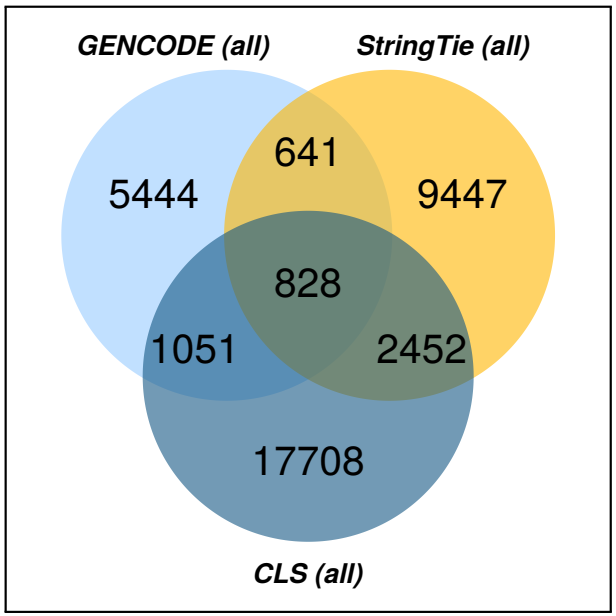
All lincRNA TMs:

CLS discovers a wealth of novel lincRNA transcript structures...

... and vastly outperforms StringTie



All lincRNA TMs:

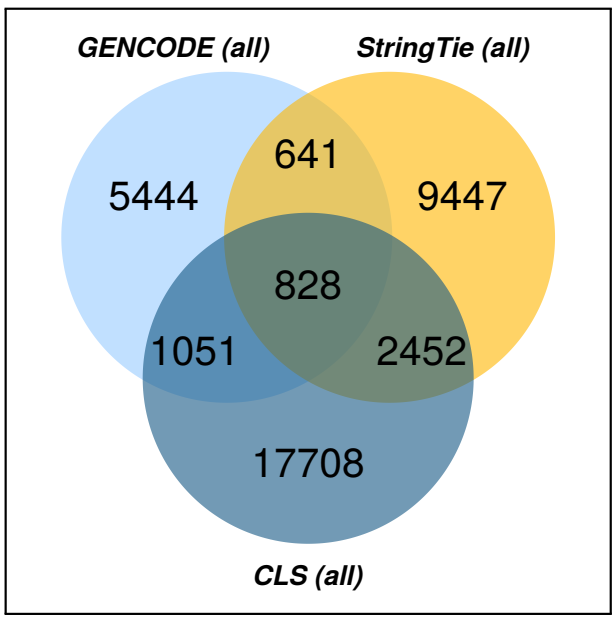


- CLS expands lincRNA transcript annotation ~**3.5x** (7,964 → 28,124)

CLS discovers a wealth of novel lincRNA transcript structures...

... and vastly outperforms StringTie

All lincRNA TMs:

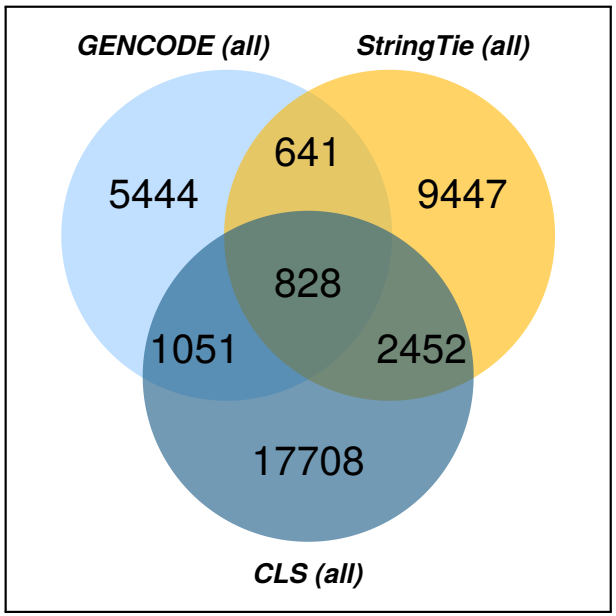


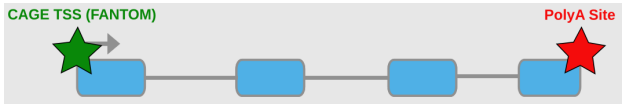
- CLS expands lincRNA transcript annotation ~**3.5x** (7,964 → 28,124)
- Novel CLS transcript structures are found in **3,574 lincRNA loci**

CLS discovers a wealth of novel lincRNA transcript structures...

... and vastly outperforms StringTie

All lincRNA TMs:



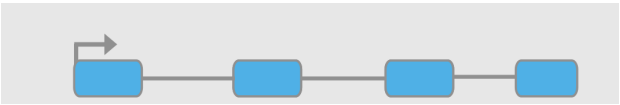


Full-length (FL) lincRNA TMs:

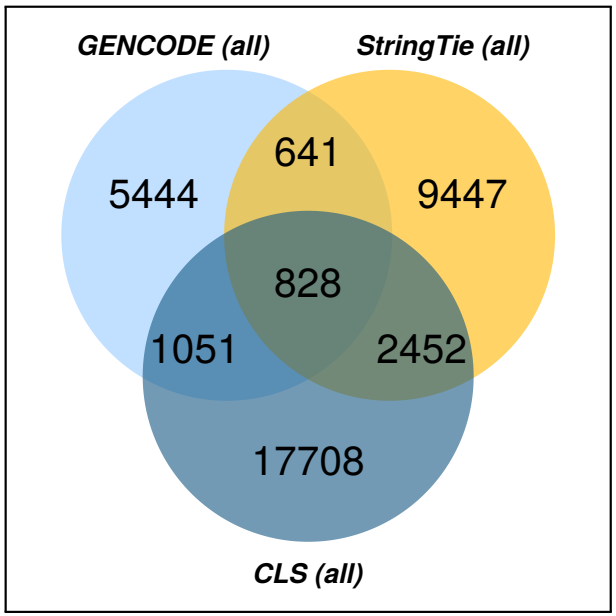
- CLS expands lincRNA transcript annotation ~**3.5x** (7,964 → 28,124)
- Novel CLS transcript structures are found in **3,574 lincRNA loci**

CLS discovers a wealth of novel lincRNA transcript structures...

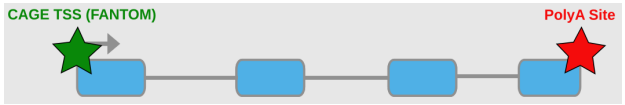
... and vastly outperforms StringTie



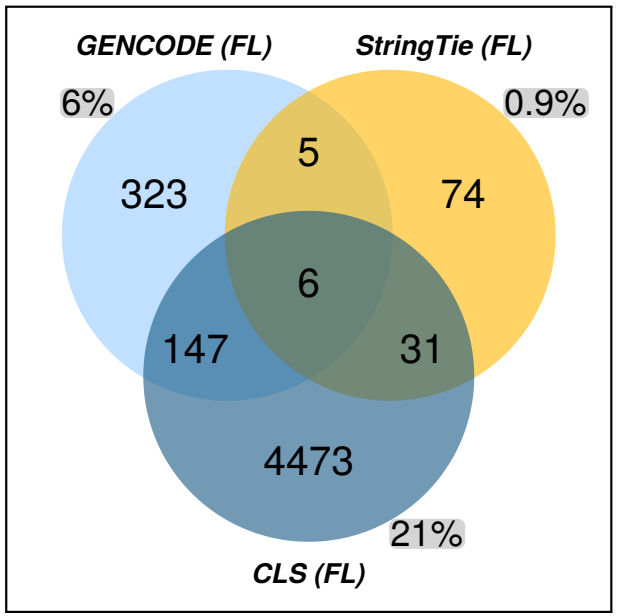
All lincRNA TMs:



- CLS expands lincRNA transcript annotation ~**3.5x** (7,964 → 28,124)
- Novel CLS transcript structures are found in **3,574 lincRNA loci**



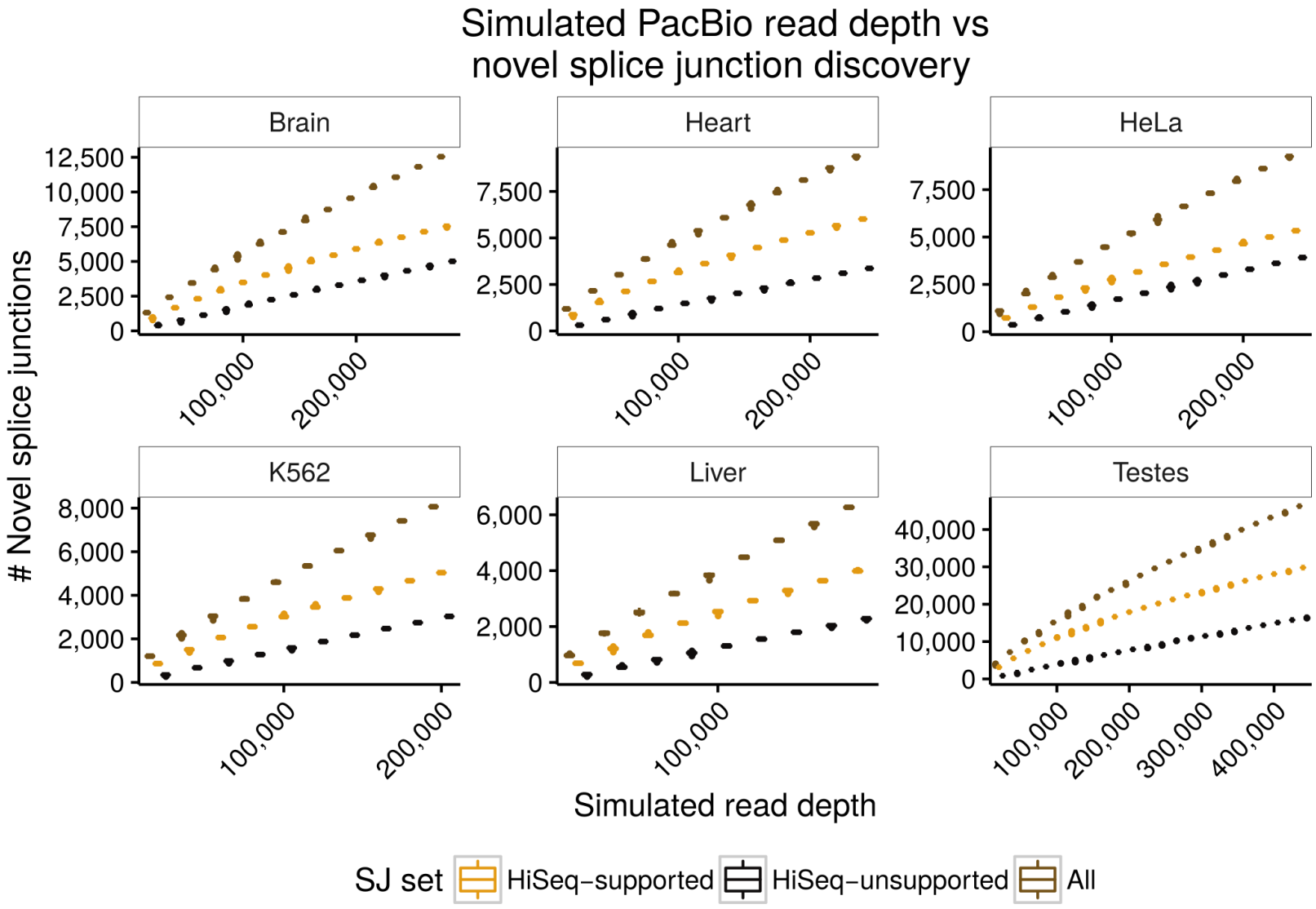
Full-length (FL) lincRNA TMs:



- CLS expands **FL** lincRNA **transcript** annotation ~**10x** (481 → 4,985)
- Novel CLS **FL** transcript structures are found in **947 lincRNA loci**

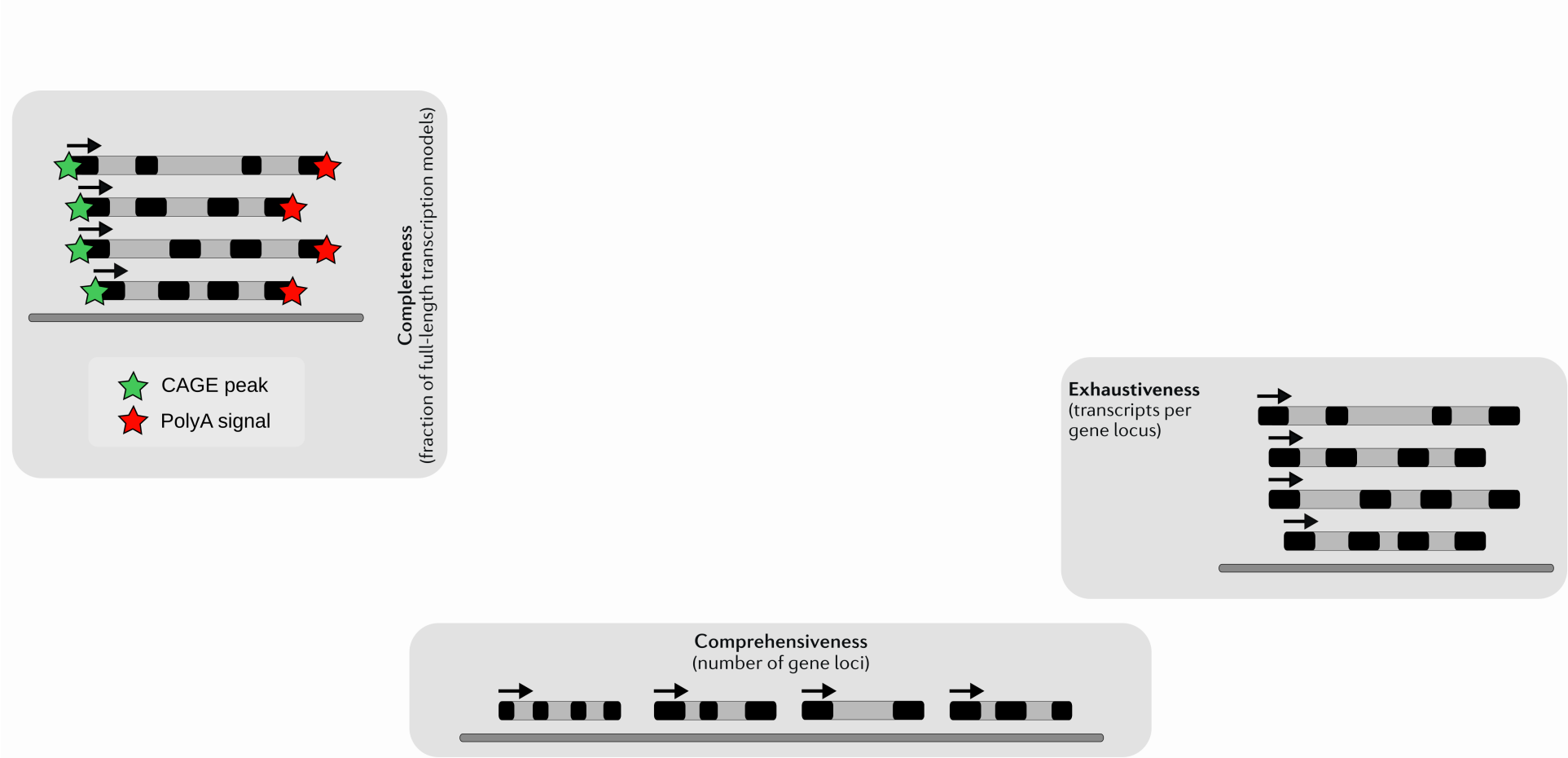
(Lagarde *et al. Nature Genetics*, 2017)

We’re still far from annotation completeness



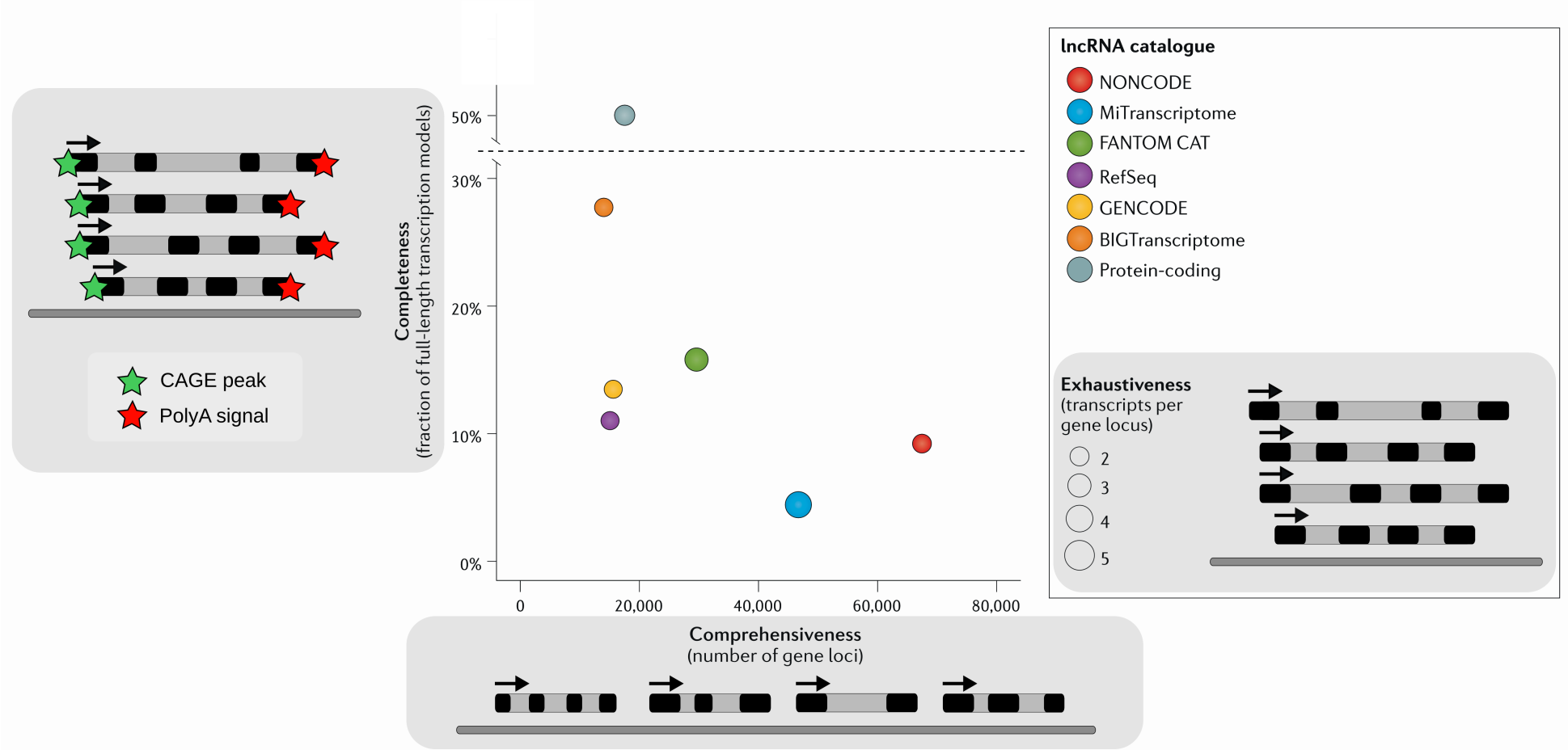
(Lagarde *et al. Nature Genetics*, 2017)

Comparison of current lncRNA gene catalogs



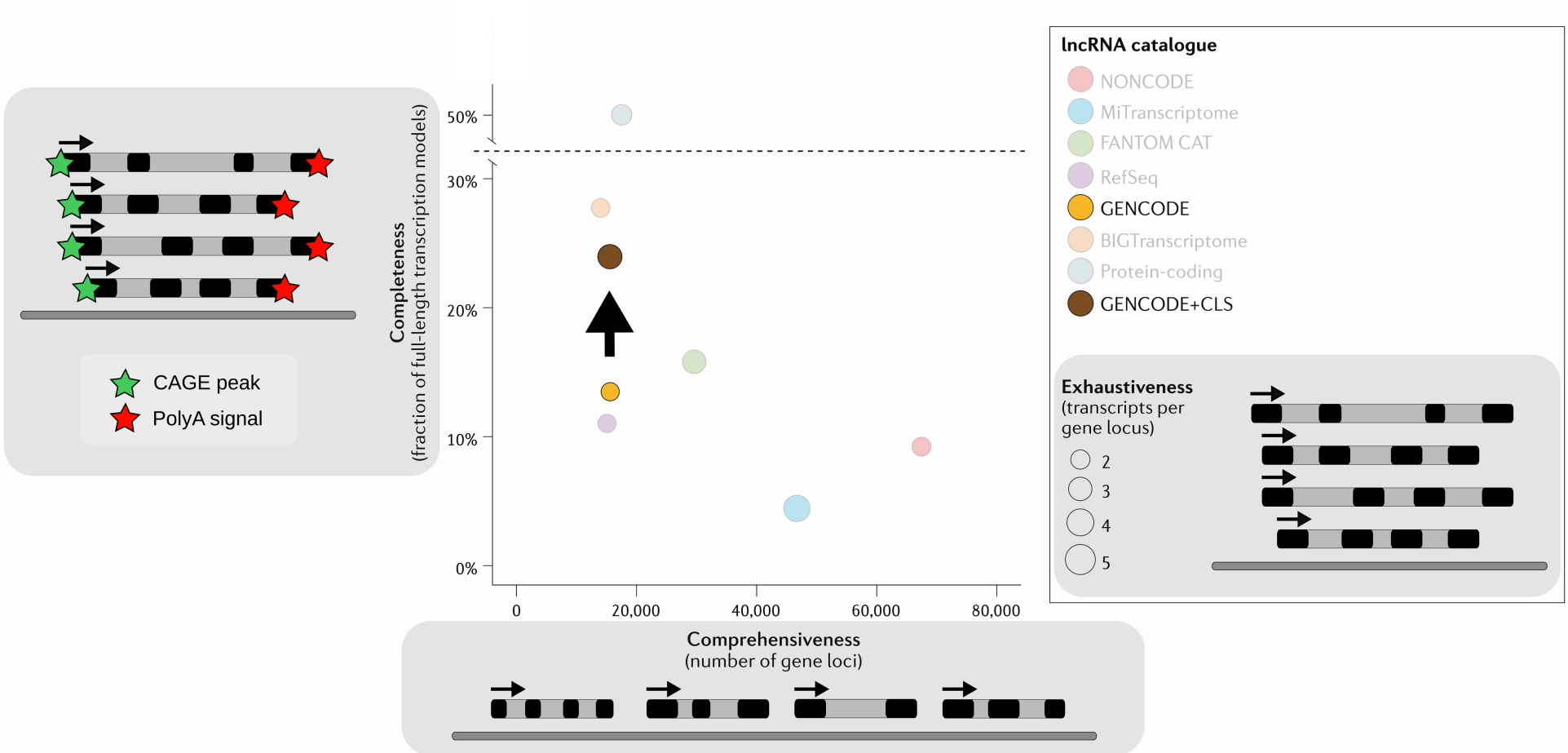
(Uszczynska-Ratajczak *et al. Nature Reviews Genetics*, 2018)

Comparison of current lncRNA gene catalogs



(Uszczynska-Ratajczak *et al. Nature Reviews Genetics*, 2018)

Comparison of current lncRNA gene catalogs

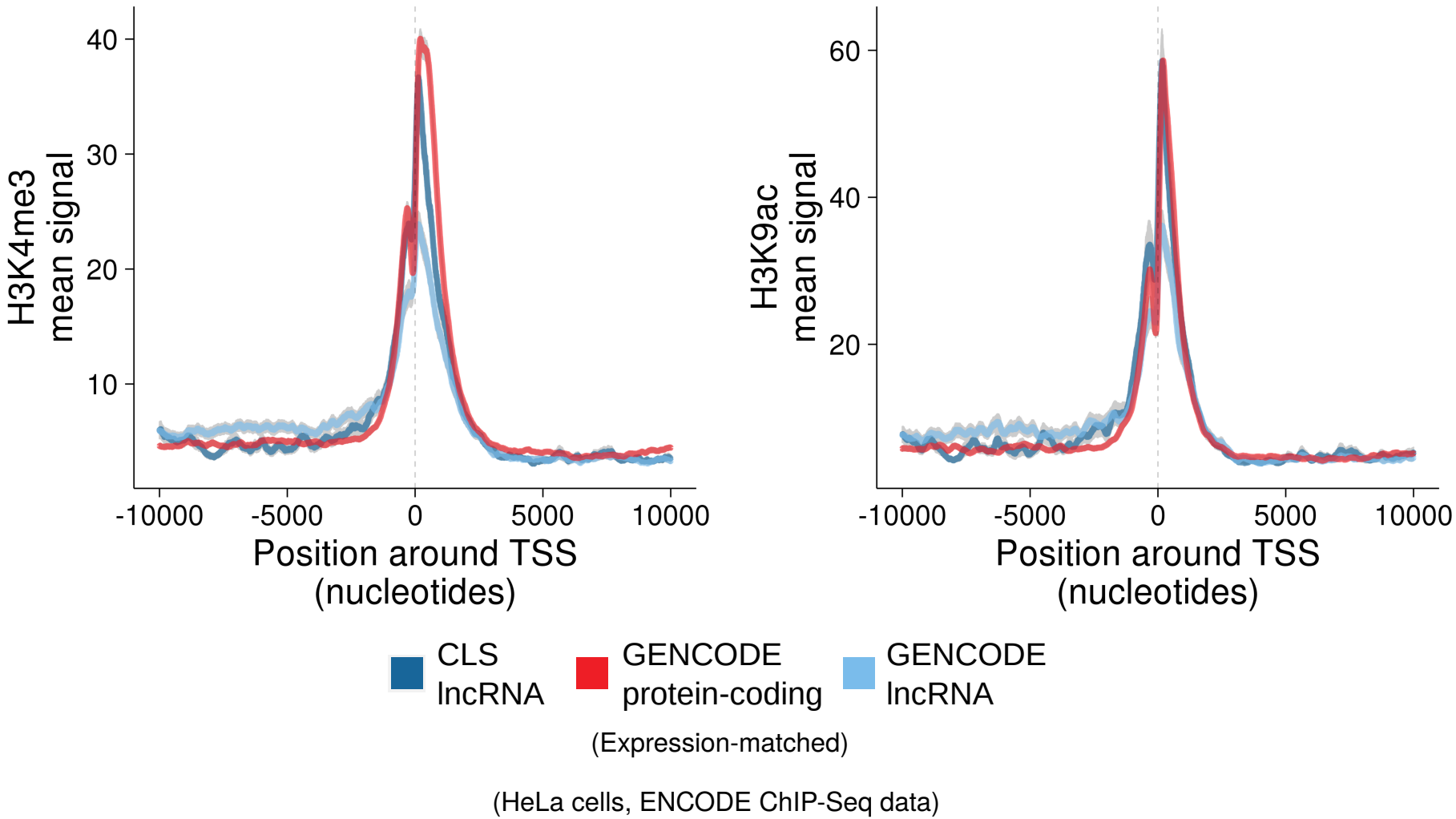


(Uszczynska-Ratajczak *et al. Nature Reviews Genetics*, 2018)

CLS improves GENCODE's

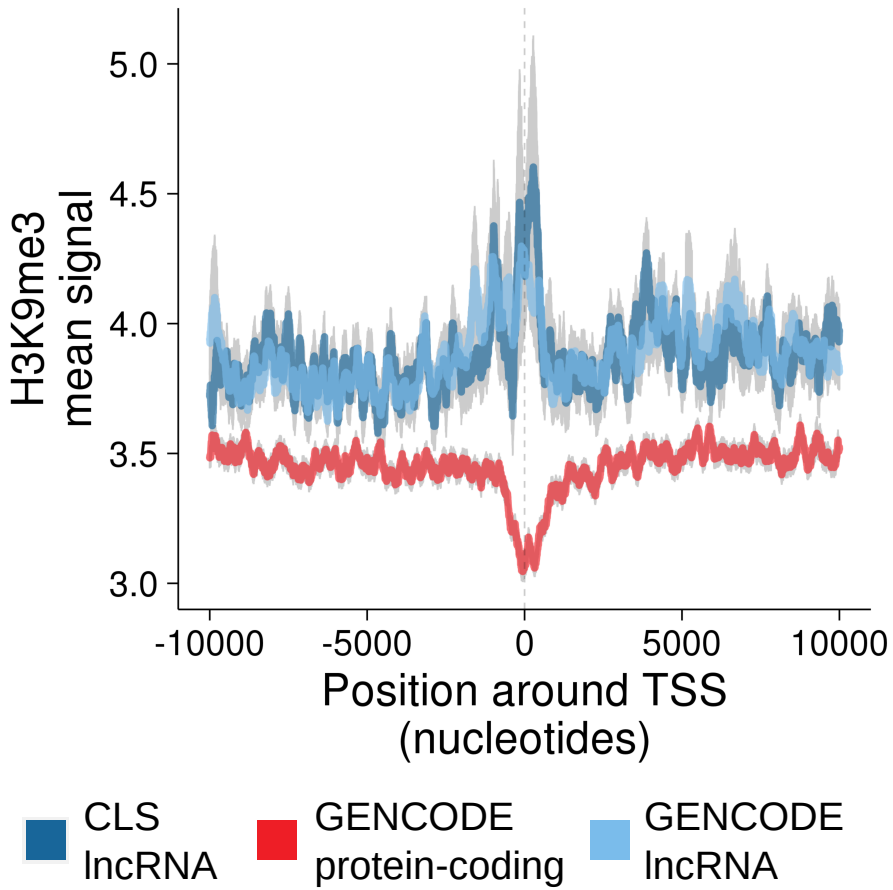
- **exhaustiveness** (1.9 → 3.3 transcripts/locus)
- **completeness** (13.5% → 24%)

LincRNA promoters share **similar levels** of **active** chromatin marks as protein-coding genes



(Lagarde *et al. Nature Genetics*, 2017)

LincRNA promoters show **higher levels** of **repressive** marks than protein-coding genes

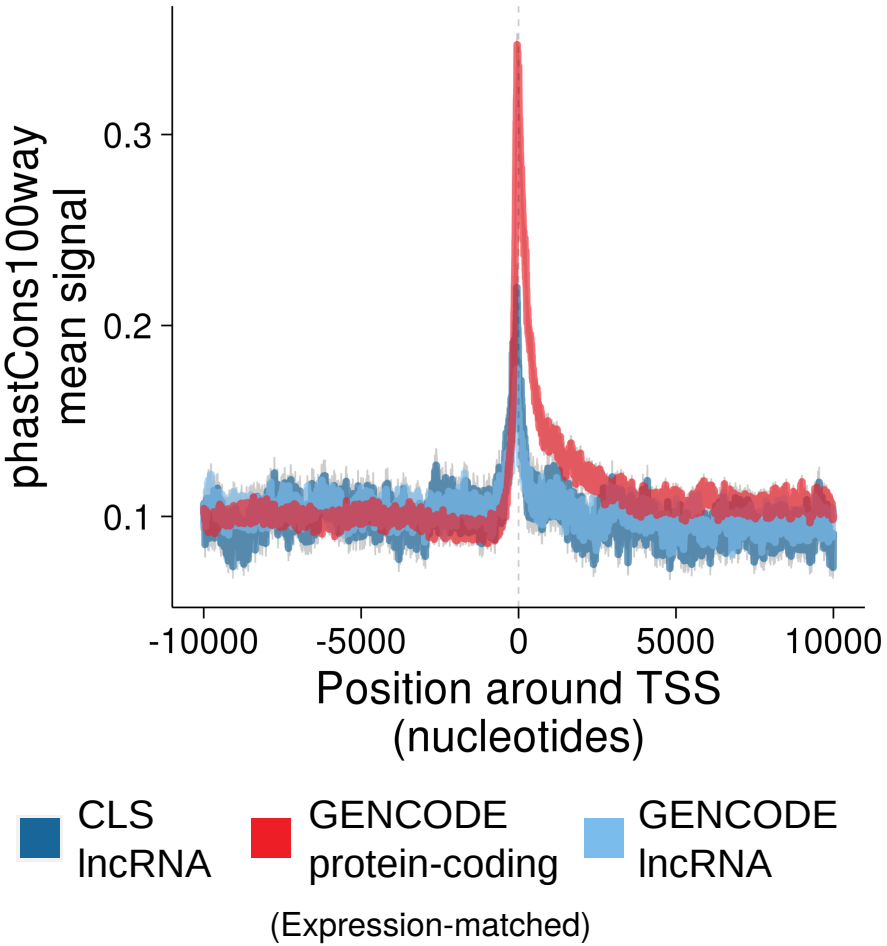


(HeLa cells, ENCODE ChIP-Seq data)

(Lagarde *et al. Nature Genetics*, 2017)

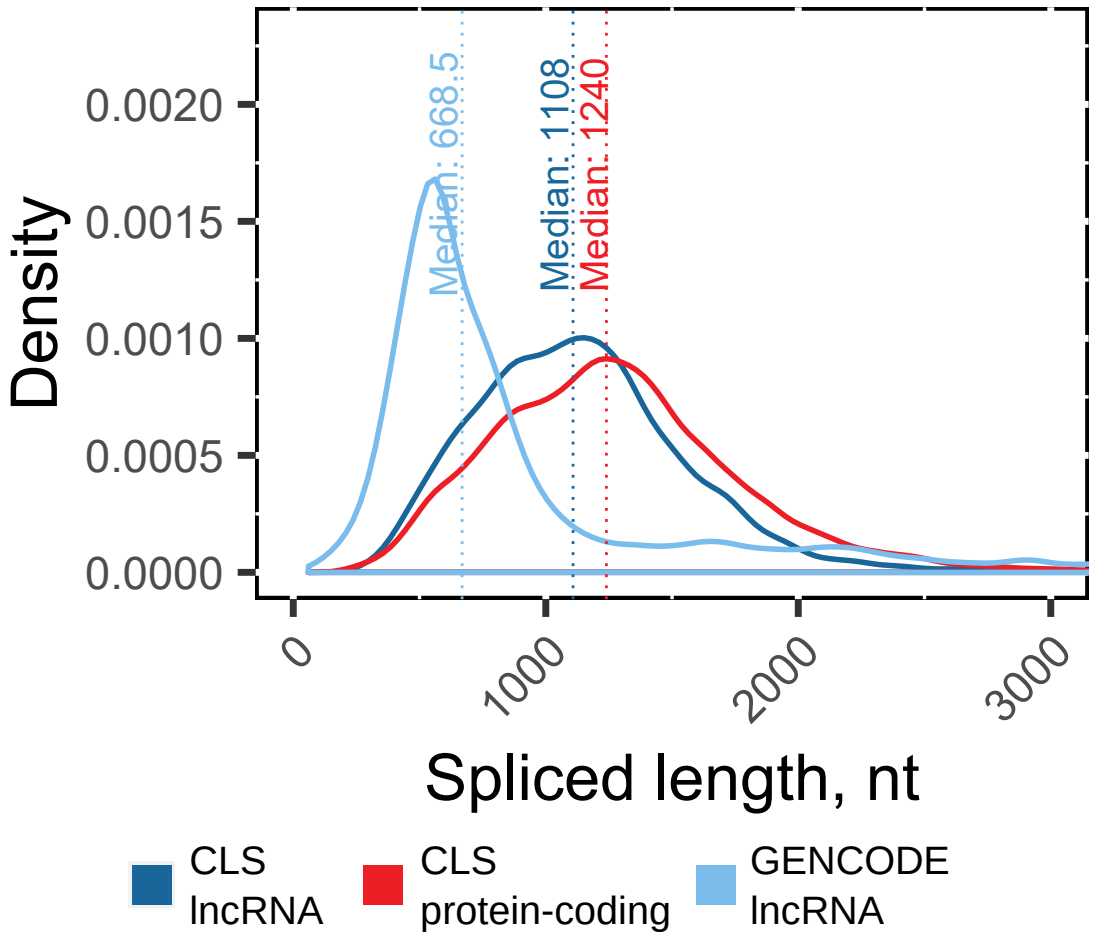
LincRNA promoters are evolutionarily conserved...

... although less than their protein-coding counterparts



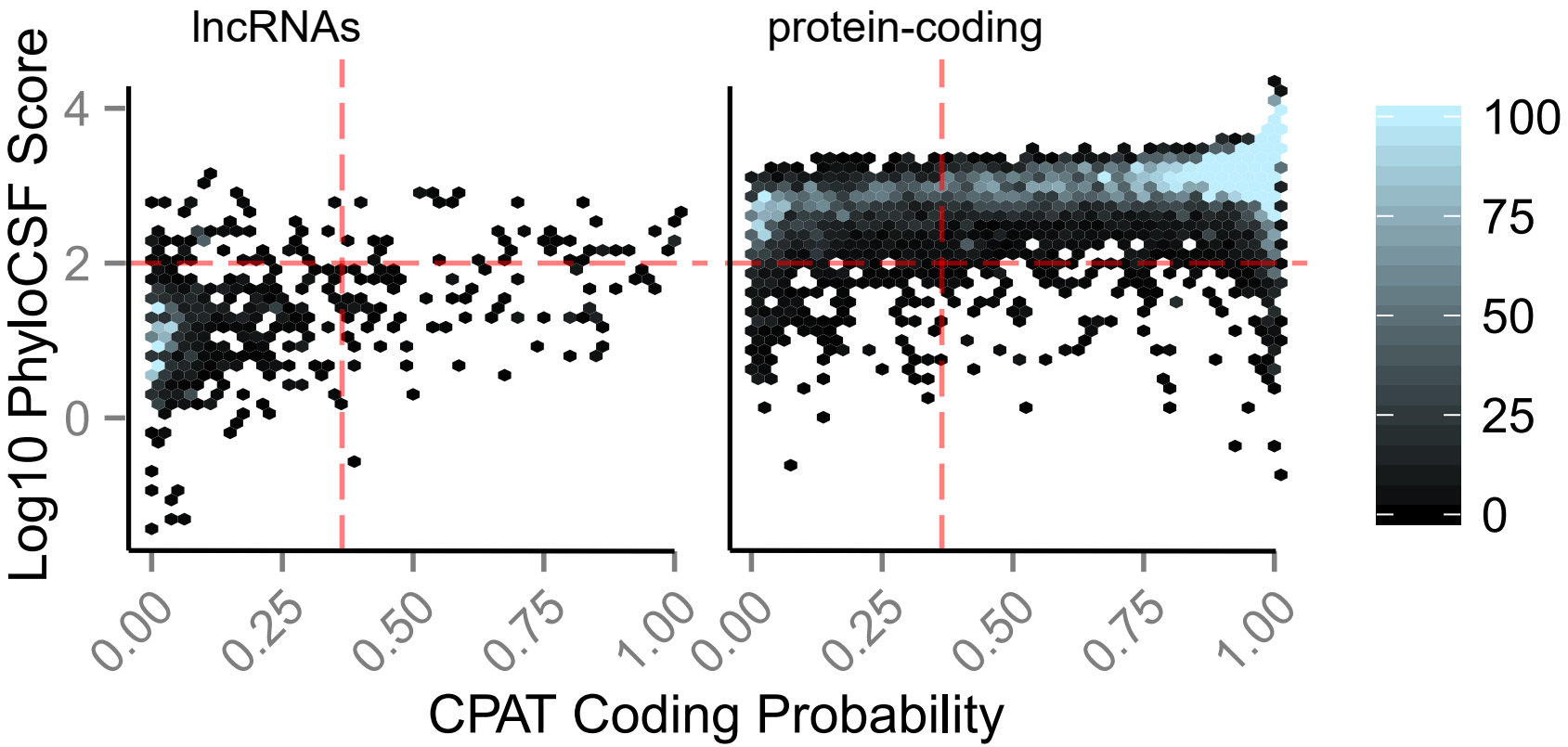
(Lagarde *et al. Nature Genetics*, 2017)

LincRNAs are not much shorter than mRNAs



(Lagarde *et al. Nature Genetics*, 2017)

LincRNAs show low coding potential, even after re-annotation



(Lagarde *et al. Nature Genetics*, 2017)

Conclusion

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases** ~**3.5x**, and shows **no signs of saturation**

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**
  - CLS **outperforms short-read-based transcriptome assembly** methods

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**
  - CLS **outperforms short-read-based transcriptome assembly** methods
- The CLS-based gene catalog allows us to **revisit confidently some lincRNA characteristics**:

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**
  - CLS **outperforms short-read-based transcriptome assembly** methods
- The CLS-based gene catalog allows us to **revisit confidently some lincRNA characteristics**:
  - LincRNAs are confirmed to bear **little coding potential**

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**
  - CLS **outperforms short-read-based transcriptome assembly** methods
- The CLS-based gene catalog allows us to **revisit confidently some lincRNA characteristics**:
  - LincRNAs are confirmed to bear **little coding potential**
  - Mature lncRNA transcripts are likely to be just **as long as coding ones**

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**
  - CLS **outperforms short-read-based transcriptome assembly** methods
- The CLS-based gene catalog allows us to **revisit confidently some lincRNA characteristics**:
  - LincRNAs are confirmed to bear **little coding potential**
  - Mature lncRNA transcripts are likely to be just **as long as coding ones**
  - LincRNA promoters show clear signs of **evolutionary conservation**

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**
  - CLS **outperforms short-read-based transcriptome assembly** methods
- The CLS-based gene catalog allows us to **revisit confidently some lincRNA characteristics**:
  - LincRNAs are confirmed to bear **little coding potential**
  - Mature lncRNA transcripts are likely to be just **as long as coding ones**
  - LincRNA promoters show clear signs of **evolutionary conservation**
  - The **histone environment** of lincRNA promoters is:
    - **Similar** to that of protein-coding ones in terms of **active marks**

Conclusion

- **CLS**, an efficient high-throughput, high-quality annotation method:
  - LncRNA transcript **complexity increases ~3.5x**, and shows **no signs of saturation**
  - CLS transcript models look **as genuine as manually-curated ones**
  - The CLS set is enriched in **full-length** lncRNA transcript models, leading to a **much-improved definition of lincRNA genome boundaries**
  - CLS **outperforms short-read-based transcriptome assembly** methods
- The CLS-based gene catalog allows us to **revisit confidently some lincRNA characteristics**:
  - LincRNAs are confirmed to bear **little coding potential**
  - Mature lncRNA transcripts are likely to be just **as long as coding ones**
  - LincRNA promoters show clear signs of **evolutionary conservation**
  - The **histone environment** of lincRNA promoters is:
    - **Similar** to that of protein-coding ones in terms of **active marks**
    - **Enriched in repressive marks** compared to protein-coding gene promoters

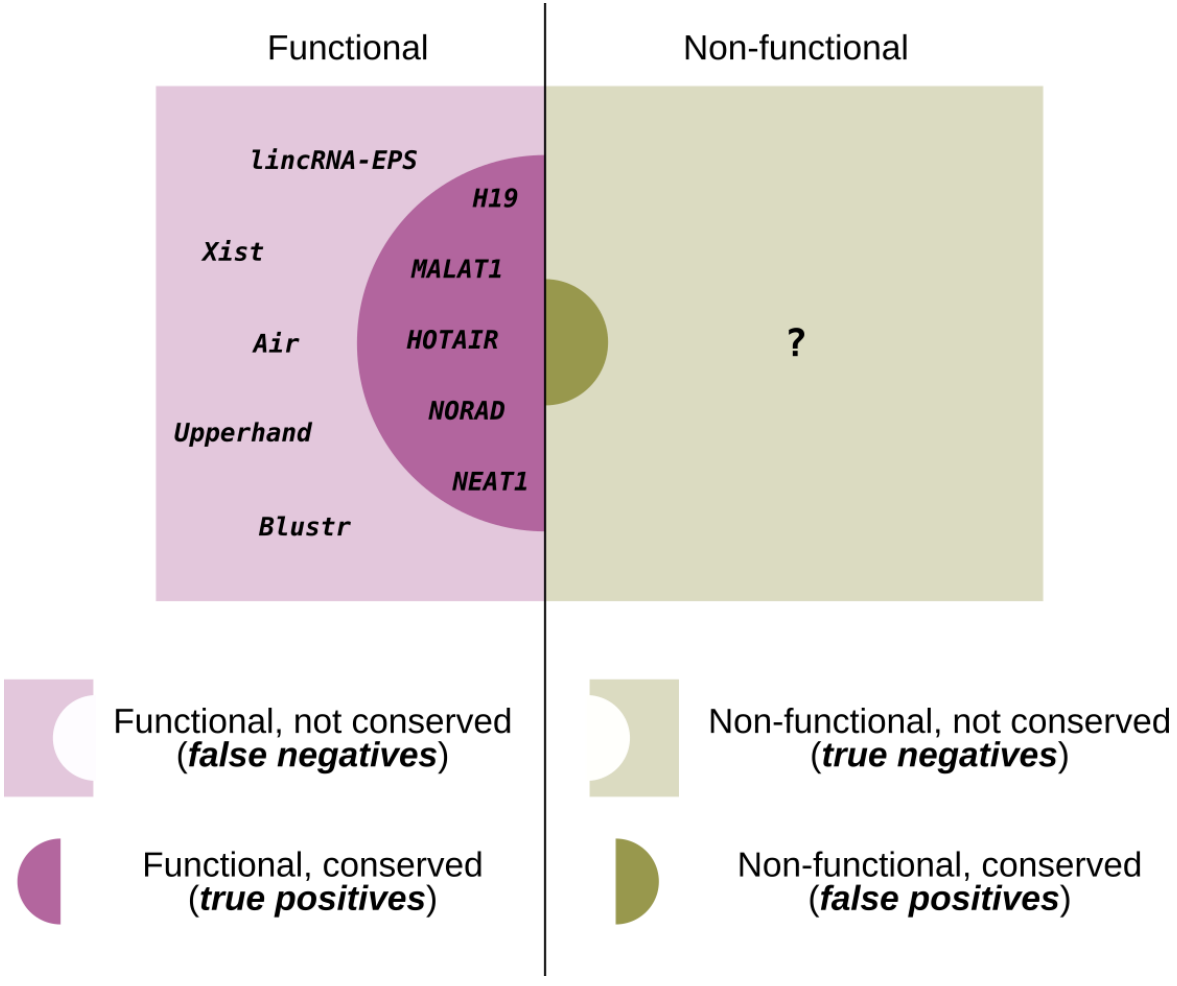
Thanks

- **CRG**
  - **Rory Johnson** (University of Bern)
  - **Barbara Uszczynska-Ratajczak** (CeNT, Warsaw)
  - **Sílvia Carbonell**
  - Carme Arnan
  - Amaya Abad
  - Sílvia Pérez-Lluch
  - Dmitri Pervouchine
  - Romina Garrido
  - **Roderic Guigó**
- **CSHL** (NY)
  - Sara Goodwin
  - Alex Dobin
  - Tom Gingeras
- **HAVANA Team** (EBI, Hinxton)
  - Jon Mudge
  - Jose-Manuel González
  - Jennifer Harrow
  - Adam Frankish
- **ICR** (London)
  - Jyoti Choudhary
  - James Wright
- **MIT** (Boston)
  - Irwin Jungreis

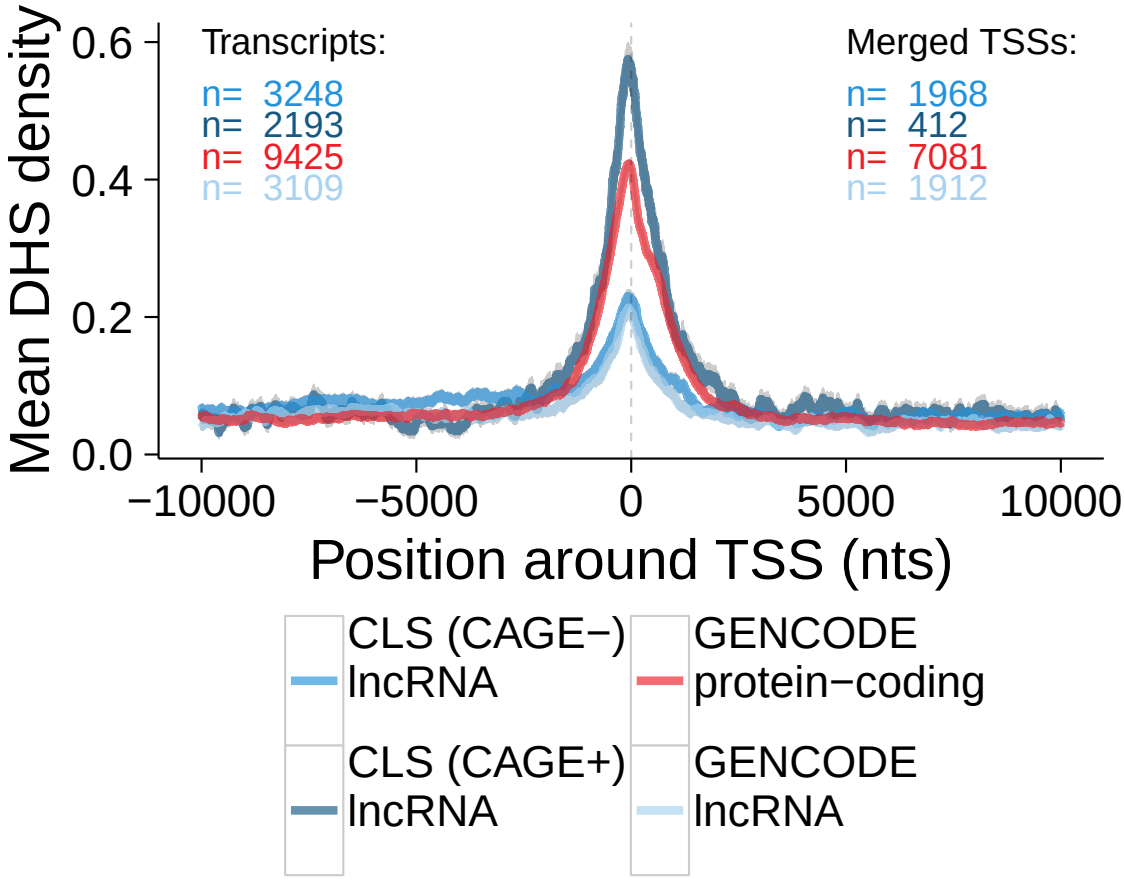
Funding:

National Human Genome Research Institute

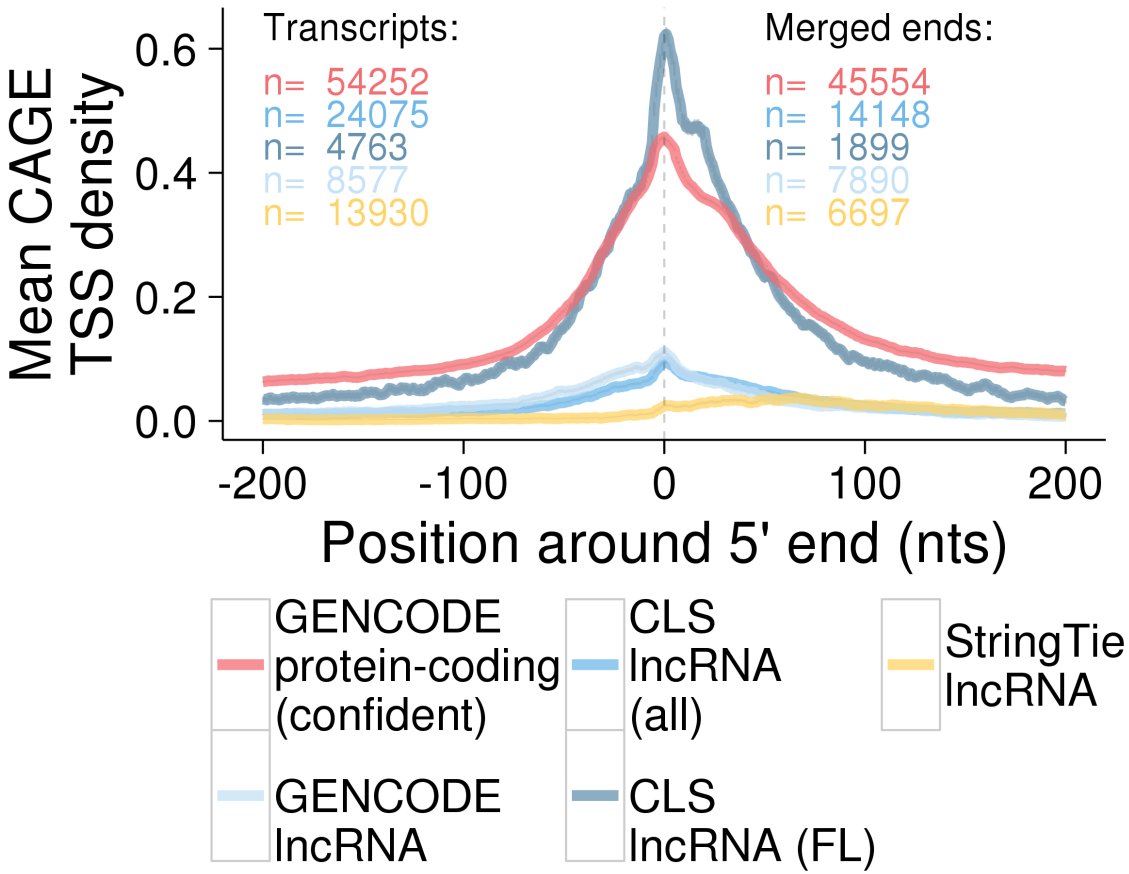
Evolutionary conservation as a predictor of lncRNA function



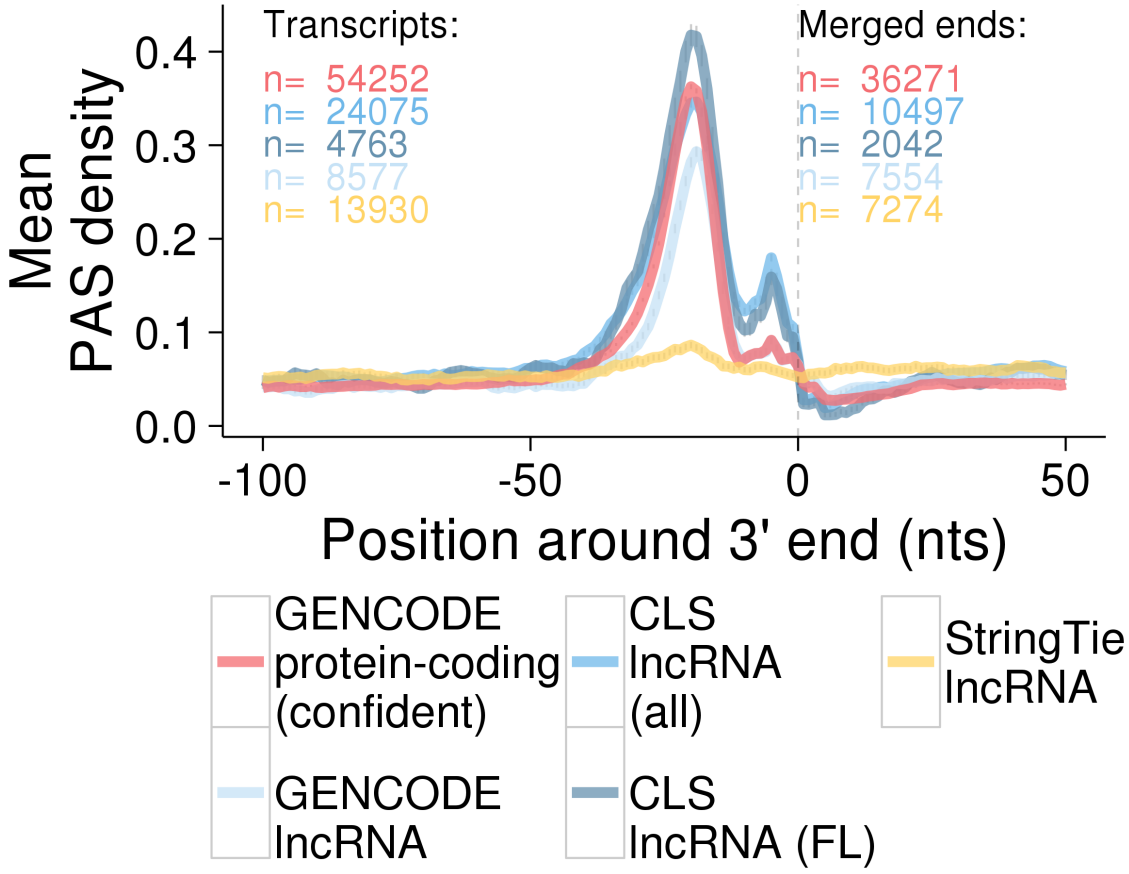
DNAse Hypersensitive sites



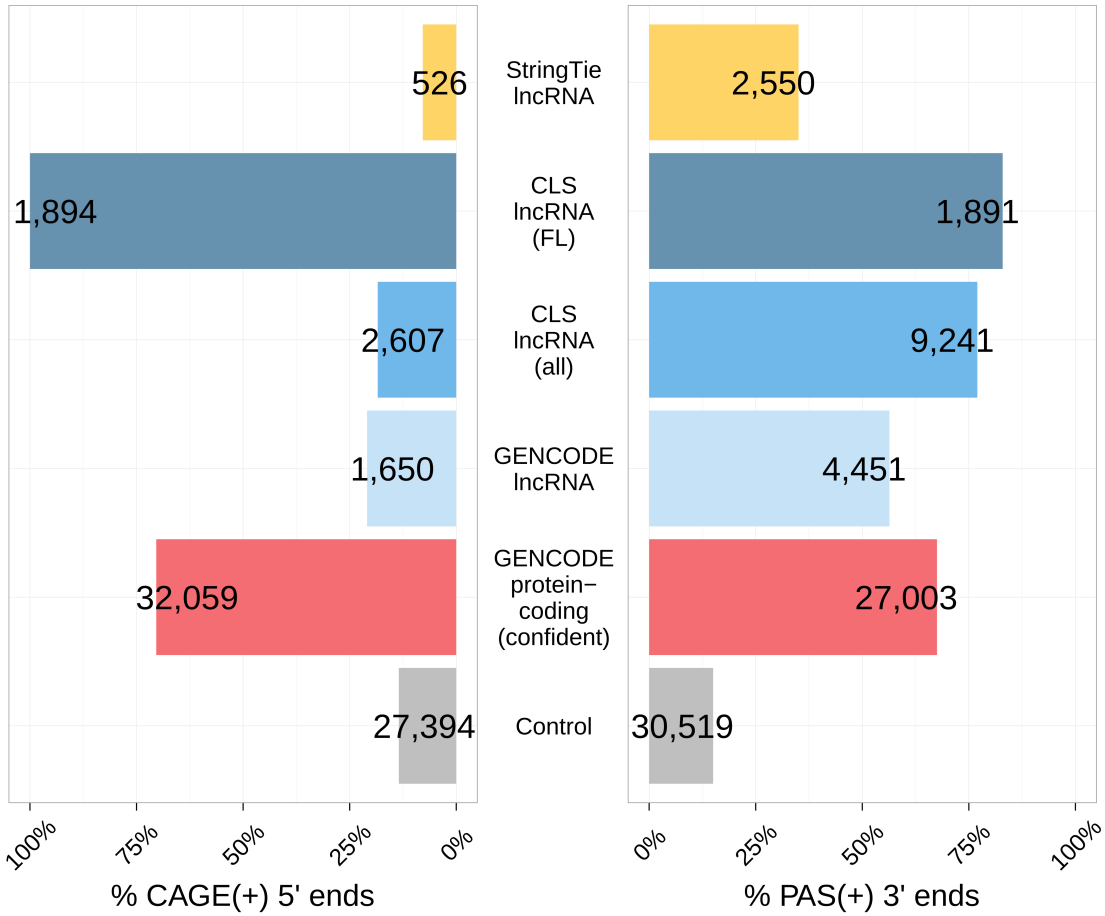
CLS/GENCODE/StringTie vs CAGE



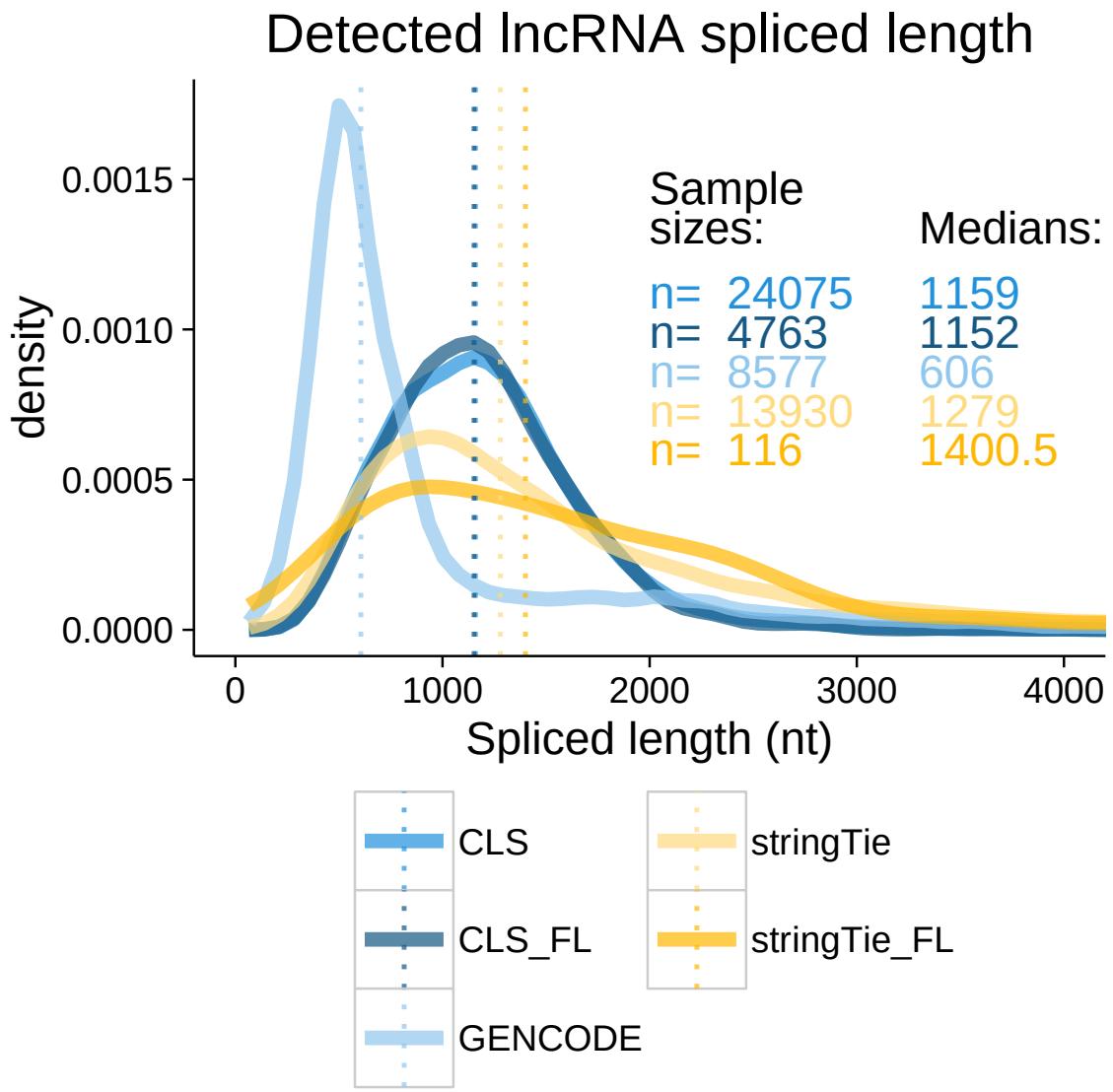
CLS/GENCODE/StringTie vs PAS



CLS/GENCODE/StringTie vs CAGE+PAS



CLS/GENCODE/StringTie TM length



HiSeq Discovery/Saturation