Lecture 1: Introduction to the Transcriptome

BIOINF3005/7160: Transcriptomics Applications

Dr Stephen Pederson

Bioinformatics Hub,
The University of Adelaide

March 2nd, 2020

University Statement

Course Details

Why Transcriptomics?

Key Definitions

Promoters

Transcription

Messenger RNA

Non-Coding RNA

THE UNIVERSITY
of ADELAIDE

University Statement

Welcome to the University of Adelaide

- The start of any academic year is an exciting time - but it can also be challenging for our students, for many reasons
- For some of our students, this year is especially challenging
- That's because many of our students who wanted to be here to study can't be here - because of COVID-19 and the travel restrictions to Australia
- This is not a situation that those students - or any of us in this room today - have any direct control over
- We may have some students watching this very lecture from overseas

Welcome to the University of Adelaide

- Before the lecture begins, there are some points worth making so that everyone understands what the current situation is here at the University of Adelaide
- We have a diverse, inclusive, and welcoming community at our University, which is something we're very proud of
- In times of difficulty, that community is more important than ever.
- This is a time when our community comes together as a whole so we can care for and support each other together
- It's particularly important for those who may be going through a more difficult time than otherwise.
- We must be respectful of other people - and that includes the many students who couldn't join us in the room today

COVID-19

- The impact of COVID-19 is being felt right around the world.
- The University's own response to COVID-19 is being shaped by the latest advice from the Government and from health authorities.
- So far, there have only been a small number of COVID-19 cases in Australia ... only three in South Australia ... and none at the University.
- There is no evidence of transmission within the general community in Australia.
- Health authorities are advising we take the same basic precautions we would when it's cold or flu season – wash your hands, cover your mouth, dispose of tissues.
- This is good advice for any time you have a cough or a cold.

COVID-19

- There are resources available for students who need someone to talk to:
  - Student Life, if you need some extra assistance or support
  - Our Safer Campus Community website has a lot of resources to help us address harassment, including ways to report and get help
  - Adelaide UniCare is our on-campus medical service, and you can make appointments online.
- The University also has an online FAQ page about COVID-19 that's being constantly updated - there's a banner on our main webpage that links to the FAQ.
- The University has made it very clear that the safety and well-being of staff and students is paramount, and we will communicate with you clearly if anything changes.
- Welcome again, and good luck with your studies.

Course Details

Key Contacts

- Prof David Adelson, *Chair of Bioinformatics, Course Coordinator*, Room 261, The Braggs
- Dr Dan Kortschak, *Program Coordinator, M Bioinformatics*, Room 260, The Braggs
- Dr Stephen Pederson, *Coordinator, Bioinformatics Hub*, Room 420, Santos Petroleum Engineering Building

Course Timetable

- Lectures:
  - Monday 9:10am - 10:00am, Room B18, Ingkarni Wardli
- Practicals:
  - Wednesday 9:10am - 11:00am, Room 111, Johnson Building
  - Friday 9:10am - 11:00am, Room 111, Johnson Building

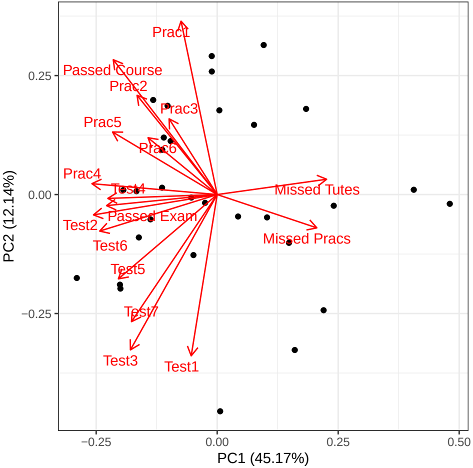
Course Outline

- Discussion of underlying biology
- Experimental approaches utilised in transcriptomic analysis
- Statistical and computational approaches utilised in transcriptomic analysis
- Practicals will be entirely computational
  - Focussed on working in `R` with some `bash`

Assessment

- Continuous Assessment (*i.e. no exam*)
- 6 assignments (+ Major Project for BIOINF7160)
- Assessment is strongly focussed on practical material
- Attendance at practicals is not compulsory, but is **strongly advised**

Results from 2019 (BIOTECH7005)



Why Transcriptomics?

The Transcriptome

Definition

The transcriptome can be defined as the complete set of (RNA) transcripts in a cell, or a population of cells, for a specific developmental stage or physiological condition[1]

Transcriptomics

Transcriptomics is simply the study of the transcriptome

[1] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nat. Rev. Genet.* 10.1 (2009), pp. 57–63.

The Transcriptome

The transcriptome can be summarised as the *RNA content of a cell*.

- Can include messenger RNA (*mRNA*), non-coding RNA (*ncRNA*), small RNA (*miRNA, piRNA etc.*)
- Can also include transfer RNA (*tRNA*) and ribosomal RNA (*rRNA*)
- Nearly all RNA is single-stranded (*ssRNA*)

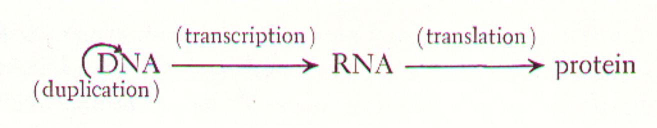
We are always dealing with a **snapshot of a dynamic process** at the time we collected data

Why Study The Transcriptome?

- *mRNA* is the intermediary step between the genome (DNA) and proteins
- small RNA and *ncRNA* play significant roles in gene regulation
- Make inference about the **biological processes driving our observations**
    - Targets for curing disease
    - Biomarkers for tumour detection
    - Understanding drought/salinity tolerance

The Central Dogma of Molecular Biology

A common (but simplistic) version of the Central Dogma:

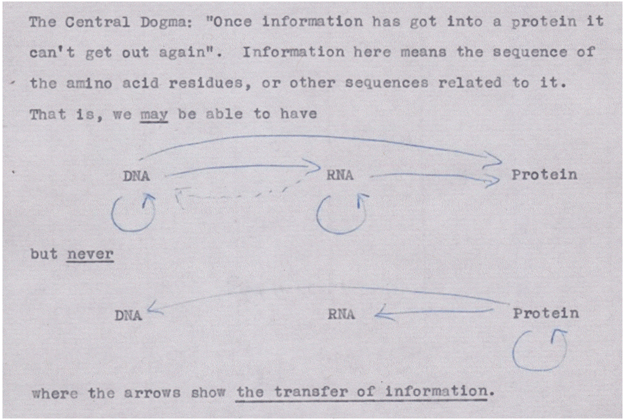


DNA makes RNA makes Protein. (Figure taken from *The Molecular Biology of the Gene*[2], p298)

[2] James Watson. *The Molecular Biology of the Gene*. W.A. Benjamin. Inc. new York, 1965.

The Central Dogma of Molecular Biology

According to Crick[3]



The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have

DNA RNA Protein

but never

DNA RNA Protein

where the arrows show the transfer of information.

[3] Frances Crick. "Ideas on protein synthesis". In: *Unpublished Note*. Wellcome Library, 1956. URL: https://wellcomelibrary.org/item/b18174139.

RNA Vs Protein

A common assumption is: RNA abundance $\propto$ Protein abundance

- This can be true, but *doesn't have to be true*
- Changes in gene expression indicate *key biological responses* to a stimulus
- Transcriptomic analysis is often about a *dynamic process*
- Often involves measuring and comparing RNA abundance (i.e. expression levels) in more than one condition

Key Definitions

What is a Gene?

Definition

The gene is the basic physical unit of inheritance. Genes are passed from parents to offspring and contain the information needed to specify traits. Genes are arranged, one after another, on structures called chromosomes. A chromosome contains a single, long DNA molecule, only a portion of which corresponds to a single gene. Humans have approximately 20,000[4] genes arranged on their chromosomes.[5]

[4] This number is clearly not correct

[5] https://www.genome.gov/genetics-glossary/Gene

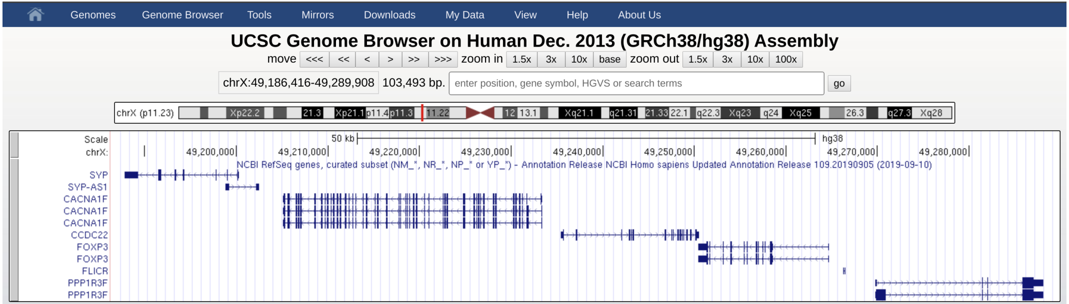
What is a Gene?

- Classically, a *gene* is a genomic locus that is **transcribed from DNA into RNA**[6]
- Some genes are protein coding, others are not
- A gene may have numerous *transcripts*, or *isoforms*
  - Some transcripts of a gene may be protein coding
  - Some transcripts from *the same gene* may not be protein coding
- Genes can range in size from dozens to millions of nucleotides

[6] NB: This definition does not include chimeric transcripts from more than one DNA locus

What is a Gene?

Here is an example region on the human X chromosome[7]



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chrX:49,186,416-49,289,908 103,493 bp.

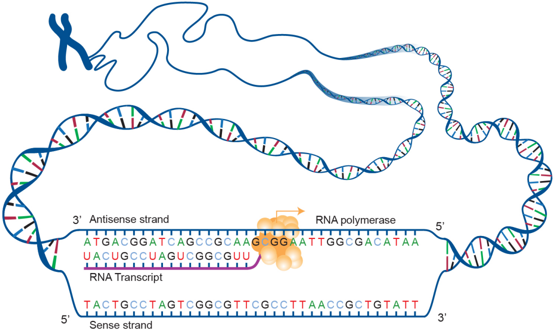
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Updated Annotation Release 109.20190905 (2019-09-10)

[7] UCSC Genome Browser

What is Transcription?

Transcription

Transcription is the process of making an RNA copy of a gene sequence[8]



[8] https://www.genome.gov/genetics-glossary/Transcription?id=197

The Common Steps of Transcription

1. RNA polymerase, together with one or more transcription factors, **binds to the promoter**
2. RNA polymerase creates a transcription bubble, which **separates the two strands of the DNA helix**, breaking the hydrogen bonds between complementary DNA nucleotides.
3. RNA polymerase **adds RNA nucleotides** complementary to the nucleotides of one DNA strand.
4. RNA sugar-phosphate backbone forms to create **single-stranded RNA**.
5. Hydrogen bonds of the RNA–DNA complex break, **freeing the newly synthesized RNA strand**.

Additional Steps of Transcription

If the cell has a nucleus (i.e. a *eukaryotic cell*):

6. The RNA may be further processed. This may include **polyadenylation**, **capping**, and **splicing**.
7. The RNA may remain in the nucleus or **exit to the cytoplasm** through the nuclear pore complex.

Before transcription even starts, the relevant sections of DNA must be unpacked from any histones. (*Beyond the scope of this course.*)

Promoters

What is a Promoter?

Definition

A promoter is a region of DNA that leads to initiation of transcription of a particular gene. Promoters are located near the transcription start sites of genes, upstream on the DNA (towards the 5' region of the sense strand). [9]

[9] https://en.wikipedia.org/wiki/Promoter_(genetics)

Structure of a Promoter

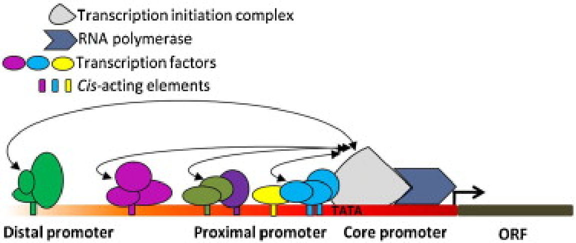


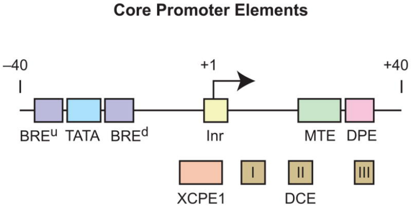
Figure taken from Hernandez-Garcia and Finer, "Identification and validation of promoters and cis-acting regulatory elements"

Key Elements of a Eukaryotic *mRNA* Promoter

- The **core promoter** is the minimal portion of the promoter required to properly initiate transcription.
  - Includes the transcription start site (TSS) and elements directly upstream
  - A binding site for RNA polymerase
  - General transcription factor binding sites, e.g. TATA box (TATAWAW[10]), B recognition element.
  - Many other elements/motifs may be present.
- **There is no set of universal elements found in every core promoter**

[10] W indicates either T or A

Key Elements of a Eukaryotic *mRNA* Promoter



- *Inr* = Initiator
- *DPE* = Downstream Core Promoter Element
- *BRE* = TFIIB Recognition Element
- *MTE* = Motif Ten Element
- *XPCE* = X Core Promoter Element 1
- *DCE* = Downstream Core Element

Figure taken from Juven-Gershon et al., "The RNA polymerase II core promoter - the gateway to transcription"

Key Elements of a Eukaryotic Promoter

- The **Proximal promoter** is the sequence upstream of the gene that tends to contain primary regulatory elements
  - Approximately 250 base pairs upstream of the transcription start site
  - Specific transcription factor binding sites (TFBS).
- Large diversity of TFBS in multiple combinations
- A very active area of research

Key Elements of a Eukaryotic Promoter

- The **Distal promoter** is the distal sequence upstream of the gene often with a weaker influence than the proximal promoter
  - Additional (weaker) regulatory elements,
  - Anything further upstream (but not an enhancer or other regulatory region whose influence is positional/orientation independent)
  - Specific transcription factor binding sites
- Often the entire region about -1500nt to +500nt is referred to as *the promoter region*. This includes distal, proximal and core elements

Promoters Function in 3-Dimensions

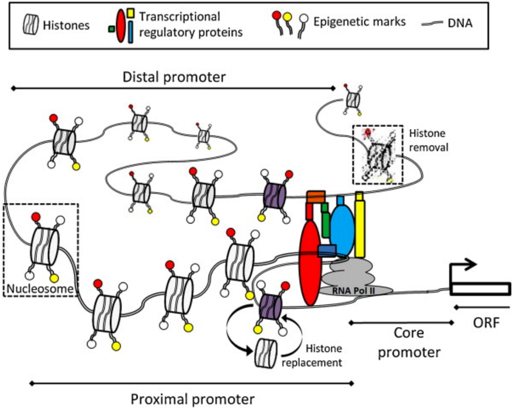


Figure taken from[11]

[11] T. Juven-Gershon et al. "The RNA polymerase II core promoter - the gateway to transcription". In: *Curr. Opin. Cell Biol.* 20.3 (2008), pp. 253–259.

Food For Thought

Given the definition of a gene as a unit of inheritance, could a promoter be considered as part of a gene?

Transcription

RNA Polymerase

When all Transcription Factors and other molecules are bound to a promoter, RNA Polymerase is recruited to complete the Transcriptional Complex and initiates transcription

- Prokaryotes
  - One form of RNAP[12] (5 subunits)
- Eukaryotes[13]
  - RNA Polymerase I: pre-*rRNA*
  - RNA Polymerase II: most *mRNA*, *miRNA* and *snRNA*
  - RNA Polymerase III: *tRNA*, 5S *rRNA*, other small RNA
- Plants
  - RNA polymerase IV/V: *siRNA* and RNAs involved in *siRNA*-directed heterochromatin formation

[12] Also very similar to that used in chloroplasts

[13] *POLMRT* is a nuclear-encoded single subunit RNA polymerase, which transcribes mitochondrial genes in many eukaryotes. More closely related to bacteriophage RNAP than nuclear RNAP

The Core Steps of Transcription

An old, but helpful animation for *mRNA* synthesis:

`https://youtu.be/J3HVVi2k2No?t=49`

(Stop at around 7:05)

Messenger RNA

Messenger RNA (*mRNA*)

- Many *mRNA* encode proteins, but many don't
- Nuclear *mRNA* are always processed immediately after (or during) transcription
- Nuclear *mRNA* have a 5' cap added[14]
  - Protects *ssRNA* from degradation
  - Regulates nuclear export
  - Promotes translation
- *mRNA* are **always** polyadenylated at the 3' end
- Are commonly spliced to remove introns

[14] Mitochondrial and chloroplastic *mRNA* are not capped

Splicing of Transcribed RNA

- RNA is transcribed from the *antisense* strand in a continuous molecule, known as a *pre-mRNA*
- Some regions within the *pre-mRNA*, known as *introns* will be removed via a mechanism known as *splicing*
- The remaining sequences (*exons*) are then joined together to form the mature *mRNA* (with 5' cap and 3' poly-A tail)

Splicing of Transcribed RNA

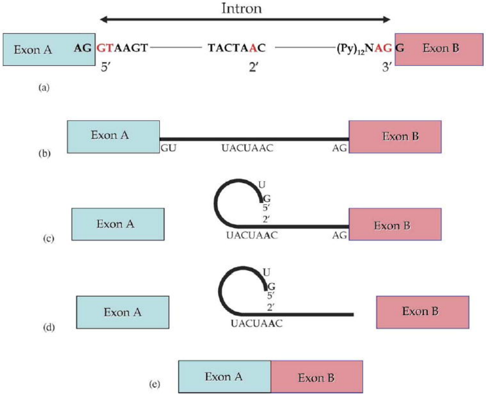


Figure taken from Wessagowit et al., "Normal and abnormal mechanisms of gene splicing and relevance to inherited skin diseases"

Alternate Transcripts and Isoforms

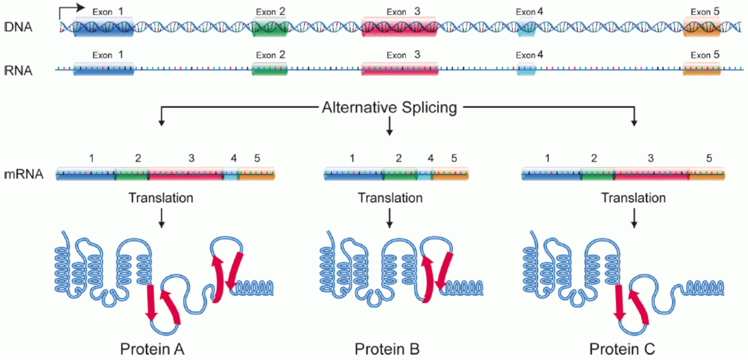
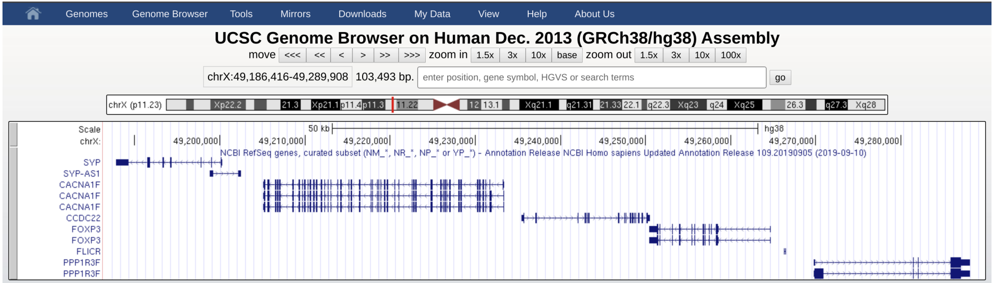


Image by National Human Genome Research Institute, via https://commons.wikimedia.org/wiki/File:DNA_alternative_splicing.gif

THE UNIVERSITY of ADELAIDE

Alternate Transcripts and Isoforms



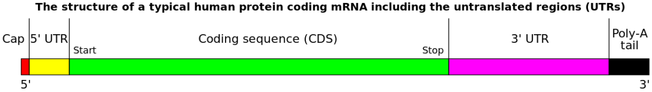
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chrX:49,186,416-49,289,908 103,493 bp.

NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Updated Annotation Release 109.20190905 (2019-09-10)

Let's also check *PSEN1* using the Ensembl data base

Non-coding Regions of an *mRNA*


The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
5'
3'

- The 5' cap and 3' polyA tail both 1) aid nuclear export, and 2) influence *mRNA* stability
- Both the 3'UTR and 5'UTR can contain secondary RNA structures
- The 5'UTR can regulate translation
- The 3'UTR can also impact *mRNA* stability (2° structure, *miRNA* binding sites)

Image sourced from https://commons.wikimedia.org/wiki/File:MRNA_structure.svg

Non-Coding RNA

Ribosomal RNA (*rRNA*)

- Ribosomes: Where translation from *mRNA* to *protein* takes place
- *rRNA* molecules make up the major ribosomal building blocks
- *rRNA* makes up about 80% of cellular RNA in eukaryotes!
- *rRNA* sequences are highly conserved between species
- Eukaryotes:
  - Cytoplasmic: *5S*, *5.8S*, *18S*, *28S*
  - Mitochondrial: *12S* and *16S*
- Prokaryotes
  - *5S*, *16S*, *23S*
  - *16S* is commonly used in *metagenomics*

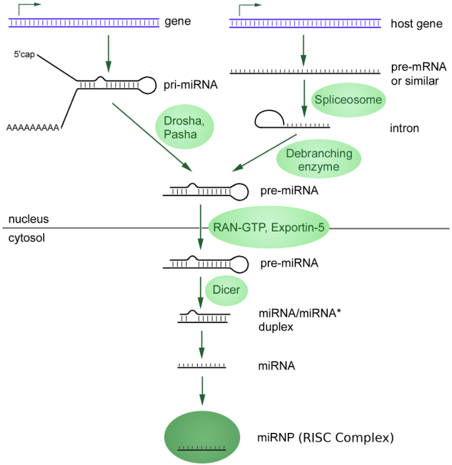
Transfer RNA (*tRNA*)

- There are 61 different types, each corresponding to a *codon* (nucleotide triplet)
- *tRNA* molecules catalyse addition of each new amino-acid during translation from *mRNA* to protein
- Each *tRNA* has a single amino-acid attached which elongates the protein one *aa* at a time
- Can undergo modification (e.g. methylation), although this is still poorly understood

micro-RNA (*miRNA*)

- Are the next most commonly studied after *mRNA*
- Can be coded as individual genes, in clusters, or within introns
- Mature *miRNA* are are about 22*nt* long
  - Begin as *primary miRNA* (*pri-miRNA*)
  - Are edited to become *precursor-miRNA* (*pre-miRNA*)
  - Processing is mainly by the RNAse proteins Drosha and Dicer
- Mostly interact with the 3'UTR of *mRNA* to suppress translation (via mRNA degradation)
  - also with 5'UTR, promoters, CDS
- Can be in cytoplasm, nucleus, external to cells, etc

miRNA Biogenesis



Modified from `https://commons.wikimedia.org/wiki/File:MiRNA-biogenesis.jpg`

Long non-coding RNA (*lncRNA*)

- $>200nt$ long
- Mainly expressed in a tissue specific manner
- Can overlap protein coding genes, $\implies$ intergenic *lncRNA* are *lincRNA*
- Currently estimated to be >270,000 *lncRNA*[15]
- Often expressed at very low-levels
- Most are poorly understood, but can play key epigenetic role

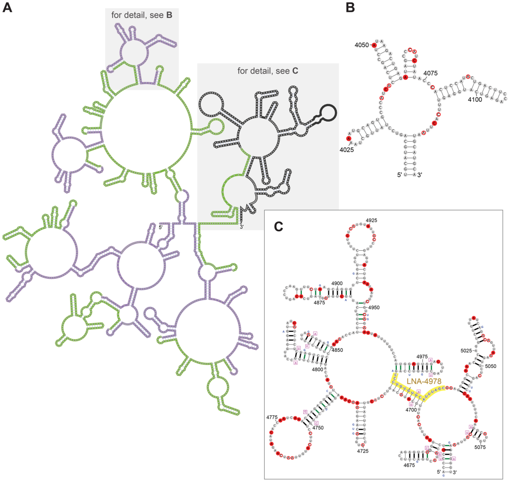
[15] L. Ma et al. "LncBook: a curated knowledgebase of human long non-coding RNAs". In: *Nucleic Acids Res.* 47.D1 (2019), pp. D128–D134.

Long non-coding RNA (*lncRNA*)

The best understood *lncRNA* is *Xist*

- 17kb RNA transcript
- Very complex 2° structure
- Key regulator of *X*-inactivation
- Expressed from the inactive *X* chromosome and ‘coats’ it
- Also has an antisense version *Tsix* which forms an interactive regulatory loop with *Xist*

Xist



A 2*kb* region of *Xist*. Taken from Fang et al., "Probing Xist RNA Structure in Cells Using Targeted Structure-Seq"

Other *ncRNA*

- Piwi-interacting RNA (*piRNA*)
  - 26-31*nt* in length
  - Biogenesis poorly understood, often found in genomic clusters
  - Form complexes with Argonaute proteins
  - Involved in **silencing of transposons**
- Small nuclear RNA (*snRNA*)
  - Usually about 150*nt* long
  - Always found with nuclear proteins (*snRNPs*)
  - Involved in processing *pre-mRNA* (e.g. splicing)
- Small nucleolar RNA (*snoRNA*)
  - Mainly guide modification of *tRNA*, *rRNA* and *snRNA*

Summary

- Transcriptomics can be the study of any or all of these molecules
- Most common is analysis of *mRNA* $\implies$ primarily quantitative
- Always looking to understand the wider biological processes in a dynamic system