Long non-coding RNAs annotation in *Strongylocentrotus purpuratus*. Studying the need of a reference genome.

Estudiante: Sonia Doblado Martín
Directora: Cinta Pegueroles Queralt
Profesor de la asignatura: Marc Maceira Duch

¿Qué son los lncRNAs?

Los lncRNAs (long non-coding RNAs) son RNAs celulares endógenos
que se asume que no codifican para proteínas, y con una longitud mayor de 200pb.

¿Cómo se anotan?

A partir de un genoma de referencia

¿Y si no hay genoma de referencia, se pueden anotar?

¿Por qué S. purpuratus?

-Tiene genoma de referencia
-Existen datos de experimentos RNA-Seq previos
-Ya tiene lncRNAs anotados

-Los lncRNAs suelen estar anotados en vertebrados u organismos modelo

Objetivos generales

1. Obtención de un transcriptoma *de novo* utilizando los experimentos RNA-Seq sobre la especie disponibles en bases de datos públicas.

2. Anotación lncRNAs a partir del *de novo assembly* usando dos metodologias: predicción de *coding potential* y predicción de lncRNAs vía *machine learning*.

3. Comparación de resultados de la anotación de lncRNAs obtenidos en este trabajo con los obtenidos mediante *pipelines* utilizadas en experimentos previos.

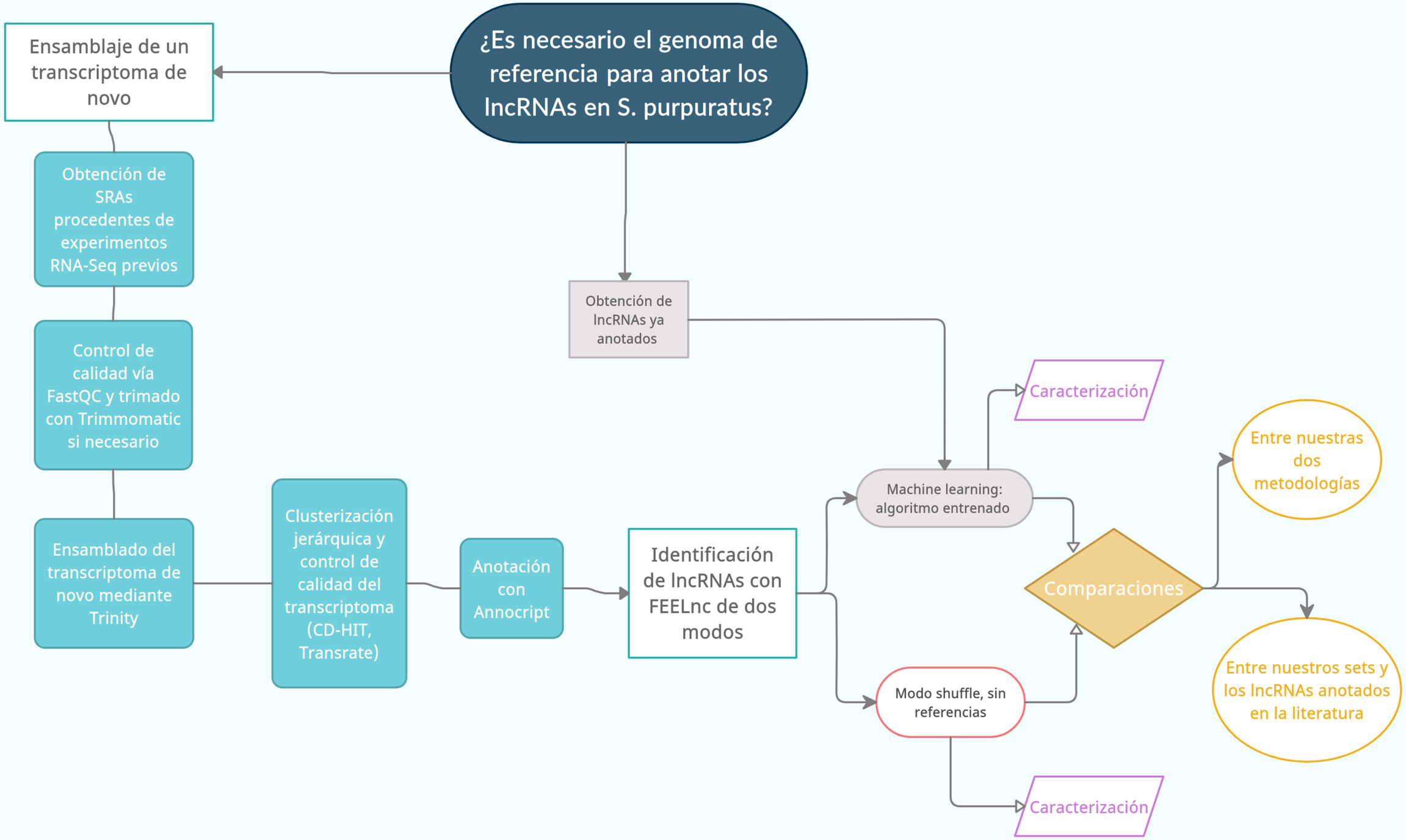
Objetivos específicos

1) Predecir el *coding potential* de las regiones anotadas *de novo*.

2) Utilizar los lncRNAs ya anotados en *S. purpuratus* como *training set*, estudiar la posible predicción de secuencias que constituyan lncRNAs en el genoma completo de *S. purpuratus*.

3) Comparativa de los lncRNAs anotados con las dos metodologías y con los lncRNAs en la literatura mapeando contra un genoma de referencia.

4) Caracterizar y validar la funcionalidad de los lncRNAs.


¿Es necesario el genoma de referencia para anotar los lncRNAs en S. purpuratus?
Ensamblaje de un transcriptoma de novo
Obtención de SRAs procedentes de experimentos RNA-Seq previos
Control de calidad vía FastQC y trimado con Trimmomatic si necesario
Ensamblado del transcriptoma de novo mediante Trinity
Clusterización jerárquica y control de calidad del transcriptoma (CD-HIT, Transrate)
Anotación con Annocript
Identificación de lncRNAs con FEELnc de dos modos
Obtención de lncRNAs ya anotados
Machine learning: algoritmo entrenado
Modo shuffle, sin referencias
Caracterización
Caracterización
Comparaciones
Entre nuestras dos metodologías
Entre nuestros sets y los lncRNAs anotados en la literatura

Ensamblaje de un transcriptoma *de novo*

Descarga de datos RNA-Seq

Control de calidad con FastQC y MultiQC

Características de nuestro transcriptoma

Número de transcritos → 907236

Número de "genes" → 534770

%GC → 38,8%

Puntuaciones

Overall alignment rate (Bowtie2) → 83,66%

Transrate Assembly Score → 0,07

BUSCO V2/V3 – Euk	Trinity
Complete BUSCOs – C	94.72%
Complete and single-copy BUSCOs – S	45.5%
Complete and duplicated BUSCOs – D	49.2%
Fragmented BUSCOs – F	5.3%
Missing BUSCOs – M	0,00 %
Total BUSCO groups searched	303

Características de nuestro transcriptoma

Número de transcritos → 90722

Número de "genes" → [illegible]70

%GC → 38,8%

Puntuaciones

Overall alignment rate (Bowti[illegible] → 83,66%

Transrate Ass[illegible] Sc[illegible] → 0,08

BUSCO V2	Trinity
Complete BU[illegible]	94.72%
Complete and [illegible]USCOs – S	45.5%
Complete and d[illegible]USCOs – D	49.2%
Fragmented BUSCO[illegible] – F	5.3%
Missing BUSCOs – M	0,00 %
Total BUSCO groups searched	303

Anotación de lncRNAs: FEELnc

Eliminación de regiones codificantes vía BLASTn

14844 transcritos resultantes candidatos a lncRNAs

Anotación de lncRNAs: FEELnc

Eliminación de regiones codificantes vía BLASTn

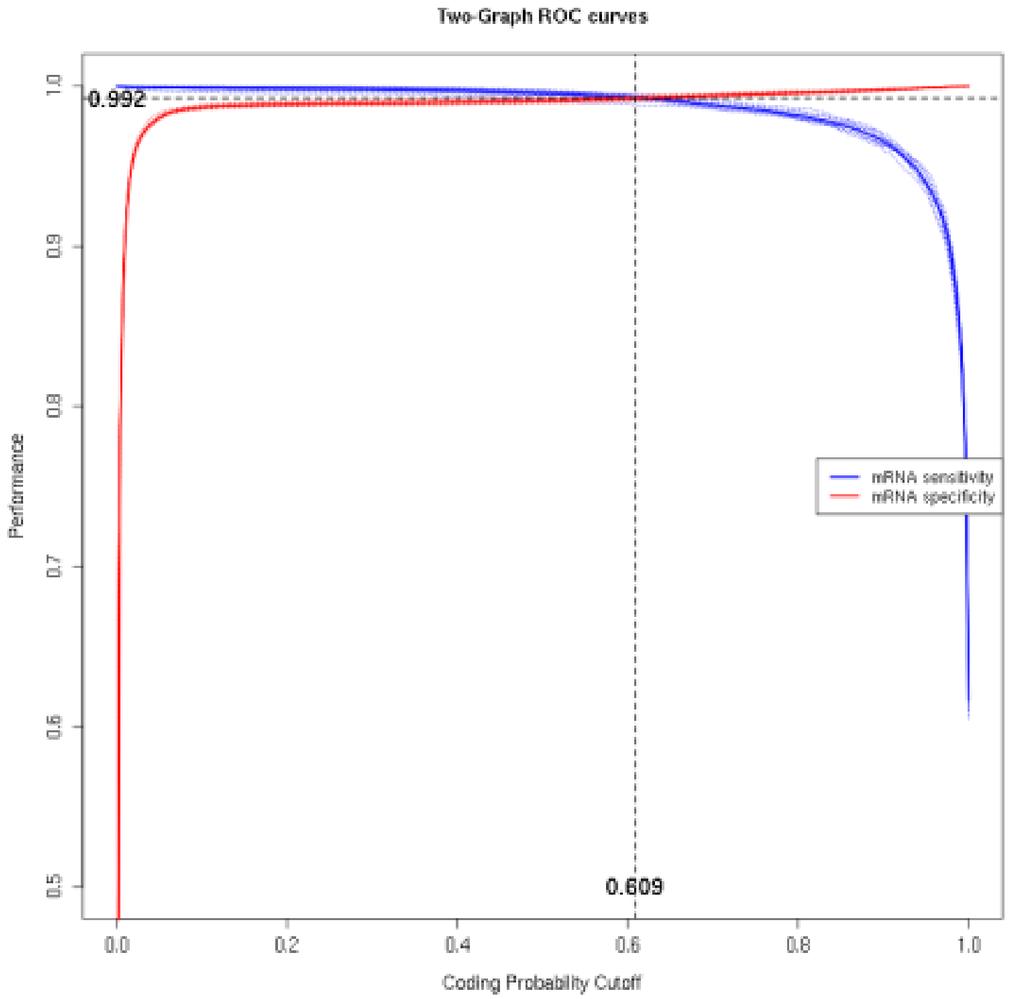
14844 transcritos resultantes candidatos a lncRNAs

Shuffle

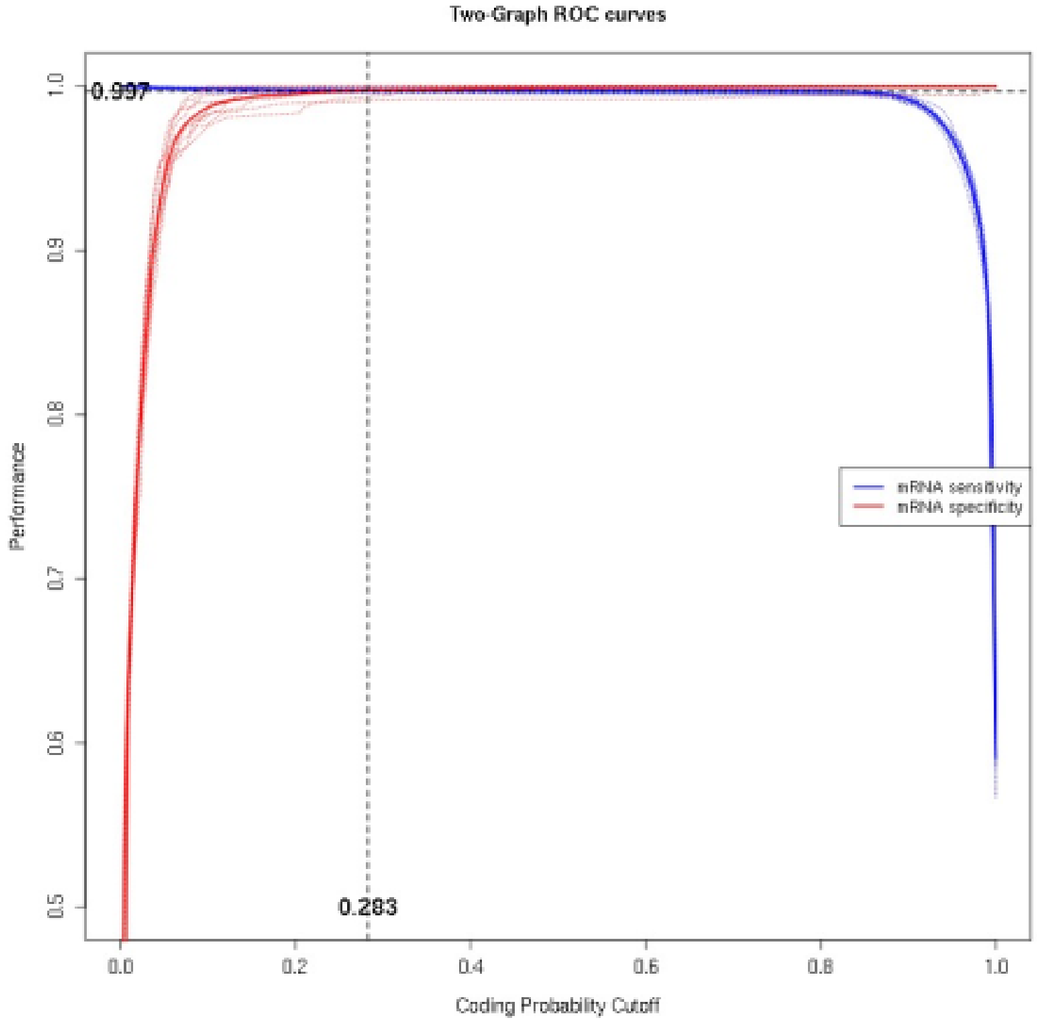
Algoritmo RF entrenado

Anotación de lncRNAs: FEELnc

Sólo con transcriptoma *de novo*



Con algoritmo entrenado



Caracterización

Protein ID	RNA ID	Z-score	Discriminative Power (%)	Interaction Strength (%)	Domain	Motif	Ranking
NP_999787.1	TRINITY_DN41917_1	-0.18	45	84	no	no	
NP_999744.1	TRINITY_DN41917_1	-0.45	22	50	yes	no	
NP_001123287.1	TRINITY_DN41917_1	-0.53	20	32	yes	no	
NP_001075435.1	TRINITY_DN41917_1	-0.57	17	30	yes	no	
NP_999704.1	TRINITY_DN41917_1	-0.66	14	16	yes	no	
NP_001123279.1	TRINITY_DN41917_1	-0.76	14	8	yes	no	
NP_999824.2	TRINITY_DN41917_1	-0.83	14	8	yes	no	
NP_001229606.1	TRINITY_DN41917_1	-0.79	14	5	yes	no	
NP_001005725.1	TRINITY_DN41917_1	-0.83	14	5	yes	no	
NP_001124190.1_826-877	TRINITY_DN41917_1	-0.84	14	3	yes	no	

Protein ID	RNA ID	Z-score	Discriminative Power (%)	Interaction Strength (%)	Domain	Motif	Ranking
NP_999787.1	TRINITY_DN41917_1	-0.16	47	85	no	no	
NP_999744.1	TRINITY_DN41917_1	-0.43	22	53	yes	no	
NP_001123287.1	TRINITY_DN41917_1	-0.49	20	36	yes	no	
NP_001075435.1	TRINITY_DN41917_1	-0.54	17	34	yes	no	
NP_999704.1	TRINITY_DN41917_1	-0.64	14	18	yes	no	
NP_999824.2	TRINITY_DN41917_1	-0.81	14	9	yes	no	
NP_001123279.1	TRINITY_DN41917_1	-0.75	14	8	yes	no	
NP_001005725.1	TRINITY_DN41917_1	-0.80	14	6	yes	no	
NP_001229606.1	TRINITY_DN41917_1	-0.78	14	5	yes	no	
NP_001124190.1_826-877	TRINITY_DN41917_1	-0.83	14	4	yes	no	

UnipProt entry	Función molecular	Proceso biológico
A2PZA6_STRPU	DNA-binding	Transcripción, regulación de la transcripción
Q26649_STRPU	DNA-binding	Regulación de la transcripción
B3FNS0_STRPU	DNA-binding, ecdysone binding	Transcripción, regulación de la transcripción, receptor en ruta de la ecdisona
A0A0B4J2U9_STRPU	Quinasa, Transferasa	Procesos metabólicos de ADP, AMP, GTP, ITP, purina
B3FNR8_STRPU	DNA-binding	Transcripción, regulación de la transcripción
B3VCG6_STRPU	Binding de ácidos nucleicos	
Q64HK6_STRPU	DNA-binding, RNA polimerasa II	Regulación positiva de la transcripción
Q9U0E3_STRPU		Movilidad celular por medio de cilios
O77156_STRPU	DNA-binding, RNA polimerasa II	Regulación positiva de la transcripción
Q6S5K1_STRPU	DNA-binding	Diferenciación celular, Regulación positiva de la transcripción

Comparamos

Entre nuestras metodologías

	Modo shuffle	Algoritmo RF entrenado
Cutoff	0,6086	0,2826
lncRNAs	14789	14637
mRNAs	27	179

Nuestros resultados contra resultado con genoma de referencia

9005 lncRNAs de estudio previo*

↓

BLASTn

↓

3750 de esos lncRNAs encontrados en nuestro set

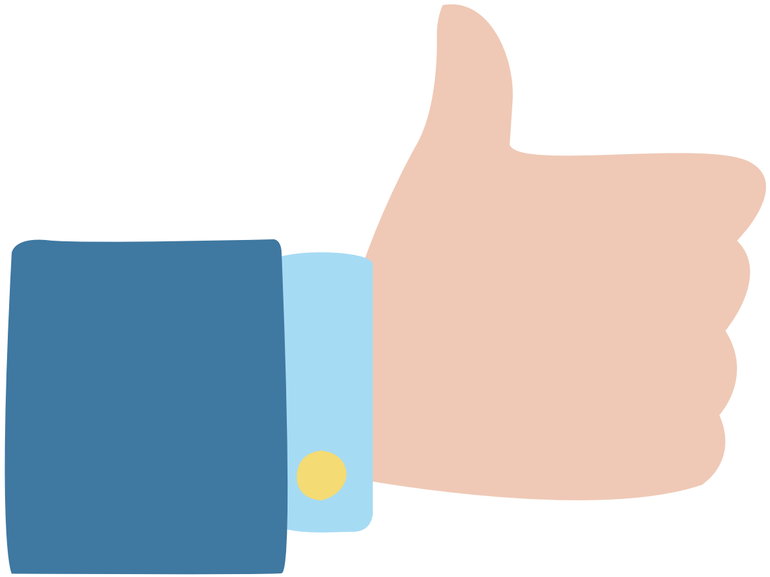
↓

41,6% de lncRNAs ya anotados presentes en nuestro transcriptoma

*https://linkinghub.elsevier.com/retrieve/pii/S2211124715004106

CONCLUSIONES

Sí sería posible anotar lncRNAs de una especie sin disponer de genoma de referencia a partir del ensamblaje *de novo* de su transcriptoma.



Sí sería posible anotar lncRNAs de una especie sin disponer de genoma de referencia a partir del ensamblaje *de novo* de su transcriptoma.

La clave estaría en conseguir un transcriptoma *de novo* de alta calidad.

Objetivos generales

1. Obtención de un transcriptoma *de novo* utilizando los experimentos RNA-Seq sobre la especie disponibles en bases de datos públicas.

2. Anotación lncRNAs a partir del *de novo assembly* usando dos metodologias: predicción de *coding potential* y predicción de lncRNAs vía *machine learning*.

3. Comparación de resultados de la anotación de lncRNAs obtenidos en este trabajo con los obtenidos mediante *pipelines* utilizadas en experimentos previos.

Objetivos específicos

1) Predecir el *coding potential* de las regiones anotadas *de novo*.

2) Utilizar los lncRNAs ya anotados en *S. purpuratus* como *training set*, estudiar la posible predicción de secuencias que constituyan lncRNAs en el genoma completo de *S. purpuratus*.

3) Comparativa de los lncRNAs anotados con las dos metodologías y con los lncRNAs en la literatura mapeando contra un genoma de referencia.

4) Caracterizar y validar la funcionalidad de los lncRNAs.

FIN