

**DESARROLLO DE UNA PLATAFORMA BIOINFORMÁTICA PARA LA
OBTENCIÓN DE INFORMACIÓN A PARTIR DE LIBRERÍAS DE ARNS
PEQUEÑOS SECUENCIADAS CON SOLEXA/ILLUMINA**

**INGENIERO EN BIOINFORMÁTICA
LEONARDO IVÁN ALMONACID CÁRDENAS**

RESUMEN

El uso de las nuevas tecnologías de secuenciación posee una tasa de crecimiento baja, en comparación con otras tecnologías, por ejemplo, los microarreglos. Este lento crecimiento se debe a que la tecnología es reciente y diferente, por ende, incomprendida en muchos aspectos. Además, el análisis de los datos obtenidos requiere de cierta experiencia Bioinformática, área que en general, la comunidad científica no posee muy desarrollada. Por otro lado, las herramientas existentes son altamente específicas y complejas de utilizar, lo que desincentiva aun más su uso. Aquí se plantea la implementación de una plataforma que permita extraer información a partir de la secuenciación con las nuevas tecnologías, mediante la intersección con lo anotado, además de realizar un análisis comparativo de los niveles de expresión de ARNs pequeños, junto con la detección de zonas transcripcionalmente activas. A partir de la secuenciación se obtuvieron 2 librerías denominadas Dorsal y Ventral. Ellas fueron filtradas y mapeadas al genoma de *Xenopus tropicalis* obteniendo un total de 9.737.900 lecturas para el conjunto Dorsal y 7.815.224 lecturas para el Ventral. Junto con el análisis de ellas se desarrolló una plataforma Bioinformática que permitió extraer grandes cantidades de información: perfiles de expresión para cada librería, distribución de tamaño de los ARNs pequeños, frecuencia genómica de los ARNs mapeados, estadística de la procedencia de los ARNs mapeados al genoma, intersección de secuencias entre las librerías, patrones de expresión diferencial, lista de genes con más ARNs mapeados. Se encontraron, *in silico*, un conjunto de ARNs pequeños que tendrían algún rol en el desarrollo de *X. tropicalis*. Además se estableció mediante dos enfoques, un conjunto de ARNs pequeños que muestran un patrón de expresión diferencial entre las dos librerías. Resumiendo, los datos obtenidos son una prueba que plataformas como éstas harán que un mayor número de científicos usen las nuevas tecnologías de secuenciación, considerando que ellas facilitan el

análisis de los datos, además de hacerlos más robustos a través de la integración con más información.

ABSTRACT

The use of next generation sequencing technologies (NGS) has a lower growth rate compared to other technologies such as microarrays. This slow growth is because the technology is recent and different, therefore, not fully understood in many ways. In addition, analysis of the data requires some Bioinformatics experience, which the scientific community not always has. Moreover, existing tools are highly specific and complex to use, which discourages its use even more. In this thesis the implementation of a platform to extract information from NGS data through the intersection with the annotation of a genome was developed, in addition to a comparative analysis of expression levels of small RNAs, together with the detection of highly transcriptionally active zones. Two libraries called Dorsal and Ventral were sequenced. They were filtered and mapped to *Xenopus tropicalis* genome, obtaining a total of 9.737.900 reads in the Dorsal set and 7,815,224 reads for the Ventral. Together with the analysis of them, a bioinformatics platform was developed that allowed the extraction of large amounts of data: expresión profiles for each library, the size distribution of small RNAs, mapped RNAs genome frequencies, statistics of the source of RNAs mapped to the genome, sequences intersection between libraries, differential expression patterns, gene list of mapped RNAs. It was found, *in silico*, a set of small RNAs that might have a role in the development of *X. tropicalis*. Also, some small RNAs that showed a differential expression pattern between the two libraries, according to two different approaches, were identified. In summary, the data obtained is a prove that platforms like these will make that a greater number of scientists use NGS, considering that they facilitate the analysis of data as well as make them more robust by integrating more information.