

TABLA DE CONTENIDOS

	página
Agradecimientos	I
Tabla de Contenidos	III
Índice de Figuras	VII
Índice de Tablas	VIII
Resumen	IX
Abstract	I
1. Introducción	1
1.1. Presentación del Problema	1
1.2. Marco Teórico	2
1.2.1. <i>Vibrio parahaemolyticus</i>	2
1.2.2. <i>Vibrio parahaemolyticus</i> pandémico.	4
1.2.3. Análisis de Secuenciación masiva y estudio genómico.	5
1.2.4. RNA-Seq	5
1.2.5. Expresión diferencial de mRNA en <i>Vibrio parahaemolyticus</i>	8
1.3. Justificación del Problema	10
2. Hipótesis	12
2.1. Hipótesis	12
3. Objetivos	13
3.1. Objetivo General	13
3.1.1. Objetivos específicos	13
4. Materiales y Métodos	14
4.1. Cultivo de bacterias y obtención de RNA.	14
4.2. Preparación de Biblioteca.	15
4.3. Visualización de la calidad de la biblioteca (FastQC).	16
4.4. Selección aleatoria de lecturas de la biblioteca.	18
4.5. Filtrado de lecturas (BBDuk).	19

4.6.	Alineamiento de lecturas (BWA).	22
4.7.	Cambio de formato de archivos de alineamiento SAM a BAM y ordenamiento de BAMs.	25
4.8.	Visualización de alineamientos (IGV).	27
4.9.	Transformación de GFF del genoma de referencia anotado a un formato compatible con HTSeq-count (GTF).	28
4.10.	Abundancia total de transcritos (HTSeq-count).	31
4.11.	Expresión relativa de genes.	34
4.11.1.	Estudio de RPKM.	34
4.11.2.	Coefficiente de determinación R^2 .	35
4.12.	Creación de matriz de lecturas.	36
4.13.	Expresión diferencial.	36
4.14.	Obtención términos GO.	38
4.15.	Implicancia biológica.	41
5.	Resultados	43
5.1.	Trabajo preliminar con Biblioteca.	43
5.1.1.	Preparación de Biblioteca.	43
5.1.2.	Análisis de calidad.	44
5.1.3.	Selección aleatoria de lecturas.	46
5.1.4.	Filtrado de lecturas.	46
5.1.5.	Alineamiento de lecturas y refinamiento de archivos.	47
5.1.6.	Visualización de alineamientos.	48
5.1.7.	Transformación de GFF a un formato GTF compatible con HTSeq-count.	49
5.2.	Abundancia total de transcritos.	50
5.3.	Expresión relativa de genes.	55
5.3.1.	RPKMO	55
5.3.2.	Coefficiente de determinación(R^2)	55
5.4.	Desarrollo de Matriz de Conteo.	56
5.5.	Expresión Diferencial.	56
5.5.1.	Genes expresados diferencialmente.	56
5.5.2.	Obtención de términos GO.	57
5.6.	Implicancia biológica.	58
5.6.1.	Datos estadísticos generales.	58
5.6.2.	Expresión del genoma en cada condición.	58

5.6.3.	Regulación de los genes altamente expresados.	59
5.6.4.	Genes con mayor expresión diferencial.	59
5.6.5.	Sistemas de secreción	60
5.6.6.	Expresión diferencial general (Gene Ontology)	61
6.	Discusión	63
6.1.	Análisis preliminar de la biblioteca	63
6.1.1.	Análisis de calidad.	63
6.1.2.	Selección aleatoria de lecturas.	64
6.1.3.	Filtrado de lecturas.	64
6.1.4.	Alineamiento de lecturas y refinamiento de archivos.	64
6.2.	Abundancia total de transcritos (Conteo de lecturas).	65
6.3.	Expresión relativa de genes.	67
6.3.1.	RPKMO.	67
6.3.2.	Coefficiente de determinación.	67
6.4.	Implicancia biológica.	68
6.4.1.	Regulación de los genes altamente expresados.	68
6.4.2.	Genes con mayor expresión diferencial.	69
6.4.3.	Sistemas de secreción.	71
6.4.4.	Expresión diferencial general (Gene Ontology).	72
7.	Conclusión	74
8.	Anexos	76
8.1.	Anexos de comandos.	76
8.1.1.	Anexo 1: Selección aleatoria de lecturas.	76
8.1.2.	Anexo 2: Filtros de calidad.	77
8.1.3.	Anexo 3: Alineamiento de lecturas	79
8.1.4.	Anexo 4: Transformación de SAM a BAM	80
8.1.5.	Anexo 5: Desarrollo de archivo GTF	81
8.1.6.	Anexo 6: Abundancia de lecturas	82
8.1.7.	Anexo 7: Expresión relativa de genes.	82
8.1.8.	Anexo 8: Expresión diferencial.	83
8.2.	Anexos de scripts.	83
8.2.1.	Script 1: Creación de GTF	83
8.2.2.	Script 2: Calculo de RPKMO	84
8.2.3.	Script 3: Script Expresión diferencial	87

8.3. Anexo: Documentos adjuntados. 89

Bibliografía **90**

ÍNDICE DE FIGURAS

	página
4.1. Modos de conteo que utiliza HTSeq-count.	33
5.1. Per sequence quality scores muestra I1	44
5.2. Per sequence quality scores muestra E1	45
5.3. Esquema de flujo de archivos obtenidos.	48
5.4. Alineamiento mostrado en interfaz IGV.	49
5.5. Porcentaje de lecturas de modo Union con calidad 10 clasificadas por HTSeq-Count.	51
5.6. Porcentaje de lecturas de modo Union con calidad 10 clasificadas por tipo de RNA.	51
5.7. Clasificaciones de genes expresados diferencialmente con expresión aumen- tada en la condición I asociados a Ontología: Proceso Biológico.	62
5.8. Clasificaciones de genes expresados diferencialmente con expresión dismi- nuida en la condición I asociados a Ontología: Proceso Biológico.	62
6.1. Formula de Phred Score.	63
6.2. Calidad de subconjuntos de lecturas mostradas en FastQC.	65

ÍNDICE DE TABLAS

	página
1.1. Especificaciones del genoma de <i>Vibrio parahaemolyticus</i> entregadas por la base de datos RefSeq de NCBI.	5
4.1. Condiciones estudiadas de <i>V. parahaemolyticus</i>	15
5.1. Cantidad de lecturas crudas obtenidas de BGI	44
5.2. Cantidad de lecturas obtenidas en la selección aleatoria.	46
5.3. Cantidad de lecturas obtenidas luego de los filtros.	47
5.4. Genes anotados de <i>Vibrio parahaemolyticus</i> en el GTF.	49
5.5. Cantidades de lecturas clasificadas por HTSeq-count para el modo Intersection-Strict	52
5.6. Cantidades de lecturas clasificadas por HTSeq-count para el modo Union .	53
5.7. Cantidades de lecturas clasificadas por HTSeq-count para el modo Interception-Nonempty	54
5.8. Rango de RPKMO para cada tipo de RNA de cada muestra.	55
5.9. Coeficiente de determinación entre muestras.	56
5.10. Cantidad de genes en <i>Vibrio parahaemolyticus</i>	58
5.11. Largo (bp) de cada tipo de RNA.	58
5.12. Porcentaje del genoma expresado para cada condición.	59
5.13. Genes expresados diferencialmente agrupados por vecindad.	59
5.14. Genes con expresión diferencial más alta	60
5.15. Genes de TSS expresados en condición aumentada y disminuida.	61