

Tabla de Contenido

1. Introducción	1
1.1. Antecedentes	1
1.2. Motivación	2
1.3. Descripción General de la Solución	4
1.4. Objetivos	5
2. Marco Teórico	7
2.1. Genética de poblaciones	7
2.2. Variant Calling Format	7
2.2.1. Meta-información	8
2.2.2. La Referencia	8
2.2.3. Líneas de datos	9
2.2.4. Recuperación de variantes simples	10
2.2.5. Variantes complejas, Rearrangements	11
2.2.6. Referenciación de variantes en individuos	13
2.3. Relative Lempel-Ziv	14
2.3.1. Factorización de Lempel-Ziv	14
2.3.2. Sets comprimidos de enteros	15
2.3.3. Formato de compresión RLZ	15
2.3.4. Autoíndice a utilizar	15
2.4. Trabajo relacionado	16

2.4.1.	Reducciones de tamaño	17
2.4.2.	Eficiencia en tiempos de consulta	17
3.	Diseño de la solución	19
3.1.	Consideraciones generales	19
3.1.1.	Alcance dentro de VCF	19
3.1.2.	Librerías a utilizar	20
3.2.	Planteamiento del problema	20
4.	Proceso de conversión e interpretación	22
4.1.	Procesamiento de la Referencia	22
4.2.	Transformación de edits a frases	23
4.2.1.	Recuperación de individuos	23
4.2.2.	Frases y su construcción	24
4.3.	Ordenamiento de las frases	26
4.4.	De frases a factores	27
4.5.	De RLZ a posiciones VCF	28
5.	Experimentación y Análisis	29
5.1.	La conversión	29
5.1.1.	Validación	29
5.1.2.	Resultados y análisis	30
5.2.	Indexación completa	32
5.2.1.	Dificultades con RLZ y soluciones	32
5.2.2.	Validación	33
5.2.3.	Resultados y análisis	34
5.3.	Propuestas de mejora	37
6.	Ejemplo de uso	38
6.1.	Indexación	38

6.2. Búsqueda de patrones	41
Conclusión	43
Bibliografía	45