



UNIVERSIDAD DE MÁLAGA

Proyecto Final

Estándares de Datos Abiertos e Integración de Datos

Autores

Achraf Ousti El Moussati
Sebastián Rozenblum
Anabel Yu Flores Moral
Gabriela Milenova Yordanova
Karen Michell Herrera Sierra

Grado en Ingeniería de la Salud
E.T.S.I. Ingeniería Informática
Universidad de Málaga

Curso 2025/2026

Abstract

HOLA. PRUEBA.

Keywords: (e.g. Mobile computing, security, complexity, algorithms, image analysis, machine learning, information retrieval)

Contents

Abstract	i
Contenidos	ii
1 Introducción	1
1.1 Diseño de una Base de Datos NoSQL para Genómica del Melanoma Acral	1
1.1.1 Colección: <code>patients</code>	1
1.1.2 Colección: <code>samples</code>	2
1.1.3 Colección: <code>variants</code>	3
2 Literature Review	4
3 Methodology	5
4 Results	6

Chapter 1

Introducción

1.1 Diseño de una Base de Datos NoSQL para Genómica del Melanoma Acral

En el campo de la bioinformática, la gestión eficiente y el análisis de grandes volúmenes de datos genómicos y clínicos son fundamentales para avanzar en la comprensión y el tratamiento de enfermedades como el **cáncer**. Este proyecto se centra en el diseño y la **implementación de una base de datos NoSQL**, utilizando **MongoDB**, para albergar y organizar los datos del estudio *Acral Melanoma (TGEN, Genome Res 2017)*. Este conjunto de datos, disponible a través del **cBioPortal for Cancer Genomics**, ofrece una rica fuente de información genómica y clínica de **pacientes con melanoma acral**, un subtipo raro y agresivo de melanoma (cáncer de piel).

Para estructurar la información de manera coherente y facilitar consultas complejas, se ha optado por un diseño que consta de **tres colecciones interconectadas**: **patients**, **samples** y **variants**. Esta estructura nos permitirá no solo capturar la información de cada paciente de forma individual, sino también trazar las relaciones entre los pacientes, las muestras biológicas obtenidas de ellos y las variantes genómicas identificadas en dichas muestras.

A continuación, se detalla la estructura propuesta para cada una de las colecciones:

1.1.1 Colección: **patients**

Esta colección almacena la información demográfica y clínica de cada paciente incluido en el estudio.

- **Nivel 1:** Información básica del paciente.
 - **patient_id**: Identificador único del paciente.
 - **sex**: Sexo del paciente.
 - **race_category**: Categoría racial del paciente.
 - **age_at_diagnosis**: Edad del paciente en el momento del diagnóstico.

- **Nivel 2:** Historial clínico y de tratamiento.
 - `clinical_history`: Objeto con información sobre el historial médico del paciente.
 - * `initial_diagnosis_date`: Fecha del diagnóstico inicial.
 - * `primary_tumor_site`: Localización del tumor primario.
 - `treatments`: Array de objetos que detalla los tratamientos recibidos.
 - * `treatment_type`: Tipo de tratamiento (e.g., "Ipilimumab", "Interferon").
 - * `start_date`: Fecha de inicio del tratamiento.
 - * `end_date`: Fecha de finalización del tratamiento.
- **Nivel 3:** Seguimiento y estado de la enfermedad.
 - `follow_up`: Objeto con información de seguimiento.
 - * `disease_free_months`: Meses libre de enfermedad.
 - * `disease_free_status`: Estado de la enfermedad (e.g., "0:DiseaseFree", "1:Recurred/Progressed").

1.1.2 Colección: `samples`

Contendrá información detallada sobre cada muestra biológica extraída de los pacientes.

- **Nivel 1:** Identificación y tipo de muestra.
 - `sample_id`: Identificador único de la muestra.
 - `patient_id`: Identificador del paciente al que pertenece la muestra (referencia a la colección `patients`).
 - `sample_type`: Tipo de muestra (e.g., "Primary", "Metastasis").
- **Nivel 2:** Detalles de la recolección y procesamiento.
 - `collection_info`: Objeto con detalles de la recolección.
 - * `collection_date`: Fecha de recolección de la muestra.
 - * `collection_method`: Método de recolección.
 - `processing_info`: Objeto con información del procesamiento.
 - * `processing_date`: Fecha de procesamiento.
 - * `sequencing_type`: Tipo de secuenciación realizada (e.g., "Whole Exome Sequencing").
- **Nivel 3:** Datos de análisis molecular.
 - `molecular_data`: Objeto que alberga datos moleculares.
 - * `mutation_count`: Número de mutaciones identificadas.
 - * `copy_number_alterations`: Array de objetos con información sobre alteraciones en el número de copias.

1.1.3 Colección: **variants**

Esta colección albergará la información específica de cada variante genómica identificada en las muestras.

- **Nivel 1:** Identificación de la variante.
 - **variant_id:** Identificador único de la variante.
 - **sample_id:** Identificador de la muestra en la que se encontró la variante (referencia a la colección **samples**).
 - **gene_symbol:** Símbolo del gen afectado.
- **Nivel 2:** Características de la variante.
 - **variant_details:** Objeto con las características de la variante.
 - * **chromosome:** Cromosoma donde se localiza la variante.
 - * **start_position:** Posición de inicio de la variante.
 - * **end_position:** Posición de finalización de la variante.
 - * **reference_allele:** Alelo de referencia.
 - * **alternate_allele:** Alelo alternativo.
- **Nivel 3:** Anotación funcional y predicciones.
 - **functional_annotation:** Objeto con la anotación funcional.
 - * **variant_classification:** Clasificación de la variante (e.g., "Missense_Mutation", "Nonsense_Mutation").
 - * **protein_change:** Cambio en la proteína resultante.
 - * **sift_prediction:** Predicción del impacto de la variante por SIFT.
 - * **polyphen_prediction:** Predicción del impacto de la variante por PolyPhen.

Este diseño proporciona una estructura robusta y flexible para el almacenamiento y consulta de los datos del estudio de melanoma acral, sentando las bases para futuros análisis que puedan desvelar nuevos conocimientos sobre esta enfermedad.

Chapter 2

Literature Review

Chapter 3

Methodology

Chapter 4

Results