

Práctica 1

1. **Contexto.** Explicar en qué contexto se ha recolectado la información.
Explicar por qué el sitio web elegido proporciona dicha información.
Indicar la dirección del sitio web.

La información que hemos recolectado usando el web scraping la hemos sacado de Nature Immunology, ya que al ser los dos Bioinformáticos, nos hemos encontrado con alguna situación en la que necesitamos extraer información de varios artículos científicos a la vez. Entonces hemos tenido la idea de automatizar la pre-selección de los artículos de interés para un campo específico de la bioinformática. Ya que los dos trabajamos en inmunología, nos hemos centrado en limpiar los últimos papers de campo, solamente para obtener artículos de inmunología computacional ya que los otros papers tienen poco valor útil para el dry lab.

Una de las maneras simples, pero efectivas de hacer una limpieza en NLP, es utilizar keywords. En nuestro caso, hemos seleccionado las keywords para el sistema inmune adaptativo, que contiene las T y B cells. Ya que queremos centrarnos en la parte computacional del ámbito, hemos añadido también términos como Machine Learning, Deep Learning ya que son muy utilizados en los últimos años en el ámbito.

La idea final es utilizar el dataset obtenido para el año 2022 (que si es necesario se puede ampliar el rango fácilmente) e filtrar usando los keywords para:

1. Ver el porcentaje de papers en el ámbito de inmunología que son computacionales
2. Ver el porcentaje de papers Open Access para ver cómo de accesible es la información del ámbito en Nature Immunology

2. **Título.** Definir un título que sea descriptivo para el dataset.

Filtraje de artículos de inmunología computacional de Nature Immunology, usando proxy based web-scraping en python.

3. **Descripción del dataset.** Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Con el web scraping obtenemos un dataset que contiene el título del artículo, un resumen, los autores, que accesibilidad tiene, la fecha de publicación, el link a la imagen representativa del estudio y el link al título entero y su correspondiente abstract.

Además, como ya hemos explicado anteriormente, hemos añadido una manera de buscar palabras de interés (keywords) que serán la herramienta de filtrado de los artículos obtenidos.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

nature immunology [View all journals](#) [Search](#) [Login](#)

[Explore content](#) [About the journal](#) [Publish with us](#) [Sign up for alerts](#) [RSS feed](#)

[nature](#) > [nature immunology](#) > research articles

Research articles

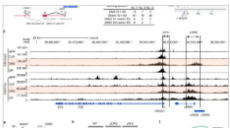
Article Type: **Article (76)** Year: **2022 (76)**

Article
31 Oct 2022

A double-negative thymocyte-specific enhancer augments Notch1 signaling to direct early T cell progenitor expansion, lineage restriction and β -selection

Notch signaling is required for T cell development. Georgopoulos and colleagues identify an enhancer that specifically boosts Notch1 expression in early thymic progenitors through the DN3 stage, expanding the less committed multipotent progenitors and preparing these cells for faithful lineage commitment.

Mariko Kashiwagi, Daniela Salgado Figueroa ... Katia Georgopoulos

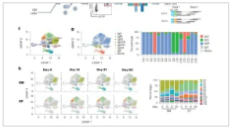


Article
31 Oct 2022

Heterogeneous plasma cells and long-lived subsets in response to immunization, autoantigen and microbiota

Durable antibody-mediated responses require long-lived plasma cells; however, these cells are difficult to identify. Hai Qi and colleagues now phenotypically identify these cells and show their heterogeneity.

Xin Liu, Jiacheng Yao ... Hai Qi

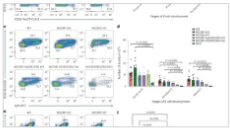


Article
31 Oct 2022

Nuclear corepressors NCOR1/NCOR2 regulate B cell development, maintain genomic integrity and prevent transformation

Farrar and colleagues perform an extensive analysis of Ncor1/2 function in B cell development. Loss of both genes results in defective pre-BCR signaling, increased accessibility of STAT5 chromatin motifs and inappropriate Rag gene expression, leading to accelerated leukemic transformation.

Robin D. Lee, Todd P. Knutson ... Michael A. Farrar



Article | [Published: 31 October 2022](#)

A double-negative thymocyte-specific enhancer augments Notch1 signaling to direct early T cell progenitor expansion, lineage restriction and β -selection

[Mariko Kashiwagi](#) , [Daniela Salgado Figueroa](#), [Ferhat Ay](#), [Bruce A. Morgan](#) & [Katia Georgopoulos](#) 
[Nature Immunology](#) (2022) | [Cite this article](#)

1338 Accesses | 7 Altmetric | [Metrics](#)

Abstract

T cell differentiation requires Notch1 signaling. In the present study, we show that an enhancer upstream of *Notch1* active in double-negative (DN) mouse thymocytes is responsible for raising Notch1 signaling intrathymically. This enhancer is required to expand multipotent progenitors intrathymically while delaying early differentiation until lineage restrictions have been established. Early thymic progenitors lacking the enhancer show accelerated differentiation through the DN stages and increased frequency of B, innate lymphoid (IL) and natural killer (NK) cell differentiation. Transcription regulators for T cell lineage restriction and commitment are expressed normally, but IL and NK cell gene expression persists after T cell lineage commitment and T cell receptor β VDJ recombination, *Cd3* expression and β -selection have been impaired. This *Notch1* enhancer is inactive in double-positive (DP) thymocytes. Its aberrant reactivation at this stage in Ikaros mutants is required for leukemogenesis. Thus, the DN-specific *Notch1* enhancer harnesses the regulatory architecture of DN and DP thymocytes to achieve carefully orchestrated changes in Notch1 signaling required for early lineage restrictions and normal T cell differentiation.

This is a preview of subscription content, [access via your institution](#)

Article	Summary
A double-negative thymocyte-specific enhancer augments Notch1 signaling to direct early T cell progenitor expansion, lineage restriction and β -selection	Notch signaling is required for T cell development. G
Heterogeneous plasma cells and long-lived subsets in response to immunization, autoantigen and microbiota	Durable antibody-mediated responses require long-l
Nuclear corepressors NCOR1/NCOR2 regulate B cell development, maintain genomic integrity and prevent transformation	Farrar and colleagues perform an extensive analysis c

5. Contenido. Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

Article: Nombre del artículo científico.

Summary: Resumen general de la investigación que se ha llevado a cabo.

Authors: Nombre de los autores del artículo científico.

Date: Fecha de publicación.

Access: Si es de acceso público o se tiene que pagar para acceder al artículo.

Figure: Una imagen representativa del estudio en cuestión.

Link paper: Link del artículo científico, los artículos de acceso no público muestran el título del artículo y el resumen general del artículo.

TCR, BCR, T CELL, B CELL, NKC, CD4, CD8, DEEP LEARNING, MACHINE

LEARNING y HLA: palabras clave que nos interesan, en caso de que el estudio contenga esa palabra clave el valor va a ser 1 y en caso de no contenerla será 0.

6. Propietario. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Teniendo en cuenta que utilizamos una página web pública, los propietarios de los artículos, primeramente son los autores de cada paper y dos, el journal Nature ya que es la plataforma donde se han publicado los estudios.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Más que inspiración, podemos decir que la razón por la que hemos desarrollado este proyecto fue, más bien, la frustración. Como investigadores, una de las partes esenciales de nuestro trabajo es nunca parar de estudiar y aprender información nueva. Para poder hacer un uso adecuado de nuestro tiempo, es necesario automatizar la búsqueda de los artículos de interés. Otros ámbitos de ciencias computacionales ya han aplicado una manera parecida para seleccionar páginas web de interés con información más relevante para el análisis en cuestión.

8. Licencia. Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección.

MIT

Usamos la licencia MIT ya que el data viene de un sitio publico y nosotros no tenemos ningún tipo de ownership. Consideramos que el código que hemos escrito no es nada groundbreaking para que esté bajo una licencia protectora, no obstante si hay algún bug, no nos hacemos responsables de las consecuencias.

9. Código. Código con el que se ha obtenido el dataset, preferiblemente en Python o, alternativamente, en R.

Link al github con el código : [mariona9906/practica1_tipologia \(github.com\)](https://github.com/mariona9906/practica1_tipologia)

La razón principal por la que hemos escogido Python como el lenguaje principal de este proyecto es la experiencia que tenemos. Los dos trabajamos con Python y BeautifulSoup ha parecido ser una librería bastante buena y fácil de aprender.


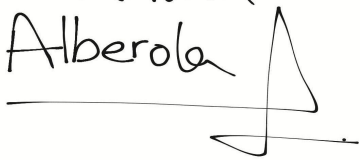

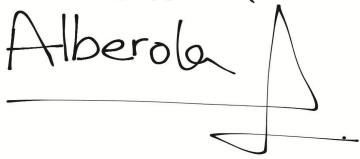

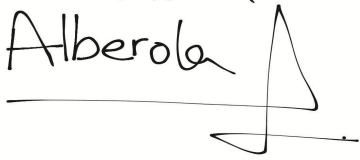

10. Dataset. Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset ([https://doi.org/...](https://doi.org/)). El dataset también deberá incluirse en la carpeta /dataset del repositorio. Si existe alguna circunstancia que impida publicar abiertamente el dataset real en Zenodo, se deberá: (1) comentar esta circunstancia y justificar el motivo en este apartado; (2) generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI; y (3) comunicar al profesor el dataset real de forma privada (p. ej., utilizando un repositorio privado).

Link Zenodo: <https://zenodo.org/record/7317220#.Y3FNF3bMJD8>

11. Vídeo. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo ([https://drive.google.com/...](https://drive.google.com/)), que deberá ubicarse en el Google Drive de la UOC.

Link video:

<https://drive.google.com/file/d/17OuWaFIVrKCL7sGpDLO-r-uJVTI82TrB/view?usp=sharing>

Contribuciones	Firma Dmytro	Firma Mariona
Investigación previa		Mariona Alberola 
Redacción de las respuestas		Mariona Alberola 
Desarrollo del código		Mariona Alberola 
Participación en el video		Mariona Alberola 