



# Using LLM applications for gene prioritization in research

---

Damien Chaussabel

April 22, 2026



# LLM use Exploration team

---

- Technology surveillance
- Workflow development
- Systematic benchmarking
- Application development (via APIs)
- Education – incorporate LLM workflows into data-to-knowledge training curriculum



Mohammed Toufiq



Taushif Khan



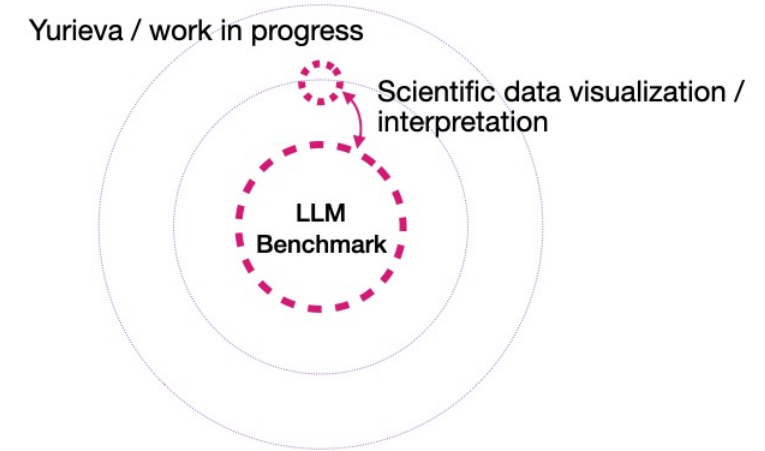
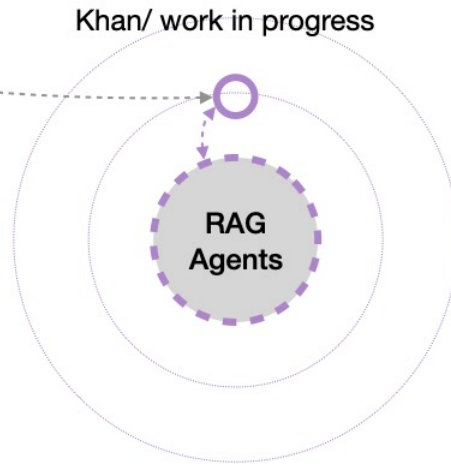
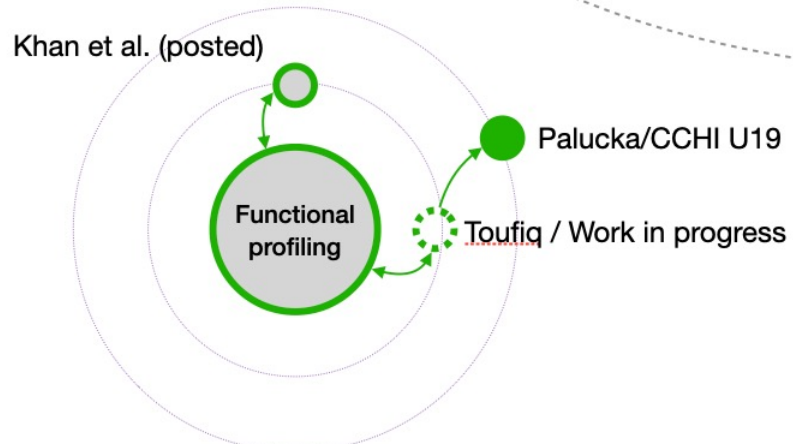
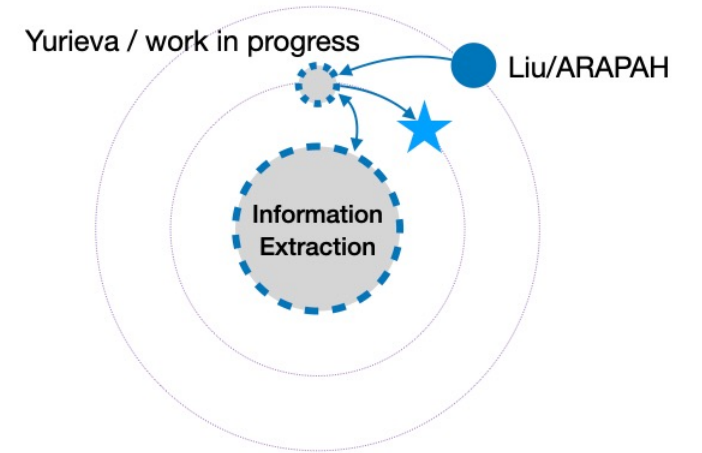
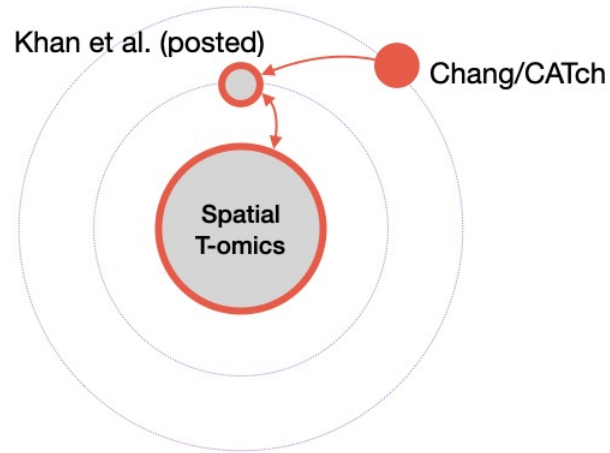
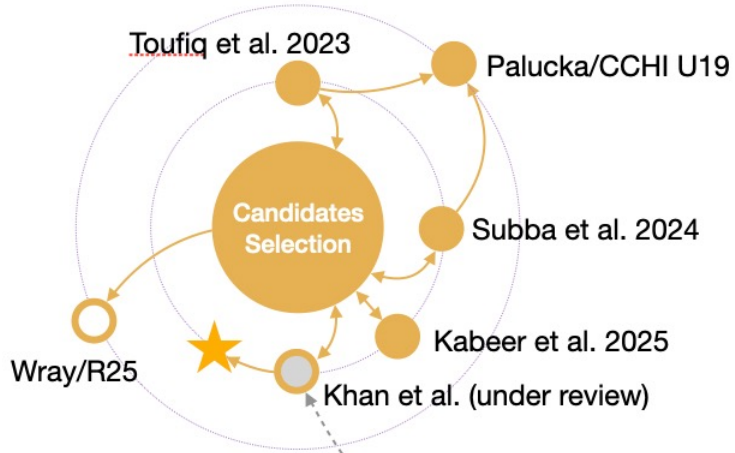
Marina Yurieva



Julius Henderson  
Joshy George

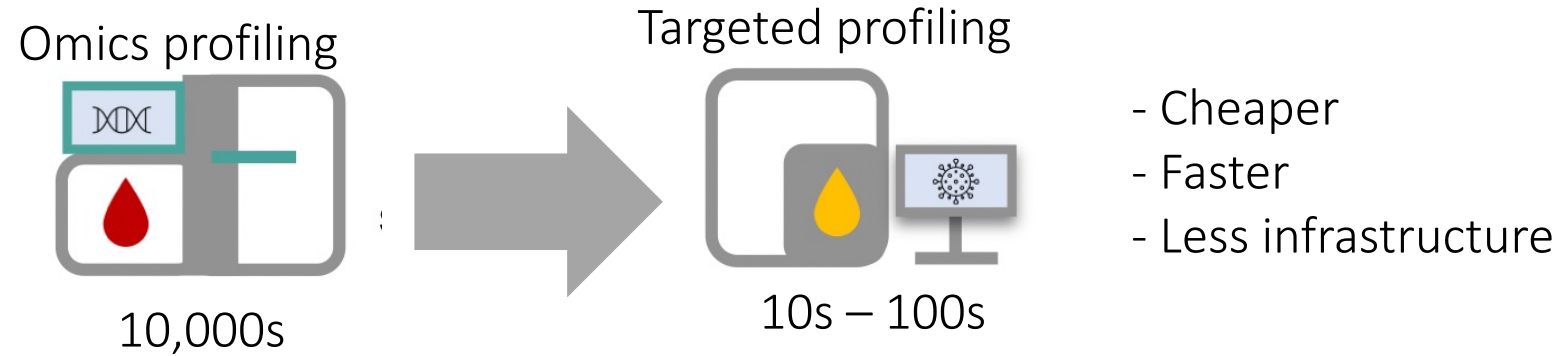
# Exploring LLM applications across biomed R&D workflows

● Workflows/Papers    ★ Applications

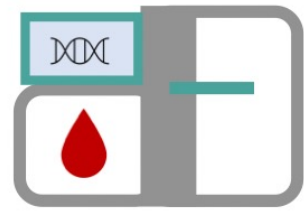


# Targeted Immune Profiling Assays

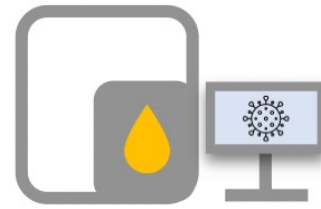
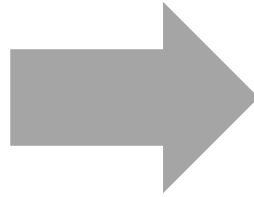
## Rationale



- Measure transcript abundance in blood for a select panel of genes (tens or hundreds).
- Research: valuable when profiling large number of samples, and when resources are limiting (assay cost but also infrastructure).
- Clinical translation: such assays can reflect pathogenesis (differential diagnosis, early detection) and response to treatment



10,000s



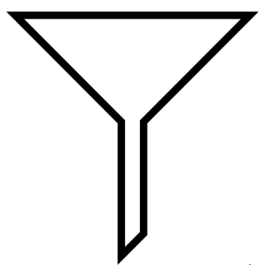
10s – 100s

## Feature selection

Data-driven

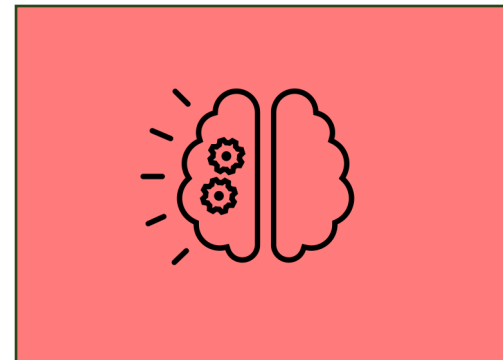
structured information

010010110



Knowledge-driven

unstructured information

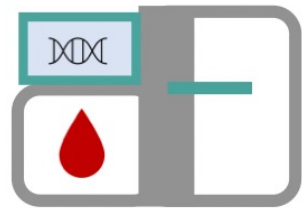


Requires sifting through vast volumes of biomedical knowledge

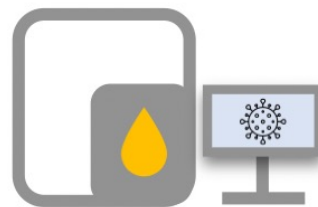
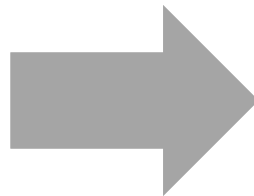


Biomarkers, Drug targets, Early-stage discoveries

Commonly used for omics profiling data



10,000s



10s – 100s

## Feature selection

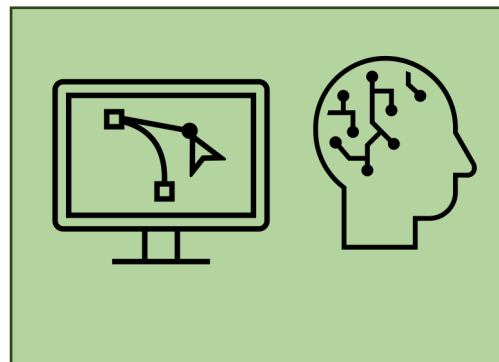
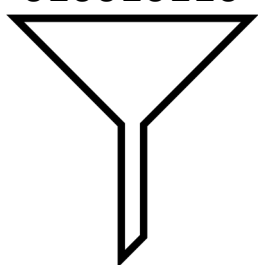
Large Language Models

Data-driven  
structured information

Knowledge-driven  
unstructured information

Commonly  
used for omics  
profiling data

010010110

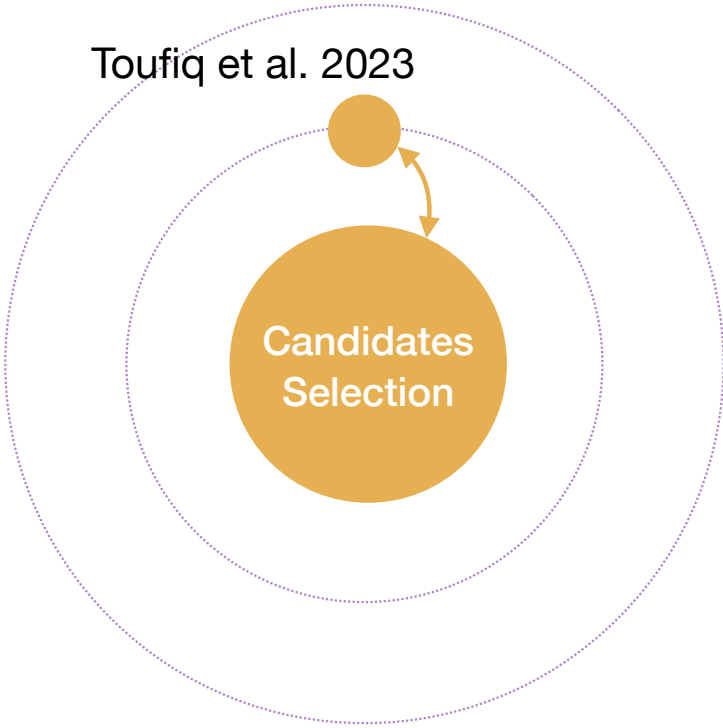


Requires sifting  
through vast  
volumes of  
biomedical  
knowledge



Biomarkers, Drug targets,  
Early-stage discoveries

# Using LLMs for Candidate prioritization: first POC




Toufiq et al. *Journal of Translational Medicine* (2023) 21:728  
<https://doi.org/10.1186/s12967-023-04576-8>

Journal of  
Translational Medicine

**RESEARCH** **Open Access**

## Harnessing large language models (LLMs) for candidate gene prioritization and selection



Mohammed Toufiq<sup>1†</sup>, Darawan Rinchai<sup>2†</sup>, Eleonore Bettacchioli<sup>3,4</sup>, Basirudeen Syed Ahamed Kabeer<sup>5</sup>, Taushif Khan<sup>1</sup>, Bishesh Subba<sup>1</sup>, Olivia White<sup>1</sup>, Marina Yurieva<sup>1</sup>, Joshy George<sup>1</sup>, Noemie Jourde-Chiche<sup>6</sup>, Laurent Chiche<sup>7</sup>, Karolina Palucka<sup>1</sup> and Damien Chaussabel<sup>1\*</sup> 



Mohammed Toufiq

Seminal work leveraging LLMs in the context of candidate gene prioritization & selection

# Gene-by-gene scoring across statements

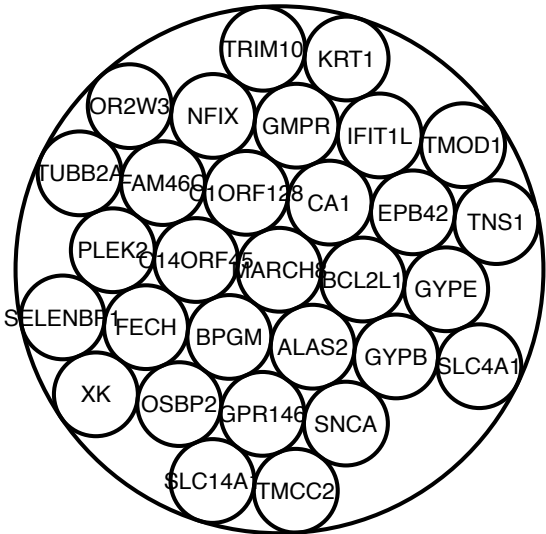
## Chat Scoring

PROMPT 3.1: I am next going to ask for [Gene Symbol] to:  
[...] Give each of the following statements a score from 0 to 10, with 0 indicating no evidence and 10 indicating very strong evidence:

- The gene is associated with erythroid cells or erythropoiesis.
- The gene is currently being used as a biomarker in clinical settings.
- The gene has potential value as a blood transcriptional biomarker.
- The gene is relevant to circulating leukocytes immune biology.
- The gene is a known drug target.
- The gene is therapeutically relevant for immune-mediated diseases.

 **OpenAI**

 **Claude**

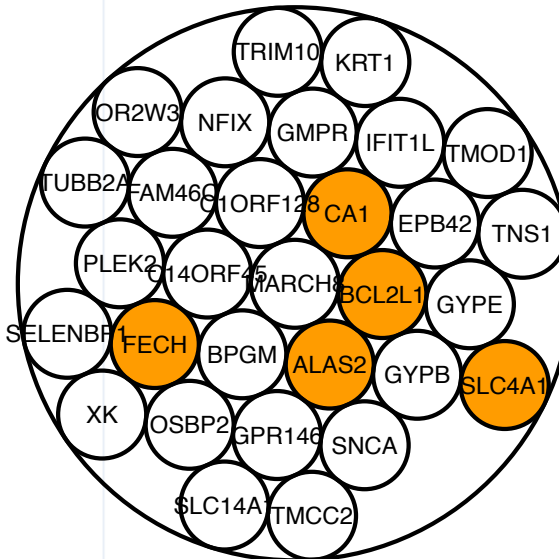
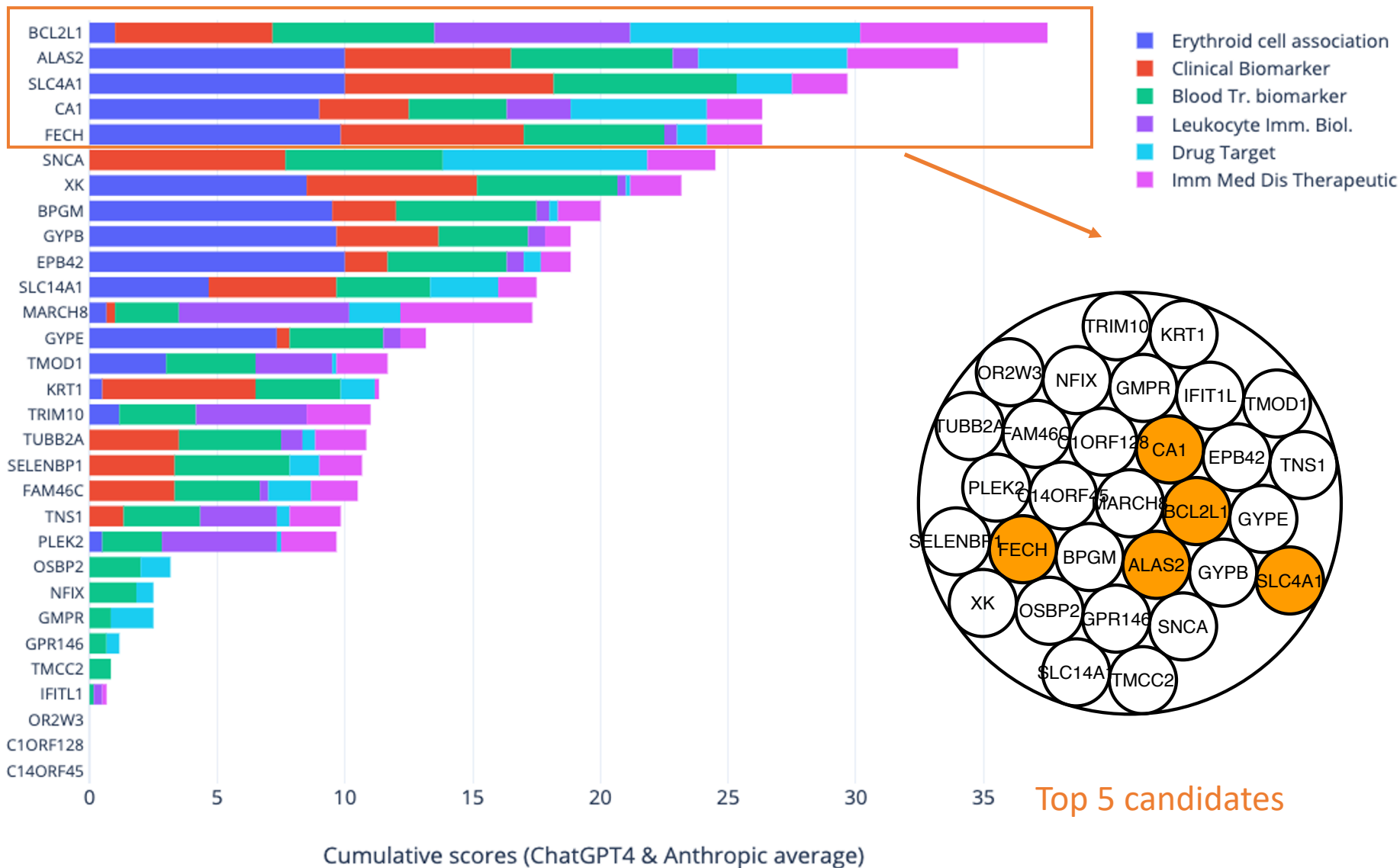


**Module M9.2**

(BloodGen3 Repertoire)

# Chat Scoring

## Module M9.2 prioritized gene list



# Fact checking justifications for the top 5 candidates

Justification for 1 statement,  
1 candidate:

Used as a biomarker in clinical settings: Galectin-3, encoded by the LGALS3 gene, is currently utilized as a clinical biomarker, particularly in the context of heart failure and fibrosis progression in conditions like liver cirrhosis, as supported by studies such as "Galectin-3 in Cardiac Remodeling and Heart Failure" (de Boer, R.A., 2020) and "Galectin-3: A biomarker and potential therapeutic target for liver cirrhosis and hepatocellular carcinoma" (Yang et al., 2019). In contrast, MAP2K3 is not widely used as a clinical biomarker and no specific references substantiate its clinical utility for this purpose. PPM1A has been investigated for its potential as a biomarker in certain types of cancer, such as lung and colorectal cancer, but it is not currently broadly applied in clinical settings, as indicated in studies like "PPM1A is a RelA phosphatase with tumor suppressor-like activity" (Lu, 2014). Similarly, SMOX has research suggesting its potential use as a cancer biomarker, specifically in "Overexpression of Spermine Oxidase SMO(PAOH1) in Breast Cancer and Lung Cancer" (Wang, H., 2009) but it is not generally employed in clinical scenarios. MARCH2 lacks evidence for its use as a clinical biomarker, with no specific references to support such an application.



Retrieving backing  
references for each  
statement.



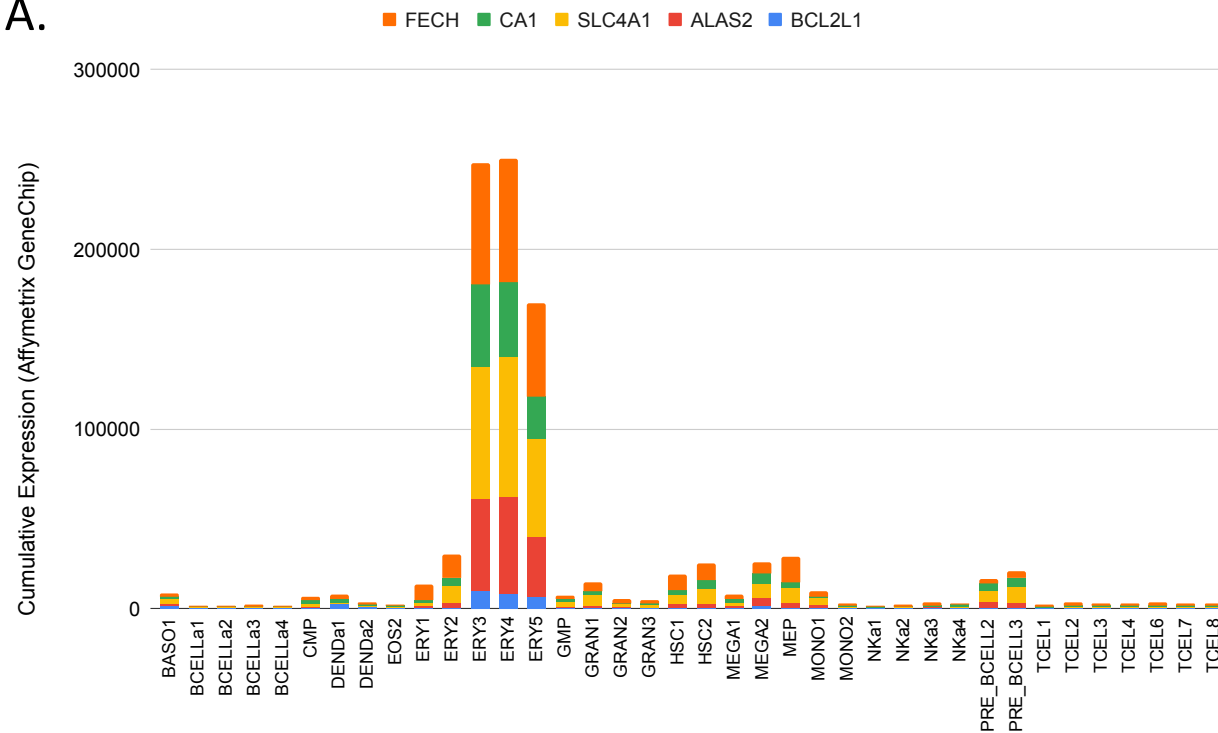
# LLMs take reference transcriptional profiling data into consideration



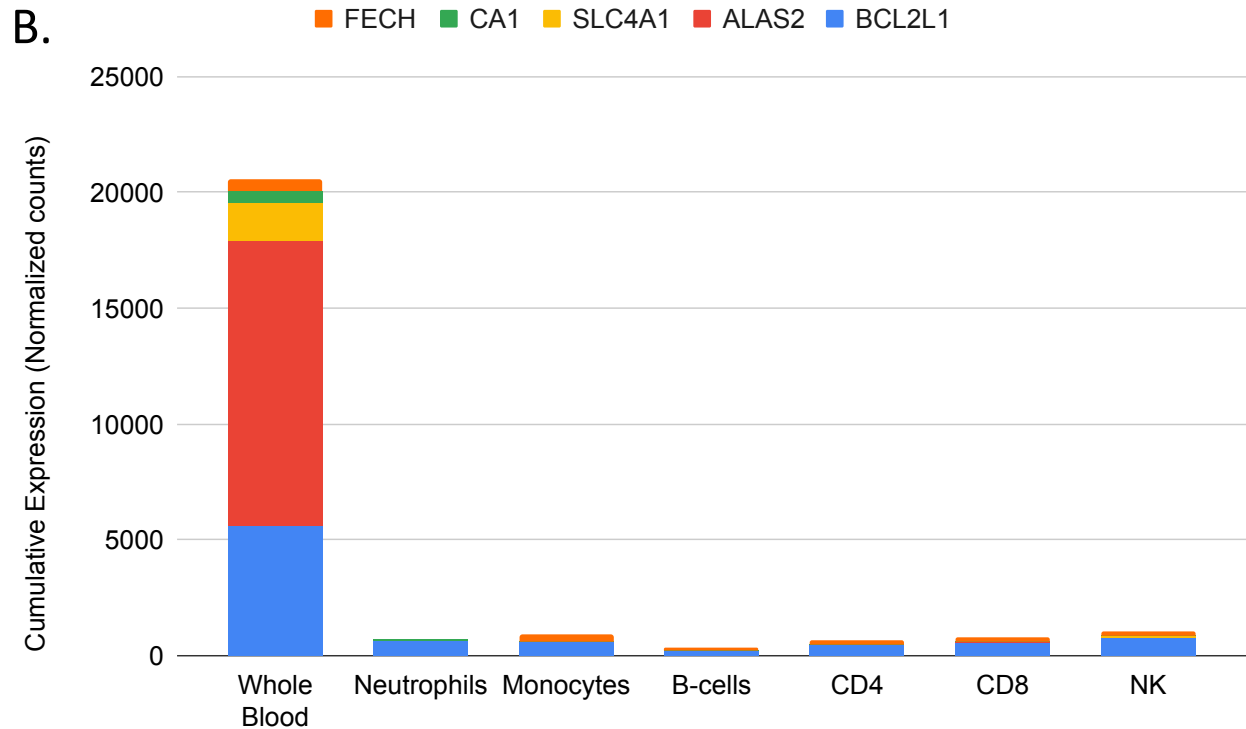
>> Reference transcriptional profiling data for the top 5 candidates is passed to the models.

>> The models interpret these data to refine their selection

A.



B.



# LLMs finalize their selection

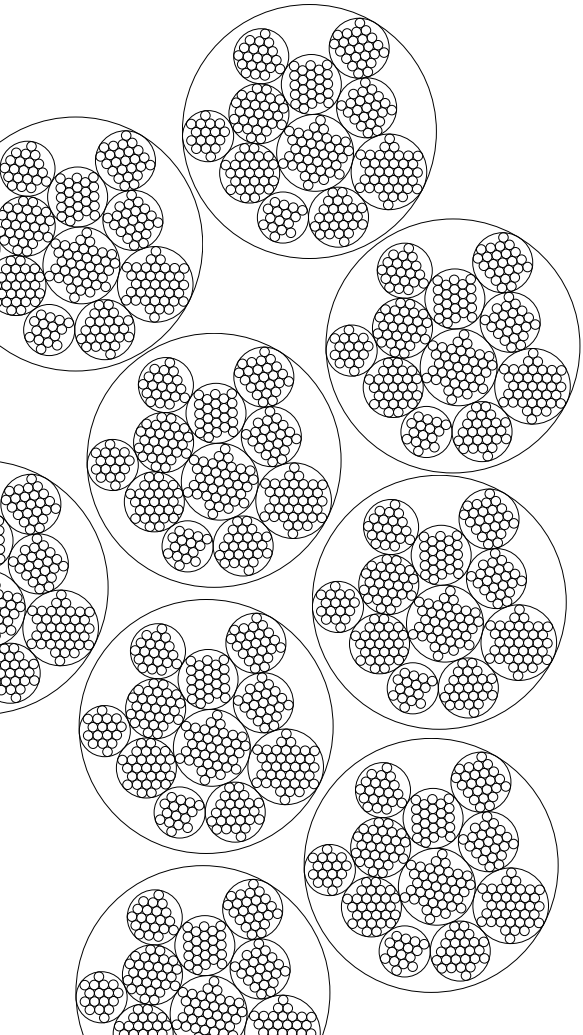
GPT-4: “With the addition of the RNA-seq data, it's clear that ALAS2 is the most abundantly expressed gene in whole blood with a count of 616. This contrasts with BCL2L1, the initially recommended gene, which exhibits a lower expression level in whole blood (280). Importantly, ALAS2 demonstrates minimal expression in all other tested immune clls (counts of 1), suggesting a strong erythroid cell-specific expression, fitting the criteria to be representative of an erythroid cell signature. BCL2L1, on the other hand, is more uniformly expressed across various cell types, which makes it less specific to erythroid cells.

Given these findings, I recommend reconsidering the selection and placing ALAS2 as the top candidate gene for inclusion in the targeted blood transcriptional profiling panel. Here's why: etc...”



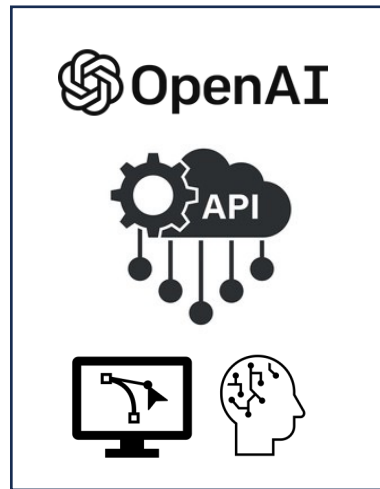
# Automating scoring via LLMs API

BloodGen3 Module  
Genes (N= 11,462)



Discovery

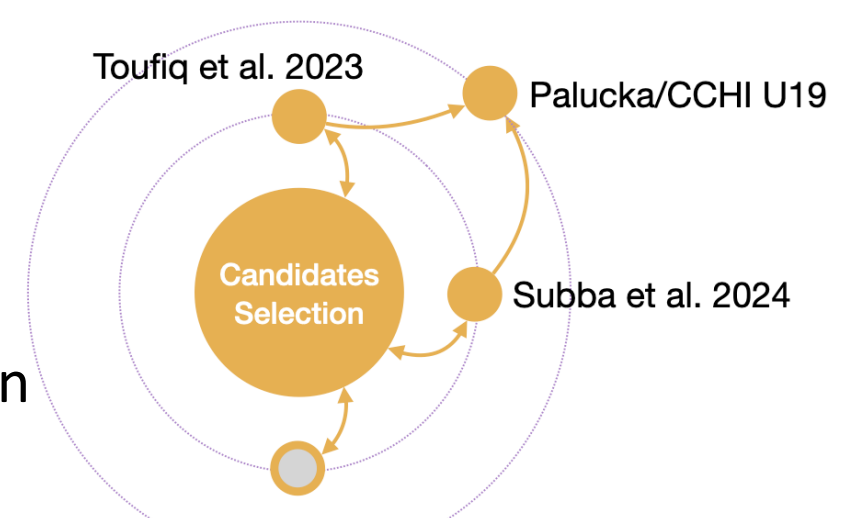
High throughput  
scoring



Custom Application



Taushif Khan



*Bioinformatics*, 2025, **41(10)**, btaf541  
<https://doi.org/10.1093/bioinformatics/btaf541>  
Advance Access Publication Date: 10 October 2025  
**Original Paper**

Use following parameters to examine your  
gene set.

**Important :** For long list of genes consider using local deployment

Choose a CSV file with genes in 'Genes' column



Drag and drop file here

Limit 200MB per file • CSV

Browse files

Choose a JSON file with DEFINED parameters



Drag and drop file here

Limit 200MB per file •  
JSON

Browse  
files

Load Example gene list

Please upload gene list

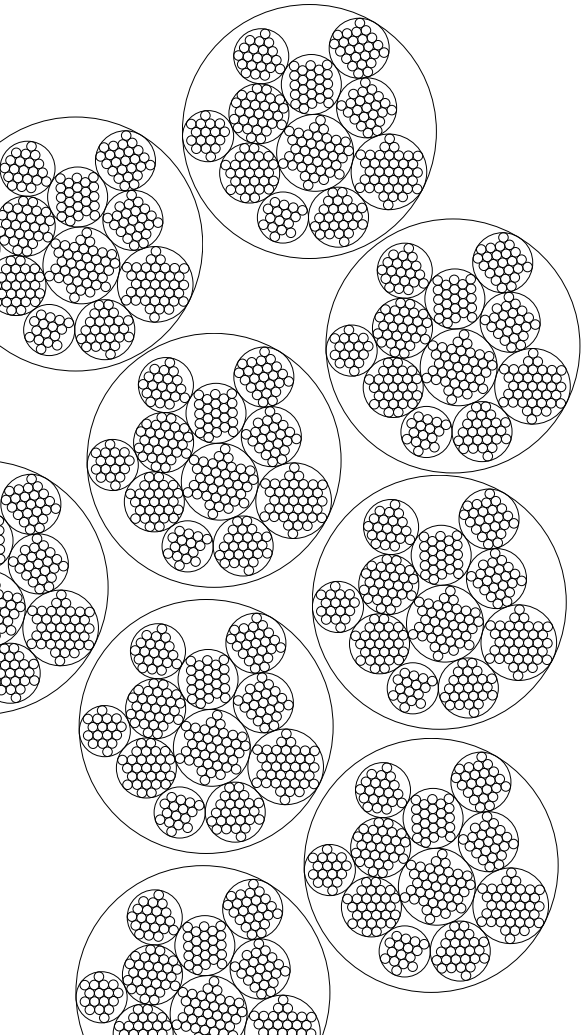
Load example parameters

Please upload parameter file

<https://genellm.streamlit.app/TestYourGeneSet>

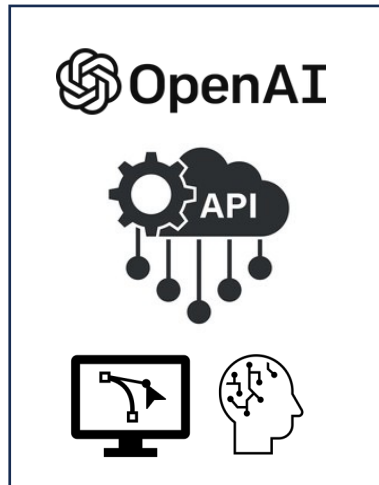
# Generating association scores for >10k genes

BloodGen3 Module  
Genes (N= 11,462)



## Discovery

High throughput  
scoring



Custom Application

The gene is associated with the pathogenesis of sepsis. Score: Based on evidence of the gene's involvement in the biological processes underlying sepsis, including but not limited to its role in the dysregulated host response to infection, organ dysfunction, or sepsis-related complications.

The gene is associated with the host immune response in sepsis. Score: Based on evidence of the gene's involvement in the immune response during sepsis, including but not limited to its role in innate or adaptive immunity, inflammation, or immunosuppression.

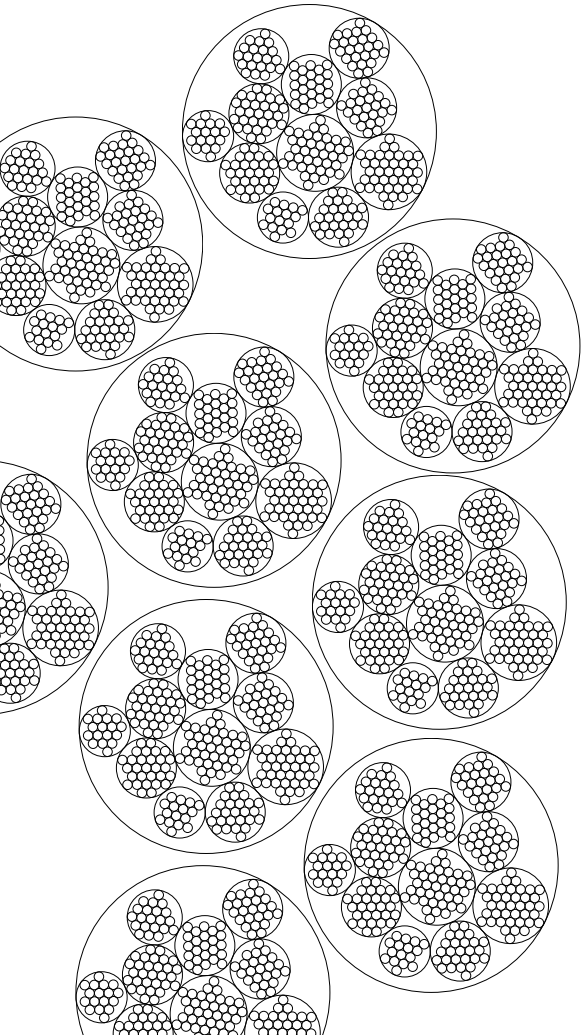
The gene is associated with sepsis-related organ dysfunction. Score: Based on evidence of the gene's involvement in the development or progression of organ dysfunction in sepsis, including but not limited to its role in cardiovascular, respiratory, renal, hepatic, or neurological dysfunction.

The gene is relevant to circulating leukocytes immune biology in sepsis. Score: Based on evidence linking the gene to the development, function, or regulation of circulating leukocytes in the context of sepsis, including impacts on leukocyte differentiation, activation, signaling, or effector functions.

Etc....

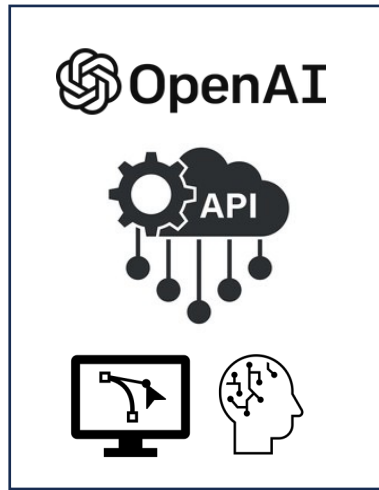
# Generating association scores for >10k genes

BloodGen3 Module  
Genes (N= 11,462)

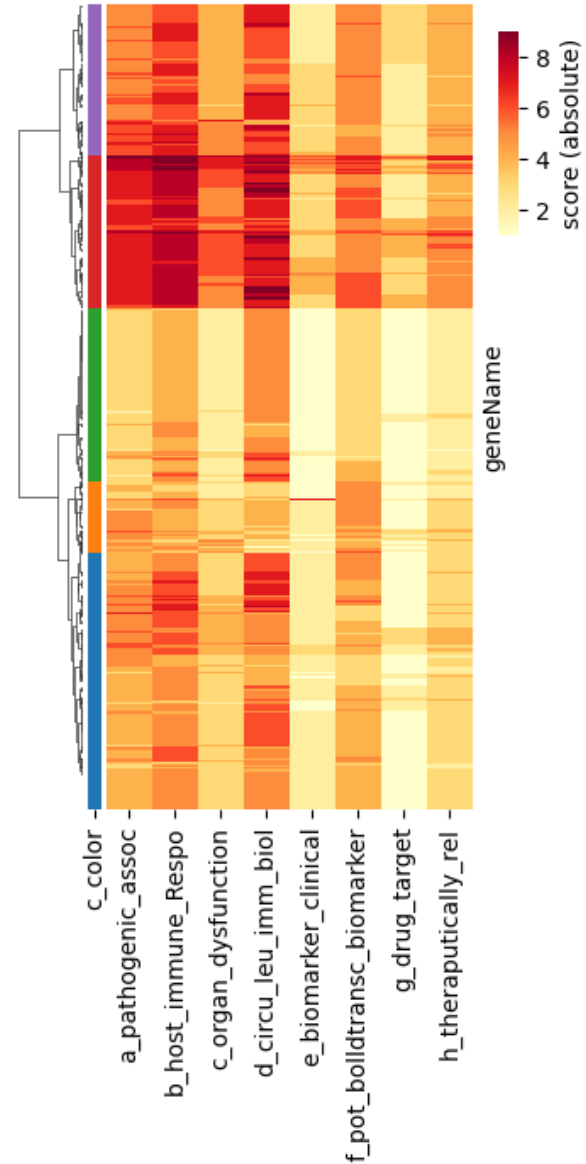


Discovery

High throughput  
scoring



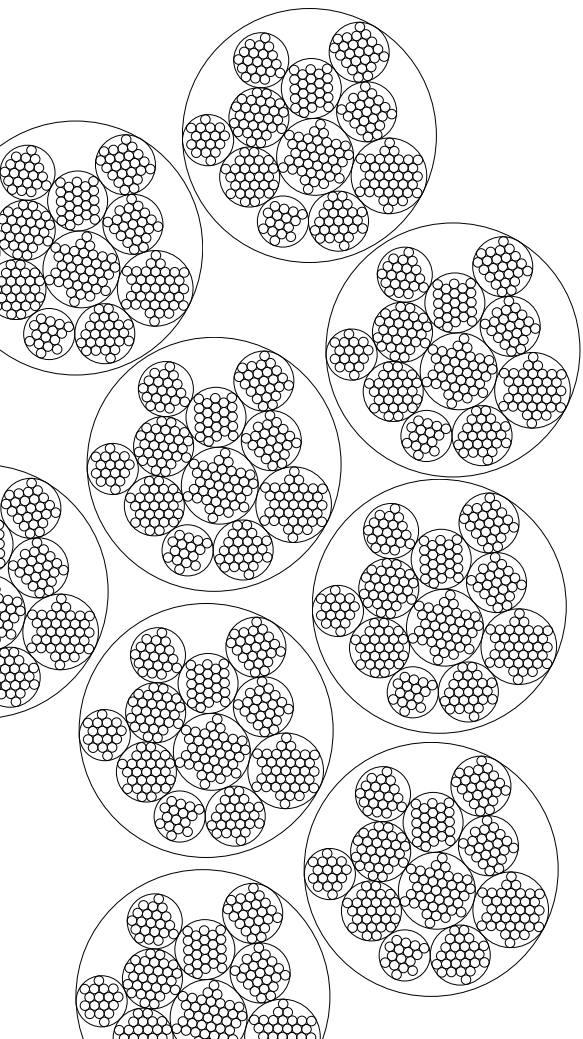
Custom Application



1,070 genes with score > 5  
for at least one of the  
statements

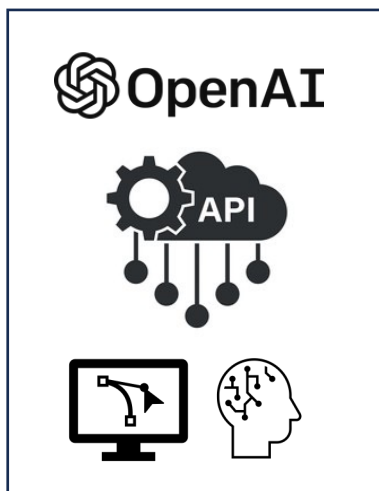
# Mapping functional relevance of the scores

BloodGen3 Module  
Genes (N= 11,462)

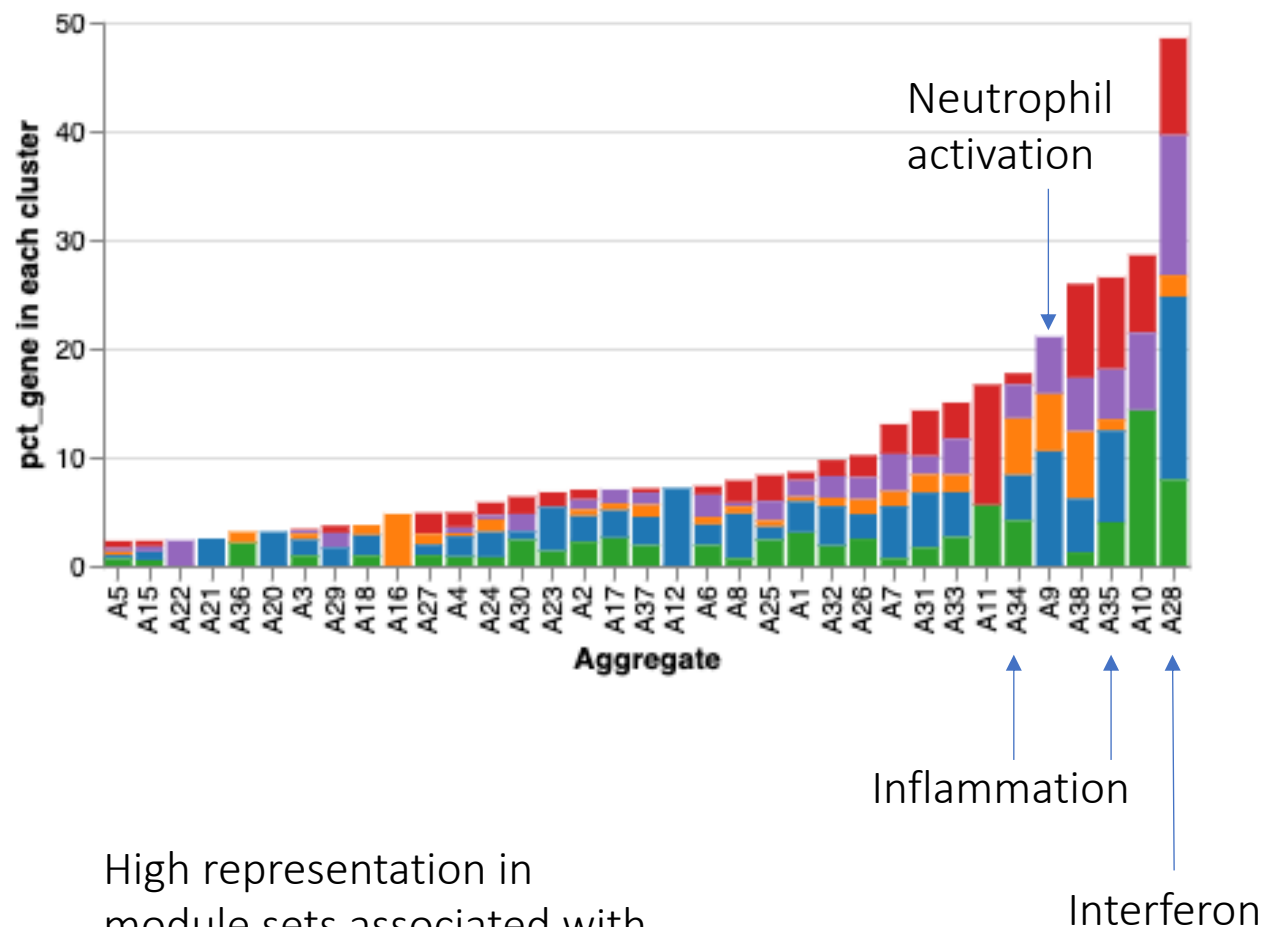


Discovery

High throughput  
scoring



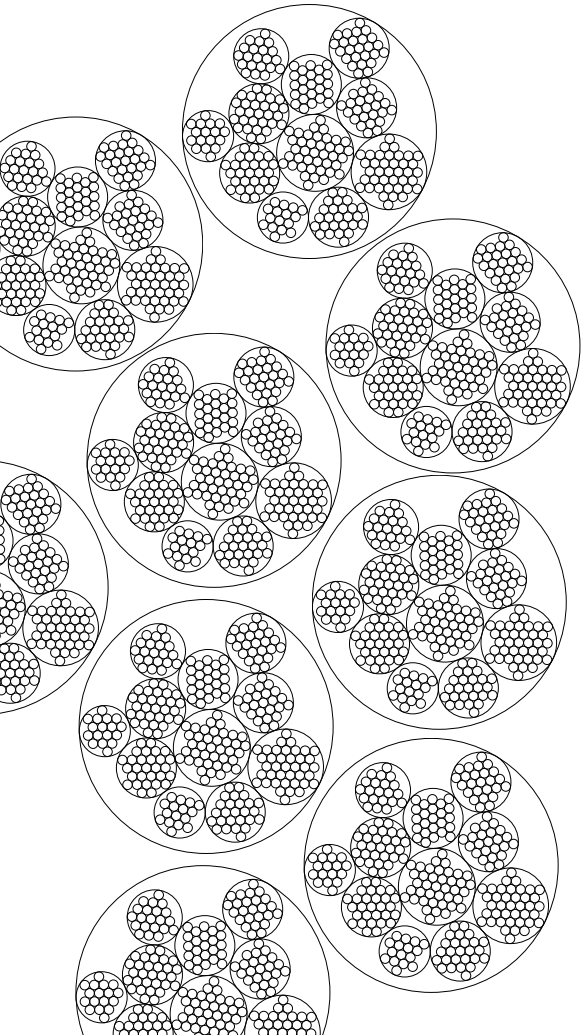
Custom Application



High representation in  
module sets associated with  
responses to infection / innate  
immunity

# Adding a RAG system as grounding layer

BloodGen3 Module  
Genes (N= 11,462)



Discovery

High throughput  
scoring

*Bioinformatics*, 2025, **41(10)**, btaf541  
<https://doi.org/10.1093/bioinformatics/btaf541>  
Advance Access Publication Date: 10 October 2025  
**Original Paper**

OXFORD

Data and text mining

## Automating candidate gene prioritization with large language models: from naive scoring to literature-grounded validation

Taushif Khan<sup>1,\*</sup> , Mohammed Toufiq<sup>1</sup>, Marina Yurieva<sup>1</sup>, Nitaya Indrawattana<sup>2,3</sup>,  
Akanitt Jittmittraphap<sup>4</sup>, Nathamon Kosoltanapiwat<sup>4</sup> , Pornpan Pumirat<sup>4</sup>,  
Passanesh Sukphopetch<sup>4</sup> , Muthita Vanaporn<sup>4</sup>, Karolina Palucka<sup>1</sup>,  
Basirudeen Syed Ahamed Kabeer<sup>5,6</sup>, Darawan Rinchai<sup>7</sup>, Damien Chaussabel<sup>1,\*</sup>

| Status    | Citation  | Title  | Scope/Application           | DOI   |
|-----------|---|--|-----------------------------|---|
| Published | Subba B, Toufiq M, Omi F, Yurieva M, Khan T, Rinchai D, Palucka K, Chaussabel D. (2024) | Human-augmented large language model-driven selection of glutathione peroxidase 4 as a candidate blood transcriptional biomarker for circulating erythroid cells | Candidate gene selection    | <a href="https://doi.org/10.1038/s41598-024-73916-5">https://doi.org/10.1038/s41598-024-73916-5</a> |
| Published | Toufiq M, Rinchai D, Bettacchioli E, et al. (2023)                                      | Harnessing large language models (LLMs) for candidate gene prioritization and selection  | Candidate gene selection    | <a href="https://doi.org/10.1186/s12967-023-04576-8">https://doi.org/10.1186/s12967-023-04576-8</a> |
| Published | Kabeer B, Subba B, Rinchai D, et al. (2024, Frontiers in Medicine)                      | From Gene Modules to Gene Markers: An Integrated AI-Human Approach Selects CD38 to Represent Plasma Cell-Associated Transcriptional Signatures                   | Candidate gene selection    | <a href="https://doi.org/10.3389/fmed.2025.1510431">https://doi.org/10.3389/fmed.2025.1510431</a>   |
| Preprint  | Khan T, Toufiq M, Yurieva M, et al. (2024, bioRxiv)                                     | Automating candidate gene prioritization with large language models: Development and benchmarking of an API-driven workflow leveraging GPT-4                     | Candidate gene selection    | <a href="https://doi.org/10.1101/2024.12.10.627808">https://doi.org/10.1101/2024.12.10.627808</a>   |
| Preprint  | Khan T, Farley CM, Wilson JJ, et al. (2024)   | Tackling the complexity of spatial transcriptomics data interpretation with large language models  | Spatial transcriptomic data | <a href="https://doi.org/10.1101/2024.11.28.625773">https://doi.org/10.1101/2024.11.28.625773</a>   |
| Preprint  | Khan T, Yurieva M, Kabeer BSA, et al. (2024, bioRxiv)                                   | Deep functional profiling of gene sets using large language models: A blueprint for tailored, context-aware functional annotation                                | Deep functional profiling   | <a href="https://doi.org/10.1101/2024.12.12.628275">https://doi.org/10.1101/2024.12.12.628275</a>   |

# Exploring LLM utility

---

- Workflow development
- Application development (using APIs)
- Systematic benchmarking
- Technology surveillance - RAG and agent-based systems



Mohammed Toufiq



Taushif Khan



Marina Yurieva



Julius Henderson

Joshy George