RESEARCH ARTICLE

REVISED Assessing the potential relevance of CEACAM6 as a blood transcriptional biomarker [version 2; peer review: 1 approved, 2 approved with reservations]

Darawan Rinchai[1], Damien Chaussabel[2]

[1]St. Giles Laboratory of Human Genetics of Infectious Diseases, The Rockefeller University, New York, New York, 10065, USA
[2]Computer Sciences Department, The Jackson Laboratory, Farmington, CT, 06032, USA

v2 **First published:** 11 Nov 2022, **11**:1294
https://doi.org/10.12688/f1000research.126721.1
Latest published: 04 Apr 2024, **11**:1294
https://doi.org/10.12688/f1000research.126721.2

Abstract

Background

Changes in blood transcript abundance levels have been associated with pathogenesis in a wide range of diseases. While next generation sequencing technology can measure transcript abundance on a genome-wide scale, downstream clinical applications often require small sets of genes to be selected for inclusion in targeted panels. Here we set out to gather information from the literature and transcriptome datasets that would help researchers determine whether to include the gene CEACAM6 in such panels.

Methods

We employed a workflow to systematically retrieve, structure, and aggregate information derived from both the literature and public transcriptome datasets. It consisted of profiling the CEACAM6 literature to identify major diseases associated with this candidate gene and establish its relevance as a biomarker. Accessing blood transcriptome datasets identified additional instances where CEACAM6 transcript levels differ in cases vs controls. Finally, the information retrieved throughout this process was captured in a structured format and aggregated in interactive circle packing plots.

Results

Although it is not routinely used clinically, the relevance of CEACAM6 as a biomarker has already been well established in the cancer field,

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 2 (revision) 04 Apr 2024		✓ view	? view
version 1 11 Nov 2022	? view	? view	

1. **Mattia Lauriola**, University of Bologna, Bologna, Italy

2. **Ritu Pandey**, University of Arizona, Tucson, USA
Daruka Mahadevan, The University of Texas Health Science Center (Ringgold ID: 12346), San Antonio, USA

3. **Akila Prashant**, JSS Academy of Higher Education & Research, Mysore, India
Deepthi V, JSS Academy of Higher Education and Research, Mysuru, India

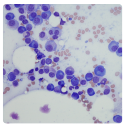
Any reports and responses or comments on the article can be found at the end of the article.

where it has invariably been found to be associated with poor prognosis. Focusing on the blood transcriptome literature, we found studies reporting elevated levels of CEACAM6 abundance across a wide range of pathologies, especially diseases where inflammation plays a dominant role, such as asthma, psoriasis, or Parkinson's disease. The screening of public blood transcriptome datasets completed this picture, showing higher abundance levels in patients with infectious diseases caused by viral and bacterial pathogens.

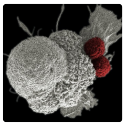
Conclusions

Targeted assays measuring CEACAM6 transcript abundance in blood may be of potential utility for the management of patients with diseases presenting with systemic inflammation and for the management of patients with cancer, where the assay could potentially be run both on blood and tumor tissues.

Keywords
Biomarkers, CEACAM6, Transcriptional profiling, Literature profiling



This article is included in the Cell & Molecular Biology gateway.



This article is included in the Oncology gateway.

Corresponding author: Damien Chaussabel (damien.chaussabel@jax.org)

Author roles: Rinchai D: Conceptualization, Data Curation, Formal Analysis, Methodology, Visualization, Writing – Review & Editing; **Chaussabel D**: Conceptualization, Data Curation, Formal Analysis, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Rinchai D and Chaussabel D. **Assessing the potential relevance of CEACAM6 as a blood transcriptional biomarker [version 2; peer review: 1 approved, 2 approved with reservations]** F1000Research 2024, **11**:1294 https://doi.org/10.12688/f1000research.126721.2

First published: 11 Nov 2022, **11**:1294 https://doi.org/10.12688/f1000research.126721.1

REVISED **Amendments from Version 1**

In this revised version of our article, we have clarified the manuscript's focus, emphasizing that the study serves as a proof of concept applying a previously established methodological framework to the CEACAM6 gene. We also address the potential of automation in data curation, discussing our exploration into the use of Large Language Models (LLMs) to enhance efficiency and accuracy. Furthermore, we have updated the discussion around CEACAM6 as a therapeutic target, acknowledging ongoing research and clinical trials that explore its potential, particularly in the context of cancer therapy. These revisions ensure the article accurately reflects current research and methodological advancements, providing a comprehensive and up-to-date overview of the subject. Additionally, we have incorporated a new reference (PMID: 32257432), shedding light on the role of CEACAM6 in individuals with positive fecal immunochemical tests but no intestinal lesions. This further elucidates CEACAM6's availability in circulating neutrophils and enhances our manuscript's precision in representing CEACAM6 as a blood biomarker.

Any further responses from the reviewers can be found at the end of the article

Introduction

Changes in blood transcript abundance can reflect differences in relative abundance of leukocyte populations as well as transcriptional regulation secondary to immune activation (for instance inflammation, interferon, and prostaglandin responses). Quantifying these changes can thus be relevant for making clinical decisions.[1,2] Robust technology platforms, such as microarrays and RNA sequencing, that enable the measurement of transcript abundance in an unbiased fashion (i.e., simultaneously measuring all RNA species that are present in a given sample) have been widely available for the past two decades. As a result, blood transcriptome studies have been conducted across a wide range of pathological or physiological states.[3–7] In addition, vast amounts of blood transcriptome profiling data have been made available in public repositories such as the NCBI Gene Expression Omnibus, or EMBL-EBI's array express.[8]

Transcriptome profiling data can be leveraged to inform the design of targeted gene panels. These panels can serve as a basis for the development of diagnostic assays for use in clinical settings. But targeted assays can also be employed in research settings, for instance when profiling of transcript abundance needs to be performed on large scales (e.g., in thousands of samples) and with a relatively short turnaround. Notably, targeted assays could also prove valuable in resource-constrained settings, where computing infrastructure, instrument, and reagents costs are limiting. The approaches employed for targeted assay design can be data-driven (e.g., applying computational models to transcriptome profiling dataset(s) to select genes based on their predictive performance) or knowledge-driven (selecting genes based on pre-existing knowledge – e.g., for the development of an "immunology panel"). However, both data and knowledge-driven approaches can also be combined. This is illustrated in recently published work in which we describe the selection of three blood transcriptional panels designed for the monitoring of responses to SARS-CoV-2.[9] Transcripts were selected first based on their membership to co-expressed gene sets, the abundance of which was found to change during COVID-19 disease (i.e., through a data-driven approach) and second based on their relevance to one of three themes, which were immunity, therapeutic development, and severe acute respiratory syndrome biology (i.e., through a knowledge-driven approach). However, the amount of information available in the literature and in public transcriptome datasets that can be leveraged for candidate gene selection can be overwhelming. Thus, we have developed an approach to identify, retrieve, structure, and aggregate such information in a manner that would support the rational selection of candidate genes for inclusion in targeted assays destined to be used in clinical or research settings.[10]

Here we decided to focus on CEACAM6, a gene encoding a protein of the carcinoembryonic antigen (CEA) family whose members are glycosylphosphatidylinositol (GPI)-linked cell surface proteins.[11,12] The methodology employed in this study is derived from our previously established "collective omics data" (COD) training curriculum,[13] as outlined in our comprehensive methods paper, "A training curriculum for retrieving, structuring, and aggregating information derived from the biomedical literature and large-scale data repositories."[10] This foundational paper provides a detailed description of our systematic approach to information curation, which we have applied in the current investigation of CEACAM6. Specifically, the study utilizes the COD1 training module workflow from this curriculum, which guides the structured retrieval and aggregation of gene-specific data for biomarker assessment. The process encompasses selecting a gene of interest, in this case, CEACAM6, to comprehensively gather and synthesize relevant information from both literature and public datasets, culminating in the creation of resources like structured data tables and interactive circle packing plots. This approach not only supports the rigorous assessment of CEACAM6's potential as a blood biomarker but also serves as a demonstrative application of our validated methodological framework, providing a practical example of how such a framework can be employed to enhance biomarker discovery efforts.

Screening the CEACAM6 literature identified a strong association with various cancers, in particular colorectal cancer where measurement of CEACAM6 blood transcript levels may be of clinical value for early detection.[14–16] Associations

were also found for pancreatic, lung, and breast cancer, as well as leukemia and inflammatory bowel disease. More in depth profiling of the literature (analyzing the full text) identified an array of conditions for which CEACAM6 abundance has been found to be significantly different from controls. This list was complemented by a screening of public blood transcriptome datasets. The tables employed to capture this information in a structured format are shared as extended data files. Another deliverable is the interactive circle packing plot that permits aggregation and seamless access to this and all underlying information. Altogether these resources supported manuscript preparation and interpretation/evaluation by the authors of the relevance of CEACAM6 as a biomarker. They may also support transcript selection efforts of members of the research community interested in designing blood transcriptional biomarker panels.

Methods

Overall literature and large-scale dataset profiling approach

The workflow implemented here to assess the potential of CEACAM6 as a blood transcriptional biomarker has been described in detail in a separate methods paper.[10] The approach was devised as part of a training module focused on the development of skills for the retrieval, structuring, aggregation, and interpretation of information derived from the literature and publicly available large-scale profiling datasets. Relevant resources that have been employed and generated in the context of this work are presented in Table 1.

Table 1. List of online resources employed for profiling CEACAM6 literature/transcriptional data, including those generated as part of the present work.

Resource name/ Description	Use	Link	Reference
CEACAM6 Interactive Circle Packing Plot	Aggregation and dissemination of information derived from the literature and transcriptional data profiling efforts	https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/	Present work
Generic information capture form	Excel spreadsheet employed for structuring relevant information captured via the screening of CEACAM6 literature or transcript abundance profiles	https://doi.org/10.6084/m9.figshare.21183718.v1	Present work, Extended Data File 1[20]
Prevalence of disease or cell type entities in the CEACAM6 literature	Prioritization of disease or cell types associated with CEACAM6	https://doi.org/10.6084/m9.figshare.21183748.v1	Present work, Extended Data File 2[30]
Information captured from the literature identifying CEACAM6 as a candidate biomarker	This information can be used as a basis for deciding whether to include CEACAM6 in a targeted panel	https://doi.org/10.6084/m9.figshare.21183832.v1	Present work, Extended Data File 3[31]
Information captured from the literature identifying instances where abundance levels of CEACAM6 blood transcripts differ between cases and controls	This information can be used as a basis for deciding whether to include CEACAM6 in a targeted panel	https://doi.org/10.6084/m9.figshare.21184357.v1	Present work, Extended Data File 4[56]
Transcript abundance measurements for CEACAM6 across 16 reference transcriptome datasets	Determining CEACAM6 differential expression and generating graphical representations	https://doi.org/10.6084/m9.figshare.21184363.v1	Present work, Extended Data File 5[57]
Information captured from reference blood transcriptional datasets for which CEACAM6 transcripts were found to differ between cases and controls	This information can be used as a basis for deciding whether to include CEACAM6 in a targeted panel	https://doi.org/10.6084/m9.figshare.21184369.v1	Present work, Extended Data File 6[58]

Table 1. *Continued*

Resource name/ Description	Use	Link	Reference
Gene Expression Browser (GXB) CD2K instance	Access CEACAM6 abundance profiles across multiple reference datasets	http://cd2k.gxbsidra.org/dm3/geneBrowser/list	21
Single Cell Portal	Identification of scRNAseq datasets where CEACAM6 expression is elevated in one or several cell clusters	https://singlecell.broadinstitute.org/single_cell	62

Briefly, the process is broken down into the following steps:

(1) Selecting a candidate gene: the most basic criterion is for transcripts for this gene to be detectable in blood. It could also be selected based on its membership in a pre-defined signature or gene set.

(2) Retrieving background information: background information about the gene is gathered from reference datasets (e.g., OMIM [https://www.omim.org/], UniProt [https://www.uniprot.org/], Entrez Gene [https://www.ncbi.nlm.nih.gov/gene]) and the introduction section of recent publications.

(3) Profiling the candidate gene's literature at a high level: the literature associated with the candidate gene is identified (see "literature profiling section" below for details). Entities corresponding to a given theme (e.g., diseases, cell types, or molecular processes) are extracted from the title of those articles ("breast cancer" is an example of a disease entity). This permits to identify the main diseases associated with the gene of interest, and, in turn, identify instances in which the candidate gene has been found to be of actual or potential utility as a biomarker for these diseases.

(4) Profiling the literature in more depth: taking advantage of Google Scholar's full text search capabilities, this step identifies publications where the abundance level of the candidate gene's transcripts in blood samples was found to be different in patients compared with appropriate controls.

(5) Profiling the abundance of the gene across multiple relevant transcriptome datasets: to complement the previous step, public blood transcriptome datasets are screened to identify instances where the abundance level of the candidate gene's transcripts in blood differs in patients in comparison with appropriate controls.

(6) Developing resources supporting manuscript preparation and evaluation of the candidate gene: the information parsed from the literature or transcriptome datasets in earlier steps is recorded in a structured format (e.g., using a standard spreadsheet template, see details below). Using the Prezi web application (Prezi Inc., San Francisco, CA, USA), this information is aggregated in interactive circle packing plots. Spreadsheets and interactive circle plots can next be used to assess the overall relevance of the gene of interest as a candidate blood transcriptional biomarker and support the writing of the manuscript. They can also serve as a resource for investigators interested in designing blood transcriptional biomarker panels.

BloodGen3 blood transcriptional module repertoire

CEACAM6 was selected based on its membership to one of the 382 modules constituting the fixed BloodGen3 module repertoire. This repertoire has been recently characterized.[17] Briefly, it was constructed based on co-expression analysis through a process that was exclusively data-driven. First, the 16 reference blood transcriptome datasets that served as input were clustered separately using K-means clustering. Co-clustering events observed across the 16 reference datasets were then recorded for each gene pair. This information served as a basis for the constitution of a large co-clustering network, with nodes representing genes and edges representing co-clustering events. A weight of 1 to 16 was attributed to the graph edges depending on the number of times co-clustering events were observed. The network was then mined using graph theory to identify densely connected subnetworks that were identified as modules and added to the repertoire. This process eventually yielded 382 non-overlapping modules (at the probe level, multiple probes mapping to the same gene could be found across different modules). Next, the repertoire was thoroughly characterized functionally and an R package was developed to support BloodGen3 module repertoire analysis and visualization.[18]

Literature profiling

The approach has been described in two published study guides: from a high-level perspective as part of the COD1 workflow[10] and in more detail in a separate study guide dedicated to literature profiling.[19] An overview of the steps implemented in the profiling of the literature associated with CEACAM6 is provided here:

(1) Literature retrieval: to identify the literature associated with the candidate gene, a PubMed query is designed by combining the official gene name and symbol along with known aliases. Troubleshooting is performed as needed to minimize false positives and false negatives. For CEACAM6 the following query was generated and, as of August 16 2022, returned 642 entries:

CEACAM6 [tiab] ORc "CEA Cell Adhesion Molecule 6" [tiab] OR CD66c [tiab] OR (NCA [tiab] AND (Carcinoembryonic OR CEACAM6 OR CD66c)) OR "Carcinoembryonic Antigen-Related Cell Adhesion Molecule 6" [tiab] OR "Carcinoembryonic Antigen Related Cell Adhesion Molecule 6" [tiab] OR "Carcinoembryonic Antigen-Related Cell Adhesion Molecule 6" [tiab] OR ("Normal Cross-Reacting Antigen" [tiab] AND (Carcinoembryonic OR CEACAM6 OR CD66c)) OR ("Non-Specific Cross-reacting Antigen" [tiab] AND (Carcinoembryonic OR CEACAM6 OR CD66c)) OR (CEAL [tiab] AND (Carcinoembryonic OR CEACAM6 OR CD66c)) NOT review [pt]

(2) Extraction of relevant concepts: the titles of the articles associated with CEACAM6 are screened for keywords associated with diseases or physiological states and with cell types. For example, if the theme is "diseases or physiological states", diseases entities such as "breast cancer", "influenza infection", "pregnancy" or "systemic lupus erythematosus" may be identified in the title of articles associated with the gene of interest.

(3) Generating literature profiles: next, the prevalence of the cell types or disease entities identified in the previous step in the candidate gene's literature is determined. Focusing on a subset of the literature, information regarding the potential relevance of the candidate gene as a biomarker can be captured in a structured format in an Excel spreadsheet.

(4) Aggregating information: the underlying literature profiling information is captured and visually represented in interactive circle packing plots using the Prezi application (Prezi Inc, San Francisco, CA, USA). This serves as a basis for generating manuscript figures and the constitution of a companion resource that can be made accessible to the community.

Information retrieval and structuring

While screening the literature and large-scale profiling datasets trainees learn to identify and extract key information from research articles or transcriptome datasets. These include basic information, as well as elements of study design (e.g., analyte name, type, species, biological samples, measurement methods, sample size) and findings (e.g., fold change, significance). The information is captured in a standard MS Excel spreadsheet template, which can be used to record information derived from both the literature and transcriptome profiling datasets (**Extended Data File 1**[20]).

Interactive circle packing plots

Information extracted from the literature and from public transcriptome datasets was aggregated in an interactive circle packing plot generated using the Prezi web application (Prezi Inc., San Francisco, CA, USA). A free basic Prezi account can be setup for this (https://prezi.com/pricing/basic/). Starting from a blank presentation, it consisted of adding and populating circles (topics) and organizing them into a hierarchy (https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/). Color-coding the circles and varying their size permitted the visualization of some of the results. Excerpts or full articles were added, as well as plots representing CEACAM6 transcriptional data profiles. Links to articles and interactive versions of the figures were also provided in order promote seamless access to information.

Transcriptome profiling data analyses and visualization

Screening of transcriptome profiling datasets consisted of determining whether differences between levels of CEACAM6 transcript abundance in patients and their respective controls were significant. The CEACAM6 profiling data were downloaded from the "CD2K" gene expression browser (GXB) instance (http://cd2k.gxbsidra.org/dm3/geneBrowser/list) for multiple blood transcriptome datasets.[21] Analyses were conducted separately for each dataset in Microsoft Excel (RRID:SCR_016137), testing for differences in variance using F-test statistics and testing for differences in expression using t-test statistics. Differences were considered significant when p was <0.05. Plots were generated using Plotly chart studio (RRID:SCR_013991, https://chart-studio.plotly.com/create/).

Results

Selection of CEACAM6

The first step consisted of selecting a gene that would be next evaluated for its potential relevance as a blood transcriptional biomarker. CEACAM6 was selected primarily based on its membership to a blood transcriptional signature of interest. This signature is part of a fixed blood transcriptional module repertoire (BloodGen3, see Ref. 17 and methods for details). The M10.4 module signature is functionally associated with neutrophil activation and comprises 11 other genes: BPI, LTF, CEACAM8, DEFA1, DEFA1B, DEFA2, DEFA4, OLFM4, ELANE, CTSG, and MPO (https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/ Step 1: candidate gene selection). In a reference collection of 16 patient cohorts,[17] abundance levels of M10.4 transcripts were the highest in subjects with *Staphylococcus aureus* infection, respiratory syncytial virus infection and bacterial sepsis (Figure 1).

General background information about CEACAM6

As part of the evaluation process, it can be useful to start by retrieving and synthesizing background information about the candidate gene. For this, summaries from different reference databases, as well as introductions from recent publications on CEACAM6, were retrieved. This information was recorded in the CEACAM6 interactive circle packing plot (https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/ Step 2: gathering background information) and used for development of the narrative below.

CEACAM6 is a glycosyl phosphatidyl inositol (GPI)-anchored cell surface glycoprotein. It is a member of the carcinoembryonic antigen (CEA) family whose members are known to play a role in cell adhesion.[22] Specifically, CEACAM6 expression has been reported in granulocytes and lung and intestinal epithelial cells.[23] In ileal epithelial cells of patients with Crohn's disease, CEACAM6 has been found to act as a receptor for adherent-invasive *Escherichia coli*.[24] It has also been found to mediate entry of *Neisseria gonorrhoeae*.[25] CEA family members are widely used as tumor

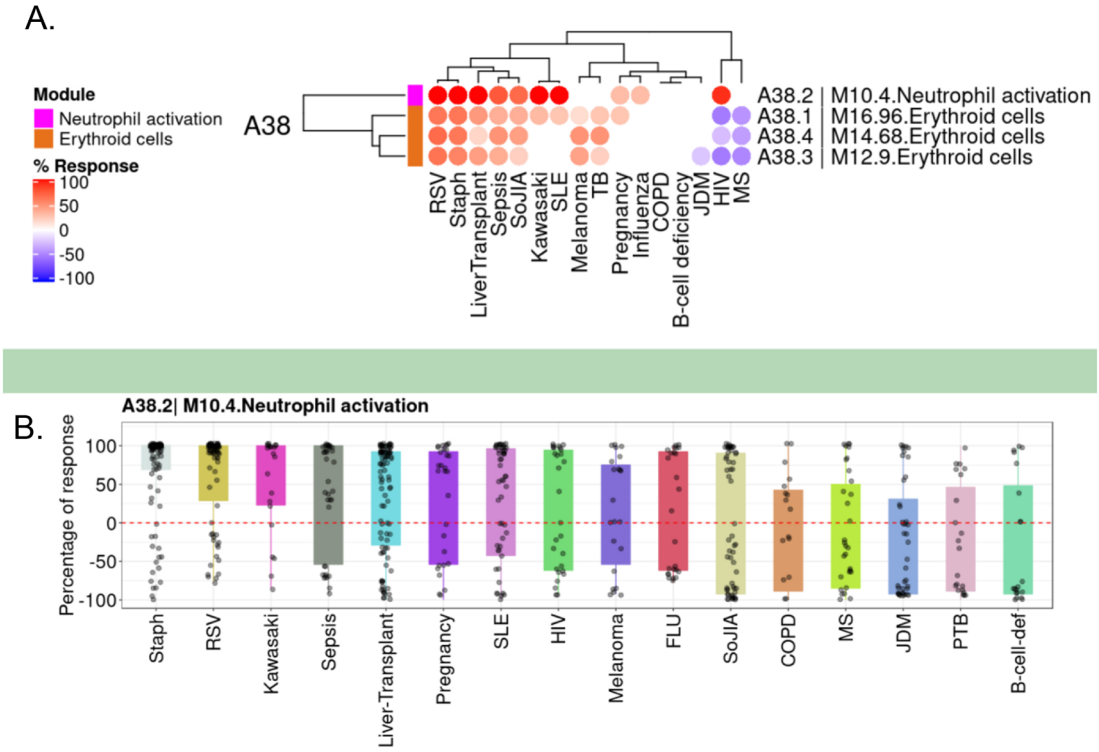


Figure 1. Differences in transcript abundance levels for BloodGen3 module M10.4 across 16 reference datasets. A. The module fingerprint heatmap represents the proportion of transcript for a given module (rows) for which abundance levels are significantly different in case subjects compared to the respective controls for a given reference dataset (columns). Values can range from +100% (solid red: abundance for all constitutive transcripts for the module are significantly higher) to −100% (solid blue: abundance for all constitutive transcripts for the module are significantly lower). Responses are shown for four modules included in the module aggregate A38 from the BloodGen3 repertoire,[17] including module M10.4 from which CEACAM6 was selected. B. The box plot represents the percentage response averaged for module M10.4, across the 16 reference datasets (we have contributed this dataset collection to GEO as part of an earlier work,[17] and it is accessible under accession number GSE100150. Plots were generated using the BloodGen3 web application: https://drinchai.shinyapps.io/BloodGen3Module/.

markers in serum as well as tumor immunoassays. CEACAM6 has been reported to act as an oncogene, promoting tumor progression and metastasis.[26] These properties may, at least in part, be effected via the role of CEACAM6 in promoting anoikis resistance, which prevents the homeostatic elimination of anchorage-dependent cells (such as epithelial cells) that are detached from the cellular matrix.[27] Since CEACAM6 membrane expression is highly specific to tumor cells, it has been suggested as a target for different cancer immunotherapies.[28] It has also recently been identified as an immune checkpoint molecule, based on its role in suppressing cytotoxic T cell responses against malignant plasma cells.[29]

Profiling the CEACAM6 literature at a high-level reveals an association with neutrophils and several types of cancers

To further our understanding of the biological significance and clinical relevance of CEACAM6, we next sought to systematically screen the literature to identify associations with cell populations and diseases or physiological states.

A query was designed to permit the retrieval of the literature associated with CEACAM6 (see methods for details). In total 642 PubMed entries were returned. Screening for names of diseases in the titles of literature associated with CEACAM6 identified 18 entities (**Extended Data File 2**[30]). Among these, "cancer" and "colorectal cancer" were found in more than 50 CEACAM6-associated articles (202 and 65, respectively, as of March 2022). "Pancreatic cancer", "lung cancer", "breast cancer", leukemia" and "inflammatory bowel disease" were found in more than 20 CEACAM6-associated articles (31, 35, 28, 35, and 30, respectively; Table 2, Figure 2A & https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/: Step 3/CEACAM6_Diseases). "Pregnancy" was found in 14 CEACAM6-associated articles. "Cholangiocarcinoma" and "myeloma" were found in more than 5 articles (7 and 9, respectively). Eight other diseases were found in only one article. Screening titles for names of cell types identified 10 entities (**Extended Data File 2**[30]). The most frequently mentioned cell types among the CEACAM6 literature were granulocytes, neutrophils, T-cells and intestinal epithelial cells (Table 2, Figure 2B & https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/: Step 3/CEACAM6_Cell Types).

Altogether, this step established that CEACAM6 is associated with a large body of literature. It also permitted the identification of the main cell types and diseases associated with this gene. This information was used in subsequent literature profiling steps.

CEACAM6 is of potential clinical relevance in the diagnosis of cancers, in particular, the early detection of colorectal carcinoma

The selection of a blood transcriptional panel could take into consideration whether a given candidate gene has already been determined to be of clinical relevance as a biomarker, whether that is at the gene, transcript, or protein level. Thus, we next sought to determine if this was the case for CEACAM6 by extracting relevant information from its literature for the main disease entities identified in the previous steps.

The approach is described in detail in the methods section. In brief, starting from the CEACAM6-associated literature we searched for publications reporting the actual or potential use of CEACAM6 as a biomarker. For this we focused more specifically on the diseases that showed the highest degree of association with CEACAM6 based on the above literature profiling results (i.e., diseases mentioned in more than 20 articles, which are listed in Table 2), namely:

Table 2. List of the most prevalent diseases/physiological states and cell types found among the CEACAM6 literature.

Themes	Entities	N articles	% CEACAM6 Literature
Diseases/physiological states	Colorectal cancer	65	10.1%
Diseases/physiological states	Pancreatic cancer	31	4.8%
Diseases/physiological states	Lung cancer	35	5.5%
Diseases/physiological states	Leukemia	35	5.5%
Diseases/physiological states	Inflammatory bowel disease	30	4.7%
Diseases/physiological states	Breast cancer	28	4.4%
Cell types	Granulocytes	66	10.3%
Cell types	Neutrophils	43	6.7%
Cell types	T-cells	27	4.2%
Cell types	Intestinal epithelial cells	26	4%

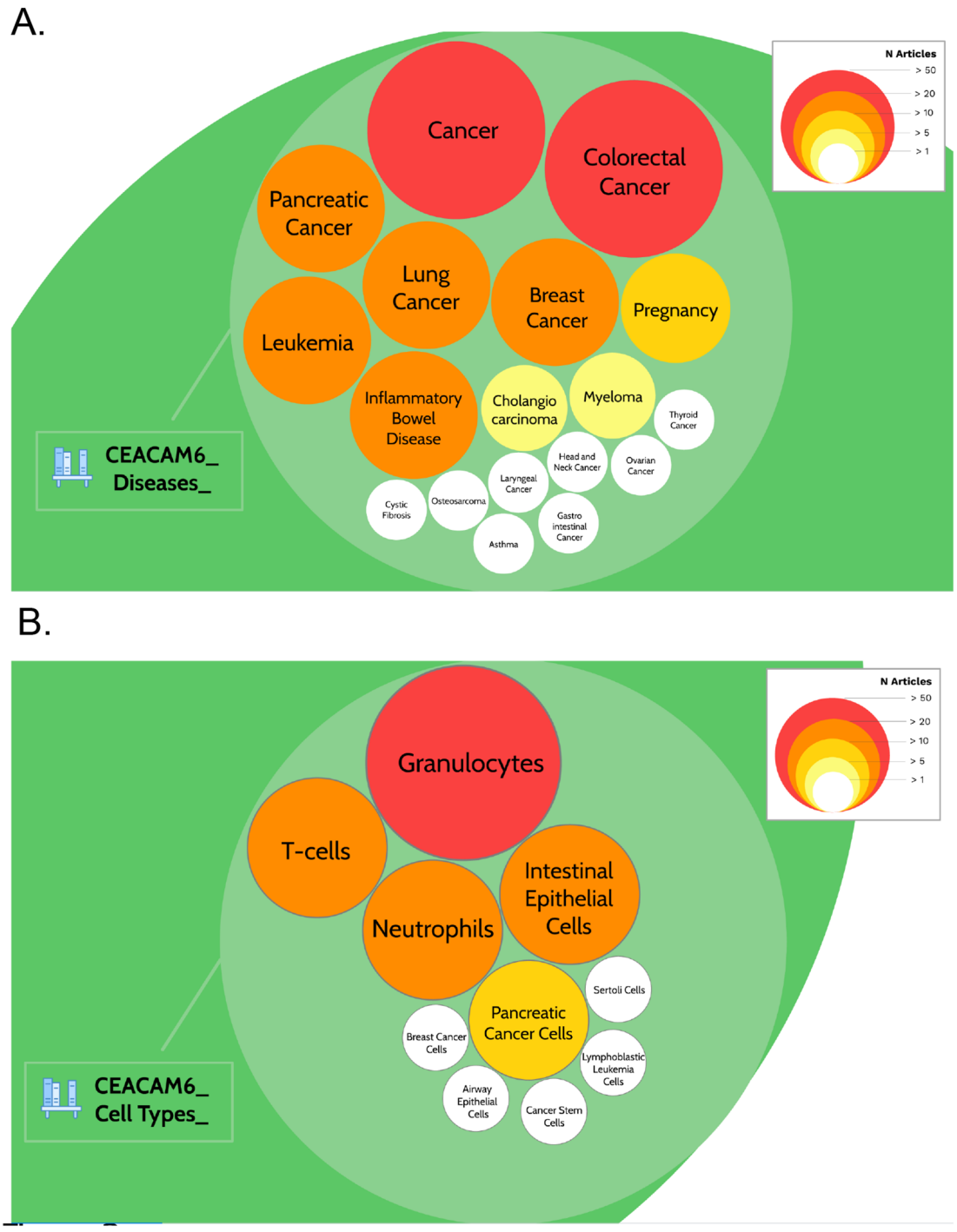


Figure 2. CEACAM6 disease and cell type literature profiles. The prevalence of articles among the literature associated with CEACAM6 for disease entities (A) or cell type entities (B) are represented by circles of different sizes and colors, corresponding to the number of associated articles. It is possible to access underlying information by zooming into each of the circles. The Prezi presentation can be accessed at this url: https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/ Step 3: background literature profiling.

leukemia, colorectal, pancreatic, lung, and breast cancers, as well as Inflammatory bowel disease. Next, articles associated with CEACAM6 and these diseases that also mentioned "biomarker", "diagnostic", "diagnosis", "prognostic" OR "prognosis" in their title or abstract were retrieved. For articles deemed to be of interest, a standard spreadsheet template was used to capture relevant information (**Extended Data File 3**[31]). Information was also aggregated in an interactive circle packing plot using the Prezi web application (https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/

Table 3. Published reports describing CEACAM6 as being of clinical relevance as a biomarker.

Immune state/ pathology	Evidence	Analyte	Sample type	Change	PMID/Ref	Clinical relevance
Colorectal cancer	Literature	mRNA	Blood	Increase	29352642, 26993598[14–16]	Early detection
Colorectal cancer	Literature	mRNA	Tumor	Positive	27042567, 22975528[33,77]	Stem cell marker, Worse prognosis
Colorectal cancer	Literature	Protein	Tumor	Positive	22975528, 14512395[32,33]	Worse prognosis
Pancreatic cancer	Literature	mRNA	Tumor	Positive	34321959[36]	Worse prognosis
Pancreatic cancer	Literature	Protein	Serum	Positive	34207784, 25409014[37,38]	Worse prognosis, distant metastases

CEACAM6/Step3: background literature profiling/CEACAM6_Diseases_Biomarker). Together, the information thus gathered served as a basis for the development of the narrative below.

As aforementioned, CEACAM6 has been noted for its oncogenic properties. Our screening of the CEACAM6 literature, which relates more specifically to its potential relevance as a biomarker in various disease settings, supports this notion. Indeed, a higher abundance of CEACAM6, whether at the transcript or protein level, in tumor tissues or serum was always associated with worse survival (in the case of colorectal,[32,33] breast,[34,35] pancreatic,[36–40] and lung cancers[41]). Other studies have found CEACAM6 to be of potential value for differential diagnosis of malignant vs benign tumors for breast cancer (with CEACAM6 protein levels measured in breast tissues[42]) and pancreatic cancer (with CEACAM6 protein levels measured in the bile[43]). Notably, and of particular relevance to this report, in the case of colorectal carcinoma, measuring the abundance of CEACAM6 at the protein and transcript levels in blood alongside TSPAN8, LGALS4, and COL1A2 has been found to be of potential value for early disease detection.[14,15] Furthermore, recently CEACAM6 was also included in a 10-gene signature predictive model for lung cancer prognosis.[44]

Altogether, this review of the literature shows that measurement of CEACAM6, whether at the transcript or protein level, in tumor tissues or in blood, is considered of potential clinical value in informing the management of different types of cancers, as summarized in Table 3.

In depth screening of the literature shows that blood levels of CEACAM6 transcripts are elevated in a wide range of diseases

More specifically we next sought to assess the relevance of CEACAM6 as a blood transcriptional biomarker. The first pass at screening the literature (above) already identified instances where measuring blood CEACAM6 transcript is deemed of potential clinical value (i.e., for the early detection of colorectal cancer[14–16] or the prognosis of lung cancer[44]). We wanted to undertake a second pass to profile the literature in more depth to identify additional studies that reported differences in the abundance of CEACAM6 transcripts in blood in patient populations.

Queries were run using Google Scholar, which supports full text search. Entries were screened manually, selecting only peer-reviewed reports where CEACAM6 levels were measured in the blood of human subjects. Relevant information was recorded in a structed format in a spreadsheet using the standard template employed in the previous step. Finally, information was aggregated in the interactive CEACAM6 Prezi circle packing plot.

Differences in CEACAM6 blood transcript levels have been reported in the literature for a wide range of pathologies. Specifically, in addition to the colorectal carcinoma and lung cancer studies described above, it was found to be part of a 13-gene disease signature which was increased in patients with Parkinson's disease as compared with asymptomatic subject.[45] It was also part of a different 13-gene disease signature that was increased in patients with severe idiopathic pulmonary fibrosis compared with patients with a mild form of the disease.[46] Notably, other members of this latter signature, including CTSG, DEFA3, and OLFM4, are also comprised in the M10.4 module that is part of the fixed BloodGen3 repertoire mentioned above. Other pathologies and states where blood CEACAM6 transcript levels were found to be increased are summarized in Table 4, and include asthma,[47] sepsis,[48] post-traumatic stress disorder,[49] psoriasis,[50] maternal anti-fetal rejection,[51] and COVID-19.[52,53] It was also found to differ based on gender (higher in male than in females)[54] and notably was also increased by steroid treatment.[55] These latter two findings suggest that in instances where demographics or use of steroids are not well-controlled for in the study design, differences in CEACAM6

Table 4. Pathological, immunological, or physiological states where CEACAM6 transcript abundance levels have been found to differ in cases vs controls.

Disease/ physiological state	Evidence	Analyte	Sample type	Abundance levels	PMID/GEO ID
Parkinson's disease	Literature	mRNA	Blood	Higher	25475535[45]
Idiopathic pulmonary fibrosis	Literature	mRNA	Blood	Higher in severe vs mild cases	22761659[46]
Psoriasis	Literature	mRNA	Blood	Higher	34639156[50]
Colorectal cancer	Literature	mRNA	Blood	Higher	29352642, 26993598[14–16]
Gender difference	Literature	mRNA	Blood	Higher in males	31722210[54]
Sepsis non-survivors	Literature	mRNA	Blood	Lower in non-survivors vs survivors	34707398[48]
Lung cancer	Literature	mRNA	Blood	Higher levels in patients with poor outcomes	34288383[44]
Post-traumatic stress disorder (PTSD)	Literature	mRNA	Blood	Higher levels in PTSD cases associated with increased inflammation vs those without	31698278[49]
Maternal anti-fetal rejection	Literature	mRNA	Blood	Lower in fetuses showing evidence of fetal inflammatory response	23905683[51]
Steroid treatment	Literature	mRNA	Blood	Higher in patients with Duchenne muscular dystrophy treated with steroids vs those who were untreated	33751844[55]
COVID-19	Literature	mRNA	Blood	Higher	35844004[53]
COVID-19	Literature	mRNA	Blood	Higher	34335605[52]
Asthma	Literature	mRNA	Blood	Higher	27925796[47]
Food-induced anaphylaxis	Literature	mRNA	Blood	Higher	26194548[78]
Early onset pre-eclampsia	Literature	mRNA	Blood	Lower in patients with early onset pre-eclampsia vs control pregnant subjects	23793063[79]
Late onset pre-eclampsia	Literature	mRNA	Blood	Lower in patients with late onset pre-eclampsia vs control pregnant subjects	23793063[79]
Female patients with Systemic onset Juvenile Idiopathic Arthritis	Literature	mRNA	Blood	Higher abundance levels in Female SoJIA patients	32794262[80]
Kawasaki disease	Public dataset	mRNA	Blood	Higher	GSE100154
Sepsis	Public dataset	mRNA	Blood	Higher	GSE100159
Systemic lupus erythematosus	Public dataset	mRNA	Blood	Higher	GSE100163
S. aureus infection	Public dataset	mRNA	Blood	Higher	GSE100165
Pregnancy	Public dataset	mRNA	Blood	Higher	GSE100157
Liver transplant recipients	Public dataset	mRNA	Blood	Higher	GSE100155

Table 4. *Continued*

Disease/ physiological state	Evidence	Analyte	Sample type	Abundance levels	PMID/GEO ID
Influenza infection	Public dataset	mRNA	Blood	Higher	GSE100160
HIV infection	Public dataset	mRNA	Blood	Higher	GSE100151
RSV infection	Public dataset	mRNA	Blood	Higher	GSE100161

transcript levels might be, at least in part, attributed to these factors rather than the underlying pathology. For reference, a full record of the information captured from the literature regarding those studies can be found in **Extended Data File 4**.[56] Additional information is also found aggregated in the CEACAM6 interactive circle packing plot (https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/ CEACAM6/Step3: background literature profiling/CEACAM6_Diseases_Biomarker).

Taken together, this in-depth review of the literature points to differences in CEACAM6 blood transcript abundance being present in patients in a wide range of diseases. Thus, suggests that assays measuring levels of CEACAM6 transcripts in blood may be employed to support biomarker development efforts across different clinical settings.

Screening of public blood transcriptome datasets to identify elevated levels of CEACAM6 in additional disease settings

Literature reports might capture only a fraction of instances where pathophysiological changes are accompanied by changes in the abundance of CEACAM6 blood transcripts. Screening publicly available transcriptome datasets could confirm published reports and help identify other instances where levels of CEACAM6 transcript abundance differ in patients relative to control subjects.

For this, we employed a data browsing web-application, the Gene eXpression Browser (GXB),[20] which provides easy access to transcriptional profiles of individual genes in curated collections of transcriptome datasets. For instance, we screened blood transcriptome data for a collection of 16 reference cohorts that were used for the construction of the BloodGen3 repertoire. These datasets are available in the CD2K instance of GXB (http://cd2k.gxbsidra.org/dm3/geneBrowser/list). CEACAM6 transcriptional profiles were retrieved for each of these cohorts and statistics run separately using MS Excel to determine the significance of changes in levels of CEACAM6 transcripts in patients vs controls (**Extended Data File 5**[57]). Changes were captured in a structured format, plotted, and aggregated in the CEACAM6 circle packing plot.

We found differences in levels of CEACAM6 transcript abundance for nine of the 16 reference BloodGen3 datasets (Table 4, **Extended Data File 6**[58]). The pathological or physiological states for which differences were observed did not overlap with those also listed in Table 4 that were identified in the previous step by in depth screening of the literature. Indeed, we found elevated abundance levels of CEACAM6 in patients with infections caused by *Staphylococcus aureus*, influenza, respiratory syncytial virus, human immunodeficiency virus, and bacterial pathogens causing sepsis, in comparison with controls (Figure 3). CEACAM6 transcript levels were not increased in patients with tuberculosis. Significant increases were also observed in non-communicable diseases such as systemic onset juvenile arthritis and Kawasaki disease but not in the context of systemic lupus erythematosus, late-stage melanoma, or chronic obstructive pulmonary disease. Finally, we also found a significant increase in abundance in the blood of liver transplant recipients under immunosuppressive therapy and in pregnant women. This transcriptome profiling dataset screen complemented our earlier literature screen, identifying nine additional diseases or physiological states in which CEACAM6 transcript is significantly changed in the blood of patients, for a total of 25 distinct diseases/states which are listed in Table 4. Plots for the nine BloodGen3 datasets are available via the GXB application and have been replotted and loaded to the CEACAM6 circle packing plot (https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/ CEACAM6/Step5: blood tx profiling/CEACAM6_Blood Tx).

Overall, the screening of a reference dataset collection indicated that differences in CEACAM6 levels could be observed in a wide range of conditions in which systemic inflammation is observed. The lack of overlap between the literature and transcriptome data profiling conducted in steps 4 and 5 suggests that expanding this search to a larger number of blood transcriptome datasets would likely significantly add to this list.

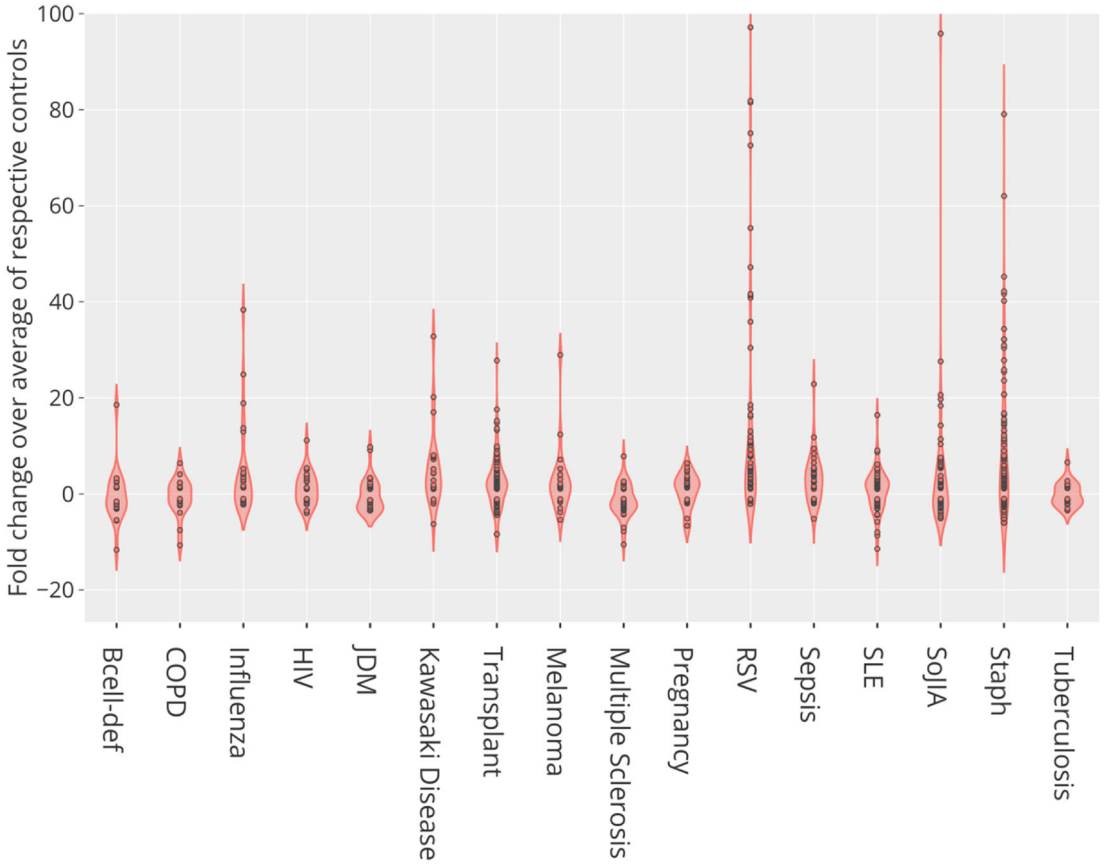


Figure 3. CEACAM6 relative abundance profiles across reference patient cohorts. This violin plot shows the fold change in abundance of CEACAM6 mRNA measured by RNA sequencing in the blood of human subjects across 16 reference disease cohorts, compared to their respective control subjects. In total blood transcriptome of 985 subjects was profiled. For details, see original work by Altman, Rinchai *et al.*[17] GEO deposition: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100150; plot: https://chart-studio.plotly.com/create/?fid=dchaussabel%3A132 interactive version with values expressed as counts: http://cd2k.gxbsidra.org/dm3/miniURL/view/NH.

To date, no drugs have been developed that target CEACAM6

Another criterion for inclusion of CEACAM6 in a focused assay could be its targeting by approved drugs or drugs currently under development. The "Open Targets" database does not report any known drugs, approved or currently under development, targeting CEACAM6 (https://platform.opentargets.org/target/ENSG00000086548). However, given its recently described role as suppressor of effector CD8 T-cells,[29] CEACAM6 is currently considered an immune checkpoint molecule and as such could be targeted by drugs designed to block its activity in cancer patients.[28] Additionally, in preclinical mouse models antibodies targeting CEACAM6 have been shown to inhibit tumor growth and metastasis.[26,59]

Profiling reference transcriptome datasets shows CEACAM6 transcript expression to be restricted to circulating neutrophils

Finally, screening of reference public transcriptome datasets can also yield insights regarding the candidate gene's regulation and restriction among circulating leukocytes. Thus, in addition to profiling 16 public blood transcriptome datasets, we examined CEACAM6 transcriptional profiles in two other reference datasets. One dataset measured transcript abundance in monocytes, neutrophils, B-cells, CD4+ T-cells, CD8+ T-cells and natural killer (NK) cells and in whole blood (GSE60424[60]). The second dataset measured changes in transcript abundance in whole blood exposed *in vitro* to a wide range of immune stimuli (toll-like receptor agonists, killed bacteria, viruses, inflammatory cytokines and interferons; GSE30101[61]). In addition, we screened the Broad Institute's single cell portal[62] for datasets in which CEACAM6 expression was elevated in one or more of the cell clusters.

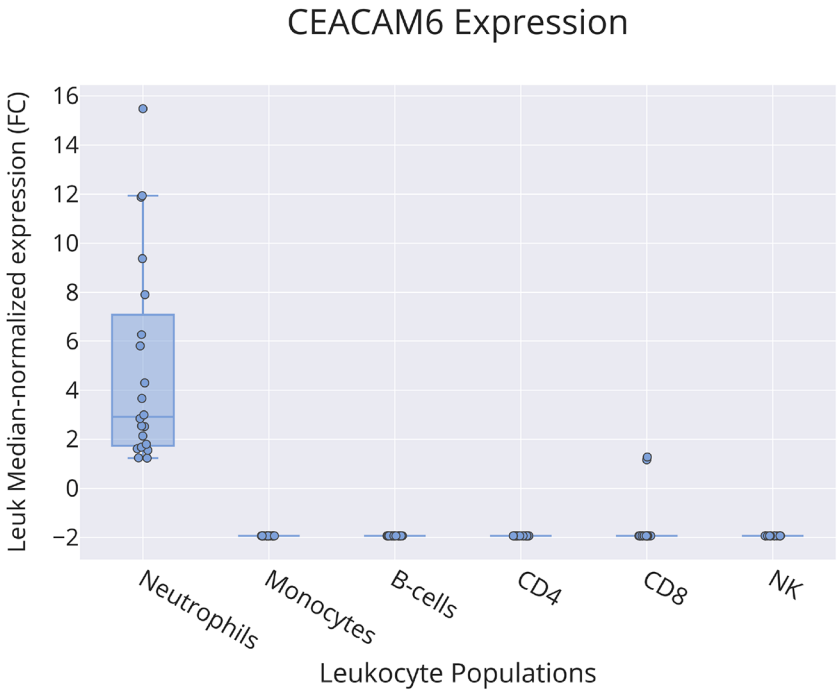


Figure 4. CEACAM6 restriction among circulating leukocyte populations. This box plot shows levels of abundance of CEACAM6 RNA measured by RNA sequencing in neutrophils, monocytes, B-cells, CD4+ T-cells, CD8+ T-cells and NK cells purified from the blood of human subjects, including patients with ALS, type 1 diabetes, multiple sclerosis (immediately before and 24 hours after initiation of beta interferon therapy) or sepsis and healthy controls. Values are normalized to the median calculated across all conditions. For details, see original work by Linsley *et al.*[60] GEO deposition: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60424 plot: https://plotly.com/~dchaussabel/171/.

Bulk leukocyte population RNAseq data showed CEACAM6 expression to be restricted to neutrophils (Figure 4) [data source: Linsley *et al.*[60]]. This observation was confirmed in a single-cell dataset in which tumor immune cell infiltrates were dissociated and profiled via RNA sequencing (Figure 5) [data source: He *et al.*[63]]. These findings were in line with the prevalence among the CEACAM6 literature of publications mentioning this cell type (https://prezi.com/view/pQ7TKEC6tgY3cuik9ckt/ CEACAM6/Step 3: background literature profiling/CEACAM6_Cell Types) (Figure 2A). However, we did not find CEACAM6 to be increased in whole blood stimulated *in vitro* (Figure 6) [data source: Obermoser *et al.*[61]]. This finding was to some extent surprising since blood signatures comprising CEACAM6 are often functionally associated with neutrophil activation.[64–66]

Taken together, further profiling of reference transcriptome datasets confirmed the close association of CEACAM6 with neutrophils, which is the most abundant circulating leukocyte population in blood. It also indicates that elevated levels of CEACAM6 transcript abundance observed across a wide range of conditions may be associated with an increase in relative abundance of cells expressing this gene, rather than regulation of its expression.

Discussion

Clinical translation of biomarker signatures obtained via transcriptome profiling technologies typically involves the development of targeted transcript panels and assays. Such assays can also prove more practical for high-temporal frequency immunological monitoring applications that require profiling of thousands of samples. They could also be more readily implemented in the context of research projects conducted in low-resource settings. Targeted panel design can be informed by both data-driven and knowledge-driven approaches. However, given the large amounts of data and knowledge available for any given candidate gene, the selection process can prove daunting. Here we employed a workflow devised for screening the literature and large-scale profiling data associated with a given candidate gene, and to retrieve and aggregate relevant information in a structured format. This information and associated resources should in turn support decision-making of investigators aiming to develop targeted panels for downstream clinical or research applications.

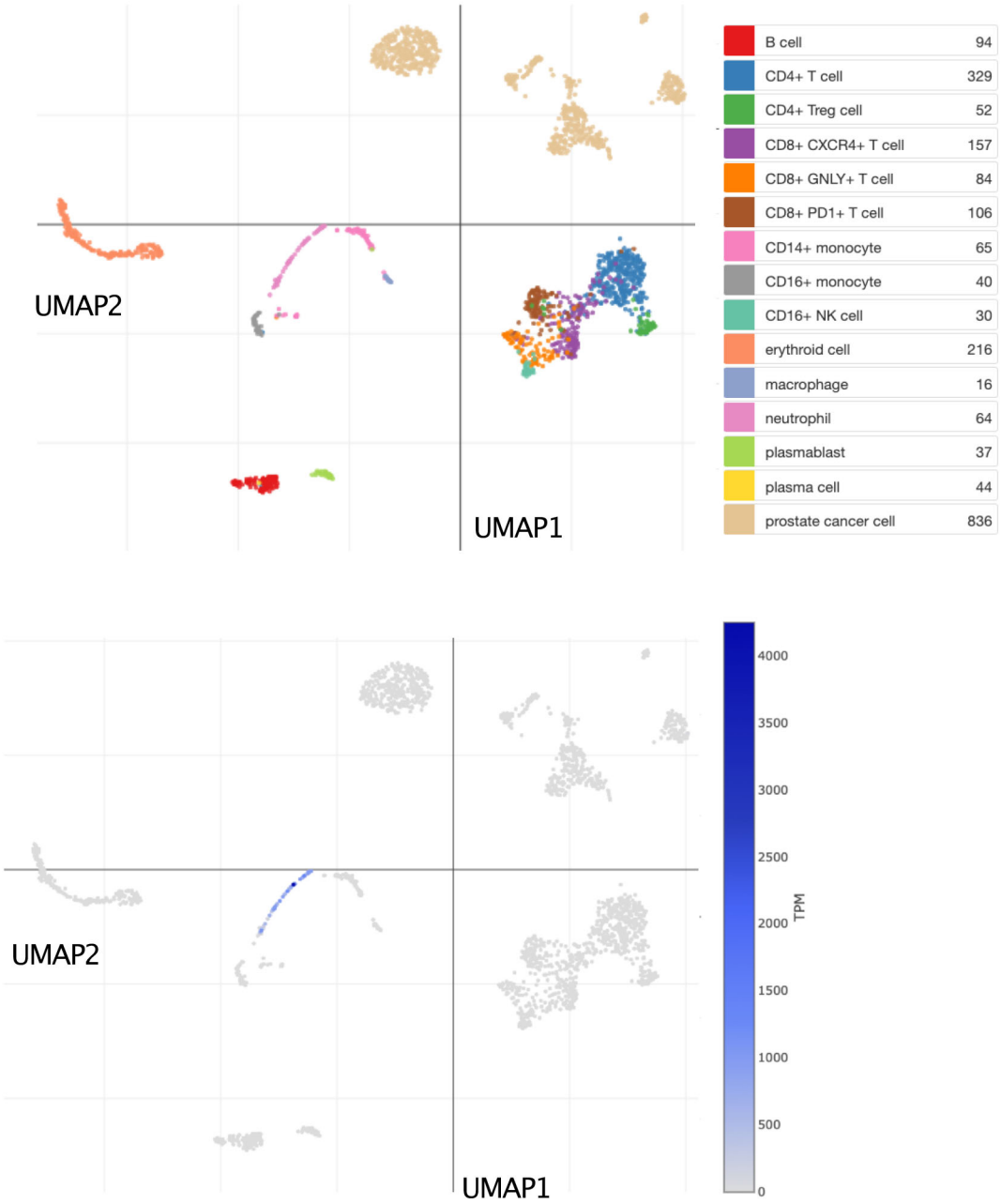


Figure 5. CEACAM6 expression at the single-cell level among dissociated prostate tumor tissue cells. This tSNE plot shows abundance levels of CEACAM6 measured by single-cell RNA sequencing among dissociated metastatic prostate tumor tissue cells. After quality control, this set consisted of 2,170 cells obtained from 14 patients and 15 biopsies. Clusters are labelled for dominant cell type based on marker gene expression on the plot above. Normalized transcript per million (TPM) counts for CEACAM6 are shown in blue on the plot below. For details, see original work by He *et al.*[63] An interactive version of this plot is accessible via the Broad Institute single cell portal: https://singlecell.broadinstitute.org/single_cell/study/SCP1244/transcriptional-mediators-of-treatment-resistance-in-lethal-prostate-cancer?genes=CEACAM6#study-visualize.

We focused on CEA cell adhesion molecule 6 (CEACAM6). This candidate is a member of blood transcriptional signatures that are often functionally associated with neutrophil activation,[64–66] which typically also includes genes encoding constituents of neutrophil granules, such as defensins (DEFA1, DEFA3, DEFA4), myeloperoxidase (MPO), bactericidal permeability increasing protein (BPI), and lactotransferrin (LTF).

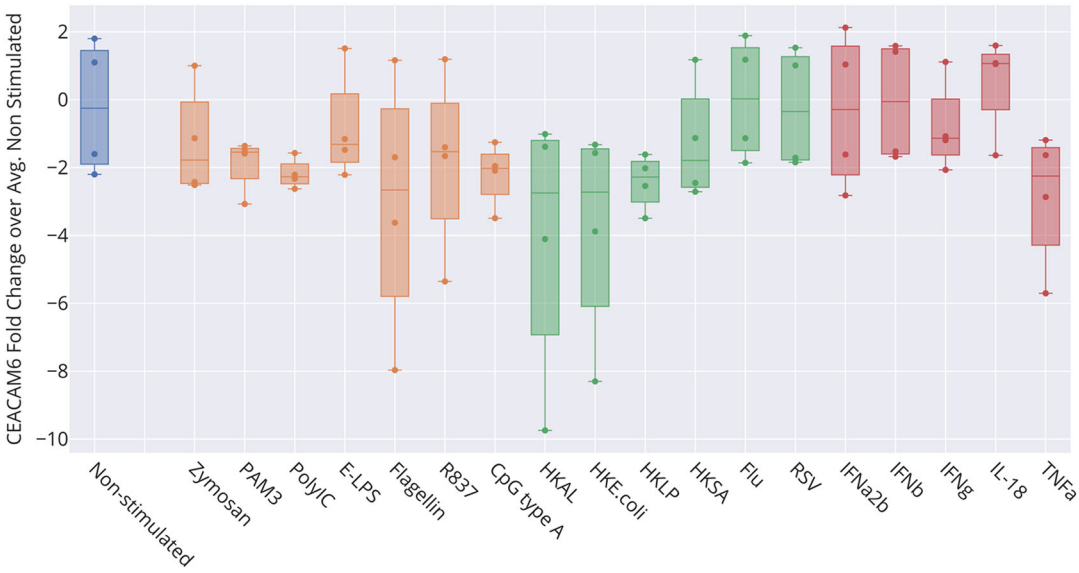


Figure 6. Levels of CEACAM6 transcript abundance in blood exposed to immune stimuli *in vitro*. The box plot represents transcript abundance levels for the CEACAM6 gene in blood samples obtained from four healthy donors, cultivated for 6 hours at 37°C in the presence of pathogens, as well as pathogen-derived and host-derived immune stimuli (HKAL=heat-killed *Acholeplasma laidlawii*, HKLP=heat-killed *Legionella pneumophila*, HKSA=heat-killed *Staphylococcus aureus*, Flu=live influenza A virus, RSV=live respiratory syncytial virus). For details, see original work by Obermoser *et al.*[61] GEO deposition: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30101 plot: https://plotly.com/~dchaussabel/173/.

Several criteria can be used when prioritizing candidate genes for inclusion in a targeted assay, which we have applied here to CEACAM6:

1) Transcripts are detectable in blood and changes can be observed across different immune states/pathologies; this criterion is met in the case of CEACAM6. An increase in levels of CEACAM6 transcripts has been reported in the literature and observed in blood transcriptome datasets for patients with infectious (e.g., bacterial sepsis), autoimmune, or inflammatory diseases (e.g., systemic lupus erythematosus, Kawasaki disease).

2) Previous reports describe the candidate as being of clinical relevance as a biomarker; this criterion is also met. Indeed, CEACAM6 is part of a family that includes members which are used routinely in clinical pathology to assess tumor specimens and inform disease prognosis and treatment.[67–69] CEACAM6 itself is deemed of potential value as a prognosis marker in different types of cancers.[33,34,38,40,41] Notably, measuring blood CEACAM6 transcript abundance is considered of potential value for the early detection of colorectal cancer.[14–16]

3) The functional relevance of the candidate gene in blood leukocytes is known; this criterion is partially met. CEACAM6 is associated with neutrophils in the literature. This was confirmed in our screen of reference transcriptome datasets, both at the bulk leukocyte population and single cell levels (Figures 4 & 5). However, the role played by CEACAM6 in neutrophils has not yet been fully elucidated. For instance, another reference dataset showed that CEACAM6 expression is not regulated in blood exposed *in vitro* to a wide range of immune stimuli (Figure 6). This finding casts some doubts on whether "neutrophil activation" should be assigned to the signature associated with CEACAM6 (by us and others). These observations may also be consistent with an earlier report that associated a "granulopoiesis signature", which comprised CEACAM6, with low density mononuclear and polymorphonuclear populations found in peripheral blood mononuclear cell fractions.[70] Furthermore, single-cell analyses recently conducted in COVID-19 patients identified a population of "developing neutrophils" that expressed neutrophil granule proteins, including module M10.4 members such as MPO, DEFA3, LTF, and ELANE, and were described as potentially being derived from plasmablasts.[71] Altogether these observations suggest that measuring levels of M10.4 transcripts might permit the monitoring of changes in abundance in this population of developing neutrophils rather than reflecting overall neutrophil abundance. However, this hypothesis and the functional relevance of this subset of neutrophils remains to be validated experimentally.

4) The candidate gene is a target for drugs that are approved or under development; Recent studies and ongoing clinical trials have explored the utility of targeting CEACAM6 in various cancers, particularly through the development of monoclonal antibodies. For instance, preclinical evaluations have demonstrated the potential of CEACAM6 as a therapy target in pancreatic adenocarcinoma, utilizing antibody-drug conjugates to effectively target and diminish CEACAM6-expressing tumors.[72] Additionally, the blocking of CEACAM6-CEACAM1 interactions has shown promise in enhancing T cell-mediated cancer cell elimination, suggesting a role for CEACAM6 in immune modulation and its potential as an immune checkpoint target.[73] The breadth of research, encompassing studies on its prognostic value and therapeutic targeting in cancers, underscores CEACAM6's significance in oncology and its emerging role as a viable therapeutic target. These investigations, reflected in various studies[74,75] and a clinical trial registered under NCT03596372, collectively indicate a growing interest in CEACAM6 as a therapeutic target, warranting further exploration and validation in clinical settings.

Alternate candidates may be found that could be selected instead of CEACAM6 for inclusion in a targeted blood transcriptional assay. CEACAM6 was chosen for this evaluation based on its membership to module M10.4, which is part of the fixed BloodGen3 repertoire.[17] Such module repertoires can be employed as a framework for the design of targeted assays, in which case only one or a few representative transcripts from a given module would usually be selected to provide coverage for the entire repertoire (those modules are formed based on co-expression and all constitutive transcripts would present with a high degree of co-linearity).[9] In the case of module M10.4, other candidates to consider would be CEACAM8, BPI, MPO, LTF, DEFA1, DEFA3, DEFA4, CTSG, OLFM4, and ELANE, since all of those genes belong to the same module as CEACAM6 (Table 5). However, to date, only CEACAM6 has been investigated in depth and thus it is not yet possible to benchmark it against these other candidates. However, it can already be noted that BPI (bactericidal/permeability-increasing protein) has been found to be of potential value as a biomarker in patients with asthma,[76] as well as chronic obstructive pulmonary disease.[77] DEFA1 and DEFA3 have been identified as potential inflammatory biomarkers for coronary heart disease.[78] CEACAM8, another member of the carcinoembryonic cell adhesion molecule family, has been found to be of potential value as a prognosis marker in patients with esophageal cancer and in patients with sepsis.[79,80]

Finally, it is worth highlighting some of the limitations of our investigation into the relevance of CEACAM6 as a blood transcriptome biomarker. For instance, it should be noted that the screen conducted among public transcriptome data is not comprehensive. Additional blood transcriptome datasets are available in GEO and other repositories that have not yet been loaded in GXB instances. As a result, the list of conditions in which CEACAM6 blood transcript abundance changes is probably conservative and will likely grow as more datasets become available for screening.

Table 5. Published gene signatures comprising CEACAM6. This table lists targeted gene sets or gene panels comprising CEACAM6. Lists of differentially expressed genes that consists of tens or hundreds of transcripts are purposedly omitted.

Disease/ physiological state	Signature name or description	Gene set	PMID/ Reference
Multiple	BloodGen3/ M10.4	MPO, LTF, BPI, CEACAM6, CEACAM8, DEFA1, DEFA3, DEFA4, CTSG, ELANE, OLFM4	34282143[17]
Parkinson's disease		ADARB2, CEACAM6, CNTNAP2, COL19A1, DEF4, DRAXIN, FCER2, HBG1, NCAPG2, PVRL2, SLC2A14, SNCA, and TCL1B	25475535[45]
Colorectal cancer	CELTiC Panel	LGALS4, CEACAM6, TSPAN8, COL1A2	29352642[14]
Idiopathic pulmonary fibrosis		CAMP, CEACAM6, CTSG, DEFA3 DEFA4, OLFM4, HLTF, PACSIN1, GABBR1, IGHM	22761659[46]
Lung cancer		HK3, SLC36A1, MSR1, CEACAM1, CEACAM6, HCG27, FXYD7, TRPLC1, NR3C2, RLN2	34288383[44]
COVID-19	Neutrophil-associated gene cluster	CEACAM6, RETN, MPO, LTF, MMP8, CEACAM8, DEFA4, OLR1, DEFA3, DEFA1B, DEFA1, ELANE	34335605[52]
COVID-19	Secretory granules signature	CEACAM8, MMP8, ELANE, LTF, CEACAM6, MPO	35844004[53]

The current methodology reliance on a systematic, manual approach to data retrieval and structuring is another limitation. We recognize the potential of automation to transform this labor-intensive process. In this respect, we are actively exploring the integration of Large Language Models (LLMs) into our data curation workflow. These advanced models show promise in streamlining the identification, extraction, and structuring of relevant information, potentially mitigating the challenges associated with the sheer volume and dynamic nature of biomedical databases. Our preliminary explorations suggest that while LLMs may not fully replace the nuanced judgment of human curators, they offer significant support by enhancing efficiency and accuracy, thereby complementing our existing methodologies. Thus, we are cautiously optimistic about the role of LLMs in enhancing our data analysis framework, aiming to improve efficiency while maintaining accuracy. This integration of LLMs is an ongoing effort and will be detailed further in upcoming publications.

In conclusion, the information presented here should help researchers decide whether to include CEACAM6 in the targeted assay they intend to develop. Some of our findings suggest that measuring abundance of CEACAM6 transcripts in blood could prove to be of value in the monitoring and management of patients with diseases associated with systemic inflammation. This would likely be true for other members of the BloodGen3 module M10.4/"neutrophil activation" gene sets. However, CEACAM6 presents with the distinct advantage of also being of potential value in the management of patients with cancer, whether the assay would be used to measure transcript abundance in blood or in tumor tissues.

Author contributions

DR and DC: Conceptualization, Data curation, Formal analysis, Visualization, Methodology Development, Writing – Review & Editing. DC: Writing – Original Draft Preparation. The contributor's roles listed above follow the Contributor Roles Taxonomy (CRediT) managed by The Consortia Advancing Standards in Research Administration Information (CASRAI) (https://casrai.org/credit/).

Data availability

Extended data

The project contains the following extended data:

- **Extended Data File 1:** a spreadsheet in the MS Excel format that is used as a template to capture relevant information from the literature and from transcriptional profiling data analysis results. Figshare: Ext Data File 1 - Information Capture Form_Generic_2022 Sept14 https://doi.org/10.6084/m9.figshare.21183718.v1.[20]

- **Extended Data File 2**: a spreadsheet in the MS Excel format listing cell type and disease entities and their prevalence in the literature associated with CEACAM6. Figshare: Ext Data File 2 CEACAM6_Lit Profiles_Entities_Step3c_2022 Sept14 https://doi.org/10.6084/m9.figshare.21183748.v1.[30]

- **Extended Data File 3**: a spreadsheet in the MS Excel format used to capture information from the CEACAM6 literature regarding its actual or potential use as a biomarker. Figshare: Ext Data File 3 CEACAM6_Articles_ Biomarker Relevance_Step3d_2022 Sept14. https://doi.org/10.6084/m9.figshare.21183832.v1.[31]

- **Extended Data File 4**: a spreadsheet in the MS Excel format used to capture information from the CEACAM6 literature reporting differences in blood transcript abundance in cases vs controls. Figshare: Ext Data File 4 CEACAM6_Articles_Blood transcript profiling_Step4c_2022 Sep14. https://doi.org/10.6084/m9.figshare.21184357.v1.[56]
- **Extended Data File 5**: a spreadsheet in the MS Excel format used to capture CEACAM6 transcriptional profiles from multiple datasets (one dataset per tab) and compute significance of differences in abundance observed between cases and controls. Figshare: Ext Data File 5 CEACAM6_Transcriptome data_ abundance profiles_Step5b_2022 Sept14. https://doi.org/10.6084/m9.figshare.21184363.v1.[57]

- **Extended Data File 6**: a spreadsheet in the MS Excel format used to capture relevant information regarding differences in CEACAM6 blood transcriptional abundance observed in multiple datasets. Figshare: Ext Data File 6 CEACAM6_Transcriptome data_diff expression_Step5c_2022 Sept14. https://doi.org/10.6084/m9.figshare.21184369.v1.[58]

References

1. Chaussabel D: **Assessment of immune status using blood transcriptomics and potential implications for global health.** *Semin. Immunol.* 2015 Feb; **27**(1): 58–66.
PubMed Abstract | Publisher Full Text

2. Li S, Todor A, Luo R: **Blood transcriptomics and metabolomics for personalized medicine.** *Comput. Struct. Biotechnol. J.* 2016; **14**: 1–7.
PubMed Abstract | Publisher Full Text

3. Devaux Y: **Transcriptome of blood cells as a reservoir of cardiovascular biomarkers.** *Biochim. Biophys. Acta, Mol. Cell Res.* 2017 Jan; **1864**(1): 209–216.
PubMed Abstract | Publisher Full Text

4. Breen MS, Stein DJ, Baldwin DS: **Systematic review of blood transcriptome profiling in neuropsychiatric disorders: guidelines for biomarker discovery.** *Hum. Psychopharmacol.* 2016 Sep; **31**(5): 373–381.
PubMed Abstract | Publisher Full Text

5. Karsten SL, Kudo LC, Bragin AJ: **Use of peripheral blood transcriptome biomarkers for epilepsy prediction.** *Neurosci. Lett.* 2011 Jun 27; **497**(3): 213–217.
PubMed Abstract | Publisher Full Text

6. Freedman JE, Vitseva O, Tanriverdi K: **The role of the blood transcriptome in innate inflammation and stroke.** *Ann. N. Y. Acad. Sci.* 2010 Oct; **1207**: 41–45.
PubMed Abstract | Publisher Full Text

7. Staratschek-Jox A, Classen S, Gaarz A, *et al.*: **Blood-based transcriptomics: leukemias and beyond.** *Expert. Rev. Mol. Diagn.* 2009 Apr; **9**(3): 271–280.
PubMed Abstract | Publisher Full Text

8. Athar A, Füllgrabe A, George N, *et al.*: **ArrayExpress update - from bulk to single-cell expression data.** *Nucleic Acids Res.* 2019 Jan 8; **47**(D1): D711–D715.
PubMed Abstract | Publisher Full Text

9. Rinchai D, Syed Ahamed Kabeer B, Toufiq M, *et al.*: **A modular framework for the development of targeted Covid-19 blood transcript profiling panels.** *J. Transl. Med.* 2020 Jul 31; **18**(1): 291.
PubMed Abstract | Publisher Full Text

10. Rinchai D, Chaussabel D: **A training curriculum for retrieving, structuring, and aggregating information derived from the biomedical literature and large-scale data repositories.** *F1000Res.* 2022 [cited 2022 Sep 7].
Publisher Full Text | Reference Source

11. Beauchemin N, Arabzadeh A: **Carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) in cancer progression and metastasis.** *Cancer Metastasis Rev.* 2013 Dec; **32**(3–4): 643–671.
PubMed Abstract | Publisher Full Text

12. Obrink B: **CEA adhesion molecules: multifunctional proteins with signal-regulatory properties.** *Curr. Opin. Cell Biol.* 1997 Oct; **9**(5): 616–626.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Chaussabel D, Rinchai D: **Using "collective omics data" for biomedical research training.** *Immunology.* 2018 Sep; **155**(1): 18–23.
PubMed Abstract | Publisher Full Text

14. Rodia MT, Solmi R, Pasini F, *et al.*: **LGALS4, CEACAM6, TSPAN8, and COL1A2: Blood Markers for Colorectal Cancer-Validation in a Cohort of Subjects With Positive Fecal Immunochemical Test Result.** *Clin. Colorectal Cancer.* 2018 Jun; **17**(2): e217–e228.
PubMed Abstract | Publisher Full Text

15. Rodia MT, Ugolini G, Mattei G, *et al.*: **Systematic large-scale meta-analysis identifies a panel of two mRNAs as blood biomarkers for colorectal cancer detection.** *Oncotarget.* 2016 May 24; **7**(21): 30295–30306.
PubMed Abstract | Publisher Full Text

16. Ferlizza E, Solmi R, Miglio R, *et al.*: **Colorectal cancer screening: Assessment of *CEACAM6*, *LGALS4*, *TSPAN8* and *COL1A2* as blood markers in faecal immunochemical test negative subjects.** *J. Adv. Res.* 2020 Mar 3; **24**: 99–107. eCollection 2020 Jul.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Altman MC, Rinchai D, Baldwin N, *et al.*: **Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data.** *Nat. Commun.* 2021 Jul 19; **12**(1): 4385.
PubMed Abstract | Publisher Full Text

18. Rinchai D, Roelands J, Toufiq M, *et al.*: **BloodGen3Module: Blood transcriptional module repertoire analysis and visualization using R.** *Bioinforma. Oxf. Engl.* 2021 Feb 24; btab121.

19. Ali FA, Marr AK, Tatari-Calderone Z, *et al.*: **Organizing gene literature retrieval, profiling, and visualization training workshops for early career researchers.** *F1000Res.* 2021 [cited 2022 Sep 2].
Publisher Full Text | Reference Source

20. Chaussabel D: **Ext Data File 1 - Information Capture Form_Generic_2022 Sept14.** 2022 Sep 21 [cited 2022 Sep 21]
Reference Source

21. Speake C, Presnell S, Domico K, *et al.*: **An interactive web application for the dissemination of human systems immunology data.** *J. Transl. Med.* 2015 Jun 19; **13**: 196.
PubMed Abstract | Publisher Full Text

22. Hammarström S: **The carcinoembryonic antigen (CEA) family: structures, suggested functions and expression in normal and malignant tissues.** *Semin. Cancer Biol.* 1999 Apr; **9**(2): 67–81.
PubMed Abstract | Publisher Full Text

23. Schölzel S, Zimmermann W, Schwarzkopf G, *et al.*: **Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas.** *Am. J. Pathol.* 2000 Feb; **156**(2): 595–605.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Barnich N, Carvalho FA, Glasser AL, *et al.*: **CEACAM6 acts as a receptor for adherent-invasive E. coli, supporting ileal mucosa colonization in Crohn disease.** *J. Clin. Invest.* 2007 Jun; **117**(6): 1566–1574.
PubMed Abstract | Publisher Full Text

25. Sarantis H, Gray-Owen SD: **Defining the roles of human carcinoembryonic antigen-related cellular adhesion molecules during neutrophil responses to Neisseria gonorrhoeae.** *Infect. Immun.* 2012 Jan; **80**(1): 345–358.
PubMed Abstract | Publisher Full Text

26. Blumenthal RD, Hansen HJ, Goldenberg DM: **Inhibition of adhesion, invasion, and metastasis by antibodies targeting CEACAM6 (NCA-90) and CEACAM5 (Carcinoembryonic Antigen).** *Cancer Res.* 2005 Oct 1; **65**(19): 8809–8817.
PubMed Abstract | Publisher Full Text

27. Duxbury MS, Ito H, Zinner MJ, *et al.*: **CEACAM6 gene silencing impairs anoikis resistance and in vivo metastatic ability of pancreatic adenocarcinoma cells.** *Oncogene.* 2004 Jan 15; **23**(2): 465–473.
PubMed Abstract | Publisher Full Text

28. Han ZW, Lyv ZW, Cui B, *et al.*: **The old CEACAMs find their new role in tumor immunotherapy.** *Investig. New Drugs.* 2020 Dec; **38**(6): 1888–1898.
PubMed Abstract | Publisher Full Text

29. Witzens-Harig M, Hose D, Jünger S, *et al.*: **Tumor cells in multiple myeloma patients inhibit myeloma-reactive T cells through carcinoembryonic antigen-related cell adhesion molecule-6.** *Blood.* 2013 May 30; **121**(22): 4493–4503.
PubMed Abstract | Publisher Full Text

30. Chaussabel D: **Ext Data File 2 CEACAM6_Lit Profiles_Entities_Step3c_2022 Sept14.** 2022 Sep 21 [cited 2022 Sep 21].
Reference Source

31. Chaussabel D: **Ext Data File 3 CEACAM6_Articles_Biomarker Relevance_Step3d_2022 Sept14.** 2022 Sep 21 [cited 2022 Sep 21].
Reference Source

32. Jantscheff P, Terracciano L, Lowy A, *et al.*: **Expression of CEACAM6 in resectable colorectal cancer: a factor of independent prognostic significance.** *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 2003 Oct 1; **21**(19): 3638–3646.
PubMed Abstract | Publisher Full Text

33. Kim KS, Kim JT, Lee SJ, *et al.*: **Overexpression and clinical significance of carcinoembryonic antigen-related cell adhesion molecule 6 in colorectal cancer.** *Clin. Chim. Acta. Int. J. Clin. Chem.* 2013 Jan 16; **415**: 12–19.
PubMed Abstract | Publisher Full Text

34. Tsang JYS, Kwok YK, Chan KW, *et al.*: **Expression and clinical significance of carcinoembryonic antigen-related cell adhesion molecule 6 in breast cancers.** *Breast Cancer Res. Treat.* 2013 Nov; **142**(2): 311–322.
PubMed Abstract | Publisher Full Text

35. Maraqa L, Cummings M, Peter MB, *et al.*: **Carcinoembryonic antigen cell adhesion molecule 6 predicts breast cancer recurrence following adjuvant tamoxifen.** *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 2008 Jan 15; **14**(2): 405–411.
PubMed Abstract | Publisher Full Text

36. Liu S, Cai Y, Changyong E, *et al.*: **Screening and Validation of Independent Predictors of Poor Survival in Pancreatic Cancer.** *Pathol. Oncol. Res. POR.* 2021; **27**: 1609868.
PubMed Abstract | Publisher Full Text

37. Kurlinkus B, Ger M, Kaupinis A, *et al.*: **CEACAM6's Role as a Chemoresistance and Prognostic Biomarker for Pancreatic Cancer: A Comparison of CEACAM6's Diagnostic and Prognostic Capabilities with Those of CA19-9 and CEA.** *Life Basel Switz.*

2021 Jun 9; **11**(6): 542.
Publisher Full Text

38. Gebauer F, Wicklein D, Horst J, *et al.*: **Carcinoembryonic antigen-related cell adhesion molecules (CEACAM) 1, 5 and 6 as biomarkers in pancreatic cancer.** *PloS One.* 2014; **9**(11): e113023.
PubMed Abstract | Publisher Full Text

39. Chen J, Li Q, An Y, *et al.*: **CEACAM6 induces epithelial-mesenchymal transition and mediates invasion and metastasis in pancreatic cancer.** *Int. J. Oncol.* 2013 Sep; **43**(3): 877–885.
PubMed Abstract | Publisher Full Text

40. Duxbury MS, Matros E, Clancy T, *et al.*: **CEACAM6 is a novel biomarker in pancreatic adenocarcinoma and PanIN lesions.** *Ann. Surg.* 2005 Mar; **241**(3): 491–496.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Kim EY, Cha YJ, Jeong S, *et al.*: **Overexpression of CEACAM6 activates Src-FAK signaling and inhibits anoikis, through homophilic interactions in lung adenocarcinomas.** *Transl. Oncol.* 2022 Jun; **20**: 101402.
PubMed Abstract | Publisher Full Text

42. Poola I, Shokrani B, Bhatnagar R, *et al.*: **Expression of carcinoembryonic antigen cell adhesion molecule 6 oncoprotein in atypical ductal hyperplastic tissues is associated with the development of invasive breast cancer.** *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 2006 Aug 1; **12**(15): 4773–4783.
PubMed Abstract | Publisher Full Text

43. Farina A, Dumonceau JM, Antinori P, *et al.*: **Bile carcinoembryonic cell adhesion molecule 6 (CEAM6) as a biomarker of malignant biliary stenoses.** *Biochim. Biophys. Acta.* 2014 May; **1844**(5): 1018–1025.
PubMed Abstract | Publisher Full Text

44. Zhang Q, Kuang M, An H, *et al.*: **Peripheral blood transcriptome heterogeneity and prognostic potential in lung cancer revealed by RNA-Seq.** *J. Cell. Mol. Med.* 2021 Sep; **25**(17): 8271–8284.
PubMed Abstract | Publisher Full Text

45. Infante J, Prieto C, Sierra M, *et al.*: **Identification of candidate genes for Parkinson's disease through blood transcriptome analysis in LRRK2-G2019S carriers, idiopathic cases, and controls.** *Neurobiol. Aging.* 2015 Feb; **36**(2): 1105–1109.
PubMed Abstract | Publisher Full Text

46. Yang IV, Luna LG, Cotter J, *et al.*: **The peripheral blood transcriptome identifies the presence and extent of disease in idiopathic pulmonary fibrosis.** *PLoS One.* 2012; **7**(6): e37708.
PubMed Abstract | Publisher Full Text

47. Bigler J, Boedigheimer M, Schofield JPR, *et al.*: **A Severe Asthma Disease Signature from Gene Expression Profiling of Peripheral Blood from U-BIOPRED Cohorts.** *Am. J. Respir. Crit. Care Med.* 2017 May 15; **195**(10): 1311–1320.
Publisher Full Text

48. Huo J, Wang L, Tian Y, *et al.*: **Gene Co-Expression Analysis Identified Preserved and Survival-Related Modules in Severe Blunt Trauma, Burns, Sepsis, and Systemic Inflammatory Response Syndrome.** *Int. J. Gen. Med.* 2021; **14**: 7065–7076.
PubMed Abstract | Publisher Full Text

49. Hori H, Yoshida F, Itoh M, *et al.*: **Proinflammatory status-stratified blood transcriptome profiling of civilian women with PTSD.** *Psychoneuroendocrinology.* 2020 Jan; **111**: 104491.
PubMed Abstract | Publisher Full Text

50. Kvist-Hansen A, Kaiser H, Wang X, *et al.*: **Neutrophil Pathways of Inflammation Characterize the Blood Transcriptomic Signature of Patients with Psoriasis and Cardiovascular Disease.** *Int. J. Mol. Sci.* 2021 Oct 6; **22**(19): 10818.
PubMed Abstract | Publisher Full Text

51. Lee J, Romero R, Chaiworapongsa T, *et al.*: **Characterization of the fetal blood transcriptome and proteome in maternal anti-fetal rejection: evidence of a distinct and novel type of human fetal systemic inflammatory response.** *Am. J. Reprod. Immunol. N Y N 1989.* 2013 Oct; **70**(4): 265–284.
Publisher Full Text

52. Prokop JW, Hartog NL, Chesla D, *et al.*: **High-Density Blood Transcriptomics Reveals Precision Immune Signatures of SARS-CoV-2 Infection in Hospitalized Individuals.** *Front. Immunol.* 2021; **12**: 694243.
PubMed Abstract | Publisher Full Text

53. Jackson H, Rivero Calle I, Broderick C, *et al.*: **Characterisation of the blood RNA host response underpinning severity in COVID-19 patients.** *Sci. Rep.* 2022 Jul 17; **12**(1): 12216.
PubMed Abstract | Publisher Full Text

54. Bongen E, Lucian H, Khatri A, *et al.*: **Sex Differences in the Blood Transcriptome Identify Robust Changes in Immune Cell Proportions with Aging and Influenza Infection.** *Cell Rep.* 2019 Nov 12; **29**(7): 1961–1973.e4.
PubMed Abstract | Publisher Full Text

55. Signorelli M, Ebrahimpoor M, Veth O, *et al.*: **Peripheral blood transcriptome profiling enables monitoring disease progression in dystrophic mice and patients.** *EMBO Mol. Med.* 2021 Apr 9; **13**(4): e13328.
PubMed Abstract | Publisher Full Text

56. Chaussabel D: **Ext Data File 4 CEACAM6_Articles_Blood transcript profiling_Step4c_2022 Sep14.** 2022 Sep 22 [cited 2022 Sep 22].
Reference Source

57. Chaussabel D: **Ext Data File 5 CEACAM6_Transcriptome data_abundance profiles_Step5b_2022 Sept14.** 2022 Sep 22 [cited 2022 Sep 22].
Reference Source

58. Chaussabel D: **Ext Data File 6 CEACAM6_Transcriptome data_diff expression_Step5c_2022 Sept14.** 2022 Sep 22 [cited 2022 Sep 22].
Reference Source

59. Riley CJ, Engelhardt KP, Saldanha JW, *et al.*: **Design and activity of a murine and humanized anti-CEACAM6 single-chain variable fragment in the treatment of pancreatic cancer.** *Cancer Res.* 2009 Mar 1; **69**(5): 1933–1940.
PubMed Abstract | Publisher Full Text

60. Linsley PS, Speake C, Whalen E, *et al.*: **Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis.** *PLoS One.* 2014; **9**(10): e109760.
Publisher Full Text

61. Obermoser G, Presnell S, Domico K, *et al.*: **Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines.** *Immunity.* 2013 Apr 18; **38**(4): 831–844.
PubMed Abstract | Publisher Full Text

62. Single Cell Portal: [cited 2022 Sep 2].
Reference Source

63. He MX, Cuoco MS, Crowdis J, *et al.*: **Transcriptional mediators of treatment resistance in lethal prostate cancer.** *Nat. Med.* 2021 Mar; **27**(3): 426–433.
PubMed Abstract | Publisher Full Text

64. Wargodsky R, Dela Cruz P, LaFleur J, *et al.*: **RNA Sequencing in COVID-19 patients identifies neutrophil activation biomarkers as a promising diagnostic platform for infections.** *PLoS One.* 2022; **17**(1): e0261679.
PubMed Abstract | Publisher Full Text

65. Leite GGF, Ferreira BL, Tashima AK, *et al.*: **Combined Transcriptome and Proteome Leukocyte's Profiling Reveals Up-Regulated Module of Genes/Proteins Related to Low Density Neutrophils and Impaired Transcription and Translation Processes in Clinical Sepsis.** *Front. Immunol.* 2021; **12**: 744799.
PubMed Abstract | Publisher Full Text

66. Rosa BA, Ahmed M, Singh DK, *et al.*: **IFN signaling and neutrophil degranulation transcriptional signatures are induced during SARS-CoV-2 infection.** *Commun. Biol.* 2021 Mar 5; **4**(1): 290.
PubMed Abstract | Publisher Full Text

67. Jessup JM, Thomas P: **Carcinoembryonic antigen: function in metastasis by human colorectal carcinoma.** *Cancer Metastasis Rev.* 1989 Dec; **8**(3): 263–280.
Publisher Full Text

68. Sikorska H, Shuster J, Gold P: **Clinical applications of carcinoembryonic antigen.** *Cancer Detect. Prev.* 1988; **12**(1–6): 321–355.
PubMed Abstract

69. Beard DB, Haskell CM: **Carcinoembryonic antigen in breast cancer. Clinical review.** *Am. J. Med.* 1986 Feb; **80**(2): 241–245.
Publisher Full Text

70. Bennett L, Palucka AK, Arce E, *et al.*: **Interferon and granulopoiesis signatures in systemic lupus erythematosus blood.** *J. Exp. Med.* 2003 Mar 17; **197**(6): 711–723.
PubMed Abstract | Publisher Full Text

71. Wilk AJ, Rustagi A, Zhao NQ, *et al.*: **A single-cell atlas of the peripheral immune response in patients with severe COVID-19.** *Nat. Med.* 2020 Jul; **26**(7): 1070–1076.
PubMed Abstract | Publisher Full Text

72. Strickland LA, Ross J, Williams S, *et al.*: **Preclinical evaluation of carcinoembryonic cell adhesion molecule (CEACAM) 6 as potential therapy target for pancreatic adenocarcinoma.** *J. Pathol.* 2009 Jul; **218**(3): 380–390.
PubMed Abstract | Publisher Full Text

73. Pinkert J, Boehm HH, Trautwein M, *et al.*: **T cell-mediated elimination of cancer cells by blocking CEACAM6-CEACAM1 interaction.** *Oncoimmunology.* 2021 Dec 30; **11**(1): 2008110. eCollection 2022.
PubMed Abstract | Publisher Full Text | Free Full Text

74. Pandey R, Zhou M, Islam S, *et al.*: **Carcinoembryonic antigen cell adhesion molecule 6 (CEACAM6) in Pancreatic**

Ductal Adenocarcinoma (PDA): An integrative analysis of a novel therapeutic target. *Sci. Rep.* 2019 Dec 4; **9**(1): 18347.
PubMed Abstract | Publisher Full Text | Free Full Text

75. Burgos M, Cavero-Redondo I, Álvarez-Bueno C, *et al.*: **Prognostic value of the immune target CEACAM6 in cancer: a meta-analysis.** *Ther Adv Med Oncol.* 2022 Jan 19; **14**: 17588359211072621. eCollection 2022.
PubMed Abstract | Publisher Full Text | Free Full Text

76. Xingyuan C, Chen Q: **Serum BPI as a novel biomarker in asthma.** *Allergy Asthma Clin. Immunol. Off. J. Can. Soc. Allergy Clin. Immunol.* 2020; **16**: 50.
PubMed Abstract | Publisher Full Text

77. Tian Y, Zeng T, Tan L, *et al.*: **BPI-ANCA in chronic obstructive pulmonary disease with pulmonary Pseudomonas aeruginosa colonisation: a novel indicator of poor prognosis.** *Br. J. Biomed. Sci.* 2018 Oct; **75**(4): 206–208.
PubMed Abstract | Publisher Full Text

78. Maneerat Y, Prasongsukarn K, Benjathummarak S, *et al.*: **PPBP and DEFA1/DEFA3 genes in hyperlipidaemia as feasible synergistic inflammatory biomarkers for coronary heart disease.** *Lipids Health Dis.* 2017 Apr 19; **16**(1): 80.
PubMed Abstract | Publisher Full Text

79. Derigs M, Heers H, Lingelbach S, *et al.*: **Soluble PD-L1 in blood correlates positively with neutrophil and negatively with lymphocyte mRNA markers and implies adverse sepsis outcome.** *Immunol. Res.* 2022 Jun 23; **70**: 698–707.
PubMed Abstract | Publisher Full Text

80. Guo C, Zeng F, Liu H, *et al.*: **Establish immune-related gene prognostic index for esophageal cancer.** *Front. Genet.* 2022; **13**: 956915.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 2

Reviewer Report 04 September 2024

https://doi.org/10.5256/f1000research.163578.r289474

?

Akila Prashant
Biochemistry, JSS Academy of Higher Education & Research, Mysore, Karnataka, India
Deepthi V
Biochemistry, JSS Academy of Higher Education and Research, Mysuru, Karnataka, India

This article offers comprehensive and detailed information on the process of screening a protein as a potential biomarker. It is assumed that this method can be applied to any protein for screening, assessment, and correlation with various diseases.
While the paper provides methods for screening a biomarker, it lacks clarity regarding its purpose. It is unclear whether the objective is to explain the screening methods, or to demonstrate CEACAM6 as a potential biomarker, or both.
If it is both:

1. The author is suggested to represent the flowchart of the methodology section in detail.
2. The paper lacks an explanation of CEACAM6 as a protein and its physiological functions/ pathways in the context of diseases since the title claims to reveal the potential relevance of CEACAM6.
3. The paper is expected to balance both the methodology of biomarker screening as well as CEACAM6 functions and potentiality as a protein as well as biomarker.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Rare Disease Genetics, Cancer Biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 13 April 2024

https://doi.org/10.5256/f1000research.163578.r263048

Ritu Pandey
Department of Cellular and Molecular Medicine, University of Arizona, Tucson, USA
Daruka Mahadevan
The University of Texas Health Science Center (Ringgold ID: 12346), San Antonio, Texas, USA

No further comments.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cancer research

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 03 October 2023

https://doi.org/10.5256/f1000research.139158.r192786

?

Ritu Pandey
Department of Cellular and Molecular Medicine, University of Arizona, Tucson, USA
Daruka Mahadevan
The University of Texas Health Science Center (Ringgold ID: 12346), San Antonio, Texas, USA

The manuscript is a summary of CEACAM6 gene as a prognostic blood biomarker across diseases and provides a guide to capture heterogeneous information on a given gene or biomarker using publicly available processed or summarized data. The work is based on prior published work by the group on analysis of blood biomarkers.

It is not clear that the main focus of the paper is showing a method or in-depth analysis of CEACAM6 gene utilizing the method, assuming both here are few comments:

1. The authors provide for the readers a pathway to gather details on a gene and summarize the information. But it is not clear how the information is getting extracted. The data captured is voluminous at times, so are there any scripts that the authors are providing or suggesting for data retrieval? There is some amount of customized data massaging and data wrangling needed for further use, but few automated steps and scripts would be useful. Data backed with evidence keeps changing such as PubMed and needs to be updated too. The strategy is based on previous work by the authors but any automation or programmatic retrieval for any of the steps should be included here for the benefit of readers.

2. The statement in the Discussion #4 that no drugs have been developed for targeting CEACAM6 could be misleading since there have been efforts, at least in cancer studies to target CEACAM6 using monoclonal antibodies. The authors do mention one myeloma study. There are several experimental approaches that have been published in preclinical models

and there are also clinical studies in cancer where this is being targeted as an immune checkpoint.

3. This study assesses changes in expression of any gene in disease cases to qualify it as a marker but equally important is what approaches have been used or what indications exist that it has been studied as a therapeutic target and if so the outcome of such study. There are several publications for studies of CEACAM6 as a therapeutic target in cancer. Examples of few of the published work from NCBI PubMed - PMID: 19334050, PMID: 35141051, PMID: 31797958, PMID: 35082925 and a clinical trial - https://clinicaltrials.gov/study/NCT03596372. These should be cited as similar approaches could be utilized for other diseases.

References

1. Strickland LA, Ross J, Williams S, Ross S, et al.: Preclinical evaluation of carcinoembryonic cell adhesion molecule (CEACAM) 6 as potential therapy target for pancreatic adenocarcinoma.*J Pathol*. 2009; **218** (3): 380-90 PubMed Abstract | Publisher Full Text
2. Pinkert J, Boehm HH, Trautwein M, Doecke WD, et al.: T cell-mediated elimination of cancer cells by blocking CEACAM6-CEACAM1 interaction.*Oncoimmunology*. 2022; **11** (1): 2008110 PubMed Abstract | Publisher Full Text
3. Pandey R, Zhou M, Islam S, Chen B, et al.: Carcinoembryonic antigen cell adhesion molecule 6 (CEACAM6) in Pancreatic Ductal Adenocarcinoma (PDA): An integrative analysis of a novel therapeutic target.*Sci Rep*. 2019; **9** (1): 18347 PubMed Abstract | Publisher Full Text
4. Burgos M, Cavero-Redondo I, Álvarez-Bueno C, Galán-Moya EM, et al.: Prognostic value of the immune target CEACAM6 in cancer: a meta-analysis.*Ther Adv Med Oncol*. 2022; **14**: 17588359211072621 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cancer research

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 08 Mar 2024

Damien Chaussabel

Thank you for your constructive review and insightful comments. We understand the concerns raised regarding the clarity of our manuscript's focus and the methodologies employed for data retrieval and analysis. Allow us to address these points with additional context that we believe will clarify our intentions and the contributions of our work.

Clarification of Manuscript Focus: Our current work is intended as a proof of concept paper, demonstrating the practical application of the methodological framework we detailed in a previously published paper ("A training curriculum for retrieving, structuring, and aggregating information derived from the biomedical literature and large-scale data repositories"). This earlier publication provides an in-depth description of the methods and serves as the foundation upon which our current study on CEACAM6 is built. We acknowledge that this may not have been made sufficiently clear in our manuscript, leading to confusion about its primary focus. We revised our introduction sections to explicitly state that this work is a demonstration of our previously published methodology applied to the CEACAM6 gene, rather than an exposition of new methodological advancements:

"The methodology employed in this study is derived from our previously established "collective omics data" (COD) training curriculum, as outlined in our comprehensive methods paper, "A training curriculum for retrieving, structuring, and aggregating information derived from the biomedical literature and large-scale data repositories."[10]. This foundational paper provides a detailed description of our systematic approach to information curation, which we have applied in the current investigation of CEACAM6. Specifically, the study utilizes the COD1 training module workflow from this curriculum, which guides the structured retrieval and aggregation of gene-specific data for biomarker assessment. The process encompasses selecting a gene of interest, in this case, CEACAM6, to comprehensively gather and synthesize relevant information from both literature and public datasets, culminating in the creation of resources like structured data tables and interactive circle packing plots. This approach not only supports the rigorous assessment of CEACAM6's potential as a blood biomarker but also serves as a demonstrative application of our validated methodological framework, providing a practical example of how such a framework can be employed to enhance biomarker discovery efforts."

Data Retrieval and Analysis: The method described in our prior paper involves a systematic but manual approach to retrieving, structuring, and aggregating information, which, as you rightly pointed out, can be labor-intensive. We recognize the importance of automation in managing the vast amount of data available in biomedical research. While our published method does not incorporate automated processes, we have, in response to the evolving needs of data curation, begun exploring the potential of Large Language Models (LLMs) to assist in manual data curation tasks. This effort is led by staff who have

recently joined our group and represents an exciting direction for enhancing the efficiency and scalability of our data curation processes. Preliminary findings suggest that while LLMs are not a complete substitute for manual curation, they can significantly aid in the process by streamlining the identification and extraction of relevant information. This ongoing work acknowledges the pertinence of your feedback regarding automation and highlights our commitment to advancing our methodologies in line with technological developments. Although the results of these explorations are not included in the current manuscript, they are part of a separate study that we plan to publish in the future. This will detail our experiences and findings regarding the integration of LLMs into our data curation workflow, providing insights that could benefit the broader research community in handling similar challenges.A new paragraph has been added to the discussion, acknowledging current limitations and potential strategies for automating the information extraction workflow:

"The current methodology reliance on a systematic, manual approach to data retrieval and structuring is another limitation. We recognize the potential of automation to transform this labor-intensive process. In this respect, we are actively exploring the integration of Large Language Models (LLMs) into our data curation workflow. These advanced models show promise in streamlining the identification, extraction, and structuring of relevant information, potentially mitigating the challenges associated with the sheer volume and dynamic nature of biomedical databases. Our preliminary explorations suggest that while LLMs may not fully replace the nuanced judgment of human curators, they offer significant support by enhancing efficiency and accuracy, thereby complementing our existing methodologies. Thus we are cautiously optimistic about the role of LLMs in enhancing our data analysis framework, aiming to improve efficiency while maintaining accuracy. This integration of LLMs is an ongoing effort and will be detailed further in upcoming publications."

Discussion on Therapeutic Targeting of CEACAM6: We appreciate your pointing out the need to correct and expand our discussion on therapeutic efforts targeting CEACAM6. It was not our intention to overlook significant research in this area. We have revised the relevant sections to accurately reflect ongoing and completed studies targeting CEACAM6 with therapeutic intent, citing the publications and clinical trials you mentioned. This will ensure our discussion acknowledges both the biomarker potential of CEACAM6 and its implications for therapeutic development:

Recent studies and ongoing clinical trials have explored the utility of targeting CEACAM6 in various cancers, particularly through the development of monoclonal antibodies. For instance, preclinical evaluations have demonstrated the potential of CEACAM6 as a therapy target in pancreatic adenocarcinoma, utilizing antibody-drug conjugates to effectively target and diminish CEACAM6-expressing tumors (PMID: 19334050). Additionally, the blocking of CEACAM6-CEACAM1 interactions has shown promise in enhancing T cell-mediated cancer cell elimination, suggesting a role for CEACAM6 in immune modulation and its potential as an immune checkpoint target (PMID: 35141051). The breadth of research, encompassing studies on its prognostic value and therapeutic targeting in cancers, underscores CEACAM6's significance in oncology and its emerging role as a viable therapeutic target. These investigations, reflected in various studies (PMID: 31797958, PMID: 35082925) and a clinical trial registered under NCT03596372, collectively indicate a growing interest in CEACAM6 as a therapeutic target, warranting further

exploration and validation in clinical settings.

In summary, we are committed to improving the clarity and utility of our manuscript based on your feedback. We believe that by clarifying the relationship between our current work and our previously published methodological framework, and by addressing the points you've raised regarding data analysis and therapeutic targeting, our manuscript will provide a clearer, more comprehensive contribution to the field. Thank you again for your valuable insights, which we are confident will strengthen our paper.

Competing Interests: No competing interests were disclosed.

Reviewer Report 09 December 2022

https://doi.org/10.5256/f1000research.139158.r155597

?

Mattia Lauriola
Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Bologna, Italy

This article is quite interesting, because the authors highlight a possible role for CEACAM6 as a blood biomarker in patients with a viral or bacterial infection, thus surrogating CEAM6 as an assay for systemic inflammation. As the authors correctly pointed out, CEACAM6 transcript levels were already described for cancer diseases including colorectal cancer. In this case, it's interesting to notice that previous publications reported that CEACM6 abundance in the blood of subjects positive for FIT (faecal immunochemical test), but negative for further intestinal lesions. This probably finds support in some of the results reported here, showing an enhanced availability of CEACAM6 in circulating neutrophils. But, the authors fail to refer to this publication PMID: 32257432[1].

References

1. Ferlizza E, Solmi R, Miglio R, Nardi E, et al.: Colorectal cancer screening: Assessment of CEACAM6, LGALS4, TSPAN8 and COL1A2 as blood markers in faecal immunochemical test negative subjects.*J Adv Res*. 2020; **24**: 99-107 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Blood biomarkers for colon cancer, EGFR, RTK signalling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 08 Mar 2024

Damien Chaussabel

Thank you for highlighting the importance of referencing the study associated with PMID: 32257432. We agree that this publication is relevant to our research on CEACAM6 as a potential blood biomarker. In light of your suggestion, we have added a reference to this study in our manuscript (now cited as reference #16). This inclusion helps to contextualize the role of CEACAM6 in subjects with a positive fecal immunochemical test (FIT) but no intestinal lesions, and supports our findings on its availability in circulating neutrophils. We appreciate your constructive feedback and believe that this addition enhances the clarity and relevance of our work.

Competing Interests: No competing interests were disclosed.