49P

Identifying and validating networks of oncology biomarkers mined from the scientific literature

Objective

To develop and to validate a full-text literature interrogation method that can help researchers to identify biomarkers of emerging scientific interest in oncology.

Conclusions

Using large-scale analytics of published literature, biomarkers across six cancer types and a cancer-agnostic network were successfully characterized in terms of their emergence in the published literature and the context in which they are described.

This novel approach could help to identify biomarkers and biomarker panels that could not be identified through traditional search methods, for expert review and exploration in a clinical setting.

Our search method effectively finds relevant literature that could be missed with keyword searches, even where full text is available, and enables users to extract relevant biological insights.

Our network analytic approach enables us to find publications based on biomarker relationships; this cannot be achieved by individual review of papers.

Although our methodology aims to reduce the incidence of false positives, biomarkers could still be mentioned in proximity without a shared biological relationship. Development is underway to optimize the utility of biomarker co-occurrence networks to identify potentially meaningful, emerging biological relationships.

Presenting author: Kim Wager

Explore the interactive tool

Email for more information

Click or scan this quick response (QR) code to download this poster.

References: 1. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. Silver Spring (MD): Food and Drug Administration (US); 2016-. 2021; **2.** Traag VA *et al. Sci Rep* 2019;9:5233.

Disclosures: Kim Wager and Tom Rees are employees of Oxford PharmaGenesis. Dheepa Chari and Steffan Ho are employees of Pfizer, Inc. Orion Penner and Bob JA Schijvenaars are employees of Digital Science. Oxford PharmaGenesis and Digital Science received funding from Pfizer, Inc.

Acknowledgments: Editorial support was provided by Oxford PharmaGenesis.

Presented at the Molecular Analysis for Precision Oncology Virtual Congress 2021 7–9 October 2021

Kim Wager,[1] Dheepa Chari,[2] Steffan Ho,[2] Tom Rees,[1] Orion Penner,[3] Bob JA Schijvenaars[3]

[1]Oxford PharmaGenesis, Oxford, UK; [2]Pfizer, Inc., New York, NY, USA; [3]Digital Science, London, UK

Introduction

- Biomarkers, as measurements of defined biological characteristics, can play a pivotal role in estimations of disease risk, early detection, differential diagnosis, assessment of disease progression and outcomes prediction.[1]
- Studies of cancer biomarkers are published daily; while some biomarkers are well characterized, others are of growing interest.
- Managing this flow of information is challenging for scientists and clinicians.
- We sought to develop a novel text-mining method employing biomarker co-occurrence processing applied to a deeply indexed full-text database to generate time-interval–delimited biomarker co-occurrence networks.

Materials and Methods

- A data set comprising 726 cancer biomarkers was obtained from the Early Detection Research Network, an initiative of the National Cancer Institute.
- Publications were identified through co-occurrence searches for these biomarkers in 20-word proximity to terms relating to six cancer types (**Table 1**).
  - Full-text publications, including proceedings and preprints, with a publication date between 1 January 2015 and 31 December 2020 were searched using the Dimensions scholarly information platform.
- To focus on biomarkers of emerging research interest, those with fewer than five or more than 1000 unique publication mentions were excluded.
- Pairwise co-occurrences (20-word proximity) of biomarkers within the full text were identified to reveal biomarker relationships.
- To identify biomarker pairs that were more likely to represent biologically relevant relationships, pairs with fewer than two publications were excluded.
- Network analysis was performed on those pairs that were mentioned more than once in the same publication.
  - Each node in the network represents a biomarker, while edges represent co-occurrence.
  - Edge weight reflects the number of unique publications in which the two biomarkers occur.
  - On the assumption that, compared with the entire network, clusters of co-occurring biomarkers are more likely to be biologically related, highly connected clusters were identified using the Leiden algorithm.[2]
- To provide a metric for publication growth rate, a linear fit of normalized publication number over time for each biomarker and mean publication growth across all biomarkers in each cluster was calculated.
- Subsets of publications (based on network clustering, publication growth rates and Mendeley library saves) were identified for exploratory investigations into the biological context for biomarker co-occurrence.
- The biological context was classed as 'successful' if one of the co-occurring biomarker pairs was found in proximity to the desired cancer type and the biomarker co-occurrence was biologically meaningful.

Results

BIOMARKER CO-OCCURRENCE NETWORKS

- The Dimensions search identified 255 942 unique full-text publications.
  - Many of these publications were relevant to more than one cancer type (**Table 1**).
- The set of pairwise biomarker co-occurrences spanned 31 550 unique pairs across all cancer types.
  - The most commonly co-occuring biomarker pairs were MMP1–MMP3, MIR21–MIR210 and MIR126–MIR21, with co-occurrences in 820, 632 and 510 publications, respectively.
- We generated biomarker co-occurrence networks for each of the six cancer types and the overall cancer type agnostic data set, accessible on the NDEx platform.

Table 1. Number of Publications Identified for Each Cancer Type

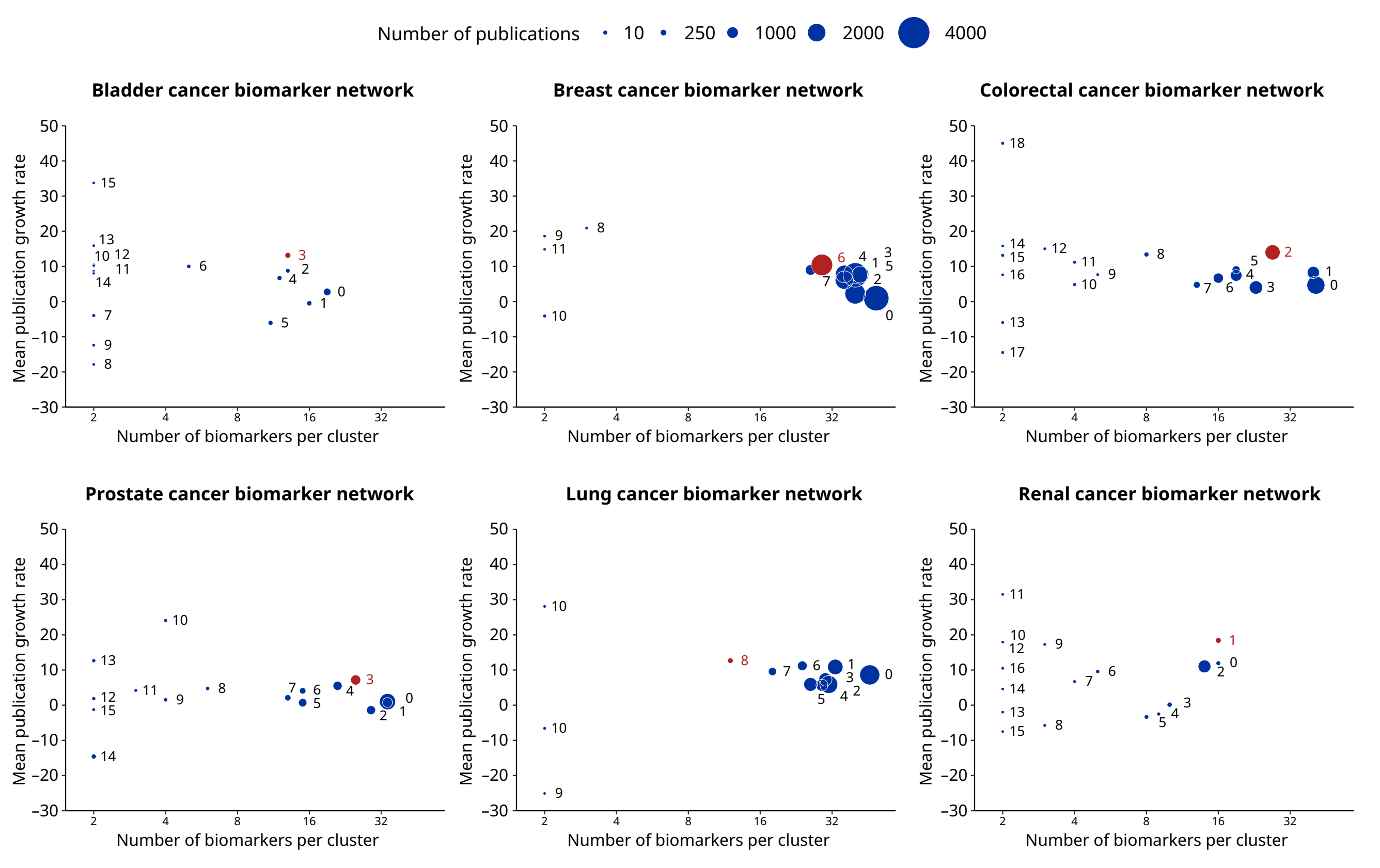
Cancer type	Breast	Lung	Colorectal	Prostate	Renal	Bladder	**Total**[a]
Number of publications	108 134	88 874	69 284	60 644	13 727	13 591	**255 942**

[a]Many publications were relevant to more than one cancer type.

PUBLICATION GROWTH RATE

- To take forward our results for validation and further analysis, we identified the clusters with the highest mean publication growth rate for each network (**Figure 1**).

Figure 1. Publication Growth Rate by Cluster for Each Cancer Type

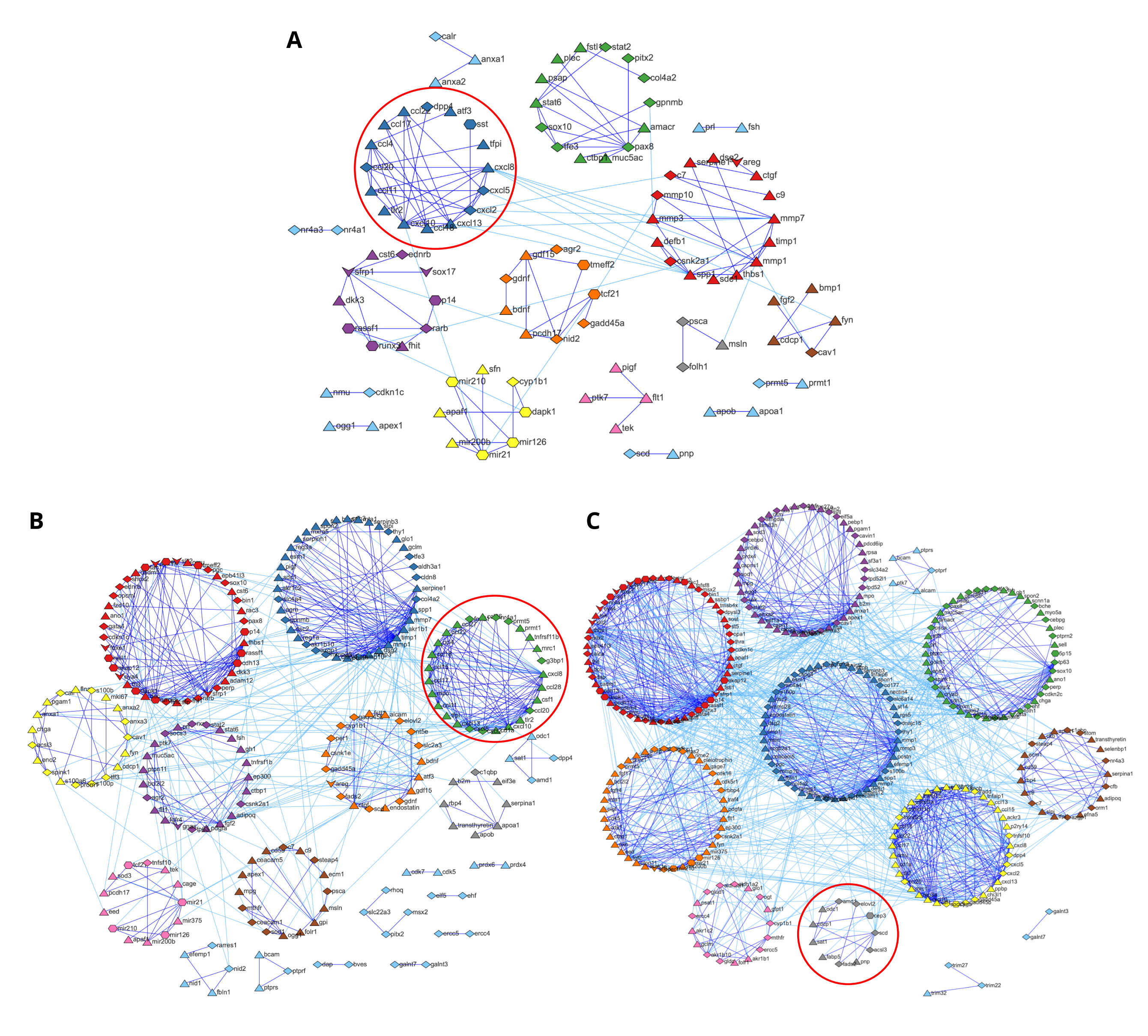


Clusters in red had the highest publication growth rate and contained at least 10 biomarkers.

BIOLOGICAL PROCESSES ACROSS A SINGLE CLUSTER

- Across all networks, we selected the cluster showing the fastest publication growth rate from those containing at least 10 biomarkers: renal cancer, cluster one (**Figure 2A**, circled in red),
  - This cluster comprised 354 unique publications, 140 of which were associated with its edges, representative of biomarker co-occurrences (**Figure 2A**).
  - The most mentioned biomarker in renal cancer cluster one was C-X-C motif chemokine ligand 5 (CXCL5), with 74 publications, while the biomarker pair with the most co-occurrences – either internal or external to the cluster – was CXCL5–CXCL2, with 122 co-mentions in 34 publications (**Figure 3A**).
- Identified biological processes were mapped to the National Cancer Institute Thesaurus and were consistent with a pro-inflammatory role for CXCL5 and CXCL2, acting through their common receptor C-X-C motif chemokine receptor 2 on neutrophils in the tumor microenvironment, influencing angiogenesis, myeloid cell infiltration and metastasis.

Figure 2. (A) Renal cancer biomarker network. (B) Colorectal cancer biomarker network. (C) Overall biomarker network.



Clusters circled in red had the highest publication growth rate and contained at least 10 biomarkers.

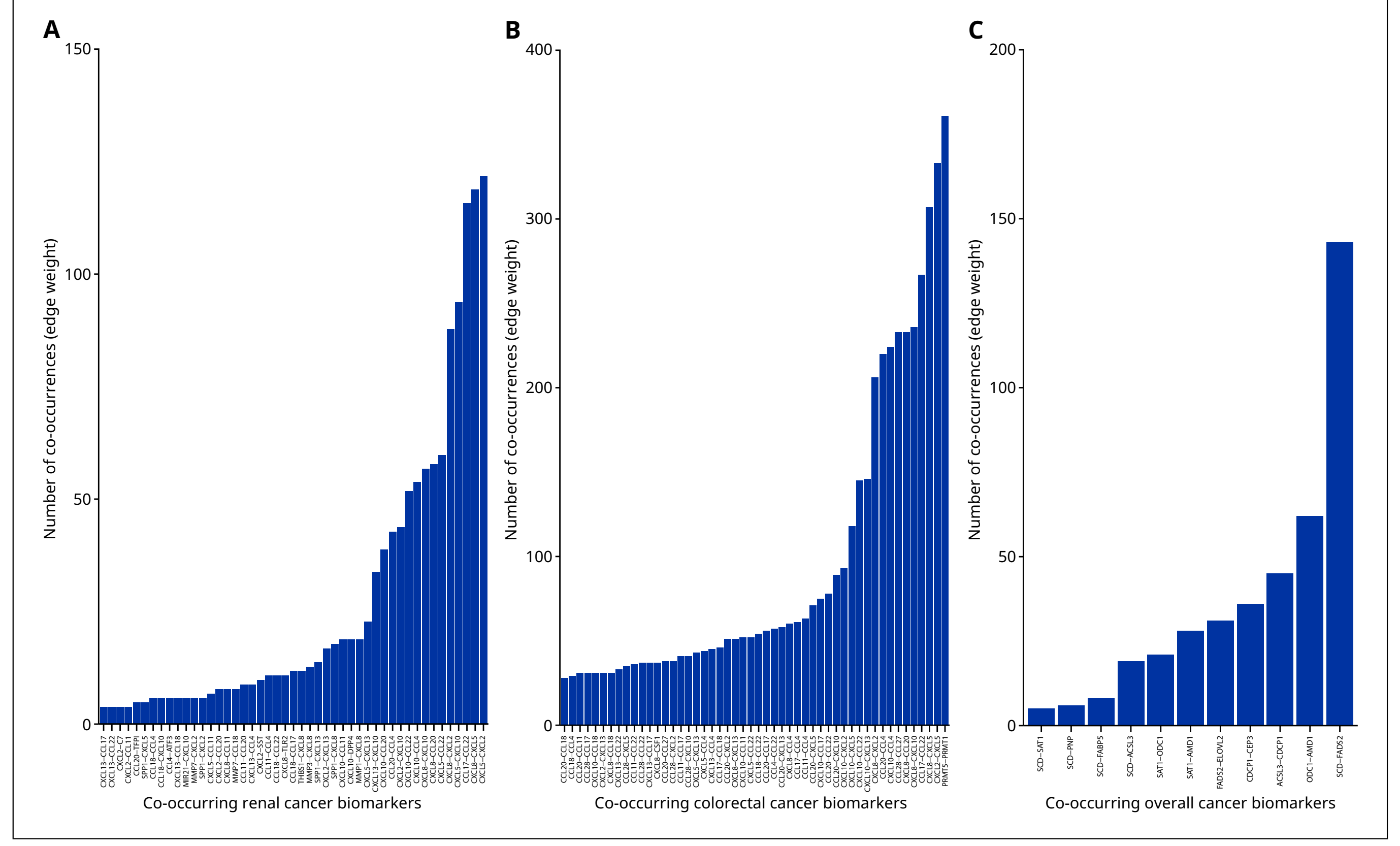
BIOLOGICAL CONTEXT OF BIOMARKER MENTIONS

- To explore alternative ways of using the output from our search methodology and network analysis, the cluster with the second highest publication growth rate – colorectal cancer, cluster two – was selected (**Figure 2B**, circled in red).
  - This cluster contained 139 edges in total, of which 89 were within the cluster (**Figure 2B**).
  - The most common pair by co-occurrence was protein arginine N-methyltransferase 5 (PRMT5)–PRMT1 with 361 co-mentions in 47 unique publications (**Figure 3B**).
- The 20 publications with the highest Mendeley saves (likely to indicate academic interest) were selected to analyze biomarker mentions.
  - Biomarkers in this colorectal cluster were mostly chemokines and were shown to be associated with processes such as cellular infiltration and chemotaxis, with a notable emphasis on chemokines that characterize M1 and M2 macrophages.

BIOLOGICAL CONTEXT WITHIN THE CANCER-AGNOSTIC NETWORK

- From the entire cancer-agnostic network (12 clusters comprising 335 nodes with 1265 edges), we chose to analyze cluster eight because it had the highest publication growth rate and at least 10 biomarkers (**Figure 2C**, circled in red).
  - This cluster contained 26 edges in total – of which 11 were within the cluster – and 418 publications.
  - The most commonly occurring biomarker pair was stearoyl-CoA desaturase–fatty acid desaturase 2 (SCD–FADS2) with 143 co-mentions (**Figure 3C**).
- Of the top 20 publications by Mendeley score, biomarker pairs in this cancer-agnostic cluster were mostly related to biogenic amine metabolism and fatty acid metabolism.

Figure 3. Number of biomarker co-occurrences in (A) renal cancer biomarker network, (B) colorectal cancer biomarker network (only top 50) and (C) overall biomarker network.


A
Number of co-occurrences (edge weight)
Co-occurring renal cancer biomarkers
B
Number of co-occurrences (edge weight)
Co-occurring colorectal cancer biomarkers
C
Number of co-occurrences (edge weight)
SCD–SAT1
SCD–PNP
SCD–FABP5
SCD–ACSL3
SAT1–ODC1
SAT1–AMD1
FADS2–ELOVL2
CDCP1–CEP3
ACSL3–CDCP1
ODC1–AMD1
SCD–FADS2
Co-occurring overall cancer biomarkers