UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Understanding Metastasis Organotropism Patterns Through Within-cell and Between-cells Molecular Interaction Networks

João André Isidoro Miranda

Mestrado em Bioquímica e Biomedicina

Dissertação orientada por:
Professor Doutor Francisco Rodrigues Pinto

2023

Acknowledgements

This is my first relevant academic work and, I would say, the true beginning of my academic career. For most people, it is an important yet small step in their careers, something that ends up at the bottom of their accomplishment list when looking back. Perhaps it will be the same for me, but I doubt it. I doubt it because the road I have taken has been a sinuous one, full of twists and turns, of setbacks and, until the last few years, I could not even fathom getting out of there.

Eventually, I was able to find a straighter and clearer path in the past few years. I have always lacked certain skills that most people take for granted, and that seem to come naturally to them. This includes socializing, mingling in groups, talking and saying what I want, what I feel, what I think, and more... Fortunately, I have been welcomed by many people, who have helped me overcome all the challenges I faced and improve on my limitations. Thus, for those that I have not seen for a long time, for those that are no longer here, for those that I can call friends, and for those who help me and support me because they are always close to me, I sincerely apologize for everything and thank you from the bottom of my heart! This is for everyone who crossed my path, and that knowingly or unknowingly, left a mark on me. Because I am bits and pieces of everyone, a mere node in the social network of humanity. And a node without edges cannot send, receive or store new information–it is irrelevant to the network.

I am not addressing anyone in particular. Some have made a greater (and still growing) contribution to my life and, of course, my acknowledgements implicitly take this into account. Hopefully, I will be able to credit you next time. However, I would like to mention two special people, who deserve my apologies and acknowledgements the most: my parents. They are the ones who have always been there for me in my best and worst moments, and who have had the most impact. I thank you for your patience and support. I know it has been a challenging journey but, at least, it has also been interesting and eventful. Not boring at all! My accomplishments are your accomplishments!

Abstract

Metastasis is responsible for the majority of cancer-related deaths. It occurs when cells from a primary tumour disseminate and initiate new tumours at distant organ sites. Metastasizing cells have to exhibit especial characteristics that allow them to surpass all barriers and bottlenecks in their way to effective colonization. Ensuring survival throughout this process depends on how those cells communicate with the surrounding environments.

Patterns of metastasis are remarkably variable between cancer types. In fact, distinct cancers seem to be predisposed to metastasize to specific organs, a feature known as metastasis organotropism. Our work is based on the hypothesis that organotropism can be partially explained by the extent of intercellular communication between metastasizing cells and cells in the secondary organ. Some proteins that establish intercellular interactions are tissue-specific and can be expressed in pre-cancerous tissue.

Using RNA-seq data from non-diseased tissue, we built networks of intercellular protein-protein interactions between cells from the primary cancer tissue and cells from a potential metastasis tissue. Controlling for other factors that affect organotropism, we found that sites where cancers metastasize more often tend to establish a larger number of intercellular interactions than sites with low incidence of metastasis. We detected 528 literature curated interactions that might play a role in metastasis formation and contribute to the observed differences in cell-cell communication, some previously known to be related to cancer and/or metastasis. Finally, using a network of signalling pathways, we observed that proteins involved in metastasis-associated interactions and their closest neighbours in the network are enriched in cancer driver genes and biological processes linked to invasion and metastasis. In conclusion, we identified intercellular interactions and proteins that drive metastasis development and help explain organotropism. These insights might constitute new research and therapeutic opportunities to treat and prevent metastasis.

Keywords: metastasis organotropism, biological network, protein-protein interaction, cell-cell communication, cancer driver gene

Resumo

Cancro, tumor ou neoplasma são sinónimos para um grupo que inclui mais de 200 doenças que podem afetar quase todos os órgãos ou tecidos humanos. É uma das principais causas de morte no mundo, responsável por mais de 10 milhões de mortes em 2020. Em Portugal, no ano de 2020, os cancros da próstata, colorretal, e do pulmão foram os mais comummente diagnosticados em homens, enquanto mama, colorretal e tiroide foram os tipos de cancro mais frequentes em mulheres. É possível classificar os cancros tanto quanto ao órgão de origem como quanto ao seu tipo celular. Os cancros com origem em células epiteliais são os mais comuns. Existem também cancros do tecido conectivo ou muscular (sarcomas), de células hematopoiéticas (leucemias e linfomas), e cancros de células do sistema nervoso.

Os organismos multicelulares complexos podem ser vistos como uma sociedade ou ecossistema, no qual os seus indivíduos (células) se organizam em comunidades (tecidos). A cooperação que garante o funcionamento ótimo destas comunidades é controlada por sinais extracelulares específicos de tecido enviados, recebidos e interpretados por cascatas de sinalização intracelulares que regulam as células durante o ciclo celular. O cancro ocorre quando, após sucessivos episódios de mutação génica e alterações cromossómicas, as células deixam de responder aos sinais de controlo, levando ao seu crescimento anómalo e proliferação descontrolada. Mutações que promovem o desenvolvimento de cancro ocorrem em genes específicos–genes carcinogénicos (*cancer driver genes*). O processo de carcinogénese leva à seleção de mutações em genes que participam em vias celulares responsáveis por controlarem processos como a proliferação, crescimento, diferenciação e apoptose. Estas alterações levam à aquisição pelas células cancerosas de fenótipos malignos, que as distinguem de células normais. Muitos destes fenótipos resultam da profunda remodelação das vias de sinalização intracelular, permitindo às células cancerosas ignorar estímulos extracelulares ou gerar uma resposta mesmo na ausência de sinal. Durante o crescimento tumoral, o ambiente específico de tecido que envolve as células é gradualmente transformado para criar um ambiente favorável ao desenvolvimento e sobrevivência das células cancerosas–o microambiente tumoral (TME). Dentro do TME, as células cancerosas interagem, colaboram e competem entre si, bem como com outros tipos de células. Este ambiente complexo de interações intercelulares estimula a progressão tumoral, levando ao aparecimento de novas populações celulares com mutações e fenótipos distintos (heterogeneidade intratumoral). Entre estes, surgem fenótipos que se distinguem pela sua agressividade, capazes de invadir o tecido

circundante ao TME, de se disseminarem pelo organismo, e de se estabelecerem e proliferarem num local distante, originando focos secundários de tumores–as metástases.

As metástases são responsáveis pela maioria das mortes relacionadas com cancro, sendo usualmente resistentes a terapias de cancro convencionais. Não obstante, o processo de formação de metástases é bastante ineficiente, e a maioria das células cancerosas que chega a um órgão distante acaba por sofrer apoptose. Acesso a todos os órgãos, excluindo os nódulos linfáticos, ocorre predominantemente através do sistema circulatório sanguíneo, sendo esta a via principal de disseminação da maioria dos cancros. Para além desta via e da disseminação linfática, alguns cancros observam uma disseminação extravascular, como no cancro dos ovários, cujas células se propagam através da cavidade peritoneal.

O processo que leva à formação de metástases pode ser dividido em várias etapas distintas, que representam obstáculos à disseminação das células cancerosas pelo organismo. Estas incluem invasão do tecido circundante ao tumor primário, passagem através dos vasos sanguíneos, sobrevivência em circulação, extravasação para o parênquima do tecido e formação de micrometástases. Células capazes de formar metástases possuem características especiais que lhes permitem ultrapassar estes obstáculos. Estas características incluem programas celulares que promovem plasticidade fenotípica, que se manifesta na capacidade de transição entre fenótipos epiteliais com elevado grau de interações de adesão celular e fenótipos mesenquimais mais móveis e agressivos. A metastização pode ocorrer com uma única célula ou em pequenos agregados celulares, que aumentam a probabilidade de sucesso da migração. Comunicar e interagir com outras células é essencial durante todo o processo de metastização, permitindo às células cancerosas evadir o sistema imunitário, captar a ajuda de células como as plaquetas e remodelar a matriz extracelular. O tumor primário tem um papel essencial para o sucesso da sua migração, ao secretar fatores que atuam principalmente no tecido onde a metástase se acabará por implantar, criando um nicho pré-metastático (PMN). O PMN é um local mais acolhedor e propício à sobrevivência das células metastizantes que o tecido circundante, no qual estas podem residir durante largos períodos até que as condições sejam ideias para proliferarem.

Os padrões de metastização não são aleatórios, existindo uma aparente "preferência" para cada cancro metastizar para determinados órgãos. Este fenómeno, denominado organotropismo, não é explicado exclusivamente por fatores físicos ou anatómicos, como padrões de circulação, vascularização e acessibilidade de um órgão. Na verdade, o PMN apresenta diferentes características, dependendo do tecido onde se localiza. Assim, os requisitos e condições para a sobrevivência de uma célula metastizante são distintos e nem todos os tecidos parecem favoráveis ao desenvolvimento de metástases. Esta dependência do sucesso da metástase no tipo de tecido de chegada parece dever-se, em parte, a fatores específicos de tecido expressos pela célula cancerosa, que permitem a adaptação a microambientes particulares.

A biologia de redes consiste na aplicação da teoria matemática de grafos, numa abordagem quantitativa para caracterizar redes que descrevem sistemas biológicos. Um grafo é um objeto matemático abstrato utilizado para representar redes, onde um conjunto de nós é ligado entre si por um conjunto de arcos. No caso de uma rede de interações proteína-proteína (PPI), os nós representam proteínas e os arcos as interações entre estas. O cancro é uma doença complexa, no qual as células apresentam diversas vias profundamente alteradas, que comunicam entre

si e com o exterior. A abordagem de redes é especialmente apropriada para caracterizar esta complexidade, permitindo ter em conta o contexto em que se insere cada proteína alterada, e prever o seu impacto e papel no fenótipo da doença.

A nossa hipótese é que o organotropismo pode ser parcialmente explicado pelo nível de comunicação intercelular estabelecido entre células metastizantes e células no PMN. Possíveis diferenças na comunicação entre células podem dever-se à expressão de fatores específicos de tecido que participam em interações intercelulares e promovem um ambiente propício ao desenvolvimento de tumores metastáticos. Além disso, esses fatores podem ser seletivamente expressos em tecidos não cancerosos. Com o intuito de avaliar o nível de comunicação entre células do tumor primário e células do tecido potencial para formação de metástases, construímos redes intercelulares de PPI específicas de tecido usando dados de expressão génica de tecido saudável. Cada rede representa a comunicação entre uma célula de metástase e uma célula do tecido que a acolhe. A comparação entre as redes intercelulares de diferentes tecidos foi realizada recorrendo a um método que controla para outros fatores que podem afetar o organotropismo para além da comunicação entre células. Os nossos resultados sugerem que, redes em locais onde os cancros metastizam mais frequentemente do que o esperado ao acaso, estabelecem um maior número de interações intercelulares que redes em locais de baixa incidência de metástase. Em seguida, usando as redes previamente estabelecidas, procedemos à identificação das interações que podem explicar as diferenças observadas na comunicação entre células. O nosso método tem a vantagem de ter apenas em conta as características específicas de tecido e de não estar enviesado para proteínas sobre-expressas em cancro. Detetámos 528 interações diretamente associadas à incidência de metástase que podem potencialmente favorecer a formação de metástases. As interações associadas a metástases apresentam-se enriquecidas em proteínas descritas na literatura e em bases de dados de associação gene-doença, como possuindo associações prévias a cancro e/ou metástases. Entre as proteínas que estabelecem estas interações, existem vários alvos de fármacos desenvolvidos para outras doenças, que podem ser potencialmente reaproveitados para o tratamento e prevenção de metástases.

Por fim, a rede de interações intercelulares foi conectada aos processos que ocorrem no interior da célula, utilizando uma rede de vias de sinalização e de interações fator de transcrição–gene alvo. Simulámos a propagação de um sinal com origem ou destino nas proteínas que participam em interações intercelulares associadas a metástases, com o intuito de encontrar os seus vizinhos na rede intracelular. Os resultados indicam que tanto o grupo de proteínas associadas a metástases como os seus vizinhos na rede estão enriquecidos em genes carcinogénicos. Adicionalmente, vários processos biológicos enriquecidos nestes dois grupos de proteínas estão diretamente relacionados com processos alterados no cancro. Entre os processos que parecem ser afetados ou estar a afetar as interações associadas a metástases, encontram-se vários que regulam a adesão e a migração celular, propriedades essenciais na invasão e metástase.

Em suma, o método desenvolvido neste trabalho permitiu-nos associar interações intercelulares específicas à formação de metástases e ao fenómeno de organotropismo. A maioria das associações detetadas são novas e não tinham uma conexão prévia à progressão das metástases. Estas interações e as proteínas que as estabelecem poderão dar azo a investigações futuras e sugerir novos alvos terapêuticos para o tratamento e prevenção das metástases.

Palavras-chave: organotropismo de metástases, rede biológica, interação proteína-proteína, comunicação intercelular, gene carcinogénico

Contents

List of Figures

List of Tables

Glossary

CAR-T cells Chimeric antigen receptor T cells. Autologous CAR-T cell therapy is a patient-specific cellular therapy in which the patient's own T cells are genetically modified to express a chimeric antigen receptor [1]. *(p. 50)*

mTOR Serine/threonine-protein kinase mTOR. The mTOR pathway is a central regulator of cellular metabolism, growth and survival in response to hormones, growth factors, nutrients, energy and stress signals [2, 3]. *(p. 51)*

Notch Group of four cell membrane receptors (NOTCH1, NOTCH2, NOTCH3, NOTCH4). The Notch pathway is important for cell-cell communication, which involves gene regulation mechanisms that control multiple cell differentiation processes [4]. *(p. 9)*

PI3K/Akt The phosphatidylinositol (PI) 3-kinases (PI3K(s)) phosphorylate PI and its phosphorylated derivatives at position 3 of the inositol ring to produce 3-phosphoinositides. Atk is a group of three closely related serine/threonine-protein kinases (AKT1, AKT2 and AKT3). The PI3K/Akt pathway regulate many processes including metabolism, proliferation, cell survival, growth and angiogenesis [3, 5, 6]. *(pp. 4, 8)*

Rac Subfamily of the Rho family of GTPases comprising the Rac1, Rac2, Rac3, and RhoG proteins. The Rac pathway regulates cellular responses such as secretory processes, phagocytosis of apoptotic cells, epithelial cell polarization, neurons adhesion, migration and differentiation, and growth-factor induced formation of membrane ruffles [3, 7]. *(p. 4)*

TGF-β The transforming growth factor beta is a family of three cytokine isoforms (TGFB1, TGFB2, TGFB3). The TGF-β pathway regulates processes such as cell growth, cell differentiation, cell migration, apoptosis, cellular homeostasis [8]. *(p. 9)*

Wnt Wnt family comprises a large number of cysteine-rich glycoproteins. The canonical Wnt pathway (Wnt/β-catennin) mainly controls cell proliferation, whereas the noncanonical Wnt pathways regulate cell polarity and migration, and the two main pathways form a network of mutual regulation [9]. *(p. 9)*

Acronyms

API application programming interface *(p. 41)*

BMDC bone marrow-derived cell(s) *(p. 8)*

CDG cancer driver gene(s) *(pp. 1–4, 44–46, 51, 52)*
ChIP-seq chromatin immunoprecipitation sequencing *(p. 28)*
CTC circulating tumour cell(s) *(pp. 5, 7, 8, 37)*

ECM extracellular matrix *(pp. 4, 6–8)*
EMT epithelial-mesenchymal transition *(p. 6)*

FDR false discovery rate *(pp. 20, 27, 29, 40, 41, 47)*

GEO Gene Expression Omnibus *(p. 19)*
GO Gene Ontology *(pp. 13, 47, 51)*
GTEx Genotype-Tissue Expression project *(pp. 13, 14, 16, 19, 22, 27, 28, 30, 32, 33, 37, 40, 52)*
GWAS genome-wide association studies *(p. 41)*

HCMDB Human Cancer Metastasis Database *(pp. 19, 27, 35, 36, 40, 52)*
HGNC HUGO Gene Nomenclature Committee *(pp. 14, 16)*
HPA The Human Protein Atlas *(pp. 13, 14)*

IQR interquartile range *(p. 16)*

KDE kernel density estimation *(pp. 17, 18)*
KNN K-nearest-neighbours *(pp. 17, 18)*
KS Kolmogorov-Smirnov test *(pp. 25, 41, 43)*

MHC major histocompatibility complex *(pp. 40, 50)*
MWU Mann-Whitney U test *(pp. 25–27, 41)*

NCG^{HD} Network of Cancer Genes & Healthy Drivers *(p. 45)*
NK natural killer *(pp. 4, 7, 8, 40, 50–52)*

OR odds ratio *(pp. 24, 26, 45, 46)*

PMN pre-metastatic niche *(pp. 8, 9, 11, 50, 51)*
PPI protein-protein interaction *(pp. 10–12, 14, 20, 22–24, 28, 30, 35, 37, 44, 46, 50–52)*
PTM post-translational modification *(p. 21)*

RNA-seq RNA sequencing *(pp. 13, 14, 22, 30, 37, 51, 52)*
RWR random walks with restart *(pp. 10, 28, 29, 44)*

scRNA-seq single-cell RNA-seq *(p. 51)*
SRCC Spearman's rank correlation coefficient *(pp. 25, 27, 32, 33, 37, 41)*

TCGA The Cancer Genome Atlas *(p. 19)*
TF transcription factor *(pp. 27, 28, 44, 51)*
TME tumour-specific microenvironment *(pp. 4–6, 8, 11)*
TMM trimmed mean of M values *(p. 14)*
TPM transcripts per million *(pp. 13, 14, 23, 31, 52)*

Symbols

APC APC regulator of WNT signaling pathway *(p. 2)*

B2M beta-2-microglobulin *(p. 40)*
BRCA1 BRCA1 DNA repair associated *(p. 2)*
BRCA2 BRCA2 DNA repair associated *(p. 2)*

CD94 Natural killer cells antigen CD94 *(pp. 40, 50–52)*
CDH1 cadherin 1 *(p. 51)*
CDH2 cadherin 2 *(p. 51)*
cDNA complementary DNA *(p. 51)*

DAP12 TYRO protein tyrosine kinase-binding protein *(p. 52)*
DNA deoxyribonucleic acid *(pp. 2, 5)*

EGFR epidermal growth factor receptor *(pp. 4, 46)*

HLA-E HLA class I histocompatibility antigen, alpha chain E *(pp. 40, 50, 51)*

KLRD1 killer cell lectin like receptor D1 *(p. 40)*

L1CAM Neural cell adhesion molecule L1 *(p. 9)*

miRNA micro RNA *(p. 4)*
mRNA messenger RNA *(pp. 30, 51)*
MYC MYC proto-oncogene, bHLH transcription factor *(pp. 1, 4, 45)*

NKG2A NKG2-A/NKG2-B type II integral membrane protein *(pp. 40, 50, 51)*
NKG2C NKG2-C type II integral membrane protein *(pp. 40, 52)*
NOTCH1 notch receptor 1 *(p. 2)*

RAC1 Rac family small GTPase 1 *(p. 45)*

RANK alias for the Tumor necrosis factor receptor superfamily member 11A protein (TNFRSF11A) *(p. 8)*

RANKL alias for the Tumor necrosis factor ligand superfamily member 11 (TNFSF11) *(p. 8)*

RB1 RB transcriptional corepressor 1 *(p. 45)*

RNA ribonucleic acid *(p. 14)*

TP53 tumor protein p53 *(pp. 2, 45)*

TYROBP transmembrane immune signaling adaptor TYROBP *(p. 52)*

VCAM-1 Vascular cell adhesion protein 1 *(p. 8)*

1 Introduction

1.1 Cancer and Tumorigenesis

Cancer is not a single disease, with a single cause or a one-size-fits-all cure. On the contrary, cancer, neoplasm or malignant tumour are synonyms for a large group of diseases that may start in almost any organ or tissue, comprising over 200 distinct disease entities in humans [10]. It ranks as a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020. The global burden is expected to grow to 30 million new cancer cases and 16.3 million cancer-related deaths based on the expected population by 2040 [11]. Cancers are traditionally classified according to the tissue and cell type from which they arise. Carcinomas are cancers arising from epithelial cells, and they are by far the most common cancers in humans (about 80% of cases). Sarcomas arise from connective tissue or muscle cells. There are also cancers that do not fit in either of these two categories, including leukaemias and lymphomas, derived from hemopoietic cells, as well as cancers derived from cells of the nervous system. In Portugal in 2020, prostate, colorectal, and lung were the most common types of cancer in men, while breast, colorectal and thyroid cancer were the most common among women [12].

Cancer is characterized by abnormal and uncontrollable cell growth. The processes and events leading to cancer development are collectively called tumorigenesis or oncogenesis. Every tumour is the product of many episodes of mutation in a multistep process analogous to Darwinian evolution. Normal human somatic cells progressively accumulate genetic mutations, some of which confer some type of growth advantage, driving the progressive transformation into highly malignant derivatives [13]. Rarely, however, does mutation in a single gene lead to the onset of cancer and not all mutations contribute to cancer phenotypes (passenger mutations). Mutations that do play a critical role in the development of cancer occur in specific genes known as cancer driver gene(s) (CDG) [14]. CDG are usually expressed in most tissues and cells, though their products coexist in different tissue-specific contexts, with distinct molecular neighbours and interactors. Genes that drive the development of cancer phenotypes by a gain-of-function (upregulation) mutation are designated as proto-oncogenes, becoming fully-fledged oncogenes after mutating. They positively control processes that, when up-regulated, give cancer cells a survival advantage against normal cells [15]. This is the case of *MYC*, a proto-oncogene that positively regulates cell proliferation [16]. On the contrary, tumour-suppressor genes drive

cancer development when they suffer a loss-of-function mutation, and so, negatively control similar processes. That is the case of *TP53*, which blocks cell division after DNA damage [15, 17]. Interestingly, some CDG, as is the case of the *NOTCH1*, can exhibit both tumour-suppressor and oncogene behaviour depending on the tissue context, cell type, and mutational signature [4].

But how do the conditions for the emergence of abnormal cell behaviour appear? How do mutations appear and drive cancer progression? A complex multicellular organism can be seen as a society or ecosystem whose individual members are cells that reproduce by cell division and organize themselves into collaborative assemblies called tissues. Cancer represents a breakdown of this multicellular cooperation, an emergence of cell-level fitness at the expense of the fitness of the organism as a whole [18]. Thus, multicellular organisms have developed comprehensive tumour suppressor mechanisms that control cell proliferation and behaviour, which are specific for each particular tissue [19]. These controls are exerted by extracellular signals sent, received, and interpreted between cells through signalling cascades, driving gene expression and directing them on how to act—resting, growing, dividing, differentiating, or dying. Any cell that escapes the controls is triggered to undergo apoptosis [15].

Nevertheless, mutations naturally accumulate throughout the lifetime of a multicellular organism. So, except for some neoplastic diseases that have a paediatric aetiology, cancer is intimately related to the ageing process [20, 21]. In each cell division, there is a probability of occurring genetic or chromosomal alterations, a risk that increases with the exposure to certain mutagens such as chemical carcinogens (e.g. asbestos, benzo[*a*]pyrene), ultraviolet (UV) light, certain viral infections (e.g. Human Papilloma Virus, Epstein-Barr Virus), chronic inflammation, amongst others factors [10, 13]. Variants of certain genes also increase the risk of developing cancer, such as in the *BRCA1* and *BRCA2* genes, of which carriers of certain variants have a significantly larger risk of developing breast cancer [21]. These acquired mutations in the genome of stem or progenitor cells may increase the fitness of their progenies, fuelling clonal expansion that populate part of a tissue, creating a patchwork of mutant clones [22]. For some cancers, research has been able to pinpoint the first mutation that provides a normal cell with an increased fitness that drives clonal expansion. These "gatekeeping" mutations are known, for example, for colorectal cancer, where they happen most often in the *APC* gene, a tumour suppressor that controls growth signals. [21] Though cancer is fundamentally a disorder driven by genetic mutations, epigenomic alterations, such as aberrant DNA methylation and remodelling of histone modifications, are also important in tumour evolution. Aberrant DNA methylation is induced by ageing and accelerated by chronic inflammation [23]. These factors, combined with the tendency for cellular repair mechanisms to be less effective as an individual ages, burden somatic tissues with driver and passenger mutations. Despite mutational signatures in normal tissues being distinct from the ones in cancer, 90% of genes driving somatic expansion in normal tissues are also known CDG [24]. Mutational loads, cancer drivers, "gatekeeping" mutations, and epigenetic changes are also tissue-specific and depends on the tissue microenvironment and cell neighbourhood [25].

So, it is the increased genomic and chromosomal instability leading to mutations in progressively more drivers, coupled with sporadic catastrophic events and epigenetic changes, what eventually leads down a path for the appearance of malignant phenotypes. But what are

these phenotypes, and how can they be identified and characterized? Hanahan and Weinberg proposed the **Hallmarks of Cancer**, a set of functional capabilities acquired by cells, as they progress from normalcy to neoplastic growth states. Besides hallmarks, the authors also suggest the existence of **Enabling Characteristics**, which consist of traits possessed by cancer cells that create the conditions for the acquisition of those functional characteristics. The Hallmarks of Cancer were first described in the year 2000 [26], and further updated in 2011 [27] to reflect new and improved knowledge about the tumorigenesis process and comprise eight Hallmarks and two enabling characteristics (Figure 1.1a). A recent publication by Hanahan [28] consolidates the knowledge supporting the hallmarks advanced in the two previous editions and proposes two new hallmarks and enabling characteristics (Figure 1.1b). Since it is a fairly new update (added during the development of this work), these have not been widely discussed and are not incorporated in the resources used in this work.

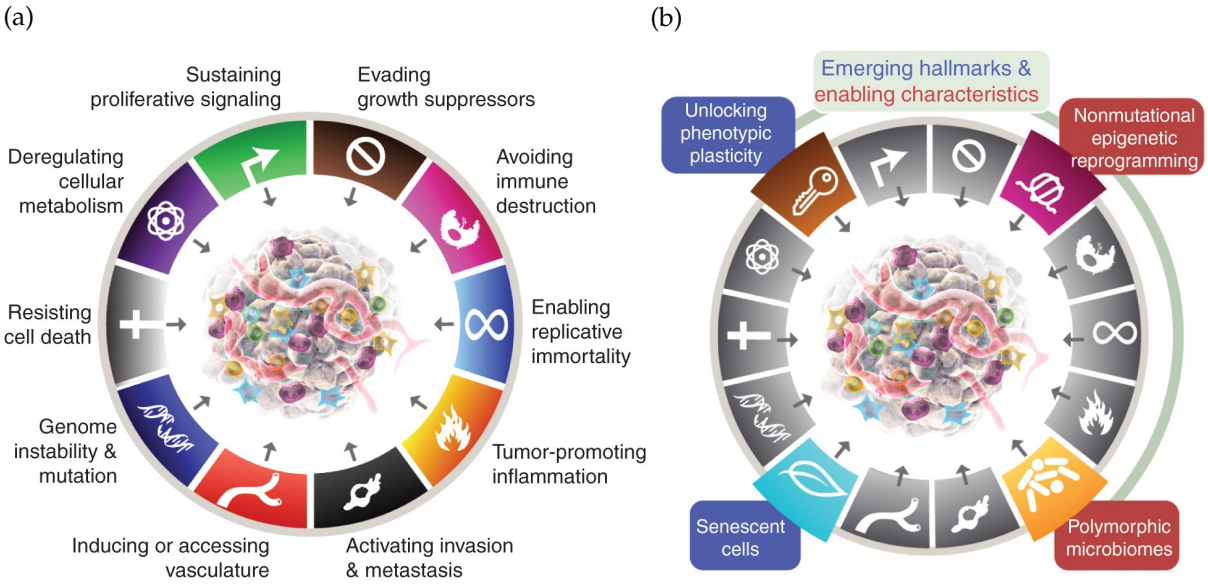


Figure 1.1: The Hallmarks of Cancer. (a) Hallmarks and enabling characteristics as defined by Hanahan and Weinberg in 2011 [27]. (b) 2022 update by Hanahan [28]. Adapted from [28].

The hallmarks *Sustaining proliferative signalling*, *evading growth suppressors*, *resisting cell death*, *enabling replicative immortality* describe characteristics that confer cancer cells their most recognized behaviour: the ability to control their own fate. For these traits to appear, cancer cells undergo an extensive remodelling of their intracellular signalling pathways, allowing them to ignore extracellular stimuli or generate a response even in the absence of a signal. But as a mutation in a single CDG is not enough to drive cancer progression, more than a single altered pathway is necessary to lead to the appearance of cancer phenotypes. Signalling pathways are not linear cascades that relay information and are not isolated inside a cell. Instead, they are highly structured and interact between each other (cross-talks) through some of their molecular component, creating a veritable network of intracellular regulatory interactions [29]. Distinct pathways might share some of their components, control the expression of the same genes, and drive the same cellular functions. In cancer, the extent of alterations in a pathway can vary widely depending on tissue and cell-specific factors, but usually mutations affect more

than one component (CDG). Due to the high degree of interconnectivity, specific mutations can change the interplay dynamics between two pathways: some might be synergetic and lead to the development of cancer phenotypes; others can be mutually exclusive and originate a lethal phenotype [30]. Another cancer hallmark that is closely related to signalling pathways is *deregulating cellular metabolism*. Oncogenic mutations in signalling pathways that control cell growth and proliferation, such as PI3K/Akt and c-*MYC*, directly affect the dynamics of metabolic pathways [31]. Such alterations in nutrient uptake and usage are what enable cancer cells to survive nutrient-poor environments, hypoxia, and drive an elevated rate of proliferation [32].

During tumour growth, the tissue-specific microenvironment surrounding cancer cells is gradually transformed to create a tumour-specific microenvironment (TME). The TME is a compartment with distinct biochemical and biomechanical properties from the surrounding normal tissue. Besides cancer cells, it comprises several cell types of various origins, including mesenchymal stromal cells (e.g. fibroblast, mesenchymal stem cells), cells of the immune system (e.g. macrophages, natural killer (NK) cells), and peripheral nerve cells [33]. It also includes highly modified extracellular matrix (ECM), which provides structural support, and an environment for cell-cell communication. Within the TME there is a highly dynamic environment, where cancer cells interact with each other, cooperate, and compete for resources. Intercellular communication is driven by a complex and dynamic network of cytokines, chemokines, growth factors, mediated by ECM proteins and matrix remodelling enzymes that sequester and control their accessibility to cell receptors [34]. Small molecules can also be delivered over larger distances by extracellular vesicles and exosomes. Direct cell-cell contact through adherens junctions mediated by cadherins stabilizes cells, and gap junctions mediated by connexins allow transfer of ions and metabolites [35]. Thin membrane tubes called tunnelling nanotubes mediate the direct connection of the cytoplasms of two cells and allow the transfer of cellular content between cells, including large molecules such as proteins and miRNAs, and even organelles such as mitochondria and lysosomes [36]. This type of communication can help cancer cells survive nutrient depleted environments inside the TME. Direct interaction between cancer cells and the ECM mediated by integrins activates several intracellular signalling pathways, as is the case of *EGFR* activation of Rac to promote cell survival [37].

The complex environment of intercellular interactions inside the TME modulates tumour behaviour and stimulates the appearance of tumour phenotypes. Clonal tumour cells in the TME secrete angiogenic factors to increase vascularization (*Inducing or accessing vasculature*), and promote inflammation by modulating tumour-associated immune cells (*Tumour-promoting inflammation*) [37, 38]. Chronic inflammation and persistence of antigen in the TME lead to T cell exhaustion, which is characterized by a progressive loss of T cell effector functions, mainly due to the expression of high levels of inhibitory receptors (*Avoiding immune destruction*) [39]. Finally, TME dynamics and interactions also play a crucial role in tumour progression and metastasis (*Activating Invasion & Metastasis*). The ability of cancer cells to invade adjoining tissue and/or spread to other organs is what sets apart cancer from benign growths. Metastasis development is the main focus of this dissertation, and we will look into what is known about this process in the next section.

1.2 The Metastatic Cascade and Organotropism

Metastasis is characterized by the dispersal and colonization of cancer cells from a primary tumour to other organs in the body. Unlike primary tumours, which can be treated by localized therapies such as surgery and radiotherapy, metastasis is a systemic disease that can affect different organ and tissue sites, and is commonly resistant to conventional cancer therapeutics. This explains why secondary tumours are responsible for the majority of cancer related deaths, and stresses the importance of elucidating the mechanisms and factors governing the metastatic process [40, 41].

Metastasis is a very inefficient process. Depending on their size, tumours can shed millions of cells each day, and some tumours start disseminating cells early in their development [42]. Even so, mouse models suggest that the majority of circulating tumour cell(s) (CTC) do not survive to form metastases, and even those that arrive at a distant organ commonly undergo apoptosis [43, 44]. To successfully colonize other organs, tumour cells must exhibit special characteristics that allow them to surpass all barriers and bottlenecks in their way. Yet, cancer cells themselves are not under positive selection to metastasize, and mutation patterns and the overall mutational burden in primary and metastatic cancers are largely concordant. Very few unique mutations associated with metastasis have been identified, though genes that regulate DNA and chromatin modifications are frequently mutated in aggressive tumours. Alterations at the epigenetic level could favour phenotypes that are fit to survive the metastatic process, but this certainly does not explain all the acquired adaptations [45]. So, what are the processes that allow cancer cells to acquire the necessary traits to successfully spread to distant organs?

The main driving forces behind the development of metastatic potential in cells are intra-tumoural heterogeneity and phenotypic plasticity. As previously explained, tumour cell evolution is driven by genetic instability. Primary tumours are composed of highly heterogeneous sub-clonal populations of cells, all direct descendants of the cell carrying the mutation which originated the first clonal expansion responsible for tumour progression. These subpopulations of cancer cells distinguish between themselves by their mutation patterns and overall mutation burden, which is constantly increasing and changing, driving cell evolution inside the TME [46, 47]. Accumulated mutations that can play a role in metastasis formation are not necessarily subjected to somatic selection inside the TME, but since they do not affect tumour progression, they are silently carried by subpopulations of cells [40]. In turn, genome heterogeneity contributes to phenotypic variability and plasticity. Some cancer cells can acquire stem-like characteristics that confers them the ability to adopt diverse phenotypic states in response to external signals and cell-intrinsic programs [41]. This suggests that primary tumours comprise some highly adaptable and competent dominant clonal subpopulations, expressing the characteristics necessary to overcome metastasis barriers and survive in foreign environments. Primary tumour cells that are only partially competent can attain full metastatic potential through stochastic events that change their epigenetic programs [48].

As already mentioned, metastasis proceeds through multiple steps and restrictive bottlenecks, collectively known as *metastatic cascade* [49]. These obstacles are different, depending on the route taken by tumour cells during dissemination throughout the body. Access to all

organs of the body (lymph nodes excluded) happens predominantly through the haematogenous circulation, making it the principal route of dissemination for most types of cancer [50]. Lymphatic dissemination and some forms of extravascular spread are also prominent routes in various types of cancer [51]. For example, ovary cancer cells spread mainly within the peritoneal cavity, invading the peritoneal mesothelium, and rarely establish haematogenous metastasis [52]. But since the haematogenous circulation is the predominant and most well studied route of metastasis, we will briefly describe the steps, molecular factors and interactions involved in that process. The metastatic cascade is commonly divided in five distinct steps, representing bottlenecks faced by metastasizing cells during migration (Figure 1.2).

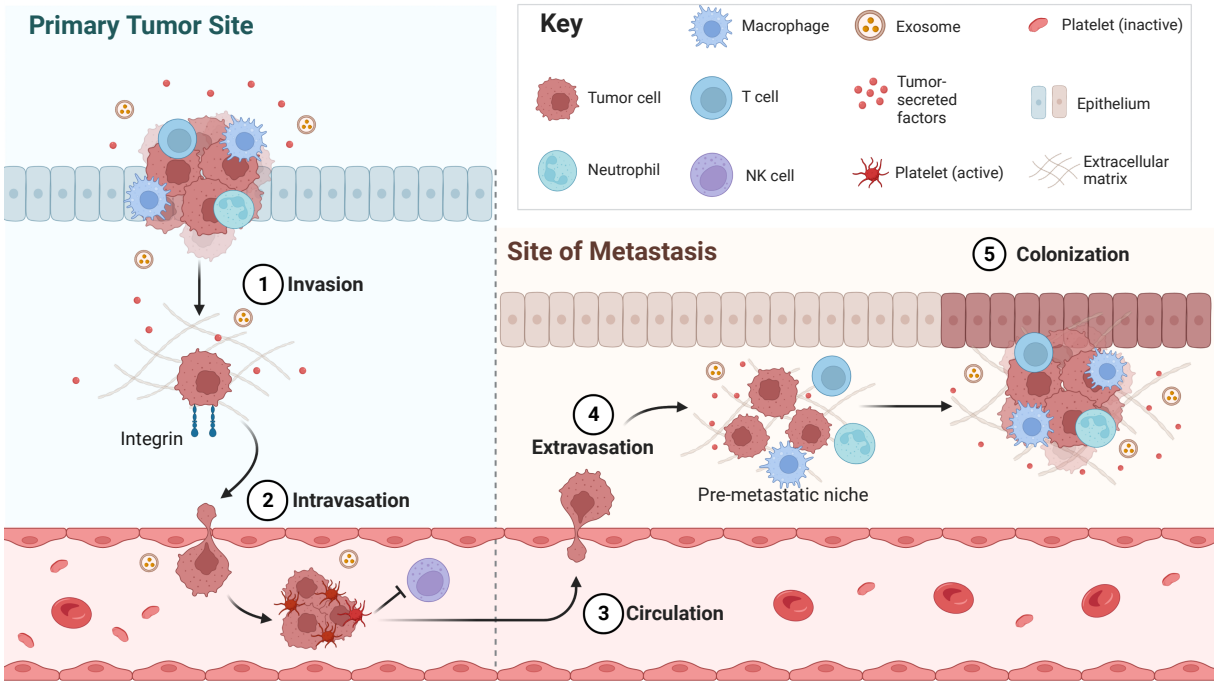


Figure 1.2: The metastatic cascade. Created with BioRender.

First, to reach circulation, metastasizing cells have to invade the surrounding tumour-associated stroma and adjacent normal tissue (1), and then squeeze through blood vessel walls (2), a process facilitated by the leaky structure of the neovasculature promoted by the TME [40]. Invasion requires the activation of signalling pathways and the expression of factors that increase motility, and invasiveness. These traits are enabled by cell-cell interactions through dynamic changes in the function of cell adhesion molecules (cadherins, IgCAM family, and selectins), ECM remodelling enzymes (matrix metalloproteinases and cathepsins), and cell-matrix adhesion molecules (integrins and syndecans). Intracellular signalling pathways activated by integrin-mediated adhesions, and by growth factors and cytokines released during ECM remodelling, influence migration success through control of cell growth, proliferation, survival, and inflammation [40, 42, 51]. A common feature of the phenotypic plasticity in carcinoma cells is the epithelial-mesenchymal transition (EMT) process. EMT is a reversible phenotypic change in which cells lose intercellular adhesion and epithelial polarization and gain mesenchymal traits that confer the needed increase in motility and invasiveness [41]. Moreover,

migration of tumour cells into the blood stream can occur through single-cell dissemination or collective migration. When tumour cells disseminate in clusters, there is a clear difference in gene expression, morphology, and function between cells. Leader cells (which interact with the surrounding environment) exhibit a high degree of plasticity and mesenchymal characteristics, whereas follower cells retain epithelial traits [53]. Though cells commonly lose expression of cell-adhesion molecules such as E-cadherin, to increase their mobility when migrating alone, in collective migration, adhesion interactions are required to tightly link clusters of cells [54].

In the bloodstream (3), to avoid damage by exposure to haemodynamic shear forces and stresses due to the loss of ECM adhesion, cancer cells associate with platelets in microaggregates. This CTC-platelet interaction is mediated by selectins, and has also the benefit of helping CTC avoid immune detection, by blocking NK cells from interacting with cell receptors [48, 55]. Eventually, CTC become entrapped in capillary beds at distant organs, extravasate into the tissue parenchyma (4), form micrometastasis and stay dormant until ideal conditions occur to start proliferation and development of a secondary tumour (5). The success of these sequential processes depend not only on physical/anatomical factors and on the intrinsic abilities of the tumour cells themselves, but also on the characteristics of the welcoming tissue. Mechanical entrapment is considered to be the main mechanism for cancer-cell arrest, but cell adhesion interactions also play a role in CTC arrest and extravasation [42]. CTC are capable of forming specific adhesive interactions in particular tissues that favour their trapping, and extravasation usually involve adhesion to the endothelium, modulation of the endothelial barrier, and transendothelial migration to invade the tissue parenchyma [55]. Alternatively, CTC lodged in the microvasculature can initiate intraluminal growth, forming a microcolony that eventually ruptures the vessel and extravasate by breaching vascular walls [40].

The blood flow patterns in the body usually determine that the first organs encountered by CTC are the lungs and liver, which explains, in part, why these are common sites of metastasis for several cancers. Permeable capillaries called sinusoids, such as those of liver and bone marrow, have gaps that might facilitate the extravasation of CTC and contribute to the high incidence of liver and bone metastasis [42]. Nevertheless, and though rapid arrest improves the chance of survival, some CTC bypass these initial filters to reach other organs through the arterial circulation. In fact, metastatic patterns are non-random and considerably different among cancer types. Most cancers display a high incidence of metastasis to specific organs. For some of these "preferred" sites, propensity for metastasis cannot be explained alone by blood flow patterns nor by easy access to the organ stroma, a phenomenon known as metastasis organotropism or organ-specific metastasis [56, 57]. In 1889, Stephen Paget was the first to propose an explanation for organotropism based on host-tumour cell interactions. His *seed and soil* hypothesis states that certain tumour cells (seeds) have an intrinsic affinity or compatibility to particular organ environments where they thrive (congenial soil), akin to how plant seeds spread everywhere but only thrive in particular welcoming soils [58]. Current evidence supports and expands this hypothesis to include the concepts of organ microenvironment and metastatic niche as factors in the outcome of metastasis [59].

The "seed and soil" hypothesis, though an appealing metaphor, misleads when it introduces the concept of congenial soil –no organ microenvironment is really compatible and welcoming

to metastasizing cells. In fact, colonization after extravasation is probably the main limiting step in the metastatic process, with less than 0.02% of cells ending up generating macroscopic metastasis [42, 43]. Tissues have a repertoire of immune surveillance mechanisms that are the first line of defence against disseminated cancer cells. Outside the protective TME, invasive cells are particularly vulnerable to tissue-resident macrophages, T cells and NK cells, as well as to non-immune cells such as astrocytes in the brain [41]. Due to the inhospitable environment of tumour-free secondary sites, many CTC end up reseeding the primary tumour, a process known as tumour self-seeding. Self-seeding requires little, if any, additional adaptation of CTC to the recipient microenvironment, and can help to select for subpopulations that are more aggressive and metastasis-ready [60].

So, how do organ microenvironments become more welcoming to disseminated cancer cells? There are two main factors that contribute to the success of a metastasizing cell in a distant organ. First, cancer cells capable of forming metastasis express, beforehand, factors that allow them to communicate, interact, and thrive in particular organ microenvironments. Many growth and survival signalling pathways that are altered in primary cancer cells are found to be amplified or have an expanded output in metastasizing cells. For example, high expression of the VCAM-1 in breast cancer cells hypersensitizes the PI3K/Akt cell survival pathway to activation by external signals such as interactions with $\alpha 4\beta 1$ integrins [61]. Besides, some factors induced by both cancer driver mutations and tissue- or cell-specific programs, might predispose cancer cells to metastasize to certain organs. For example, both normal mammary epithelial cells and breast cancer cells express RANK, which is the receptor for the cytokine RANKL. RANKL, an osteoclast differentiation factor highly expressed in the bone marrow, triggers migration of cancer cells that express the RANK receptor [50, 62]. Second, cross-talk between primary tumour cells and distant organ microenvironments start even before CTC arrive at the organ site. Primary tumours and CTC secrete factors such as cytokines, chemokines and hormones, soluble or packed in extracellular vesicle, that prepare a welcoming niche to receive metastasizing cells—the pre-metastatic niche (PMN). The PMN is a recent concept, developed in the past decade, that can be defined as a supportive and receptive tissue microenvironment with the conditions for the survival and proliferation of metastasizing cells at the distant organ [63]. Survival of metastasizing cells in these niches depends on stromal signals, cell adhesion, extracellular matrix interactions, and metabolic cues. The PMN is tissue-specific, thus, many determinants of metastatic tropism differ between cancer types and secondary organ sites. Nevertheless, some characteristics of the PMN are common in every organ niche and are promoted by factors secreted by primary tumour cells. They include inducing immune suppression to avoid detection of metastasizing cells by tissue-resident T cells and NK cells; promoting inflammation with cytokines that regulate tumour growth; recruiting non-resident cells such as bone marrow-derived cell(s) (BMDC) that secrete CTC attracting factors; remodelling the ECM to allow the adhesion of CTC and BMDC; and inducing vascular permeability and angiogenesis to facilitate the invasion of the tissue stroma by CTC [63, 64]. Conditions that allow the survival of a metastasizing cell when arriving at a particular PMN might not be ideal to promote proliferation. So, they might enter a phase of proliferative quiescence known as protective dormancy. Tumour cells can remain as dormant single cells or in

micrometastasis clusters for months and even years. Barriers that prevent tumour development include oxidative stress, immune surveillance, and physical condition in the stroma. Intercellular interactions play an essential role in promoting tumour cell survival, proliferation arrest, and eventually in stimulating their exit from dormancy. Entry into dormancy is promoted by growth-inhibitory signals in the PMN, such as TGF-β. Interactions that might stimulate escape from dormancy include integrin-mediated cell signalling, cell-adhesion interactions through L1CAM, and activation of Wnt and Notch signalling by factors produced by fibroblasts [51, 65].

1.3 Network Biology

The traditional reductionism paradigm, which dominated biochemistry and molecular biology research during the twentieth century, involves isolating and studying each cellular component separately [66, 67]. However, a biological system such as a cell is not just an assembly of its components. Instead, most biological characteristics (phenotypes) arise from complex interactions between each molecular component. Aided by the advent of high-throughput technologies and the increase in computational power, Systems Biology emerged in the past decades as an answer to the shortcomings of the reductionist approach. Applying systems thinking to study complex biological systems means taking into account all components, their status (at a particular moment in time or their change through time), and their interactions [67, 68]. Computational Systems Biology employs this holistic approach to extract patterns in large experimental datasets, and study the dynamics of a system that give rise to emergent behaviours not explained by the characteristics of each component alone [69].

A major field in computational biology is network biology, which applies graph theory as a quantitative approach to describe networks that characterize biological systems. A graph $G = (N, E)$ is an abstract mathematical object used to represent networks, where $n \in N$ is the set of nodes (components) and $e \in E$ the set of edges (interactions) between nodes [70]. There are several types of graph categories. Depending on the nature of the interactions, networks can be directed or undirected. In directed networks, the interaction between any two nodes has a well-defined direction, which represents, for example, the direction of information flow in a signalling cascade (Figure 1.3a). In undirected networks, the edges do not have an assigned direction (Figure 1.3b). The interactome, which represents the physical binding relationship between proteins, is an example of an undirected network [67]. A multigraph (Figure 1.3c), in opposition to a simple graph, allows for multiple edges (edges that connect the same nodes with the same direction) and loops (edges that joins a node to itself). Finally, it is possible to assign numerical values (weights) to each edge to form a weighted graph (Figure 1.3d) [71].

The structure, also known as topology, is the pattern of connectivity between nodes in the graph. Studying networks concerns understanding the impact of the individual characteristics of each node in the global properties of its structure. Centrality measures, such as the degree—the number of edges connected to a node—give a way to quantify the importance of each node in a network [71]. The clustering coefficient C_i, which is defined as the ratio between the number of edges linking nodes adjacent to node i and the total possible number of edges among them, quantifies how close the local neighbourhood of a node is to being part of a clique—a region

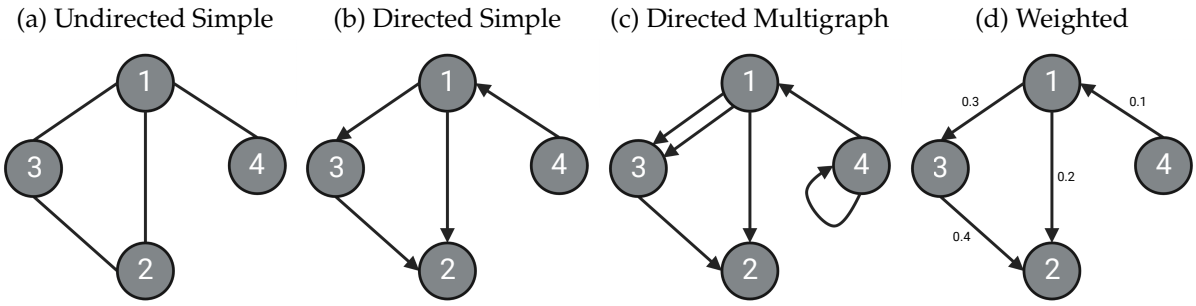


Figure 1.3: Example of possible graph types. Circles represent nodes and lines/arrows represent edges. Created with BioRender.

of the graph where every node is connected to each other [67]. Network diffusion algorithms, including random walks with restart (RWR), assume that information located on nodes can be transmitted to their neighbours through the edges connecting them. This process can be applied, for example, to simulate the propagation of a signal through a network of regulatory interactions and find the pathways and nodes more affected by a perturbation such as a disease [72]. Taken together, these methods can be used to characterize the global topological properties of a network, the similarity, or connectivity between nodes and define different network models.

Random network models are constructed by randomly connecting N nodes to E edges with a probability p. The node degrees follow a Poisson distribution, which indicates that most nodes have approximately the same number of edges. Random networks are statistically homogeneous because extreme values of degree and large values of C_i are very rare. Also, the mean path length between any two nodes is proportional to the logarithm of the network size, which indicates relatively short paths between any pair of nodes—a property known as small-world. On the contrary, biological networks such as protein-protein interaction networks have a *scale-free* topology. The degree distribution of scale-free networks follows a power law $P(k)\ k^{-\gamma}$, where γ is the degree exponent. This means that the network is characterized by a few highly connected nodes known as hubs, while the majority of nodes has a small number of edges. Besides, biological networks show a high degree of clustering, which suggests the existence of highly connected local regions—modules. Biological networks also have the small-world property, but their average path length is smaller than what is predicted for random networks. This means that information flow is highly efficient and that any local perturbation reaches the whole network very quickly [67, 73]. Those characteristics are conserved throughout evolution because they confer biological networks one of their most important property: robustness, i.e., the ability of a network to respond to changes while maintaining normal function and behaviour. Hub proteins are central nodes in a protein-protein interaction (PPI) networks, thus the deletion of any hub would be catastrophic for the cell. Nonetheless, as only a few proteins are hubs, most deletions, or change of function mutations affect non-hub proteins. This means that the cell can still perform its function by using alternate paths that do not pass by the affected node. Modularity also contributes to network robustness. Proteins with the same specific function and those that form complexes tend to cluster into several functional communities that are highly conserved. Modules are highly connected, which facilitate the rewiring of a pathway upon

failure of a node. Additionally, highly connected proteins tend to be linked with low-connected proteins, which weakens the communication between modules, probably limiting the effects of local perturbations in cellular networks [67, 70].

Several types of cellular networks can be constructed. PPI network characterize physical interactions between proteins; metabolic networks track the chemical conversions between metabolites; gene regulatory networks represent the relationship between transcription factors and their target genes. All cellular networks have similar properties and can be affected by diseases that limit or change the function of a node. The ability of networks to reproduce the complex interplay between molecules inside a cell, makes them a well suited tool for studying complex diseases. By understanding the context of an altered gene in the network, one can predict the phenotypic impact of the defect and its role in disease. This impact is observed in the local and global properties of the network, which shift as networks suffer a rewiring during disease progression. Network medicine is the application of network biology to characterize human disease. Disease proteins interact closely with each other, forming connected sub-graphs called disease modules, that usually overlap with the topological and functional modules in biological networks. This suggests that by looking in the neighbourhood of disease-associated proteins, one can identify new disease proteins and the processes in which they participate [74].

1.4 Hypothesis and Goals

Our hypothesis builds upon five premises already described in the introduction:

1. Metastatic patterns and driver mutations are mainly identical between primary tumour and secondary foci, with the differences between the cells of these foci commonly ascribed to epigenetic changes that lead to the amplification of certain cell programs.

2. Intercellular interactions are essential to the success of metastasis. The chances of a tumour cell surviving the metastatic process depend on how it responds and communicate with the surrounding environment, for example, in avoiding immune surveillance or co-opting other cells to its cause.

3. Both the TME and the PMN are tissue-specific. Tumour cells from different tissues of origin express distinct proteins that participate in cell-cell interactions. Also, the requirements for a cell to survive inside the PMN are different between tissues.

4. Some factors expressed in non-diseased cells play a role in the metastatic process and are specific to some tissues and/or cell-types.

5. Network biology tools are well suited for studying complex diseases such as cancer. With networks, one can capture the complex interplay between proteins and pathways affected in cancer and metastasis and identify new metastasis-associated genes.

Our hypothesis is that organotropism, i.e., the apparent preference for cancers to metastasize to specific tissues and organs, can be partially explained by the level of communication established between metastasizing cells and cells in the PMN. Differences in cell-cell communication

can be attributed to the tissue-specific expression of factors involved in intercellular interactions, and that these factors can be selectively expressed in normal tissues. With this in mind, we will first build tissue-specific intercellular PPI networks between pairs of cancer and metastasis tissues. We want to quantify and compare the level of communication between pairs with a high incidence of metastasis to those with a low incidence. Second, we will use these networks to identify interactions and proteins that might influence the success of metastasis to specific organs. Ultimately, we will establish a connection between the metastasis-associated proteins and an intracellular signalling network in order to ascertain whether cancer-specific genes and processes are associated with them.

2

Materials and Methods

2.1 Programming Environments and Packages

Most of this work was done in the Python programming language on the Jupyter notebook environment, employing the following packages: *NumPy* [75] and *Pandas* [76] for data structures and manipulation; *SciPy* [77], *Statsmodels* [78] and *Scikit-learn* [79] for statistics and machine learning; *iGraph* [80] for network building and analysis; and Plotly [81] for data visualization. The package *clusterProfiler* [82] for the R programming language was used to perform Gene Ontology (GO) enrichment analysis. This document was created with the (pdf/Xe/Lua) LaTeX processor and the novathesis template (v6.10.5) [83].

The code developed during this project is fully available at code repository. The raw data files available at data repository.

2.2 Analysis of Gene Expression Data

2.2.1 Gene Expression Datasets

We used 2 distinct datasets of gene expression across healthy human tissues to build tissue-specific intercellular PPI networks.

The Genotype-Tissue Expression project (GTEx) is a project established to study tissue-specific gene expression and characterize genetic effects on the transcriptome across human tissues [84]. At the time of this work, the GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2) provides the most up-to-date RNA sequencing (RNA-seq) datasets, based on data generated from 17382 non-diseased tissue samples collected from 948 postmortem donors. We used the gene median levels *GTEx dataset* available at the GTEx Portal and obtained on 2021/11/16 at 12:12:24. This preprocessed RNA-seq dataset contains median transcripts per million (transcripts per million (TPM)) levels for 56200 unique Ensembl gene identifiers [85] across 54 tissues and 2 cell lines.

The The Human Protein Atlas (HPA) project aims to map human protein expression in cells using a multi-omics approach, including antibody-based imaging, mass spectrometry-based proteomics, and transcriptomics [86]. We used the transcript expression levels *Consensus dataset* based on the HPA version 21.1 and downloaded from the HPA portal on 2021/12/10

at 12:59:08. This dataset summarizes HPA and GTEx transcriptomics data across 55 human non-diseased tissues, normalized using trimmed mean of M values (TMM) to allow for between-sample comparisons. To combine the two data sources, a consensus normalized expression value (nTPM) is calculated as the maximum normalized value for each gene [86]. In total, the Consensus dataset comprises nTPM levels for 20090 Ensembl gene identifiers across 55 tissues.

As mentioned above, both datasets present expression values in TPM. Converting raw counts to TPM levels is a common normalization procedure in RNA-seq data analysis that corrects for both library size and gene length to allow within-sample comparison of gene expression levels [87]. The Consensus dataset employs TMM as an additional normalization step. TMM can be used to remove some bias of TPM values for between samples comparison, or can be used instead of TPM directly in raw counts data [88]. Since this step is specially useful when performing differential expression analysis, we reason that the median levels provided in the GTEx dataset can be used as is to build the PPI networks.

Distributions of RNA-seq data are highly skewed to large values (there is no upper limit of expression). So, as an additional normalization procedure, a log_2 transformation was performed in both datasets as $\log_2(\text{TPM} + 1)$. Though this procedure might reduce the biological signal of the data [89], decreasing data dispersion is needed to compare expression values between tissues and build tissue-specific networks (Supplementary Figure A.1).

Finally, we introduced an expression cut-off of 1 TPM: genes with expression below the cut-off in all tissues were removed.

2.2.2 Gene identifier (ID) Mapping

The Ensembl gene IDs provided in the GTEx and Consensus datasets identify not only protein-coding genes, but also noncoding RNA genes, pseudogenes and gene variants, among others. Despite each dataset also providing generic gene symbols for each entry, we found many duplicated symbols, old symbols and alias, and non-approved IDs. Since we needed consistent and unique protein-coding genes to build PPI networks and compare the results for both datasets, we mapped and translated all Ensembl gene IDs into unique HUGO Gene Nomenclature Committee (HGNC) names [90].

Gene ID mapping was performed using the complete HGNC approved dataset, downloaded from HGNC database on 2021/11/26 at 11:34:43. The developed algorithm not only maps Ensembl gene IDs, but also checks the gene symbols provided in each gene expression dataset to allow mapping of entries with old or deprecated Ensembl IDs (Figure 2.1). As Ensembl genome assembly models are used to align and map RNA-seq reads, gene ID mapping takes precedence over gene symbols mapping. This means that entries mapped using a provided gene symbol are dropped if that symbol is already used in an entry mapped with the Ensembl ID. To further ensure consistency, all gene expression entries with the same Ensembl IDs were discarded.

2.2.3 Tissue and Organ Identifiers

The GTEx and Consensus datasets have distinct ways of identifying human tissues, organs and anatomical locations. Some tissues correspond to different locations on the same organ. For

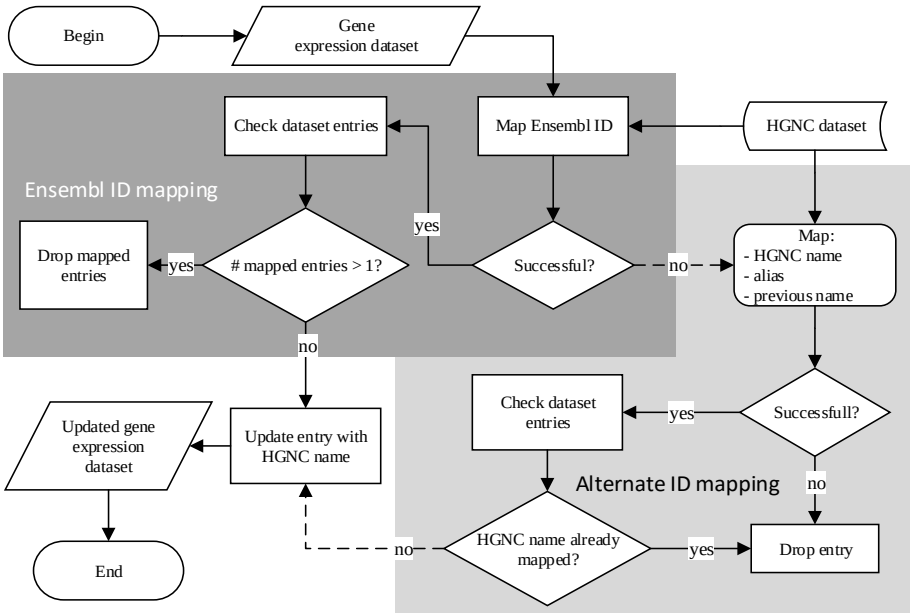


Figure 2.1: Flowchart of the Ensembl ID mapping algorithm: diagram describing the processing of a database entry. The algorithm iterates over all entries. Ensembl ID mapping and Alternate ID mapping are not concurrent processes, which means that all entries are processed before the latter process begins.

example, several brain regions appear in both datasets with distinct gene expression signatures (Table 2.1). Moreover, datasets of metastasis frequencies omit details about cancer or metastasis location on a particular organ.

Table 2.1: Brain regions in GTEx and Consensus datasets.

GTEx Name	Consensus Name
Brain - Amygdala	amygdala
Brain - Anterior cingulate cortex (BA24)	basal ganglia
Brain - Caudate (basal ganglia)	cerebellum
Brain - Cerebellar Hemisphere	cerebral cortex
Brain - Cerebellum	choroid plexus
Brain - Cortex	hippocampal formation
Brain - Frontal Cortex (BA9)	hypothalamus
Brain - Hippocampus	medulla oblongata
Brain - Hypothalamus	midbrain
Brain - Nucleus accumbens (basal ganglia)	pons
Brain - Putamen (basal ganglia)	thalamus
Brain - Substantia nigra	white matter

To standardize tissue labels and designations across all datasets, we manually curated a list that matches the tissue names in each dataset to a unique tissue ID (Supplementary Table B.1). Different tissues/regions of the same organ were matched to the same tissue ID.

2.3 Tissue-specificity

We used the Tau Tissue Specificity Index [91] to characterize the tissue-specificity of genes across tissues on the GTEx and Consensus datasets. The tau index (τ), is defined as

$$\tau = \frac{\sum_{i=1}^{n}(1-\widehat{x_i})}{n-1}; \quad \widehat{x_i} = \frac{x_i}{\max\limits_{1\leq i\leq n}(x_i)}, \tag{2.1}$$

where n is the number of tissues and x_i is the expression of the gene in tissue i.

Two lists of housekeeping and tissue-specific genes determined by Dezső et al. [92] were used to evaluate the τ distribution of genes. The lists were downloaded on 2021/12/01 from Housekeeping Genes and Tissue-specific Genes, respectively. The genes in the two lists are identified by their NCBI Entrez Gene ID [93] and were converted to HGNC Gene Names using the complete HGNC approved dataset, as described in Section 2.2.2.

2.4 Outlier Detection

An outlier or anomaly can be defined as a point (or set of points) that has an extreme value compared to the remaining data. Another common definition for outlier is a point that was generated through distinct mechanisms and belongs to a different distribution [94]. Outliers can appear due to experimental errors, noise, or arise from natural variation in the underlying data. Since this anomalous points do not conform to the expected behaviour of the distribution, they are normally treated differently from the rest. Several methods exist to detect outliers, including statistical methods and tests, clustering-based methods and classification-based methods [95].

The *Tukey's Fences* is a simple non-parametric statistical method to detect outliers, commonly used to determine the whiskers in box plots [96]. A data instance with a value lower than the lower fence ($Q1 - k \times \text{IQR}$) or higher than the upper fence ($Q3 + k \times \text{IQR}$) is considered an outlier, where IQR is the interquartile range (Figure 2.2).

Two values are commonly used to set k:

- $k = 1.5$: the value used to build box plot whiskers, determines the inner fence. The region between fences contains 99.3% of observations, which means this method is equivalent to the 3σ rule for normal distributions (Figure 2.2).
- $k = 3$: determines the outer fence. The region between fences contains $> 99.999\%$ of observations.

2.5 Clustering Analysis

Machine or statistical learning is a field of statistics and computer science concerning the development of self-learning algorithms (models) for exploratory data analysis, pattern-finding, and prediction of new features [98].

Clustering is a type of unsupervised classification system used to find patterns in unstructured data. Unlike supervised learning classification, where mathematical models are trained

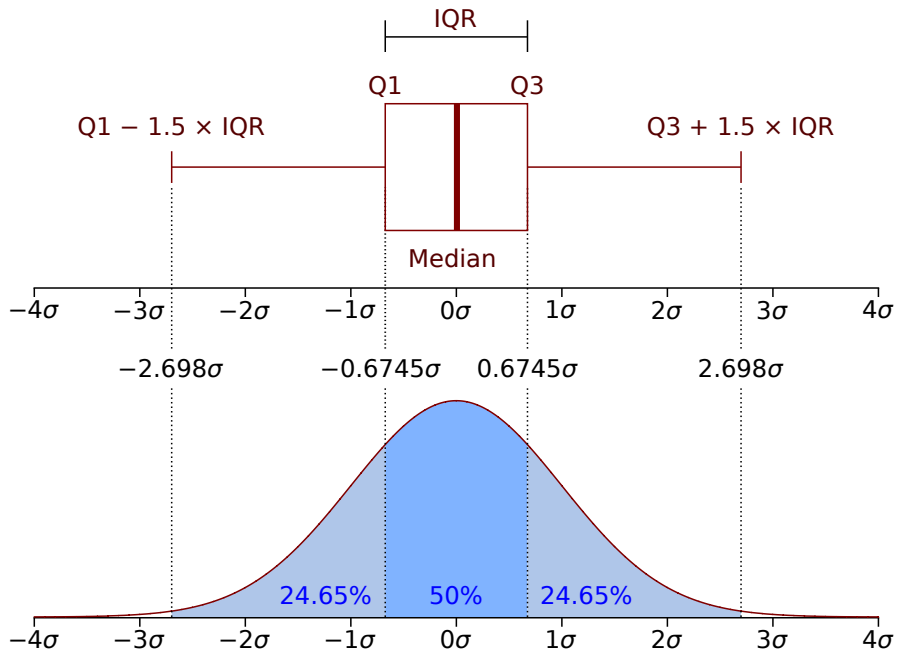


Figure 2.2: Box plot and probability density function of a normal $N(0, 1\sigma)$ population. The IQR is defined as IQR = $Q3 - Q1$, where Q1 is the first quartile and Q3 is the third quartile. Adapted from [97].

with labelled data examples (prior knowledge), clustering tries to find a set of finite and discrete data structures (clusters) in a finite and unlabelled dataset [99, 100]. Data partition is performed in a way to ensure that clusters are "homogeneous" and separated between each other. This requires using measures of similarity/distance, so that observations within each group are more similar/close to each other than to observations in other groups [101].

Clustering analysis can be divided into two major approaches:

- **Hierarchical or linkage-based**: Data points are hierarchically linked and grouped using a measure of distance, forming a dendrogram. Algorithms find nested clusters either in *agglomerative mode* (each point starts in its own cluster and pairs of similar clusters are successively merged as one moves up the hierarchy) or in *divisive mode* (starting with all data points in one cluster and recursively dividing each cluster into smaller ones as it goes down the hierarchy) [99, 102].

- **Partitional or cost minimization**: Data is assigned into clusters by optimizing a similarity criterion (cost function). Centroid-based algorithms use the distance (usually Euclidean) to the cluster centre (centroid) to assign points to a cluster. In density-based algorithms, clusters are areas with a higher density of points than the rest of the dataset. Density can be defined using the K-nearest-neighbours (KNN) method or kernel density estimation (KDE). Finally, in clustering using mixture models, clusters are formed of points likely to belong to the same probability distribution [99, 102].

2.5.1 Mean Shift Clustering

Mean Shift is a density-based clustering algorithm that aims to discover *blobs* in a smooth density of samples. It is non-parametric, so it does not require prior assumption about the shape of the clusters [103].

The mean shift procedure uses KDE estimate density. In this procedure, regions of high density are determined by a kernel function K. The weighted mean of the density in the window determined by K is,

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)} \tag{2.2}$$

where $N(x)$ is the neighbourhood of samples within a given distance around x and $K(x_i - x) \neq 0$. The difference $m(x) - x$ is the mean shift. In each iteration of the mean shift algorithm, data points move (shift) to the regional mean ($x \leftarrow m(x)$) until it converges. Regional means correspond to the cluster centres, which are updated in each iteration and filtered in the final step to remove near-duplicates [103, 104].

The Mean Shift algorithm has only one parameter, the *bandwidth*, which dictates the distance/size of the kernel function, i.e., the considered region to calculate the mean. For each clustering procedure, the bandwidth was estimated using the `estimate_bandwidth` function provided in Scikit-learn. This function needs the input of a quantile to apply the KNN method. The optimal quantile was chosen by evaluating the clustering performance for each gene (see Section 2.5.2).

2.5.2 Clustering performance evaluation

The Davies-Bouldin index is an internal evaluation scheme of clustering results. It uses similarity as a measure that compares the distance between clusters with the size of the clusters themselves.

Similarity is defined as a measure R_{ij} that trades off:

- s_i, the average distance between each point of the cluster and the centroid of that cluster (cluster diameter).
- d_{ij}, the distance between cluster centroids i and j.

R_{ij} can be constructed to be non-negative and symmetric:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{2.3}$$

Then the Davies-Bouldin index is defined as the average similarity between each cluster C_i for $i = 1, \ldots, k$ and its most similar one C_j.:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{2.4}$$

A lower Davies-Bouldin index relates to a model with better separation between the clusters, with zero being the lowest possible score [105].

2.6 Organotropism Pairs of Tissues

An organotropism pair of tissues is composed of a cancer tissue and a tissue where that cancer is likely to metastasize. To determine organotropism pairs, we used metastasis frequency data from two sources.

2.6.1 Metastasis Frequency Datasets

The Human Cancer Metastasis Database (HCMDB) integrates expression data of cancer metastasis from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) [106]. Data is grouped by experiments performed by the authors comparing the transcriptomes of primary and metastasis tumours. Each entry has information (when available), about primary and metastasis tumour sites. We used the first version of the database, available at HCMDB and obtained on 2021/10/12 at 16:58:28. The current version of HCMDB comprises 29 cancer types derived from more than 455 experiments.

The **Autopsy Study** dataset was assembled using the *Data table for analysed cases* from [107]. This study performs quantitative analyses of metastasis patterns using archival data from postmortem tissue analysis. The final dataset comprises review data from 3827 autopsies that included examination of all organ systems, performed between the years 1914 and 1943 on patients from 5 US medical centres. None of these patients received chemotherapy or radiation treatment. In total, 41 primary tumours and 30 metastatic sites were considered.

Correspondence between cancer/tissues names in the datasets and the tissue identifiers created for GTEx and Consensus datasets was performed after computing the organotropism pairs. This was done to avoid influencing statistics results by removing relevant frequency data. Primary and metastasis sites to which no match was found were removed from the final datasets. The correspondence between tissues and tissue IDs annotated with information justifying the procedure is available in Supplementary tables B.2, B.3, B.4, and B.5.

2.6.2 Hypergeometric Test-based Organotropism Pairs

The hypergeometric discrete random variable with parameters N, n, M counts the number of k objects with a specific characteristic (successes) in a sample of size N chosen without replacement from a population of M objects, where n is the number of objects with the specific characteristic in the total population. The probability mass function is defined as

$$p(k, N, n, M) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}} \tag{2.5}$$

for $k \in [\max(0, M - M + n), \min(n, N)]$.

The hypergeometric test uses the hypergeometric distribution to measure the statistical significance of a drawn sample. In a test for over-representation of successes in the sample, the hypergeometric p-value is calculated as the probability of randomly drawing k or more

successes from the population in N total draws [108]. The test was applied to the two metastasis frequency datasets to find overrepresented pairs of tissues.

Correction for multiple testing was performed using the Benjamini-Hochberg procedure for controlling the false discovery rate (FDR) [109]. This method corrects for type I errors in a set of $H_1 \ldots H_m$ null hypothesis tested with corresponding $P_1 \ldots P_m$ p-values. For a given α:

1. find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$
2. Reject the null hypothesis for all $H_{(i)}$ for $i = 1, \ldots, k$

Pairs of tissues with a p-value < 0.05 after FDR correction were considered organotropism pairs. Pairs were filtered with the tissue ID correspondence and organotropism pairs involving the same tissue as cancer and metastasis site were removed. The final output is a table, where each row represents a primary site and each column a metastasis site, filled with zeros and ones, where ones represent organotropism pairs.

2.6.3 Controlled Comparison Algorithm

A control pair is a pair of tissues that was not classified as an organotropism pair using the hypergeometric test, and which can be used to compare the distribution of PPI. The controlled comparison algorithm was developed to choose control pairs, so that a tissue appears the same number of times in control pairs as in organotropism pairs.

The algorithm takes as input the table of organotropism pairs generated in 2.6.2 and inserts control pairs with the value of -1. That way, a simple cost function (cf) was devised based on the sums of row and column elements of the input table. The goal of the algorithm is to minimize cf to optimize the distribution of control pairs.

For a bidimensional array $A_{m \times n}$ (table of organotropism pairs), cf is defined as

$$cf = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} \right)^2 + \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_{ij} \right)^2, \quad cf \geq 0 \tag{2.6}$$

where a_{ij} denotes the entry on the ith row and jth column of A. $cf = 0$ corresponds to the ideal proportion of pairs (# of organotropism pairs = # of control pairs).

The algorithm works by iterating the rows and assigning a probability of inserting a control pair in each row entry. It tries, when possible, to always insert the exact number of control pairs for the sum of elements in the row to be zero.

Let $w_i = (w_{i1}, w_{i2}, \ldots, w_{in})$ be the weights vector of row i and $S_j = \sum_{i=1}^{m} a_{ij}$ the sum of the elements of column j. The weight for entry a_{ij} is determined as

$$w_{ij} = \begin{cases} \frac{|S_j| + S_j}{2} & \text{if } a_{ij} = 0 \\ 0 & \text{if } a_{ij} = 1 \end{cases} \tag{2.7}$$

When $\sum_{j=1}^{n} w_{ij} = 0$ there are no ideal entries in row i to insert control pairs and the path taken will not lead to a global minimum. In this case, the algorithm recomputes the weights vector,

with the weight for entry a_{ij} determined as

$$w_{ij} = \begin{cases} 1 & \text{if } a_{ij} = 0 \\ 0 & \text{if } a_{ij} = 1 \end{cases} \tag{2.8}$$

The probability of assigning a control pair to entry a_{ij} is

$$p_{ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}} \tag{2.9}$$

where $\sum_{j=1}^{n} w_{ij}$ is the sum of the elements of weights vector i. The entries to insert control pairs in row i are randomly chosen, taking into account the probabilities vector $p_i = (p_{i1}, p_{i2}, \ldots, p_{in})$.

Control pairs with the same cancer and metastasis tissue are checked and removed when they exist. After assigning all control pairs, the cf is computed. The algorithm stops when $cf = 0$ or if it reaches the maximum number of iterations. On each iteration loop, rows are shuffled to increase the search space of possible combinations of control pairs.

2.7 Tissue-specific Intercellular PPI Networks

2.7.1 Intercellular Interactions Datasets

The OmniPath database combines data from more than 100 resources and contains protein-protein and gene regulatory interactions, enzyme–post-translational modification (PTM) relationships, protein complexes, protein annotations and intercellular communication [110]. The Python client for the OmniPath web service [111] was used to download the intercellular interaction network in table format. Each entry describes an interaction, including the source and target proteins, and data about the direction, effect, type of interaction, cellular location of the interacting proteins, sources, and references.

The complete network was processed using two filters. First, interacting proteins were required to be present in one of three locations: secreted (extracellular), plasma membrane (transmembrane), and plasma membrane (peripheral). Second, a literature curated network was created by removing the records with no references. The non-curated network has more interactions and proteins but has also more noise, i.e., interactions that might not happen in physiological conditions. For example, it includes interactions only caught in high throughput screenings. The two datasets were used in parallel to build the intercellular interactions networks.

The network includes interactions between a protein and a protein complex (composed of two or more interacting proteins) or between two protein complexes. To simplify the analysis, each complex was unfolded, i.e. separated into its components. Interactions were assigned between each protein in a complex and each protein in the interaction partner, being it a single protein or another unfolded complex.

2.7.2 Grouping Tissues by Tissue ID

The tissue IDs map different tissues/regions of the same organ, which means that the same tissue ID might have different patterns of gene expression. This is not an issue in the controlled

comparison, where tissues have the same proportion in the organotropism and control groups. However, it can influence results, for example, when correlating the frequency of metastases with the number of interactions in each tissue pair or when searching for metastasis-associated intercellular interactions.

To circumvent this potential problem, we defined new datasets of gene expression using only the tissue IDs and ditching GTEx and Consensus labels. Since PPI networks were constructed using both presence/absence calls and the TPM values (weighted networks), the different regions of the same organ were grouped using two criteria. For TPM values, the median of the TPM distribution of all regions was used. For the datasets with the presence/absence calls, the grouping was performed by majority voting: if the gene is present in the majority of the regions, the gene is deemed to be expressed in the tissue. In case of a tie, the gene is considered as absent.

2.7.3 Network Construction

Networks of intercellular interactions can be represented as a bipartite graph $G = (U, V, E)$. A bipartite graph can be divided into two disjoint and independent sets of nodes, U and V, such that each edge (E) connects a node in U to one in V [112]. In intercellular PPI networks, each set represents a cell, and each node a protein. The interactions in the whole graph, as downloaded from OmniPath, have implicit directions, with *source* and *target* genes. That includes bidirectional interactions, which can encompass physical interactions without an apparent direction or with both directions, and those detected with methods that are not able to assign a direction (Figure 2.3a). So, to take into account this possibly ambiguous information, two different types of graphs were built: a simple undirected graph, where bidirectional interactions are treated as a single interaction (Figure 2.3b); and a simple directed graph which can be integrated in signalling pathways, where bidirectional interactions are removed (Figure 2.3c). In both graphs, the interaction $A_U \rightarrow B_V$ is considered distinct from $A_V \rightarrow B_U$, since the two interacting nodes are present in both sets. In other words, despite representing the same protein, A_U and A_V are expressed in different cells. Thus, even when not considering the direction of the interaction, the contribution from this single interaction to the communication between cells is bigger than, for example, interaction $C_U \rightarrow D_V$.

Tissues are comprised of different cells types, each with distinct patterns of gene expression. Since bulk RNA-seq data is being used to build tissue-specific networks, it is not possible to distinguish between different cell types. And so, instead of U and V representing two cells, it is more accurate to say that they represent generic cells in tissues U and V, respectively. In other words, they represent approximately all possible interactions a cell originating in tissue U can establish with a cell in tissue V.

To build each intercellular PPI network, two graphs are created: one that comprises interactions that flow from a cell in tissue U to a cell in tissue V and another with the opposite direction. The two graphs are merged, and the resulting graph is simplified—multiple edges and loops are removed. For directed networks, interactions with multiple edges (bidirectional interactions) are completely removed, as previously mentioned.

Two different approaches were used to build tissue-specific intercellular PPI networks. The

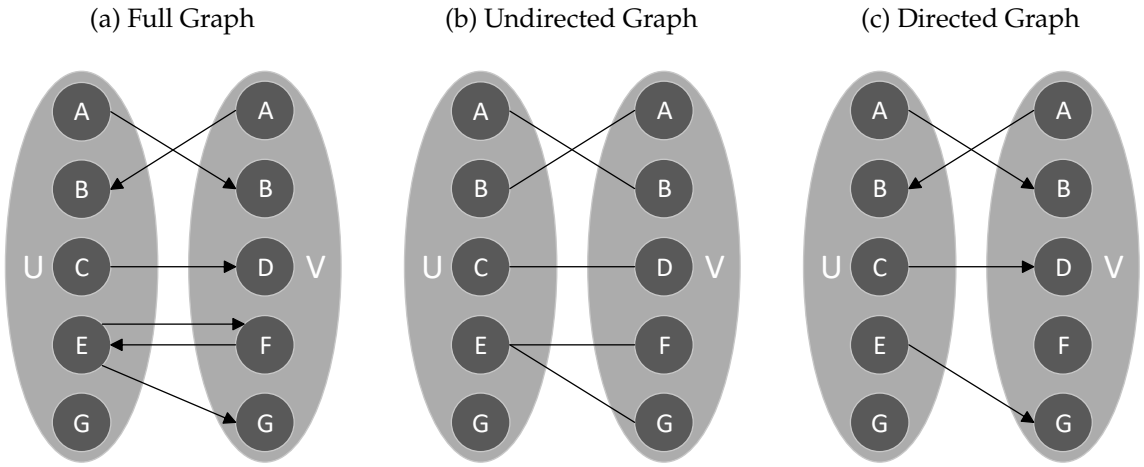


Figure 2.3: Abstract representation of the different graph types used to represent intercellular PPI networks. Circles represent nodes (proteins) present in two distinct sets (cells) U and V, with identifiers $A \ldots G$. Arrows and lines represent edges (interactions).

first approach uses the presence/absence calls previously computed to distinguish between tissues. This means that each tissue has a distinct number of genes and so, networks will differ in the number of interactions. The second approach creates weighted networks using directly the expression value (TPM) of each gene. The weight of a gene in a certain tissue is defined as the expression value normalized by the maximum expression value in all tissues. The weight of an interaction is the product of the two gene weights. In these networks, all possible interactions are present, and they differ in the sum of all interaction weights.

2.7.4 Jaccard Index

The Jaccard Similarity Index or Coefficient is a measure of similarity and diversity between two finite sets. It is useful for binary data such as presence/absence sets. The Jaccard Index is defined as the ratio of the size of the intersection with the size of the union of the two sets [113]:

$$J(U,V) = \frac{|U \cap V|}{|U \cup V|} = \frac{|U \cap V|}{|U| + |V| - |U \cap V|} \tag{2.10}$$

The Jaccard Index was applied to the intercellular PPI networks built using gene calls. For this case, the two sets U and V represent tissues. $|U \cap V|$ represents the number of interactions in the network and $|U \cup V|$ the total number of intercellular interactions that can be established by cells from tissues U and V. This measure gives an idea of the similarity/compatibility between the two tissues, since it takes into account the potential of each tissue to form intercellular interactions.

2.7.5 Z-score

The z-score or standard score measures the distance between an observation (x) and the mean (μ) in units of standard deviation (σ):

$$\text{z-score} = \frac{x - \mu}{\sigma} \tag{2.11}$$

Observations above the mean have positive z-scores, while those below the mean have negative scores. z-scores were generated for each intercellular PPI network using the mean and standard deviation from a distribution of random network values.

For networks built using gene presence/absence calls, random networks were created by sampling intercellular genes from 3 different sets: source-only genes, target-only genes and genes that can appear as both source and target in an intercellular interaction. The original number of genes present in the tissue for each set was kept, and each gene has a probability of being picked proportional to its expression value. Keeping the number of genes preserves the tissue-specific signal of gene expression. z-scores were calculated using the number of interactions in the network.

In the case of weighted networks, the interaction weights were randomized by shuffling gene weights after constructing the network. That means the expression values used in the randomized networks are the ones specific for the tissue, only they are assigned to different genes. This preserves the tissue-specific signal of gene expression. z-scores were calculated using the sum of interaction weights.

2.8 Selection of Intercellular PPI Interactions

2.8.1 Statistical Analysis

2.8.1.1 Fisher's Exact Test and Odds Ratio

In an experiment with two random variables x and y with a binary response, the frequency distribution of the variables can be represented by a 2×2 contingency table (Table 2.2). The Fisher's exact test is a commonly used test to analyse contingency tables. In this test, the population distribution from which the observed table is taken is conditioned on the margins, i.e. row and column totals are fixed. The null hypothesis is that the contingency table is from the hypergeometric distribution with parameters (M, n, N), making it identical to a hypergeometric test (see Section 2.6.2) [114].

Table 2.2: Example of a contingency table for two random variables x and y with a binary response. The marginal totals are defined with the parameters of a hypergeometric distribution $p(M, n, N)$.

	y	$\bar{y}$	total
x	a	b	n
$\bar{x}$	c	d	$M - n$
total	N	$M - N$	M

The odds ratio (OR) is a measure of association between two binary variables. It expresses the odds of an event occurring in one group relative to the odds of it occurring in another group. The OR for variables x and y in the contingency table 2.2 is defined as

$$OR = \frac{a/b}{c/d} \tag{2.12}$$

2.8.1.2 Mann-Whitney U Test

The Mann-Whitney U test (MWU) test is a non-parametric version of the t-test for independent samples. Although it compares the distributions of two random variables, for identically shaped distributions it can be seen as a test for comparing medians. The two-tailed formulation for the null hypothesis is that the distributions of both groups are the same [115, 116]. The test statistic (U statistic) for two samples x and y is defined as:

$$U = n_x n_y + \frac{n_x(n_x + 1)}{2} - W \tag{2.13}$$

where n_x and n_y are the number of observations from samples x and y, respectively, and W is the rank sum for sample x. The distribution of U values can be determined exactly by computing its value for all possible permutations of the observations, numerically approximated by using a large sample of the possible permutations. When the number of observations is large, the U statistics converges to the normal distribution.

2.8.1.3 Spearman Rank Correlation Coefficient

The Spearman's rank correlation coefficient (SRCC) is a non-parametric measure used to determine the strength and direction of the relationship between two variables, where the relationship is monotonic, i.e., increases or decreases consistently, but not necessarily at a constant rate [117]. It is similar to the Pearson's Correlation for ranked data. The SRCC for two ranked variables x and y is

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \tag{2.14}$$

where S_{xy} is the covariance of the rank variables and S_{xx} and S_{yy} are the variances of the rank variables x and y, respectively.

The coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. The SRCC can be used to test the hypothesis that there is no association between the two samples. The p-value indicates the probability that r_s assumes a large absolute value solely due to chance.

2.8.1.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (KS) test is a non-parametric test of the equality of continuous, one-dimensional probability distributions. It can be used as a goodness-of-fit test to determine if a sample was drawn from a reference probability distribution (one-sample KS), or used to determine whether two samples were drawn from the same distribution (two-sample KS) [118]. Consider two independent random samples: $X_1, X_2, \ldots, X_n$, and $Y_1, Y_2, \ldots, Y_m$, with sizes n and m, respectively. The KS statistic ($T_{n,m}$) for the two-sample test is the largest distance between the

empirical distribution functions of the first and second samples, $H_1(x)$ and $H_2(x)$ respectively, across all values of x:

$$T_{n,m} = \sup_x |H_{1,n}(x) - H_{2,m}(x)| \tag{2.15}$$

where sup is the supremum function.

2.8.1.5 Shannon Entropy

The Shannon entropy is a measure of the uncertainty or randomness in a set of data. It quantifies the amount of information required to identify an element in the set. In biology, the Shannon entropy is also called the Shannon diversity index and is used to quantify species diversity. The information content (or surprise) of an event x_i is a function which increases as the probability $p(x_i)$ of an event decreases [119]. This relationship is described by the function

$$h(x_i) = -log_2 p(x_i) \tag{2.16}$$

The entropy of the ensemble X is defined to be the Shannon information content of

$$H(X) = \sum_{i=1}^{n} p(x_i) h(x_i) \tag{2.17}$$

where n is the number of events in X.

The Shannon entropy was applied to assess the diversity of tissues in intercellular interactions. Two populations of tissues were defined for each interaction: a cancer tissue population and a metastasis tissue population. $p(x_i)$ is the probability of choosing the ith tissue from population X. Thus, $H(X)$ quantifies the uncertainty in predicting the identity of a randomly chosen tissue. The larger the uncertainty, the more diverse is the population.

2.8.2 Interaction Selection Workflow

Interactions that might be associated with metastasis formation and organotropism were found by establishing a relationship between the intercellular networks where the interaction is present (or the weight of the interaction in the network) and the frequency of metastasis (or label –organotropism/control) of that pair of tissues. Distinct statistical methods were used, depending on the combination of network type/metastasis data:

1. **Networks built using gene presence/absence calls with organotropism pairs of tissues**: Both datasets consist of variables with a binary response. The Fisher's Exact test was used to find out if an interaction appears more or less frequently in organotropism than in control pairs. The log(OR) was computed to find the sign of the association between the interaction and metastasis formation.

2. **Networks built using gene presence/absence calls with metastasis frequency**: The MWU test was used to test if the group of networks where an interaction is present had a higher or lower distribution of metastasis frequency than the group where the interaction is absent.

The difference between the U statistic of the two groups was used to find the sign of the association between the interaction and metastasis formation.

3. **Weighted networks with organotropism pairs of tissues**: The MWU was used to find if the pairs of organotropism had a larger or smaller distribution of the sums of interaction weights than the control pairs group. The difference between the U statistic of the two groups was used to find the sign of the association between the interaction and metastasis formation.

4. **Weighted networks with metastasis frequency**: SRCC was used to determine if an interaction has statistically significant positive or negative correlation with metastasis formation. The sign of the interaction corresponds to the sign of the correlation.

Interactions were searched in both GTEx and Consensus datasets of gene expression (with grouped tissues) and in HCMDB and Autopsy Study datasets of metastasis frequency. Only the undirected intercellular interactions graph was used. After computing the tests for all conditions, results were subjected to FDR correction (see Section 2.6.2) and statistically significant interactions were aggregated in one dataset. An interaction sign was computed for each condition. A positive sign means that the interaction contributes to metastasis formation (is a driver of metastasis) and negative signals that the interaction hinders metastasis formation.

Next, the tissue specificity of each interaction was evaluated. Ideally, a metastasis-associated intercellular interaction can be established between cells from many distinct tissues. Interactions specific to one or few tissues can also be associated with metastasis formation, but their relevance might be masked by other factors specific to that or those few tissues that also drive organotropism, such as organ accessibility, vascularization and anatomic location. Two methods were used to assess the tissue specificity/diversity of an interaction:

- **Shannon entropy**: A low entropy indicates that the interaction is established between cells from few distinct tissues. Two distinct Shannon entropies were calculated for each interaction: one to assess cancer tissue diversity and the other for metastasis tissues. Only interactions where both entropies are > 0 were selected.

- **Interaction tau**: the tau (τ) of an interaction between genes A and B is defined as $\max(\tau(A), \tau(B))$. Only interactions with a $\tau < 0.9$ were selected.

Finally, two further filters were imposed: interactions with a low correlation ($-0.25 \leq r_s \leq 0.25$) were dropped and interactions were required to be statistically significant in both datasets of gene expression (GTEx and Consensus).

2.9 Intracellular PPI Network

2.9.1 Intracellular Interactions Datasets

The Python client for the OmniPath web service was used to download retrieve intracellular interactions from two datasets: signalling interactions from Omnipath and transcription factor

(TF)—target gene interactions from the DoRothEA database. DoRothEA curates and assigns a confidence level to each regulon (collection of a TF and its targets) based on the available supporting evidence, ranging from A (highest confidence) to E (lowest confidence). Evidence includes literature-curated resources, chromatin immunoprecipitation sequencing (ChIP-seq) experimental data, computational prediction of TF binding motifs, and inference from gene expression data (GTEx) [120]. We choose confidence levels A to D to build the intracellular network, leaving out interactions that are only supported by computational predictions (level E). The interaction data was retrieved in table format, where each entry describes an interaction, including the source and target proteins, and data about the direction, effect, type of interaction, sources, and references.

The network includes interactions involving protein complexes composed of two or more interacting proteins. To simplify the analysis, each complex was unfolded as described in the intercellular PPI network methods (see Methods 2.7.1). Finally, the network was filtered using the GTEx and Consensus datasets, keeping only genes present on, at least, one of the datasets.

2.9.2 Random Walks with Restart (RWR)

A random walk is a process that describes a path consisting of a succession of random steps on some mathematical space. When applied to a graph, a random walk describes the path taken by a "walker" that moves from one node to a neighbouring node with a probability that is proportional to the weight of the connecting edge (in unweighted graphs, all edges have the same weight) [74]. RWR is a random-walk-based propagation process used to identify nodes that are closest to some node of interest (starting node). In the RWR implementation, the walker has a probability of returning to the starting node at each step of the walk [70].

Let $G = (N, E)$ be a directed graph representing a PPI network, where N is the set of nodes (proteins), and E is the set of edges (interactions). G can be represented as an adjacency matrix $A = a_{i,j}$, where $a_{i,j} = 1$ when there is an edge e_{ij} between node n_i and n_j, and $a_{i,j} = 0$ otherwise. A RWR over G is a process that starts from node n_i, and at each time step t moves to one randomly selected neighbour of the current node n_j, with a probability α of restarting at every time step:

$$F_{t+1} = (1 - \alpha)WF_t + \alpha F_0 \tag{2.18}$$

where $W = AD^{-1}$ and D is the diagonal degree matrix of A. F_t is a vector in which the ith element holds the probability of being at node i at time step t and F_0 is the initial probabilities vector. The steady state probability vector F_∞ gives a measure of proximity of each node to the starting node n_i [70, 121].

The RWR algorithm was applied to the intracellular network to simulate the propagation of a signal inside a cell. Two distinct signals were considered: a signal that reflects incoming communication from neighbouring cells, starting in intercellular receptors (target proteins); and a signal that simulates outgoing communication, ending in emitters/ligands (source proteins). A RWR was performed for each intercellular protein (query protein) with a restart probability of 0.25. To simulate the signal that ends in a source protein, the direction of the edges in the

intracellular graph was reversed. In practice, this means that the outgoing edges of a node become its incoming edges. That way, the random walk can start in the source protein and trace a path backwards that ends in the protein that potentially initiated the outgoing signal. The result is a vector for each intercellular protein, containing the probability of visiting each intracellular protein.

2.9.3 Permutation Test

A permutation test is a non-parametric test in which the null distribution of a test statistic is estimated by randomly permuting the group labels of the observations. Permutation tests can be exact tests if samples are small and allow the computation of every possible permutation, but for large sample sizes the null distribution is estimated by Monte Carlo sampling (pseudo-random sampling). Permutation tests assume that the observations are independent and identically distributed under the null hypothesis [122].

Permutation tests were performed to find the proteins in the intracellular network closer to the intercellular proteins that establish metastasis-associated interactions. For each intracellular protein, the probabilities vector of the RWR was divided into two main groups: RWR probabilities of walks starting in proteins that appear in interactions with a positive association with metastasis A, and RWR probabilities of walks starting in proteins absent from metastasis-associated interactions B. Let X_A and X_B be two random variables for each individual from the two groups A and B. The null hypothesis is that all observations are sampled from the same underlying distribution and that they have been assigned to one of the samples at random. First, the observed value of the test statistic is computed for each intracellular protein and for each subgroup of metastasis-associated proteins:

$$T_{obs} = \bar{x}_A - \bar{x}_B \tag{2.19}$$

where $\bar{x}_A$, $\bar{x}_B$ are the sample means of A and B, respectively. A higher value of T_{obs} means that this intracellular node is more likely to be visited in random walks starting in intercellular proteins with a positive association with metastasis than walks starting in other intercellular proteins. Next, the observations of groups A and B are pooled, and the difference in sample means is calculated and recorded for every permutation of the group labels A and B, keeping constant the group sample size. The one-sided p-value of the test is calculated as the proportion of sampled permutations where the difference in means is greater than T_{obs}.

The intracellular network includes intercellular gene, some of which might have significant RWR probabilities. Thus, all intercellular genes were removed before submitting the results to FDR correction (see Section 2.6.2). Intracellular proteins were considered as close to intercellular interactions associated with metastasis if they are statistically significant (FDR < 0.05) and if their T_{obs} is an outlier in the distribution of T_{obs} values using the Tuckey's Fences method (see Methods 2.4). This filter is imposed to ensure that only proteins with larger RWR probabilities are chosen as being associated with metastasis development.

3

Results

3.1 Calls of Presence/Absence of Gene Expression

Proteins are usually considered the main effectors of phenotype in cells, responsible for performing essential catalytic, structural, signalling functions, among others. Protein levels in a cell are regulated not only by mRNA abundance (transcription rate) but also by mRNA transport and stability, translation rates and protein stability (turnover rates). Nevertheless, mRNA levels are still a good approximation to characterize tissue-specific signatures in gene expression, and can account for more than 50% of protein level variability [123, 124]. As the first step in constructing our tissue-specific PPI networks, we used tissue-specific RNA-seq data to determine where each gene is expressed. First, we determined the tissue-specificity of each gene, and defined specificity thresholds to group genes. Second, we devised two strategies to perform presence/absence calls for each gene, taking into account the different levels of tissue-specificity.

3.1.1 The Tissue-Specificity of Genes

The pattern of gene expression varies significantly across human tissues. It is common to classify genes in two major categories, based on their patterns of expression: housekeeping and tissue-specific genes. Housekeeping genes are required for maintenance of basal cellular functions, and thus are present in all cells, regardless of tissue and cell type [125]. Tissue-specific genes are only expressed in one tissue and are usually associated with specific cell and tissue phenotypes and functions [92]. However, many genes may have a midrange expression profile that does not fit any of the above-mentioned groups. This third group contains genes highly expressed in a subset of tissues, with a much lower level in the remaining tissues [91]. Furthermore, as high-throughput methods with improved sensitivity to measure gene expression emerge, genes thought to be specific to one tissue might be found to be present in other tissues. In fact, most genes have some background expression in all tissues [126].

We used the Tau Tissue Specificity Index to characterize the patterns of gene expression across GTEx and Consensus dataset tissues (see Section 2.3). The tau index (τ) was found to perform better than other methods, such as the Gini coefficient and simple expression thresholds [127]. τ values vary between 0 and 1. Genes with $\tau = 0$ are expressed in all tissues, with approximately the same expression level. Genes with $\tau = 1$ are only expressed in one tissue. We evaluated the

τ distribution of lists of housekeeping and tissue-specific genes [92] (see Methods 2.3) and set three tissue-specificity groups:

- $\tau \leq 0.4$: includes most housekeeping genes (Figure 3.1a). These genes are expected to be present in all tissues.
- $\tau \geq 0.9$: set to include the main peak of tissue-specific genes (Figure 3.1b). These genes are expected to be present in only one tissue.
- $0.4 < \tau < 0.9$: intermediate τ levels. Genes in this group feature a broad range of presence/absence patterns.

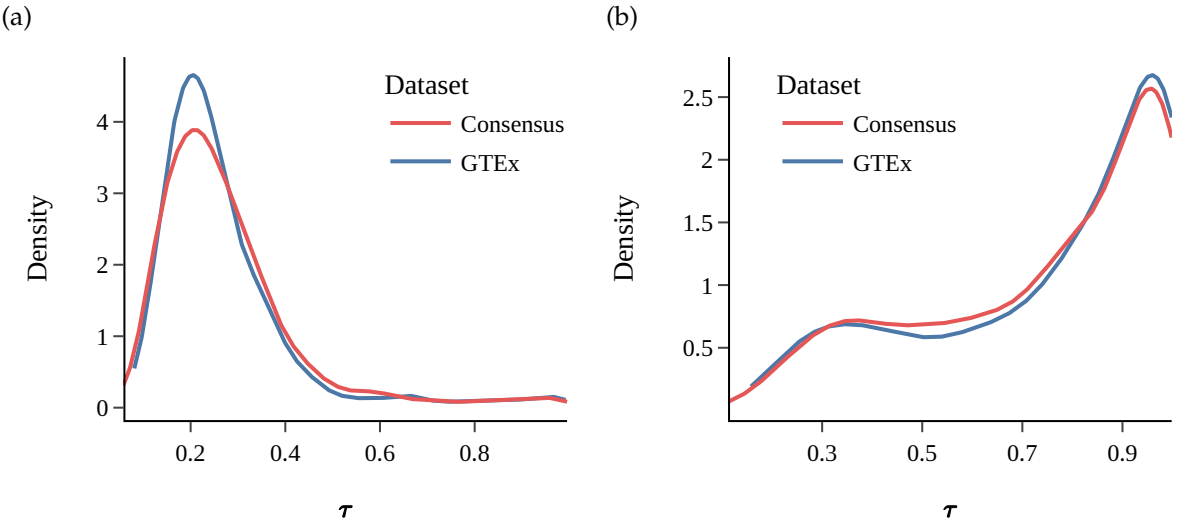


Figure 3.1: KDE plot of the distribution of tau values in (a) housekeeping and (b) tissue-specific genes.

The τ index is a good way to characterize tissue-specificity but does not allow us to assign presence/absence calls, unless we consider only genes very close to the endpoints of the τ interval. That means we expect that not all genes in the housekeeping group are expressed in all tissues, and that some genes in the tissue-specific group are expressed in more than one tissue. Thus, we applied two distinct methods to perform the present/absent calls: an outlier detection method on the housekeeping and tissue-specific groups, and clustering analysis for the intermediate τ group. We also set two expression cut-offs as a fail-safe for these methods. Expression values above 5 TPM ($\log_2$(TPM+1) $\approx$ 2.6) always mean that the gene is expressed, and values below 1 TPM ($\log_2$(TPM+1) = 1) signal that a gene is not present in a tissue. These thresholds are used to ensure that large values of expression are never discarded, and that values close to background noise are not considered.

3.1.2 Presence/Absence Calls using Outlier Detection

To perform calls of gene expression, we applied a genewise method for outlier detection based on the Tukey's Fences (see Section 2.4). A call is decided as follows:

- $\tau \leq 0.4$: only lower outliers are considered. If the expression in a tissue is lower than the lower fence, the gene is considered as not expressed in that tissue.

- $\tau \geq 0.9$: only upper outliers are considered. If the expression in a tissue is higher than the upper fence, the gene is considered as present in that tissue.

The results of the outlier expression show that, as τ increases, the number of tissues where a gene is present drops (Figure 3.2). This is consistent with the expected behaviour of τ, which is directly proportional to the tissue-specificity of the gene, and suggests that the outlier detection is performing as expected. The correlation is stronger on the group $\tau \geq 0.9$ with a SRCC below -0.65 and significantly different from 0 (p-value ≈ 0) for both GTEx and Consensus datasets (see Methods 2.8.1.3). The correlation is smaller for $\tau \leq 0.4$, -0.33 for GTEx and -0.37 for Consensus, but still significant (p-value ≈ 0).

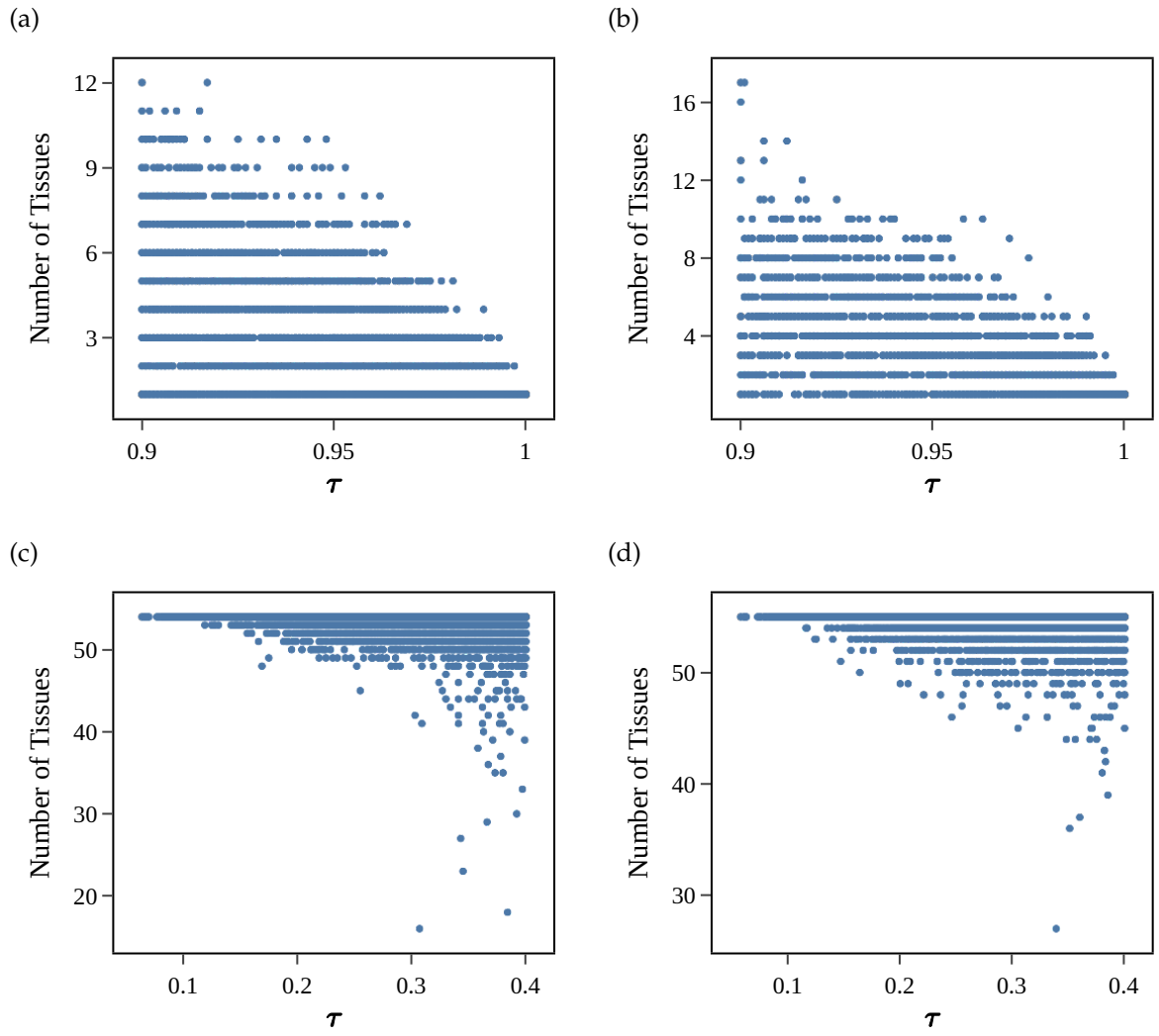
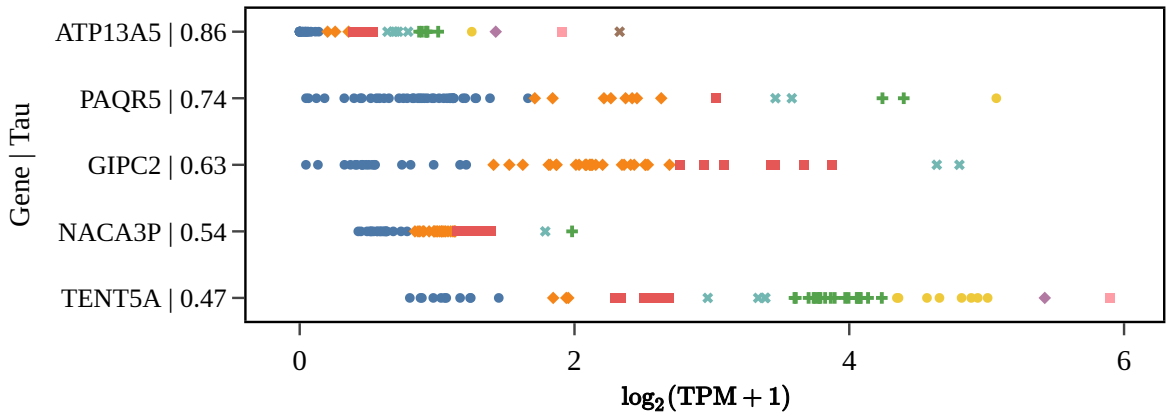


Figure 3.2: Presence/absence calls using the outlier detection method. Relationship between the number of tissues where each gene is expressed and its τ. Upper row: results for $\tau \geq 0.9$ in (a) GTEx and (b) Consensus. Lower row: results for $\tau \leq 0.4$ in (a) GTEx and (b) Consensus.

3.1.3 Presence/Absence Calls with Clustering Analysis

The $0.4 < \tau < 0.9$ set includes genes with widely disparate distributions of gene expression. At these values of τ, we expect to find genes expressed in a subgroup of tissues that is smaller the larger is τ. We used clustering analysis (see Methods 2.5.1) to perform presence/absence calls for each genes in the group. The Mean Shift algorithm finds clusters of points by defining high density regions in the expression dataset (Figure 3.3). We determine that a gene is not present in a tissue if the corresponding expression value belongs to the cluster with the lowest value (represented by the blue circles in Figure 3.3).

(a)



(b)

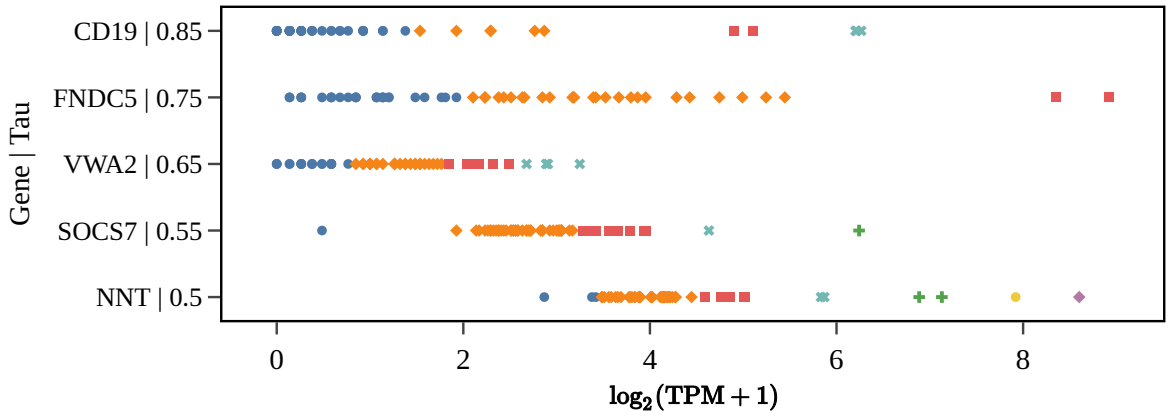


Figure 3.3: Presence/absence calls using the clustering analysis method. Example for a selected group of genes in (a) GTEx and (b) Consensus datasets. Each colour/marker represents a cluster.

Results from the clustering analysis show an expected inverse relationship between the τ and the number of tissues where a gene is expressed (Figure 3.4). The correlation is stronger than what was observed in the gene calls using the outlier detection, with a SRCC of -0.78 for GTEx and -0.81 for Consensus (p-value ≈ 0).

The way we split the distribution of τ values does not ensure we are using the ideal separation to perform the calls, since we are applying different methods to each group. To test

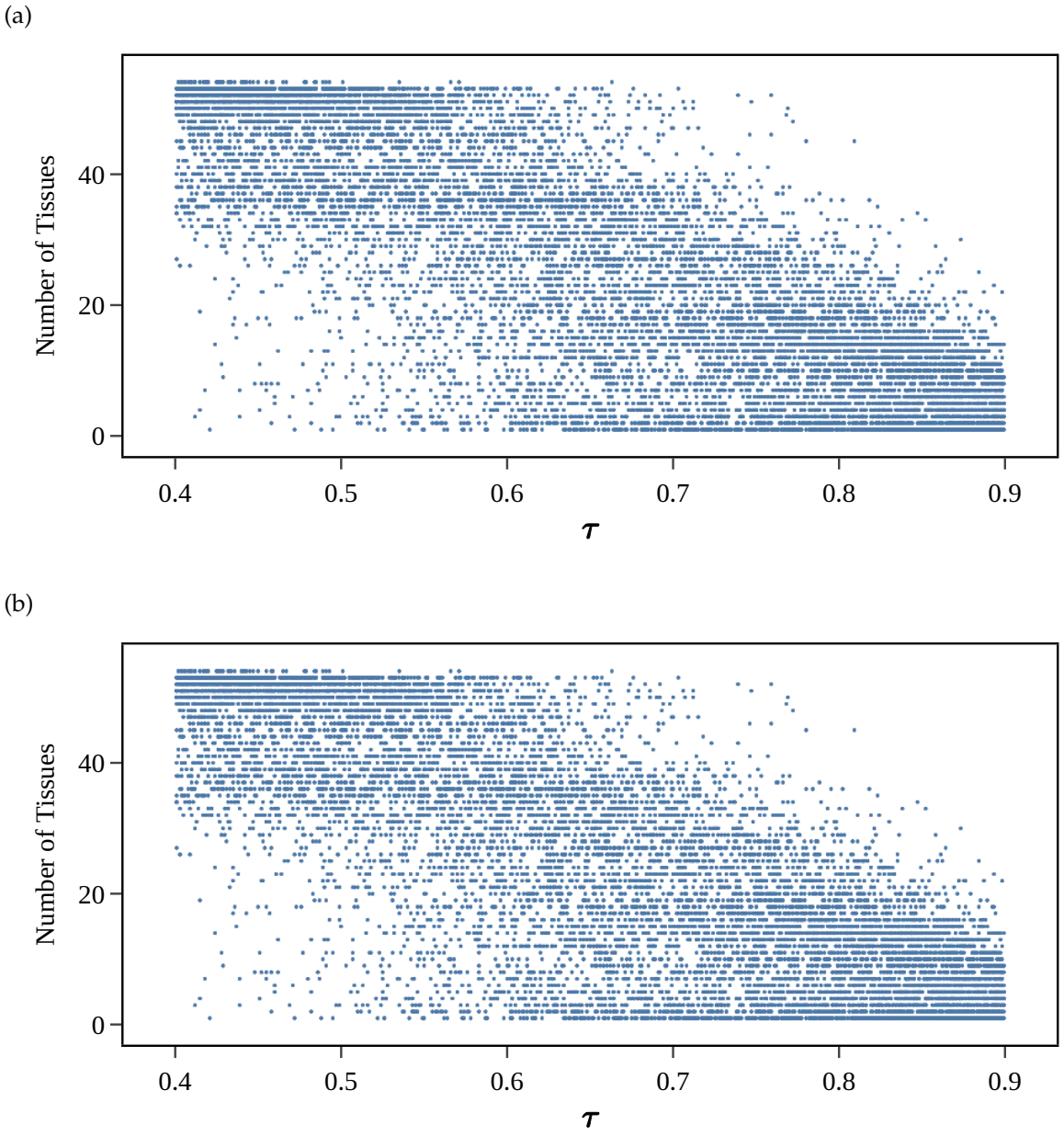


Figure 3.4: Presence/absence calls using clustering analysis method. Relationship between the number of tissues where each gene is expressed and its τ in (a) GTEx and (b) Consensus datasets.

this assumption, we varied each τ threshold by ± 0.5 and reapplied each method. For genes previously called by outlier detection, using clustering does not significantly change the number of tissues per gene. The difference is larger when we apply the outlier detection method for genes previously called by clustering, with most genes being called present in less tissues. This seems to imply that the outlier detection method is more sensitive and best suited for genes with more extreme values of τ.

3.2 Organotropism Pairs of Tissues

The primary objective of this work is to identify potential determinants of organotropic metastases by using intercellular PPI networks to characterize the communication between a metastasis seed cell and cells in the metastatic site. This includes comparing intercellular PPI networks in order to identify relevant differences in cell-cell communication. Therefore, we defined the concept of organotropism pairs of tissues. An organotropism pair of tissues is composed of a primary tumour tissue and a tissue where that tumour is likely to develop metastases. We applied the Hypergeometric Test to find pairs of tissues with enriched frequency in two metastasis datasets: HCMDB and Autopsy Study (see Methods 2.6.2). The differences seen in the number of organotropism pairs between conditions (Figure 3.5) are mainly due to the size and available tissues in the datasets. The Autopsy Study has more primary tumour sites, and the Consensus dataset has a greater variety of tissues. Thus, the combination of the two data sources originates more organotropism pairs.

Distinct tissues have distinct patterns of gene expression across cell types, which directly influences the number of membrane receptors, adhesion and secreted proteins each cell expresses. As a consequence, we reason that cancer cells from distinct tissues have different interaction-establishing potentials, which can impact their ability to colonize other tissues. Also, other factors such as blood flow patterns and organ accessibility influence metastatic patterns. These factors have to be taken into account when comparing the intercellular PPI networks between organotropism pairs and the remaining pairs. One way of correcting for this issue, is for us to have each tissue appearing exactly the same amount of times in the organotropism pairs and in the remaining pairs, which we will call our control pairs.

We developed an algorithm that adds control pairs by taking into account the proportion $\frac{\text{\# of organotropism pairs}}{\text{\# of control pairs}}$ (see Methods 2.6.3). The algorithm is heuristic, so it may not find the global minimum, and the ideal proportion might not be a possible solution for every input table of organotropism pairs. Also, there may exist more than one global minima, with different combinations of control pairs and, consequently, different distributions of intercellular interactions. In order to avoid a possible bias in the comparison, we computed 1000 sets of control pairs for each condition (dataset of metastasis frequency + dataset of gene expression).

Searching for pairs of tissues with enriched frequency is not a common approach used to describe and study the organotropism phenomena. Metastatic patterns are usually described in the clinic and in epidemiological studies. In these settings, organotropic metastasis sites usually refer to the most commonly observed secondary tumour sites. To find out if the organotropism pairs we obtained using the hypergeometric test are in accordance with what is referenced in clinical practice, we defined two new sets of organotropism pairs. First, we applied the outlier detection method (see Methods 2.4) to the two datasets of metastasis frequency to find the most frequent metastasis sites for each primary tumour site. Second, we performed a literature search and curation of studies that describe metastasis incidences (Supplementary Table B.6).

When comparing the hypergeometric test-based organotropism pairs with the results of the literature curation, we see a significantly better agreement with the HCMDB dataset (Table 3.1). As previously mentioned, the Autopsy Study dataset is larger and has more tissues, with many

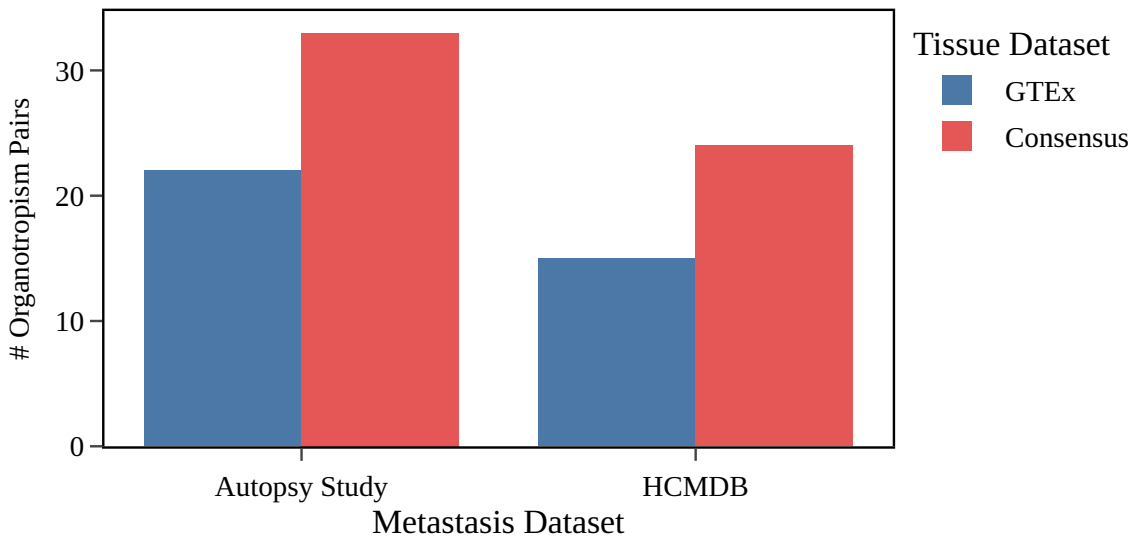


Figure 3.5: Hypergeometric test-based organotropism pairs. Number of organotropism pairs per dataset combination.

of them being uncommon primary and secondary sites. For example, in this dataset, some cancers have a high incidence of metastases to the heart, which is an extremely rare site of metastasis. The Autopsy Study uses data gathered until the year 1943, a period before the development of modern cancer treatments, such as radiation therapy and chemotherapy. This means that, at the time, cancer spread was unimpeded, so it could potentially form metastases in what are considered uncommon locations.

The hypergeometric test may be determining pairs as significant which are enriched but have a very low frequency. This suspicion is confirmed when we do the intersection between outlier detection-based and hypergeometric test-based pairs. We see that close to 90% of organotropism pairs defined using the hypergeometric test in the HCMDB dataset are also pairs with a large metastasis frequency (Supplementary Table B.7). On the contrary, this percentage does not reach 50% with the Autopsy Study dataset.

Table 3.1: Intersection between organotropism pairs determined using the hypergeometric test and literature curation. The **Ratio** represents the percentage of hypergeometric test-based organotropism pairs which are also found in the literature.

Metastasis Dataset	Tissue Dataset	Jaccard	Ratio (%)
HCMDB	GTEx	0.23	73.3
	Consensus	0.20	75.0
Autopsy Study	GTEx	0.14	36.4
	Consensus	0.14	42.4

3.3 Analysis of Intercellular PPI Networks

Cell-cell communication is of paramount importance for the development of metastasis. When arresting at a novel and adverse environment, the chance of survival of CTC is dictated by how they can successfully respond to, hijack, send or ignore signals from neighbouring cells, avoid immune response, remaining undetected and staying dormant until conditions are ideal to proliferate. Our hypothesis is that, in tissues where a cancer is more likely to metastasize, cancer cells establish a distinct pattern of intercellular communication. For detecting this difference, we used intercellular PPI networks to measure the level of communication between cells from different tissues of origin and establish a link with the emergence of organ-specific metastasis.

We built tissue-specific intercellular PPI networks for each pair of tissues in GTEx and Consensus datasets, using both gene presence/absence calls and gene expression value (see Methods 2.7). In the first case, where networks are unweighted, an intercellular interaction is established if both genes involved are present in the pair of tissues. In the second case, where network are weighted, every interaction has a weight computed from the expression values of the genes involved. Since we are using bulk RNA-seq to define patterns of expression, the networks are a representation of the possible cell-cell interactions that might be established between cells from the two tissues. The level of intercellular communication was assessed using the number of interactions and the jaccard index in unweighted networks, and the sum of interaction weights in weighted networks.

3.3.1 Cancer-Wise Analysis of Metastatic Patterns

We started by using intercellular PPI networks to find a connection between the level of cell-cell communication and the metastatic patterns of each cancer. Since organotropism is directly related to the frequency of metastasis, our hypothesis is that cells from pairs of tissues with higher frequencies of metastasis establish a significantly different number of interactions (or interaction weights) than pairs unlikely to form metastasis. Intercellular interactions may both hinder or improve the chances of survival and proliferation of a metastasizing cell. Therefore, it is not possible to assume if high frequency pairs will have a higher or lower level of intercellular communication.

Each tissue has a distinct pattern of gene expression, which makes cells from these tissues more likely to establish a larger number of interactions or interactions with more weight. To correct for this issue, for each network statistic, we computed a z-score from a distribution of random network values (see Methods 2.7.5). When plotting the results of the undirected intercellular networks built using gene calls, there seems to be no clear relationship between the z-score and the frequency of metastasis for all cancers in each studied condition (Figure 3.6). This stands for networks constructed using only curated interactions (Supplementary Figure A.2) and for directed networks, where there are no visible differences between the two possible directions (Supplementary Figure A.3 and A.4). The same absence of pattern exists for weighted networks (Supplementary Figures A.5, A.6, and A.7).

Computing the SRCC confirms the apparent lack of relationship between the frequency

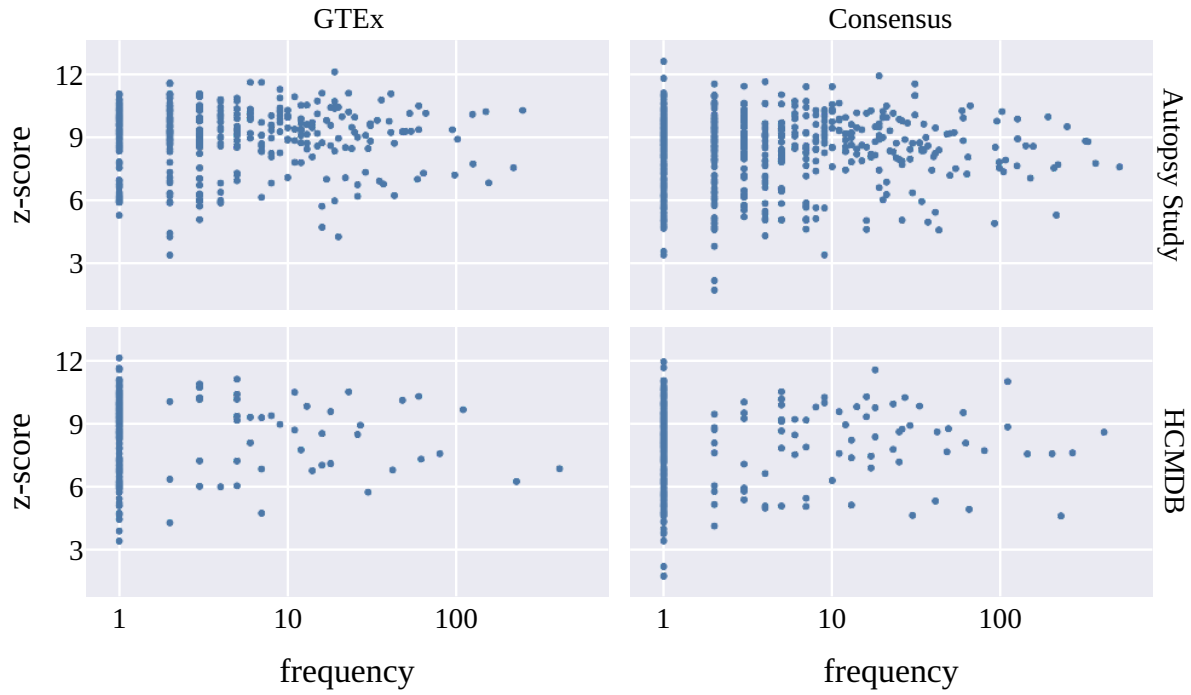


Figure 3.6: Cancer-wise analysis of metastatic patterns for **undirected networks built with gene expression calls**. Relationship between the z-score and the frequency of metastasis in log scale. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair).

of metastasis and the intercellular interaction z-score (Supplementary Tables B.9-B.12). Some conditions with distinct datasets show a positive correlation, whereas others show a negative correlation. But the correlations are weak in all conditions and, most of the time, statistically non-significant (p-value > 0.05). When computing the correlation coefficient for each cancer separately, we find a higher prevalence of negative correlations, but only a small part is statistically significant. Since the results lack statistical significance and consistency between the conditions, we cannot infer any kind of relationship between the intercellular communication (measured by the z-score) and the pattern of metastasis.

3.3.2 Controlled Comparison of Intercellular Networks

Besides the pattern of gene expression particular to each tissue, metastasis formation is also conditioned by factors such as blood circulation pattern, organ accessibility or anatomical location, which are not taken into account when using the interaction z-score. Thus, we must account for these physical factors when studying the intrinsic compatibilities between metastasizing cells and welcoming tissues, which can be attributed to cell-cell communication. We used the organotropism and control pairs of tissues defined in Section 3.2 to compare and find differences in intercellular communication between the two groups. By requiring that tissues appear the same number of times in both groups, we take into account the physical factors that might influence the sites of metastasis formation, i.e., features besides the network statistic used to measure intercellular communication are cancelled out.

Organotropism pairs seem to establish a larger number of intercellular interactions compared

to control pairs in networks using both all or only literature curated intercellular interactions (Figure 3.7 and Supplementary Figure A.8). This tendency is also present when using the jaccard index to compare groups (Supplementary Figure A.9) and in directed networks, where there seems to be no significant differences in the number of interactions in each direction (Supplementary Figures A.10 and A.11). For weighted networks, we observe a similar pattern, with organotropism pairs having a consistently larger sum of weights than the control group (Supplementary Figures A.12, A.13).

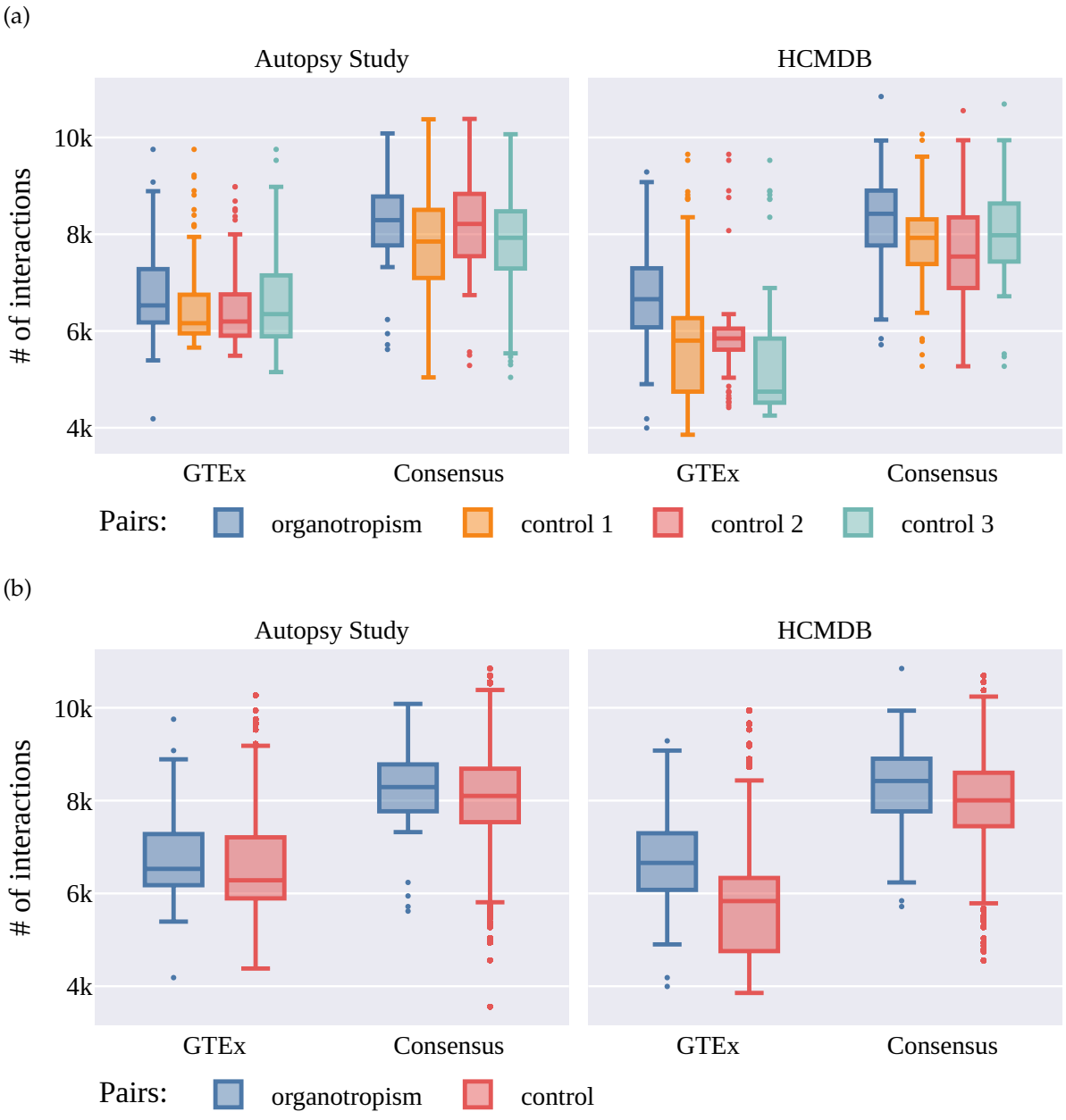


Figure 3.7: Controlled comparison between organotropism pairs and control pairs for **undirected networks built with gene expression calls**. (a) Example with three randomly selected distributions of control pairs; (b) Comparison with all generated controls. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). Cell-cell communication evaluated using the number of interactions.

Across all conditions, the difference between groups appear to be more pronounced in the HCMDB dataset. To find out if this difference is statistically significant, we perform a simple permutation-based statistical test. We counted the frequency of occurrences where a randomly-chosen control group has a median above the organotropism group to compute a p-value for each condition. The difference between organotropism and control groups in the HCMDB dataset is always statistically significant (p-value ≤ 0.05) across all conditions in both presence/absence calls and weighted networks (Supplementary Tables B.13-B.16). On the contrary, the difference between groups in the Autopsy Study is statistically significant in only 4 of the 36 tested conditions (all with weighted networks using Consensus/Curated/Undirected, GTEx/All/C $\rightarrow$ M, Consensus/All/M $\rightarrow$ C and Consensus/Curated/M $\rightarrow$ C interactions).

3.4 Metastasis-associated Intercellular Interactions

Our previous analysis suggests that, after taking into account other factors at play in metastasis development, cells from organotropism pairs of tissues tend to establish a larger number of intercellular interactions than cells from control pairs. Our interest now shifts to understanding which intercellular interactions may have a role in the metastatic process, and possible uncover proteins that might be suitable targets to prevent the development of metastasis tumours. We detected metastasis-associated interactions by determining a link between the presence/absence of each interaction in intercellular networks (or the weight of the interaction in each network) and the frequency of metastasis (or organotropism/control pair label) of each network pair of tissues (see Methods 2.8). We recorded the occurrences where the relation was deemed statistically significant, and the corresponding signal of the association. A positive signal indicates that the interaction contributes to metastasis formation, whereas a negative signal suggests the interaction may have a preventive role in tumour development.

In overall, we found 1095 metastasis-associated interactions with 607 unique genes. 881 interactions were positively associated with metastasis. Within those results, 528 interactions are present in the curated graph (363 positive associations), which corresponds to 453 unique genes. In the Table 3.2, we include the most significant interactions found in the different tests. We found some interactions that are only established between proteins in the same cell (not true cell-cell interactions) and decided to exclude them from the table. The table with all detected interactions is available at metastasis-associated intercellular interactions.

To exemplify how the method works, we show the results for the curated interaction between the β_2 microglobulin (*B2M* gene) protein and the antigen CD94 (*KLRD1* gene). The β_2 microglobulin is a component of the class I major histocompatibility complex (MHC), expressed in almost all tissues and cell types [128]. CD94 is a NK cell receptor, involved in self–nonself discrimination. Upon recognition of the MHC class Ib molecule HLA-E, CD94 can act as an inhibitor of NK cell activity (CD94–NKG2A receptor complex) or as an activating receptor of NK cell-mediated cell lysis (CD94–NKG2C receptor complex) [129]. We found a statistically significant and positive relation of this interaction with metastasis development in three of the four combinations of network type/metastasis data: presence/absence calls networks with organotropism pairs (fisher's exact test, FDR = 4.79×10^{-2}, Figure 3.8a), presence/absence calls

Table 3.2: Top intercellular interactions associated with metastasis ordered by test statistic. **median diff**: difference of medians for between the two groups. Signal refers to how the interaction affects metastasis development. **(+)**: promotes metastasis formation. **(-)**: prevents metastasis formation.

Source	Target	Value	Statistic	Test	Signal	FDR
B2M	KLRD1	5.613	OR	fisher's exact	+	4.79e-02
IFNG	IFNGR1	5.613	OR	fisher's exact	+	4.79e-02
IFNG	IFNGR2	5.613	OR	fisher's exact	+	4.79e-02
CSPG4	ITGA2	0.198	OR	fisher's exact	-	4.79e-02
CD33	SIGLEC10	2.000	median diff	MWU	+	3.56e-11
CCL4	CCR5	3.000	median diff	MWU	+	3.20e-11
ORM1	CCR5	2.000	median diff	MWU	+	7.17e-10
TNF	VSIR	2.000	median diff	MWU	+	1.38e-09
TNF	FLT4	2.000	median diff	MWU	+	1.38e-09
IFNG	IFNGR2	0.353	SRCC	SRCC	+	1.01e-08
MXRA5	SIGLEC7	0.344	SRCC	SRCC	+	1.79e-08
IFNG	IFNGR1	0.342	SRCC	SRCC	+	2.12e-08
HLA-B	KLRD1	0.324	SRCC	SRCC	+	1.23e-07
HLA-E	KLRD1	0.323	SRCC	SRCC	+	1.27e-07

networks with metastasis frequency (MWU test, FDR = 3.46×10^{-6}, Figure 3.8b), and weighted networks with metastasis frequency (SRCC test, FDR = 9.6×10^{-11}, Figure 3.8c).

We used Biopython package [130] to access the PubMed application programming interface (API) [93] and search for titles and abstracts containing the following query: "(*Gene*) AND (metastasis OR invasion)", where *Gene* represents the gene name of the protein of interest. To correct for genes that might be overrepresented in research, we also searched for publications that simply mention each gene. Then, we computed the ratio between the number of PubMed IDs matching the query with the association to metastasis or invasion, and the total number of PubMed IDs retrieved for each gene. Both searches were limited to 1000 references per gene. Proteins participating in metastasis-associated intercellular interactions seem to appear in more publications containing the queried term than non-associated proteins (Figure 3.9a and Supplementary Figure A.14a). The distribution of the two groups is statistically different at a significance level of 0.05 (KS test, p-value $< 7.5 \times 10^{-7}$), both when using all intercellular proteins or only proteins from curated interactions (see Methods 2.8.1.4). These results suggest that the group of proteins participating in metastasis-associated interactions is enriched in proteins with a known association with invasion/metastasis.

DisGeNET (v7.0) is a resource aiming to provide genotype-phenotype relationships. It integrates data from diverse sources, including literature text mining, genome-wide association studies (GWAS), and curated repositories, into a dataset of gene associations to human diseases [131]. Gene-disease associations are ranked by a score that takes into account the number and type of sources, and the number of publications supporting the association. Using DisGeNET SQLite database, we added up the scores of all associations each gene that codes for proteins establishing intercellular interactions has with different neoplasmic diseases (*Neoplasm*

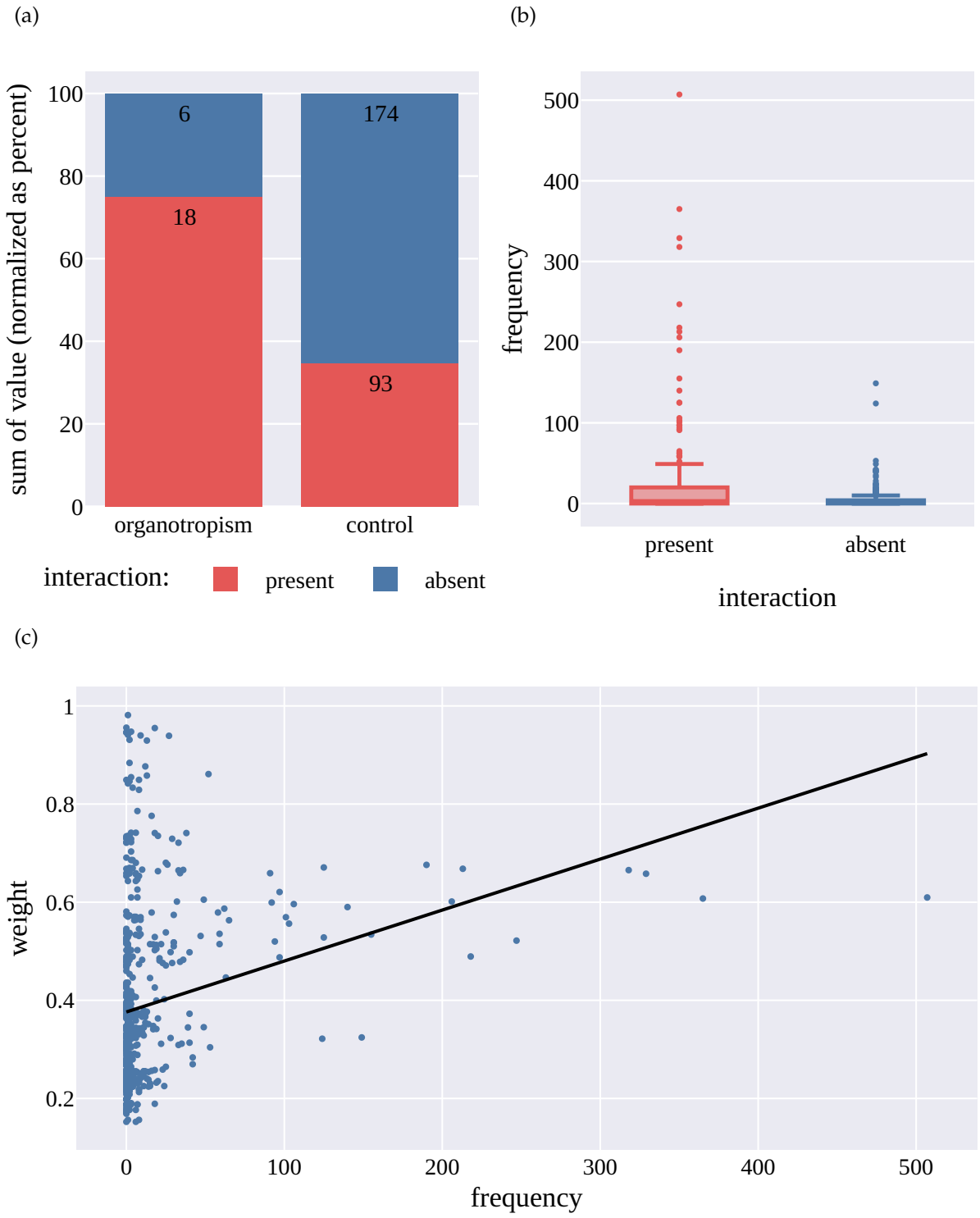


Figure 3.8: Method to associate intercellular interactions with metastasis and organotropism. Example for the *B2M-KLRD1* interaction. (a) Networks built with gene expression calls (Consensus) vs organotropism pairs (HCMDB)–evaluated with the Fisher's exact test. Numbers inside the bars correspond to the size (number of pairs) in each group. (b) Networks built with gene expression calls (Consensus) vs frequency of metastasis (Autopsy Study)–evaluated with the MWU. (c) Networks built with gene weights (Consensus) vs frequency of metastasis (Autopsy Study)–evaluated with SRCC and illustrated with an ordinary least squares regression line. **present**: pairs which establish the interaction. **absent**: pairs which without the interaction.

disease class name). We found that the group of genes from metastasis-associated interactions has a larger sum of association scores than genes with no association (Figure 3.9b and Supplementary Figure A.14b). The distribution of the two groups is statistically different (KS test, p-value $< 5.9 \times 10^{-10}$) both when using genes from the complete intercellular network or only genes from curated interactions.

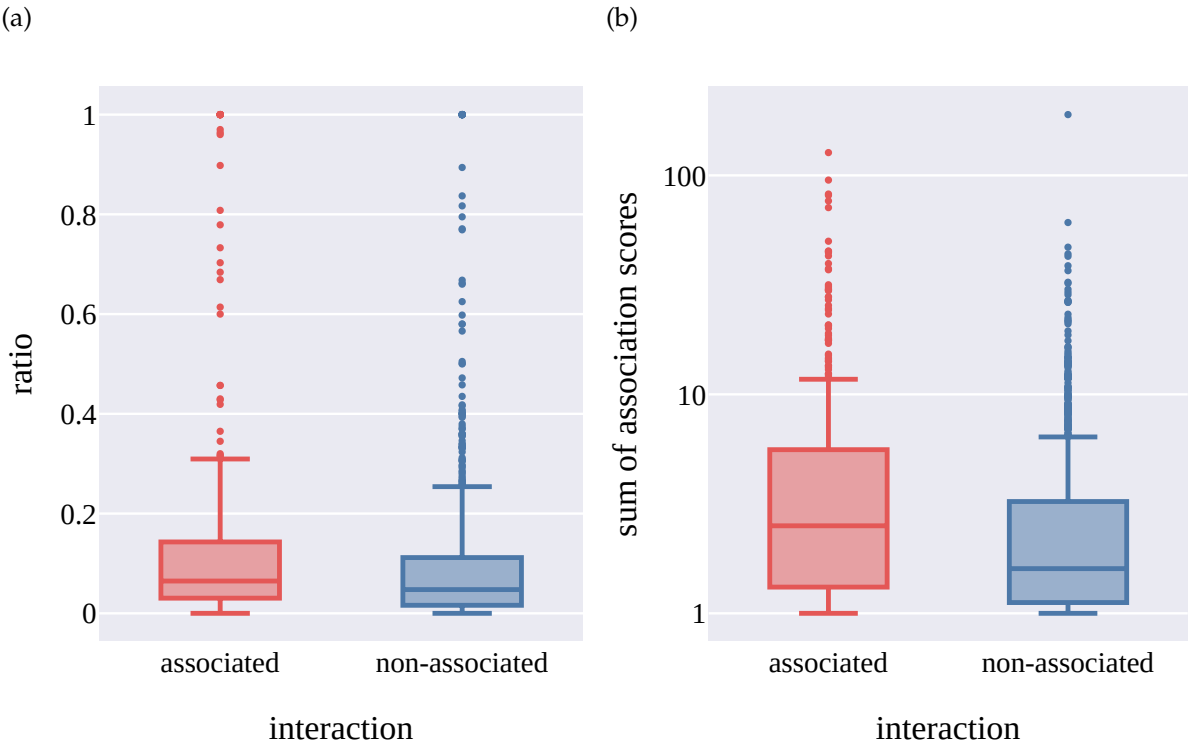


Figure 3.9: Prior known connections with cancer or metastasis in genes that participate in curated metastasis-associated interactions. (a) PubMed search for titles and abstracts containing the query: "(Gene) AND (metastasis OR invasion)". Distribution of the ratio between of the number of PubMed IDs matching the queried term and the total number of PubMed IDs that mention each gene. (b) DisGeNET association with the *Neoplasm* disease class. Distribution in log scale of the sum of association scores for each gene.

The Open Targets Platform integrates evidence from omics experiments, drugs, animal models, and scientific literature to score and rank target-disease associations. It also records target-drug and disease-drug associations, which allow for the identification and prioritization of potential therapeutic drug targets [132]. We downloaded the Open Targets Platform (v22.04) in parquet format to uncover target-disease-drug associations for metastasis-associated proteins. We found 222 proteins (160 from curated interactions) that are targeted by, at least, one known drug (results available at Drug Targets). All identified targets also have, at least, one disease association with a cancer or benign tumour (MONDO_0045024 identifier). Drug-disease associations do not discriminate between drugs that treat cancer and drugs that merely target symptoms resulting from disease or treatment. So, we could not accurately split drugs into anticancer and non-anticancer groups. Nevertheless, we found 53 proteins (39 from curated interactions) that are not targeted by drugs associated with neoplastic diseases. Table 3.3 shows the top ten targets, ordered by number of targeting drugs. All these proteins were identified as having an

association with metastasis development, and so might be relevant for drug repurposing efforts.

Table 3.3: Top 10 targets with only non cancer-associated drugs. The **Curated** column signals if the gene is present in the curated graph.

Target Gene Symbol	# of known drugs	Curated
APP	11	Yes
MC4R	5	Yes
CD40LG	5	Yes
SELP	4	Yes
PTH1R	4	Yes
BACE1	4	No
FCGRT	4	Yes
F2R	4	Yes
C3	3	No
IFNG	3	Yes

3.5 Intracellular Network Analysis

We have seen that intercellular networks between organotropism pairs of tissues appear to have a larger number of interactions than control pairs. Furthermore, we identified some interactions that may play a role in metastasis development and showed that a significant part of the proteins involved in those interactions have a prior documented relationship to cancer or metastasis processes. But how does cell-cell communication impact signalling and gene expression inside each cell? Which intracellular genes are affecting or affected by this communication, and what is their connection to the Hallmarks of Cancer? Do CDG play a role?

To try answering these questions, we build a PPI network that includes signalling pathways and TF-target gene data (see Methods 2.9.1). The complete intracellular network is represented by a directed graph containing 18215 nodes (proteins) and 311021 edges (interactions). We applied RWR to the intracellular network to simulate the propagation of two distinct signals inside the cell: one starting in proteins that function as receptors (targets) in intercellular interactions, diffuses through the network and eventually influences gene expression (target network); and the other that starts anywhere in the network and ends in proteins that send out signals in cell-cell communication (source network). Figure 3.10 illustrates this process; for details, see Methods 2.9.2. Using a permutation test, we selected intracellular proteins that are close to the proteins that take part in intercellular interactions with a positive association with metastasis development (see Methods 2.9.3). In total, we found 1041 intracellular proteins near target proteins and 27 near source proteins in the network (747 and 27 with curated intercellular interactions, respectively). We only used intercellular proteins involved in interactions with a positive association with metastasis, as opposed to proteins establishing interactions with a negative sign or even proteins that take part in both positive and negative interactions. Since we are not using the effect of the interaction (inhibition or activation), detecting intracellular proteins associated with positive interactions simplifies the interpretation of their possible role

in the network and in intercellular communication. This is particularly true if those proteins are classified as CDG and/or present in pathways related to the Hallmarks of Cancer.

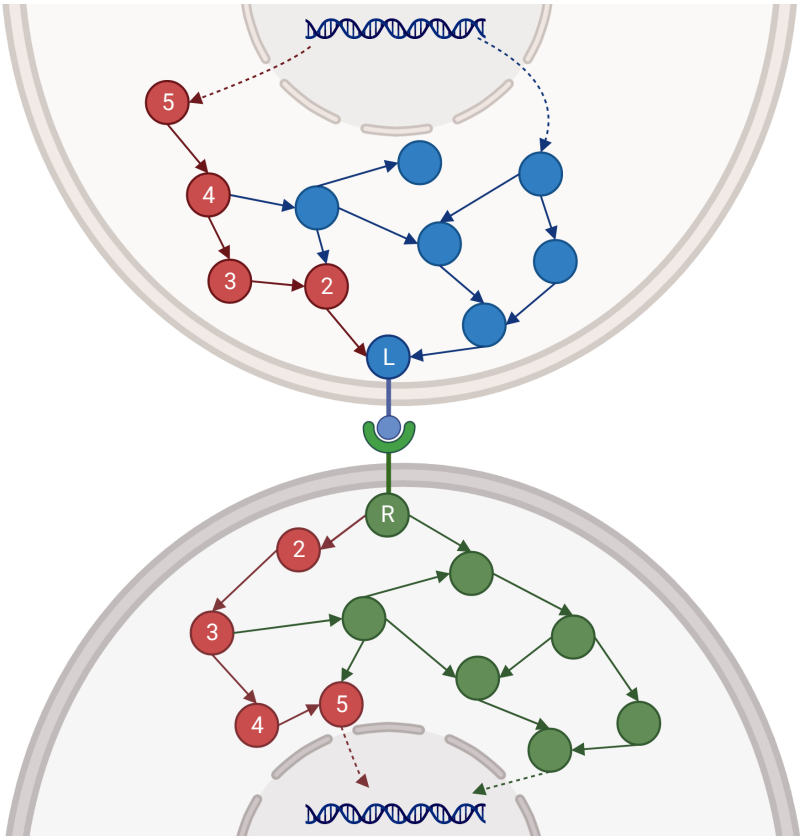


Figure 3.10: The two different types of intracellular signalling networks: upper cell in blue–source network; lower cell in green–target network. Nodes and edges in red represent a possible path taken by a RWR starting in an intercellular protein. Node numbers represent the order by which they are visited by the RWR, where the first node is either **L**–ligand, or **R**–receptor.

3.5.1 Cancer Driver Gene Enrichment Analysis

The Network of Cancer Genes & Healthy Drivers (NCG^{HD}) available at NCG^{HD} portal is a manually curated collection of cancer genes, healthy drivers and their properties. NCG^{HD} distinguishes between experimentally validated (canonical) CDG and candidate CDG which were identified in cancer sequencing screens and have only statistical support [133]. We performed an over-representation analysis using Fisher's Exact test on the metastasis-associated intracellular proteins using version 7.0 of the NCG^{HD} database using canonical and candidate CDG. Proteins close in the network to intercellular interactions that promote metastasis development are significantly enriched in CDG, with OR above 2 for the target dataset and above 5 for the source dataset (Table 3.4). Amongst the canonical CDG enriched in our list, we found known proto-oncogenes such as *MYC* and *RAC1* and tumour suppressor genes such as *TP53* and *RB1*. This suggests that CDG may have a direct influence in cell-cell signalling, both in pathways responsible for sending information and responding to stimuli. Complete lists of CDG are available at intracellular CDG and intercellular CDG.

As explained in Methods 2.9.3, we removed all intercellular proteins present in the intracellular network from the results of the permutation test. This allowed us to clearly link intracellular protein to intercellular interactions associated with the metastasis process. Notwithstanding, known protein products of proto-oncogenes such as the *EGFR* take part in intercellular interactions. Thus, it is worth finding out if genes that code for intercellular protein are also enriched in CDG. The dataset of intercellular proteins is a subset of the intracellular PPI network, and might be naturally enriched in CDG. So, for intercellular proteins, we performed CDG enrichment analysis both on proteins associated and non-associated with metastasis. We observed a statistically significant enrichment of CDG in the group of metastasis-associated proteins, though with smaller OR than in the intracellular analysis (Table 3.5). On the contrary, the group of non-associated proteins does not have more CDG than what is expected by chance, with p-values ≈ 1. These results not only indicate that the set of proteins involved in intercellular interactions is not naturally enriched in CDG, but they also provide further evidence for the role of the intercellular interactions we detected in the metastatic process.

3.5.2 Cancer Hallmarks Enrichment Analysis

CDG have a direct influence in cancer development and progression. Mutations in these genes alter the dynamics of processes and pathways in which their protein products participate, affecting neighbouring proteins and leading to the appearance of the characteristic cancer phenotypes. As previously mentioned, the Hallmarks of Cancer are a conceptual framework for identifying and describing the cellular processes that are altered in cancer cells. However, the

Table 3.4: Enrichment analysis of CDG in intracellular genes.

Gene Type	Network	# of Genes	# of CDGs	p-value	Odds Ratio
source	complete	27	16	1.80E-06	6.72
	curated	27	15	1.14E-05	5.78
target	complete	1041	335	5.82E-31	2.32
	curated	708	239	2.23E-25	2.45

Table 3.5: Enrichment analysis of CDG in intercellular genes associated (Yes) and non-associated (No) with metastasis development.

Gene Type	Network	Associated	# of Genes	# of CDGs	p-value	Odds Ratio
source	complete	No	1055	239	9.94E-01	0.72
		Yes	384	111	9.31E-03	1.39
	curated	No	1078	250	9.79E-01	0.74
		Yes	277	80	3.07E-02	1.34
target	complete	No	824	202	9.95E-01	0.70
		Yes	364	115	7.09E-03	1.42
	curated	No	806	205	9.89E-01	0.71
		Yes	265	86	1.66E-02	1.41

interpretation of these concepts and how they relate to specific biological processes as described by GO terms vary between studies. Chen et al. examined the semantic similarity between annotations and the gene set overlap and established a consensus among four different schemes mapping GO terms to Cancer Hallmarks [134]. Consensus terms are GO terms selected by more than 2 mapping schemes and are available at Consensus mapping scheme file.

We started by performing a GO term enrichment analysis on the set of intracellular genes associated with metastasis. We then used the hallmarks to GO terms mapping file to search for Cancer Hallmarks within enriched GO terms. With the dataset of curated intercellular interactions, we found 336 enriched GO terms in the group of genes associated to source proteins, and 1550 in the group associated to target proteins (Supplementary Table B.17). Five GO terms were mapped to four hallmarks in the source dataset, and thirteen GO terms were mapped to eight Hallmarks in the target dataset (Figure 3.11). Only the target dataset had genes associated with the hallmark *Activating Invasion and Metastasis*, corresponding to the GO term *negative regulation of cell adhesion* with 29 intracellular genes. Results for the complete intercellular network are similar and can be found in Supplementary Figure A.15.

For intercellular proteins associated with metastasis, we performed GO enrichment analysis in both metastasis-associated and non-associated proteins. The enrichment yield only one statistically significant GO term in the non-associated dataset (FDR < 0.05), which did not map to any hallmark of cancer (Supplementary Table B.18). This clearly contrasts with the results for the metastasis-associated group of proteins. In the curated network, we uncover 361 enriched GO terms in the group of source proteins, and 229 in the group of target proteins (Supplementary Table B.18). 12 GO terms were mapped to four hallmarks in the source dataset, and 9 GO terms were mapped to six hallmarks in the target dataset (Figure 3.12). The *Activating Invasion and Metastasis* hallmark is present in both datasets, mapping to the GO terms *Regulation of cell adhesion* (79 genes) and *Cell migration* (113 genes) in source proteins, and to *Cell migration* (107 genes) in target proteins. The enrichment analysis using the complete intercellular network also show a similar pattern, with even more GO terms mapped to the *Activating Invasion and Metastasis* hallmark (Supplementary Figure A.16). These findings suggest that proteins involved in intercellular interactions with an association with metastasis progress are particularly enriched in cellular processes that enable cell invasion and migration. In addition, it emphasizes the importance of intercellular communication in the successful progression of metastases and in the emergence of organotropism patterns.

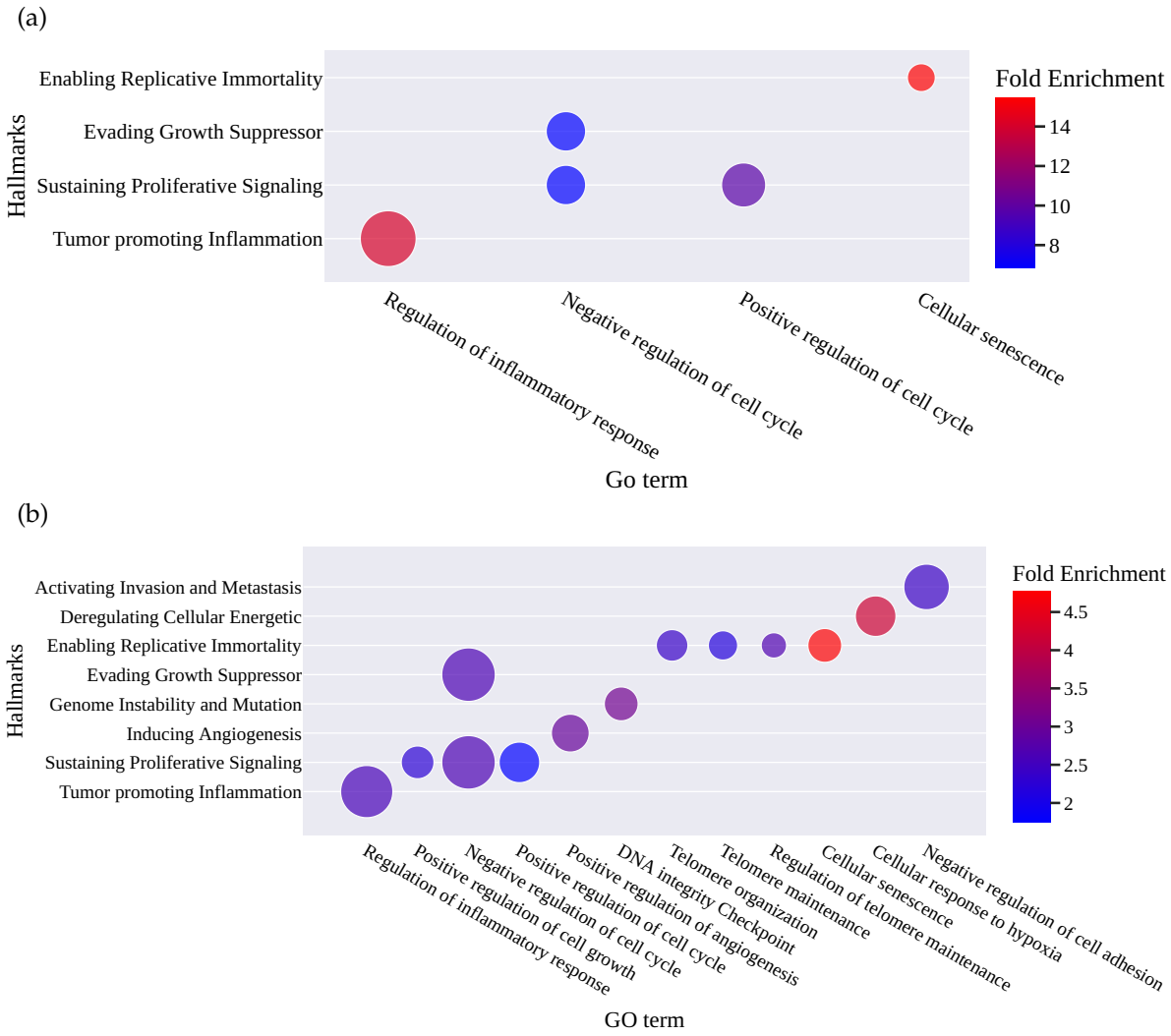
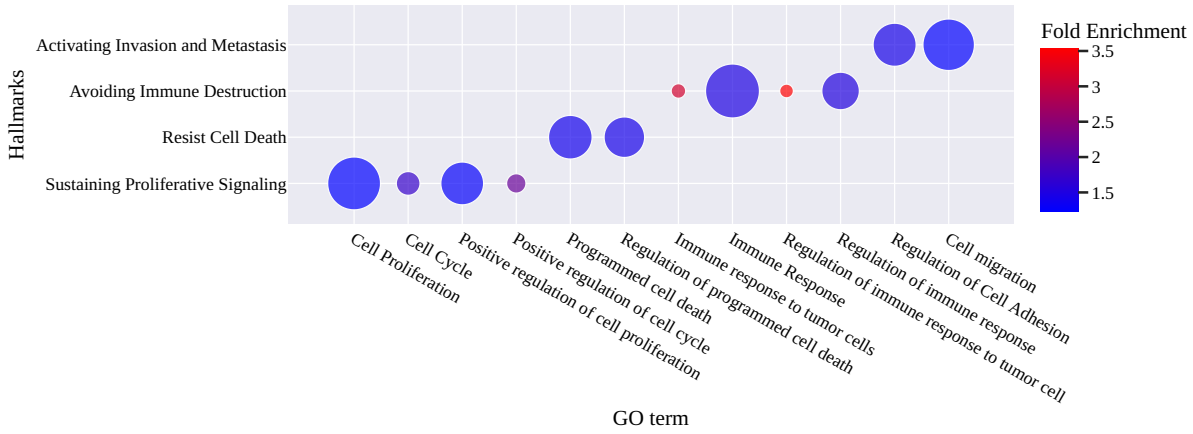


Figure 3.11: GO terms Hallmarks enrichment for intracellular genes linked to curated intercellular interactions. (a) source network genes; (b) target network genes. The size of the circle corresponds to the number of genes in each GO term.

(a)



(b)

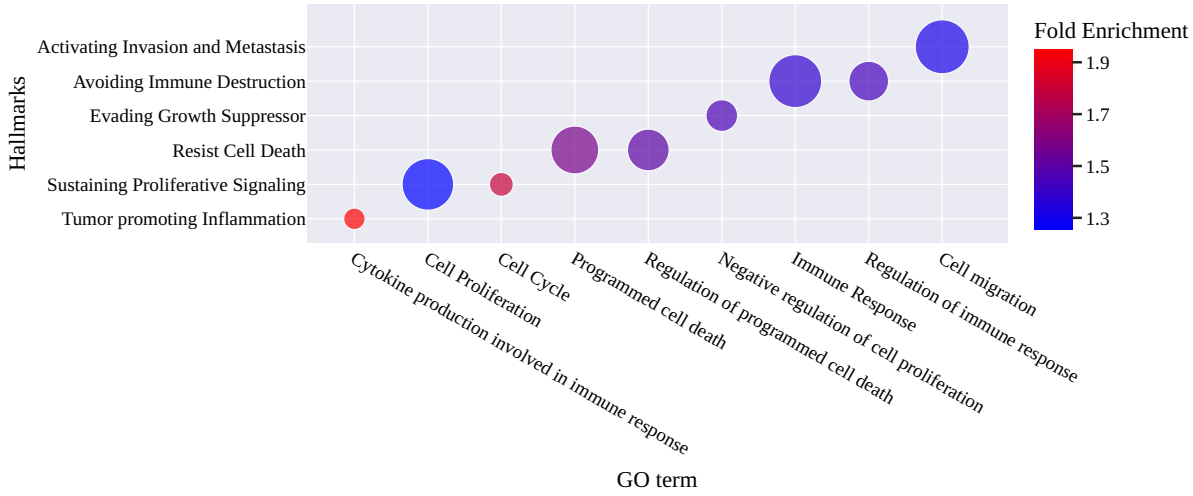


Figure 3.12: GO terms Hallmarks enrichment for intercellular genes from curated intercellular interactions. (a) source genes; (b) target genes. The size of the circle corresponds to the number of genes in each GO term.

4

Discussion

The hypothesis that cell-cell communication plays a crucial role in the success of the metastatic process has been extensively discussed in the literature [48, 135]. Moreover, factors selectively expressed both by primary tumour cells and the PMN, contribute to the creation of a welcoming environment for the survival and proliferation of metastasizing cells, which helps explain the tissue-specificity of metastases [57, 65, 136]. Taking into account these premises, we built tissue-specific intercellular PPI networks to measure the level of communication between cells from the primary cancer tissue and cells from a tissue to where that cancer might metastasize. Using a method of comparison that controls for other tissue-related factors that affect organotropism besides cell-cell communication, we found that sites where cancers metastasize more often than what is expected at random (organotropism pairs) establish a larger number of intercellular interactions than sites with low incidence of metastasis. This was observed in all conditions tested, including in both networks built with gene presence/absence calls and weighted networks. We then investigated metastasis-associated interactions, which may be driving metastasis and contributing to the observed differences in cell-cell communication. Since our method uses intercellular networks built using gene expression from non-diseased tissue, our results are not biased towards proteins that are overexpressed in cancer. Furthermore, it allows us to pinpoint determinants that might explain organotropism, taking into account only the specificities of the cancer tissue-of-origin. We detected 528 curated interactions that could play a role in metastasis formation, some of which already described in literature (PubMed) and in gene-disease association resources (DisGeNET) as playing a role in cancer and/or metastasis. One of those is the interaction already described between the β_2 microglobulin, a component of the HLA-E complex, and the NK cell antigen CD94. The NKG2A–CD94 heterodimeric receptor recognizes peptides presented by HLA-E and leads to inhibition of NK cell effector functions, including NK cell-mediated cell killing [137]. In fact, we also detected in our analysis both the HLA-E–NKG2A and the HLA-E–CD94 as metastasis-associated interactions. Contrary to classical MHC class I molecules, HLA-E is usually upregulated in several cancers, which correlates with a poor prognosis [129]. Tumour cell lines with knockout of HLA-E have increased NK cell sensitivity [138], and downregulation of NKG2A in NK cells increased their antitumour activity [139]. The use of natural killer cells as cancer immunotherapy agents is gaining interest, as they are easier to prepare and safer than CAR-T cells [1]. Besides using modified NK cells

with enhanced cytotoxic activity, it is possible to enhance their antitumour activity by targeting immune checkpoints such as the HLA-E–NKG2A/CD94 axis. We found monalizumab in the OpenTargets search, an anticancer drug that targets NKG2A. Monalizumab is a monoclonal antibody currently being tested in phase three clinical trials. It is an immune checkpoint inhibitor that enhances NK cell-mediated killing of tumour cells by blocking NKG2A receptor [137, 140]. Another example found with our method is the heterotypic adherens juction between N-cadherin (*CDH2* gene) and E-cadherin (*CDH1* gene). There is evidence that this interaction promotes tumour cell migration and metastasis through, at least, two known mechanisms. First, E-cadherin expressed in cancer cells promotes their coupling with cancer-associated fibroblasts expressing N-cadherin, which enables cooperation and drives collective invasion [141]. Second, cell adhesion in the bone-specific microenvironment between metastasizing cells expressing E-cadherin and osteoblasts expressing N-cadherin activates the mTOR pathway, which stimulates cell growth and proliferation, driving the formation of micrometastases [142]. Finally, we connected the intercellular interactions graph with processes inside the cell using a network depicting signalling pathways and TF-target gene interactions. Both intercellular metastasis-associated proteins and their closest neighbours in the intracellular PPI network are enriched in CDG. We also found several enriched GO biological processes related to the hallmarks of cancer which are directly affected or affecting metastasis-associated interactions. This includes the terms *Regulation of cell adhesion* and *Cell migration* that are linked to the hallmark *Activating Invasion and Metastasis*. Taken together, these results give strength to our method of associating specific interactions to organotropism and metastasis formation. Most associations we found are new and have not previously been connected to metastasis progression. These intercellular interactions and their components might motivate new research and suggest new therapeutic opportunities to treat and prevent metastasis.

The analysis and integration of gene expression data in intercellular PPI networks constituted a central part of our work and included two major steps–choosing and processing RNA-seq data, and performing gene expression calls. The decision to use bulk RNA-seq data introduces a limitation: instead of characterizing the intercellular communication between two specific individual cells, our networks comprise averages of gene expression signals from individual cells in two different tissues [143]. One way to improve upon this issue would be to use single-cell RNA-seq (scRNA-seq) data to build intercellular PPI networks. scRNA-seq technology measures the transcriptome at a single-cell resolution, and thus can be used to resolve intratumoural heterogeneity, classify different cell types and cell states [144]. This would allow us to account for cancer subtypes and characterize specific interactions that metastasizing cells establish with different cell types inside the PMN. However, scRNA-seq also introduces some technical limitations, and data processing comes with new challenges. For example, the small amount of mRNA per cell coupled with low capturing efficiency may cause some mRNAs to be missed during reverse transcription and cDNA amplification, thus preventing them from being detected in the sequencing step. These *dropout* events generate a significant between-cell technical variability, specially for genes with low or moderate expression, resulting in very sparse data [145]. So, it is possible that by using scRNA-seq gene expression, we could lose some metastasis-associated interactions that were found with bulk RNA-seq data. This means that scRNA-seq

data does not fully replace the analysis we performed, but can complement it, giving it a better context and higher resolution.

A common approach to perform gene presence/absence calls is to consider all genes with expression values above a specific TPM threshold as present in a tissue. This has the disadvantage of not considering the distinct patterns of expression for different genes. We tried to solve this issue by using the tissue-specificity of each gene as an additional factor to classify genes as expressed or not in a tissue. Several parameters in our method were set based on observation of distributions and through internal validation, including the tau values used to assign genes to different groups and parameters adjusted both in outlier detection and clustering analysis. Additionally, other factors not taken into account, such as protein turnover and chromatin accessibility, impact protein levels. These issues can be overcome by comparing the results we obtained with calls made using other methods. For example, Bgee [146], a database for retrieval and comparison of gene expression patterns across multiple animal, provides gene expression calls for GTEx tissue that can be directly compared with our method.

It is worth to mention three small issues we encountered throughout our work, which could not be immediately solved. One concerns the HCMDB resource that integrates RNA-seq data from cancer and metastasis tumours available at public repositories, used to extract information about metastasis frequency. We found that some pairs of entries refer to the same subject and experiment, but account for different tumours (primary or metastasis) and therefore correspond to duplicate data about metastasis frequency for the same cancer and metastasis site. The observed impact of this issue on metastatic patterns appears to be minimal and, in fact, the HCMDB dataset is highly concordant with what is observed in the clinical literature. Nonetheless, we already have identified other sources of metastasis frequency, which will be used to validate our results in this work. The other has to do with the dataset of intercellular interactions we retrieve from the Omnipath database. We found that some cell-cell interactions, annotated as such, occur between membrane and transmembrane components within the same cell. One example is the interaction between CD94 and DAP12 (*TYROBP* gene). DAP12 is a transmembrane adapter protein that associates with activating receptors found on the surface of immune cells. It associates with the CD94-NKG2C receptor, which mediates signalling and cell activation in NK cells [129]. Some of those interactions were identified as associated with metastasis and are present in the final list. For future analyses, we will try to use Omnipath annotations to filter out these interactions. The third and last issue concern the step of complex unfolding we performed when processing the intercellular interaction dataset. This step assigns interactions to each protein in the complex individually. Though it simplifies the analysis of the intercellular PPI networks, it introduces an artefact in our work: if a certain complex protein is not expressed in a tissue, that complex will not form and all interactions with that complex will not exist.

Summing up, we developed a novel approach to uncover organotropism determinants using intercellular PPI networks. We showed that we could capture relevant proteins and interactions which are enriched in CDG and associated to known processes altered in cancer. Contrary to other association studies, we did not use genomic signatures (mutation or gene expression patterns) of metastasis tumours to find metastasis driver genes. As mentioned, this has the

advantage of allowing us to associate the pre-tumoural variation of gene expression with organotropism. Going forward, we would like to address some of the known issues in our method and extend it with new data to improve the strength of our associations. A clear step would be validating our findings using gene expression data from cancer patients to determine if the proteins and interactions detected are overexpressed in the subsets of patients that have metastasis. Lastly, it would also be interesting to focus on metastatic sites to find tissue-specific interactions that can help explain why certain cancers prefer to spread to those sites.

Bibliography

[1] T. J. Laskowski, A. Biederstädt, and K. Rezvani. "Natural Killer Cells in Antitumour Adoptive Cell Immunotherapy". In: *Nature Reviews Cancer* 22.10 (2022-10), pp. 557–575. ISSN: 1474-1768. DOI: 10.1038/s41568-022-00491-0. (Visited on 2023-02-11) (cit. on pp. xv, 50).

[2] *MTOR - Serine/Threonine-Protein Kinase mTOR - Homo Sapiens (Human) | UniProtKB | UniProt*. https://www.uniprot.org/uniprotkb/P42345/entry. (Visited on 2023-05-19) (cit. on p. xv).

[3] G. S. Martin. "Cell Signaling and Cancer". In: *Cancer Cell* 4.3 (2003-09), pp. 167–174. ISSN: 1535-6108. DOI: 10.1016/S1535-6108(03)00216-2. (Visited on 2023-03-15) (cit. on p. xv).

[4] P. Ranganathan, K. L. Weaver, and A. J. Capobianco. "Notch Signalling in Solid Tumours: A Little Bit of Everything but Not All the Time". In: *Nature Reviews Cancer* 11.5 (2011-05), pp. 338–351. ISSN: 1474-1768. DOI: 10.1038/nrc3035. (Visited on 2023-03-23) (cit. on pp. xv, 2).

[5] *PIK3CA - Phosphatidylinositol 4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha Isoform - Homo Sapiens (Human) | UniProtKB | UniProt*. https://www.uniprot.org/uniprotkb/P42336/entry. (Visited on 2023-05-19) (cit. on p. xv).

[6] *AKT1 - RAC-alpha Serine/Threonine-Protein Kinase - Homo Sapiens (Human) | UniProtKB | UniProt*. https://www.uniprot.org/uniprotkb/P31749/entry. (Visited on 2023-05-19) (cit. on p. xv).

[7] *RAC1 - Ras-related C3 Botulinum Toxin Substrate 1 - Homo Sapiens (Human) | UniProtKB | UniProt*. https://www.uniprot.org/uniprotkb/P63000/entry. (Visited on 2023-05-19) (cit. on p. xv).

[8] *TGFB1 - Transforming Growth Factor Beta-1 Proprotein - Homo Sapiens (Human) | UniProtKB | UniProt*. https://www.uniprot.org/uniprotkb/P01137/entry. (Visited on 2023-05-19) (cit. on p. xv).

[9] J. Liu et al. "Wnt/β-Catenin Signalling: Function, Biological Mechanisms, and Therapeutic Opportunities". In: *Signal Transduction and Targeted Therapy* 7.1 (2022-01), pp. 1–23. ISSN: 2059-3635. DOI: 10.1038/s41392-021-00762-6. (Visited on 2023-03-15) (cit. on p. xvi).

[10] *What Is Cancer? - NCI*. https://www.cancer.gov/about-cancer/understanding/what-is-cancer. cgvArticle. 2007. (Visited on 2023-03-06) (cit. on pp. 1, 2).

[11] W. H. O. (WHO). *Cancer*. https://www.who.int/news-room/fact-sheets/detail/cancer. 2021-03. (Visited on 2021-09-20) (cit. on p. 1).

[12] H. Sung et al. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249. ISSN: 1542-4863. DOI: 10.3322/caac.21660. (Visited on 2023-03-06) (cit. on p. 1).

[13] M. R. Stratton, P. J. Campbell, and P. A. Futreal. "The Cancer Genome". In: *Nature* 458.7239 (2009-04), pp. 719–724. ISSN: 1476-4687. DOI: 10.1038/nature07943. (Visited on 2023-03-08) (cit. on pp. 1, 2).

[14] F. Martínez-Jiménez et al. "A Compendium of Mutational Cancer Driver Genes". In: *Nature Reviews Cancer* 20.10 (2020-10), pp. 555–572. ISSN: 1474-1768. DOI: 10.1038/s41568-020-0290-x. (Visited on 2023-03-20) (cit. on p. 1).

[15] B. Alberts et al. *Molecular Biology of the Cell*. Sixth edition. New York, NY: Garland Science, Taylor and Francis Group, 2015. ISBN: 978-0-8153-4524-4 (cit. on pp. 1, 2).

[16] I. M. de Alboran et al. "Analysis of C-MYC Function in Normal Cells via Conditional Gene-Targeted Mutation". In: *Immunity* 14.1 (2001-01), pp. 45–55. ISSN: 1074-7613. DOI: 10.1016/S1074-7613(01)00088-7. (Visited on 2023-05-19) (cit. on p. 1).

[17] E. Toufektchan and F. Toledo. "The Guardian of the Genome Revisited: P53 Downregulates Genes Required for Telomere Maintenance, DNA Repair, and Centromere Structure". In: *Cancers* 10.5 (2018-05), p. 135. ISSN: 2072-6694. DOI: 10.3390/cancers10050135. (Visited on 2023-05-19) (cit. on p. 2).

[18] C. A. Aktipis et al. "Cancer across the Tree of Life: Cooperation and Cheating in Multicellularity". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1673 (2015-07), p. 20140219. DOI: 10.1098/rstb.2014.0219. (Visited on 2023-03-08) (cit. on p. 2).

[19] J. Boutry et al. "Ecological and Evolutionary Consequences of Anticancer Adaptations". In: *iScience* 23.11 (2020-11), p. 101716. ISSN: 2589-0042. DOI: 10.1016/j.isci.2020.101716. (Visited on 2023-03-08) (cit. on p. 2).

[20] J. Campisi. "Aging, Cellular Senescence, and Cancer". In: *Annual Review of Physiology* 75.1 (2013), pp. 685–705. DOI: 10.1146/annurev-physiol-030212-183653. (Visited on 2023-03-01) (cit. on p. 2).

[21] B. Vogelstein et al. "Cancer Genome Landscapes". In: *Science* 339.6127 (2013-03), pp. 1546–1558. DOI: 10.1126/science.1235122. (Visited on 2021-09-20) (cit. on p. 2).

[22] L. Moore et al. "The Mutational Landscape of Human Somatic and Germline Cells". In: *Nature* 597.7876 (2021-09), pp. 381–386. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03822-7. (Visited on 2023-03-22) (cit. on p. 2).

[23] T. Ushijima, S. J. Clark, and P. Tan. "Mapping Genomic and Epigenomic Evolution in Cancer Ecosystems". In: *Science* 373.6562 (2021-09), pp. 1474–1479. DOI: 10.1126/science.abh1645. (Visited on 2023-03-08) (cit. on p. 2).

[24] A. Acha-Sagredo, P. Ganguli, and F. D. Ciccarelli. "Somatic Variation in Normal Tissues: Friend or Foe of Cancer Early Detection?" In: *Annals of Oncology* 33.12 (2022-12), pp. 1239–1249. ISSN: 0923-7534. DOI: 10.1016/j.annonc.2022.09.156. (Visited on 2023-03-08) (cit. on p. 2).

[25] J. J. Bianchi et al. "Not All Cancers Are Created Equal: Tissue Specificity in Cancer Genes and Pathways". In: *Current Opinion in Cell Biology*. Cell Signalling (2020) 63 (2020-04), pp. 135–143. ISSN: 0955-0674. DOI: 10.1016/j.ceb.2020.01.005. (Visited on 2021-09-17) (cit. on p. 2).

[26] D. Hanahan and R. A. Weinberg. "The Hallmarks of Cancer". In: *Cell* 100.1 (2000-01), pp. 57–70. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/S0092-8674(00)81683-9. (Visited on 2023-03-29) (cit. on p. 3).

[27] D. Hanahan and R. A. Weinberg. "Hallmarks of Cancer: The Next Generation". In: *Cell* 144.5 (2011-03), pp. 646–674. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.02.013. (Visited on 2021-09-20) (cit. on p. 3).

[28] D. Hanahan. "Hallmarks of Cancer: New Dimensions". In: *Cancer Discovery* 12.1 (2022-01), pp. 31–46. ISSN: 2159-8274, 2159-8290. DOI: 10.1158/2159-8290.CD-21-1059. (Visited on 2022-02-01) (cit. on p. 3).

[29] T. Korcsmáros et al. "Uniformly Curated Signaling Pathways Reveal Tissue-Specific Cross-Talks and Support Drug Target Discovery". In: *Bioinformatics* 26.16 (2010-08), pp. 2042–2050. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq310. (Visited on 2023-03-15) (cit. on p. 3).

[30] F. Sanchez-Vega et al. "Oncogenic Signaling Pathways in The Cancer Genome Atlas". In: *Cell* 173.2 (2018-04), 321–337.e10. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.03.035. (Visited on 2023-03-15) (cit. on p. 4).

[31] R. G. Jones and C. B. Thompson. "Tumor Suppressors and Cell Metabolism: A Recipe for Cancer Growth". In: *Genes & Development* 23.5 (2009-01), pp. 537–548. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.1756509. (Visited on 2023-03-20) (cit. on p. 4).

[32] N. N. Pavlova and C. B. Thompson. "The Emerging Hallmarks of Cancer Metabolism". In: *Cell Metabolism* 23.1 (2016-01), pp. 27–47. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2015.12.006. (Visited on 2023-03-20) (cit. on p. 4).

[33] M. Rafaeva and J. T. Erler. "Framing Cancer Progression: Influence of the Organ- and Tumour-Specific Matrisome". In: *The FEBS Journal* 287.8 (2020), pp. 1454–1477. ISSN: 1742-4658. DOI: 10.1111/febs.15223. (Visited on 2021-11-16) (cit. on p. 4).

[34] F. R. Balkwill, M. Capasso, and T. Hagemann. "The Tumor Microenvironment at a Glance". In: *Journal of Cell Science* 125.23 (2012-12), pp. 5591–5596. ISSN: 0021-9533. DOI: 10.1242/jcs.116392. (Visited on 2023-03-29) (cit. on p. 4).

[35] M.-F. Pang and C. M. Nelson. "Intercellular Communication, the Tumor Microenvironment, and Tumor Progression". In: *Intercellular Communication in Cancer*. Ed. by M. Kandouz. Dordrecht: Springer Netherlands, 2015, pp. 343–362. ISBN: 978-94-017-7380-5. DOI: 10.1007/978-94-017-7380-5_13. (Visited on 2022-05-25) (cit. on p. 4).

[36] G. Pinto, C. Brou, and C. Zurzolo. "Tunneling Nanotubes: The Fuel of Tumor Progression?" In: *Trends in Cancer* 6.10 (2020-10), pp. 874–888. ISSN: 2405-8033. DOI: 10.1016/j.trecan.2020.04.012. (Visited on 2022-12-22) (cit. on p. 4).

[37] M. W. Pickup, J. K. Mouw, and V. M. Weaver. "The Extracellular Matrix Modulates the Hallmarks of Cancer". In: *EMBO reports* 15.12 (2014-12), pp. 1243–1253. ISSN: 1469-221X. DOI: 10.15252/embr.201439246. (Visited on 2023-03-01) (cit. on p. 4).

[38] M. Binnewies et al. "Understanding the Tumor Immune Microenvironment (TIME) for Effective Therapy". In: *Nature Medicine* 24.5 (2018-05), pp. 541–550. ISSN: 1546-170X. DOI: 10.1038/s41591-018-0014-x. (Visited on 2023-03-08) (cit. on p. 4).

[39] K. E. Pauken and E. J. Wherry. "Overcoming T Cell Exhaustion in Infection and Cancer". In: *Trends in Immunology* 36.4 (2015-04), pp. 265–276. ISSN: 1471-4906. DOI: 10.1016/j.it.2015.02.008. (Visited on 2023-03-30) (cit. on p. 4).

[40] S. Valastyan and R. A. Weinberg. "Tumor Metastasis: Molecular Insights and Evolving Paradigms". In: *Cell* 147.2 (2011-10), pp. 275–292. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.09.024. (Visited on 2021-11-03) (cit. on pp. 5–7).

[41] K. Ganesh and J. Massagué. "Targeting Metastatic Cancer". In: *Nature Medicine* 27.1 (2021-01), pp. 34–44. ISSN: 1546-170X. DOI: 10.1038/s41591-020-01195-4. (Visited on 2023-03-20) (cit. on pp. 5, 6, 8).

[42] J. Massagué and A. C. Obenauf. "Metastatic Colonization by Circulating Tumour Cells". In: *Nature* 529.7586 (2016-01), pp. 298–306. ISSN: 1476-4687. DOI: 10.1038/nature17038. (Visited on 2021-11-04) (cit. on pp. 5–8).

[43] K. J. Luzzi et al. "Multistep Nature of Metastatic Inefficiency: Dormancy of Solitary Cells after Successful Extravasation and Limited Survival of Early Micrometastases". In: *The American Journal of Pathology* 153.3 (1998-09), pp. 865–873. ISSN: 0002-9440. DOI: 10.1016/S0002-9440(10)65628-3. (Visited on 2023-04-05) (cit. on pp. 5, 8).

[44] C. Heyn et al. "In Vivo MRI of Cancer Cell Fate at the Single-Cell Level in a Mouse Model of Breast Cancer Metastasis to the Brain". In: *Magnetic Resonance in Medicine* 56.5 (2006), pp. 1001–1010. ISSN: 1522-2594. DOI: 10.1002/mrm.21029. (Visited on 2023-04-05) (cit. on p. 5).

[45] N. J. Birkbak and N. McGranahan. "Cancer Genome Evolutionary Trajectories in Metastasis". In: *Cancer Cell* 37.1 (2020-01), pp. 8–19. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2019.12.004. (Visited on 2023-03-20) (cit. on p. 5).

[46] R. A. Burrell et al. "The Causes and Consequences of Genetic Heterogeneity in Cancer Evolution". In: *Nature* 501.7467 (2013-09), pp. 338–345. ISSN: 1476-4687. DOI: 10.1038/nature12625. (Visited on 2023-03-09) (cit. on p. 5).

[47] S. Turajlic et al. "Resolving Genetic Heterogeneity in Cancer". In: *Nature Reviews Genetics* 20.7 (2019-07), pp. 404–416. ISSN: 1471-0064. DOI: 10.1038/s41576-019-0114-6. (Visited on 2023-04-02) (cit. on p. 5).

[48] A. W. Lambert, D. R. Pattabiraman, and R. A. Weinberg. "Emerging Biological Principles of Metastasis". In: *Cell* 168.4 (2017-02), pp. 670–691. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.11.037. (Visited on 2021-09-17) (cit. on pp. 5, 7, 50).

[49] A. C. Obenauf and J. Massagué. "Surviving at a Distance: Organ-Specific Metastasis". In: *Trends in Cancer* 1.1 (2015-09), pp. 76–91. ISSN: 2405-8033. DOI: 10.1016/j.trecan.2015.07.009. (Visited on 2021-09-17) (cit. on p. 5).

[50] G. P. Gupta and J. Massagué. "Cancer Metastasis: Building a Framework". In: *Cell* 127.4 (2006-11), pp. 679–695. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.11.001. (Visited on 2023-03-31) (cit. on pp. 6, 8).

[51] J. Massagué and K. Ganesh. "Metastasis-Initiating Cells and Ecosystems". In: *Cancer Discovery* 11.4 (2021-04), pp. 971–994. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-21-0010. (Visited on 2023-04-02) (cit. on pp. 6, 9).

[52] E. Lengyel. "Ovarian Cancer Development and Metastasis". In: *The American Journal of Pathology* 177.3 (2010-09), pp. 1053–1064. ISSN: 0002-9440. DOI: 10.2353/ajpath.2010.100105. (Visited on 2021-11-03) (cit. on pp. 6, 91).

[53] M. Castaneda et al. "Mechanisms of Cancer Metastasis". In: *Seminars in Cancer Biology* 87 (2022-12), pp. 17–31. ISSN: 1044-579X. DOI: 10.1016/j.semcancer.2022.10.006. (Visited on 2023-04-08) (cit. on p. 7).

[54] M. Janiszewska, M. C. Primi, and T. Izard. "Cell Adhesion in Cancer: Beyond the Migration of Single Cells". In: *Journal of Biological Chemistry* 295.8 (2020-02), pp. 2495–2505. ISSN: 0021-9258. DOI: 10.1074/jbc.REV119.007759. (Visited on 2023-04-10) (cit. on p. 7).

[55] B. Strilic and S. Offermanns. "Intravascular Survival and Extravasation of Tumor Cells". In: *Cancer Cell* 32.3 (2017-09), pp. 282–293. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2017.07.001. (Visited on 2023-04-10) (cit. on p. 7).

[56] A. F. Chambers, A. C. Groom, and I. C. MacDonald. "Dissemination and Growth of Cancer Cells in Metastatic Sites". In: *Nature Reviews Cancer* 2.8 (2002-08), pp. 563–572. ISSN: 1474-1768. DOI: 10.1038/nrc865. (Visited on 2023-03-31) (cit. on p. 7).

[57] H. A. Smith and Y. Kang. "Determinants of Organotropic Metastasis". In: *Annual Review of Cancer Biology* 1.1 (2017-03), pp. 403–423. ISSN: 2472-3428. DOI: 10.1146/annurev-cancerbio-041916-064715. (Visited on 2021-09-15) (cit. on pp. 7, 50).

[58] S. Paget. "THE DISTRIBUTION OF SECONDARY GROWTHS IN CANCER OF THE BREAST." In: *The Lancet*. Originally Published as Volume 1, Issue 3421 133.3421 (1889-03), pp. 571–573. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(00)49915-0. (Visited on 2023-04-12) (cit. on p. 7).

[59] J. E. Talmadge and I. J. Fidler. "AACR Centennial Series: The Biology of Cancer Metastasis: Historical Perspective". In: *Cancer Research* 70.14 (2010-07), pp. 5649–5669. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-10-1040. (Visited on 2021-11-15) (cit. on p. 7).

[60] M.-Y. Kim et al. "Tumor Self-Seeding by Circulating Cancer Cells". In: *Cell* 139.7 (2009-12), pp. 1315–1326. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.11.025. (Visited on 2021-11-16) (cit. on p. 8).

[61] S. Vanharanta and J. Massagué. "Origins of Metastatic Traits". In: *Cancer Cell* 24.4 (2013-10), pp. 410–421. ISSN: 1535-6108. DOI: 10.1016/j.ccr.2013.09.007. (Visited on 2023-03-09) (cit. on p. 8).

[62] D. H. Jones et al. "Regulation of Cancer Cell Migration and Bone Metastasis by RANKL". In: *Nature* 440.7084 (2006-03), pp. 692–696. ISSN: 1476-4687. DOI: 10.1038/nature04524. (Visited on 2023-04-13) (cit. on p. 8).

[63] Y. Liu and X. Cao. "Characteristics and Significance of the Pre-metastatic Niche". In: *Cancer Cell* 30.5 (2016-11), pp. 668–681. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2016.09.011. (Visited on 2023-03-08) (cit. on p. 8).

[64] A. R. Chin and S. E. Wang. "Cancer Tills the Premetastatic Field: Mechanistic Basis and Clinical Implications". In: *Clinical Cancer Research* 22.15 (2016-07), pp. 3725–3733. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-16-0028. (Visited on 2023-04-13) (cit. on p. 8).

[65] H. Peinado et al. "Pre-Metastatic Niches: Organ-Specific Homes for Metastases". In: *Nature Reviews Cancer* 17.5 (2017-05), pp. 302–317. ISSN: 1474-1768. DOI: 10.1038/nrc.2017.6. (Visited on 2023-03-07) (cit. on pp. 9, 50).

[66] P. W. Anderson. "More Is Different". In: *Science* 177.4047 (1972-08), pp. 393–396. DOI: 10.1126/science.177.4047.393. (Visited on 2023-04-16) (cit. on p. 9).

[67] A.-L. Barabási and Z. N. Oltvai. "Network Biology: Understanding the Cell's Functional Organization". In: *Nature Reviews Genetics* 5.2 (2004-02), pp. 101–113. ISSN: 1471-0064. DOI: 10.1038/nrg1272. (Visited on 2023-04-16) (cit. on pp. 9–11).

[68] H. Kitano. "Systems Biology: A Brief Overview". In: *Science* 295.5560 (2002-03), pp. 1662–1664. DOI: 10.1126/science.1069492. (Visited on 2023-04-16) (cit. on p. 9).

[69] H. Kitano. "Computational Systems Biology". In: *Nature* 420.6912 (2002-11), pp. 206–210. ISSN: 1476-4687. DOI: 10.1038/nature01254. (Visited on 2023-04-16) (cit. on p. 9).

[70] C. Liu et al. "Computational Network Biology: Data, Models, and Applications". In: *Physics Reports*. Computational Network Biology: Data, Models, and Applications 846 (2020-03), pp. 1–66. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2019.12.004. (Visited on 2021-09-17) (cit. on pp. 9, 11, 28).

[71] M. Koutrouli et al. "A Guide to Conquer the Biological Network Era Using Graph Theory". In: *Frontiers in Bioengineering and Biotechnology* 8 (2020). ISSN: 2296-4185. (Visited on 2023-03-01) (cit. on p. 9).

[72] N. Masuda, M. A. Porter, and R. Lambiotte. "Random Walks and Diffusion on Networks". In: *Physics Reports*. Random Walks and Diffusion on Networks 716–717 (2017-11), pp. 1–58. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2017.07.007. (Visited on 2023-04-14) (cit. on p. 10).

[73] R. Albert. "Scale-Free Networks in Cell Biology". In: *Journal of Cell Science* 118.21 (2005-11), pp. 4947–4957. ISSN: 0021-9533. DOI: 10.1242/jcs.02714. (Visited on 2023-04-18) (cit. on p. 10).

[74] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. "Network Medicine: A Network-based Approach to Human Disease". In: *Nature reviews. Genetics* 12.1 (2011-01), pp. 56–68. ISSN: 1471-0056. DOI: 10.1038/nrg2918. (Visited on 2021-09-17) (cit. on pp. 11, 28).

[75] C. R. Harris et al. "Array Programming with NumPy". In: *Nature* 585.7825 (2020-09), pp. 357–362. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2649-2. (Visited on 2022-11-23) (cit. on p. 13).

[76] W. McKinney. "Data Structures for Statistical Computing in Python". In: *Python in Science Conference*. Austin, Texas, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a. (Visited on 2022-11-23) (cit. on p. 13).

[77] P. Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17.3 (2020-03), pp. 261–272. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0686-2. (Visited on 2022-11-23) (cit. on p. 13).

[78] S. Seabold and J. Perktold. "Statsmodels: Econometric and Statistical Modeling with Python". In: *Python in Science Conference*. Austin, Texas, 2010, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011. (Visited on 2022-11-23) (cit. on p. 13).

[79] F. Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928. (Visited on 2022-11-23) (cit. on p. 13).

[80] G. Csardi and T. Nepusz. "The Igraph Software Package for Complex Network Research". In: *InterJournal* Complex Systems (2005-11), p. 1695 (cit. on p. 13).

[81] *Plotly: Open Source Graphing Library for Python*. https://plotly.com/python/. (Visited on 2022-11-23) (cit. on p. 13).

[82] T. Wu et al. "clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data". In: *The Innovation* 2.3 (2021-08), p. 100141. ISSN: 2666-6758. DOI: 10.1016/j.xinn.2021.100141. (Visited on 2023-03-07) (cit. on p. 13).

[83] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. 13).

[84] THE GTEX CONSORTIUM. "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues". In: *Science* 369.6509 (2020-09), pp. 1318–1330. DOI: 10.1126/science.aaz1776. (Visited on 2021-11-22) (cit. on p. 13).

[85] B. L. Aken et al. "The Ensembl Gene Annotation System". In: *Database* 2016 (2016-01), baw093. ISSN: 1758-0463. DOI: 10.1093/database/baw093. (Visited on 2022-10-17) (cit. on p. 13).

[86] M. Uhlén et al. "Tissue-Based Map of the Human Proteome". In: *Science* 347.6220 (2015-01), p. 1260419. DOI: 10.1126/science.1260419. (Visited on 2022-10-17) (cit. on pp. 13, 14).

[87] A. Conesa et al. "A Survey of Best Practices for RNA-seq Data Analysis". In: *Genome Biology* 17.1 (2016-01), p. 13. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8. (Visited on 2022-10-27) (cit. on p. 14).

[88] M. D. Robinson and A. Oshlack. "A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data". In: *Genome Biology* 11.3 (2010-03), R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25. (Visited on 2022-10-27) (cit. on p. 14).

[89] Z. B. Abrams et al. "A Protocol to Evaluate RNA Sequencing Normalization Methods". In: *BMC Bioinformatics* 20.24 (2019-12), p. 679. ISSN: 1471-2105. DOI: 10.1186/s12859-019-3247-x. (Visited on 2022-10-27) (cit. on p. 14).

[90] S. Tweedie et al. "Genenames.Org: The HGNC and VGNC Resources in 2021". In: *Nucleic Acids Research* 49.D1 (2021-01), pp. D939–D946. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa980. (Visited on 2022-10-18) (cit. on p. 14).

[91] I. Yanai et al. "Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification". In: *Bioinformatics* 21.5 (2005-03), pp. 650–659. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti042. (Visited on 2022-09-02) (cit. on pp. 16, 30).

[92] Z. Dezső et al. "A Comprehensive Functional Analysis of Tissue Specificity of Human Gene Expression". In: *BMC Biology* 6.1 (2008-11), p. 49. ISSN: 1741-7007. DOI: 10.1186/1741-7007-6-49. (Visited on 2021-12-06) (cit. on pp. 16, 30, 31).

[93] E. W. Sayers et al. "Database Resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 49.D1 (2021-01), pp. D10–D17. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa892. (Visited on 2022-10-31) (cit. on pp. 16, 41).

[94] A. Zimek and P. Filzmoser. "There and Back Again: Outlier Detection between Statistical Reasoning and Data Mining Algorithms". In: *WIREs Data Mining and Knowledge Discovery* 8.6 (2018), e1280. ISSN: 1942-4795. DOI: 10.1002/widm.1280. (Visited on 2022-11-08) (cit. on p. 16).

[95] V. Chandola, A. Banerjee, and V. Kumar. "Anomaly Detection: A Survey". In: *ACM Computing Surveys* 41.3 (2009-07), 15:1–15:58. ISSN: 0360-0300. DOI: 10.1145/1541880.1541882. (Visited on 2022-11-08) (cit. on p. 16).

[96] P. J. Rousseeuw and M. Hubert. "Robust Statistics for Outlier Detection". In: *WIREs Data Mining and Knowledge Discovery* 1.1 (2011), pp. 73–79. ISSN: 1942-4795. DOI: 10.1002/widm.2. (Visited on 2022-11-08) (cit. on p. 16).

[97] *English: Boxplot and a Probability Density Function (Pdf) of a Normal N(0,1σ2) Population.* 2011-03. (Visited on 2022-11-07) (cit. on p. 17).

[98] S. Raschka and V. Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2*. Third edition. Expert Insight. Birmingham Mumbai: Packt, 2019. ISBN: 978-1-78995-575-0 (cit. on p. 16).

[99] A. Saxena et al. "A Review of Clustering Techniques and Developments". In: *Neurocomputing* 267 (2017-12), pp. 664–681. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2017.06.053. (Visited on 2022-11-10) (cit. on p. 17).

[100] R. Xu and D. Wunsch. "Survey of Clustering Algorithms". In: *IEEE Transactions on Neural Networks* 16.3 (2005-05), pp. 645–678. ISSN: 1941-0093. DOI: 10.1109/TNN.2005.845141 (cit. on p. 17).

[101] A. K. Jain. "Data Clustering: 50 Years beyond K-means". In: *Pattern Recognition Letters*. Award Winning Papers from the 19th International Conference on Pattern Recognition (ICPR) 31.8 (2010-06), pp. 651–666. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2009.09.011. (Visited on 2022-11-14) (cit. on p. 17).

[102] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. First. Cambridge University Press, 2014-05. ISBN: 978-1-107-05713-5. DOI: 10.1017/CBO9781107298019. (Visited on 2022-11-25) (cit. on p. 17).

[103] D. Comaniciu and P. Meer. "Mean Shift: A Robust Approach toward Feature Space Analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (2002-05), pp. 603–619. ISSN: 1939-3539. DOI: 10.1109/34.1000236 (cit. on p. 18).

[104] Y. Cheng. "Mean Shift, Mode Seeking, and Clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.8 (1995-08), pp. 790–799. ISSN: 1939-3539. DOI: 10.1109/34.400568 (cit. on p. 18).

[105] D. L. Davies and D. W. Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979-04), pp. 224–227. ISSN: 1939-3539. DOI: 10.1109/TPAMI.1979.4766909 (cit. on p. 18).

[106] G. Zheng et al. "HCMDB: The Human Cancer Metastasis Database". In: *Nucleic Acids Research* 46.D1 (2018-01), pp. D950–D955. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1008. (Visited on 2021-10-18) (cit. on p. 19).

[107] G. diSibio and S. W. French. "Metastatic Patterns of Cancers: Results From a Large Autopsy Study". In: *Archives of Pathology & Laboratory Medicine* 132.6 (2008-06), pp. 931–939. ISSN: 0003-9985. DOI: 10.5858/2008-132-931-MPOCRF. (Visited on 2021-11-03) (cit. on p. 19).

[108] I. Rivals et al. "Enrichment or Depletion of a GO Category within a Class of Genes: Which Test?" In: *Bioinformatics* 23.4 (2007-02), pp. 401–407. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl633. (Visited on 2022-11-30) (cit. on p. 20).

[109] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B. Methodological* 57.1 (1995), pp. 289–300. ISSN: 0035-9246. (Visited on 2022-11-28) (cit. on p. 20).

[110] D. Türei et al. "Integrated Intra- and Intercellular Signaling Knowledge for Multicellular Omics Analysis". In: *Molecular Systems Biology* 17.3 (2021-03), e9923. ISSN: 1744-4292. DOI: 10.15252/msb.20209923. (Visited on 2022-10-12) (cit. on p. 21).

[111] M. K. Türei Dénes. *Omnipath: Python Client for the OmniPath Web Service*. (Visited on 2022-12-20) (cit. on p. 21).

[112] G. A. Pavlopoulos et al. "Bipartite Graphs in Systems Biology and Medicine: A Survey of Methods and Applications". In: *GigaScience* 7.4 (2018-04), giy014. ISSN: 2047-217X. DOI: 10.1093/gigascience/giy014. (Visited on 2022-10-25) (cit. on p. 22).

[113] N. C. Chung et al. "Jaccard/Tanimoto Similarity Test and Estimation Methods for Biological Presence-Absence Data". In: *BMC Bioinformatics* 20.15 (2019-12), p. 644. ISSN: 1471-2105. DOI: 10.1186/s12859-019-3118-5. (Visited on 2023-01-10) (cit. on p. 23).

[114] L. Choi, J. D. Blume, and W. D. Dupont. "Elucidating the Foundations of Statistical Inference with 2 x 2 Tables". In: *PLOS ONE* 10.4 (2015-04), e0121263. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0121263. (Visited on 2023-01-13) (cit. on p. 24).

[115] H. B. Mann and D. R. Whitney. "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1 (1947-03), pp. 50–60. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177730491. (Visited on 2023-01-27) (cit. on p. 25).

[116] G. W. Divine et al. "The Wilcoxon–Mann–Whitney Procedure Fails as a Test of Medians". In: *The American Statistician* 72.3 (2018-07), pp. 278–286. ISSN: 0003-1305. DOI: 10.1080/00031305.2017.1305291. (Visited on 2023-01-27) (cit. on p. 25).

[117] W. Mendenhall, R. J. Beaver, and B. M. Beaver. *Introduction to Probabilities and Statistics*. Thirteenth. Belmont, CA: Brooks/Cole, Cengage Learning, 2009 (cit. on p. 25).

[118] "Kolmogorov–Smirnov Test". In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer, 2008, pp. 283–287. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_214. (Visited on 2023-02-09) (cit. on p. 25).

[119] J. Unpingco. *Python for Probability, Statistics, and Machine Learning*. Cham: Springer International Publishing, 2022. ISBN: 978-3-031-04648-3. DOI: 10.1007/978-3-031-04648-3. (Visited on 2023-01-30) (cit. on p. 26).

[120] L. Garcia-Alonso et al. "Benchmark and Integration of Resources for the Estimation of Human Transcription Factor Activities". In: *Genome Research* 29.8 (2019-01), pp. 1363–1375. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.240663.118. (Visited on 2023-02-20) (cit. on p. 28).

[121] S. Köhler et al. "Walking the Interactome for Prioritization of Candidate Disease Genes". In: *The American Journal of Human Genetics* 82.4 (2008-04), pp. 949–958. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2008.02.013. (Visited on 2023-02-14) (cit. on p. 28).

[122] B. Phipson and G. K. Smyth. "Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn". In: *Statistical Applications in Genetics and Molecular Biology* 9.1 (2010-10). ISSN: 1544-6115. DOI: 10.2202/1544-6115.1585. (Visited on 2023-02-16) (cit. on p. 29).

[123] Y. Liu, A. Beyer, and R. Aebersold. "On the Dependency of Cellular Protein Levels on mRNA Abundance". In: *Cell* 165.3 (2016-04), pp. 535–550. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.03.014. (Visited on 2022-12-08) (cit. on p. 30).

[124] C. Buccitelli and M. Selbach. "mRNAs, Proteins and the Emerging Principles of Gene Expression Control". In: *Nature Reviews Genetics* 21.10 (2020-10), pp. 630–644. ISSN: 1471-0064. DOI: 10.1038/s41576-020-0258-4. (Visited on 2022-04-13) (cit. on p. 30).

[125] E. Eisenberg and E. Y. Levanon. "Human Housekeeping Genes, Revisited". In: *Trends in Genetics* 29.10 (2013-10), pp. 569–574. ISSN: 0168-9525. DOI: 10.1016/j.tig.2013.05.010. (Visited on 2022-01-25) (cit. on p. 30).

[126] P. Kapranov, A. T. Willingham, and T. R. Gingeras. "Genome-Wide Transcription and the Implications for Genomic Organization". In: *Nature Reviews Genetics* 8.6 (2007-06), pp. 413–423. ISSN: 1471-0064. DOI: 10.1038/nrg2083. (Visited on 2022-12-07) (cit. on p. 30).

[127] N. Kryuchkova-Mostacci and M. Robinson-Rechavi. "A Benchmark of Gene Expression Tissue-Specificity Metrics". In: *Briefings in Bioinformatics* 18.2 (2017-03), pp. 205–214. ISSN: 1467-5463. DOI: 10.1093/bib/bbw008. (Visited on 2021-11-02) (cit. on p. 30).

[128] W.-C. Huang et al. "*B*2-Microglobulin Is a Signaling and Growth-Promoting Factor for Human Prostate Cancer Bone Metastasis". In: *Cancer Research* 66.18 (2006-09), pp. 9108–9116. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-06-1996. (Visited on 2023-02-10) (cit. on p. 40).

[129] L. Borst, S. H. van der Burg, and T. van Hall. "The NKG2A–HLA-E Axis as a Novel Checkpoint in the Tumor Microenvironment". In: *Clinical Cancer Research* 26.21 (2020-11), pp. 5549–5556. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-19-2095. (Visited on 2023-02-10) (cit. on pp. 40, 50, 52).

[130] P. J. A. Cock et al. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics". In: *Bioinformatics* 25.11 (2009-06), pp. 1422–1423. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp163. (Visited on 2023-02-07) (cit. on p. 41).

[131] J. Piñero et al. "The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update". In: *Nucleic Acids Research* 48.D1 (2020-01), pp. D845–D855. ISSN: 0305-1048. DOI: 10.1093/nar/gkz1021. (Visited on 2023-02-07) (cit. on p. 41).

[132] D. Ochoa et al. "The Next-Generation Open Targets Platform: Reimagined, Redesigned, Rebuilt". In: *Nucleic Acids Research* 51.D1 (2023-01), pp. D1353–D1359. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1046. (Visited on 2023-02-16) (cit. on p. 43).

[133] L. Dressler et al. "Comparative Assessment of Genes Driving Cancer and Somatic Evolution in Non-Cancer Tissues: An Update of the Network of Cancer Genes (NCG) Resource". In: *Genome Biology* 23.1 (2022-01), p. 35. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02607-z. (Visited on 2023-02-23) (cit. on p. 45).

[134] Y. Chen, F. J. Verbeek, and K. Wolstencroft. "Establishing a Consensus for the Hallmarks of Cancer Based on Gene Ontology and Pathway Annotations". In: *BMC Bioinformatics* 22.1 (2021-04), p. 178. ISSN: 1471-2105. DOI: 10.1186/s12859-021-04105-8. (Visited on 2023-03-09) (cit. on p. 47).

[135] K. Kamińska et al. "The Role of the Cell–Cell Interactions in Cancer Progression". In: *Journal of Cellular and Molecular Medicine* 19.2 (2015), pp. 283–296. ISSN: 1582-4934. DOI: 10.1111/jcmm.12408. (Visited on 2022-05-25) (cit. on p. 50).

[136] R. R. Langley and I. J. Fidler. "Tumor Cell-Organ Microenvironment Interactions in the Pathogenesis of Cancer Metastasis". In: *Endocrine Reviews* 28.3 (2007-05), pp. 297–321. ISSN: 0163-769X. DOI: 10.1210/er.2006-0027. (Visited on 2022-05-25) (cit. on p. 50).

[137] N. K. Wolf, D. U. Kissiov, and D. H. Raulet. "Roles of Natural Killer Cells in Immunity to Cancer, and Applications to Immunotherapy". In: *Nature Reviews Immunology* 23.2 (2023-02), pp. 90–105. ISSN: 1474-1741. DOI: 10.1038/s41577-022-00732-1. (Visited on 2023-02-10) (cit. on pp. 50, 51).

[138] M. Sheffer et al. "Genome-Scale Screens Identify Factors Regulating Tumor Cell Responses to Natural Killer Cells". In: *Nature Genetics* 53.8 (2021-08), pp. 1196–1206. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00889-w. (Visited on 2023-02-11) (cit. on p. 50).

[139] T. Kamiya et al. "Blocking Expression of Inhibitory Receptor NKG2A Overcomes Tumor Resistance to NK Cells". In: *The Journal of Clinical Investigation* 129.5 (2019-05), pp. 2094–2106. ISSN: 0021-9738. DOI: 10.1172/JCI123955. (Visited on 2023-02-10) (cit. on p. 50).

[140] T. van Hall et al. "Monalizumab: Inhibiting the Novel Immune Checkpoint NKG2A". In: *Journal for ImmunoTherapy of Cancer* 7.1 (2019-10), p. 263. ISSN: 2051-1426. DOI: 10.1186/s40425-019-0761-3. (Visited on 2023-05-02) (cit. on p. 51).

[141] A. Labernadie et al. "A Mechanically Active Heterotypic E-cadherin/N-cadherin Adhesion Enables Fibroblasts to Drive Cancer Cell Invasion". In: *Nature Cell Biology* 19.3 (2017-03), pp. 224–237. ISSN: 1476-4679. DOI: 10.1038/ncb3478. (Visited on 2023-03-30) (cit. on p. 51).

[142] H. Wang et al. "The Osteogenic Niche Promotes Early-Stage Bone Colonization of Disseminated Breast Cancer Cells". In: *Cancer Cell* 27.2 (2015-02), pp. 193–210. ISSN: 1535-6108. DOI: 10.1016/j.ccell.2014.11.017. (Visited on 2023-05-03) (cit. on p. 51).

[143] C. Trapnell. "Defining Cell Types and States with Single-Cell Genomics". In: *Genome Research* 25.10 (2015-01), pp. 1491–1498. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.190595.115. (Visited on 2023-04-10) (cit. on p. 51).

[144] T. Baslan and J. Hicks. "Unravelling Biology and Shifting Paradigms in Cancer with Single-Cell Sequencing". In: *Nature Reviews Cancer* 17.9 (2017-09), pp. 557–569. ISSN: 1474-1768. DOI: 10.1038/nrc.2017.58. (Visited on 2023-05-05) (cit. on p. 51).

[145] O. Stegle, S. A. Teichmann, and J. C. Marioni. "Computational and Analytical Challenges in Single-Cell Transcriptomics". In: *Nature Reviews Genetics* 16.3 (2015-03), pp. 133–145. ISSN: 1471-0064. DOI: 10.1038/nrg3833. (Visited on 2023-05-07) (cit. on p. 51).

[146] F. B. Bastian et al. "The Bgee Suite: Integrated Curated Expression Atlas and Comparative Transcriptomics in Animals". In: *Nucleic Acids Research* 49.D1 (2021-01), pp. D831–D847. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa793. (Visited on 2023-05-04) (cit. on p. 52).

[147] T. Else et al. "Adrenocortical Carcinoma". In: *Endocrine Reviews* 35.2 (2014-04), pp. 282–326. ISSN: 0163-769X. DOI: 10.1210/er.2013-1029. (Visited on 2021-11-15) (cit. on p. 91).

[148] A. B. Shinagare et al. "Metastatic Pattern of Bladder Cancer: Correlation With the Characteristics of the Primary Tumor". In: *American Journal of Roentgenology* 196.1 (2011-01), pp. 117–122. ISSN: 0361-803X. DOI: 10.2214/AJR.10.5036. (Visited on 2022-04-11) (cit. on p. 91).

[149] J. L. Ferguson and S. P. Turner. "Bone Cancer: Diagnosis and Treatment Principles". In: *American Family Physician* 98.4 (2018-08), pp. 205–213. ISSN: 0002-838X, 1532-0650. (Visited on 2021-11-15) (cit. on p. 91).

[150] T. T. Lah, M. Novak, and B. Breznik. "Brain Malignancies: Glioblastoma and Brain Metastases". In: *Seminars in Cancer Biology*. Enigmatic Tumor Properties Associated with Metastatic Spread 60 (2020-02), pp. 262–273. ISSN: 1044-579X. DOI: 10.1016/j.semcancer.2019.10.010. (Visited on 2022-05-04) (cit. on p. 91).

[151] W. Chen et al. "Organotropism: New Insights into Molecular Mechanisms of Breast Cancer Metastasis". In: *npj Precision Oncology* 2.1 (2018-02), pp. 1–12. ISSN: 2397-768X. DOI: 10.1038/s41698-018-0047-0. (Visited on 2021-09-17) (cit. on p. 91).

[152] H. Li, X. Wu, and X. Cheng. "Advances in Diagnosis and Treatment of Metastatic Cervical Cancer". In: *Journal of Gynecologic Oncology* 27.4 (2016-07). DOI: 10.3802/jgo.2016.27.e43. (Visited on 2021-11-16) (cit. on p. 91).

[153] N. Hugen et al. "Metastatic Pattern in Colorectal Cancer Is Strongly Influenced by Histological Subtype". In: *Annals of Oncology* 25.3 (2014-03), pp. 651–657. ISSN: 0923-7534. DOI: 10.1093/annonc/mdt591. (Visited on 2021-11-12) (cit. on p. 91).

[154] A. Maheshwari and P. T. Finger. "Cancers of the Eye". In: *Cancer and Metastasis Reviews* 37.4 (2018-12), pp. 677–690. ISSN: 1573-7233. DOI: 10.1007/s10555-018-9762-9. (Visited on 2022-05-04) (cit. on p. 91).

[155] R. Hundal and E. A. Shaffer. "Gallbladder Cancer: Epidemiology and Outcome". In: *Clinical Epidemiology* 6 (2014), p. 99. DOI: 10.2147/CLEP.S37357. (Visited on 2021-11-16) (cit. on p. 91).

[156] R. J. Motzer et al. "Kidney Cancer, Version 2.2017, NCCN Clinical Practice Guidelines in Oncology". In: *Journal of the National Comprehensive Cancer Network* 15.6 (2017-06), pp. 804–834. ISSN: 1540-1405, 1540-1413. DOI: 10.6004/jnccn.2017.0100. (Visited on 2021-11-16) (cit. on p. 91).

[157] K. Uchino et al. "Hepatocellular Carcinoma with Extrahepatic Metastasis". In: *Cancer* 117.19 (2011), pp. 4475–4483. ISSN: 1097-0142. DOI: 10.1002/cncr.25960. (Visited on 2022-05-04) (cit. on p. 91).

[158] M. Riihimäki et al. "Metastatic Sites and Survival in Lung Cancer". In: *Lung Cancer* 86.1 (2014-10), pp. 78–84. ISSN: 0169-5002. DOI: 10.1016/j.lungcan.2014.07.020. (Visited on 2021-11-03) (cit. on p. 91).

[159] S.-G. Wu et al. "Sites of Metastasis and Overall Survival in Esophageal Cancer: A Population-Based Study". In: *Cancer Management and Research* 9 (2017-12), pp. 781–788. DOI: 10.2147/CMAR.S150350. (Visited on 2022-05-24) (cit. on p. 91).

[160] M. Blastik, É. Plavecz, and A. Zalatnai. "Pancreatic Carcinomas in a 60-Year, Institute-Based Autopsy Material With Special Emphasis of Metastatic Pattern". In: *Pancreas* 40.3 (2011-04), pp. 478–480. ISSN: 0885-3177. DOI: 10.1097/MPA.0b013e318205e332. (Visited on 2022-05-05) (cit. on p. 92).

[161] L. Bubendorf et al. "Metastatic Patterns of Prostate Cancer: An Autopsy Study of 1,589 Patients". In: *Human Pathology* 31.5 (2000-05), pp. 578–583. ISSN: 0046-8177. DOI: 10.1053/hp.2000.6698. (Visited on 2022-05-05) (cit. on p. 92).

[162] A. F. Jerant et al. "Early Detection and Treatment of Skin Cancer". In: *American Family Physician* 62.2 (2000-07), pp. 357–368. (Visited on 2022-05-30) (cit. on p. 92).

[163] I. J. Fidler. "Melanoma Metastasis". In: *Cancer Control* 2.5 (1995-09), pp. 398–404. ISSN: 1073-2748. DOI: 10.1177/107327489500200503. (Visited on 2022-05-30) (cit. on p. 92).

[164] M. Riihimäki et al. "Metastatic Spread in Patients with Gastric Cancer". In: *Oncotarget* 7.32 (2016-07), pp. 52307–52316. ISSN: 1949-2553. DOI: 10.18632/oncotarget.10740. (Visited on 2022-05-05) (cit. on p. 92).

[165] J. Shaw. "Diagnosis and Treatment of Testicular Cancer". In: *American Family Physician* 77.4 (2008-02), pp. 469–474. (Visited on 2022-12-16) (cit. on p. 92).

[166] M. E. Cabanillas, D. G. McFadden, and C. Durante. "Thyroid Cancer". In: *The Lancet* 388.10061 (2016-12), pp. 2783–2795. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(16)30172-6. (Visited on 2021-11-16) (cit. on p. 92).

[167] F. Duprez et al. "Distant Metastases in Head and Neck Cancer". In: *Head & Neck* 39.9 (2017), pp. 1733–1743. ISSN: 1097-0347. DOI: 10.1002/hed.24687. (Visited on 2022-06-13) (cit. on p. 92).

[168] J. C. Liu et al. "Patterns of Distant Metastasis in Head and Neck Cancer at Presentation: Implications for Initial Evaluation". In: *Oral Oncology* 88 (2019-01), pp. 131–136. ISSN: 1368-8375. DOI: 10.1016/j.oraloncology.2018.11.023. (Visited on 2022-06-13) (cit. on p. 92).

[169] S. A. Sohaib et al. "Recurrent Endometrial Cancer: Patterns of Recurrent Disease and Assessment of Prognosis". In: *Clinical Radiology* 62.1 (2007-01), pp. 28–34. ISSN: 0009-9260. DOI: 10.1016/j.crad.2006.06.015. (Visited on 2022-12-16) (cit. on p. 92).

A

Supplementary Figures

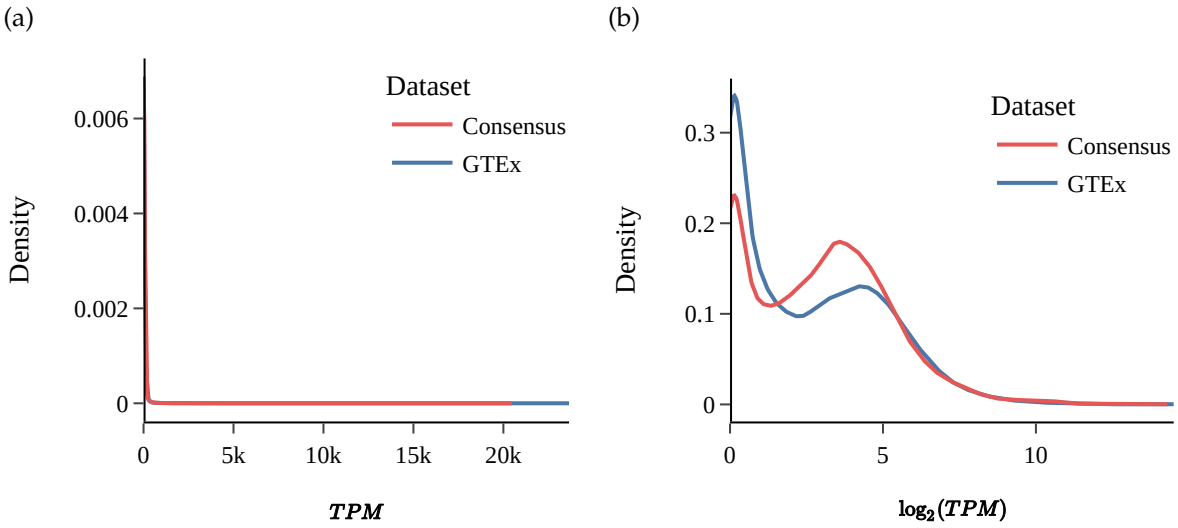


Figure A.1: KDE plot of the distribution of gene expression values in *Lung* before (a) and (b) after $\log_2$ transformation.

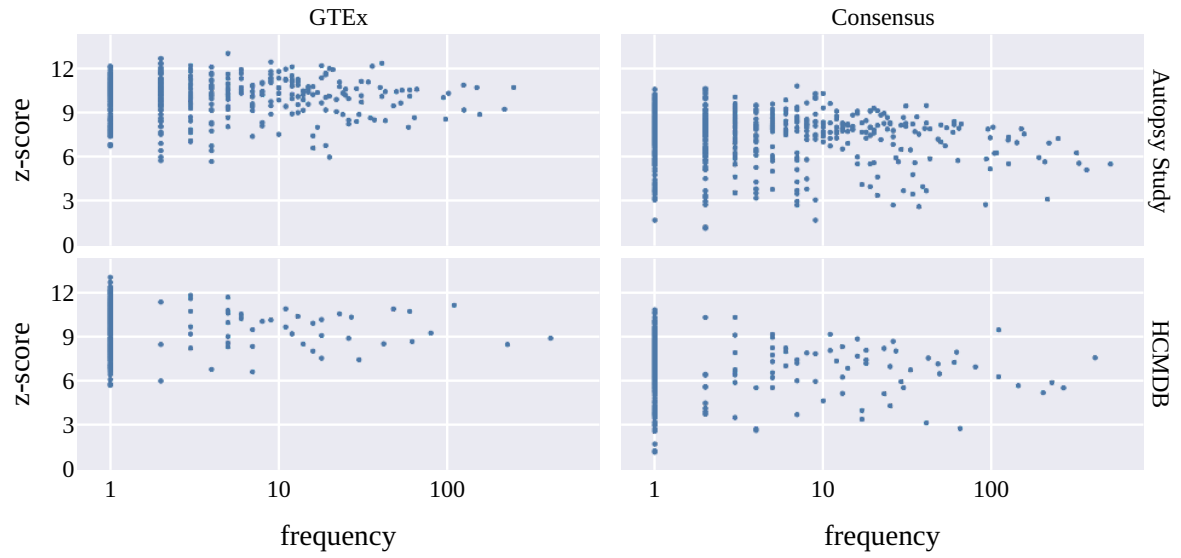


Figure A.2: Cancer-wise analysis of metastatic patterns for **undirected curated networks built with gene expression calls**. Relationship between the z-score and the frequency of metastasis in log scale. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair).

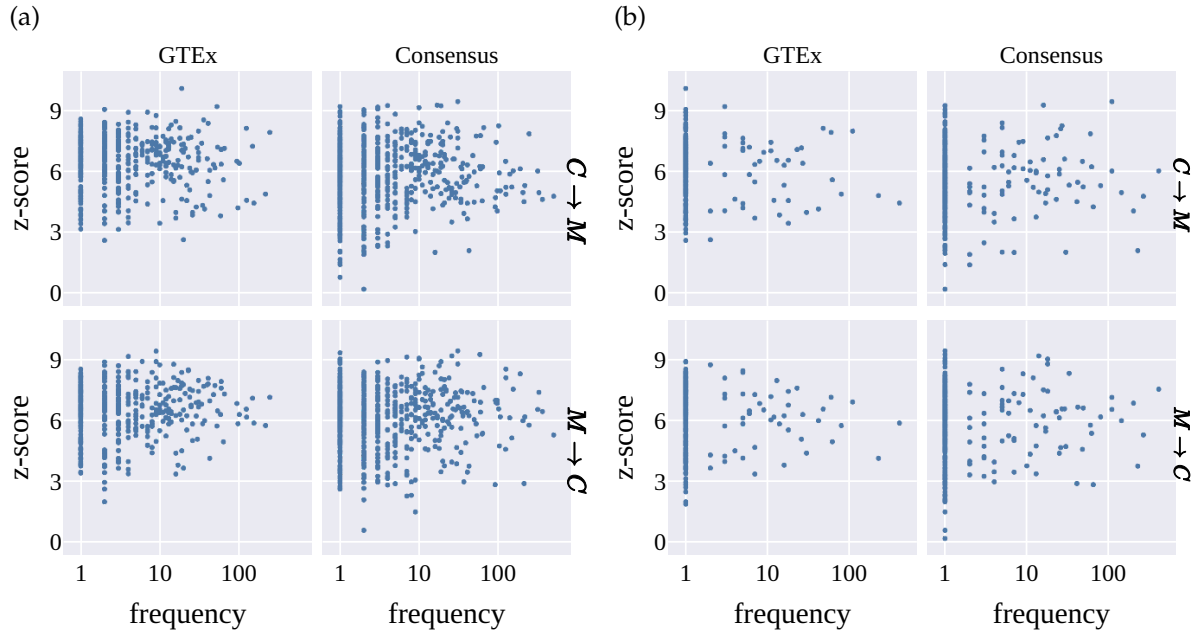


Figure A.3: Cancer-wise analysis of metastatic patterns for **directed networks built with gene expression calls** in (a) Autopsy Study and (b) HCMDB. Relationship between the z-score and the frequency of metastasis in log scale. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

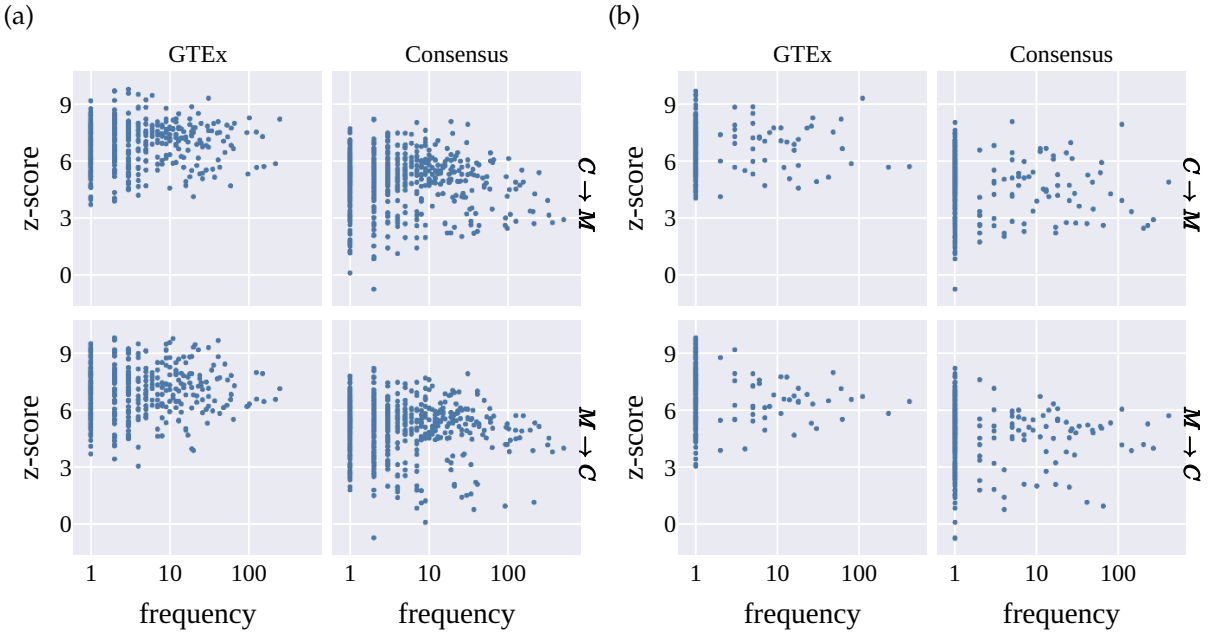


Figure A.4: Cancer-wise analysis of metastatic patterns for **directed curated networks built with gene expression calls** in (a) Autopsy Study and (b) HCMDB. Relationship between the z-score and the frequency of metastasis in log scale. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

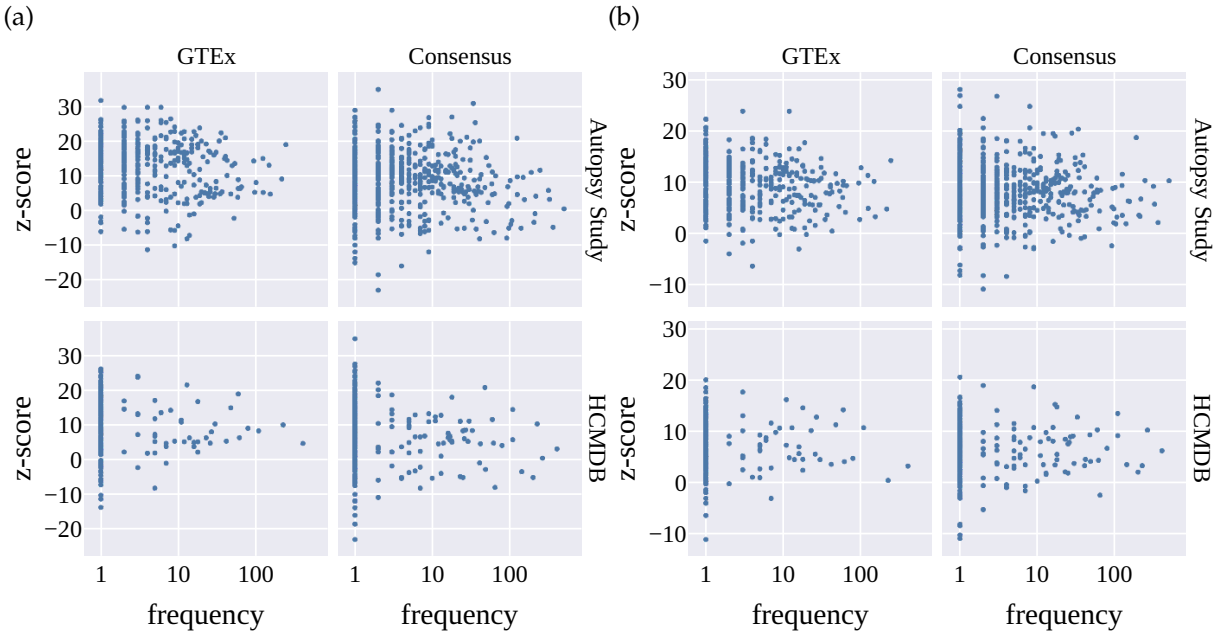


Figure A.5: Cancer-wise analysis of metastatic patterns for **undirected networks built with gene weights** using (a) complete graph and (b) curated graph. Relationship between the z-score and the frequency of metastasis in log scale. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair).

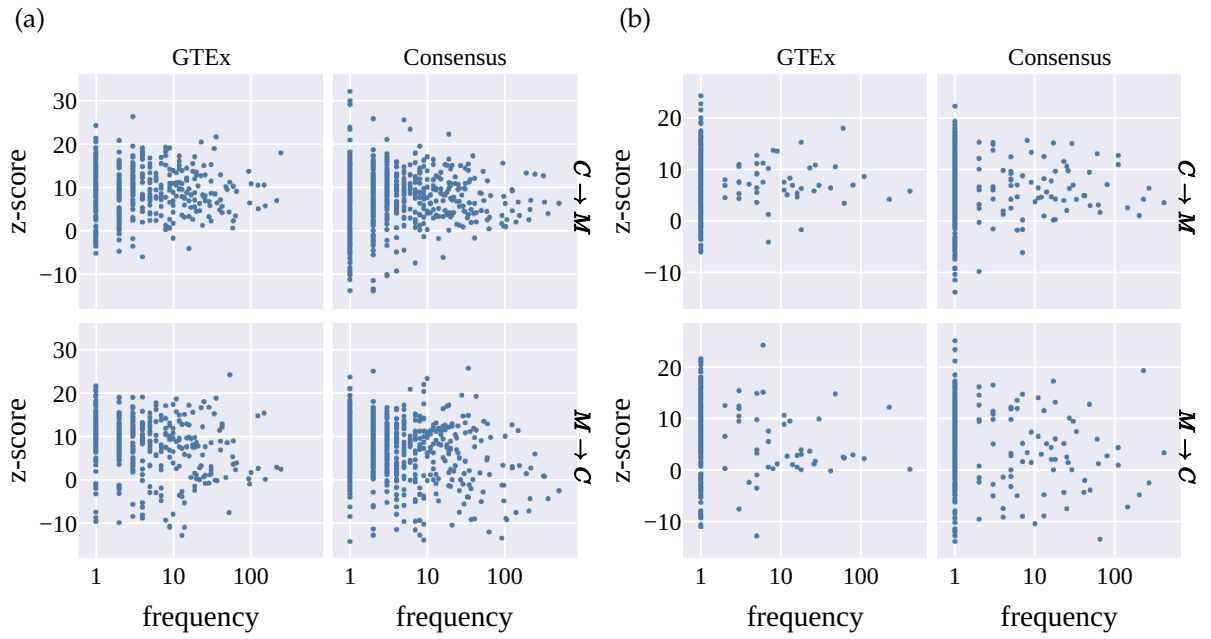


Figure A.6: Cancer-wise analysis of metastatic patterns for **directed networks built with gene weights** in (a) Autopsy Study and (b) HCMDB. Relationship between the z-score and the frequency of metastasis in log scale. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

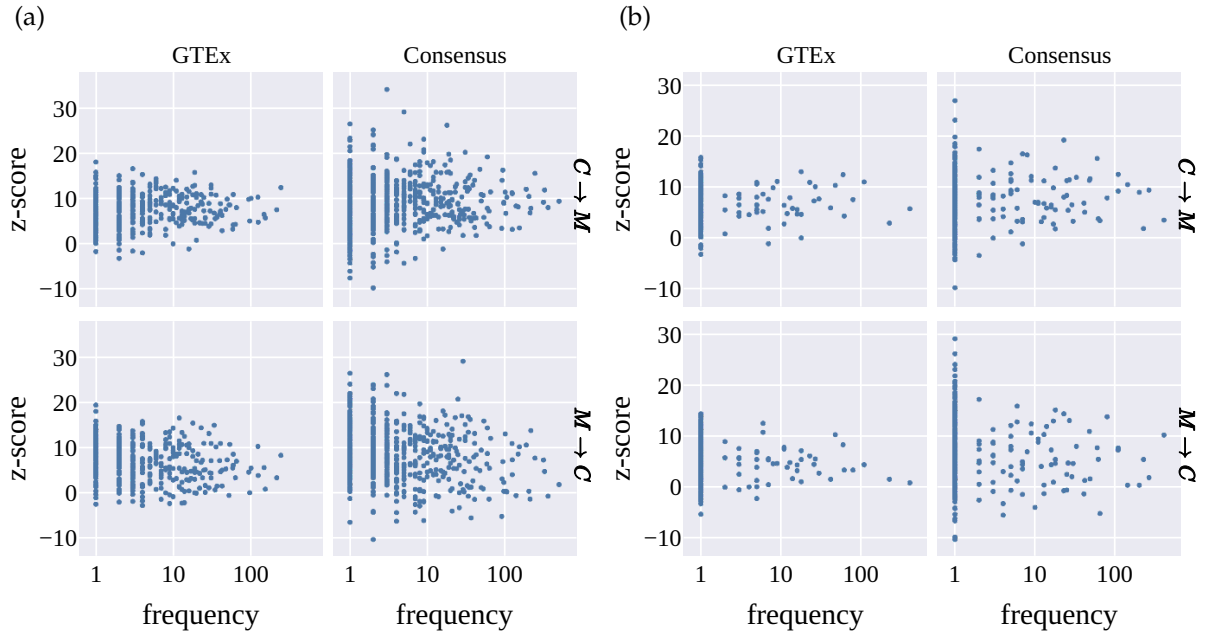


Figure A.7: Cancer-wise analysis of metastatic patterns for **directed curated networks built with gene weights** in (a) Autopsy Study and (b) HCMDB. Relationship between the z-score and the frequency of metastasis in log scale. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

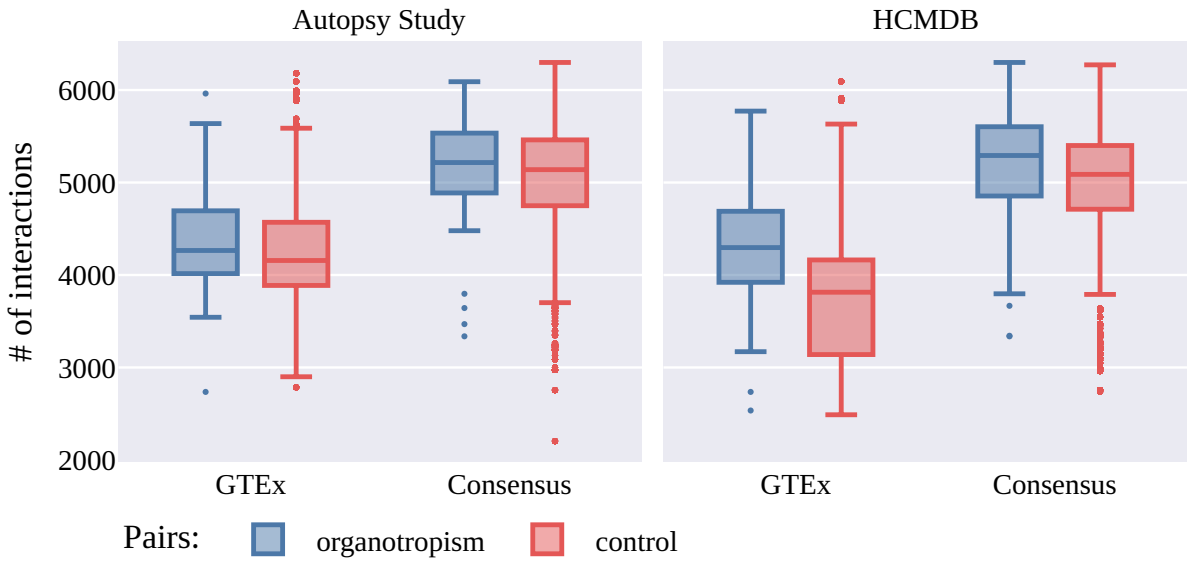


Figure A.8: Controlled comparison between organotropism pairs and control pairs for **undirected curated networks built with gene expression calls**. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). Cell-cell communication evaluated using the number of interactions.

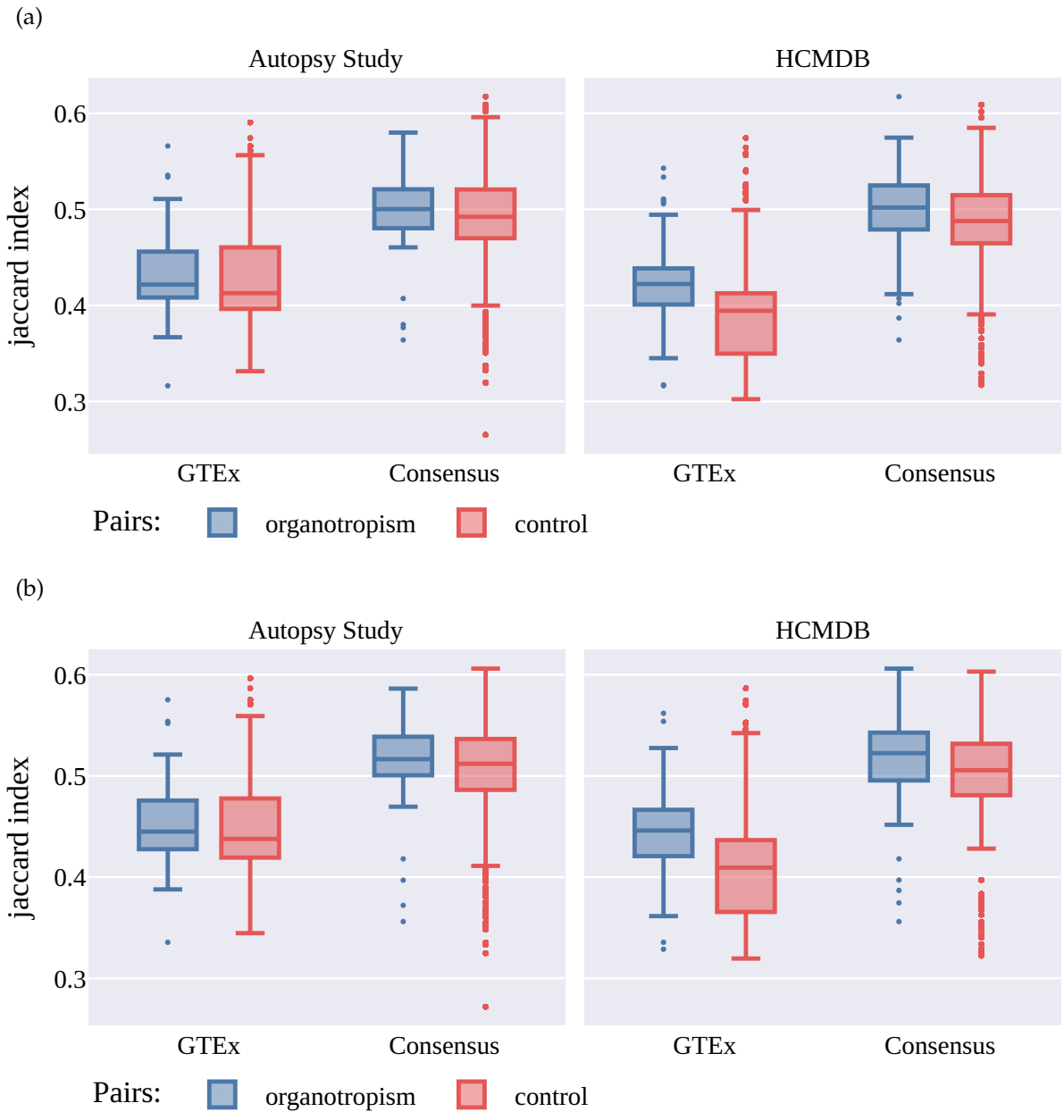


Figure A.9: Controlled comparison between organotropism pairs and control pairs for **undirected networks built with gene expression calls** using (a) complete graph; (b) curated graph. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). Cell-cell communication evaluated using the jaccard index.

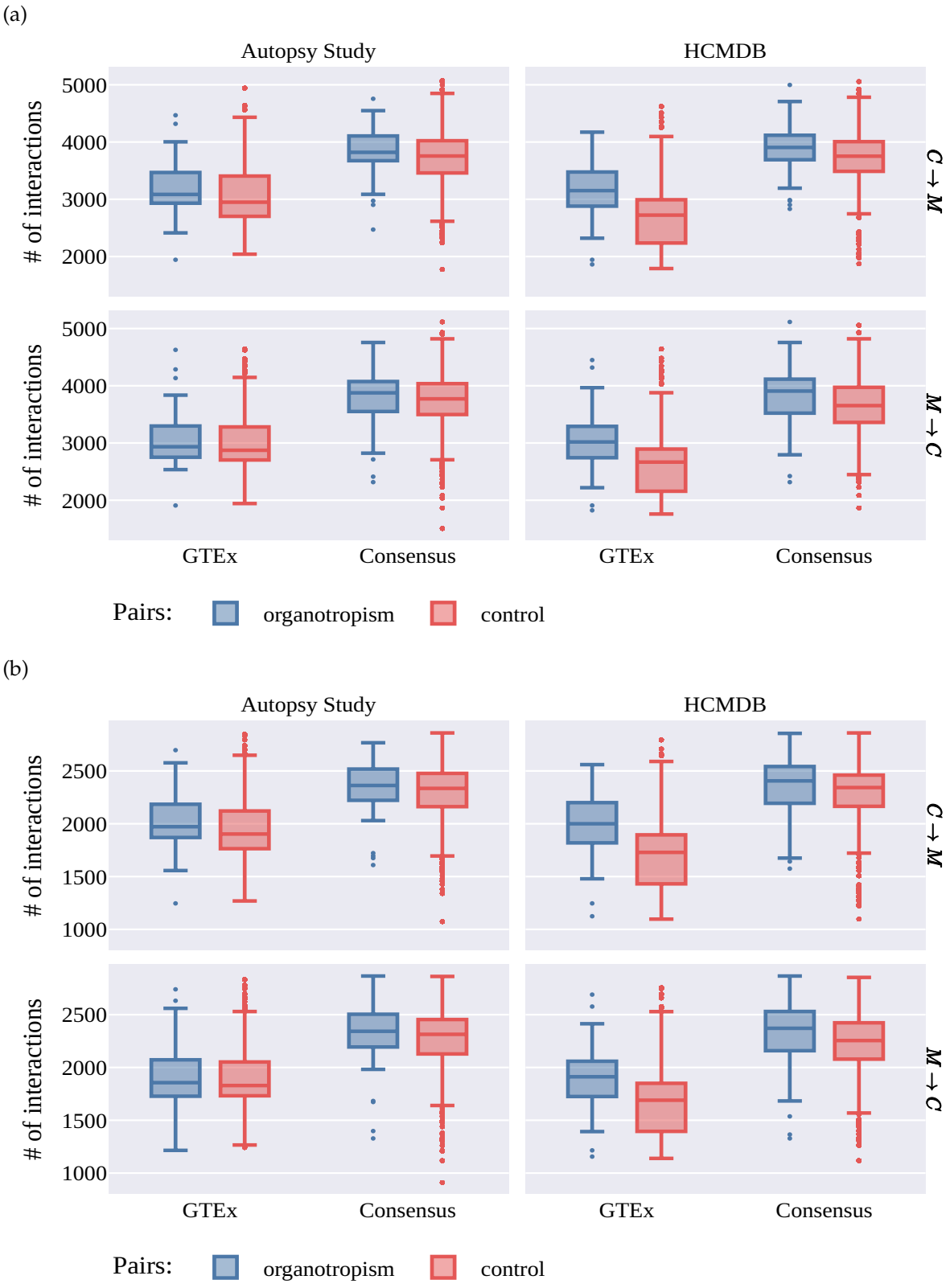


Figure A.10: Controlled comparison between organotropism pairs and control pairs for **directed networks built with gene expression calls** using (a) complete graph; (b) curated graph. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). Cell-cell communication evaluated using the number of interactions. **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

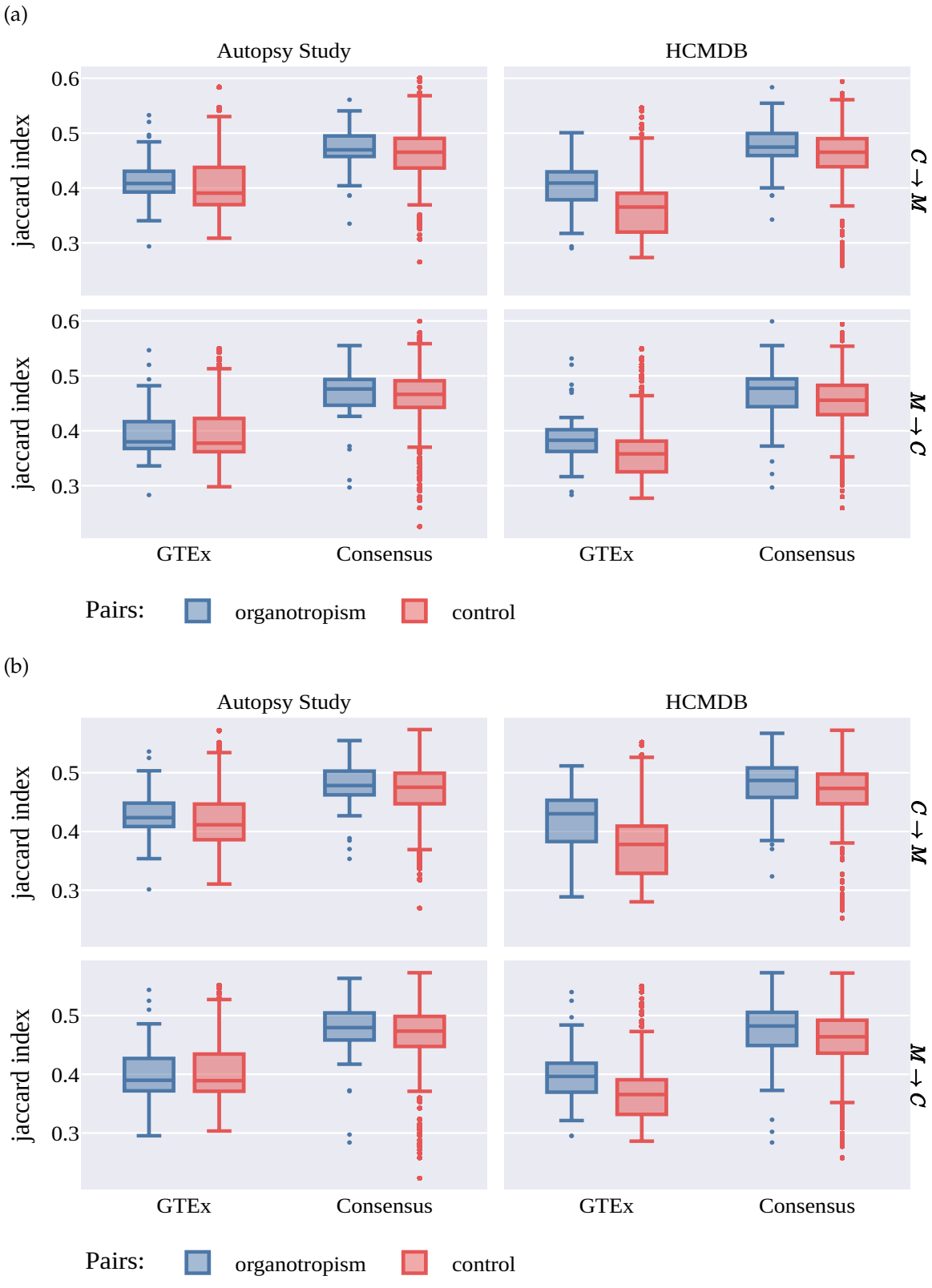


Figure A.11: Controlled comparison between organotropism pairs and control pairs for **directed networks built with gene expression calls** using (a) complete graph; (b) curated graph. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). Cell-cell communication evaluated using the jaccard index. **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

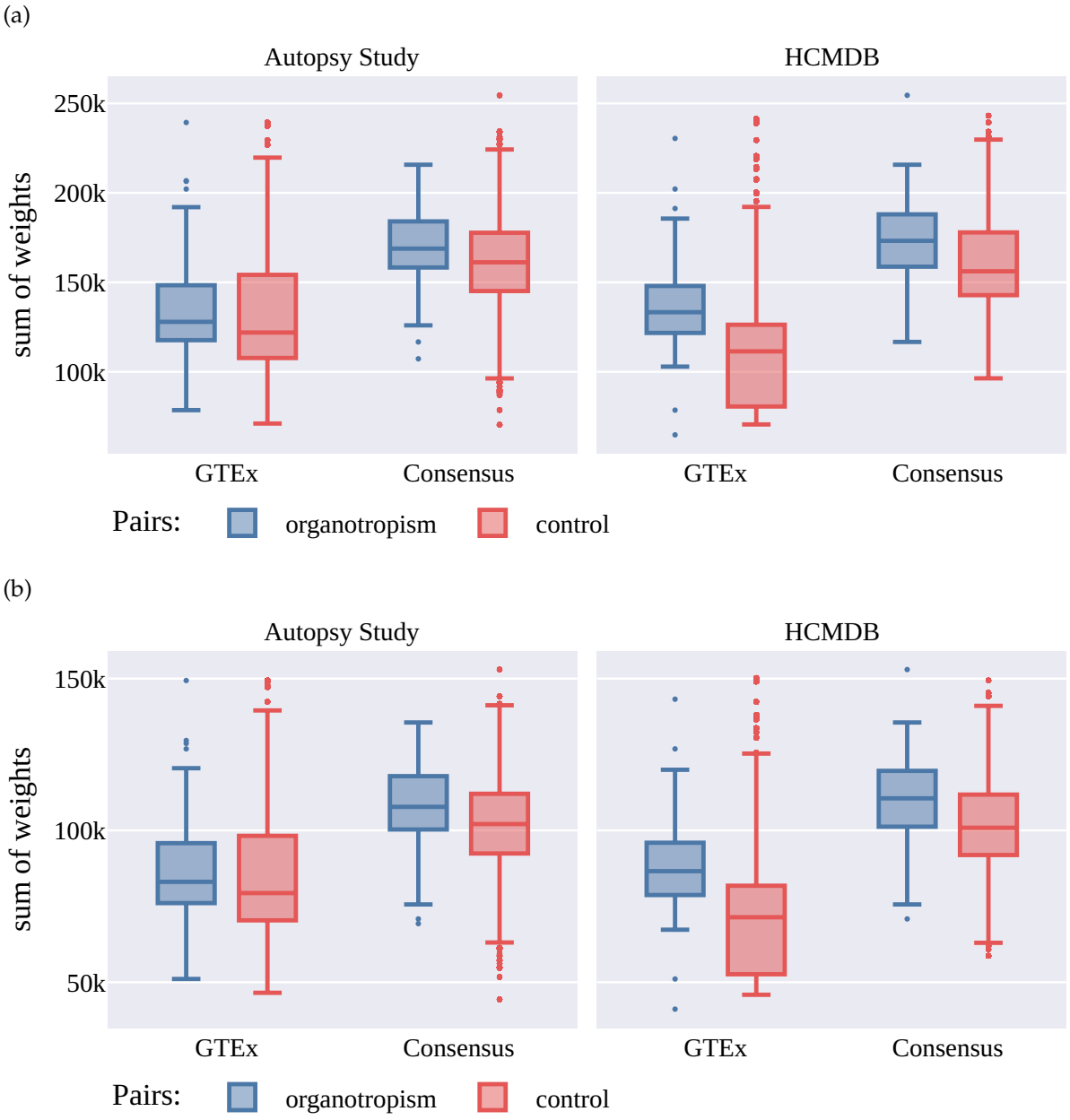


Figure A.12: Controlled comparison between organotropism pairs and control pairs for **undirected networks built with gene weights** using (a) complete graph; (b) curated graph. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair).

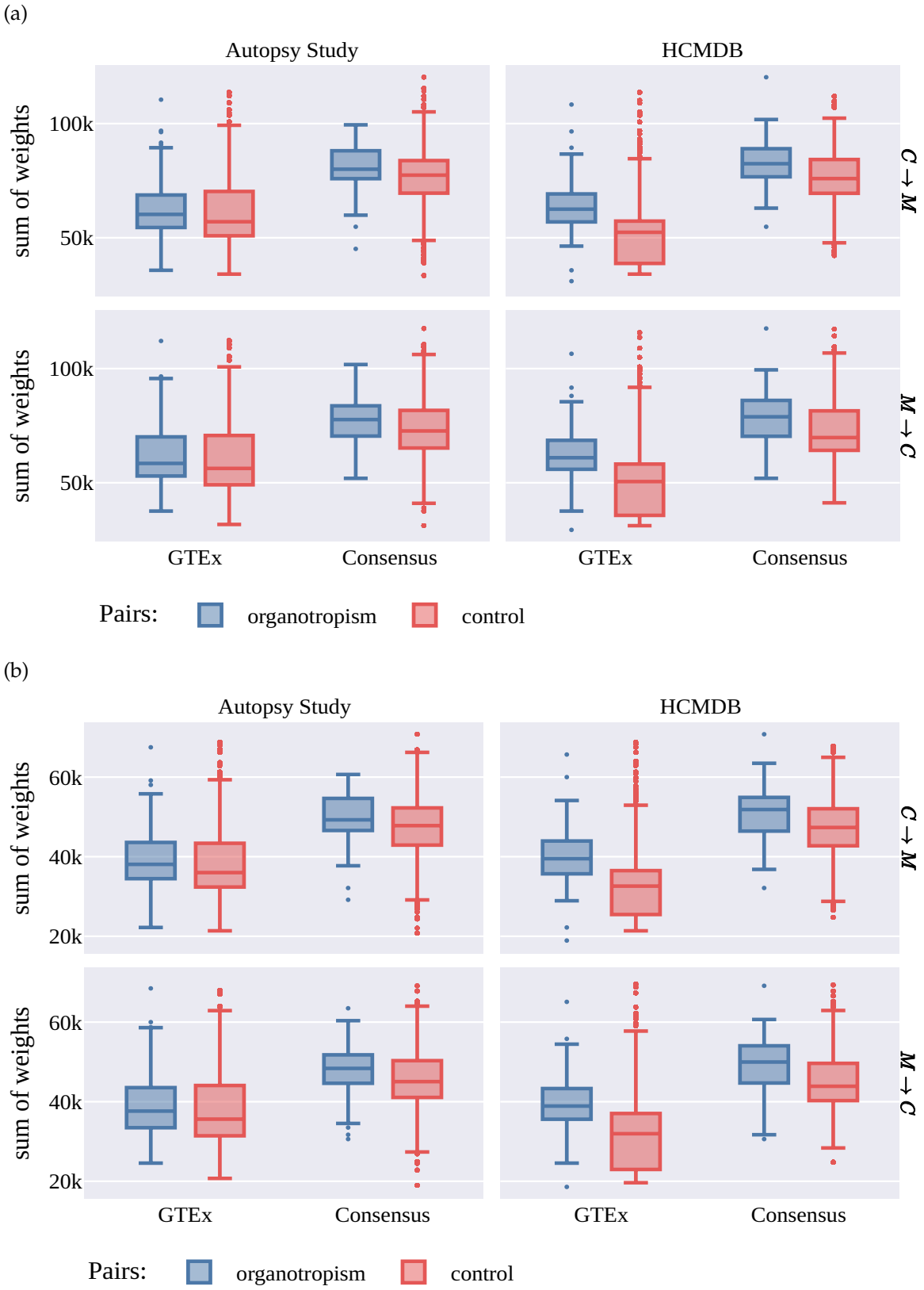


Figure A.13: Controlled comparison between organotropism pairs and control pairs for **directed networks built with gene weights** using (a) complete graph; (b) curated graph. Each data point represents an intercellular PPI networks (cancer–metastasis tissue pair). **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

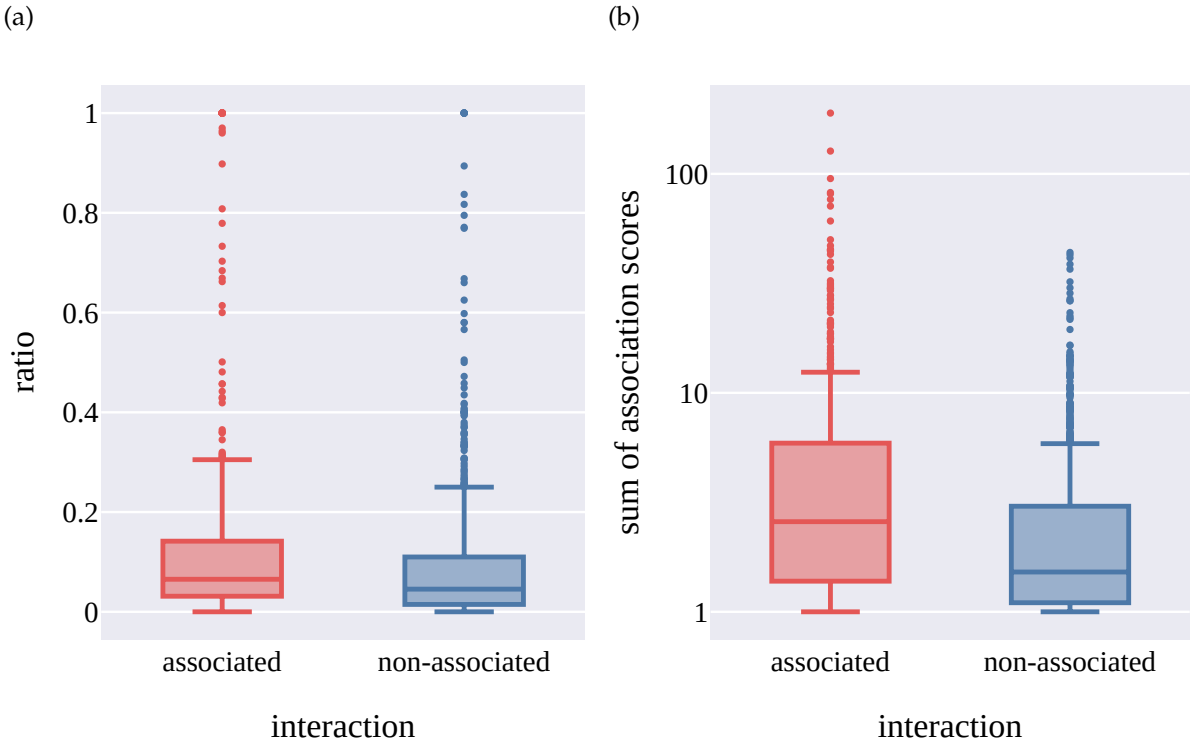


Figure A.14: Prior known connections with cancer or metastasis in genes that participate in metastasis-associated interactions. (a) PubMed search for titles and abstracts containing the query: "(Gene) AND (metastasis OR invasion)". Distribution of the ratio between of the number of PubMed IDs matching the queried term and the total number of PubMed IDs that mention each gene. (b) DisGeNET association with the *Neoplasm* disease class. Distribution in log scale of the sum of association scores for each gene.

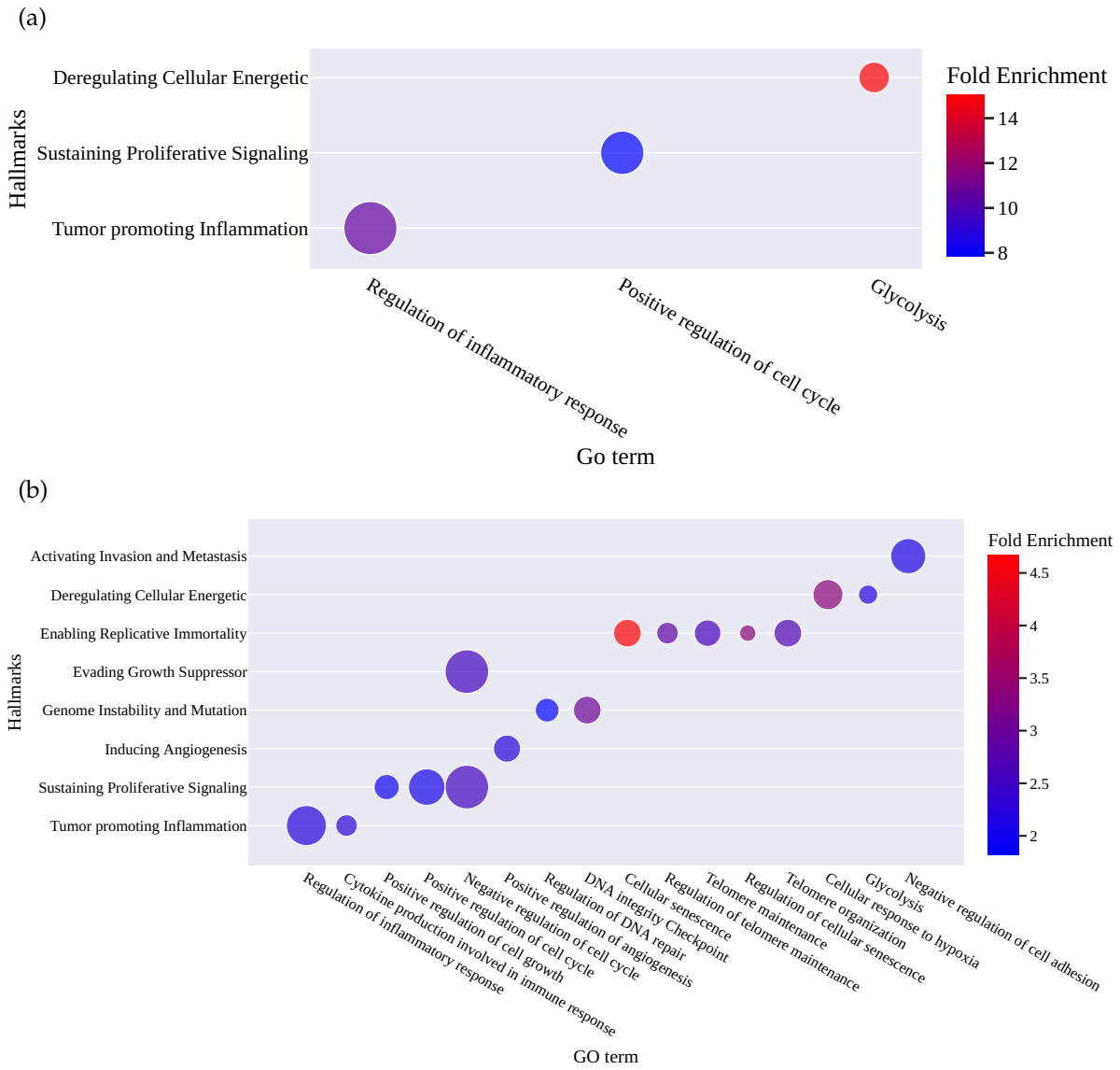
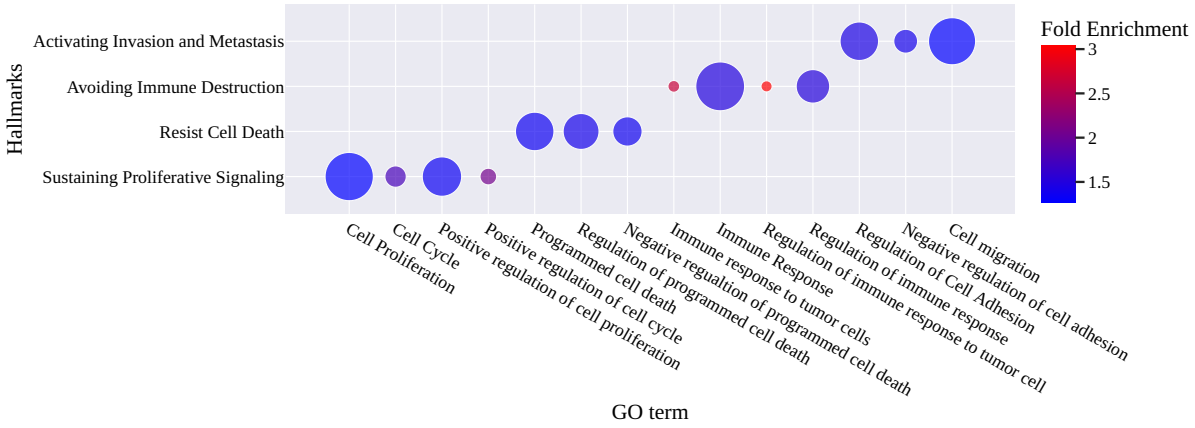


Figure A.15: GO terms Hallmarks enrichment for intracellular genes linked to intercellular interactions (complete graph). (a) source graph; (b) target graph. The size of the circle corresponds to the number of genes in each GO term.

(a)



(b)

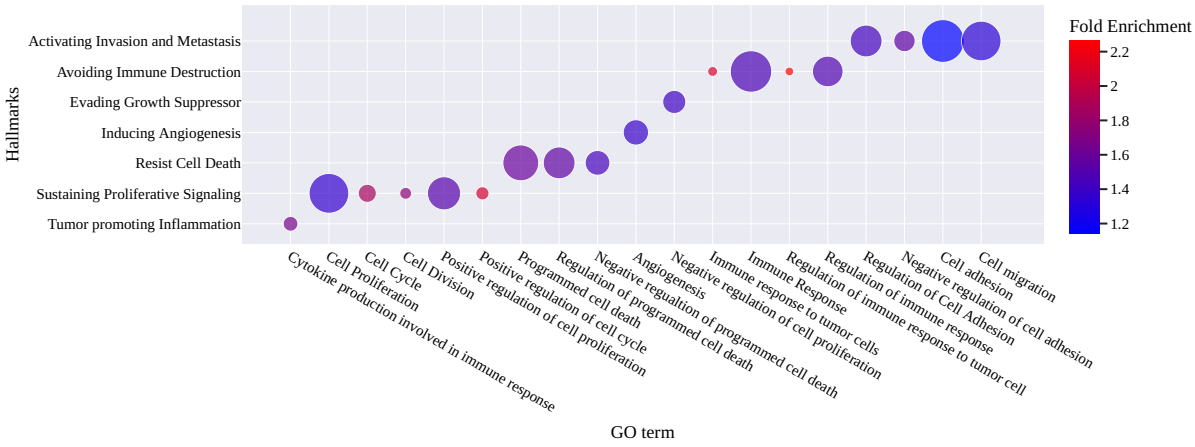


Figure A.16: GO terms Hallmarks enrichment for intercellular genes (complete graph). (a) source graph; (b) target graph. The size of the circle corresponds to the number of genes in each GO term.

B

Supplementary Tables

Table B.1: Correspondence between tissue ID and tissue names in GTEx and Consensus datasets.

Tissue ID	GTEx Name	Consensus Name
adipose_tissue	Adipose - Subcutaneous	adipose tissue
	Adipose - Visceral (Omentum)	-
adrenal_gland	Adrenal Gland	adrenal gland
appendix	-	appendix
artery	Artery - Aorta	-
	Artery - Coronary	-
	Artery - Tibial	-
bladder	Bladder	urinary bladder
blood	Whole Blood	-
bone	-	bone marrow
brain	Brain - Amygdala	amygdala
	Brain - Anterior cingulate cortex (BA24)	basal ganglia
	Brain - Caudate (basal ganglia)	cerebellum
	Brain - Cerebellar Hemisphere	cerebral cortex
	Brain - Cerebellum	choroid plexus
	Brain - Cortex	hippocampal formation
	Brain - Frontal Cortex (BA9)	hypothalamus
	Brain - Hippocampus	medulla oblongata
	Brain - Hypothalamus	midbrain
	Brain - Nucleus accumbens (basal ganglia)	pons
	Brain - Putamen (basal ganglia)	thalamus
	Brain - Substantia nigra	white matter
breast	Breast - Mammary Tissue	breast
cervix	Cervix - Ectocervix	cervix

Continued on next page

Table B.1: Correspondence between tissue ID and tissue names in GTEx and Consensus datasets.

Tissue ID	GTEx Name	Consensus Name
	Cervix - Endocervix	-
colorectum	Colon - Sigmoid	colon
	Colon - Transverse	rectum
epididymis	-	epididymis
eye	-	retina
fallopian_tube	Fallopian Tube	fallopian tube
fibroblasts	Cells - Cultured fibroblasts	-
gallbladder	-	gallbladder
heart	Heart - Atrial Appendage	heart muscle
	Heart - Left Ventricle	-
kidney	Kidney - Cortex	kidney
	Kidney - Medulla	-
liver	Liver	liver
lung	Lung	lung
lymph_node	-	lymph node
lymphocytes	Cells - EBV-transformed lymphocytes	-
nerve	Nerve - Tibial	-
oesophagus	Esophagus - Gastroesophageal Junction	esophagus
	Esophagus - Mucosa	-
	Esophagus - Muscularis	-
olfactory_bulb	-	olfactory bulb
ovary	Ovary	ovary
pancreas	Pancreas	pancreas
parathyroid_gland	-	parathyroid gland
pituitary_gland	Pituitary	pituitary gland
placenta	-	placenta
prostate	Prostate	prostate
salivary_gland	Minor Salivary Gland	salivary gland
seminal_vesicle	-	seminal vesicle
skeletal_muscle	Muscle - Skeletal	skeletal muscle
skin	Skin - Not Sun Exposed (Suprapubic)	skin
	Skin - Sun Exposed (Lower leg)	-
small_intestine	Small Intestine - Terminal Ileum	small intestine
	-	duodenum
smooth_muscle	-	smooth muscle
spinal_cord	Brain - Spinal cord (cervical c-1)	spinal cord

Continued on next page

Table B.1: Correspondence between tissue ID and tissue names in GTEx and Consensus datasets.

Tissue ID	GTEx Name	Consensus Name
spleen	Spleen	spleen
stomach	Stomach	stomach
testis	Testis	testis
thymus	-	thymus
thyroid	Thyroid	thyroid gland
tongue	-	tongue
tonsil	-	tonsil
uterus	Uterus	endometrium
vagina	Vagina	vagina

Table B.2: Correspondence between the cancer labels in the HCMDB dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Cancer Label	Tissue ID	Description
bladder cancer	bladder	-
brain cancer	brain	-
breast cancer	breast	-
cervical cancer	cervix	-
colorectal cancer	colorectum	-
esophagus cancer	oesophagus	-
ewing's sarcoma	-	Ewing sarcoma is a type of cancer that may be a bone sarcoma or a soft-tissue sarcoma. Removed – encompasses several types of tissues.
eye cancer	eye	-
gastric cancer	stomach	-
head and neck cancer	-	Broad spectrum of cancers arising from distinct types of tissues. Removed – encompasses several types of tissues.
kindey cancer	kidney	*kindey* was considered a typo of "kidney".
laryngeal cancer	-	It is a head & neck cancer. Removed – no tissue correspondence.
liver cancer	liver	-
lung cancer	lung	-

Continued on next page

Table B.2: Correspondence between the cancer labels in the HCMDB dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Cancer Label	Tissue ID	Description
midgut carcinoid tumor	-	Carcinoid tumours are of neuroendocrine origin and derived from primitive stem cells in the gut wall, especially the appendix. Removed – no tissue correspondence.
nasopharynx cancer	-	It is a head & neck cancer. Removed – no tissue correspondence.
oral cancer	tongue	It is a head & neck cancer. Oral squamous cell carcinoma arises from mucous basal cells. We will use *tongue* gene expression data since it is the most common type of oral cancer.
osteosarcoma	bone	-
ovarian cancer	ovary	-
pancreatic cancer	pancreas	-
pancreatic neuroendocrine tumor	-	Removed – no tissue correspondence.
penis cancer	-	-
prostate cancer	prostate	-
skin cancer	skin	-
small intestine cancer	small_intestine	-
synovial sarcoma	-	-
testicular cancer	testis	-
thymoma	thymus	-
thyroid cancer	thyroid	-

Table B.3: Correspondence between the labels of metastasis organs in the HCMDB dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Tissue Label	Tissue ID	Description
adrenal gland	adrenal_gland	-
bone	bone	-
brain	brain	-

Continued on next page

Table B.3: Correspondence between the labels of metastasis organs in the HCMDB dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Tissue Label	**Tissue ID**	**Description**
breast	breast	-
caudaequina	spinal_cord	Bundle of spinal nerves roots arising from inferior end of the adult spinal cord . Merged with *spinal chord*.
chest wall	-	Removed – encompasses several types of tissues.
colorectum	colorectum	-
fat	adipose_tissue	-
head & neck	-	Removed – encompasses several types of tissues.
kindey	kidney	*kindey* was considered a typo of "kidney".
liver	liver	-
lung	lung	-
lymph node	lymph_node	-
mediastinal	-	Removed – no tissue correspondence.
muscle	skeletal_muscle	Muscular tissue is classified into three types: skeletal – forms the large muscles responsible for movement; cardiac – heart muscle; smooth – located in the walls of blood vessels and visceral organs. *muscle* was interpreted as referring to skeletal muscle since skeletal muscle is what is commonly referred as "muscle" and the other types are part of major organs with their own specific tissue.
non-regional / distant lymph nodes	lymph_node	The location of the lymph nodes was not taked into account. Merged with *lymph node*.
omentum	-	The omenta are folds of peritoneal membrane which is also a fat deposit. The tissue in GTEx named *Adipose – Visceral (Omentum)* concernes only the adipose portion of the omenta. Metastasis occur in the membrane tissue. Removed.
other	-	Removed – undefined.
ovary	ovary	-
pancreas	pancreas	-

Continued on next page

Table B.3: Correspondence between the labels of metastasis organs in the HCMDB dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Tissue Label	Tissue ID	Description
parotis	-	The term "parotis" might be referring to the parotid glands (pair salivary glands) or to a term that describes a tumour near the ear. Removed – ambiguous term.
pelvis	-	Removed – no tissue correspondence.
peritoneal surfaces	-	see *peritoneum*.
peritoneum	-	Membrane that envelops the abdominal wall and its organs. Removed – no tissue correspondence.
pleura	-	Membrane that lines the pleural cavities and covers the lungs. Removed – no tissue correspondence.
pleura/pleural effusion	-	see *pleura*.
posterior peritoneum	-	see *peritoneum*.
renal	kidney	*renal* was considered as a synonym of kidney. Merged with *kindey*.
skeleton	bone	Merged with *bone*.
skin	skin	-
small intestine	small_intestine	-
soft tissue	-	Soft tissue refers to tissues that are not hardened by processes of ossification or calcification. Removed – encompasses several types of tissues.
spinal cord	spinal_cord	-
spleen	spleen	-
subcutaneous	-	The subcutaneous layer separates the skin from underlying tissues and organs. It consists of adipose and areolar tissue. Removed – encompasses several types of tissues.
subcutanious soft tissue	-	see *subcutaneous* and *soft tissue*.
unknown	-	Removed – undefined.
viscera	-	Viscera is a broad term for organs inside the thoracic and abdominopelvic cavities. Removed – encompasses several types of tissues.

Table B.4: Correspondence between the labels of primary tumour sites in the Autopsy Study dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Tissue Label	**Tissue ID**	**Description**
adrenal	adrenal_gland	-
anus	-	Removed – no tissue correspondence.
appendix	appendix	Removed – no tissue correspondence.
bile duct	-	Formed by the union of the cystic duct (from the gallbladder) to the hepatic duct (from the liver). Removed – no tissue correspondence.
bladder	bladder	-
bone	bone	-
branchial cyst	-	Removed – no tissue correspondence.
breast	breast	-
cervix	cervix	Term that designates the neck or any necklike part of an organ. In this case, it corresponds to the inferior narrow portion of the uterus that opens into the vagina.
colon	colorectum	-
duodenum	small_intestine	The first part of the small intestine. Merged with *small intestine*.
esophagus	oesophagus	-
eye	eye	-
gallbladder	gallbladder	-
kidney	kidney	-
larynx	-	Removed – no tissue correspondence.
lip	-	Removed – no tissue correspondence.
liver	liver	-
lung	lung	-
ovary	ovary	-
pancreas	pancreas	-
penis	-	Removed – no tissue correspondence.
pharynx	-	Removed – no tissue correspondence.
pleura	-	Membrane that lines the pleural cavities and covers the lungs. Removed – no tissue correspondence.
prostate	prostate	-
rectum	colorectum	-

Continued on next page

Table B.4: Correspondence between the labels of primary tumour sites in the Autopsy Study dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Tissue Label	Tissue ID	Description
retro-peritoneum	-	The peritoneum is the membrane that envelops the abdominal wall and its organs. Removed – no tissue correspondence.
salivary gland	salivary_gland	-
skin (body)	skin	The location of the skin sample was not taked into account. Merged with *skin (lower face)* and *skin (upper face)*.
skin (lower face)	skin	The location of the skin sample was not taked into account. Merged with *skin (body)* and *skin (upper face)*.
skin (upper face)	skin	The location of the skin sample was not taked into account. Merged with *skin (body)* and *skin (lower face)*.
small intestine	small_intestine	-
stomach	stomach	-
testis	testis	-
thyroid	thyroid	-
tongue	tongue	-
tonsil	-	Removed – no tissue correspondence.
unknown	-	Removed – undefined.
uterus	uterus	-
vagina	vagina	-
vulva	-	Removed – no tissue correspondence.

Table B.5: Correspondence between the labels of metastasis tumour sites in the Autopsy Study dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Tissue Label	Tissue ID	Description
adrenal	adrenal_gland	-
bone	bone	-
bladder	bladder	-
brain	brain	-
breast	breast	-
diaphragm	-	Skeletal muscle between the thoracic and abdominal cavities. Removed – no tissue correspondence.

Continued on next page

Table B.5: Correspondence between the labels of metastasis tumour sites in the Autopsy Study dataset and tissue IDs. The *Description* column shows the reasoning behind some of the decisions.

Tissue Label	**Tissue ID**	**Description**
gallbladder	gallbladder	-
heart	heart	-
kidney	kidney	-
lung	lung	-
colon	colorectum	-
liver	liver	-
lymph node (reg)	lymph_node	The location of the lymph nodes was not taked into account. Merged with *lymph node (dist)*.
lymph node (dist)	lymph_node	The location of the lymph nodes was not taked into account. Merged with *lymph node (reg)*.
omentum	-	The omenta are folds of peritoneal membrane which is also a fat deposit. The tissue in GTEx named *Adipose – Visceral (Omentum)* concernes only the adipose portion of the omenta. Metastasis occur in the membrane tissue. Removed.
ovary	ovary	-
pancreas	pancreas	-
pericardium	-	Membrane tha lines the heart. Removed – no tissue correspondence.
peritoneum	-	Membrane that envelops the abdominal wall and its organs. Removed – no tissue correspondence.
pleura	-	Membrane that lines the pleural cavities and covers the lungs. Removed – no tissue correspondence.
prostate	prostate	-
skeletal muscle	skeletal_muscle	-
skin	skin	-
small intestine	small_intestine	-
spleen	spleen	-
stomach	stomach	-
testis	testis	-
thyroid	thyroid	-
uterus	uterus	-
vagina	vagina	-

Table B.6: Common metastasis sites found in literature for each cancer.

Primary Site	Metastasis Sites	Observations	References
Adrenal Gland	Liver; Lung; Lymph Node; Bone	Data for Adrenocortical carcinoma (ACC)	[147]
Bladder	Lymph Node; Bone; Lung; Liver	Most BCs (70%–80%) are low grade, non-muscle invasive papillary ("superficial") tumours (NMIBCs) that rarely progress.	[148]
Bone	Lung	Osteosarcoma, Chondrosarcoma, Ewing's Sarcoma	[149]
Brain	Lymph Node; Lung; Bone; Liver; Skin	Data for glioblastoma (most common malignant brain tumour) are not described.	[150]
Breast	Bone; Liver; Brain; Lung; Lymph Node	-	[151]
Cervix	Lymph Node; Lung; Bone; Liver; Brain	-	[152]
Colorectum	Liver; Lung; Peritoneum; Bone; Brain	-	[153]
Eye	Liver; Lung; Bone; Skin	Data for Choroidal melanomas – the most common primary intraocular malignancy in adults. Represent 5% of all melanomas	[154]
Gallbladder	Lymph Node; Liver	-	[155]
Kidney	Lung; Lymph Node; Bone; Liver; Adrenal Gland; Brain	-	[156]
Liver	Lung; Lymph Node; Bone; Adrenal Gland	Data for hepatocelullar carcinoma – most common liver cancer (75%)	[157]
Lung	Bone; Liver; Brain; Adrenal Gland	-	[158]
Oesophagus	Liver; Lymph Node; Lung; Bone; Brain	-	[159]
Ovary	Fallopian Tube; Peritoneum; Omentum	-	[52]

Continued on next page

Table B.6: Frequent metastasis sites found in literature.

Primary Site	Metastasis Sites	Observations	References
Pancreas	Liver; Lymph Node; Lung; Peritoneum; Adrenal Gland; Bone	Autopsy study in Hungary. Liver (61.3%), lymph node (57%), peritoneum (23.7%), lung (22.1%), adrenal gland (4.7%), bone (4.5%),	[160]
Prostate	Bone; Lung; Liver; Adrenal Gland; Peritoneum	Autopsy study in Switzerland. Bone (90%), lung (46%), liver (25%), pleura (21%), adrenal gland (13%), peritoneum (7%)	[161]
Skin	Lymph Node; Lung; Liver; Brain	Data for squamous cell carcinoma and melanoma. Brain metastasis are particularly common in melanoma	[162][163]
Stomach	Liver; Peritoneum; Lung; Bone; Lymph Node	Registry study in Sweden. Liver (48%), peritoneum (32%), lung (15%), and bone (12%), lymph node (11%)	[164]
Testis	Lymph Node; Lung	-	[165]
Thyroid	Lymph Node; Lung; Bone	Papillary thyroid cancer (most common subtype): lymph nodes, lung; Follicular thyroid cancer, Hurthle cell thyroid cancer, and poorly differentiated thyroid cancers: lung and bone.	[166]
Tongue	Lung; Bone; Lymph Node; Liver	Data for oral cavity cancers.	[167][168]
Uterus	Lymph Node; Vagina; Peritoneum; Lung	Data for endometrial cancer.	[169]

Table B.7: Intersection between organotropism pairs determined using the hypergeometric test and outlier detection. The ratio represents the percentage of hypergeometric test-based organotropism pairs which are also found with the outlier detection.

Metastasis Dataset	**Tissue Dataset**	**Jaccard**	**Ratio (%)**
HCMDB	GTEx	0.36	86.7
	Consensus	0.34	91.7
Autopsy Study	GTEx	0.12	31.8
	Consensus	0.16	48.5

Table B.8: Intersection between organotropism pairs determined using outlier detection and literature curation. The ratio represents the percentage of outlier detection-based organotropism pairs which are also found in the literature.

Metastasis Dataset	**Tissue Dataset**	**Jaccard**	**Ratio (%)**
HCMDB	GTEx	0.56	82.4
	Consensus	0.49	77.4
Autopsy Study	GTEx	0.42	61.0
	Consensus	0.43	59.5

Table B.9: Spearman Rank Correlation Coefficient results for undirected networks built using gene present/absence calls.

Metastasis Dataset	Tissue Dataset	Interactions	Coefficient	p-value
HCMDB	GTEx	all	-0.053034	0.409
		curated	-0.082554	0.198
	Consensus	all	0.009396	0.862
		curated	-0.200076	0.000
Autopsy Study	GTEx	all	0.030411	0.543
		curated	0.036260	0.468
	Consensus	all	0.015307	0.707
		curated	-0.072838	0.073

Table B.10: SRCC results for directed networks built using gene present/absence calls. **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

Metastasis Dataset	Tissue Dataset	Interactions	Direction	Coefficient	p-value
HCMDB	GTEx	all	$C \rightarrow M$	-0.023832	0.711
			$M \rightarrow C$	-0.072274	0.260
		curated	$C \rightarrow M$	0.038780	0.546
			$M \rightarrow C$	-0.132626	0.038
	Consensus	all	$C \rightarrow M$	0.020683	0.701
			$M \rightarrow C$	0.028143	0.602
		curated	$C \rightarrow M$	-0.187525	0.000
			$M \rightarrow C$	-0.123228	0.022
Autopsy Study	GTEx	all	$C \rightarrow M$	0.038881	0.436
			$M \rightarrow C$	0.052990	0.289
		curated	$C \rightarrow M$	0.120259	0.016
			$M \rightarrow C$	0.025249	0.613
	Consensus	all	$C \rightarrow M$	0.052126	0.200
			$M \rightarrow C$	0.026975	0.508
		curated	$C \rightarrow M$	-0.016083	0.693
			$M \rightarrow C$	-0.094005	0.021

Table B.11: SRCC results for undirected weighted networks.

Metastasis Dataset	Tissue Dataset	Interactions	Coefficient	p-value
HCMDB	GTEx	all	-0.121467	0.058
		curated	-0.051335	0.424
	Consensus	all	-0.165930	0.002
		curated	-0.062821	0.244
Autopsy Study	GTEx	all	-0.087368	0.080
		curated	-0.108180	0.030
	Consensus	all	-0.113788	0.005
		curated	-0.073063	0.073

Table B.12: SRCC results for directed weighted networks. **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

Metastasis Dataset	Tissue Dataset	Interactions	Direction	Coefficient	p-value
HCMDB	GTEx	all	$C \rightarrow M$	-0.072707	0.257
			$M \rightarrow C$	-0.175010	0.006
		curated	$C \rightarrow M$	-0.013224	0.837
			$M \rightarrow C$	-0.137924	0.031
	Consensus	all	$C \rightarrow M$	-0.018284	0.735
			$M \rightarrow C$	-0.167934	0.002
		curated	$C \rightarrow M$	0.043678	0.418
			$M \rightarrow C$	-0.154551	0.004
Autopsy Study	GTEx	all	$C \rightarrow M$	-0.006454	0.897
			$M \rightarrow C$	-0.272989	0.000
		curated	$C \rightarrow M$	0.133911	0.007
			$M \rightarrow C$	-0.245210	0.000
	Consensus	all	$C \rightarrow M$	0.081737	0.044
			$M \rightarrow C$	-0.216922	0.000
		curated	$C \rightarrow M$	0.122763	0.002
			$M \rightarrow C$	-0.205191	0.000

Table B.13: Results for the median test in undirected networks built using gene presence/absence calls.

Metastasis Dataset	Tissue Dataset	Interactions	p-value (# of interactions)	p-value (jaccard index)
HCMDB	GTEx	all	0.000	0.000
		curated	0.000	0.000
	Consensus	all	0.011	0.011
		curated	0.006	0.023
Autopsy Study	GTEx	all	0.191	0.190
		curated	0.171	0.220
	Consensus	all	0.135	0.204
		curated	0.261	0.294

Table B.14: Results for the median test in directed networks built using gene presence/absence calls. **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

Metastasis Dataset	Tissue Dataset	Interactions	Direction	p-value (# of interactions)	p-value (jaccard index)
HCMDB	GTEx	all	$C \rightarrow M$	0.000	0.000
			$M \rightarrow C$	0.000	0.000
		curated	$C \rightarrow M$	0.002	0.000
			$M \rightarrow C$	0.000	0.000
	Consensus	all	$C \rightarrow M$	0.025	0.040
			$M \rightarrow C$	0.001	0.000
		curated	$C \rightarrow M$	0.024	0.046
			$M \rightarrow C$	0.000	0.001
Autopsy Study	GTEx	all	$C \rightarrow M$	0.348	0.162
			$M \rightarrow C$	0.221	0.251
		curated	$C \rightarrow M$	0.313	0.319
			$M \rightarrow C$	0.264	0.354
	Consensus	all	$C \rightarrow M$	0.160	0.257
			$M \rightarrow C$	0.069	0.052
		curated	$C \rightarrow M$	0.285	0.310
			$M \rightarrow C$	0.205	0.256

Table B.15: Results for the median test in undirected weighted networks.

Metastasis Dataset	Tissue Dataset	Interactions	p-value
HCMDB	GTEx	all	0.000
		curated	0.000
	Consensus	all	0.001
		curated	0.001
Autopsy Study	GTEx	all	0.115
		curated	0.126
	Consensus	all	0.062
		curated	0.049

Table B.16: Results for the median test in directed weighted networks. **C→M**: interactions from cancer to metastasis. **M→C**: interactions from metastasis to cancer.

Metastasis Dataset	Tissue Dataset	Interactions	Direction	p-value
HCMDB	GTEx	all	$C \rightarrow M$	0.000
			$M \rightarrow C$	0.000
		curated	$C \rightarrow M$	0.000
			$M \rightarrow C$	0.000
	Consensus	all	$C \rightarrow M$	0.013
			$M \rightarrow C$	0.000
		curated	$C \rightarrow M$	0.001
			$M \rightarrow C$	0.000
Autopsy Study	GTEx	all	$C \rightarrow M$	0.027
			$M \rightarrow C$	0.107
		curated	$C \rightarrow M$	0.070
			$M \rightarrow C$	0.061
	Consensus	all	$C \rightarrow M$	0.148
			$M \rightarrow C$	0.000
		curated	$C \rightarrow M$	0.216
			$M \rightarrow C$	0.001

Table B.17: GO terms and Cancer Hallmarks enrichment - Intracellular Genes.

Gene Type	Network	GO Terms	# of Mapped Hallmarks
source	complete	242	3
	curated	336	4
target	complete	1695	8
	curated	1550	8

Table B.18: GO terms and Cancer Hallmarks enrichment - Intercellular Genes.

Gene Type	Network	Associated	GO Terms	# of Mapped Hallmarks
source	complete	Yes	556	4
	curated	Yes	361	4
target	complete	Yes	688	7
	curated	Yes	229	6
source	complete	No	0	0
	curated	No	0	0
target	complete	No	1	0
	curated	No	0	0