HELMHOLTZ ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE UNIVERSITÄT MÜNCHEN

Elucidating protein glycosylation mechanisms by combining network-based pathway analysis with prior knowledge

Elisa Benedetti

August 2018

TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Elucidating protein glycosylation mechanisms by combining network-based pathway analysis with prior knowledge

Elisa Benedetti

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Jan Baumbach

Prüfer der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Dmitrij Frishman

Die Dissertation wurde am 05.09.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 07.11.2019 angenommen.

Acknowledgements

My PhD would not have been possible without my advisor Prof. Dr. Dr. Fabian J. Theis, who allowed me to be part of his institute and provided invaluable input throughout my journey.

My deepest gratitude and appreciation goes to Dr. Jan Krumsiek, who was an incredible mentor and guide. Thank you for taking a chance on me in the first place and for making of the supervision your people your priority. Thank you for your work ethics, for never stopping challenging my ideas and never backing down from asking questions. You allowed me to grow and follow my own path, finding ambitions I did not even know I had, and I will always feel in your debt for that. From you I learned what good leadership looks like.

These years would not have been nearly as pleaseant without the people who shared this experience with me. Thank you to all ICB members for the lively atmosphere and the collaborative environment and to the wonderful office team, Elisabeth, Marianne, Sabine and Anna, for always helping out with a smile despite their overwhelming schedules.

A special mention goes to the SysDiab family: Mustafa, Parviz, Helena, Michl, Alida, and Kiki. I feel incredibly lucky to have had the pleasure to share these years with all of you and I could not have dreamed of better colleagues. Thank you for all the fruitful discussions, for your constructive criticism, and for all the time we spent together inside and outside the office. Thank you to my roomies Mustafa and Parviz for all the conversations and laughs, which alleviated the stress of the last months. To Alida and Kiki, who have now moved on, but played an essential role in making these years so incredibly enjoyable and fun. You both were amazing colleagues and I am proud to be able to call you friends. Alida, I am so grateful for every experience we shared together. Thank you for your patience in teaching me how to squat and deadlift, and for being the best gym buddy I could wish for. Kiki, thank you for your kindness and your generosity, for your selfless support and honest advice.

I am deeply thankful to my collaborators in the MIMOmics consortium, in particular to

the team at Genos and to Prof. Gordan Lauc, who infected me with the glycomics bug through his contageous enthusiasm and provided invaluable advice and assistance during my PhD.

Thank you to my extended family and friends for the endless support, and in particular to my sister Silvia, and my parents, Paola and Piero: I would not be here without you and I am incredibly proud and grateful for having you always by my side, no matter how many countries or continents are between us. Thank you for being excited with me anytime I take on a new challenge, and for encouraging me to chase new opportuinities without even blinking, even when it means being further away.

And finally, to Marco. Thank you for your support during the last intense months, for enduring my long hours and for always being the place I want to go back to after a though day. Thank you for knowing me better than anyone else, for challenging my views and calling me out when necessary. *Sed magis amica veritas*, always. I am proud of where we are and I am looking forward to our next adventure together.

Abstract

Protein glycosylation, namely the covalent attachment of monosaccharide chains called glycans to the protein backbone, is the most frequent and structurally diverse post-translational modification. This enzymatic process is not directly encoded in the genome, but is the result of the activity of hundreds of enzymes that catalyze and regulate the attachment and removal of monosaccharides on proteins in a concerted fashion. Alternative glycosylation, namely the presence of different glycan structures on a given glycosylation site on a protein, can have substantial effects on the protein structure and function, sometimes completely reverting its activity. Moreover, the final glycosylation state of any protein or cell heavily depends on the physiological and environmental conditions of the organism, and alterations in glycosylation profiles of many proteins have been described for a variety of human diseases.

Due to the overall complexity of the glycosylation process and the absence of a genetic template to manipulate, a proper in vivo investigation of the underlying molecular mechanisms is, to this day, experimentally infeasible. Current technologies allow to establish the in vitro substrate specificities of all major glycosylation enzymes, although their in vivo cell-, protein-, or site-specific activities cannot be predicted with this approach. Current measurement technologies are able to measure glycans in large population cohorts, either from isolated proteins or from a mixture of proteins, enabling systematic statistical data analysis, which can help gain insight into the regulation of protein glycosylation in absence of direct experimental evidence.

In this doctoral thesis, I investigate large-scale glycomics datasets to elucidate the molecular mechanisms of glycan synthesis. My main aims are: (i) inferring new biochemical reactions taking part in glycan synthesis; (ii) predicting the structural details of glycans from mass spectrometry measurements; (iii) establishing the best preprocessing strategy for glycomics data; (iv) optimizing the inference of correlation networks. Each of these points is addressed in a specific project.

First, we estimate a partial correlation network, or Gaussian Graphical Model (GGM), from four large-scale Immunoglobulin G (IgG) glycomics cohorts. Here, only glycans from a specific glycosylation site, on a specific protein from human plasma are quantified. We demonstrate that statistically significant partial correlations among glycan pairs mostly represent known glycosylation synthesis reactions. Our analysis also shows evidence of previously unknown substrate specificities for two glycosylation enzymes, which would allow additional steps in the glycan synthesis pathway. We validate our predictions using data from a GWAS and results from three in vitro experiments. Our findings demonstrate that GGMs are able to recover single biochemical steps in glycosylation pathways and can drive the discovery of new synthesis steps.

Second, we consider a mixture glycomics dataset, where glycans from all proteins in human plasma are quantified via Mass Spectrometry (MS). This platform allows to identify molecular masses, which could correspond to different glycan structures. We intersect the data-driven GGM with the prior knowledge available on the synthesis pathway to infer the most abundant structure contributing to each measured molecular mass. Our predictions are validated with previously published datasets and demonstrate high accuracy. This approach could contribute to the characterization of complex glycomics datasets thereby help in reducing the cost of additional fragmentation experiments for the identification of glycan structural features.

Third, we exploit the strong relationship between GGMs and glycosylation pathways to evaluate different preprocessing strategies for glycomics data. We quantify the quality of any given normalization through the ability of the corresponding data-driven GGM to reconstruct known biochemical synthesis steps. This is an innovative approach to the problem of normalization evaluation, as it is based on a biological quality measure rather than on purely statistical criteria. We consider six glycomics datasets and three different measurements platforms. We are able to identify an optimal preprocessing strategy that holds for any of the considered glycomics platform and data type.

Finally, we use the overlap between GGM and glycosylation pathways to address a major problem in correlation network inference, namely the determination of a biologically meaningful correlation cutoff. That is, we search for the cutoff that produced the network that best reproduces known molecular interactions. We show that even a coarse, incomplete, or partly incorrect prior knowledge is suitable for this approach, as long as a sufficient amount of correct information is included. We first prove the validity of the approach on glycomics measurements, for which we have a well-characterized, supposedly complete biochemical synthesis pathway. We then apply the algorithm to urine metabolomics and TCGA RNA-sequencing data, where our method is able to identify an optimal network

and to outperform regular statistical cutoffs.

Taken together, we demonstrate that GGMs are able to reconstruct biochemical pathways of glycan synthesis from large scale glycomics data, as well as to predict true but unknown enzymatic steps. Moreover, GGMs and prior knowledge can be successfully exploited to infer glycan structural features from mass spectrometry measurements, and to optimize glycomics data preprocessing and GGM estimation. In conclusion, this thesis provided new insights into protein glycosylation, as well as new statistical tools for the analysis and interpretation of glycomics data.

Zusammenfassung

Die Proteinglykosylierung - die kovalente Bindung von Monosaccharidketten (Glykane) an das Proteinrückgrat - ist die häufigste und strukturell facettenreichste post-translationale Modifikation. Dieser enzymatische Prozess ist nicht direkt im Genom kodiert, sondern ist das Ergebnis der Aktivitäten von Hunderten von Enzymen, die die Bindung und Spaltung der Monosaccharide von Proteinen katalysieren und regulieren. Alternative Glykosylierung, die Anwesenheit von verschiedenen Glykanstrukturen an einer gegebenen Glykosylierungsstelle eines Proteins, kann wesentliche Auswirkungen auf die Proteinstruktur und -funktion haben; beispielsweise kann die Funktionalität eines Proteins vollkommen verändert oder sogar umgekehrt werden. Außerdem hängt der Endglykosylierungszustand eines Proteins oder einer Zelle stark von den physiologischen Bedingungen und den Umweltfaktoren des Organismus ab. Veränderungen der Glykosylierungsprofile vieler Proteine wurden bereits mit einer Vielzahl von menschlichen Erkrankungen assoziiert.

Aufgrund der Komplexität des Glykosylierungsprozesses und der Abwesenheit einer manipulierbaren genetischen Komponente ist eine tiefgehende In-vivo-Untersuchung der zugrundeliegenden molekularen Mechanismen bis heute experimentell nicht durchführbar. Aktuelle Technologien erlauben es, die In-vitro-Substratspezifitäten aller wichtigen Glykosylierungsenzyme zu bestimmen, allerdings können daraus keine Vorhersagen bezüglich ihrer In-vivo zell-, protein- oder ortsspezifischen Aktivitäten gemacht werden. Einen Schritt in Richtung Verständnis der Regulierung und Funktionalität nativer Glykoproteine ohne direkte experimentelle Beweise bieten gegenwärtige Messtechnologien, die in der Lage sind, Glykane in großen Populationskohorten entweder aus isolierten Proteinen oder aus einer Mischung von Proteinen zu messen, was eine systematische, statistische Datenanalyse ermöglicht.

In dieser Dissertation untersuche ich large-scale Glycomics-Datensätze, um die molekularen Mechanismen der Glykansynthese zu untersuchen. Meine Hauptziele sind: (i) Erschließen neuer biochemischer Reaktionen, die an der Glykansynthese beteiligt sind; (ii) Vorhersagen der strukturellen Details von Glykanen aus Massenspektrometrie-Messungen;

(iii) Untersuchen von geeigneten Preprocessing-Strategien für MS-basierte Glycomics-Daten; (iv) Optimierung der Inferenz von Korrelationsnetzwerken. Jeder dieser Punkte wird in einem der nachfolgenden Projekte beschrieben.

Um neue biochemische Reaktionen in der Glykansynthese zu identifizieren, generieren wir ein partielles Korrelationsnetzwerk, oder Gaussian Graphical Model (GGM), basierend auf vier großen Immunglobulin G (IgG) Glycomics Kohorten. In diesen Kohorten wurden nur Glykane von einer spezifischen Glykosylierungsstelle an einem spezifischen Protein im menschlichen Plasma quantifiziert. Wir zeigen, dass statistisch signifikante partielle Korrelationen zwischen Glykanpaaren meist bekannte Reaktionen aus der Glykansynthese darstellen. Unsere Analyse gibt auch Hinweise auf bisher unbekannte Substratspezifitäten für zwei Glykosylierungsenzyme, die den bisherigen Glykansyntheseweg um zwei zusätzliche Schritte erweitern würden. Wir validieren unsere Vorhersagen mit Daten aus einem GWAS und Ergebnissen von drei In-vitro-Experimenten. Unsere Resultate zeigen, dass GGMs in der Lage sind, einzelne biochemische Schritte in Glykosylierungswegen zu rekonstruieren und neue Syntheseschritte zu identifizieren.

Im zweiten Teilprojekt zielen wir darauf ab, Glykan-Strukturen basierend auf MS-Daten genauer vorherzusagen. Hierzu verwenden wir einen gemischten Glycomics-Datensatz, in dem Glykane aus allen Proteinen im menschlichen Plasma mittels Massenspektrometrie (MS) quantifiziert werden. Diese Plattform ermöglicht es, molekulare Massen zu identifizieren, die verschiedenen Glykanstrukturen entsprechen könnten. Wir überlappen das datengesteuerte GGM mit dem experimentell-basierten Syntheseweg, um auf die am häufigsten vorkommende Struktur zu schließen, die zu jeder gemessenen molekularen Masse beiträgt. Unsere Vorhersagen werden mit zuvor veröffentlichten Datensätzen validiert und zeigen eine hohe Sensitivität. Dieser Ansatz könnte zur Charakterisierung von komplexen Glycomics-Datensätzen beitragen und somit die Kosten zusätzlicher Fragmentierungsexperimente zur Identifizierung von Strukturmerkmalen von Glykanen erheblich senken.

Im dritten Teilprojekt nutzen wir die enge Beziehung zwischen statistischen GGMs und experimentell identifizierten Glykosylierungswegen, um verschiedene Preprocessing-SchritteD für Glycomics-Daten zu analysieren. Dazu evaluieren wir die Qualität einer gegebenen Normalisierungsmethode anhand der Fähigkeit des daraus resultierenden datengetriebenen GGMs, bekannte biochemische Syntheseschritte zu rekonstruieren. Dies ist ein innovativer Ansatz für das Problem der Bewertung von Normalisierungsmethoden, da er auf einem biologischen Qualitätsmaß statt auf rein statistischen Kriterien beruht. Unter Betrachtung von sechs Glycomics-Datensätzen und drei verschiedene Messplattformen sind wir in der Lage, eine optimale Preprocessing-Strategie zu bestimmen, die für jede der

betrachteten Glycomics-Plattformen und -Datentypen gilt

Schließlich verwenden wir die Überlappung zwischen GGM und Glykosylierungswegen, um ein Hauptproblem in der Korrelationsnetzwerk-Inferenz anzugehen: die Bestimmung einer biologisch sinnvollen Korrelationsgrenze. Das heißt, wir suchen nach dem Cutoff, der das Netzwerk generiert, das die bekannten Glykosylierungswege (biologische Regerenz) am besten reproduziert. Wir zeigen, dass auch eine grobe, unvollständige oder teilweise inkorrekte Referenz für diesen Ansatz geeignet ist, solange eine ausreichende Menge an korrekter Information enthalten ist. Wir beweisen zunächst die Gültigkeit des Ansatzes für Glycomics-Messungen, für die wir einen gut charakterisierten, vermutlich vollständigen biochemischen Syntheseweg haben. Wir wenden den Algorithmus dann auf Metabolomics-Messungen von Urin-Proben und TCGA-RNA-Sequenzierungsdaten an, wobei unsere Methode in der Lage ist, ein optimales Netzwerk zu identifizieren und übliche statistische Cutoffs zu übertreffen.

Zusammenfassend demonstrieren wir, dass GGMs in der Lage sind, biochemische Wege der Glykan-Synthese aus Glycomics-Daten in großem Maßstab zu rekonstruieren, sowie echte, aber unbekannte enzymatische Schritte vorherzusagen. Darüber hinaus können GGMs und experimentelles Vorwissen erfolgreich genutzt werden, um Strukturmerkmale von Glykanen aus Massenspektrometrie-Messungen abzuleiten und das Preprocessing von Glycomics-MEssungen sowie GGM-Schätzung zu optimieren. Zusammenfassend liefert diese Arbeit neue Einblicke in die Proteinglykosylierung sowie neue statistische Werkzeuge für die Analyse und Interpretation von Glycomics-Daten.

List of contributed articles

The following list shows peer-reviewed publications, manuscripts in submission or revision relevant for each chapter of this thesis.

Chapter 3

- **Benedetti, E.**, Pučić-Baković, M., Keser, T., Wahl, A., Hassinen, A., Yang, J-Y., Liu, L., Trbojević-Akmačić, I., Razdorov, G., Štambuk, J., Klarić, L., Ugrina, I., Selman, M.H.J., Wuhrer, M., Rudan, I., Polasek, O., Hayward, C., Grallert, H., Strauch, K., Peters, A., Meitinger, T., Gieger, C., Vilaj, M., Boons, G-J., Moremen, K.W., Ovchinnikova, T., Bovin, N., Kellokumpu, S., Theis*, F.J., Lauc*, G., Krumsiek*, J., Network inference from glycoproteomics data reveals new reactions in the IgG glycosylation pathway, *Nat. Commun.*, 8(1):1483, 2017.

Chapter 4

- **Benedetti, E.**, Frick, A., Reiding, K.R., Ruhaak, L. R., Beekman, M., Slagboom, E., Wuhrer, M., Krumsiek, J., Network-driven structural inference on the human total plasma N-glycome. *Manuscript in preparation.*

Chapter 5

- **Benedetti, E.**, Gerstner, N., Pučić-Baković, M., Keser, T., Reiding, K.R., Ruhaak, L. R., Pavić, T., Selman, M.H.J., Rudan, I., Polašek, O., Hayward, C., Beekman, M., Slagboom, E., Wuhrer, M., Dunlop, M.G., Lauc, G., Krumsiek, J., Systematic evaluation of normalization methods for glycomics data based on biological network inference. *Manuscript in preparation.*

Chapter 6

- **Benedetti, E.**, Pučić-Baković, M., Keser, T., Gerstner, N., Büyüközkan, M., Pavić, T., Selman, M.H.J., Rudan, I., Polašek, O., Hayward, C., Al-Amin, H., Suhre, K., Kastenmüller, G., Lauc, G., Krumsiek, J., Using prior knowledge to optimize correlation network cutoffs. *Submitted.*

Further publications

During the course of my PhD I was involved in further projects which are not specifically discussed in this thesis.

- Wahl, A., Kasela, S., Carnero-Montoro, E., van Iterson, M., Štambuk, J., Sharma, S., van den Akker, E., Klaric, L., **Benedetti, E.**, Razdorov, G., Trbojević-Akmačić, I., Vučković, F., Ugrina, I., Beekman, M., Deelen, J., van Heemst, D., Heijmans, B.T., B.I.O.S. Consortium, Wuhrer, M., Plomp, R., Keser, T., Šimurina, M., Pavić, T., Gudelj, I., Krištić, J., Grallert, H., Kunze, S., Peters, A., Bell, J.T., Spector, T.D., Milani, L., Slagboom, P.E., Lauc, G., Gieger, C., IgG glycosylation and DNA methylation are interconnected with smoking, *Biochim Biophys Acta*, 1862(3):637-648, 2018.

- Wahl, A., van den Akker, E., Klaric, L., Štambuk, J., **Benedetti, E.**, Plomp, R., Razdorov, G., Trbojević-Akmačić, I., Deelen, J., van Heemst, D., Slagboom, P.E., Vučković, F., Grallert, H., Krumsiek, J., Strauch, K., Peters, A., Meitinger, T., Hayward, C., Wuhrer, M., Beekman, M., Lauc, G., Gieger, C., Genome-Wide Association Study on Immunoglobulin G Glycosylation Patterns, *Front Immunol.*, 9:277, 2018.

Contents

Chapter 1

Introduction

In this chapter, we first illustrate the basic concepts of glycobiology, we then describe the most common technologies for data acquisition and we present the data analysis approaches exploited in this work. We conclude the chapter introducing the research questions addressed here, as well as a summary of the content of this thesis.

1.1 Glycobiology

Glycosylation is the enzymatic process that covalently binds sugar chains, called *glycans*, to proteins and lipids. Glycosylated molecules are referred to as *glycoproteins* and *glycolipids*, or *glycoconjugates* in general. This thesis will focus on protein glycosylation exclusively.

It is estimated that at least 50% of all human proteins are glycosylated [1], most of which are found on the extracellular surface of the plasma cell membrane or as secreted proteins [2]. The proportion to which glycans contribute to the overall mass of their glycoconjugate can vary greatly, but in many cases they constitute a substantial portion. For example, the glycans bound to the *major glycoprotein* on human erythrocyte cell membranes have been estimated to contribute to the total molecular weight for more than 50% [3]. Moreover, the surface of all cells is covered with a dense layer of glycans, referred to as *glycocalyx*. This glycocalyx is a feature of *all* cells, with no known exception. It is worth noting that nature was able to create human cells *without a nucleus* (for example corneocytes), but none without a sugar coating. Despite this prominent feature, the role and effect of glycosylation is often ignored when describing the plasma membrane of cells, where glycans are depicted as trees sparsely decorating the mostly flat surface of the lipid bilayer (Figure 1.1A). In reality, the glycocalyx is an extremely dense layer that can be

over ten times as thick as the plasma membrane itself (Figure 1.1B). This is just one of the many examples that illustrate the relevance of glycans, which have by now been recognized to take part in virtually all biological processes [4].

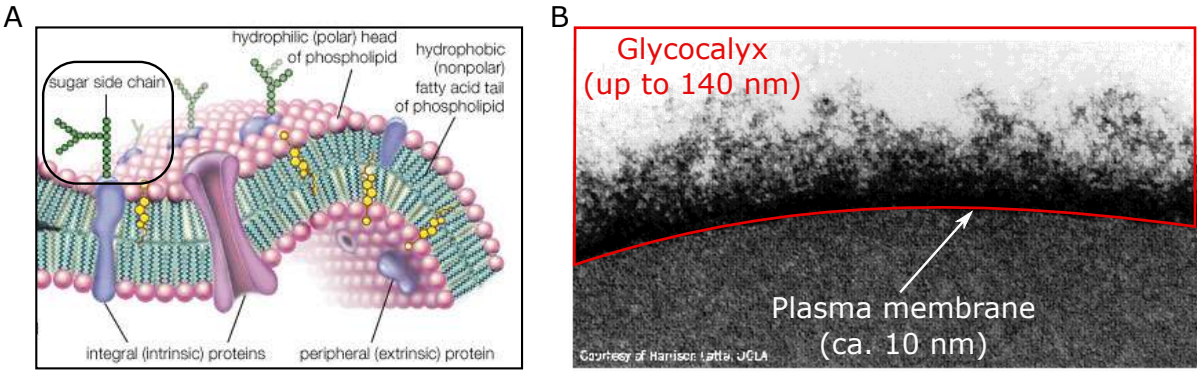


Figure 1.1: Glycocalyx. **A** Typical textbook image of the plasma membrane. In this picture, glycans appear sparsely on the surface of the lipid bilayer. Taken from the Encyclopedia Britannica [5]. **B** Electron microscopy image of a human erythrocyte. The cell is densely covered by a thick layer of glycans, referred to as *glycocalyx*. The glycocalyx can be up to 140 nm thick, more than ten times the thickness of the plasma membrane (typically ca. 10 nm), which is barely visible in the image. Adapted from Voet and Voet, Biochemistry [6].

In this section, we will introduce some general concepts of protein glycosylation, starting from the description of the basic building blocks of vertebrate glycans and their synthesis, highlighting then the role of glycans in modulating protein functions and interactions, and concluding with an overview of the documented involvement of glycans in human diseases.

1.1.1 Glycan building blocks and linkage

There are nine basic monosaccharides found in vertebrate glycoconjugates [7] (Figure 1.2), although a variety of other monosaccharides exist in other species [8]. The most common constituents of vertebrate glycans are hexoses, six-carbons sugars, which are found on glycoconjugates in three forms: **Glucose** (Glc), **Galactose** (Gal), and **Mannose** (Man). These sugars are all epimers, i.e., they are made by the same atoms and differ only in the configuration. The same holds for the two hexosamines **N-Acetylglucosamine** (GlcNAc) and **N-Acetylgalactosamine** (GalNAc), where the 2-hydroxyl group of the corresponding hexose is substituted by an acetylated amino group. **Glucuronic acid** (GlcA) is made from oxidation to a carboxyl group of the C6 atom of glucose, while the removal of the C6 atom of glucose generates the pentose **Xylose** (Xyl). **Fucose** (Fuc) is created by the loss of the 6-hydroxyl group from galactose, but it is found in a different configuration (L instead of D, as for all other monosaccharides). The last known building block is a sialic acid called **N-Acetylneuraminate** (Neu5Ac), a negatively charged nine-

carbon sugar acid. In this thesis, we will encounter seven of these nine monosaccharides (gray boxes in Figure 1.2).

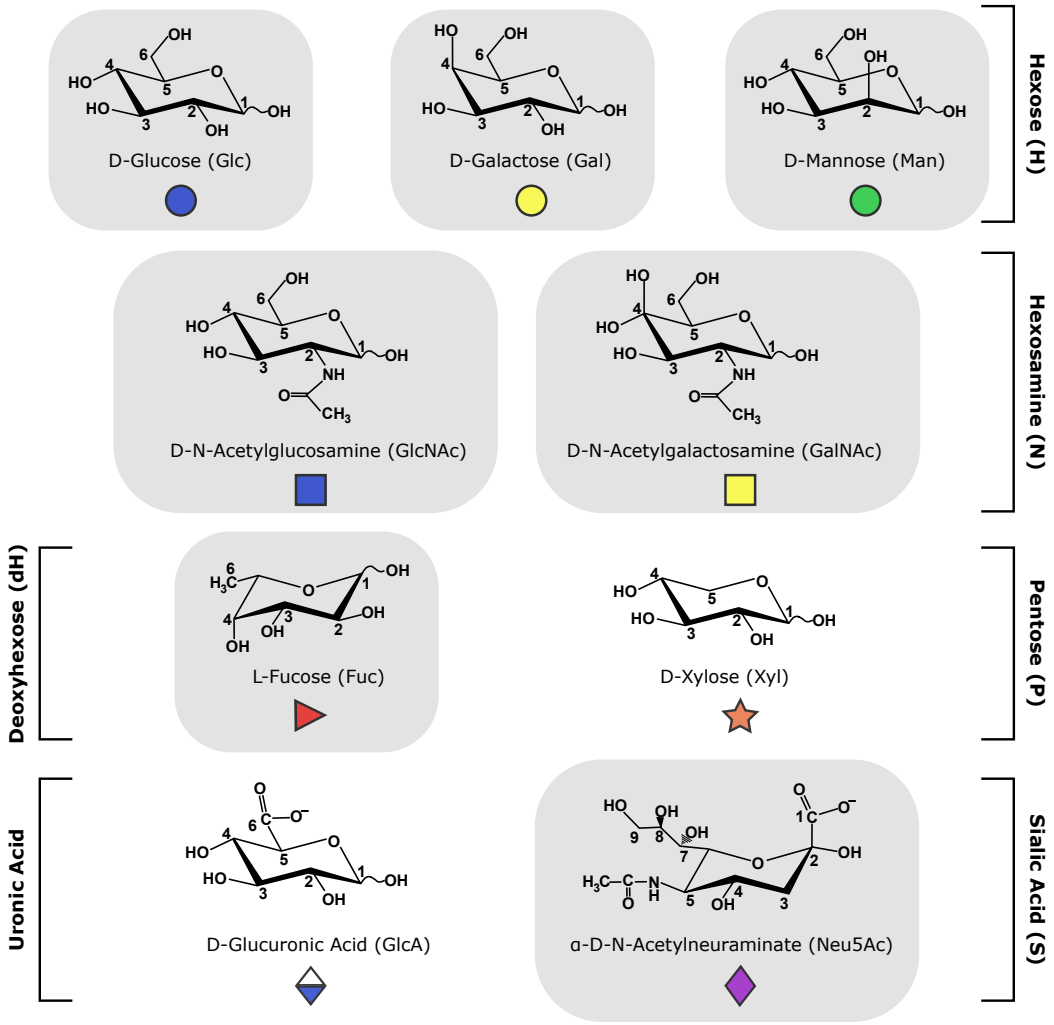


Figure 1.2: Common monosaccharides found in human glycoconjugates. A colored shape is associated to each monosaccharide, and this symbolic nomenclature, defined according to the Consortium for Functional Glycomics (CFG) [9], will be used to represent glycan structures throughout this thesis. Gray boxes represent monosaccharides that will be discussed in this work.

In glycans, monosaccharides are linked together via a *glycosidic bond*, which is formed between the C1 carbon, also known as *anomeric carbon*, of one building block and the hydroxyl group (OH) of the other. This same chemical bond can also be created between a glycan and a protein, if the latter is a hydroxyl amino acid, such as serine (Ser) or threonine (Thr) (see Subsection 1.1.2). The glycosidic linkage can occur in either α or β stereoisomeric form (Figure 1.3).

Hexoses have four hydroxyl groups attached to the ring carbons, so each one of these

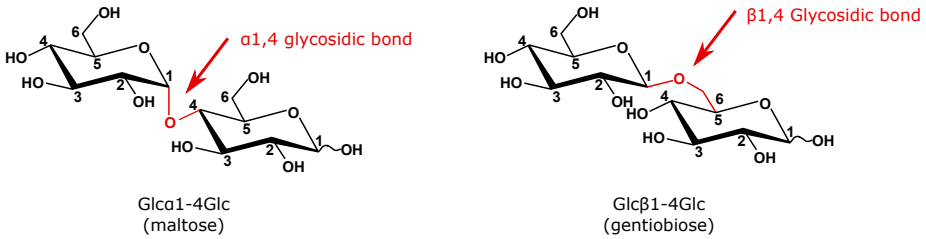


Figure 1.3: Glycosidic bond.

monosaccharide can have up to four glycosidic linkage sites, each with two linkage stereoisomers. This means that, given two hexoses, they can produce eight different *disaccharides*. If we consider three different hexoses, we can have up to 384 combinations[1]. This is a completely different level of complexity in comparison to other macromolecules, like DNA and proteins, where three different nucleotides or amino acids can only produce six different trimers.

Given that we have nine basic monosaccharides, N-glycans typically include between ten and twenty monosaccharides each as well as branching points (see Figure 1.6), the space of possible configurations explodes into an untreatable number. However, only a few thousands of different glycan structures have so far been discovered in biological systems [7,10], which strongly indicates that glycan synthesis is highly regulated at a cellular level.

1.1.2 Biosynthesis

Principles of glycan biosynthesis All glycans are the product of chemical reactions catalyzed by enzymes that add (*glycosyltransferases*) or remove (*glycosidases*) single monosaccharides. Most glycosylation reactions use *activated forms* of monosaccharides as donors (typically nucleotide sugars [11]), where different monosaccharides require different nucleotides (Figure 1.4). Glycosyltransferases have a high specificity for sugar donor and acceptor and their activity is usually also specific to the glycosylation site and substrate configuration [12]. Because of this strict specificity, most glycosyltransferases can only add one type of monosaccharide in a specific linkage configuration [12], and often different enzymes are necessary to catalyze the same reaction on different substrates and proteins [13]. To generate all the diverse glycan structures observed in human glycoconjugates, roughly 250 different glycosyltransferases are estimated to be coded in the genome [14]. Glycosi-

[1]In a sequence of three hexoses we have two glycosidic bonds with two linkage possibilities each. The hexoses can be chosen from three different molecules (Glucose, Mannose, Galactose) and each pair of hexoses leads to eight different combinations. Therefore, for any sequence of three hexoses, the total number of possible combinations becomes $8 \cdot 8 \cdot (3 \cdot 2 \cdot 1) = 384$.

dases have similar specificity properties, with approximately 100 different enzymes coded in the human genome [15, 16].

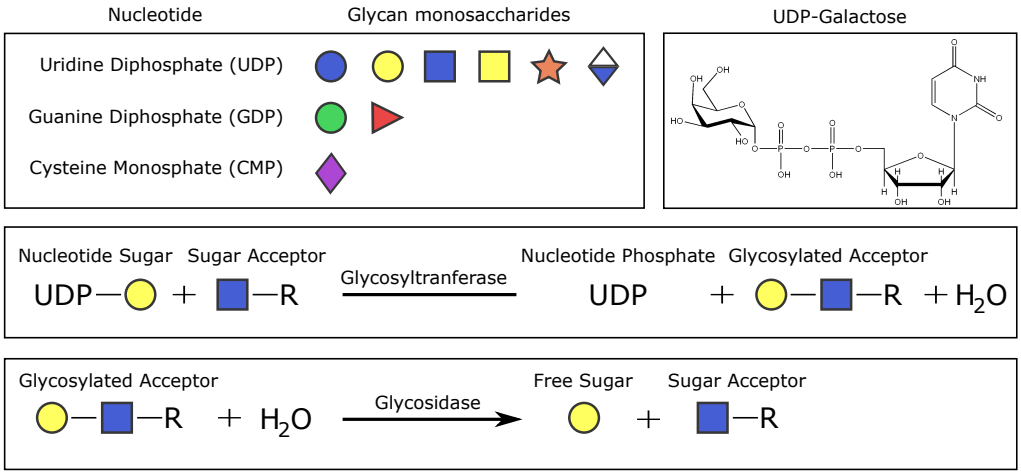


Figure 1.4: Activated sugars and example reactions of glycosylation. Glycosyltransferases need nucleotide sugars as donors. Different monosaccharides need different nucleotides: Man and Fuc need GDP, Neu5Ac needs CMP, and all others need UDP. During a typical glycosylation reaction, a glycosyltransferases catalyzes the attachment or removal of a monosaccharide to or from an N-glycan (R).

Glycan synthesis pathways Glycans found on proteins are usually classified in two major classes, according to how they are linked to the protein backbone:

1. **N-linked glycans**: are linked to the amide nitrogen atom of the side chain of an asparagine (Asn) in the motif Asn-X-Ser/Thr, where X must not be proline (Pro) [17]. In animal cells, the monosaccharide linked to the asparagine residue is almost exclusively GlcNAc [18].
2. **O-linked glycans**: are linked to the oxygen atom of a serine (Ser) or threonine (Thr) residue in a polypeptide. In this case the connecting monosaccharide is often a GalNAc [19].

Other than their linkage to the protein, N- and O-glycans differ significantly in their synthesis processes and hence in their structures. While O-linked glycans are synthesized by adding monosaccharides one at a time directly on the protein, the biosynthetic pathway of N-linked glycans is more complex, and can be divided into three spatially separated stages:

1. *Formation of a lipid-linked precursor oligosaccharide.* A characteristic 14-monosaccharide-chain (Glc2Man9GlcNAc2) highly conserved across species [20], is built onto a dolichol molecule in the Endoplasmatic Reticulum (ER) (Figure 1.5A).

2. *En bloc transfer of the oligosaccharide to the protein.* The precursor is co-translationally transferred *en bloc* to the protein during folding (Figure 1.5B, black oval).
3. *Processing of the oligosaccharide on the protein.* The protein travels through the Golgi apparatus, where glycosyltransferases and glycosidases further modify the attached glycan (Figure 1.5B). In the *cis*-Golgi, the glycan is first trimmed to produce Man5GlcNAc2. Glycans escaping this trimming will result in high mannose glycan structures (Figure 1.6, left). Structures processed to Man5GalNAc2 can be further modified to produce hybrid and truncated structures (Figure 1.6, center). In the *medial*-Golgi, the GlcNAc of the first antenna is added to Man5GlcNAc2. This allows further trimming of the mannoses and the initiation of the formation of branching. In the *trans*-Golgi, the structures are extended into complex glycans (Figure 1.6, right).

1.1.3 Regulation

For glycosylation there is **no direct genetic template**, and it is estimated that roughly 2% of the human genome encodes proteins involved in glycan biosynthesis, degradation or transport [21]. The final glycosylation state of a protein or cell is therefore determined by numerous factors, including (i) the availability of nucleotide sugar donors in the ER and Golgi apparatus, and (ii) the expression, activity and localization of the glycosylation enzymes, which often compete for the same substrates and can be site-, protein- and tissue-specific. Moreover, protein glycosylation is a highly dynamic process that quickly adapts to changes in the surrounding environment [22], and hence the same protein or cell can express very different glycosylation profiles in response to different stimuli.

For all these reasons, the final glycosylation state of a given site on a given protein synthesized by a particular cell type is not unique, and, to the contrary, all glycosylation sites exhibit a variety of different attached glycan structures. The extent of this effect, known as *microheterogeneity*, varies considerably from one glycosylation site to another and from glycoprotein to glycoprotein, and in some cases it can have prominent consequences on the protein function. For example, different glycans attached to the Fc region of Immunoglobulin G can have opposite effects, i.e., leading to pro- or anti-inflammatory protein activity (see Section 1.1.6 for a more detailed description).

Although the general mechanism of protein N-glycosylation is quite well established, the investigation of the site-, protein-, or cell-specific pathways of glycosylation *in vivo* is, to this day, experimentally infeasible due to the enormous complexity of the process. *In vitro* experiments have allowed to identify the substrate specificities of all major glycosyltransferases [23] and localization experiments on cell lines have helped in determining the localization of many enzymes within the ER and Golgi-apparatus [24–28].

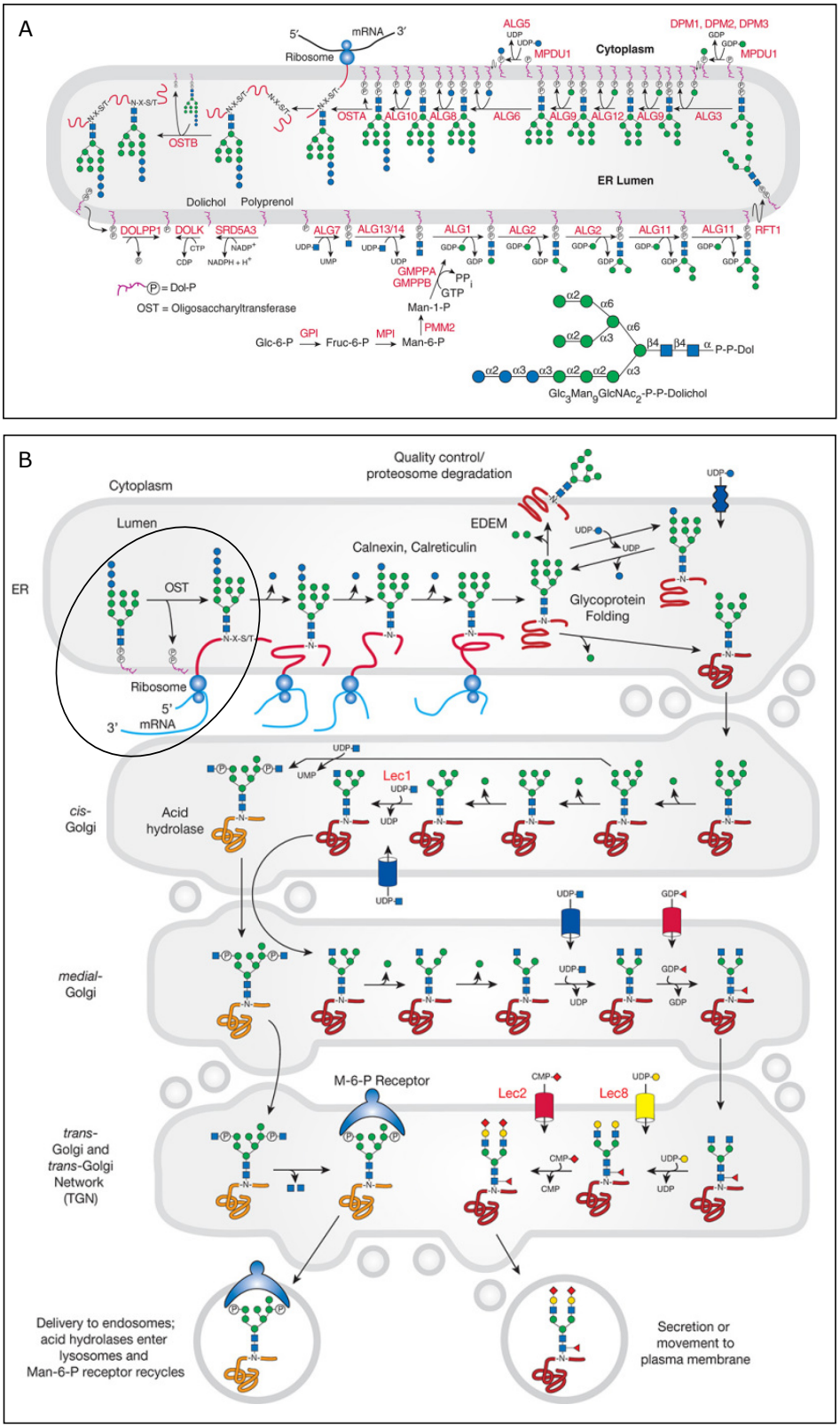


Figure 1.5: N-glycan biosynthesis. Adapted from "Essentials of Glycobiology" [2]. **A** Formation of the lipid-linked oligosaccharide. **B** *En bloc* transfer of the oligosaccharide to the protein and further processing.

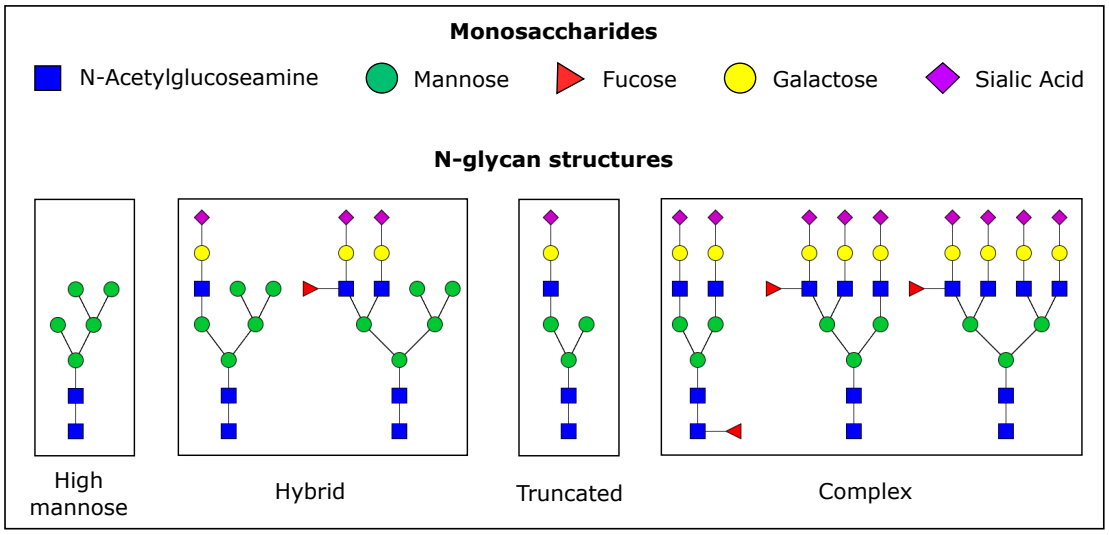


Figure 1.6: Common N-glycan structures.

However, given that glycosylation is highly site-, protein- and cell-specific, it is possible that results obtained on a particular cell type or protein do not generalize well to other systems. For example, the *in vitro* specificities of glycosylation enzymes might not correctly represent the *in vivo* ability of a given glycosyltransferase to act on specific proteins. A valuable tool to investigate both the synthesis and the function of protein glycans is the analysis of mice where one or more glycosyltransferase genes have been knocked out [29, 30]. However, given that glycosylation is highly site-, protein- and cell-specific, it is possible that results obtained on a particular cell type or protein do not generalize well to other systems. For example, the *in vitro* specificities of glycosylation enzymes might not correctly represent the *in vivo* ability of a given glycosyltransferase to act on specific proteins. A valuable tool to investigate both the synthesis and the function of protein glycans is the analysis of mice where one or more glycosyltransferase genes have been knocked out [29, 30].

However, even in the very early stages of development, severe alterations to the normal glycosylation of proteins is fatal to the organism [31], and hence a systematic analysis approach is again infeasible.

In conclusion, although indications of the specificities and activities of enzymes involved in protein glycosylation can be derived from *in vitro* assays and model systems, a proper *in vivo* validation on the site-, protein- and cell-specific activities of these enzymes is lacking.

To overcome this *enpasse*, we generated hypotheses on site- and protein-specific synthesis

pathway from the analysis of population omics data. This type of approach has proven to be a powerful tool, in particular in the field of metabolomics [32]. In this thesis, we tackled the problem of determining protein- and site- specific glycosylation pathways by analyzing four large-scale glycomics cohorts (**Research question I**, Section 1.4). In the project described in Chapter 3, we were able to identify known biochemical steps, as well as validate new ones with *in vitro* experiments.

1.1.4 Function

Given their ubiquitous and diverse nature, it is not surprising that glycan functions are diverse, and their effect on the activity of the protein they are bound to can vary from very subtle to critical for the development, growth and survival of an organism [31]. It is therefore difficult to establish a general structure-function relationship for glycans, considering that (a) there is no direct genetic template, (b) each protein can have multiple glycosylation sites with different structures attached, and (c) the same structure on different proteins can have different functions. The analysis of specific glycan structures on specific proteins has nevertheless allowed to establish so far five main glycan functions [2], which can be classified into two broad categories, according to whether they affect the carrying protein (*intrinsic*), or mediate the interaction with other molecules (*extrinsic*):

- Providing structural components (cell walls, extracellular matrix)
- Modifying protein properties (stability, solubility)

} Intrinsic Functions

- Directing glycoconjugates trafficking (intra- and extra-cellular)
- Mediating and modulating signaling (intra- and extra-cellular)
- Self identification (immune response)

} Extrinsic Functions

N-linked glycosylation can, for example, affect protein conformation, solubility, signaling, antigenicity and protein-protein interactions. A concrete example of the prominence of these effects is given in Section 1.1.6.

1.1.5 Relevance in human physiology and disease

In a report aimed at articulating "*a unified vision for the field on glycoscience and glycomics*", the National Academy of Science (USA) recognized in 2012 that ***glycans are directly involved in the pathophysiology of every major disease*** [4]. Although for

most diseases the molecular mechanisms are still unknown, changes in the glycosylation profiles of proteins and cells have been observed in a fast increasing number of pathologies, including but not limited to: Congenital Disorders of Glycosylation (CDG) [33], cardiovascular diseases [34,35], Alzheimer's disease [36–38], rheumatoid arthritis [39], inflammatory bowel disease [40], lupus [41], diabetes [42–45], HIV [46–48], and cancer [49–54]. Moreover, since glycosylation adapts quickly to reflect changes in the cell state, glycans and *glycan-binding-proteins* (GBPs) are being investigated more frequently as possible drug targets, in particular for HIV [55–58] and different types of cancer [59, 60]. Despite this, many branches of the biological sciences still ignore the presence of glycans when describing disease etiologies and mechanisms.

Since providing an exhaustive list of glycans' involvement in the development and progression of diseases is out of the scope of this thesis, we describe in this section two concrete examples that illustrate how understanding the molecular mechanism of glycosylation and protein-glycan interactions can improve medical procedures and treatment of diseases.

ABO blood groups. Probably one of the most important discoveries in the history of glycobiology, the identification of different groups in human blood dates back to the beginning of the 20th century [61] and allowed the development of the blood transfusion, a widely used procedure in modern medicine. The ABO blood groups are controlled by a single gene (the *ABO gene*), which can have three alleles (i, I^A and I^B), each of which codes for a different glycosyltranferase. These glycosyltransferases act on the terminal part of the glycans attached to several proteins and lipids on the surface of erythrocytes. As a consequence, depending on which allele is expressed, these cells will exhibit different glycan structures on their glycocalyx.

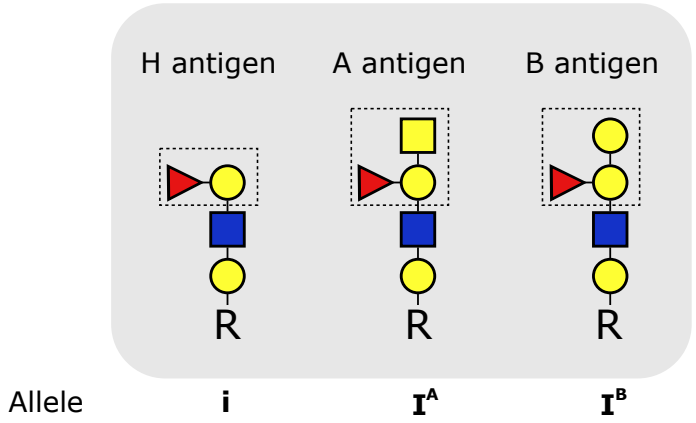


Figure 1.7: ABO blood groups antigens on N-glycans (R).

Allele i produces antigens H, which translates to blood group O, allele I^A gives rise to antigen and blood type A, while allele I^B creates antigen and blood type B (Figure 1.7). Allele i is recessive, and alleles I^A and I^B are codominant. Therefore, blood group O needs two i alleles, while the presence of both A and B alleles will give rise to both antigens (blood type AB). Note that antigens A and B only differ for a single chemical functional group (the N-Acetylgalactosamine in antigen A has an N-acetyle group whereas the galactose in antigen B has a hydroxyl group, see Figure 1.2) and yet the human immune system is so sensitive to structural differences that individuals with blood type A cannot receive blood from a type B donor and viceversa.

Influenza virus. Influenza is still a major global health problem: more than *70 million deaths* have been attributed to the influenza pandemic of 1918 [63]. One of the most alarming properties of influenza is that it has the ability to be transmitted across species, for example from swine to human, as it recently happened with the strain H1N1 [64].

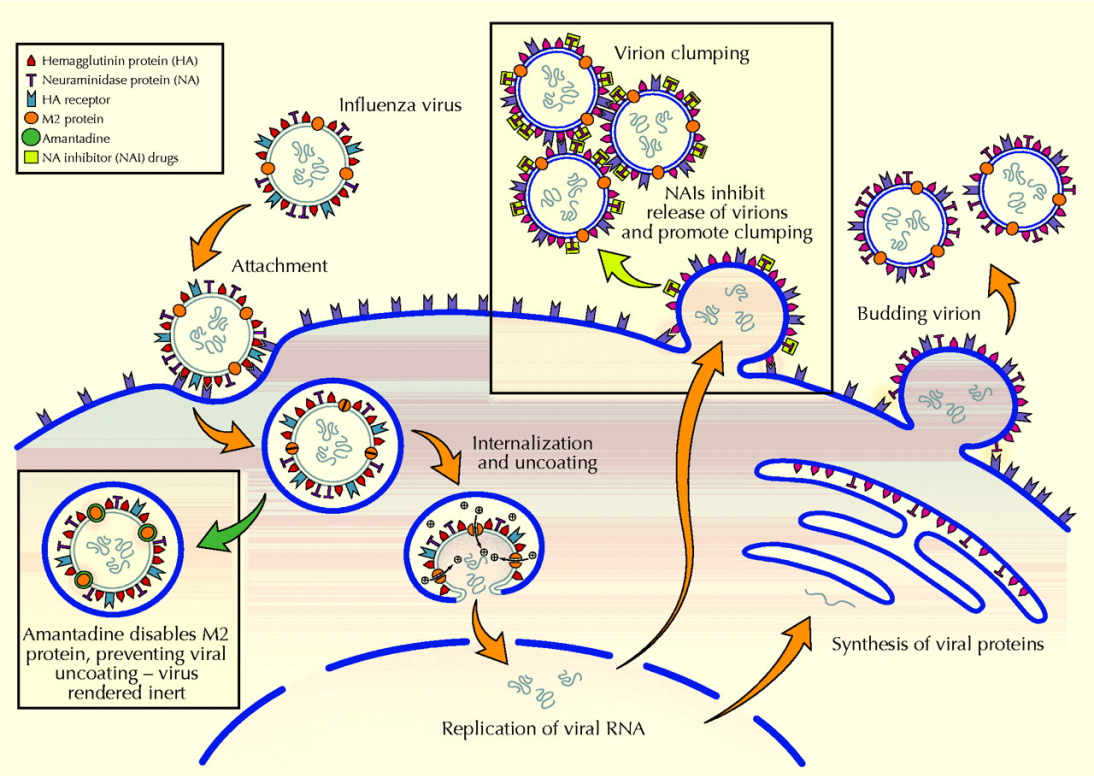


Figure 1.8: Life cycle of the influenza virus. Hemagglutinins on the membrane of the virus bind to the sialic acids on the glycocalyx of the host cell and triggers endocytosis. After viral DNA replication, new viral particles bud off the host cell and neuraminidases cleave the sialic acids bound to hemagglutinins, enabling the new viruses to infect other cells. Taken from Stiver (2003) [62].

Influenza is a membrane enclosed virus and has two important proteins protruding from its membrane: Hemagglutinin (H), a receptor for sialic acid, and Neuraminidase (N), an enzyme that catalyzes the cleavage of sialic acid off of the host cell[2]. In the very first stages of the infection, as the virus approaches the host cell, hemagglutinin attaches to sialic acids on the host cell surface, enabling the docking of the viral particle. This triggers the *endocytosis* of the viral particle into the host cell. The virus then is able to release its native nucleic acid into the cell and initiate its replication. Eventually, new particles will bud off the cell surface of the host cell. However, hemagglutinin keeps the newly replicated virions bound to the surface of the host cell. In order for the virions to release themselves and start their own cycle of replication, neuraminidases cleave off the sialic acids bound to the hemagglutinins on the virus particles (Figure 1.8). Thanks to the in detail understanding of the molecular mechanism of influenza infection, pharmaceutical industries are now able to develop antiflu drugs based on synthetic compounds that mimic the structure of sialic acid and inhibit the action of the neuraminidase, therefore limiting the spread of the infection.

1.1.6 Immunoglobulin G

Due to their accessibility, secreted glycoproteins are the best characterized glycoconjugates. Among them, *Immunoglobulin G* (**IgG**) is the most abundant and investigated. It is a large molecule of about 150 kDa made of four peptide chains: two identical heavy chains and two identical light chains. The two heavy chains are linked to each other and to a light chain each by disulfide bonds. The resulting tetramer has two identical halves, which together form the typical antibody Y-shape (Figure 1.9). IgG has four isoforms or *subclasses*. Like all antibodies, it is produced and secreted by B lymphocytes and has two functional domains, namely an *antigen-binding fragment* (Fab), which is responsible for recognizing antigens on foreign pathogens and infected cells, and a *crystallizable fragment* (Fc), which triggers the immune response by interacting with various Fc receptors [65].

Although secreted proteins can have up to 20 different glycosylation sites, IgG only has four main ones: one on each side of the antigen-binding portion of the protein, which are however only actively glycosylated in 15 to 20% of the cases [66–68], and one highly conserved glycosylation site in each heavy chain of the Fc region (at Asn 297) (Figure 1.9) [69]. While the functional effects behind Fab glycosylation are still unclear [70], Fc glycosylation is well-characterized. From a structural point of view, glycans are mostly of the complex type and biantennary, with the possible addition of a *core fucose*, namely a fucose attached to the first GlcNAc of the glycan structure, or a *bisecting GlcNAc*, i.e.,

[2]The flu strains owe their name to the particular forms of hemagglutinin and neuraminidase that are found on the membrane of the virus, for example H1N1.

a GlcNAc attached to the first mannose of the glycan structure, to which no additional monosaccharides can be added (Figure 1.9). Contrary to most other glycoproteins, where the glycans are protruding from the surface, Fc glycans on IgG are buried within the hydrophobic core of the protein (Figure 1.9) and therefore even small changes in the sugar composition can have prominent effects on the structure of the protein and its affinity to Fc receptors [71, 72]. For example, the presence of a core fucose modifies the structure of the Fc region and dramatically reduces its ability to bind to the receptor FcγRIIIa [73], which triggers the initiation of *Antibody-Dependent Cellular Cytotoxicity* (ADCC), which, in turn, results in the destruction of target cells. IgG with glycans lacking a core-fucose are more than 100 times more effective in initiating ADCC through binding to FcγRIIIa [74, 75]. However, on average, roughly 95% of IgG Fc glycans are core-fucosylated [76], and this indicates that the mechanism is tightly regulated. Moreover, the addition of sialic acid is able to actively revert IgG's functionality from pro-inflammatory to anti-inflammatory [77]. Therefore, alternative glycosylation of IgG is exploited to effectively modulate the protein structure and enable vastly different functionalities.

Furthermore, alterations in the expected glycosylation profiles of IgG have been linked to numerous diseases, including autoimmune diseases [41, 78], rheumatoid arthritis [39], diabetes [79], and some types of cancer [50, 53, 80, 81], although the involvement of the protein in the etiology of the diseases remains unclear. Elucidating the molecular mechanisms of IgG synthesis and regulation would help in better understanding how IgG contributes to the antibody-based immune response.

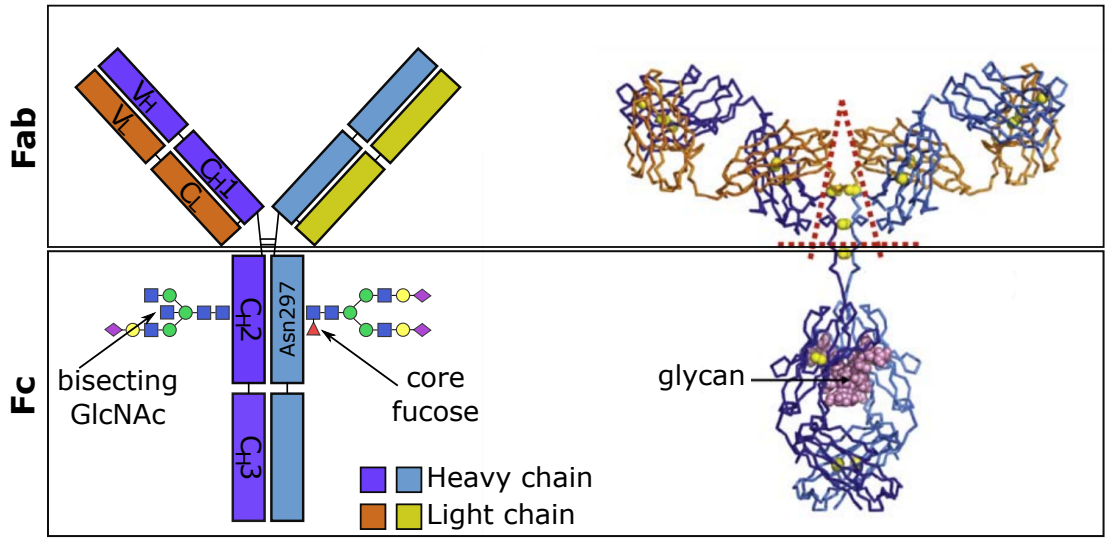


Figure 1.9: Structure of Immunoglobulin G. Adapted from Jefferis (2009) [82].

1.2 Glycomics measurement

Given the ubiquitous nature of protein glycosylation and its potentially critical effect on the structure and function of glycoproteins, increasing attention is being directed towards the systematic analysis of glycan structures. Since the functional effect on any given glycan is site- and protein-specific, the ultimate goal of *glycomics*, namely the study of all glycoforms on a protein, a cell or tissue, would require to somehow retain the information about the site and protein of origin of each glycan structure measured. Unfortunately, this is to this day still technologically infeasible. Two alternative approaches have been developed to investigate glycan structures on proteins on a large-scale basis: (i) For isolated proteins, glycan structures can be analyzed in a *site-specific* fashion by measuring a small peptide around the glycosylation site (*glycoproteomics*), which in most cases differs from site to site, together with each glycan; (ii) N-glycans can be released from all glycosylation sites on either a purified protein or a mixture of proteins. In this thesis we will analyze data from both scenarios.

1.2.1 Measurement platforms

The three approaches to quantify protein glycosylation considered in this thesis (UPLC, LC-ESI-MS and MALDI-TOF-MS) are based on two widely used measurement technologies: *Liquid Chromatography* (**LC**) [83] and *Mass Spectrometry* (**MS**) [84], which will be briefly discussed below.

Liquid Chromatography. LC sample analysis relies on pumps to pass a pressurized liquid solvent containing the sample through a column filled with a solid adsorbent material [83]. Due to their different chemical and physical properties, each component in the sample will interact slightly differently with the adsorbent material, and each component will therefore have different flow rates, or *elution times*, leading to their separation as they flow through a column.

LC performed at high pressure (about 1200 atmospheres) is referred to as *Ultra High-Performance Liquid Chromatography* (**UPLC**) [85].

Mass Spectrometry. The MS analysis of the compounds in a sample consists of three main steps: first, the molecules are ionized; the resulting ions are then separated by a magnetic field according to their charge and mass; finally, a detector captures and quantifies the separated ions, producing a spectrum of intensities as a function of the mass-to-charge ratio (m/z) [84]. Molecular ionization can be achieved with the *Matrix Assisted*

Laser Desorption Ionization (**MALDI**) [86] technology, which involves crystallizing the sample on a metal target and ionizing it with a laser pulse.

MALDI sources are usually attached to *Time of Flight* (**TOF**) [87] analyzers, which amplify the ions flight path through an electric field. This combination allows the identification of ions with a very high molecular mass (> 200 kDa). Alternatively, the *ElectroSpray Ionization* (**ESI**) [88] technique provides another efficient protocol to create molecular ions for mass spectrometry. Here, ionization is achieved by applying a high voltage to a liquid solution containing the sample. ESI-MS can be coupled to LC allowing *on-line* LC-ESI-MS analysis [89], where the liquid eluting from the LC column is directly fed into the ESI machine.

The costs associated to the platforms considered in this thesis are summarized in Table 1.1.

Table 1.1: Cost of glycomics measurements. Adapted from Huffman et al. (2014) [90].

	UPLC	**LC-ESI-MS**	**MALDI-TOF-MS**
Throughput per instrument	Medium low (∼50 samples/day)	Medium high (∼100 samples/day)	Very high (<1 minute/sample)
Cost of equipment	40k - 70k Euros	200k - 500k Euros	100k - 500k Euros
Cost per sample	Rather high due to low throughput and cost of consumables	very high due to expensive equipment and relatively low throughput	low due to high throughput

1.2.2 Measured glycans

The three platforms described above have been used to measure the three types of glycomics data investigated in this thesis and presented in Chapters 3, 4, 5, and 6. Each of these techniques has advantages and disadvantages, which are outlined below.

- **UPLC: Total IgG glycans** N-glycans from both the Fc and Fab region of IgG are measured.

  Advantages: Provides branch-specific information, i.e., it is able to differentiate between the 3-arm and 6-arm (corresponding to the left and right antenna in our graphical representation, respectively) IgG glycan isomers , due to a slightly higher retention time of the 3-arm isomer [90].

Disadvantages: 1. The information about the glycosylation site is lost [90]. 2. Peaks in the chromatogram do not necessarily correspond to single glycan structures, as different structures with similar elution times contribute to the same peak [76].

- **LC-ESI-MS: IgG Fc glycopeptides** IgG Fc glycans are measured together with a short peptide in proximity of the glycosylation site.

  Advantages: 1. Provides structural information on the measured glycoforms [90]. 2. Since the peptide sequence in proximity of the Fc glycosylation site is different for different IgG subclasses, it provides site- and subclass-specific glycosylation profiles [91].

  Disadvantages: Difficult to perform for more than one glycosylation site.

- **MALDI-TOF-MS: Total plasma N-glycome (TPNG)** N-glycans from all proteins in plasma are measured together.

  Advantages: Provides the masses of a large number of glycans.

  Disadvantages: 1. The information about the protein of origin and the corresponding glycosylation site is lost. 2. Given the presence of epimers in the glycan building blocks, different structures could contribute to the same mass, or *composition*, and therefore structural information is not directly available from the spectra [92].

Glycosylation measurements from isolated proteins are valuable to understand protein-specific synthesis and regulation. However, a more general footprint of the overall glycosylation profile of an organism, like the total plasma N-glycome, can provide relevant information about alterations in the system's homeostasis, for example in the presence of diseases [93] and inflammation [94]. Given the strong regulation of glycan biosynthesis, the structural details of the plasma N-glycome are essential in order to be able to track alterations in the observed glycosylation profiles back to molecular mechanisms. In this thesis, we contribute to a better understanding of the structural properties of the TPNG by providing a data-driven approach to infer structural details from glycan compositions quantified by MALDI-TOF-MS (**Research question II**, Section 1.4).

1.3 Glycomics analysis

The main objective of glycomics analysis is to infer biologically relevant information on the structure, synthesis, regulation and function of glycans from the analysis of large-scale datasets. Typically, measured and derived glycan traits are correlated to disease phenotypes to identify accessible biomarkers [95–100] and predictors of disease outcomes [101–103]. However, the observed associated traits are often difficult to link to specific

glycan structures and functions without available prior information [104]. In this thesis, we study the available glycomics data focusing on proper data processing prior to analysis, as well as glycan pathway and structure inference, which provide valuable information for a better characterization of glycobiology processes.

1.3.1 Data preprocessing

Regardless of the chosen measurement platform, glycan data are susceptible to a *systematic bias*, due to technical variations from non-biological sources, originating for example from small alterations in experimental conditions, sample preparation, temperature, or instrument calibration. The exact sources of the bias are usually unknown, but its presence can significantly affect the outcome of any downstream analysis. If the goal is to extract biologically meaningful information from the data, the effect of technical variations should, as far as possible, be corrected for prior to analysis. The process that aims at reducing such bias from the measured data is referred to as *normalization*.

One of the most common procedures to normalize glycan data is *total area normalization* [105], which requires dividing the intensity of each glycan by the sum of the intensities of all measured glycans. Once normalized, each entry represents what *percentage* of the total sample intensity the corresponding glycan contributes. This allows for an intuitive interpretation when, for example, comparing the glycome composition of different individuals [106], but imposes strong constraints on the correlation structure of the data (see Subsection 2.1.3 for details). Briefly, the constant sum constraint (each sample's percentage always sums up to 100%), introduces spurious values in the correlation structure, and, therefore, any observed correlation among variable pairs might be a result of the constraint introduced by the normalization method and not represent true biological associations.

Even regarding this approach as unsuitable for correlation analysis, however, the problem of which normalization method to select from the several available remains. Different approaches have different underlying assumptions and can introduce additional bias in the data if not suitably chosen. Systematic comparisons of commonly implemented normalization approaches have been performed in recent years for many omics data, e.g. transcriptomics [107], proteomics [108], as well as metabolomics [109–111], but an analogous study for glycomics data is still missing. In this thesis, we addressed the question of glycan normalization using an innovative approach, based on a *biological measure* of quality. The idea is to rank each method according to its ability to preserve known biochemical interactions among the variables (**Research question III**, Section 1.4).

1.3.2 Network inference

The inference of biological information from large-scale omics measurements has been one of the big challenges in systems biology [112]. The observed inter-individual variation in the concentration of various types of molecules, from proteins [113] to metabolites [114, 115], has been exploited to gain insights into their synthesis [32, 116] and regulation [117], as well as into their involvement in clinical phenotypes [118, 119].

One popular approach to extract biologically relevant interactions from measured molecular concentrations is based on the computation and subsequent analysis of the *data correlation structure*. Correlations quantify dependencies among pairs of variables and can be computed with a number of different approaches. The most common method, known as **Pearson correlation** [120], estimates the linear associations between variables. When dealing with high-dimensional omics data, however, this usually leads to a highly correlated system, where single coefficients are difficult to associate to specific biochemical or regulatory effects. An efficient alternative to this approach is **partial correlation** [121], which accounts for the presence of confounding variables and covariates when estimating pairwise correlations, therefore filtering out *indirect* interactions, i.e., those that are exclusively due to the mediating effect of one or more other variables (see Subsection 2.3.1 for details).

Since population-based measurements of biological molecules are intrinsically noisy, calculated pairwise correlation coefficients must undergo a *selection criterion* to determine which coefficients are to be considered *significantly* different from zero. This criterion is usually based on *statistical* principles, which allow to control in different ways the amount of expected **false positives**, i.e., correlation coefficients that arise because of noise and that do not represent true molecular associations. The two most commonly used statistical approaches to perform this selection, known as **multiple testing correction**, control either (i) the *proportion* of false positives, or **False Discovery Rate** (FDR) [122], or (ii) the *probability* of including at least one false positive, or **Family-wise Error Rate** (FWER) [123]. Only those coefficients that pass the chosen selection criterion will be defined as *significant*, i.e., representing true molecular interactions, and will be considered for further analysis.

The data correlation structure can be easily visualized as a **network**, where nodes represent measured variables and edges the significant correlation coefficients among them. In the case of partial correlation, these networks are also known as **Gaussian Graphical Models** (GGMs) and have been shown to represent known molecular interactions in different data types [115, 124–129]. In metabolomics data, for example, GGMs selectively identify single biochemical steps in metabolic synthesis pathways [32]. In this thesis, we

will apply GGMs to glycomics data for the first time and show that significant partial correlation coefficients represent known biochemical steps in the glycan synthesis pathways (Chapter 3 and 4).

Since, however, different statistical selection criteria have different basic assumptions, the resulting networks could be vastly different. All of these network are *statistically* correct, but do not necessarily convey the most accurate representation of the underlying biological mechanisms. Therefore, instead of selecting correlation coefficients based on statistical significance, we define a *biological* selection criterion, which *maximizes* the biological information contained in the resulting correlation network (**Research question IV**, Section 1.4).

1.4 Research questions

Due to the lack of a direct genetic template and the huge number of molecules involved in its regulation, glycosylation is an extremely complex biological process. Found on the great majority of membrane and secreted proteins, glycans can dramatically alter protein structure and interactions. Different measurement technologies now enable the quantification of glycan moieties from thousands of samples, and the analysis of large-scale glycomics datasets can help elucidating several aspects of protein glycosylation that could be difficult to test at a cellular level. Pairwise correlations are a powerful tool for the identification of single enzymatic steps in synthesis pathways, but a selection criterion needs to be defined to determine which correlation coefficients are to be considered significant.

However, several aspects need to be better characterized to fully unlock the potential of the currently available glycomics data technology. Within the scope of this doctoral thesis, four main research questions were addressed, which contribute to a better understanding of the biology of protein glycosylation, as well as to the development of a more efficient inference pipeline for glycomics analysis.

I *Can we infer new site- and protein-specific enzymatic reactions of glycan synthesis from large population glycomics data?*

II *Which glycan structures contributing to the Total Plasma N-glycome can be inferred from mass spectrometry measurements?*

III *What is the most appropriate normalization approach for glycan data?*

IV *Is there a more biologically meaningful cutoff for correlation networks than the ones obtained with statistical methods?*

1.5 Thesis overview

In the following, a short outline of the content of this thesis is provided. A graphic representation is provided in Figure 1.10.

The Materials and Methods chapter (**Chapter 2**) first presents the glycomics datasets analyzed in this work, highlighting their features and differences. In the second part, we describe the main statistical approaches used in this thesis. We introduce the basis for network inference, starting with correlation measures and multiple testing correction, and the approaches to analyze the resulting networks, like pathway analysis and network modularity. In the last section, we illustrate how a Genome-Wide Association Study is performed.

All the projects presented in this thesis are the result of collaborations with several laboratories and institutions. While data were provided by these partners, I contributed the vast majority of the statistical and computational analysis, as well as the result interpretation and discussion. I appeared or will appear as sole first author of all publications deriving from each of the project listed below.

The four main results chapters are divided according to their goal: while the first two (Chapters 3 and 4) demonstrate how the data correlation structure can be used to infer biological information from glycomics population studies, the last two (Chapters 5 and 6) use this correlation to investigate technical aspects of the network inference pipeline.

In **Chapter 3**, we infer new enzymatic reactions in the IgG glycosylation pathway. We analyze the correlation structure of four large-scale glycomics cohorts and show that partial correlation networks selectively identify known biochemical steps in the IgG glycan synthesis pathway. We then use this relationship to generate hypotheses on possible new enzymatic steps based on the data-driven network and validate the findings in a set of experiments, using *in vitro* enzymatic assays and enzyme localization in cell lines.

Chapter 4 deals with the analysis of glycomics measurements from all plasma proteins in a large human population, wherein the information about the glycosylation site or protein of origin is unknown. Moreover, only the masses of the glycan moieties are available, which means that multiple structures can contribute to the same mass, or *composition*. Here, we use the data correlation structure and prior knowledge on the synthesis pathway of

glycans to infer the single glycan structures within the compositions that contribute most to the observed correlation structure. Our predictions are then validated with external data and demonstrate that, in most cases, our data-driven approach is reliable.

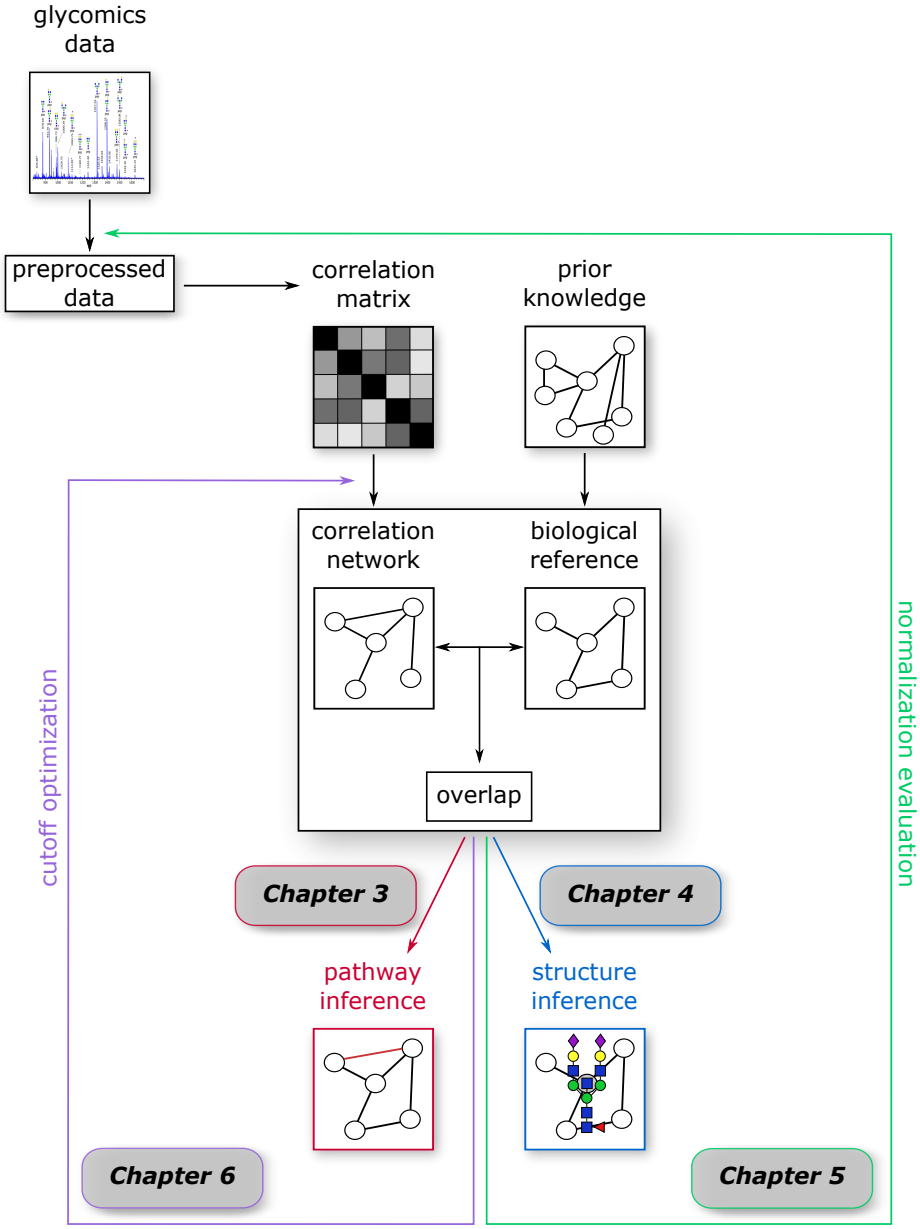


Figure 1.10: Overview of the content of this thesis. We first describe how we can compute data-driven correlation networks and compare them to the available prior knowledge by defining a quantitative measure of overlap. This measure is then exploited to either extract information about the underlying biological system, such as new biochemical reactions (**Chapter 3**) and structural details (**Chapter 4**), or to improve the pipeline of network inference, in particular the data preprocessing (**Chapter 5**) and significance selection (**Chapter 6**).

In **Chapter 5** we show how the relationship between the data-driven correlation network and prior knowledge can be exploited to evaluate different preprocessing approaches. We apply different normalization strategies to several glycomics data and evaluated which normalization method leads to the best biological network. Specifically, from each normalized dataset we infer a correlation network and subsequently quantify the performance of each normalization approach according to how well the inferred correlation network is able to recover single enzymatic steps in the known glycosylation pathway. In this way, we have a *biological measure* of quality for the assessment of performance. By applying the approach to six different glycomics datasets across three platforms, we are able to identify the methodology that performs best overall.

Chapter 6 tackles the question of determining a correlation cutoff for network reconstruction. This step is usually determined *statistically* via multiple testing correction methods. However, statistical significance is not necessarily related to the underlying biology. We therefore investigate whether we can define a significance cutoff based on the ability of the network to correctly represent biological interactions. We once again use the comparison between data-driven correlation networks and prior knowledge to optimize the network correlation cutoff and show that the approach works even when the prior knowledge is incomplete or partially incorrect. To demonstrate the generalizability of these findings, we replicate the results in one metabolomics and one transcriptomics dataset.

In the final **Chapter 7**, the main results of this thesis are summarized and the scientific contributions of this work in the context of the field are discussed, together with possible extensions and potential future projects.

Chapter 2

Materials and Methods

In this chapter, we introduce the data and statistical analysis tools that will appear throughout the thesis. In the first section, we present the different glycomics dataset analyzed in this thesis (IgG Fc glycopeptides, IgG total N-glycans, total plasma N-glycome), together with their corresponding measurement platform (LC-ESI-MS, UPLC, MALDI-TOF-MS). The data preprocessing protocol is then described, as well as the biochemical pathway of glycan synthesis specific to each dataset. The second section focuses on the core statistical and computational methods developed and used in the rest of this thesis.

2.1 Glycomics data

The glycomics data analyzed in this work were measured by our collaboration partners at Genos Laboratories and Leiden University Medical Center. In the following, we briefly describe the considered cohorts and measurement platforms, and provide an overview of the features of each considered dataset, as well as of the differences between the measured glycan profiles.

2.1.1 Cohorts and platforms

IgG Fc glycopeptides - LC-ESI-MS

Cohort Human plasma samples were obtained from four Croatian population studies: two from the Croatian islands of Vis and Korčula, which are part of the "10001 Dalmatians" biobank [130], and a second cohort from Korčula and one cohort from Split collected separately a few years later (see Table 2.1 for details). Overall, the four cohorts includes 2,548 samples from individuals ranging from 18 to 91 years of age.

Data acquisition To generate the glycomics data from these cohorts, IgG was isolated [90, 131] and glycopeptides from the Fc region of the protein were extracted through trypsin digestion and measured by Liquid Chromatography - Electrospray Ionization - Mass Spectrometry (LC-ESI-MS). Because different IgG subclasses have different amino acid sequences around the glycosylation site [82,91], this platform allows the measurement of subclass- specific IgG Fc glycosylation. However, in Caucasian populations, the tryptic Fc glycopeptides of IgG2 and IgG3 have identical peptide moieties [82,91] and hence were not distinguishable with this profiling method. The final spectra included 50 different structures: 20 for IgG1, 20 for IgG2 and IgG3, and 10 for IgG4 (due to low abundance).

Total IgG N-glycans - UPLC

Cohort Human plasma samples were obtained from the Study of Colorectal Cancer in Scotland (SOCCS), a case–control study designed to identify genetic and environmental factors associated with nonhereditary colorectal cancer risk and survival outcomes [132]. For the purpose of this analysis, only the 535 control samples were considered (see Table 2.1 for details).

Data acquisition To generate the glycomics data from this cohort, N-glycans were first released from isolated IgG with peptide N-glycosidase F (PNGase F), an enzyme able to specifically cleave N-linked glycans from their corresponding Asn [133]. This procedure allows for the release of *all* IgG glycans, both from the Fc and the Fab region of the protein (see Figure 1.9). Released glycans were labelled with 2-aminobenzamide (2-AB) and separated by hydrophilic interaction via Ultra high-Performance Liquid Chromatography (UPLC) . Although this methodology loses the information about the glycosylation site, UPLC-based glycomics has the advantage of providing branch-specific information, i.e., it is able to differentiate between the 3-arm and 6-arm glycan isomers due to a slightly higher retention time of the 3-arm isomer. The final chromatogram included 24 distinct peaks.

Total plasma N-glycome - MALDI-TOF-MS

Cohort Human plasma samples were obtained from the Leiden Longevity Study (LLS), a family-based study comprising the offspring (and their partners) of 421 nonagenarians sibling pairs of Dutch descent [134]. A total of 2,056 individuals were included in the analyses described in this thesis (see Table 2.1 for details).

Data acquisition To generate the glycomics data from this cohort, N-glycans were released from all N-glycosylation sites on all plasma proteins (total plasma N-glycome, TPNG) with PNGase F and analyzed via Matrix Assisted Laser Desorption/Ionization - Time Of Flight - Mass Spectrometry (MALDI-TOF-MS) [134]. This technique al-

Table 2.1: Glycomics data.

	Korcula 2013	Korcula 2010	Split	Vis	CRC controls	LLS
Measurement Platform	LC-ESI-MS	LC-ESI-MS	LC-ESI-MS	LC-ESI-MS	UPLC	MALDI-TOF-MS
Type of glycans measured	IgG Fc	IgG Fc	IgG Fc	IgG Fc	IgG total	total plasma
Number of peaks	50	50	50	50	24	61
IgG1 | IgG2 | IgG4	20 | 20 | 10	20 | 20 | 10	20 | 20 | 10	20 | 20 | 10	-	-
Number of samples	669	504	980	395	535	2,056
Males | Females	271 | 398	156 | 348	386 | 594	152 | 243	288 | 247	936 | 1120
Age range	18-88	18-90	18-85	18-91	21-74	30-80
(mean ± standard deviation)	(53, 16)	(56, 14)	(50, 14)	(55, 15)	(52, 6)	(59, 7)

lows to precisely identify the mass of the molecules, which can be decomposed into the corresponding *composition*, i.e., the number and type of monosaccharides (Hexose, N-Acetylhexosamine, Fucose, etc.) necessary to produce the observed mass. The final spectra in this cohort included 61 distinct masses, or *compositions*.

2.1.2 Preprocessing

First, related individuals (first cousins or closer) and samples with missing values were excluded from all analyses. Glycan abundances were then normalized with the *Probabilistic Quotient* method [135] and subsequently log-transformed to improve normality [136]; finally, the normalized values were corrected for age and gender prior to statistical analysis. In the case of glycoproteomics measurements, where IgG glycans were separated by subclass, the normalization procedure was applied on each subclass separately.

The idea behind the Probabilistic Quotient normalization procedure, which was first introduced for metabolomics data, is to eliminate the effect of different *dilution factors* from the data. Let us consider a data matrix $X = x_{ij}$, where $i = 1, 2, ...n$ indicates the samples and $j = 1, 2, ...k$ the measured glycans. Therefore, $x_{i\cdot}$ represents sample i and $x_{\cdot j}$ indicates glycan j across all samples. Probabilistic Quotient normalization works as follows:

1. Define a *reference sample* as a vector r whose entries are the median values of each glycan measurement across all samples:

$$r = \{\text{med}(x_{\cdot 1}), \text{med}(x_{\cdot 2}), ...\text{med}(x_{\cdot k})\} = \{r_1, r_2, ...r_k\} \tag{2.1}$$

2. For a given sample $x_{i\cdot}$, compute a quotient vector by dividing each entry of sample $x_{i\cdot}$ by the corresponding entry of the reference sample r:

$$q_i = \{\frac{x_{i1}}{r_1}, \frac{x_{i2}}{r_2}, ...\frac{x_{ik}}{r_k}\} = \{q_{i1}, q_{i2}, ...q_{ik}\} \tag{2.2}$$

3. Compute the *dilution factor* Q_i of sample $x_{i\cdot}$ as the median of elements in q_i:

$$Q_i = \text{med}(q_i) \tag{2.3}$$

4. Divide each entry of sample $x_{i\cdot}$ by the dilution factor Q_i to obtain the normalized data sample $\hat{x}_{i\cdot}$:

$$\hat{x}_{i\cdot} = \{\frac{x_{i1}}{Q_i}, \frac{x_{i2}}{Q_i}, ...\frac{x_{ik}}{Q_i}\} = \{\hat{x}_{i1}, \hat{x}_{i2}, ...\hat{x}_{ik}\} \tag{2.4}$$

5. Repeat 2-4 for all samples.

2.1.3 Normalizations

In Chapter 5, where the goal was to test the effect of different normalization strategies on the investigation outcome, we considered several additional alternatives to Probabilistic Quotient for glycomics data normalization and compared their performance. The considered normalizations are described in this subsection.

Raw: These are the unprocessed spectra.

Median Centering: To each glycan value in the dataset, the value of that glycan median is subtracted. The underlying assumption is that the samples have a constant offset.

Total Area: The intensity of each glycan is normalized to the total area of the spectrum. This preserves the relative intensities of each peak within the sample, at the cost of losing one degree of freedom due to the constant sum constrain and giving rise to a so-called "compositional dataset" [137]. The underlying assumption here is that only relative intensities are biologically relevant. This type of normalization introduces spurious effects in the data covariance matrix, altering therefore the original structure. Let us consider three variables 1, 2, and 3 whose values x_1, x_2, and x_3 have been normalized with this technique. This means

$$\sum_{i=1}^{3} x_i = \kappa, \text{ with } \kappa \text{ constant.} \tag{2.5}$$

Therefore, if we compute the covariance of x_1 with the sum of all variables, we have

$$\text{cov}(x_1, x_1 + x_2 + x_3) = \text{cov}(x_1, \kappa) = 0. \tag{2.6}$$

But we also have that

$$cov(x_1, x_1 + x_2 + x_3) = \text{cov}(x_1, x_1) + \text{cov}(x_1, x_2) + \text{cov}(x_1, x_3). \tag{2.7}$$

Therefore

$$\text{cov}(x_1, x_2) + \text{cov}(x_1, x_3) = -\text{var}(x_1) \leq 0. \tag{2.8}$$

This means that, *just because of the constraint introduced by the normalization*, each row of the variance-covariance matrix will have at least one negative value.

Quantile: This method forces the distributions of the glycans (columns) to be the same with respect to the quantiles. It requires replacing each point of a glycan with the mean of the corresponding quantile [138].

Rank: Values are replaced with the corresponding ranks [139].

Log-transformation: Biological data have been observed to often follow a log-normal distribution [136]. Since our correlation estimator assumes normally distributed data, for each considered normalization (except the median centering) we included in the comparison both the non-transformed, as well as the log-transformed data.

Subclass-specific normalization: LC-ESI-MS IgG glycans are measured at the glycopeptides level, which means that the information about the IgG isoform is preserved (see Subsection 2.1.1). For this platform, each normalization method was applied both on the 50 glycan measurements together, as well as separately per each IgG subclass.

2.1.4 Glycosylation synthesis pathways

As we mentioned in Chapter 1, the *in vivo* investigation of protein-specific glycosylation pathways is experimentally infeasible. Nevertheless, details on enzyme specificities are available from *in vitro* experiments, and this information was used as **prior knowledge** to build the glycosylation pathways used in this thesis.

IgG Secreted IgG has been observed to mainly carry 26 different glycoforms, which are synthesized by the stepwise modifications of one monosaccharide at a time. The glycosylation reactions supported by *in vitro* experimental evidence are shown in Figure 2.1 (colored arrows). This model will be extended in Chapter 3 (Figure 2.1, black arrows) and this will then be used in all subsequent chapters. Depending on the chosen measurement platform, not all IgG glycoforms might be identified, and hence the pathway needs to be adjusted to match the measured data. Figure 2.2 shows the IgG glycosylation pathways for LC-ESI-MS and UPLC data, where only the synthesis reactions among glycoforms measured in the corresponding platform are included. Note that in glycoproteomics data we have different IgG subclasses: in absence of experimental evidence that suggested otherwise, in this thesis the pathways for all subclasses were assumed to be the same. For IgG4 only the fucosylated glycan structures were measured and hence only the corresponding part of the pathway was considered.

Total Plasma N-Glycome For the total plasma N-glycome, the construction of one unified glycosylation pathway is a substantially more complex task, as these data include glycans from different proteins and glycosylation sites. We first created a *theoretical pathway* describing the synthesis of all known human protein glycan structures, including all single monosaccharide modifications against which there was no experimental evidence (Figure 2.3). Since in our TPNG data only compositions were measured, we reduced the theoretical pathway to only include structures whose composition appeared in the dataset (Figure 2.4). We then merged all the nodes representing structures with the same com-

position. The resulting *compositional pathway* included the same composition covered by the considered dataset (see Chapter 4).

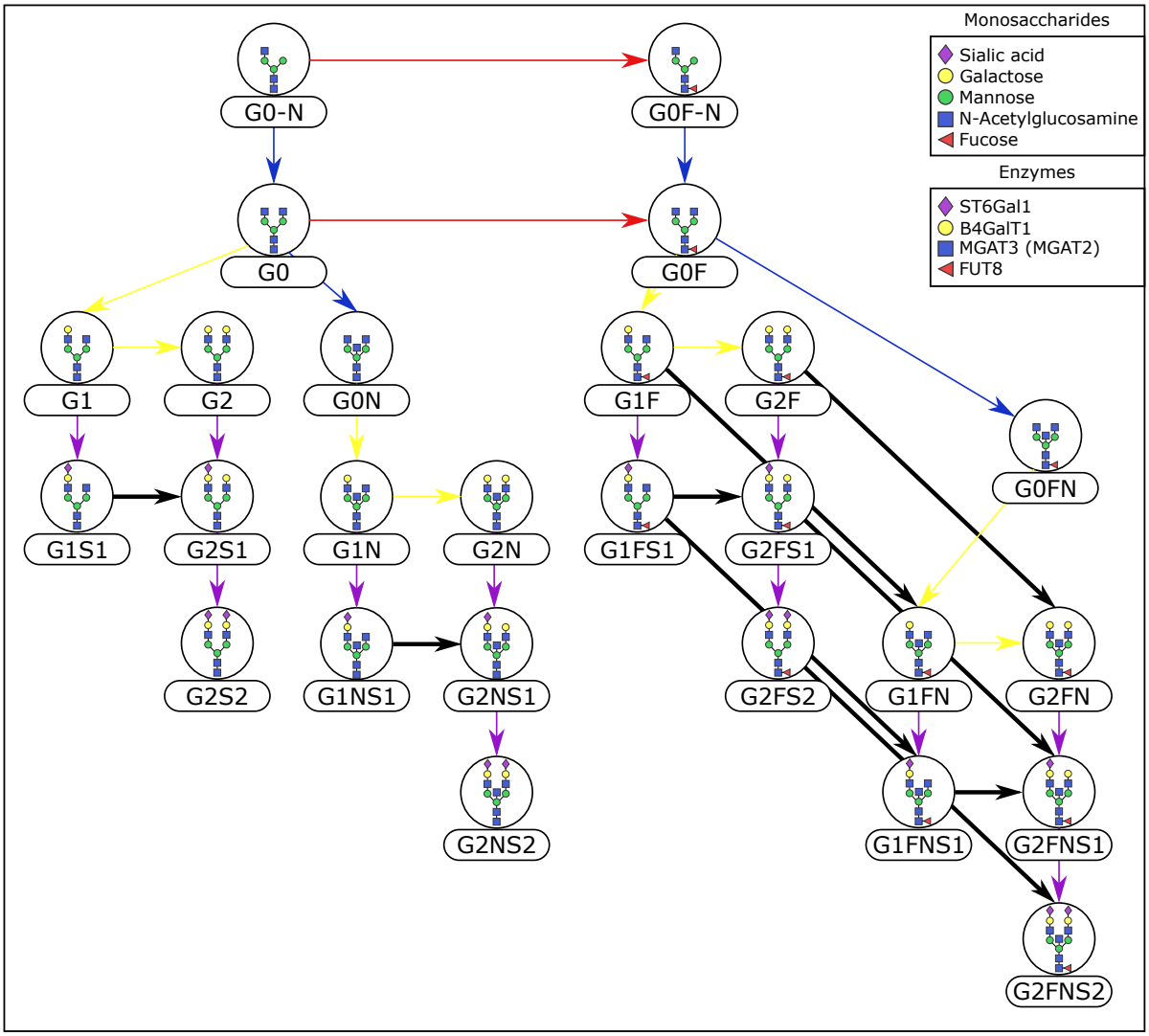


Figure 2.1: IgG glycan structures and their glycosylation pathway. IgG glycans are biantennary complex type structures. The glycan nomenclature describes how many galactoses (G0/G1/G2) are present, whether there is a core fucose (F) or a bisecting N-Acetylglucosamine (N), and whether the structure includes one or two sialic acids (S1/S2). Nodes represent glycan structures and arrows represent single enzymatic reactions in the synthesis process. Colored arrows represent enzymatic reactions reported in literature, while black arrows indicate new enzymatic steps inferred from the data and validated experimentally in this thesis (see Chapter 3). MGAT3 is responsible for the addition of the bisecting GlcNAc, while MGAT2 adds the GlcNAc to the second antenna of the glycan.

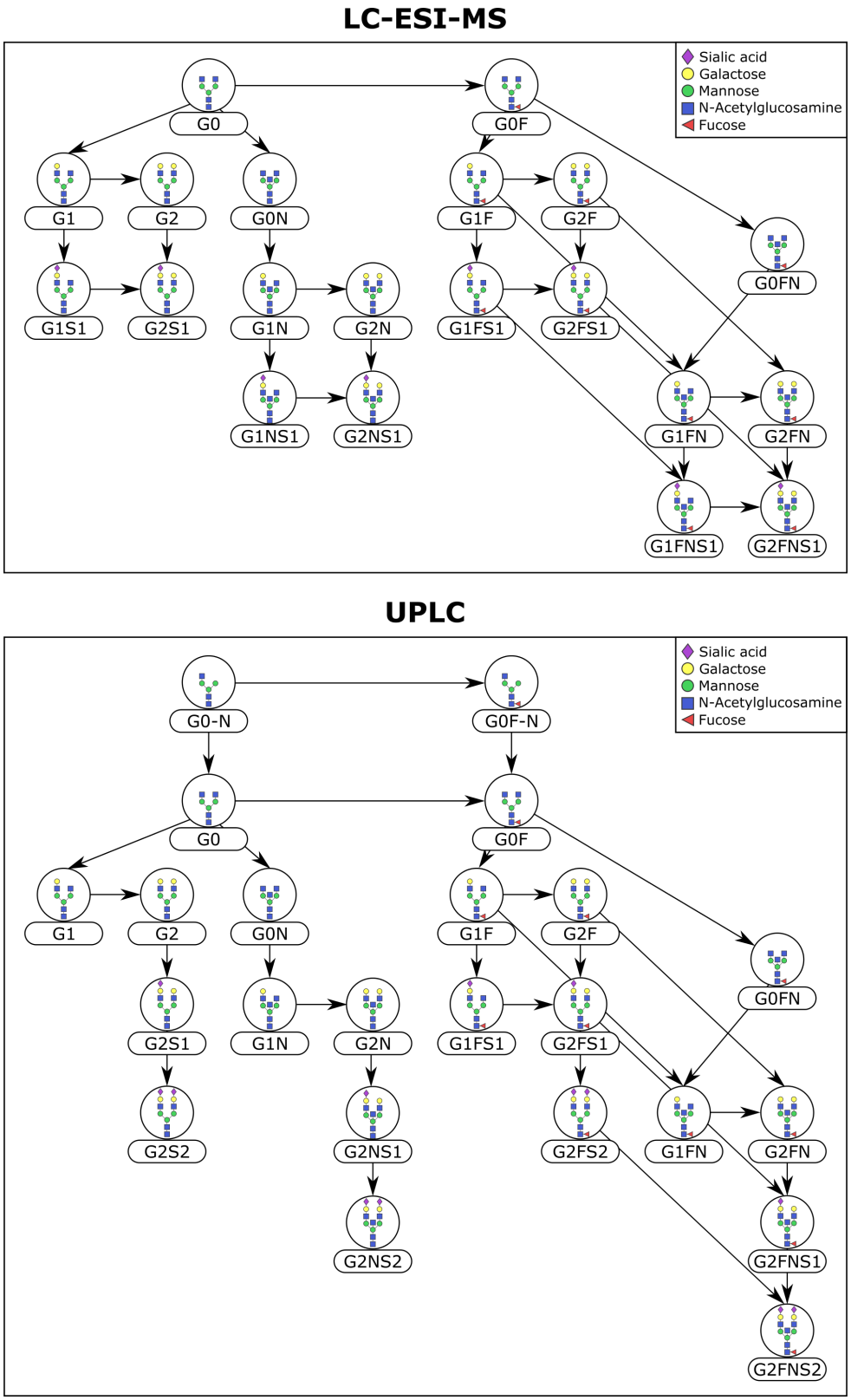


Figure 2.2: IgG glycosylation pathway adapted to the available data.

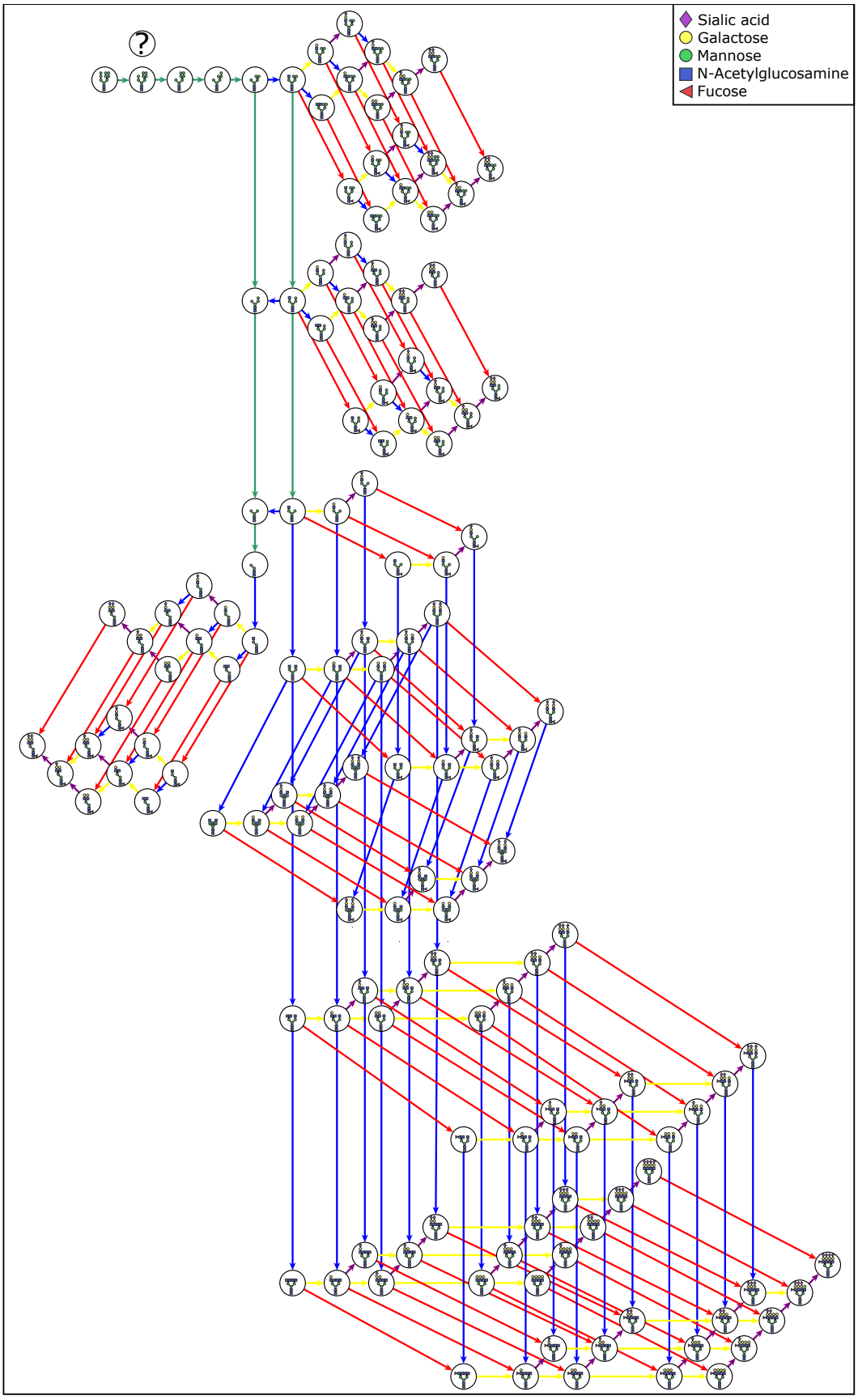


Figure 2.3: TPNG theoretical pathway.

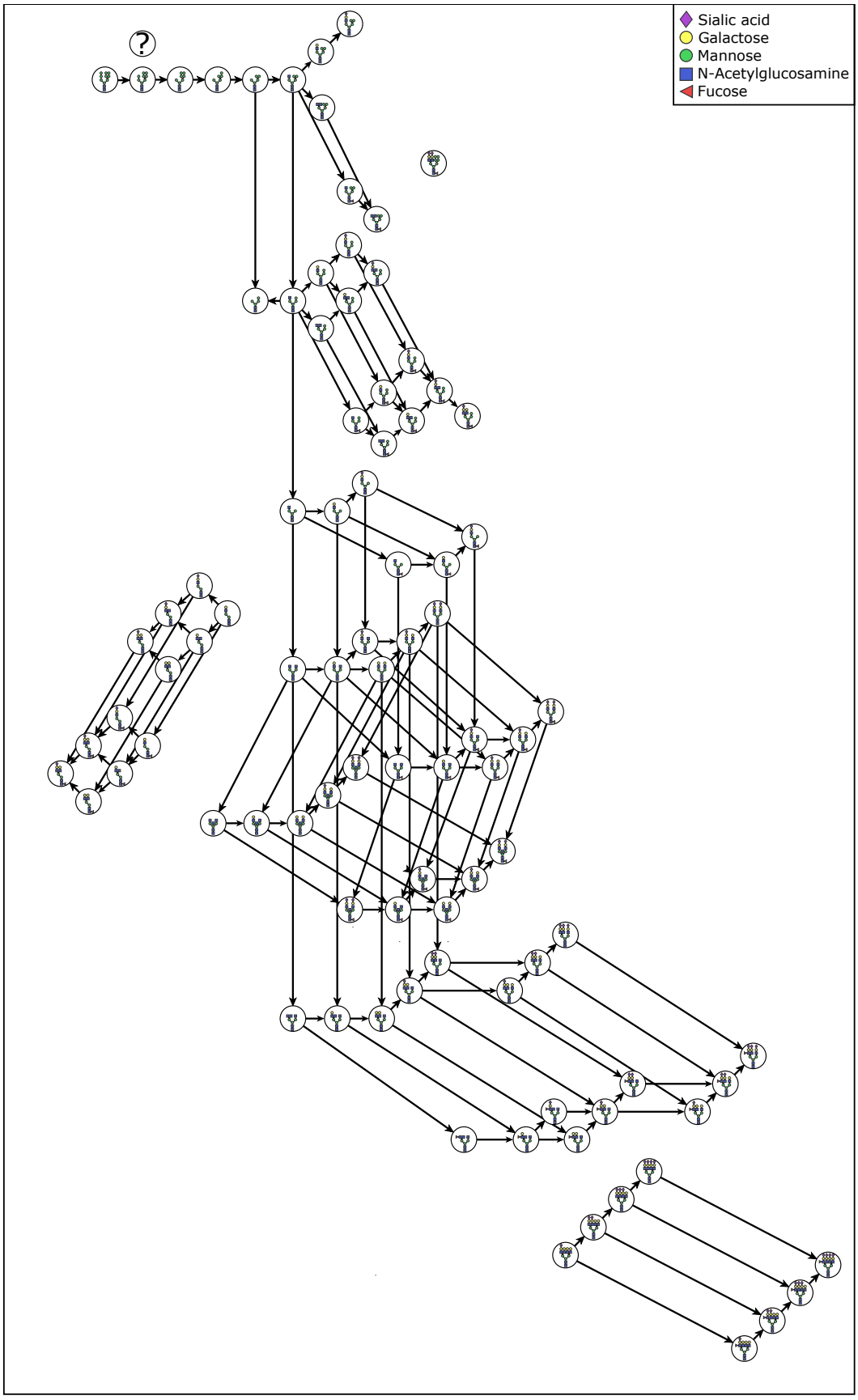


Figure 2.4: TPNG theoretical pathway adapted to the available data.

2.2 Other omics data

2.2.1 Genomics data

The German study population KORA ("Kooperative Gesundheitsforschung in der Region Augsburg" [140]) included 3,788 DNA samples with 18,185,628 SNPs (after QC) and 1,887 glycomics samples. 1,823 samples included both gene information and glycan concentrations. Samples with mismatched phenotypic and genetic genders were excluded, leaving 1,641 samples and 18,185,628 SNPs to be analyzed (Table 2.2).

Table 2.2: Glycomics data in the KORA F4 cohort.

	KORA F4
Measurement Platform	LC-ESI-MS
Type of glycans measured	IgG Fc
Number of peaks	50
IgG1 | IgG2 | IgG4	20 | 20 | 10
Number of samples	1,641
Males | Females	793 | 848
Age range	32-81
(mean ± standard deviation)	(61, 9)

2.2.2 Metabolomics data

Metabolomics samples were taken from an antipsychotics study conducted in Qatar (Al-Amin et al., manuscript in preparation). Urine samples were analyzed using ultra-high-performance liquid-phase chromatography and gas-chromatography separation, coupled with tandem mass spectrometry by Metabolon, Inc. Data were runday-median scaled, normalized using probabilistic quotient normalization [135] and log-transformed. From the original data matrix, we first excluded metabolites with more than 20% missing values, and then samples with more than 10% missing values. Samples with missing covariates were subsequently excluded from the analysis. The filtered data matrix contained 97 samples and 1,021 metabolites (527 known structures and 494 unknown), see Table 2.3. Remaining missing values were imputed with a KNN-based method with variable preselection [141]. Data were corrected for age, gender and BMI prior to analysis. All participants have given written informed consent and the local ethics committees have approved the studies.

Table 2.3: Metabolomics data in the Qatar cohort.

	urine
Platform	Metabolon
Number of peaks	1,021
Known | Unknown	527 | 494
Number of samples	95
Males | Females	35 | 60
Age range	21-60
(mean ± standard deviation)	(36, 8)
BMI range	17-46
(mean ± standard deviation)	(28, 5)

2.2.3 Transcriptomics data

RNA-seq data were downloaded from The Cancer Genome Atlas (TCGA) [142] and initially included 2,726 samples and 16,115 genes from 11 cancer types: acute myeloid leukemia, bladder urothelial carcinoma, colon adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, kidney clear cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serous cystadenocarcinoma, rectum adenocarcinoma, and uterine corpus endometrioid carcinoma. For each cancer type, genes with more than 20% of missing values were excluded. Missing values in the remaining genes were imputed using a KNN-based method with variable preselection [141]. Values were corrected for age and gender. The final dataset was obtained by considering only genes present in all cancer types after preprocessing (12,005). Cancer type was further corrected for prior to statistical analysis.

Table 2.4: Transcriptomics data in the PANCAN cohort.

	PANCAN
Platform	RNA-seq
Number of transcripts	12,005
Number of samples	2,726
Males | Females	1294 | 1432
Age range	18-90
(mean ± standard deviation)	(63, 12)
Cancer types	11

2.3 Network inference and analysis

Since biological data have been shown to often follow a log-normal distribution [136, 143], i.e., a continuous probability distribution whose logarithm is normally distributed, in the description below we will only consider the case of normally distributed variables.

2.3.1 Correlation measures

Throughout this thesis, we considered three measure of linear dependence to describe the relationship between molecules in our data:

1. Pearson correlation
2. Analytical partial correlation, referred to as "*parcor*"
3. Regularized partial correlation, referred to as "*GeneNet*"

This section briefly introduces the mathematical formulation of these measures and elaborates on their advantages and limitations.

Given two normally distributed random variables X_i and X_j, their covariance is:

$$\mathrm{cov}(i,j) = c_{ij} = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])], \tag{2.9}$$

where $\mathrm{E}[\cdot]$ is the expectation value. The matrix $\Sigma = (c_{ij})$ is referred to as the *variance-covariance matrix*. The pairwise **Pearson correlation** coefficient r_{ij} between the two variables X_i and X_j is given by their covariance c_{ij} divided by the product of their standard deviation σ_i and σ_j [120]:

$$r_{i,j} = \frac{c_{ij}}{\sigma_i \sigma_j}, \tag{2.10}$$

with $r \in [-1, 1]$. The resulting coefficient quantifies the linear dependence between X_i and X_j, but it does not account for the presence of covariates or confounders that may mediate the interaction. When dealing with highly correlated variables, this might lead to an inaccurate interpretation of the underlying dependency structure. To overcome this problem, **partial correlation** can be used to compute the linear association between two variables *conditioned against* one or more other variables in the dataset. The idea is to remove from the computed correlation between X_i and X_j the effect of the correlation of the two variables with other variables in the dataset. This can be obtained from the inverse of the covariance matrix $\Sigma^{-1} = (w_{ij})$ as [121]:

$$z_{ij} = \frac{-w_{ij}}{\sqrt{w_{ii}w_{jj}}}, \tag{2.11}$$

where z_{ij} is the pairwise partial correlation coefficient of variables X_i and X_j. This analytical derivation of partial correlation, from now on referred to as **parcor**, allows for an efficient estimation of the coefficients. However, since it involves a matrix inversion operation, it is unstable for small sample sizes and cannot be computed directly if the number of samples is less than the number of variables ($n < p$).

This issue can be overcome by considering a *regularized* formulation of partial correlation. For example, Schäfer and Strimmer proposed a regularized covariance estimator, called **GeneNet** [144], based on the Ledoit-Wolf lemma [145]. Briefly, to correct the covariance matrix prior to inversion, a shrinkage parameter λ^* is defined as

$$\lambda^* = \frac{\sum\limits_{i \neq j} \mathrm{var}(r_{ij})}{\sum\limits_{i \neq j} r_{ij}^2}, \tag{2.12}$$

where r_{ij} is the empirical correlation between random variables X_i and X_j. From this optimized shrinkage value, the empirical correlation and covariance coefficients r_{ij} and c_{ij} are redefined, respectively, as

$$r_{ij}^* = \begin{cases} 1 & \text{if } i = j \\ r_{ij}\min(1, \max(0, 1 - \lambda^*)) & \text{if } i \neq j \end{cases} \tag{2.13}$$

and

$$c_{ij}^* = \begin{cases} c_{ii}^2 & \text{if } i = j \\ r_{ij}^* \sqrt{c_{ii}c_{jj}} & \text{if } i \neq j. \end{cases} \tag{2.14}$$

This shrinkage-based estimation of the covariance matrix $\Sigma^* = (c_{ij}^*)$ is always positive-definite and well-conditioned, therefore allowing for a more stable estimation of partial correlation coefficients when applying formula 2.11.

To assess the statistical significance, a mixture model is fitted to the partial correlation matrix to estimate the corresponding p-values [144], which results in a more robust p-value estimation.

2.3.2 Multiple testing correction

Correlation matrices were corrected for multiple testing by controlling the False Discovery Rate (FDR) at a significance level α (usually 0.01 or 0.05) using the Benjamini–Hochberg method [122]. This approach allows for the control of the expected *proportion* of false positives.

For a set of null hypotheses $H_1, H_2, ...H_n$ and their associated p-values $P_1, P_2, ...P_n$, the approach works as follows:

1. Rank the p-values $P_1, P_2, ...P_n$ associated to the correlation coefficients in ascending order
2. For a given significance level α, find the largest k such that $P_{(k)} \leq \frac{k}{n}\alpha$
3. Reject the null-hypothesis for (i.e., declare *statistically significant*) all $H_{(i)}$ with associated p-values $P_{(i)} \leq P_{(k)}$ and $i = 1, ...k$.

Another common multiple testing correction approach is the Bonferroni correction [146], which differs from the FDR approach in that it controls for the Family-wise Error Rate (FWER), i.e., for the probability of having *at least one* false positive. In this case, the null-hypothesis is rejected for all p-values that satisfy $P_k \leq \alpha/n$.

In Chapter 6, we will use a biological measure to optimize the correlation cutoff for network inference, and compare the GGMs obtained with our approach to those obtained using the aforementioned statistical cutoffs.

2.3.3 Network representation

In general, a network is a graph defined by a set of nodes and edges. Pairwise correlations are well suited for visualization as undirected weighted graphs, where the nodes represent the variables in the dataset and the edges the strength of their correlation.

For the purpose of this thesis, we define a **correlation network** as the network representation of a data-driven correlation matrix of n variables, where a selection criterion has been applied to establish which of the computed $n(n-1)/2$ correlation coefficients is to be considered *true*. For example, if we choose the *statistical significance* as a selection criterion, the network representation will depict all correlation coefficients whose p-values passed the multiple testing correction.

The **adjacency matrix** $A = (a_{ij})$ associated to any given network describes which nodes are *adjacent* in the network, i.e., share a connection. Since correlation matrices are symmetric, their corresponding network adjacency matrix will be as well:

$$a_{ij} = \begin{cases} 1 & \text{if node } i \text{ and node } j \text{ are connected by an edge} \\ 0 & \text{otherwise.} \end{cases} \tag{2.15}$$

A **graphical model** is a probabilistic model in which the conditional dependence structure between a set of random variables is visualized as a graph. In particular, a *Gaussian Graphical Model* (GGM) is an undirected graph where edges represent the conditional dependence structure among normally distributed variables: the presence of a connecting edge between two variable nodes indicates that such variables are still correlated even once the confounding effect of all other variables has been corrected for. Therefore, GGMs correspond to partial correlation networks for multivariate normally distributed data and have been widely exploited in systems biology to infer molecular interaction networks, from gene regulatory networks [125, 147, 148] to metabolic pathways [32, 116]. In this thesis, we apply for the first time this methodology to glycomics data.

2.3.4 Biological references

Glycomics data. The biological reference reflects the current understanding of the IgG glycosylation pathway, as established in Chapter 3. Glycans can be modified by the addition of one monosaccharide at a time, but only selected reactions are enzymatically feasible, as shown in Figure 2.2.

Metabolomics data. There is no established complete biochemical pathway to consider as biological reference for metabolomics data. Known metabolic reactions were imported from the RECON2 database [149] and included in one of the adjacencies. As a more coarse type of biological references, we used sub- and super-pathway annotations provided with the metabolites measurements by Metabolon, Inc., from which adjacency matrices were created by connecting all metabolites within the same sub- or super-pathway, respectively.

Transcriptomics data. Pathway annotations were imported from the Reactome database [150, 151]. We restricted the analysis to pathways containing at least 50 genes and with at least 30% of the genes in the pathway measured in the TCGA data. These constraints led to a total of 469 Reactome pathways being selected. For each pathway, protein-protein interactions were downloaded from the STRING database [152, 153] and used as biological reference for the optimization.

2.3.5 Pathway analysis

In order to better characterize what edges in a data-driven network represent biologically, we relate the network to the available prior knowledge on synthesis pathways.

To quantitatively estimate the agreement between a data-driven correlation networks and a biochemical pathway or, more generally, any prior-knowledge-based *biological reference*, we employed Fisher's exact test [154]. This statistical test evaluates whether two categorical variables are statistically independent, with low p-values indicating lack of independence.

For the purpose of the analyses presented in this thesis, we treated the Fisher's p-value as a quantitative measure of the overlap between the available biological reference and the calculated correlation networks. In this case, lower p-values indicate a better overlap between the biological reference and the data-driven network.

In order to compute the Fisher's test p-value, all computed correlation coefficients are first classified in a contingency table (Table 2.5), according to their statistical significance and whether the corresponding variable pair is directly connected in the biological reference.

Table 2.5: Contingency table for pathway analysis

	in reference	not in reference
significant correlation	a	b
non significant correlation	c	d

In other words, we classify which correlation coefficients are *true positives* (present in both the data-driven network and the reference), *false positives* (in the network but not in the reference), *false negatives* (in the reference but not in the network), and *true negatives* (neither in the network nor the reference). The p-value of the Fisher's exact test is calculated according to the hypergeometric distribution as:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}, \tag{2.16}$$

where $n = a + b + c + d$.

2.3.6 Modularity

The network modularity algorithm was adapted from the widely used community detection clustering method of Newman [155], which optimizes a modularity Q to determine clusters. In this thesis, we used the Q measure to assess the modularity of predefined clusters,

given by the three IgG subclasses. To this end, subclass-based network modularity was calculated as the relative out-degree from each subclass to all other subclasses for all significantly positive edges. Let V_1, V_2, V_3 indicate the sets of glycans belonging to each of the three measured IgG subclasses. The subclass-based modularity is mathematically described as

$$Q := \sum_{i=1}^{3} \left[\frac{A(V_i, V_i)}{A(V, V} - \left(\frac{A(V_i, V)}{A(V, V)} \right)^2 \right], \tag{2.17}$$

where $A(V_i, V_j)$ represents the total number of edges between glycan sets V_i and V_j, and V represents all glycans in the network [156].

To assess the significance of the observed modularity, we performed graph randomization via edge rewiring [157,158]. In this process, two edges in the original data-driven network are randomly selected and the end nodes of each edge are swapped. The operation was repeated 10 times the number of edges to reach sufficient randomization. The entire randomization was repeated 105 times to obtain a sufficient number of null model networks.

2.3.7 Resampling techniques

To improve the generalizability of the results presented in this thesis, we resampled the original datasets multiple times with replacement (*bootstrapping*) and repeated the whole analysis pipeline. Thus, for each bootstrapped dataset a new set of results was obtained. We then computed the median and 95% confidence intervals from the overall distribution of the bootstrapping results and used them to provide an estimate of the robustness of our findings.

The resampling idea can be applied also when trying to simulate the effects of smaller sample sizes on the outcomes of the analysis (see for example Section 6.1 in Chapter 6). In this case, the dataset is resampled multiple times without replacement (*subsampling*) to produce new datasets of the chosen sample size. Again, the analysis pipeline is repeated on the subsampled data and, for each considered sample size, the median and 95% confidence intervals are used to provide an estimate of the robustness of the results to variations in sample size.

2.4 Genome-wide association study

Genotyping was performed using the Affymetrix GeneChip array 6.0 with prephasing by SHAPEIT v2 and imputation by IMPUTE v2.3.0, using 1000 Genomes (phase 1 integrated haplotypes CEU) as a reference panel. We limited our analysis to non-monomorphic SNPs that had a minor allele frequency >1%, a high genotyping quality (call rate >97%), and did not significantly deviate from the Hardy–Weinberg equilibrium (HWE; $pHWE \geq 5 \times 10^6$).

The glycan measurements were preprocessed using a similar pipeline as that for the Croatian data in the pathway analysis (see Subsection 2.1.2). Samples from each IgG subclass were log-transformed and batch-corrected using the ComBat algorithm of the R package "sva" (R package version 3.14.0). The data were exponentiated to retrieve the original scale and then normalized using the probabilistic quotient algorithm [135]. Glycan ratios were calculated as the product–substrate ratios of all possible reactions in the IgG glycosylation pathway, as shown in Figure 3.5A, and then log-transformed and regressed against age and sex. A rank-based inverse normal transformation was applied to the residuals.

For the purposes of this study, we only focused on SNPs located in the regions of the known glycosylation enzymes—ST6GAL1 (chr.3), B4GALT1 (chr.9), FUT8 (chr.14), and MGAT3 (chr.22)—and with a linkage disequilibrium (LD) and $R2 \geq 0.8$ (see Supplementary Table 1 in the original publication). Genomic positions were retrieved from the UCSC Genomic Browser (GRCh37/hg19) [159], while LD information was obtained using the software SNIPA [160].

GWAS was performed with snptest software v2.5.1 [161] using an additive genetic model. We used an established GWAS significance threshold [162] corrected for the number of considered ratios, i.e., $5 \cdot 10^{-8}/95 = 5.26 \cdot 10^{-10}$. For suggestive hits, we used a relaxed threshold of 10^{-7}, as also suggested in Panagiotou and Ioannidis (2012) [162].

P-gains were introduced to describe the increase in strength of the association of a ratio compared to the corresponding single glycans. We assume that a significant p-value combined with a high p-gain indicates that the two glycans are functionally linked in a biochemical reaction involving the gene of the associating SNP. P-gains were defined as the ratio between the minimum of the association p-values of single glycans and the association p-value of the corresponding ratio [163, 164]. The gain in the p-value of the ratio was considered to be significant if it was greater than or equal to 10, as this value indicates gains of one order of magnitude. This threshold was taken as suggestive also in previous studies, e.g. in Suhre et al., 2011 [165] and Shin et al., 2014 [166]. See Supplementary Data 3 in the original publication for a full list of results.

Chapter 3

Network inference from glycoproteomics data reveals new reactions in the IgG glycosylation pathway

In this chapter, we investigate the IgG glycosylation pathway using plasma IgG glycomics LC–ESI-MS measurements from four independent cohorts measured by our collaboration partners at Genos (see Subsection 2.1.1). We first generate a partial correlation network, or Gaussian graphical model (GGM), where the nodes represent individual glycans and the edges represent their pairwise correlations, corrected for the confounding effects of all other glycans and clinical covariates.

Previous studies using serum metabolomics data have shown that highly correlated pairs in GGMs represent enzymatic reactions [32,166]. This is the first study to apply GGMs to large-scale IgG glycomics data from four independent populations. We find that significant partial correlations predominantly occur between glycan structures that are one enzymatic step apart in the known IgG glycosylation pathway shown in Figure 2.1, demonstrating that network statistics on quantitative glycoprotein measurements allow us to detect true enzymatic reaction steps in the glycosylation pathway.

Based on this result, we expect edges in the GGM that did not appear in the known pathway to represent true but hitherto unknown enzymatic steps, i.e., unknown substrate specificities of the enzymes in the pathway. To investigate this hypothesis, we develop

a rule-based inference approach to test alternative pathway models. This shows that additional reactions are supported by the data for all four cohorts. More in detail, we predict that bisection of fucosylated, galactosylated glycans, as well as galactosylation of monosialylated glycans occur during IgG glycan synthesis.

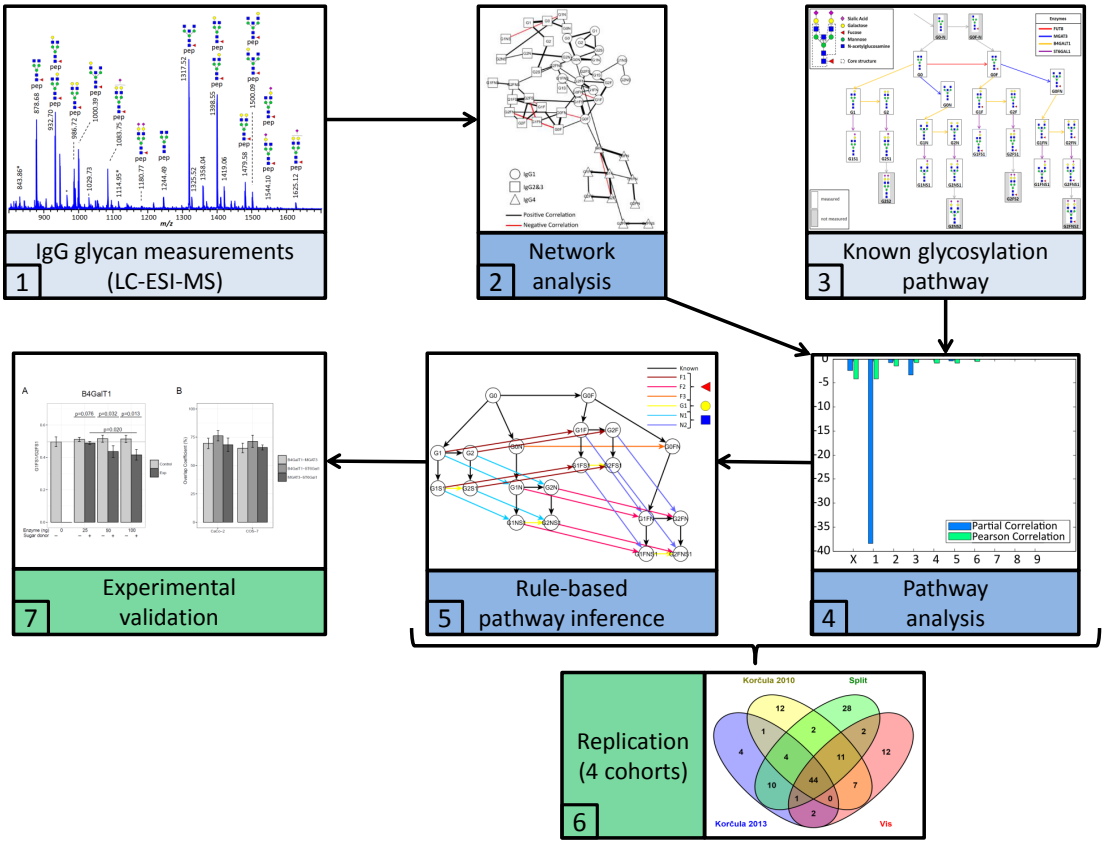


Figure 3.1: Analytical procedure. Starting from the immunoglobulin G (IgG) glycan abundances measured using liquid chromatography coupled with electrospray mass (LC-ESI-MS) (1), we calculated a correlation-based network (2) and mapped it to the known IgG glycosylation pathway (3). We found that most edges in the network corresponded to single enzymatic steps in the pathway (4). Based on this finding, we inferred unknown enzymatic reactions that were putatively involved in the synthesis of IgG glycans using a rule-based approach (5). We then replicated the findings using four cohorts (6) and performed different in vitro validation experiments to confirm the predicted reactions (7).

As direct experimental validation is considered infeasible for the reasons outlined above, we validate our findings with two different approaches (Figure 3.1). First, we use a genome-wide association study (GWAS) in a fifth cohort. It has previously been shown that the substrate–product ratios of metabolites are associated with their enzymes in GWAS [163, 165]. Therefore, we consider the ratios of substrate–product pairs of the predicted reactions as quantitative traits, with which we can confirm several of our predicted reactions across

the IgG subclasses. Second, we perform three sets of in vitro experiments to confirm the predicted enzymes substrate specificities, as well as their colocalization inside the Golgi apparatus. Our results show that at least one of the inferred reactions occurs *in vitro*, that one rejected rejection does not occur, and that the glycosyltransferases involved in the predicted reactions are colocalized in the Golgi stacks of two different cell lines.

All results reported in this chapter are part of the following publication:

- **Benedetti, E.**, Pučić-Baković, M., Keser, T., Wahl, A., Hassinen, A., Yang, J-Y., Liu, L., Trbojević-Akmačić, I., Razdorov, G., Štambuk, J., Klarić, L., Ugrina, I., Selman, M.H.J., Wuhrer, M., Rudan, I., Polasek, O., Hayward, C., Grallert, H., Strauch, K., Peters, A., Meitinger, T., Gieger, C., Vilaj, M., Boons, G-J., Moremen, K.W., Ovchinnikova, T., Bovin, N., Kellokumpu, S., Theis*, F.J., Lauc*, G., Krumsiek*, J., Network inference from glycoproteomics data reveals new reactions in the IgG glycosylation pathway, *Nat. Commun.*, 8(1):1483, 2017.

I contributed all data analysis on the Croatian glycomics cohorts, as well as all results interpretation and discussion.

3.1 IgG glycomics correlation networks

We analyzed four glycomics datasets, where IgG Fc glycans were quantified by LC-ESI-MS (Table 2.1). Data preprocessing and normalization was performed as described in Subsection 2.1.2. In the following, the Korčula 2013 cohort was selected for use in the discovery analysis. The results for all other cohorts are discussed in the replication section below.

We used both regular Pearson correlation and partial correlation analysis to make comparisons. The partial correlation analysis tested the conditional dependency between two variables when accounting for the confounding effects of all other glycans, as well as age and gender. In total, 905 Pearson correlation coefficients were significantly different from zero following multiple testing correction (FDR 0.01 [122]). This significance level corresponded to an absolute correlation cutoff of 0.105, with coefficients approximately symmetrically distributed around zero (Figure 3.2A). Partial correlation coefficients are, by nature, much lower in absolute value than Pearson coefficients, and so only 66 of the total 1,275 coefficients were found to be significant, the majority of which were positive (Figure 3.2B).

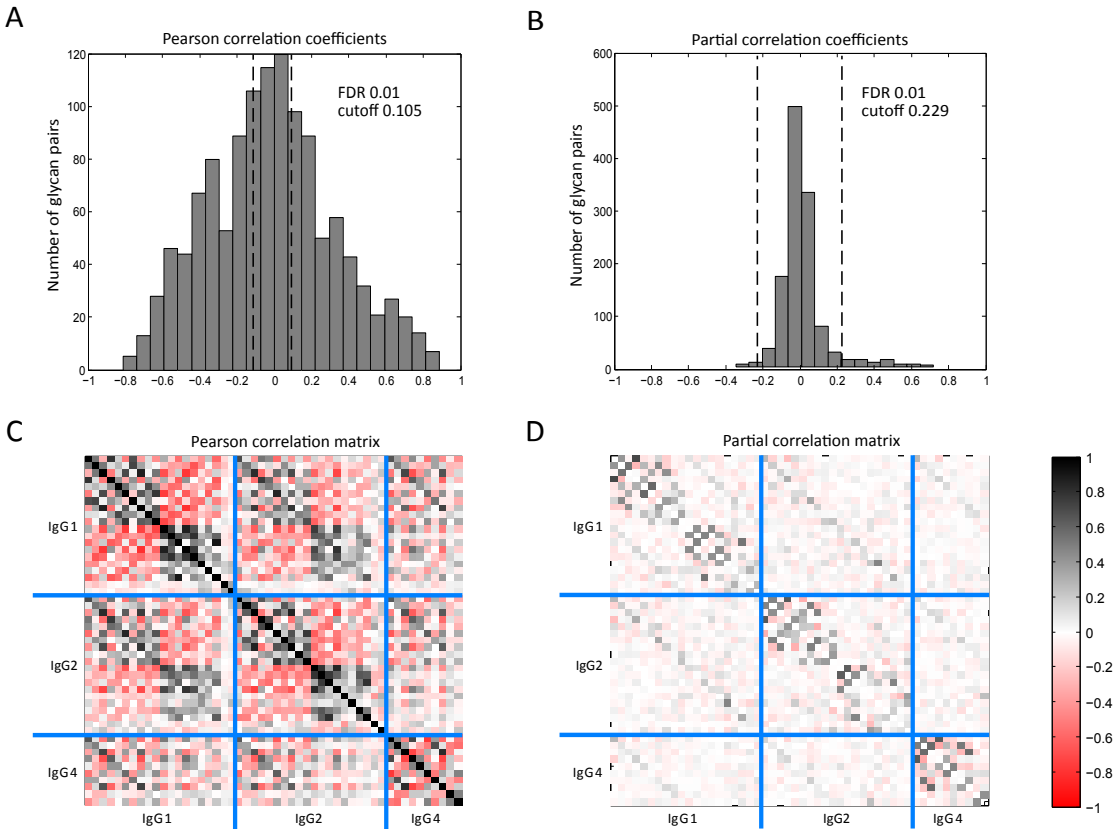


Figure 3.2: Pearson and partial correlations of IgG glycans. **A**, **B** Histograms of all pairwise Pearson and partial correlation coefficients, respectively. Black dashed lines indicate the significance cutoff (FDR = 0.01). The Pearson correlation matrix contains a large number of significant coefficients, which are evenly distributed around zero. Partial correlations are generally lower, and a much smaller proportion was statistically significant. Moreover, most significant partial correlations were positive. **C**, **D** Pearson and partial correlation matrices, respectively. Black and red indicate positive and negative coefficients, respectively. Blue lines separate the different IgG subclasses. The ordering of the glycans is the same for all subclasses. The stronger signal around the inter-subclass diagonals represents connections between glycans with the same structure in different IgG subclasses.

Upon inspection of the correlation matrices, we observed a remarkably similar structure between the different IgG subclasses (Figure 3.2C and D)—that is, glycoforms that were strongly correlated in one subclass also tended to be strongly correlated in the other subclasses. Moreover, there were only a few significant correlation coefficients for cross-subclass glycan pairs (off-diagonal blocks of the matrix). This suggests that the regulation of IgG is highly conserved across subclasses. Interestingly, seven of the nine cross-subclass pairs involved glycans with the same structure. For a full list of the partial correlations, see Supplementary Data 1 in the original publication.

The Pearson and partial correlation matrices were represented as networks (i.e., weighted

graphs), with the nodes representing glycans and the edges indicating coefficients that are statistically significant (Figure 3.3A and B).

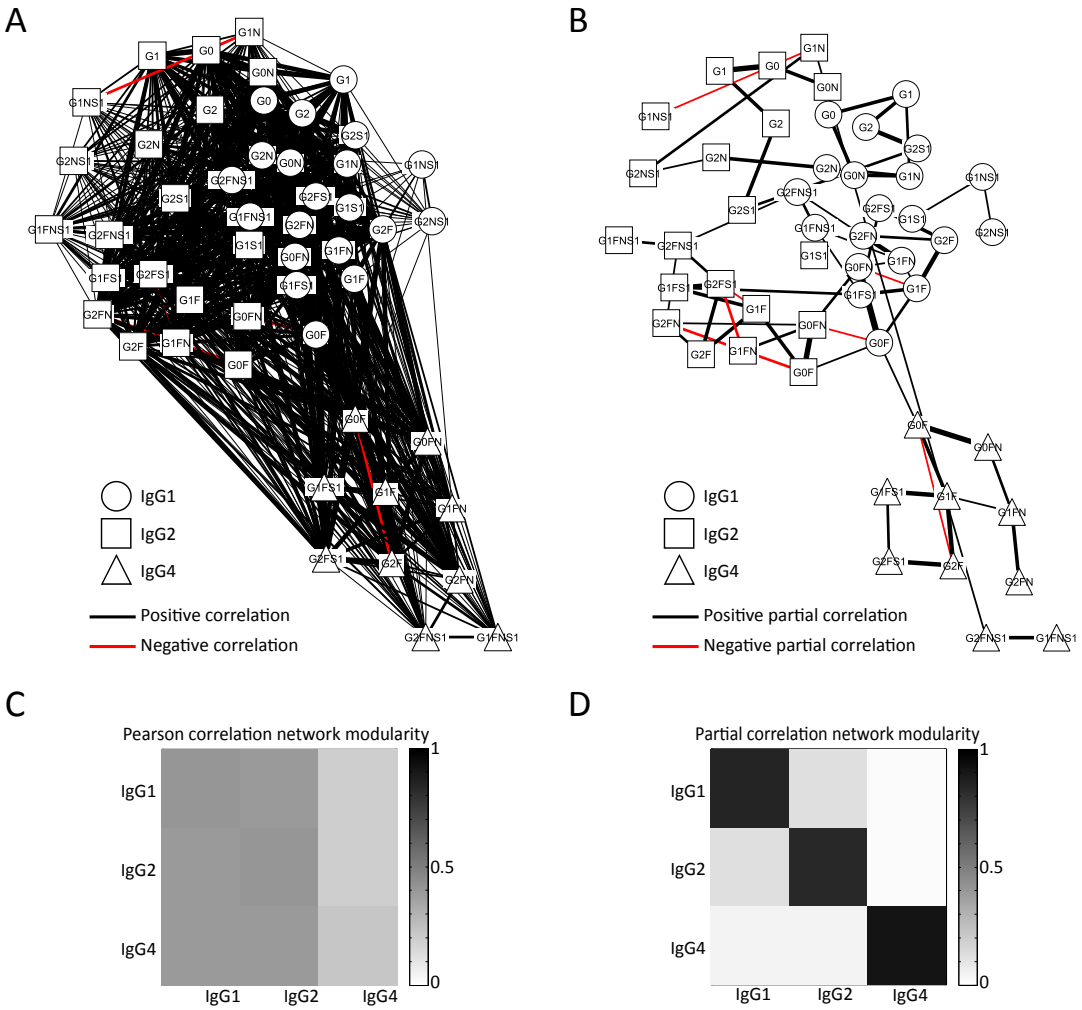


Figure 3.3: Network representation and modularity. **A**, **B** Pearson and partial correlation matrices, respectively, visualized as networks, where the nodes represent different glycoforms, and the edges indicate significant positive (black) and negative (red) correlations. Different node shapes correspond to different immunoglobulin G (IgG) subclasses, while the thickness of each edge corresponds to the magnitude of the respective correlation. **C**, **D** Pearson and partial correlation modularity, respectively, between IgG subclasses, measured as the relative out-degree from each subclass (row) to each other subclass (column). A Pearson correlation modularity analysis showed that all subclasses were highly interconnected. By contrast, the GGM showed high subclass modularity, indicating that associations between glycans mostly occurred within each IgG subclass. Furthermore, while the first two IgG subclasses were slightly interconnected, the IgG4 subclass was mostly isolated in the network.

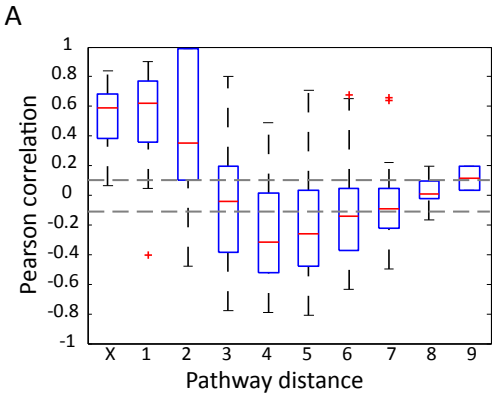
Most of the network edges connected glycan pairs that differed by only a single monosaccharide residue. This directly reflects the underlying glycan synthesis pathway, whereby

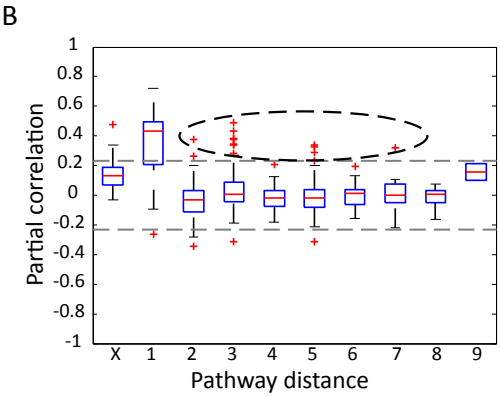
monosaccharides are added one at a time and in a given order to create the final glycoform (Figure 2.2). Furthermore, unlike the GGM, which showed a strong modularity with respect to the IgG subclasses, the Pearson correlation network did not show a clear separation between the subclasses (Figure 3.3C and D). This indicates that significant partial correlations were mostly found between glycans belonging to the same IgG subclass, with few significant correlations between glycans with different IgG isoforms. To investigate this observation quantitatively, we calculated a subclass-based network modularity for all significantly positive edges based on the method by Newman et al. (2004), and as previously adapted by Krumsiek et al. (2011). We used degree-preserving random edge rewiring as a null model to assess the statistical significance (see Subsection 2.3.6). The computed modularity for the original network was $Q = 0.495$ with an empirical p-value of $< 10^{-5}$, proving a high level of subclass-specific modularity.

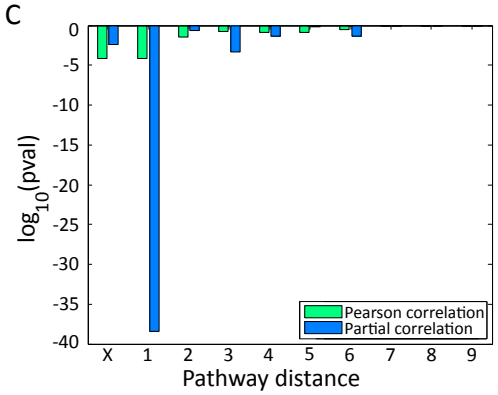
3.2 Overlap of GGM with known IgG glycosylation pathway

We systematically investigated the relationship between the known IgG glycosylation pathway (Figure 2.2) and the data-driven GGM (Figure 3.3B). To do this, we defined the "pathway distance" between any pair of glycans as the minimum number of enzymatic steps separating the two structures—for example, two glycans that corresponded to the reactant and product of a single enzymatic reaction in the IgG glycosylation pathway had a pathway distance of 1, whereas the shortest path from G0 to G2S1 includes three enzymatic steps, giving them a pathway distance of 3. We could not interpret correlations between glycans with the same structure belonging to different IgG subclasses in terms of the enzymatic reactions because they are bound to different proteins, and so we labeled these "X" (Figure 3.4). All other cross-subclass glycan pairs were ignored in our analysis.

Significant Pearson correlation coefficients were found for both short and longer pathway distances (Figure 3.4A); however, there were far more significant partial correlation coefficients at a pathway distance of 1 (Figure 3.4B) than at any other pathway distance, demonstrating that significant partial correlations tend to occur between glycans that are directly connected in the pathway. To assess whether significant partial correlations occurred more often at a given pathway distance than expected by chance, we performed a Fisher's exact test. The results of the test were highly significant ($P = 3.41 \cdot 10^{-39}$; Figure 3.4C and D), proving that there is a strong relationship between the data-driven GGM and the known IgG glycosylation pathway.







D

	Path Dist =1	Path Dist ≠1	Total
Significant correlations	35	31	66
Insignificant correlations	12	1147	1159
Total	47	1178	1225

Figure 3.4: Systematic comparison of correlations and pathway distances. **A**, **B** Pearson and partial correlation coefficients, respectively, versus pathway distance. Gray dashed lines represent the significance threshold (FDR = 0.01). On each box, the central mark indicates the median, and the whiskers indicate the 25th and 75th percentiles, respectively. The label "X" represents correlations between the same glycoforms across different immunoglobulin G (IgG) subclasses. In line with the network visualization, we observed significant Pearson correlation coefficients across all pathway distances, suggesting that Pearson correlations are non-specific with respect to the IgG glycosylation pathway. By contrast, significant partial correlation coefficients accumulated at a pathway distance of 1. The black dashed oval highlights significant partial correlations for pathway distances >1. **C** P-values for Fisher's exact tests for both Pearson and partial correlations at different pathway distances. There were significantly more significant partial correlations between glycans with a pathway distance of 1, demonstrating a close relationship between the IgG glycosylation pathway and the reconstructed GGM. The log10 p-values can be interpreted as a variance-normalized measure of the effect size. **D** Contingency table for the partial correlations at a pathway distance of 1. Entries represent the numbers of partial correlations satisfying the corresponding conditions.

3.3 Rule-based prediction of new enzymatic reactions

Above, we demonstrated that significant partial correlation coefficients represent pairs of glycans that are directly linked in the known IgG glycosylation pathway. Interestingly, however, there were also 22 significant partial correlations for pathway distances greater than one (and not contained in the "X" group), as indicated by the black oval in Figure 3.4B. Therefore, given the strong evidence for a relationship between the GGM and the IgG glycosylation pathway, we hypothesized that these correlations represented true but yet unknown pathway reactions. In principle, all glycans that differ in structure by a single monosaccharide could be connected by a reaction performed by one of the four enzymes involved in the glycosylation pathway shown in Figure 2.1; and among these 22 significant correlations, 15 (68%) differed by only one sugar residue. Furthermore, if we discard the seven negative partial correlations, whose interpretation has been shown to be problematic [32], this increases to 88% (see Supplementary Table 2 in the original publication for details). Thus, all 15 of these glycan pairs are candidates for direct enzymatic reactions.

To analyze this quantitatively, we tested whether these unexplained partial correlations could be attributed to missing steps in the known pathway. To do this, we first created a list of all possible novel pathway reactions, i.e., all connections between glycan structures that only differed by a single sugar unit and that were not present in the known IgG glycosylation pathway. Since we followed an unbiased approach, this included reactions for which in vitro experiments showed evidence of inhibition, e.g. the addition of fucose to the G0N structure [167]. We then divided these initial reactions into sets of "rules" according to the features of the hypothetical substrate and the corresponding enzyme performing the reaction (Figure 3.5A and Table 3.1)— i.e., we built the rules to account for previously undescribed substrate specificities for the four glycosyltransferases involved in IgG glycosylation. For example, the first rule (F1) describes the fucosylation of galactosylated, non-bisected glycans, as these reactions are not included in the known pathway. In this way, starting from 22 single potential new reactions, we defined six rules, as described in Figure 3.5A and Table 3.1.

The rationale for our inference method and model selection technique was that the pathway model that contains the greatest proportion of true reactions should produce the lowest p-value with a Fisher's exact test, as seen in Figure 3.4C. In total, we considered 63 pathway models that extended the known glycosylation pathway using all combinations of the six rules described above. To obtain a robust model fit, we performed bootstrapping with 10,000 resamplings and calculated 95% confidence intervals for each p-value distribution.

We considered a pathway model to fit the data significantly better than the known pathway

Table 3.1: Rules for pathway inference.

Substrate	Enzyme	Product	Rule name
galactosylated non-bisected non-fucosylated	FUT8	galactosylated non-bisected fucosylated	F1
galactosylated bisected non-fucosylated	FUT8	galactosylated bisected fucosylated	F2
non-galactosylated bisected non-fucosylated	FUT8	non-galactosylated bisected fucosylated	F3
mono-galactosylated mono-sialylated	B4GalT1	di-galactosylated mono-sialylated	G1
galactosylated non-bisected non-fucosylated	MGAT3	galactosylated bisected non-fucosylated	N1
galactosylated non-bisected fucosylated	MGAT3	galactosylated bisected fucosylated	N2

if it had a lower Fisher's test p-value and its 95% confidence interval did not overlap with that of the known pathway. Where several proposed pathway models were found to perform significantly better than the known pathway, we chose the simpler model, i.e., the one that included the fewest rules. Note that for this analysis we used p-values as variance-normalized measures of effect size for model comparison, rather than as the probability of an event occurring by chance. Figure 3.5B shows a comparison of the p-values for the known pathway, the known pathway extended with any one of the six defined rules, and all combinations that gave a significantly better p-value than the known pathway alone. A list with the results for all 64 (26) pathway models, including the known pathway for reference, can be found in Supplementary Data 2 of the original publication.

In the selected pathway model from this analysis, rules G1 and N2 were added to the known pathway (Figure 3.5C), which resulted in the inclusion of eight new enzymatic steps in the IgG glycosylation pathway. By considering this selected model as the ground truth and reclassifying all partial correlations according to the pathway distances derived from this extended model, we found that most of the significant partial correlations that had longer distances in the original IgG glycosylation pathway (Figure 3.4B) had a pathway distance of 1 in the modified IgG glycosylation pathway (Figure 3.5D). Note that the pathway model that included all possible enzymatic reactions (model "F1F2F3G1N1N2" in Figure

3.1) did not yield the lowest p-value, indicating that the addition of more reactions than required to provide the optimal pathway model impaired the result.

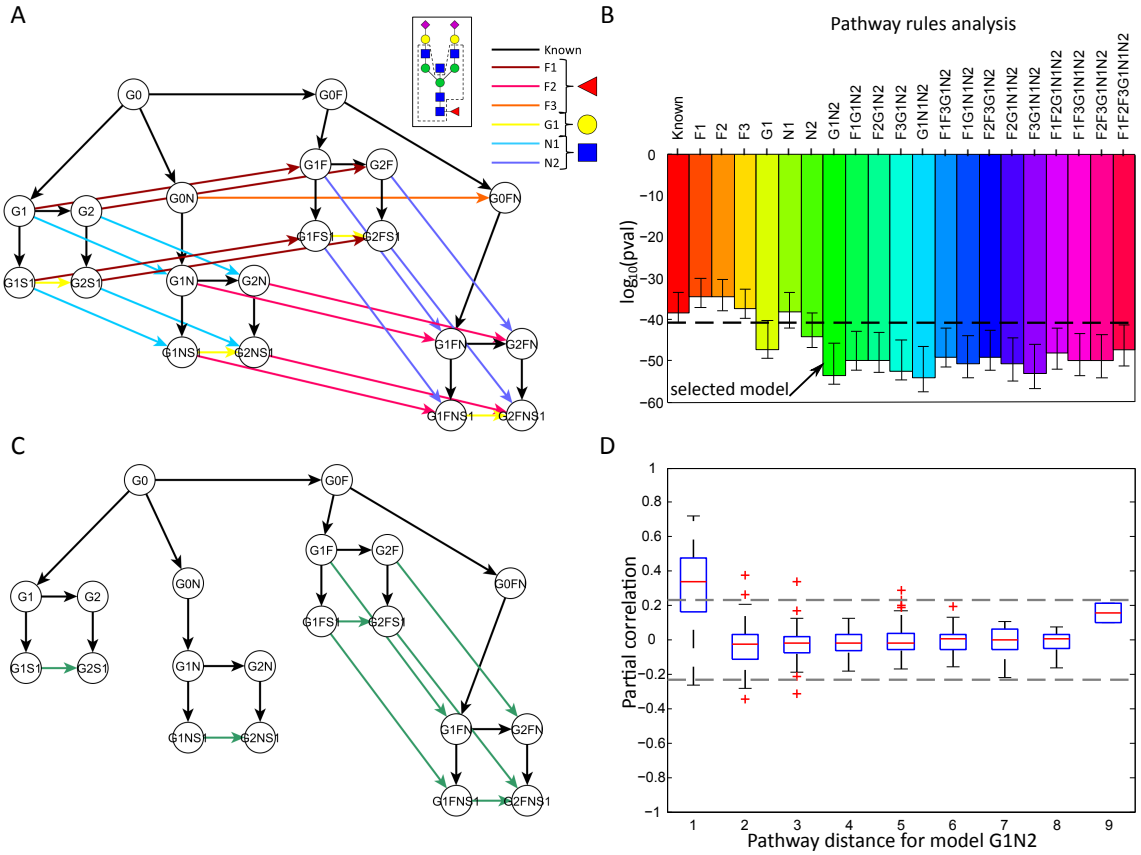


Figure 3.5: Rule-based approach for pathway inference. **A** Sketch of all single-monosaccharide additions in the immunoglobulin G (IgG) glycosylation pathway. The black network represents the known IgG glycosylation pathway, while arrows with the same color describe a rule. Shades of the same color represent reactions performed by the same enzyme. **B** Fisher's exact test results for the addition of different combinations of rules to the known pathway. The pathway model that most resembles the biological truth is expected to have the best overlap with the calculated GGM and hence yield a lower p-value. The black dashed line represents the lower end of the 95% confidence interval of the p-value for the known pathway obtained by bootstrapping. The simplest model that was significantly more accurate than the known pathway is indicated by a black arrow and includes rules G1 and N2. For a full list of results, see Supplementary Data 2 in the original publication. **C** Pathway model inferred by our approach. **D** Partial correlation coefficients for different pathway distances in the selected model. On each box, the central mark indicates the median, and the whiskers indicate the 25th and 75th percentiles, respectively. Most of the significant partial correlations that were a long distance apart in the known pathway are now at a distance of 1 (cf. Figure 3.4B).

3.4 Replication in three additional cohorts

We replicated these findings using IgG glycomics data measured on the same platform in three independent Croatian cohorts (Table 2.1). We again observed that most partial correlation coefficients between glycans were positive (Figure 3.6) and that the calculated GGM displayed a highly modular structure with respect to the IgG subclasses (Figure 3.7).

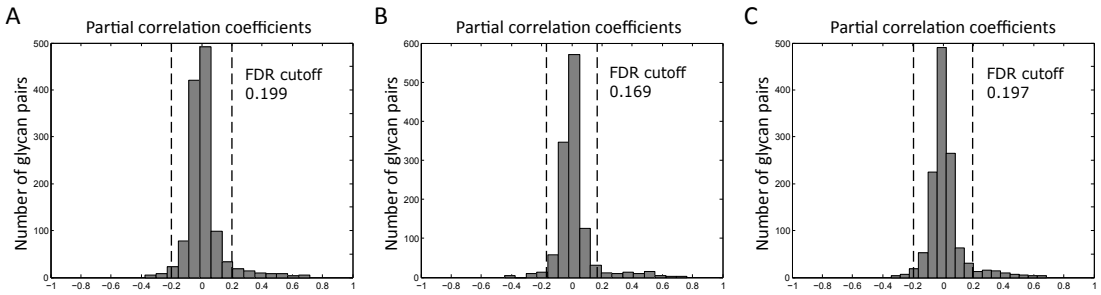


Figure 3.6: Distributions of partial correlation coefficients in the replication cohorts: Korčula 2010 (A), Split (B), Vis (C).

Moreover, pathway analysis showed that the edges represented single enzymatic steps in the IgG glycosylation pathway in all GGM networks (Figure 3.8). When inferring possible additional enzymatic steps, we again found that the addition of rules G1 and N2 to the known pathway gave a significantly better overlap with the GGM (Figure 3.9), providing further evidence that the enzymatic reactions included in these rules represent true steps in the IgG glycosylation pathway. The GGMs for all cohorts can be found in Supplementary Software. To quantitatively evaluate the agreement between GGMs

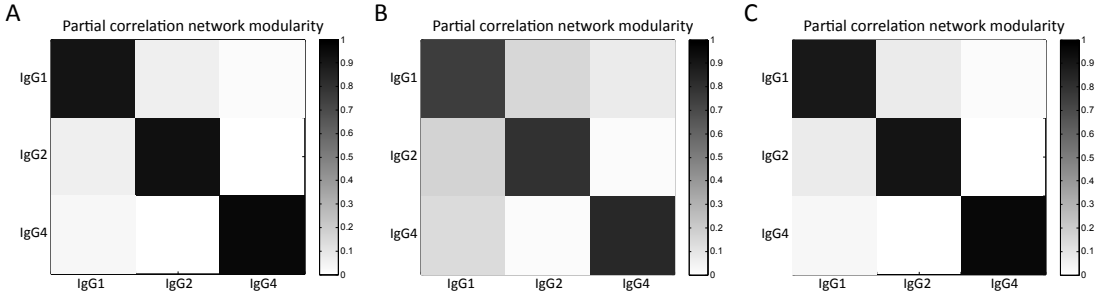


Figure 3.7: Network modularity of the partial correlation coefficients in the replication cohorts: Korčula 2010 (A), Split (B), Vis (C).

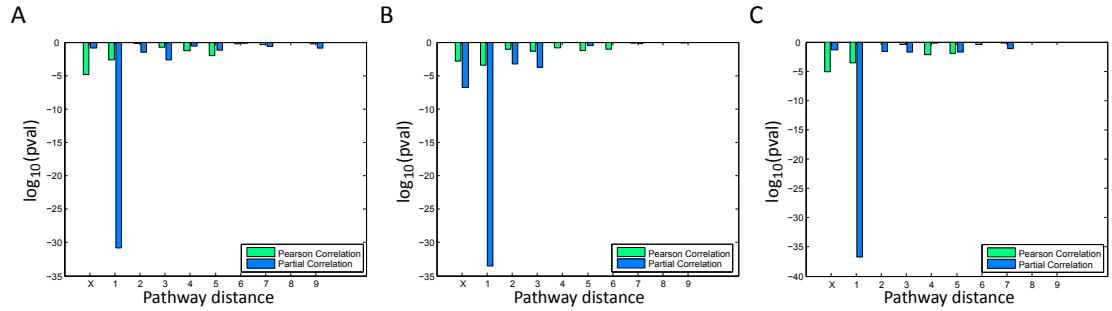


Figure 3.8: Pathway analysis in the replication cohorts: Korčula 2010 (A), Split (B), Vis (C).

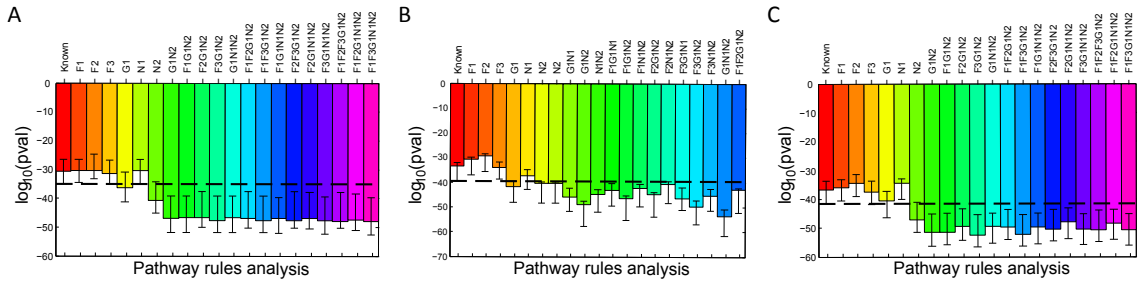
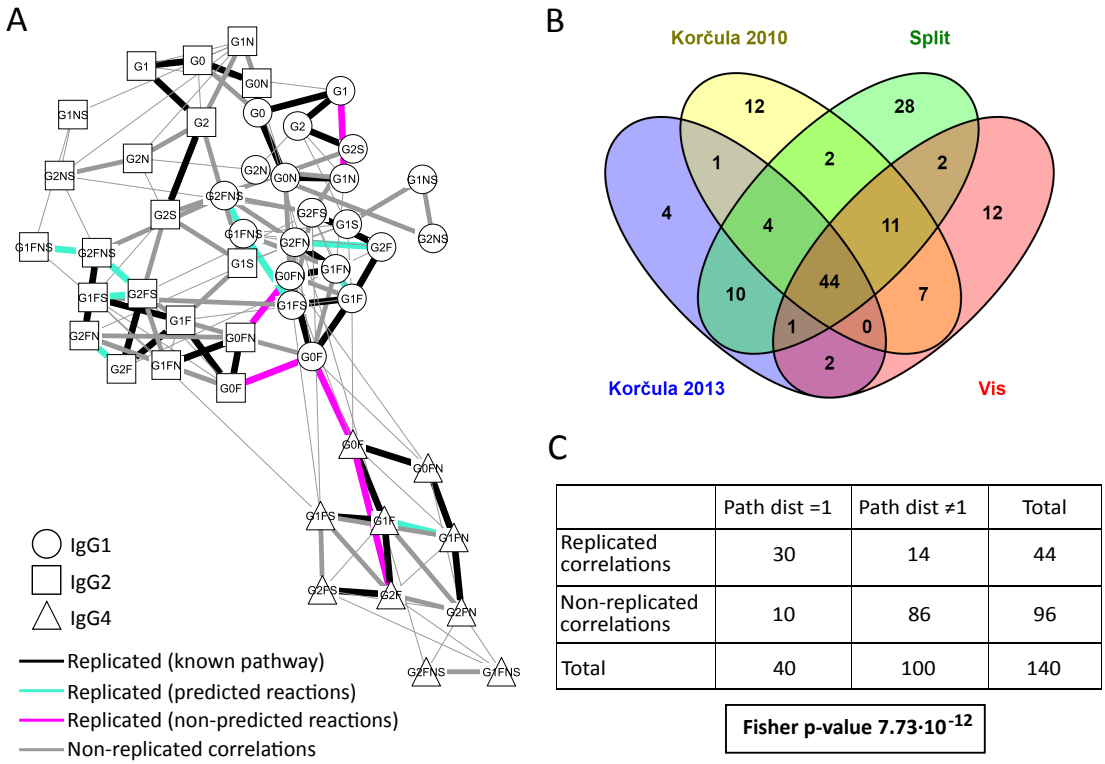


Figure 3.9: Rule-based pathway inference in the replication cohorts: Korčula 2010 (A), Split (B), Vis (C).

across the four cohorts, we generated a consensus network that represented the overlap between the networks (Figure 3.10A). We considered an edge to be "replicated" if it was significant in all four cohorts. This showed that 44 of the 140 significant correlations were replicated across all four cohorts (Figure 3.10B). To investigate how these edges related to the IgG glycosylation pathway, we again performed a Fisher's exact test. We only considered partial correlations that were found to be significant in at least one cohort, and built a contingency table that classified these according to their replication status and pathway distance (Figure 3.10). The highly significant result of this test ($P = 7.73 \cdot 10^{-12}$) indicates that the replicated edges tend to represent true pathway reactions even more strongly than the non-replicated edges do, demonstrating that partial correlations corresponding to single enzymatic reactions in the IgG glycosylation pathway were robustly identified in all cohorts.

3.5 GWAS evidence for predicted reactions

We applied a GWAS-based approach on an independent cohort to provide evidence-based validation, assuming that significant associations between glycan ratios and single nucleotide polymorphisms (SNPs) in the IgG glycosyltransferase genes indicate that the



	Path dist =1	Path dist ≠1	Total
Replicated correlations	30	14	44
Non-replicated correlations	10	86	96
Total	40	100	140

Fisher p-value $7.73 \cdot 10^{-12}$

Figure 3.10: Replication. **A** Consensus network. Black edges represent replicated partial correlations that correspond to direct enzymatic steps in the known immunoglobulin G (IgG) glycosylation pathway, green edges represent replicated edges matching the reactions predicted by our approach, and red edges represent replicated correlations corresponding to reactions that were not predicted to take part in IgG glycosylation. Replicated edges were defined as partial correlations that were significant in all four cohorts. Gray edges represent partial correlations that were significant in at least one cohort but not in all four. Note that three of the five replicated but non-predicted edges linked the same glycan structure in different IgG subclasses, which we did not consider in our inference approach. Thus, there are only two edges that are truly non-predicted. **B** Venn diagram of the significant partial correlations in the four cohorts. In total, 44 edges were shared among all four cohorts. **C** Contingency table for the partial correlation coefficients that were found to be significant in at least one of the four considered cohorts. The classification variables in this case are replication status and pathway distance. Here, we considered edges that were significant in at least one of the four cohorts, and we considered an edge to be replicated if it occurred in all four cohorts. The resulting p-value was very low, indicating that replicated edges are more likely to represent enzymatic reactions than non-replicated edges.

underlying reactions truly exist. This rationale is based on previous studies on blood metabolomics data, in which ratios of two metabolites were frequently found to be associated with genetic variation in the gene region of their catalyzing enzymes (see e.g. Gieger et al., 2008 [163]; Suhre et al., 2011 [165]; Shin et al., 2014 [166]). To quantify the increase of association strength of the ratio with respect to the single glycans, p-gains as defined in Petersen et al. (2012) [164] were used. Only significantly associated ratios with a sufficient

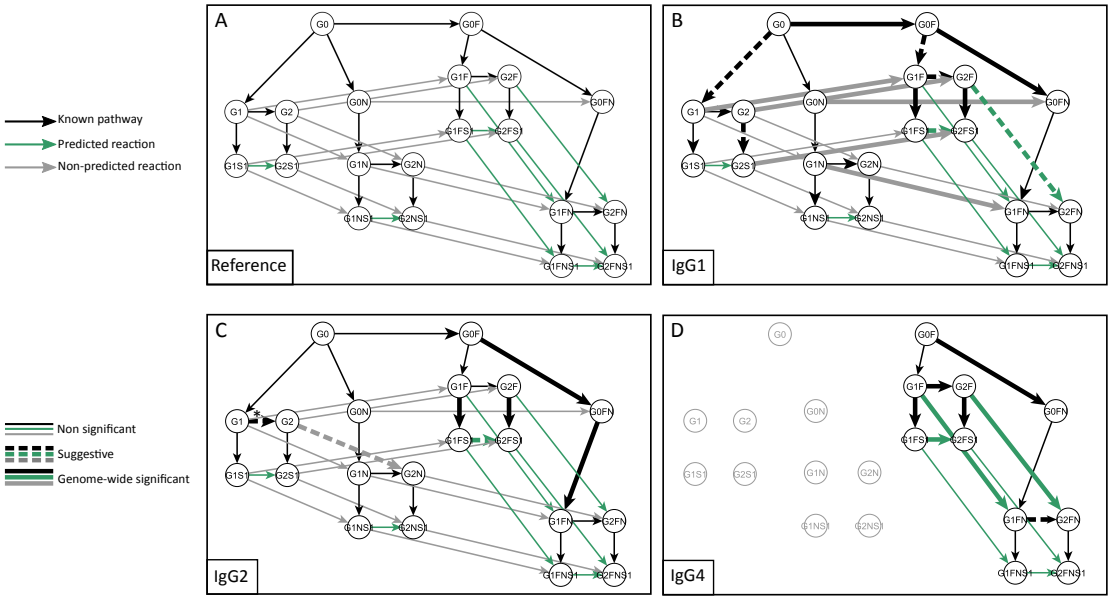


Figure 3.11: Genome-wide association study (GWAS) results for glycan ratios. **A** Reference pathway for interpreting the GWAS results. Black lines represent reactions in the known immunoglobulin G (IgG) glycosylation pathway, green lines represent reactions associated with the predicted rules, and gray lines represent possible reactions that were not selected by our approach. **B**, **C**, **D** GWAS results for IgG1, IgG2, and IgG4, respectively. Solid thick arrows represent ratios that were significantly associated with single nucleotide polymorphisms (SNPs) in the regions coding for an IgG glycosylation enzyme ($P < 5.26 \cdot 10^{-10}$ and p-gain>10). Dashed arrows represent suggestive associations ($5.26 \cdot 10^{-10} \leq P \leq 10 - 7$ and p-gain>10). Gray nodes in the IgG4 plot represent glycoforms that were not measured. The asterisk (*) indicates that the ratio was unexpectedly associated with SNPs in the FUT8 gene region.

p-gain (see Methods) were considered to confirm a given enzymatic reaction. For this analysis, we used glycomics data from the German population study KORA F4 [140]. Plasma IgG Fc N-glycopeptide measurements were obtained using the same LC-ESI-MS platform as for the discovery and replication cohorts described above, and included the same 50 measured glycoforms. Linear associations with genetic variants were calculated using the logarithm of all glycan product–substrate ratios defined in Figure 3.5A (see Methods). We considered SNPs in the four glycosyltransferase genes involved in IgG glycosylation, namely ST6GAL1, B4GALT1, FUT8, and MGAT3 (see Supplementary Table 1 in the original publication for details). As a positive control, we first verified that the glycan product–substrate ratios in the known pathway were significantly associated with loci in the regions coding for the enzymes that are catalyzing the reactions. We found that 12 out of 47 ratios were genome-wide significant, while another five met a suggestive p-value of 10^{-7} (Figure 3.11, thick black lines). Interestingly, we also found one ratio (G2/G1 in IgG2) that was associated with genetic variants in the region of the enzyme FUT8,

which is responsible for the addition of core fucose (Figure 3.11, arrow with asterisks). This was unexpected as neither of the structures in the ratio are fucosylated. For our 22 predicted reactions, we found three significant and three suggestive hits (Figure 3.11, thick green lines). Importantly, these significantly associated ratios tended to be the same across the three IgG subclasses and were equally distributed across the predicted rules. We found three confirmations for the rule G1 and three for the rule N2. By contrast, five significant associations and one suggestive hit were observed among the 26 ratios that were not predicted by our approach, but these did not replicate across subclasses (Figure 3.11, thick gray lines). In particular, three out of these six non-predicted reactions originated from rule F1 exclusively in IgG1 and could not be replicated in IgG2, while the other three hits spread across three different rules. Overall, we found at least one genome-wide significant association for all of the considered genes, providing evidence that we could indeed investigate all four glycosyltransferase enzymes involved in IgG glycosylation. In the supplementary information of the original publication, the complete GWAS results are provided in Supplementary Data 3, while the regional association plots in Supplementary Figure 7.

3.6 Experimental validation by enzymatic assays

To address different aspects of our predictions, our collaborators performed three sets of in vitro experiments: two enzymatic assays and one colocalization experiment.

In a first experiment, we aimed to verify whether GalT1 and MGAT3 exhibited the predicted, previously unknown substrate specificities. To this end, Prof. Gordan Lauc from Genos and Prof. Kelley W. Moremen from the University of Georgia compared UPLC spectra of pooled IgG glycans before and after exposure to the two enzymes. We considered seven different experimental conditions, covering various combinations of enzyme concentrations as well as negative controls (lacking sugar donors) that are not expected to show any reaction (Figure 3.12A). As expected, GalT1 efficiently galactosylated a number of glycans in the IgG glycome (see Supplementary Data 4 in the original publication for details). To investigate our inferred reactions, we focused on the ratio of the substrate G1FS1 and product G2FS1. With increasing concentrations of added GalT1 enzyme (25, 50 and 100 ng), this ratio drops significantly compared to the respective negative controls (Figure 3.12A), directly confirming one of the predicted reactions in rule G1 in a concentration-dependent manner.

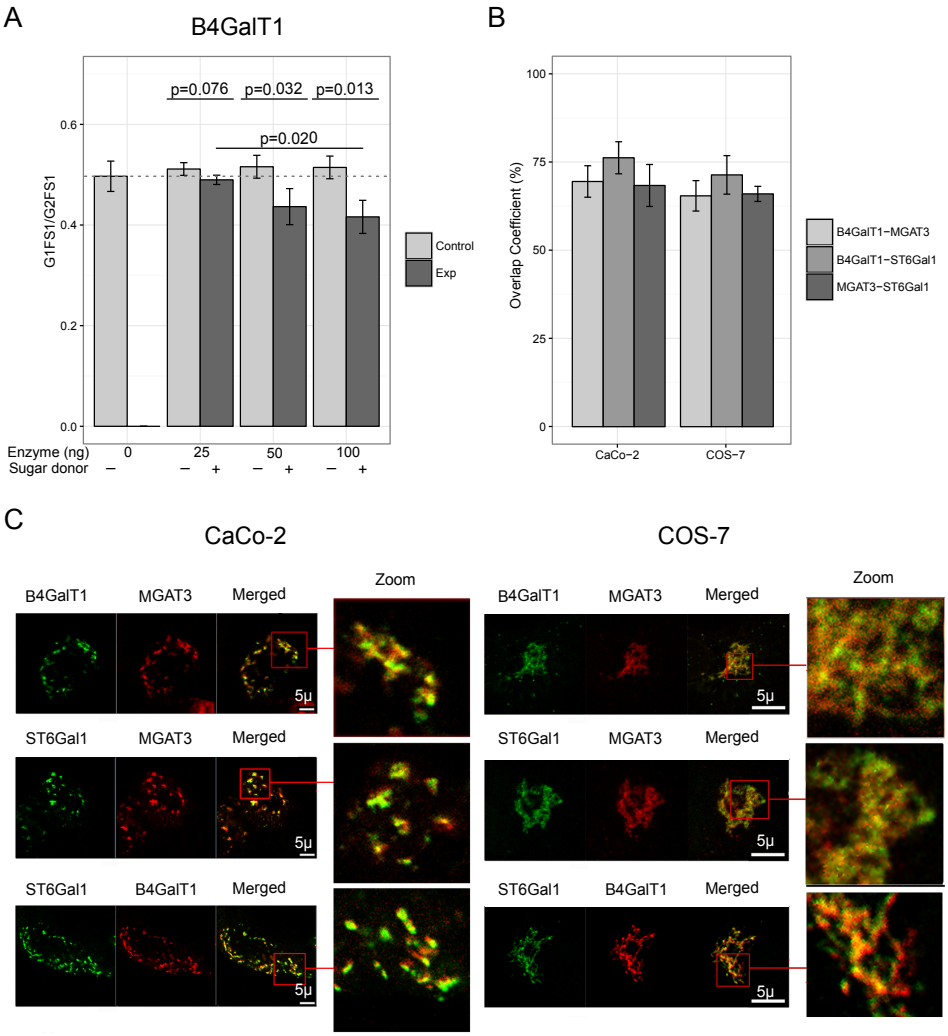


Figure 3.12: Experimental validation results. **A** In vitro enzymatic assay. The figure illustrates the ratio of G1FS1 over G2FS1 across different concentrations of the enzyme (B4GalT1), and in presence or absence of sugar donors. Bars represent the average value over triplicates, while error bars represent standard deviations. With increasing enzyme concentrations, the glycan ratio decreases significantly with respect to the corresponding negative control, confirming the occurrence of the predicted reaction. p-values were obtained from a two sample t-test. **B** Quantitative overlap between the localization of the three enzymes. The overall colocalization of each enzyme pair is expressed as an overlap coefficient percentage (mean % ± standard deviation). We observe substantial colocalization of all enzyme pairs in both cell lines. **C** Exemplary colocalization images of B4GalT1, ST6Gal-1 and MGAT3 in CaCo-2 (left) and COS-7 (right) cells, used for the overlap quantification in B. The individual figures represent a typical view from 5 different Golgi areas examined. In the images labeled as "Merged" and "Zoom", yellow areas represent enzymatic overlap. Due to the dispersed Golgi stacks throughout the cytoplasm in CaCo-2 cells, the overlap can be observed clearly in separated cisternae, proving that localization of the glycosyltransferases is not limited to cis-, medial-, or trans-Golgi areas. Bar represents 5μm.

When performing the analogous experiment for MGAT3, however, we were not able to see

addition of the bisecting GlcNAc to any of the glycans (not even reactions in the known IgG glycosylation pathway) (Supplementary Data 4 in the original publication). This might indicate that the fluorescent label attached to the IgG glycans interferes with the enzymatic reaction. For this reason, we were not able to experimentally prove or disprove any enzymatic reaction in rule N2.

In a second experiment, we focused on a reaction from an excluded rule, namely the addition of bisecting GlcNAc to galactosylated non-fucosylated glycans (rule N1). To this end, Prof. Nicolai Bovin from the Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry in Moscow performed the enzymatic reaction with a pure G2 synthetic glycopeptide. For positive control, we also considered a G0 glycopeptide, a known substrate for MGAT3.

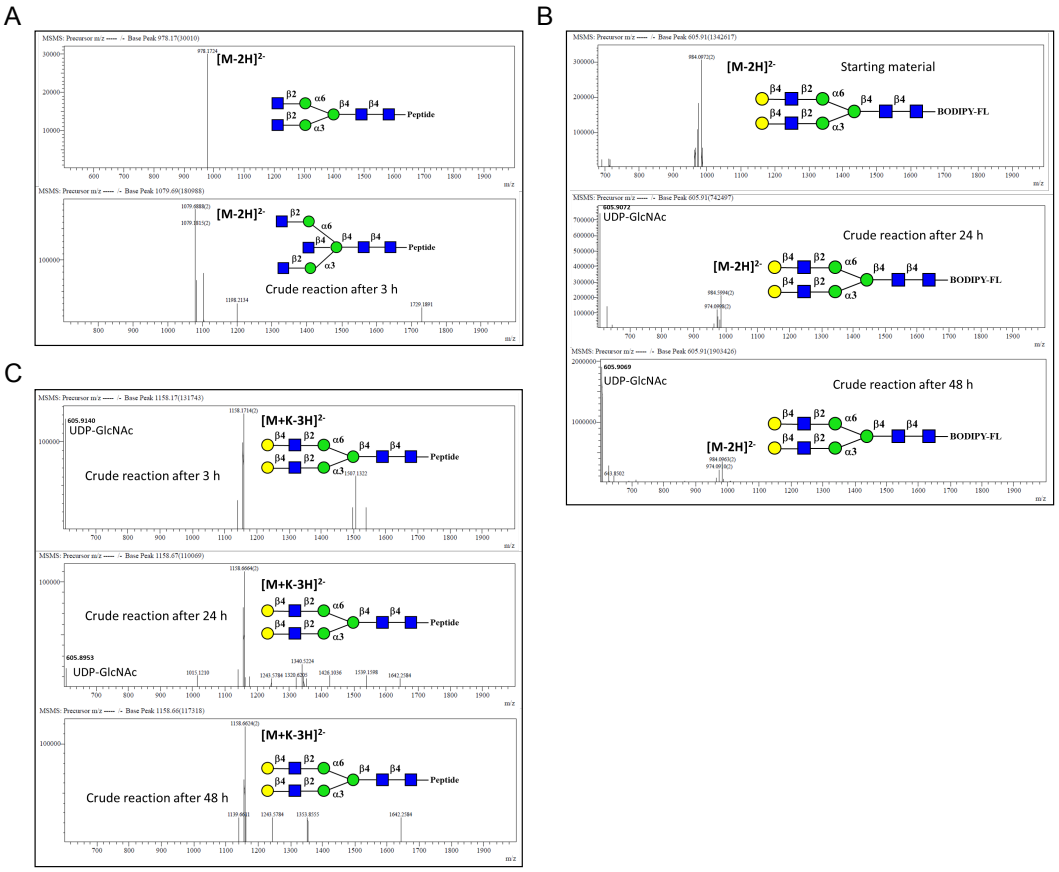


Figure 3.13: Enzymatic assay on synthetic substrates for MGAT3. **A** Initial glycopeptide: G0. The structure is a known substrate for MGAT3 and after three hours the substrate was completely converted into product (G0N). **B** Initial glycopeptide: G2. The substrate did not show any conversion to the product after 24 or 48 hours. This is compatible with our prediction that rule N1 is infeasible. **C** Initial structure: fluorescently labeled G2. The previous result was confirmed also with this alternative substrate, which did not show any conversion after up to 48 hours.

While the G0 structure was completely converted to product within 3 hours (Figure 3.13A), there was no measurable addition of bisecting GlcNAc to the G2 structure even after 48 hours (Figure 3.13B). The latter result was further confirmed using a fluorescently labelled G2 glycan structure, which also did not show any conversion to G2N (Figure 3.13C). These experiments are thus supportive of our prediction that rule N1 is not enzymatically feasible.

3.7 Enzyme colocalization experiments in cell lines

In vitro evidence for enzymatic reactions does not necessarily translate to in vivo conditions. A general consensus in the field has been that Golgi glycosyltransferases mainly localize in the stack of cisternae according to their expected order of functioning [168,169]. In particular, this is expected to prohibit the bisection of galactosylated glycans due to the different localization of the enzymes. In contrast, our predictions suggest that the addition of bisecting GlcNAc can occur also on galactosylated, fucosylated glycans (rule N2).

To address this aspect, Prof. Sakari Kellokumpu from the University of Oulu performed colocalization experiments of the enzymes involved in our predicted reactions, namely B4GalT1, MGAT3 and ST6Gal1, in kidney COS-7 cells and CaCo-2 colorectal cancer cells. Evidence of such a colocalization between the three glycosyltransferases would indicate that our predictions are, in fact, not impossible. Localization of of the enzymes in the Golgi stacks of cisternae (cis-, -medial, -trans) was assessed using confocal microscopy and Z-stack imaging with Venus- or Cherry-tagged enzyme constructs expressed at modest levels both in COS-7 and CaCo-2 cells. The latter have the advantage of having Golgi stacks dispersed throughout the cytoplasm, facilitating colocalization analyses at the level of individual Golgi stacks and, thereby, aiding interpretation of the imaging data. In addition, cells were stained with anti-GM130 cis-Golgi marker antibody.

The overlap between the enzymes and GM130 was on average 62%/73% (MGAT3), 55%/67% (B4GalT1) and 51%/67% (ST6Gal1) in CaCo-2/COS-7 cells, respectively (Figure 3.14).

This means that, to different degrees, all three enzymes can be found in the cis-part of the Golgi. Higher overlap percentages were detected in COS-7 cells due to the more compact Golgi architecture in the cells. Comparing the overall localization of the three enzymes, we observed prominent colocalization. The overlaps were quantified as 69%/65% (B4GalT1-MGAT3), 76%/71% (B4GalT1-ST6Gal1) and 68%/66% (MGAT3-ST6Gal1) in

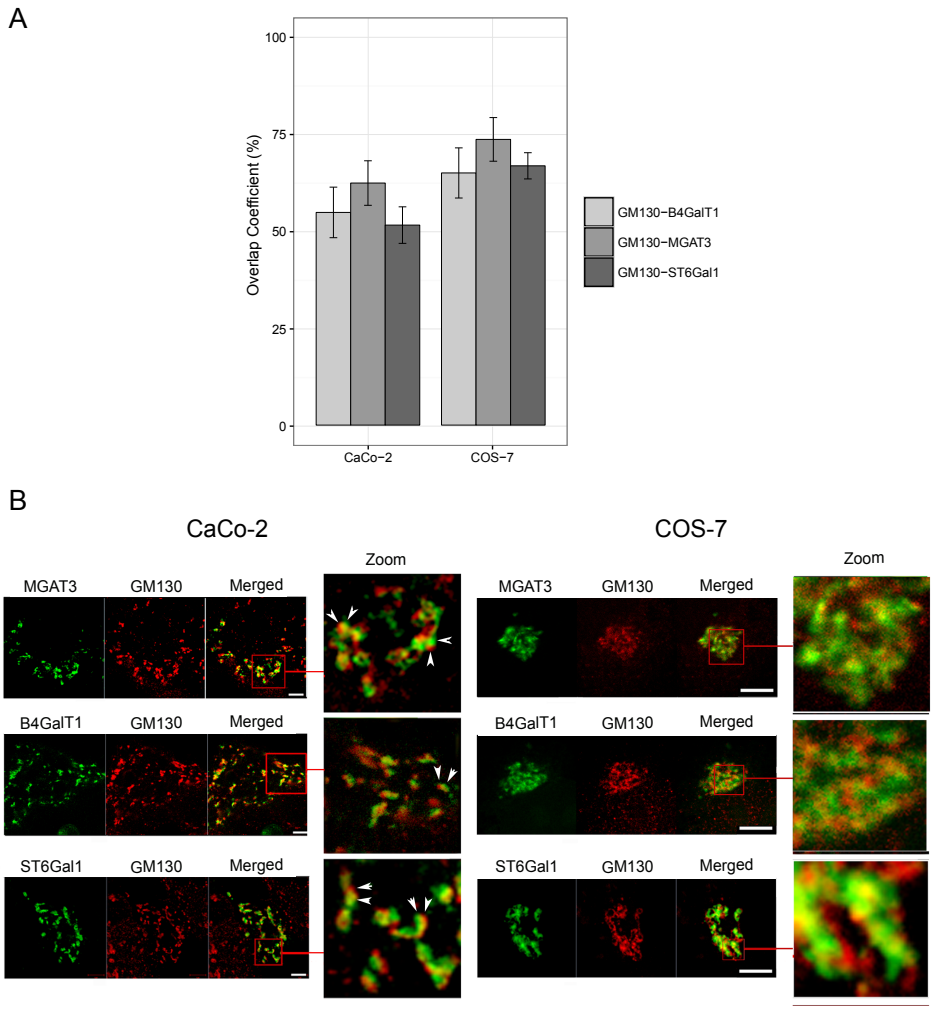


Figure 3.14: Colocalization experiment. **A** Quantitative overlap between the three enzymes and the cis-Golgi marker GM130. The overall colocalization of each enzyme pair was expressed as overlap coefficient percentage (mean % ± standard deviation) obtained using pixel per pixel comparison of each Z-stack image and by combining the values across 5 cells each. **B** Colocalization images of B4GalT1, ST6Gal-1 and MGAT3 with GM130 in CaCo-2 (left) and COS-7 (right) cells. COS-7 cells and CaCo-2 cells were grown on cell culture plates and transfected with the defined plasmids, and processed for confocal Z-stack imaging (0.3 μm sections). The individual figures represent a typical view from 5 different Golgi areas examined. In the images labeled as "Merged" and "Zoom", yellow areas represent localization overlap. The cis-Golgi GM130 marker protein and the three enzymes showed differential localization in CaCo-2 cells, indicating that the dispersed Golgi stacks are functionally polarized in this colorectal cancer cell line. This was evident from the observed differential distribution of the green and red colours within each Golgi stack of cisternae (arrowheads in figure). Yet, each enzyme showed partial colocalization with the cis-Golgi-marker GM130 based on coalescence of the two colours (yellow areas in figure) in the middle of the Golgi stack. Higher overlap percentages were detected in COS-7 cells due to the more compact Golgi architecture in the cells. Bar represents 5μm.

CaCo-2/COS-7 cells, respectively (Figure 3.12B,C). These data indicate that, unexpectedly, a substantial proportion of all these three enzymes are present in the same Golgi compartments, indicating that our newly proposed reactions described by rules G1 and N2 are compatible with enzyme localization inside the Golgi stacks of cisternae.

3.8 Conclusion

In this study, we demonstrated for the first time that GGMs can be used to reconstruct single enzymatic reaction steps in the glycan synthesis pathway using IgG Fc glycan measurements from human plasma. We also found that additional glycosylation reactions can be inferred from the calculated network, with the pathway rules G1 and N2 (Table 3.1) likely representing real biochemical steps in the IgG glycosylation pathway. Rule G1 represents the galactosylation of sialylated glycans. The current standard glycosylation pathway is based on immunohistochemical studies that were performed over 30 years ago, which suggested different subcellular localization of galactosyltransferases and sialyltransferases in the Golgi apparatus [170, 171].

Galactosylation is a prerequisite for sialylation, and so the hypothesis of physically separated enzymes implied that galactosylation could only occur prior to sialylation. However, it has recently been shown that these two enzymes are colocalized in COS7 cells and are likely to act as a complex [172]. Our results directly support this hypothesis, as rule G1 represents sialylation of IgG prior to further galactosylation. Rule N2 suggests that fucosylated, galactosylated glycans can be modified by adding a bisecting GlcNAc through MGAT3. Again, the standard glycosylation pathway assumes differential localization of the B4GalT1 and MGAT3 enzymes, and thus that the addition of bisecting GlcNAc could only occur prior to galactosylation. Previous studies have moreover indicated that overexpression of the B4GalT1 enzyme decreases the amount of bisecting GlcNAc [173], suggesting that the two enzymes might mutually inhibit each other's activity by competing for the same substrate. In contrast, our results suggest that the two enzymes may be colocalized and that galactosylated glycans could be direct substrates for MGAT3, hinting at a different regulation of the enzymatic activity than previously described.

As a limitation, it is to be noted that partial correlations calculated from glycomics data might not represent true biological processes in all cases. For example, we observed cross IgG-subclass correlations of the same glycan structures, which might be attributed to overall sugar or glycan abundances, rather than single enzymatic steps. Vice versa, not

all glycan pairs that share a biochemical reaction will necessarily show correlation in the data. Reasons for this could be too low concentrations of glycans or high turnover rates of the IgG antibodies in blood. However, we used the correlation-based methodology to generate novel pathway hypothesis for experimental testing, which does not require a perfect reconstruction of the pathway.

Our findings were replicated across the analyzed cohorts, suggesting that the mechanisms that regulate IgG glycosylation are conserved across different Croatian populations. To validate our hypothesis for new enzymatic reactions in the IgG glycan synthesis pathway, we performed GWAS on glycan product-substrate ratios. Previous GWAS analyses on total IgG glycans measured with ultra performance liquid chromatography (UPLC) in the Vis and Korčula 2010 cohorts revealed statistically significant associations between traits describing fucosylated, non-bisected glycans and the MGAT3 gene 9. Here, we used specific glycan product–substrate ratios as quantitative traits, allowing us to analyze individual reactions at an IgG subclass-specific level. Six ratios that corresponded to our predicted reactions were found to be significantly associated with SNPs in the gene regions coding for the enzymes involved in the putative reactions (three for rule G1 and three for rule N2, see Supplementary Table 4Data 3), further supporting our hypothesis of these novel pathway steps occurring in vivo.

The GWAS evidence stems from an in vivo system; however, it is an indirect association and does not provide proof for the predicted reactions. Therefore, we performed in vitro enzyme assays probing specific reactions from the inferred pathway model. We found evidence that the addition of galactose to monosialylated glycans via B4GalT1 is indeed possible (rule G1). Confirming or disproving reactions in rule N2 was not possible due to experimental limitations. Moreover, we were able to show that a reaction from the rejected rule N1 did indeed not occur in the experiment. In addition to substrate specificity, current knowledge of the physical distribution of enzymes across the Golgi apparatus implied a directed order of the enzymes involved in the glycosylation process, thus preventing our predicted reactions from occurring in cells. In contrast to this, we found that the three enzymes involved in our predictions (B4GalT1, MGAT3, ST6Gal1) strongly colocalize across the Golgi in two different cell lines, suggesting that, in fact, the reactions are not infeasible. Taken together, while full in vivo validation of the new reactions is out of reach at this point, we found substantial evidence supporting our prediction in in vitro experiments.

Future studies could build on our findings in several ways. (1) The predicted rules could be investigated at a single-reaction level to determine whether all or only some of the enzymatic steps described in rules G1 and N2 are included in the IgG glycan synthesis pathway.

(2) In addition, a single-reaction pathway inference approach could be used to explore the subclass-specific pathways suggested by some of our GWAS results. (3) The approach described in this thesis could also be used to analyze other glycomics datasets, obtained from different platforms (e.g. UPLC fluorescence [FLR], matrix-assisted laser desorption ionization–time-of-flight–MS [MALDI-TOF-MS], or multiplexed capillary gel electrophoresis with laser induced fluorescence detection [xCGE-LIF]; to investigate whether the same reconstructed pathways are produced. (4) Measurement techniques for total plasma glycomes, including glycoforms with extremely heterogeneous structures (i.e., high mannose, hybrid, truncated, and complex glycans) from approximately 24 glycoproteins in blood, have recently become available [174]. Therefore, it would be of major interest to apply our methodology to these more complex datasets, to determine whether partial correlations can be used to reconstruct single enzymatic reactions even when dealing with a heterogeneous set of glycoproteins. (5) Replication of the results should also be verified in a non-Croatian cohort, as population-specific effects may have gone undetected in this analysis. (6) From a theoretical perspective, an analytical formulation of the likelihood function of the different pathway models based on information criteria such as the AIC (Akaike) or BIC (Bayesian) would lead to more rigorous model selection.

In conclusion, in this study we demonstrated for the first time that GGMs based on large IgG glycomics datasets contain strong footprints of biochemical reactions in the IgG glycosylation pathway. We proposed an inference algorithm based on the accordance of GGMs and the candidate pathways, to improve our understanding of the complex process of protein glycosylation. Novel reaction steps could be partially validated using GWAS data and in vitro experiments. In general, the finding that GGMs can be used to represent single steps in glycan synthesis indicates that it may be possible to compare the GGMs from healthy and sick individuals to detect alterations in enzymatic activity of the glycosyltransferases, shedding light on the molecular mechanisms that regulate IgG glycosylation.

Chapter 4

Network-driven structural inference on the human total plasma N-glycome

After establishing the ability of GGMs to retrieve pathway steps from site- and protein-specific glycomics data, in this chapter we consider glycomics data from different plasma proteins.

Given the complexity and diversity of glycan structures, accurate measurements of large-scale glycomics datasets from human plasma are difficult to obtain. The optimal scenario would involve the quantification of all glycan structures in a protein- and site-specific fashion, so that associations to phenotypes and profile changes can be traced back to specific molecular entities and pathways. This is to this day experimentally infeasible. The overall glycome of plasma proteins can currently only be measured as a mixture, i.e., without retaining the information about the protein or site of origin.

Quantification is usually performed either via UPLC, or MALDI-TOF-MS (see Subsection 2.1.1). UPLC separates glycans according to their chemical-physical properties. The resulting chromatographic peaks, therefore, cannot be directly associated to specific glycan structures without further experiments, and substantially different structures could have very similar retention times, hence occurring within the same peak (see Subsection 2.1.1). MALDI-TOF-MS, on the other hand, identifies different molecular masses, which, per se, do not give any information about the order with which the glycan monosaccharides are linked together. Fortunately, prior knowledge allows for a significant reduction of the number of possibilities. For example, it has been shown that N-glycans in human

proteins always have two N-Acetylglucosamines at the core followed by two or three mannoses [175] (Figure 1.6). However, since human glycan building blocks include several epimers, i.e., monosaccharides made of the same atoms but in different configurations (e.g., glucose, mannose and galactose, which are all hexoses), different structures could contribute to the same mass, referred to as *composition* (Figure 4.1), and would therefore not be distinguishable by the measurement platform.

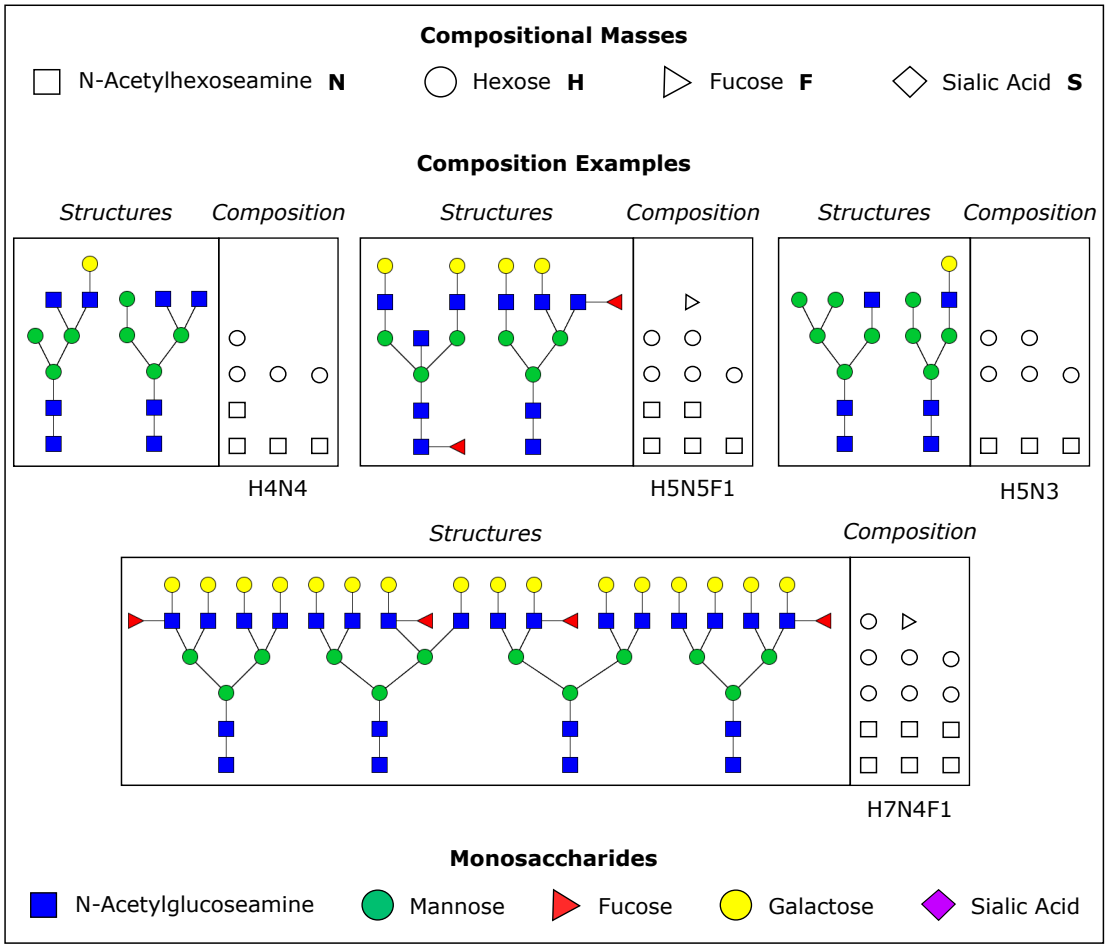


Figure 4.1: Example of the glycan structures within *compositions* measured via MALDI-TOF-MS. Due to the presence of epimers, very different glycan structures can have the same mass and hence not be distinguishable with the used platform.

Since a correct characterization of glycan structures is essential to understand their function and effect on protein activity, in this chapter we address the question of identifying the most abundant glycan structures within each composition measured in the human total plasma N-Glycome (TPNG) via MALDI-TOF-MS from the Leiden Longevity Study (Table 2.1) by exploiting the correlations among measured compositions combined with the available prior knowledge on glycan synthesis.

First, we design a *theoretical pathway* of glycosylation, considering known glycan structures from plasma proteins and allowing, among them, all synthetic steps that have not been proven enzymatically infeasible. In order to be able to relate this pathway to the measured data, we subsequently convert the structures to the corresponding compositions, generating a *compositional pathway*. We then compute data partial correlation coefficients and show that significant coefficients mainly correspond to single synthetic steps in the compositional pathway. We use this strong relationship between data-driven significant partial correlations and compositional pathway to predict the most abundant glycan structure within any composition. Our predictions are then validated using external data, and show that our data-driven approach is able to correctly identify the majority of structures.

In conclusion, we demonstrate that data-driven correlations and prior knowledge can be successfully integrated to gain insights into glycan structural details when they are not directly available from the data, reducing the costs of further fragmentation experiments.

4.1 Creation of the compositional pathway

Before proceeding with the data analysis, we investigated the biochemical pathway of the TPNG synthesis. While some of the enzymatic steps are well established, like the processing from high mannose to hybrid and complex glycans [18], others, like those among bi- tri- and tetra-antennary complex glycans, are still not well characterized. Moreover, different proteins might have different enzymatic affinities and thus potentially different pathways [13]. In order to group all the available information into one unified model, we built a theoretical pathway including all possible enzymatic reactions (i.e., all single monosaccharide modifications) against which there was no strong experimental evidence. Therefore, the resulting model describes all potential reactions of N-glycan synthesis across all proteins (Figure 4.2).

As mentioned above, the measured TPNG dataset included glycan masses, where each of these masses could correspond to one or more glycan structures. In order to relate the constructed pathway of glycan synthesis to the measured data, we created a pathway where the nodes corresponded to the 61 measured compositions.

Starting from the theoretical pathway (Figure 4.3A), we mapped each structure to its corresponding composition (Figure 4.3B). We subsequently merged all the nodes in the theoretical pathway with the same composition, and obtained a compositional pathway (Figure 4.3C). As not all compositions that appeared in the compositional pathway were measured in the available dataset, we deleted all unmeasured compositions together with

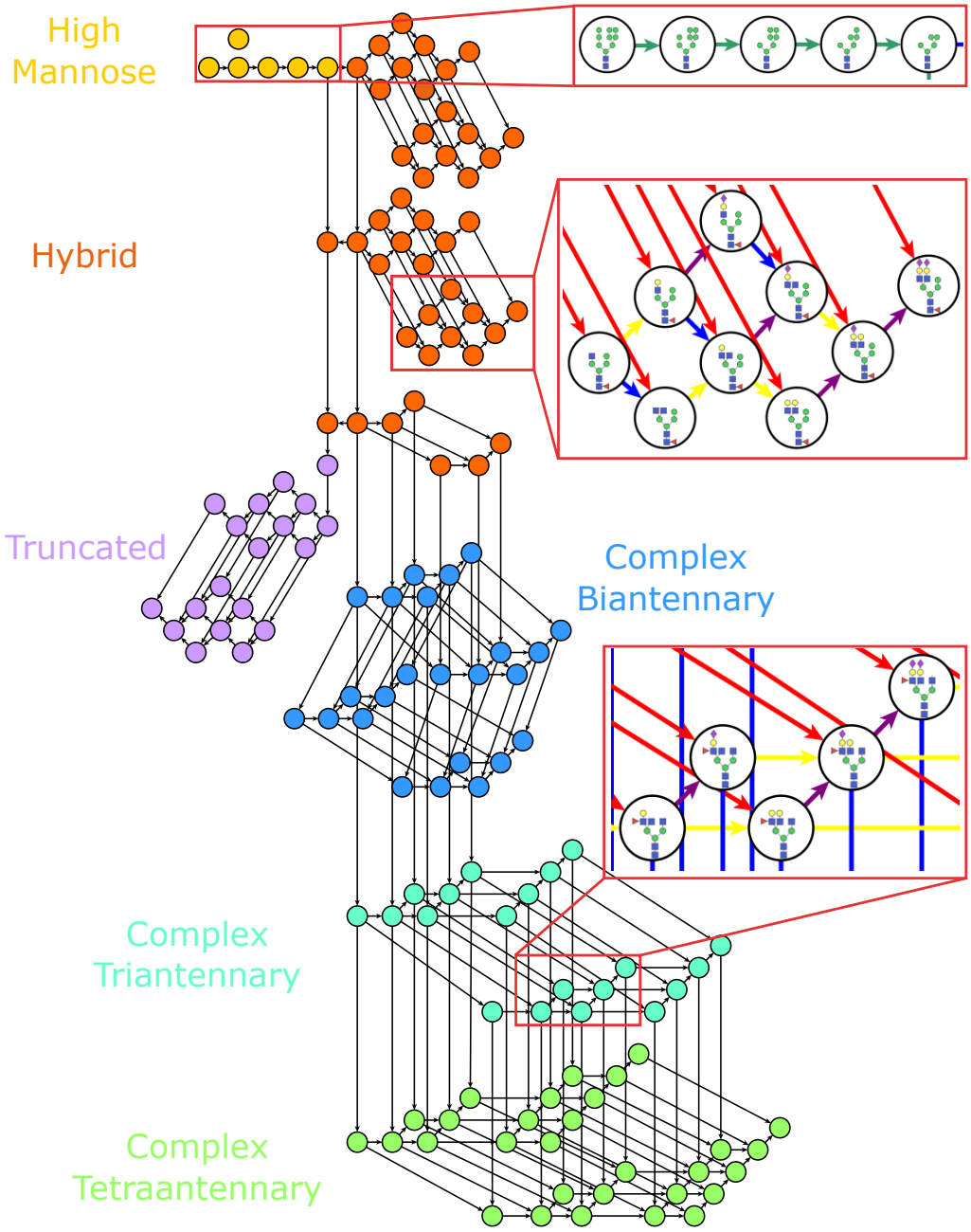


Figure 4.2: Theoretical pathway of TPNG synthesis. Each node corresponds to a glycan structure. Edges represent possible enzymatic reactions of glycan synthesis, namely all single monosaccharide modifications against which there is no strong experimental evidence against. Node colors represent the glycan classes (high mannose, hybrid, truncated, complex). The insets show examples of the synthesis pathway. A representation of the theoretical pathway with all corresponding glycan structures is reported in Figure 2.3.

their edges to have a model compatible with the available data (Figure 4.3D). The resulting compositional data pathway is shown in Figure 4.4 and contains the same 61 compositions

that were measured in the dataset considered in this thesis.

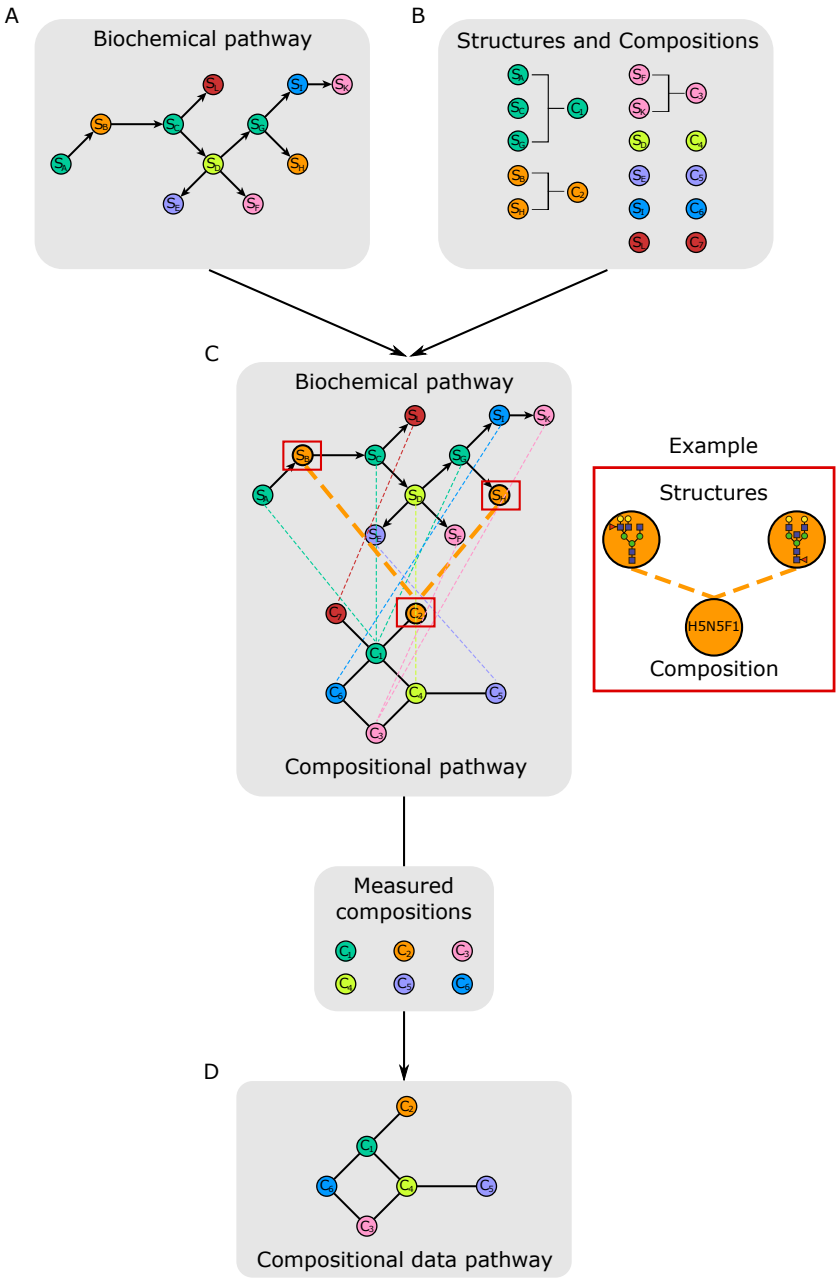


Figure 4.3: Creation of the compositional pathway. **A** Theoretical biochemical pathway. **B** Mapping between structures in the theoretical pathway and the corresponding compositional name. **C** Structures with the same mass in the theoretical pathway were merged into one single compositional node. **D** Since not all compositions in the theoretical pathway were measured in the available dataset, the unmeasured compositions were removed from the pathway together with all their edges.

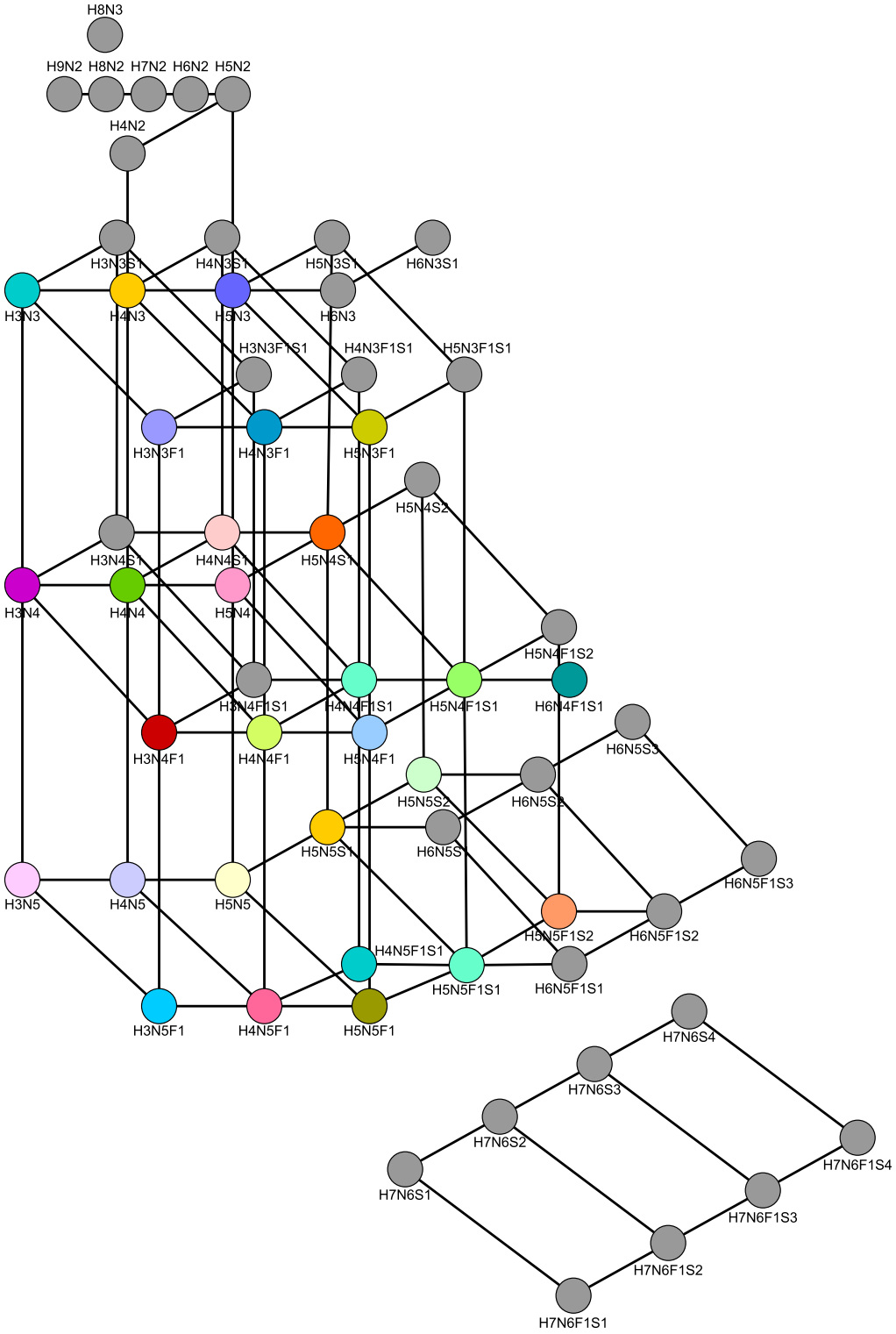


Figure 4.4: Compositional data pathway. This compositional data pathway included the same 61 compositions measured in the considered dataset. Grey nodes represent compositions composed by a single structure, while colored nodes represent compositions made of more than one structure.

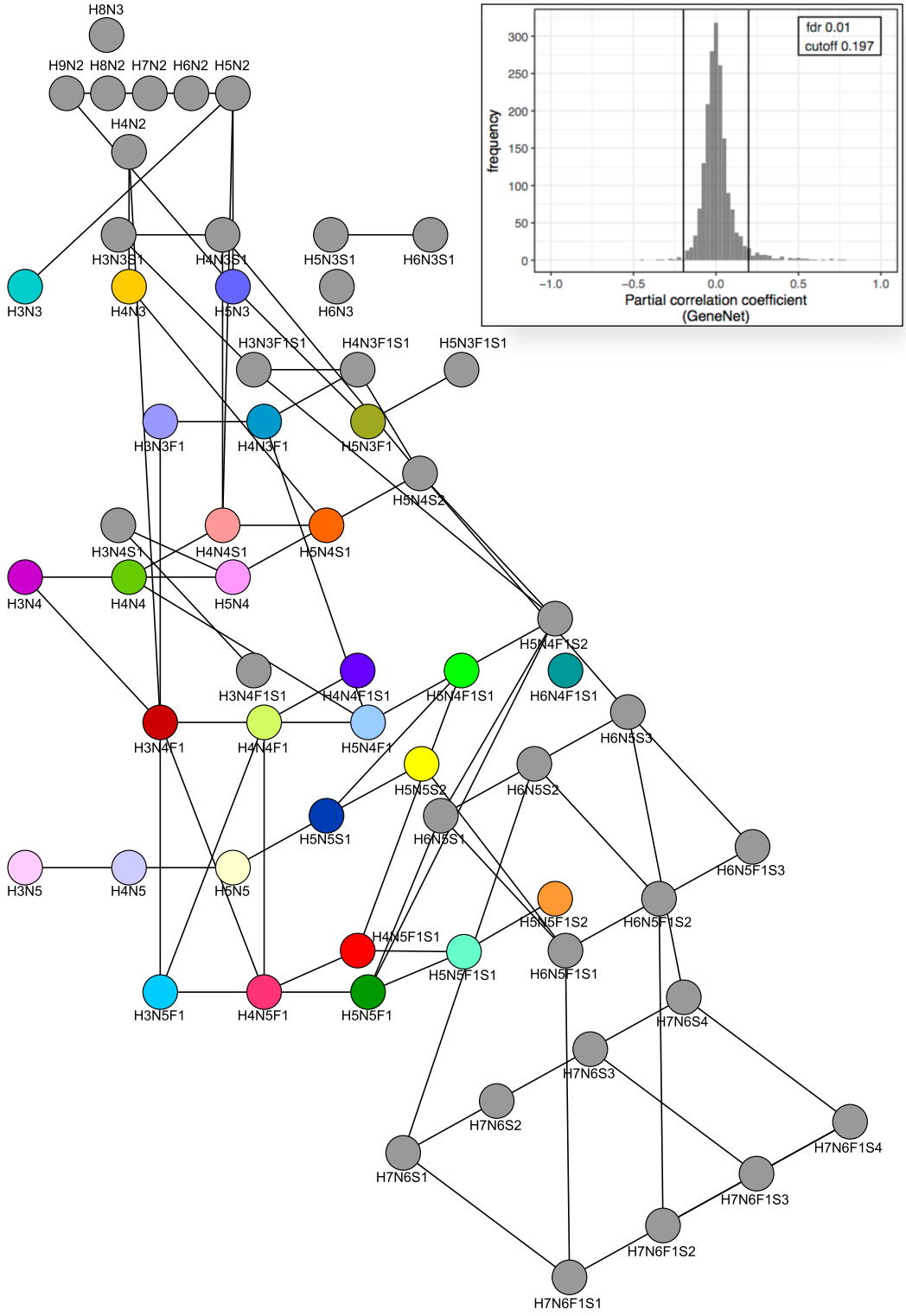


Figure 4.5: Data driven GGM. Nodes represent compositions and edges significant partial correlations estimated with GeneNet. The inset shows the partial correlation coefficients distribution. Black lines indicate the significance cutoff (FDR 0.01).

4.2 Pathway analysis

To investigate the correlation among compositions, we considered a regularized partial correlation (GeneNet) measure, which has previously been shown to identify single enzymatic reactions in Immunoglobulin G glycomics data. Values were adjusted for multiple testing (FDR 0.01 using Benjamini-Hochberg [122]). Only roughly 4% of the coefficients resulted significant (83 out of 1,830), with a strong prevalence of positive correlations (Figure 4.5, side panel). We visualized significant correlations as a network (Figure 4.5), where nodes indicate compositions and edges significant correlations.

To understand what the computed correlations represent in biological terms, we systematically compared them to the compositional data pathway. Briefly, we computed a contingency table by classifying all composition pairs according to the significance of their correlation and the presence of an edge in the compositional pathway. We used the values in this table to perform a Fisher's exact test, where the resulting p-value quantifies the statistical dependency between the correlation network and the compositional pathway: a low p-value indicates lack of independence. In this case, the computed Fisher's p-value was $1.14 \cdot 10^{-53}$, which demonstrated that significant partial correlations do identify single enzymatic reactions in the pathway of glycan synthesis. Note that the ability of partial correlation in detecting biochemical steps in biological network has already been shown for metabolomics [32] and IgG glycomics data (see Chapter 3 [176]), so the result is not completely unexpected. However, a systematic correlation analysis of TPNG data had never been performed before.

4.3 Structural inference

We have shown that the data-driven partial correlation network was able to selectively identify biochemical relations among the measured variables. However, since the variables in our data were compositions, we could not make any a priori statement on which glycan structures were contributing to the observed correlations. To infer this information, we exploited the strong relationship between GGM and compositional data pathway.

The approach was as follows: once the data driven GGM was computed from the data (Figure 4.6A), we intersected it with the compositional data pathway (Figure 4.6B). This was done to exclusively select the edges representing single enzymatic reactions of glycan

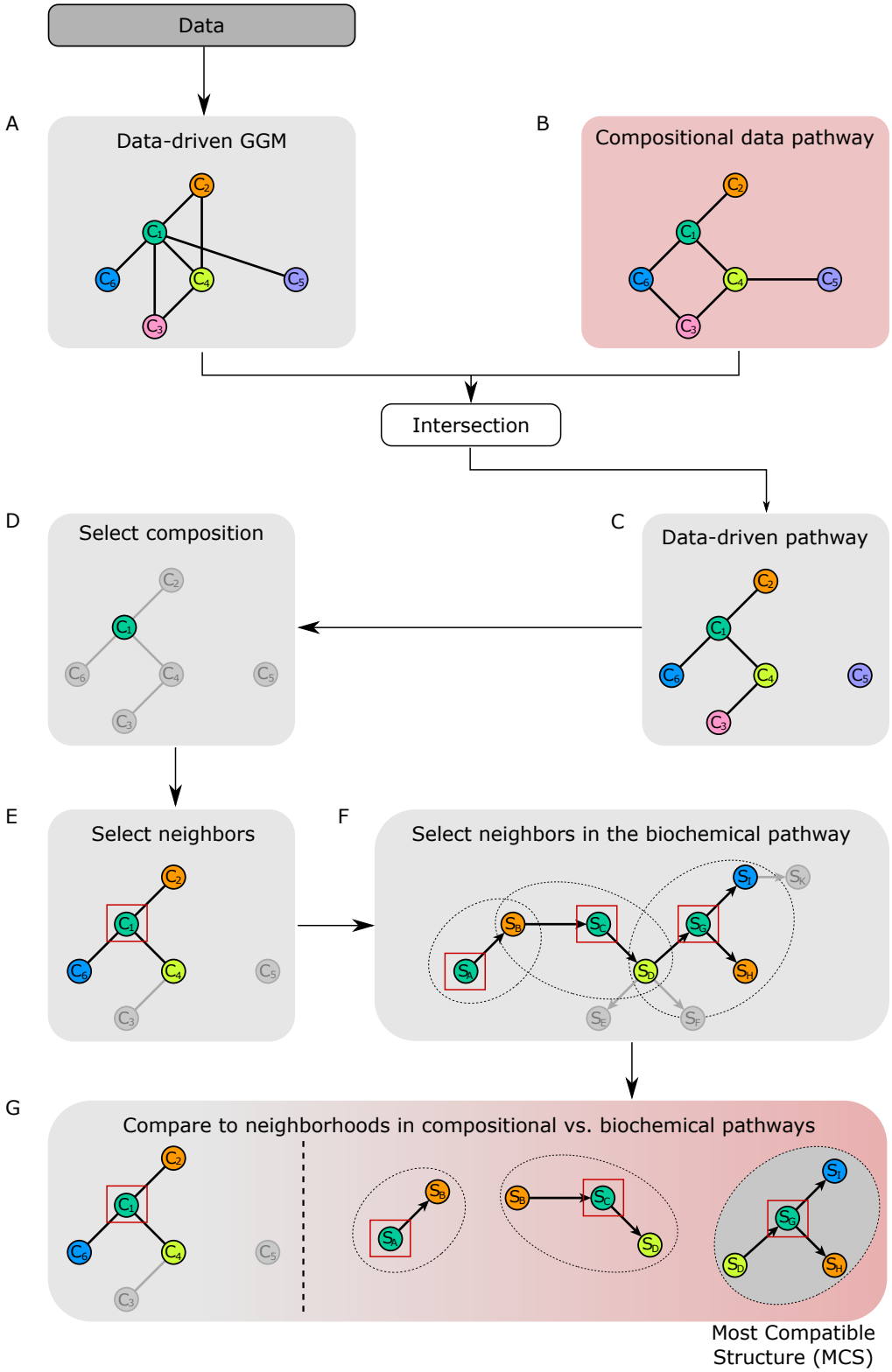


Figure 4.6: Most Compatible Structure (MCS) identification approach.

synthesis that also showed a significant partial correlation among the corresponding compositions, giving rise to a data-driven pathway (Figure 4.6C). At this point, for any given composition C (Figure 4.6D), we select the nearest neighbors (Figure 4.6E) and mapped them back to the theoretical biochemical pathway, where a single composition could correspond to several structures (Figure 4.6F). Finally, we selected the structure with composition C that maintained the most neighbors to be the Most Compatible Structure (MCS, Figure 4.6G). Notice that in case of structures with the same number of nearest neighbors, a clear decision could not be made, and therefore the structures were identified as equally contributing. The procedure was repeated for all compositions in the dataset.

In our dataset, 28 of the 61 compositions were made of more than one glycan structure (Figure 4.7). Out of these, for 14 of them our approach could identify a single MCS, for 2 we could at least exclude one structure and for 12 we could not make any prediction.

In order to validate our findings, we compared our results to glycomics data obtained from the TPNG analyzed with different platforms [106, 177]. Among those compositions for which we could identify a single MCS, 11 out of 14 were correctly inferred, two were incorrectly predicted and for one we could not find validation data. Moreover, among those compositions for which no single structure could be identified with our algorithm, but for which at least one of the possible structures could be excluded (H5N4 and H5N4F1), the most abundant structure was in the set of structures selected by our approach.

Overall, our approach demonstrated to be a cheap and effective tool to predict glycan structures from compositions.

4.4 Conclusion

The overall glycosylation state of an organism is a powerful biomarker to identify physiological changes or diseases [10, 51, 102, 178–180]. The MALDI-TOF-MS platform allows to measure N-glycans released from all human plasma proteins in large-scale cohorts. The resulting spectrographic peaks represent different molecular masses (or compositions), which, however, could be associated to multiple glycan structures. Since the structure of a glycan is often essential to modulate the function of the corresponding glycoprotein [181], we addressed the problem of inferring the glycan structures contributing most to the identified masses in the human total plasma N-glycome from a large cohort.

We first introduced a unified theoretical biochemical pathway, which we based on to the

Figure 4.7: MCS analysis results. In the figure, all compositions made of more than one structure are shown. MCS as inferred with our approach are shown with a gray background. Structures found in previously published studies on the TPNG [106, 177] are indicated with a check mark.

available literature on glycan synthesis. We then adapted this pathway to match the compositions measured in the available dataset. We showed that significant partial correlation among measured compositions represent in large part single enzymatic steps in the pathway. This relationship between partial correlation and glycosylation pathways was already observed for IgG Fc glycans (see Chapter 3 [176]); however, this was a surprising result, as the glycomics data are not extracted from a single isolated protein, but from a mixture of proteins with vastly different functions. This result therefore indicates that a substantial part of the glycosylation pathway is shared among different proteins. Protein-specific synthesis steps are still likely to occur [13], but protein-specific data would be needed to investigate them in detail.

We used the intersection between significant partial correlations and biochemical pathway to predict which structure within a composition, referred to as the Most Compatible Structure (MCS), was the one contributing most to the observed correlation matrix. We then validated our prediction with previously published data [106, 177]. Out of the 14 compositions for which we could identify a single MCS, 11 were predicted correctly, 2 incorrectly, and for one prediction we did not have any validation data. In two additional cases we could correctly exclude one of the structures within the composition.

For 12 compositions we could not make any prediction: out of these, two (H3N3 and H6N4F1S1) did not show any significant correlation in the GGM, and were therefore not suitable to be analyzed with our approach; 8 out of the remaining 10 were compositions including bisected biantennary and triantennary glycans. Since tetraantennary structures are isolated from the rest of the theoretical pathway due to unmeasured nodes (see Figure 2.4), bisected biantennary and triantennary glycans have the same neighbors in the theoretical pathway, namely biantennary non bisected structures, and were therefore not distinguishable by our approach.

This problem could be reduced once measurement platforms will be able to more accurately quantify low abundant glycans, e.g., incompletely glycosylated tetraantennary structures. However, it is also possible that those structures do not appear at all on plasma proteins: recent studies demonstrated that incomplete glycosylation on plasma proteins could trigger degradation [182], and hence proteins carrying such structures would be more likely removed from circulation. Should this be the case, the problem could be addressed by modifying the theoretical pathway to account for unmeasured structures. One way to accomplish this would be to also connect structures that are only separated by unmeasured nodes. This approach would however require careful analysis, as multiple paths could be possible, leading to multiple pathway models. However, the same pathway analysis

employed in this thesis could be used to test these different models and select the one with the highest overlap to the data correlation matrix (i.e., with the lowest Fisher's test p-value).

One intrinsic limitation of the approach presented here is that it needs a highly correlated set of variables to be able to predict the MCS. In a biological system with only few significant partial correlation coefficients, it would be hard to discriminate between different structures according to their neighbors in the data-driven pathway. Furthermore, it is worth noticing that the MCS our methodology identifies is technically the structure within a composition that most contributes to the structure of the partial correlation. For the considered dataset, we proved that in most cases this corresponds to the most abundant structure in that composition, but this could be not true in other cases: should the more abundant structures participate in substantially fewer pathway reactions than the less abundant molecules, it would be difficult for our approach to select the former over the latter. This drawback could be reduced by considering the inference results on neighboring nodes when predicting the MCS on a new composition: by allowing a multivariate inference approach, prediction accuracy should improve.

In conclusion, in this chapter, we presented an efficient data-driven approach for the inference of glycan structures from mass spectrometry large-scale datasets. The results of our analysis could be of concrete help in reducing the costs of further fragmentation experiments, only selecting those compositions for which our approach was not able to make a clear prediction.

Chapter 5

Systematic evaluation of normalization methods for glycomics data based on performance of network inference

Similar to all other omics data types, glycomics samples need to be preprocessed prior to statistical analysis in order to minimize intrinsic, non-biological variation. This variation can arise, for example, from alterations in the experimental setup, temperature, or instrument conditions. The process that aims at reducing technical variations from the data is referred to in this thesis as *normalization*. Different normalization procedures may have substantially different assumptions regarding the nature of the non-biological signal, which in most practical cases is unknown. Systematic comparisons of commonly implemented preprocessing strategies have been performed in recent years for many omics data, e.g. transcriptomics [107], proteomics [108], as well as metabolomics [109–111], but an analogous study for glycomics data is, to the best of our knowledge, currently unavailable.

This need for a glycomics-specific evaluation is further supported by the observation that the de facto standard for large-scale glycomics data preprocessing is the Total Area (TA) normalization [105], which describes each glycan intensity in a sample as a percentage of the total. Following this transformation, the normalized intensities of a sample sum up to one (or 100%) by definition, therefore implying the loss of one degree of freedom. The division of each value by the sum of all values in a sample is referred to as a closure operation, and the resulting dataset is known as a *compositional dataset* [137]. Notably,

these type of data alter the structure of the covariance matrix, subsequently affecting any downstream correlation-based analysis (for details on this phenomenon, see Section 2.1.3).

Compositional datasets are not unique to glycomics, but are widely used in other fields, in particular in microbiome analysis [183], where percentages are used to describe the relative abundance of different microbial species. However, regular multivariate methods are not appropriate to treat these types of data, and specific statistical techniques need to be employed [184–188]. Most of these methods require for the analysis the establishment of new variables, typically defined as ratios between the original compositional values [189–191]. This makes interpretation of the results in terms of the original quantities challenging, if not impossible [192, 193].

In order to be able to infer specific molecular interactions from the analysis of large-scale glycomics data, the selection of a more suitable alternative to TA normalization is therefore necessary. Given the variety of possible preprocessing strategies available, we need to define an evaluation criterion to quantitatively assess the quality of different normalization methods.

Common evaluation schemes for the performance of preprocessing strategies are mostly based on two approaches: 1. Minimizing the variation between technical replicates [194, 195]; 2. Maximizing the variation across groups [111, 196]. Consistency across technical replicates is a desirable outcome, but alone is not sufficient to guarantee good data quality, and technical replicates might not always be available. The maximization of phenotypic associations, on the other hand, is based on the assumption that the measured variables associate strongly to an arbitrarily chosen phenotype, which might or might not be the case for specific data. This criterion does therefore not necessarily reflect the true underlying biology.

In this chapter, we address the question of evaluating different normalization strategies for glycomics data with an unconventional and innovative approach. Specifically, we assess the quality of a normalized dataset through its ability to reconstruct a biochemically correct pathway using statistical network inference. The idea is based on the observation that Gaussian Graphical Models are able to selectively identify single enzymatic steps in metabolic pathways [32]. Here, we compare the GGMs inferred from data normalized with different approaches to the known biochemical pathway of glycan synthesis and we evaluate the quality of each normalization according to how well the corresponding GGM retrieves known synthesis reactions (Figure 5.1). By computing the quantitative overlap between estimated GGM and glycosylation pathway, we rely on a *biological* measure of quality, as a higher overlap indicates data whose correlations are able to better reflect known

biochemical interactions. Hence, the normalization that produces the highest overlap is defined as the best. Glycomics data provide an ideal test case to demonstrate the validity of this approach, as the known biochemical pathway of synthesis is well characterized.

In the following, we compared the performance of different variations of seven commonly implemented normalization methods applied to glycan data from six cohorts and across three different glycomics platforms, including measurements of IgG Fc, total IgG or total plasma N-glycans.

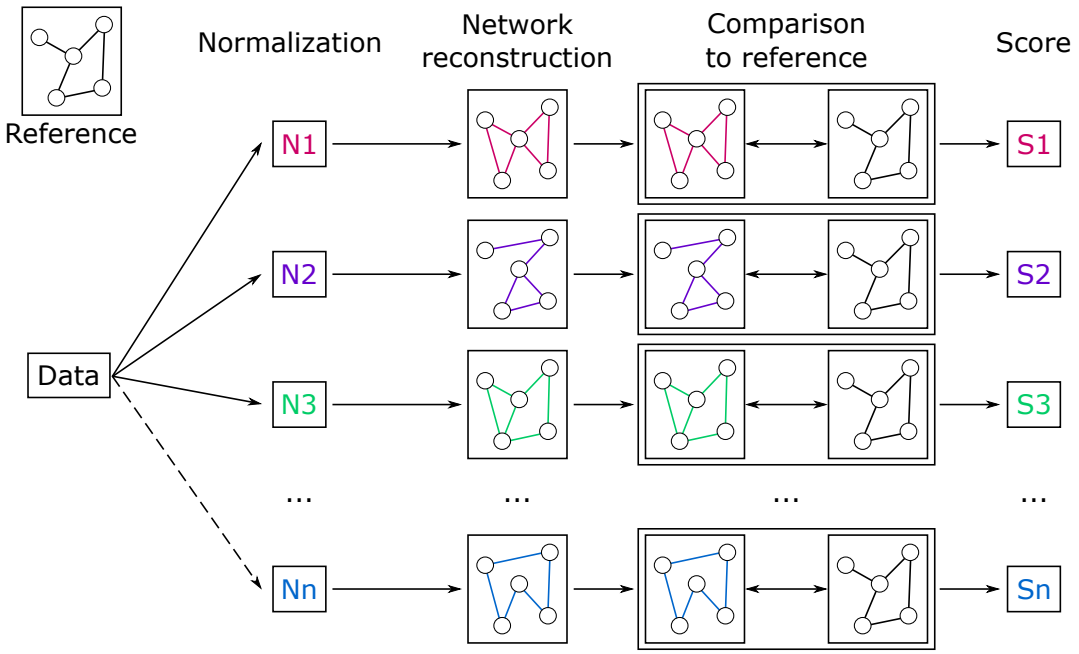


Figure 5.1: Pipeline for the evaluation of different normalization methods for glycomics data. First, data are normalized with various approaches. From each processed dataset, a GGM is inferred and compared to the available prior knowledge on the biochemical pathway of glycan synthesis. The result of this comparison is a quantitative overlap that describes how well the estimated GGM represents known synthesis reactions. This overlap is the used for the evaluation of the normalization approach, where higher overlap corresponds to a better normalization.

5.1 Considered normalizations

For all six cohorts presented in Section 2.1.1 (Table 2.1), seven basic preprocessing approaches were considered (Table 5.1). All methods are commonly used in omics data analysis: (1) Raw, i.e., unprocessed, data were included for comparison; (2) Quantile [197] and (3) Rank [139] normalization are widely used in microarray data analysis; (4) Total

Table 5.1: Considered Normalizations. Where possible, two variants were additionally considered: a log-transformation subsequent to normalization and, for LC-ESI-MS measurements, normalization per IgG subclass instead of total IgG.

	Normalization	Log	Subclass
1	Raw	✓	—
2	Quantile per glycan [197]	✓	✓
3	Rank per glycan [139]	✓	✓
4	Total Area (TA) [198]	✓	✓
5	Median Centering [199]	—	—
6	Probabilistic Quotient [200]	✓	✓
7	TAProbabilistic Quotient [135]	✓	✓

Area (TA) is often used to treat large-scale glycomics [198] and microbiome data [183]; (5) Median centering [199], (6) Probabilistic Quotient applied to raw and (7) to TA normalized data are popular methods for the preprocessing of metabolomics data [135, 200].

Since omics data have been observed to often follow a log-normal distribution [136, 143], and since GGMs assume normally distributed data, log-transformation on normalized data was also included in the analysis when applicable (indicated with a check mark in the second column of Table 5.1), resulting in 13 different preprocessing strategies. For LC-ESI-MS data, 10 additional variations were included, as in this case data normalization can be performed over the full dataset or per each IgG subclass separately (third column in Table 5.1). A detailed description of each normalization procedure can be found in Section 2.1.3.

5.2 Prior knowledge based evaluation

Once all normalizations were applied to the data, partial correlation coefficients were computed with the GeneNet algorithm [144], which has proven to give more reliable and stable estimates of partial correlation coefficients than the analytical solution (see Section 6.1). Statistical significance of coefficients was determined by applying a False Discovery Rate (FDR) of 0.01. The resulting partial correlation network, or Gaussian Graphical Model (GGM), was then compared to the biochemical pathway of glycan synthesis. As quantitative measure of overlap between the calculated GGM and the pathway, we chose the Fisher's p-value (see Methods), where lower p-values correspond to a higher overlap, and, in the context of this thesis, to a better normalization. Schematics of the pathways used for the evaluation can be found in Figures 2.2 and 2.4.

5.2.1 LC-ESI-MS

For the LC-ESI-MS platform, most methodologies performed well (Figure 5.2, left). Interestingly, the unprocessed data (Raw) were among the best-performing methods, which indicates that for this platform data transformation is not essential for correlation analysis.

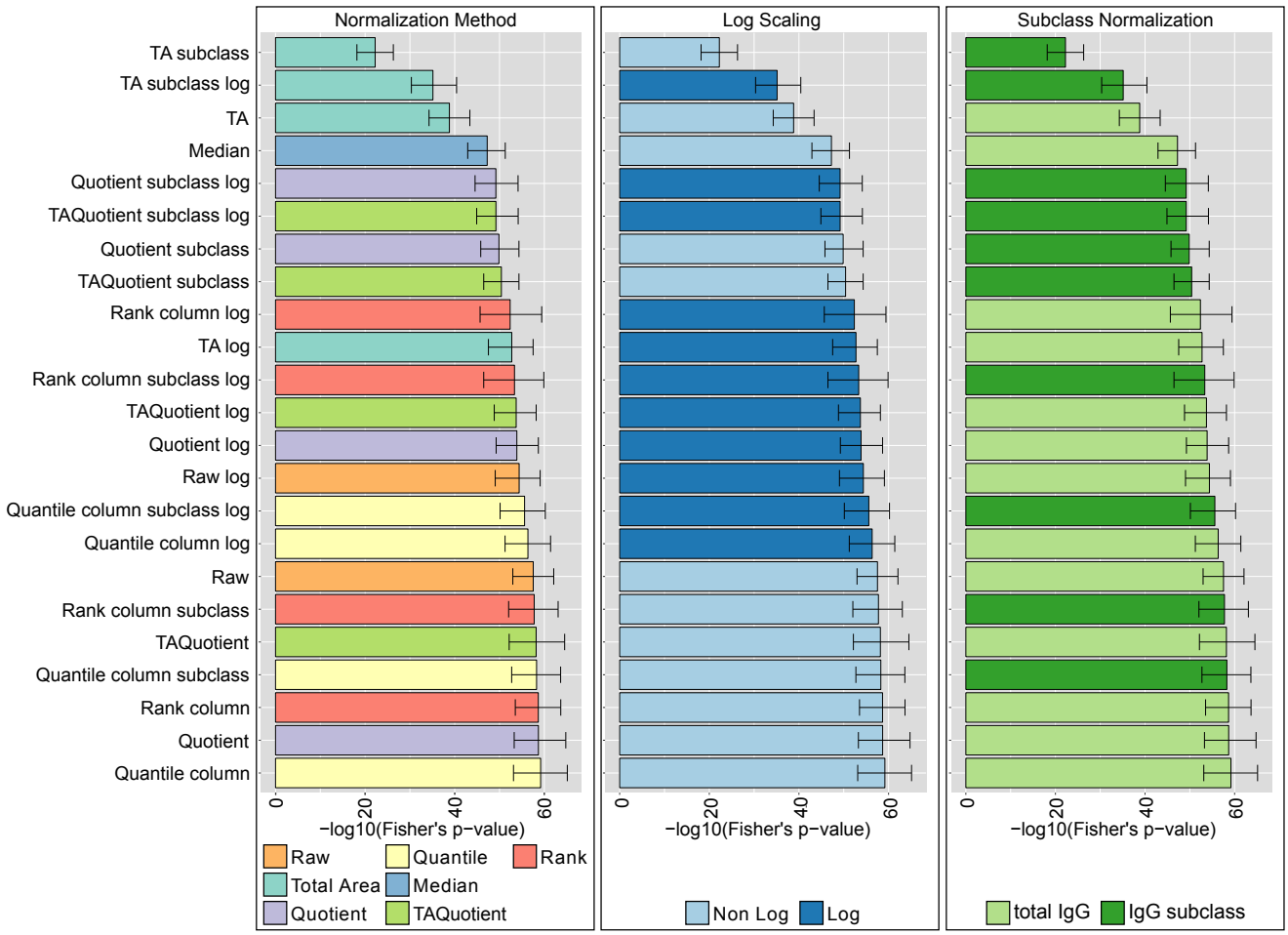


Figure 5.2: Normalization analysis results for the Korčula 2013 cohort. Results in the panels are colored according to type of normalization (left), log-transformation (center), or normalization per IgG subclass or total IgG (right). Bars represent the median of the Fisher's exact test p-values over 1,000 bootstrapping, and error bars the corresponding 95% confidence intervals.

As expected, TA-based normalizations performed significantly worse than all other considered preprocessing. Given the assumption of normality of the considered correlation measure, we expected Log-transformed data to perform better than their non-transformed counterparts. However, we observed that Log-transformation seemed to even worsen performance, although not significantly (Figure 5.2, center).

An exception to this rule is the TA-log normalization, for which the logarithm appears to

neutralize the constraints imposed by TA. Normalizing per total IgG or per IgG subclass did not result in substantial differences in performance (Figure 5.2, right).

The results of the evaluation were consistent across all cohorts (Figures 5.3-5.5).

In conclusion, we showed that for LC-ESI-MS IgG Fc glycomics data, all considered preprocessing performed comparably except TA, which was significantly worse than the rest. Surprisingly, non log-transformed data did not perform worse than the transformed data, and normalizing per total IgG or per IgG subclass did not make a significant difference.

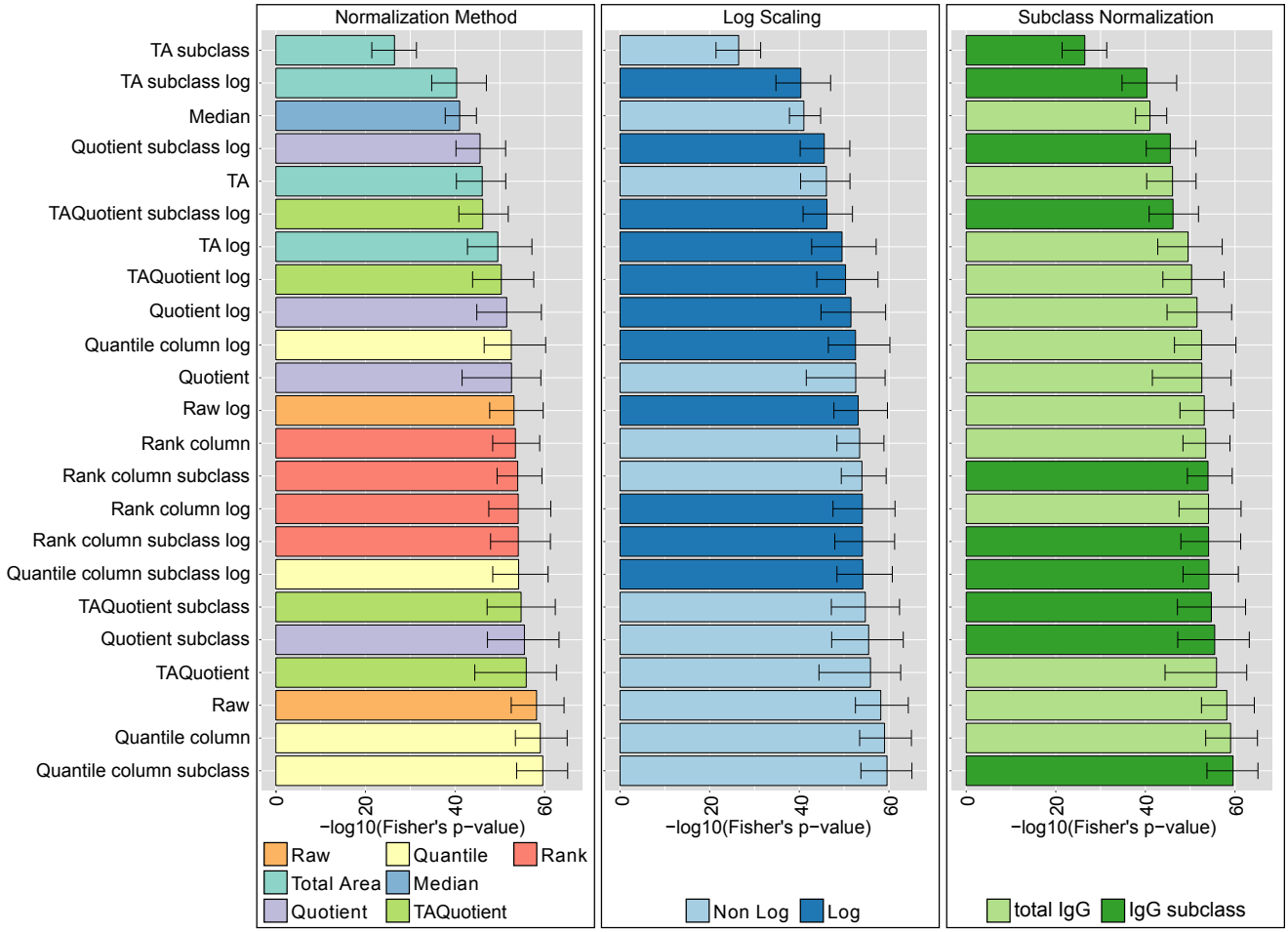


Figure 5.3: Normalization analysis results for the Korčula 2010 cohort. Results in the panels are colored according to type of normalization (left), log-transformation (center), or normalization per IgG subclass or total IgG (right). Bars represent the median of the Fisher's exact test p-values over 1,000 bootstrapping, and error bars the corresponding 95% confidence intervals.

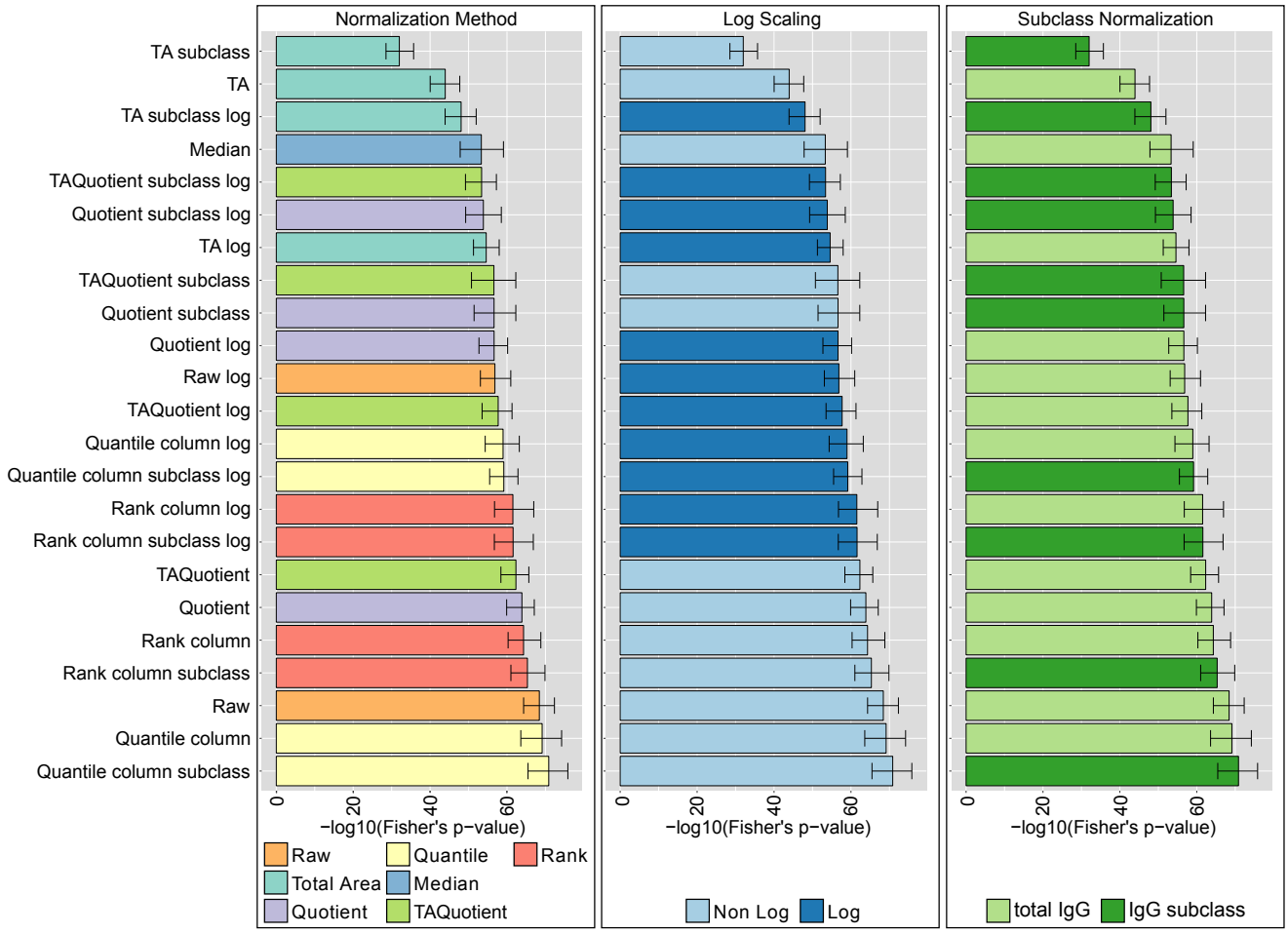


Figure 5.4: Normalization analysis results for the Split cohort. Results in the panels are colored according to type of normalization (left), log-transformation (center), or normalization per IgG subclass or total IgG (right). Bars represent the median of the Fisher's exact test p-values over 1,000 bootstrapping, and error bars the corresponding 95% confidence intervals.

5.2.2 UPLC

In the available IgG UPLC dataset, only 24 glycan peaks were measured, versus the 50 structures of the LC-ESI-MS data. Just because of this intrinsic difference, the overall p-values of the analysis were expected to be higher in this case. However, this does not affect our analysis, as our approach evaluates normalizations according to their ranking and not absolute performance. Contrary to the previous case, here the performance was highly affected by the chosen normalization method (Figure 5.6, left), with TA Probabilistic Quotient and simple Probabilistic Quotient being the top ranking methods. In this case, the unprocessed data did not perform well at all.

Moreover, in contrast to what was observed in the LC-ESI-MS case, for UPLC data, the log-transformation had a significant impact on the performance of normalizations, although with opposite effects depending on the methodology: for some it substantially enhanced performance (Quantile, Total Area), while for others it was detrimental (Rank, Raw data) (Figure 5.6, right).

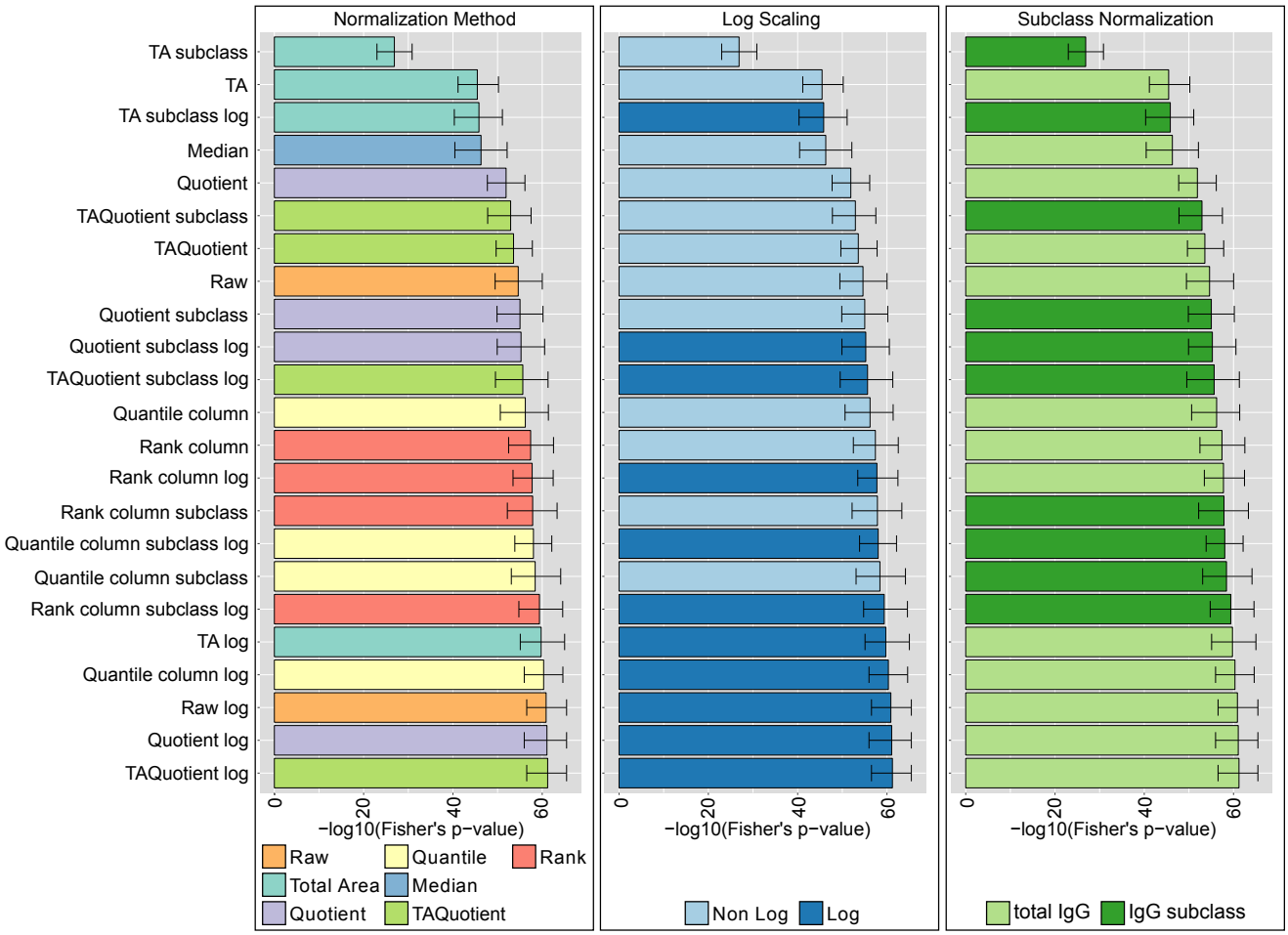


Figure 5.5: Normalization analysis results for the Vis cohort. Results in the panels are colored according to type of normalization (left), log-transformation (center), or normalization per IgG subclass or total IgG (right). Bars represent the median of the Fisher's exact test p-values over 1,000 bootstrapping, and error bars the corresponding 95% confidence intervals.

In conclusion, for UPLC data, log-transformation did significantly affect the normalization performance, while Probabilistic Quotient-based normalization were the overall best performing method.

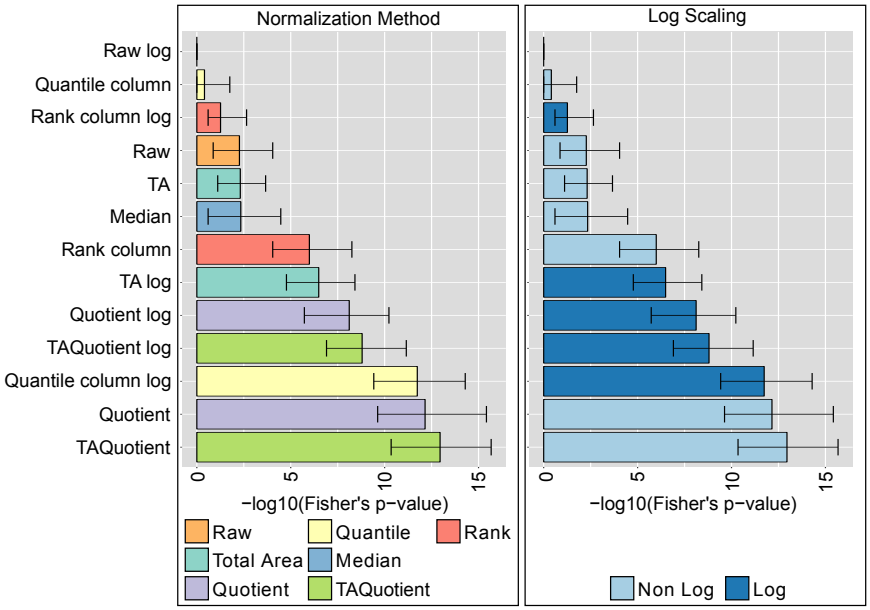


Figure 5.6: Normalization analysis results for the CRC cohort. Results in the panels are colored according to type of normalization (left), or log-transformation (right). Bars represent the median of the Fisher's exact test p-values over 1,000 bootstrapping, and error bars the corresponding 95% confidence intervals.

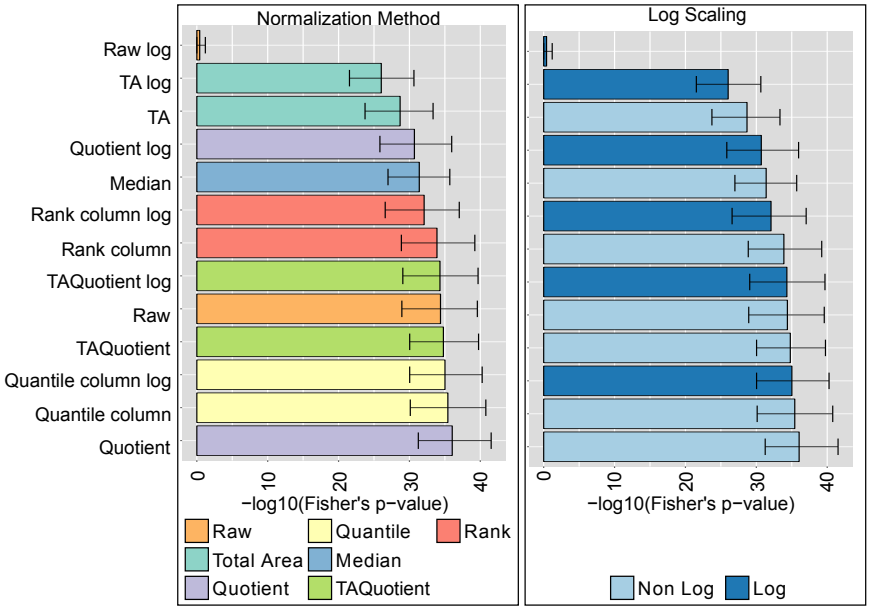


Figure 5.7: Normalization analysis results for the LLS cohort. Results in the panels are colored according to type of normalization (left), or log-transformation (right). Bars represent the median of the Fisher's exact test p-values over 1,000 bootstrapping, and error bars the corresponding 95% confidence intervals.

5.2.3 MALDI-TOF-MS

The MALDI dataset included 61 glycan variables, so we expected Fisher's p-values of the same order as the LC-ESI-MS for the best performing methods. Similar to the LC-ESI-MS case, we found a multitude of methods that performed well (Figure 5.7, left). Log-transformed unprocessed data delivered the worse performance, followed by TA normalization variations. Except for these cases, Log-transformation did not significantly affect the normalization performance (Figure 5.7, right).

In conclusion, for MALDI data several normalizations were found appropriate. Log-transformation did not significantly alter performance, except when considering unprocessed data.

5.3 Conclusion

Little attention is often paid to data preprocessing, although it can highly affect the reliability and reproducibility of any downstream data analysis. Several systematic evaluations of preprocessing methodologies have been recently published for different omics data types, but glycomics has so far been ignored in this regard.

In order to address this problem, we developed an innovative approach to assess the quality of different normalization strategies applied to glycomics data. The main feature of our procedure lies in the definition of a biological measure of quality, i.e., we quantify how well significant correlations in the data normalized with a given technique represent known biochemical reactions in the pathway of glycan synthesis. Our quantitative measure of choice for this evaluation was the Fisher's test p-value, which allows for an intuitive interpretation of overlap between correlations and biochemical pathway.

In this chapter, we performed a systematic analysis of 23 preprocessing strategies applied to six large-scale cohorts across three platforms, with measurements ranging from single protein and single glycosylation site (LC-ESI-MS), to total plasma N-glycome (MALDI-TOF-MS). The observed normalization ranking was highly compatible across platforms: overall, the Probabilistic Quotient resulted the most reliable method, as all variations of this procedure ranked consistently in the top performers in all cohorts and across platforms. Log-transformation and normalization per IgG subclass or per total IgG did not seem to significantly affect the ability of this method to correctly retrieve the glycan synthesis pathway. Interestingly, while Total Area normalization did not rank high in comparison to other methods (as expected), the log-transformed Total Area preprocessing was a well-performing method, and TA Probabilistic Quotient was among the best per-

forming approaches overall, suggesting that additional transformations on TA normalized data can neutralize the constraints imposed by on the data correlation structure.

One interesting finding of our investigation was the substantial difference of the evaluation results between MS- and UPLC-based platforms: while for the former most normalization approaches performed comparably, for the latter the variance among the considered strategies was considerable. The origin of this discrepancy is not easy to track, but it could be due to the fact that UPLC does not separate glycans according to their mass, like MS-based techniques do, but according to their chemical and physical properties, which determine different retention times. This leads to most chromatographic peaks to represent a mixture of glycan structures. Although it has been shown that there is a predominant structure in the vast majority of IgG chromatographic peaks24, this contamination could make the data correlation structure noisier and thus more sensitive to different normalizations.

It is important to note that the results obtained in this chapter do not necessarily hold for types of analyses other than correlation studies. An analogous systematic evaluation would be therefore needed for other statistical questions; however, while in this case we could define a well-characterized biological measure of quality, such a measure would be harder to define for other types of analysis. Moreover, while the results presented here seem to suggest that log-transformation is not necessary, it should be considered that data normality is an assumption of several other statistical tests and approaches, and thus we still recommend, in general, to log-transform the data after normalization. The results obtained in this analysis were highly consistent across the considered platforms. Nevertheless, investigating the performance of these normalization methods on data obtained from other common measurement platforms, e.g., Multiplex Capillary Gel Electrophoresis with Laser-induced Fluorescence (xCGE-LIF) [90], would be valuable to improve the generalizability of our findings. The same approach described here could moreover be employed to evaluate other preprocessing steps, for example batch correction, which aims at reducing the variance due to samples being measured at different times, or missing values imputation.

In conclusion, we recommend to normalize glycan data with the Probabilistic Quotient normalization followed by log-transformation. This technique has demonstrated to be robust and reliable regardless of the measurement platform.

Chapter 6

Using prior knowledge to optimize correlation network cutoffs

Network inference, i.e., the reconstruction of biological networks from high-throughput data, has become a booming field in systems biology [201–203] . Interactions among biomolecules extracted from the analysis of large datasets have been shown to represent known and predict novel biological mechanisms [204,205], in particular enzymatic reactions in molecular pathways [32, 116].

Virtually all network inference methodologies require the definition of a parameter that determines which molecular interactions should be included in the network and which should be discarded. The construction of correlation-based networks commonly requires a series of simple steps (Figure 6.1A). First, pairwise correlations between variables are estimated from the data, for which a wide variety of methods is available. The next step is to determine which correlation coefficients are statistically different from zero using a hypothesis test, which produces p-values associated with each correlation coefficient. These p-values are then compared to a given significance level α, typically 0.01 or 0.05, with appropriate multiple hypothesis testing correction. Finally, significant correlations can be visualized and further analyzed as a network, where nodes represent the variables in the dataset and edges represent significant correlations.

However, this straightforward network inference pipeline has two major pitfalls that are usually overlooked and substantially affect the robustness and reproducibility of correlation-based network inference. First, for most correlation measures, the resulting network will vary substantially depending on the number of observations available in the dataset. In general, the bigger the sample size, the lower the p-values. This means that with increas-

ing sample size, weaker correlations become significant and the corresponding network becomes denser (Figure 6.1B). Second, different multiple testing methods (e.g., Bonferroni [146] or Benjamini-Hochberg [122]) have different underlying assumptions, such as controlling for the family-wise error rate (FWER) versus the false discovery rate (FDR), respectively. However, in practice, the choice of one method over another is usually not scrutinized adequately. Thus, depending on the arbitrary choice for error correction and significance level, one may obtain vastly different networks (Figure 6.1B) which are all statistically sound, but that do not necessarily represent relevant underlying biological mechanisms.

In this chapter, we address the problem of correlation-based network inference from a different perspective. Instead of a statistically-driven cutoff selection, we propose to choose the correlation cutoff that produces the correlation network with the highest overlap to

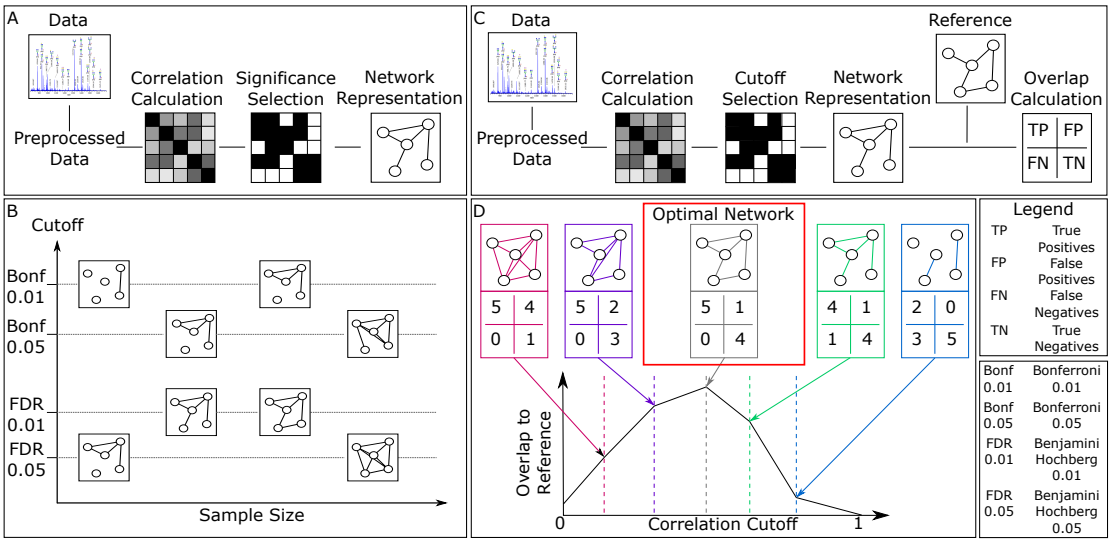


Figure 6.1: Pipeline of network inference and workflow of the new approach. **A** Typical pipeline of correlation network inference. A correlation matrix is estimated from the preprocessed data. A significance selection step identifies correlations that are statistically different from zero. Significant correlations are commonly visualized as a network. **B** Schematic representation of the dependence of the correlation network on sample size and statistical cutoff. Note that, despite looking substantially different, all resulting networks can be considered statistically correct. **C** Prior knowledge-based network overlap estimation. The correlation network is compared to a prior knowledge network, where the overlap is quantified using true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Based on these values, a quality overlap measure between data-driven correlation matrix and biological reference is computed. **D** Prior knowledge-based network inference approach. We discard the p-value-based significance selection, and instead analyze how the overlap between correlation network and biological reference varies depending on the correlation cutoff. We then define optimal the correlation cutoff at the point where the overlap is maximal.

a given ground truth (Figure 6.1C and D), hereafter referred to as 'biological reference'. That is, we search for the network that shows the highest overlap with the known underlying biology, thereby avoiding the above-mentioned arbitrarily determined cutoffs for p-values. We postulate that even a coarse, incomplete, or partly incorrect biological reference is suitable for this approach, as long as a sufficient amount of correct biological knowledge is covered. In many cases, the molecular networks regulating the system under study are not fully known, which results in an only partial biological reference being available. For example, often only few of the synthesis pathways of the system under study are well characterized, and for some systems, detailed biochemical information is not available at all. In these cases, we argue that one can still use the available prior knowledge as a biological reference and obtain a cutoff that is close to the global optimum. We first show that statistical significance selection is indeed substantially influenced by the dataset size. We then apply the prior-knowledge based cutoff optimization approach to plasma IgG glycomics measurements. In this particular case, we have a well-characterized, supposedly complete biochemical synthesis pathway, which we use as gold standard biological reference to test our optimization approach. We show that the optimal correlation cutoff is unique and sample size independent. Moreover, even when the optimization procedure is performed with only a fraction of the original biological reference, the resulting optimum remains the same. Finally, we apply the algorithm to urine metabolomics and TCGA [142] RNA-sequencing data, where our method is able to identify an optimal network and to outperform regular statistical cutoffs.

All results reported in this chapter are part of the following publication:

- **Benedetti, E.**, Pučić-Baković, M., Keser, T., Gerstner, N., Büyüközkan, M., Tamara, P., Selman, M.H.J., Rudan, I., Polašek, O., Hayward, C., Al-Amin, H., Suhre, K., Kastenmüller, G., Lauc, G., Krumsiek, J., Using prior knowledge to optimize correlation network cutoffs. *Submitted.*

6.1 Statistical correlation cutoffs depend on sample size

For most correlation measures, the larger the sample size, the lower the resulting correlation cutoff at a given significance level. In other words, increasing the number of subjects measured in a study automatically results in a denser correlation network. To quantitatively investigate this effect, we analyzed IgG glycomics measurements from four large Croatian cohorts (see Subsection 2.1.1). In the following, the results for one of the four cohorts (Korčula 2013) are shown, while the other three cohorts were used for replication.

The discovery dataset included 669 samples and 50 glycan structures measured. Data were normalized, log-transformed and corrected for age and gender prior to analysis.

We subsampled the glycomics dataset without replacement to simulate different sample sizes, from 10 to 669 samples. For each subsample, we computed the glycan correlation matrix and applied a 0.01 FDR cutoff using the Benjamini-Hochberg method as an exemplary approach for multiple testing correction. Results would be qualitatively identical with other methods (e.g. Bonferroni) and α levels. We considered two correlation measures commonly used in the field of computational biology: classical pairwise Pearson correlation and partial correlation, which accounts for the presence of confounders. We included two different estimators for partial correlation: Exact partial correlations obtained from the inversion of the covariance matrix (referred to as parcor), and a shrinkage-based regularization approach, which has been shown to give a more stable estimate and still work in datasets with less samples than variables (GeneNet) (see Subsection 2.3.1).

As expected, for both Pearson correlation and parcor, the significance cutoff decreases with increasing sample size and does not converge even for larger sample sizes (Figure 6.2A, red and blue curves, respectively).

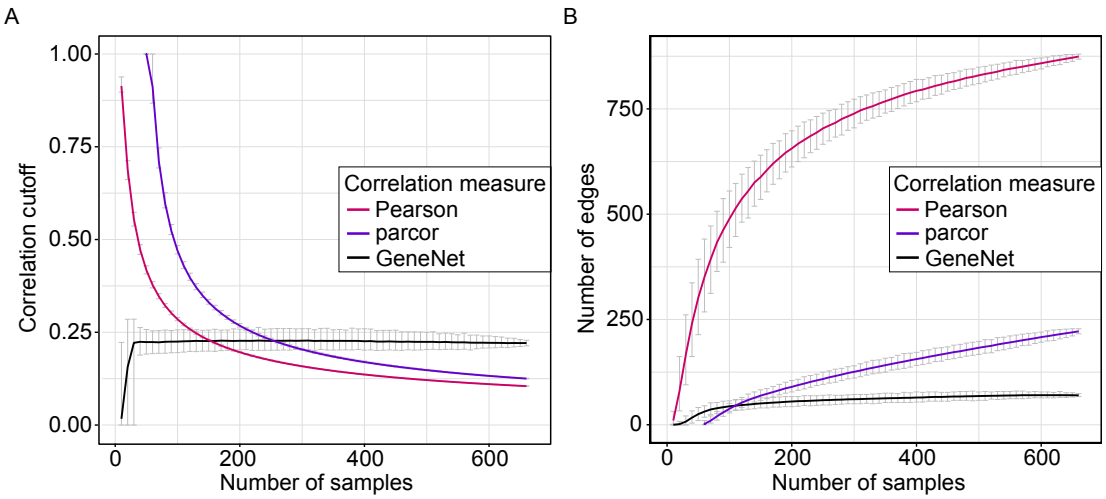


Figure 6.2: Correlation cutoff as a function of the sample size. **A** Correlation cutoff (0.01 FDR) as a function of the dataset sample size for the three correlation measures considered: Pearson correlation (red), exact partial correlation (purple), GeneNet partial correlation (black). Error bars represent 95% confidence intervals from 1,000 bootstrapping samples. **B** Number of edges in the correlation network after applying a 0.01 FDR cutoff as a function of the dataset sample size. Error bars represent 95% confidence intervals of 1,000 bootstrapping samples. Note that for parcor, correlation coefficients can only be estimated for a sample size greater than or equal to the number of variables, in this case 50.

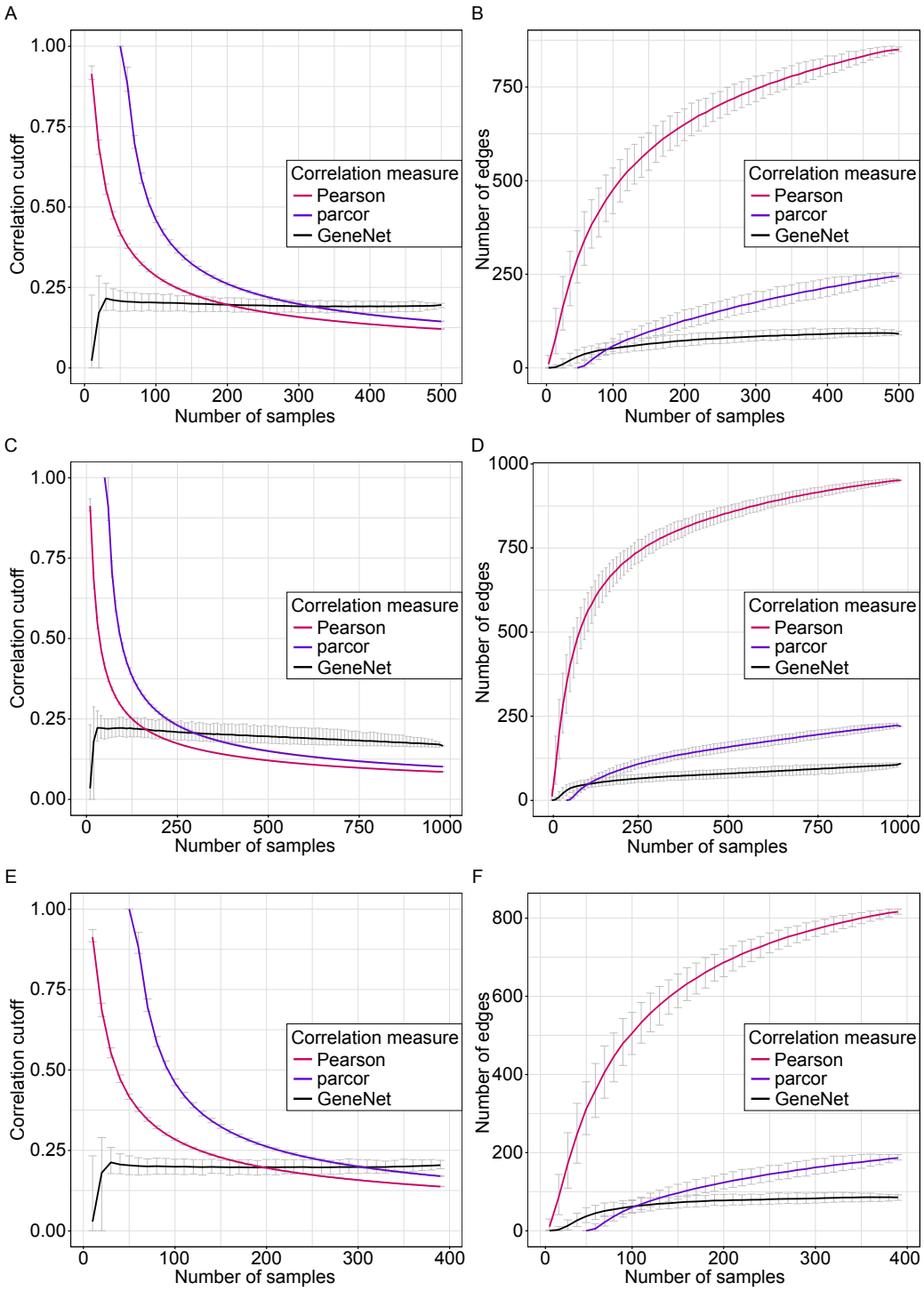


Figure 6.3: Size dependence of statistical cutoffs in the glycomics replication cohorts. **A**, **C**, **E** Correlation cutoff (0.01 FDR, Benjamini-Hochberg) as a function of the dataset sample size for the three correlation measures considered in the Korčula 2010, Split and Vis cohorts, respectively. Error bars represent 95% confidence intervals from 1,000 bootstrapping samples. **B**, **D**, **F** Number of edges in the correlation network after applying a 0.01 FDR cutoff as a function of the dataset

sample size in the Korčula 2010, Split and Vis cohorts, respectively. Error bars represent 95% confidence intervals of 1,000 bootstrapping samples. Note that for parcor, correlation values can only be estimated for a sample size greater or equal to the number of variables, in this case 50.

Interestingly, partial correlations estimated with GeneNet do not show the same behavior, as the statistical correlation cutoff is fairly stable across the considered sample sizes (Figure 6.2A, black line). This is also reflected in the total number of edges in the resulting network: while for Pearson correlation and parcor the number of significant coefficients included in the network systematically increases with the sample size, the network estimated with GeneNet maintains a roughly constant number of edges (Figure 6.2B). As a quantitative example, when considering twice as many samples, from 200 to 400, the GeneNet network remains stable with around 60 edges, while the Pearson correlation network increases by a factor of roughly 1.2 (from 655 to 790) and the parcor network increases by a factor 1.5 (from 95 to 155). Analogous results were obtained in the three replication cohorts (Figure 6.3).

This first analysis showed that indeed there is a strong dependence of network density (number of significant correlation) on sample size of the dataset for both Pearson and partial correlations. GeneNet did not show this behavior, which is most likely an effect of the p-value estimation method used in the algorithm (see Methods), and it gave rise to a considerably more stable network, almost independent of the sample size.

6.2 Reference-based cutoff optimization

We then applied our reference-based network inference approach to IgG glycomics data, for which the pathway of synthesis is well characterized (Figure 6.4A). We have previously shown that edges in a partial correlation network represent single enzymatic reaction in the IgG glycosylation pathway8.

First, we tested how our method compares to regular statistical cutoffs. As a quantitative measure of overlap, we used Fisher's exact test based on the overlap contingency table, which classifies glycan-glycan pairs depending on whether an edge between them appears both in the correlation network and in the biological reference (true positives), only in the correlation network (false positives), only in the biological reference (false negatives) or in neither (true negatives). This p-value will be lower the higher the overlap between correlation network and biological reference is (see Methods). The cutoff that produces the maximum overlap to the biological reference is hereafter referred to as the "optimal cutoff"

and the corresponding network as the "optimal network". Regular Pearson correlation performed poorly in comparison to parcor, while GeneNet was the overall best performing method (Figure 6.4B).

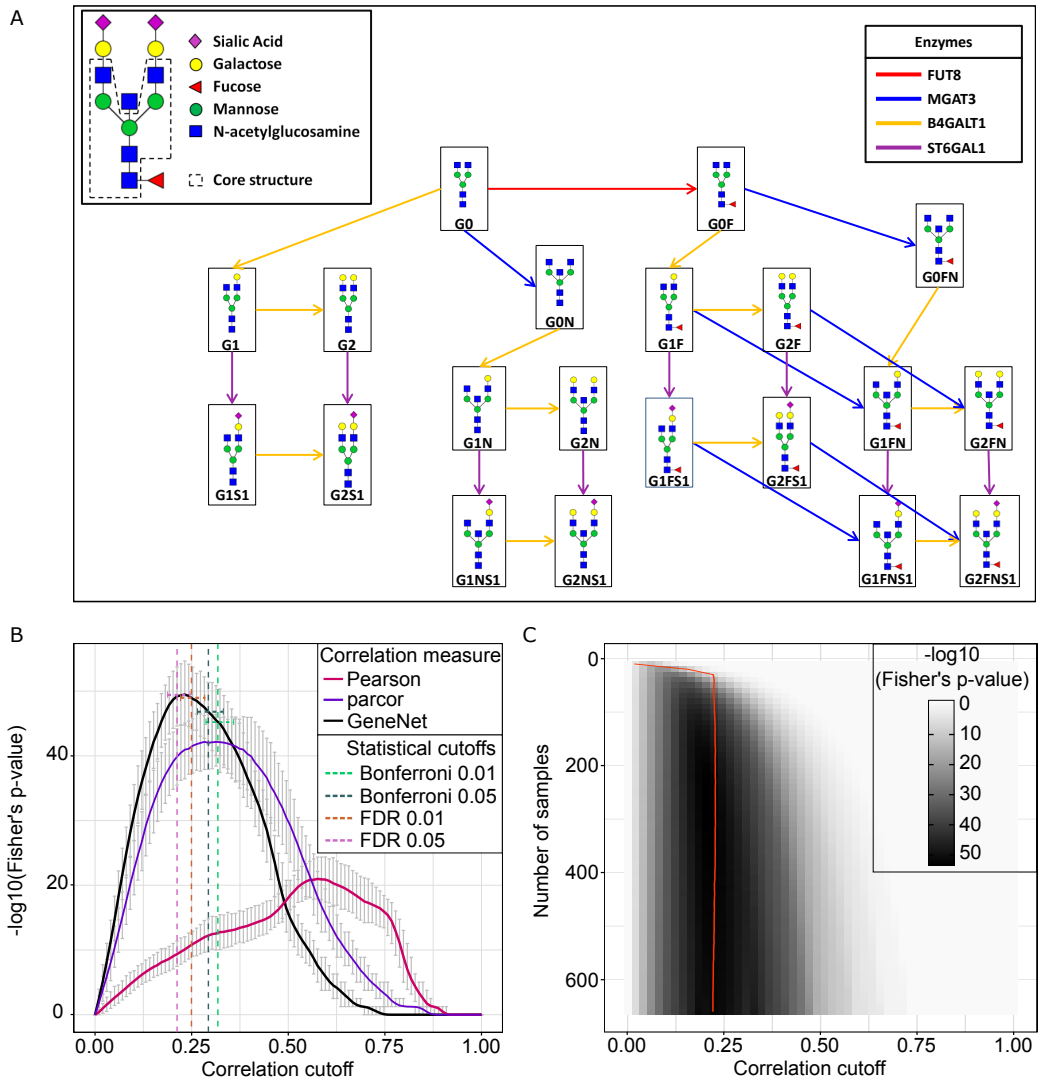


Figure 6.4: Network quality as a function of the correlation cutoff. **A** IgG glycan structures and synthesis pathway. The figure was adapted from Benedetti et al. (2017) and represents the IgG glycosylation pathway, with nodes representing glycan structures and arrows representing single enzymatic reactions in the synthesis process. **B** Fisher's exact test p-values as a function of the correlation cutoff calculated for three correlation estimators: Pearson correlation (pink), exact partial correlation (purple), GeneNet partial correlation (black). For each correlation cutoff, the original dataset was bootstrapped 1,000 times. Error bars represent the 95% confidence intervals of the bootstrapping results. Dashed lines represent the mean of the bootstrapped statistical cutoffs for GeneNet. **C** Fisher's exact test p-value for partial correlations estimated with GeneNet, as a function of both sample size and correlation cutoff. Shade represents the mean across 1,000 bootstrapping samples, while the red line represents the mean of the 0.01 FDR cutoff.

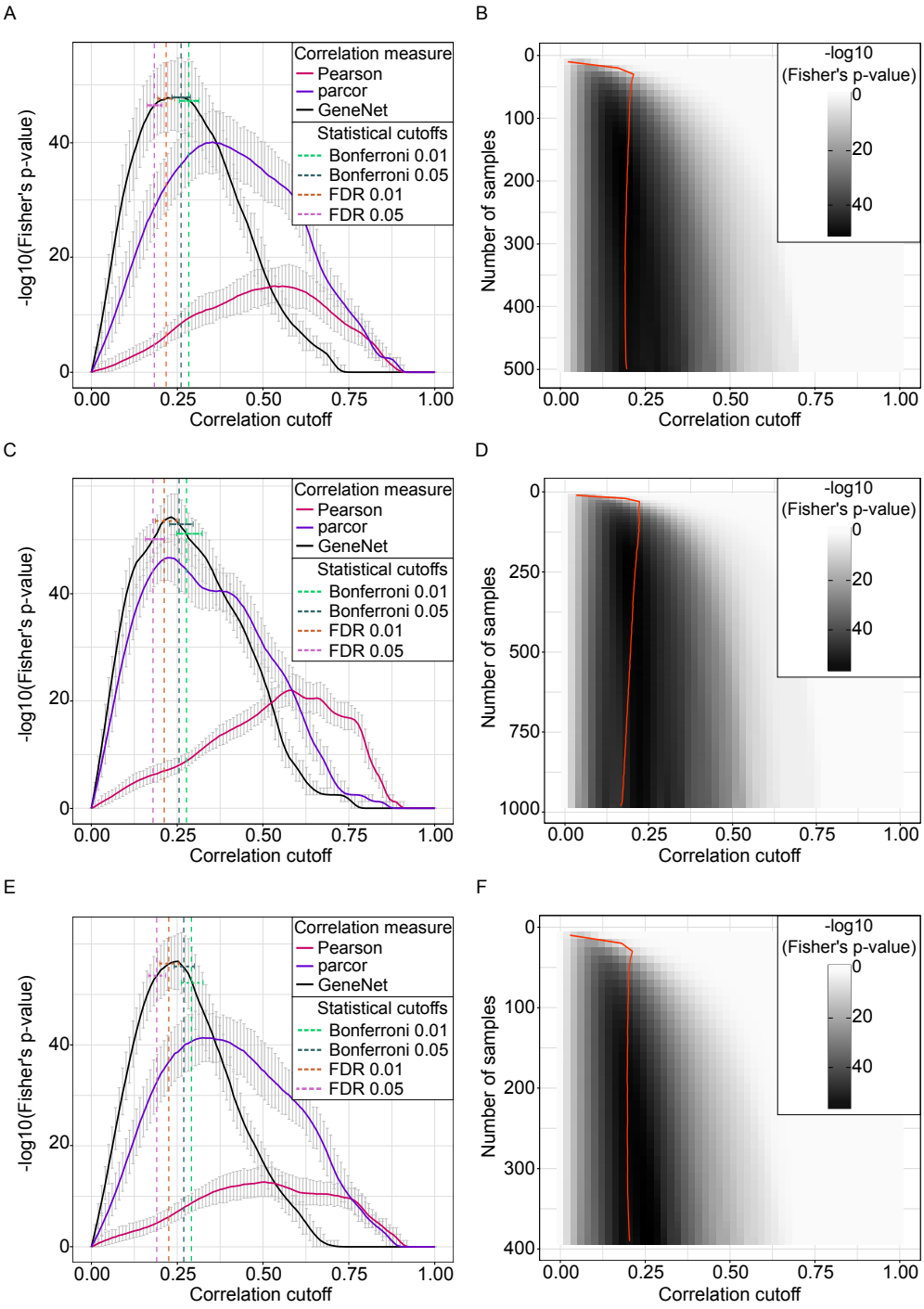


Figure 6.5: Network quality as a function of the correlation cutoff in the glycomics replication cohorts. **A**, **C**, **E** Fisher's exact test p-values as a function of the correlation cutoff calculated in the Korčula 2010, Split and Vis cohorts, respectively, for three correlation estimators: Pearson correlation (red), exact partial correlation (purple), GeneNet partial correlation (black). For each correlation cutoff, the original dataset was bootstrapped 1,000 times.

Error bars represent the 95% confidence intervals of the bootstrapping results. Dashed lines represent the mean of the bootstrapped statistical cutoffs for GeneNet. Interestingly, the maxima of the GeneNet curve are fairly similar across cohorts: 0.24 (Korčula 2010), 0.23 (Split), 0.24 (Vis). **B**, **D**, **F** Fisher's exact test p-value for partial correlations estimated with GeneNet in the Korčula 2010, Split and Vis cohorts, respectively, as a function of both sample size and correlation cutoff. Colors represent the mean across 1,000 bootstrapping samples, while the red line represents the mean of the 0.01 FDR cutoff.

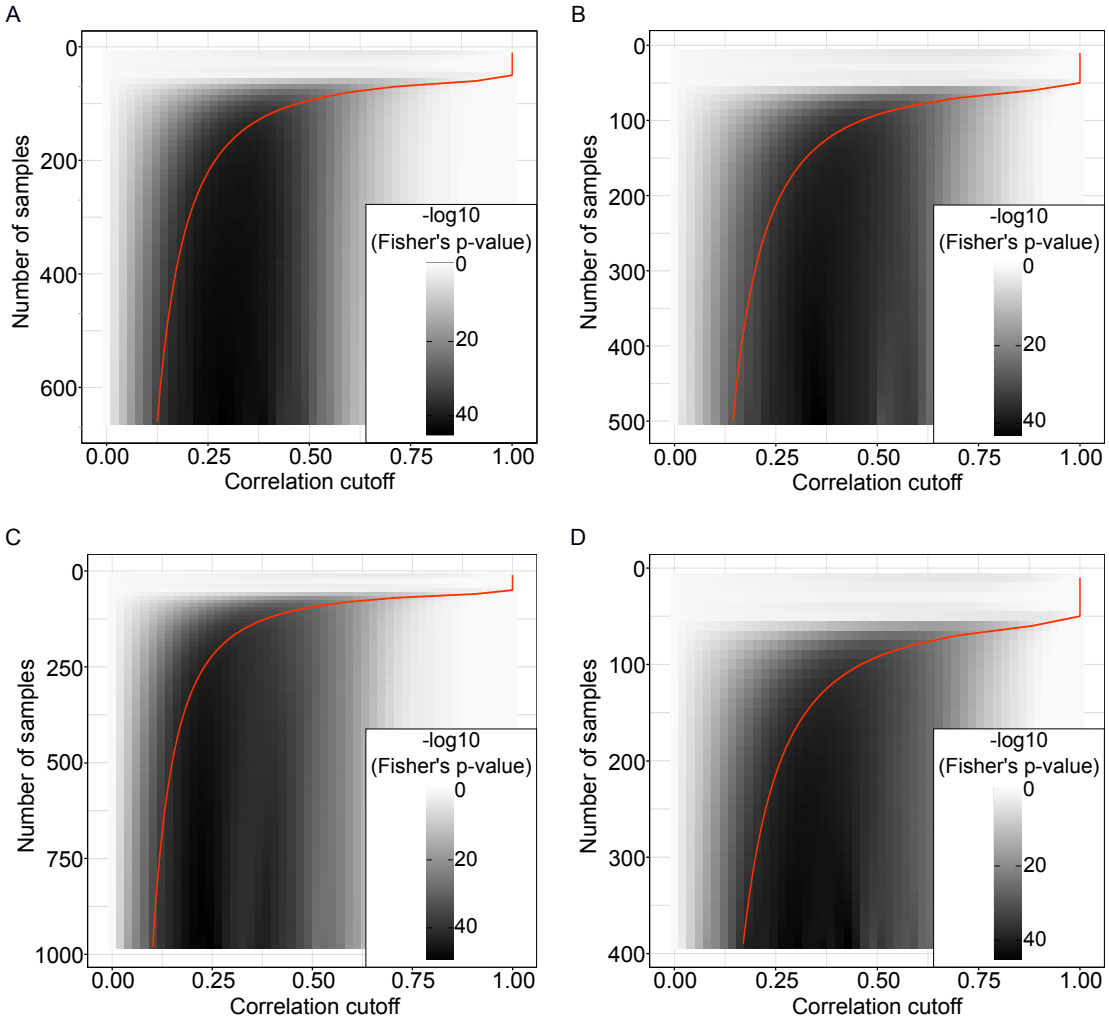


Figure 6.6: Cutoff optimization as a function of the sample size for parcor correlation in the glycomics replication cohorts. **A**, **B**, **C**, **D** Fisher's exact test p-value for partial correlations estimated with parcor in the Korčula 2013 (A), Korčula 2010 (B), Split (C) and Vis (D) cohorts, respectively, as a function of both sample size and correlation cutoff. Color represents the mean across 1,000 bootstrapping samples, while the red line represents the mean of the 0.01 FDR cutoff.

Notably, the optimal GeneNet network outperforms the GeneNet networks obtained with most statistical cutoffs. The analysis of the replication cohorts showed similar results

(Figure 6.5). This proves that biological prior knowledge improves the choice of a network cutoff, and that the optimal network is identifiable and unique for all correlation measures considered.

To assess whether the optimal network obtained with our procedure depends on sample size, as statistical cutoffs, we reperformed the optimization procedure on subsamples of the original dataset (Figure 6.4C). For GeneNet, the optimal cutoff turned out to be size-independent, as expected. This indicates that, by optimizing the cutoff with our approach even with a relatively small sample size (roughly 160 observations), we still obtain the

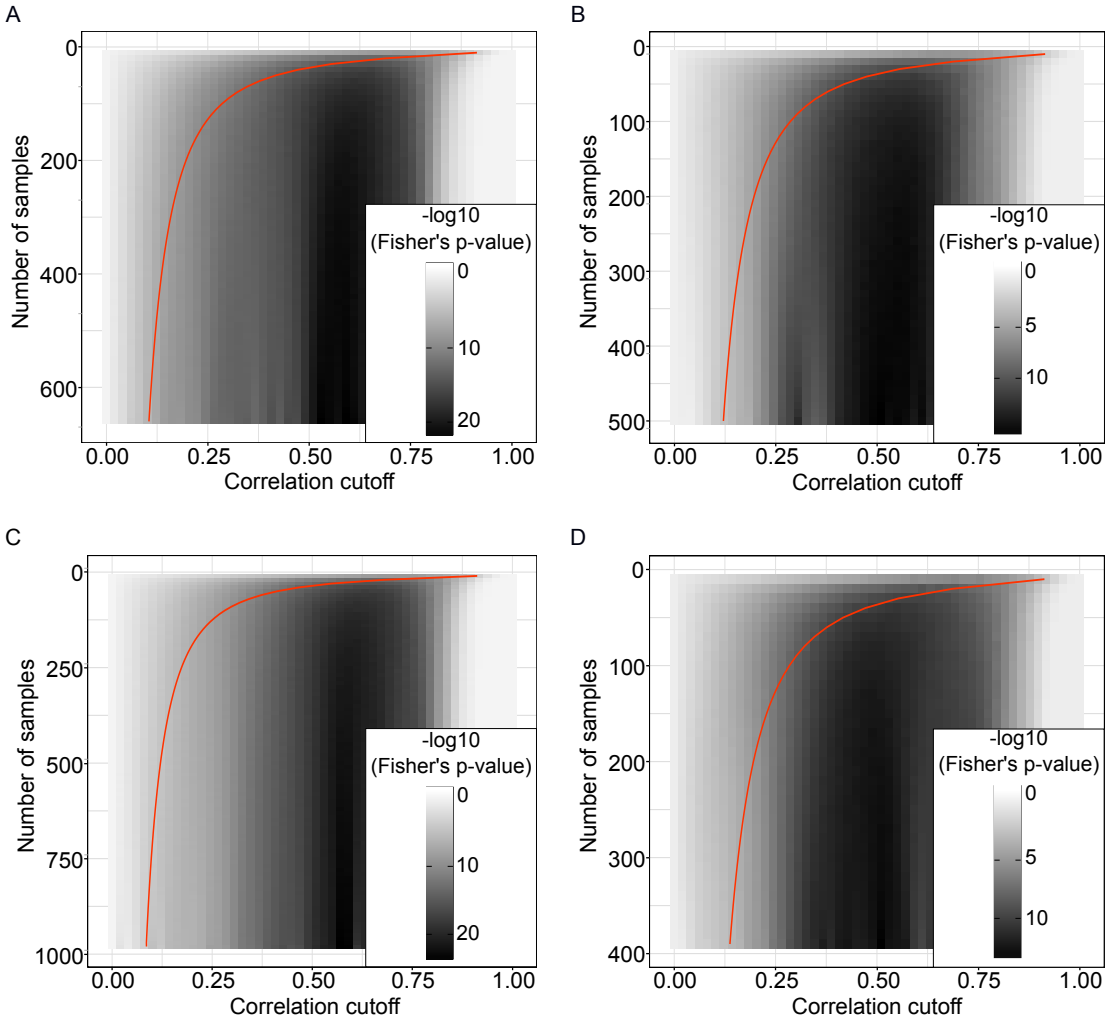


Figure 6.7: Cutoff optimization as a function of the sample size for Pearson correlation in the glycomics replication cohorts. **A**, **B**, **C**, **D** Fisher's exact test p-value for Pearson correlations in the Korčula 2013 (A), Korčula 2010 (B), Split (C) and Vis (D) cohorts, respectively, as a function of both sample size and correlation cutoff. Color represents the mean across 1,000 bootstrapping samples, while the red line represents the mean of the 0.01 FDR cutoff.

same optimal network that we would get with a much larger dataset (669 observations). Strikingly, even for Pearson and parcor correlation, for which statistical cutoffs showed strong sample size dependence, the optimal cutoff appeared to be sample size independent over 300 samples (Figure 6.6 and 6.7, respectively), although the overall performance was lower than GeneNet. In conclusion, using prior information to optimize the correlation cutoff allowed to infer the same optimal network regardless of the sample size of the considered dataset.

6.3 Incomplete, incorrect, or coarse biological references

Our optimization approach determines the correlation cutoff at which the data-driven network best represents the biological reference. However, IgG glycan synthesis is a very well-characterized process, while in most other practical cases a reference that describes the system in detail is not available. We postulate that even with an incomplete or partially incorrect biological reference, we will obtain an optimal network. To this end, we considered the performance obtained from the optimization procedure when comparing the full biological reference with an artificially incomplete, incorrect or coarse version of it, as described in the following.

Scenario 1: incomplete biological reference. Since for many biological systems the full biochemical pathway of synthesis is not available, we simulated a case in which only a given percentage of the IgG glycosylation pathway is known. To this end, we randomly constructed incomplete pathways by selecting a fraction (10 to 90% in increments of 10%) of the edges in the IgG glycosylation pathway shown in Figure 6.4A. For each percentage, we generated 100 different incomplete pathways and used each of them to optimize the correlation cutoff (Figure 6.8A). Obviously, due to the increase in false positives, the fewer edges from the original reference we consider, the lower the overlap to the correlation network becomes. Importantly, however, the optimum is highly conserved across the curves, yielding the same optimal cutoff (0.23) regardless of the amount of prior information available. This means that if we only knew, e.g., 50% of the reactions in the IgG glycosylation pathway shown in Figure 2.2, we would still obtain the identical optimal network as we would by using the full pathway.

Scenario 2: partially wrong biological reference. In many cases, our understanding of how a biological system works might be partially incorrect. Therefore, we considered

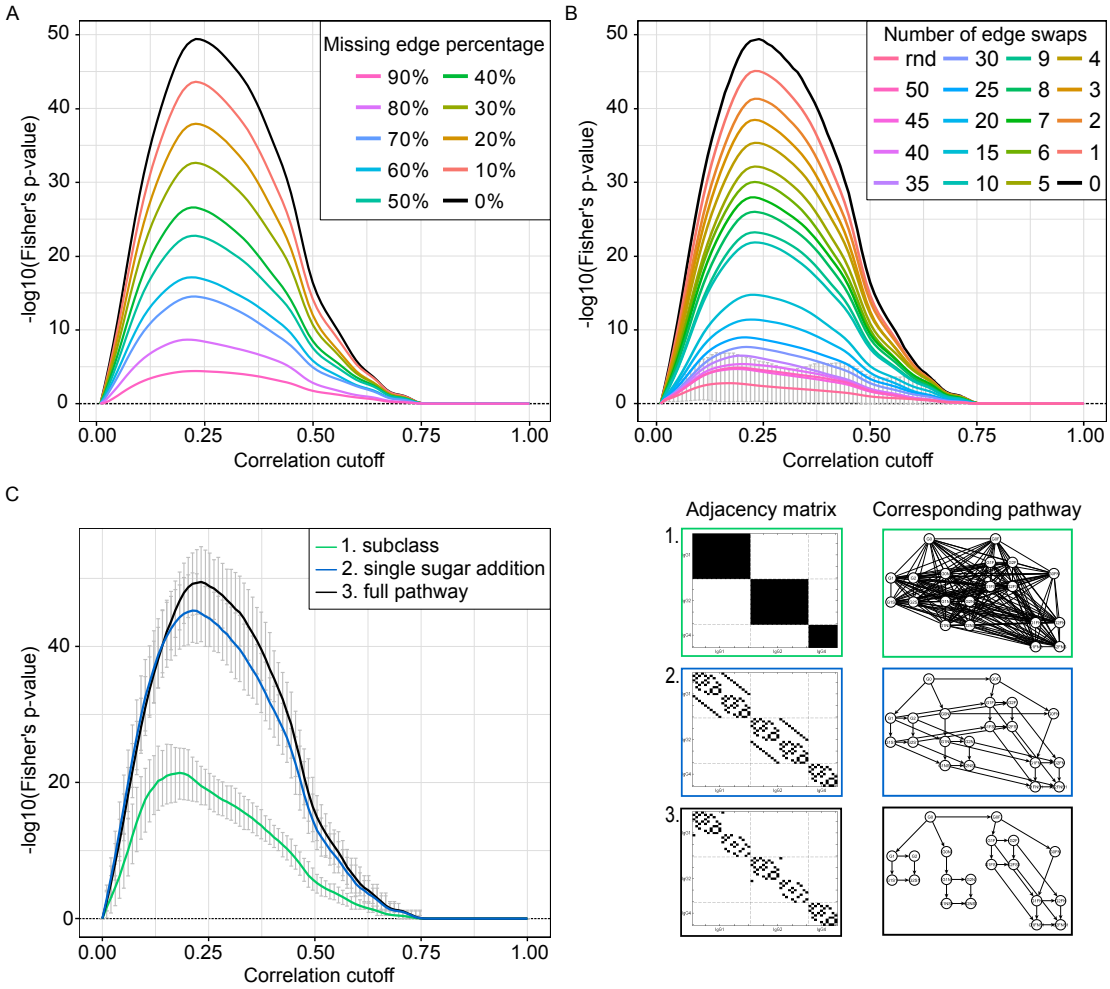


Figure 6.8: Cutoff optimization with partial knowledge. **A** Incomplete biological reference. For each percentage, 100 different adjacency matrices were generated by randomly selecting edges from the IgG glycosylation pathway. The curves in the figure represent the means of the 1,000 bootstrapping resamplings on each adjacency matrix. **B** Incorrect biological reference. Edges in the IgG glycosylation pathway were randomly swapped to simulate incorrect information in the biological reference. For each considered number of swaps, 100 adjacency matrices were generated and the averages over those curves and over the 1,000 bootstrapping resamplings are shown. Here, the red curve represents 100 fully randomized adjacency matrices. The error bars on this curve represent the 95% confidence interval of the bootstrapping. Any signal that falls within these intervals should be regarded as noise. **C** Coarse biological reference. For IgG glycomics data we know that only enzymatic reactions between glycans attached to the same IgG isoform are feasible (adjacency matrix 1) and, in addition, that only they can be modified by the addition of one sugar unit at a time (adjacency matrix 2). The black curve corresponds to the optimization performed on the full reference (adjacency matrix 3) for comparison. The curves in the figure represent the means of the 1,000 bootstrapping resamplings and the different considered adjacencies. In all plots, the black curve corresponds to the optimization performed on the full reference. The error bars represent the 95% confidence interval of the bootstrapping.

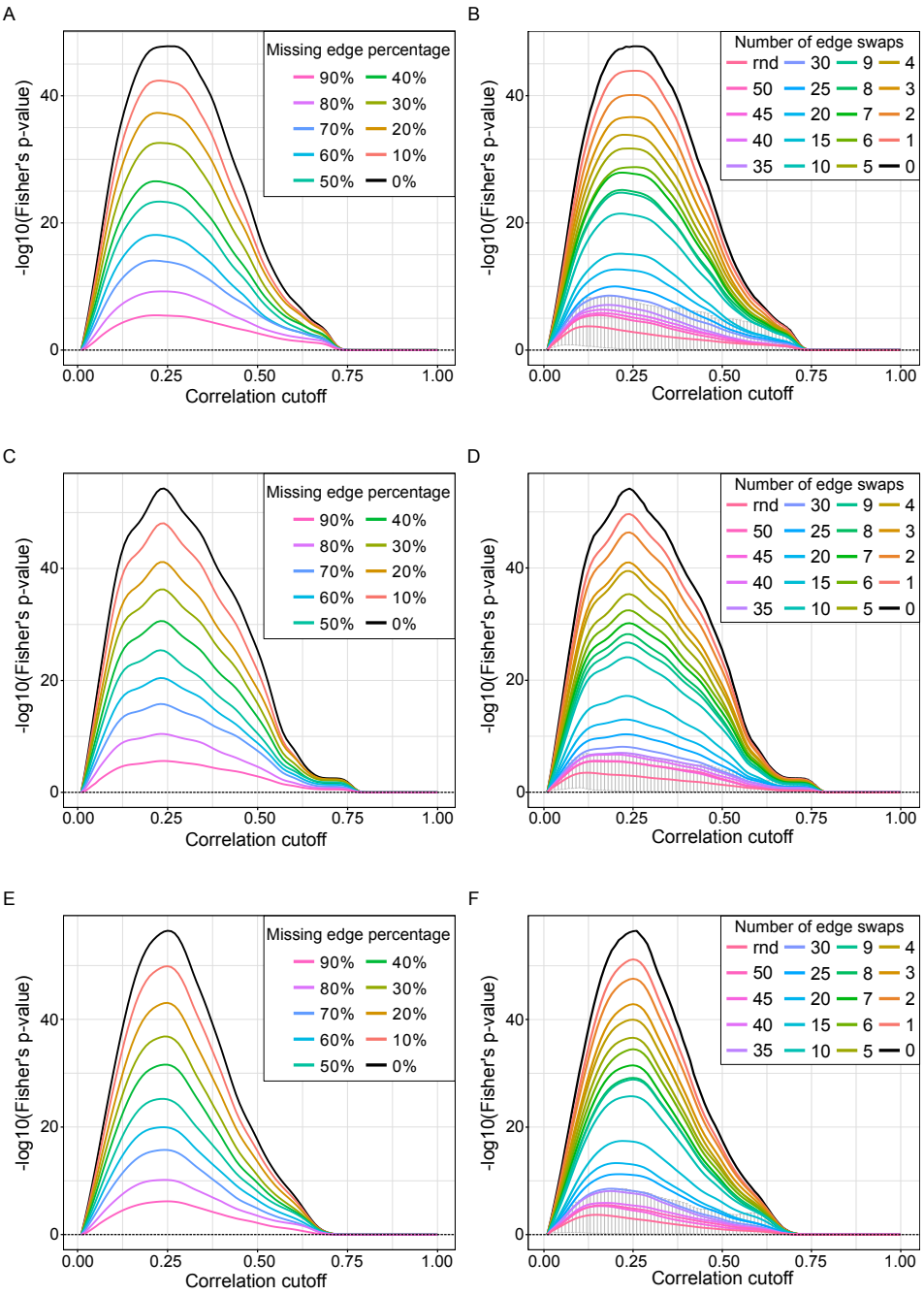


Figure 6.9: Cutoff optimization with partial knowledge in the glycomics replication cohorts. **A**, **C**, **E** Incomplete biological reference in the Korčula 2010 (A), Split (C) and Vis (E) cohorts. For each percentage, 100 different adjacency matrices were generated by randomly selecting edges from the IgG glycosylation pathway. The curves in the figure represent the means of the 1,000 bootstrapping resamplings on each adjacency matrix. **B**, **D**, **F** Incorrect biological reference in the Korčula 2010 (B), Split (D) and Vis (F) cohorts. Edges in the IgG glycosylation pathway were randomly swapped to simulate incorrect information in the biological reference.

For each considered number of swaps, 100 adjacency matrices were generated and the averages over those curves and over the 1,000 bootstrapping resamplings are shown. Here, the red curve represents 100 fully randomized adjacency matrices. The error bars on this curve represent the 95% confidence interval of the bootstrapping. Any signal that falls within these intervals should be regarded as noise.

the possibility of our reference to include wrong information, i.e., a given number of wrong edges.

We simulated an increasing number of edge swaps in the IgG glycosylation pathway until we reached full randomization. For each condition, we generated 100 different pathways and performed the optimization procedure on them (Figure 6.8B). Again, while the overall performance decreased as expected, the shape of the curve clearly leads to the same optimal cutoff as the original pathway for up to 20 swaps. This means that even when starting with a substantially incorrect prior, as long as partial truth is contained in the reference network, the optimized network will still produce the same network as the one obtained with the complete biological reference.

Scenario 3: coarse biological reference. Sometimes no detailed biochemical mechanisms of synthesis are known, but only general biological properties of the molecules in the dataset. For example, we know that glycan processing occurs when the sugar chain is already bound to the protein. In our datasets, we have the measurements of three different protein isoforms (IgG1, IgG2 and IgG3 together, and IgG4). Therefore, we can constrain the set of possible biochemical reactions only to glycans pairs within the same IgG isoform (adjacency matrix 1 in Figure 6.8C). Moreover, we know that glycosylation enzymes can only add a single monosaccharide at a time during glycan synthesis. Hence, we can further reduce the possible reactions to those between glycan pairs that differ of a single sugar unit (adjacency matrix 2 in Figure 6.8C). When comparing the optimization results carried out starting from these biological references to that of the full biochemical pathway (adjacency matrix 3 in Figure 6.8C), we observe that, while the overall performance varies substantially, the optimal values are close to each other, thus producing similar networks. Therefore, even when biochemical details are not available for the system under study, other sources of information can be used for the optimization and lead to the same optimum as the complete biological reference.

The three scenarios' results could be replicated for the other cohorts (Figure 6.9 and 6.10).

In conclusion, for various cases of incomplete prior knowledge, our approach still leads to a close to globally optimal network.

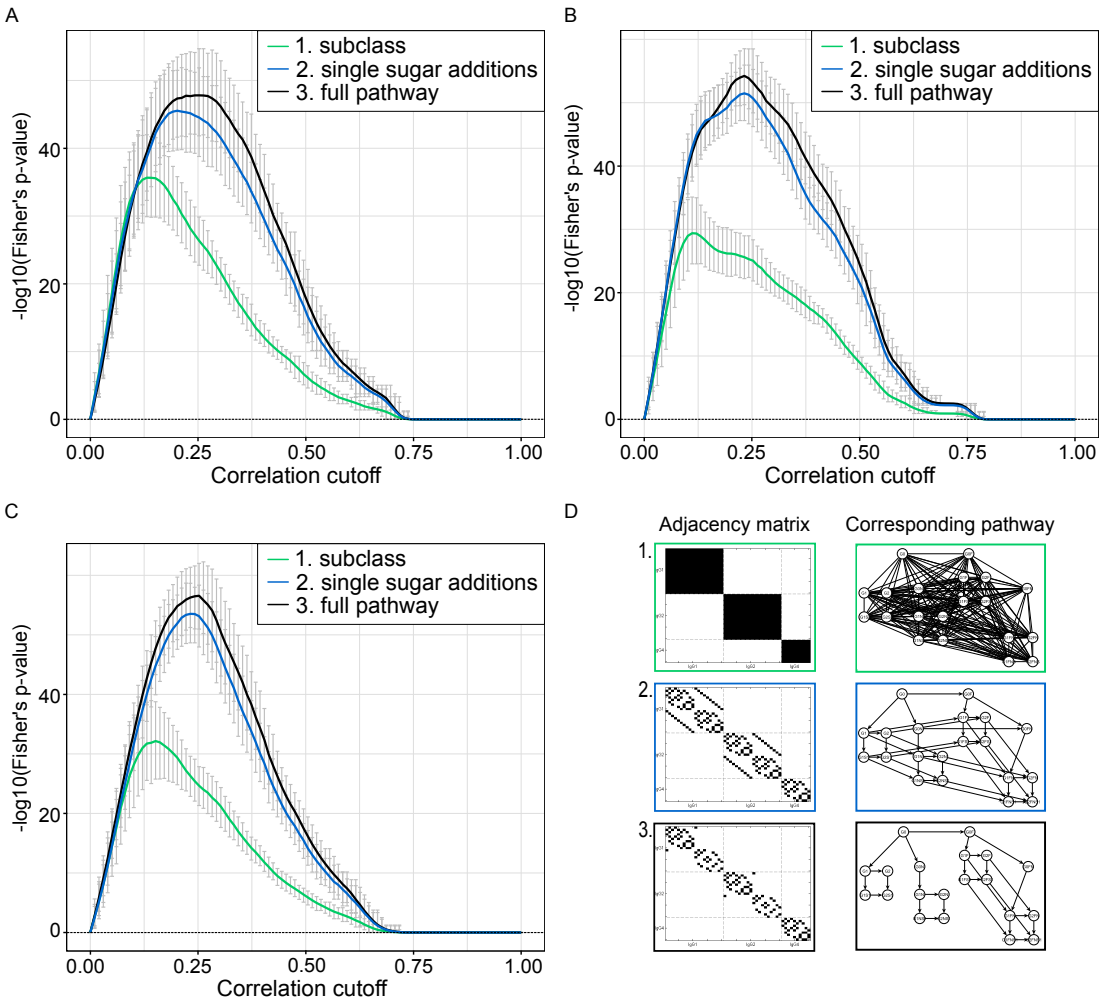


Figure 6.10: Cutoff optimization with coarse prior knowledge in the glycomics replication cohorts. **A**, **B**, **C** Coarse biological reference in the Korčula 2010 (A), Split (B) and Vis (C) cohorts. The curves in the figure represent the means of the 1,000 bootstrapping resamplings and the different considered adjacencies. The black curve corresponds to the optimization performed on the full reference for comparison. The error bars represent the 95% confidence interval of the bootstrapping. **D** References considered in the optimization. For IgG glycomics data we know that only enzymatic reactions between glycans attached to the same IgG isoform are feasible (adjacency matrix 1) and, in addition, that only they can be modified by the addition of one sugar unit at a time (adjacency matrix 2). The black curve corresponds to the optimization performed on the full reference (adjacency matrix 3) for comparison.

6.4 Application to metabolomics data

In order to test whether our approach can be generalized to other data types, we applied the algorithm to untargeted urine metabolomics dataset (Antipsychotics cohort, Table 2.3). The dataset consisted of 95 samples with 1,021 measured metabolites. Data were normalized, log-transformed, imputed and corrected for age, gender, and BMI prior to analysis (see Methods). Since current pathway databases cover only a part of the metabolites measured in a typical mass-spectrometry-based analysis, we had to rely on partial prior information: (1) Enzymatic reactions connecting the measured metabolites were obtained from the RECON2 database13. In addition, we created adjacency matrices from metabolite annotations. We constrained reactions between metabolites to only those among molecules (2) within the same biological pathway (in the following referred to as sub-pathway) or (3) within the same molecular class (referred to as super-pathway).

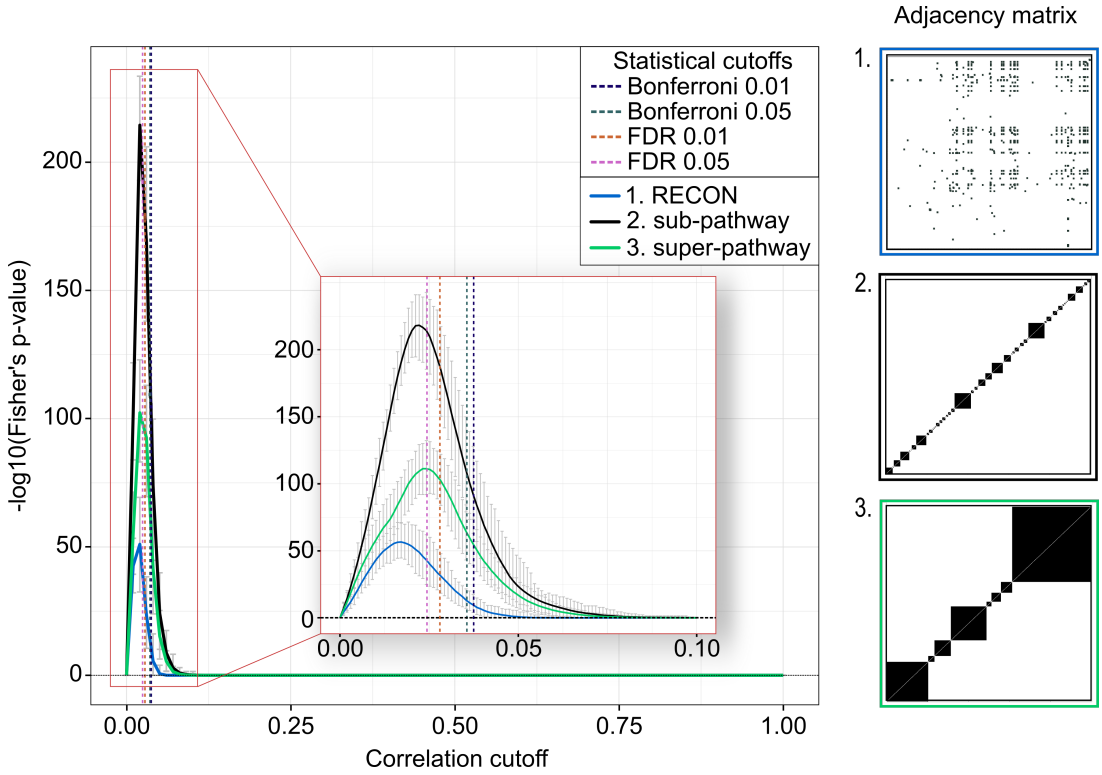


Figure 6.11: Cutoff optimization for metabolomics data. We used biochemical reactions from the RECON database as partial prior knowledge (adjacency matrix 1), as well as sub- and super-pathway annotations (adjacency matrices 2 and 3, respectively). Curves in the figure represent the average over 100 bootstrapping resamplings, and error bars show the corresponding 95% confidence intervals. Vertical lines indicate the mean of the statistical cutoffs, and the horizontal error bars the corresponding 95% confidence intervals over the bootstrapping.

We inferred GeneNet-based networks using these three priors as biological references (Figure 6.11).

Although the absolute performances vary significantly depending on the chosen prior, the maxima are still remarkably close to each other. This means that the corresponding resulting optimal networks will be similar. Similar to the glycomics case, the statistical cutoff of FDR 0.05 was found to consistently perform comparably to the optimized cutoff. The performance of Pearson and parcor correlation measures can be found in Figure 6.12.

In conclusion, we demonstrated that our approach can be generalized to metabolomics data, where a full biological reference is unavailable. Partial prior information can be used from different sources and the optima obtained with different priors are highly consistent.

6.5 Application to transcriptomics data

To evaluate the approach on a substantially different type of omics data, we analyzed RNA-sequencing data from The Cancer Genome Atlas [142] (TCGA, Table 2.4). After preprocessing, the dataset included expression measurements of 12,005 genes from 2,726 samples across 11 different cancer types (see Subsection 2.2.3). Values were corrected for age, gender and cancer type prior to analysis.

Gene expression measurements were analyzed in subsets according to their pathway annotations from the Reactome database [150, 151]. For each pathway, we used the corresponding STRING [152, 153] subnetwork as biological reference.

Since we tested 469 pathways in this analysis, we used a conservative significance threshold for the Fisher's p-value of $0.01/469 = 2.13 \cdot 10^{-5}$, yielding 129 pathways with a significant optimum (see Figure 6.14 for detailed results). In 102 of these cases, our optimized cutoff clearly outperformed the statistical cutoffs, which tend to produce too sparse networks. As a showcase of how this difference in correlation cutoff translates into differences in the inferred network, we compared the partial correlation networks, or GGMs, obtained with FDR 0.05 to our optimization procedure for the "MAPK1/MAPK3 signaling" pathway (Figure 6.13). The statistically inferred network is substantially too sparse and thus only shows limited overlap to the biological reference (median of Fisher's test p-value $> 10^{-5}$). The optimized network achieved a p-value of 10^{-14}, therefore substantially better resembling the biological reference than the statistically inferred network. Notice that PPI networks are much denser than biochemical pathways (from the glycomics and

metabolomics analysis), and therefore the estimated GGMs was expected to reflect this property.

In conclusion, the analysis of the TCGA data demonstrated that the performance of statistical cutoffs is highly unpredictable and thus statistically inferred GGMs cannot be assumed to well represents the underlying molecular mechanisms.

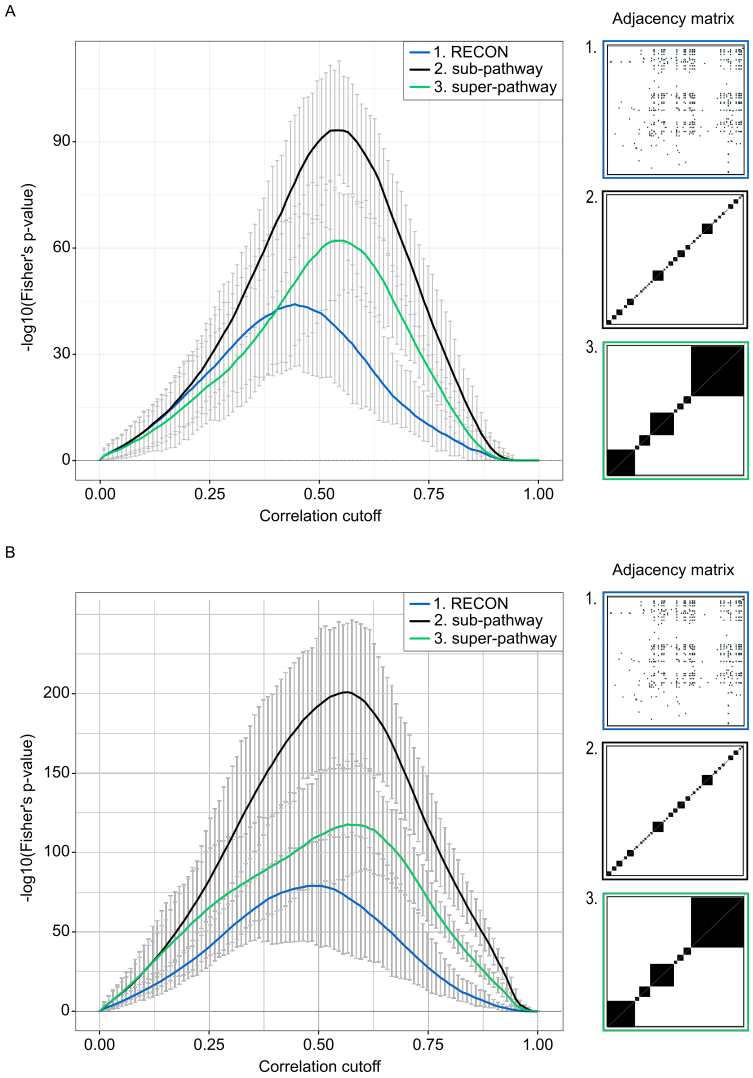


Figure 6.12: Parcor (A) and Pearson (B) correlation cutoff optimization with partial and coarse knowledge in the metabolomics cohort. As prior knowledge, we used biochemical reactions from the RECON database (adjacency matrix 1), as well as sub- and super-pathway annotations (adjacency matrices 2 and 3, respectively). Curves in the figure represent the average over 100 bootstrapping resamplings, and error bars show the corresponding 95% confidence intervals. Vertical lines indicate the mean of the statistical cutoffs, and the horizontal error bars the corresponding 95% confidence intervals over the bootstrapping.

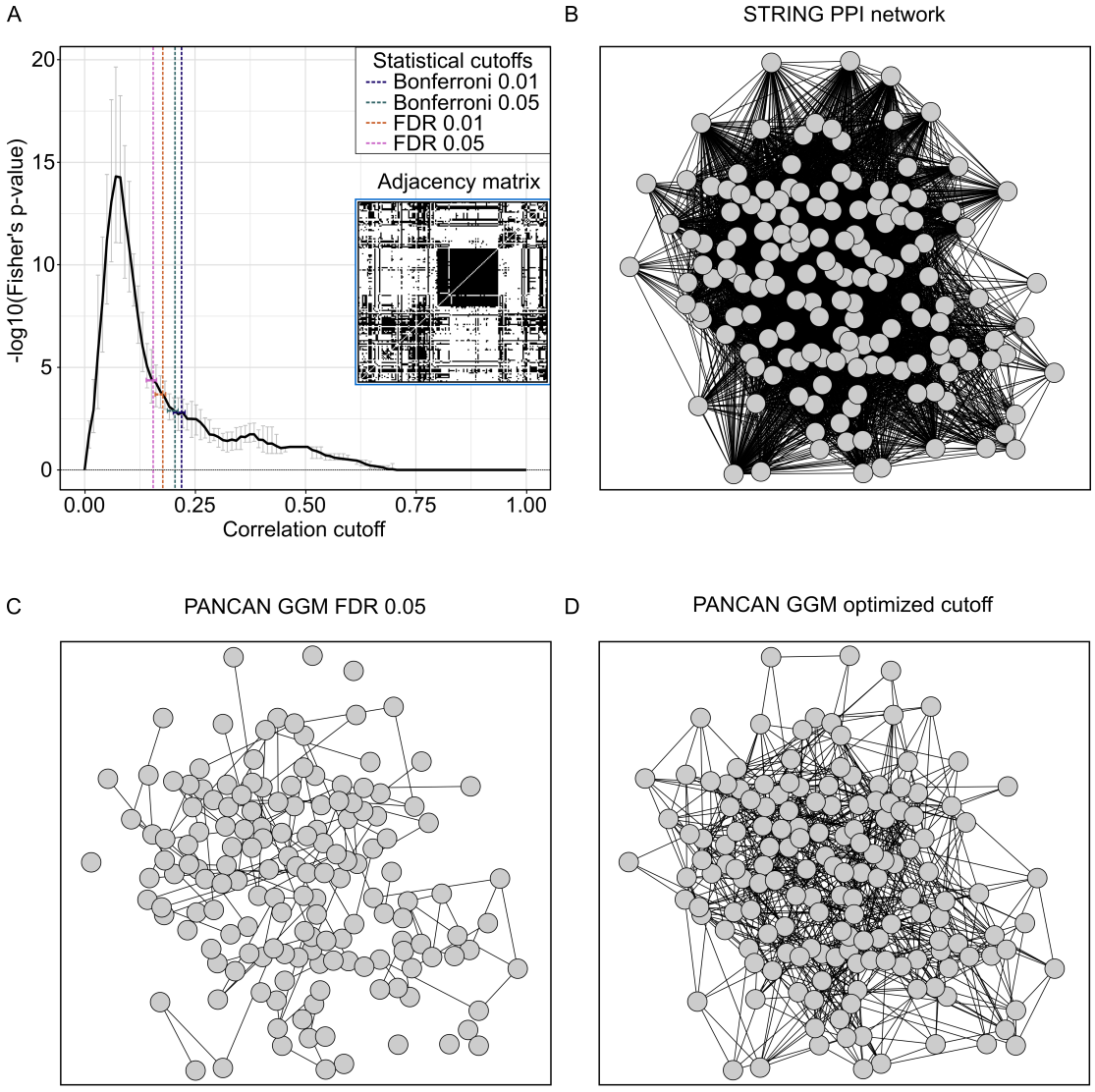


Figure 6.13: TCGA transcriptomics analysis results for the MAPK1/MAPK3 signaling pathway. **A** Cutoff optimization for the transcriptomics data. Protein-protein interaction networks from STRING were used as reference. The black curve represents the average over 100 bootstrapping resamplings, and the error bars show the corresponding 95% confidence intervals. Vertical lines indicate the mean of the statistical cutoffs, and the horizontal error bars the corresponding 95% confidence intervals over the bootstrapping. This example illustrates how statistical cutoffs can completely fail to identify a biological optimum. **B** Biological reference (PPI network from STRING). **C** GGM obtained with a 0.05 FDR cutoff (correlation cutoff 0.155). **D** GGM obtained with our optimization procedure (correlation cutoff 0.071).

6.6 Conclusion

In this chapter, we addressed a problem that has not received much attention in the field so far. Correlation network inference relies on statistical correlation cutoffs, which suffer from sample size dependence and are subject to an arbitrary choice of significance level and multiple testing correction procedure. We showed that an exception to this general observation is GeneNet, which exhibits a remarkable robustness to sample size, but is still subject to the statistical cutoff problem.

The approach presented here overcomes this problem by establishing a biologically optimal correlation cutoff for network inference. The procedure ranges over the correlation cutoff value until an optimal overlap with a given, possibly incomplete, biological reference is achieved. We benchmarked the approach on LC-ESI-MS IgG glycosylation data from four large Croatian cohorts. For this type of data, the full synthesis pathway has been established and could thus serve as a gold standard for method evaluation. We showed that for the GeneNet partial correlation method, the resulting optimization curve leads to a well-determined and unique optimum, regardless of sample size and p-value cutoffs. Other correlation-based methods performed inferior compared to GeneNet.

The approach was then applied to the more realistic case of partial prior knowledge, i.e., the case where a detailed and correct biological reference is not available. We considered three different scenarios: 1. Only a fraction of the biochemical pathway of synthesis is known; 2. The biochemical pathway contains incorrect information; 3. Only relations between classes of variables are known. In all three cases, we obtained nearly optimal networks even with little biological knowledge available. This means that even only marginally informative priors are sufficient to obtain a reasonable approximation of the true network optimum. We further demonstrated the applicability of the approach on metabolomics and transcriptomics data, for which only partial prior knowledge is available. The three partial biological references used for the metabolomics data, based either on metabolic reactions or molecular annotations, yielded very similar optima, supporting the claim that partial knowledge from different sources can be used to optimize the correlation cutoff. Our approach was further validated on transcriptomics data, where we used protein-protein interaction networks as references.

Interestingly, for the metabolomics dataset statistical, the 0.05 FDR cutoff was very close to the optimum, just like in the glycomics case. However, we argue that this good performance of FDR is purely coincidental and cannot be generalized to other datasets or data types. This was corroborated by the analysis of transcriptomics data, for which FDR cut-

offs were found to significantly overestimate the optimal cutoffs in many cases. Therefore, we conclude that FDR is not a real competitor of our approach, as its performance cannot be predicted a priori and varies substantially depending on the data type.

The procedure described in this chapter requires a quantitative overlap measure to perform the cutoff optimization. We chose Fisher's exact test p-value as a proxy for the agreement between calculated correlation network and prior knowledge. It is to be noted that more conventional machine learning measures exist for classification problems. The popular F1-score [206], however, does not account for true negatives and was therefore disregarded here. Interestingly, Matthews correlation coefficient [207], another popular measure that uses all values in the contingency table, is actually related to the Fisher's p-value. Its absolute value is proportional to the square root of the chi-square statistic, which is asymptotically equivalent to that of the Fisher's exact test [208].

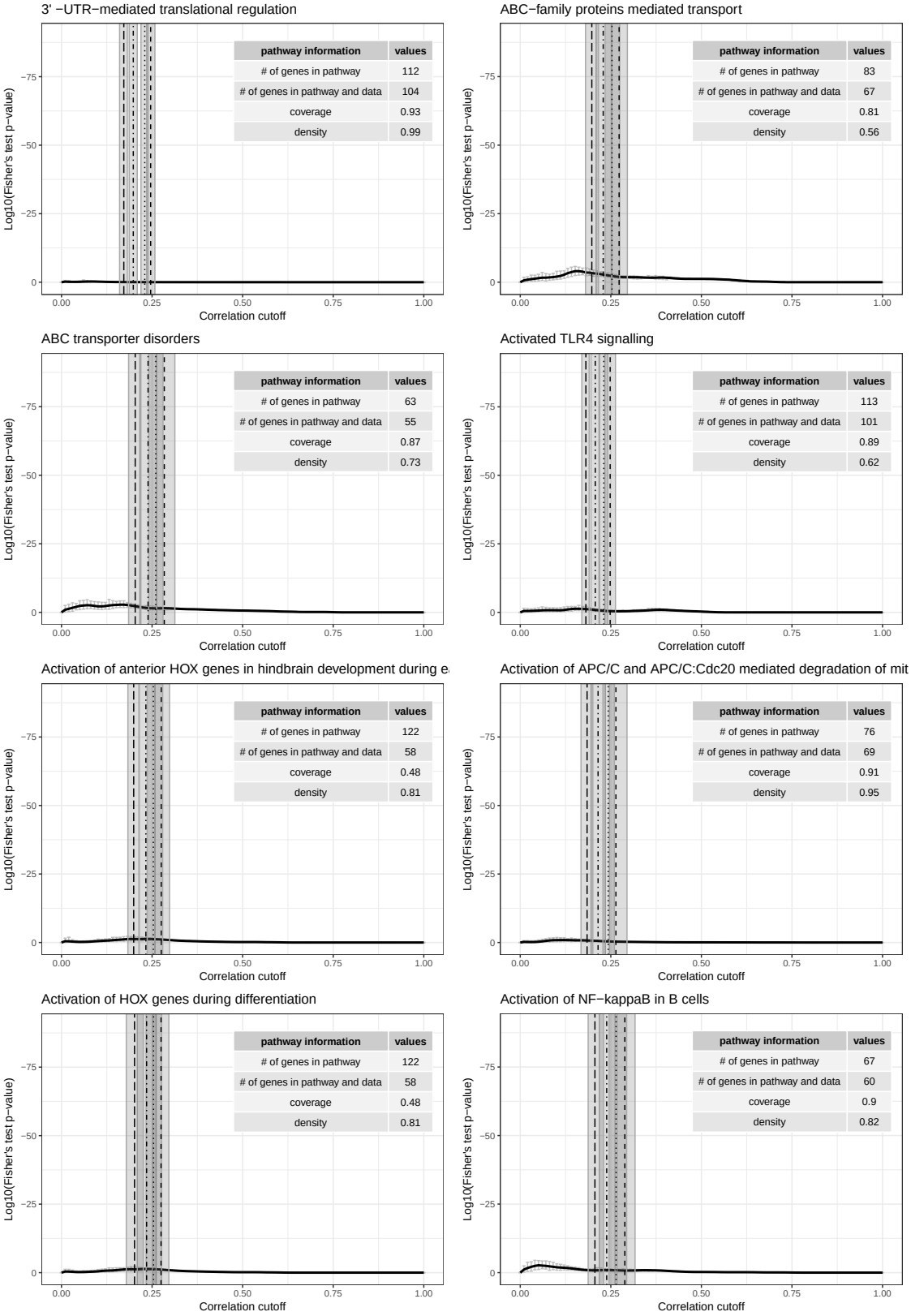
It is worth mentioning that our approach fails if there is a part of the network not covered by the biological reference (missing information) where a different optimal cutoff holds. Since the unknown part does not factor into the optimization process, the identified cutoff would be off and inference on that part of the network would be inaccurate. This is an issue that can conceptually not be overcome unless we get better coverage of the biological reference.

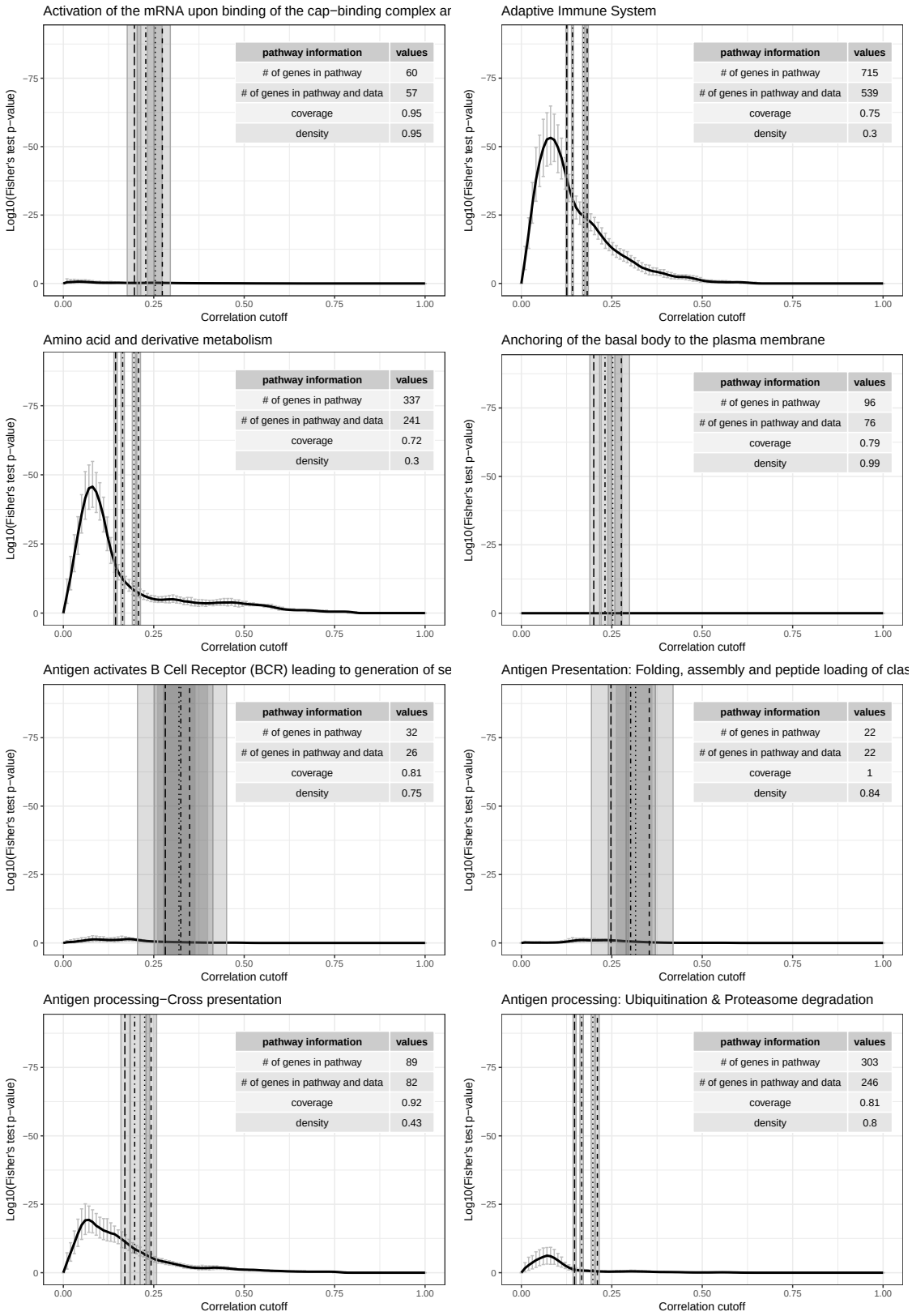
As an inference tool, cutoff optimization as presented in this chapter is a very flexible and generalizable strategy. Most of the network inference methods that attempt to account for prior knowledge integrate the biological reference directly into a specific network inference or regression framework [209–214], for example by penalizing or enhancing specific edges according to the biological reference. On the contrary, our approach uses prior knowledge as an external reference system to optimize the purely data-driven association matrix. This will allow applying the same concept to different association measures, for example mutual information [215] or other non-linear association quantities, in future studies.

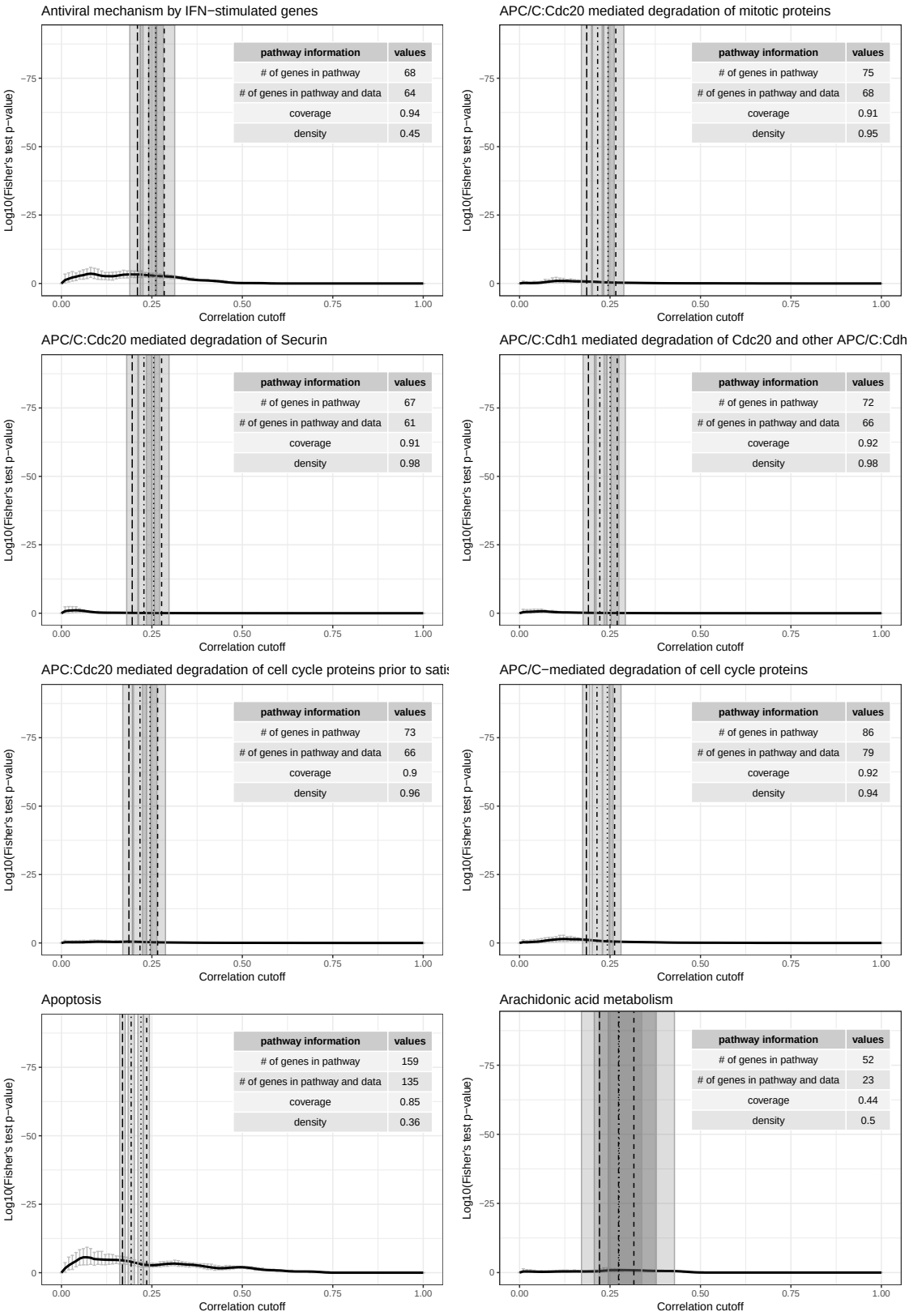
A central feature of our approach is that our formulation is robust to partial and incorrect information in the biological reference, which makes it applicable to a wide variety of omics data, in particular when the prior knowledge of the system is sparse and possibly of low quality.

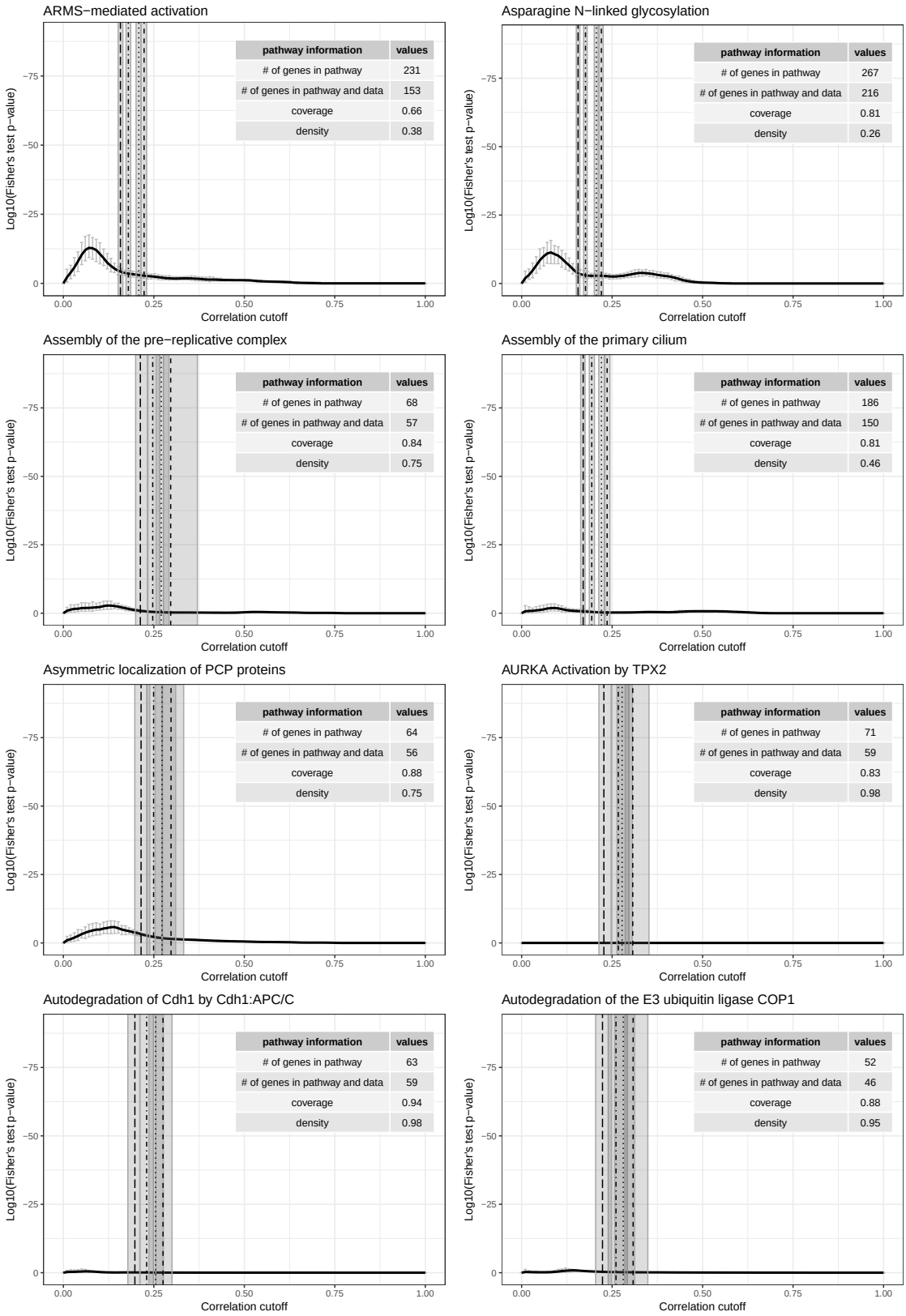
In conclusion, we have shown that even when dealing with data types for which a well-characterized biochemical pathway or interaction network is not fully established, priors based on the known biology of the system can be used for optimization. We have demonstrated that our optimal network outperforms those obtained with common statistical

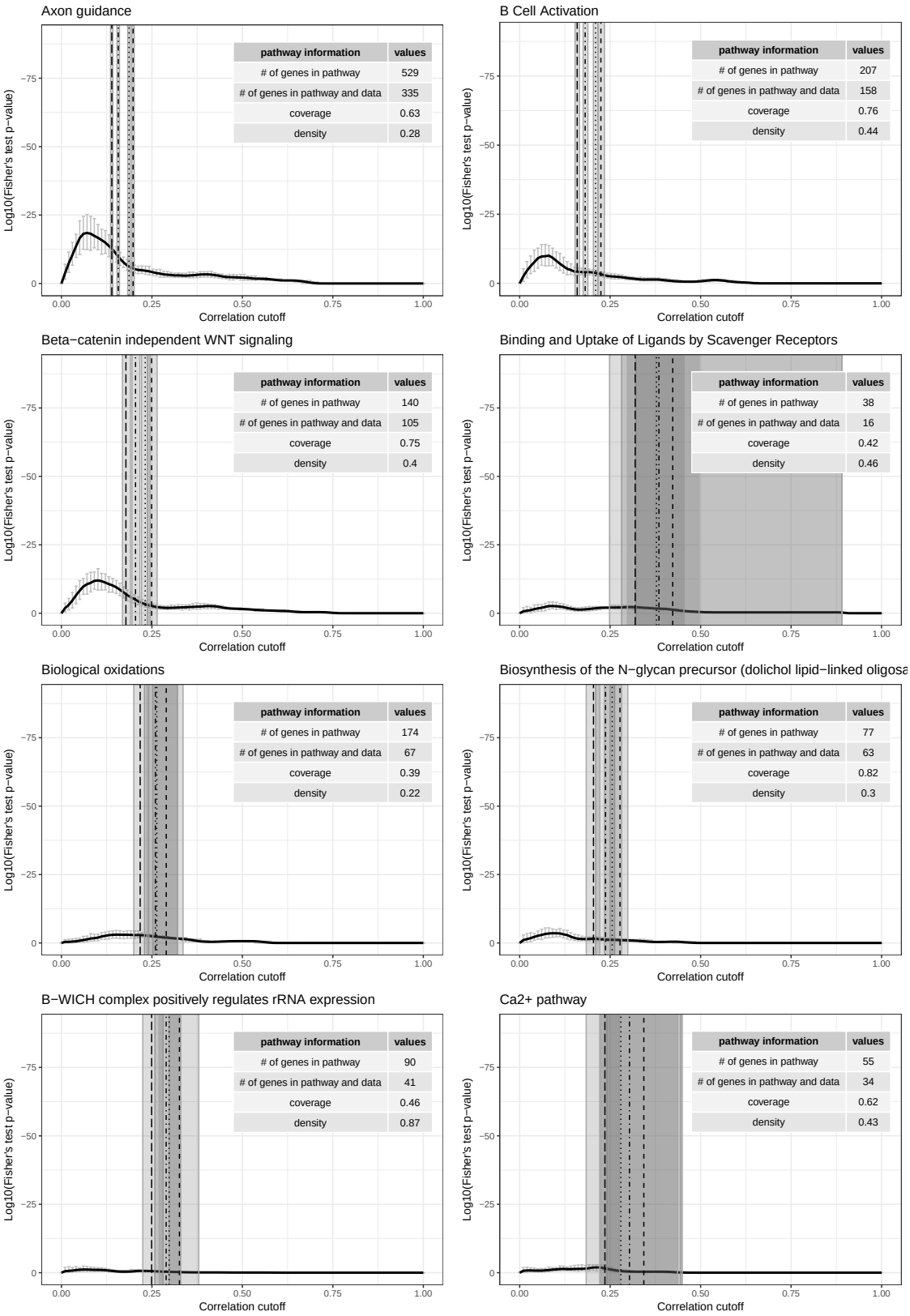
cutoffs, and therefore recommend using our pipeline in cases where at least some biological information on the system under study is available.

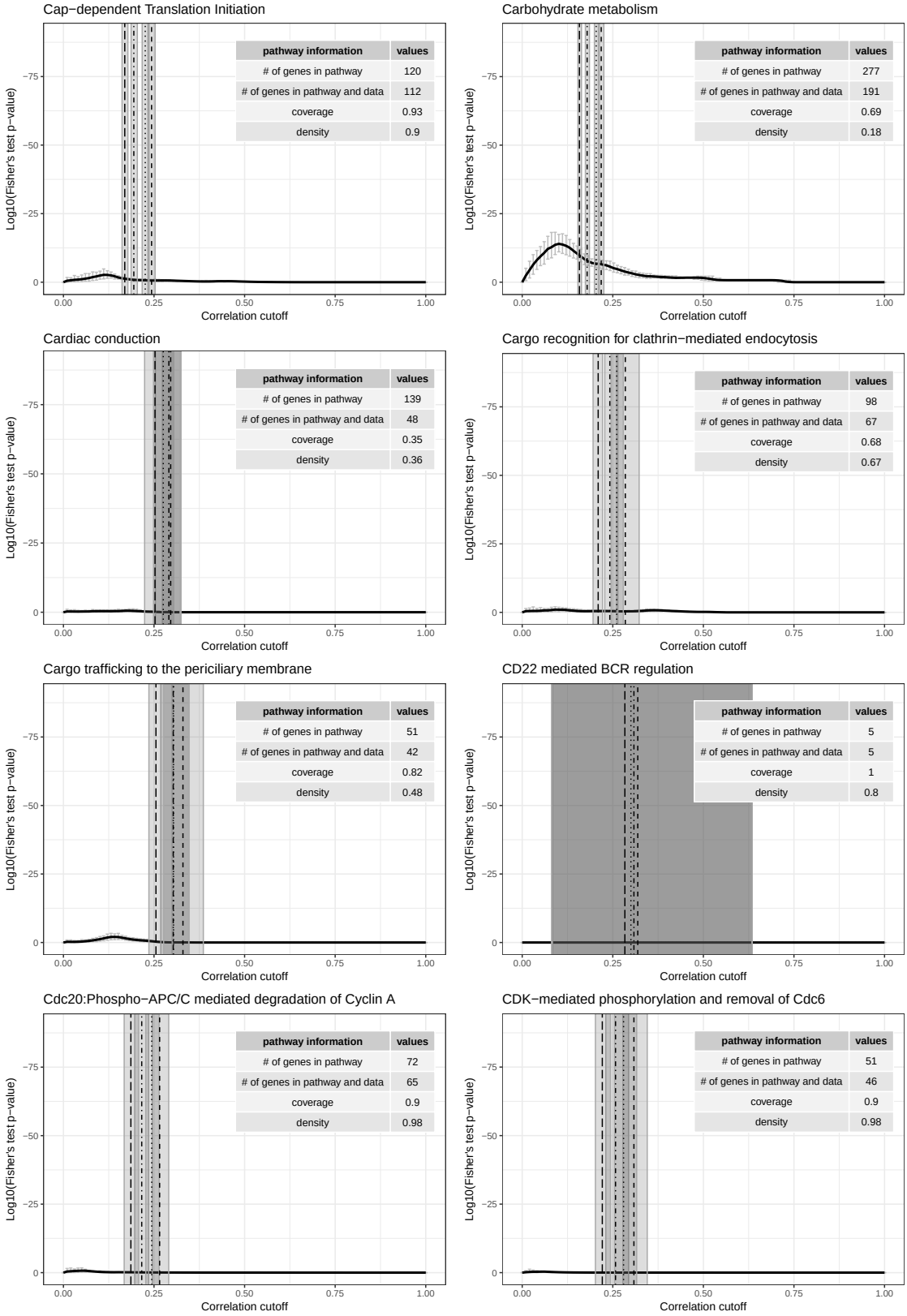

3' −UTR−mediated translational regulation
ABC−family proteins mediated transport
ABC transporter disorders
Activated TLR4 signalling
Activation of anterior HOX genes in hindbrain development during e
Activation of APC/C and APC/C:Cdc20 mediated degradation of mit
Activation of HOX genes during differentiation
Activation of NF−kappaB in B cells
pathway information
values
# of genes in pathway
# of genes in pathway and data
coverage
density
112
104
0.93
0.99
83
67
0.81
0.56
63
55
0.87
0.73
113
101
0.89
0.62
122
58
0.48
0.81
76
69
0.91
0.95
122
58
0.48
0.81
67
60
0.9
0.82
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

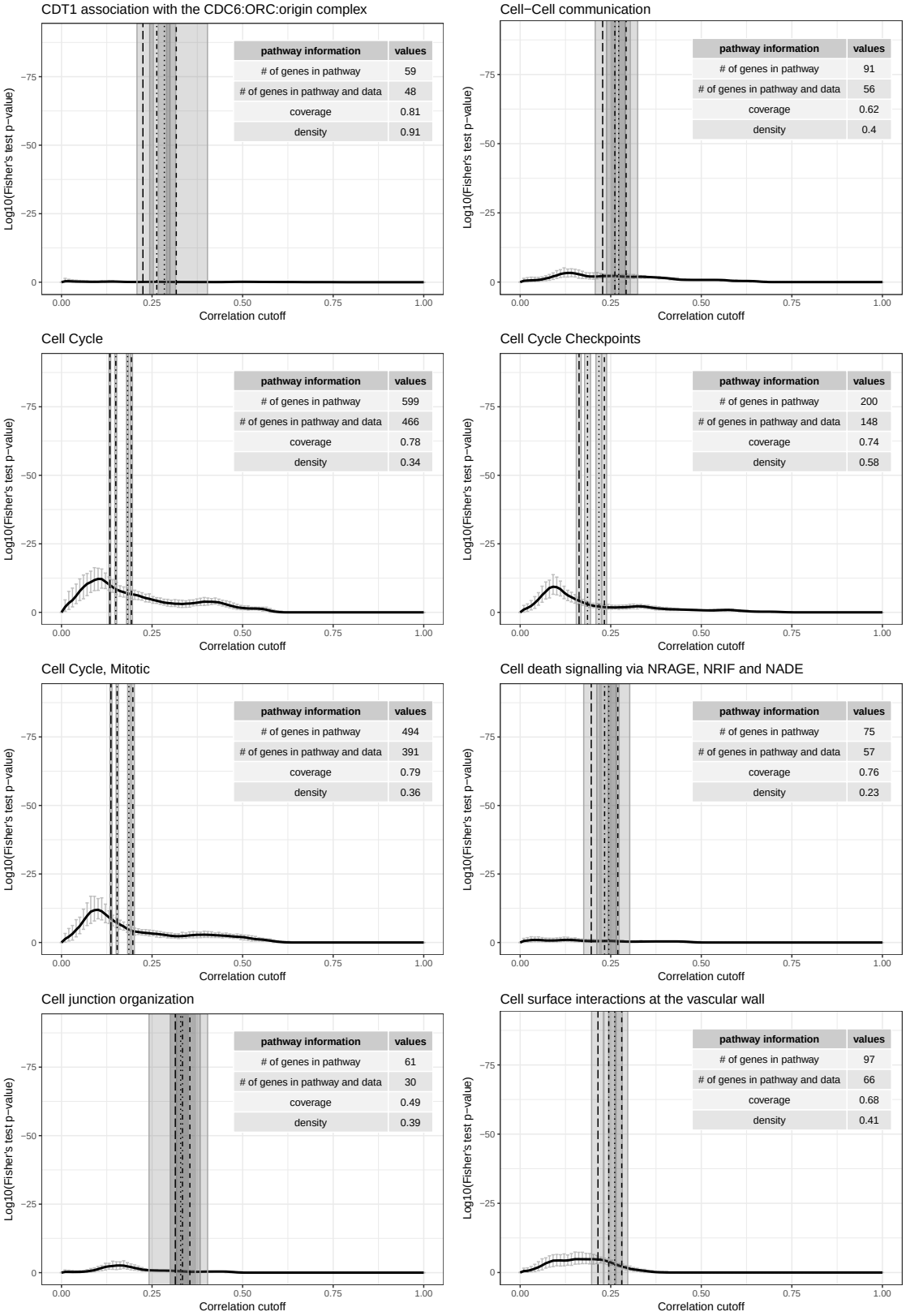

Activation of the mRNA upon binding of the cap−binding complex ar
Adaptive Immune System
Amino acid and derivative metabolism
Anchoring of the basal body to the plasma membrane
Antigen activates B Cell Receptor (BCR) leading to generation of se
Antigen Presentation: Folding, assembly and peptide loading of clas
Antigen processing−Cross presentation
Antigen processing: Ubiquitination & Proteasome degradation
Log10(Fisher's test p−value)
Correlation cutoff
pathway information
values
# of genes in pathway
# of genes in pathway and data
coverage
density
60
57
0.95
0.95
715
539
0.75
0.3
337
241
0.72
0.3
96
76
0.79
0.99
32
26
0.81
0.75
22
22
1
0.84
89
82
0.92
0.43
303
246
0.81
0.8

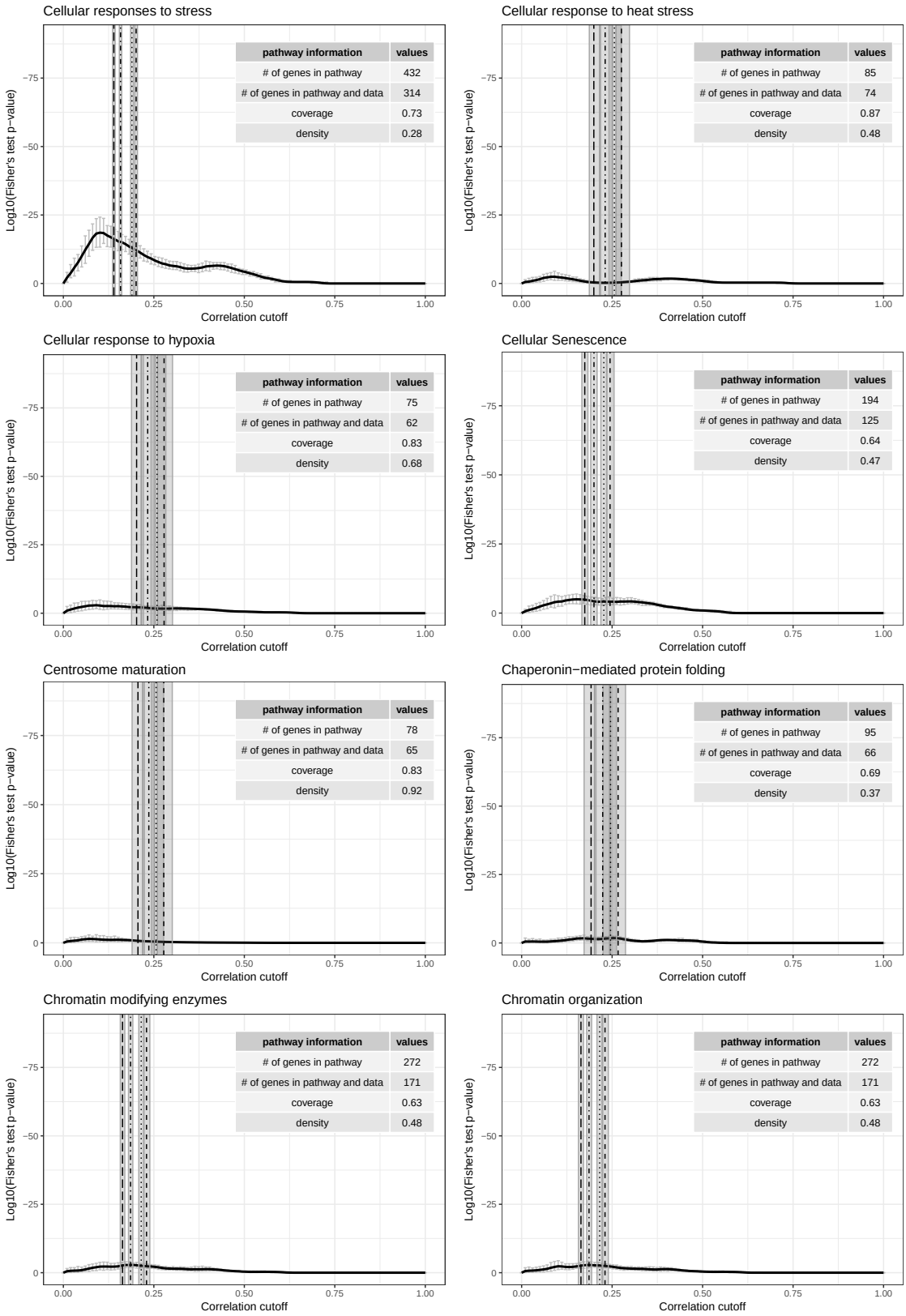

Antiviral mechanism by IFN−stimulated genes
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 64
coverage | 0.94
density | 0.45
APC/C:Cdc20 mediated degradation of mitotic proteins
pathway information | values
# of genes in pathway | 75
# of genes in pathway and data | 68
coverage | 0.91
density | 0.95
APC/C:Cdc20 mediated degradation of Securin
pathway information | values
# of genes in pathway | 67
# of genes in pathway and data | 61
coverage | 0.91
density | 0.98
APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh
pathway information | values
# of genes in pathway | 72
# of genes in pathway and data | 66
coverage | 0.92
density | 0.98
APC:Cdc20 mediated degradation of cell cycle proteins prior to satis
pathway information | values
# of genes in pathway | 73
# of genes in pathway and data | 66
coverage | 0.9
density | 0.96
APC/C−mediated degradation of cell cycle proteins
pathway information | values
# of genes in pathway | 86
# of genes in pathway and data | 79
coverage | 0.92
density | 0.94
Apoptosis
pathway information | values
# of genes in pathway | 159
# of genes in pathway and data | 135
coverage | 0.85
density | 0.36
Arachidonic acid metabolism
pathway information | values
# of genes in pathway | 52
# of genes in pathway and data | 23
coverage | 0.44
density | 0.5
Log10(Fisher's test p−value)
Correlation cutoff

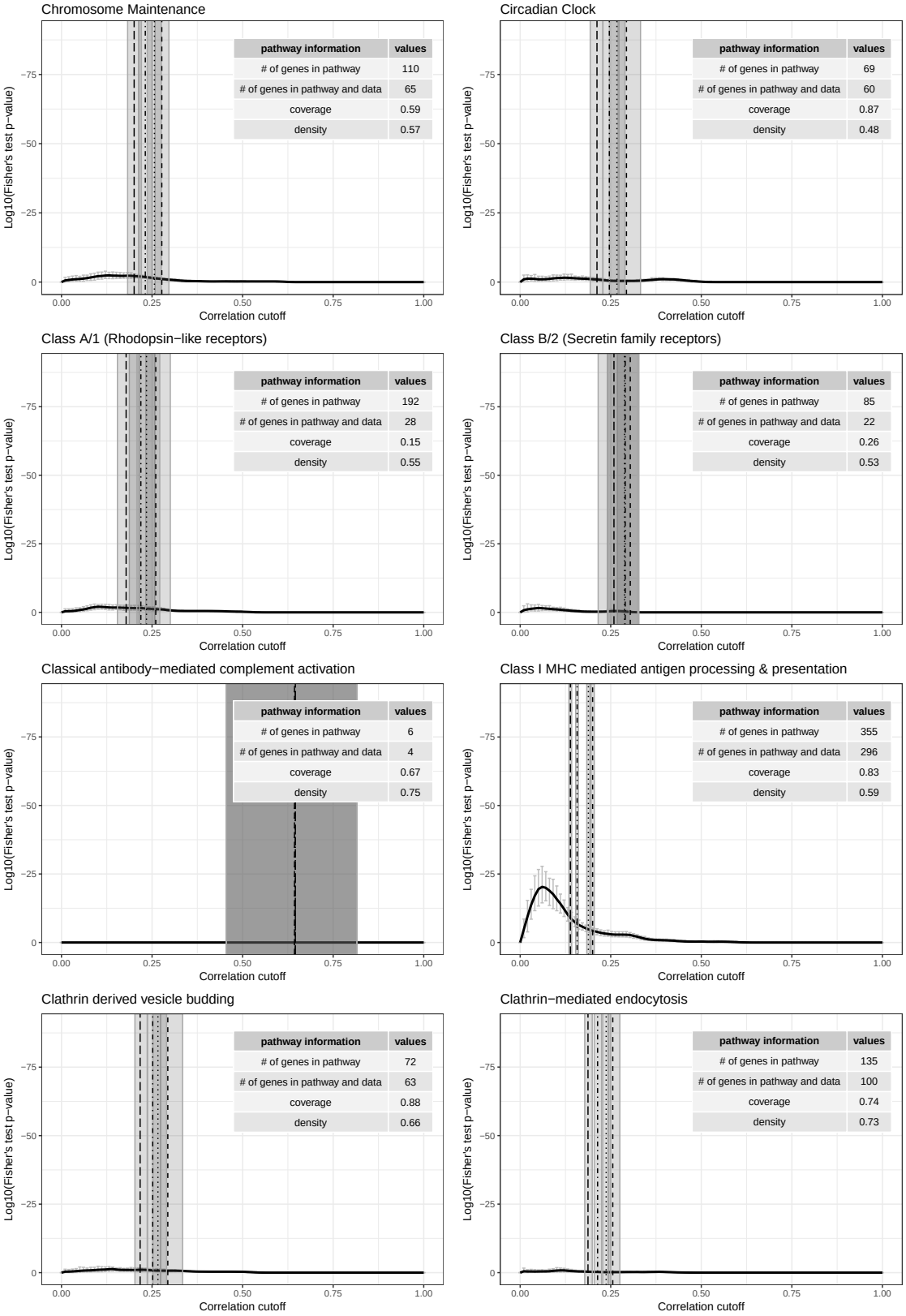

ARMS−mediated activation
pathway information | values
# of genes in pathway | 231
# of genes in pathway and data | 153
coverage | 0.66
density | 0.38
Asparagine N−linked glycosylation
pathway information | values
# of genes in pathway | 267
# of genes in pathway and data | 216
coverage | 0.81
density | 0.26
Assembly of the pre−replicative complex
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 57
coverage | 0.84
density | 0.75
Assembly of the primary cilium
pathway information | values
# of genes in pathway | 186
# of genes in pathway and data | 150
coverage | 0.81
density | 0.46
Asymmetric localization of PCP proteins
pathway information | values
# of genes in pathway | 64
# of genes in pathway and data | 56
coverage | 0.88
density | 0.75
AURKA Activation by TPX2
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 59
coverage | 0.83
density | 0.98
Autodegradation of Cdh1 by Cdh1:APC/C
pathway information | values
# of genes in pathway | 63
# of genes in pathway and data | 59
coverage | 0.94
density | 0.98
Autodegradation of the E3 ubiquitin ligase COP1
pathway information | values
# of genes in pathway | 52
# of genes in pathway and data | 46
coverage | 0.88
density | 0.95
Log10(Fisher's test p−value)
Correlation cutoff


Axon guidance
pathway information | values
# of genes in pathway | 529
# of genes in pathway and data | 335
coverage | 0.63
density | 0.28
B Cell Activation
pathway information | values
# of genes in pathway | 207
# of genes in pathway and data | 158
coverage | 0.76
density | 0.44
Beta−catenin independent WNT signaling
pathway information | values
# of genes in pathway | 140
# of genes in pathway and data | 105
coverage | 0.75
density | 0.4
Binding and Uptake of Ligands by Scavenger Receptors
pathway information | values
# of genes in pathway | 38
# of genes in pathway and data | 16
coverage | 0.42
density | 0.46
Biological oxidations
pathway information | values
# of genes in pathway | 174
# of genes in pathway and data | 67
coverage | 0.39
density | 0.22
Biosynthesis of the N−glycan precursor (dolichol lipid−linked oligosa
pathway information | values
# of genes in pathway | 77
# of genes in pathway and data | 63
coverage | 0.82
density | 0.3
B−WICH complex positively regulates rRNA expression
pathway information | values
# of genes in pathway | 90
# of genes in pathway and data | 41
coverage | 0.46
density | 0.87
Ca2+ pathway
pathway information | values
# of genes in pathway | 55
# of genes in pathway and data | 34
coverage | 0.62
density | 0.43
Log10(Fisher's test p−value)
Correlation cutoff

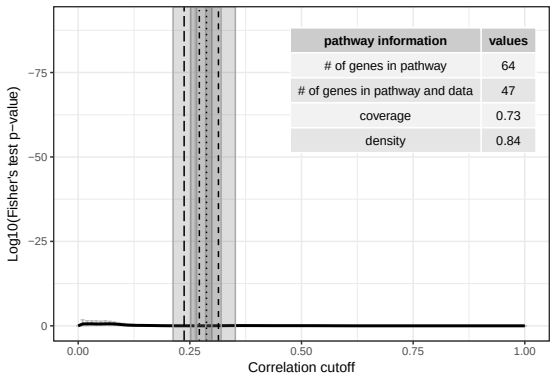

Cap−dependent Translation Initiation
pathway information | values
# of genes in pathway | 120
# of genes in pathway and data | 112
coverage | 0.93
density | 0.9
Carbohydrate metabolism
pathway information | values
# of genes in pathway | 277
# of genes in pathway and data | 191
coverage | 0.69
density | 0.18
Cardiac conduction
pathway information | values
# of genes in pathway | 139
# of genes in pathway and data | 48
coverage | 0.35
density | 0.36
Cargo recognition for clathrin−mediated endocytosis
pathway information | values
# of genes in pathway | 98
# of genes in pathway and data | 67
coverage | 0.68
density | 0.67
Cargo trafficking to the periciliary membrane
pathway information | values
# of genes in pathway | 51
# of genes in pathway and data | 42
coverage | 0.82
density | 0.48
CD22 mediated BCR regulation
pathway information | values
# of genes in pathway | 5
# of genes in pathway and data | 5
coverage | 1
density | 0.8
Cdc20:Phospho−APC/C mediated degradation of Cyclin A
pathway information | values
# of genes in pathway | 72
# of genes in pathway and data | 65
coverage | 0.9
density | 0.98
CDK−mediated phosphorylation and removal of Cdc6
pathway information | values
# of genes in pathway | 51
# of genes in pathway and data | 46
coverage | 0.9
density | 0.98
Log10(Fisher's test p−value)
Correlation cutoff


CDT1 association with the CDC6:ORC:origin complex
pathway information | values
# of genes in pathway | 59
# of genes in pathway and data | 48
coverage | 0.81
density | 0.91
Cell−Cell communication
pathway information | values
# of genes in pathway | 91
# of genes in pathway and data | 56
coverage | 0.62
density | 0.4
Cell Cycle
pathway information | values
# of genes in pathway | 599
# of genes in pathway and data | 466
coverage | 0.78
density | 0.34
Cell Cycle Checkpoints
pathway information | values
# of genes in pathway | 200
# of genes in pathway and data | 148
coverage | 0.74
density | 0.58
Cell Cycle, Mitotic
pathway information | values
# of genes in pathway | 494
# of genes in pathway and data | 391
coverage | 0.79
density | 0.36
Cell death signalling via NRAGE, NRIF and NADE
pathway information | values
# of genes in pathway | 75
# of genes in pathway and data | 57
coverage | 0.76
density | 0.23
Cell junction organization
pathway information | values
# of genes in pathway | 61
# of genes in pathway and data | 30
coverage | 0.49
density | 0.39
Cell surface interactions at the vascular wall
pathway information | values
# of genes in pathway | 97
# of genes in pathway and data | 66
coverage | 0.68
density | 0.41
Log10(Fisher's test p−value)
Correlation cutoff

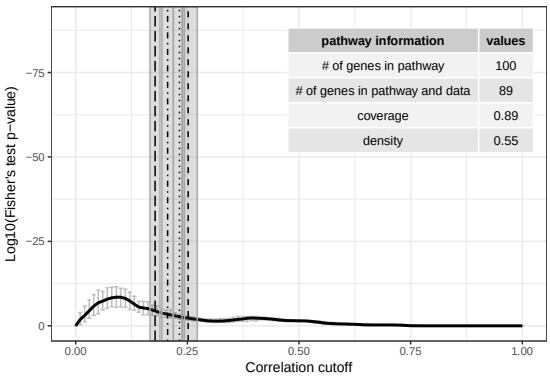

Cellular responses to stress
pathway information | values
# of genes in pathway | 432
# of genes in pathway and data | 314
coverage | 0.73
density | 0.28
Cellular response to heat stress
pathway information | values
# of genes in pathway | 85
# of genes in pathway and data | 74
coverage | 0.87
density | 0.48
Cellular response to hypoxia
pathway information | values
# of genes in pathway | 75
# of genes in pathway and data | 62
coverage | 0.83
density | 0.68
Cellular Senescence
pathway information | values
# of genes in pathway | 194
# of genes in pathway and data | 125
coverage | 0.64
density | 0.47
Centrosome maturation
pathway information | values
# of genes in pathway | 78
# of genes in pathway and data | 65
coverage | 0.83
density | 0.92
Chaperonin−mediated protein folding
pathway information | values
# of genes in pathway | 95
# of genes in pathway and data | 66
coverage | 0.69
density | 0.37
Chromatin modifying enzymes
pathway information | values
# of genes in pathway | 272
# of genes in pathway and data | 171
coverage | 0.63
density | 0.48
Chromatin organization
pathway information | values
# of genes in pathway | 272
# of genes in pathway and data | 171
coverage | 0.63
density | 0.48
Log10(Fisher's test p−value)
Correlation cutoff


Chromosome Maintenance
pathway information | values
# of genes in pathway | 110
# of genes in pathway and data | 65
coverage | 0.59
density | 0.57
Circadian Clock
pathway information | values
# of genes in pathway | 69
# of genes in pathway and data | 60
coverage | 0.87
density | 0.48
Class A/1 (Rhodopsin−like receptors)
pathway information | values
# of genes in pathway | 192
# of genes in pathway and data | 28
coverage | 0.15
density | 0.55
Class B/2 (Secretin family receptors)
pathway information | values
# of genes in pathway | 85
# of genes in pathway and data | 22
coverage | 0.26
density | 0.53
Classical antibody−mediated complement activation
pathway information | values
# of genes in pathway | 6
# of genes in pathway and data | 4
coverage | 0.67
density | 0.75
Class I MHC mediated antigen processing & presentation
pathway information | values
# of genes in pathway | 355
# of genes in pathway and data | 296
coverage | 0.83
density | 0.59
Clathrin derived vesicle budding
pathway information | values
# of genes in pathway | 72
# of genes in pathway and data | 63
coverage | 0.88
density | 0.66
Clathrin−mediated endocytosis
pathway information | values
# of genes in pathway | 135
# of genes in pathway and data | 100
coverage | 0.74
density | 0.73
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

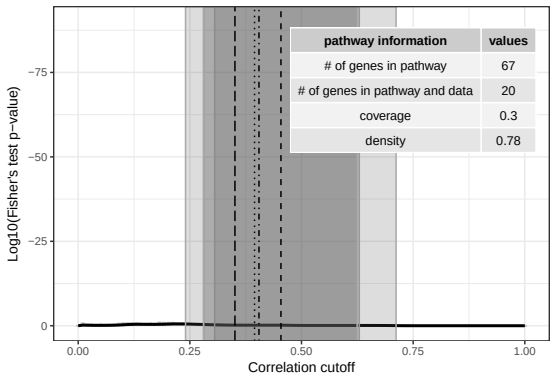
Cleavage of Growing Transcript in the Termination Region



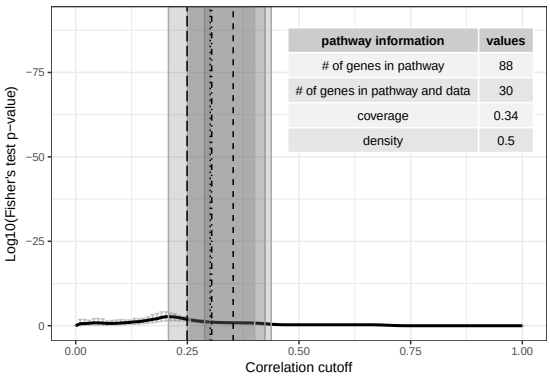
CLEC7A (Dectin−1) signaling



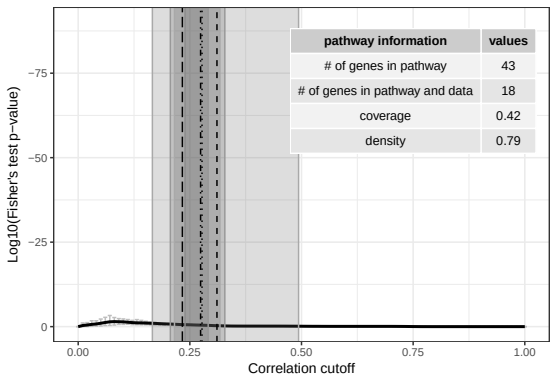
Collagen biosynthesis and modifying enzymes



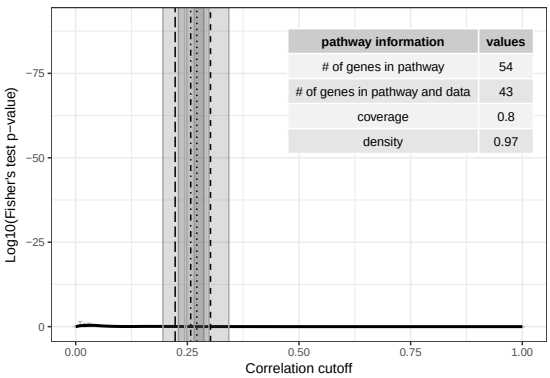
Collagen formation



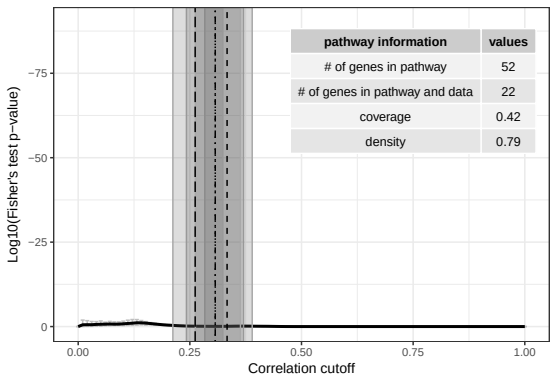
Complement cascade



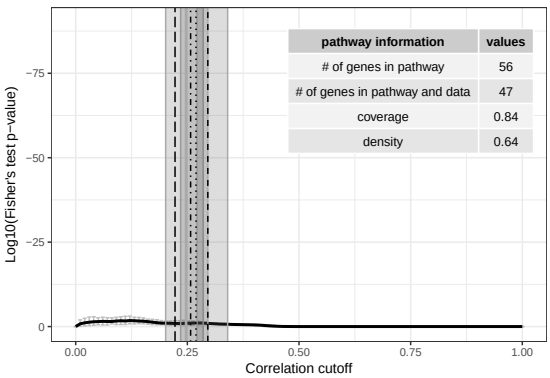
Complex I biogenesis

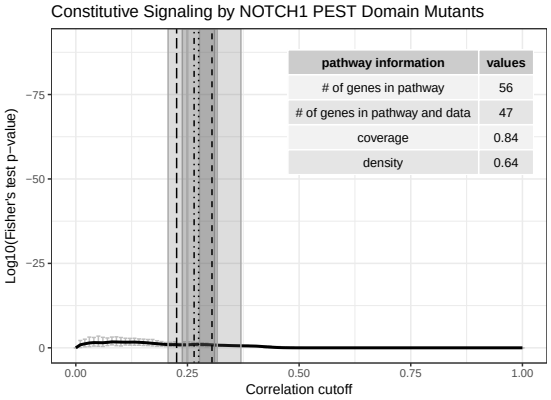


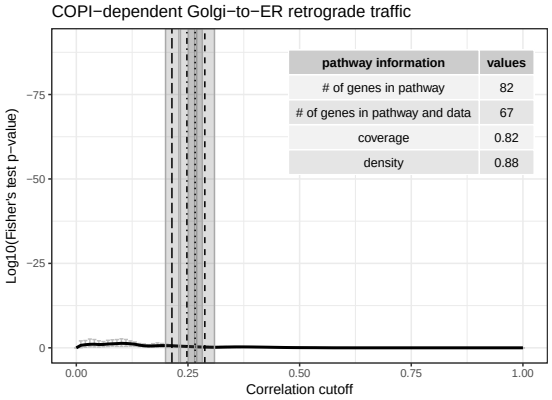
Constitutive Signaling by Aberrant PI3K in Cancer

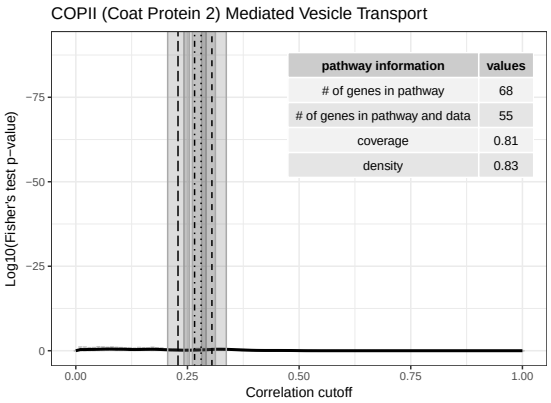


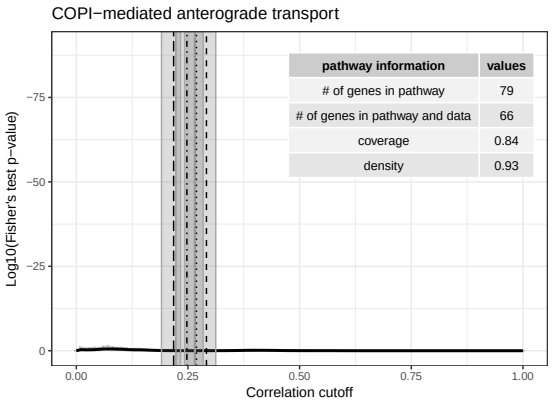
Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants

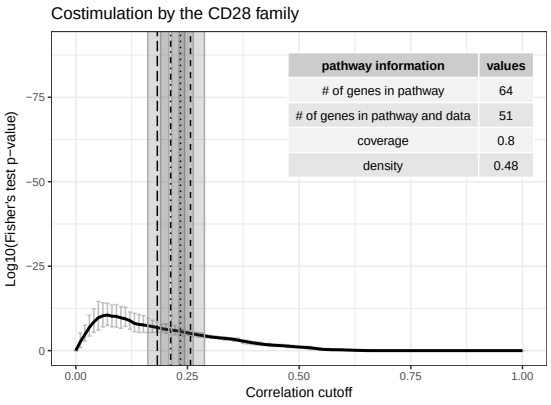


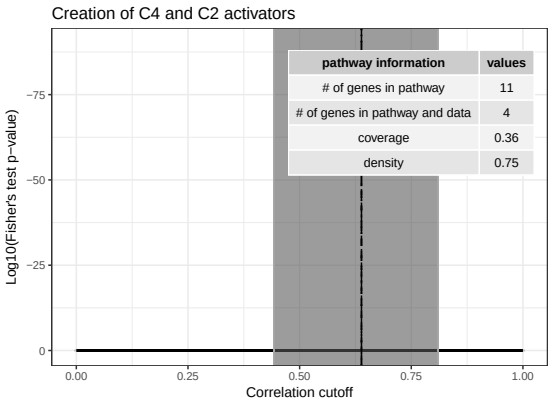

Constitutive Signaling by NOTCH1 PEST Domain Mutants
pathway information | values
# of genes in pathway | 56
# of genes in pathway and data | 47
coverage | 0.84
density | 0.64
Log10(Fisher's test p−value)
Correlation cutoff

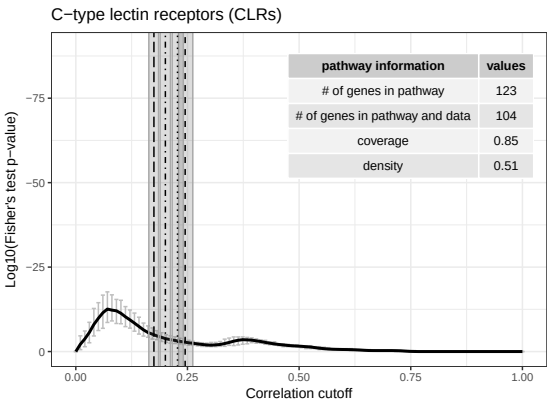

COPI−dependent Golgi−to−ER retrograde traffic
pathway information | values
# of genes in pathway | 82
# of genes in pathway and data | 67
coverage | 0.82
density | 0.88
Log10(Fisher's test p−value)
Correlation cutoff

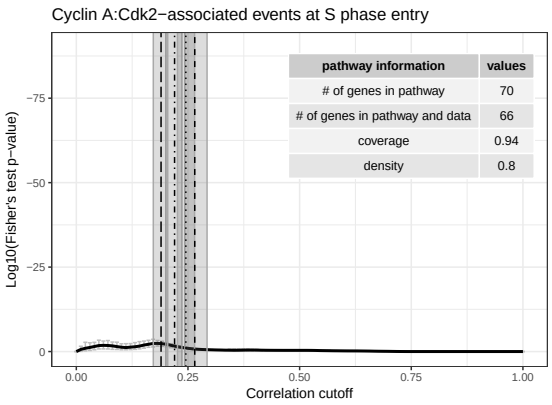

COPII (Coat Protein 2) Mediated Vesicle Transport
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 55
coverage | 0.81
density | 0.83
Log10(Fisher's test p−value)
Correlation cutoff

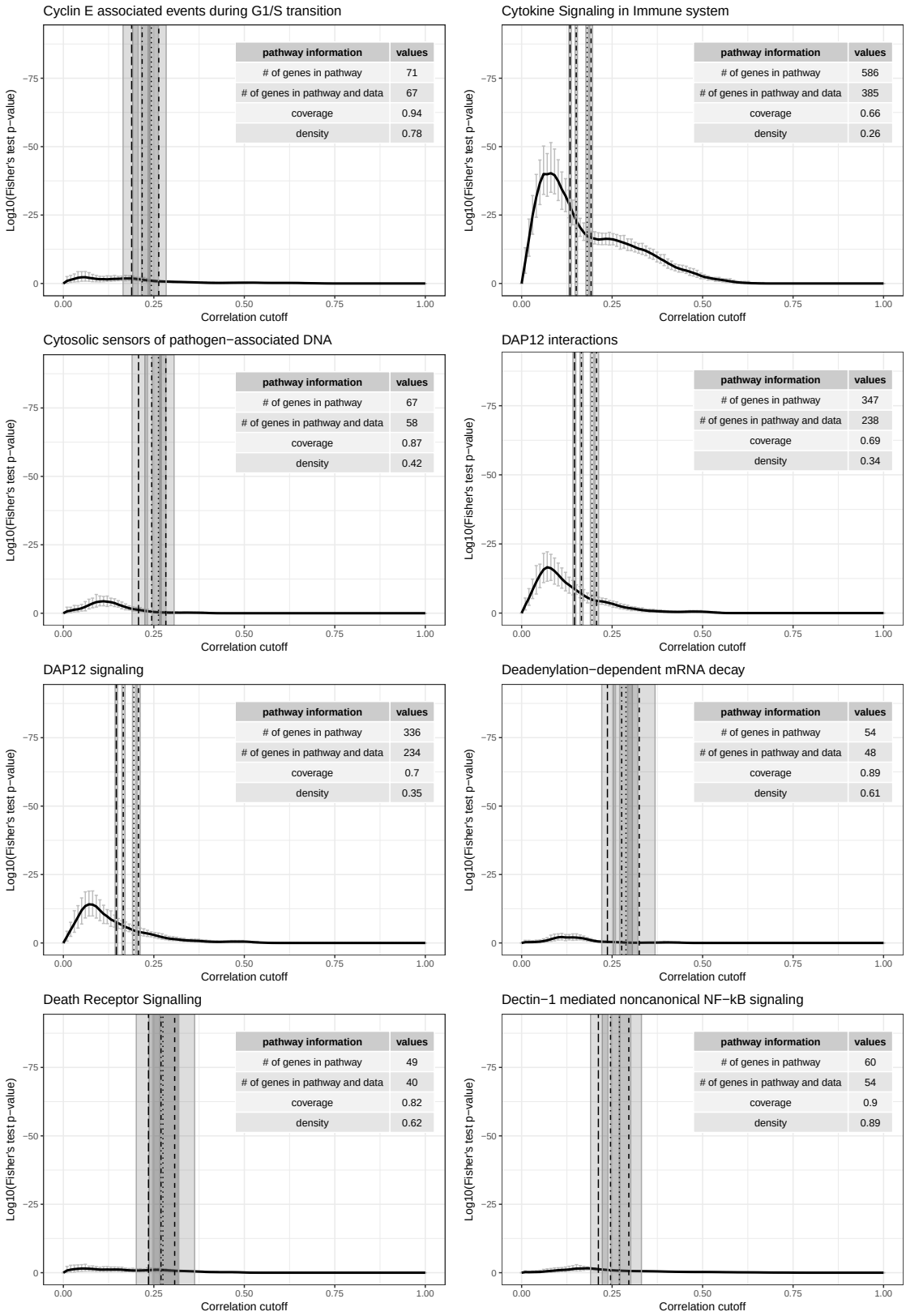

COPI−mediated anterograde transport
pathway information | values
# of genes in pathway | 79
# of genes in pathway and data | 66
coverage | 0.84
density | 0.93
Log10(Fisher's test p−value)
Correlation cutoff

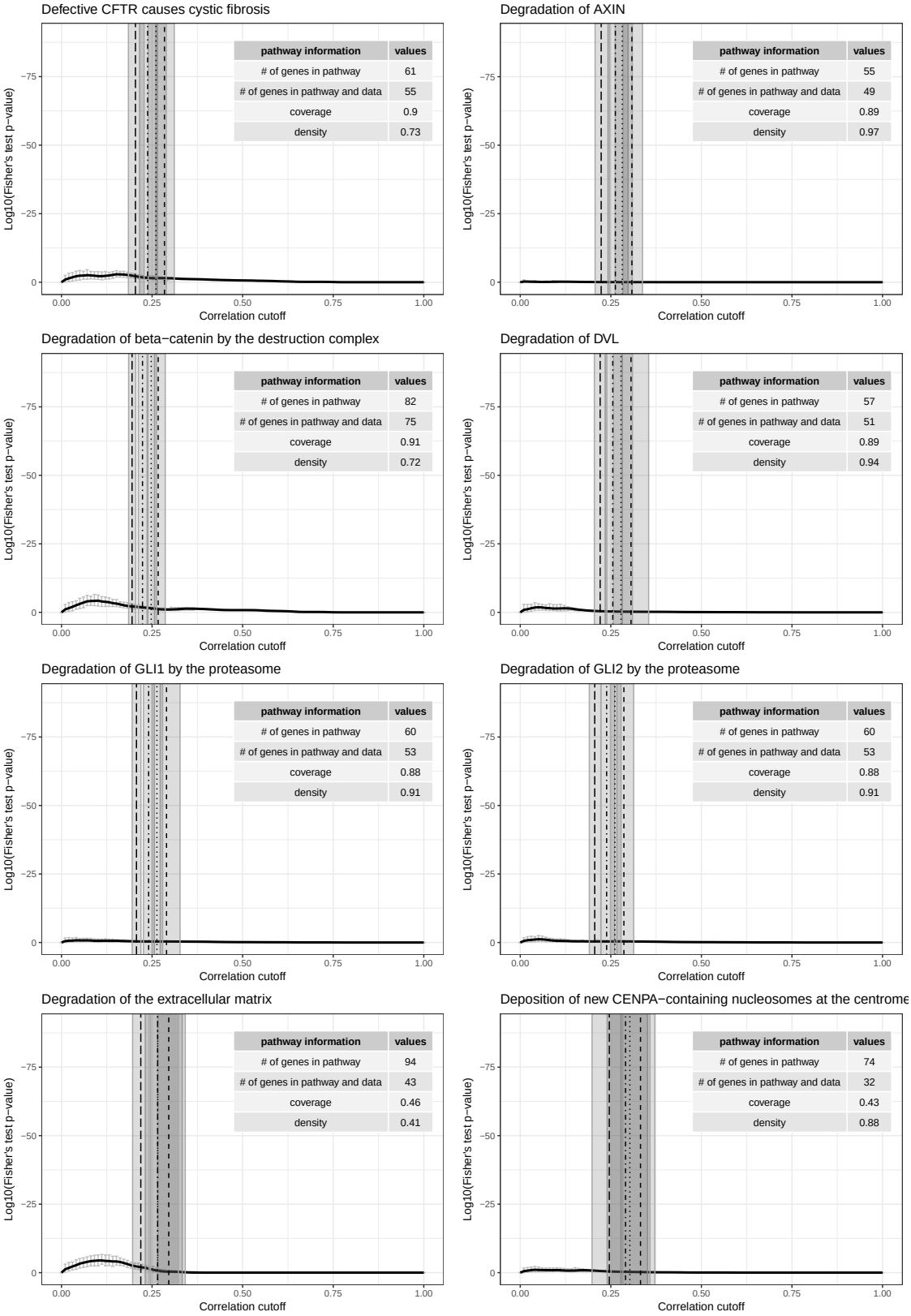

Costimulation by the CD28 family
pathway information | values
# of genes in pathway | 64
# of genes in pathway and data | 51
coverage | 0.8
density | 0.48
Log10(Fisher's test p−value)
Correlation cutoff

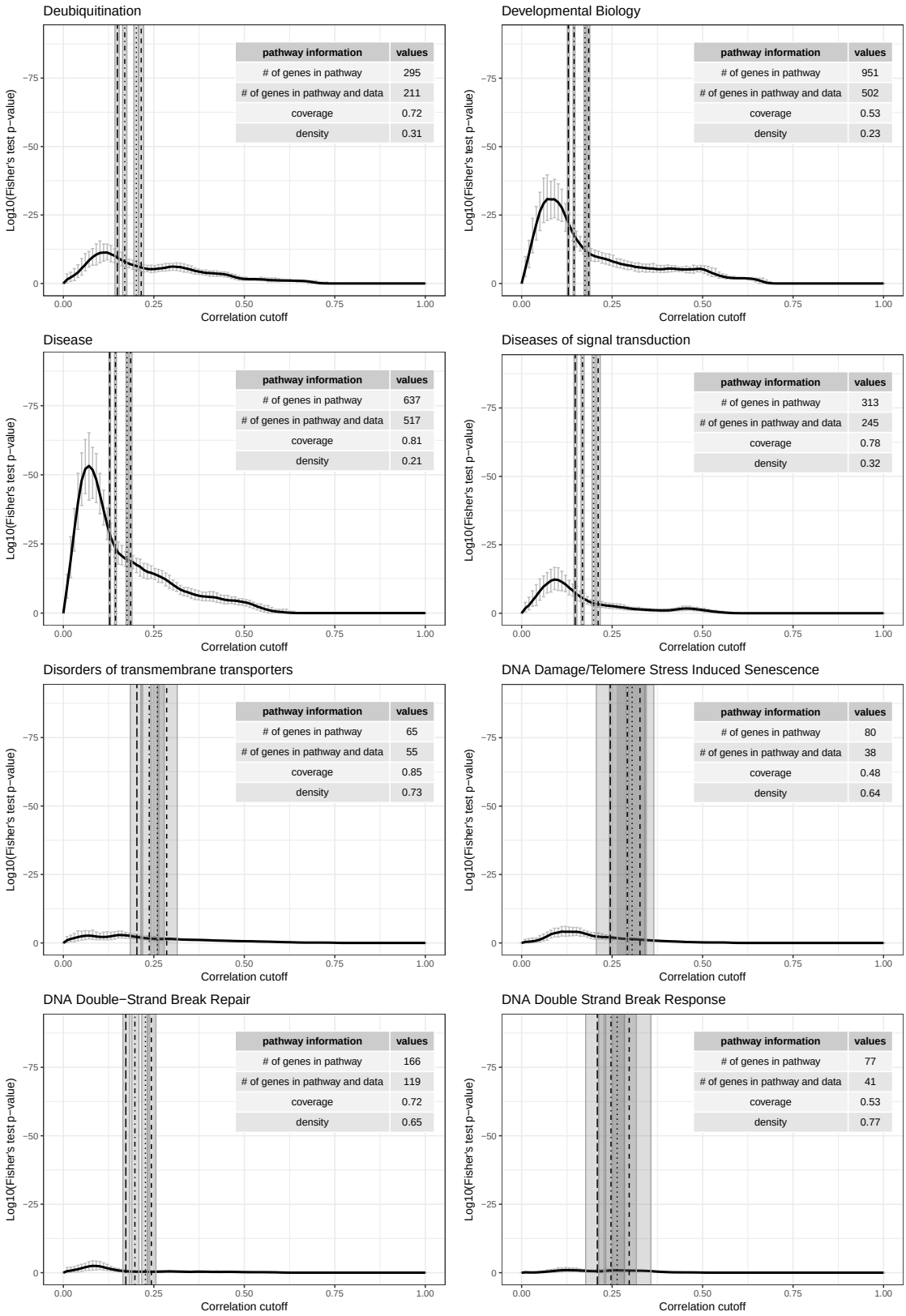

Creation of C4 and C2 activators
pathway information | values
# of genes in pathway | 11
# of genes in pathway and data | 4
coverage | 0.36
density | 0.75
Log10(Fisher's test p−value)
Correlation cutoff

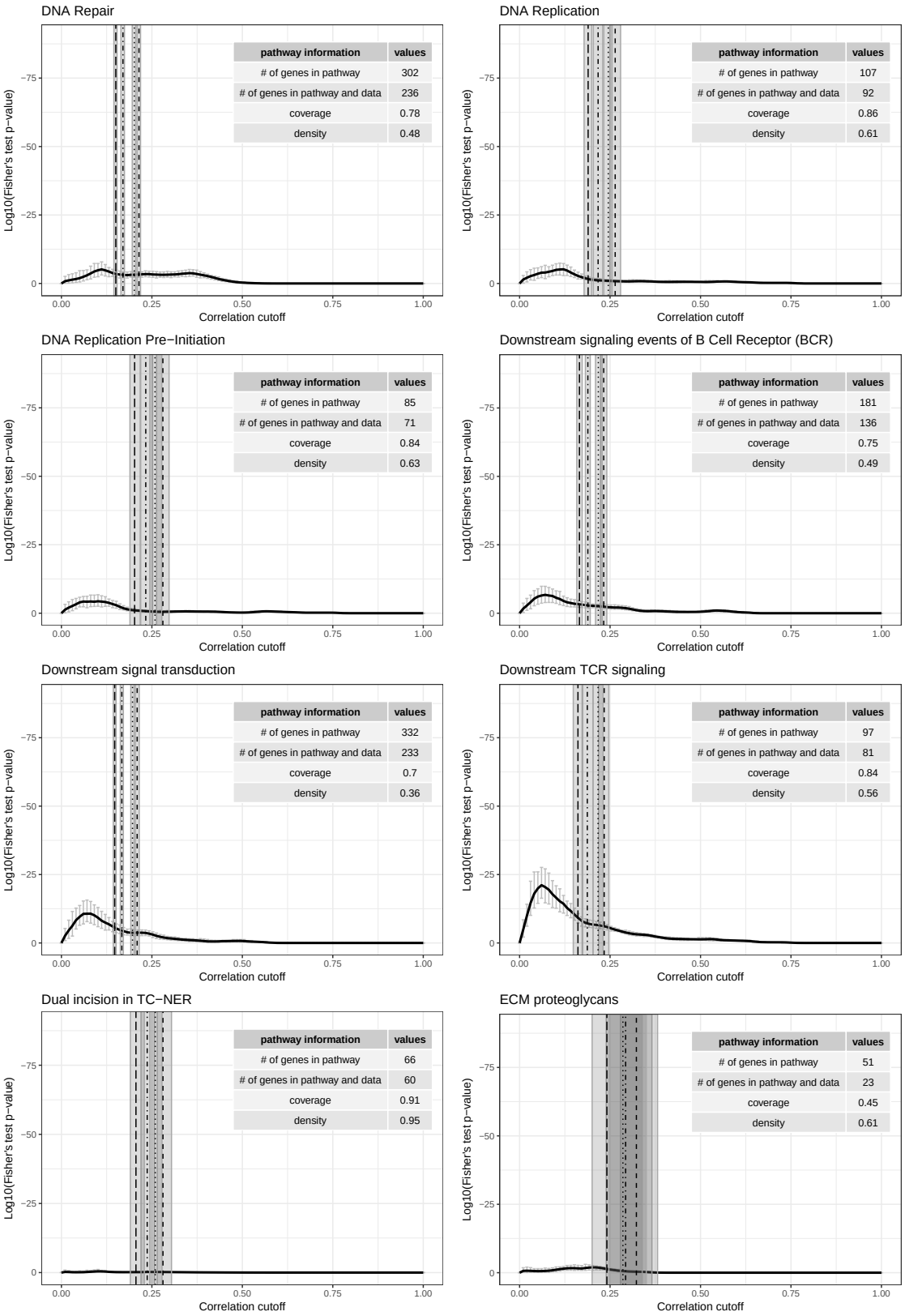

C−type lectin receptors (CLRs)
pathway information | values
# of genes in pathway | 123
# of genes in pathway and data | 104
coverage | 0.85
density | 0.51
Log10(Fisher's test p−value)
Correlation cutoff

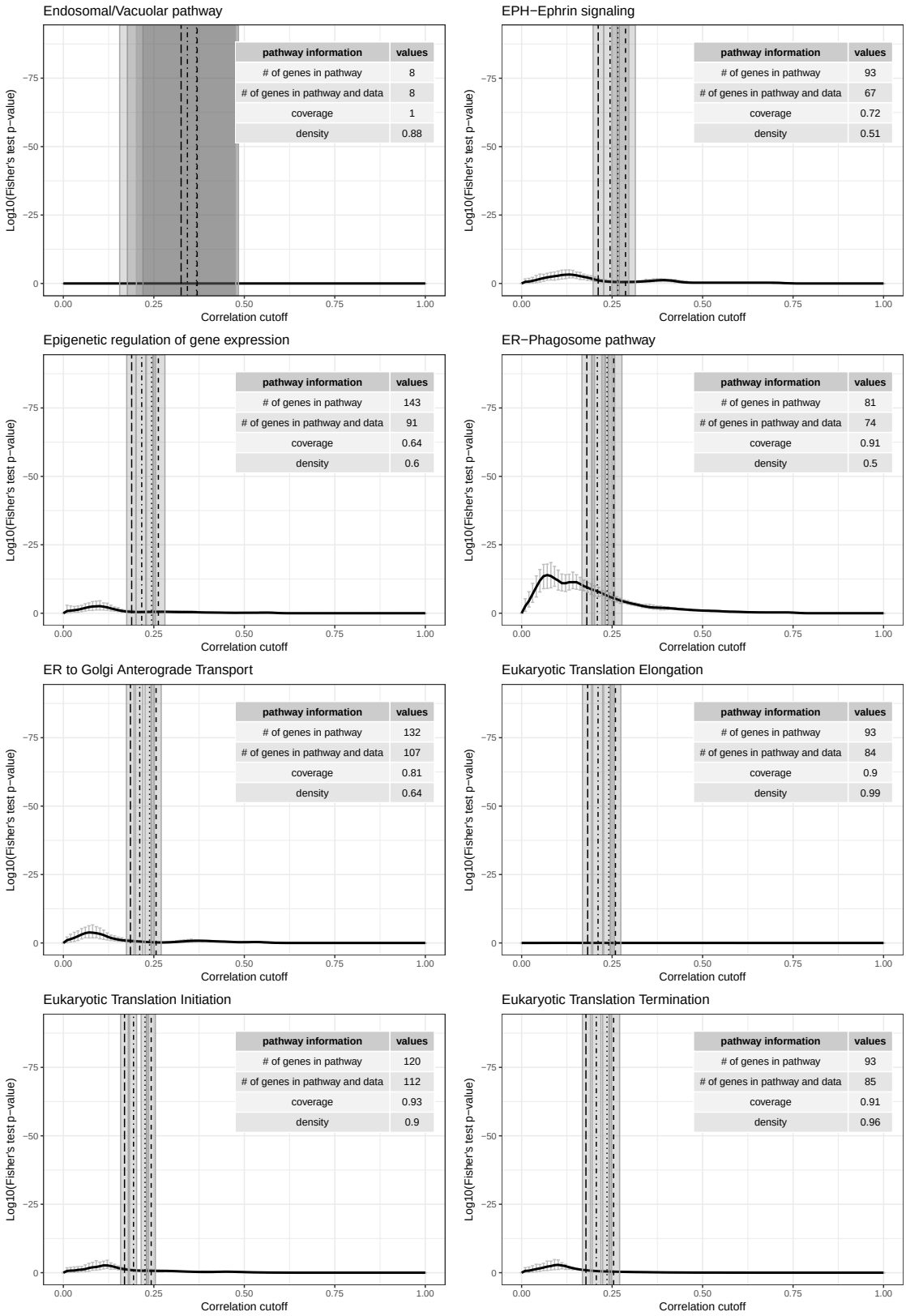

Cyclin A:Cdk2−associated events at S phase entry
pathway information | values
# of genes in pathway | 70
# of genes in pathway and data | 66
coverage | 0.94
density | 0.8
Log10(Fisher's test p−value)
Correlation cutoff

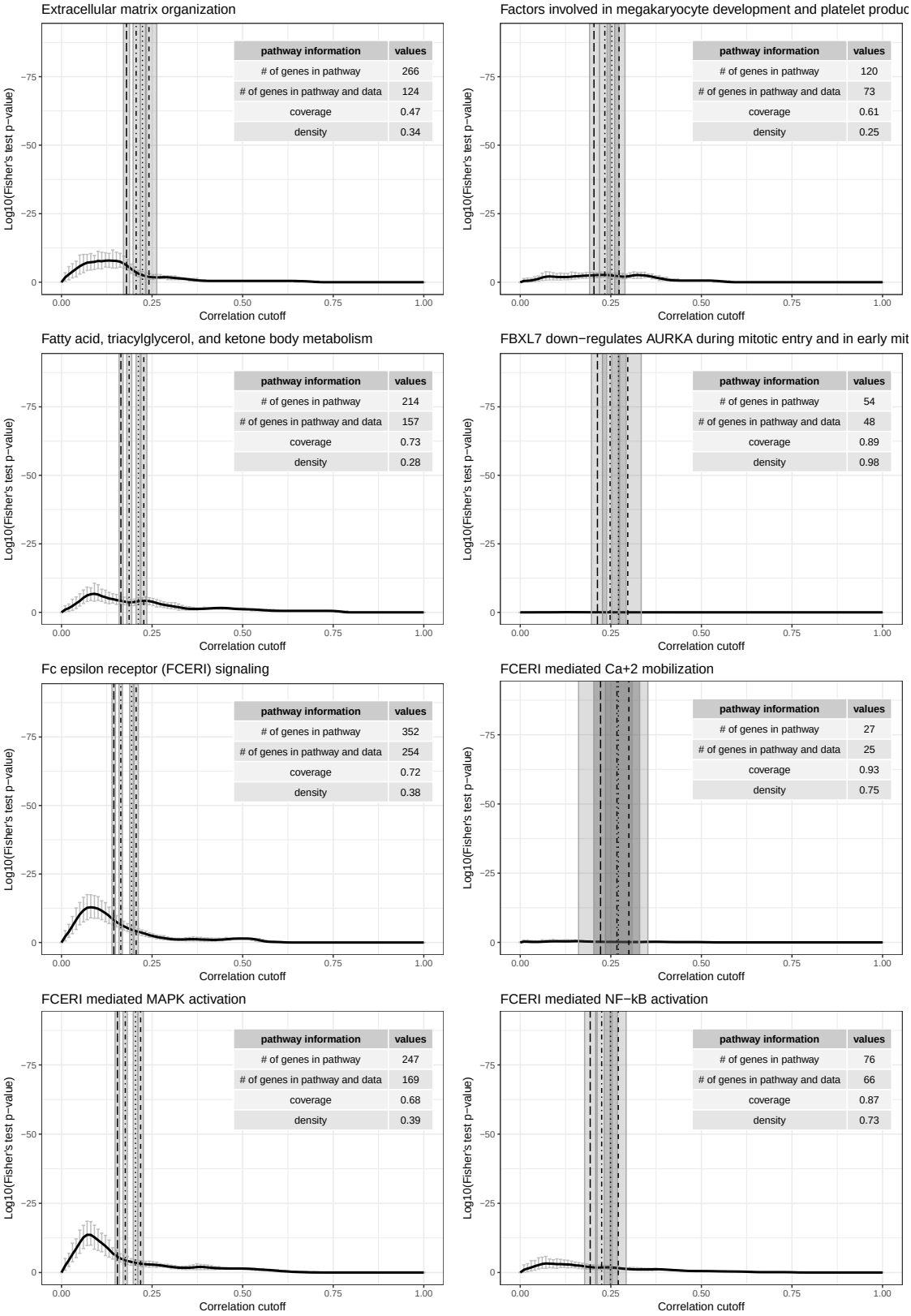

Cyclin E associated events during G1/S transition
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 67
coverage | 0.94
density | 0.78
Cytokine Signaling in Immune system
pathway information | values
# of genes in pathway | 586
# of genes in pathway and data | 385
coverage | 0.66
density | 0.26
Cytosolic sensors of pathogen−associated DNA
pathway information | values
# of genes in pathway | 67
# of genes in pathway and data | 58
coverage | 0.87
density | 0.42
DAP12 interactions
pathway information | values
# of genes in pathway | 347
# of genes in pathway and data | 238
coverage | 0.69
density | 0.34
DAP12 signaling
pathway information | values
# of genes in pathway | 336
# of genes in pathway and data | 234
coverage | 0.7
density | 0.35
Deadenylation−dependent mRNA decay
pathway information | values
# of genes in pathway | 54
# of genes in pathway and data | 48
coverage | 0.89
density | 0.61
Death Receptor Signalling
pathway information | values
# of genes in pathway | 49
# of genes in pathway and data | 40
coverage | 0.82
density | 0.62
Dectin−1 mediated noncanonical NF−kB signaling
pathway information | values
# of genes in pathway | 60
# of genes in pathway and data | 54
coverage | 0.9
density | 0.89
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

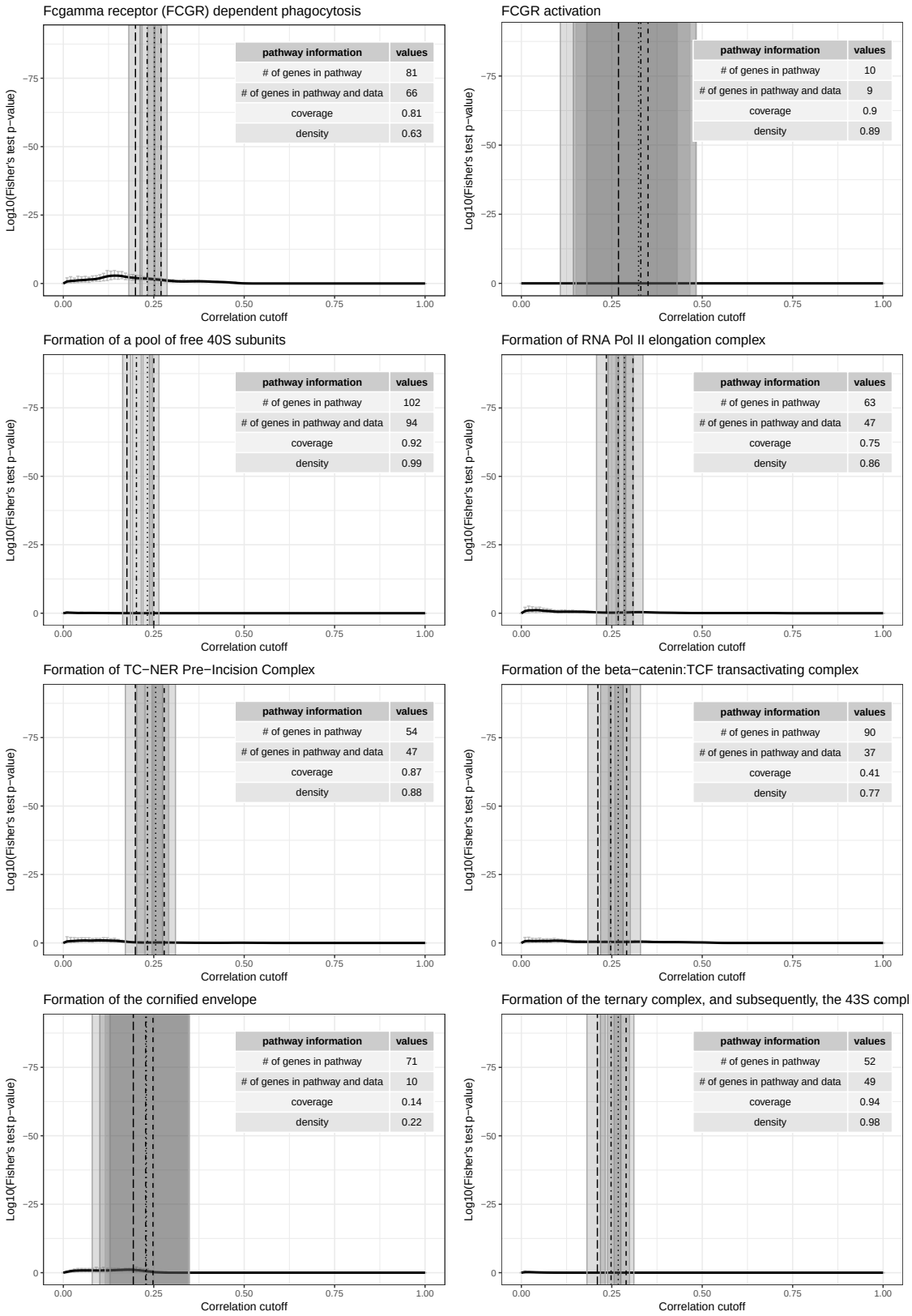

Defective CFTR causes cystic fibrosis
pathway information | values
# of genes in pathway | 61
# of genes in pathway and data | 55
coverage | 0.9
density | 0.73
Degradation of AXIN
pathway information | values
# of genes in pathway | 55
# of genes in pathway and data | 49
coverage | 0.89
density | 0.97
Degradation of beta−catenin by the destruction complex
pathway information | values
# of genes in pathway | 82
# of genes in pathway and data | 75
coverage | 0.91
density | 0.72
Degradation of DVL
pathway information | values
# of genes in pathway | 57
# of genes in pathway and data | 51
coverage | 0.89
density | 0.94
Degradation of GLI1 by the proteasome
pathway information | values
# of genes in pathway | 60
# of genes in pathway and data | 53
coverage | 0.88
density | 0.91
Degradation of GLI2 by the proteasome
pathway information | values
# of genes in pathway | 60
# of genes in pathway and data | 53
coverage | 0.88
density | 0.91
Degradation of the extracellular matrix
pathway information | values
# of genes in pathway | 94
# of genes in pathway and data | 43
coverage | 0.46
density | 0.41
Deposition of new CENPA−containing nucleosomes at the centrome
pathway information | values
# of genes in pathway | 74
# of genes in pathway and data | 32
coverage | 0.43
density | 0.88
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

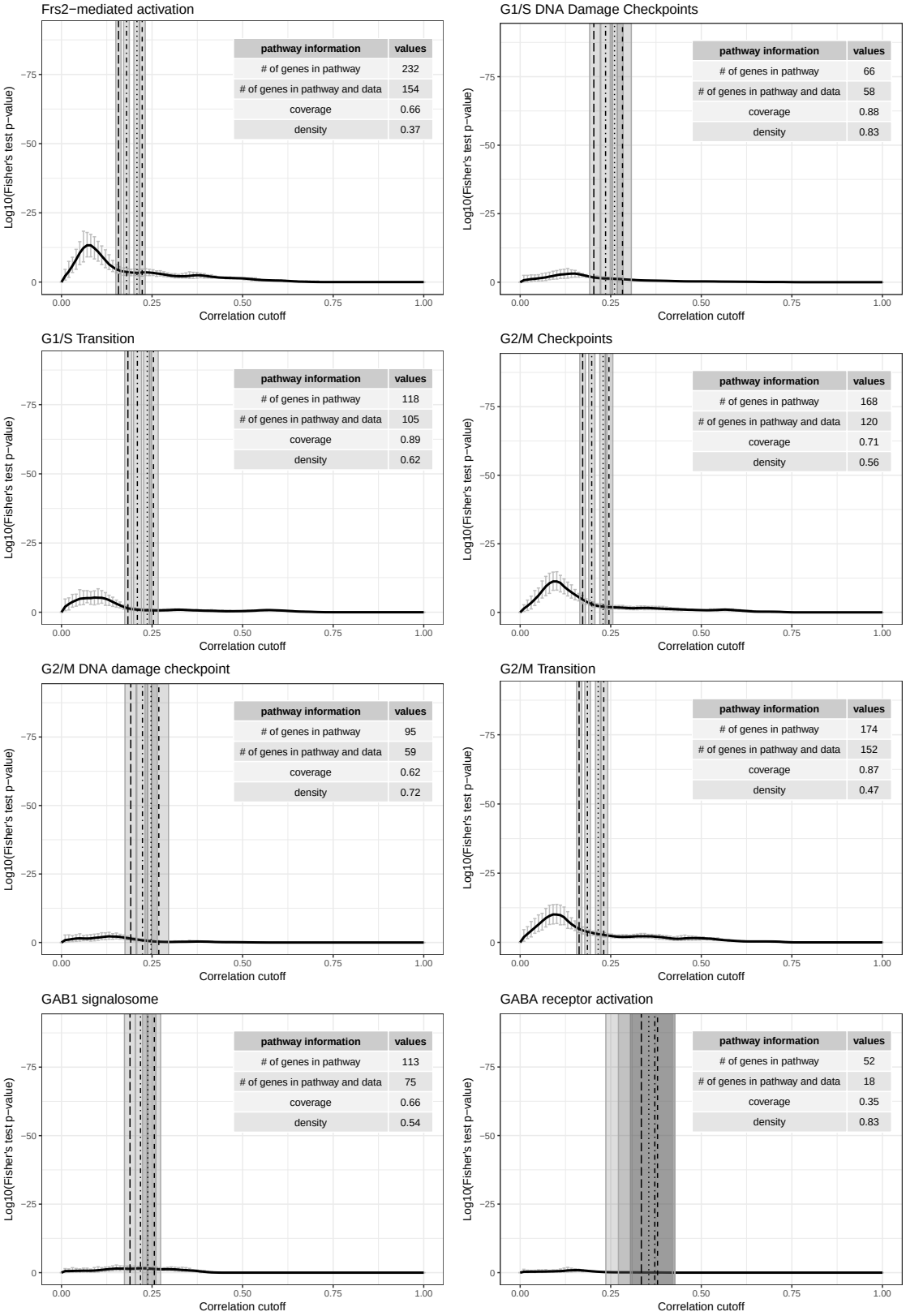

Deubiquitination
pathway information | values
# of genes in pathway | 295
# of genes in pathway and data | 211
coverage | 0.72
density | 0.31
Developmental Biology
pathway information | values
# of genes in pathway | 951
# of genes in pathway and data | 502
coverage | 0.53
density | 0.23
Disease
pathway information | values
# of genes in pathway | 637
# of genes in pathway and data | 517
coverage | 0.81
density | 0.21
Diseases of signal transduction
pathway information | values
# of genes in pathway | 313
# of genes in pathway and data | 245
coverage | 0.78
density | 0.32
Disorders of transmembrane transporters
pathway information | values
# of genes in pathway | 65
# of genes in pathway and data | 55
coverage | 0.85
density | 0.73
DNA Damage/Telomere Stress Induced Senescence
pathway information | values
# of genes in pathway | 80
# of genes in pathway and data | 38
coverage | 0.48
density | 0.64
DNA Double−Strand Break Repair
pathway information | values
# of genes in pathway | 166
# of genes in pathway and data | 119
coverage | 0.72
density | 0.65
DNA Double Strand Break Response
pathway information | values
# of genes in pathway | 77
# of genes in pathway and data | 41
coverage | 0.53
density | 0.77
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

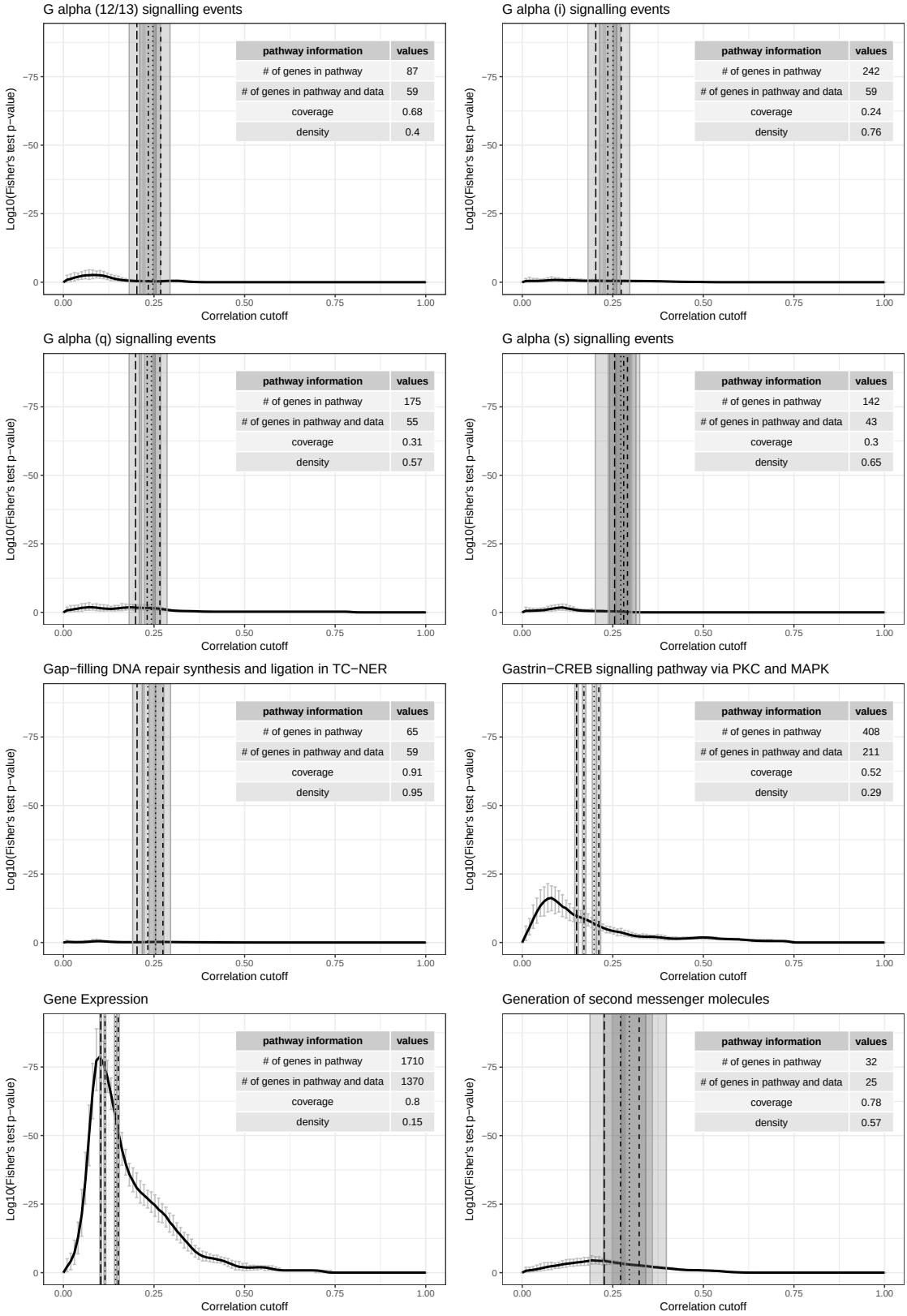

DNA Repair
pathway information | values
# of genes in pathway | 302
# of genes in pathway and data | 236
coverage | 0.78
density | 0.48
DNA Replication
pathway information | values
# of genes in pathway | 107
# of genes in pathway and data | 92
coverage | 0.86
density | 0.61
DNA Replication Pre−Initiation
pathway information | values
# of genes in pathway | 85
# of genes in pathway and data | 71
coverage | 0.84
density | 0.63
Downstream signaling events of B Cell Receptor (BCR)
pathway information | values
# of genes in pathway | 181
# of genes in pathway and data | 136
coverage | 0.75
density | 0.49
Downstream signal transduction
pathway information | values
# of genes in pathway | 332
# of genes in pathway and data | 233
coverage | 0.7
density | 0.36
Downstream TCR signaling
pathway information | values
# of genes in pathway | 97
# of genes in pathway and data | 81
coverage | 0.84
density | 0.56
Dual incision in TC−NER
pathway information | values
# of genes in pathway | 66
# of genes in pathway and data | 60
coverage | 0.91
density | 0.95
ECM proteoglycans
pathway information | values
# of genes in pathway | 51
# of genes in pathway and data | 23
coverage | 0.45
density | 0.61
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

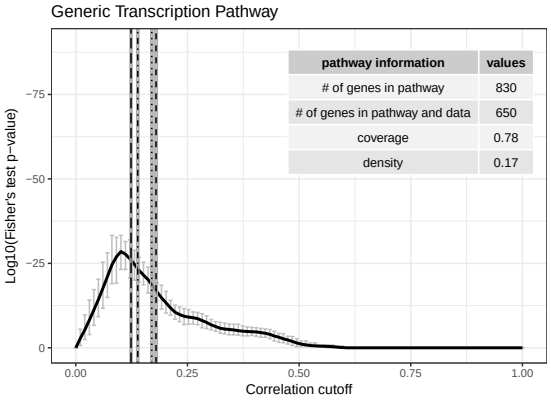

Endosomal/Vacuolar pathway
pathway information | values
# of genes in pathway | 8
# of genes in pathway and data | 8
coverage | 1
density | 0.88
EPH−Ephrin signaling
pathway information | values
# of genes in pathway | 93
# of genes in pathway and data | 67
coverage | 0.72
density | 0.51
Epigenetic regulation of gene expression
pathway information | values
# of genes in pathway | 143
# of genes in pathway and data | 91
coverage | 0.64
density | 0.6
ER−Phagosome pathway
pathway information | values
# of genes in pathway | 81
# of genes in pathway and data | 74
coverage | 0.91
density | 0.5
ER to Golgi Anterograde Transport
pathway information | values
# of genes in pathway | 132
# of genes in pathway and data | 107
coverage | 0.81
density | 0.64
Eukaryotic Translation Elongation
pathway information | values
# of genes in pathway | 93
# of genes in pathway and data | 84
coverage | 0.9
density | 0.99
Eukaryotic Translation Initiation
pathway information | values
# of genes in pathway | 120
# of genes in pathway and data | 112
coverage | 0.93
density | 0.9
Eukaryotic Translation Termination
pathway information | values
# of genes in pathway | 93
# of genes in pathway and data | 85
coverage | 0.91
density | 0.96
Log10(Fisher's test p−value)
Correlation cutoff

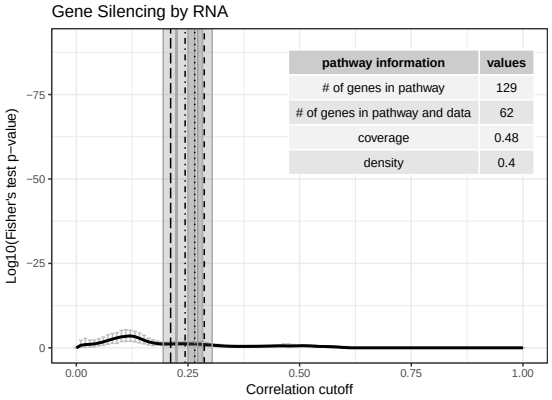

Extracellular matrix organization
pathway information | values
# of genes in pathway | 266
# of genes in pathway and data | 124
coverage | 0.47
density | 0.34
Factors involved in megakaryocyte development and platelet produc
pathway information | values
# of genes in pathway | 120
# of genes in pathway and data | 73
coverage | 0.61
density | 0.25
Fatty acid, triacylglycerol, and ketone body metabolism
pathway information | values
# of genes in pathway | 214
# of genes in pathway and data | 157
coverage | 0.73
density | 0.28
FBXL7 down−regulates AURKA during mitotic entry and in early mit
pathway information | values
# of genes in pathway | 54
# of genes in pathway and data | 48
coverage | 0.89
density | 0.98
Fc epsilon receptor (FCERI) signaling
pathway information | values
# of genes in pathway | 352
# of genes in pathway and data | 254
coverage | 0.72
density | 0.38
FCERI mediated Ca+2 mobilization
pathway information | values
# of genes in pathway | 27
# of genes in pathway and data | 25
coverage | 0.93
density | 0.75
FCERI mediated MAPK activation
pathway information | values
# of genes in pathway | 247
# of genes in pathway and data | 169
coverage | 0.68
density | 0.39
FCERI mediated NF−kB activation
pathway information | values
# of genes in pathway | 76
# of genes in pathway and data | 66
coverage | 0.87
density | 0.73
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

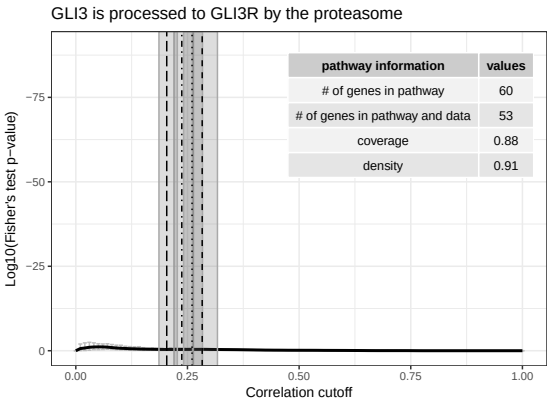

Fcgamma receptor (FCGR) dependent phagocytosis
pathway information | values
# of genes in pathway | 81
# of genes in pathway and data | 66
coverage | 0.81
density | 0.63
FCGR activation
pathway information | values
# of genes in pathway | 10
# of genes in pathway and data | 9
coverage | 0.9
density | 0.89
Formation of a pool of free 40S subunits
pathway information | values
# of genes in pathway | 102
# of genes in pathway and data | 94
coverage | 0.92
density | 0.99
Formation of RNA Pol II elongation complex
pathway information | values
# of genes in pathway | 63
# of genes in pathway and data | 47
coverage | 0.75
density | 0.86
Formation of TC−NER Pre−Incision Complex
pathway information | values
# of genes in pathway | 54
# of genes in pathway and data | 47
coverage | 0.87
density | 0.88
Formation of the beta−catenin:TCF transactivating complex
pathway information | values
# of genes in pathway | 90
# of genes in pathway and data | 37
coverage | 0.41
density | 0.77
Formation of the cornified envelope
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 10
coverage | 0.14
density | 0.22
Formation of the ternary complex, and subsequently, the 43S compl
pathway information | values
# of genes in pathway | 52
# of genes in pathway and data | 49
coverage | 0.94
density | 0.98
Log10(Fisher's test p−value)
Correlation cutoff

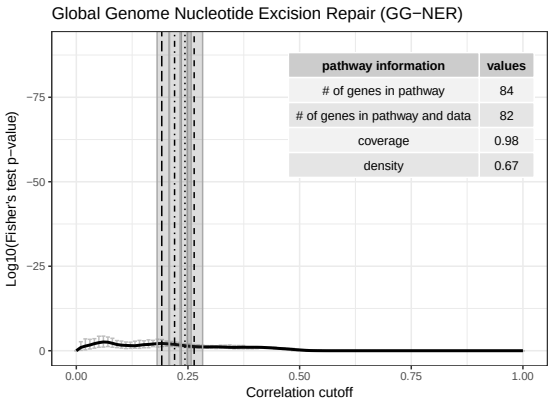

Frs2−mediated activation
pathway information | values
# of genes in pathway | 232
# of genes in pathway and data | 154
coverage | 0.66
density | 0.37
G1/S DNA Damage Checkpoints
pathway information | values
# of genes in pathway | 66
# of genes in pathway and data | 58
coverage | 0.88
density | 0.83
G1/S Transition
pathway information | values
# of genes in pathway | 118
# of genes in pathway and data | 105
coverage | 0.89
density | 0.62
G2/M Checkpoints
pathway information | values
# of genes in pathway | 168
# of genes in pathway and data | 120
coverage | 0.71
density | 0.56
G2/M DNA damage checkpoint
pathway information | values
# of genes in pathway | 95
# of genes in pathway and data | 59
coverage | 0.62
density | 0.72
G2/M Transition
pathway information | values
# of genes in pathway | 174
# of genes in pathway and data | 152
coverage | 0.87
density | 0.47
GAB1 signalosome
pathway information | values
# of genes in pathway | 113
# of genes in pathway and data | 75
coverage | 0.66
density | 0.54
GABA receptor activation
pathway information | values
# of genes in pathway | 52
# of genes in pathway and data | 18
coverage | 0.35
density | 0.83
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

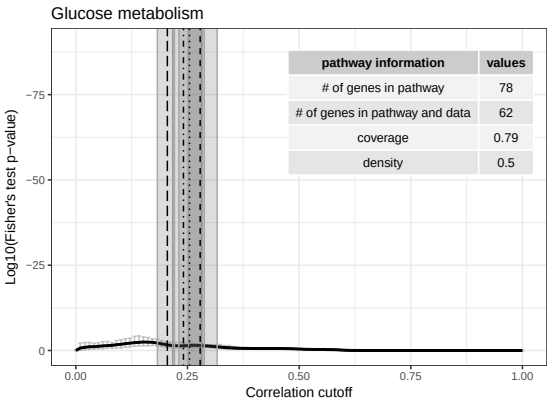

G alpha (12/13) signalling events
pathway information | values
# of genes in pathway | 87
# of genes in pathway and data | 59
coverage | 0.68
density | 0.4
G alpha (i) signalling events
pathway information | values
# of genes in pathway | 242
# of genes in pathway and data | 59
coverage | 0.24
density | 0.76
G alpha (q) signalling events
pathway information | values
# of genes in pathway | 175
# of genes in pathway and data | 55
coverage | 0.31
density | 0.57
G alpha (s) signalling events
pathway information | values
# of genes in pathway | 142
# of genes in pathway and data | 43
coverage | 0.3
density | 0.65
Gap−filling DNA repair synthesis and ligation in TC−NER
pathway information | values
# of genes in pathway | 65
# of genes in pathway and data | 59
coverage | 0.91
density | 0.95
Gastrin−CREB signalling pathway via PKC and MAPK
pathway information | values
# of genes in pathway | 408
# of genes in pathway and data | 211
coverage | 0.52
density | 0.29
Gene Expression
pathway information | values
# of genes in pathway | 1710
# of genes in pathway and data | 1370
coverage | 0.8
density | 0.15
Generation of second messenger molecules
pathway information | values
# of genes in pathway | 32
# of genes in pathway and data | 25
coverage | 0.78
density | 0.57
Log10(Fisher's test p−value)
Correlation cutoff

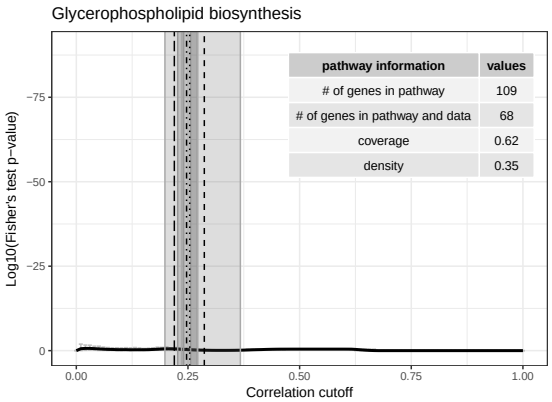

Generic Transcription Pathway
pathway information | values
# of genes in pathway | 830
# of genes in pathway and data | 650
coverage | 0.78
density | 0.17
Log10(Fisher's test p−value)
Correlation cutoff

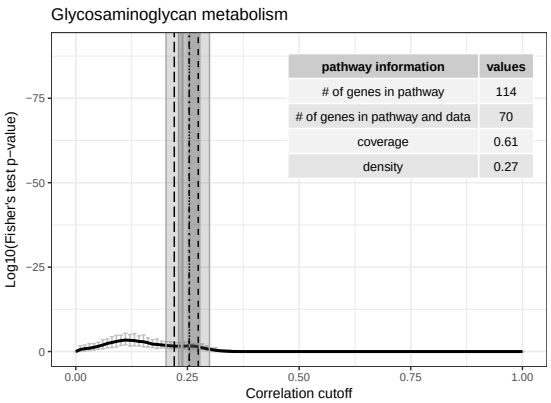

Gene Silencing by RNA
pathway information | values
# of genes in pathway | 129
# of genes in pathway and data | 62
coverage | 0.48
density | 0.4
Log10(Fisher's test p−value)
Correlation cutoff

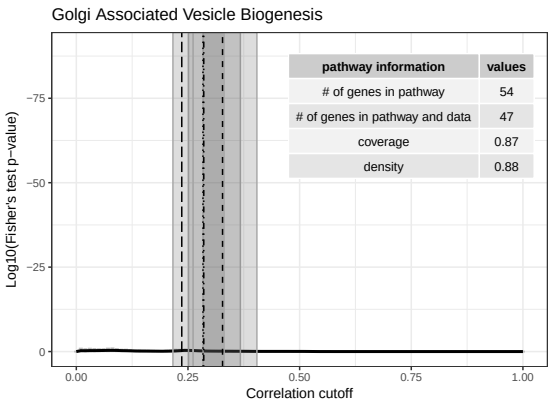

GLI3 is processed to GLI3R by the proteasome
pathway information | values
# of genes in pathway | 60
# of genes in pathway and data | 53
coverage | 0.88
density | 0.91
Log10(Fisher's test p−value)
Correlation cutoff

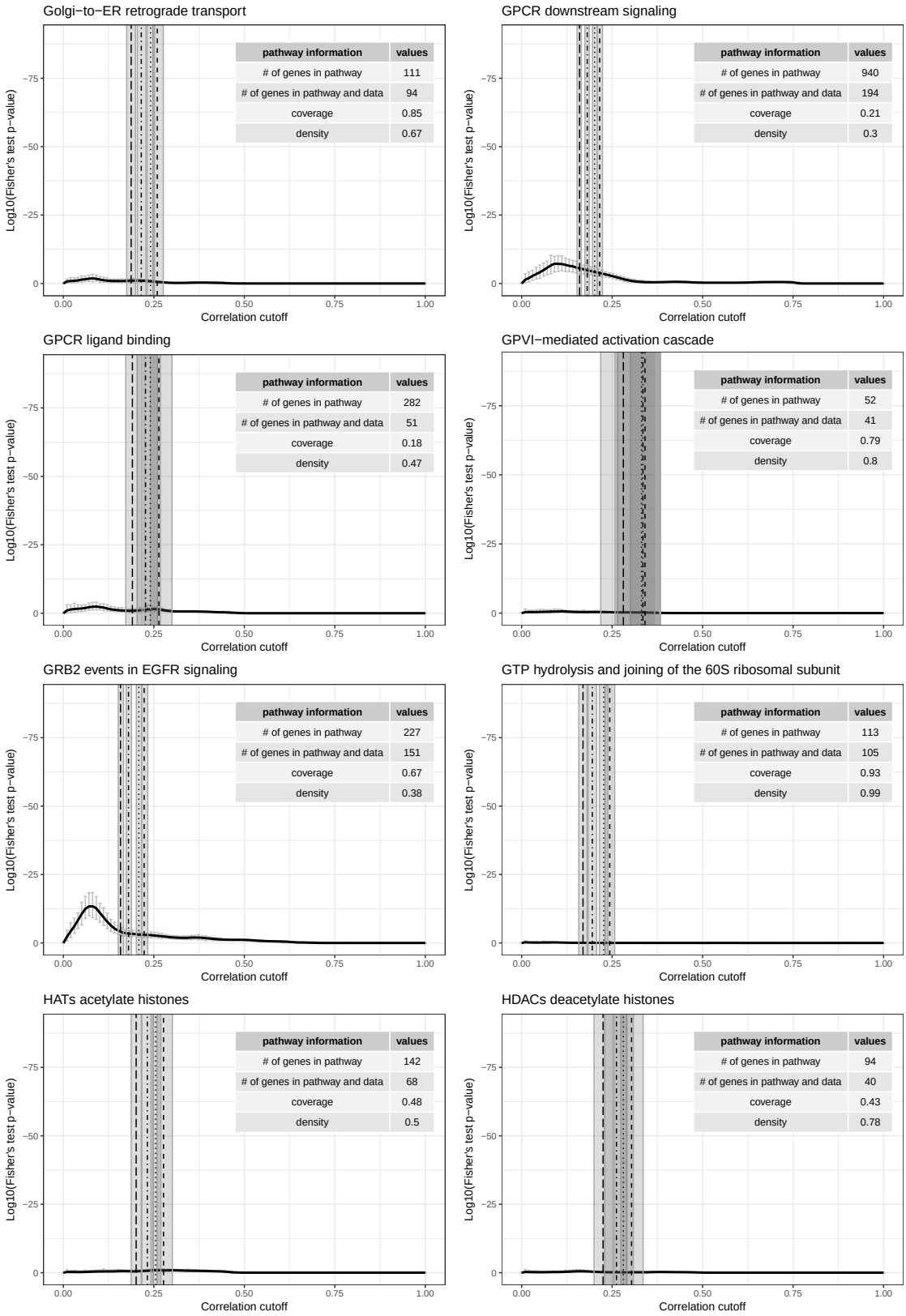

Global Genome Nucleotide Excision Repair (GG−NER)
pathway information | values
# of genes in pathway | 84
# of genes in pathway and data | 82
coverage | 0.98
density | 0.67
Log10(Fisher's test p−value)
Correlation cutoff

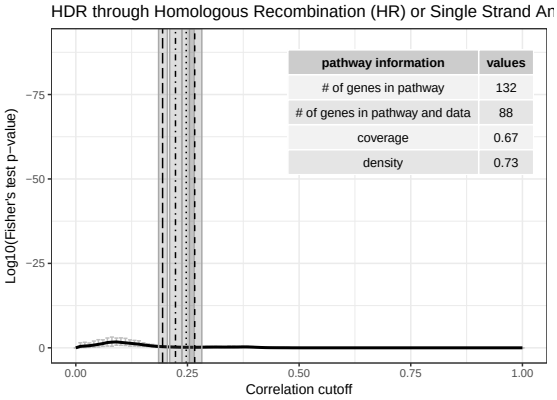

Glucose metabolism
pathway information | values
# of genes in pathway | 78
# of genes in pathway and data | 62
coverage | 0.79
density | 0.5
Log10(Fisher's test p−value)
Correlation cutoff

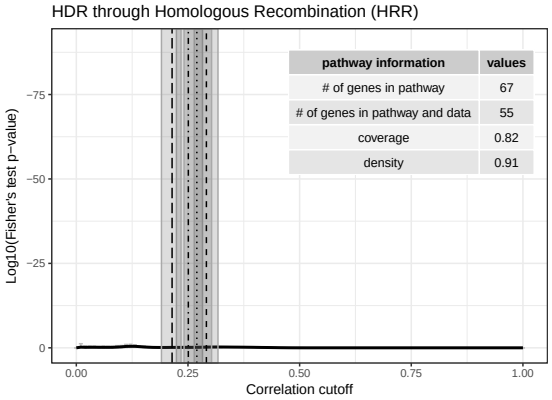

Glycerophospholipid biosynthesis
pathway information | values
# of genes in pathway | 109
# of genes in pathway and data | 68
coverage | 0.62
density | 0.35
Log10(Fisher's test p−value)
Correlation cutoff

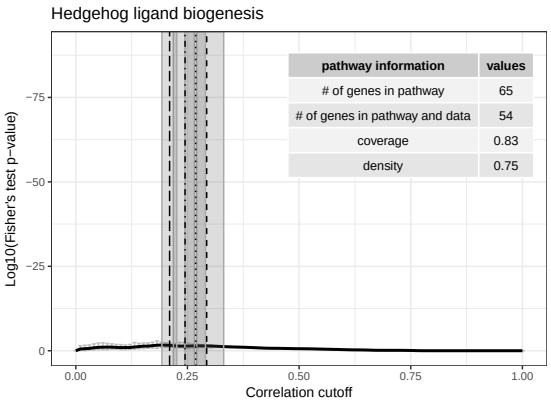

Glycosaminoglycan metabolism
pathway information | values
# of genes in pathway | 114
# of genes in pathway and data | 70
coverage | 0.61
density | 0.27
Log10(Fisher's test p−value)
Correlation cutoff

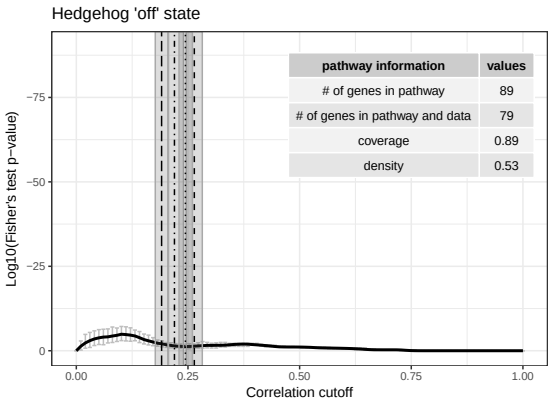

Golgi Associated Vesicle Biogenesis
pathway information | values
# of genes in pathway | 54
# of genes in pathway and data | 47
coverage | 0.87
density | 0.88
Log10(Fisher's test p−value)
Correlation cutoff

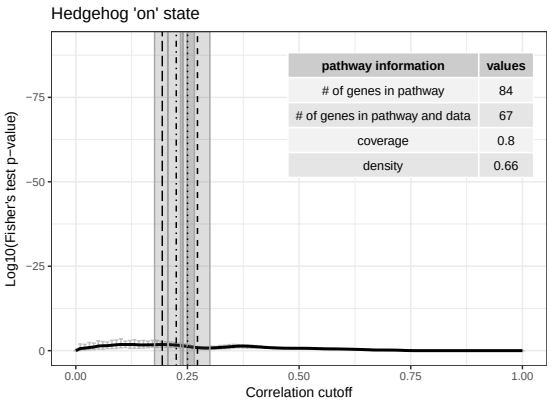

Golgi−to−ER retrograde transport
pathway information | values
# of genes in pathway | 111
# of genes in pathway and data | 94
coverage | 0.85
density | 0.67
GPCR downstream signaling
pathway information | values
# of genes in pathway | 940
# of genes in pathway and data | 194
coverage | 0.21
density | 0.3
GPCR ligand binding
pathway information | values
# of genes in pathway | 282
# of genes in pathway and data | 51
coverage | 0.18
density | 0.47
GPVI−mediated activation cascade
pathway information | values
# of genes in pathway | 52
# of genes in pathway and data | 41
coverage | 0.79
density | 0.8
GRB2 events in EGFR signaling
pathway information | values
# of genes in pathway | 227
# of genes in pathway and data | 151
coverage | 0.67
density | 0.38
GTP hydrolysis and joining of the 60S ribosomal subunit
pathway information | values
# of genes in pathway | 113
# of genes in pathway and data | 105
coverage | 0.93
density | 0.99
HATs acetylate histones
pathway information | values
# of genes in pathway | 142
# of genes in pathway and data | 68
coverage | 0.48
density | 0.5
HDACs deacetylate histones
pathway information | values
# of genes in pathway | 94
# of genes in pathway and data | 40
coverage | 0.43
density | 0.78
Log10(Fisher's test p−value)
Correlation cutoff

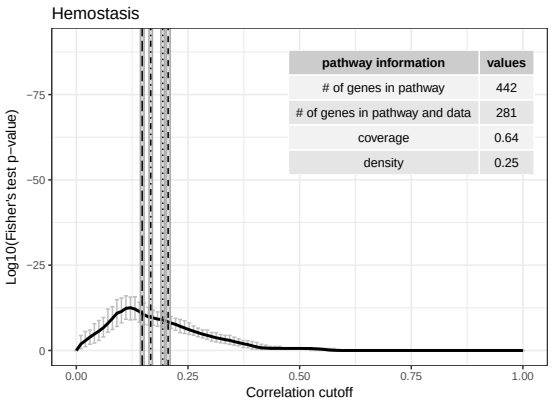

HDR through Homologous Recombination (HR) or Single Strand An
pathway information | values
# of genes in pathway | 132
# of genes in pathway and data | 88
coverage | 0.67
density | 0.73
Log10(Fisher's test p−value)
Correlation cutoff

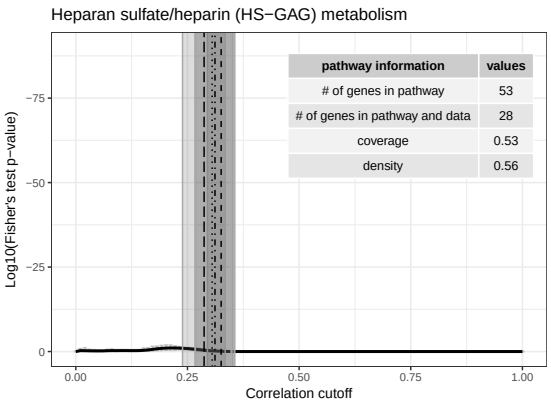

HDR through Homologous Recombination (HRR)
pathway information | values
# of genes in pathway | 67
# of genes in pathway and data | 55
coverage | 0.82
density | 0.91
Log10(Fisher's test p−value)
Correlation cutoff

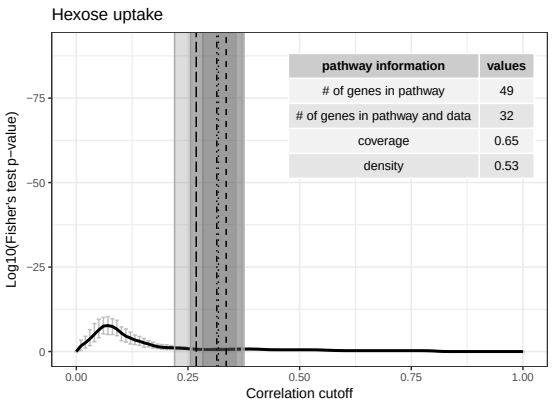

Hedgehog ligand biogenesis
pathway information | values
# of genes in pathway | 65
# of genes in pathway and data | 54
coverage | 0.83
density | 0.75
Log10(Fisher's test p−value)
Correlation cutoff

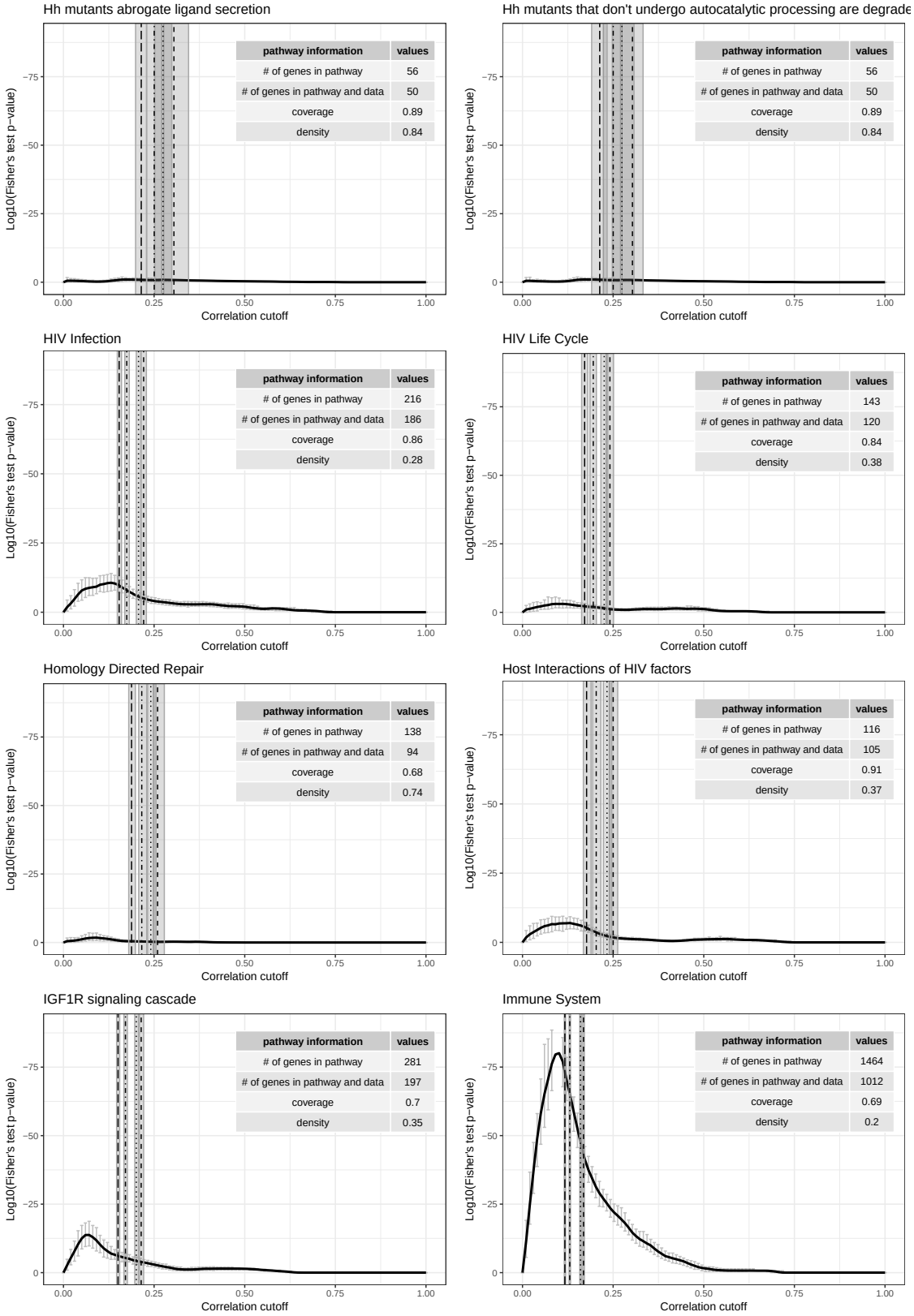

Hedgehog 'off' state
pathway information | values
# of genes in pathway | 89
# of genes in pathway and data | 79
coverage | 0.89
density | 0.53
Log10(Fisher's test p−value)
Correlation cutoff

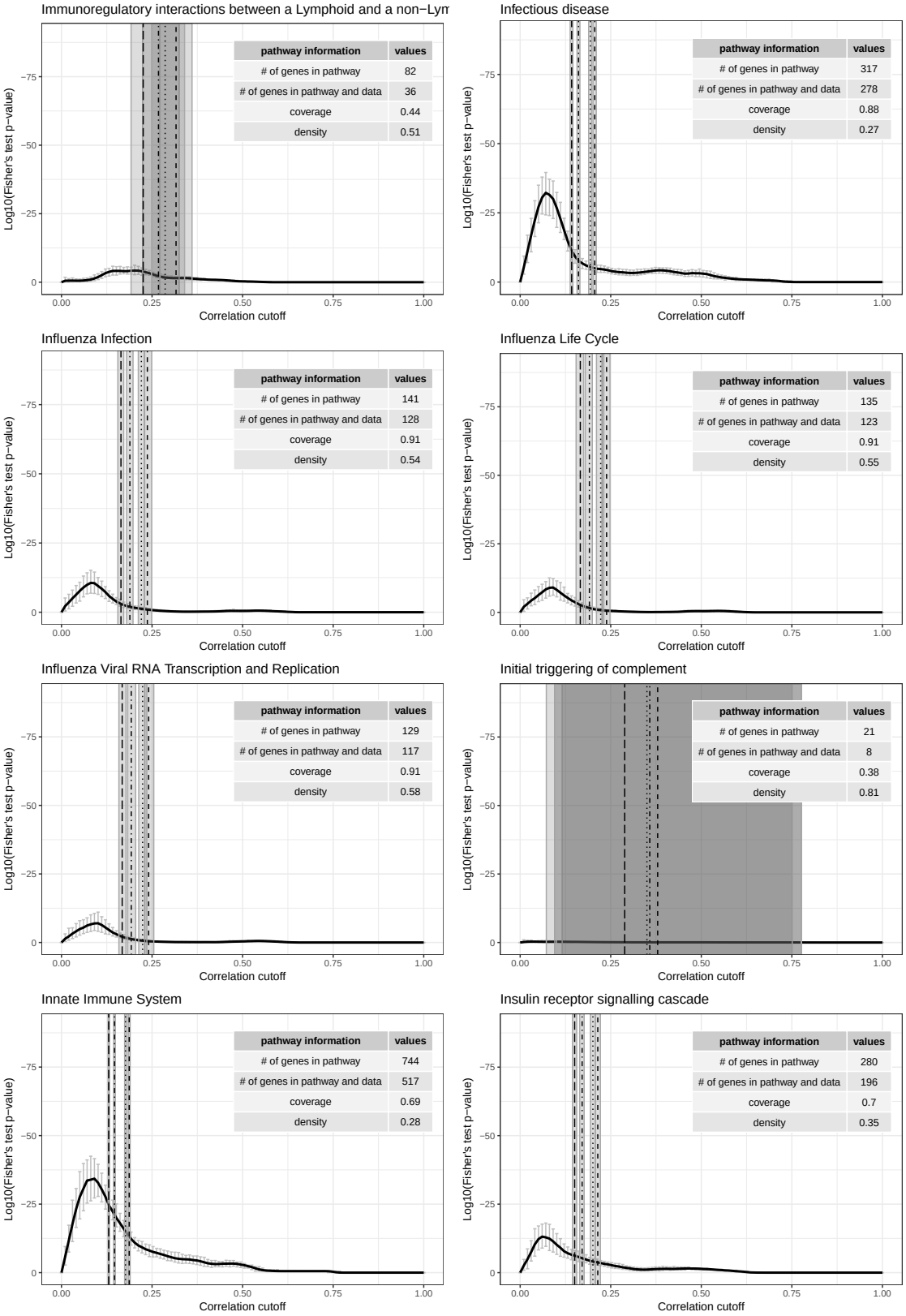

Hedgehog 'on' state
pathway information | values
# of genes in pathway | 84
# of genes in pathway and data | 67
coverage | 0.8
density | 0.66
Log10(Fisher's test p−value)
Correlation cutoff

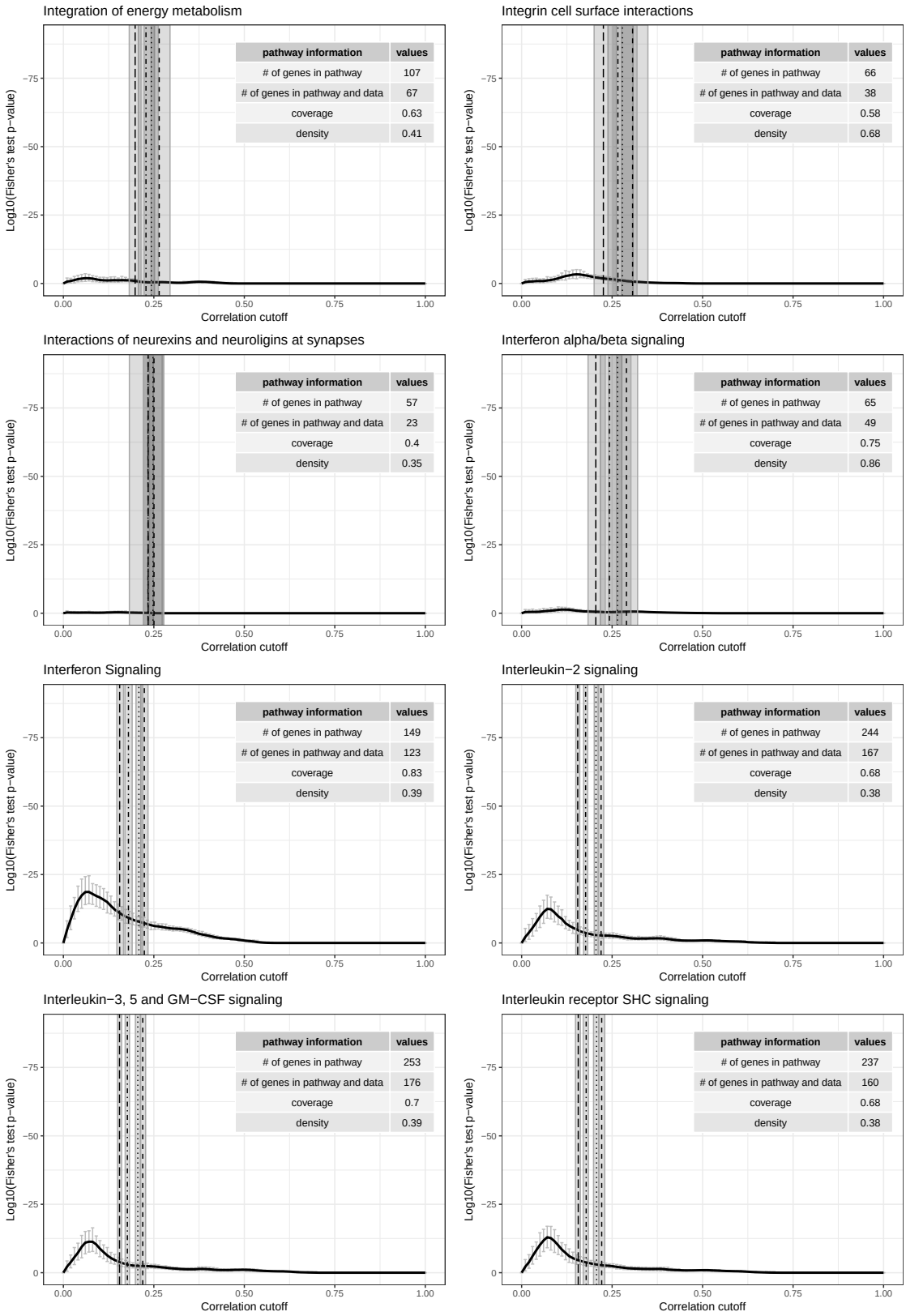

Hemostasis
pathway information | values
# of genes in pathway | 442
# of genes in pathway and data | 281
coverage | 0.64
density | 0.25
Log10(Fisher's test p−value)
Correlation cutoff

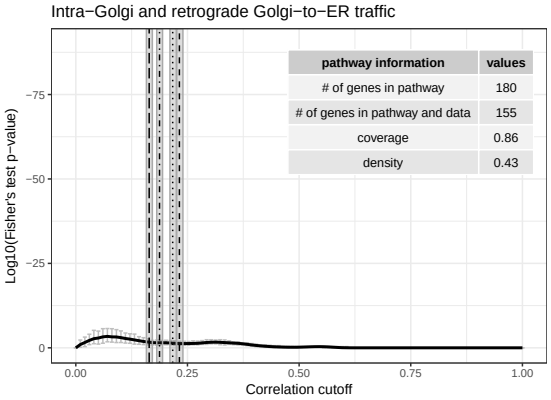

Heparan sulfate/heparin (HS−GAG) metabolism
pathway information | values
# of genes in pathway | 53
# of genes in pathway and data | 28
coverage | 0.53
density | 0.56
Log10(Fisher's test p−value)
Correlation cutoff

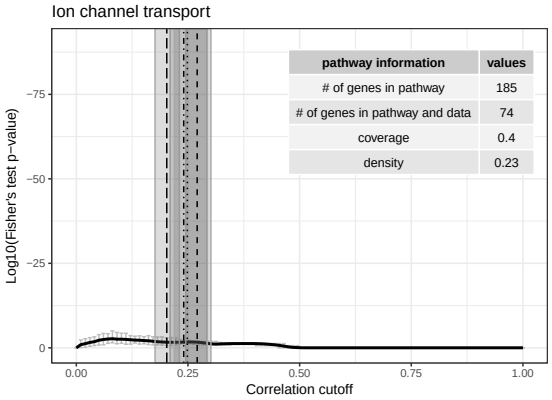

Hexose uptake
pathway information | values
# of genes in pathway | 49
# of genes in pathway and data | 32
coverage | 0.65
density | 0.53
Log10(Fisher's test p−value)
Correlation cutoff

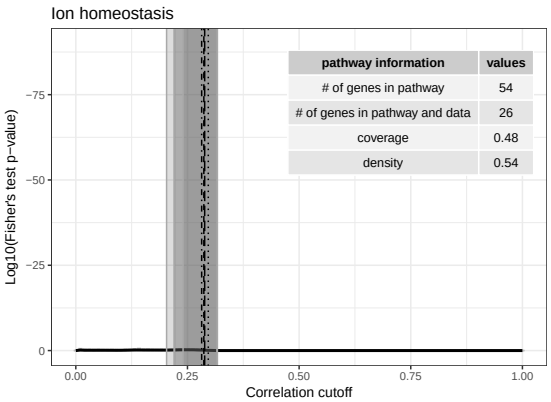

Hh mutants abrogate ligand secretion
Hh mutants that don't undergo autocatalytic processing are degrade
HIV Infection
HIV Life Cycle
Homology Directed Repair
Host Interactions of HIV factors
IGF1R signaling cascade
Immune System
pathway information
values
# of genes in pathway
# of genes in pathway and data
coverage
density
56
50
0.89
0.84
216
186
0.86
0.28
143
120
0.84
0.38
138
94
0.68
0.74
116
105
0.91
0.37
281
197
0.7
0.35
1464
1012
0.69
0.2
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

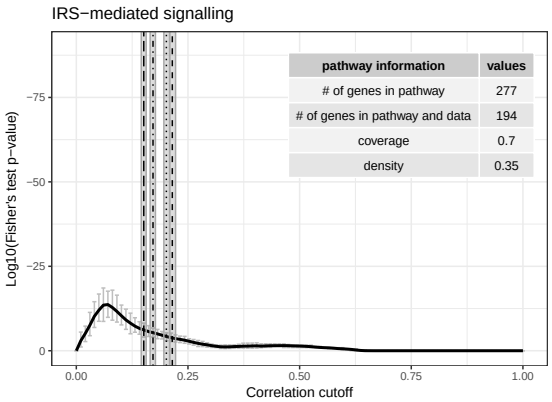

Immunoregulatory interactions between a Lymphoid and a non−Lym
pathway information | values
# of genes in pathway | 82
# of genes in pathway and data | 36
coverage | 0.44
density | 0.51
Infectious disease
pathway information | values
# of genes in pathway | 317
# of genes in pathway and data | 278
coverage | 0.88
density | 0.27
Influenza Infection
pathway information | values
# of genes in pathway | 141
# of genes in pathway and data | 128
coverage | 0.91
density | 0.54
Influenza Life Cycle
pathway information | values
# of genes in pathway | 135
# of genes in pathway and data | 123
coverage | 0.91
density | 0.55
Influenza Viral RNA Transcription and Replication
pathway information | values
# of genes in pathway | 129
# of genes in pathway and data | 117
coverage | 0.91
density | 0.58
Initial triggering of complement
pathway information | values
# of genes in pathway | 21
# of genes in pathway and data | 8
coverage | 0.38
density | 0.81
Innate Immune System
pathway information | values
# of genes in pathway | 744
# of genes in pathway and data | 517
coverage | 0.69
density | 0.28
Insulin receptor signalling cascade
pathway information | values
# of genes in pathway | 280
# of genes in pathway and data | 196
coverage | 0.7
density | 0.35
Log10(Fisher's test p−value)
Correlation cutoff

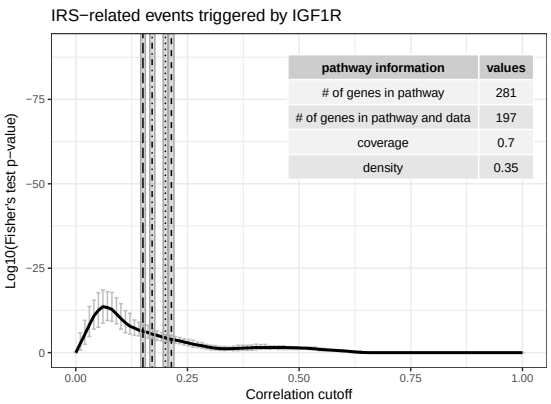

Integration of energy metabolism
pathway information | values
# of genes in pathway | 107
# of genes in pathway and data | 67
coverage | 0.63
density | 0.41
Integrin cell surface interactions
pathway information | values
# of genes in pathway | 66
# of genes in pathway and data | 38
coverage | 0.58
density | 0.68
Interactions of neurexins and neuroligins at synapses
pathway information | values
# of genes in pathway | 57
# of genes in pathway and data | 23
coverage | 0.4
density | 0.35
Interferon alpha/beta signaling
pathway information | values
# of genes in pathway | 65
# of genes in pathway and data | 49
coverage | 0.75
density | 0.86
Interferon Signaling
pathway information | values
# of genes in pathway | 149
# of genes in pathway and data | 123
coverage | 0.83
density | 0.39
Interleukin−2 signaling
pathway information | values
# of genes in pathway | 244
# of genes in pathway and data | 167
coverage | 0.68
density | 0.38
Interleukin−3, 5 and GM−CSF signaling
pathway information | values
# of genes in pathway | 253
# of genes in pathway and data | 176
coverage | 0.7
density | 0.39
Interleukin receptor SHC signaling
pathway information | values
# of genes in pathway | 237
# of genes in pathway and data | 160
coverage | 0.68
density | 0.38
Log10(Fisher's test p−value)
Correlation cutoff

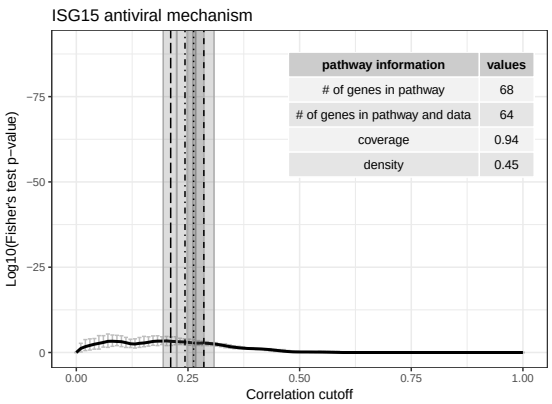

Intra−Golgi and retrograde Golgi−to−ER traffic
pathway information | values
# of genes in pathway | 180
# of genes in pathway and data | 155
coverage | 0.86
density | 0.43
Log10(Fisher's test p−value)
Correlation cutoff

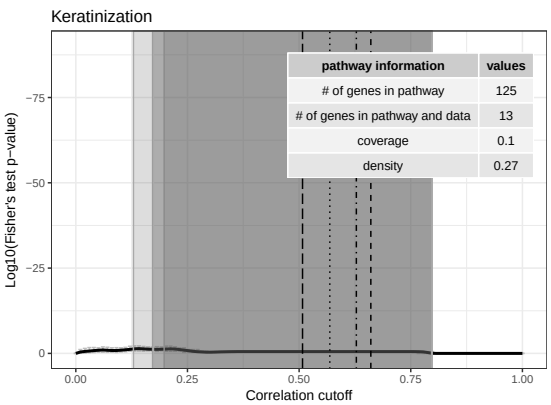

Ion channel transport
pathway information | values
# of genes in pathway | 185
# of genes in pathway and data | 74
coverage | 0.4
density | 0.23
Log10(Fisher's test p−value)
Correlation cutoff

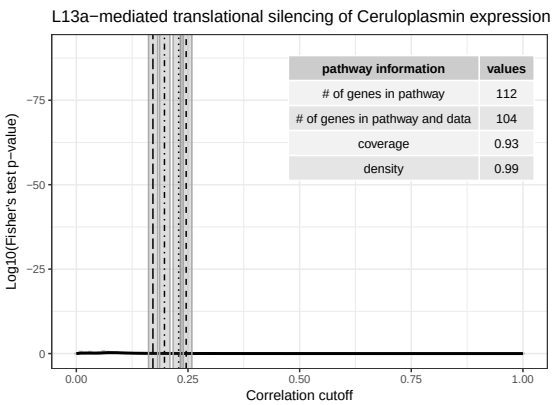

Ion homeostasis
pathway information | values
# of genes in pathway | 54
# of genes in pathway and data | 26
coverage | 0.48
density | 0.54
Log10(Fisher's test p−value)
Correlation cutoff

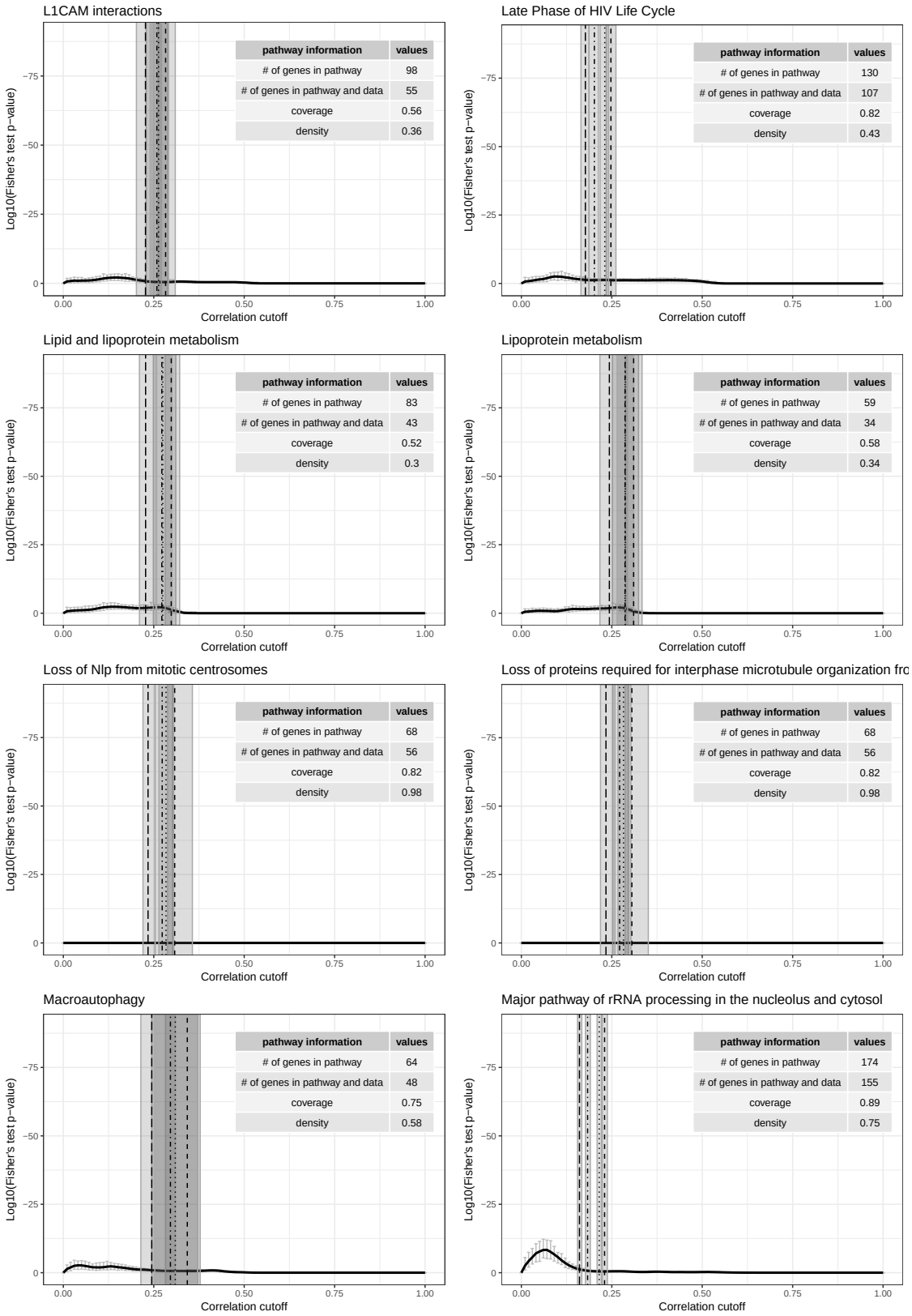

IRS−mediated signalling
pathway information | values
# of genes in pathway | 277
# of genes in pathway and data | 194
coverage | 0.7
density | 0.35
Log10(Fisher's test p−value)
Correlation cutoff


IRS−related events triggered by IGF1R
pathway information | values
# of genes in pathway | 281
# of genes in pathway and data | 197
coverage | 0.7
density | 0.35
Log10(Fisher's test p−value)
Correlation cutoff

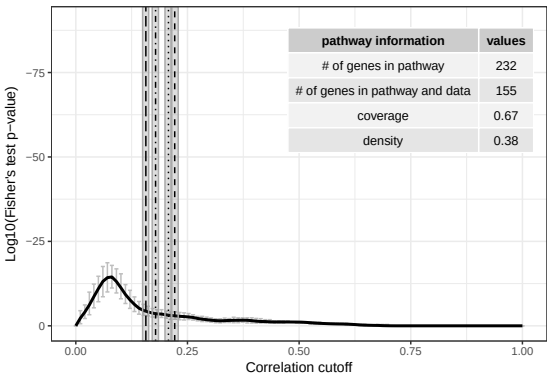

ISG15 antiviral mechanism
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 64
coverage | 0.94
density | 0.45
Log10(Fisher's test p−value)
Correlation cutoff


Keratinization
pathway information | values
# of genes in pathway | 125
# of genes in pathway and data | 13
coverage | 0.1
density | 0.27
Log10(Fisher's test p−value)
Correlation cutoff

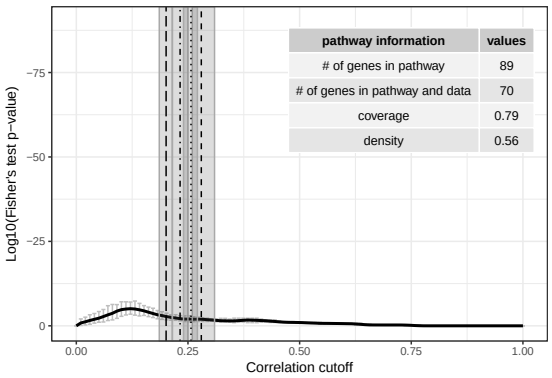

L13a−mediated translational silencing of Ceruloplasmin expression
pathway information | values
# of genes in pathway | 112
# of genes in pathway and data | 104
coverage | 0.93
density | 0.99
Log10(Fisher's test p−value)
Correlation cutoff


L1CAM interactions
pathway information | values
# of genes in pathway | 98
# of genes in pathway and data | 55
coverage | 0.56
density | 0.36
Late Phase of HIV Life Cycle
pathway information | values
# of genes in pathway | 130
# of genes in pathway and data | 107
coverage | 0.82
density | 0.43
Lipid and lipoprotein metabolism
pathway information | values
# of genes in pathway | 83
# of genes in pathway and data | 43
coverage | 0.52
density | 0.3
Lipoprotein metabolism
pathway information | values
# of genes in pathway | 59
# of genes in pathway and data | 34
coverage | 0.58
density | 0.34
Loss of Nlp from mitotic centrosomes
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 56
coverage | 0.82
density | 0.98
Loss of proteins required for interphase microtubule organization fro
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 56
coverage | 0.82
density | 0.98
Macroautophagy
pathway information | values
# of genes in pathway | 64
# of genes in pathway and data | 48
coverage | 0.75
density | 0.58
Major pathway of rRNA processing in the nucleolus and cytosol
pathway information | values
# of genes in pathway | 174
# of genes in pathway and data | 155
coverage | 0.89
density | 0.75
Log10(Fisher's test p−value)
Correlation cutoff

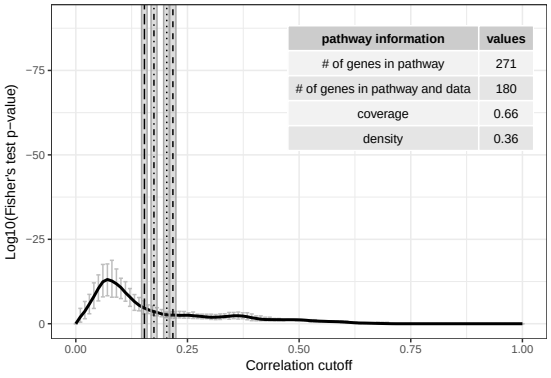
MAPK1/MAPK3 signaling



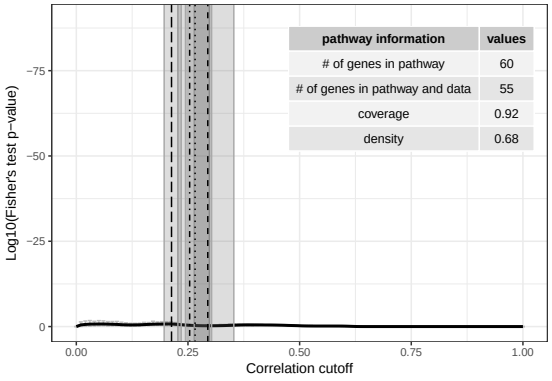
MAPK6/MAPK4 signaling



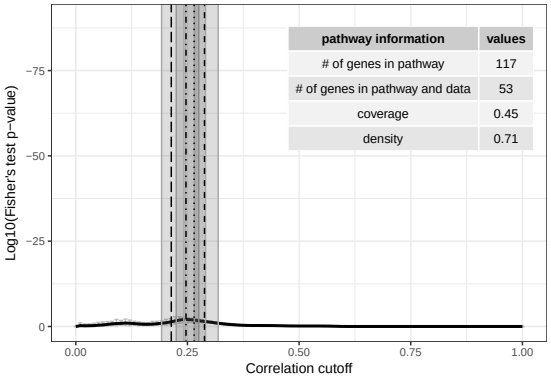
MAPK family signaling cascades



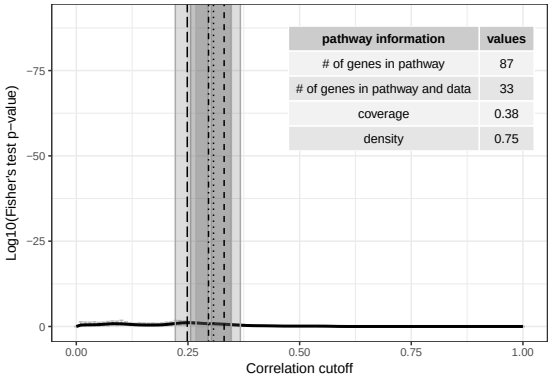
MAP kinase activation in TLR cascade



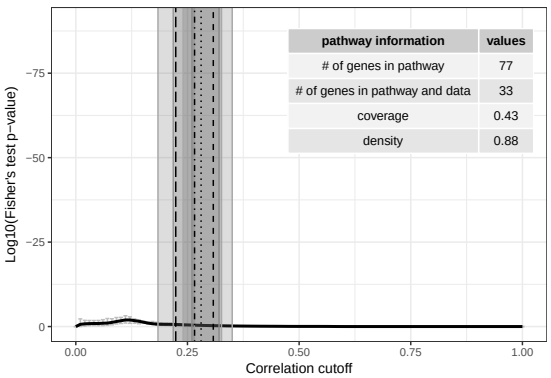
Meiosis



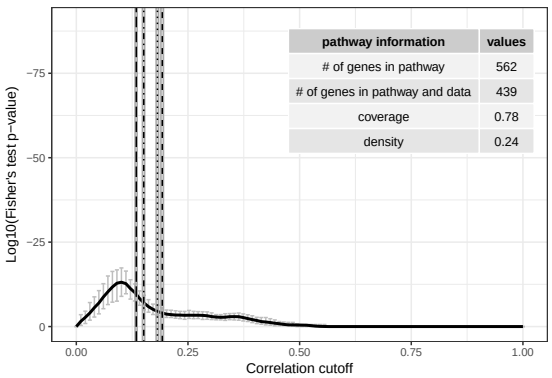
Meiotic recombination

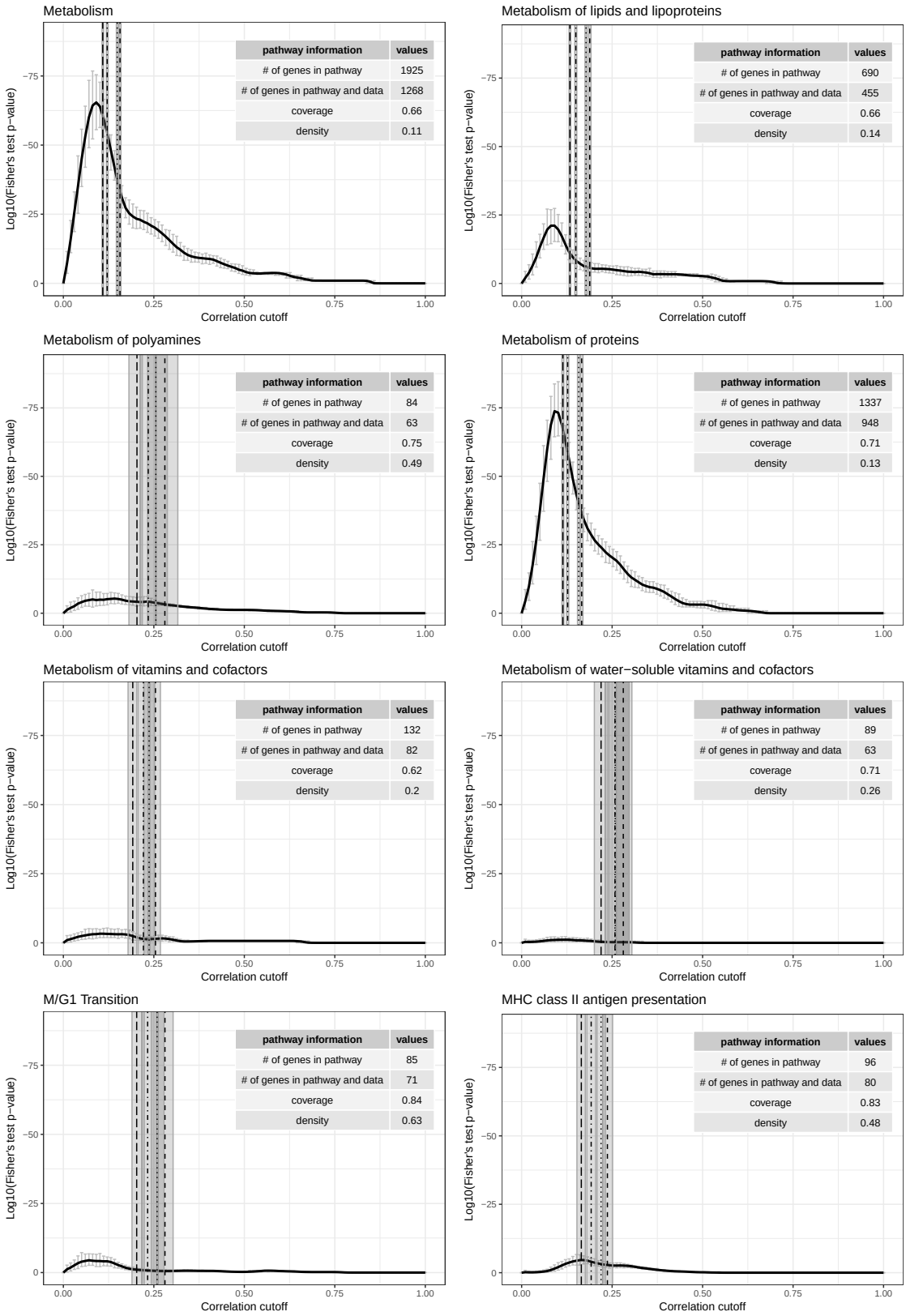


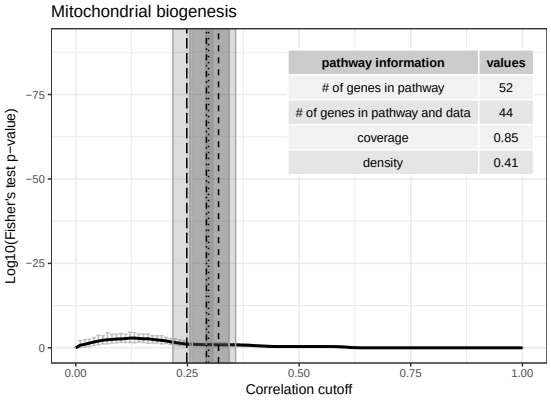
Meiotic synapsis

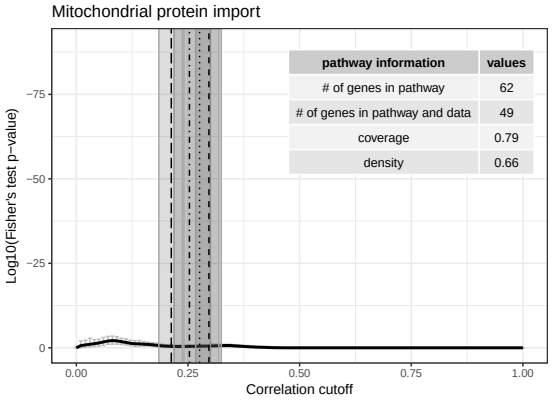


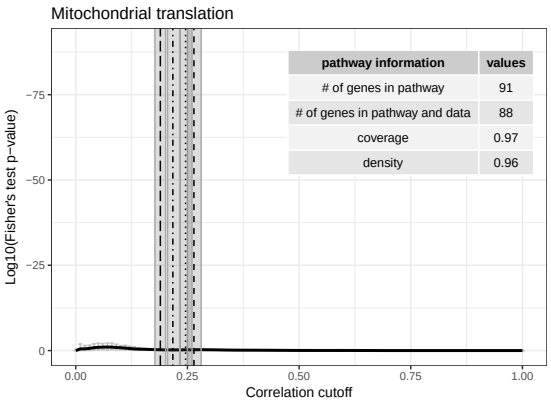
Membrane Trafficking

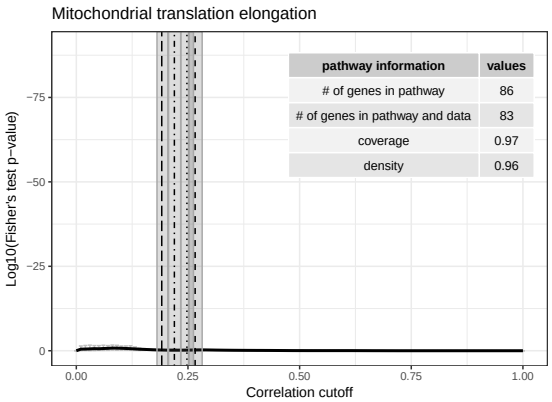


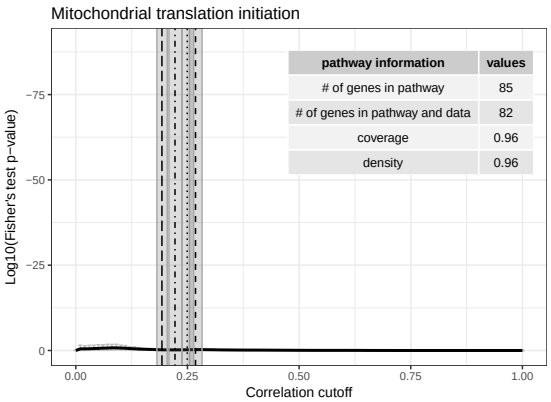

Metabolism
pathway information | values
# of genes in pathway | 1925
# of genes in pathway and data | 1268
coverage | 0.66
density | 0.11
Metabolism of lipids and lipoproteins
pathway information | values
# of genes in pathway | 690
# of genes in pathway and data | 455
coverage | 0.66
density | 0.14
Metabolism of polyamines
pathway information | values
# of genes in pathway | 84
# of genes in pathway and data | 63
coverage | 0.75
density | 0.49
Metabolism of proteins
pathway information | values
# of genes in pathway | 1337
# of genes in pathway and data | 948
coverage | 0.71
density | 0.13
Metabolism of vitamins and cofactors
pathway information | values
# of genes in pathway | 132
# of genes in pathway and data | 82
coverage | 0.62
density | 0.2
Metabolism of water−soluble vitamins and cofactors
pathway information | values
# of genes in pathway | 89
# of genes in pathway and data | 63
coverage | 0.71
density | 0.26
M/G1 Transition
pathway information | values
# of genes in pathway | 85
# of genes in pathway and data | 71
coverage | 0.84
density | 0.63
MHC class II antigen presentation
pathway information | values
# of genes in pathway | 96
# of genes in pathway and data | 80
coverage | 0.83
density | 0.48
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

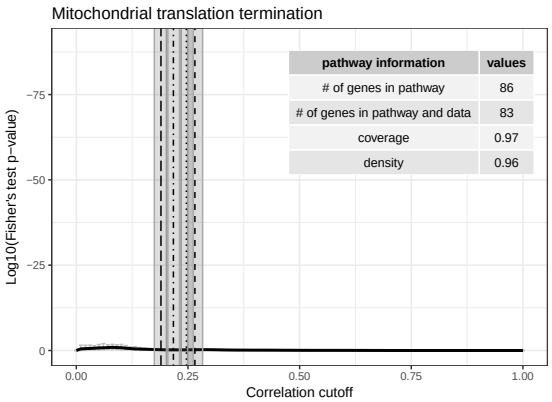

Mitochondrial biogenesis
pathway information | values
# of genes in pathway | 52
# of genes in pathway and data | 44
coverage | 0.85
density | 0.41
Log10(Fisher's test p−value)
Correlation cutoff

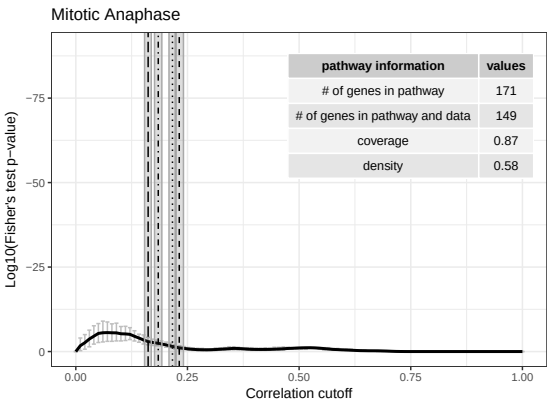

Mitochondrial protein import
pathway information | values
# of genes in pathway | 62
# of genes in pathway and data | 49
coverage | 0.79
density | 0.66
Log10(Fisher's test p−value)
Correlation cutoff

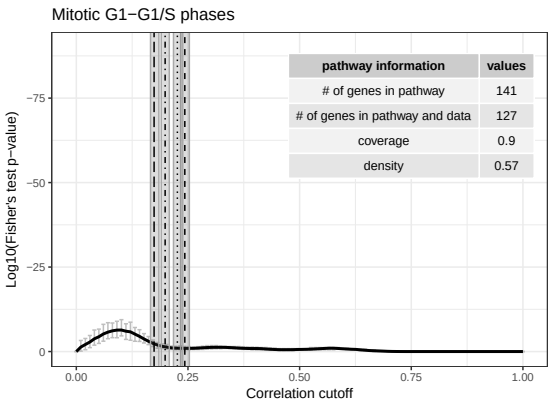

Mitochondrial translation
pathway information | values
# of genes in pathway | 91
# of genes in pathway and data | 88
coverage | 0.97
density | 0.96
Log10(Fisher's test p−value)
Correlation cutoff

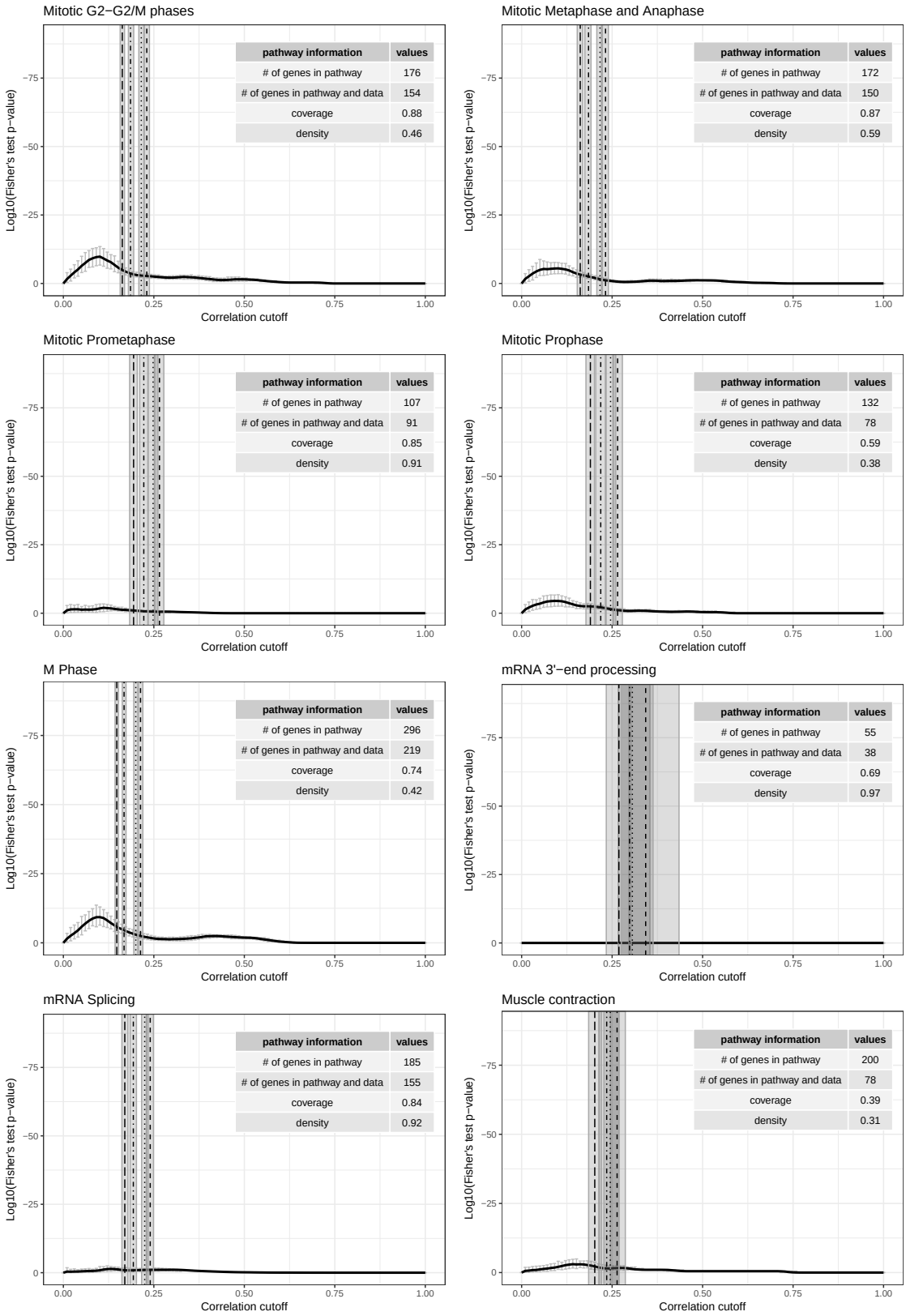

Mitochondrial translation elongation
pathway information | values
# of genes in pathway | 86
# of genes in pathway and data | 83
coverage | 0.97
density | 0.96
Log10(Fisher's test p−value)
Correlation cutoff

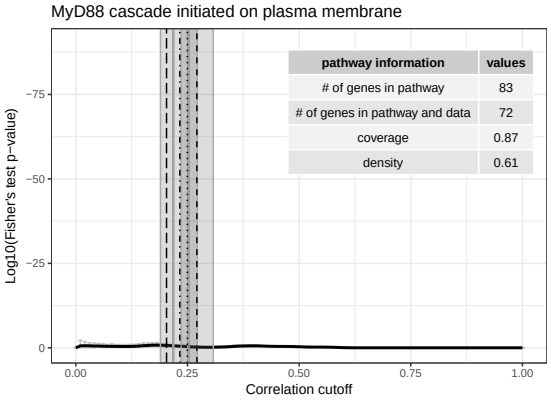

Mitochondrial translation initiation
pathway information | values
# of genes in pathway | 85
# of genes in pathway and data | 82
coverage | 0.96
density | 0.96
Log10(Fisher's test p−value)
Correlation cutoff

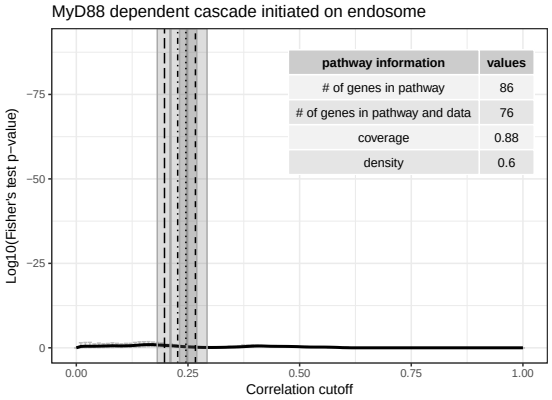

Mitochondrial translation termination
pathway information | values
# of genes in pathway | 86
# of genes in pathway and data | 83
coverage | 0.97
density | 0.96
Log10(Fisher's test p−value)
Correlation cutoff

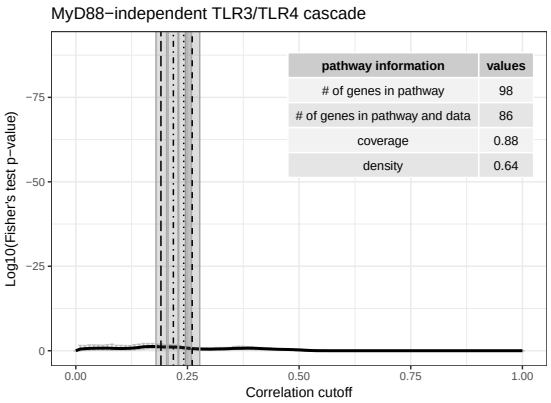

Mitotic Anaphase
pathway information | values
# of genes in pathway | 171
# of genes in pathway and data | 149
coverage | 0.87
density | 0.58
Log10(Fisher's test p−value)
Correlation cutoff

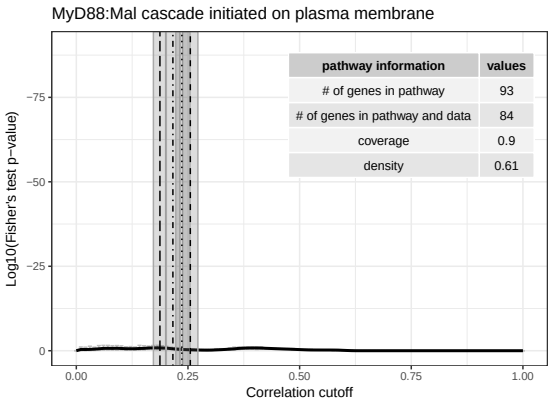

Mitotic G1−G1/S phases
pathway information | values
# of genes in pathway | 141
# of genes in pathway and data | 127
coverage | 0.9
density | 0.57
Log10(Fisher's test p−value)
Correlation cutoff

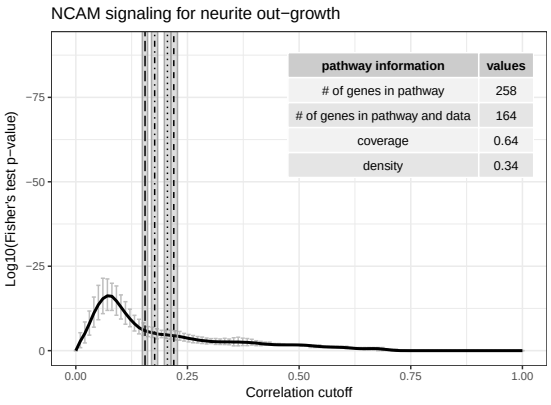

Mitotic G2−G2/M phases
pathway information | values
# of genes in pathway | 176
# of genes in pathway and data | 154
coverage | 0.88
density | 0.46
Mitotic Metaphase and Anaphase
pathway information | values
# of genes in pathway | 172
# of genes in pathway and data | 150
coverage | 0.87
density | 0.59
Mitotic Prometaphase
pathway information | values
# of genes in pathway | 107
# of genes in pathway and data | 91
coverage | 0.85
density | 0.91
Mitotic Prophase
pathway information | values
# of genes in pathway | 132
# of genes in pathway and data | 78
coverage | 0.59
density | 0.38
M Phase
pathway information | values
# of genes in pathway | 296
# of genes in pathway and data | 219
coverage | 0.74
density | 0.42
mRNA 3'−end processing
pathway information | values
# of genes in pathway | 55
# of genes in pathway and data | 38
coverage | 0.69
density | 0.97
mRNA Splicing
pathway information | values
# of genes in pathway | 185
# of genes in pathway and data | 155
coverage | 0.84
density | 0.92
Muscle contraction
pathway information | values
# of genes in pathway | 200
# of genes in pathway and data | 78
coverage | 0.39
density | 0.31
Log10(Fisher's test p−value)
Correlation cutoff

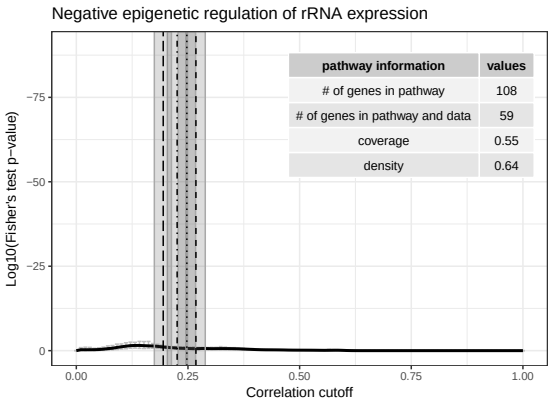

MyD88 cascade initiated on plasma membrane
pathway information | values
# of genes in pathway | 83
# of genes in pathway and data | 72
coverage | 0.87
density | 0.61
Log10(Fisher's test p−value)
Correlation cutoff

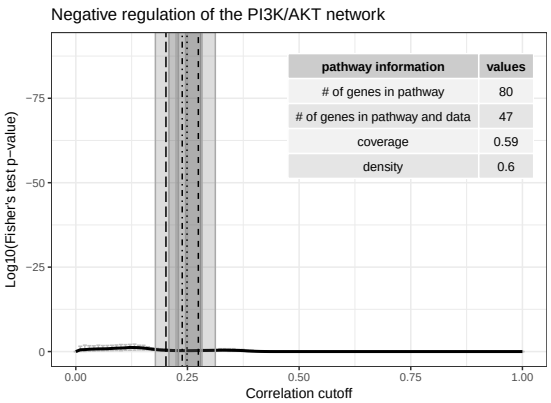

MyD88 dependent cascade initiated on endosome
pathway information | values
# of genes in pathway | 86
# of genes in pathway and data | 76
coverage | 0.88
density | 0.6
Log10(Fisher's test p−value)
Correlation cutoff

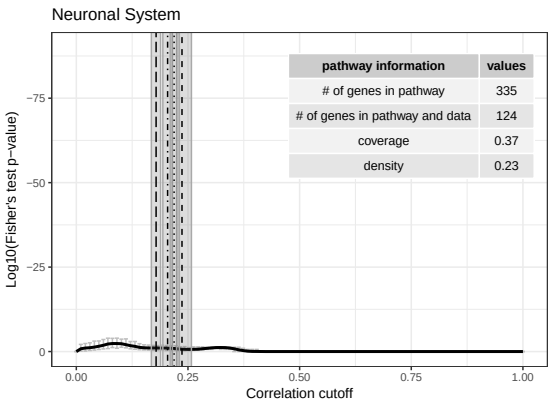

MyD88−independent TLR3/TLR4 cascade
pathway information | values
# of genes in pathway | 98
# of genes in pathway and data | 86
coverage | 0.88
density | 0.64
Log10(Fisher's test p−value)
Correlation cutoff

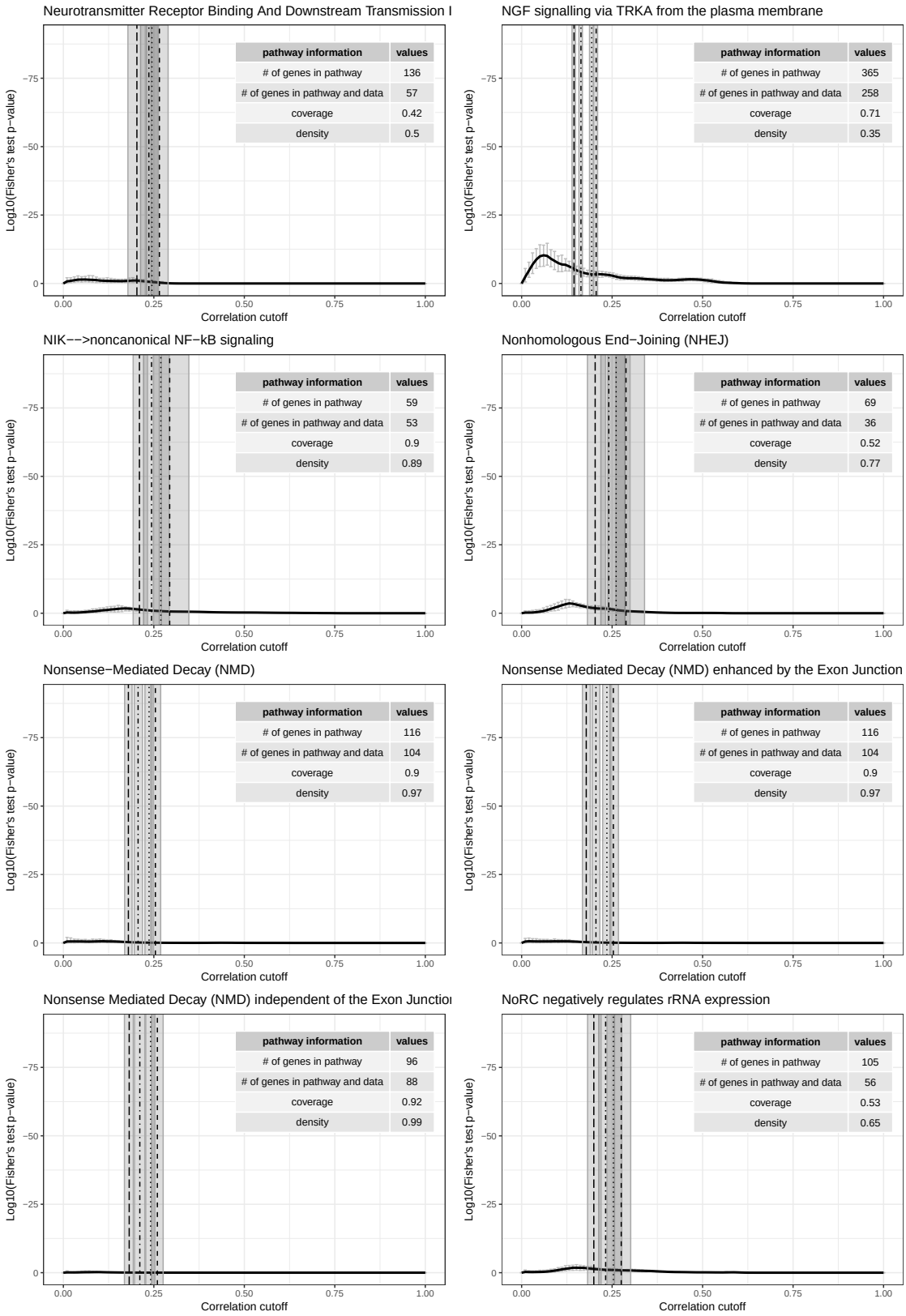

MyD88:Mal cascade initiated on plasma membrane
pathway information | values
# of genes in pathway | 93
# of genes in pathway and data | 84
coverage | 0.9
density | 0.61
Log10(Fisher's test p−value)
Correlation cutoff

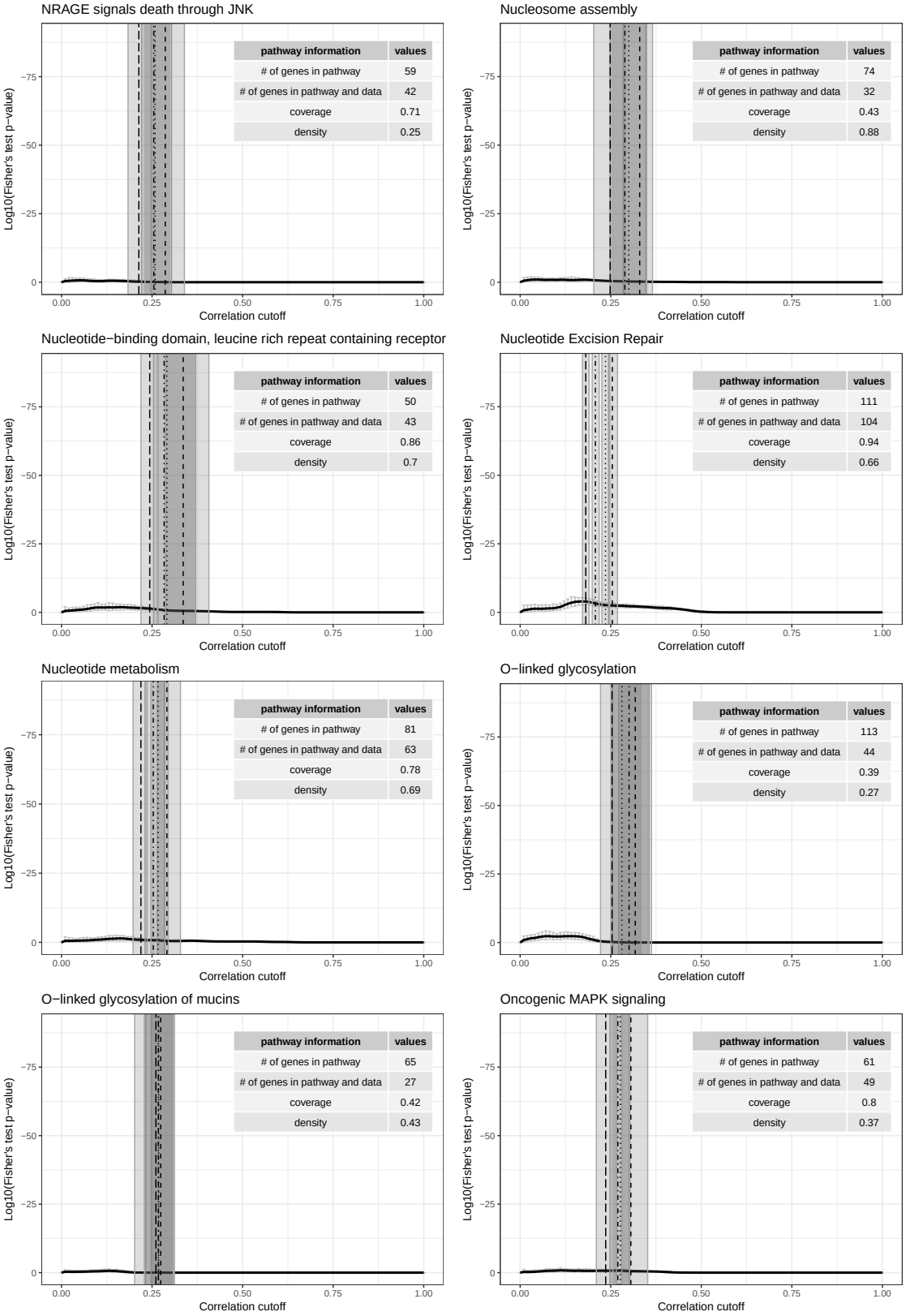

NCAM signaling for neurite out−growth
pathway information | values
# of genes in pathway | 258
# of genes in pathway and data | 164
coverage | 0.64
density | 0.34
Log10(Fisher's test p−value)
Correlation cutoff

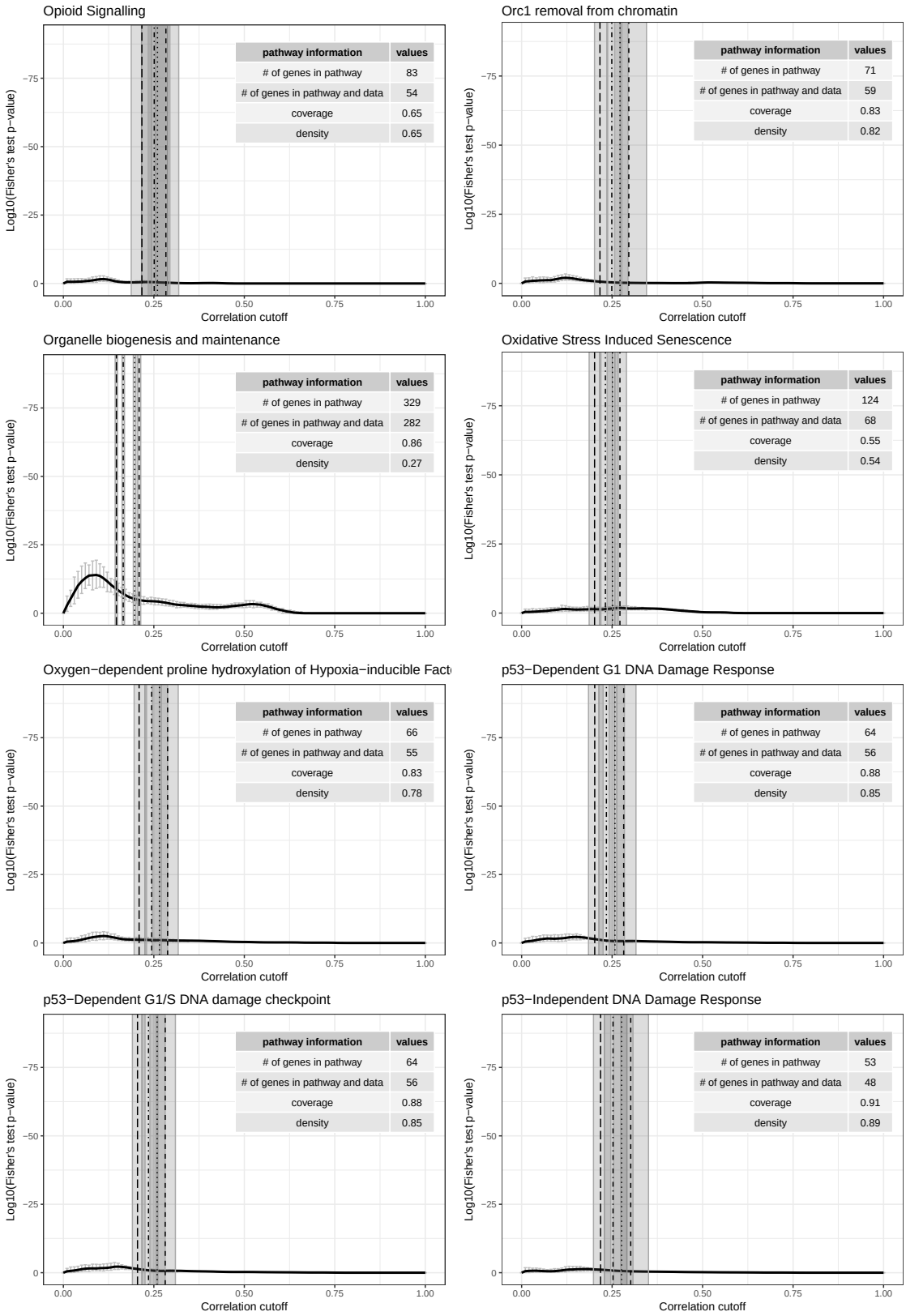

Negative epigenetic regulation of rRNA expression
pathway information | values
# of genes in pathway | 108
# of genes in pathway and data | 59
coverage | 0.55
density | 0.64
Log10(Fisher's test p−value)
Correlation cutoff

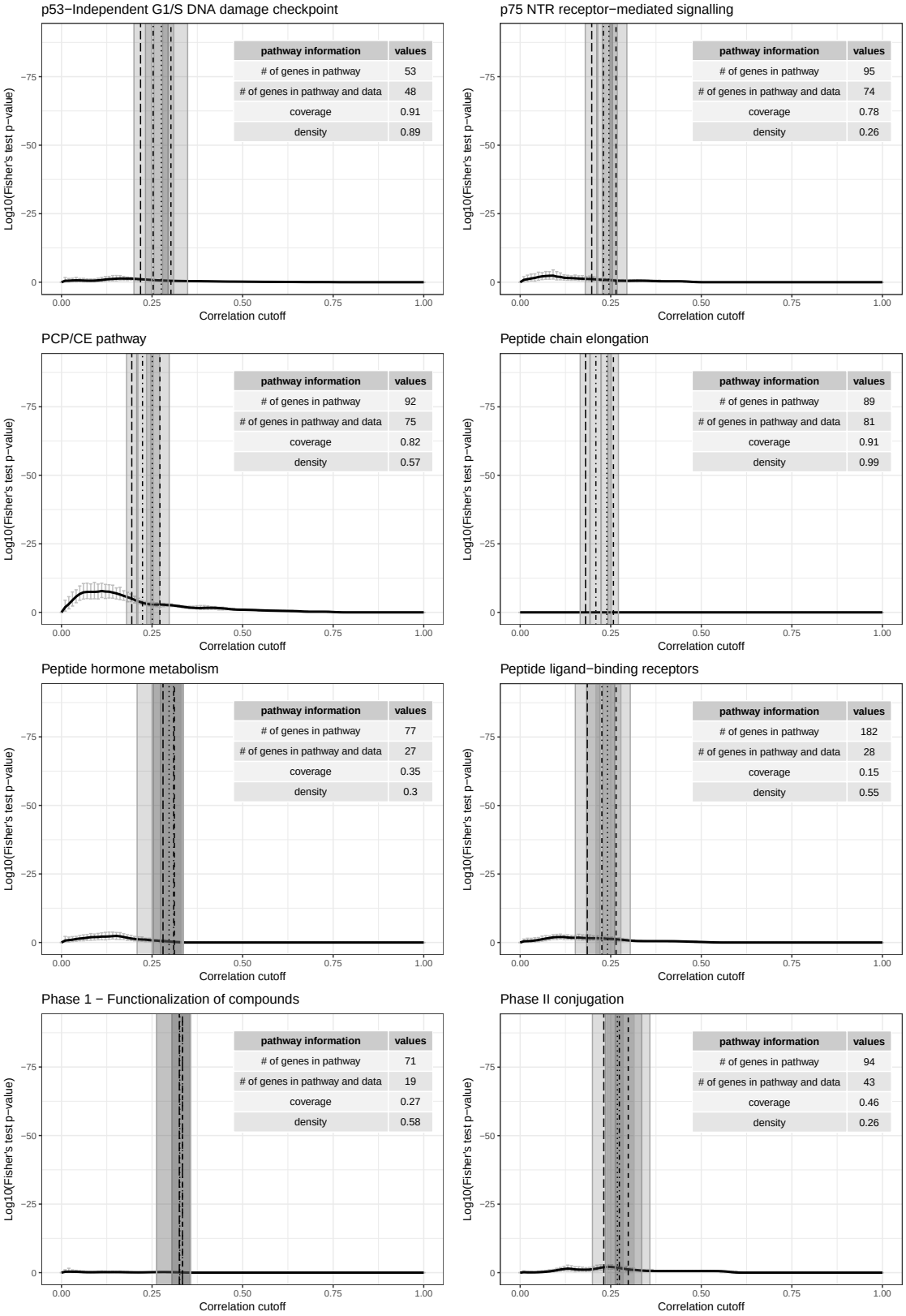

Negative regulation of the PI3K/AKT network
pathway information | values
# of genes in pathway | 80
# of genes in pathway and data | 47
coverage | 0.59
density | 0.6
Log10(Fisher's test p−value)
Correlation cutoff


Neuronal System
pathway information | values
# of genes in pathway | 335
# of genes in pathway and data | 124
coverage | 0.37
density | 0.23
Log10(Fisher's test p−value)
Correlation cutoff

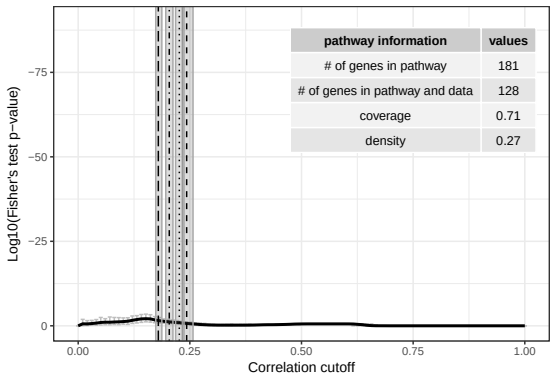

Neurotransmitter Receptor Binding And Downstream Transmission I
NGF signalling via TRKA from the plasma membrane
NIK−−>noncanonical NF−kB signaling
Nonhomologous End−Joining (NHEJ)
Nonsense−Mediated Decay (NMD)
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction (
Nonsense Mediated Decay (NMD) independent of the Exon Junctio
NoRC negatively regulates rRNA expression
pathway information
values
# of genes in pathway
# of genes in pathway and data
coverage
density
136
57
0.42
0.5
365
258
0.71
0.35
59
53
0.9
0.89
69
36
0.52
0.77
116
104
0.9
0.97
116
104
0.9
0.97
96
88
0.92
0.99
105
56
0.53
0.65
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75


NRAGE signals death through JNK
pathway information | values
# of genes in pathway | 59
# of genes in pathway and data | 42
coverage | 0.71
density | 0.25
Nucleosome assembly
pathway information | values
# of genes in pathway | 74
# of genes in pathway and data | 32
coverage | 0.43
density | 0.88
Nucleotide−binding domain, leucine rich repeat containing receptor
pathway information | values
# of genes in pathway | 50
# of genes in pathway and data | 43
coverage | 0.86
density | 0.7
Nucleotide Excision Repair
pathway information | values
# of genes in pathway | 111
# of genes in pathway and data | 104
coverage | 0.94
density | 0.66
Nucleotide metabolism
pathway information | values
# of genes in pathway | 81
# of genes in pathway and data | 63
coverage | 0.78
density | 0.69
O−linked glycosylation
pathway information | values
# of genes in pathway | 113
# of genes in pathway and data | 44
coverage | 0.39
density | 0.27
O−linked glycosylation of mucins
pathway information | values
# of genes in pathway | 65
# of genes in pathway and data | 27
coverage | 0.42
density | 0.43
Oncogenic MAPK signaling
pathway information | values
# of genes in pathway | 61
# of genes in pathway and data | 49
coverage | 0.8
density | 0.37
Log10(Fisher's test p−value)
Correlation cutoff

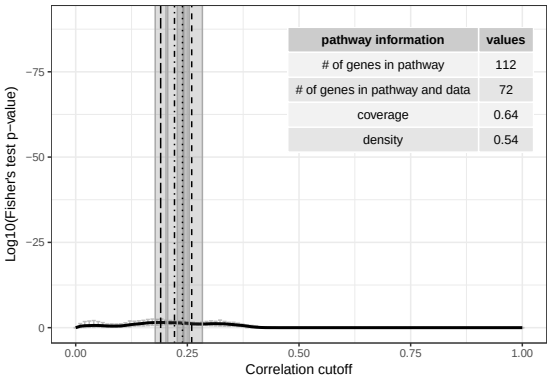

Opioid Signalling
pathway information | values
# of genes in pathway | 83
# of genes in pathway and data | 54
coverage | 0.65
density | 0.65
Orc1 removal from chromatin
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 59
coverage | 0.83
density | 0.82
Organelle biogenesis and maintenance
pathway information | values
# of genes in pathway | 329
# of genes in pathway and data | 282
coverage | 0.86
density | 0.27
Oxidative Stress Induced Senescence
pathway information | values
# of genes in pathway | 124
# of genes in pathway and data | 68
coverage | 0.55
density | 0.54
Oxygen−dependent proline hydroxylation of Hypoxia−inducible Fact
pathway information | values
# of genes in pathway | 66
# of genes in pathway and data | 55
coverage | 0.83
density | 0.78
p53−Dependent G1 DNA Damage Response
pathway information | values
# of genes in pathway | 64
# of genes in pathway and data | 56
coverage | 0.88
density | 0.85
p53−Dependent G1/S DNA damage checkpoint
pathway information | values
# of genes in pathway | 64
# of genes in pathway and data | 56
coverage | 0.88
density | 0.85
p53−Independent DNA Damage Response
pathway information | values
# of genes in pathway | 53
# of genes in pathway and data | 48
coverage | 0.91
density | 0.89
Log10(Fisher's test p−value)
Correlation cutoff


p53−Independent G1/S DNA damage checkpoint
pathway information | values
# of genes in pathway | 53
# of genes in pathway and data | 48
coverage | 0.91
density | 0.89
p75 NTR receptor−mediated signalling
pathway information | values
# of genes in pathway | 95
# of genes in pathway and data | 74
coverage | 0.78
density | 0.26
PCP/CE pathway
pathway information | values
# of genes in pathway | 92
# of genes in pathway and data | 75
coverage | 0.82
density | 0.57
Peptide chain elongation
pathway information | values
# of genes in pathway | 89
# of genes in pathway and data | 81
coverage | 0.91
density | 0.99
Peptide hormone metabolism
pathway information | values
# of genes in pathway | 77
# of genes in pathway and data | 27
coverage | 0.35
density | 0.3
Peptide ligand−binding receptors
pathway information | values
# of genes in pathway | 182
# of genes in pathway and data | 28
coverage | 0.15
density | 0.55
Phase 1 − Functionalization of compounds
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 19
coverage | 0.27
density | 0.58
Phase II conjugation
pathway information | values
# of genes in pathway | 94
# of genes in pathway and data | 43
coverage | 0.46
density | 0.26
Log10(Fisher's test p−value)
Correlation cutoff

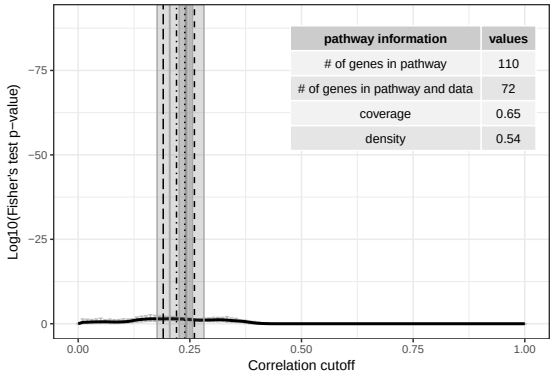
Phospholipid metabolism



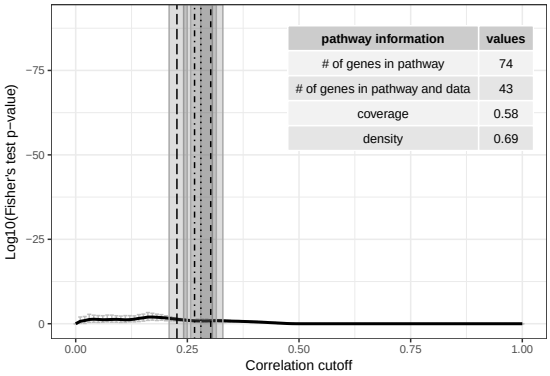
PI3K/AKT activation



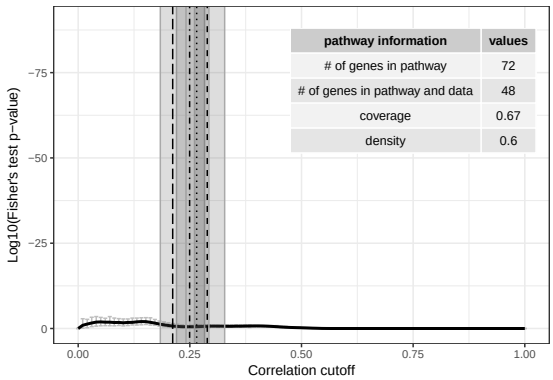
PI3K/AKT Signaling



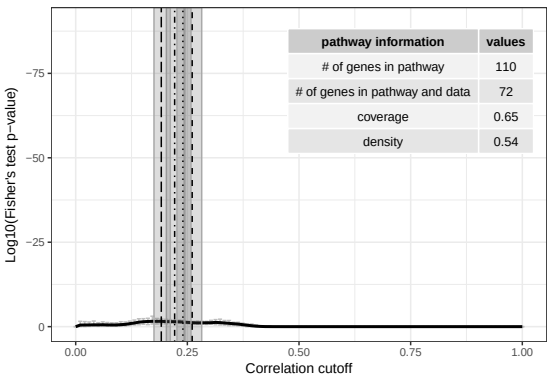
PI3K/AKT Signaling in Cancer



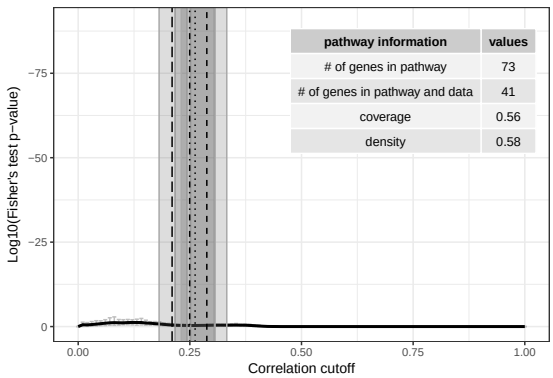
PI3K Cascade



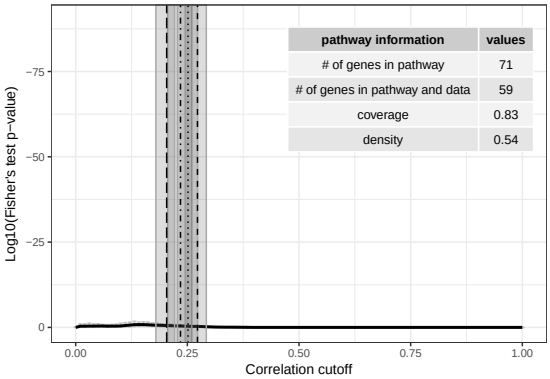
PI3K events in ERBB4 signaling

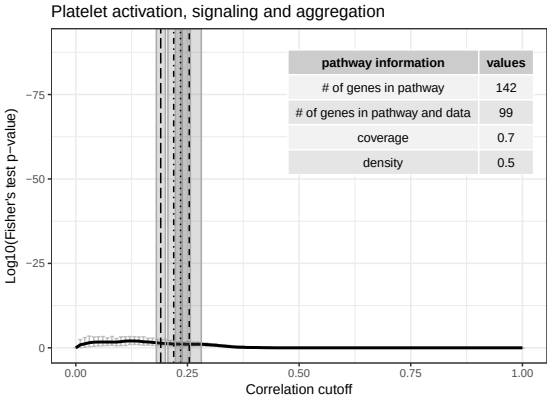


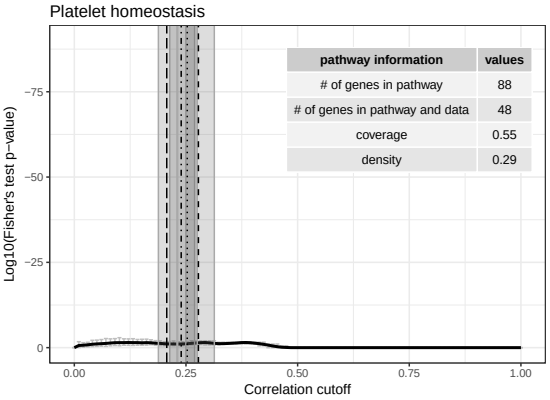
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling

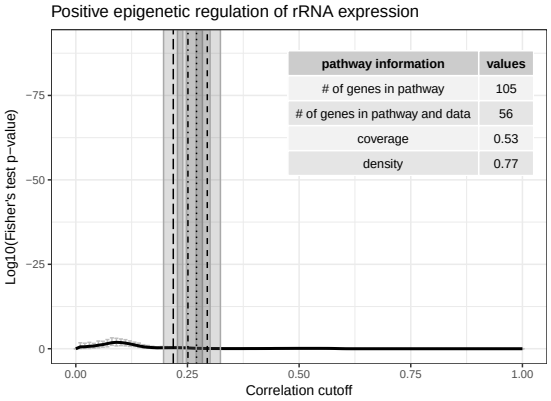


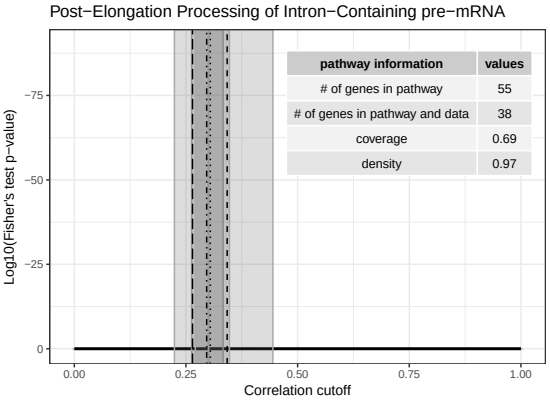
PI Metabolism

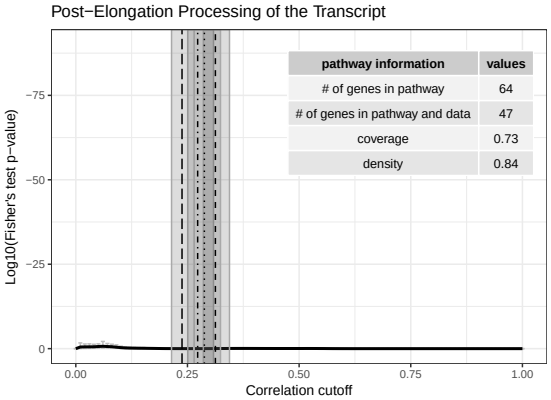


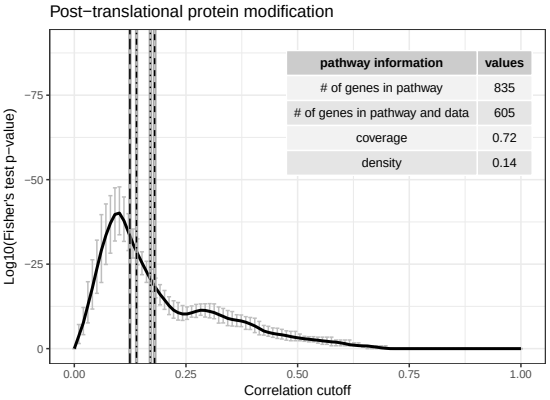

Platelet activation, signaling and aggregation
pathway information | values
# of genes in pathway | 142
# of genes in pathway and data | 99
coverage | 0.7
density | 0.5
Log10(Fisher's test p−value)
Correlation cutoff

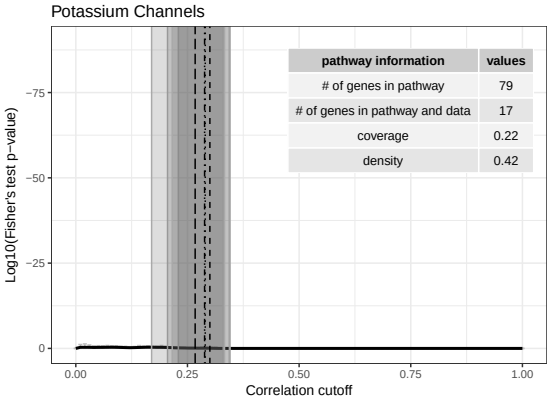

Platelet homeostasis
pathway information | values
# of genes in pathway | 88
# of genes in pathway and data | 48
coverage | 0.55
density | 0.29
Log10(Fisher's test p−value)
Correlation cutoff

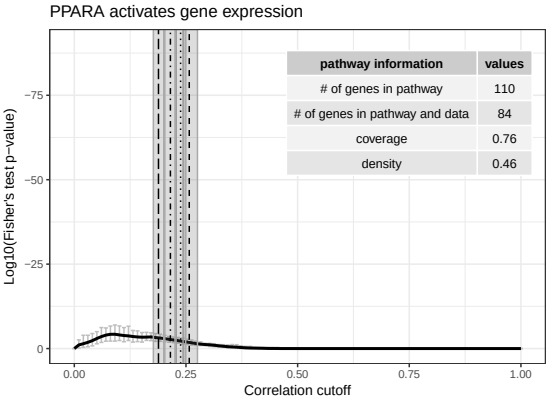

Positive epigenetic regulation of rRNA expression
pathway information | values
# of genes in pathway | 105
# of genes in pathway and data | 56
coverage | 0.53
density | 0.77
Log10(Fisher's test p−value)
Correlation cutoff

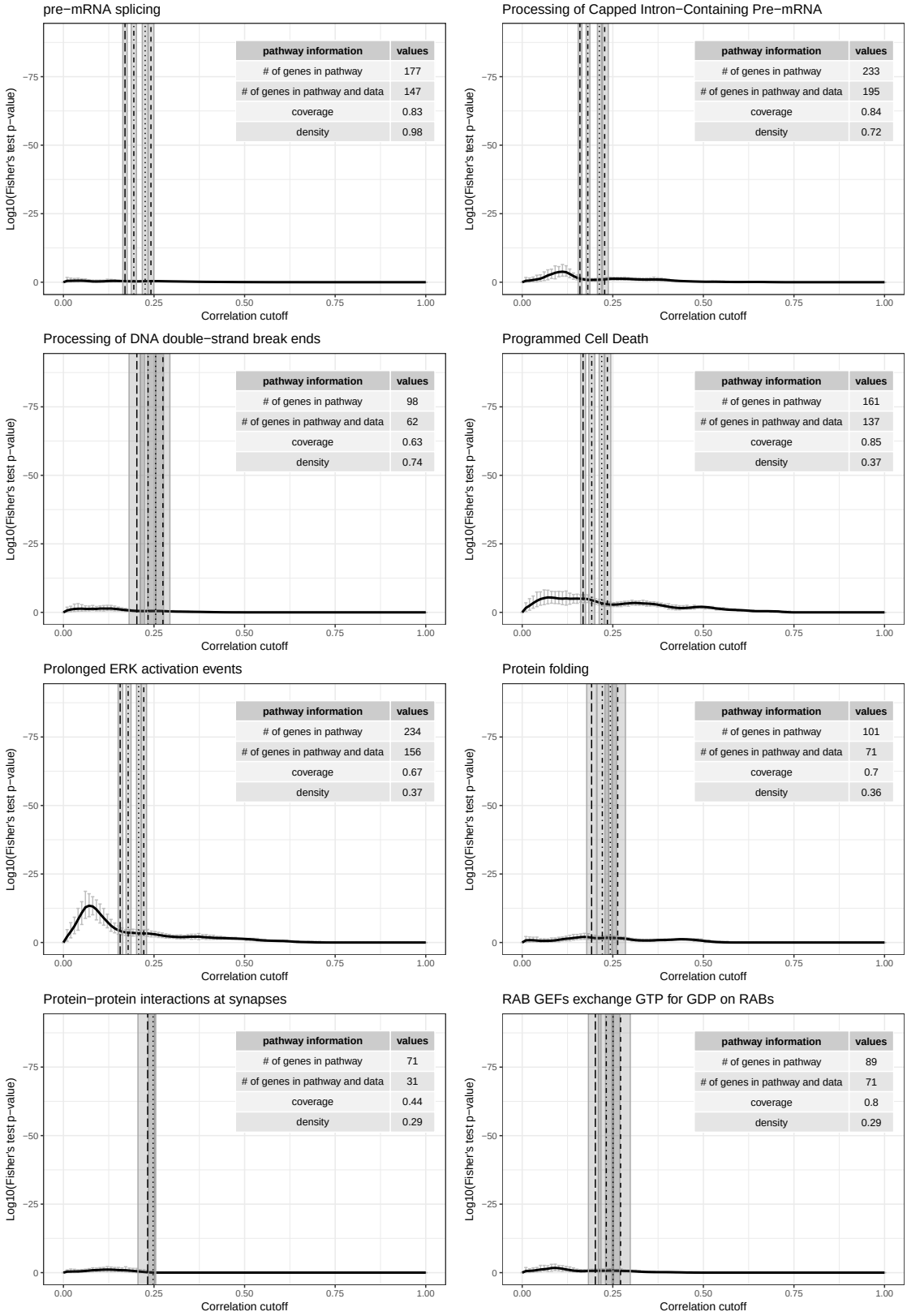

Post−Elongation Processing of Intron−Containing pre−mRNA
pathway information | values
# of genes in pathway | 55
# of genes in pathway and data | 38
coverage | 0.69
density | 0.97
Log10(Fisher's test p−value)
Correlation cutoff


Post−Elongation Processing of the Transcript
pathway information | values
# of genes in pathway | 64
# of genes in pathway and data | 47
coverage | 0.73
density | 0.84
Log10(Fisher's test p−value)
Correlation cutoff

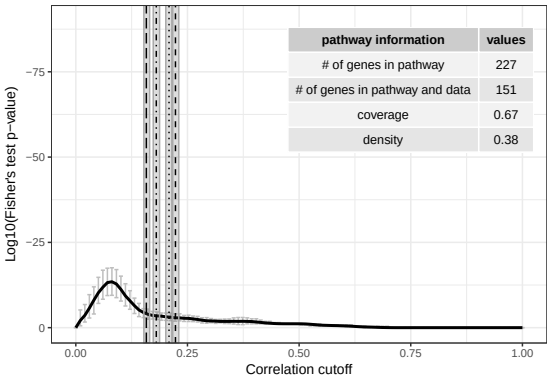

Post−translational protein modification
pathway information | values
# of genes in pathway | 835
# of genes in pathway and data | 605
coverage | 0.72
density | 0.14
Log10(Fisher's test p−value)
Correlation cutoff


Potassium Channels
pathway information | values
# of genes in pathway | 79
# of genes in pathway and data | 17
coverage | 0.22
density | 0.42
Log10(Fisher's test p−value)
Correlation cutoff

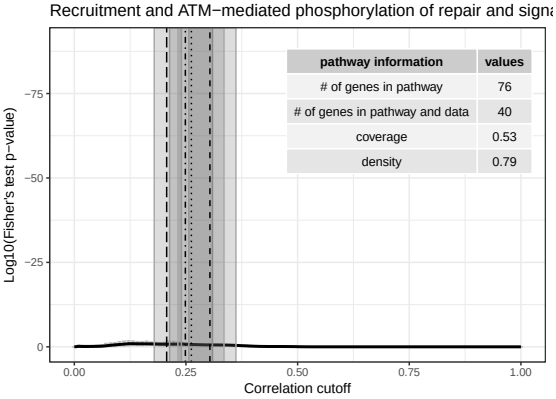

PPARA activates gene expression
pathway information | values
# of genes in pathway | 110
# of genes in pathway and data | 84
coverage | 0.76
density | 0.46
Log10(Fisher's test p−value)
Correlation cutoff


pre−mRNA splicing
pathway information | values
# of genes in pathway | 177
# of genes in pathway and data | 147
coverage | 0.83
density | 0.98
Processing of Capped Intron−Containing Pre−mRNA
pathway information | values
# of genes in pathway | 233
# of genes in pathway and data | 195
coverage | 0.84
density | 0.72
Processing of DNA double−strand break ends
pathway information | values
# of genes in pathway | 98
# of genes in pathway and data | 62
coverage | 0.63
density | 0.74
Programmed Cell Death
pathway information | values
# of genes in pathway | 161
# of genes in pathway and data | 137
coverage | 0.85
density | 0.37
Prolonged ERK activation events
pathway information | values
# of genes in pathway | 234
# of genes in pathway and data | 156
coverage | 0.67
density | 0.37
Protein folding
pathway information | values
# of genes in pathway | 101
# of genes in pathway and data | 71
coverage | 0.7
density | 0.36
Protein−protein interactions at synapses
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 31
coverage | 0.44
density | 0.29
RAB GEFs exchange GTP for GDP on RABs
pathway information | values
# of genes in pathway | 89
# of genes in pathway and data | 71
coverage | 0.8
density | 0.29
Log10(Fisher's test p−value)
Correlation cutoff

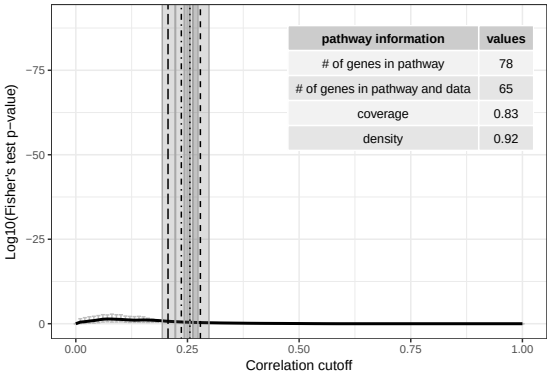
RAF/MAP kinase cascade



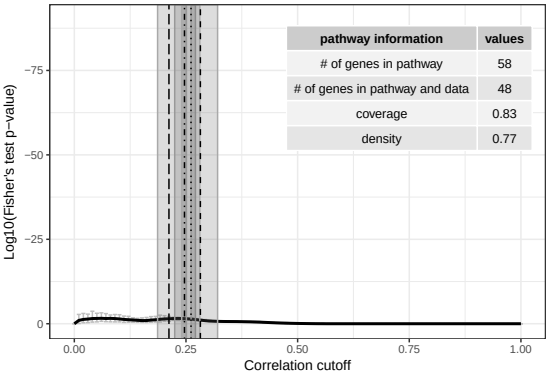
Recruitment and ATM−mediated phosphorylation of repair and signa



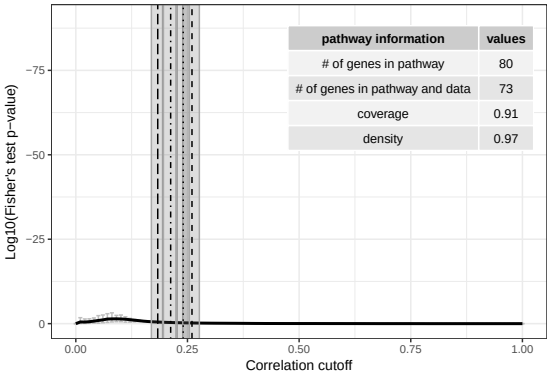
Recruitment of mitotic centrosome proteins and complexes



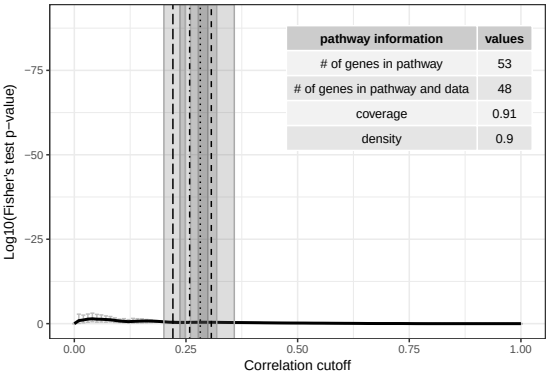
Regulation of actin dynamics for phagocytic cup formation



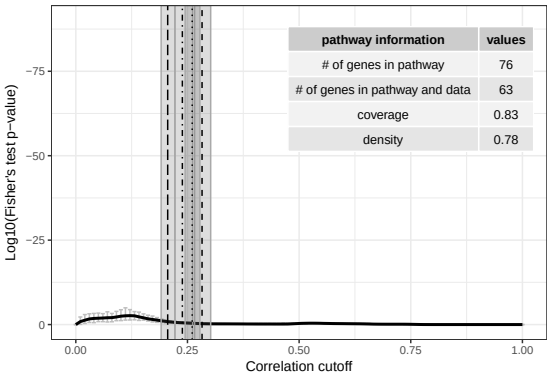
Regulation of APC/C activators between G1/S and early anaphase



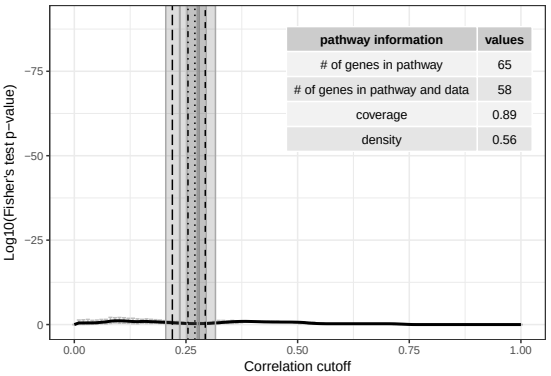
Regulation of Apoptosis

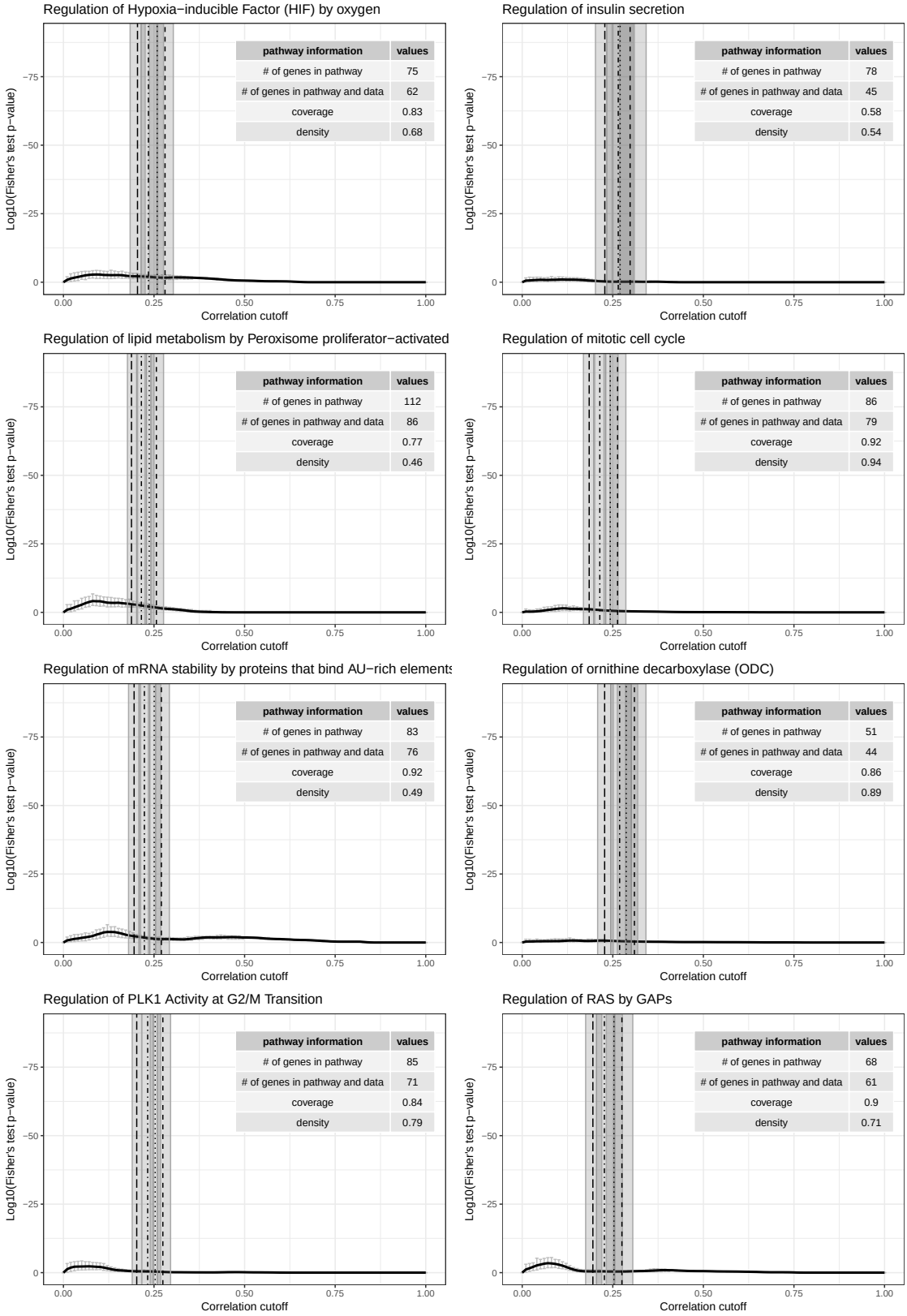


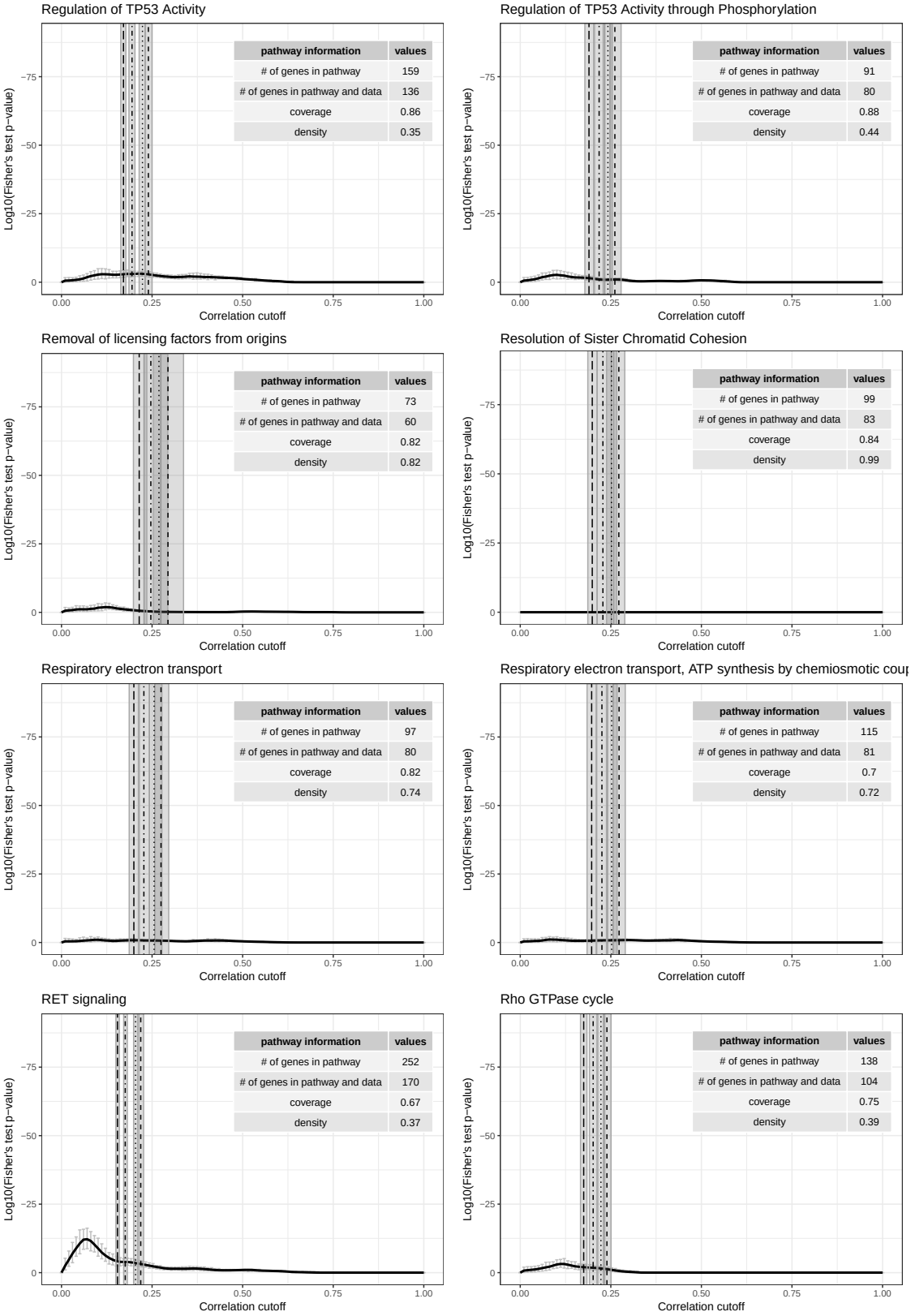
Regulation of DNA replication

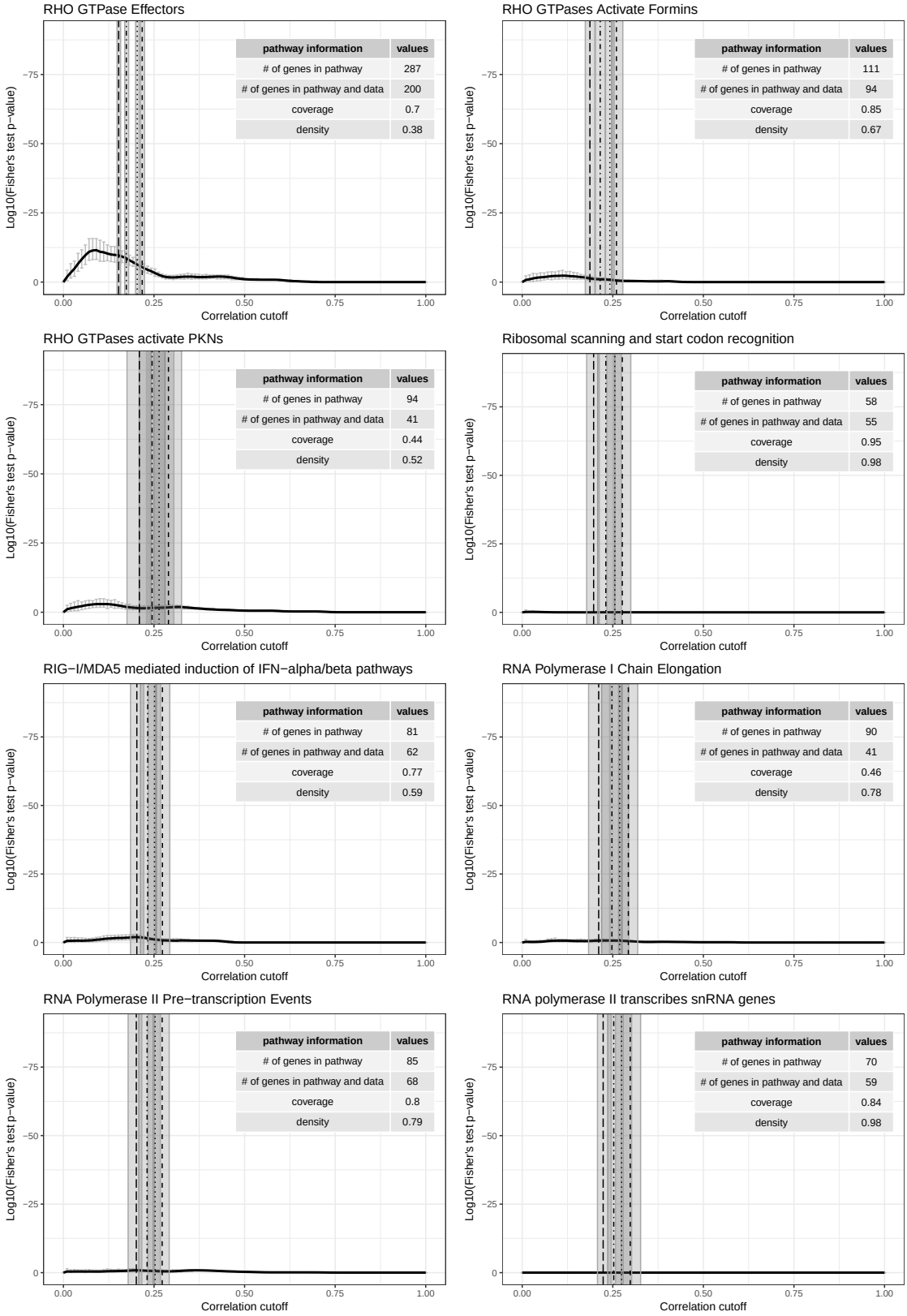


Regulation of HSF1−mediated heat shock response

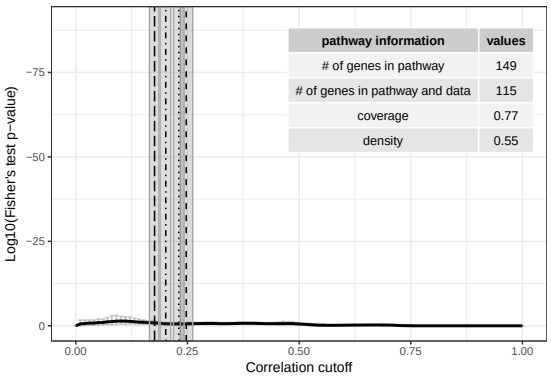



Regulation of Hypoxia−inducible Factor (HIF) by oxygen
pathway information | values
# of genes in pathway | 75
# of genes in pathway and data | 62
coverage | 0.83
density | 0.68
Regulation of insulin secretion
pathway information | values
# of genes in pathway | 78
# of genes in pathway and data | 45
coverage | 0.58
density | 0.54
Regulation of lipid metabolism by Peroxisome proliferator−activated
pathway information | values
# of genes in pathway | 112
# of genes in pathway and data | 86
coverage | 0.77
density | 0.46
Regulation of mitotic cell cycle
pathway information | values
# of genes in pathway | 86
# of genes in pathway and data | 79
coverage | 0.92
density | 0.94
Regulation of mRNA stability by proteins that bind AU−rich elements
pathway information | values
# of genes in pathway | 83
# of genes in pathway and data | 76
coverage | 0.92
density | 0.49
Regulation of ornithine decarboxylase (ODC)
pathway information | values
# of genes in pathway | 51
# of genes in pathway and data | 44
coverage | 0.86
density | 0.89
Regulation of PLK1 Activity at G2/M Transition
pathway information | values
# of genes in pathway | 85
# of genes in pathway and data | 71
coverage | 0.84
density | 0.79
Regulation of RAS by GAPs
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 61
coverage | 0.9
density | 0.71
Log10(Fisher's test p−value)
Correlation cutoff

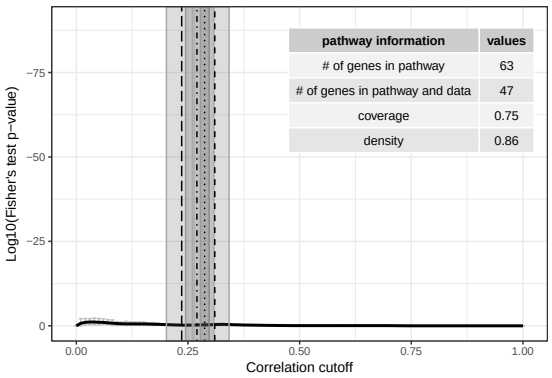

Regulation of TP53 Activity
pathway information | values
# of genes in pathway | 159
# of genes in pathway and data | 136
coverage | 0.86
density | 0.35
Regulation of TP53 Activity through Phosphorylation
pathway information | values
# of genes in pathway | 91
# of genes in pathway and data | 80
coverage | 0.88
density | 0.44
Removal of licensing factors from origins
pathway information | values
# of genes in pathway | 73
# of genes in pathway and data | 60
coverage | 0.82
density | 0.82
Resolution of Sister Chromatid Cohesion
pathway information | values
# of genes in pathway | 99
# of genes in pathway and data | 83
coverage | 0.84
density | 0.99
Respiratory electron transport
pathway information | values
# of genes in pathway | 97
# of genes in pathway and data | 80
coverage | 0.82
density | 0.74
Respiratory electron transport, ATP synthesis by chemiosmotic coup
pathway information | values
# of genes in pathway | 115
# of genes in pathway and data | 81
coverage | 0.7
density | 0.72
RET signaling
pathway information | values
# of genes in pathway | 252
# of genes in pathway and data | 170
coverage | 0.67
density | 0.37
Rho GTPase cycle
pathway information | values
# of genes in pathway | 138
# of genes in pathway and data | 104
coverage | 0.75
density | 0.39
Log10(Fisher's test p−value)
Correlation cutoff


RHO GTPase Effectors
pathway information | values
# of genes in pathway | 287
# of genes in pathway and data | 200
coverage | 0.7
density | 0.38
RHO GTPases Activate Formins
pathway information | values
# of genes in pathway | 111
# of genes in pathway and data | 94
coverage | 0.85
density | 0.67
RHO GTPases activate PKNs
pathway information | values
# of genes in pathway | 94
# of genes in pathway and data | 41
coverage | 0.44
density | 0.52
Ribosomal scanning and start codon recognition
pathway information | values
# of genes in pathway | 58
# of genes in pathway and data | 55
coverage | 0.95
density | 0.98
RIG−I/MDA5 mediated induction of IFN−alpha/beta pathways
pathway information | values
# of genes in pathway | 81
# of genes in pathway and data | 62
coverage | 0.77
density | 0.59
RNA Polymerase I Chain Elongation
pathway information | values
# of genes in pathway | 90
# of genes in pathway and data | 41
coverage | 0.46
density | 0.78
RNA Polymerase II Pre−transcription Events
pathway information | values
# of genes in pathway | 85
# of genes in pathway and data | 68
coverage | 0.8
density | 0.79
RNA polymerase II transcribes snRNA genes
pathway information | values
# of genes in pathway | 70
# of genes in pathway and data | 59
coverage | 0.84
density | 0.98
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

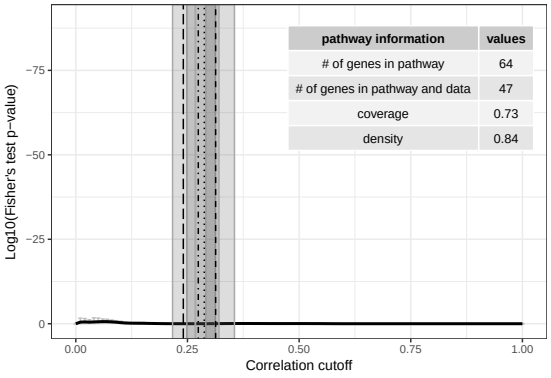
RNA Polymerase II Transcription



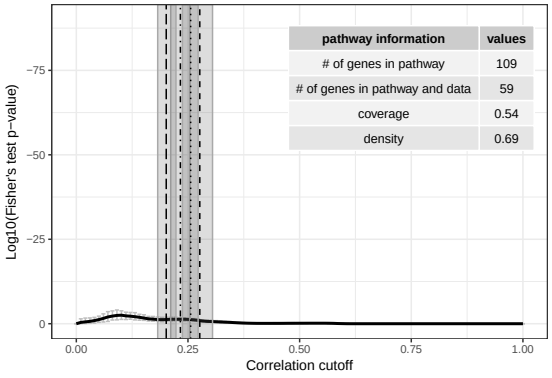
RNA Polymerase II Transcription Elongation



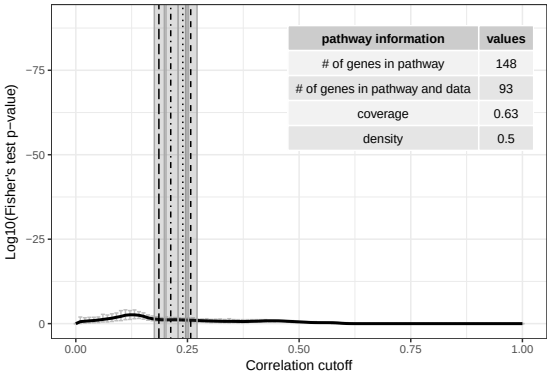
RNA Polymerase II Transcription Termination



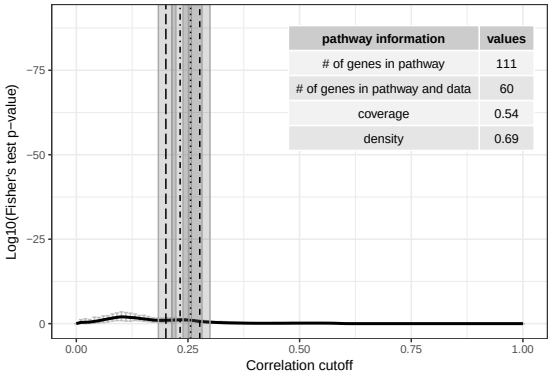
RNA Polymerase I Promoter Clearance



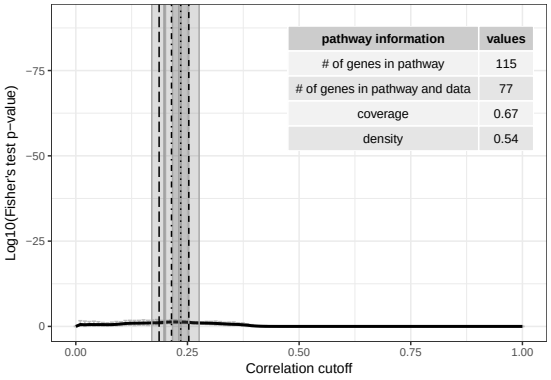
RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcri



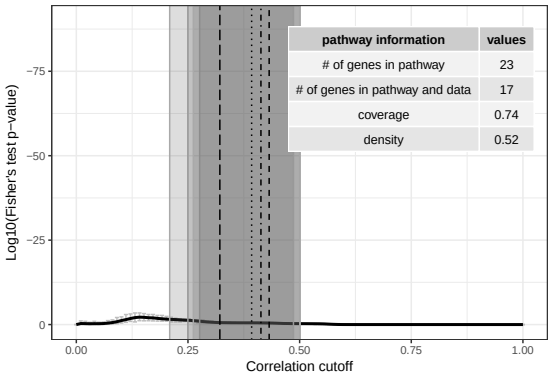
RNA Polymerase I Transcription



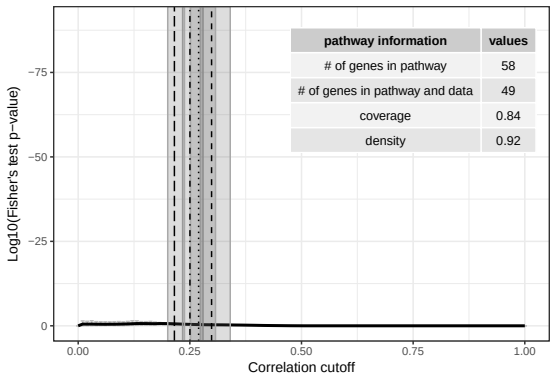
Role of LAT2/NTAL/LAB on calcium mobilization



Role of phospholipids in phagocytosis

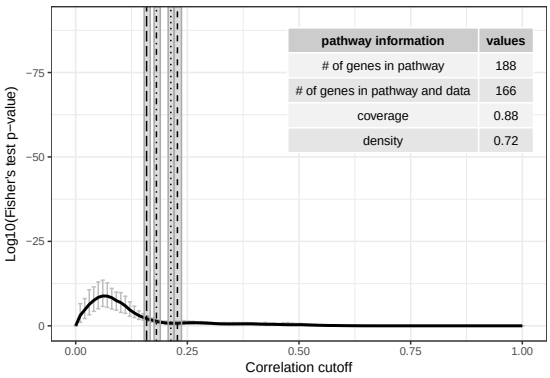


rRNA modification in the nucleus and cytosol



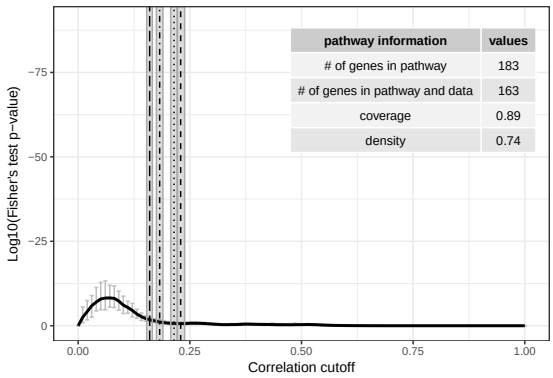
pathway information	values
# of genes in pathway	58
# of genes in pathway and data	49
coverage	0.84
density	0.92

rRNA processing



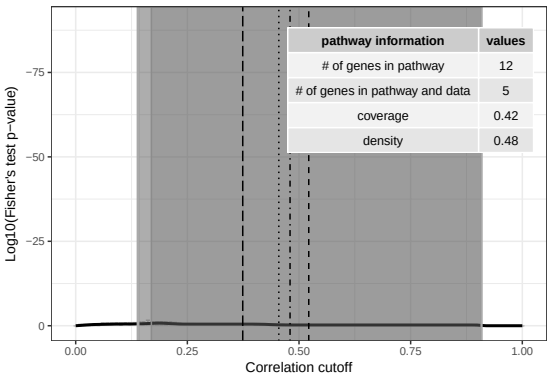
pathway information	values
# of genes in pathway	188
# of genes in pathway and data	166
coverage	0.88
density	0.72

rRNA processing in the nucleus and cytosol



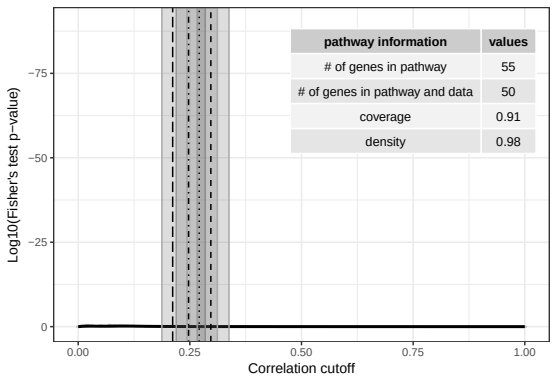
pathway information	values
# of genes in pathway	183
# of genes in pathway and data	163
coverage	0.89
density	0.74

Scavenging of heme from plasma



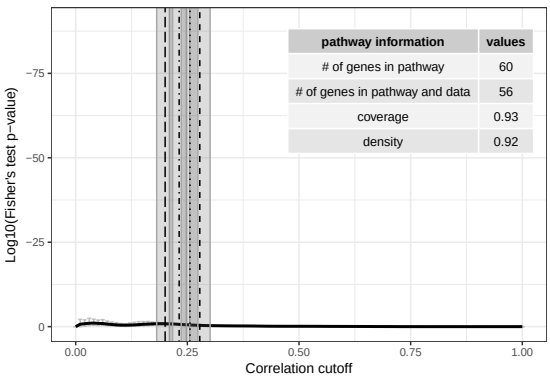
pathway information	values
# of genes in pathway	12
# of genes in pathway and data	5
coverage	0.42
density	0.48

SCF−beta−TrCP mediated degradation of Emi1



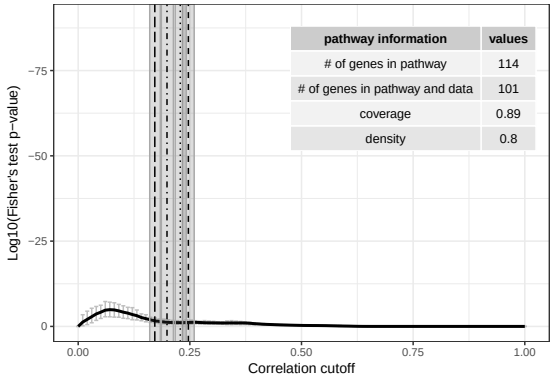
pathway information	values
# of genes in pathway	55
# of genes in pathway and data	50
coverage	0.91
density	0.98

SCF(Skp2)−mediated degradation of p27/p21



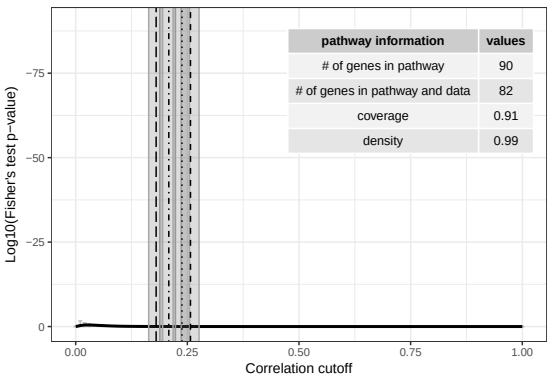
pathway information	values
# of genes in pathway	60
# of genes in pathway and data	56
coverage	0.93
density	0.92

Selenoamino acid metabolism

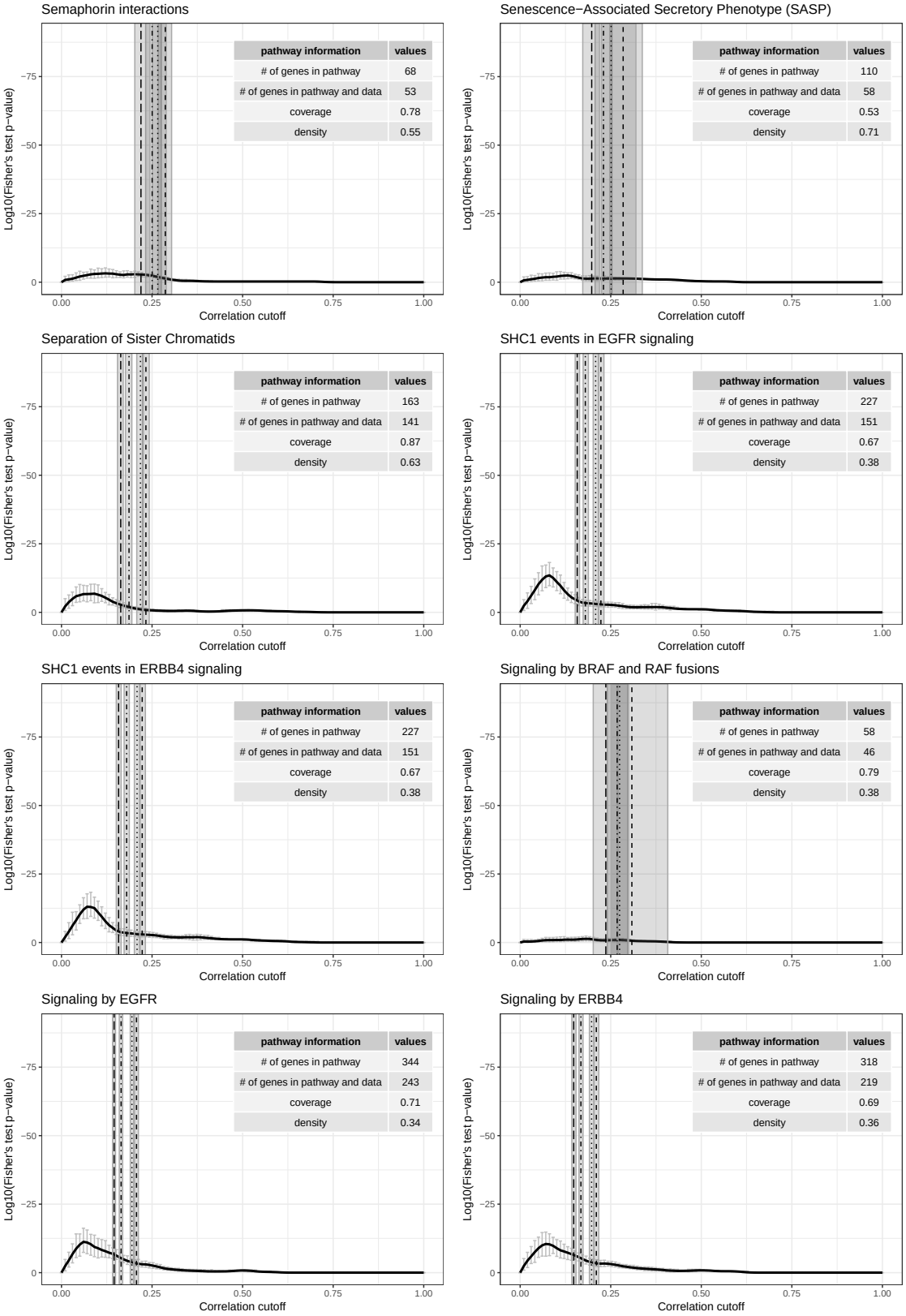


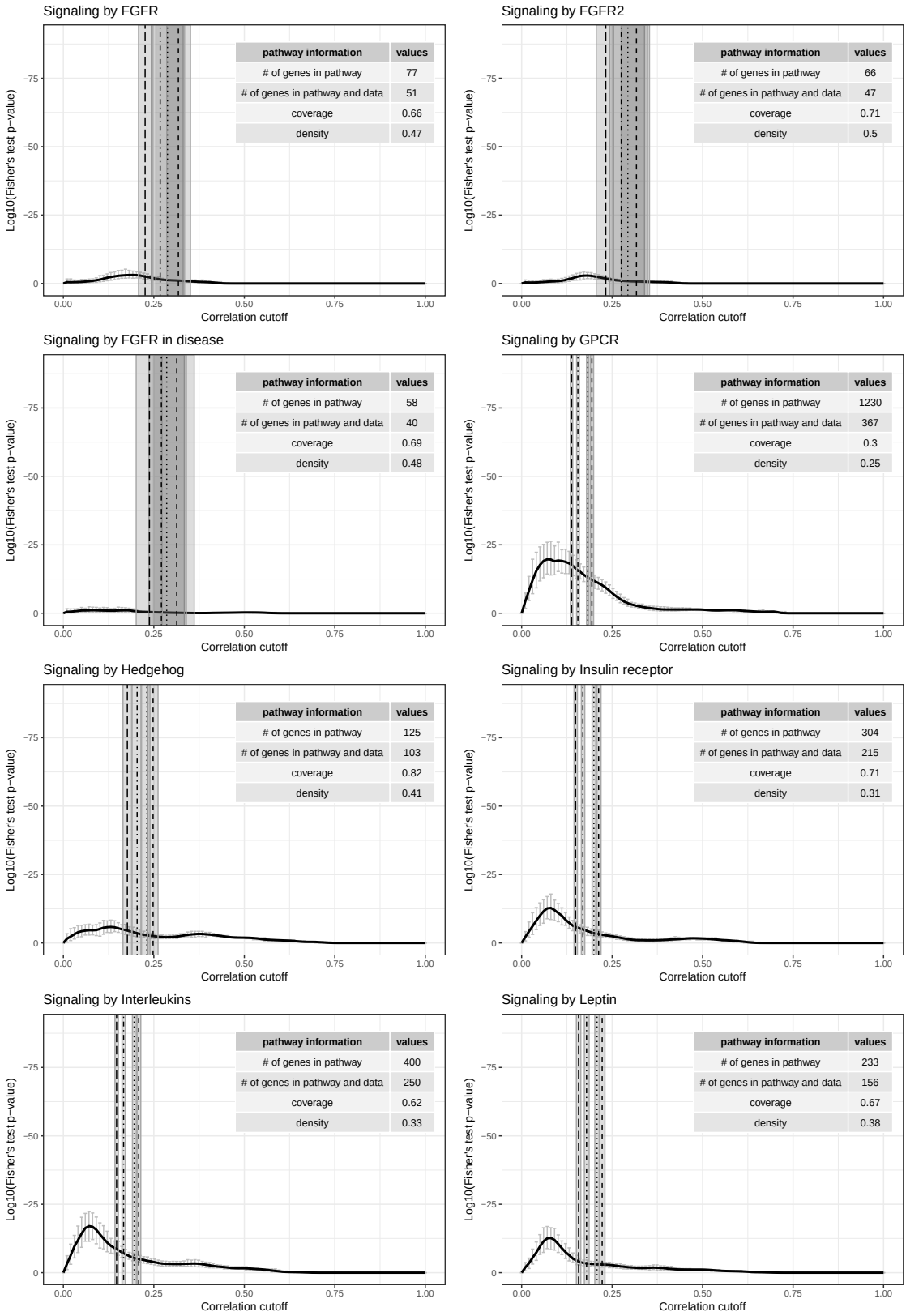
pathway information	values
# of genes in pathway	114
# of genes in pathway and data	101
coverage	0.89
density	0.8

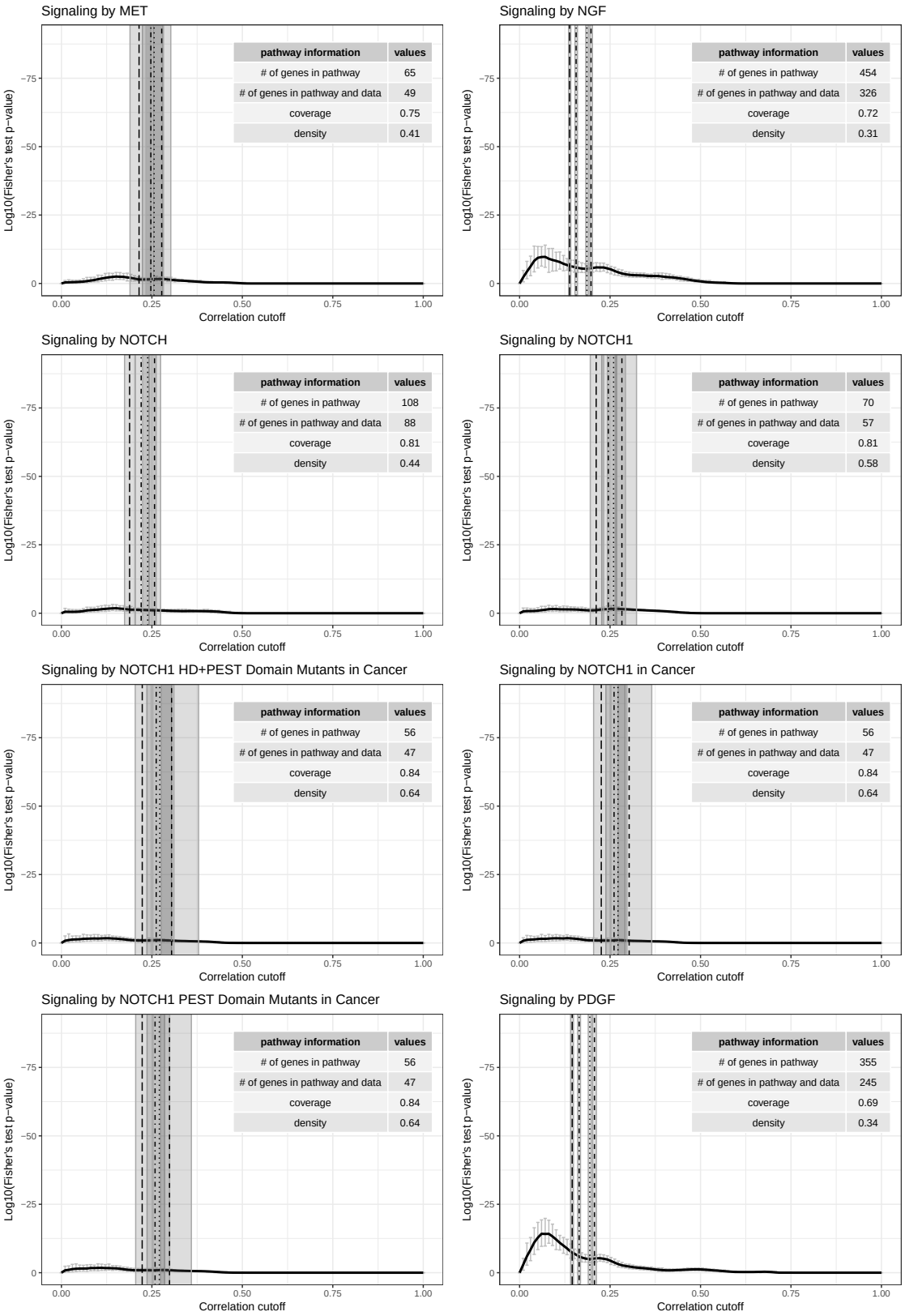
Selenocysteine synthesis

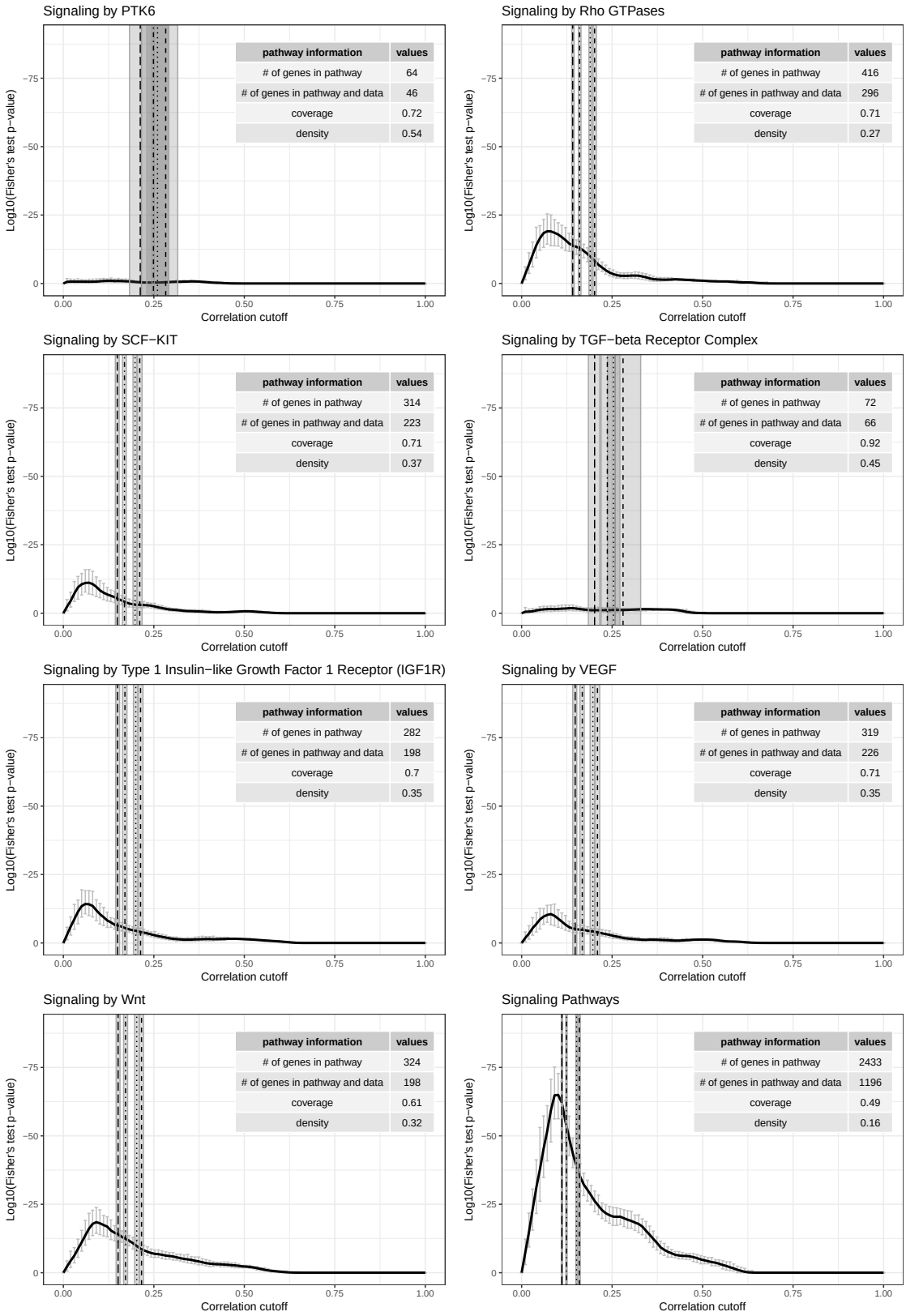


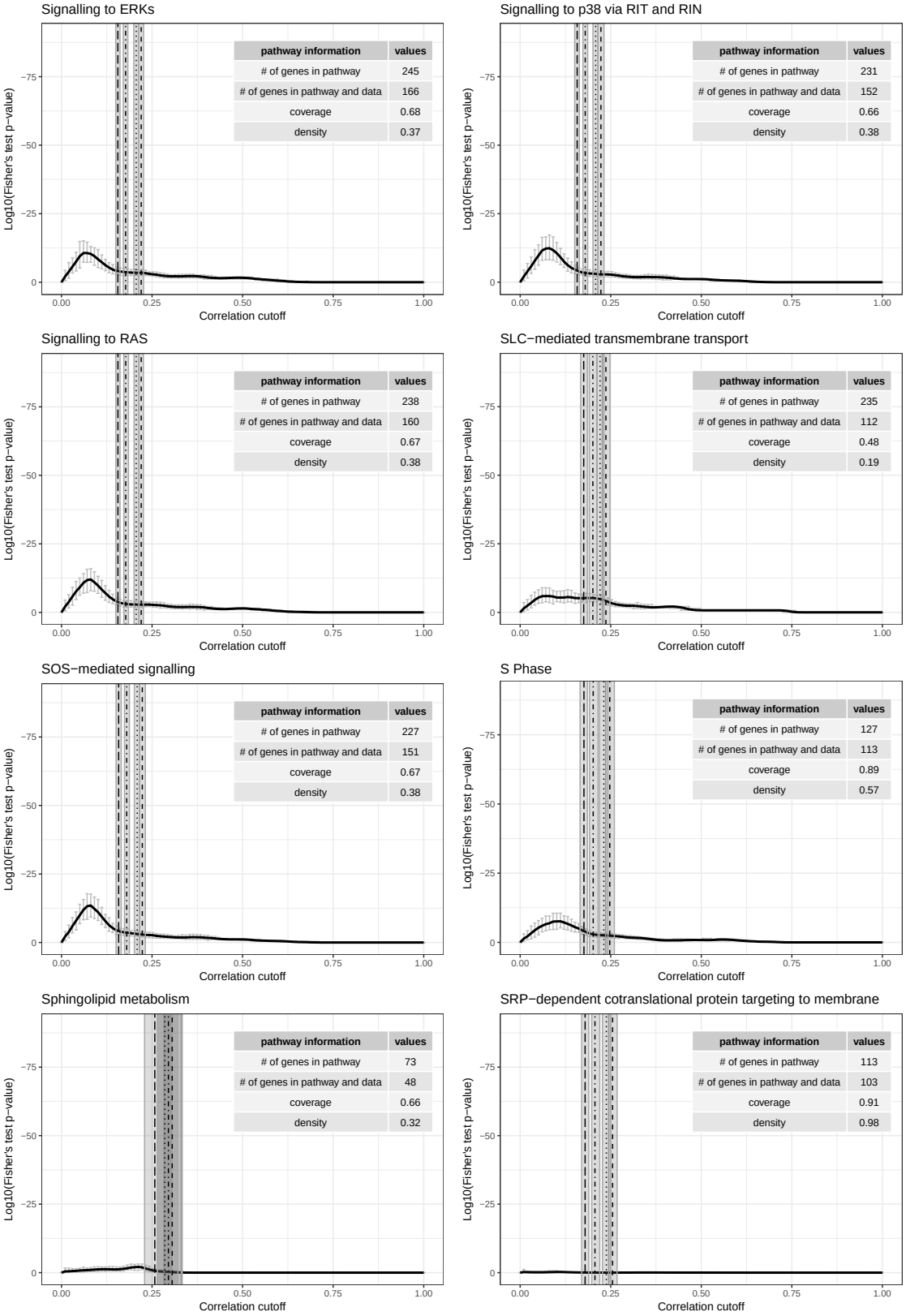
pathway information	values
# of genes in pathway	90
# of genes in pathway and data	82
coverage	0.91
density	0.99

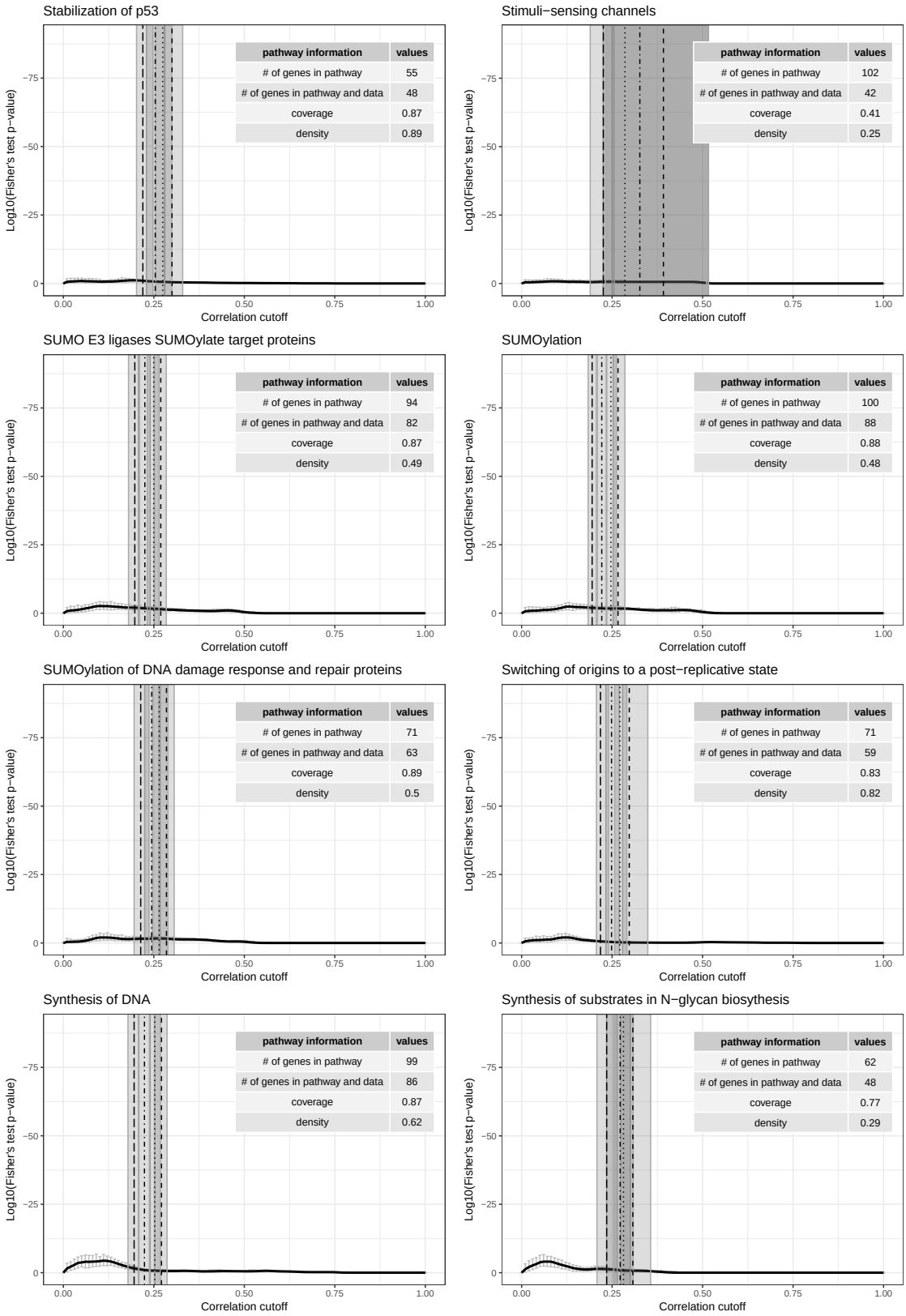

Semaphorin interactions
pathway information | values
# of genes in pathway | 68
# of genes in pathway and data | 53
coverage | 0.78
density | 0.55
Senescence−Associated Secretory Phenotype (SASP)
pathway information | values
# of genes in pathway | 110
# of genes in pathway and data | 58
coverage | 0.53
density | 0.71
Separation of Sister Chromatids
pathway information | values
# of genes in pathway | 163
# of genes in pathway and data | 141
coverage | 0.87
density | 0.63
SHC1 events in EGFR signaling
pathway information | values
# of genes in pathway | 227
# of genes in pathway and data | 151
coverage | 0.67
density | 0.38
SHC1 events in ERBB4 signaling
pathway information | values
# of genes in pathway | 227
# of genes in pathway and data | 151
coverage | 0.67
density | 0.38
Signaling by BRAF and RAF fusions
pathway information | values
# of genes in pathway | 58
# of genes in pathway and data | 46
coverage | 0.79
density | 0.38
Signaling by EGFR
pathway information | values
# of genes in pathway | 344
# of genes in pathway and data | 243
coverage | 0.71
density | 0.34
Signaling by ERBB4
pathway information | values
# of genes in pathway | 318
# of genes in pathway and data | 219
coverage | 0.69
density | 0.36
Log10(Fisher's test p−value)
Correlation cutoff

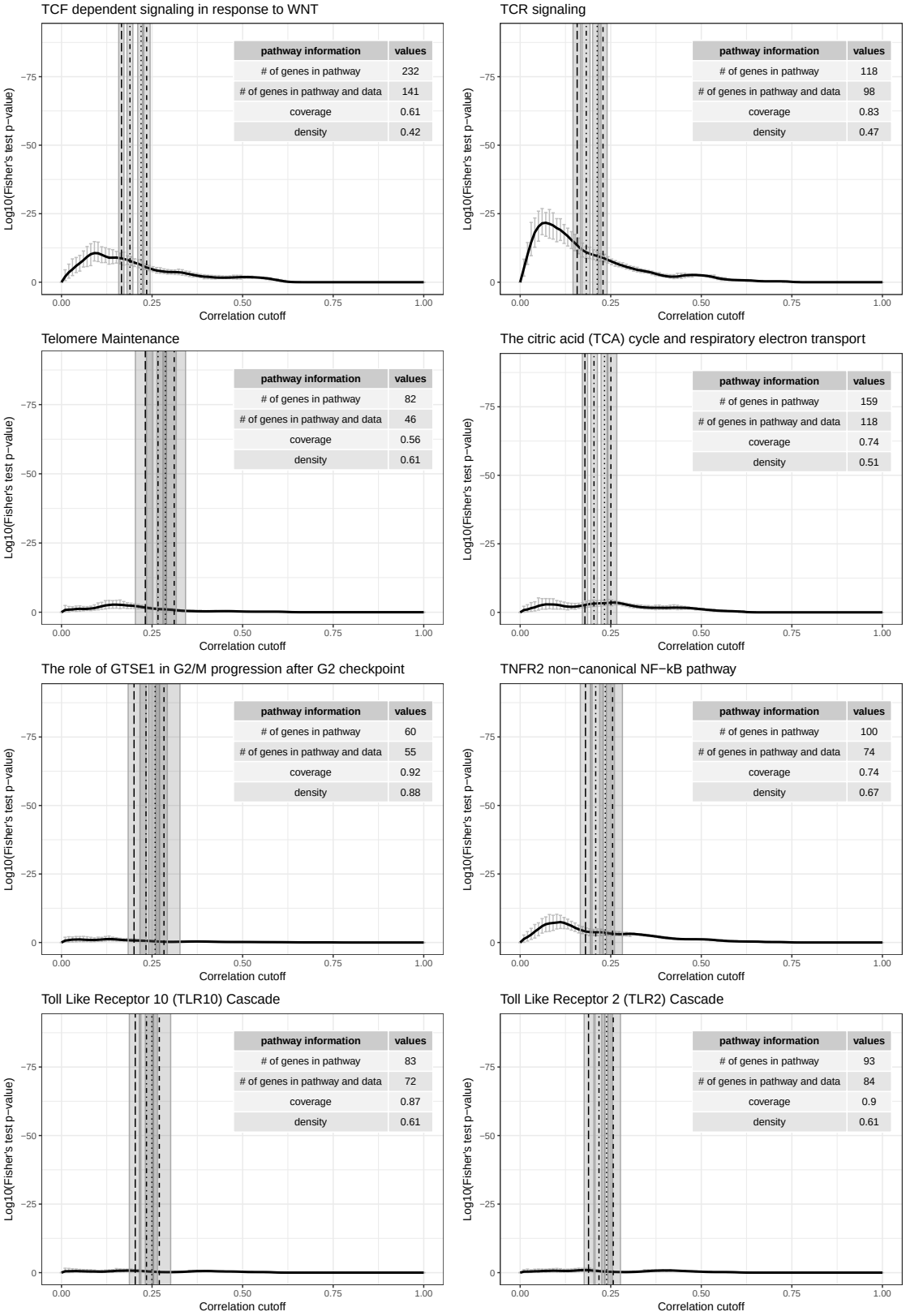

Signaling by FGFR
pathway information | values
# of genes in pathway | 77
# of genes in pathway and data | 51
coverage | 0.66
density | 0.47
Signaling by FGFR2
pathway information | values
# of genes in pathway | 66
# of genes in pathway and data | 47
coverage | 0.71
density | 0.5
Signaling by FGFR in disease
pathway information | values
# of genes in pathway | 58
# of genes in pathway and data | 40
coverage | 0.69
density | 0.48
Signaling by GPCR
pathway information | values
# of genes in pathway | 1230
# of genes in pathway and data | 367
coverage | 0.3
density | 0.25
Signaling by Hedgehog
pathway information | values
# of genes in pathway | 125
# of genes in pathway and data | 103
coverage | 0.82
density | 0.41
Signaling by Insulin receptor
pathway information | values
# of genes in pathway | 304
# of genes in pathway and data | 215
coverage | 0.71
density | 0.31
Signaling by Interleukins
pathway information | values
# of genes in pathway | 400
# of genes in pathway and data | 250
coverage | 0.62
density | 0.33
Signaling by Leptin
pathway information | values
# of genes in pathway | 233
# of genes in pathway and data | 156
coverage | 0.67
density | 0.38
Log10(Fisher's test p−value)
Correlation cutoff


Signaling by MET
pathway information | values
# of genes in pathway | 65
# of genes in pathway and data | 49
coverage | 0.75
density | 0.41
Signaling by NGF
pathway information | values
# of genes in pathway | 454
# of genes in pathway and data | 326
coverage | 0.72
density | 0.31
Signaling by NOTCH
pathway information | values
# of genes in pathway | 108
# of genes in pathway and data | 88
coverage | 0.81
density | 0.44
Signaling by NOTCH1
pathway information | values
# of genes in pathway | 70
# of genes in pathway and data | 57
coverage | 0.81
density | 0.58
Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer
pathway information | values
# of genes in pathway | 56
# of genes in pathway and data | 47
coverage | 0.84
density | 0.64
Signaling by NOTCH1 in Cancer
pathway information | values
# of genes in pathway | 56
# of genes in pathway and data | 47
coverage | 0.84
density | 0.64
Signaling by NOTCH1 PEST Domain Mutants in Cancer
pathway information | values
# of genes in pathway | 56
# of genes in pathway and data | 47
coverage | 0.84
density | 0.64
Signaling by PDGF
pathway information | values
# of genes in pathway | 355
# of genes in pathway and data | 245
coverage | 0.69
density | 0.34
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75

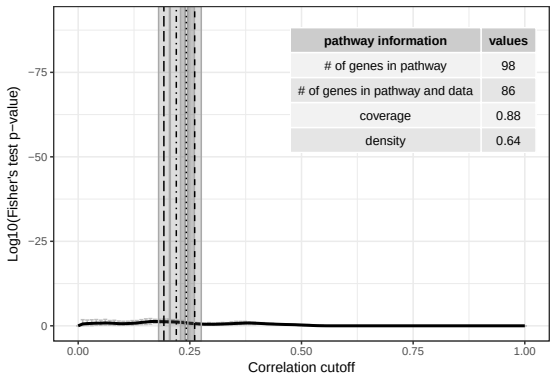

Signaling by PTK6
pathway information | values
# of genes in pathway | 64
# of genes in pathway and data | 46
coverage | 0.72
density | 0.54
Signaling by Rho GTPases
pathway information | values
# of genes in pathway | 416
# of genes in pathway and data | 296
coverage | 0.71
density | 0.27
Signaling by SCF−KIT
pathway information | values
# of genes in pathway | 314
# of genes in pathway and data | 223
coverage | 0.71
density | 0.37
Signaling by TGF−beta Receptor Complex
pathway information | values
# of genes in pathway | 72
# of genes in pathway and data | 66
coverage | 0.92
density | 0.45
Signaling by Type 1 Insulin−like Growth Factor 1 Receptor (IGF1R)
pathway information | values
# of genes in pathway | 282
# of genes in pathway and data | 198
coverage | 0.7
density | 0.35
Signaling by VEGF
pathway information | values
# of genes in pathway | 319
# of genes in pathway and data | 226
coverage | 0.71
density | 0.35
Signaling by Wnt
pathway information | values
# of genes in pathway | 324
# of genes in pathway and data | 198
coverage | 0.61
density | 0.32
Signaling Pathways
pathway information | values
# of genes in pathway | 2433
# of genes in pathway and data | 1196
coverage | 0.49
density | 0.16
Log10(Fisher's test p−value)
Correlation cutoff


Signalling to ERKs
pathway information | values
# of genes in pathway | 245
# of genes in pathway and data | 166
coverage | 0.68
density | 0.37
Signalling to p38 via RIT and RIN
pathway information | values
# of genes in pathway | 231
# of genes in pathway and data | 152
coverage | 0.66
density | 0.38
Signalling to RAS
pathway information | values
# of genes in pathway | 238
# of genes in pathway and data | 160
coverage | 0.67
density | 0.38
SLC−mediated transmembrane transport
pathway information | values
# of genes in pathway | 235
# of genes in pathway and data | 112
coverage | 0.48
density | 0.19
SOS−mediated signalling
pathway information | values
# of genes in pathway | 227
# of genes in pathway and data | 151
coverage | 0.67
density | 0.38
S Phase
pathway information | values
# of genes in pathway | 127
# of genes in pathway and data | 113
coverage | 0.89
density | 0.57
Sphingolipid metabolism
pathway information | values
# of genes in pathway | 73
# of genes in pathway and data | 48
coverage | 0.66
density | 0.32
SRP−dependent cotranslational protein targeting to membrane
pathway information | values
# of genes in pathway | 113
# of genes in pathway and data | 103
coverage | 0.91
density | 0.98
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75


Stabilization of p53
pathway information | values
# of genes in pathway | 55
# of genes in pathway and data | 48
coverage | 0.87
density | 0.89
Stimuli−sensing channels
pathway information | values
# of genes in pathway | 102
# of genes in pathway and data | 42
coverage | 0.41
density | 0.25
SUMO E3 ligases SUMOylate target proteins
pathway information | values
# of genes in pathway | 94
# of genes in pathway and data | 82
coverage | 0.87
density | 0.49
SUMOylation
pathway information | values
# of genes in pathway | 100
# of genes in pathway and data | 88
coverage | 0.88
density | 0.48
SUMOylation of DNA damage response and repair proteins
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 63
coverage | 0.89
density | 0.5
Switching of origins to a post−replicative state
pathway information | values
# of genes in pathway | 71
# of genes in pathway and data | 59
coverage | 0.83
density | 0.82
Synthesis of DNA
pathway information | values
# of genes in pathway | 99
# of genes in pathway and data | 86
coverage | 0.87
density | 0.62
Synthesis of substrates in N−glycan biosythesis
pathway information | values
# of genes in pathway | 62
# of genes in pathway and data | 48
coverage | 0.77
density | 0.29
Log10(Fisher's test p−value)
0
−25
−50
−75
0.00
0.25
0.50
0.75
1.00
Correlation cutoff

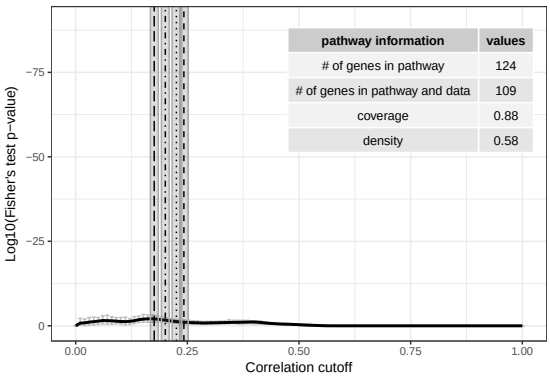

TCF dependent signaling in response to WNT
pathway information | values
# of genes in pathway | 232
# of genes in pathway and data | 141
coverage | 0.61
density | 0.42
TCR signaling
pathway information | values
# of genes in pathway | 118
# of genes in pathway and data | 98
coverage | 0.83
density | 0.47
Telomere Maintenance
pathway information | values
# of genes in pathway | 82
# of genes in pathway and data | 46
coverage | 0.56
density | 0.61
The citric acid (TCA) cycle and respiratory electron transport
pathway information | values
# of genes in pathway | 159
# of genes in pathway and data | 118
coverage | 0.74
density | 0.51
The role of GTSE1 in G2/M progression after G2 checkpoint
pathway information | values
# of genes in pathway | 60
# of genes in pathway and data | 55
coverage | 0.92
density | 0.88
TNFR2 non−canonical NF−kB pathway
pathway information | values
# of genes in pathway | 100
# of genes in pathway and data | 74
coverage | 0.74
density | 0.67
Toll Like Receptor 10 (TLR10) Cascade
pathway information | values
# of genes in pathway | 83
# of genes in pathway and data | 72
coverage | 0.87
density | 0.61
Toll Like Receptor 2 (TLR2) Cascade
pathway information | values
# of genes in pathway | 93
# of genes in pathway and data | 84
coverage | 0.9
density | 0.61
Log10(Fisher's test p−value)
Correlation cutoff

Toll Like Receptor 3 (TLR3) Cascade



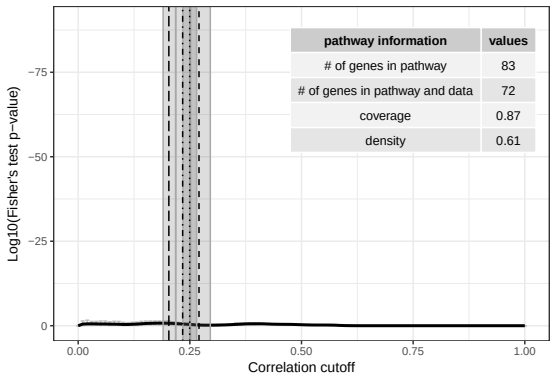
pathway information	values
# of genes in pathway	98
# of genes in pathway and data	86
coverage	0.88
density	0.64

Toll Like Receptor 4 (TLR4) Cascade



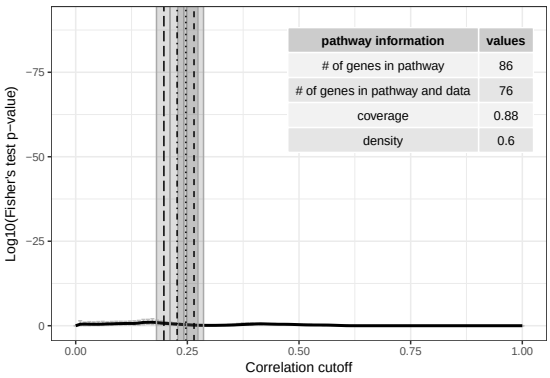
pathway information	values
# of genes in pathway	124
# of genes in pathway and data	109
coverage	0.88
density	0.58

Toll Like Receptor 5 (TLR5) Cascade



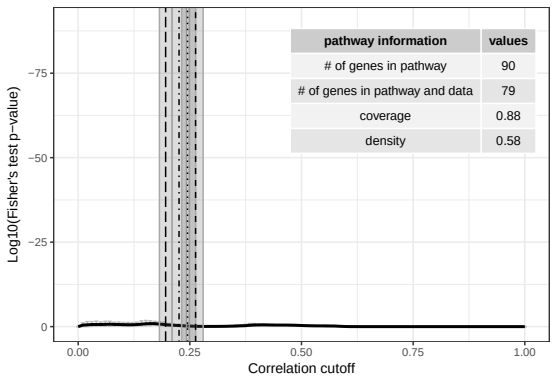
pathway information	values
# of genes in pathway	83
# of genes in pathway and data	72
coverage	0.87
density	0.61

Toll Like Receptor 7/8 (TLR7/8) Cascade



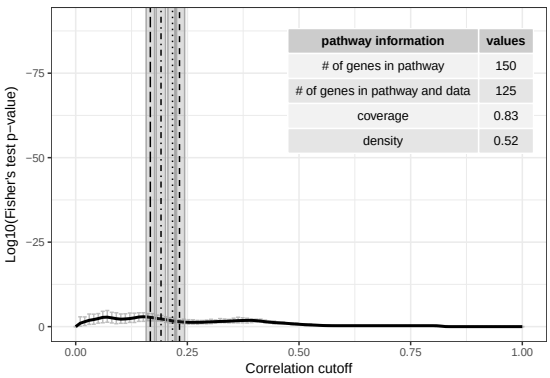
pathway information	values
# of genes in pathway	86
# of genes in pathway and data	76
coverage	0.88
density	0.6

Toll Like Receptor 9 (TLR9) Cascade



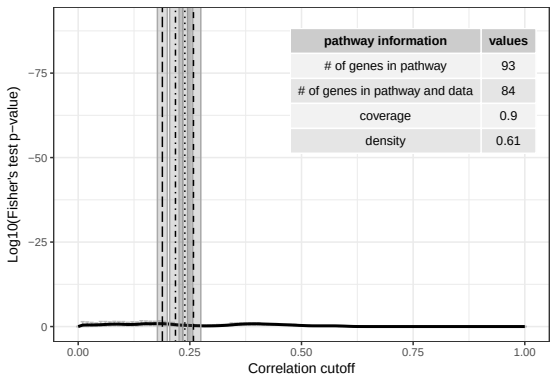
pathway information	values
# of genes in pathway	90
# of genes in pathway and data	79
coverage	0.88
density	0.58

Toll−Like Receptors Cascades



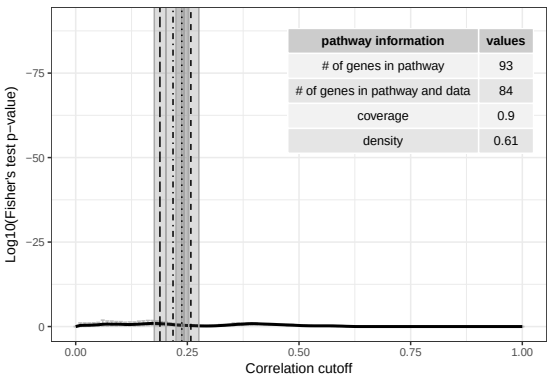
pathway information	values
# of genes in pathway	150
# of genes in pathway and data	125
coverage	0.83
density	0.52

Toll Like Receptor TLR1:TLR2 Cascade

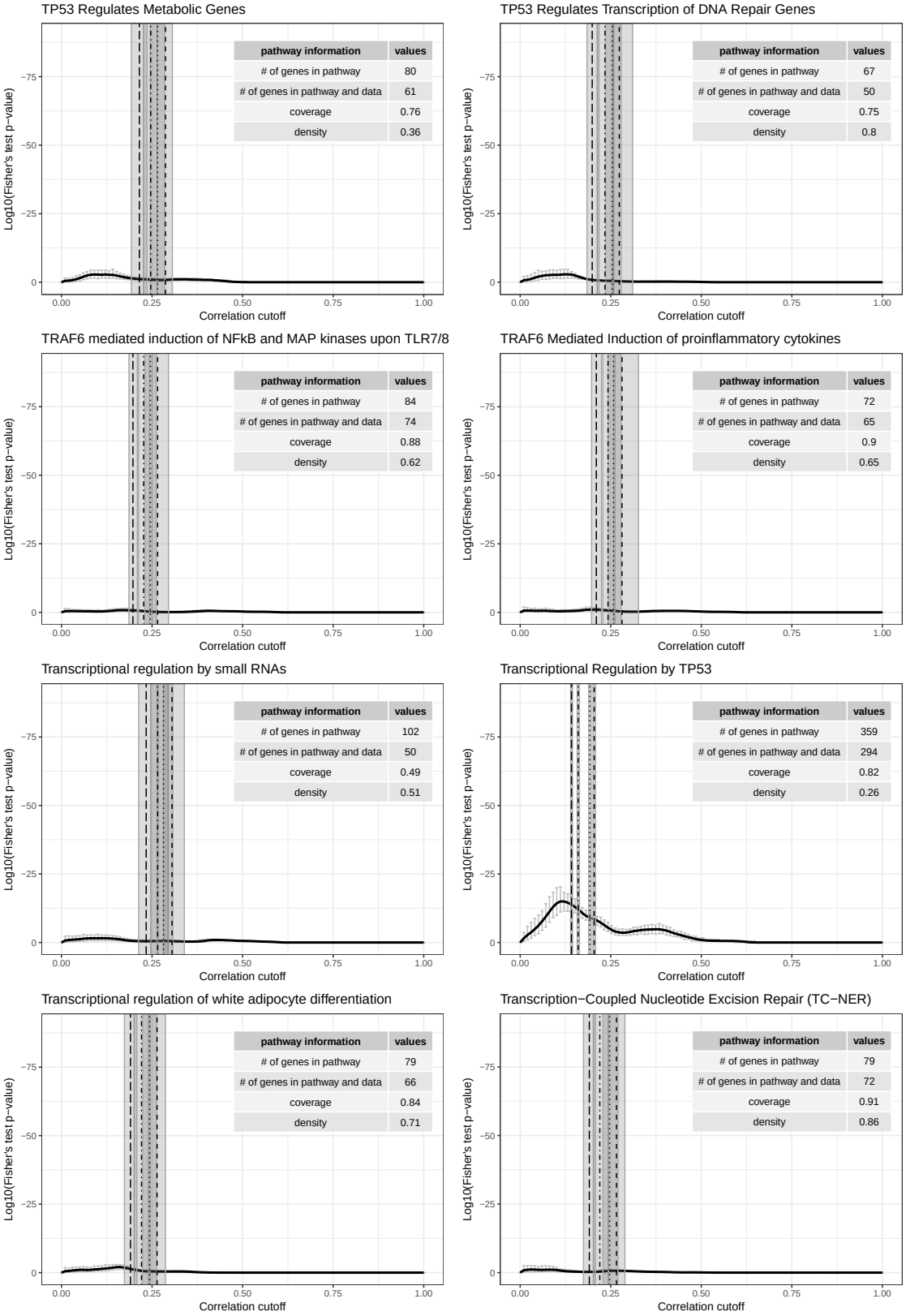


pathway information	values
# of genes in pathway	93
# of genes in pathway and data	84
coverage	0.9
density	0.61

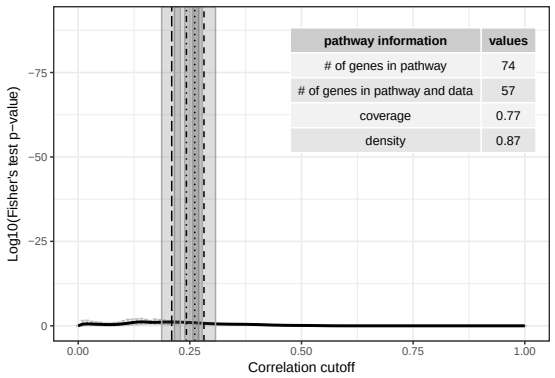
Toll Like Receptor TLR6:TLR2 Cascade



pathway information	values
# of genes in pathway	93
# of genes in pathway and data	84
coverage	0.9
density	0.61

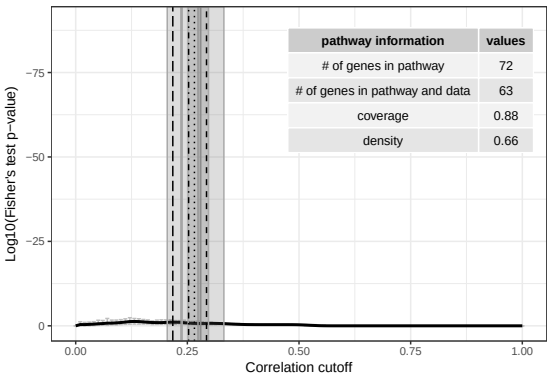

TP53 Regulates Metabolic Genes
pathway information | values
# of genes in pathway | 80
# of genes in pathway and data | 61
coverage | 0.76
density | 0.36
TP53 Regulates Transcription of DNA Repair Genes
pathway information | values
# of genes in pathway | 67
# of genes in pathway and data | 50
coverage | 0.75
density | 0.8
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8
pathway information | values
# of genes in pathway | 84
# of genes in pathway and data | 74
coverage | 0.88
density | 0.62
TRAF6 Mediated Induction of proinflammatory cytokines
pathway information | values
# of genes in pathway | 72
# of genes in pathway and data | 65
coverage | 0.9
density | 0.65
Transcriptional regulation by small RNAs
pathway information | values
# of genes in pathway | 102
# of genes in pathway and data | 50
coverage | 0.49
density | 0.51
Transcriptional Regulation by TP53
pathway information | values
# of genes in pathway | 359
# of genes in pathway and data | 294
coverage | 0.82
density | 0.26
Transcriptional regulation of white adipocyte differentiation
pathway information | values
# of genes in pathway | 79
# of genes in pathway and data | 66
coverage | 0.84
density | 0.71
Transcription−Coupled Nucleotide Excision Repair (TC−NER)
pathway information | values
# of genes in pathway | 79
# of genes in pathway and data | 72
coverage | 0.91
density | 0.86
Log10(Fisher's test p−value)
Correlation cutoff

Transcription of the HIV genome



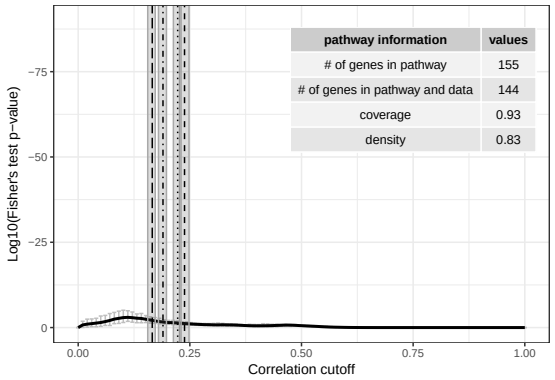
pathway information	values
# of genes in pathway	74
# of genes in pathway and data	57
coverage	0.77
density	0.87

trans−Golgi Network Vesicle Budding



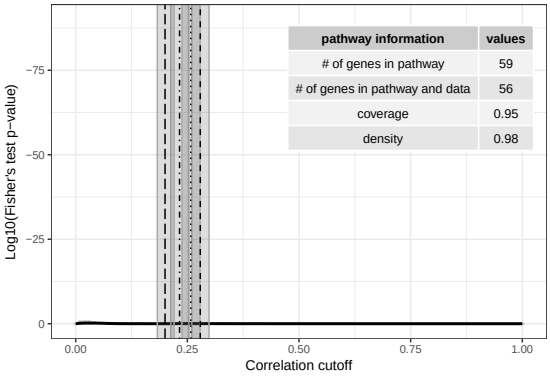
pathway information	values
# of genes in pathway	72
# of genes in pathway and data	63
coverage	0.88
density	0.66

Translation



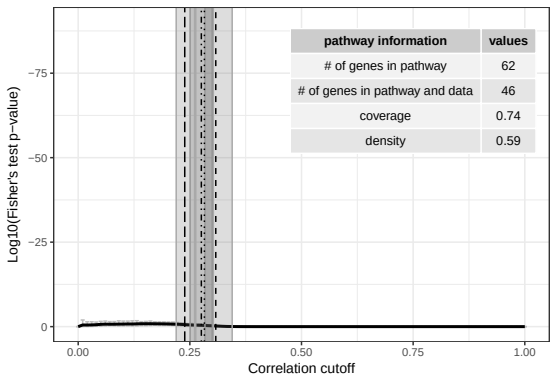
pathway information	values
# of genes in pathway	155
# of genes in pathway and data	144
coverage	0.93
density	0.83

Translation initiation complex formation



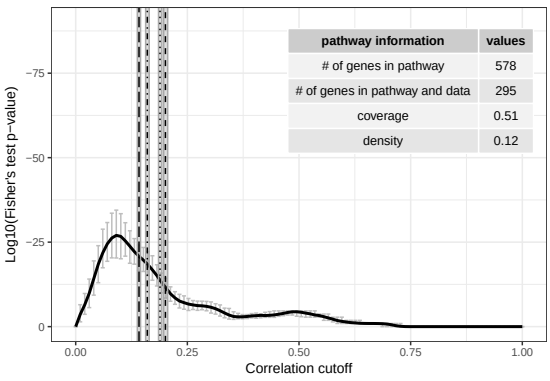
pathway information	values
# of genes in pathway	59
# of genes in pathway and data	56
coverage	0.95
density	0.98

Translocation of GLUT4 to the plasma membrane



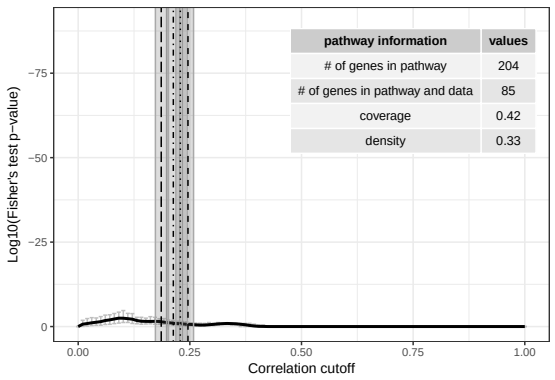
pathway information	values
# of genes in pathway	62
# of genes in pathway and data	46
coverage	0.74
density	0.59

Transmembrane transport of small molecules



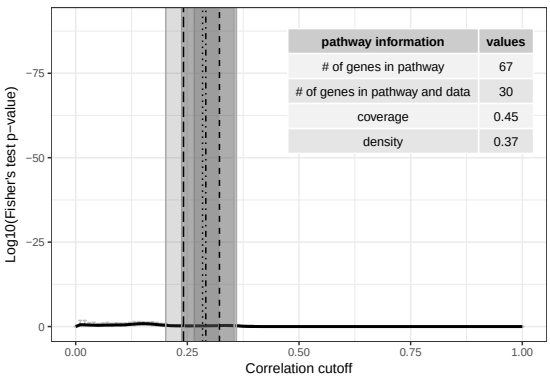
pathway information	values
# of genes in pathway	578
# of genes in pathway and data	295
coverage	0.51
density	0.12

Transmission across Chemical Synapses

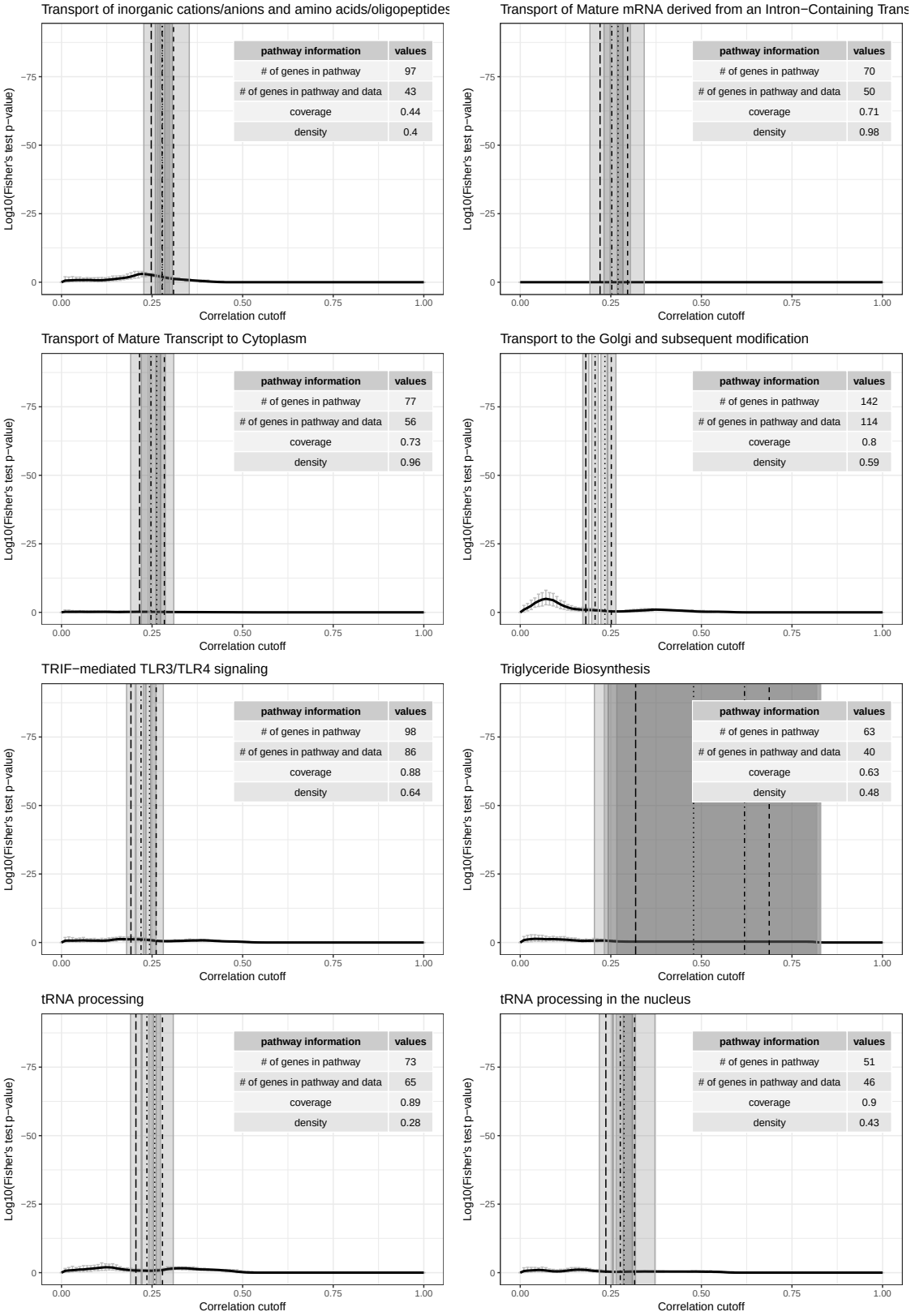


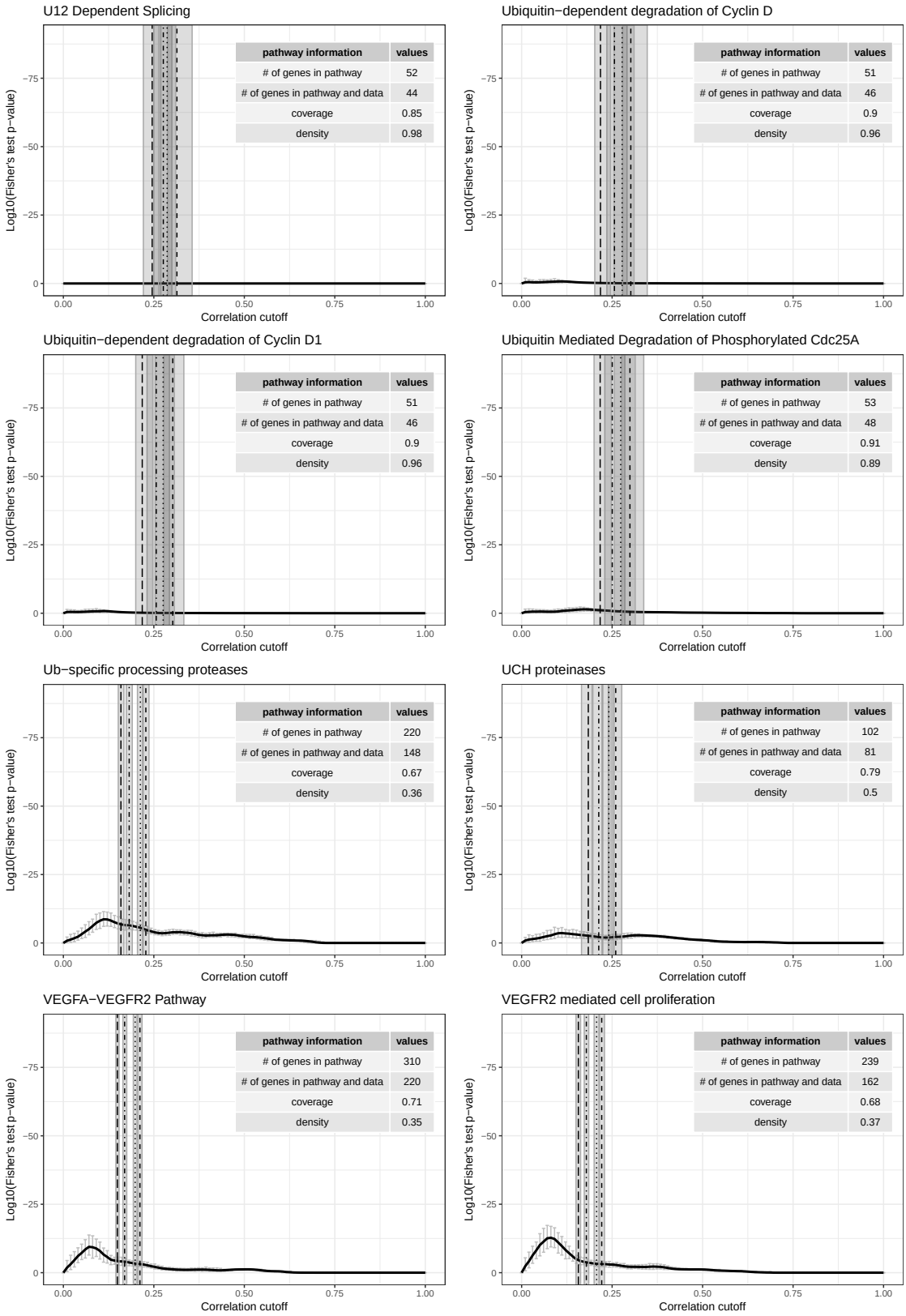
pathway information	values
# of genes in pathway	204
# of genes in pathway and data	85
coverage	0.42
density	0.33

Transport of glucose and other sugars, bile salts and organic acids,



pathway information	values
# of genes in pathway	67
# of genes in pathway and data	30
coverage	0.45
density	0.37


Transport of inorganic cations/anions and amino acids/oligopeptides
pathway information | values
# of genes in pathway | 97
# of genes in pathway and data | 43
coverage | 0.44
density | 0.4
Transport of Mature mRNA derived from an Intron−Containing Trans
pathway information | values
# of genes in pathway | 70
# of genes in pathway and data | 50
coverage | 0.71
density | 0.98
Transport of Mature Transcript to Cytoplasm
pathway information | values
# of genes in pathway | 77
# of genes in pathway and data | 56
coverage | 0.73
density | 0.96
Transport to the Golgi and subsequent modification
pathway information | values
# of genes in pathway | 142
# of genes in pathway and data | 114
coverage | 0.8
density | 0.59
TRIF−mediated TLR3/TLR4 signaling
pathway information | values
# of genes in pathway | 98
# of genes in pathway and data | 86
coverage | 0.88
density | 0.64
Triglyceride Biosynthesis
pathway information | values
# of genes in pathway | 63
# of genes in pathway and data | 40
coverage | 0.63
density | 0.48
tRNA processing
pathway information | values
# of genes in pathway | 73
# of genes in pathway and data | 65
coverage | 0.89
density | 0.28
tRNA processing in the nucleus
pathway information | values
# of genes in pathway | 51
# of genes in pathway and data | 46
coverage | 0.9
density | 0.43
Log10(Fisher's test p−value)
Correlation cutoff
0.00
0.25
0.50
0.75
1.00
0
−25
−50
−75


U12 Dependent Splicing
pathway information | values
# of genes in pathway | 52
# of genes in pathway and data | 44
coverage | 0.85
density | 0.98
Ubiquitin−dependent degradation of Cyclin D
pathway information | values
# of genes in pathway | 51
# of genes in pathway and data | 46
coverage | 0.9
density | 0.96
Ubiquitin−dependent degradation of Cyclin D1
pathway information | values
# of genes in pathway | 51
# of genes in pathway and data | 46
coverage | 0.9
density | 0.96
Ubiquitin Mediated Degradation of Phosphorylated Cdc25A
pathway information | values
# of genes in pathway | 53
# of genes in pathway and data | 48
coverage | 0.91
density | 0.89
Ub−specific processing proteases
pathway information | values
# of genes in pathway | 220
# of genes in pathway and data | 148
coverage | 0.67
density | 0.36
UCH proteinases
pathway information | values
# of genes in pathway | 102
# of genes in pathway and data | 81
coverage | 0.79
density | 0.5
VEGFA−VEGFR2 Pathway
pathway information | values
# of genes in pathway | 310
# of genes in pathway and data | 220
coverage | 0.71
density | 0.35
VEGFR2 mediated cell proliferation
pathway information | values
# of genes in pathway | 239
# of genes in pathway and data | 162
coverage | 0.68
density | 0.37
Log10(Fisher's test p−value)
Correlation cutoff

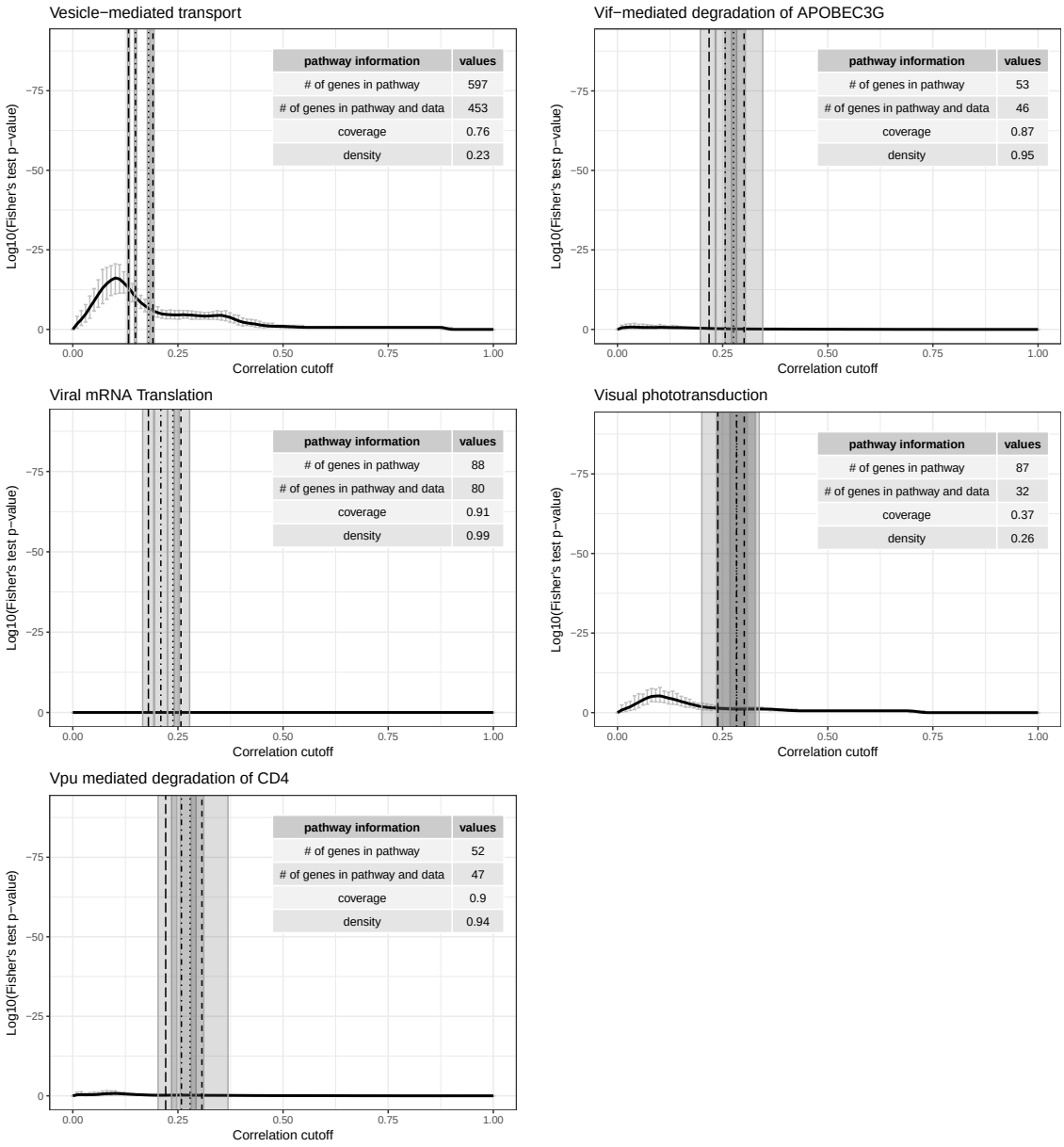


Figure 6.14: In the following pages we report the cutoff optimization results for the transcriptomics data. For each of the 469 pathway considered, protein-protein interaction networks from STRING were used as reference. The black curve represents the average over 100 bootstrapping resamplings, and the error bars show the corresponding 95% confidence intervals. Vertical lines indicate the mean of the statistical cutoffs, and the areas the corresponding 95% confidence intervals over the bootstrapping. Pathways are shown in alphabetical order.

Chapter 7

Discussion and Outlook

The investigation of the molecular mechanisms regulating protein glycosylation is revealing essential to better understand and describe physiological processes and diseases at a molecular level. The absence of a direct genetic template, together with the site-, protein- and cell-specificity of glycosylation, have made the characterization of the biochemical pathways of glycan synthesis extremely difficult. The advancement of measurement technologies, however, allow today to quantify glycan structures from different proteins and fluids in large-scale datasets, enabling systematic statistical analyses.

In this thesis, we contributed to a better understanding of protein glycan structures and biosynthesis, as well as to a more biologically meaningful pipeline of analysis for glycomics data using Gaussian graphical models on large-scale glycomics data and the available prior knowledge on the glycan synthesis pathways.

Scientific achievements

The novel scientific findings included in this work are listed below.

- **GGMs on glycomics data identify known glycosylation reactions.** Gaussian graphical models are able to identify single enzymatic reactions in the synthesis pathway of protein glycans, in the case of both protein-specific (IgG Fc, Chapter 3) and protein mixture (TPNG, Chapter 4) glycomics data. This means that partial correlations among measured glycans reflect the underlying biochemical synthesis pathway. The ability of GGMs to reflect synthesis pathways had already been shown for metabolomics data [32], and there already it was pointed out that it is surprising,

as these molecules are not measured directly where their production occurs, e.g., in B-cells or in liver, but in circulating blood, where other processes, like transport and degradation, take place.

- **GGMs identify new IgG glycosylation reactions.** As a consequence of the previous finding, edges in the GGM that are not present in the known biochemical pathway can correspond to true but previously unknown enzymatic reactions taking part in glycan synthesis (Chapter 3). We formulated hypotheses of new synthetic steps in the IgG glycosylation pathway based on the data-driven partial correlation network and validated them experimentally. We therefore proved for the first time that statistical analysis and computational modelling on large-scale glycomics data can provide concrete biological insights into the intracellular processes of glycosylation and that systems biology can drive biological experiments.

- **GGMs and prior knowledge can be used for structural inference.** MALDI-TOF-MS measurements of the total plasma N-glycome are valuable to investigate changes in the overall glycosylation profile of plasma proteins, but do only resolve molecular masses and not single glycan structures. We used data-driven partial correlation networks coupled with prior knowledge to infer structural details of the glycans within each mass spectrometry peak (Chapter 4). Our predictions on the glycan structures were validated using previously published independent data and showed high accuracy. This approach could be used in the future to limit the need for expensive fragmentation experiments to resolve the structural profile of mass spectrometry-based glycomics measurements.

- **GGMs and prior knowledge can be used for glycomics preprocessing evaluation.** The strong relationship between calculated GGMs and prior knowledge on the glycan synthesis pathways can be used to evaluate different preprocessing strategies for glycomics data (Chapter 5). We estimated the quality of different normalization approaches according to the overlap between the inferred GGM and the available prior knowledge, where higher overlap corresponded to higher quality. We defined *best* the normalization strategy that allowed to reproduce the most known biochemical relations among glycan structures, therefore introducing a biological criterion for quality assessment. Results were replicated across different cohorts and measurement platform, demonstrating the generalizability of our findings. According to our analysis, the 'Probabilistic Quotient' normalization followed by log-transformation is the most reliable approach for glycan preprocessing.

- **Prior knowledge can be used for correlation networks cutoff optimization.** Statistically determined correlation cutoffs for network inference are not necessarily

the best option to obtain a network that accurately represents the underlying biological system. Exploiting once again the relationship between inferred GGM and prior knowledge, we developed an approach for the *optimization* of the correlation cutoff for network inference (Chapter 6). Briefly, we varied the correlation cutoff and computed the corresponding network-prior knowledge overlap. The *optimal cutoff* was defined as the one that produced the network with maximal overlap to the available prior knowledge. Note that there is a fundamental difference between our procedure and most other prior knowledge-based network inference approaches [209–214]: while the latter combine prior information and data-driven GGM to statistically regularize the inference problem, we use the biological reference exclusively for comparison to the data-driven GGM. This means that, in our setting, incomplete or incorrect prior knowledge from different sources can be employed as well for the optimization. Importantly, in such cases, the optimized network can be further exploited for inference of new molecular associations. To prove generalizability, the approach was applied to two completely different omics data, namely a metabolomics and a transcriptomics dataset. In these cases, no complete prior knowledge is available, and, therefore, only partial biological references can be used in the optimization. We show that even in this scenario our approach is successful in determining an optimal cutoff and that the corresponding optimal network is superior to the statistically inferred networks in identifying meaningful molecular interactions.

Taken together, human glycomics data contain a strong and stable footprint of the underlying biochemical pathways of synthesis, which can be reconstructed by partial correlation networks in large-scale datasets. The inferred networks can be compared to the available prior knowledge to either infer new biology, or to optimize the pipeline of glycomics data analysis.

Extensions and future directions

Data analysis approaches

From a methodological point of view, the results presented in this thesis could be extended in several directions, which are discussed below.

1. Prediction

The first aspect addressed by our analysis that could be improved in the near future concerns the accuracy of our predictions.

- **Accounting for glycosidases in pathway inference.** So far, in our pathway models, we assumed synthesis reactions to go in one direction only, mostly representing the addition of monosaccharides to glycan structures (see Figure 2.3). Cross-sectional omics data only allow for non-directional inference; however, any potential new synthesis reactions identified by the data-driven GGM could in principle be catalyzed by either a glycosyltransferases or a glycosidases. For IgG, we have proved experimentally that the new inferred reactions are performed by glycosyltransferases, but this cannot be assumed in general. For TPNG, for example, the action of glycosidases could be relevant, as most glycans are exposed on the surface of the proteins [2], not buried within the hydrophobic core like in IgG, and hence easily accessible. Recent studies on glycosidases are showing that these enzymes are also active extracellularly and could modify the glycan structures on a protein even once it has been secreted and it is circulating in blood [216], for example to modulate the protein's activity and degradation [182]. In a generalized model for pathway inference, therefore, each new biochemical link inferred from a correlation network should be experimentally tested as a possible synthesis (i.e., performed by a glycosyltranferase) or degradation (i.e., performed by a glycosidase) reaction.

- **Multivariate TPNG structural inference.** Our approach to infer the structural details of TPNG is univariate, as it resolves compositions one at a time, without accounting for the previous results when inferring the structure of a new node. A natural extension would therefore investigate a global solution, where each inference step has to account for the results on all other compositions. This would avoid possible inconsistencies in the inference results and provide more accurate predictions.

2. Glycosylation regulation

One of the most interesting open issues in glycobiology is the understanding of the molecular mechanisms mediating the regulation of protein glycosylation, namely what makes the cell *decide* to attach the observed glycans to a given site on a given protein. A variety of different aspects play a role in the final glycosylation profile of different proteins, from stochastic elements, like the availability of activated sugar donors [217–219] and the partial competition of different enzymes [93], to (epi-)genetic factors [220, 221], which are extremely difficult to reproduce and manipulate in a controlled experimental setting.

In future projects, we could address the investigation of the regulatory aspect of protein glycosylation with several statistical approaches.

- **Probabilistic pathway inference.** In our pathway analysis and inference approach, we treated GGMs as binary adjacency matrices, namely only considering

whether a given edge was present in the network or not. A natural but non-trivial extension of this model would actively consider the strength of the partial correlation between glycan pairs to infer the *likelihood* of the corresponding enzymatic reaction, where higher correlation coefficients would correspond to more probable reactions. A quantitative pathway inference could provide valuable insights into the regulation of different intracellular enzymatic steps. For example, if a given structure can be synthesized from two different glycan substrates, the partial correlation coefficients corresponding to the two synthesis steps could indicate which one is the more enzymatically favorable *in vivo*, as opposed to the *in vitro* enzymatic assay experiments. Alternatively, this approach could be used to optimize the experimental validation of new inferred enzymatic reaction, prioritizing those with a higher partial correlation coefficient.

- **Differential GGMs.** The strong relationship between glycan synthesis pathways and data-driven GGMs could also be exploited to infer the potential disruption of particular glycosylation steps in specific physiological conditions. We have shown that edges in the GGMs represent glycan synthesis reactions, and that glycosylation pathways are highly conserved across populations (Figure 3.10). Therefore, a significant difference between the partial correlation coefficients in the GGMs inferred from two physiologically different groups of individuals could reflect a different regulation of the corresponding enzymatic step in the two conditions. Since is known that strong aberrations in the glycosylation pathways are often fatal already in the early stage of development [31, 222], to maximize the chances of observing statistically significant differences, one should compare the GGM inferred from controls to that representing a disease that strongly affect cellular activities like cancer.

- **Protein-specific glycosylation pathways.** When more protein- and site-specific glycomics data will become available in the future, the aforementioned approaches could also contribute to the determination of differences in the protein- and site-specific glycosylation pathways. Given the results of the TPNG analysis in Chapter 4, we expect a substantial part of the pathway to be shared among different proteins. However, protein-specific steps are more than likely to occur, and these would be hard to recognize from total plasma glycomics datasets. However, when comparing the GGMs inferred from different protein-specific glycomics data, significant differences in the strength of the partial correlation coefficients could reflect different enzymatic activities of glycosyltransferases on the considered proteins. The cause of this effect could be due, for example, to different stereochemical properties, which might result in impaired or enhanced accessibility of the enzymes to the glycan

substrates, and provide therefore a valuable starting point for further experimental investigations.

- **Glyco-proteomics analysis.** Protein-specific glycomics data are currently only available for a handful of proteins, mostly due to the lack of high affinity antibodies for isolation [223], or to the too low protein abundance, which makes quantification with the current technology infeasible. This is likely to change in the very near future [224, 225], but protein-glycan relationships could, in the meantime, be investigated by analyzing protein-glycan associations from large-scale data. The idea is here to compute linear associations between, for example, mixture protein glycomics data, like the TPNG, and the corresponding plasma proteomics data. A first analysis between total plasma N-glycome and plasma proteomics data has been performed in the QMDiab cohort [226]. However, as the TPNG data were there measured via UPLC, observed associations were not easily traceable to protein-glycan structure pairs (see Section 2.1.1). Using MALDI-TOF-MS TPNG data together with the results of our structural inference, more specific protein-glycan associations could be identified, which could help better understanding modulator activity of glycans for protein functions.

- **Gene-glycan molecular tracing.** Several Genome Wide Association Studies (GWAS) have identified genes associated with the abundances of various glycan structures, both in IgG [79] and in plasma proteins [227]. Interestingly, most of those genes are not glycosylation enzymes (i.e., glycosyltransferases or glycosidases) and hence might be indirectly involved in the regulation of the glycosylation process, rather than in the actual synthesis. For example, the BACH2 gene on chromosome 6 has been observed to significantly associate to monogalactosylated glycans in IgG [79]. This gene is not a glycosylation enzyme, but is responsible, among other things, for the transcriptional activation of B-cells [228], and it has been found to associate with a variety of diseases related to autoimmunity, i.e., type 1 diabetes [229–232], celiac disease [233], Crohn's disease [234] and multiple sclerosis [235]. In order to better understand how the genes interfere with protein glycosylation, we could use available databases and prior knowledge to systematically *trace* the molecular interactions connecting the gene-glycan pairs found to be significantly associated in the cohort of interest. A similar tracing approach has been shown to be able to explain the molecular steps leading to observed gene-metabolites interactions [236]. Our group is currently developing efficient algorithms for the automation of this molecular tracing approach for metabolomics data, integrating biochemical pathways from different databases (e.g., KEGG [237], RECON 3D [238], Reactome [150, 151])

with data-driven statistical associations. An extension of this model would also include glycosylation pathways.

Future experiments

In addition to what discussed above, there are other important issues that could be addressed once measurement technologies will allow for the quantification of more specific glycomics data. In particular, the work presented here raised two major points.

- **IgG antigen-specific Fc glycosylation.** Given the wide variety of pathogens that humans, on average, encounter throughout life, the immune system evolved to elicit different responses according to the danger level of the invading pathogen. In the IgG molecule, the Fc region, where a highly conserved glycosylation site is present, is responsible to initiate the immune response through interaction with Fc-receptors. A major question is therefore whether nature *optimized* IgG Fc glycosylation to produce an immune response specific to the antigen-carrying pathogen. In order to test this, one would need to quantify the Fc glycan structure from isolated IgGs with a given antigen-specificity, and compare the obtained glycan profile to that of bulk IgG or IgG with a different specificity. Should this comparison identify significant differences, an interesting analysis would be the clustering of individuals according to their antigen-specific IgG glycosylation profiles. Each cluster could then be associated to the efficacy of the immune response in neutralizing the antigen-carrying pathogen. The outcome of this analysis could lead to the characterization of an *optimal* glycosylation profile for any given pathogen.
- **Antigen-specific Ig glycome optimization.** The natural extension of the previous point, which is unfortunately still infeasible, would be the artificial engineering of the glycosylation profile on therapeutic antibodies used, for example, for Intra-Venous Immunoglobulin (IVIG) therapy [239, 240]. This would allow to provide patients with only the most effective antibodies against the specific pathogen they are infected with, optimizing immune response and recovery.

Outlook

In this thesis, we have highlighted how glycans are involved in a wide variety of biological processes, despite the specific molecular mechanisms of their regulation being mostly unclear. This is in part due to the intrinsic complexity of the glycosylation process, which, in contrast to most other macromolecules, does not have a direct genetic template and therefore cannot be easily manipulated under controlled experimental conditions [241]. On the

other hand, measurement technologies are still not able to handle the full complexity of an organism's glycome, namely its site-, protein- and cell-specificity, and currently allow the quantification of glycans from either a handful of isolated proteins or protein mixtures.

The situation is however likely to change soon, as measurement technologies are becoming more versatile and precise and as more people are investigating the role of glycans in diseases. In particular, due to the dynamic and adaptive nature of glycosylation, glycans are finding increasing success as potential early stage biomarkers [242] and drug targets, especially for diseases where other omics have failed, like cancer [98, 100, 101, 243, 244] and HIV [56, 245]. Moreover, glyco-engineering, i.e., the optimization of the glycosylation on synthetic glycoproteins, is proving valuable to improve the efficacy of widely used drugs: since a major proportion of biotherapeutics products, from antibodies to cytokines, are glycoproteins [246], modulating the protein function through its glycosylation will allow to meet more specific and personalized functional requirements [106, 247].

Conclusion

In this thesis, we investigated large-scale glycomics datasets by means of Gaussian graphical models. The core idea was based on the observation that edges in the computed GGM correspond to single enzymatic steps in the glycan synthesis pathways. The quantitative overlap between data-driven correlation networks and the available prior knowledge was used to address specific but substantially different questions, ranging from the prediction of new enzymatic reactions in the pathway of glycan synthesis, to the optimization of preprocessing and network inference strategies for glycomics downstream analysis. In all cases, our findings were either validated experimentally or replicated in different cohorts and using different data types to prove generalizability. In conclusion, we have shown that the pairing of GGMs with prior knowledge is a powerful tool to investigate the synthesis and regulation of protein glycans in humans.

Bibliography

[1] R. Apweiler, H. Hermjakob, and N. Sharon, "On the frequency of protein glycosylation, as deduced from analysis of the swiss-prot database1," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1473, no. 1, pp. 4–8, 1999.

[2] A. Varki, R. Cummings, J. Esko, H. Freeze, G. Hart, and J. Marth, *Essential of glycobiology. 653.* Cold Spring Harbor Laboratory Press, 1999.

[3] S. P. Grefrath and J. A. Reynolds, "The molecular weight of the major glycoprotein from the human erythrocyte membrane," *Proceedings of the National Academy of Sciences*, vol. 71, no. 10, pp. 3913–3916, 1974.

[4] D. Walt, K. Aoki-Kinoshita, B. Bendiak, C. Bertozzi, G. Boons, A. Darvill, G. Hart, L. Kiessling, J. Lowe, R. Moon, J. Paulson, R. Sasisekharan, A. Varki, and C. Wong, "Transforming Glycoscience: A Roadmap for the Future," *Nantional Academy of Sciences*, pp. 1–209, 2012.

[5] *Encyclopedia Britannica.* Encyclopædia Britannica, Inc., 2007.

[6] D. Voet and J. G. Voet, *Biochemistry.* John Wiley & Sons, 2004.

[7] R. D. Cummings, "The repertoire of glycan determinants in the human glycome," *Molecular BioSystems*, vol. 5, no. 10, pp. 1087–1104, 2009.

[8] A. Dell, A. Galadari, F. Sastre, and P. Hitchen, "Similarities and differences in the glycosylation mechanisms in prokaryotes and eukaryotes," *International journal of microbiology*, vol. 2010, 2011.

[9] A. Varki, R. D. Cummings, M. Aebi, N. H. Packer, P. H. Seeberger, J. D. Esko, P. Stanley, G. Hart, A. Darvill, T. Kinoshita, *et al.*, "Symbol nomenclature for graphical representations of glycans," *Glycobiology*, vol. 25, no. 12, pp. 1323–1324, 2015.

[10] A. Knezevic, O. Polasek, O. Gornik, I. Rudan, H. Campbell, C. Hayward, A. Wright, I. Kolcic, N. O'Donoghue, J. Bones, *et al.*, "Variability, heritability and environmental determinants of human plasma n-glycome," *Journal of proteome research*, vol. 8, no. 2, pp. 694–701, 2008.

[11] E. Cabib and L. F. Leloir, "Guanosine diphosphate mannose," *Journal of Biological Chemistry*, vol. 206, no. 2, pp. 779–790, 1954.

[12] N. Taniguchi, K. Honke, and M. Fukuda, *Handbook of glycosyltransferases and related genes*. Springer Science & Business Media, 2011.

[13] J. Baenziger, "Protein-specific glycosyltransferases: how and why they do it!," *The FASEB Journal*, vol. 8, no. 13, pp. 1019–1025, 1994.

[14] P. Stanley, "Golgi glycosylation," *Cold Spring Harbor perspectives in biology*, p. a005199, 2011.

[15] G. Davies and B. Henrissat, "Structures and mechanisms of glycosyl hydrolases," *Structure*, vol. 3, no. 9, pp. 853–859, 1995.

[16] B. Henrissat and G. Davies, "Structural and sequence-based classification of glycoside hydrolases," *Current opinion in structural biology*, vol. 7, no. 5, pp. 637–644, 1997.

[17] R. Marshall, "Glycoproteins," *Annual review of biochemistry*, vol. 41, no. 1, pp. 673–702, 1972.

[18] M. Aebi, R. Bernasconi, S. Clerc, and M. Molinari, "N-glycan structures: recognition and processing in the er," *Trends in biochemical sciences*, vol. 35, no. 2, pp. 74–82, 2010.

[19] P. V. d. Steen, P. M. Rudd, R. A. Dwek, and G. Opdenakker, "Concepts and principles of o-linked glycosylation," *Critical reviews in biochemistry and molecular biology*, vol. 33, no. 3, pp. 151–208, 1998.

[20] A. Helenius and M. Aebi, "Roles of n-linked glycans in the endoplasmic reticulum," *Annual review of biochemistry*, vol. 73, no. 1, pp. 1019–1049, 2004.

[21] H. Schachter and H. H. Freeze, "Glycosylation diseases: quo vadis?," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1792, no. 9, pp. 925–930, 2009.

[22] G. Lauc, A. Vojta, and V. Zoldoš, "Epigenetic regulation of glycosylation is the quantum mechanics of biology," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1840, no. 1, pp. 65–70, 2014.

[23] N. Taniguchi, T. Endo, G. W. Hart, P. H. Seeberger, and C. H. Wong, *Glycoscience: biology and medicine*. Springer Japan, 2015.

[24] J. Roth and E. G. Berger, "Immunocytochemical localization of galactosyltransferase in hela cells: codistribution with thiamine pyrophosphatase in trans-golgi cisternae.," *The Journal of Cell Biology*, vol. 93, no. 1, pp. 223–229, 1982.

[25] A. Sturm, K. D. Johnson, T. Szumilo, A. D. Elbein, and M. J. Chrispeels, "Subcellular localization of glycosidases and glycosyltransferases involved in the processing of n-linked oligosaccharides," *Plant Physiology*, vol. 85, no. 3, pp. 741–745, 1987.

[26] J. C. Paulson and K. J. Colley, "Glycosyltransferases. structure, localization, and control of cell type-specific glycosylation.," *Journal of Biological Chemistry*, vol. 264, no. 30, pp. 17615–17618, 1989.

[27] K. J. Colley, "Golgi localization of glycosyltransferases: more questions than answers," *Glycobiology*, vol. 7, no. 1, pp. 1–13, 1997.

[28] L. Tu and D. K. Banfield, "Localization of golgi-resident glycosyltransferases," *Cellular and molecular life sciences*, vol. 67, no. 1, pp. 29–41, 2010.

[29] T. A. Shinkel, C.-G. Chen, E. Salvaris, T. R. Henion, H. Barlow, U. Galili, M. J. Pearse, and A. J. d'Apice, "Changes in cell surface glycosylation in α1, 3-galactosyltransferase knockout and α1, 2-fucosyltransferase transgenic mice," *Transplantation*, vol. 64, no. 2, pp. 197–204, 1997.

[30] M. Asano, K. Furukawa, M. Kido, S. Matsumoto, Y. Umesaki, N. Kochibe, and Y. Iwakura, "Growth retardation and early death of β-1, 4-galactosyltransferase knockout mice with augmented proliferation and abnormal differentiation of epithelial cells," *The EMBO Journal*, vol. 16, no. 8, pp. 1850–1857, 1997.

[31] K. W. Marek, I. K. Vijay, and J. D. Marth, "A recessive deletion in the glcnac-1-phosphotransferase gene results in peri-implantation embryonic lethality," *Glycobiology*, vol. 9, no. 11, pp. 1263–1271, 1999.

[32] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis, "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data," *BMC Systems Biology*, vol. 5, no. 1, p. 21, 2011.

[33] H. H. Freeze and M. Aebi, "Altered glycan structures: the molecular basis of congenital disorders of glycosylation," *Current opinion in structural biology*, vol. 15, no. 5, pp. 490–498, 2005.

[34] A. O. Akinkuolie, J. E. Buring, P. M. Ridker, and S. Mora, "A novel protein glycan biomarker and future cardiovascular disease events," *Journal of the American Heart Association*, vol. 3, no. 5, p. e001221, 2014.

[35] R. W. McGarrah, J. P. Kelly, D. M. Craig, C. Haynes, R. C. Jessee, K. M. Huffman, W. E. Kraus, and S. H. Shah, "A novel protein glycan–derived inflammation biomarker independently predicts cardiovascular disease and modifies the association of hdl subclasses with mortality," *Clinical chemistry*, pp. clinchem–2016, 2016.

[36] J.-Z. Wang, I. Grundke-Iqbal, and K. Iqbal, "Glycosylation of microtubule–associated protein tau: An abnormal posttranslational modification in alzheimer's disease," *Nature medicine*, vol. 2, no. 8, p. 871, 1996.

[37] S. L. Lundström, H. Yang, Y. Lyutvinskiy, D. Rutishauser, S.-K. Herukka, H. Soininen, and R. A. Zubarev, "Blood plasma igg fc glycans are significantly altered in alzheimer's disease and progressive mild cognitive impairment," *Journal of Alzheimer's Disease*, vol. 38, no. 3, pp. 567–579, 2014.

[38] S. Schedin-Weiss, B. Winblad, and L. O. Tjernberg, "The role of protein glycosylation in alzheimer disease," *The FEBS journal*, vol. 281, no. 1, pp. 46–62, 2014.

[39] R. B. Parekh, R. a. Dwek, B. J. Sutton, D. L. Fernandes, a. Leung, D. Stanworth, T. W. Rademacher, T. Mizuochi, T. Taniguchi, and K. Matsuta, "Association of rheumatoid arthritis and primary osteoarthritis with changes in the glycosylation pattern of total serum IgG.," *Nature*, vol. 316, no. 6027, pp. 452–457, 1985.

[40] B. J. Campbell, L.-G. Yu, and J. M. Rhodes, "Altered glycosylation in inflammatory bowel disease: a possible role in cancer development," *Glycoconjugate journal*, vol. 18, no. 11-12, pp. 851–858, 2001.

[41] F. Vučković, J. Krištić, I. Gudelj, M. Teruel, T. Keser, M. Pezer, M. Pučić-Baković, J. Štambuk, I. Trbojević-Akmačić, C. Barrios, *et al.*, "Association of systemic lupus erythematosus with decreased immunosuppressive potential of the igg glycome," *Arthritis & rheumatology*, vol. 67, no. 11, pp. 2978–2989, 2015.

[42] D. A. McClain, W. A. Lubas, R. C. Cooksey, M. Hazel, G. J. Parker, D. C. Love, and J. A. Hanover, "Altered glycan-dependent signaling induces insulin resistance and

hyperleptinemia," *Proceedings of the National Academy of Sciences*, vol. 99, no. 16, pp. 10695–10699, 2002.

[43] N. Itoh, S. Sakaue, H. Nakagawa, M. Kurogochi, H. Ohira, K. Deguchi, S.-I. Nishimura, and M. Nishimura, "Analysis of n-glycan in serum glycoproteins from db/db mice and humans with type 2 diabetes," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 293, no. 4, pp. E1069–E1077, 2007.

[44] G. Thanabalasingham, J. E. Huffman, J. J. Kattla, M. Novokmet, I. Rudan, A. L. Gloyn, C. Hayward, B. Adamczyk, R. M. Reynolds, A. Muzinic, *et al.*, "Mutations in hnf1a result in marked alterations of plasma glycan profile," *Diabetes*, p. DB_120880, 2012.

[45] A. O. Akinkuolie, A. D. Pradhan, J. E. Buring, P. M. Ridker, and S. Mora, "Novel protein glycan side-chain biomarker and risk of incident type 2 diabetes mellitus," *Arteriosclerosis, thrombosis, and vascular biology*, pp. ATVBAHA–115, 2015.

[46] M. Lantéri, V. Giordanengo, N. Hiraoka, J.-G. Fuzibet, P. Auberger, M. Fukuda, L. G. Baum, and J.-C. Lefebvre, "Altered t cell surface glycosylation in hiv-1 infection results in increased susceptibility to galectin-1-induced cell death," *Glycobiology*, vol. 13, no. 12, pp. 909–918, 2003.

[47] J. S. Moore, X. Wu, R. Kulhavy, M. Tomana, J. Novak, Z. Moldoveanu, R. Brown, P. A. Goepfert, and J. Mestecky, "Increased levels of galactose-deficient igg in sera of hiv-1-infected individuals," *Aids*, vol. 19, no. 4, pp. 381–389, 2005.

[48] M. E. Ackerman, M. Crispin, X. Yu, K. Baruah, A. W. Boesch, D. J. Harvey, A.-S. Dugast, E. L. Heizen, A. Ercan, I. Choi, *et al.*, "Natural variation in fc glycosylation of hiv-specific antibodies impacts antiviral activity," *The Journal of clinical investigation*, vol. 123, no. 5, pp. 2183–2192, 2013.

[49] S. Hakomori, "Glycosylation defining cancer malignancy: new wine in an old bottle," *Proceedings of the National Academy of Sciences*, vol. 99, no. 16, pp. 10231–10233, 2002.

[50] R. Saldova, L. Royle, C. M. Radcliffe, U. M. Abd Hamid, R. Evans, J. N. Arnold, R. E. Banks, R. Hutson, D. J. Harvey, R. Antrobus, *et al.*, "Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and igg," *Glycobiology*, vol. 17, no. 12, pp. 1344–1356, 2007.

[51] J. N. Arnold, R. Saldova, U. M. A. Hamid, and P. M. Rudd, "Evaluation of the serum n-linked glycome for the diagnosis of cancer and chronic inflammation," *Proteomics*, vol. 8, no. 16, pp. 3284–3293, 2008.

[52] R. Saldova, M. R. Wormald, R. A. Dwek, and P. M. Rudd, "Glycosylation changes on serum glycoproteins in ovarian cancer may contribute to disease pathogenesis," *Disease markers*, vol. 25, no. 4-5, pp. 219–232, 2008.

[53] G. Chen, Y. Wang, L. Qiu, X. Qin, H. Liu, X. Wang, Y. Wang, G. Song, F. Li, Y. Guo, *et al.*, "Human igg fc-glycosylation profiling reveals associations with age, sex, female sex hormones and thyroid cancer," *Journal of proteomics*, vol. 75, no. 10, pp. 2824–2834, 2012.

[54] M. N. Christiansen, J. Chik, L. Lee, M. Anugraham, J. L. Abrahams, and N. H. Packer, "Cell surface protein glycosylation in cancer," *Proteomics*, vol. 14, no. 4-5, pp. 525–546, 2014.

[55] E. W. Adams, D. M. Ratner, H. R. Bokesch, J. B. McMahon, B. R. O'Keefe, and P. H. Seeberger, "Oligosaccharide and glycoprotein microarrays as tools in hiv glycobiology: glycan-dependent gp120/protein interactions," *Chemistry & biology*, vol. 11, no. 6, pp. 875–881, 2004.

[56] R. Pejchal, K. J. Doores, L. M. Walker, R. Khayat, P.-S. Huang, S.-K. Wang, R. L. Stanfield, J.-P. Julien, A. Ramos, M. Crispin, *et al.*, "A potent and broad neutralizing antibody recognizes and penetrates the hiv glycan shield," *Science*, p. 1213256, 2011.

[57] H. Mouquet, L. Scharf, Z. Euler, Y. Liu, C. Eden, J. F. Scheid, A. Halper-Stromberg, P. N. Gnanapragasam, D. I. Spencer, M. S. Seaman, *et al.*, "Complex-type n-glycan recognition by potent broadly neutralizing hiv antibodies," *Proceedings of the National Academy of Sciences*, vol. 109, no. 47, pp. E3268–E3277, 2012.

[58] P. L. Moore, E. S. Gray, C. K. Wibmer, J. N. Bhiman, M. Nonyane, D. J. Sheward, T. Hermanus, S. Bajimaya, N. L. Tumba, M.-R. Abrahams, *et al.*, "Evolution of an hiv glycan–dependent broadly neutralizing antibody epitope through immune escape," *Nature medicine*, vol. 18, no. 11, p. 1688, 2012.

[59] S. J. Danishefsky and J. R. Allen, "From the laboratory to the clinic: a retrospective on fully synthetic carbohydrate-based anticancer vaccines," *Angewandte Chemie International Edition*, vol. 39, no. 5, pp. 836–863, 2000.

[60] S. J. Danishefsky, Y.-K. Shue, M. N. Chang, and C.-H. Wong, "Development of globo-h cancer vaccine," *Accounts of chemical research*, vol. 48, no. 3, pp. 643–652, 2015.

[61] K. Landsteiner, "Zur kenntnis der antifermentativen, lytischen und agglutinierenden wirkungen des blutserums und der lymphe," *Zentralbl. Bakteriol.*, vol. 27, pp. 357–362, 1900.

[62] G. Stiver, "The treatment of influenza with antiviral drugs," *Canadian Medical Association Journal*, vol. 168, no. 1, pp. 49–57, 2003.

[63] N. P. Johnson and J. Mueller, "Updating the accounts: global mortality of the 1918-1920" spanish" influenza pandemic," *Bulletin of the History of Medicine*, pp. 105–115, 2002.

[64] R. Jain and R. D. Goldman, "Novel influenza a (h1n1): clinical presentation, diagnosis, and management," *Pediatric emergency care*, vol. 25, no. 11, pp. 791–796, 2009.

[65] F. Nimmerjahn and J. V. Ravetch, "Fcγ receptors as regulators of immune responses," *Nature Reviews Immunology*, vol. 8, no. 1, pp. 34–47, 2008.

[66] J. Stadlmann, M. Pabst, and F. Altmann, "Analytical and functional aspects of antibody sialylation," *Journal of clinical immunology*, vol. 30, no. 1, pp. 15–19, 2010.

[67] K. R. Anumula, "Quantitative glycan profiling of normal human plasma derived immunoglobulin and its fragments fab and fc," *Journal of immunological methods*, vol. 382, no. 1-2, pp. 167–176, 2012.

[68] A. Bondt, Y. Rombouts, M. H. Selman, P. J. Hensbergen, K. R. Reiding, J. M. Hazes, R. J. Dolhain, and M. Wuhrer, "Igg fab glycosylation analysis using a new mass spectrometric high-throughput profiling method reveals pregnancy-associated changes," *Molecular & Cellular Proteomics*, pp. mcp–M114, 2014.

[69] J. N. Arnold, M. R. Wormald, R. B. Sim, P. M. Rudd, and R. a. Dwek, "The impact of glycosylation on the biological function and structure of human immunoglobulins.," *Annual review of immunology*, vol. 25, pp. 21–50, 2007.

[70] F. S. van de Bovenkamp, L. Hafkenscheid, T. Rispens, and Y. Rombouts, "The emerging importance of igg fab glycosylation in immunity," *The Journal of Immunology*, vol. 196, no. 4, pp. 1435–1441, 2016.

[71] G. P. Subedi and A. W. Barb, "The Structural Role of Antibody N-Glycosylation in Receptor Interactions," *Structure*, vol. 23, pp. 1573–1583, sep 2015.

[72] G. P. Subedi and A. W. Barb, "The immunoglobulin G1 N-glycan composition affects binding to each low affinity Fc gamma receptor.," *MAbs*, pp. 1–13, 2016.

[73] C. N. Scanlan, D. R. Burton, and R. A. Dwek, "Making autoantibodies safe," *Proceedings of the National Academy of Sciences*, vol. 105, no. 11, pp. 4081–4082, 2008.

[74] S. Iida, H. Misaka, M. Inoue, M. Shibata, R. Nakano, N. Yamane-Ohnuki, M. Wakitani, K. Yano, K. Shitara, and M. Satoh, "Nonfucosylated therapeutic igg1 antibody can evade the inhibitory effect of serum immunoglobulin g on antibody-dependent cellular cytotoxicity through its high binding to fcγriiia," *Clinical Cancer Research*, vol. 12, no. 9, pp. 2879–2887, 2006.

[75] T. Shinkawa, K. Nakamura, N. Yamane, E. Shoji-Hosaka, Y. Kanda, M. Sakurada, K. Uchida, H. Anazawa, M. Satoh, M. Yamasaki, *et al.*, "The absence of fucose but not the presence of galactose or bisecting n-acetylglucosamine of human igg1 complex-type oligosaccharides shows the critical role of enhancing antibody-dependent cellular cytotoxicity," *Journal of Biological Chemistry*, vol. 278, no. 5, pp. 3466–3473, 2003.

[76] M. Pucić, A. Knezević, J. Vidic, B. Adamczyk, M. Novokmet, O. Polasek, O. Gornik, S. Supraha-Goreta, M. R. Wormald, I. Redzić, H. Campbell, A. Wright, N. D. Hastie, J. F. Wilson, I. Rudan, M. Wuhrer, P. M. Rudd, D. Josić, G. Lauc, M. Pucic, A. Knezevic, J. Vidic, B. Adamczyk, M. Novokmet, O. Polasek, O. Gornik, S. Supraha-Goreta, M. R. Wormald, I. Redzic, H. Campbell, A. Wright, N. D. Hastie, J. F. Wilson, I. Rudan, M. Wuhrer, P. M. Rudd, D. Josic, and G. Lauc, "High throughput isolation and glycosylation analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations," *Mol Cell Proteomics*, vol. 10, p. M111.010090, oct 2011.

[77] A. Samuelsson, T. L. Towers, and J. V. Ravetch, "Anti-inflammatory activity of ivig mediated through the inhibitory fc receptor," *Science*, vol. 291, no. 5503, pp. 484–486, 2001.

[78] H. Albert, M. Collin, D. Dudziak, J. V. Ravetch, and F. Nimmerjahn, "In vivo enzymatic modulation of igg glycosylation inhibits autoimmune disease in an igg subclass-dependent manner," *Proceedings of the National Academy of Sciences*, vol. 105, no. 39, pp. 15005–15009, 2008.

[79] G. Lauc, J. E. Huffman, M. Pučić, L. Zgaga, B. Adamczyk, A. Mužinić, M. Novokmet, O. Polašek, O. Gornik, J. Krištić, T. Keser, V. Vitart, B. Scheijen, H. W. Uh, M. Molokhia, A. L. Patrick, P. McKeigue, I. Kolčić, I. K. Lukić, O. Swann, F. N. van Leeuwen, L. R. Ruhaak, J. J. Houwing-Duistermaat, P. E. Slagboom, M. Beekman, A. J. M. de Craen, A. M. Deelder, Q. Zeng, W. Wang, N. D. Hastie, U. Gyllensten, J. F. Wilson, M. Wuhrer, A. F. Wright, P. M. Rudd, C. Hayward, Y. Aulchenko, H. Campbell, and I. Rudan, "Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers," *PLoS Genetics*, vol. 9, no. 1, p. e1003225, 2013.

[80] K. Kodar, J. Stadlmann, K. Klaamas, B. Sergeyev, and O. Kurtenkov, "Immunoglobulin g fc n-glycan profiling in patients with gastric cancer by lc-esi-ms: relation to tumor progression and survival," *Glycoconjugate journal*, vol. 29, no. 1, pp. 57–66, 2012.

[81] G. Lauc, J. E. Huffman, M. Pučić, L. Zgaga, B. Adamczyk, A. Mužinić, M. Novokmet, O. Polašek, O. Gornik, J. Krištić, *et al.*, "Loci associated with n-glycosylation of human immunoglobulin g show pleiotropy with autoimmune diseases and haematological cancers," *PLoS genetics*, vol. 9, no. 1, p. e1003225, 2013.

[82] R. Jefferis and M.-P. Lefranc, "Human immunoglobulin allotypes: possible implications for immunogenicity," *MAbs*, vol. 1, no. 4, pp. 332–338, 2009.

[83] L. R. Snyder, J. J. Kirkland, and J. W. Dolan, *Introduction to modern liquid chromatography*. John Wiley & Sons, 2011.

[84] F. W. Aston *et al.*, *Mass spectra and isotopes*. Edward Arnold London, 1942.

[85] M. E. Swartz, "Uplc$^{\text{TM}}$: an introduction and review," *Journal of Liquid Chromatography & Related Technologies*, vol. 28, no. 7-8, pp. 1253–1263, 2005.

[86] F. Hillenkamp and J. Peter-Katalinic, *MALDI MS: a practical guide to instrumentation, methods and applications*. John Wiley & Sons, 2013.

[87] M. Karas, D. Bachmann, and F. Hillenkamp, "Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules," *Analytical chemistry*, vol. 57, no. 14, pp. 2935–2939, 1985.

[88] J. B. Fenn, "Electrospray wings for molecular elephants (nobel lecture)," *Angewandte chemie international edition*, vol. 42, no. 33, pp. 3871–3894, 2003.

[89] S. Souverain, S. Rudaz, and J.-L. Veuthey, "Matrix effect in lc-esi-ms and lc-apci-ms with off-line and on-line extraction procedures," *Journal of Chromatography A*, vol. 1058, no. 1-2, pp. 61–66, 2004.

[90] J. E. Huffman, M. Pučić-Baković, L. Klarić, R. Hennig, M. H. J. Selman, F. Vučković, M. Novokmet, J. Krištić, M. Borowiak, T. Muth, O. Polašek, G. Razdorov, O. Gornik, R. Plomp, E. Theodoratou, A. F. Wright, I. Rudan, C. Hayward, H. Campbell, A. M. Deelder, U. Reichl, Y. S. Aulchenko, E. Rapp, M. Wuhrer, and G. Lauc, "Comparative performance of four methods for high-throughput glycosylation analysis of immunoglobulin G in genetic and epidemiological research.," *Molecular & cellular proteomics : MCP*, vol. 13, no. 6, pp. 1598–610, 2014.

[91] M. Balbin, A. Grubb, G. G. de Lange, and R. Grubb, "DNA sequences specific for Caucasian G3m(b) and (g) allotypes: allotyping at the genomic level," *Immunogenetics*, vol. 39, pp. 187–193, 1994.

[92] D. J. Harvey, "Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates," *Mass Spectrometry Reviews*, vol. 18, no. 6, pp. 349–450, 1999.

[93] K. Ohtsubo and J. D. Marth, "Glycosylation in cellular mechanisms of health and disease," *Cell*, vol. 126, no. 5, pp. 855–867, 2006.

[94] O. Gornik and G. Lauc, "Glycosylation of serum proteins in inflammatory diseases," *Disease markers*, vol. 25, no. 4, 5, pp. 267–278, 2008.

[95] C. Kirmiz, B. Li, H. J. An, B. H. Clowers, H. K. Chew, K. S. Lam, A. Ferrige, R. Alecio, A. D. Borowsky, S. Sulaimon, *et al.*, "A serum glycomics approach to breast cancer biomarkers," *Molecular & Cellular Proteomics*, vol. 6, no. 1, pp. 43–55, 2007.

[96] C. B. Lebrilla and H. J. An, "The prospects of glycan biomarkers for the diagnosis of diseases," *Molecular bioSystems*, vol. 5, no. 1, pp. 17–20, 2009.

[97] Y. Mechref, Y. Hu, A. Garcia, and A. Hussein, "Identifying cancer biomarkers by mass spectrometry-based glycomics," *Electrophoresis*, vol. 33, no. 12, pp. 1755–1767, 2012.

[98] B. Adamczyk, T. Tharmalingam, and P. M. Rudd, "Glycans as cancer biomarkers," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1820, no. 9, pp. 1347–1353, 2012.

[99] F. Dall'Olio, V. Vanhooren, C. C. Chen, P. E. Slagboom, M. Wuhrer, and C. Franceschi, "N-glycomic biomarkers of biological aging and longevity: a link with inflammaging," *Ageing research reviews*, vol. 12, no. 2, pp. 685–698, 2013.

[100] L. R. Ruhaak, S. Miyamoto, and C. B. Lebrilla, "Developments in the identification of glycan biomarkers for the detection of cancer," *Molecular & Cellular Proteomics*, pp. mcp–R112, 2013.

[101] D. H. Dube and C. R. Bertozzi, "Glycans in cancer and inflammation—potential for therapeutics and diagnostics," *Nature reviews Drug discovery*, vol. 4, no. 6, p. 477, 2005.

[102] V. Vanhooren, L. Desmyter, X.-E. Liu, M. Cardelli, C. Franceschi, A. Federico, C. Libert, W. Laroy, S. Dewaele, R. Contreras, and C. Chen, "N-Glycomic Changes in Serum Proteins During Human Aging," *Rejuvenation Research*, vol. 10, pp. 521–531a, dec 2007.

[103] Z. Kyselova, Y. Mechref, P. Kang, J. A. Goetz, L. E. Dobrolecki, G. W. Sledge, L. Schnaper, R. J. Hickey, L. H. Malkas, and M. V. Novotny, "Breast cancer diagnosis and prognosis through quantitative measurements of serum glycan profiles," *Clinical chemistry*, vol. 54, no. 7, pp. 1166–1175, 2008.

[104] G. Liu and S. Neelamegham, "Integration of systems glycobiology with bioinformatics toolboxes, glycoinformatics resources, and glycoproteomics data," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 7, no. 4, pp. 163–181, 2015.

[105] J. J. Houwing-Duistermaat, H. W. Uh, and A. Gusnanto, "Discussion on the paper 'statistical contributions to bioinformatics: Design, modelling, structure learning and integration'by jeffrey s. morris and veerabhadran baladandayuthapani," *Statistical Modelling*, vol. 17, no. 4-5, pp. 319–326, 2017.

[106] R. Hennig, S. Cajic, M. Borowiak, M. Hoffmann, R. Kottler, U. Reichl, and E. Rapp, "Towards personalized diagnostics via longitudinal study of the human plasma n-glycome," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1860, no. 8, pp. 1728–1738, 2016.

[107] D. M. Johnstone, C. Riveros, M. Heidari, R. M. Graham, D. Trinder, R. Berretta, J. K. Olynyk, R. J. Scott, P. Moscato, and E. A. Milward, "Evaluation of Different Normalization and Analysis Procedures for Illumina Gene Expression Microarray Data Involving Small Changes.," *Microarrays (Basel, Switzerland)*, vol. 2, pp. 131–52, may 2013.

[108] T. Välikangas, T. Suomi, and L. L. Elo, "A systematic evaluation of normalization methods in quantitative label-free proteomics," *Briefings in Bioinformatics*, vol. 19, p. bbw095, oct 2016.

[109] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data," *BMC Genomics*, vol. 7, p. 142, jun 2006.

[110] S. M. Kohl, M. S. Klein, J. Hochrein, P. J. Oefner, R. Spang, and W. Gronwald, "State-of-the art data normalization methods improve NMR-based metabolomic analysis," *Metabolomics*, vol. 8, pp. 146–160, jun 2012.

[111] B. Li, J. Tang, Q. Yang, X. Cui, S. Li, S. Chen, Q. Cao, W. Xue, N. Chen, and F. Zhu, "Performance evaluation and online realization of data-driven normalization methods used in lc/ms based untargeted metabolomics analysis," *Scientific reports*, vol. 6, p. 38881, 2016.

[112] H. V. Westerhoff and B. O. Palsson, "The evolution of molecular biology into systems biology," *Nature biotechnology*, vol. 22, no. 10, p. 1249, 2004.

[113] L. Wu, S. I. Candille, Y. Choi, D. Xie, L. Jiang, J. Li-Pook-Than, H. Tang, and M. Snyder, "Variation and genetic control of protein abundance in humans," *Nature*, vol. 499, no. 7456, p. 79, 2013.

[114] K. A. Lawton, A. Berger, M. Mitchell, K. E. Milgram, A. M. Evans, L. Guo, R. W. Hanson, S. C. Kalhan, J. A. Ryals, and M. V. Milburn, "Analysis of the adult human plasma metabolome," 2008.

[115] J. Bartel, J. Krumsiek, and F. J. Theis, "Statistical methods for the analysis of high-throughput metabolomics data," *Computational and structural biotechnology journal*, vol. 4, no. 5, p. e201301009, 2013.

[116] J. Krumsiek, K. Suhre, A. M. Evans, M. W. Mitchell, R. P. Mohney, M. V. Milburn, B. Wägele, W. Römisch-Margl, T. Illig, J. Adamski, C. Gieger, F. J. Theis, and G. Kastenmüller, "Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information," *PLoS Genetics*, vol. 8, p. e1003005, oct 2012.

[117] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models—a review," *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.

[118] C. P. Wild, "The exposome: from concept to utility," *International journal of epidemiology*, vol. 41, no. 1, pp. 24–32, 2012.

[119] V. Canuel, B. Rance, P. Avillach, P. Degoulet, and A. Burgun, "Translational research platforms integrating clinical and omics data: a review of publicly available solutions," *Briefings in bioinformatics*, vol. 16, no. 2, pp. 280–290, 2014.

[120] R. J. Muirhead, *Aspects of multivariate statistical theory*, vol. 197. John Wiley & Sons, 2009.

[121] S. L. Lauritzen, *Graphical models*, vol. 17. Clarendon Press, 1996.

[122] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.

[123] Z. Šidák, "Rectangular confidence regions for the means of multivariate normal distributions," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 626–633, 1967.

[124] A. De La Fuente, N. Bing, I. Hoeschele, and P. Mendes, "Discovery of meaningful associations in genomic data using partial correlation coefficients," *Bioinformatics*, vol. 20, no. 18, pp. 3565–3574, 2004.

[125] A. V. Werhli, M. Grzegorczyk, and D. Husmeier, "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks," *Bioinformatics*, vol. 22, no. 20, pp. 2523–2531, 2006.

[126] D. Veiga, F. Vicente, M. Grivet, A. De la Fuente, and A. Vasconcelos, "Genome-wide partial correlation analysis of escherichia coli microarray data," *Genet Mol Res*, vol. 6, no. 4, pp. 730–742, 2007.

[127] A. Adourian, E. Jennings, R. Balasubramanian, W. M. Hines, D. Damian, T. N. Plasterer, C. B. Clish, P. Stroobant, R. McBurney, E. R. Verheij, *et al.*, "Correlation network analysis for data integration and biomarker selection," *Molecular BioSystems*, vol. 4, no. 3, pp. 249–259, 2008.

[128] S. Andorf, J. Selbig, T. Altmann, K. Poos, H. Witucka-Wall, and D. Repsilber, "Enriched partial correlations in genome-wide gene expression profiles of hybrids (a. thaliana): a systems biological approach towards the molecular basis of heterosis," *Theoretical and applied genetics*, vol. 120, no. 2, p. 249, 2010.

[129] M. Layeghifard, D. M. Hwang, and D. S. Guttman, "Disentangling interactions in the microbiome: a network perspective," *Trends in microbiology*, vol. 25, no. 3, pp. 217–228, 2017.

[130] I. Rudan, A. Marušić, S. Janković, K. Rotim, M. Boban, G. Lauc, I. Grković, Z. Đogaš, T. Zemunik, Z. Vatavuk, G. Benčić, D. Rudan, R. Mulić, V. Krželj, J. Terzić, D. Stojanović, D. Puntarić, E. Bilić, D. Ropac, A. Vorko-Jović, A. Znaor, R. Stevanović, Z. Biloglav, and O. Polašek, ""10+001 Dalmatians:" Croatia Launches Its National Biobank," *Croatian Medical Journal*, vol. 50, no. 1, pp. 4–6, 2009.

[131] M. H. Selman, R. J. Derks, A. Bondt, M. Palmblad, B. Schoenmaker, C. A. Koeleman, F. E. van de Geijn, R. J. Dolhain, A. M. Deelder, and M. Wuhrer, "Fc specific IgG glycosylation profiling by robust nano-reverse phase HPLC-MS using a sheath-flow ESI sprayer interface," *Journal of Proteomics*, vol. 75, no. 4, pp. 1318–1329, 2012.

[132] F. Vučković, E. Theodoratou, K. Thaçi, M. Timofeeva, A. Vojta, J. Štambuk, M. Pučić-Baković, P. M. Rudd, L. Đerek, D. Servis, A. Wennerström, S. M. Farrington, M. Perola, Y. Aulchenko, M. G. Dunlop, H. Campbell, and G. Lauc, "IgG Glycome in Colorectal Cancer.," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 22, pp. 3078–86, jun 2016.

[133] A. L. Tarentino, C. M. Gomez, and T. H. Plummer Jr, "Deglycosylation of asparagine-linked glycans by peptide: N-glycosidase f," *Biochemistry*, vol. 24, no. 17, pp. 4665–4671, 1985.

[134] K. R. Reiding, L. R. Ruhaak, H.-W. Uh, S. El Bouhaddani, E. B. van den Akker, R. Plomp, L. A. McDonnell, J. J. Houwing-Duistermaat, P. E. Slagboom, M. Beekman, *et al.*, "Human plasma n-glycosylation as analyzed by matrix-assisted laser desorption/ionization-fourier transform ion cyclotron resonance-ms associates with markers of inflammation and metabolic health," *Molecular & Cellular Proteomics*, vol. 16, no. 2, pp. 228–242, 2017.

[135] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1H NMR metabonomics," *Analytical Chemistry*, vol. 78, no. 13, pp. 4281–4290, 2006.

[136] A. L. Koch, "The logarithm in biology 1. Mechanisms generating the log-normal distribution exactly," *Journal of Theoretical Biology*, vol. 12, pp. 276–290, nov 1966.

[137] J. Aitchison, "The Statistical Analysis of Compositional Data," *Chapman and Hall*, 1986.

[138] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[139] A. Tsodikov, A. Szabo, and D. Jones, "Adjustments and measures of differential expression for microarray data," *Bioinformatics*, vol. 18, no. 2, pp. 251–260, 2002.

[140] H. E. Wichmann, C. Gieger, and T. Illig, "KORA-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes," 2005.

[141] K. T. Do, S. Wahl, J. Raffler, S. Molnos, M. Laimighofer, J. Adamsky, K. Suhre, K. Strauch, A. Peters, C. Gieger, C. Langenberg, I. Stewart, F. J. Theis, H. Grallert, G. Kastenmueller, and J. Krumsiek, "Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies," *bioRxiv*, p. 260281, mar 2018.

[142] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, J. Zhang, C. Kandoth, R. Akbani, H. Shen, L. Omberg, A. Chu, A. A. Margolin, L. J. van't Veer, N. Lopez-Bigas, P. W. Laird, B. J. Raphael, L. Ding, A. G. Robertson, L. A. Byers, G. B. Mills, J. N. Weinstein, C. Van Waes, Z. Chen, E. A. Collisson, C. C. Benz, C. M. Perou, and J. M. Stuart, "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin," *Cell*, vol. 158, pp. 929–944, aug 2014.

[143] C. Furusawa, T. Suzuki, A. Kashiwagi, T. Yomo, and K. Kaneko, "Ubiquity of log-normal distributions in intra-cellular reaction dynamics," *Biophysics*, vol. 1, pp. 25–31, 2005.

[144] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.," *Statistical applications in genetics and molecular biology*, vol. 4, p. Article32, 2005.

[145] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, pp. 365–411, feb 2004.

[146] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, vol. 8, pp. 3–62, 1936.

[147] J. Schäfer, R. Opgen-Rhein, and K. Strimmer, "Reverse engineering genetic networks using the genenet package," *The Newsletter of the R Project Volume 6/5, December 2006*, vol. 1, no. 9, p. 50, 2006.

[148] J. Yin and H. Li, "A sparse conditional gaussian graphical model for analysis of genetical genomics data," *The annals of applied statistics*, vol. 5, no. 4, p. 2630, 2011.

[149] N. Swainston, K. Smallbone, H. Hefzi, P. D. Dobson, J. Brewer, M. Hanscho, D. C. Zielinski, K. S. Ang, N. J. Gardiner, J. M. Gutierrez, S. Kyriakopoulos, M. Lakshmanan, S. Li, J. K. Liu, V. S. Martínez, C. A. Orellana, L.-E. Quek, A. Thomas, J. Zanghellini, N. Borth, D.-Y. Lee, L. K. Nielsen, D. B. Kell, N. E. Lewis, and P. Mendes, "Recon 2.2: from reconstruction to model of human metabolism.," *Metabolomics : Official journal of the Metabolomic Society*, vol. 12, p. 109, 2016.

[150] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio, "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 42, pp. D472–D477, jan 2014.

[151] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio, "The Reactome Pathway Knowledgebase," *Nucleic Acids Research*, vol. 46, pp. D649–D655, jan 2018.

[152] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, "STRING v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, jan 2015.

[153] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Research*, vol. 45, pp. D362–D368, jan 2017.

[154] R. A. Fisher, "On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, vol. 85, p. 87, jan 1922.

[155] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, p. 026113, feb 2004.

[156] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *Proceedings of the 2005 SIAM international conference on data mining*, pp. 274–285, SIAM, 2005.

[157] S. Maslov and K. Sneppen, "Specificity and Stability in Topology of Protein Networks," *Science*, vol. 296, pp. 910–913, may 2002.

[158] P. Wong, S. Althammer, A. Hildebrand, A. Kirschner, P. Pagel, B. Geissler, P. Smialowski, F. Blöchl, M. Oesterheld, T. Schmidt, N. Strack, F. J. Theis, A. Ruepp, and D. Frishman, "An evolutionary and structural characterization of mammalian protein complex organization," *BMC Genomics*, vol. 9, p. 629, dec 2008.

[159] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC.," *Genome research*, vol. 12, pp. 996–1006, jun 2002.

[160] M. Arnold, J. Raffler, A. Pfeufer, K. Suhre, and G. Kastenmüller, "SNiPA: an interactive, genetic variant-centered annotation browser.," *Bioinformatics (Oxford, England)*, vol. 31, pp. 1334–6, apr 2015.

[161] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies.," *Nature reviews. Genetics*, vol. 11, pp. 499–511, jul 2010.

[162] O. A. Panagiotou and J. P. A. Ioannidis, "What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations.," *International journal of epidemiology*, vol. 41, pp. 273–86, feb 2012.

[163] C. Gieger, L. Geistlinger, E. Altmaier, M. Hrabé de Angelis, F. Kronenberg, T. Meitinger, H.-W. Mewes, H.-E. Wichmann, K. M. Weinberger, J. Adamski, T. Illig, and K. Suhre, "Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum.," *PLoS genetics*, vol. 4, p. e1000282, nov 2008.

[164] A.-K. Petersen, J. Krumsiek, B. Wägele, F. J. Theis, H.-E. Wichmann, C. Gieger, and K. Suhre, "On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies.," *BMC bioinformatics*, vol. 13, no. 1, p. 120, 2012.

[165] K. Suhre, S.-Y. Shin, A.-K. Petersen, R. P. Mohney, D. Meredith, B. Wägele, E. Altmaier, CARDIoGRAM, P. Deloukas, J. Erdmann, E. Grundberg, C. J. Hammond, M. H. de Angelis, G. Kastenmüller, A. Köttgen, F. Kronenberg, M. Mangino, C. Meisinger, T. Meitinger, H.-W. Mewes, M. V. Milburn, C. Prehn, J. Raffler, J. S.

Ried, W. Römisch-Margl, N. J. Samani, K. S. Small, H. -Erich Wichmann, G. Zhai, T. Illig, T. D. Spector, J. Adamski, N. Soranzo, and C. Gieger, "Human metabolic individuality in biomedical and pharmaceutical research," *Nature*, vol. 477, no. 7362, pp. 54–60, 2011.

[166] S.-Y. Shin, E. B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T.-P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A. M. Valdes, C. L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, M. Waldenberger, J. B. Richards, R. P. Mohney, M. V. Milburn, S. L. John, J. Trimmer, F. J. Theis, J. P. Overington, K. Suhre, M. J. Brosnan, C. Gieger, G. Kastenmüller, T. D. Spector, and N. Soranzo, "An atlas of genetic influences on human blood metabolites," *Nature Genetics*, vol. 46, no. 6, pp. 543–550, 2014.

[167] J. A. Voynow, R. S. Kaiser, T. F. Scanlin, and M. C. Glick, "Purification and characterization of GDP-L-fucose-N-acetyl beta-D-glucosaminide alpha 1-6fucosyltransferase from cultured human skin fibroblasts. Requirement of a specific biantennary oligosaccharide as substrate.," *The Journal of biological chemistry*, vol. 266, pp. 21572–7, nov 1991.

[168] T. Nilsson, M. Pypaert, M. H. Hoe, P. Slusarewicz, E. G. Berger, and G. Warren, "Overlapping distribution of two glycosyltransferases in the Golgi apparatus of HeLa cells.," *The Journal of cell biology*, vol. 120, pp. 5–13, jan 1993.

[169] C. Rabouille, N. Hui, F. Hunte, R. Kieckbusch, E. G. Berger, G. Warren, and T. Nilsson, "Mapping the distribution of Golgi enzymes involved in the construction of complex oligosaccharides.," *Journal of cell science*, pp. 1617–27, apr 1995.

[170] E. G. Berger and F. J. Hesford, "Localization of galactosyl- and sialyltransferase by immunofluorescence: evidence for different sites.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 14, pp. 4736–9, 1985.

[171] E. G. Berger, M. Thurnher, and U. Müller, "Galactosyltransferase and sialyltransferase are located in different subcellular compartments in HeLa cells," *Exp Cell Res*, vol. 173, pp. 267–273, 1987.

[172] A. Hassinen and S. Kellokumpu, "Organizational Interplay of Golgi N-Glycosyltransferases Involves Organelle Microenvironment-Dependent Transitions between Enzyme Homo- and Heteromers," *Journal of Biological Chemistry*, vol. 289, pp. 26937–26948, sep 2014.

[173] N. Taniguchi, K. Honke, and M. Fukuda, *Handbook of glycosyltransferases and related genes*. Springer Science & Business Media, 2012.

[174] F. Clerc, K. R. Reiding, B. C. Jansen, G. S. M. Kammeijer, A. Bondt, and M. Wuhrer, "Human plasma protein N-glycosylation.," *Glycoconjugate journal*, nov 2015.

[175] A. Kobata, "The n-linked sugar chains of human immunoglobulin g: their unique pattern, and their functional roles," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1780, no. 3, pp. 472–478, 2008.

[176] E. Benedetti, M. Pučić-Baković, T. Keser, A. Wahl, A. Hassinen, J.-Y. Yang, L. Liu, I. Trbojević-Akmačić, G. Razdorov, J. Štambuk, L. Klarić, I. Ugrina, M. H. J. Selman, M. Wuhrer, I. Rudan, O. Polasek, C. Hayward, H. Grallert, K. Strauch, A. Peters, T. Meitinger, C. Gieger, M. Vilaj, G.-J. Boons, K. W. Moremen, T. Ovchinnikova, N. Bovin, S. Kellokumpu, F. J. Theis, G. Lauc, and J. Krumsiek, "Network inference from glycoproteomics data reveals new reactions in the IgG glycosylation pathway," *Nature Communications*, vol. 8, p. 1483, dec 2017.

[177] R. Saldova, A. Asadi Shehni, V. D. Haakensen, I. Steinfeld, M. Hilliard, I. Kifer, Å. Helland, Z. Yakhini, A.-L. Børresen-Dale, and P. M. Rudd, "Association of n-glycosylation with breast carcinoma and systemic features using high-resolution quantitative uplc," *Journal of proteome research*, vol. 13, no. 5, pp. 2314–2327, 2014.

[178] M. Pučić, A. Mužinić, M. Novokmet, M. Škledar, N. Pivac, G. Lauc, and O. Gornik, "Changes in plasma and igg n-glycome during childhood and adolescence," *Glycobiology*, vol. 22, no. 7, pp. 975–982, 2012.

[179] M. Novokmet, E. Lukić, F. Vučković, T. Keser, K. Rajšl, D. Remondini, G. Castellani, H. Gašparović, O. Gornik, G. Lauc, *et al.*, "Changes in igg and total plasma protein glycomes in acute systemic inflammation," *Scientific reports*, vol. 4, p. 4347, 2014.

[180] L. R. Ruhaak, H.-W. Uh, A. M. Deelder, R. E. Dolhain, and M. Wuhrer, "Total plasma n-glycome changes during pregnancy," *Journal of proteome research*, vol. 13, no. 3, pp. 1657–1668, 2014.

[181] R. Raman, S. Raguram, G. Venkataraman, J. C. Paulson, and R. Sasisekharan, "Glycomics: an integrated systems approach to structure-function relationships of glycans," *Nature Methods*, vol. 2, no. 11, p. 817, 2005.

[182] W. H. Yang, P. V. Aziz, D. M. Heithoff, M. J. Mahan, J. W. Smith, and J. D. Marth, "An intrinsic mechanism of secreted protein aging and turnover," *Proceedings of the National Academy of Sciences*, vol. 112, no. 44, pp. 13657–13662, 2015.

[183] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: and this is not optional," *Frontiers in microbiology*, vol. 8, p. 2224, 2017.

[184] F. Xia, J. Chen, W. K. Fung, and H. Li, "A logistic normal multinomial regression model for microbiome compositional data analysis," *Biometrics*, vol. 69, no. 4, pp. 1053–1063, 2013.

[185] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, and S. D. Peddada, "Analysis of composition of microbiomes: a novel method for studying microbial composition," *Microbial ecology in health and disease*, vol. 26, no. 1, p. 27663, 2015.

[186] G. B. Gloor and G. Reid, "Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data," *Canadian journal of microbiology*, vol. 62, no. 8, pp. 692–703, 2016.

[187] E. Z. Chen and H. Li, "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data," *Bioinformatics*, vol. 32, no. 17, pp. 2611–2617, 2016.

[188] P. Shi, A. Zhang, H. Li, *et al.*, "Regression analysis for microbiome compositional data," *The Annals of Applied Statistics*, vol. 10, no. 2, pp. 1019–1040, 2016.

[189] J. Aitchison, "Logratios and natural laws in compositional data analysis," *Mathematical Geology*, vol. 31, no. 5, pp. 563–580, 1999.

[190] J. Aitchison, C. Barceló-Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn, "Logratio analysis and compositional distance," *Mathematical Geology*, vol. 32, no. 3, pp. 271–275, 2000.

[191] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal, "Isometric logratio transformations for compositional data analysis," *Mathematical Geology*, vol. 35, no. 3, pp. 279–300, 2003.

[192] J. Aitchison and J. J. Egozcue, "Compositional data analysis: where are we and where should we be heading?," *Mathematical Geology*, vol. 37, no. 7, pp. 829–850, 2005.

[193] M. C. Tsilimigras and A. A. Fodor, "Compositional data analysis of the microbiome: fundamentals, tools, and challenges," *Annals of epidemiology*, vol. 26, no. 5, pp. 330–335, 2016.

[194] S. J. Callister, R. C. Barry, J. N. Adkins, E. T. Johnson, W.-j. Qian, B.-J. M. Webb-Robertson, R. D. Smith, and M. S. Lipton, "Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics," *Journal of proteome research*, vol. 5, no. 2, pp. 277–286, 2006.

[195] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for rna-seq data," *Genome biology*, vol. 14, no. 9, p. 3158, 2013.

[196] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon, "Scaling and normalization effects in nmr spectroscopic metabonomic data sets," *Analytical chemistry*, vol. 78, no. 7, pp. 2262–2267, 2006.

[197] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in rna-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, pp. 204–216, 2012.

[198] E. S. Moh, M. Thaysen-Andersen, and N. H. Packer, "Relative versus absolute quantitation in disease glycomics," *PROTEOMICS–Clinical Applications*, vol. 9, no. 3-4, pp. 368–382, 2015.

[199] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the biological information content of metabolomics data," *BMC genomics*, vol. 7, no. 1, p. 142, 2006.

[200] K. T. Do, M. Pietzner, D. J. Rasp, N. Friedrich, M. Nauck, T. Kocher, K. Suhre, D. O. Mook-Kanamori, G. Kastenmüller, and J. Krumsiek, "Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations," *NPJ systems biology and applications*, vol. 3, no. 1, p. 28, 2017.

[201] R. Albert, "Network inference, analysis, and modeling in systems biology.," *The Plant cell*, vol. 19, pp. 3327–38, nov 2007.

[202] H. Carter, M. Hofree, and T. Ideker, "Genotype to phenotype via network analysis," *Current Opinion in Genetics & Development*, vol. 23, pp. 611–621, dec 2013.

[203] A. K. Rider, T. Milenković, G. H. Siwo, R. S. Pinapati, S. J. Emrich, M. T. Ferdig, and N. V. Chawla, "Networks' Characteristics Matter for Systems Biology HHS Public Access," *Netw Sci*, vol. 213, no. 2, 2014.

[204] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, pp. 56–68, jan 2011.

[205] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types," *Nature Communications*, vol. 5, p. 3231, feb 2014.

[206] D. M. W. Powers, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION," *Journal of Machine Learning Technologies ISSN*, vol. 2, no. 1, pp. 2229–3981, 2011.

[207] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, pp. 442–451, oct 1975.

[208] G. Camilli, "The relationship between fisher's exact test and pearson's chi-square test: A bayesian perspective," *Psychometrika*, vol. 60, no. 2, pp. 305–312, 1995.

[209] Z. Wang, W. Xu, F. A. San Lucas, and Y. Liu, "Incorporating prior knowledge into Gene Network Study," *Bioinformatics*, vol. 29, pp. 2633–2640, oct 2013.

[210] J. Linde, S. Schulze, S. G. Henkel, and R. Guthke, "Data- and knowledge-based modeling of gene regulatory networks: an update," *EXCLI Journal*, vol. 14, p. 346, 2015.

[211] B. Pei and D.-G. Shin, "Reconstruction of Biological Networks by Incorporating Prior Knowledge into Bayesian Network Models," *Journal of Computational Biology*, vol. 19, pp. 1324–1334, dec 2012.

[212] Y. Zuo, Y. Cui, G. Yu, R. Li, and H. W. Ressom, "Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO," *BMC Bioinformatics*, vol. 18, p. 99, dec 2017.

[213] M. Ante, E. Wingender, and M. Fuchs, "Integration of gene expression data with prior knowledge for network analysis and validation," *BMC Research Notes*, vol. 4, no. 1, p. 520, 2011.

[214] V. Stavrakas, I. N. Melas, T. Sakellaropoulos, and L. G. Alexopoulos, "Network Reconstruction Based on Proteomic Data and Prior Knowledge of Protein Connectivity Using Graph Theory," *PLOS ONE*, vol. 10, p. e0128411, may 2015.

[215] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, jul 1948.

[216] M. J. Gramer, C. F. Goochee, V. Y. Chock, D. T. Brousseau, and M. B. Sliwkowski, "Removal of sialic acid from a glycoprotein in cho cell culture supernatant by action of an extracellular cho cell sialidase," *Nature Biotechnology*, vol. 13, no. 7, p. 692, 1995.

[217] P. M. Rudd and R. A. Dwek, "Glycosylation: heterogeneity and the 3d structure of proteins," *Critical reviews in biochemistry and molecular biology*, vol. 32, no. 1, pp. 1–100, 1997.

[218] N. S. Wong, L. Wati, P. M. Nissom, H. Feng, M. Lee, and M. G. Yap, "An investigation of intracellular glycosylation activities in cho cells: effects of nucleotide sugar precursor feeding," *Biotechnology and bioengineering*, vol. 107, no. 2, pp. 321–336, 2010.

[219] W. P. Rijcken, B. Overdijk, D. Van den Eijnden, and W. Ferwerda, "The effect of increasing nucleotide-sugar concentrations on the incorporation of sugars into glycoconjugates in rat hepatocytes," *Biochemical Journal*, vol. 305, no. 3, pp. 865–870, 1995.

[220] G. Lauc and V. Zoldoš, "Epigenetic regulation of glycosylation could be a mechanism used by complex organisms to compete with microbes on an evolutionary scale," *Medical hypotheses*, vol. 73, no. 4, pp. 510–512, 2009.

[221] G. Lauc, A. Vojta, and V. Zoldoš, "Epigenetic regulation of glycosylation is the quantum mechanics of biology," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1840, no. 1, pp. 65–70, 2014.

[222] J. Lübbehusen, C. Thiel, N. Rind, D. Ungar, B. H. Prinsen, T. J. de Koning, P. M. van Hasselt, and C. Körner, "Fatal outcome due to deficiency of subunit 6 of the conserved oligomeric golgi complex leading to a new type of congenital disorders of glycosylation," *Human molecular genetics*, vol. 19, no. 18, pp. 3623–3633, 2010.

[223] C. T. Campbell and K. J. Yarema, "Large-scale approaches for glycobiology," *Genome biology*, vol. 6, no. 11, p. 236, 2005.

[224] L. Krishnamoorthy and L. K. Mahal, "Glycomic analysis: an array of technologies," *ACS chemical biology*, vol. 4, no. 9, pp. 715–732, 2009.

[225] J. Stadlmann, J. Taubenschmid, D. Wenzel, A. Gattinger, G. Dürnberger, F. Dusberger, U. Elling, L. Mach, K. Mechtler, and J. M. Penninger, "Comparative glycoproteomics of stem cells identifies new players in ricin toxicity," *Nature*, vol. 549, no. 7673, p. 538, 2017.

[226] K. Suhre, M. Arnold, A. M. Bhagwat, R. J. Cotton, R. Engelke, J. Raffler, H. Sarwath, G. Thareja, A. Wahl, R. K. DeLisle, *et al.*, "Connecting genetic risk to disease end points through the human blood plasma proteome," *Nature communications*, vol. 8, p. 14357, 2017.

[227] G. Lauc, A. Essafi, J. E. Huffman, C. Hayward, A. Knežević, J. J. Kattla, O. Polašek, O. Gornik, V. Vitart, J. L. Abrahams, *et al.*, "Genomics meets glycomics—the first gwas study of human n-glycome identifies hnf1α as a master regulator of plasma protein fucosylation," *PLoS genetics*, vol. 6, no. 12, p. e1001256, 2010.

[228] T. Kamio, T. Toki, R. Kanezaki, S. Sasaki, S. Tandai, K. Terui, D. Ikebe, K. Igarashi, and E. Ito, "B-cell–specific transcription factor bach2 modifies the cytotoxic effects of anticancer drugs," *Blood*, vol. 102, no. 9, pp. 3317–3322, 2003.

[229] S. F. Grant, H.-Q. Qu, J. P. Bradfield, L. Marchand, C. E. Kim, J. T. Glessner, R. Grabs, S. P. Taback, E. C. Frackelton, A. W. Eckert, *et al.*, "Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes," *Diabetes*, vol. 58, no. 1, pp. 290–295, 2009.

[230] J. D. Cooper, D. J. Smyth, A. M. Smiles, V. Plagnol, N. M. Walker, J. E. Allen, K. Downes, J. C. Barrett, B. C. Healy, J. C. Mychaleckyj, *et al.*, "Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci," *Nature genetics*, vol. 40, no. 12, p. 1399, 2008.

[231] J. C. Barrett, D. G. Clayton, P. Concannon, B. Akolkar, J. D. Cooper, H. A. Erlich, C. Julier, G. Morahan, J. Nerup, C. Nierras, *et al.*, "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes," *Nature genetics*, vol. 41, no. 6, p. 703, 2009.

[232] V. Plagnol, J. M. Howson, D. J. Smyth, N. Walker, J. P. Hafler, C. Wallace, H. Stevens, L. Jackson, M. J. Simmonds, P. J. Bingley, *et al.*, "Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases," *PLoS genetics*, vol. 7, no. 8, p. e1002216, 2011.

[233] P. C. Dubois, G. Trynka, L. Franke, K. A. Hunt, J. Romanos, A. Curtotti, A. Zhernakova, G. A. Heap, R. Ádány, A. Aromaa, *et al.*, "Multiple common variants for

celiac disease influencing immune gene expression," *Nature genetics*, vol. 42, no. 4, p. 295, 2010.

[234] A. Franke, D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, *et al.*, "Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci," *Nature genetics*, vol. 42, no. 12, p. 1118, 2010.

[235] S. Sawcer, G. Hellenthal, M. Pirinen, C. C. Spencer, N. A. Patsopoulos, L. Moutsianas, A. Dilthey, Z. Su, C. Freeman, S. E. Hunt, *et al.*, "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis," *Nature*, vol. 476, no. 7359, p. 214, 2011.

[236] B. Pera, J. Krumsiek, S. E. Assouline, R. Marullo, J. Patel, J. M. Phillip, L. Román, K. K. Mann, and L. Cerchietti, "Metabolomic profiling reveals cellular reprogramming of b-cell lymphoma by a lysine deacetylase inhibitor through the choline pathway," *EBioMedicine*, vol. 28, pp. 80–89, 2018.

[237] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[238] G. A. P. Gonzalez, L. R. El Assal, A. Noronha, I. Thiele, H. S. Haraldsdóttir, and R. M. Fleming, "Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to recon 3d," *Journal of cheminformatics*, vol. 9, no. 1, p. 39, 2017.

[239] N. Washburn, I. Schwab, D. Ortiz, N. Bhatnagar, J. C. Lansing, A. Medeiros, S. Tyler, D. Mekala, E. Cochran, H. Sarvaiya, *et al.*, "Controlled tetra-fc sialylation of ivig results in a drug candidate with consistent enhanced anti-inflammatory activity," *Proceedings of the National Academy of Sciences*, p. 201422481, 2015.

[240] T. B. Parsons, W. B. Struwe, J. Gault, K. Yamamoto, T. A. Taylor, R. Raj, K. Wals, S. Mohammed, C. V. Robinson, J. L. Benesch, *et al.*, "Optimal synthetic glycosylation of a therapeutic antibody," *Angewandte Chemie International Edition*, vol. 55, no. 7, pp. 2361–2367, 2016.

[241] A. Varki, "Factors controlling the glycosylation," *trends in CELL BIOLOGY (Vol. 8)*, 1998.

[242] H. J. An, S. R. Kronewitter, M. L. A. de Leoz, and C. B. Lebrilla, "Glycomics and disease markers," *Current opinion in chemical biology*, vol. 13, no. 5-6, pp. 601–607, 2009.

[243] V. Padler-Karavani, "Aiming at the sweet side of cancer: aberrant glycosylation as possible target for personalized-medicine," *Cancer letters*, vol. 352, no. 1, pp. 102–112, 2014.

[244] N. Taniguchi and Y. Kizuka, "Glycans and cancer: role of n-glycans in cancer biomarker, progression and metastasis, and therapeutics," in *Advances in cancer research*, vol. 126, pp. 11–51, Elsevier, 2015.

[245] S. M. Muthana and J. C. Gildersleeve, "Glycan microarrays: powerful tools for biomarker discovery," *Cancer Biomarkers*, vol. 14, no. 1, pp. 29–41, 2014.

[246] L. Zhang, S. Luo, and B. Zhang, "Glycan analysis of therapeutic glycoproteins," in *MAbs*, vol. 8, pp. 205–215, Taylor & Francis, 2016.

[247] A. Russell, E. Adua, I. Ugrina, S. Laws, and W. Wang, "Unravelling immunoglobulin g fc n-glycosylation: A dynamic marker potentiating predictive, preventive and personalised medicine," *International journal of molecular sciences*, vol. 19, no. 2, p. 390, 2018.