Summary

Foreword

The candidate's PhD project has covered different research areas, ranging from gene expression and epigenetics to the ontological modeling of biomedical data. For this reason, the thesis is structured into three main chapters. Chapter 1 is devoted to the ontology for neurovascular conditions developed for the NEUROWEB Project. Chapter 2 is devoted to the integrative analysis of epigenetic and transcriptional data to elucidate the regulatory role of the transcription factor NF-Y. Chapter 3 is devoted to the transcript and miRNA profiling of the R9C transgenic mouse model for dilated cardiomyopathy. Every chapter is structured into the same sections: Background, Material and Methods, Results, Discussion, Acknowledgments and Reference.

The work described in the final chapter is currently in progress; therefore results should be regarded as preliminary.

The NEUROWEB Project

Modeling Clinical Conditions to Support Data Integration and Association Studies for Neurovascular Medicine.

Abstract

The NEUROWEB Project was started in 2006 to support neurovascular researchers performing *genotype-phenotype association studies*, intended as the search for statistical correlations between the phenotype and the genotype of patients affected by neurovascular disorders. In this project, the phenotype refers to the patients' pathological condition, and needs to be formulated on the basis of the clinical data collected during the diagnostic activity.

The availability of large data cohorts is an essential factor for the successfulness of association studies; for that reason, the project involves four different clinical institutions, which are located in different countries of the European Union, and which grant access to their clinical repositories to the other NEUROWEB Consortium members. Although all sites comply with common international guidelines, they adopt specific ways to organize the diagnostic activity and the patient data collected. As a consequence, the content of the different repositories cannot be directly integrated, unless misalignment problems are addressed.

To that end, we developed an ontological model for neurovascular pathological conditions, the *NEUROWEB Reference Ontology (NW-RO)*. The NW-RO, coupled to the NEUROWEB query system and user interface, enables to query the local repositories, and to retrieve clusters of patients characterized by the input clinical condition. This core functionality can be exploited not only for genotype-phenotype association studies, but also for epidemiological studies relating a clinical condition to environmental and demographic factors, or looking for associations between different clinical indicators. In addition, the terminological base attached to the NW-RO enables search functionalities on external resources, such as the retrieval of scientific papers from PubMed/Medline.

Bridging the clinical domain to biomolecular research is a crucial endeavor for biomedical ontology modeling, supporting the data integration and data analysis needs of the emerging field of *Translational Research*. Even though the final solution to this problem is larger than NEUROWEB Project reach, the NW-RO is endued with a prototypal extension to handle *biomolecular entities*, showing the flexibility and potentiality of the framework adopted.

Background

Biomedical informatics is a discipline including diverse applications such as clinical data storage and integration, telemedicine, decision support, text-mining, knowledge management and representation, not to mention real-time systems and embedded computing. This section will be devoted to introducing general concepts of specific interest for clinical data modeling and ontology design.

The Electronic Health Record

The notion of electronic health record (EHR)[1] [7, 8] is of primary importance to the field of medical informatics. A patient interacting with a health-care institution typically undergoes a process of information collection, generation and elaboration; these information items (i.e. clinical features) span different areas, such as identification and demographics, medical findings, life style, health history; they are organized into a general structure, the health record, which is instantiated for every patient; therefore, health records of different patients differ by values, rather than by content[2]. The importance of that information corpus predates the computer revolution, when health records were stored in a paper medium. As a matter of fact, the availability of health records enables an array of essential functions, such as clinical activity (diagnosis, prognosis and treatment), research, and administrative tasks.

The availability of information technology solutions for data storage and elaboration offers great opportunities to clinical medicine, but requires mastering the complexity of biomedical data; moreover, different tasks often pose specific issues.

The data stored in the EHR can be used for different goals, with specific inherent needs and perspectives. Four broad groups can be identified:

a. data access in the daily clinical practice;
b. decision support systems;
c. clinical trials, association studies, translational research;
d. administration and management;
e. document management and cooperative work support.

The NEUROWEB Project is primarily committed to the use group (c), clinical trials, association studies, translational research; however, the system was designed to offer at least some functionality even for the use group (a), data access in the daily clinical practice. In addition, it is useful to consider the differences between these two use cases, as health records adopted by single health-care institutions are often designed to primarily satisfy the use group (a), and later restructured, or coupled to additional information infrastructures[3], in order to be able to satisfy other use cases. In the next subsections the use groups (a) and (c) will be briefly described.

Data Access in the Daily Clinical Practice

In the first place, the health record is meant to be accessed by the clinician, in order to retrieve essential information required by his daily activity with patients.

The typical user-health record interaction is on a by-patient basis, rather than on a by-field basis; that is, the clinician is often more interested in accessing the value of sev-

---

[1] Electronic Medical Record (EMR) is often used as a synonym of EHR.

[2] Unless the data of the patients are stored according to different health record models.

[3] This is typically the case of the NEUROWEB Reference Ontology, that enables the federation of different databases for genotype-phenotype association studies.

eral fields from the health record of the same patient, rather than in the value of one or few fields across different patients[4].
The preference for the patient-based access also explains why clinicians largely resort to fields with natural language (NL) content to summarize the patient status in an expressive and intuitive way. The opposite access mode (on a field-basis) is more common for other use cases, such as in clinical trials and association studies.
The challenge here is to meet both these objectives:

a. maximize the exploitation of the electronic medium (for queries, statistical analysis, etc…);
b. offer to the clinical user a view of the data being meaningful and intuitive, and also compatible with his mental model and work organization.

Clinical Trials, Association Studies, Translational Research

Clinical trials, association studies and translational research require the use of clinical data in a context other than the clinical practice; the final end is not the health-care of a single patient, but rather the discovery or testing of some biomedical relation or property, of general interest and validity in the medical domain. The relations and properties assessed in these studies contribute to the implementation of Evidence-Based Medicine (EBM), which is the use of available evidence from systematic research in the clinical practice[5] [9]. Specifically,

- *clinical trials* consist in the testing of a drug or medication device, checking for its safety for the patients' health, and for its efficacy as a treatment (e.g.: is aspirin a good treatment for influenza?);
- *association studies* are performed to identify correlations between (a) one or more clinical features, including the genetic make-up of individuals, and (b) a clinical or pre-clinical disease condition, or disease progression behavior (e.g. is the sodium level in the diet associated to hypertension? Is low-meat diet slowing the progression of prostate neoplasm from benign to malignant?);
- *translational research* refers to the combination of biological research, at the molecular and cellular level, to clinical ends (e.g. identification of biomolecular disease markers for breast cancer).

Clinical trials and association studies present similar requisites:

- cluster the patients into groups with common clinical conditions (i.e. segmentation);
- large cohorts are necessary to grant statistical reliability; due to the rising costs of patient enrollment into ad-hoc cohorts, integration of existing data silos is an appealing solution; however, it is prone to flaws and biases, unless the content of different resources is aligned on a semantic basis, preserving the meaning and the methodological coherence [1, 4].

Translational research [10] is a broader term, generally referring to studies that connect the clinical realm to biomolecular research; specific types of clinical trials and as-

---

[4] For instance, a clinician may first record the patient's symptoms and signs on admission, and schedule a first array of exams; when the exam results are available, he will need to enter them into the health record, then go back to the symptoms and signs previously stored, consider other patient's data (e.g. demographics and medical history), and eventually formulate an initial diagnostic hypothesis. Considering another example, when the clinician formulates the treatment, he may have to check for the patient's drug intolerances.
Even when looking for *similar cases* (i.e. patients with a clinical profile similar to the one under examination), the health record is still accessed mostly on a by-patient basis. This use-case is usually triggered by the encounter of a clinical case not fitting into the clinician's experience, or the established guidelines (for diagnosis, treatment, etc…).

[5] "The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research" [9].

sociation studies (e.g. phenotype-genotype associations) are included as well. The common denominator here is the existence of discrepancies between the clinical and biomolecular domain [2], typically in terms of:

- *methodology*: clinical medicine is committed to curing disease and preserving human health; it is subject to very tight constraints in the way the object of study can be investigated and sampled (e.g. Alzheimer cannot be diagnosed extracting a sample from the patient's brain, and then performing an array of lab tests); on the contrary, biomolecular research is committed to the elucidation of biological mechanisms, and can exploit experimental designs causing the death or disability of non-human organisms (e.g. systematic gene deletion is used to identify genes essential for the organism's survival);
- *structural granularity*: biomolecular research principally operates at the level of molecules and cells, with a special accent laid on the genetic information encoded by nucleic acids; clinical medicine spans all organization levels, with traditional practice operating mainly at the organ and organ-system level;
- *target organism*: clinical medicine is devoted to the human organism, whereas biomolecular research largely exploits model organisms (e.g. yeast, fly, mouse), exhibiting variable levels of similarity to homo sapiens in terms of biological structures and processes.

Modeling frameworks supporting translational research must address these challenges, providing a bridge between the two disciplines [2].

The Contribution of Ontology to Clinical Modeling

Apparently, the *data* have a central role in health applications: data are transferred from existing records when a patient is admitted, they are generated ex-novo during the diagnostic profiling, they are accessed and elaborated during the formulation of the diagnosis, prognosis and treatment, or during the execution of other tasks; data from different repositories need to be integrated and inter-operated to enable certain applications.

However, data themselves do not provide any explicit model of their semantics and the rationale of their organization [11]. Ontologies, intended as formal representations of entities and their inter-relations, enable to overcome those limitations. Biomedical ontologies are often perceived just as a tool for terminological standardization; however, a more articulated perspective will be adopted here.

An Introduction to Ontologies

Ontologies are an important research topic in diverse communities of computer and information sciences, such as natural language processing, cooperative information systems, intelligent information integration, and knowledge management; they are also gaining momentum in the biomedical community [12, 13]. Originally, ontology is a branch of philosophy that deals with the theory of being, and considers questions about *what is* and *what is not*. The word derives from the Greek *onto* (being) and *logia* (written or spoken discourse) [14]. In computer science, ontologies are computational models representing the kind of entities, properties of entities, and relations between entities that can be ascribed to a specific domain of knowledge or application [14-17]; under an AI perspective, they can be regarded as *content theories*[6], in opposition to *mechanism theories* [18]. Ontologies provide a shared and common understanding of a domain that can be communicated across people and computer application systems; they are successfully used for integrating heterogeneous databases, enabling interoperability

[6] A catalogue of items in a store is the simplest example of a content theory.

among disparate systems, and specifying interfaces to knowledge-based services [11, 17].

Different Perspectives on Ontologies

The definition of ontology introduced in the previous sub-section is simple and universally accepted in the world of ontology design and development. However, different schools of thought, implicitly or explicitly following different philosophical approaches (e.g. nominalism, realism, conceptualism, conceptual realism, etc…), have proposed different perspectives on ontologies[7]. Discussing in detail the articulation and merit of such different positions is beyond the scope of this work. The perspective adopted here is that ontology is a *representational artifact*, and it has *reality* as referent; in other words, the ensemble of entities, properties and relations composing an ontology is committed to the representation of (a portion of) reality. However, reality is *not* understandable by human agents per se, i.e. in a purely objective fashion: the understanding of reality is always mediated by *interpretation schemes* (cf. the notion of *scientific paradigm* [19]), *mental models* [20], *perspectives* [21] and *contexts* [22]. The way reality is represented within ontology always depends on the epistemological perspective on reality inherently adopted[8]. This paradigm for ontology design was deemed optimal for scientific applications[9].

Ontology Meta-Models

An important methodological distinction exists between language-oriented ontologies (for which the name *terminologies* is more appropriate), and ontologies committed to a more comprehensive representation of semantics. In terminologies, the accent is laid on the catalogue of the linguistic terms used in a domain, their inter-relations, and their synonyms; typical applications are (a) content indexing for document retrieval, (b) information extraction, (c) data integration based on terminological matching. Under the paradigm adopted here, ontologies committed to a more comprehensive representation of semantics require an accurate analysis of knowledge sources, and, more specifically, how the inherent cognitive models, adopted in the domain of interest, should influence ontology design. In the most simple case of ontology, i.e. a taxonomy (more formally, a subsumption[10] tree), that analysis translates to the identification of taxonomical criteria, i.e. the reasons why certain entities are grouped into certain classes, and why certain classes are different than others. In the case of more elaborate ontologies, the same analysis has to be extended to the semantic of the other relations beyond subsumption; the formalization of such classificatory criteria enables the definition of an ontology meta-model[11].

If a meta-model succeeds in capturing the cognitive models common to different domains and applications, it can act as a foundational model of general validity[12]; such

---

[7] Compare the explicitly realist position exposed in [14] to the definition proposed by [15-17], making explicit reference to concepts and conceptualizations.

[8] This position can be related to *Conceptual Realism* (cf. [25]).

[9] Simplifying the problem, a scientific theory is often regarded as a working model of a portion of reality. Under this perspective, the (abstract) components of the theory (e.g. the notion of *electromagnetic field* in *Maxwell's Electromagnetism*, or the notion of *gene* in contemporary *genenetics*) don't exist in reality, yet are valid instruments to represent and interact with reality, enabling prediction, manipulation, and control.

[10] The *subsumption relation (IS-A)*, relates a subclass to its class, in the same way the inclusion relation relates a set to its super-set in set theory.

[11] In computer science, *meta-modeling* is the definition of building blocks and rules to build models.

[12] One of the fundamental aspects to be taken into account is: how are categories formulated? Refer to [26] for a systematic analysis of this topic.

meta-models are commonly termed *Foundational Ontology* or *Formal Upper Ontology*[13] (cf. [23]).
In the context of the NEUROWEB Project, the issue of meta-modeling was addressed by analyzing the expert classificatory criteria adopted by the neurovascular clinicians, and the relations between the expert knowledge specifically detained by the neurovascular communities and the general biomedicine knowledge.

Local and Global Domain Ontologies

Under an IT perspective, the fast proliferation of numerous *local ontologies*, developed for specific projects or applications, but incapable to work as general model for the domain of interest, is a major factor hindering data exchange, system inter-operability and resource integration. Specifically, the misalignment among these models is not on the syntactic/formal language level (cf. the success of the formal language OWL-DL [24]); terminological misalignment is a relevant factor, yet the major problem lies in the absence of a common meta-model and design methodology.
The solution to overcome this problem is not the radical elimination of local ontologies, and replacement with a one-fits-all model; indeed, some of the local ontologies actually capture valuable aspects of the knowledge schemes, specifically characterizing local communities and institutions. This is particularly true in the clinical field, where local communities of experts often develop specialized practices and approaches.
A more reasonable scenario entails, on a global level[14]:

- the establishment of a common ontology design methodology and meta-model;
- the establishment of a (global) *Reference Ontology*, composed by modular and orthogonal sub-ontologies, capturing the entities and relations pertaining to the general domain knowledge;

And, on the local level:

- the reception of as many elements as possible from the Reference Ontology, the meta-model and the design methodology, to the extent that they fit into the local needs;

Even in case of incomplete alignment, a partial mapping between the local ontologies and the Reference Ontology can be very probably established, supporting integration tasks.

Modeling Issues in the Clinical Domain

Clinical Indicators

The term *clinical indicator* is adopted to identify any clinically-relevant patient feature, computationally represented as a field of the health record; the term indicator is preferred to datum, as it does not already imply the computational representation.
A clinical indicator can refer to different aspects of the patient; such differences are important for the modeling activity. A non-exhaustive but representative list is:

- *personal identification* (e.g. name and surname, social insurance number);
- *demographics* (e.g. age, education, ethnic group);
- *life-style* (e.g. diet, smoking, alcohol consumption);
- elementary, common-use *biometrics* (e.g. weight, height);
- *vital parameters* (e.g. body temperature, pressure, pulse);
- *signs*, which are assessed by the medical doctor through a direct examination of the patient (e.g. pallor, swollen lymph nodes);

---

[13] An important example of Foundational Ontology for scientific research in the chemical, biological and medical domains is BFO (Basic Formal Ontology) [27].
[14] The OBO Foundry initiative [28] is moving along these lines.

- *symptoms*, which are self reported by the patients (e.g. headache, confusion);
- *lab tests*, which are usually performed on a *sample*[15], and which can be further grouped into genetic, biochemical, biomolecular, cellular, histological, etc… (e.g. hemoglobin level, blood leukocitary formula, blood group);
- *instrumental recordings* (e.g. electrocardiogram);
- *imaging* (e.g. MRI);
- *expert reports*, further elaborating instrumental recordings or imaging, usually formulated in natural language (for instance, the radiologist's report[16]);
- undergoing or past *treatment* (e.g. drug treatment: active ingredient and dose schedule; prostheses: ball and cage artificial heart valve);
- *family history* (e.g. father died of myocardial infarction);
- *diagnosis*, referring to an aspect or the whole clinical condition of the patients; these can be expressed by a short textual description, and/or by categorical values from disease classification systems (e.g. Diabetes Mellitus Type-2).

Clinical Indicators: Modeling Issues

Different aspects need to be handled when modeling clinical indicators and their values, which can be summarized as:

- *data type*: numerical (e.g. number of daily smoked cigarettes: 8), categorical (e.g. sex: male, female), serial code (e.g. social insurance number: 123-456-789), free text (e.g. "The absence of symptoms implies a benign lesion."), multimedia (e.g. a picture);
- *scale or metric* (e.g. height: meters; cognitive function: Rancho Level of Cognitive Functioning Scale);
- *time-dependence*, as the same indicator can be assessed multiple times;
- *normality range*: normality ranges are usually applied to determine whether an indicator has a clinically abnormal value (e.g. normal body temperature, measured in the armpit, is between 35 and 37 °C);
- *methodology and instrumental technology*; in this category we include the procedure followed by the medical expert to perform the clinical assessment (e.g. temperature measurement: rectal) and the information relative to the measure device itself (e.g. CT-scan technology: electron beam; CT-scan device manufacturer: Siemens); additional methodological issues arise when instrumental results are elaborated using some formal and reproducible method (e.g. an algorithm encoded in a software tool) or are interpreted by an expert; the latter case introduces an additional problem of subjectivity;
- *dependence on pre-conditions*: a specific issue with methodology is the existence of pre-conditions for the validity of the indicator determination (e.g. blood pressure higher than the reference threshold translates into hypertension only if it is not measured after psychological or physical stress);
- *degree of objectivity or subjectivity*; the latter can be interpreted in a negative sense, implying lack of standardization and methodological alignment, or in a positive sense, intending the contribution of the expert's implicit knowledge to problem solving;
- *granularity and inter-dependence of clinical conditions*: every indicator, with its specific value in the patient, can be regarded as a component of the global condition; indicators with a more global scope (e.g. diabetes mellitus type-2) usually rely on aggregation of narrower indicators (e.g. fasting glycemia), and

[15] A sample is a limited portion of the patient's body (e.g. a blood sample).

[16] E.g. "Appearances are consistent with a sessile osteochondroma of the proximal humeral diaphysis involving the medial cortex. The absence of symptoms implies a benign lesion."

the application of normality ranges (e.g. over-simplifying the case, influenza is diagnosed when temperature is T > 38°C, the patients reports a general sense of malaise, and he has runny nose and/or sore throat); the dependence on the context can be analytically broken down into relations to other indicators, but only with a varying degree of completeness; the harder cases occur when these relations are established through implicit cognitive heuristics of the medical expert;
- *directly observed or inferred*: indicators may reflect, to a different extent, the "pure" observation[17] of some physical reality or the inferential processes[18] guiding the diagnostic activity (relying on medical theory, guidelines and best practices, personal experience); indicators belonging to the personal identification, life-style and demographics typically have no interpretative content, whereas the final diagnosis has the maximal interpretative content; this issue is strictly connected to the *subjectivity/objectivity* issue, as the inference criteria may be explicit/objective to a variable extent;
- *granularity of physical structures*, stratified at different scales of the organic organization (e.g. molecule, intracellular or extracellular structure, cell or extracellular matrix, tissue, organ, system, individual, population).

Neurovascular Disorders

Ischemic Stroke

A *stroke* is the rapidly developing loss of brain functions due to a disturbance in the blood vessels supplying blood to the brain. This can be due to *ischemia* (lack of blood supply) or due to a *hemorrhage* (rupture of a blood vessel or an abnormal vascular structure). As a result, the affected area of the brain is unable to function, leading to cognitive and motor deficits. Persistent interruption of blood supply eventually leads to tissue necrosis and permanent brain lesions.
Stroke is a major health problem: it causes 9% of all deaths around the world, and it is the second most common cause of death after ischemic heart disease; in particular, considering the worldwide improvement of living conditions, and the aging of population in the developed countries, stroke may soon become the leading cause of death worldwide. About 80% of strokes are ischemic, defining a prominent position for Ischemic Stroke in the neurovascular field.
Ischemic Stroke can have different causes, implying
- different pathobiological mechanisms leading to the generation of the ischemia;
- different treatment strategies;
- differet prognosis;
- different prevention strategies.

As far as the etiology is concerned, *Atherosclerotic* and *Cardioembolic* Ischemic Strokes are both caused by an embolus obstructing a brain artery; in the case of the Atherosclerotic etiology, the obstructing body is a thromboembolus generated by a ruptured atherosclerotic plaque (i.e. ruptured *atheroma*) located in a large brain-afferent artery, either intra-cranial or extra-cranial; in the case of the Cardiomebolic etiology, the embolus is generated in the heart, and then carried to the brain by the bloodstream. The *Lacunar* Ischemic Strokes are characterized by small-artery occlusion, and they are caused by the presence of *microatheromas* in the small brain arteries, a condition

[17] No observation is truly *pure*: what is observed depends on a theory specifying what is relevant, and how it is observed influences the result. In addition, human contributions, such as instrument calibration, or discarding meaningless values, are often hidden behind apparently objective data.
[18] The *inferential process* mentioned here is intended as a human cognitive process.

termed *Small Vessel Disease*; therefore, in the lacunar subtype, the occlusion is not determined by an embolus generated in a different site.
[29]

The TOAST Classification System

The *TOAST Classification System* (Trial of Org 10172 in Acute Stroke Treatment) [30-32] originally introduced the etiology as the main criterion for stroke subtype classification, identifying *Atherosclerotic*, *Cardioembolic* and *Lacunar* as the main etiologies [30]. The TOAST was further improved by introducing classificatory categories to manage the degree of certainty in the etiology assessment, introducing for every etiology the subgroups *Evident*, *Probable* and *Possible* [32].
Although the TOAST Classification System includes a natural language description of the criteria for the diagnosis of the different subtypes, often including decision thresholds for quantitative clinical indicators[19], it is not a *formal* diagnostic system[20]. In other words, it is not a decision algorithm in the declension of computer science: it cannot be directly implemented in an automatic decision system.

Materials and Methods

The NEUROWEB Reference Ontology Development and Implementation

The Knowledge Acquisition Activity

The initial stages of the NEUROWEB Project were devoted to the definition of a data integration strategy and to the definition of typical use-cases. An effort was started by the clinicians to define a standard set of indicators (later termed Core Data-Set); meanwhile, the knowledge engineering group proposed the clinical phenotype[21] as a foundational concept, and hence the need for an ontological model. Once a consensus was reached for this important decision, a phenotype description template was sent to the clinicians to collect phenotype formulations according to the CDS (see *Supplementary Materials*, *Phenotype Description Form for Remote Knowledge Acquisition*). This initiative was only partial successful, though it helped to identify several improvements of the CDS, as well as the TOAST as a standard source of phenotype definitions. The initial ontology meta-model was defined according to the TOAST content and feedback from the clinicians at INNCB.
The knowledge acquisition activity was then performed through direct interviews with experienced clinicians, mainly at the AOK-OPNI and the UOP; the CDS and TOAST categories were explicitly used as reference concepts. The ontology meta-model was improved, reaching a state very similar to the final one. At that stage, MI-EMC contributed other phenotypes using the phenotype description template. The prototypal phenotype formulations and the ontology meta-model were thoroughly discussed with clinicians at plenary project meetings and through email exchanges.
The NW-RO was finally completed adding the biomolecular entities, and the terminological extension.

---

[19] E.g. *stenosis degree* ≥ *50%* for *Atherosclerotic Evident*. *Stenosis* is the restriction of a blood vessel caused by the presence of an atheroma on the vessel surface.
[20] Of course, this does not imply that the TOAST is not scientifically rigorous or irreproducible or inconsistent.
[21] For simplicity, clinical *phenotype*, *state* and *condition* will be treated as synonyms.

The OWL-DL Implementation of the NEUROWEB Reference Ontology

Description Logics (DL) [33-35] are a family of logic-based knowledge representation formalisms designed to represent and reason about the knowledge of an application domain in a structured and well-understood way. The basic notions in description logics are *atomic concepts* and *atomic roles* (*unary* and *binary predicates* in the terminology of *first order language*, respectively). In order to distinguish the function of each concept in the *relation* (represented by a role), the individual object that corresponds to the second argument of the role is called *role filler*. For instance, *hasPart.Wheel* is an expression describing the relation of having wheels as parts; the individual objects belonging to the concept *Wheel* are fillers of the role *hasPart*.
A specific description logic is mainly characterized by the constructors it provides to form complex concepts and roles from the atomic ones. The language used to formalize the ontological clinical knowledge is *SHOIN* [35], which is an extension of the basic description logic. In order to develop the Reference Ontology computational model we have adopted the *OWL DL* version [36]. The editor adopted for the OWL files generation is *Protégé* [37], where the Reference Ontology concepts are represented as T-Box (Terminological Box) entities.

The NEUROWEB System

The NEUROWEB Query System

The NEUROWEB phenotypes are defined in the NEUROWEB Reference Ontology (NW-RO) as *DL axioms*; in this formulation, phenotypes cannot be directly exploited to query the local repositories, since the latter are typically implemented as *relational databases*. To retrieve patient clusters on a phenotypic basis, the axioms need to be translated into the local database queries, exploiting the mapping between the NW-RO and the local database content.
Two software components were developed to perform this task: the *Phenotype Converter* and the *NEUROWEB to Local (N2L) Mapper*. The Phenotype Converter translates a SQL query composed by NW-RO entities into a query composed only by NW-RO entities mapped to local databases (mostly CDS Indicators), which is then dispatched to the N2L Mapper. The N2L Mapper exploits the mapping to generate the actual queries on the local database fields. The architecture of the query system is displayed in the figure M.1.
The Phenotype Converter was implemented in Java, exploiting the Jena programming interface to navigate the ontology and extract the class names and axioms. The conversion process essentially navigates the references from the Top to the Low Phenotypes axioms, and finally to conditions on CDS elements, which are the leaf nodes of the phenotype trees. For this navigation, the has-cause, has-evidence and by-means-of relations are used. By combining the simple conditions with the operators and quantifiers of the axiom, the top level phenotype can be represented as a *nested AND/OR expression*.

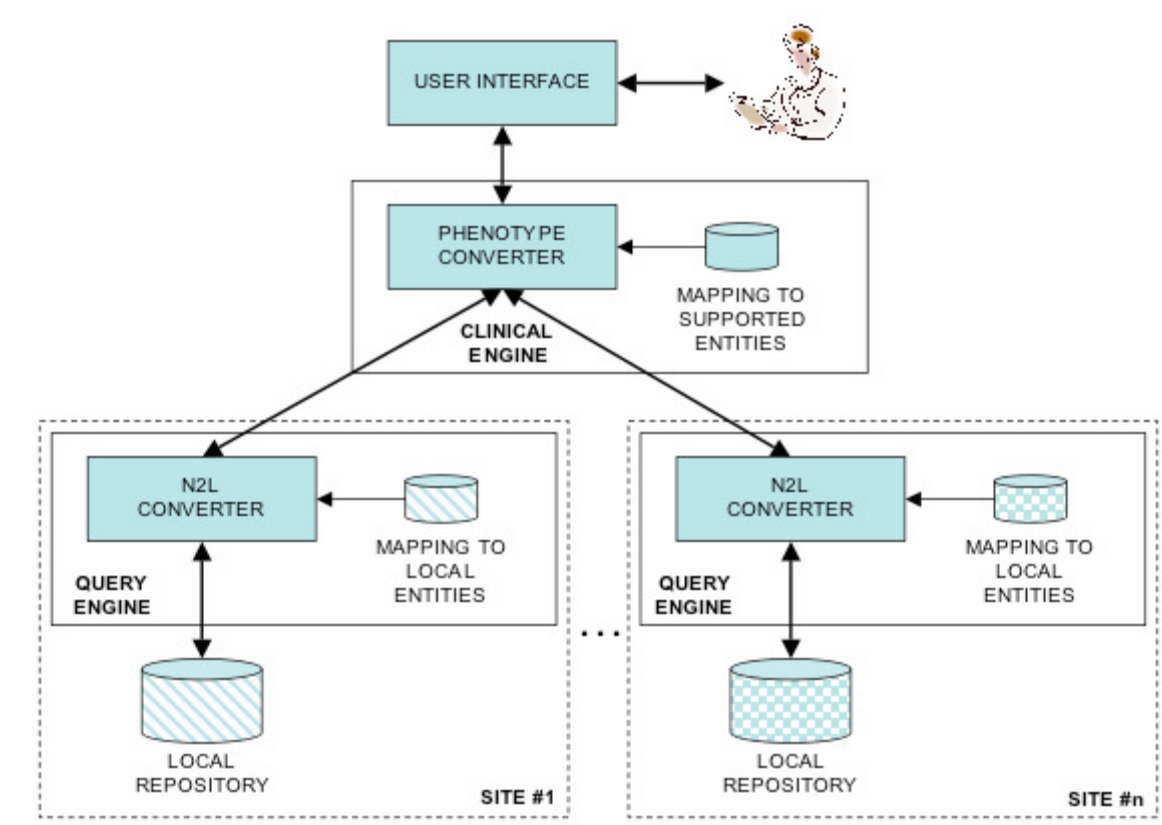


Figure M.1 - The figure displays the information flow of a clinical query in the NEUROWEB system. The software modules are displayed as solid boxes, whereas the data items are displayed as cylinders. The dashed boxes group the components of a local site (only two sites were depicted for compact-ness reasons). Thick arrows identify the information flow elicited by each user's query.

The NEUROWEB to Local Mapping

Local medical experts, which are familiar with the actual meaning, coding, etc. of the local repository logical structure, were requested to create the mapping between the elements in the local repository and the ones in the Reference Ontology and the CDS. In order to facilitate this task, a graphical interactive tool, named *Ontology Mapper*, was developed as a Protégé plugin.
The Ontology Mapper enables to map a local database field to a CDS Indicator or Low Phenotype. In addition, it is also possible to transform the values of the local field (e.g. algebraic transformations of a quantitative indicator, conversion of the categorical scale). The transformation can be specified by the user as pseudocode.
For instance, the algebraic transformation is handled as follows:

```
Case
when Local.Field.X is null then NA
else Local.Field.X * 0.88 + 4.5
end as "CDS.Indicator.X"
```

Whereas the conversion of the categorical scale is handled as follows:

```
Case
when Local.Field.X is null then else Category.B0
when Local.Field.X is Category.A1 then Category.B1
when Local.Field.X is Category.A2 then Category.B2
else Category.B0
end as "CDS.Indicator.X"
```

The NEUROWEB Database and Web Service Technology

The NEUROWEB system exploits the Web Service technology to decouple the central components from the local sites and facilitate the addition of new clinical partners. Each partner is free to choose the preferred technology to implement the local components. However, to deliver a low-cost and effective sample solution, the central technological committee has developed and made available a reference implementation, which exploits the popular open-source technologies *Glassfish* (for the communication tasks) and *Postgres* (for the database view). Glassfish safely manages call wrappings to expose the local interface as a *WSDL* document. The *Clinical Query Application* runs as a web application on the Glassfish server, processes the incoming web service call, translates it into a SQL query, runs the query on the database view, wraps it and returns the resulting record set.

Results

Introduction: Modeling Clinical Conditions to Support Association Studies

Genotype-phenotype association studies essentially require to:

1. define a phenotype of interest;
2. select the individuals characterized by the phenotype of interest;
3. test pre-selected genotypes for association, or perform an unbiased, genome-wide search for associated genotypes.

In the NEUROWEB Project, the phenotype of interest is a pathological neurovascular condition, and the phenotype must be formulated on the basis of the clinical indicators stored in the local repositories (hence the term *clinical* phenotype). In the most simplistic scenario, the researcher defines a database query, specifying what should be the values for an array of clinical indicators, and then lets the system retrieve the corresponding patients from the local databases. What are the inherent problems?

a. The local databases were developed independently; as a consequence the fields corresponding to the same clinical indicator most probably have different names.
b. In case a certain clinical indicator is represented in all the databases, it may have values according to different metric units or qualitative scales.
c. In case a certain clinical indicator is represented in all the databases, it may have been assessed using a different methodology or instrumentation, without the warranty of an equivalent reliability in the assessment.
d. Since the local sites adopt specific diagnostic procedures, a certain clinical condition may be defined at different granularity levels in different databases.
e. Clinical Indicators providing a global characterization of the patient's condition (such as the stroke diagnosis according to the TOAST classification) may have been defined according to diverging criteria, and therefore cannot be reliably integrated.

A framework is required to address all these problems. The solution adopted was to design a model (*the NEUROWEB Reference Ontology, NW-RO*) for clinical conditions. The clinical phenotypes defined in the NW-RO are grounded on the content of local repositories, and are constituted by (a) built-in phenotypes of most common use, (b) deconstructed into more elementary units; these units can be exploited by the user in a compositional fashion, in order to modify the formulation of the built-in phenotypes or to assemble new ones. As a whole, the integration problem is addressed by reconciling the content of the local databases into a central semantic model, the NW-RO.

With respect to the NW-RO, the notion of clinical phenotype is equivalent to clinical condition: it is the representation of the patient's state according to clinical indicators, which reflect the examinations performed during the diagnostic activity and the final diagnostic judgment expressed by an expert clinician.

The definition of the NW-RO meta-model was achieved through the interaction with the neurovascular experts from the different partner sites, and required an intensive effort of knowledge acquisition, analysis and elaboration. The next sub-section is devoted to the description of the results of such activities.

Knowledge Analysis

The Strategic Role of Knowledge Analysis

Analyzing knowledge models and knowledge-based social organizations is a hard and time-consuming task. Under an engineering perspective, undertaking knowledge analysis must generate advantages for the design of the ontological model; otherwise it is a theoretically intriguing but practically useless effort. Several elements suggest that the knowledge analysis task is critical for the NEUROWEB Project:

- the identification and classification of clinical conditions is not a problem with a formally set and universally accepted solution[22]; on the contrary, the topic is object of substantial research and discussions among experts; it is natural for different clinical communities to devise and adopt specialized solutions, and then to engage into experience-sharing and benchmarking efforts to identify and disseminate the best-practices;
- the retrieval of patients conforming to a certain clinical condition requires the highest level of methodological reliability, otherwise the results of association studies are totally devoid of scientific value[23]; consequently, the model of clinical conditions must be formal, expressive and scientifically accurate;
- a model of clinical conditions incapable to reflect the clinicians' classificatory criteria would be easily rejected by its users and hence fail its goal;
- the modeling challenges posed to ontology design by translational research often depend on the different cognitive models characterizing the experts in clinical medicine and biomolecular research; these models are implicitly reflected by the data organization and content; a system supporting data exchange and their integrative analysis must be grounded on a clear understanding of the differences and the common ground between the two disciplines.

Disciplines

To describe the knowledge landscape in the NEUROWEB Project we need a few foundational concepts. The notion of *scientific discipline*, though apparently intuitive, was not formalized by prominent Philosophy of Science scholars, such as [19, 38-40]. Here, a discipline is characterized by:

a. the domain of reality being the object of study (e.g. neurovascular pathologies);
b. the aims (e.g. preserve the patient's health);
c. the methodology (e.g. how to perform medical examinations, how to interpret diagnostic results, how to generate diagnostic hypotheses from initial evidences and how to refine them until the final diagnosis is formulated, how to determine the treatment);
d. the content theory (e.g. histology, anatomy, physiology, disease classifications, etc...).

[22] A typical counter-example is the classification of chemical elements by the periodic table, which is universally accepted by chemists, and it is not object of discussion or further research.

[23] A typical counter-example is a query system for scientific papers: the system retrieves potentially interesting papers, but the final solution concerning whether the paper is relevant or not to the topic of interest is determined by the user. The presence of partially-unrelated papers requires a greater effort on the user's side, but does not undermine the validity of the result.

These criteria can be effectively used to understand the different aggregations of scientific communities proposed by different philosophers of science [19, 38-40].
Neurovascular Medicine can be suitably conceived as a discipline.

Neurovascular Communities and the TOAST Classification

The NEUROWEB clinical communities can be conveniently regarded as *Research Communities*, which are a form of social group standing in between the traditional *Scientific Community* (as it is conceived in the philosophy of science) and the *Community of Practice* (as it is conceived in knowledge engineering [41-43]). Whereas a scientific community adheres to a hypothetical-deductive paradigm of scientific thinking, heavily relying on formal constructs (e.g., in physics, Newton's Law of Mechanics; in Chemistry, the Periodic Table), a community of practice is characterized by some form of *implicit knowledge* [44], which may be encoded, to a variable extent, as a semi-formal *Representational Artifact*.
NEUROWEB clinical sites are communities of neurovascular experts, whose daily work is the examination of patients; in each center, the patients have been previously diagnosed stroke within an ordinary hospital, and they are admitted to the center in order to undergo a more detailed diagnostic process, as well as to receive the most suitable treatment. Each clinical community encodes its own procedural rules into written documents, termed protocols. The local protocols have to comply with best practices, often encoded by written documents termed guidelines; best practices are determined through interaction at international congresses and scientific publications.
Therefore, protocols capture elements at two different levels: the minimal requirements prescribed by the guidelines, and characteristic elements of the local communities' customs and expertise. These elements can be clearly identified in the NEUROWEB clinical communities:

- the examination of patients involves collective forms of working (e.g. advises from elder colleagues, collective discussions over anomalous cases);
- the examination of patients, followed by the diagnosis formulation, and the treatment decision, is a problem-solving activity; each problem-solving case undergoes reification into a clinical record;
- the protocol can be regarded as a procedural manual, and it is the final product of the negotiation phase.

As remarked in [39] the knowledge cycle of a Research Community typically originates a domain conceptualization (a domain vision), usually implicit and not formalized [45-46].
The domain conceptualization of a community may remain implicitly encoded in its routines and written documents, or it may undergo a process of structuring into a Representational Artifact; examples of Representational Artifacts are, on the side of the content theories, classifications of pathologies, and on the side of mechanism theories, diagnostic algorithms, i.e. decision trees in which the if-then-else clauses are formulated in Natural Language (NL) [47-49].
The capital difference between a narration, or a manual, and a Representational Artifact is the presence of an explicit structure; the structuring process consists in transferring the knowledge content from the NL-encoding to the structure of the artifact. Though the Representational Artifact is not directly encoded in a formal language, its construction is the necessary condition for knowledge formalization. Representational Artifacts always have residual semantic elements, encoded as NL, and their structure may be partial or unsystematic [49]. Therefore, it is not practicable to treat them as the sole knowledge asset. That is in agreement with consolidated KE literature, stating that expert knowledge is primarily detained by subjects rather than artifacts or written documents [44-45], and consequently postulating the necessity of knowledge acquisition (KA) from the experts preliminary to the knowledge representation (KR) phase [51].

We argue that Representational Artifacts are valuable assets to guide the KA activity, but, as knowledge sources, they cannot replace the experts' knowledge. In order to support association studies, we are specifically interested in Representational Artifacts able to guide the clustering of patients, in other words, classificatory systems for clinical phenotypes.
In the context of the NEUROWEB consortium, the structuring process does not occur within the local communities, but rather as a negotiation among different research communities. The TOAST classificatory system is the Representational Artifact fulfilling at best the NEUROWEB needs. It was developed by the international neurovascular community to classify stroke patients on a diagnostic basis, and it was already adopted by most of the clinical sites before the existence of the NEUROWEB Consortium. The adoption of the same Representational Artifact (the TOAST) testifies for the presence of a common conceptual and methodological ground across the different NEUROWEB Research Communities. As a matter of fact, the TOAST assumes specific diagnostic criteria grounded not only on a common conceptualization of clinical phenotypes (content theory component), but also on shared methodological principles (mechanism theory component). So far, two major modeling issues have to be managed:

- the TOAST is a classificatory system with a strong NL-component; to be represented with a formal language, it requires a further structuring effort (i.e. refinement of the conceptual model); that activity cannot be carried out without involving the clinical experts from the Research Communities;
- managing the difference among the Research Communities, which requires to conjugate two different objectives:
    - federating their clinical repositories, and coherently formulating the TOAST clinical phenotypes on that basis;
    - besides the TOAST phenotypes, allowing also the formulation of customized clinical phenotypes, in order to convey the Research Communities specificities.

From Neurovascular Medicine to the Biomolecular Domain: What Challenges

To support genotype-phenotype association studies it is valuable to connect clinical phenotypes to genomic entities; the researchers in the NEUROWEB Consortium are specifically interested in:

- comparing the genotype-phenotype associations generated within the NEUROWEB System to the ones stored in publicly accessible databases;
- identifying possible relations between the function of a gene involved in a genotype-phenotype relation, and the nature of the associated phenotype (e.g. what's the relation between a LDL Receptor and Atherosclerotic Ischemic Stroke?).

We argue that such an operation requires to bridge two knowledge domains belonging to different disciplines. In this case, the two disciplines are clinical medicine, specifically vascular neurology, and molecular biosciences (the convergence of molecular biology, molecular genetics and genomics). The divergence between the two disciplines concerns multiple epistemological dimensions:

- The object of study is the same, the human body (though biosciences are not restricted to the study of humans, but also other life forms as well). However, the focus is on different facets, occupying different *granularity levels*: medicine mainly addresses the level of organs, anatomical parts, and global body parameters, though it takes into account also the tissue, cell and molecule levels; molecular biosciences addresses the molecular level, with a particular focus on the genetic information encoded by DNA, and the processes centered on it. Therefore, clinical medicine and molecular biosciences have different reference levels of granularity.

- The two disciplines have different aims: clinical medicine aims at determining and treating the pathological states of patients – it is concerned with the patient's health state; molecular biosciences aim at elucidating the molecular mechanisms of biological organisms.
- The methodology is different as well (e.g. different experimental techniques), as a consequence of the different facets in the object of study, and the different aims.
- Finally, also the content theory (the catalog of entities, their properties and their relations) is different. This difference is influenced by the object of study and the methodology. The genomic information of interest for NEUROWEB (gene functions and phenotypes) is typically stored within general-purpose repositories, which are not the expression of a research community with a specific methodology. For instance, HGMD [52] is a repository for genotype-phenotype associations, which includes phenotypes referring to pathologies, thus underpinning a clinical perspective, altogether with phenotypes referring to the alteration of specific molecular functions and processes, thus underpinning a molecular bioscience perspective; the former, however, are not treated with the level of detail adopted in the NEUROWEB project, and the methodological problem of how the pathology is diagnosed is not addressed. Beyond methodology, other problems arise in relation to the granularity levels of the object of study.

The strategy adopted to address this problem is exactly the one adopted to reconcile different classification systems: identify suitable building blocks of the clinical phenotypes, which can relate to biomolecular entities. In the next section we will discuss how such building blocks were identified, and how they fit in into the ontology meta-model.

From Knowledge Structuring to the Ontology Meta-Model

Step 1: The Two-Layer Model

How to reconcile the methodological specificities of the local neurovascular communities into the common framework of the TOAST Classification System?

In the beginning, we devised a two-layer model. The lower layer, later termed *Core Data Set (CDS)* was constituted by selected clinical indicators, which were identified by the clinicians on the basis of their importance for stroke diagnosis. Under an IT perspective, the CDS, coupled to a mapping between CDS indicators and the content of local databases, would be enough for a query system to operate on the four clinical repositories. In that setting, the task of phenotype formulation would be performed by the user assembling different CDS indicators into a query formula. Since the clinical communities accept the TOAST as a standard, it is natural to add a further layer, constituted by preassembled CDS queries that represent standard phenotypes. In this two-layer model, a phenotype is formulated as a conjunction/disjunction of criteria on the CDS indicators, expressed as equalities, inequalities, or quantitative ranges to be satisfied[24]. This layer is present in the final model as well, and is termed *Top Phenotypes*.

Let's briefly discuss the merits and limitations of this approach. The existence of a negotiated, but centrally encoded, phenotype formulation establishes a methodologically-coherent federation of the different communities' resources. Local practices are reconciled within this model, as different sets of CDS indicators can be used to identify a common phenotype. The easiest case occurs when the same clinical evidence can be assessed using different technologies but achieving the same degree of probatory strength (e.g. Severe Stenosis in the Internal Carotid Artery can be determined by a Duplex AND a Computed Tomography Angiography Scan, OR by a Duplex AND a Magnetic

[24] E.g. Severe Stenosis = (CTA Degree of Stenosis > 60%) OR (MRI Degree of Stenosis > 60%).

Resonance Angiogram Scan). The way phenotypes are encoded strictly resembles the typical formulation of a query, and thus can be easily handled by a clinical expert; for that reason, the two-layer model was successfully used for the Knowledge Acquisition activity.

Step 2: Identifying the Building Blocks of Clinical Phenotypes – the Three-Layer Model

The two-layer model is not enough to support NEUROWEB System functionalities for the following reasons:

- The CDS indicators do not represent general concepts of the medical domain; CDS indicators should be rather regarded as the minimal elements of the diagnostic practice, as it is exerted in the NEUROWEB clinical communities; for example, anatomical parts are often referred in the CDS indicators, but they are not represented as stand-alone entities of the CDS. For the same reason, CDS indicators are not suitable units to bridge the clinical practice and biomolecular research.
- In several cases, the actual indicators in the local databases do not perfectly match the CDS indicators, because of granularity discrepancies[25]; for example, one database reports the stenosis degree (as a number) and the stenosis side, whereas another database only reports the presence/absence of a severe stenosis. A hierarchy of phenotypes at different granularity levels is required to handle this problem, and it is not provided by CDS indicators.

To overcome these problems, we decided to introduce an additional layer. We deconstructed the TOAST phenotypes into more elementary components (i.e. building blocks) holding general validity in the medical domain. To convey the fact that these are more elementary phenotypes, the layer was termed *Low Phenotypes*.

To define the entities and relations constituting the *Low Phenotypes* layer, we analyzed what are the TOAST classification criteria:

1. *etiology* (Atherosclerotic, Cardioembolic, Lacunar Stroke);
2. *confidence* of the etiological assessment (Evident, Probable, Possible), depending on the strength of the diagnostic evidence for the most-probable etiology;

In addition, *anatomy* (i.e. the location of the lesion) is not used by the TOAST, but could be suitably used to extend it, and is consistently used in the different clinical communities.

These criteria are explicitly recognized by the clinicians, and are generally used in clinical medicine.

Another major point to consider is that the diagnostic activity is always characterized by the acquisition of diagnostic evidences, enabling the reconstruction of the undergoing patho-physiological processes and structures, even if these *are not directly observed*. For instance, the ischemic stroke is caused by the lesion of an atherosclerotic plaque, triggering the coagulation cascade, the release of a clot particle into the bloodstream (embolization), and the obstruction of a brain artery; however, the state of the plaque is not required by the current diagnostic guidelines; its capability to cause embolization is rather inferred from the presence of the ischemic lesion, the presence of the atherosclerotic plaque in a clinically relevant afferent artery, and the absence of alternative explanations. However, from a biological standpoint, it is important to represent not only the diagnostic evidences, but also the inferred state of the biological parts and processes. In the specific case of ischemic stroke, there is a consistent partition between the evidences for the ischemic damage (typically brain imaging displaying the damaged tissue, and the cognitive / motor impairment of the patient) and the evidences for the cause of the occlusion causing ischemia; the latter is usually a persis-

[25] Specifically, *clinical state granularity*, refer to the description in the Introduction.

tent or progressive state of the patient's organism (i.e. with a time-span usually in order of years or decades), whereas the latter occurs after a chain of point-events (i.e. with a very compact time-span) leading to trauma.
To represent all these aspects, we structured the ontology meta-model into five large groups of entities:

- *Top Phenotypes* (Ischemic Stroke types, according to the TOAST or user-defined, representing a neurovascular clinical state);
- *Low Phenotypes* (building blocks of neurovascular clinical states);
- *Topo-Anatomical Entities*, comprising Anatomical Parts and Topological Concepts (building blocks, necessary but not sufficient to represent a clinical state);
- *CDS Indicators* and *Diagnostic Values* (defining what values of the clinical indicators determine the presence of a certain phenotype);
- *Biomolecular Entities*, currently comprising only *Biomolecular Processes* and their *Participants* (typical entities of the biomolecular world).

In particular, the Low Phenotypes are divided into:

- *Durative Etiological Background* (e.g. *Atherosclerosis*);
- *Traumatic Point-Event* (typically *Ischemic Traumatic Event*);
- *Durative Diagnostic Evidence* (e.g. *Stenosis*);
- *Point-event Diagnostic Evidence* (e.g. *Relevant Lesion*).

A Top Phenotype is primarily decomposed into a Durative Etiological Background – with its diagnostic evidence – and a Traumatic Point-Event through the *Has-Cause-Durative* and *Has-Cause-PointEvent* relations respectively. The reason for grouping together Durative Etiological Background and its diagnostic evidence as targets of the Has-Cause-Durative relation is that the nature of the diagnostic evidence influences the etiological assessment confidence characterizing the Top Phenotype (Evident, Possible, Probable). The Durative Etiological Background and the Traumatic Point-Event are connected to their diagnostic evidences via the *Has-Diagnostic-Evidence* relation. A Diagnostic Evidence can be broken down into Diagnostic Evidences at a smaller granularity level via the same *Has-Diagnostic-Evidence* relation.
The reason for introducing the *Has-Diagnostic-Evidence* relation is that the Durative Etiological Background and the Traumatic Point-Event are never observed per se, but are rather inferred from observables evidences, interpreted according to some general *diagnostic theory*. However, it is important to go beyond the mere diagnostic evidences, identifying phenotypes that can fully correlate to biomolecular entities. For this reason, the Durative Etiological Background and the Traumatic Point-Event, but not their diagnostic evidences, are connected to the Biomolecular Processes via the *Involves* relation. When the definition of a Low Phenotype implies the reference to an Anatomical Part or a Topological Concept, this relation is represented by *Has-Location* and *Has-Site* respectively.
Via the *By-Means-Of* relation, the Diagnostic Evidences are mapped to the *CDS Indicator* and its value range (termed *Diagnostic Value*, connected via the *Has-Value* relation), which are required for the assessment of the phenotype presence.

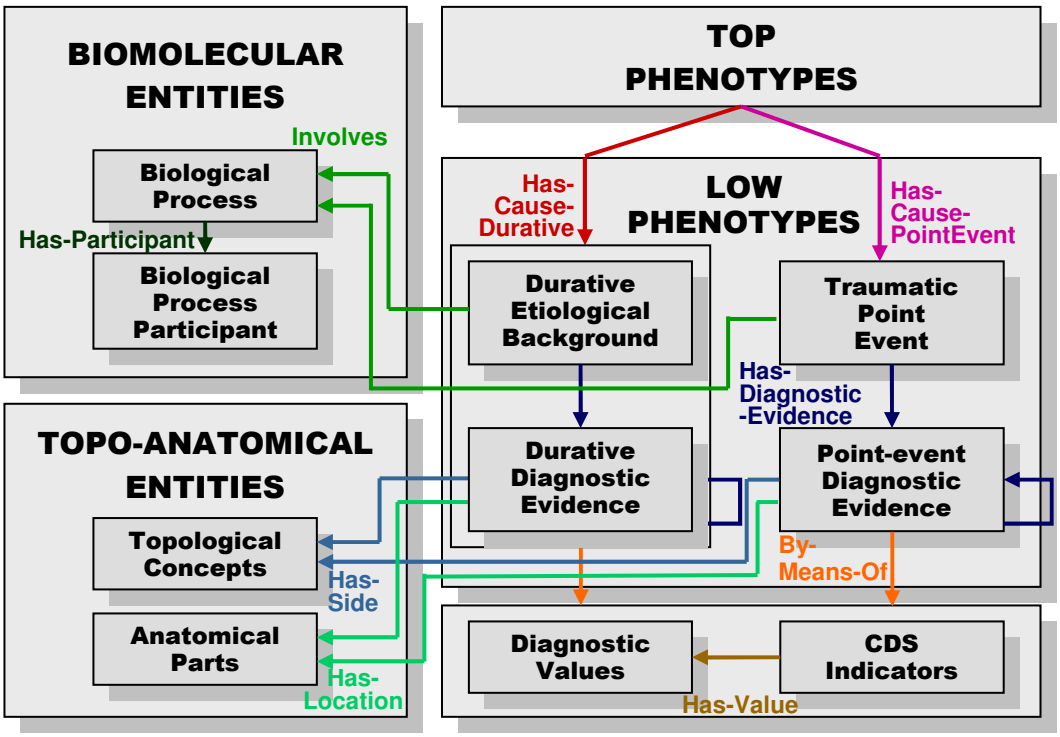


Figure R.1. Has-Diagnostic-Evidence is not a reflexive relation. The loop on Diagnostic Evidences should be interpreted as the capability to connect a more specific to a more general diagnostic evidence.

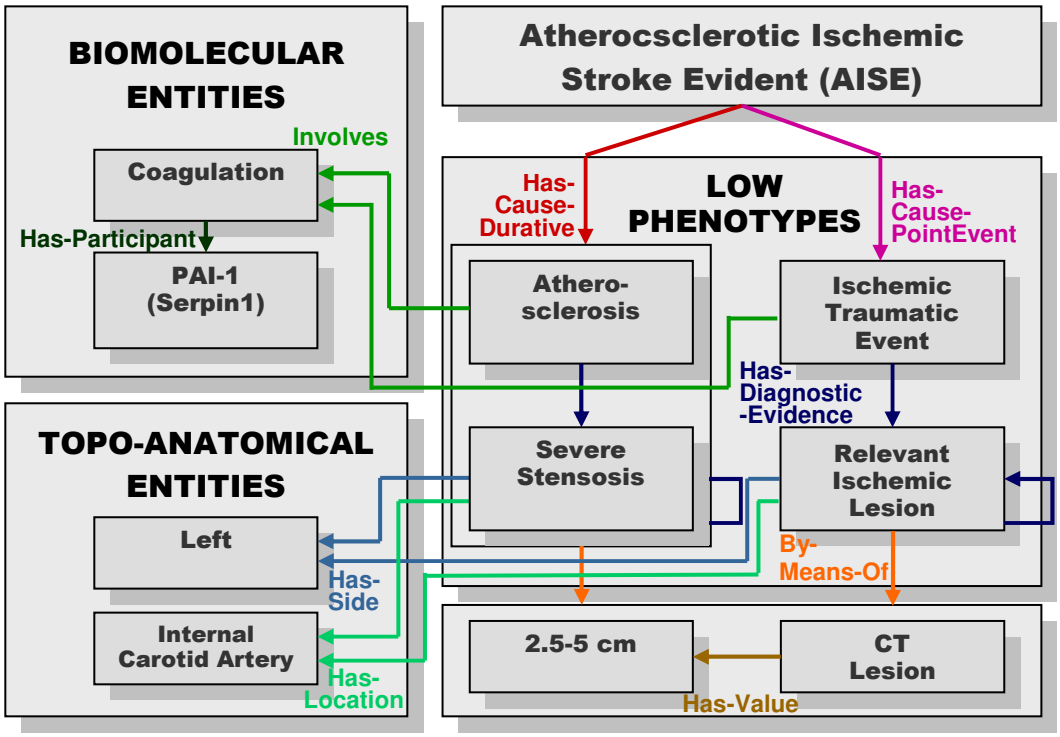


Figure R.2. The meta-model instantiated with specific classes.

The entities from the Top Phenotypes, Low Phenotypes and Topo-Anatomical Entities can be mapped to entities and terms from other biomedical ontologies and terminologies. Biomolecular Processes are mapped to Gene Ontology [53], and Biomolecular Process Participants to the Entrez-Gene NCBI Database [54][26].
The meta-model is displayed in figure R.1. Figure R.2 displays the same meta-model, only with specific entities replacing the main classes.

Step 3: The Logic Formalization

The NEUROWEB Reference Ontology is implemented using the formal language *Description Logics (DL)*. A phenotype is typically formulated as a logic axiom, capturing the structure of entities and relations. This aspect will be addressed with more details in the next section, *A Closer Look at the NEUROWEB Ontology*, exploring the representation of a typical NEUROWEB clinical phenotype. A brief description of Description Logics as a knowledge representation formalism can be found in the *Material and Methods / The OWL-DL Implementation of the NW-RO* section.

Step 4: The Mapping from CDS Indicators to Local Databases

The NEUROWEB Reference Ontology needs to be mapped to local databases; otherwise no patient retrieval functionality can be supported. In most of the cases, the CDS Indicators can be mapped to the local database fields, and the mapping equals to a linguistic translation into the local terminology. In a few cases, however, the granularity of the clinical indicators represented by the local database fields is too large for the CDS mapping. In these cases it is still possible to map the local database fields to the NW-RO, but only at the expense of by-passing the CDS and part of the lower-granularity Low Phenotypes. On the one hand, this solution offers a higher coverage; however, on the other hand, it exposes to the risk of only partially granting methodological coherence in phenotype-based searches. Indeed, the methodological coherence depends on the application of the same phenotype formulations, as specified by the NW-RO; incomplete definitions, stopping at a certain granularity level, offer a limited warranty of reproducibility and homogeneity.
For this reason, a decision was taken to establish two different mapping & query modes:

- *standard mode*: only the mapping to the CDS is accepted as valid; if a field in a local database cannot be mapped to the CDS, the corresponding indicator is assumed to be missing; as a consequence, the patients from that repository may not match certain phenotypes; the methodological coherence is maximally granted, but false negatives may be generated in a large number;
- *flexible mode*: the mapping to the CDS and to all of the Diagnostic Evidence Low Phenotypes is accepted as valid; the methodological coherence is only partially granted, but false negatives are minimized.

For an example of how the flexible mapping works, please refer to the section *A Closer Look at the NEUROWEB Reference Ontology / The Flexible Mapping in Action: the Carotid Stenosis*.
A future scenario entails the use of local ontologies for the mapping onto the NW-RO; this solution enables to express a more sophisticated relationship and exploit the full power of the NEUROWEB framework.

What Role for the Existing Medical Ontologies

A major decision to be taken was to choose between the development of a specific Reference Ontology and the adoption of an existing one [55]. The second solution

[26] Please refer to *A Closer Look at the NEUROWEB Reference Ontology / The Terminological Base*.

would apparently offer superior advantages, by granting interoperability with external resources, and facilitating the involvement of new partners, in case they are already complying with an existing ontology. However, there are crucial problems undermining this solution. First, the phenotype ontologies developed within the biological community are oriented to high-throughput genetic experiments in model organisms [12, 57-58], and hence are not suitable for clinical applications. Second, no publicly-available medical ontology is committed to the representation of clinical finding and clinical conditions or phenotypes [56]. From a pragmatic perspective, one may argue that, even if an optimal solution is not available, the re-use of an existing general-purpose medical ontology may still offer significant advantages. As a case in point, we did consider SNOMED-CT [59-61] as a possible candidate, but decided against it because of several semantic shortcomings that would be encountered by adopting it[27] [28].

- Considering the NEUROWEB stroke type taxonomy (Top Phenotypes), most of the concepts are either missing from SNOMED-CT, or they are formulated in an unsuitable way. For instance, the definition of SNOMED-CT *Atherosclerotic Occlusive Disease* clearly implies an atherosclerotic etiology, but not the specific features of stroke, which are part of the NEUROWEB Ischemic Stroke definition.
- SNOMED-CT offers qualitative scales for clinical findings but does not provide quantitative criteria to assign them (e.g. no *stenosis percentage ranges* are associated to the previously mentioned scale), nor it resolves inter-dependencies among different indicators. In addition, several CDS Indicators do not have a corresponding term in SNOMED-CT (e.g. *relevant scan lesion*, where relevance is determined by co-axiality of stenosis and ischemic lesion).

We also considered the *Disease Ontology (DisO)* [64], a general-purpose classification of pathologies. DisO was developed to annotate biological-samples within genetic databanks (in the context of the *NUgene Project*). DisO includes concepts which have terminological correspondence in the NEUROWEB Reference Ontology (e.g. *Stroke*, *Atherosclerosis*, *Subarachnoid hemorrhage*, *Cerebral embolism*, *Cerebral thrombosis*, *Occlusion and stenosis of carotid artery*, etc…); however, there are major problems preventing its adoption; just to make a few, representative examples:

1. the DisO disease taxonomy does not follow the taxonomy adopted by the NEUROWEB clinican communities;
2. no criteria are provided to assign disease classes on the basis of clinical data;
3. the concepts are organized by adopting only is-a, part-of, inverse-of, union-of and disjoint-from relations, lacking any specification of causality, which is a fundamental criterion for stroke categorization in the NEUROWEB Project; the failure to depict the relation between ischemic stroke and its etiological background hampers also the genotype-phenotype associations studies (e.g. genotypes associated to atherosclerosis would not necessarily relate to ischemic stroke).

Although general-purpose medical ontologies cannot replace the NW-RO, the NW-RO is connected to the best-matching terms from such ontologies. This terminological extension constitutes a valuable resource for keyword-based searches in external resources. For more details on the topic, please refer to *A Closer Look at the NEUROWEB Reference Ontology / The Terminological Base*.

[27] Concerning the shortcomings related to the formal structure of SNOMED-CT, see [62].

[28] For an assessment of SNOMED-CT applicability to stroke-related clinical data and diagnostic categories see [63].

A Closer Look at the NEUROWEB Reference Ontology

The NW-RO Overall Architecture

The NW-RO is composed by three layers. The bottom layer is the *Core Data Set (CDS)*, a set of clinical indicators which are mandatory for the diagnosis, and are mapped to the content of local repositories. The top layer, the *Top Phenotypes*, is a neurovascular disease taxonomy, representing the classificatory categories used by neurovascular experts. The middle layer, the *Low Phenotypes*, enables the deconstruction of Top Phenotypes into more elementary units. The Low Phenotypes are connected to the CDS values required for their occurrence, thus closing the loop. Connections are present only among adjacent layers; therefore no connections exist between the CDS and the Top Phenotypes. The overall architecture is displayed in figure R.3.

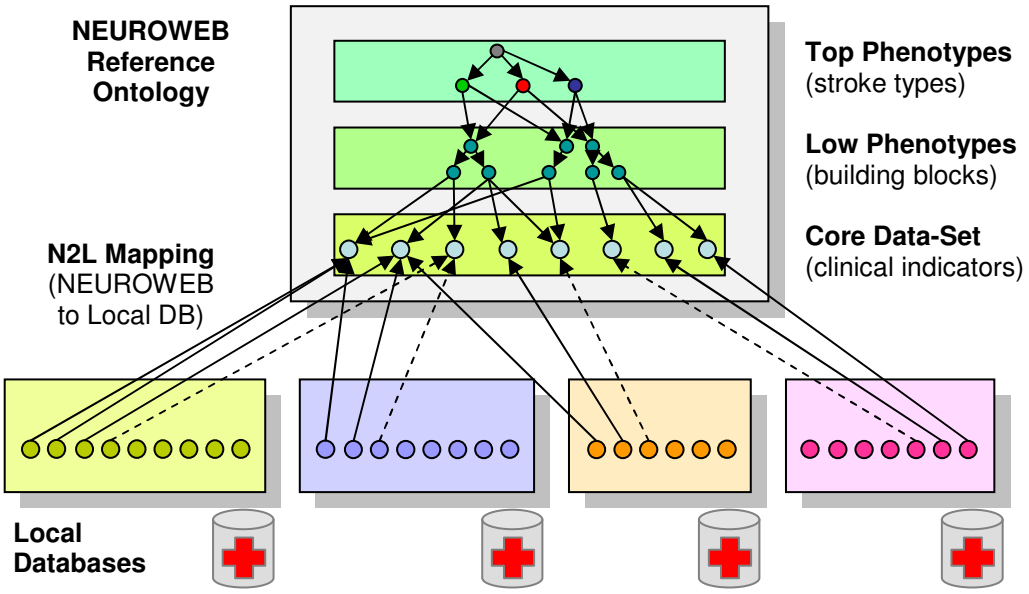


Figure R.3. The Architecture of the NEUROWEB Reference Ontology.

The NW-RO in Action: the Atherosclerotic Ischemic Stroke Evident

Atherosclerotic Ischemic Stroke Evident (AISE) is a Top Phenotype. This sub-section is devoted to displaying how it is encoded by Description Logics (DL) formula (figure R.4), and how this representation can be mapped to a graphical diagram (figure R.5). The diagram notation is composed by phenotype-boxes (equivalent to tokens), transition-rectangles (transitions) and arcs connecting only phenotype-boxes to rectangles, and vice-versa. If a phenotype-box A is in-connected to a single transition T1, all the phenotype-boxes out-connected to transition T1 are necessary for A to occur; if multiple transitions (T1, T2, etc...) are in-connected to A, at least one is necessary for A to occur. Within a phenotype box, a dashed line represents a relation between two entities; if only a component of the box occurs, but the relation is not present, the whole phenotype in the box does not occur. The arcs out-going a phenotype box have an attached relation name. The overall diagram depicts the inter-dependencies among phenotypes, a specified by the DL formula.

$$(1)\ \mathtt{AISE} \equiv \exists \mathtt{hasCausePointEvent}.\mathtt{IschemicTraumaticEvent} \sqcap \exists \mathtt{hasCauseDurative}.(\mathtt{Atherosclerosis} \sqcap \exists \mathtt{hasDiagnosticEvidence}.\mathtt{SevereStenosis})$$

$$(2)\ \mathtt{IschemicTraumaticEvent} \equiv \exists \mathtt{hasDiagnosticEvidence}.\mathtt{RelevantLesion}$$

$$(3)\ \mathtt{RelevantLesion} \equiv \exists \mathtt{hasDiagnosticEvidence}.\mathtt{LeftRelevantLesion} \sqcup \exists \mathtt{hasDiagnosticEvidenceRightRelevantLesion}$$

$$(4)\ \mathtt{LeftRelevantLesion} \equiv \exists \mathtt{hasDiagnosticEvidence}.((\mathtt{ModerateLesion} \sqcup \mathtt{SevereLesion}) \sqcap \exists \mathtt{hasSide}.\mathtt{Right}) \sqcap \exists \mathtt{hasDiagnosticEvidence}.(\mathtt{SevereStenosis} \sqcap \exists \mathtt{hasSide}.\mathtt{Right})$$

Figure R.4. Fragment of the DL Formula defining AISE (Atherosclerotic Ischemic Stroke Evident). The first axiom states (a) the existence of a relation from AISE to Ischemic Traumatic Event via the Has-Cause-Point-Event relation, and (b) the existence of a relation from AISE to Atherosclerosis via the Has-Cause-Durative; in addition, Atherosclerosis must have Severe Stenosis as Diagnostic Evidence (this is a specific requirement of Atherosclerotic Ischemic Stroke when it is Evident). According to the first axioms, the presence of AISE requires the presence of Ischemic Traumatic Event, Atherosclerosis and Severe Stenosis. According to the second axiom, Ischemic Traumatic Event requires the presence of a Relevant Lesion, to which it is connected via the Has Diagnostic Evidence Relation. The formula further decomposes Relevant Lesion into Left and Right Relevant Lesion.

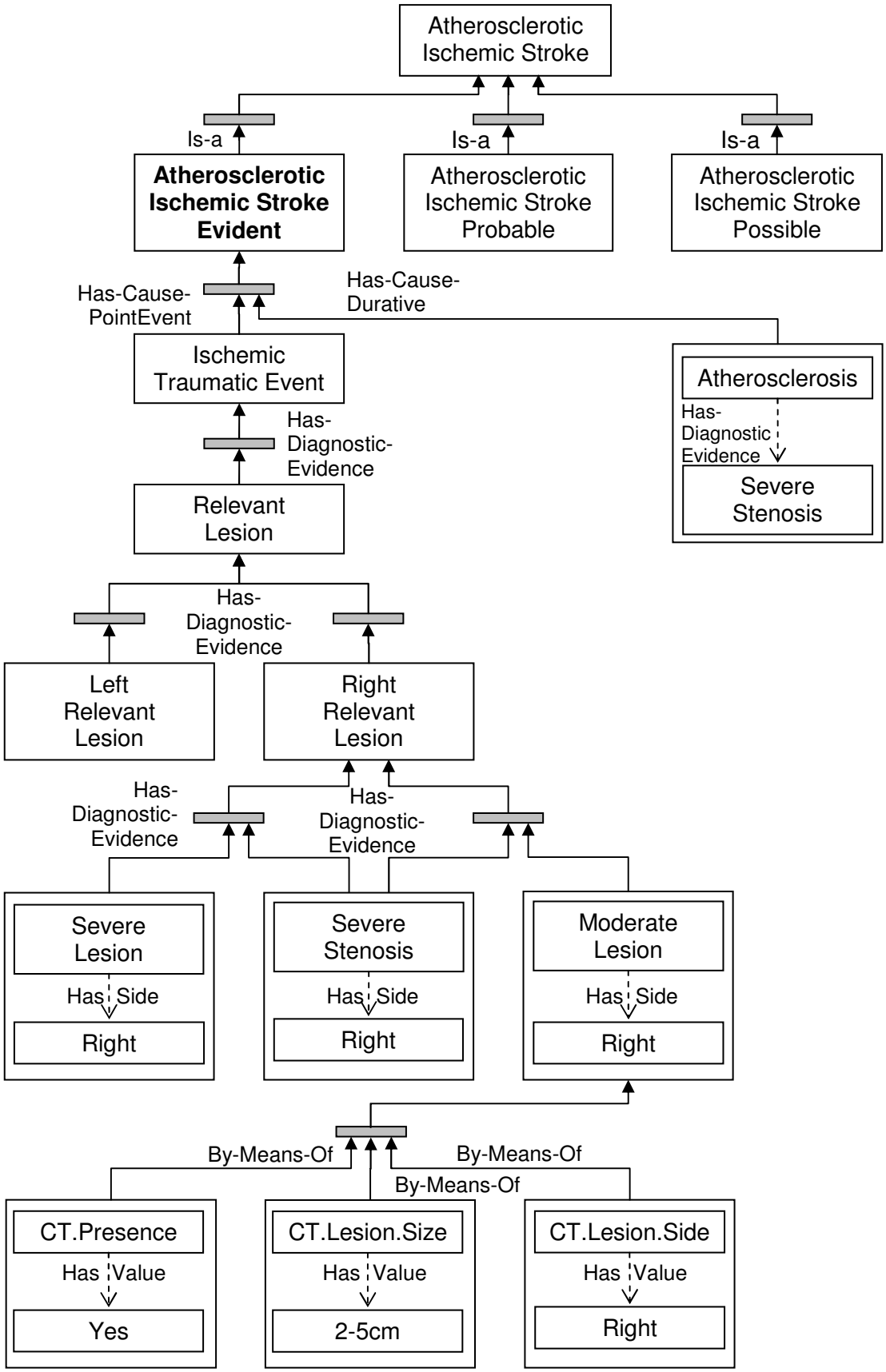


Figure R.5. The decomposition of Atherosclerotic Ischemic Stroke Evident (AISE) into more elementary phenotypes, graphical notation. AISE was not fully decomposed just for space reasons.

The Flexible Mapping in Action: the Carotid Stenosis

Presence of Carotid Stenosis is a Low Phenotype composed by a number of narrower Low Phenotypes via the Has-Diagnostic-Evidence relation. The increasing granularity is due to:

- anatomical part specification (*anatomy*);
- side specification (*topology*);
- degree of occlusion (*quantitative scale*).

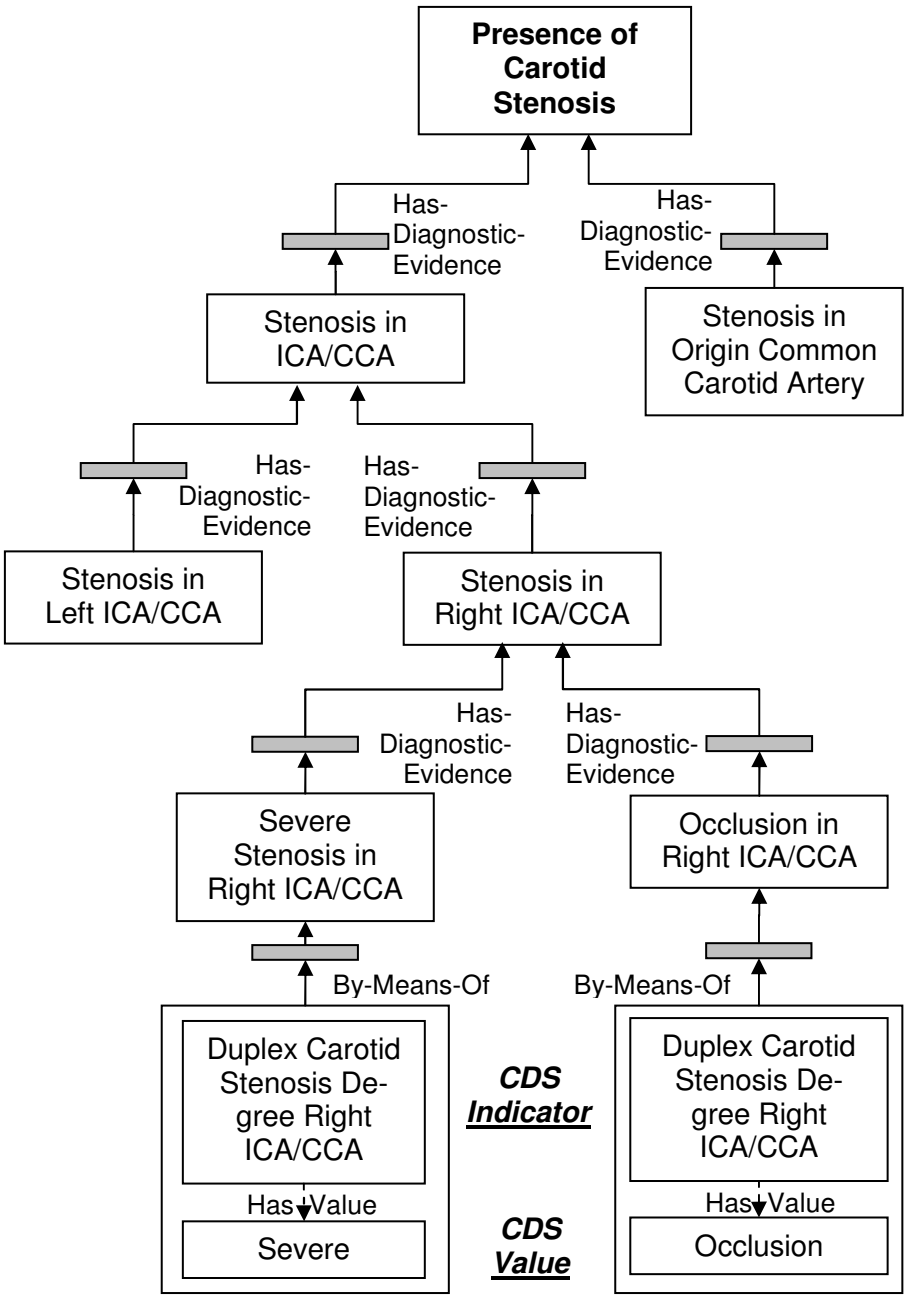


Fig R.6. The Low Phenotype Presence of Carotid Stenosis: the decomposition into narrower phenotypes can be exploited by the flexible mapping.

Indeed, the presence of a stenosis in the carotid artery may be diagnosed if there is an evidence of it in the segment called Internal/Common Carotid Artery (ICA-CCA) or in the Origin of the Common Carotid Artery; the ICA-CCA Stenosis can be diagnosed if there is an evidence of it in the right or in the left Internal/Common Carotid Artery and in particular if there is an evidence of a severe stenosis or a complete occlusion of the vessel.

The occlusion must be diagnosed by means of a specific exam called Duplex, with result equal to occlusion in the ordinal scale. These phenotypes at increasing granularity are all suitable connection points for the flexible mapping.

The Terminological Base

As an additional feature, the NEUROWEB System supports semantic searches over biomedical literature. To support this functionality, the NW-RO was mapped to other ontological/terminological resources, at the present stage MeSH[29] [65-66] and OpenGALEN[30] [67-69][31].
The mapping from the NW-RO to MeSH and OpenGALEN was generated by visual inspection and manual comparison of terms/concepts. The properties `owl:equivalentClass` and `rdfs:subClassOf` were used to map the terms. For each mapping (NW-RO – MeSH and NW-RO – OpenGALEN), a separate OWL file was generated containing only the mappings. The mapping files can be updated anytime.
The NEUROWEB – MeSH mapping currently is constituted by 60 `owl:equivalentClass` mappings and 14 `rdfs:subClassOf` mappings; the NEUROWEB – OpenGALEN mapping is constituted by 18 `owl:equivalentClass` mappings and 6 `rdfs:subClassOf` mappings.

Functionalities to be Supported by the Biomolecular Extension

The NEUROWEB consortium is committed to carrying out high-throughput *Single Nucleotide Polymorphism (SNP)* genotyping campaigns [70-71] exploiting DNA samples of previously hospitalized or newly admitted patients; subsequently, the experimentally collected data will be analyzed, in order to identify statistical genotype-phenotype associations. The goal is to identify non-Mendelian genetic factors that can act as risk or protective factors for the ischemic stroke and related clinical conditions; the NEUROWEB research approach is typically multi-genic. This research activity is supported by the NEUROWEB Reference Ontology (NW-RO) through the formulation of clinical phenotypes, and does not require any biomolecular extension.
In relation to genotype-phenotype association studies, two additional tasks were envisioned for the NEUROWEB System; they are currently not supported, but the NW-RO is endued with a prototypal extension to biomolecular entities.
The first task to be supported is the retrieval and integration of genotype-phenotype relations from publicly available resources; this functionality enables to cross-check associations found in NEUROWEB patients, or can provide interesting associations to be further tested by querying the NEUROWEB databases. In order to define the ontology requirements for this functionality, it is first necessary to identify the discrepancies between NEUROWEB phenotypes and the externally-defined phenotypes. Currently, only a handful of publicly-available databases provide genotype-phenotype associations, such as the *Human Gene Mutation Database (HGMD)* [HGMD] and the *NIH Genetic Association Database (GAD)* [72][32]. The amount of available data, its accessibility, and the

---

[29] MeSH (Medical Subject Headings) is a medical thesaurus developed at the US National Library of Medicine (NLM) containing a controlled vocabulary for content indexing of MEDLINE/PubMed paper abstracts.
[30] OpenGALEN is an open-source description-logic based medical ontology and is provided by the Open-GALEN Foundation. The aim of OpenGALEN is to represent the relation among medical domain entities of general use.
[31] SNOMED-CT was not included only for financial reasons. Outside North America, SNOMED-CT is not available for exploitation without commercial license.
[32] *NCBI OMIM*, a robust and informative resource, should not be taken into account for two reasons: (1) most of its content is devoted to mono-genic diseases with Mendelian inheritance, whereas

richness of the information provided are expected to grow in the future, as more and more genome-wide association studies (GWA) are performed and made available to the research community. Given the current situation, we analyzed several HGMD and GAD phenotypes, to identify the correspondences and discrepancies with NEUROWEB phenotypes. Some of the phenotypes typically refer to disease and other clinical conditions (e.g. Diabetes Mellitus Type-II, Coronary Disease, Hypertension, Ischemic Stroke, Atherosclerosis); these phenotypes, when they are of neurovascular interest, can be mapped to some level of the NW-RO. We refer to the reconciliation of these discrepancies as *horizontal integration*; such discrepancies are due to the different criteria applied to categorize clinical conditions. Unlike the case of horizontal integration, other phenotypes are described making explicit reference to genes, gene families, molecular and cellular processes (e.g. *Apolipoprotein A1 Deficiency*, *HDL Deficiency*, *Increased Lipid Metabolism*). The latter cannot be handled unless the corresponding biomolecular entities are introduced into the NW-RO and mapped to NEUROWEB phenotypes. We refer to the reconciliation of these discrepancies as *vertical integration*. We expect this problem to be dominant when genotype-phenotype relations are imported from model organisms, where the observation of anomalies at the molecular and cellular level is a standard research methodology.

The second task to be supported consists in offering possible explanations of the biological mechanisms underlying a statistically-relevant genotype-phenotype association. This task requires the ontology to decompose a phenotype into its underlying biological mechanisms, finally referring to biomolecular processes and parts. This scenario entails the same vertical integration we mentioned for the first functionality.

The solution of the horizontal and vertical integration problems goes beyond the NEUROWEB Project scope:

- the horizontal integration should be addressed by defining a general representation meta-modeling for clinical phenotypes and pathological conditions (cf [56]), with specific communities adopting the meta-model and defining specialized solutions when required; on the contrary, the current situation for human genotype-phenotype resources is rather messy: different players develop repositories without using a formal model (or at least robust criteria) for phenotype annotation;
- the vertical integration should be addressed by providing *orthogonal reference ontologies* of biological parts and processes at different granularity levels (OBO [28] has undertaken such a mission); coupled to the previous resource, this would enable to efficiently represent the biomolecular basis of clinical phenotypes, when known.

In the outlined scenario, the biomolecular extension of the NW-RO should be regarded as a prototypal solution, to be tested and enriched in the project follow-up, and to be compared to alternative solutions being proposed in the community.

Discussion

The Discussion section comprises two main parts. In the first part, we consider the scientific contributions of the NEUROWEB Reference Ontology (NW-RO), beyond the narrow scope of NEUROWEB System functionalities; namely, we will consider how the NW-RO can act as a catalyst, prompting clinicians to share and discuss different diagnostic protocols on a more formal basis, and how it can contribute to the ontology design methodology in the biomedical arena.

NEUROWEB's commitment is to non-Mendelian multi-genic effects; (2) OMIM content is organized as free text, a feature that severely impairs the computational exploitation of the information content.

Value of the NEUROWEB Reference Ontology beyond the NEUROWEB Project

The First Mereology of Neurovascular Conditions

We argue that the NUROWEB Reference Ontology (NW-RO) is the first ontology proposing a meta-model for neurovascular conditions (i.e. neurovascular phenotypes). Most of the existing representation resources are terminologies (e.g. SNOMED-CT, MeSH) or disease taxonomies (e.g. DisO). THE NW-RO is committed to represent the *mereology* of phenotypes, which is the formally structured build-up of parts and sub-parts required for the phenotype in order to be characterized. The part of a phenotype is another phenotype, depicting a narrower clinical condition, until the minimal elements of diagnostic practice (i.e. the CDS elements) are reached. This feature enables the NW-RO to formulate the phenotypes on the basis of database content, granting methodological coherence in association studies.

A Formal and Computable Model of the TOAST Classification System

The TOAST Classification System is an important resource in the neurovascular community; the improvement of stroke classification criteria is a focus of current research in the neurovascular community. The contribution offered by the NW-RO is a formalization of the TOAST concepts, enabling the clinician to organize their discussions, and the structuring of their resources, on a more formal and structured basis.

Towards a General Meta-Model for Clinical Findings and Clinical Phenotypes

A cooperative project has been recently started, under the auspices of the NCBO (National Center for Biomedical Ontology of the NIH, National Institutes of Health) and OBO (Open Biomedical Ontology) Foundry, in order to formulate a general-validity meta-model for symptoms, signs, clinical findings and clinical phenotype [56]. The NEUROWEB modeling experience will be a valuable contribution to this effort, whose undertaking reveals the importance of this modeling field in present-day biomedical informatics.

Further Challenges of Clinical Condition Modeling

Though effective in supporting the NEUROWEB System Functionality, the NEUROWEB Reference Ontology does not wear out all the issues and challenges arising when modeling clinical conditions. Two topics will be briefly sketched in the following subsections: the relation to the very different clinical state paradigm in machine learning approaches, and the possibility of an “operative system” for association studies, i.e. a computational system supporting the formulation and iterative refinement of clinical phenotypes in association studies.

Clinical States: the Ontological Model and the Data Space Paradigm

A very interesting aspect is the representation of a clinical state as a portion of a multi-dimensional space, a paradigm typically adopted in machine learning approaches to decision support and automated diagnosis. Such a representation typically requires an array of independent and quantitative (or at least ordinal categorical) indicators, and may fall short when the indicators have a rich semantic structure inter-relating them. However, these approaches offer the advantages of crossing the borders of pre-set categories and work on a continuous space. The advent of more powerful and systematic diagnostic devices, in the fields of imaging and biomarkers, may empower this scenario, effectively grounding every phenotype on a space of quantitative features.

An Operating System for Association Studies?

The NEUROWEB Reference Ontology is a content theory of clinical indicators and clinical conditions pertinent to ischemic stroke clinical medicine. Compared to other dis-

ease "ontologies", the NW-RO offers the advantage of going beyond taxonomical disease classification by identifying building-blocks. Such a system enables the user to reformulate pre-built clinical phenotypes or to build them ex-novo. There are several reasons why this is desirable for association studies. Most important, different types of association studies may require different criteria to aggregate indicators into clinical phenotypes; for instance, genotype-phenotype association studies require the homogeneity of the biological processes underlying the disease condition; for instance, cardioembolic and atherosclerotic ischemic stroke may require similar anti-coagulant treatments to minimized the ischemic damage, but the pre-disposing genetic factors may be very different due to the diverse underlying etiologies; moreover cardioembolic stroke itself may be a reasonable unitary etiology when considering the origin of the embolus, but the embolization process can be caused by a very heterogeneous array of factors triggering the coagulation cascade (e.g. coronary disease, atrial fibrillation due to congenital anomalies, prosthetic heart valves), that it may have to be broken down into subtypes in order to mine more meaningful genotype-phentype relations.
Another interesting aspect is the generation of totally new phenotypes, going beyond the traditional diagnostic categories, exploiting biomolecular criteria (e.g. similarity of the perturbed pathways according to gene expression or mutation studies).
All these issue require an in-depth analysis (going beyond the current project commitments) and also a stratified experience with the use of compositional ontologies for clinical conditions that is not yet available at the present stage. The scenario is the development of an "operating system"[33] for association studies, i.e. a computational system supporting the formulation and iterative refinement of clinical phenotypes in association studies.

The Challenges Posed by the Biomolecular Extension

The strategy adopted for the NW-RO biomolecular extension relied on

1. decomposing the phenotypes in order to distinguish diagnostic evidences from biological processes at the organ and tissue level, such as the durative etiological background and the traumatic point-event (ischemic or hemorrhagic);
2. mapping the biological processes from the organ and tissue level to the cellular and molecular scale.

Clearly, this is not the only possible approach. The foundational notion of process may be replaced by a different one. As far as we know, it's the only published proposal[34].
The most problematic aspect concerning the notion of biomolecular process is the boundary problem: since biomolecular processes can be regarded as networks of molecular transformation reactions (such as formation of complexes, post-translational modification of proteins, transport of biomolecules, binding to DNA promoters, template-dependent nucleic acid polymerization, etc…), it has hard to identify robust boundaries for processes; it is not even clear whether such boundaries are, to a certain extent, natural, or they mostly are mere fiat boundaries imposed to organize knowledge. This problem is typically reflected by the complicated structure of the Gene Ontology / Biological Process ontology: higher-level processes are decomposed into narrower and narrower sub-processes, generating a multi-level hierarchy.

---

[33] The computer oprative system provides the user a high-level representation of the programs running and the data stored in memory supports (file system), and it provides the user an array of fucntionality to interact with them.
[34] A different proposal, yet unpublished, was presented by Barry Smith and Richard Schuermann at the Dallas Workshop for Symptoms, Signs, Clinical Findings and Clinical Phenotypes [56].

Acknowledgments

The NEUROWEB Project was a cooperative effort of the NEUROWEB Consortium partners (clinical, technological, and public administration).
The clinical partners were:

- AOK-OPNI: Orszagos Pszichiatriai es Neurologiai Intezet;
- INNCB: Fondazione IRCCS Istituto Neurologico "Carlo Besta";
- MI-EMC: Erasmus Universitair Medisch Centrum Rotterdam;
- UOP: University of Patras.

The technological partners were:

- UNIMIB-DISCo: Università degli Studi di Milano-Bicocca (UNIMIB), Dipartimento di Informatica Sistemistica e Comunicazione (DISCo);
- PU: Pannon University;
- VELTI SA.: Velti A.E. Software products and services;
- MICROSYSTEMS: Microsystems S.r.l.;
- SIRSE: SIRSE-NET S.p.a.;
- CNR-ITB: Consiglio Nazionale delle Ricerche.

The public administration partners were:

- REGLOM: Regione Lombardia.

The following part is devoted to the special acknowledgments concerning the work described in this chapter. Within the UNIMIB-DISCo work team, Dr. Gianluca Colombo was the lead person for the design and implementation of the NEUROWEB Reference Ontology (NW-RO), and also for the knowledge acquisition activity. Giuseppe Frisoni and Prof. Flavio De Paoli were the lead persons for the definition of the NEUROWEB data integration strategy and its computational framework; they also supervised the NW-RO exploitation within the NEUROWEB computational framework, and its terminological extension; Giuseppe Frisoni contributed significantly also to the knowledge acquisition activity. Daniele Merico was the lead person for the analysis of the biomolecular resources and knowledge structures; he also significantly contributed to the NW-RO design and to the discussion of the overall project methodology. Michaela Guendel was the lead person for the terminological extension of the NW-RO. Prof. Giancarlo Mauri acted as supervisor and coordinator of the UNIMIB-DISCo work team. In addition, Prof. Yanis Ellul (UOP) pioneered the work on the Core Data Set, which was then refined with the contribution of other clinical partners. Prof. Zoltan Nagy (AOK-OPNI) provided thoughtful insights concerning the classification criteria in neurovascular medicine, an essential step for the definition of the NW-RO meta-model. Dr. Giorgio Boncoraglio (INNCB) provided a comprehensive description of the TOAST Classificatory System. Prof. Istvan Vassanyi (PU) significantly contributed to the computational exploitation of the NW-RO and its mapping to the local databases. Last but not least, Prof. Eugenio Parati provided overall guidance of the project, and formulated intriguing future scenarios for phenotype reformulation.
Special thanks also to Prof. Maria Luisa Lavitrano (Dept. Surgical Sciences, Azienda Ospedaliera San Gerardo, Monza, Italy) for elucidating the state-of-the-art understanding of the molecular and cellular mechanisms of atherosclerosis.

Publications

Journals

[1] Colombo G*, Merico D*, Frisoni G, Vassanyi I, Antoniotti M, De Paoli F, Mauri G.
(* equal contrib.)
An ontological modeling approach to neurovascular disease study: the NEUROWEB case.
(submitted to Biomedical Informatics Journal, September 2008)

[2] Colombo G, Merico D, Nagy Z, De Paoli F, Antoniotti M, Mauri G.
Ontological modeling at a domain interface: bridging clinical and biomolecular knowledge.
(accepted by The Knowledge Engineering Review, May 2008)

Conference Proceedings

[3] Colombo G, Antoniotti M, Merico D, De Paoli F, Mauri G.
Ontological modelling for neurovascular disease study: issues in the adoption of Description Logic.
(International Workshop on DL; June 2007; Brixen, Italy; published on CEUR-WS vol. 250)

[4] Colombo G, Merico D, Mauri G.
Reference Ontology Design for a Neurovascular Knowledge Network.
(MTSR 2007; 11-12 October 2007; Corfu, Greece; to appear in Advances in Metadata and Semantics Research (Springer))

[5] Bonomi A, Colombo G, Merico D, Palmonari M, Vizzari G.
Sharing the NEUROWEB Ontology through the web: the NavEditOW Approach.
(accepted for MultiAgent Systems & Bioinformatics 2008; 13 September 2008; Cagliari, Italy; to appear on LNCS)

Book Chapters

[6] Colombo G, Merico D, Gündel M.
Metadata and Ontologies for Health.
(to appear in Handbook of Metadata, Semantics and Ontologies, Sicilia MA (Ed.), World Scientific Publishing Co., Hackensack, NJ, USA)

Reference

[7] McLendon K.
Electronic medical record systems as a basis for computer-based patient records.
J AHIMA. 1993 Sep;64(9):50, 52, 54-5.

[8] Mantas J.
Electronic health record.
Stud Health Technol Inform. 2002;65:250-7.

[9] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS.
"Evidence based medicine: what it is and what it isn't".
BMJ 1996; 312 (7023): 71–2.

[10] Woolf, SH.
The Meaning of Translational Research and Why It Matters.
JAMA 2008;299;211-213.

[11] Berners-Lee T, Hendler J, Lassila O.
The Semantic Web.
Scientific American, May 2001.

[12] Bard SY, Rhee JBL.
Ontologies in biology: design, applications and future challenges.
Nature Reviews Genetics 2004, 6(5): 213-222.

[13] Rubin DL, Shah NH, Noy NF.
Biomedical ontologies: a functional perspective.
Brief Bioinform. 2008 Jan;9(1):75-90.

[14] Smith B, Welty C.
Ontology---towards a new synthesis.
Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS2001). ACM Press, 2001.

[15] Gruber T.
Toward Principles for the Design of Ontologies Used for Knowledge Sharing.
International Journal Human-Computer Studies Vol. 43, Issues 5-6, November 1995, p.907-928.

[16] Guarino N.
Formal Ontology, Conceptual Analysis and Knowledge Representation.
International Journal of Human-Computer Studies, 43(5-6):625–640, 1995.

[17] Gruber T.
Ontology.
Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008.

[18] Chandrasekaran B, Josephson JR, Benjamins VR.

What are ontologies, and why do we need them?
Intell. Syst. and Their Appl., IEEE; Vol. 14(1), Jan/Feb 1999 Page(s):20-26.
[19] Kuhn T.
The Structure of Scientific Revolutions.
Chicago, University of Chicago Press, 1970.
[20] Johnson-Laird PN.
Mental models: Towards a cognitive science of language, inference, and consciousness.
Cambridge, MA, Harvard University Press, 1983.
[21] Boland J, Tenkasi RV.
Perspective making and perspective taking in communities of knowing.
Organizational Science 6(4), 1995, pp 350-372.
[22] Benerecetti M, Bouquet P, Ghidini C.
Contextual Reasoning Distilled.
Journal of Theoretical and Artificial Intelligence (JETAI), 12(2000), pp 279-305.
[23] Smith B.
The Basic Tools of Formal Ontology.
in Nicola Guarino (ed.), Formal Ontology in Information Systems. Amsterdam, Oxford, Tokyo, Washington, DC: IOS Press (Frontiers in Artificial Intelligence and Applications), 1998, 19–28.
[24] Cuenca Grau B, Horrocks I, Parsia B, Patel-Schneider P, Sattler U.
Next Steps for OWL.
In Proc. of OWL: Experiences and Directions, CEUR Proceedings, 2006.
[25] Cocchiarella N.
Conceptual Realism as a Formal Ontology.
In Formal Ontology, Poli Roberto and Simons Peter (eds). Dordrecht: Kluwer 1996. pp. 27-60
[26] Thomasson AL.
Methods of Categorization,
Formal Ontology in Information Systems - Proceedings of the third International Conference, 2004 (FOIS-2004). IOS Press, pp 3-16.
[27] Arp R, Smith B.
Function, Role, and Disposition in Basic Formal Ontology.
Available from Nature Precedings <http://hdl.handle.net/10101/npre.2008.1941.1> (2008).
[28] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S.
The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.
Nat Biotechnol. 2007 Nov;25(11):1251-5.
[29] Donnan GA, Fisher M, Macleod M, Davis SM.
Stroke.
Lancet May 2008, 371 (9624): 1612–23.
[30] Adams HP Jr, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, Marsh EE 3rd.
Classification of Subtype of Acute Ischemic Stroke, Definition for Use in a Multicenter Clinical Trial, TOAST. Trial of Org 10172 in Acute Stroke Treatment.
Stroke 1993, 24: 35-41.
[31] Goldstein LB, Jones MR, Matchar DB, Edwards LJ, Hoff J, Chilukuri V, Armstrong SB, Horner RD, Bamford J.
Improving the reliability of Stroke subgroup classification using the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) criteria (Commentary).
Stroke 2001; 32: 1091-1097.
[32] Ay H, Furie KL, Singhal A, Smith WS, Sorensen AG, Koroshetz WJ.
An evidence-Based Causative Classification System for Acute Ischemic Stroke.
Ann. Neurol. 2005; 58: 688–697.
[33] Horrocks I, Sattler U, Tobies S.
Practical Reasoning for Expressive Description Logics.
In LPAR: 6-10 September 1999; Tbilisi, Georgia. Harald Ganzinger, David McAllester and Andrei Voronkov (Eds): Springer-Verlag; 1999: 161-180.
[34] Sattler U.

Description Logics for the Representation of Aggregated Objects.
In Proceedings of the 14th European Conference on Artificial Intelligence. 2000; Amsterdam. W. Horn: IOS Press; 2000.
[35] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (Eds).
The description logic handbook: theory, implementation, and applications.
New York, NY, USA: Cambridge University Press; 2003.
[36] W3C - Web Ontology Language.
[http://www.w3.org/TR/owl-guide/]
[37] Protégé User Documentation.
[http://protege.stanford.edu/doc/users.html]
[38] Feyerabend P.
Against Method: Outline of an Anarchistic Theory of Knowledge.
London, Verso, 1975.
[39] Laudan L.
Progress and its problems: Toward a theory of scientific growth.
Berkeley, University of California at Berkeley, 1977.
[40] Lakatos I.
The Methodology of Scientific Research Programmes.
Philosophical Papers Volume 1, Cambridge, Cambridge University Press, 1978.
[41] Brown JS, Duguid P.
Organizational Learning and Community of Practice: Toward a Unified View of Working.
Organization Science, 1991, pp 40-57.
[42] Wenger E.
Community of Practice: Learning, Meaning and Identity.
Cambridge, Cambridge University Press, 1998.
[43] Hildreth P, Kimble C, Wright P.
Communities of Practice in the Distributed International Environment.
Journal of Knowledge Management, 2000.
[44] Nonaka H, Takeuchi I.
The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation.
Oxford University Press, 1995.
[45] Davenport T, Prusack L.
Working Knowledge - How Organizations Manage What They Know.
HBS Press, 1998.
[46] Holsapple CW, Joshi KD.
Organizational knowledge resources.
Decision Support Systems, 2001, pp 39-54.
[47] Bowker GC, Turner W, Star SL, Gasser L.
Social science, technical systems, and cooperative work: Beyond the great divide.
Lawrence Erlbaum Associates, 1997.
[48] Bandini S, Colombo E, Colombo G, Sartori F, Simone C.
The Role of Knowledge Artifacts in Innovation Management: the Case of a Chemical Compound Designer.
CoP, International Conference on Communities and Technologies, Kluwer Academic Publishers, Dordrecht, 2003, pp 327-345.
[49] Smith B, Kusnierczyk W, Schober D, Ceusters W.
Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain.
Proceedings of KR-MED 2006.
[50] Studer R, Benjamins RV, Fensel D.
Knowledge Engineering: Principles and Methods.
Data Knowledge Engineering Journal, 1998, pp 161-197.
[51] Preece A, Sleeman DH, Flett AN, et al.
Better Knowledge Management through Knowledge Engineering.
IEEE Intelligent Systems, 2001, pp 36-43.
[52] Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN.
Human Gene Mutation Database (HGMD): 2003 update.
Hum Mutat 2003, 21(6):577-81.

[53] The Gene Ontology Consortium.
Gene Ontology: tool for the unification of biology.
Nat. Genet. 2000, 25: 25-29.
[54] Maglott D, Pruitt K, Tatusova T.
Entrez Gene: A Directory of Genes
NCBI Handbook (created 2005).
[http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch19]
[55] Gomez-Perez A, Corcho-Garcia O, Fernandez-Lopez M.
Ontological Engineering.
Springer-Verlag, New York, NY, U.S.A., 2003.
[56] [www.bioontology.org/wiki/index.php/DallasWorkshop]
[57] Smith C, Goldsmith CA, and Eppig J.
The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.
Genome Biology, 6(1):R7, 2004.
[58] PATO – Phenotypic Quality Ontology.
[www.bioontology.org/wiki/index.php/PATO:Main Page]
[59] Cimino JJ.
Coding systems in health care.
Methods of Information in Medicine, 35(4–5):273–284, December 1996.
[60] Donnelly K.
SNOMED-CT: The advanced terminology and coding system for eHealth.
Stud Health Technol Inform 2006, 121:279-290.
[61] Giannangelo K, Fenton SH.
SNOMED CT survey: an assessment of implementation in EMR/EHR applications.
Perspect Health Inf Manag. 2008 May 20;5:7.
[62] Bodenreider O, Smith B, Kumar A, Burgun A.
Investigating subsumption in SNOMED-CT: An Exploration into Large Description Logic-based Biomedical Terminologies.
Artificial Intelligence in Medicine, 39(3):183–195, 2007.
[63] van der Kooij J, Goossen WT, Goossen-Baremans AT, de Jong-Fintelman M, van Beek L.
Using SNOMED CT codes for coding information in electronic health records for stroke patients.
Stud Health Technol Inform. 2006;124:815-23.
[64] Disease Ontology (DisO) – The NUgene Project.
[diseaseontology.sf.net]
[65] US National Library of Medicine: Fact Sheet - Medical Subject Headings (MeSH).
[http://www.nlm.nih.gov/pubs/factsheets/mesh.html]
[66] US National Library of Medicine: Medical Subject Headings – Qualifiers
[http://www.nlm.nih.gov/mesh/topsubscope2008.html]
[67] Rogers J, Rector A.
GALEN's Model of Parts and Wholes: Experience and Comparisons.
Proc AMIA Symp. 2000; 714-718.
[68] Rector AL, Rogers JE, Zanstra PE, Van Der Haring E
OpenGALEN: Open Source Medical Terminology and Tools.
Proc AMIA Symp. 2003; :982 14728486.
[69] OpenGALEN: Background and GALEN Model.
[http://www.opengalen.org.]
[70] International HapMap Consortium.
A haplotype map of the human genome.
Nature, 437(7063):1299–1320, 2005.
[71] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K.
dbSNP: the NCBI database of genetic variation.
Nucleic Acids Research, 29:308–311, 2001.
[72] Becker K, Barnes KC, Bright TJ, Wang SA.
The Genetic Association Database.
Nature Genetics, 36(5):431–432, May 2004.

Supplementary Material

Phenotype Description Template for Remote Knowledge Acquisition

Authors

Daniele Merico, Giuseppe Frisoni, Marco Antoniotti

Date

31/01/2007

Introduction

This is a template for the description of "phenotypes", as they are defined in the NEUROWEB Glossary. Please try to fill-in most of the fields while avoiding free text unless explicitly allowed.

In the last NEUROWEB meeting, we have agreed that "phenotypes" are central to the system architecture, in order to group patients for clinical and genetic association analyses. A list of phenotypes has recently circulated among the partners. The group at Erasmus-Rotterdam (MI-EMC) provided a sizable textual description of the "atherosclerotic ischemic stroke", with some implicit reference to the TOAST etiological criterion. The group at Budapest (AOK-OPNI) provided a list of "new stroke phenotypes" without further characterization.

In order to complete this initial part of the modeling work, making sure that the descriptions roundtrips are minimized, we are proposing to formulate a phenotype according to the following hierarchical structure.

Please note that at this stage we are collecting only the NEUROWEB phenotypes (that is, phenotypes derived from standard clinical protocols, such as those defined by the TOAST, or anyway considered a standard in medical practice - as defined in the NEUROWEB Glossary).

How to deal with the USER-DEFINED phenotypes will be the subject of an upcoming document.

Phenotype Description Template

The following is a description of the template parts. The actual template follows.

1. Name:
   a simple, short and clear, symbolic name for the phenotype.
2. Description:
   a brief textual description of the phenotype and its relationship with other phenotypes and current medical practice.
3. Core Data Set (CDS) Mapping:
   In brief, a set of parameter ranges (constraints) based on the fields contained in the CDS initially compiled by Prof. Yanis Ellul, which must be satisfied in order for the phenotype occurrence to be valid.
   3.1. list of clinical data to be used (a list of names, referring to Core Dataset fields);
   3.2. list of constraint ranges on each clinical data value;
   3.3. a boolean formula combining the constraints from the previous list (using AND, OR, XOR, NOT, IMPLIES…); this formula is not compulsory, although an effort to produce it would be highly appreciated. If you have a hard time depicting the combination of constraints required for phenotype occurrence validity, please at least describe the phenotype in plain text within the section "description";
   3.4. additional comments (e.g. to what degree is the formulation satisfactory? Are there elements missing from the CDS?);

Note: You may find it natural to think at a phenotype as a combination of more elementary phenotypes, which you have already characterized with this form. In that case, you do not have to repeat the list of clinical data and their constraints: simply put a link to the elementary phenotypes in which those lists are reported, but do remember to put the phenotype combination in the formula.

(Fictional) Example of Template Use
NAME: occipital atherosclerotic stroke
DESCRIPTION: Occipital atherosclerotic stroke is …
CDM:
Clinical data to be used:
1. Conscious level
2. Type of brain imaging
    a. Relevant scan lesion (generic)
    b. Relevant scan lesion location
3. Type of vessel study
    a. Left carotid degree of stenosis
    b. Right carotid degree of stenosis

Constraints:
1. Conscious level > 2
2. Type of brain imaging = MRI
    a. Relevant scan lesion (generic) = yes
    b. Relevant scan lesion location = occipital
3. Type of vessel study = angiography
    a. Left carotid degree of stenosis > 50%
    b. Right carotid degree of stenosis > 50%

Formula: AND(1, AND(2, 2.a, 2.b), AND(3, OR(3.1, 3.b)))
Additional comments: as you wish…

NF-Y Targets

Characterization of NF-Y Transcription Factor Targets: Function and Transcriptional State.

Abstract

NF-Y is a trimeric transcription factor containing H2A/H2B-like subunits, which specifically binds to the CCAAT box, a common eukaryotic promoter element. In this work, the NF-Y binding sites were assessed by ChIP-chip (chromatin immunoprecipitation on chip) on chromosome 20, 21, 22. The majority of the target genes are bound by NF-Y in the promoter and/or within the coding region.

To gain insights into NF-Y-dependent transcriptional regulation, we assessed the relationships between NF-Y binding and (1) positive histone marks (H3K9-14ac and H3K4me3), as detected by ChIP-chip, (2) the transcriptional state of the corresponding targets. As a result, NF-Y loci can be divided in two distinct clusters: (i) a large cohort contains H3K9-14ac and H3K4me3 marks and correlates with expression and (ii) a sizeable group is devoid of these marks and is found on transcriptionally silent genes. Within this class, we find that NF-Y binding is associated with negative histone marks, such as H4K20me3 and H3K27me3.

NF-Y removal by a dominant negative NF-YA leads to a decrease in the transcription of expressed genes associated with H3K4me3 and H3K9-14ac, while increasing the levels of many inactive genes. These data indicate that NF-Y is embedded in positive as well as in negative methyl histone marks, serving a dual function in transcriptional regulation, as an activator or as a repressor.

Background

The NF-Y Transcription Factor Binds the CCAAT Box

Transcription initiation by RNA polymerase II at class II gene promoters is a finely regulated process requiring the interplay of many different transcription factors [1-2]. General transcription factors (GTFs), namely TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, recognize specifically the core promoter, recruit the RNA polymerase, and help melt the DNA, thus enabling the initiation of transcription at the correct start site. Assembly of this preinitiation complex is controlled by a large set of transcriptional activators and repressors that recognize, in a sequence-specific way, DNA sequence motifs located on proximal or distal enhancer regions of the promoters and function by contacting either directly or indirectly, through co-activators and co-repressors, the GTFs.
Among such sequence motifs, the CCAAT box is known to be one of the most frequent. This has been illustrated by several unbiased bioinformatic studies of large sets of vertebrate promoters [3-10]. Testing has shown that the CCAAT box significantly contributes to promoter activity [11].
Different entities contain the word CCAAT in their acronyms, but several types of evidence indicate that NF-Y, also termed CBF and HAP2/3/5 for Saccharomyces cerevisiae, is the CCAAT regulator. (i) Highly specific antibodies were used in supershift electrophoretic mobility shift assays and chromatin immunoprecipitations (ChIPs) with a plethora of different promoters [12-13]. (ii) Specific dominant negative NF-YA vectors were employed in cotransfection and adenovirus infection experiments. (iii) Nucleotides flanking the CCAAT box emerged in the bioinformatic studies cited above, with perfect matches to NF-Y preferences, as assayed by in vitro binding studies [14]. It is therefore reasonable to conclude that NF-Y is by far the major protein that regulates this element.

The Protein Structure of the NF-Y Complex

NF-Y is a ubiquitous heteromeric transcription factor (TF) composed of three subunits, NF-YA, NF-YB, and NF-YC, all necessary for DNA binding [11]. NF-YB and NF-YC contain conserved core regions, which display very high structural similarity to histone folds composed of three alpha helices separated by short loop/strand regions (figure B.1.b) [15]. NF-YB/NF-YC association is essential for NF-YA binding and sequence-specific DNA interactions (figure B.1.c).

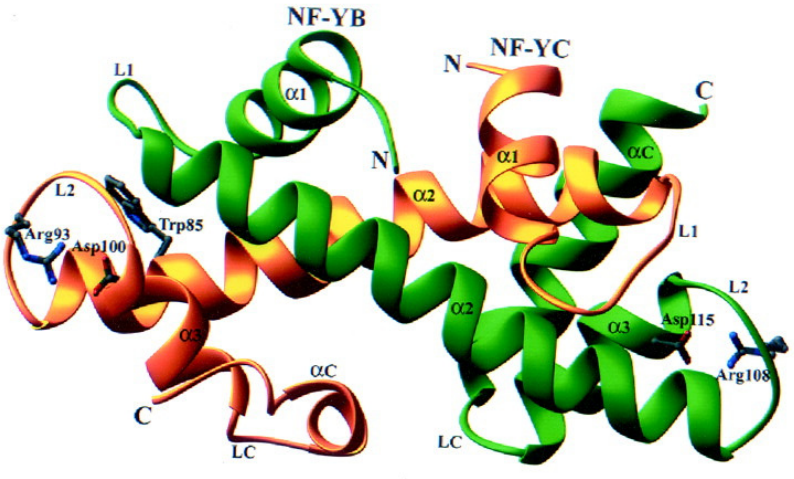


Figure B.1.a. The NF-Y B-C dimer (ribbon representation).

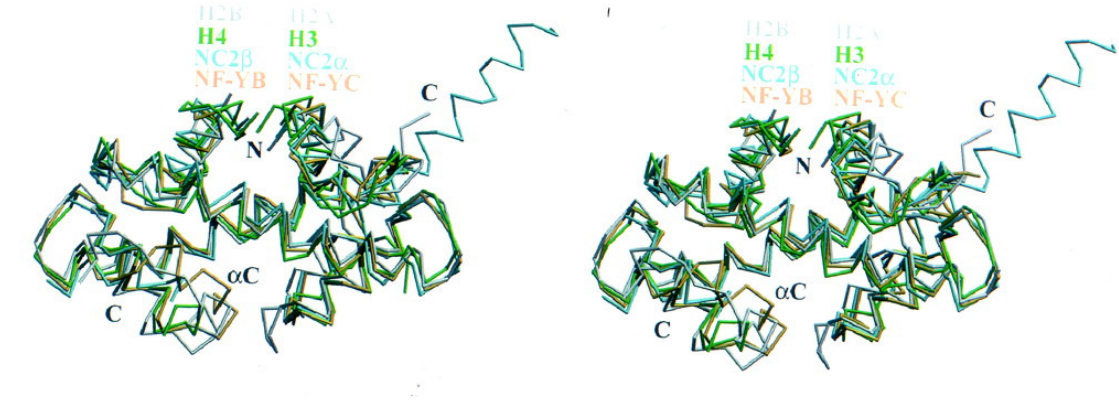


Figure B.1.b. Stereo C-alpha traces of the superimposition of the NF-YC/NF-YB (orange), H2A/H2B (gray), NC2/NC2 (blue), and H3/H4 (green) histone pairs.

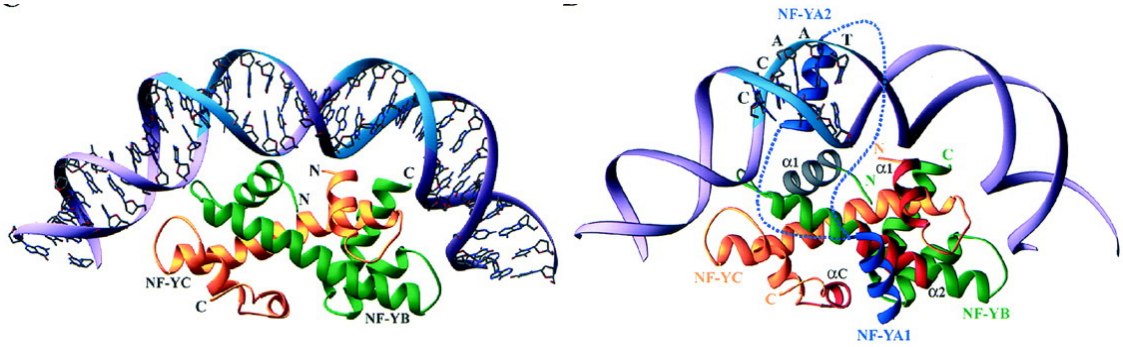


Figure B.1.c. Model of the complex of NF-YC/NF-YB with the CCAAT element: the DNA backbone is shown as ribbons (purple) with the bases displayed; the two possible locations of the CCAAT box, according to the modeling, have been colored cyan. NF-YA is colored blue. The two alternative positions for the linker connecting NF-YA1 and NF-YA2 sub-domains are shown as blue dotted lines.

How does NF-Y Regulate Transcription?

Interestingly, CCAAT boxes are present at a specific location within promoters, typically between -60 and -100 bp from the transcriptional start site (TSS). Within this context, NF-Y is not a powerful activator but rather a promoter organizer that cooperates with the activity of neighboring TFs. The emergence of genome-wide technologies now allows a look at TF binding in a more systematic and unbiased way. Previous studies, performed with CpG island arrays and with an oligonucleotide array representative of a small set of human promoters, left us with an inconclusive picture as to the widespread NF-Y distribution in vivo [16].

A more comprehensive understanding of a TF mechanism of action can be attained only by considering the chromatin environment, defined by patterns of histone post-translational modifications (for a full review, please refer to [17-19]). In particular, H3K4 trimethylation (H3K4me3) and H3K9-14 acetylations (H3K9-14ac) are associated with an active chromatin environment in regions and promoters that are transcribed or poised for rapid induction by external stimuli [20-23]. Their presence in vivo has been detailed at the single-gene level, and location analysis confirmed their widespread distribution in the proximity of promoters [21, 24-32]. Interestingly, while the presence of these marks precedes gene activation, an increase in their levels is generally noticed in systems of inducible transcription. In NF-Y-dependent endoplasmic reticulum stress promoters, for example, a substantial increase in H3 acetylation and H3K4me3 was seen after induction, while NF-Y binding was detailed before [33-34]. On the other hand, posttranslational modifications such as H3K9, H3K27, and H4K20 methylations are known to be associated with inactive or actively repressed areas of the genome [35-37]).

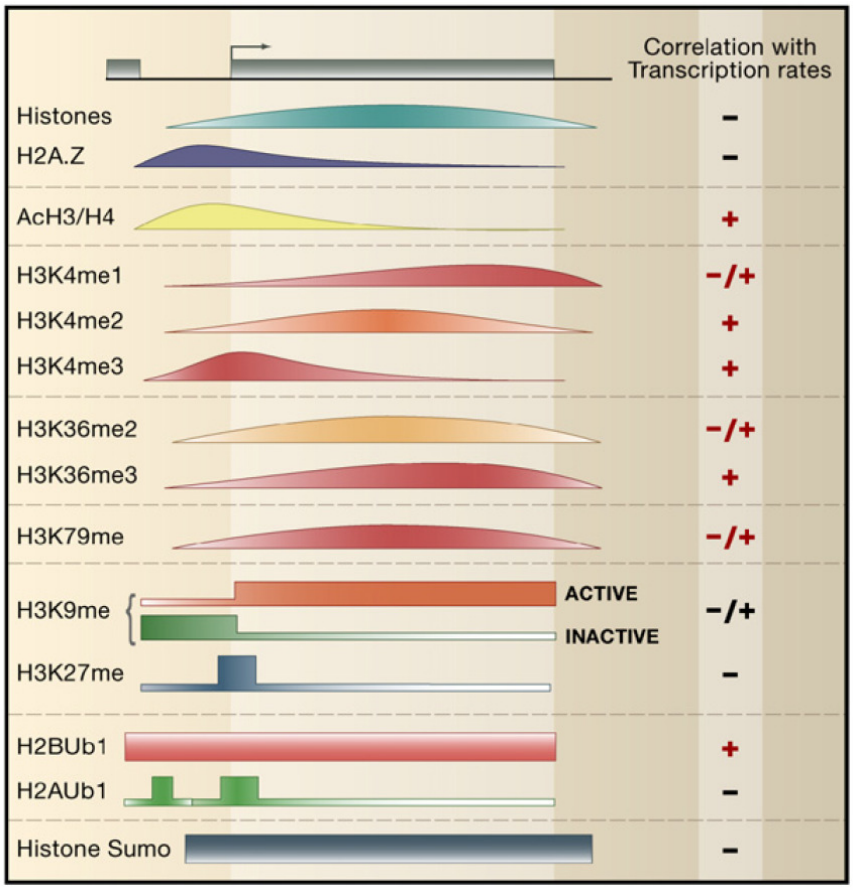


Figure B.2. The different histone modifications, their locations on the (transcribed) gene sequence, and their role of activation or repression of transcription.

In general, understanding the relationship between histone modifications and TF binding is quite relevant; it is not completely clear, at the moment, whether the binding of certain factors is required for specific modifications to be brought in through recruitment of histone-modifying enzymatic machines or whether the binding of TFs is allowed only by a preexisting nucleosomal environment with an appropriate pattern of histone modifications.
To shed light on the relationship between NF-Y and active histone marks, we used high-density tiling arrays of chromosomes 20, 21, and 22 in ChIP-on-chip experiments.

Materials and Methods

Cells, infections, and PCR analysis

HeLa cells were grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal calf serum, 1% antibiotics (penicillin and streptomycin), and L-glutamine in 5% CO_2. HCT116 $p53^{-/-}$ cells were grown in McCoy's medium. Infection of HCT116$p53^{-/-}$ and HeLa cells with Ad-YAm29, Ad-NF-YA, and Ad-GFP adenoviruses was carried out as described previously [38]. Semiquantitative ChIP PCR and reverse transcription-PCR (RT-PCR) analyses were performed according to standard procedures, in the linear range of amplification, essentially as previously described [39].

ChIP, amplicon generation, and ChIP on chip

ChIPs were performed essentially as described previously [16]. Briefly, 5 x 10^6 cell equivalents of chromatin, < 0.5 kb, were immunoprecipitated with 15 µg of anti-YB,

anti-H3K4me3 (Abcam), and anti-H3K9-14ac (Upstate) antibodies. Immunoprecipitation-enriched DNAs were used to generate amplicons for hybridization experiments. Parallel ChIPs were run with a Flag control antibody (Sigma), and bona fide NF-Y targets were used to check enrichments before and after the amplification steps. The generation of amplicons from the individual ChIPs was performed by following the protocol of ligation-mediated PCR previously described [40-41]. Design of the oligonucleotides, preparation of the slides, hybridization, and scanning of the fluorescence intensities were performed by Nimblegen. Validations of the results by semiquantitative PCR were performed on independent, nonamplified ChIPs using primers listed in File S1 in the supplemental material [A1]. Negative-histone-mark ChIPs were also performed as described previously [16] using anti-H3K9me3 (Abcam), -H3K27me3 (Abcam), and -H4K20me3 (Upstate) antibodies.

ChIP-on-chip data analysis

The Cy5 and Cy3 raw data obtained from each experiment were normalized by subtracting the median and then adding back the average median of the two channels. The independent median normalization was chosen since channel distributions were not parametric and therefore the mean was not suitable. This approach does not perturb the detection of biologically relevant peaks, while allowing efficient removal of noise due to unbalance between the two channels (not shown). After single-channel normalization, the Cy5/Cy3 ratio was calculated for each probe and then converted to the corresponding Z score. This kind of transformation is useful when seeking to compare the relative standings of items from distributions with different means and/or standard deviations and fitted our requirement of performing comparison between different arrays. Box-and-whisker plots clearly showed that Z-score conversion allowed for a direct comparison of independent data sets derived from biological replicates (not shown). Peak search was then performed essentially as previously reported [40], with the so called “peak first” approach. A given percentile threshold was chosen for all experiments, and all probes with enrichment values greater than the percentile were selected. An NF-Y “peak” was defined as a stretch of at least three adjacent probes (spaced by no more than 150 bp on the genomic sequence) exceeding the percentile threshold. In H3K9-14ac and H3K4me3 experiments, five adjacent probes were considered to define a peak. A hypergeometric P-value-based distribution was then associated with the peaks obtained. Finally, peaks obtained from the different experiments were merged, and a given peak was predicted as biologically relevant if present in at least two out of four experiments for NF-Y or two out of three experiments for active histone marks, with corresponding probability values computed with a binomial distribution.

Transcription Microarray Analysis

The GEO GSE6022 data set, containing a total of three untreated HeLa cell samples, each being a biological replicate, was used for HeLa profiling experiments. The GEO GSE6207 data set, containing seven untreated samples, was used for HepG2 profiling experiments. The data set from Sato et al. [42], containing three untreated samples, each being a biological replicate (Affymetrix HG-U133-A platform), was used for human embryonic stem (ES) cell profiling experiments. The GEO GSE8884 (Affimetrix HG-U133 plus 2.0 platform) was also used for the latter analysis, yielding essentially the same results despite the larger coverage of the platform (not shown). Expression signals for microarray data were calculated according to the rma algorithm and used for quality control. Absent (A) and present (P) calls were also calculated according to the open-source version of the MAS5 algorithm and used both for quality control and comparisons. A gene was called P if it was present at least in two out of three of the samples. The same rules were applied for A calls. A gene was called UN if it did not fall in the previous categories. The A, P, and UN groups have no overlaps.

Results

Confirming NF-Y CCAAT Specificity

To assay the prevalence of NF-Y binding on CCAAT promoters, we performed ChIP analysis with an anti-NF-YB antibody on chromatins derived from seven different cell lines.

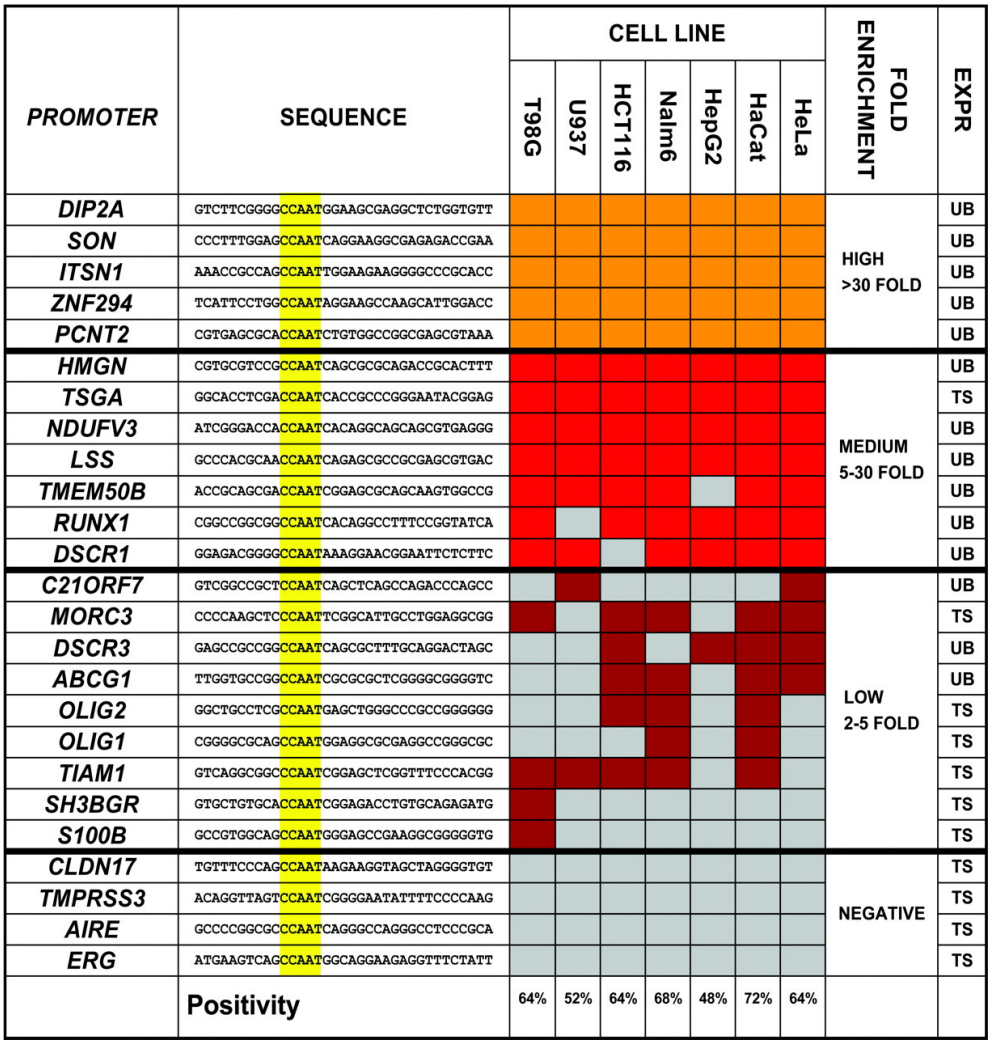


Figure R.1. NF-Y binding to CCAAT promoters in vivo. Twenty-five randomly picked CCAAT promoters from chromosome 21 were analyzed by ChIP with anti-NF-YB antibody on chromatin derived from seven different cell lines. Enrichment over that for an irrelevant Flag control antibody was assessed by semiquantitative PCR and classified as high (> 30-fold), medium (5- to 30-fold), or low (2- to 5-fold). Colored boxes indicate positivity and gray boxes indicate no enrichment in ChIPs. The bottom line refers to percentages of NF-Y-positive sites for each cell line. The expression patterns of the respective genes were analyzed according to available transcriptome data (UniGene build no. 186) and reported on the right. UB (ubiquitous) refers to genes expressed in at least 10 out of 45 body sites of UniGene's EST Profile Viewer, while TS (tissue specific) refers to genes showing expression in less than 10 body sites.

We checked 25 randomly picked CCAAT-containing promoters of chromosome 21 genes. Figure R.1 indicates that most of the promoters were indeed bound by NF-Y in the majority of cells. However, the degree of enrichment over a control Flag antibody varied substantially and was arbitrarily divided into three levels: high (> 30-fold), me-

dium (5- to 30-fold), and low (2- to 5-fold). The expression patterns of these genes were then analyzed according to available Unigene transcriptome data and divided into two broad categories: genes that were ubiquitously expressed and those that had a tissue or cell type preference. With few exceptions, TMEM50B, RUNX1, and DSCR1, the high- and medium-enrichment genes, were all in the ubiquitous category and bound by NF-Y in all cell lines. Four genes, CLDN17, TMPRSS3, AIRE, and ERG, all tissue specific, were negative in all cell lines. Finally, weakly enriched sites were mostly found on tissue-specific genes. We conclude that (i) NF-Y is bound to the majority of CCAAT-containing promoters and (ii) the highest enrichment correlates with a ubiquitous expression profile.

NF-Y ChIP on chip on tiling arrays

ChIP-chip Experiment, Data Analysis and Stringency Selection

To define in an unbiased way the "landscape" of NF-Y binding in vivo, we performed NF-YB ChIP-on-chip analysis in HeLa cells with a tiling array containing all non-repetitive sequences of the entire chromosome 21 and large parts of chromosomes 20 and 22. Fifty-mer oligonucleotides are tiled every 50 bp, giving a high degree of resolution and an overall representation of 2.2% of the human genome. To obtain maximal definition of the locations, ChIPs were performed using chromatin shorter than 500 bp, with a mean size of 250 to 300 bp. The enrichments present in the starting ChIPs were assayed against a Flag control antibody and were routinely ~ 100-fold when checked on bona fide NF-Y target genes [A1: File S2].

Four NF-YB probes derived from independent ChIPs were Cy5 labeled and hybridized together with the corresponding Cy3 input DNA, used as an internal control. After normalization of single channels, ratios between NF-YB and input probes were calculated and converted to Z scores (see Materials and Methods for details). An NF-Y peak was defined as a set of at least three consecutive probes whose Z scores deviated significantly from the average of normalized data. Typically, selected ratios were above 1.8 to 2.3 depending on the replicate. Only peaks present in at least two out of four replicates were collected for further analysis.

Overall, the number of NF-Y binding sites ranged from 757 to 2120, depending on the stringency applied (P values between 1.6×10^{-5} and 1.4×10^{-4}). The ChIPs performed for figure R.1 served as a guide for data analysis: none of the NF-Y-negative promoters was scored, while a majority of the positives were retrieved (table RT.1), and, with the fourth stringency considered, 14/16 positives were recovered.

NF-Y Binding Site Classification by Genomic Location

Binding sites were then classified and divided into three main categories: (i) "promoters" (PR), describing NF-Y locations residing from -2 kb to +0.5 kb relative to the TSS of a RefSeq sequence and representing 9 to 11% of the total depending on the stringency; (ii) "genes" (GE), indicating NF-Y locations residing within RefSeq-annotated genes; and (iii) "elsewhere" (EL), referring to locations external to promoters and intergenic regions. The last two categories each accounted for 42 to 48% of the sites.

Independent ChIP validations were performed on 35 loci derived from the less-stringent criteria: all NF-Y promoter loci tested scored positive in standard ChIPs and so did 11/15 sites among the "genes" cohort ([A1: File S3]). Thus, based on the pre-validation and validation ChIPs, the highest stringency reported in table RT.1 highly underscored the extent of NF-Y binding and the 1.43×10^{-4} stringency reflected more closely the actual targets, with 10 to 15% false positives/negatives.

Stringency (P)	Binding Sites				RefSeq TUs (n = 907)			Pre-validated promoters	
	Total	% (number)			% (number)		All	NF-Y [+]	NF-Y [-]
		PR	GE	EL	PR	GE			
1 (1.06E-05)	757	11.1 (84)	45.4 (344)	43.5 (329)	9.8 (89)	26.1 (237)	298	9/16	0/9
2 (2.88E-05)	1,176	9.7 (114)	46.9 (551)	43.5 (511)	13.6 (123)	34.8 (316)	386	11/16	0/9
3 (6.79E-05)	1,533	8.9 (137)	47.4 (727)	43.6 (699)	16.0 (145)	40.4 (366)	446	11/16	0/9
4 (1.43E-04)	2,120	9.0 (191)	48.1 (1,020)	42.9 (909)	21.6 (196)	47.4 (430)	519	14/16	0/9

Table RT.1: NF-Y location analysis: binding sites and TUs.

Category	% (number) of sites containing a CCAAT box (±150 bp)			
	1 (P = 1.06E-05)	2 (P = 2.88E-0.5)	3 (P = 6.79E-05)	4 (P = 1.43E-04)
Total	63.3 (479)	62.8 (739)	61.8 (947)	61.6 (1,306)
PR	70.2 (59)	71.1 (81)	66.4 (91)	63.4 (121)
GE	59.6 (205)	59.3 (327)	58.5 (425)	58.6 (598)
EL	66.3 (218)	64.8 (331)	64.4 (431)	64.6 (587)

Table RT.2: NF-Y sites containing at least one CCAAT box.

We next analyzed NF-Y sites in terms of transcriptional units (TUs), defined as University of California, Santa Cruz, RefSeq-annotated genes (hg17 assembly) with their respective promoters; overall, there are 907 TUs (with no redundant promoter) within the regions considered here. As expected, by comparing the number of NF-Y promoter locations with that of the corresponding TUs (191 versus 196), we could recover, on average, one location per positive promoter.
A different picture emerged for NF-Y GE category sites, since the number of NF-Y$^+$ TUs was significantly lower than the overall number of locations: 430 versus 1,020. Considering that some of these locations referred to overlapping or divergent units, this implies that on average two or three NF-Y binding sites were found per positive TU.
We also analyzed in greater detail the distribution of NF-Y sites (figure R.2). Within PR, most sites reside in the core area, as expected. Within GE, the site distribution shows a relatively steep decline as a function of the distance from the TSS, in agreement with a relevant role for NF-Y at the 5' ends of genes. Interestingly, we found that more than 50% of NF-Y$^+$ promoters possess an additional NF-Y site within the body of the gene (figure R.2.b). Within the EL category, one-third of the locations overlap (20%) or are nearby (9%+4%) GenBank annotated mRNAs with no RefSeq definition, which are most likely sites of Pol II activity (figure R.2.b).

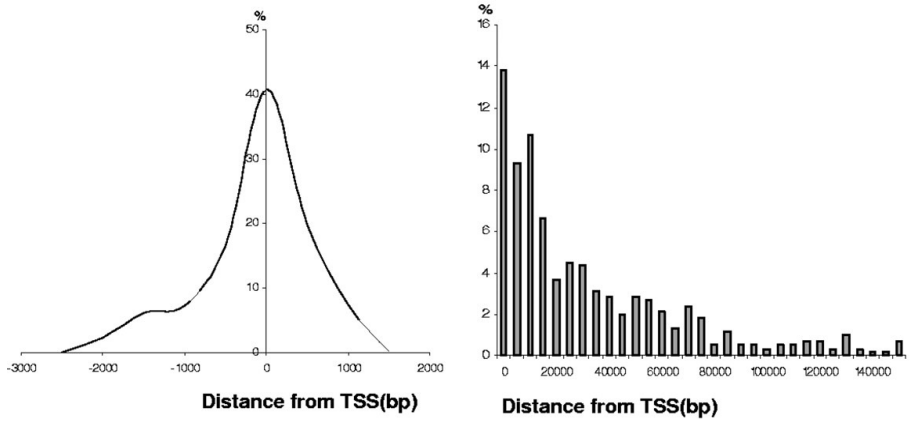


Figure R.2.a. Location of NF-Y binding sites. NF-Y sites were plotted according to their positions with respect to the TSS. Left: NF-Y promoter sites. Right: NF-Y GE sites.

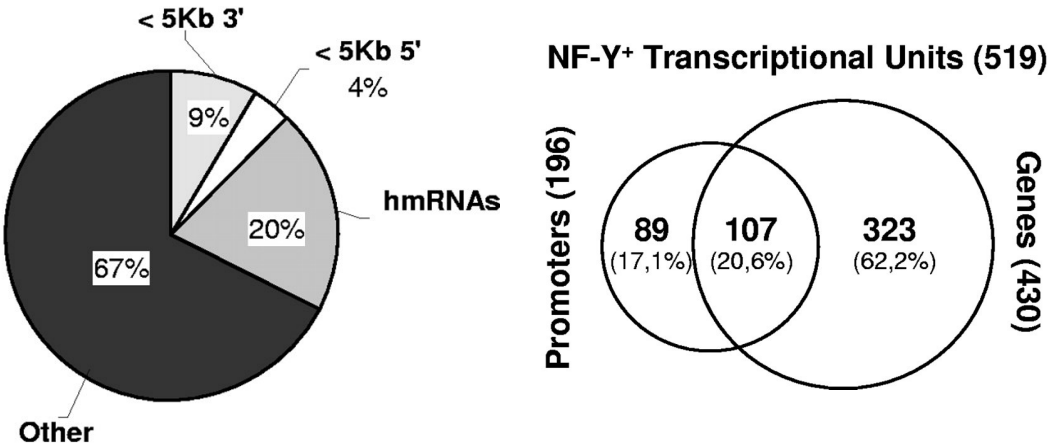


Figure R.2.b. Categories of NF-Y binding sites. Left: EL sites were classified as present within 5 kb to the 5' or 3' end of an annotated gene, as present within a GenBank-annotated human mRNA not corresponding to a RefSeq (hmRNAs), or far from annotated genes (other). Right: Venn diagrams showing the overlap between

NF-Y TUs containing a site in the promoter and those containing a site within the body of the gene.

Presence of CCAAT boxes in NF-Y locations.

It was of interest to assess whether NF-Y locations contain CCAAT boxes. For this analysis, we considered a short interval of 500 bp centered on the peak, which is most likely a stringent but reliable window considering the length of the original chromatin [A1: File S2]. The pentanucleotide was found in 61.6% of total locations, with a higher percentage in PR and lower in GE (table RT.2). We also calculated the overall number of CCAAT boxes per location. Based on purely statistical considerations, the expected frequency is around 1 CCAAT box/location; a specific enrichment was clearly evident in PR, since 2 to 2.5 CCAAT boxes per promoter were present (figure R.3). For the GE and EL locations, a lower enrichment of 1.5 to 1.7 CCAAT boxes per location was scored. Hence, we conclude that CCAAT boxes are present and overrepresented in most NF-Y locations.

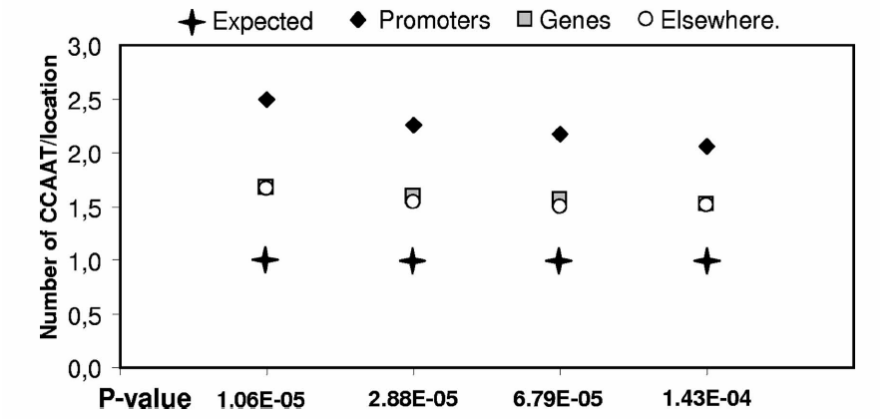


Figure R.3. CCAAT boxes in NF-Y locations. Average numbers of CCAAT boxes present in the different NF-Y location categories were plotted for the four stringencies initially considered. This number was calculated based on the actual mean dimensions of the identified NF-Y binding sites, ranging from 460 to 560 bp depending on the stringency, divided by the theoretical occurrence of the pentanucleotide CCAAT (once per 512 bp).

NF-Y Binding and Transcriptional State

H3K9-14ac and H3K4me3 ChIP-chip Detection

To characterize NF-Y sites in terms of the chromatin environment, we analyzed H3K9-14 acetylations and H3K4 trimethylation on the same tiling platform (table RT.3). The data, derived from triplicate experiments, were analyzed essentially with the same stringency criteria used for NF-Y, except that a peak was defined as a stretch of at least five consecutive probes. This window was selected since these modification were supposed to cover larger part of a genomic sequence compared to the punctuate nature of TFs binding events. We validated a number of locations by independent ChIPs: with one exception, all Promoters were confirmed, whereas for Genes, adherence to ChIP on chip data was high (14/15 for H3K9-K14ac, 13/15 for H3K4me3).

H3K9-14ac sites are more abundant than H3K4me3, on average three times. The distribution is also different: nearly 50% of H3K9-14ac is found within the GE category, and a minority in PR; on the contrary, most of the H3K4me3 locations are found in PR, in agreement with this mark being more abundant at the very 5' of genes.

We also analyzed H3K9-14ac and H3K4me3 sites in terms of Transcriptional Units identified (out of the 907 present on the tiling platform), initially focusing our attention to promoter regions. 30-45% of the total PRs scored positive for H3K4me3, 33-51% for H3-K9-14ac, depending on the stringency.

We pursued further analysis on the 4th stringency considered. The vast majority of the analyzed PR regions was either positive (39%) or negative (42%) for both marks, with a minority (18%) that was either H3K9-14ac+ or H3K4me3+; 87% of H3K4me3 promoters were also positive for H3K9-14ac (figure R.3.a). The same analysis was repeated for GE sites; as expected, the larger class was H3K9-14ac+/H3K4me3-, most likely due to the high rate of H3K4me3 false negatives; once again, the vast majority -90%- of H3K4me3 sites, was also positive for H3K9-14ac (figure R.4.a).
We then analyzed microarray data of HeLa cells (untreated state). All genes were classified as P (Present) or A (Absent), whit an A calls indicating a complete absence, or very low levels of expression. Overall, 59% of Hela TUs were in the P set and 39% in the A calls. As expected, active histone marks were clearly enriched in P calls, with an extremely robust correlation, in the PR group (figure R.3.b). The correlation was almost absent in the GE group (figure R.4.b). A possible yet speculative reason could be that GE sites are actual promoters controlling alternative transcripts that are not specifically recognized by current Affymetrix platforms.
As a control, we compared our data with H3K9-14ac and H3K4me3 locations previously reported in HepG2 cells [27], also analyzing Affymetrix profiling data of the same cell line. As expected, a strong correlation between active histone marks PRs and gene expression was present in HepG2 cells as well, with very significant p-values (figure R.3.b). Less so it was in the GE list, particularly for H3K4me3 in Hela (figure R.4.b). Interestingly, overlap of the profiling data of Hela and HepG2 was extremely high, with HepG2 cells being the more restricted: in any case most of P and A calls were conserved between the two epithelial lines. This allowed us to match the H3K9-14ac+ and H3K4me3+ TUs in Hela with those determined in HepG2: indeed the intersection was greater than 85% for both marks in PRs (figure R.3.c) and in GE, with the exception of H3K4me3 in the latter category (figure R.4.c).
As a second control, we also compared our data with the genome-wide analysis of H3K4me3 islands performed by Guenther et al. [32] in human embryonic stem cells, obtaining similar results (data not shown).
In conclusion, our analysis is robust and consistent with previous results, showing a strong correlation between H3K4me3 and H3K9-14ac, and between these marks and expression.

String.	H3K9-14ac					
	Peaks				TUs (n = 907)	
	Total	% (number)			% (number)	
		PR	GE	EL	Promoters	Genes
1.06E-05	1,116	29.6 (330)	45.3 (505)	25.2 (281)	33.2 (301)	37.0 (336)
2.88E-05	1,525	25.0 (381)	47.1 (719)	27.9 (425)	38.8 (352)	44.4 (403)
6.79E-05	2,386	18.7 (445)	50.2 (1,198)	31.1 (743)	45.2 (410)	54.9 (498)
1.43E-04	3,148	16.6 (521)	48.2 (1,517)	35.3 (1,110)	51.6 (468)	62.2 (564)

String.	H3K4me3					
	Peaks				TUs (n = 907)	
	Total	% (number)			% (number)	
		PR	GE	EL	Promoters	Genes
1.06E-05	348	71.3 (248)	14.9 (52)	13.8 (48)	30.2 (274)	6.9 (63)
2.88E-05	458	65.7 (301)	18.1 (83)	16.2 (74)	35.7 (324)	9.5 (86)
6.79E-05	823	44.1 (363)	25.6 (211)	30.3 (249)	42.4 (385)	23.3 (211)
1.43E-04	1,012	38.0 (385)	29.6 (300)	32.3 (327)	44.8 (406)	26.8 (243)

Table R.T3. Active histone mark location analysis.

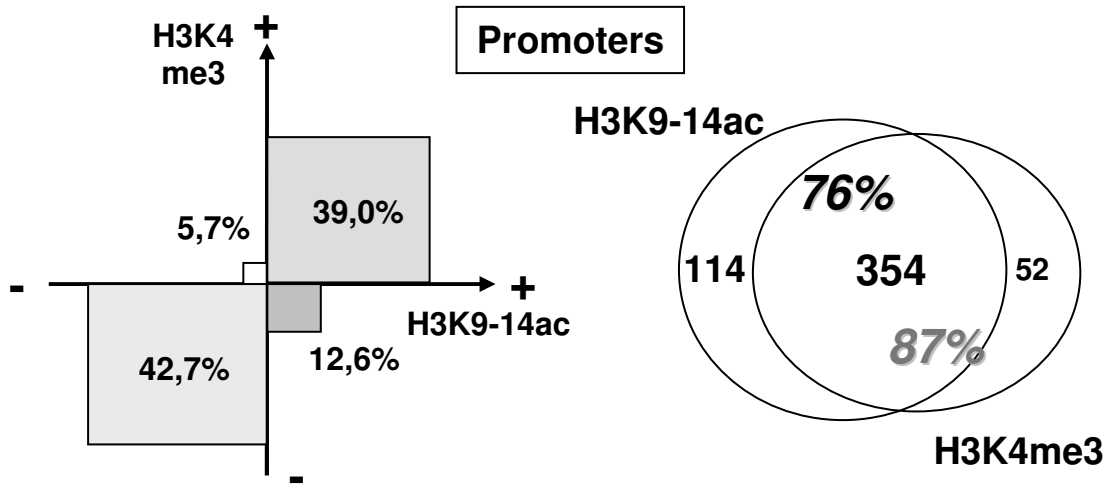


Figure R.3.a. Promoter-class: active histone mark distribution. The presence or absence of both histone marks is highly correlated.

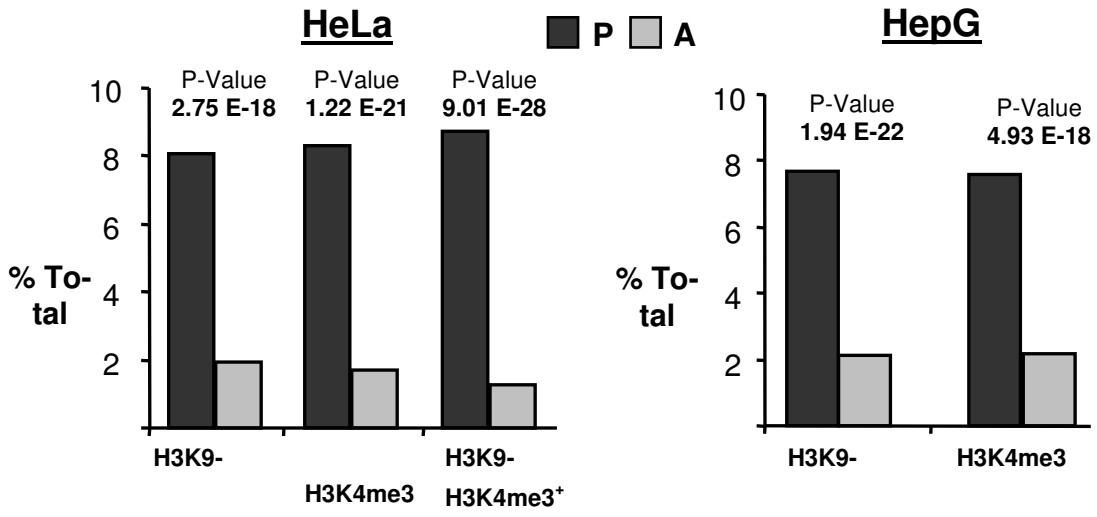


Figure R.3.b. Promoter-class: transcriptional state (Absent/present, A/P) by histone mark distribution. Transcriptional state and histone marks were determined for HeLa cells (left) and HepG2 cells (right), with similar results. Positive histone marks strongly correlate with Present status.

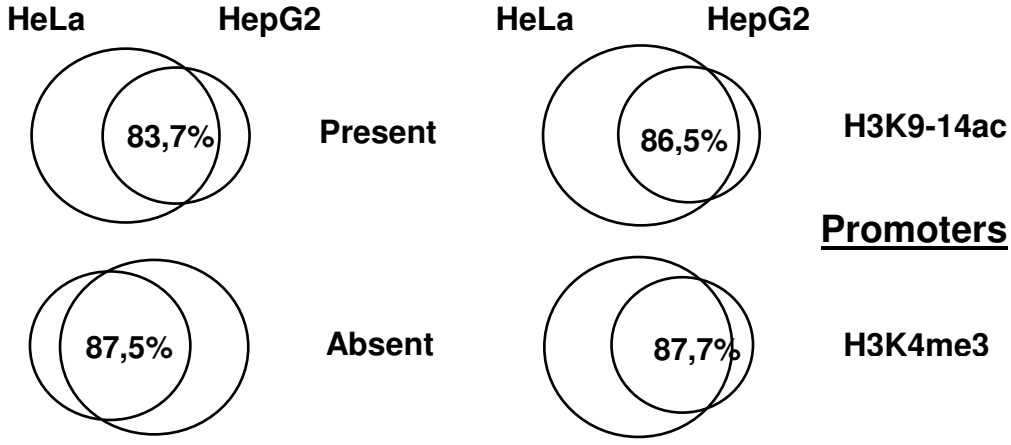


Figure R.3.c. Promoter-class: overlap of the transcriptional status and histone marks presence comparing HeLa and HepG2 cells. The two cell lines display a satisfactory overlap.

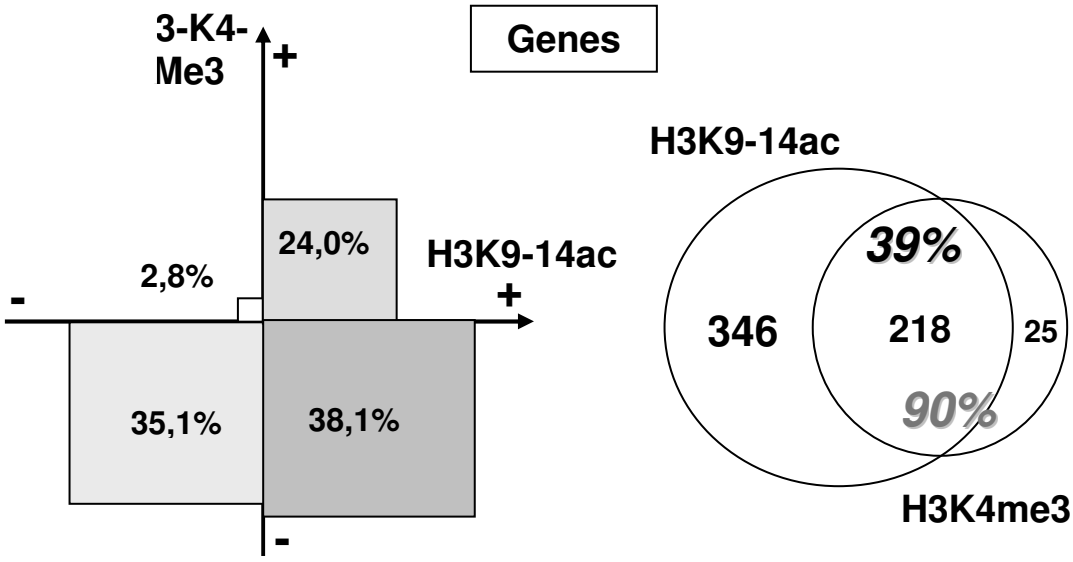


Figure R.4.a. Gene-class: active histone mark distribution. H3K4 trimethylation is affected by a negative bias.

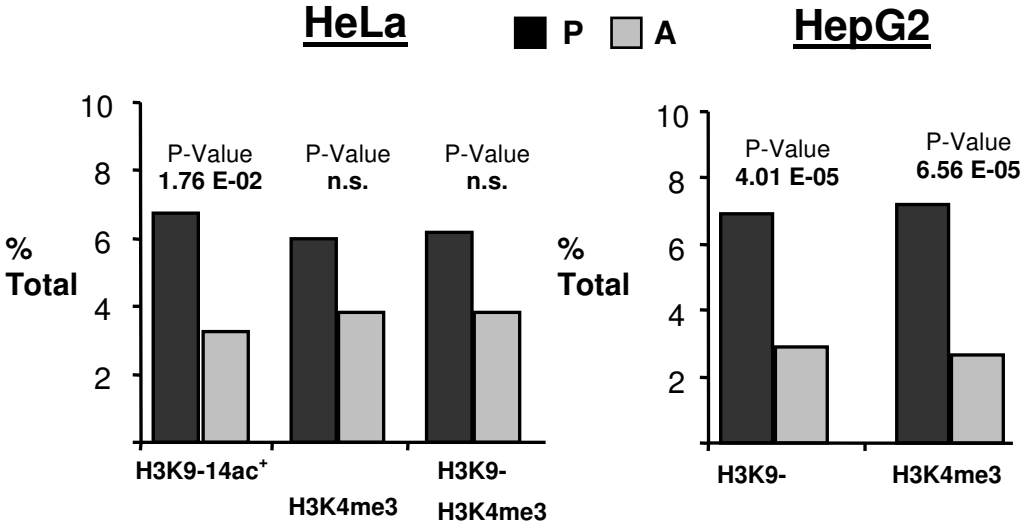


Figure R.4.b. Gene-class: transcriptional state (Absent/present, A/P) by histone mark distribution. Transcriptional state and histone marks were determined for HeLa cells (left) and HepG2 cells (right), with correlation between positive histone marks and Present status stronger in HepG2 than in HeLa. However, even in HepG2, the correlation is weaker in Gene than in Promoter class.

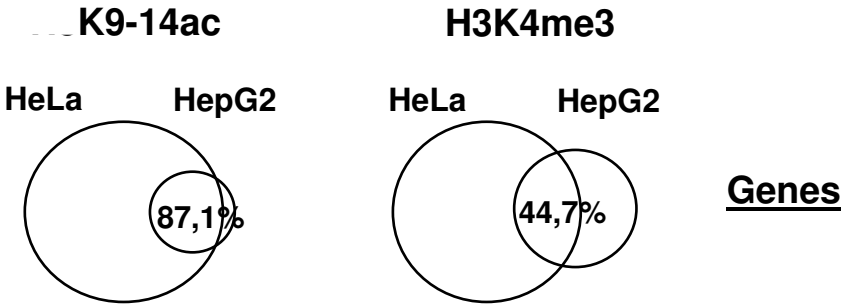


Figure R.4.c. Promoter-class NF-Y targets: overlap of the transcriptional status and histone marks presence comparing HeLa and HepG2 cells. The two cell lines display a satisfactory overlap, except for trimethylations.

Correlation between NF-Y and H3K9-14ac, H3K4me3 Histone Marks

By combining the three different ChIP-on-chip data sets, we observed that the majority of NF-Y+ promoters were also H3K9-14ac+ or H3K4me3+ (68% and 61% respectively, figure R.5); a good proportion, 57%, contained both marks. These proportions are remarkably larger than the baseline, as displayed by the highly significant correlative P values.

Within NF-Y genes, the percentage of H3K9-14ac+ GE sites was higher, at 85%, while the percentage of H3K4me3+ GE sites was lower, at 42%, and, consequently, the percentage of NF-Y+ genes double positive for histone marks was also around 40% (figure R.5). However, the correlation between NF-Y and active histone marks within this category was stronger, with a P value of 1.3×10^{-27} for NF-Y+ H3K9-14ac+ H3K4me3+ genes, indicating that essentially all of $NF\text{-}Y^{+}/H3K4me3^{+}$ sites were also acetylated.

Somewhat surprisingly, we noticed that a sizeable set of NF-Y+ TUs were neither acetylated nor trimethylated, both within promoters (27%) and within the bodies of the genes (18%). This cluster was unexpected, at least in these proportions, and was further analyzed below.

When we reversed the analysis, asking how many H3K9-14ac+ and/or H3K4me3+ TUs were bound by NF-Y, we found that 28% of H3K9-14ac+ and 30% of H3K4me3+ promoters were also NF-Y+ (figure R.5); this is quite significant, considering the promoters that are not active in HeLa cells and considering that the overall percentage of NF-Y+ promoters was 21% (table RT.1). Within the GE category the percentage of H3K9-14ac+ TUs positive for NF-Y was 64% and the percentages of double positives were as high as 79%.

Altogether, these data indicate a significant correlation between NF-Y binding and these histone modifications, either within promoters or within the bodies of the genes, and highlight a smaller population of NF-Y-positive locations that are neither H3K9-14 acetylated nor H3K4 trimethylated.

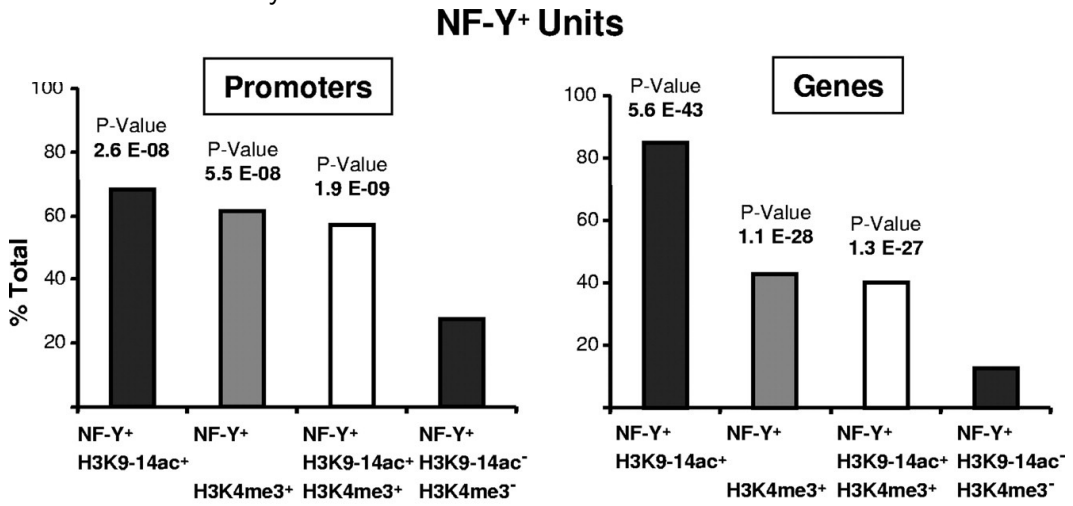


Figure R.5.a. Correlation between NF-Y binding and active histone marks. Percentages of NF-Y+ TUs that scored positive for either H3K9-K14ac or H3K4me3 or for both histone marks.

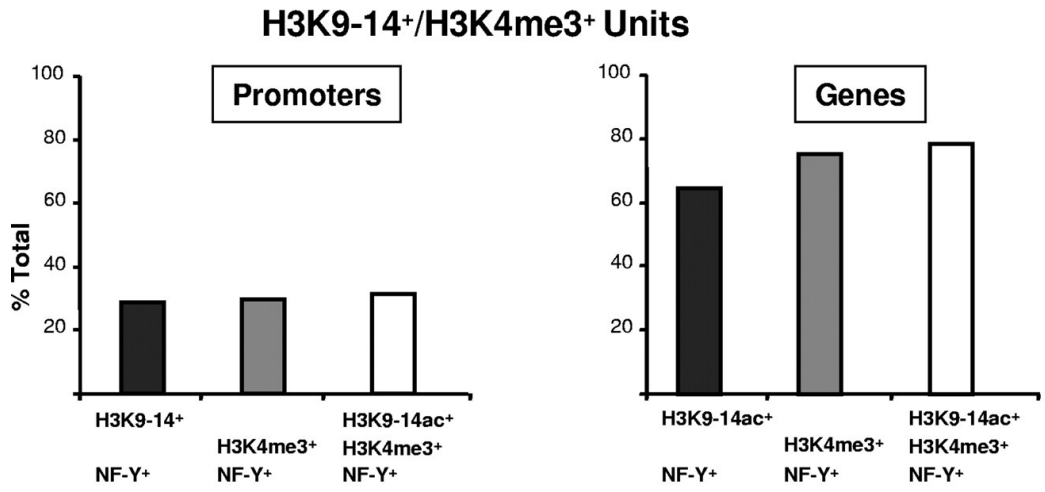
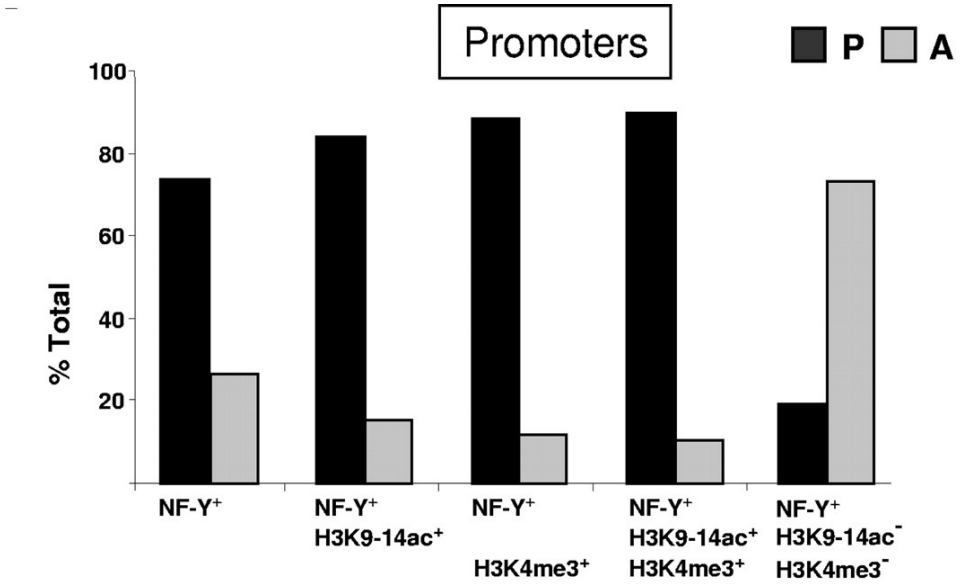


Figure R.5.b. Percentages of histone mark-positive TUs bound by NF-Y.

Correlation between NF-Y, H3K9-14ac, H3K4me3 Histone Marks, and Gene Expression

Next, we interrogated the list of NF-Y+ TUs with respect to both histone marks and expression using available HeLa profiling data sets. Overall, 59% of the 907 TUs analyzed here were in the P set and 39% in the A set. As expected, the majority, 75%, of NF-Y+ promoters followed P calls (figure R.6). NF-Y+ genes showed modest skewing toward expression, with P calls at 67% (figure R.6). We then integrated the active histone mark data into this analysis. Within promoters, the NF-Y+ H3K9-14ac+ and NF-Y+/H3K4me3+ clusters showed clear skewing toward P calls, and even more did the triple-positive NF-Y+/H3K9-14ac+/H3K4me3+ clusters; only a residual 10% of TUs were scored as A in the latter class (figure R.6).

On the other hand, the NF-Y+/H3K9-14ac-/H3K4me3- promoters were strongly enriched in A calls (figure R.6), indicating (i) that NF-Y is bound at these promoters in the absence of active histone marks and (ii) that this cohort of genes is not expressed. A somewhat similar situation emerged for NF-Y+ genes: enrichment of NF-Y+/H3K9-14ac+/H3K4me3+ clusters toward expression was less evident, but once again the majority of NF-Y+/H3K9-14ac-/H3K4me3- genes were found in A calls (figure R.6).



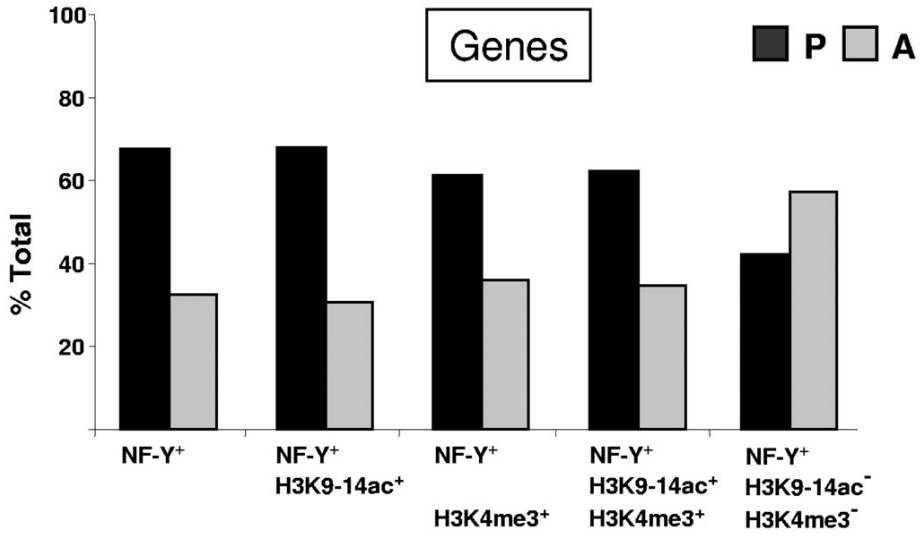


Figure R.6. Distribution of A/P transcriptional status for Promoter-class and Genes-class NF-Y targets, in relation to histone marks. Absence of both histone marks, even in presence of NF-Y, is strongly associated to transcriptional repression.

To gain insights into this dual behavior, NF-Y-positive TUs were subjected to Gene Ontology enrichment analysis after dividing them into those containing at least one active histone mark and those devoid of both modifications. Functional categories expected to be positively affected by NF-Y (cell cycle, cell signaling) were present in the NF-Y+/H3K9-14ac+/H3K4me3+ cohort of genes, but no single function was overwhelmingly enriched; this is consistent with the notion that NF-Y is a broad transcriptional regulator, with specific roles in certain cellular functions (figure R.7.a). A very similar picture was obtained for the down-regulated genes in a NF-YB knockdown experiment (figure R.7.b) [A2]. On the contrary, a different picture emerged with NF-Y+ TUs devoid of active marks, since functions expected to be repressed in HeLa cells, such as sensory perception and immune response, were weakly enriched (figure R.7.c).

NF-Y+ & Active Histone Marks+ Transcriptional Units

GO.ID	*Gene Ontology Category*	*# genes*	*P-Value*	*Enrich. Ratio*
GO:0007049	**cell cycle**	23	1,41E-03	1,46
GO:0051726	regulation of cell cycle	13	8,28E-04	1,65
GO:0000278	mitotic cell cycle	6	0,00E+00	1,9
GO:0008092	**cytoskeletal protein binding**	8	2,88E-03	1,69
GO:0006260	**DNA replication**	5	0,00E+00	1,9
GO:0003712	**transcription cofactor activity**	7	5,55E-03	1,67
GO:0016044	**membrane organization and biogenesis**	9	0,00E+00	1,9
GO:0006897	endocytosis	7	0,00E+00	1,9
GO:0048667	**neuron morphogenesis during differentiation**	7	5,55E-03	1,67
GO:0007169	**transmembrane receptor protein tyrosine kinase signaling pathway**	4	3,93E-02	1,52
GO:0022613	**ribonucleoprotein complex biogenesis and assembly**	5	0,00E+00	1,9
GO:0006928	cell motility	9	0,00E+00	1,9
GO:0004857	enzyme inhibitor activity	6	4,66E-02	1,43
GO:0065003	macromolecule complex assembly	15	2,47E-04	1,68
GO:0006520	amino acid metabolic process	11	3,99E-04	1,75

Figure R.7.a. Functional enrichment (Gene Ontology) of NF-Y targets with at least one active histone mark.

NF-Y siRNA Microarray - DOWN

GO.ID	*Gene Ontology Category*	*# genes*	*P-value*	*Enrich ratio*
GO:0007049	**cell cycle**	60	4.65E-22	4.2
GO:0000279	M phase	31	1.31E-19	7.6
GO:0005694	chromosome	30	2.15E-15	5.7
GO:0006260	DNA replication	23	2.24E-14	7.0
GO:0006259	**DNA metabolism**	44	2.47E-14	3.7
GO:0006281	DNA repair	19	1.30E-08	4.4
GO:0006974	**resp. to DNA damage stimulus**	23	2.35E-10	4.6
GO:0015630	**microtubule cytoskeleton**	25	8.89E-12	4.9
GO:0005856	cytoskeleton	41	1.16E-08	2.6
GO:0046907	**intracellular transport**	37	1.69E-10	3.2
GO:0015031	protein transport	31	9.01E-08	2.8
GO:0005783	endoplasmic reticulum	26	2.00E-05	2.4
GO:0048193	Golgi vesicle transport	8	3.14E-04	3.7
GO:0006457	**protein folding**	12	3.35E-04	2.9
GO:0051082	unfolded protein binding	10	3.89E-04	3.1
GO:0005739	**mitochondrion**	28	2.61E-05	2.2
GO:0009055	electron carrier activity	12	5.47E-05	3.4
GO:0016071	mRNA metabolism	15	4.34E-05	3.0
GO:0006396	**RNA processing**	18	5.33E-04	2.2

Figure R.7.b. Functional enrichment (Gene Ontology) of downregulated genes after a NF-Y interference (siRNA).

NF-Y+ & Active Histone Marks- Transcriptional Units

GO.ID	*Gene Ontology Category*	*# genes*	*P-Value*	*Enrich. Ratio*
GO:0007600	**sensory perception**	5	2,55E-02	2,05
GO:0065003	**macromolecule complex assembly**	4	5,49E-02	1,83
GO:0007155	**cell adhesion**	5	6,21E-02	1,69
GO:0006955	**immune response**	4	8,39E-02	1,64
GO:0006915	**apoptosis**	5	8,83E-02	1,56

Figure R.7.c. Functional enrichment (Gene Ontology) of NF-Y targets with both negative histone marks.

Altogether, these data indicate that NF-Y binding is associated with two different chromatin states and opposite functional outcomes: in the presence of the two active histone marks, NF-Y is bound to expressed loci; binding within areas devoid of these modifications is coupled to inactive genes.

NF-Y Associates with Chromatin with Negative Histone Marks

We further investigated the cluster of nonexpressed NFY+/H3K9-14ac-/H3K4me3-TUs. To assay the possibility that negative histone marks were present in these areas, we performed ChIP experiments with antibodies directed against H3K9me3, H3K27me3, and H4K20me3, which have been associated, in various terms, with inactive or partially inactive chromatin environments [19]. As shown in figure R.8, with one exception, these loci were confirmed to be NF-Y positive, most of them with a low or medium enrichment level compared to genes targeted by NF-Y and transcribed, such as the SON gene (figure R.8, top).
Interestingly, all 12 tested sites scored positive for H4K20me3, as did 7 out of 12 for H3K27me3; among these SUHW1, encoding the human homologue of Drosophila Suppressor of Hairy-wing insulator binding protein, and the vitamin D receptor-activated CYP24A1; only three sites, APOBEC3, FTCD, and C21orf81, were H3K9me3+. Reassuringly, with just one exception, ChIPs confirmed negativity for the active histone mark

H3K4me3 (not shown). These results are consistent with the idea that NF-Y can be associated in vivo with areas of the genome containing negative histone marks.

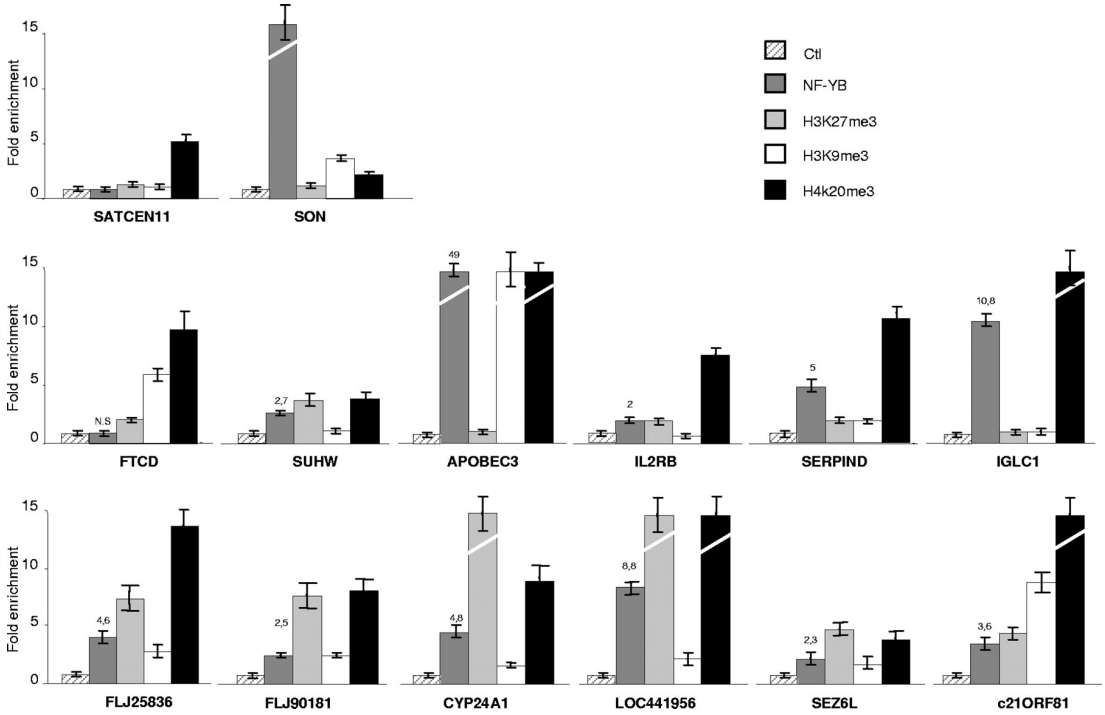


Figure R.8. ChIP with negative histone marks in nonexpressed NF-Y+ loci. Twelve NF-Y+/H3K9-K14ac-/H3K4me3- sites, randomly selected from the list of targets not expressed in HeLa cells were analyzed by ChIP with anti-NF-YB, anti-H3K27me3, anti-H3K9me3, and anti-H4K20me3 antibodies. Enrichment over that for an irrelevant Flag control antibody was assessed by semiquantitative PCR in duplicate experiments; enrichments greater than twofold were considered significant. The chromosome 11 satellite centromeric region (SATCEN11) and the promoter of the transcribed NF-Y target SON were used as internal controls (top row).

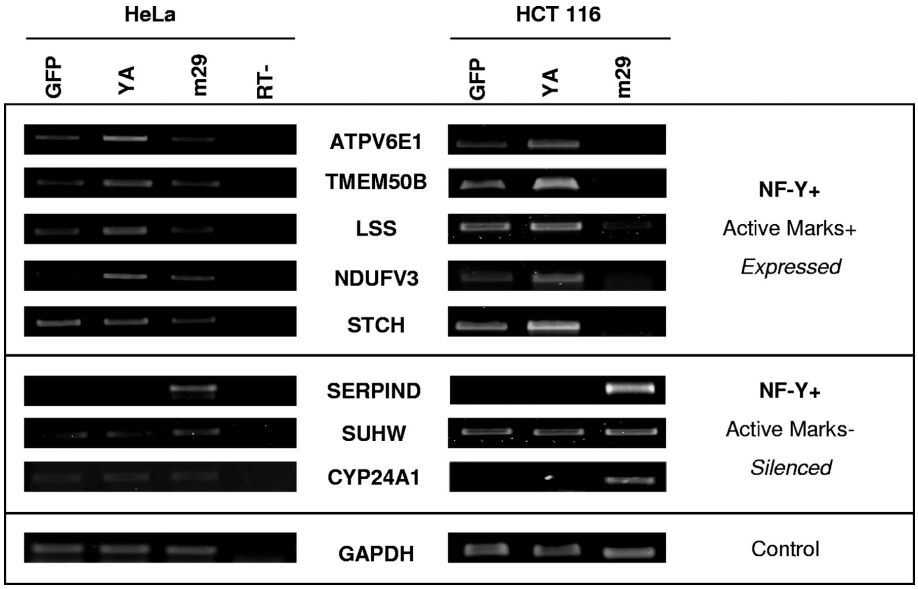


Figure R.9. Effects of NF-Y removal on gene expression. HeLa cells (left) were infected in parallel with control Ad-GFP, wild-type Ad-NF-YA, and the dominant negative Ad-YAm29 adenovirus. RT-PCR analyses of infected cells were performed in the linear range of amplification for the indicated genes. HCT116 cells (right) were analyzed under the same conditions.

NF-Y is an activator as well as a repressor

The binding of a TF to a target DNA sequence does not necessarily imply direct transcriptional regulation. To assay the role of NF-Y binding in gene expression, we infected HeLa and HCT116 cells with adenoviral vectors coding for an NF-YA dominant negative protein (Ad-YA-m29) containing a mutation in the DNA binding domain; the mutant still associates with the histone fold NF-YB/NF-YC dimer but renders the trimer incapable of CCAAT association ([38] and references therein). In parallel, we infected cells with wild-type Ad-NF-YA and Ad-GFP viruses. Under these conditions, expression of CCAAT-containing, NF-Y-dependent promoters was previously shown to be crippled [16]. Figure R.9 shows RT-PCR analysis of expressed NF-Y targets associated with positive histone marks (top); dominant negative YA treatment led to decrease in expression of the genes analyzed, while little effects were observed in the control wild-type Ad-GFP or Ad-NF-YA infections. The mRNA levels of control, NF-Y-independent genes such as GAPDH were unaffected. Similar results were obtained with the HCT116 cells (Fig. R.9, top right). In parallel, we also analyzed NF-Y loci not expressed and associated with negative histone marks; some of these genes, notably CYP24A1 and SERPIND, were clearly and specifically up-regulated by the YA-m29 treatment, both in HeLa and in HCT116 cells.
Note that the former had high levels of H3K27me3 and low H4K20me3; SERPIND showed the opposite pattern. SUHW1, which was low on both these marks, was very modestly induced by YA-m29. Interestingly, all these genes contain prototypical CCAAT boxes in their promoters. We conclude that the presence of an active NF-Y trimer is required both for the expression of active genes and for repression of some of the NF-Y targets associated with negative histone marks.

Discussion

Summary

In this work, we correlated the in vivo binding of NF-Y with the presence of H3-K9-14 acetylations and H3-K4 trimethylation using high-density arrays covering chromosome 21 and large parts of chromosomes 20 and 22. We came to the following relevant conclusions. A majority of genes analyzed here are targeted by NF-Y, either in the promoter or in the 5' end region of the gene. Functionally, two types of NF-Y loci exist: one has active histone marks and correlates with transcription, and the other is found in sites that are transcriptionally silent and loaded with negative histone marks. NF-Y serves as an activator of the former and as a repressor of the latter.

NF-Y Binding Location and Motif Specificity

Several bioinformatic studies have found CCAAT boxes in > 60% of human promoters [3, 5-7]; specific functional groups, such as those involving cell cycle and endoplasmic reticulum stress-regulated genes, were found to be quite enriched for CCAAT-containing genes [4, 8, 43]. Interestingly, analysis of microRNA control sequences identifies NF-Y as one of the few TFs involved in the regulation of all of them [10].
The first issue tackled here is how frequent are NF-Y binding sites. The answer to this depends, to some extent, upon the stringency of the analysis. Monitoring 25 promoters and validations of additional loci helped us fine-tune our analysis, so that we “lose” only 10 to 15% of the positive locations, keeping the false-positive sites at the same level. The two extremes range from 757 to 2,120 sites; the former is an absolute minimum that considers only very-high-affinity sites, the latter is a wider look at low-affinity sites or at sites present only in a subcategory of cells, as ChIPs were made from a heterogeneous population. Extrapolation of these numbers to the whole genome brings NF-Y sites to 35,000 to 80,000, which is high compared to extrapolations performed for other TFs such as E2F1 (20,000 to 30,000 binding sites [40]). However, it should be noted that there are at least 200,000 NF-YB molecules within the nuclei of various cell

lines (R. Mantovani, unpublished data). These data are consistent with the idea that NF-Y is involved in a wide range of RNA production procedures, both positive and negative; indeed, the TUs that contained at least one NF-Y site are a majority, 519 out of 907. However, it should be noted that there are at least 200,000 NF-YB molecules within the nuclei of various cell lines (R. Mantovani, unpublished data). These data are consistent with the idea that NF-Y is involved in a wide range of RNA production procedures, both positive and negative; indeed, the TUs that contained at least one NF-Y site are a majority, 519 out of 907.

The second issue tackled by this work concerns the presence of the CCAAT consensus in the identified NF-Y sites. Previous location analysis experiments performed with similar platforms reported a low rate of consensus sites near locations of TFs [40, 44-45], while others suggested significant variation from the in vitro-derived consensus [46]. It was found, using stringent criteria, that about 60% of NF-Y binding sites contain a CCAAT consensus, particularly promoter sites. Furthermore, 1.5 to 2.5 CCAAT boxes per locus are found; thus, the presence of one or more CCAAT boxes is in general important for NF-Y binding, in agreement with biochemical in vitro work [14]. Note that the total number of CCAAT-containing promoters in the cluster analyzed here is 463 out of 907 (51%), which is lower than the 60 to 67% derived from other studies [3]. Nonetheless, around 35% of the identified NF-Y loci are apparently devoid of CCAAT boxes. There are explanations for this finding. (i) Variation of a single nucleotide in the CCAAT pentanucleotide was reported to be compatible with NF-Y binding, especially when other functionally important overlapping sites are involved [47-48]. A degenerate CCAAT box would be missed in our stringent analysis. (ii) The ChIP-on-chip experiments were performed with an anti-NF-YB antibody; since NF-YB and NF-YC subunits are in excess with respect to NF-YA (Mantovani, unpublished), some CCAAT-less sites devoid of the sequence-specific NF-YA might have emerged.

The last point raised by our analysis is related to the genomic location of the identified NF-Y binding sites. The CCAAT box has long been considered almost exclusively as a promoter element crucial for Pol II recruitment [49]. In keeping with expectations [12], the vast majority of NF-Y PR sites, 70%, are positioned near the TSS, and there is clear enrichment of sites near the TSS also within the GE cohort (figure R.2.a). However, NF-Y locations appear to be more scattered, and roughly 50% of NF-Y-positive promoters contain an extra site within the body of the gene (figure R.2.b). It is possible that their role in many such cases is related to promoter-enhancer connections, while other GE locations could be alternative promoters of the same gene or sites of divergent RNA production. Moreover, within the EL cohort, a large number of sites are located in mRNA expressing areas, sites of Pol II activity. This accounts for 20% of the 909 EL peaks (table RT.1 and figure R.2.b). If we apply to these Pol II/NF-Y+ units the same ratio of 2 sites/unit, as measured in bona fide NF-Y RefSeq GE sites, at least 100 additional Pol II units would be added. Extrapolating to the whole genome, we estimate 4,500 NF-Y+ units outside of the annotated RefSeq genes. The remaining EL peaks could either be enhancer sequences located at a distance from transcribed regions or TUs whose RNAs have not been annotated yet.

NF-Y: Activator and Repressor Role

It is clear that H3K4me3 acetylation and H3K9-14 acetylation are hallmarks of active genes [23, 27-28]. Our data for HeLa cells confirm previous results for HepG2, obtained using similar criteria of analysis [27]. First of all we found an excellent correlation between active mark islands and gene expression in HeLa cells (Figure R.3.a-b). Second, as the overlap between expressed and not expressed genes in HeLa and HepG2 cells is quite high (85%), we found a remarkable degree of coincidence between histone mark sites in the two cell lines, with deviations only for H3K4me3 within genes, an effect due to cell type-specific patterns and/or to our underestimation of such a category (Figure R.3.a-b). A similar picture emerged also by comparison with the data

set of Guenther et al. [32], obtained with human ES cells [A1: File S5I to K]. Skewing toward expression was less evident for H3K4me3-positive promoters in ES cells, in agreement with previous observations [29].
NF-Y binding has been so far mostly associated with two possible transcriptional states: an actively transcribing gene and a poised promoter, with a pre-bound NF-Y ready to help recruiting whichever TF is specifically responsible for full activation following a stimulus. Indeed, the majority of NF-Y-bound units contains H3K9-14ac and H3K4me3 sites and is expressed. Note that active genes are expected to be readily discernible in Affymetrix profiling data, while the levels of the inducible ones might not, thus underscoring this already significant correlation.
It has long been known that the binding of TFs and cofactors to promoters is a hallmark of expression, by signaling to the Pol II transcription machinery the positional coordinates [2]. It would seem logical, therefore, to postulate that positive histone marks are positioned according to a code determined by the sequence-specific TFs. Reassuringly, several studies have confirmed that TF binding indeed correlates with the presence of "active" marks and with gene expression. In a thorough study performed with quantitative PCR on the correlation between > 30 histone modifications and MYC sites, Guccione et al. [50] concluded that the presence of
H3K4me2/3, H3K9-14ac, and H3K79me2 is a prerequisite for MYC binding. Interestingly, elimination of MYC had little effect on the levels of these modifications, unlike what was found for H4 acetylations, which were decreased [50]. Clearly, this is not the case for NF-Y, since our unbiased analysis identified NF-Y regions devoid of these active histone marks.
The most surprising result is the identification of the discrete cohort of NF-Y+ TUs in which H3K9-14ac and H3K4me3 islands are absent and specifically enriched in A calls (almost 80%); this indicates that NF-Y is not always associated with active/inducible transcription. This cluster comprises mostly tissue-specific and developmentally regulated genes (figure R.7.a-c) and is associated with repressive histone marks such as H4K20me3 and H3K27me3. While hints at a negative role for NF-Y in transcription had previously been reported [38, 51-53], we are intrigued by the extent of this phenomenon, which involves 20 to 25% of the total binding sites. Significantly, we find that removal of the NF-Y trimer leads to activation of these loci. Among the de-repressed, CYP24A1, but not SERPIND, is H3K27me3+. It is known that this modification is associated with Polycomb [37]; hints to a possible mechanism of repression through deposition of this negative mark come from recent experiments with *Caenorhabditis elegans*, in which ceNF-Y function has been genetically linked to Polycomb through direct interaction with the ESC/E(Z) component [54].
The structural resemblance of NF-YB–NF-YC to H2A-H2B should be remembered when considering the bifunctional behavior of NF-Y. The first sign of an opening chromatin cluster, and one specifically required for H3K4me3, is monoubiquitination of H2B on K123 [55]; we noted that lysines are present in the corresponding region of the H2B-like NF-YB, and indeed NF-YB is monoubiquitinated (G. Donati and R. Mantovani, data not shown). On the other hand, H2A is monoubiquitinated by Polycomb components, and the functional significance of this modification is opposite to that of H2Bubiquitin, leading to repression of transcription [56]. Thus, H2A-ubiquitin could serve as a signal to recruit H3K27 methyltransferase.
Given the dual dominant NF-Y function uncovered here, we are tempted to speculate that a code of posttranslational modifications may exist for the histone-like NF-Y. The bivalent behavior of NF-Y and its independence from a specific pattern of activating versus repressive histone modifications might have important consequences for transcriptional regulation at bivalent loci that have been recently mapped in both ES and lineage-committed cells. At these sites, in fact, the activating H3K4me3 and the repressive H3K27me3 are present together on areas of the genome "poised" for alternative developmental fates [57-59]. Interestingly, the role of NF-Y in stem cells has been

highlighted by two recent studies: (i) NF-YA was shown to be required for maintenance of hematopoietic stem cells [60] and (ii) the CCAAT box was found to be specifically enriched in conserved regions of genes highly expressed in mouse and human ES cells, with NF-Y being required both for regulation of these elements and for cell survival [9]. Moreover, a switch in the two major isoforms of NF-YA was noticed upon differentiation, with the "short" form being highest in stem cells and decreasing in embryonic bodies. Note that this isoform is specifically required for stemness in the hematopoietic system [60]. These results, considered together with the widespread bifunctional role of NF-Y and its independence from specific methyl marks shown here in committed HeLa cells, suggest that NF-Y might be associated with bivalent sites in ES cells, possibly even regulating their positioning. Because of its structure and histone interactions [61], NF-Y would be ideal to maintain nucleosome-free areas, accommodate accessibility of nearby TFs, and recruit modifying complexes, positive or negative. This hypothesis and the cause-and-effect relationships between the positioning of histone marks and NF-Y binding will now be investigated with appropriate genetic experiments.

Acknowledgments

This work was a cooperative effort of the Mantovani Lab (Dr. Michele Ceribelli, Diletta Dolfini, Raffaella Gatta, Dr. Alessandra Maria Viganò, Prof. Roberto Mantovani), Daniele Merico and Prof. Giulio Pavesi.
Dr Michele Ceribelli was the lead person for the collection and organization of the experimental results, and significantly contributed to their interpretation. Diletta Dolfini curated the mapping of the NF-Y binding sites to the transcriptional units, and the categorization of NF-Y binding sites according to the genomic location; she also performed the experiments assessing the NF-Y target expression in different cell lines, and contributed to the result interpretation. Daniele Merico performed the correlation analysis between NF-Y binding sites, the positive histone modifications and the transcriptional state of the targets; he also performed the functional profiling of the targets (including the NF-Y knockdown experiment), and contributed to the result interpretation. Raffaella Gatta performed the experiments with the dominant negative NF-Y construct. Dr. Alessandra Maria Viganò contributed to the design of the experiments and to the result interpretation. Dr. Giulio Pavesi designed and developed the algorithm for ChIP-chip data analysis. Prof. Roberto Mantovani provided overall guidance for the design of the experiments and the interpretation of results.
The NF-YB knockdown experiment was performed by Prof. Carol Imbriano's group and collaborators; Daniele Merico contributed to the microarray analysis, by identifying and functionally profiling the differentially expressed genes.

Publications

[A1] Ceribelli M, Dolfini D, Merico D, Gatta R, Pavesi G, Vigano A, Mantovani R.
The histone like NF-Y is a bifunctional transcription factor.
Molecular Cell Biology (MCB); 2008 Mar; 28(6): 2047-58.

[A2] Benatti P, Basile V, Merico D, Fantoni LI, Tagliafico E, Imbriano C.
A balance between NF-Y and p53 governs the pro- and anti-apoptotic transcriptional response.
Nucleic Acids Research (NAR); 2008 Mar; 36(5): 1415-28

Reference

[1] Hampsey M.
Molecular Genetics of the RNA Polymerase II General Transcriptional Machinery
Microbiol Mol Biol Rev, June 1998, p. 465-503, Vol. 62, No. 2

[2] Thomas MC, Chiang CM.
The general transcription machinery and general cofactors.
Crit. Rev. Biochem. Mol. Biol. 2006 41:105–178.

[3] Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S.
Identification and characterization of the potential promoter regions of 1031 kinds of human genes.
Genome Res., 2001 11:677–684.
[4] Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y.
Genomewide in silico identification of transcriptional regulators controlling the cell cycle in human cells.
Genome Res., 2003 13:773–780.
[5] FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C. 2004.
Clustering of DNA sequences in human promoters.
Genome Res., 2004 14:1562–1574.
[6] Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D.
Statistical analysis of over-represented words in human promoter sequences.
Nucleic Acids Res., 2004 32:949–958.
[7] Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Kel AE, Arakawa T, Carninci P, Kawai J, Hayashizaki Y, Takagi T, Nakai K, Sugano S.
Large-scale collection and characterization of promoters of human and mouse genes.
In Silico Biol. 2004 4:429–444.
[8] Zhu Z, Shendure J, Church GM.
Discovering functional transcription-factor combinations in the human cell cycle.
Genome Res., 2005 15:848–855.
[9] Grskovic M, Chaivorapol C, Gaspar-Maia A, Li H, Ramalho-Santos M.
Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells.
PLoS Genet., 2007 3:e145
[10] Lee J, Li Z, Brower-Sinning R, John B.
Regulatory circuit of human microRNA biogenesis.
PLoS Comput. Biol., 2007 3:e67.
[11] Mantovani R.
The molecular biology of the CCAAT-binding factor NF-Y.
Gene, 1999 239:15–27.
[12] Mantovani R.
A survey of 178 NF-Y binding CCAAT boxes.
Nucleic Acids Res. 1998, 26:1135–1143.
[13] Dolfini D, Mantovani R, unpublished data.
[14] Dorn A, Bollekens J, Staub A, Benoist C, Mathis D.
A multiplicity of CCAAT binding proteins.
Cell, 1987 50:863–872.
[15] Romier C, Cocchiarella F, Mantovani R, Moras D.
The crystal structure of the NF-YB/NF-YC heterodimer gives insight into transcription regulation and DNA binding and bending by transcription factor NF-Y.
J. Biol. Chem., 2003 278:1336–1345.
[16] Testa A, Donati G, Yan P, Romani F, Huang TH, Vigano MA, Mantovani R.
ChIP on chip experiments uncover a widespread distribution of NF-Y binding CCAAT sites outside of core promoters.
J. Biol. Chem., 2005 280:13606–13615.
[17] Berger SL.
The complex language of chromatin regulation during transcription.
Nature, 2007 447:407–412.
[18] Li B, Carey M, Workman JL.
The role of chromatin during transcription.
Cell. 2007 Feb 23;128(4):707-19.
[19] Bernstein BE, Meissner A, Lander ES.
The mammalian epigenome.
Cell 2007 128:669–681.
[20] Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T.
Active genes are tri-methylated at K4 of histone H3.
Nature, 2002 419:407–411.

[21] Kurdistani SK, Tavazoie S, Grunstein M.
Mapping global histone acetylation patterns to gene expression.
Cell, 2004 117:721–733.
[22] Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T.
Histone H3 lysine 4 methylation patterns in higher eukaryotic genes.
Nat. Cell Biol., 2004 6:73–77.
[23] Ruthenburg AJ, Allis CD, Wysocka J.
Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark.
Mol. Cell, 2007 25:15–30.
[24] Liang G, Lin JC, Wei V, Yoo C, Cheng JC, Nguyen CT, Weisenberger DJ, Egger G, Takai D, Gonzales FA, Jones PA.
Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome.
Proc. Natl. Acad. Sci., 2004 101: 7357–7362.
[25] Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ.
Single-nucleosome mapping of histone modifications in S. cerevisiae.
PLoS Biol., 2005 3:e328.
[26] Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA.
Genome-wide map of nucleosome acetylation and methylation in yeast.
Cell, 2005 122:517–527.
[27] Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ III, Gingeras TR, Schreiber SL, Lander ES.
Genomic maps and comparative analysis of histone modifications in human and mouse.
Cell, 2005 120:169–181.
[28] Roh TY, Cuddapah S, Zhao K.
Active chromatin domains are defined by acetylation islands revealed by genomewide mapping.
Genes Dev., 2005 19:542–552.
[29] Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, and Ren B.
A high-resolution map of active promoters in the human genome.
Nature, 205 436:876–880.
[30] Roh TY, Cuddapah S, Cui K, Zhao K.
The genomic landscape of histone modifications in human T cells.
Proc. Natl. Acad. Sci. USA, 2006 103:15782–15787.
[31] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, I. Chepelev, and K. Zhao. 2007.
High-resolution profiling of histone methylations in the human genome.
Cell, 2007 129:823–837.
[32] Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA.
A chromatin landmark and transcription initiation at most promoters in human cells.
Cell, 2007 130:77–88.
[33] Baumeister P, Luo S, Skarnes WC, Sui G, Seto E, Shi Y, Lee AS.
Endoplasmic reticulum stress induction of the Grp78/BiP promoter: activating mechanisms mediated by YY1 and its interactive chromatin modifiers.
Mol. Cell. Biol., 2005 25:4529–4540.
[34] Donati G, Imbriano C, Mantovani R.
Dynamic recruitment of transcription factors and epigenetic changes on the ER stress response gene promoters.
Nucleic Acids Res., 2006 34:3116–3127.
[35] Fang J, Feng Q, Ketel CS, Wang H, Cao R, Xia L, Erdjument-Bromage H, Tempst P, Simon JA, Zhang Y.
Purification and functional characterization of SET8, a nucleosomal histone H4-lysine 20-specific methyltransferase.
Curr. Biol., 2002 12:1086–1099.
[36] Nishioka K, Rice JC, Sarma K, Erdjument-Bromage H, Werner J, Wang Y, Chuikov S, Valenzuela P, Tempst P, Steward R, Lis JT, Allis CD, Reinberg D. PR-Set7 is a nucleosome-specific methyltransferase that modifies lysine 20 of histone H4 and is associated with silent chromatin.

Mol. Cell, 2002 9:1201–1213.
[37] Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, Jang SW, Margueron R, Reinberg D, Green R, Farnham PJ.
Suz12 binds to silenced regions of the genome in a cell-type-specific manner.
Genome Res., 2006 16:890–900
[38] Imbriano C, Gurtner A, Cocchiarella F, Di Agostino S, Basile V, Gostissa M, Dobbelstein M, Del Sal G, Piaggio G, Mantovani R.
Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters.
Mol. Cell. Biol. 2005 25:3737–3751.
[39] Ceribelli M., Alcalay M, Vigano MA, Mantovani R.
Repression of new p53 targets revealed by ChIP on chip experiments.
Cell Cycle 2006 5:1102–1110.
[40] Bieda M, Xu X, Singer M, Green R, Farnham PJ.
Unbiased location analysis of E2F1 binding sites suggests a widespread role for E2F1 in the human genome.
Genome Res. 2006 16:595–605.
[41] Viganò MA, Lamartine J, Testoni B, Merico D, Alotto D, Castagnoli C, Robert A, Candi E, Melino G, Gidrol X, Mantovani R.
New p63 targets in keratinocytes identified by a genome-wide approach.
EMBO J. 2006 25:5105–5116.
[42] Sato N, Sanjuan IM, Heke M, Uchida M, Naef F, Brivanlou AH.
Molecular signature of human embryonic stem cells and its comparison with the mouse.
Dev. Biol. 2003 260:404–413.
[43] Tabach Y, Milyavsky M, Shats I, Brosh R, Zuk O, Yitzhaky A, Mantovani R, Domany E, Rotter V, Pilpel Y.
The promoters of human cell cycle genes integrate signals from two tumor suppressive pathways during cellular transformation.
Mol. Syst. Biol. 2005 1:E1–E15.
[44] Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M.
Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1.
Cell 2005 122:33–43.
[45] Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, Weissman S, Snyder M.
CREB binds to multiple loci on human chromosome 22.
Mol. Cell. Biol. 2004 24:3804–3814.
[46] Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y.
A global map of p53 transcription-factor binding sites in the human genome.
Cell 2006 124:207–219.
[47] Gilthorpe, J., M. Vandromme, T. Brend, A. Gutman, D. Summerbell, N. Totty, and P. W. Rigby.
Spatially specific expression of Hoxb4 is dependent on the ubiquitous transcription factor NFY.
Development 2002 129:3887–3899.
[48] Milos, P. M., and K. S. Zaret.
A ubiquitous factor is required for C/EBP-related proteins to form stable transcription complexes on an albumin promoter segment in vitro.
Genes Dev, 1992 6:991–1004.
[49] Kabe Y, Yamada J, Uga H, Yamaguchi Y, Wada T, Handa H.
NF-Y is essential for the recruitment of RNA polymerase II and inducible transcription of several CCAAT box-containing genes.
Mol. Cell. Biol. 2005 25:512–522
[50] Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall' Olio V, Zardo G, Nervi C, Bernard L, Amati B.
Myc-binding-site recognition in the human genome is determined by chromatin context.
Nat. Cell Biol. 2006 8:764–770.

[51] Manni I, Mazzero G, Gurtner A, Mantovani R, Haugwitz U, Krause K, Engeland K, Sacchi A, Soddu S, Piaggio G.
NF-Y mediates the transcriptional inhibition of the cyclin B1, cyclin B2, and Cdc25C promoters upon induced G2 arrest.
J. Biol. Chem. 2001 276:5570–5576.
[52] Uramoto H, Wetterskog D, Hackzell A, Matsumoto Y, Funa K.
p73 competes with co-activators and recruits histone deacetylase to NF-Y in the repression of PDGF beta-receptor.
J. Cell Sci. 2004 117:5323–5331.
[53] Bernadt CT, Nowling T, Wiebe MS, Rizzino A.
NF-Y behaves as a bifunctional transcription factor that can stimulate or repress the FGF-4 promoter in an enhancer-dependent manner.
Gene Expr. 2005 12:193–212.
[54] Deng H, Sun Y, Zhang Y, Luo X, Hou W, Yan L, Chen Y, Tian E, Han J, Zhang H.
Transcription factor NFY globally represses the expression of the C. elegans Hox gene Abdominal-B homolog egl-5.
Dev. Biol. 2007 308:583–592.
[55] Laribee RN, Fuchs SM, Strahl BD.
H2B ubiquitylation in transcriptional control: a FACT-finding mission.
Genes Dev. 2007 21:737–743.
[56] Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P, Jones RS, Zhang Y.
Role of histone H2A ubiquitination in Polycomb silencing.
Nature 2004 431:873–878.
[57] Azuara, V., P. Perry, S. Sauer, M. Spivakov, H. F. Jorgensen, R. M. John, M. Gouti, M. Casanova, G. Warnes, M. Merkenschlager, and A. G. Fisher.
Chromatin signatures of pluripotent cell lines.
Nat. Cell Biol. 2006 8:532–538.
[58] Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R,
Schreiber SL, Lander ES.
A bivalent chromatin structure marks key developmental genes in embryonic stem cells.
Cell 2006 125:315–326.
[59] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE.
Genomewide maps of chromatin state in pluripotent and lineage-committed cells.
Nature 2007 448:553–560.
[60] Zhu J, Zhang Y, Joe GJ, Pompetti R, Emerson SG.
NF-YA activates multiple hematopoietic stem cell (HSC) regulatory genes and promotes HSC self-renewal.
Proc. Natl. Acad. Sci. 2005 USA 102:11728–11733.
[61] Caretti, G., M. C. Motta, and R. Mantovani.
NF-Y associates H3-H4 tetramers and octamers by multiple mechanisms.
Mol. Cell. Biol. 1999 19:8591–8603.

R9C Model Expression Profiling

Gene Expression Profiling
of the R9C Transgenic Mouse Model for Dilated Cardiomyopathy.

Abstract

Cardiomyopathies of diverse etiology impair cardiac muscle function and frequently progress to a convergence point where they induce heart dilatation and overt failure. Although heart failure is a major source of global morbidity and death in the developed world, afflicted patients are typically diagnosed with end stage disease when few effective avenues for restorative intervention remain and clinical outcomes are poor.
In order to elucidate the molecular basis of cardiomyopathy pathogenesis, we performed an extensive gene expression survey of cardiac ventricle samples, isolated from a transgenic mouse model of cardiomyopathy, which is characterized by the overexpression of the phospholamban (PLN) mutant R9C. PLN-R9C interferes with Ca^{++} cycling and cardiomyocyte relaxation, eventually leading to death by heart failure between 16 and 24 weeks of age.
In addition to traditional mRNA microarray profiling, we also profiled microRNA (miRNA) transcriptional patterns, as these non-coding RNAs play an important role as gene expression modulators. Samples were extracted at 8, 16 and 24 weeks of age, from transgenic mice and normal control littermates.
The analysis of global transcriptional patterns displayed a very high correlation with heart functionality indexes, confirming the validity of the study. Functional profiling enabled to generate a map of cellular processes differentially regulated in R9C-induced DCM. Activated processes included immune response and inflammation, apoptosis, cytoskeleton remodeling, extracellular matrix deposition, growth factor signaling and other regulatory pathways; inhibited process included oxidative metabolism, mitochondrial components and peroxisome activity/biogenesis. To further understand the regulatory pathways controlling the pathogenetic process, we identified differentially expressed miRNA whose predicted targets were enriched by differential transcripts (upregulated targets for down-regulated miRNA, and vice-versa). This analysis enabled to identify miR-1 as an outstanding regulator in DCM, a finding that was confirmed by comparing the transcriptome of R9C mice to miR-1 knockout mice.
Further study is being devoted to (1) a more comprehensive characterization of transcriptionally active regulatory pathways, (2) the identification of other miRNA involved in the DCM pathogenesis, and (3) the comparison of these results to other cardiomyopathy models (e.g. hypertrophic) and profiling data from human patients.

Background

Heart Failure and Dilated Cardiomyopathy

Heart failure (HF) is a condition in which a problem with the structure or function of the heart impairs its ability to supply sufficient blood flow to meet the body's needs. It is one of the most common, costly, disabling, and deadly medical conditions encountered by a wide range of physicians and surgeons in both primary and secondary care [1].
Cardiomyopathies of diverse etiology impair cardiac muscle function and frequently progress to a convergence point where they induce heart dilatation and overt failure. Dilated cardiomyopathy (DCM) is characterized by dilatation and impaired contraction of 1 or both ventricles in the absence of significant coronary artery disease. The incidence of DCM has been estimated to be 5 to 8 cases per 100,000 individuals, with a prevalence of 36 per 100,000 [2]. Thus, DCM is a leading cause of heart failure and cardiac transplantation in Western countries [3]. The high morbidity and mortality associated with DCM underscore the need for a better understanding of the underlying molecular events leading to heart failure in DCM.

The R9C-PLN Transgenic Mouse Model of Dilated Cardiomyopathy

Reduced contractile function and pathological remodeling are recognized clinical hallmarks of heart failure, but the critical early events that impair myocyte performance are largely undefined [4].
Intracellular Ca^{++} handling is the central coordinator of cardiac contraction and relaxation [5]. Contraction begins with sarcoplasmic reticulum (SR) release of Ca^{++} into the cytosol via the ryanodine receptor; relaxation occurs with SR Ca^{++} reuptake through the Ca^{++} adenosine triphosphatase (ATPase) SERCA2a pump. Phospholamban (PLN), an abundant, 52–amino acid transmembrane SR phosphoprotein [6], regulates the Ca^{++} ATPase SERCA2a. In particular, PLN acts as SERCA2a inhibitor, unless it is phosphorylated by Protein Kinase A (PKA).
A phospholamban mutation at nucleotide 25, which encodes an Arg 3 Cys substitution at codon 9 (R9C), causes inherited human dilated cardiomyopathy, with dominant effect; the onset of dilated cardiomyopathy in affected patients typically commences during adolescence followed by progressive deterioration in cardiac function leading to crisis and mortality [7]. A transgenic mouse model of this mutation (tagged throughout the text as R9C) showed a remarkably similar cardiac phenotype, with the afflicted mice presenting with early onset dilated cardiomyopathy characterized by decreased cardiac contractility and premature death [7]. Cellular and biochemical studies revealed that, unlike wild-type PLN, PLN-R9C did not directly inhibit SERCA2a. Rather, PLNR9C PKA, consequently blocking PKA-mediated phosphorylation of wild-type PLN, and preventing the modulation of its SERCA2a inhibitory activity [7].

microRNAs

MicroRNAs (miRNA) are single-stranded RNA molecules of about 21–23 nucleotides in length, involved in gene regulation. miRNAs are encoded by genes that are transcribed from DNA but not translated into protein (non-coding RNA, cf. figure B.1). miRNAs are transcribed to primary transcripts (pri-miRNA), usually by Pol-II, and fold into a typical stem-loop secondary structure; pri-miRNA are then cleaved, generating a shorter stem-loop structure (pre-miRNA), which is exported to the cytosol, and further cleaved to a short RNA duplex. miRNA enters the RISC (RNA-Induced Silencing Complex) as ssRNA, and binds a complementary sequence in the 3' UTR of the target-transcript they activate target transcript degradation (mediated by a RISC endonuclease), and/or inhibit its translation [8].

Since target-recognition is guided by sequence complementarity, different methods have been proposed to predict miRNA target transcripts (e.g, TargetScan [9]).

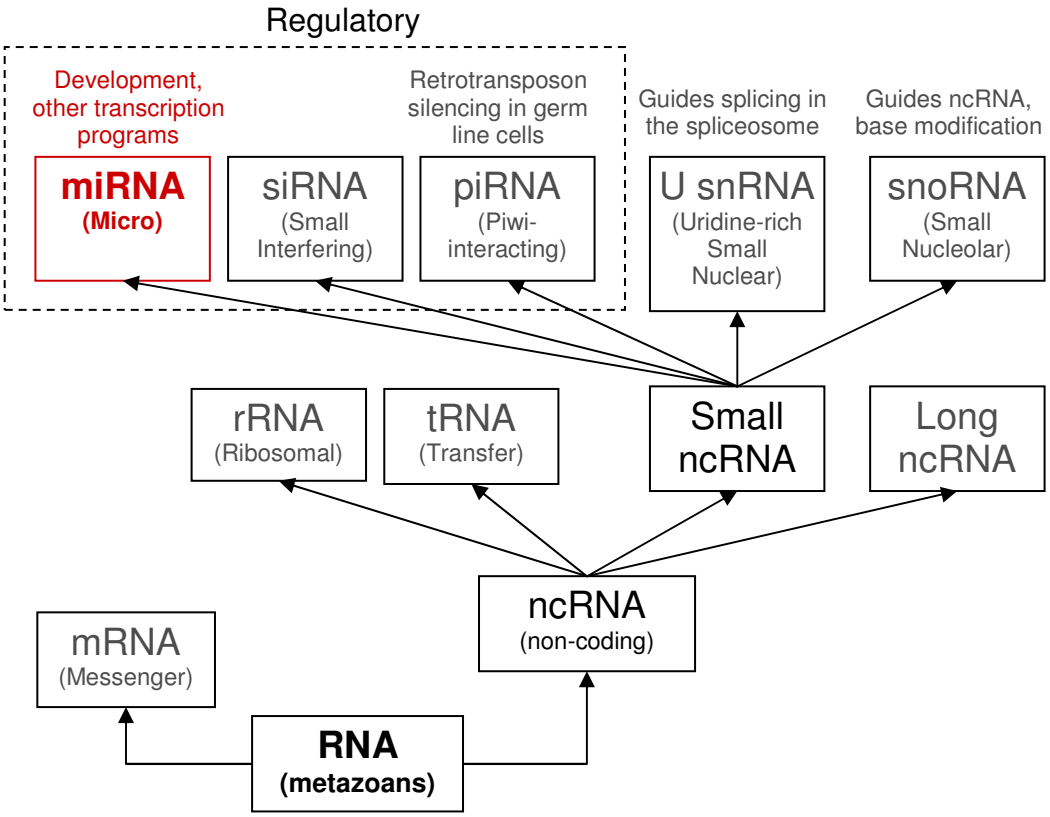


Figure B.1. A comprehensive classification of non-coding RNA types in metazoans.

Materials and Methods

Sample Preparation

Male and female mice were analyzed at 8, 16, and 24 weeks by M-mode and Doppler echocardiography for non-invasive assessment of left ventricular function and dimensions. Immediately prior to the preparation of cardiac tissue samples, at least six mice in each category were CO_2-asphyxiated, and the ventricle muscle was collected rapidly and rinsed in ice-cold PBS. For pathology and histological analyses, the hearts were washed extensively in ice-cold PBS and fixed immediately with ice-cold 4% paraformaldehyde in PBS. Cardiomyocytes were isolated, and intracellular Ca^{++} measurements were performed.

Microarray Technology and Data Analysis

Microarray-based global mRNA profiling experiments were performed using the Affymetrix Mouse 430 2.0 full-genome array chips. Raw image data were analyzed using the Affymetrix MAS5 software package; resulting .CEL files were then imported (using the affy R/Bioconductor [10] package) and processed using the rma algorithm, as implemented in the rma R/Bioconductor package; rma inlcudes a sample normalization step. miRNA expression profiles were generated using the Exiqon miRNA microarray platform. PCA (Principal Component Analysis) was performed using the Ade4 R package; the rows of the data matrix were normalized subtracting the mean and dividing by standard deviation, a common pre-processing step when performing PCA. miRNA were ranked using the ratio of the mean signal within each class (R9C-DCM 16 weeks, wild-

type 16 weeks); miRNA with one or more expression value not available across the replicates were discarded from the analysis. Kolomogorv-Smirnov tests were performed using R.

Functional Enrichment Map

The functional enrichment map was generated as follows. Gene Ontology (GO) functional enrichments were computed using the GSEA test [11], with Pearson correlation to fractional shortening as differentiality metric. GO was previously filtered to exclude terms with less than 10 or more than 500 annotated genes. Enriched terms were selected if they passed an uncorrected p-value threshold of 0.01 and an estimated False Discovery Rate lower than 10%. In order to visualize the terms in the functional enrichment map, an interaction network was first generated among gene ontology terms. Terms were regarded as set of genes, and the overlap among these sets was computed using a modified version of the Jaccard coefficient (Jc). Given two sets A and B, the Jc is usually computed as the ratio between the size of the intersection and the union of set A and B; since GO-derived gene sets typically have very different sizes, the size of the union was replaced by the size of the smallest set; only interactions with modified Jc larger than 0.5 were retained. The resulting network was visualized using Cytoscape [12], applying the organic layout. The network was additionally edited manually, to remove terms without a specific biological scope (e.g. Cellular Macromolecular Complex Subunit Organization), which were not previously removed by the size-dependent filter. Node color, accounting for term state (up-regulation, down-regulation) was assigned using GSEA uncorrected p-value.

Results

Transgenic Model Phenotype

We established a survival curve for the transgenic line in which 44 mice overexpressing the R9C transgene under control of the myosin heavy chain, cardiac-specific promoter and 79 littermate controls were analyzed. The PLN-R9C mice had a median survival of only 20 weeks with fewer than 15% persisting past 24 weeks. The first recorded deaths in the PLN-R9C line were observed between 12 and 16 weeks of age, whereas only one wild-type control mouse died over the entire 24-week period. For our subsequent gene expression analyses, 24 weeks was established as end stage human dilated cardiomyopathy due to the high mortality, 8 weeks was established as a time point representative of early stage disease prior to the first recorded mortality, and 16 weeks was established as a midpoint in disease progression. However, enlargement of both ventricle and atria was evident by 8 weeks of age in the PLN-R9C mice. Likewise cross-sections of the myocardium stained with hematoxylin and eosin showed thinning of the ventricular wall and evidence of left ventricular dilatation in the 8-week transgenic animals with continued progression of dilatation with age. High power magnification also indicated obvious regions of fat and connective tissue infiltration and muscle degeneration in the PLN-R9C hearts even at the 8-week time point (figure R.1).

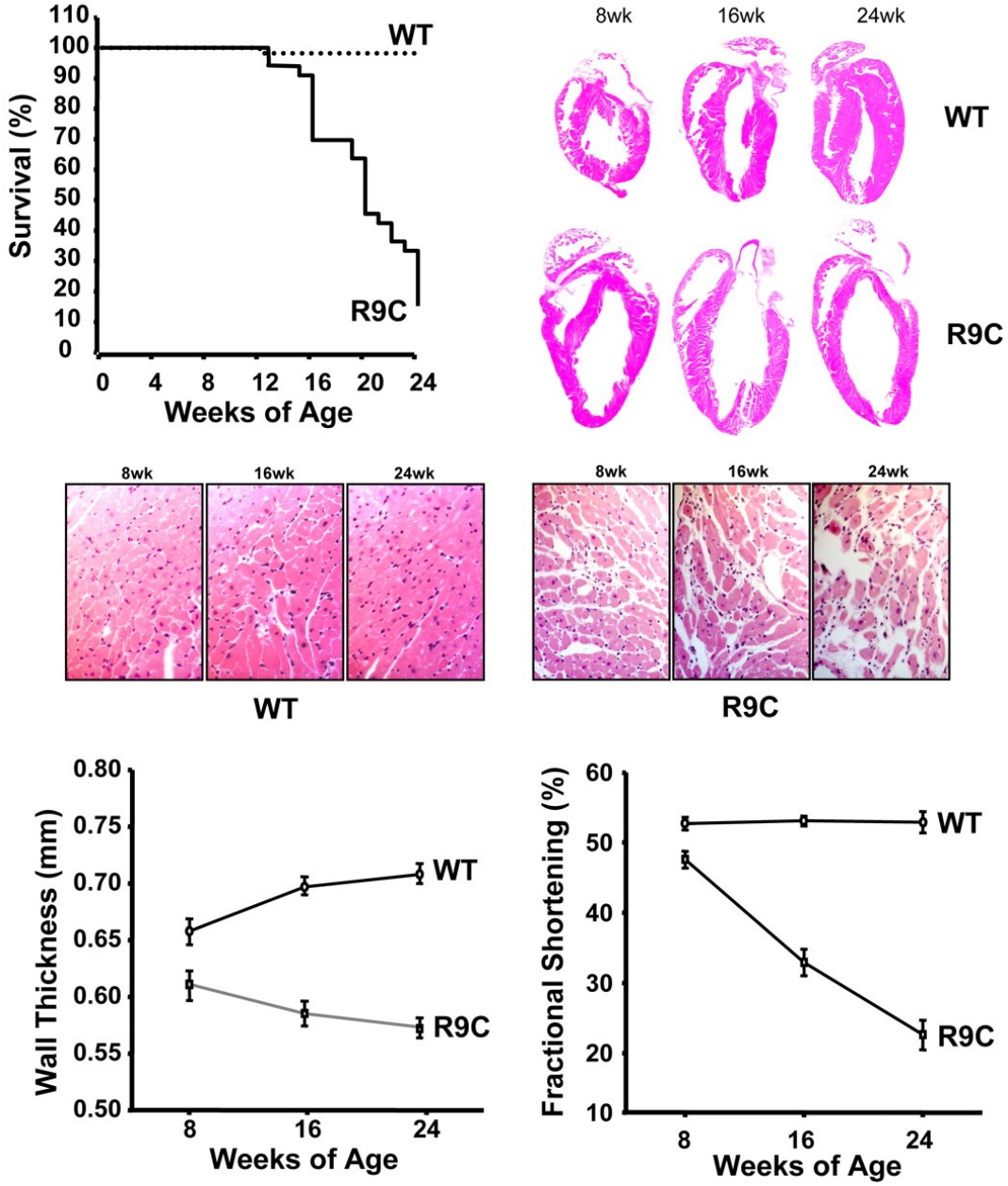


Figure R.1. Phenotypic analyses of wild-type and R9C mice. Top left: survival curves for wild-type and R9C mice. Top right: cardiac sections show significant cardiac enlargement in the R9C mice even at the earliest time point of 8 weeks of age. Middle: histological sections with higher power micrographs show evidence of cardiac disease from 8 weeks of age. Bottom left: measurement of anterior and posterior wall thickness in wild-type and R9C mice. Bottom right: cardiac shortening assessed by echocardiography; significant functional impairment in the R9C transgenic animals begins as early as 8 weeks of age.

Global Transcription Patterns

Transcripts were profiled at 8, 16 and 24 weeks for wild-type and transgenic animals. To understand the global behavior of gene expression, the expression data matrix was analyzed using Principal Component Analysis (PCA). PCA is a dimensionality reduction techniques that projects the data into a new space, whose dimensions are ranked according the amount of original data variation explained; its typical application to gene expression data consists in the projection of the samples into a new “meta-gene”

space, where dimensions correspond to distinct patterns of gene expression. As displayed by figure R.2, the first two Principal Components account for a relevant share of original variation (evaluated as the sum of their eigenvalues, which is larger than 40% of the total sum). PC-1 is characterized by particular biological significance, as it explains the disease progression: wild-type samples have negative values, whereas transgenic (R9C-DCM) samples have increasingly positive values from 8w to 16w, reaching saturation at 24w. Replicates display a very limited dispersion along PC-1, with the only exception of R9C.hd.w16.4. We confirmed the biological significance of PC-1 by computing the Pearson correlation between PC-1 sample coordinates and three quantitative phenotypes of DCM (Dilated Cardiomyopathy): wall thickness, fractional shortening and survival rate. Correlation was always larger than 0.8, and maximal for fractional shortening.

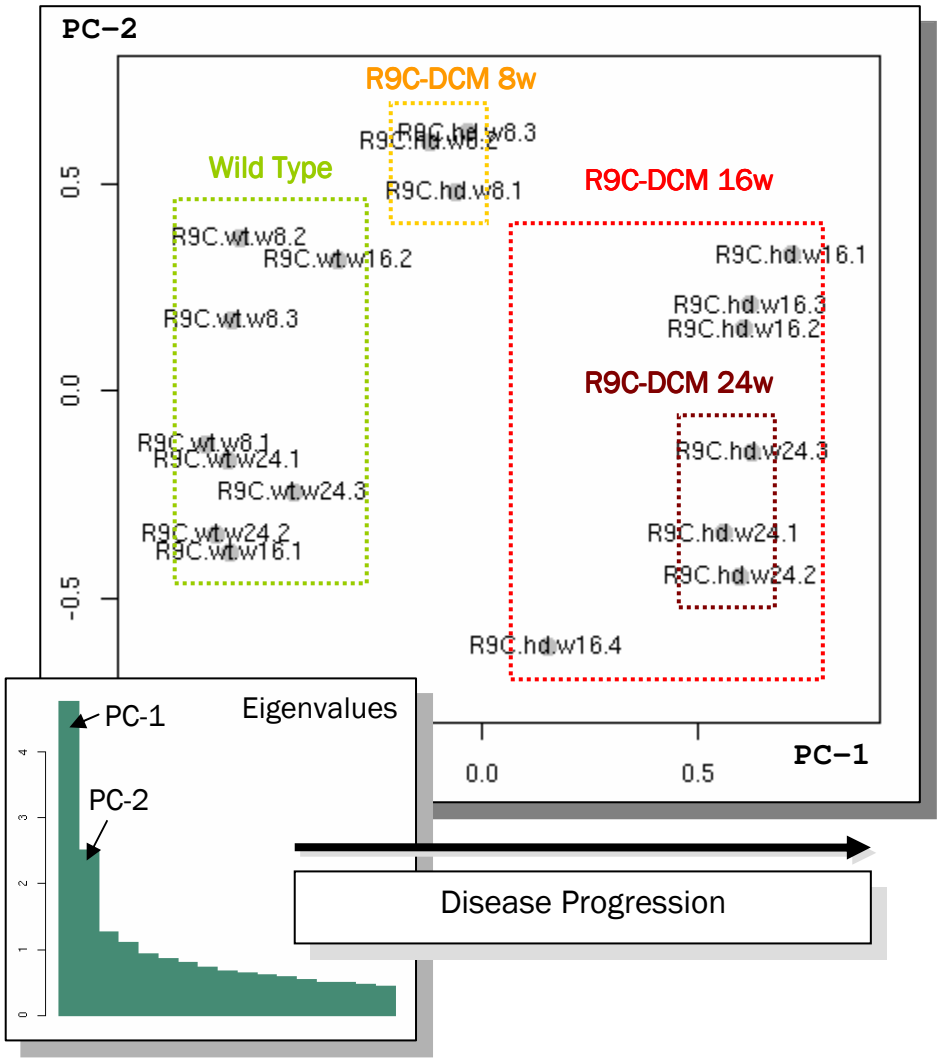


Figure R.2. Principal Component Analysis. The central diagram is a bi-plot, displaying the sample in the new "meta-gene" space; the coordinates are Principal Component (PC) 1 and 2. Profiled samples are displayed as grey dots with a text label specifying their identity; colored dotted box group samples belonging to the same class. The bottom-left diagram is the eigenvalue barplot; the intensity of each eigenvalue is related to the amount of original variation explained by each PC. PC 1 and 2 have cognate eigenvalues standing out of the distribution.

PC-2 is less straightforward to explain biologically. Apparently, it discriminates R9C-DCM 8 weeks from both wild-type and R9C-DCM 16-24 weeks; however, replicates show a

significant dispersion along PC-2, suggesting some caution in its interpretation. Preliminary evidence (data not shown) suggests that PC-2 may account for an initial stress response, occurring in R9C-DCM at 8 weeks, but then declining at 16 and 24 weeks. This hypothesis will be further investigated.

Functional Enrichment Map

To identify the processes differentially regulated in R9C-DCM we adopted a novel visualization approach, termed Functional Enrichment Map. First of all, we decided to focus our attention on the gene expression pattern displayed by the first principal component, as it strongly relates to disease phenotypes. An enrichment test was performed on Gene Ontology [13] terms, in order to select the ones enriched by genes highly correlated to fractional shortening (fractional shortening was specifically picked as it better relates to gene expression patterns). The selected terms were then re-organized in a network, termed Functional Enrichment Map, which captures term regulation state (up-regulation, down-regulation) and term inter-dependence. Term inter-dependence is specifically modeled considering the number of genes co-annotated under two different terms: if two terms annotate the same genes, or if the genes annotated by one term form a completely included subset of the genes annotated by the other term, the interaction score is one; more in general, the interaction score is computed using a modified version of the Jaccard coefficient (see Materials and Methods).

The Functional Enrichment Map clearly displays large clusters of highly inter-related terms, with the same activity state. Outstanding biological functions, previously related to Dilated Cardiomyopathy can be identified; for instance, Immune Response, Apoptosis, Actin Cytoskeleton Remodeling, and Extracellular Matrix / Collagen Biosynthesis are all up-regulated; Oxidative Metabolism is down-regulated. Additional biological functions offer promising insights into the regulatory pathways controlling DCM pathogenesis, such as Ras/Rho Signaling, Growth Factor Signaling and Embryonic Developmental Processes.

The global, manually annotated view of the Functional Enrichment Map is displayed in figure R.4.a. Local, magnified views of local areas of the map are provided by figures R.4.Q1-11, according to the break-down scheme displayed in figure R.4.b.

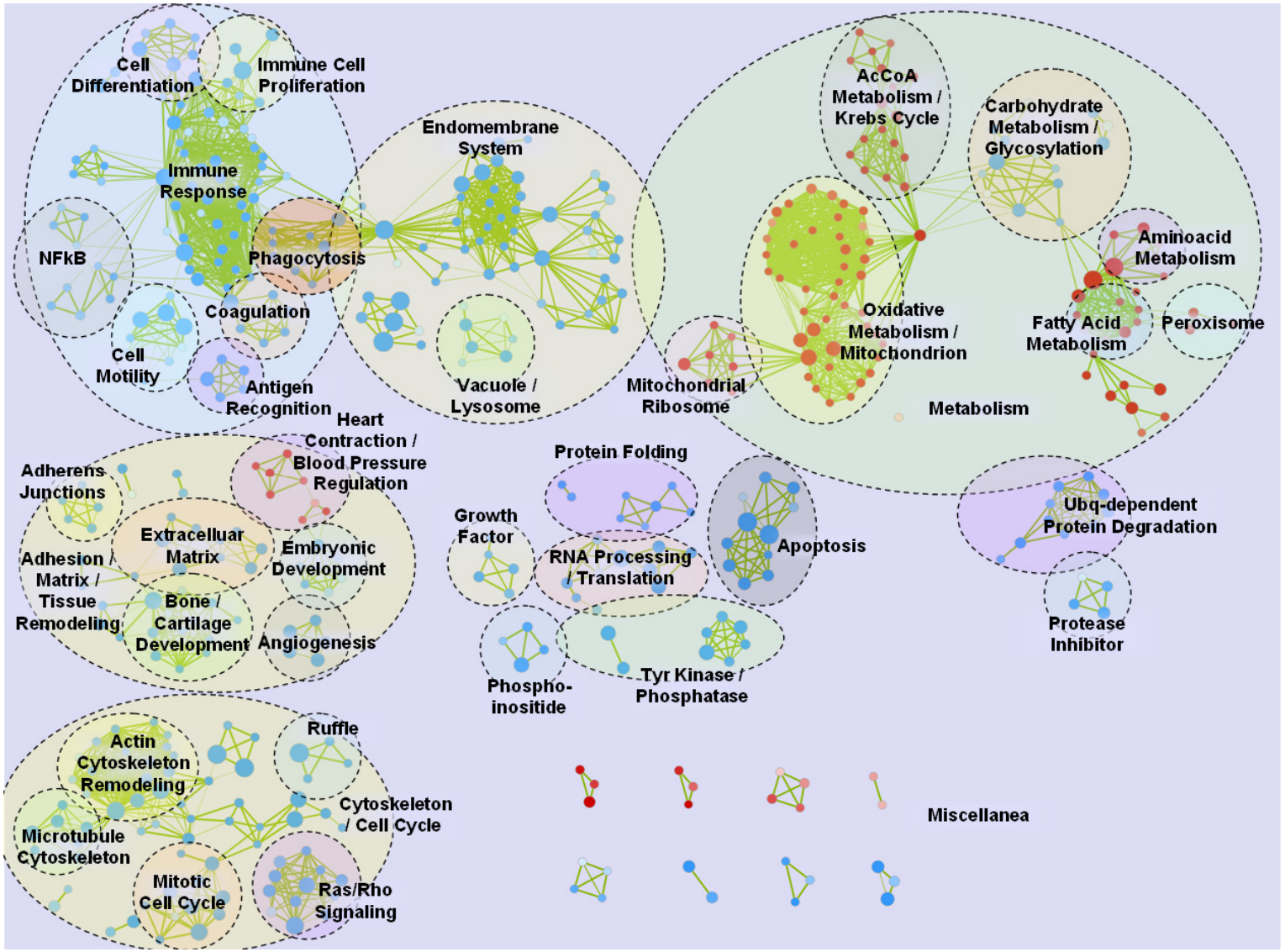


Figure R.4.a. The global view of the Functional Enrichment Map for R9C transcriptomics. Clusters of Gene Ontology terms are annotated according to the aggregate function. Blue corresponds to up-regulation and red to down-regulation. Interaction thickness scales with term inter-relatedness (modified Jc); node size scales with the number of genes annotated under the term.

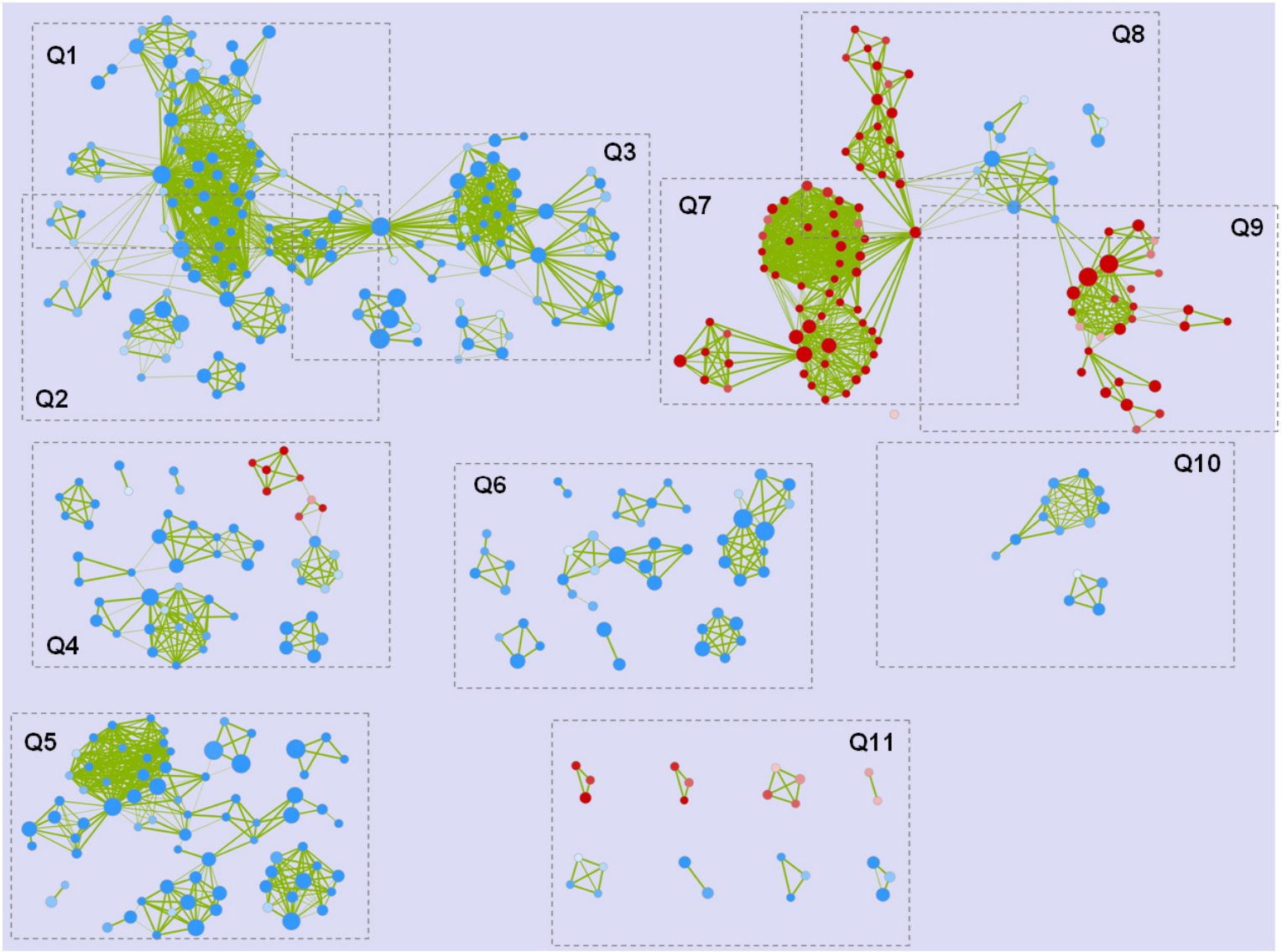


Figure R.4.b. The R9C Enrichment Map divided into areas, which are displayed in full magnification in the next pages.

Figure R.4.Q1. The Immune Response cluster, part 1.

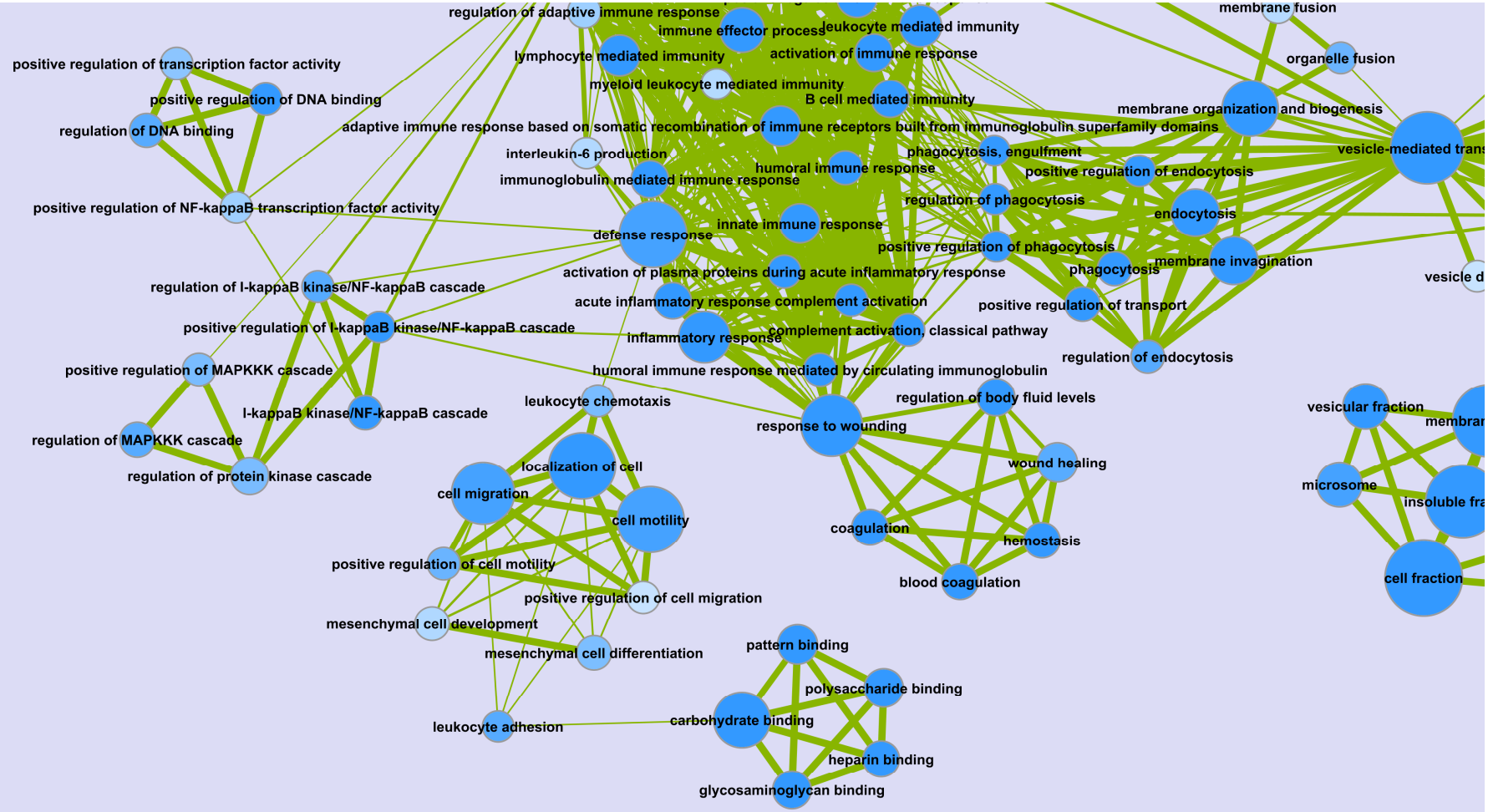


Figure R.4.Q2. The Immune Response cluster, part 2.

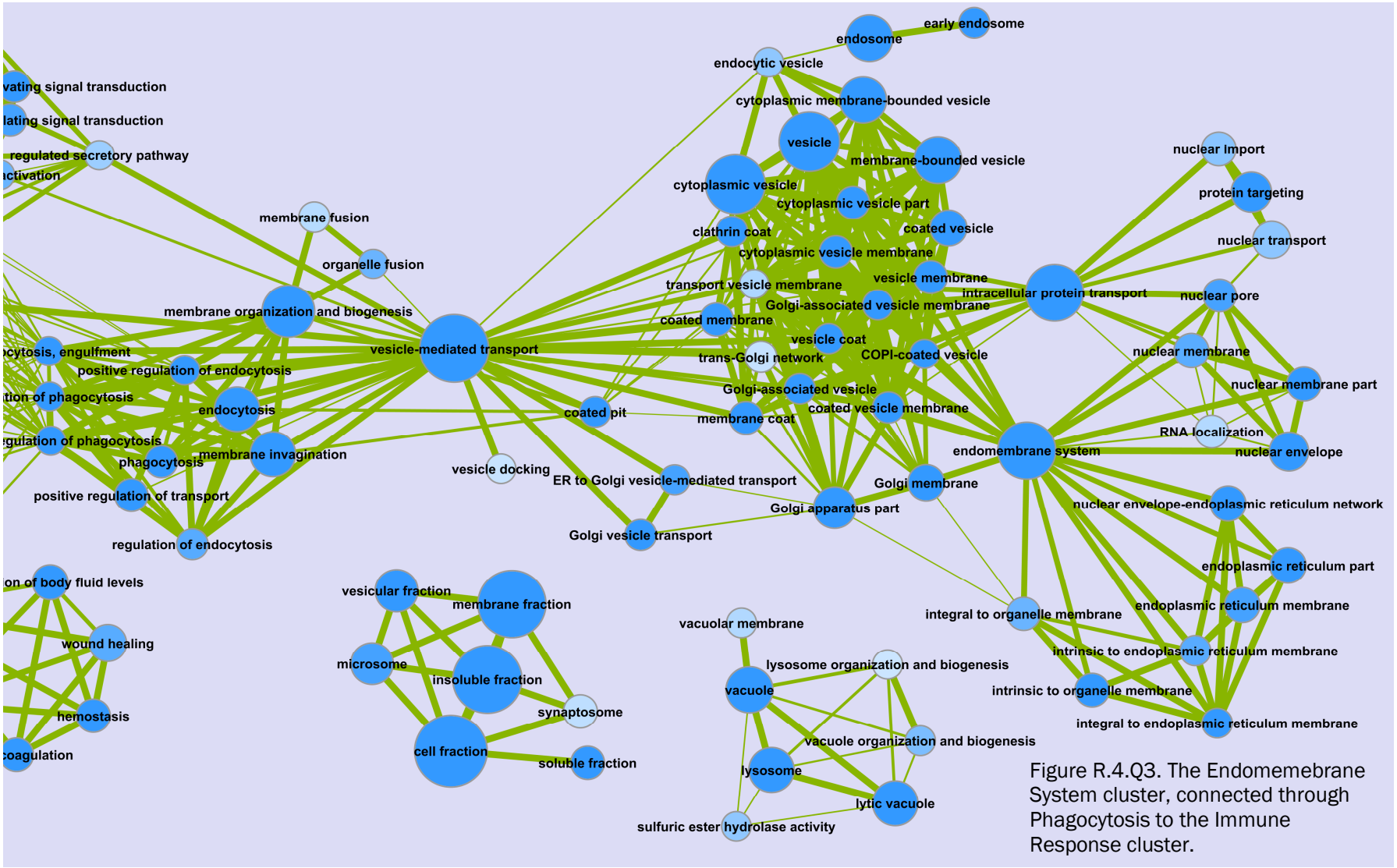


Figure R.4.Q3. The Endomemebrane System cluster, connected through Phagocytosis to the Immune Response cluster.

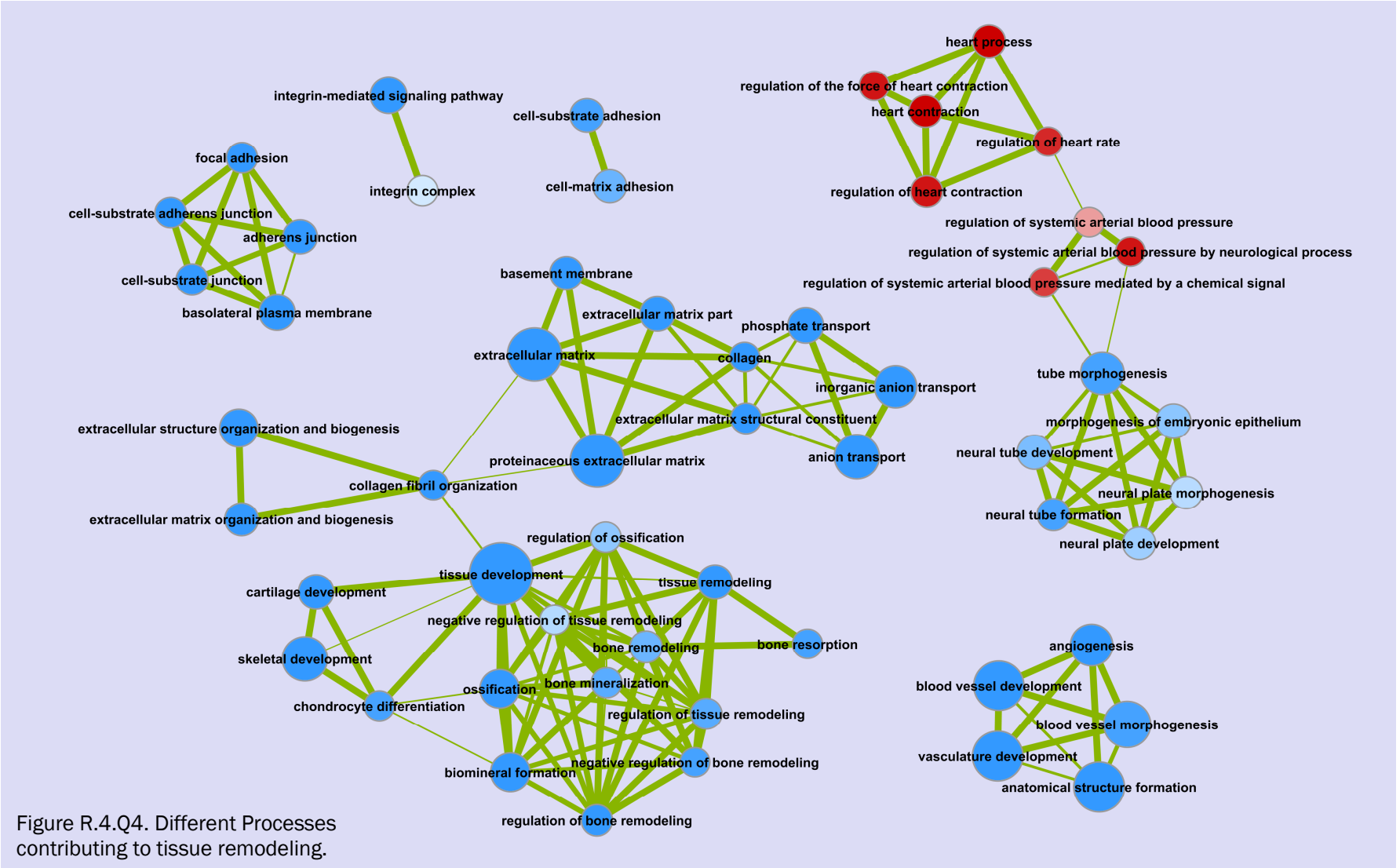


Figure R.4.Q4. Different Processes contributing to tissue remodeling.

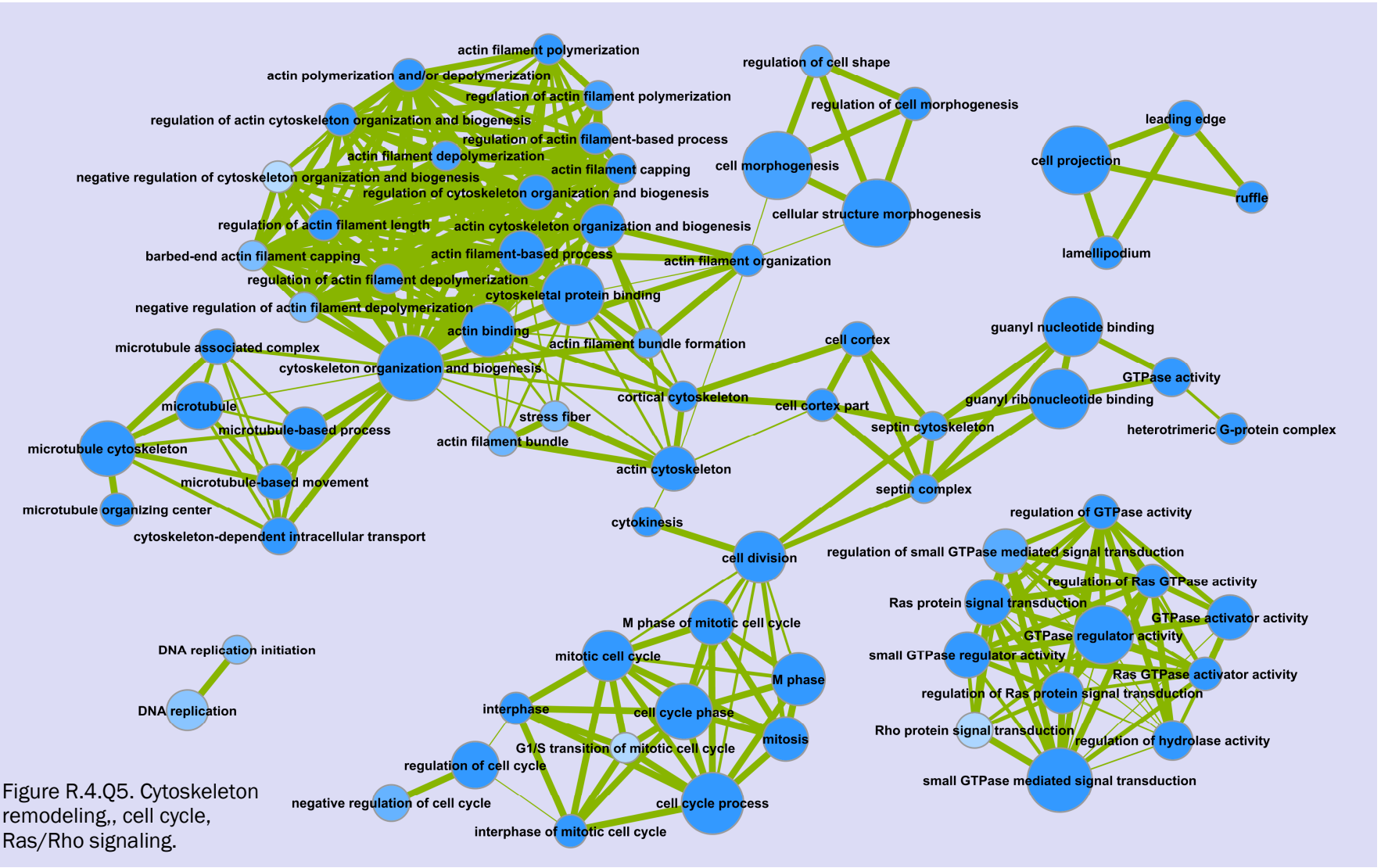


Figure R.4.Q5. Cytoskeleton remodeling,, cell cycle, Ras/Rho signaling.

Figure R.4.Q6. Growth and cell fate pathways.

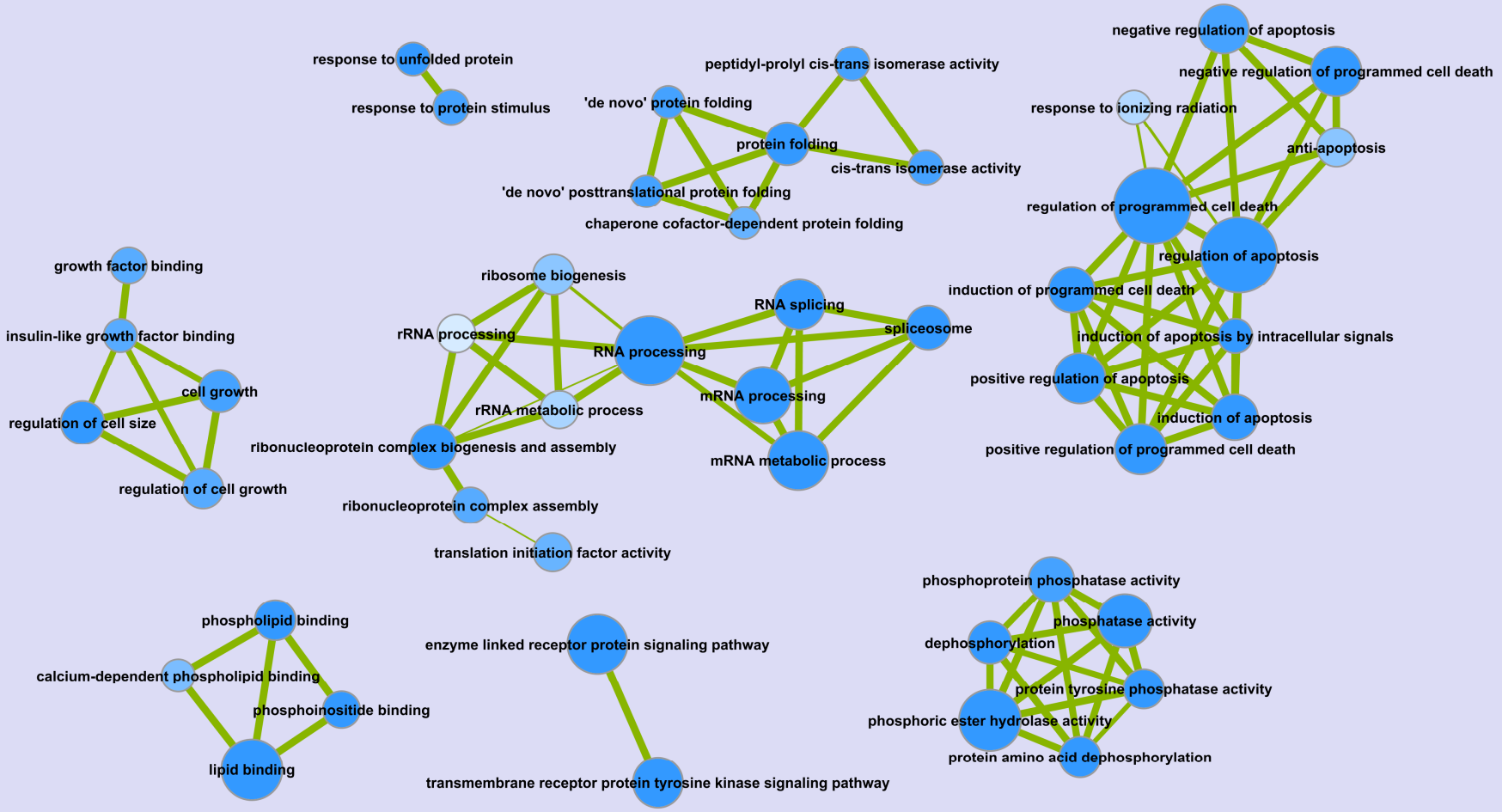


Figure R.4.Q7. Oxidative metabolism cluster, part 1.

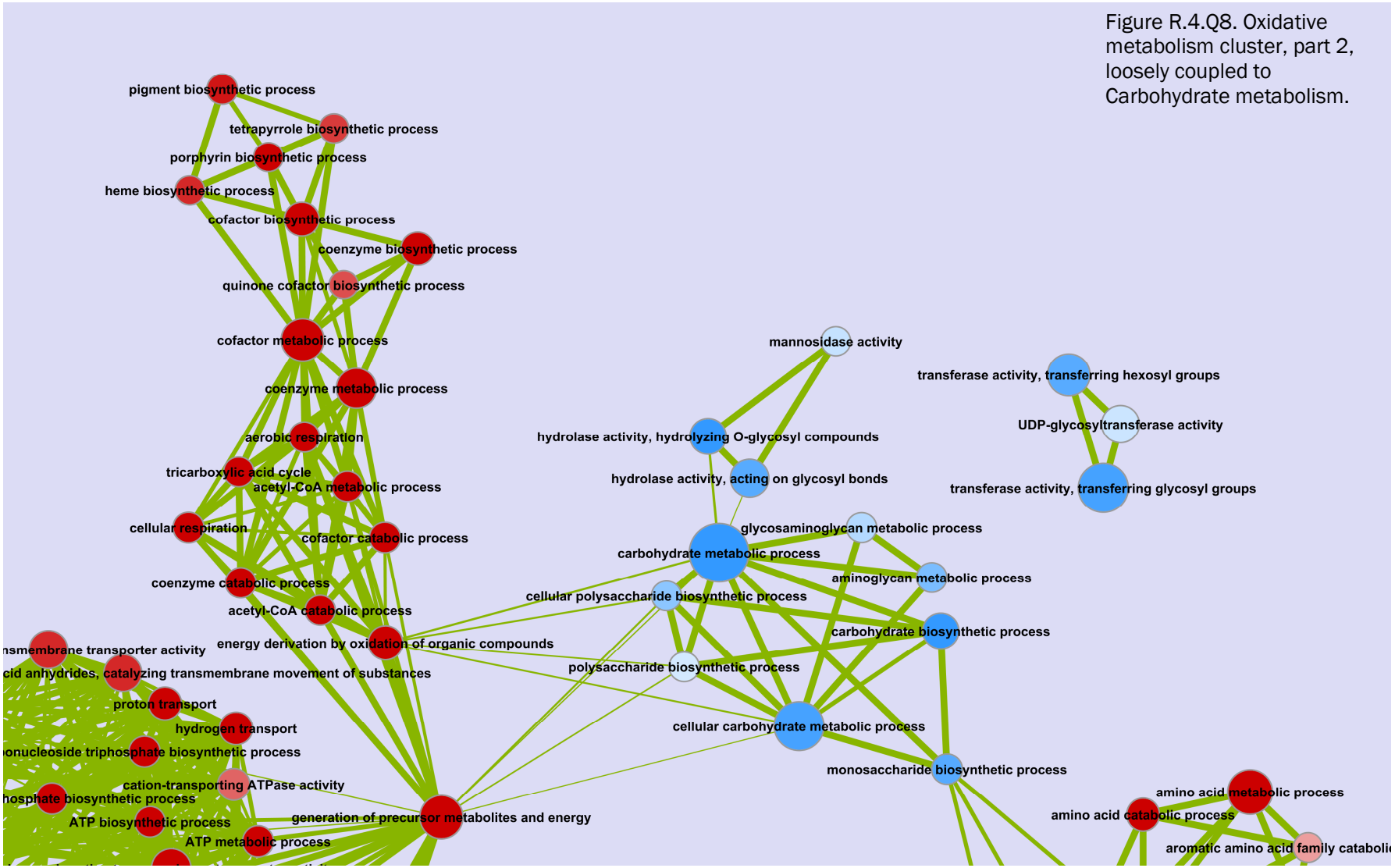


Figure R.4.Q8. Oxidative metabolism cluster, part 2, loosely coupled to Carbohydrate metabolism.

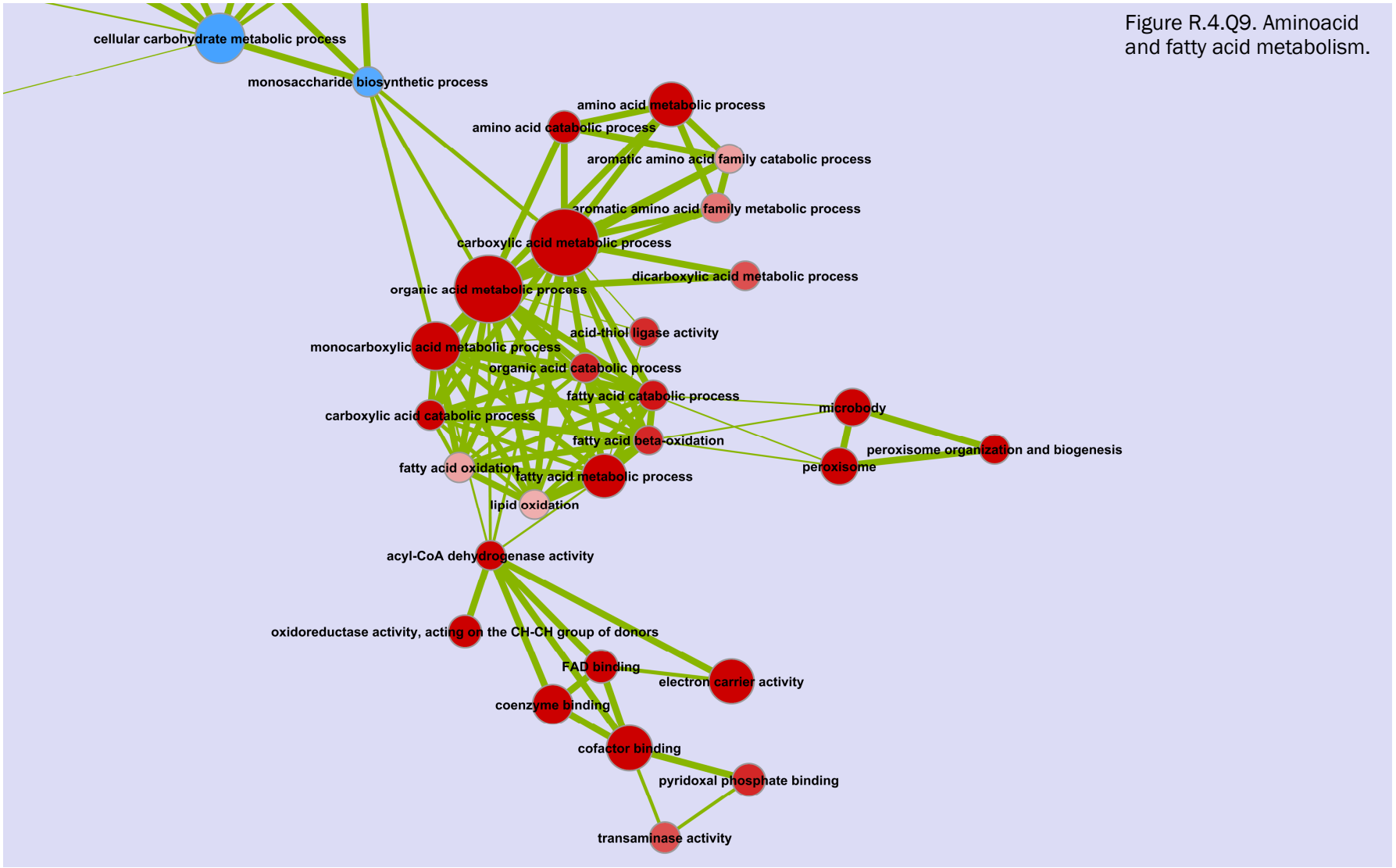


Figure R.4.Q9. Aminoacid and fatty acid metabolism.

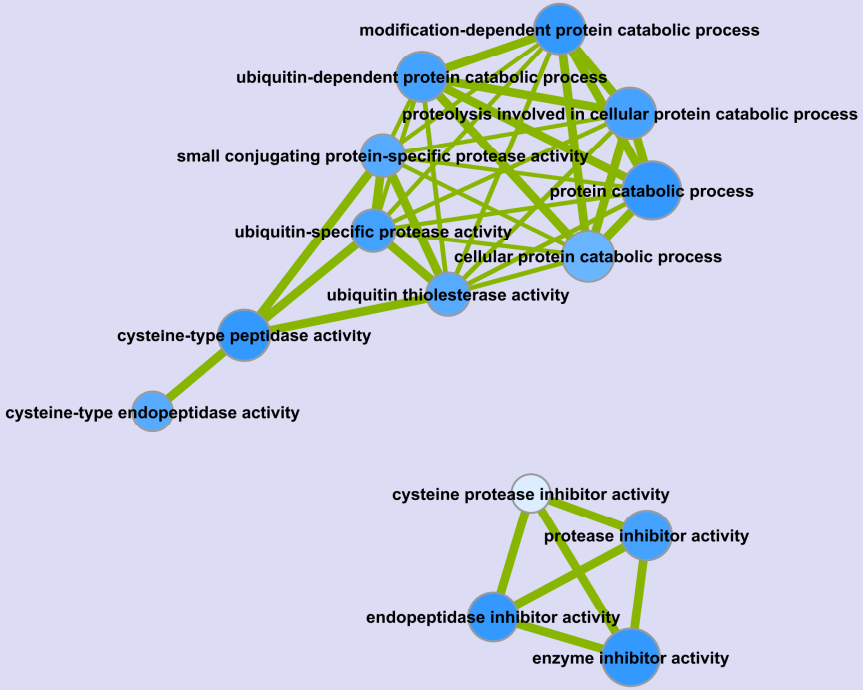


Figure R.4.Q10. Protein degradation and protease inhibition.

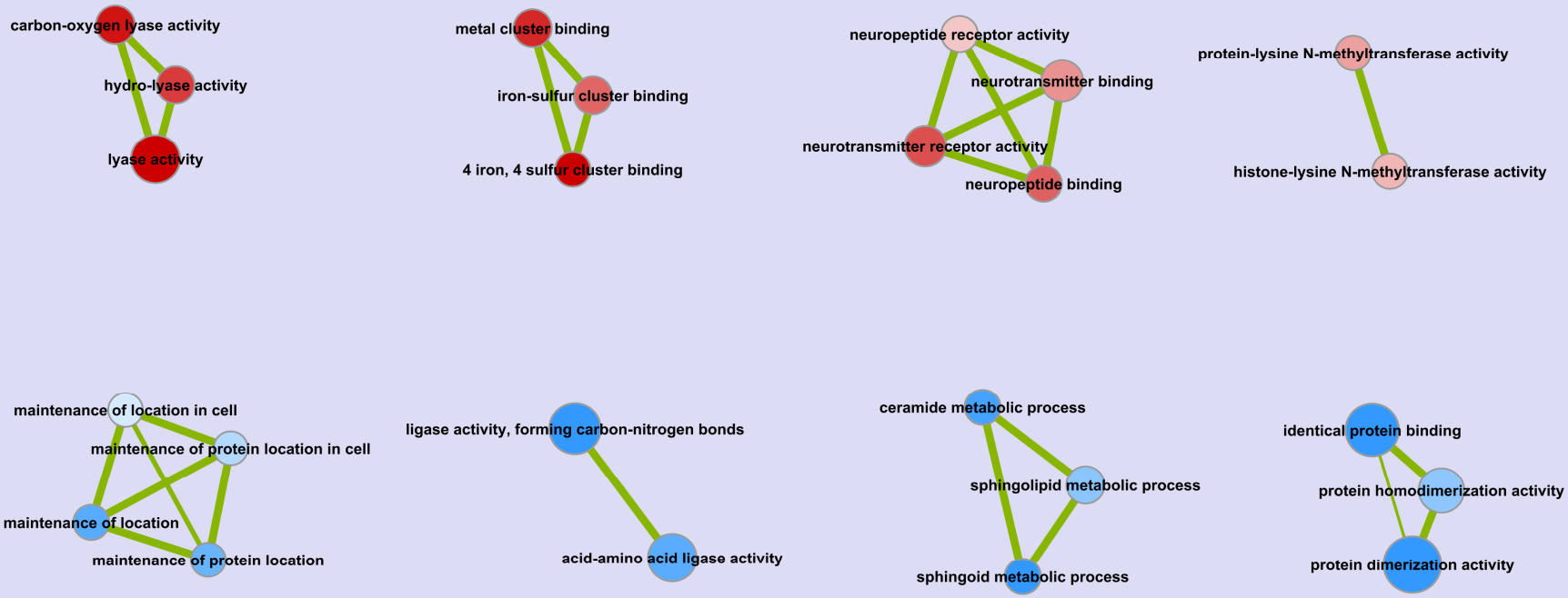


Figure R.4.Q11. Miscellanea.

miRNA Analysis

miRNA Selection

miRNAs were profiled at 16 weeks for wild-type and transgenic animals; the completion of the design, in order to include the 8 and 24 weeks time-points, is currently under course.

Considering the 16 week miRNA profiling, we designed the following test, to identify miRNA actively involved in R9C-DCM pathogenesis. We first ranked miRNA by differential gene expression, and filtered the ones with robust expression signals and mapped to public databases. Then, we tested their target sets (according to TargetScan 4.2. sequence-based predictions [9]) for enrichment in up-regulated or down-regulated genes, using the Kolmogorov-Smirnov test (KS-test) for distribution inequality. miRNA that are actively involved in R9C-DCM pathogenesis are expected to be differential and to have targets enriched in differential genes, but with opposite sign, since miRNA act as post-transcriptional inhibitors. In other words, if a miRNA is up-regulated, its targets are expected to be down-regulated, and vice-versa. Specifically, we selected only those miRNA whose target-set enrichment p-value is large enough to be associated only with one of the two ends of the miRNA ranking (figure R.5.a-b).

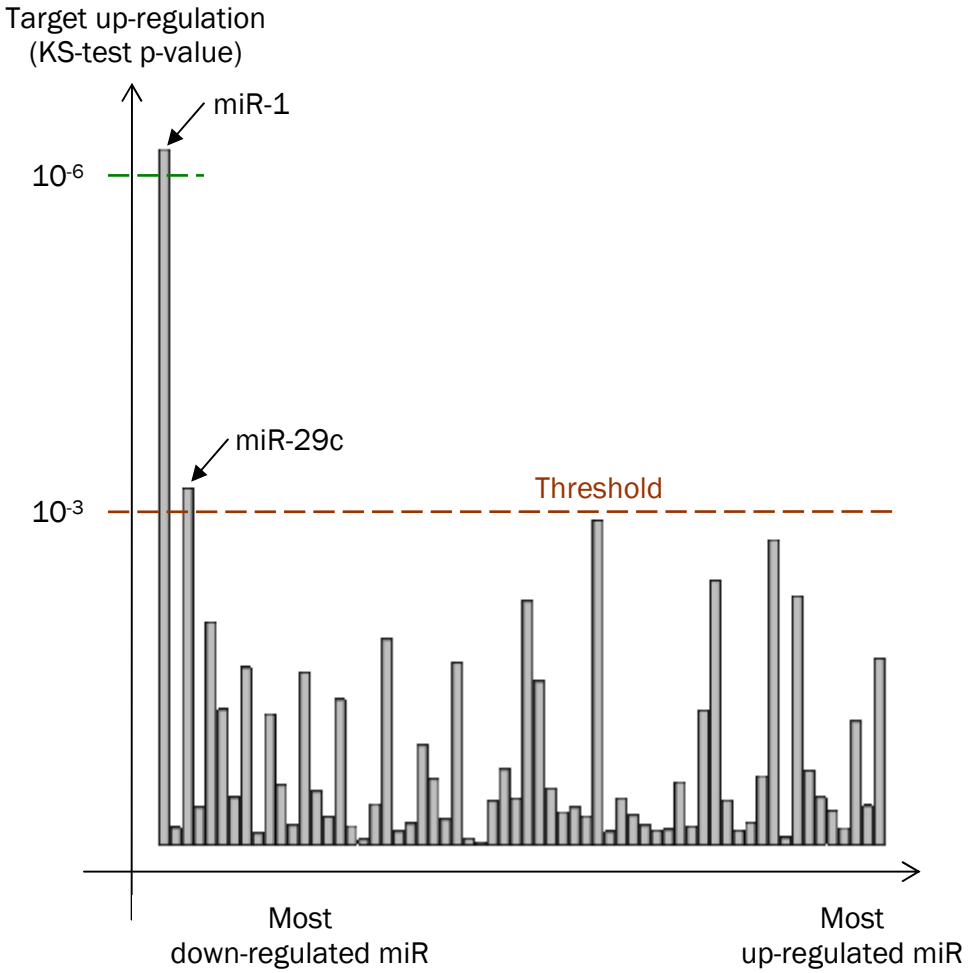


Figure R.5.a. Differentially expressed miRNA and associated target-set enrichment p-value for up-regulation. miR-1 is the only miRNA displaying a p-value larger than the confusion threshold (brown); miR-29c is borderline. P-values below the confusion threshold can be associated to the target-sets of both down-regulated and up-regulated miRNA, and therefore are deemed not significant.

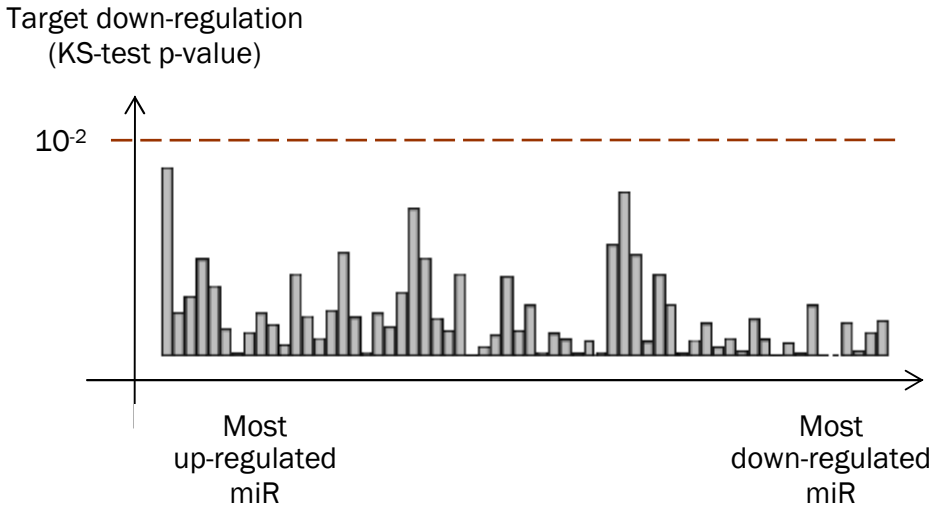


Figure R.5.a. Differentially expressed miRNA and associated target-set enrichment p-value for down-regulation. None of the miRNA passes the confusion threshold.

Surprisingly, one down-regulated miRNA passes this very stringent test, but none of the up-regulated. The selected miRNA is miR-1, a well known regulator of cardiac development and function [14].

Computational miR-1 Validation

In order to validate miR-1 role in R9C-DCM, we performed an extensive comparison between the transcriptomes of R9C-DCM and miR-1 KO [14].

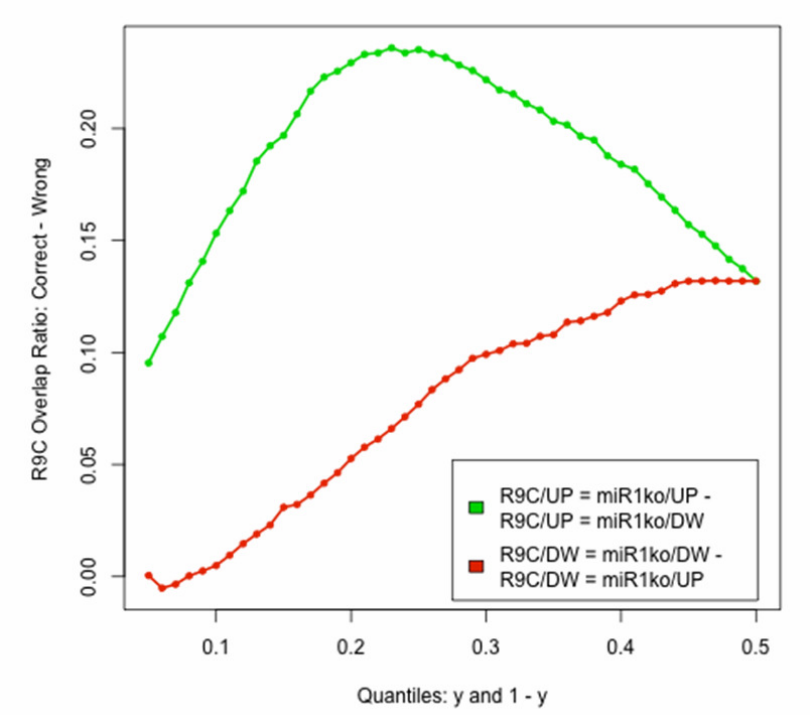


Figure R.6. Genes up-regulated in miR-1 KO display a larger than randomly expected overlap with genes upregulated in R9C-DCM. The y axis represents the fractional overlap between a running selection of up-regulated or down-regulated genes in R9C-DCM and miR-1 KO, after correction for the expected random overlap. The x axis represents the threshold used for the running selection (top *x* fraction of up-regulated and bottom *x* fraction of down-regulated genes). Green and red represent the overlap of up-regulated and down-regulated genes, respectively.

miR-1 KO data is referred to a very similar system (mouse, heart ventricular samples), and thus very limited discrepancy due to experimental factors is expected. As displayed by figure R.6, up-regulated genes in R9C-DCM and miR-1 KO are characterized by a consistent overlap which, even after correction for the expected random overlap, ranges between 10 and 25%.

miR-1 Putative Pathway

miR-1 is positively regulated by the transcription factors MyoD and the Mef2 complex [15]. These transcription factors also control the promoter of miR-133[35], a miRNA involved in muscle differentiation [14-15]. As displayed by figure R.7, the differential transcriptional states at 16 weeks of miR-1, miR-133 and its transcriptional activators Mef2 and MyoD are highly consistent.

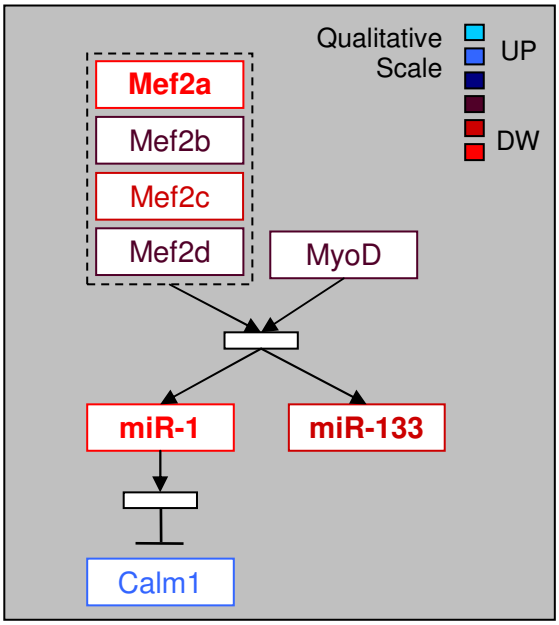


Figure R.7. miR-1 putative pathway. Red represents down-regulation and blue represents up-regulation.

Discussion

The results insofar are promising, but require additional work, both in terms of experimental validation and computational analysis. Additional computational work is currently being devoted to the in-depth interpretation and validation of the active processes identified through the Functional Enrichment Map. The same visualization approach can be exploited also to compare the R9C-DCM model to other murine models of cardiomyopathy, or to transcriptional data from human patients; another interesting application would be to identify and visualize which processes are specifically regulated by miR-1, exploiting the miR-1 KO transcriptional signature and miR-1 predicted targets. The efficacy of the miRNA selection criterion, combining miRNA differentiality and target-set differentiality, may be disputed, given the extremely limited number of candidates found; to overcome this problem, data from different miRNA microarray platforms will be evaluated and pooled together if necessary, and alternative tests for target en-

[35] miR-133 is differential at 16 weeks, and has a target-set significantly enriched in up-regulation, yet it is not selected after the previous analysis, as it does not pass ours stringent test on the robustness of gene expression signals.

richment in up-regulated or down-regulated genes will be taken into account and benchmarked, beyond the KS-test.

Acknowledgements

This work was a cooperative effort of the Bader Lab, Emili Lab and Gramolini Lab (University of Toronto, Banting & Best Dept. of Medical Research). In particular, Prof. Anthony Gramolini and collaborators developed the transgenic model and carried out the phenotype characterization and sample extraction. Daniele Merico, supervised by profs. Gary Bader and Andrew Emili carried out the transcriptomics pattern analysis, developed and applied the functional enrichment map visualization, carried out the miRNA selection and computationally validated miR-1 role. Special thanks to Dr. Ruth Isserlin who carried out a preliminary analysis of miRNA data, and was responsible for the data management.

Reference

[1] McMurray JJ, Pfeffer MA.
Heart failure.
Lancet. 2005 May 28-Jun 3;365(9474):1877-89.

[2] Dec GW, Fuster V.
Idiopathic dilated cardiomyopathy.
N Engl J Med 1994;331:1564–75.

[3] Roger VL, Weston SA, Redfield MM, Hellermann-Homan JP, Killian J, Yawn BP, Jacobsen SJ.
Trends in heart failure incidence and survival in a community-based population.
JAMA 2004;292:344 –50.

[4] Rockman HA, Koch WJ, Lefkowitz RJ.
Seven-transmembrane-spanning receptors and heart function.
Nature. 2002 Jan 10;415(6868):206-12.

[5] Solaro RJ.
Is calcium the 'cure' for dilated cardiomyopathy?
Nat Med. 1999 Dec;5(12):1353-4.

[6] Frank K, Kranias EG.
Phospholamban and cardiac contractility.
Ann Med. 2000 Nov;32(8):572-8.

[7] Schmitt JP, Kamisago M, Asahi M, Li GH, Ahmad F, Mende U, Kranias EG, MacLennan DH, Seidman JG, Seidman CE.
Dilated cardiomyopathy and heart failure caused by a mutation in phospholamban.
Science 2003; 299:1410–1413

[8] Boyd SD.
Everything you wanted to know about small RNA but were afraid to ask.
Lab Invest. 2008 Jun;88(6):569-78.

[9] Lewis BP, Burge CB, Bartel DP.
Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets.
Cell 2005; 120:15-20.

[10] [http://www.bioconductor.org/]

[11] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.
Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.
Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.

[12] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD.
Integration of biological networks and gene expression data using Cytoscape.
Nat Protoc. 2007;2(10):2366-82.

[13] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.
Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
Nat Genet. 2000 May;25(1):25-9.
[14] Zhao Y, Ransom JF, Li A, Vedantham V, von Drehle M, Muth AN, Tsuchihashi T, McManus MT, Schwartz RJ, Srivastava D.
Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2.
Cell. 2007 Apr 20;129(2):303-17.
[15] Zhao Y, Samal E, Srivastava D.
Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis.
Nature. 2005 Jul 14;436(7048):214-20.