Multi-species genomic study of oxidative stress response and adaptation

Luc Thomès

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE : SCIENCES DE LA VIE ET DE LA SANTE

UPR 9002 du CNRS, Architecture et Réactivité de l'arN, Institut de Biologie Moléculaire et Cellulaire, Strasbourg

THÈSE

présentée par :

Luc THOMES

soutenue le : **04 septembre 2020**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Bioinformatique

Étude génomique multi-espèces de la réponse adaptative au stress oxydatif

THÈSE dirigée par :

M JOSSINET Fabrice	Maître de conférences, Université de Strasbourg
M LESCURE Alain	Chargé de recherche, CNRS

RAPPORTEURS :

M FEELISCH Martin	Professeur, Université de Southampton
Mme GASPIN Christine	Directeur de recherche, INRAe

AUTRES MEMBRES DU JURY :

M BRIENS Mickaël	Tuteur scientifique, Société Adisseo S.A.S
Mme CALENGE Fanny	Chargé de recherche, INRAe
Mme LECOMPTE Odile	Professeur, Université de Strasbourg

"Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts some of them interesting or curious but making no meaningful picture as a whole."

Theodosius Grigorievich Dobzhansky

Remerciements

Tout d'abord, j'adresse mes premiers remerciements à Pascale Romby pour m'avoir accueilli au sein de l'UPR 9002 du CNRS dans l'unité ARN – Architecture et Réactivité de l'arN – de l'Institut de Biologie Moléculaire et Cellulaire (IBMC) de Strasbourg. Je remercie également l'entreprise Adisseo qui m'a permis de réaliser ce projet grâce à un financement CIFRE – ANRT ainsi que nos collaborateurs INRAe dans ce projet : l'UMR PEGASE de Rennes et l'URA de Tours.

Je souhaite également remercier les membres du jury pour avoir accepté d'examiner mon travail : Christine Gaspin, Odile Lecompte, Fanny Calenge et Martin Feelisch.

De plus, je voudrais exprimer ma gratitude aux trois personnes qui ont joué un rôle central au cours de cette thèse: Mickaël Briens, Alain Lescure et Fabrice Jossinet. Merci Mickaël de m'avoir soutenu durant ces années et de m'avoir fait découvrir le monde de la recherche chez Adisseo. Je garderai un très bon souvenir de cette expérience, et tu n'y es clairement pas pour rien ! Un grand merci à toi également Alain, avec qui j'ai passé le plus de temps durant cette thèse. J'ai été marqué par ta capacité à voir les choses toujours positivement, ce qui m'a poussé à aller toujours plus loin et à ne jamais baisser les bras. Merci pour ta patience et pour ton soutien durant ces années, j'ai beaucoup appris en travaillant au sein de ton équipe. Enfin, merci à toi Fabrice, en particulier pour tes coups de pouce en bioinfo lorsque j'en avais besoin. Même si je n'ai pas eu l'occasion d'échanger avec toi autant qu'avec Mickaël et Alain je dois dire que sans toi, je ne me serai sans doute jamais lancé dans la bioinformatique. Merci de m'avoir transmis cette passion de par tes enseignements à la fac.

De façon plus générale, j'aimerais également remercier les personnes qui ont façonné mon parcours universitaire. Merci à toi Anne Friedrich pour m'avoir donné l'opportunité de dispenser quelques séances de TP ainsi que pour ton soutien, merci à toi Philippe Dumas pour avoir instillé en moi un intérêt pour la biologie des systèmes et le langage Mathematica, et merci à vous deux, Julie Thompson et Olivier Poch, pour avoir considérablement renforcé mon intérêt dans l'analyse de séquences et l'élaboration d'outils bioinformatiques pour mener ces analyses.

J'en profite pour remercier également les membres passés et présents de l'équipe "Régulations post-transcriptionnelles et nutrition". Un grand merci à toi Mélanie pour ta bonne humeur quotidienne, nos pauses café, nos discussions et surtout pour m'avoir fait découvrir la "Danse du burritos" ! Merci aussi d'avoir tenté de me former à quelques manips, j'aurais aimé être un aussi bon disciple que tu as été une excellente formatrice. Un énorme merci à toi également Mireille, ma voisine de bureau durant la majorité de ma thèse. J'ai vraiment apprécié travailler en ta compagnie. Merci pour ta sympathie, pour toutes ces discussions et réflexions et pour tout ce que tu as pu faire pour

moi. Je n'aurais pas pu rêver d'une meilleure voisine ! Enfin, Nedaa et Ahmad, c'est à vous à présent de prendre la relève. Merci pour votre compagnie quotidienne et pour votre sympathie. Je vous souhaite tout le meilleur pour la suite de vos projets respectifs. Enfin, je ne peux terminer sans te remercier toi aussi, Laurence, qui malgré que tu ne sois pas au même étage que moi dans l'institut, m'a fait me sentir un peu moins seul comme bioinformaticien dans l'équipe !

Je sais que je vais en oublier certain(e)s mais je veux également remercier les stagiaires passés et présents du labo pour leur bonne humeur, nos échanges et nos discussions : Mélissa, Victor, Léonie, Stefan, Lana et Adrien. Plus particulièrement, merci Mélissa pour le dépaysement que tu as apporté au labo, j'ai grave kiffé wesh. Merci également à toi Léonie. C'est agréable de pouvoir côtoyer des personnes toujours souriantes comme toi, y compris pour cette fois où une personne a confondu le congélateur et le frigo pour tes bactéries ! Enfin Victor, un très gros merci à toi également. J'ai vraiment eu de la chance que tu sois mon premier stagiaire car tes capacités et ton sérieux m'ont grandement facilité la tâche ! Merci beaucoup pour nos discussions (scientifiques et musicales) ainsi que pour ton implication au labo.

Outre l'équipe de Strasbourg, je tiens à adresser des remerciement à l'autre équipe à laquelle j'étais rattaché chez Adisseo pendant ma thèse, le CERN. Merci aux personnes que j'ai rencontrées et avec qui j'ai pu échanger, en particulier Yves Mercier, Vincent Jacquier, Estelle Devillard, Amine Hachemi et évidement Mickaël Briens. Vous avez largement contribués à ce que mon expérience chez Adisseo soit la plus agréable possible.

Tout aussi important que le soutient des membres de l'équipe, je veux remercier les personnes des autres labo qui ont également contribuées à m'apporter du courage pendant ces années. Laura, Marta, Raphaël, Kévin, Roberto, José, merci à vous tous, vous avez été géniaux et avez rendu cette période bien plus agréable à traverser. Je te remercie tout particulièrement, José, car tu es pour moi la meilleure rencontre que j'ai faite à l'institut. Tu es quelqu'un de passionné, curieux et j'adore ton humour. Merci pour nos conversations passionnantes et tous ses moments passé avec toi ainsi que Raphaël, Laura et Marta. J'espère que notre prochaine bière, on la boira au Mexique ! En cuanto a ti, Roberto, quiero agradecerte tu simpatía y realmente espero poder verte de nuevo para la "hora del duelo" ;-).

Je tiens enfin à remercier ma famille, en particulier mes parents, grands-parents et ma cousine Vanessa pour leur soutien depuis toutes ces années. Un merci particulier pour ma mère qui a toujours cru en moi y compris dans les moments difficiles ainsi que pour mon père qui m'a aidé à garder confiance en mes capacités et à ne pas relâcher mes efforts. J'espère que tout ce chemin parcouru depuis la maternelle, en grande partie grâce à vous, vous rendra fiers de moi. J'en profite pour remercier mon oncle Jean-Claude et sa femme Chantal pour leurs encouragements toutes ces années.

Merci aussi d'avoir initié chez moi le désir de faire de la recherche ainsi que pour tous ces ouvrages de biologie qui me sont encore régulièrement utiles.

Mention spéciale à mon ancien prof de batterie, Didier Lemaitre, pour m'avoir encouragé depuis tout petit en musique et pour le reste. Je pense que ma capacité à ne pas trembler de peur devant un public je te la dois, après toutes ces fois où j'ai dû surpasser ma timidité lors de mes auditions de batterie.

Pour terminer, je tiens à remercier mes amis, passés et présents, qui ont tous au moins contribués en partie à faire de moi ce que je suis aujourd'hui. Flora, Jordan, Romain, David, Yannis, Émilie, Nassera, Julien, Nicolas, Cassandre, Joshua, Cindy, Evelyne, Vanessa, merci d'être comme vous êtes et de votre présence dans ma vie. Un merci tout particulier à toi, "Monsieur Fidèle", sur qui je sais que je pourrai toujours compter ; à toi, Nicolas, pour toutes nos discussions passionnantes sur à peu près tout et n'importe quoi ; et à ma meilleure amie Vanessa qui, plus qu'une source de motivation, est aussi l'une de mes sources d'inspiration.

Le chemin qui m'a conduit jusqu'ici n'a pas toujours été des plus agréables à parcourir, mais quand je vois le résultat je me dis que malgré tout, ça en valait quand même mille fois la peine.

Contents

Tables index

Figures index

Abreviations

ACTH	Adrenocorticotropic hormone
AKT	Protein kinase B
BCAA	Branched-chain amino acid
BCAT1	Branched-chain amino acid transaminase 1
BLAST	Basic local alignment search tool
CAT	Catalase
CD44	Hyaluronate receptor
DNA	Deoxyribonucleic acid
DUOX	Dual oxidase
ECM	Extracellular matrix
GAS	General adaptation syndrome
GO	Gene ontology
GPX	Glutathione peroxidase
GRX	Glutaredoxin
GSH	Glutathione
H2O2	Hydrogene peroxide
HA	Hyaluronic acid
HGT	Horizontal gene transfert
HIF-1	Hypoxia-inducible factor
KEAP1	Kelch-like ECH associated protein
LAS	Local adaptation syndrome
lncRNA	Long non-coding RNA
miRNA	Micro RNA
mm Hg	Millimetre of mercury
mRNA	Messenger RNA
NAD	Nicotinamide adenine dinucleotide
NADP	Nicotinamide adenine dinucleotide phosphate
ncRNA	Non-coding RNA
NGS	Next-generation sequencing
NOX	NADPH oxidase
NRF2	Nuclear factor-erythroid 2 related factor 2
O2	Dioxygene
PI3K	Phosphatidylinositol-3-kinase

PRX	Peroxiredoxin
redox	Reduction-oxidation
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
RT-qPCR	Real-time polymerase chain reaction
SelenoN	Selenoprotein N (protein)
SELENON	Selenoprotein N (gene)
SOD	Superoxide dismutase
SRC	Proto-oncogene tyrosine-protein kinase
TGFß	Transforming growth factor beta
TR	Thioredoxin reductase
tRNA	Transfert RNA
TRX	Thioredoxin reductase

Avant-propos

Travailler à l'interface entre la biologie moléculaire et la bioinformatique n'est pas une situation confortable et requiert une expertise ainsi qu'une connaissance spécifique de chacun de ces domaines. À cause de cette interdisciplinarité, les experts de chaque spécialité montrent parfois une certaine appréhension et doutent d'être "la bonne personne" pour évaluer mon travail. Considérant cela, j'ai fait le choix d'introduire les concepts principaux de ces deux domaines – la biologie du stress et la bioinformatique – ainsi que du domaine à l'interface des deux – l'évolution –, en les replaçant dans un contexte général et historique. Ceci m'a permis de présenter les approches et méthodes conceptuelles qui constituent les bases de mes études.

Foreword

Working at the interface between molecular biology and bioinformatics is not a comfortable situation, requiring specific expertise and knowledge in both fields. Because of this interdisciplinarity, experts from each specialty sometimes show apprehension and doubt to be "the right person" to evaluate my work. Considering this, I made the choice to introduce the principal concepts of these two domains – stress biology and bioinformatics – and one domain at the interface between the two – evolution –, by placing them in a general and historical context. This allowed me to present the conceptual approaches and methods that constitute the basics of my studies.

INTRODUCTION

1) Stress

During their life, organisms will experience a wide range of environmental components recently defined as the "exposome" (Vermeulen et al., 2020) (Figure 1). Chemicals, dietary constituents, psychosocial situations or even physical factors and biotic aggressions are potential challenges susceptible to induce a physiological stress to an organism. The term "stress" is generally used in common parlance to describe a state of psychic perturbation similar to anxiety. Initially proposed by Hans Selye and defined in his book from 1956 entitled "The Stress of Life", the term "stress" was coined to refer to a physiological state. Etymologically, it derives from the English word “distress”, itself originating from the old French word “destresse”. Despite its common use, few people agree on a precise and consensual definition of stress. It is actually understood in our society as the feeling of being anxious or worried about something, but it was initially coined to consider all manifestations of organisms during their response against environmental disruptors. This complex situation needs to be clarified by coming back to the origins of the stress notion.

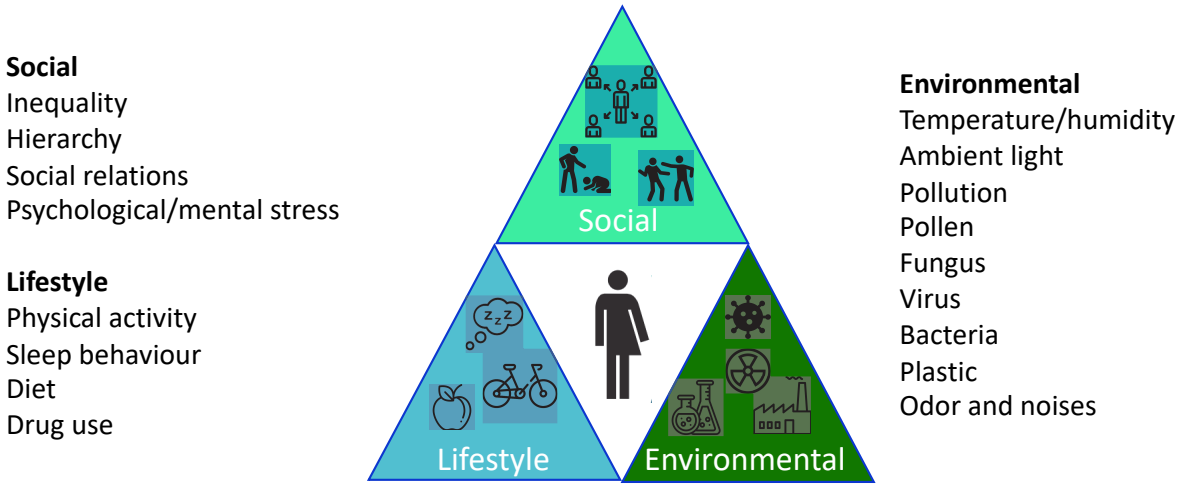


Figure 1 - The exposome concept. The exposome corresponds to all the challenges humans can face during their life. These elements can be extended to all living organisms and are considered as stress factors. Adapted from (Vermeulen et al., 2020).

1.1) Origin of stress

One of the main goal of medicine is to determine the specific manifestations that characterize diseases: the symptoms. If a disease is caused by a pathogen (bacteria, virus, parasite) or by a toxic, depending on its nature and its mode of damage, it is expected to display specific effects on a patient. Identifying these specific clinical descriptions is the key to assure a proper diagnostic and so to provide the most appropriated treatment. To this aim, the method that is employed consists in systematically ignoring all symptoms that are shared between different pathologies, as they are not discriminating of the causes. Around 1930, Hans Selye, a graduated student in medicine at Prague

University, postulated that if a wide range of diseases exacerbates similar reactions, this could be the sign of a global biological response to many environmental perturbations. This observation was the foundation of his stress theory.

1.1.1) Non-specific symptoms

When patients are exposed to a pathogenic or toxic agent, they develop a list of symptoms, some of them independent of the nature of the agent. Such symptoms include the feeling of discomfort, muscle and articular pain, intestinal disturbances, loss of appetite and general weight loss. Strikingly, it was known at that time that these symptoms are also manifested after exposition to non-pathogenic agents. For example, surprisingly, patients that endured severe burns often develop gastrointestinal ulcers. This was challenging to understand, as it was no apparent direct connection between skin burns and the inflammation of the digestive tract. This observation raised the hypothesis of a common response mechanism that uses the same biological pathway, ending up in similar symptoms. Hans Selye decided to name “stress” this specific physiological state manifested after exposition to any disruptor. Firstly, he proposed to define stress by stating what it is not: it is not a nervous tension as it can also be observed in animals without nervous system, excluding it from being confounded with anxiety; nor it is only the nonspecific results of damages, as some normal activities (i.e. sport) can also lead to stress without causing any observable damage; it is neither any deviation from a stable state (homeostasis) as normal activities such as muscle contraction or feeding also causes deviations from the resting state in the concerned organs. Hans Selye then enounced two important characteristic features of stress: on one hand, stress is not a non-specific reaction as the pattern of stress reaction is very specific and it affects define organs in a selective manner. On the other hand, stress is not a specific reaction as it can results from a variety of pathogenic or non-pathogenic agents. Finally, Hans Selye ended up with a notion of stress that he defined as "a physiologic state manifested by a specific syndrome which consists of all non-specifically induced changes within a biologic system" (Figure 2). In other words, "stress has its own characteristic form and composition, but no particular cause".

Here is the distinction between a non-specifically formed change and a non-specifically caused change. The first situation corresponds to a mechanism that affects most or all parts of an organism without specificity. The second one describes a mechanism that can be induced by many factors, here corresponding to stress. In that way, stress and its manifestations are non-specifically induced, but its stereotypical manifestations are by definition, specific.

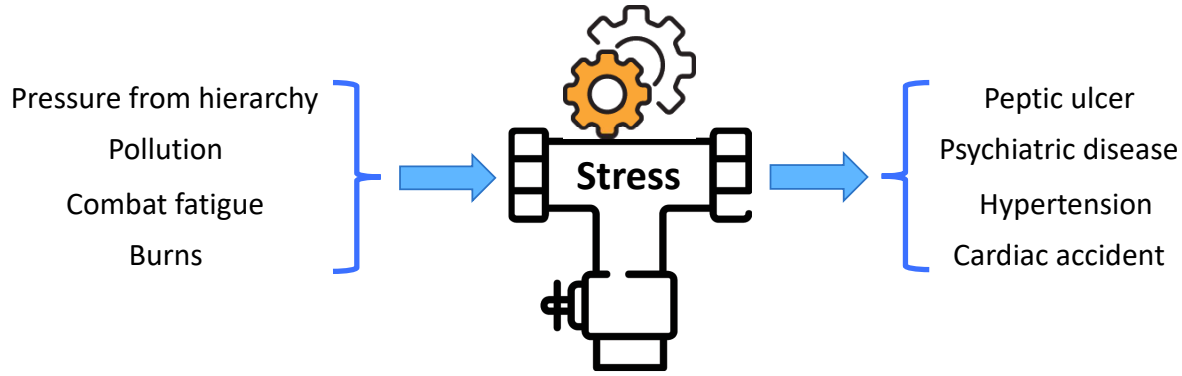


Figure 2 - Non-specificity of stress response. Each cause represented on the left is specific. Similarly, each result represented on the right is also specific. However, it is no possible direct connection between a cause and a result. Each cause can produce the same set of non-specifically induced results through a common pathway, here stress. Adapted from (Selye, 1976).

From this definition, stress constitutes a physiologic state as it concerns an organism as a whole. In contrary to the specific effects caused by bacteria growing in lungs and leading to pulmonary problems, stress affects central organs (the central nervous system), peripheral organs (muscles, stomach, intestine) as well as the blood circulation. Even when a perturbation acts locally, stress always constitutes a global response.

Stress is manifested by a specific set of symptoms, referred in medicine as a syndrome. These symptoms are the physiologic manifestations of a response that allows stress identification and quantification. They are specific in the way that they are always detected together during response to a stress factor. Consequently, they do not depend on the agent nature that can be biologic, chemical or physical.

1.1.2) Stress misunderstandings: a cause, a state or a consequence?

The definition of stress proposed by Hans Selye led to major criticisms and misunderstandings. The term stress as used in his explanations seemed to be too much confusing and misleading, being interpreted as a cause, a state or a consequence simultaneously. Even today, 65 years after publication of Selye's seminal work, this remark persists as this term is yet misused. Some people talks about organisms exposed to stress, assuming that "stress" is a cause, the agent from which originates a response. Some others consider stress as a consequence, saying that organisms are stressed. Finally, the last ones consider stress as a state characterized by a typical response: the stress response.

In his definition of stress as a physiologic state, Hans Selye highlighted the clear distinction between the state (the stress) and the factor able to produce this state, called the "stressor" and also referred as stress factor, agent or inductor. This point makes possible to describe an organism stressed

by exposure to a stressor that triggers an appropriated stress response. For example, nutrient deprivation can be considered as a stressor triggering a nutritional stress response.

After many years of observations, Hans Selye clarified the organization of stress response at the organ level, underlying the importance of the hypothalamus pituitary adrenal (HPA) axis (Figure 3). Hypothalamus is a region of the brain that connects the nervous system to the endocrine system through the pituitary gland intermediate. This gland is a protrusion of the hypothalamus responsible for hormones synthesis and its release within the blood circulation, such as growth hormone (GH), prolactin (PRL), luteinizing hormone (LH), follicle stimulating hormone (FSH), thyrotropin (TSH) or adrenocorticotropin (ACTH) (Perez-Castro et al., 2012).

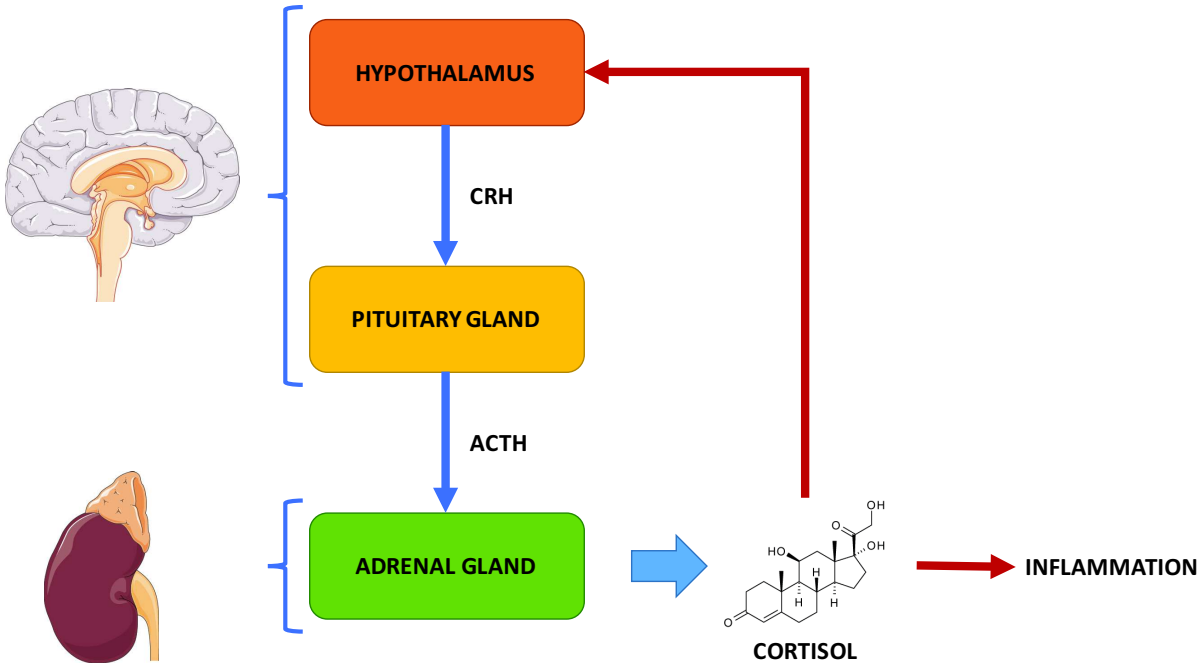


Figure 3 - The human HPA axis, an early model for stress response. Hypothalamus is tightly connected to the pituitary gland in the brain. Hypothalamus releases the corticotrophin-releasing hormone (CRH) to the pituitary gland that, consequently, secretes the adrenocorticotropic hormone (ACTH). ACTH is released in the blood circulation to reach the adrenal gland localized on the top of kidneys. Upon ACTH stimulation, the adrenal gland secretes cortisol, an anti-inflammatory steroid hormone. Then, the hypothalamus responds to cortisol level through a feedback inhibition mechanism.

1.2) Physiological response

In his book, Hans Selye reported that during a response to a stressor, multiples organs undergo specific and reproducible changes. These observations were made within a wide range of vertebrate species, from humans to chickens. Independently of the species and the nature of the stressor, the stress state presented many physiological manifestations, including a hypertrophy of the adrenal cortex, an atrophy of lymphatic organs scattered throughout the body, gastrointestinal ulcers and a general weight loss. Dissecting with more details these changes over time, he proposed two new

concepts validated in various vertebrate species: the general and the local adaptive syndromes, referred as GAS and LAS respectively.

1.2.1) Mechanism of the General Adaptive Syndrome

GAS is a dynamic syndrome produced by agents that have a general effect on the whole body such as an intense heat wave or a long exposure to sun rays. The general stress response evolves over time through three different steps when exposure to the stressor is maintained:

- The first step called “alarm reaction” corresponds to the initial response after exposure to a stress factor. This stage is characterized by a deterioration of many physiologic parameters such as a decreases in muscular tone, blood pressure (hypotension), body temperature (hypothermia), blood sugar concentration (hypoglycaemia) and general body mass. Conversely, consequently to hormonal secretions from the adrenal cortex, hormones concentration in blood increases. When survival to the stressor is possible, this state allows the organism to prepare an appropriated response, aiming to recover the pre-stress state. If it failed to respond adequately, the stressed organism will die within hours to few days after initial exposure.
- Alarm reaction is followed by a "stage of resistance". This step is mainly characterized by the opposite manifestations: hormones accumulation in the adrenal cortex, blood sugar and hormones concentration decrease and recover of the initial individual weight. Resistance allows local adaptation to the stressor: multiplication of cells to enhance tissue capacities and activation of a local inflammation to defend the injured area against infection. During this adaptive transition, stressed organisms generally show a higher resistance against the causative stressor due to the optimal development of an appropriated response. In return, resistance capacities against most of other stress factors are lowered because of this specialization.
- In cases where stressor exposure is maintained and the organism is unable to return to the pre-stress state, a last stage takes place: the "stage of exhaustion". The acquired resistance fades and symptoms from the alarm reaction are manifested again. Once this stage is reached, if stressor exposure is maintained, the stressed organism will die.

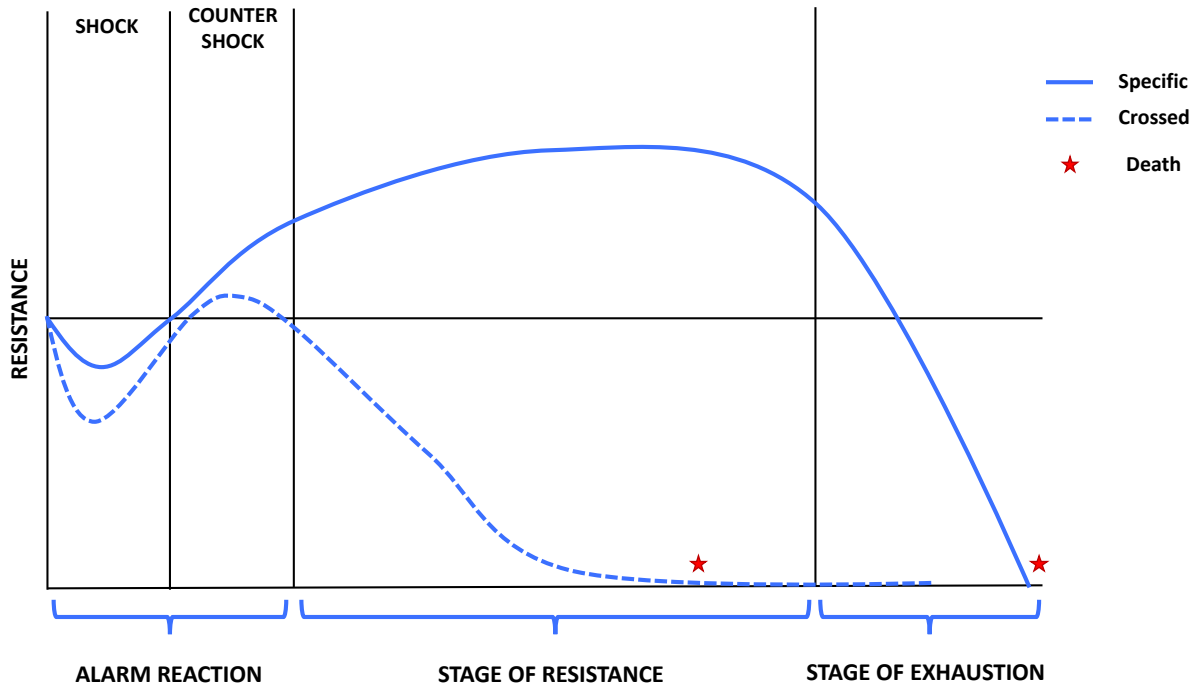


Figure 4 - Acquisition of adaptation during GAS. After stressor exposure, three stages follow in an organism to face the challenge. After the initial step of alarm reaction, the stage of resistance allows an organism to activate specific mechanisms against the causative stressor (straight line). In some cases, stressor exposure can lead to a cross-adaptation where the organism develops inappropriate resistance mechanisms (dotted line). In that case, the organism, not being able to cope with the environmental perturbation, will develop severe conditions. Adapted from (Selye, 1956).

During GAS, the second step will determine the ability to face a perturbation: the ability of adaptation (Figure 4). This determination is based on several physiological mechanisms and the probability to succeed depends on multiple parameters.

1.2.2) The adaptive process

Adaptation, the process leading to resistance capacities during the second phase of GAS, is a complex mechanism. It is generally defined as all changes needed in a living being to accommodate against environmental conditions. It allows exposed organisms to survive by increasing their specific resistance against specific stressors. Basically, adaptation allows to transiently explore an alternative stable state, temporarily disrupting homeostasis. Two different but convergent processes permit this: adaptive homeostasis and hormesis. Adaptive homeostasis is a notion defined by Kelvin J. A. Davies in 2016 as "the transient expansion or contraction of the homeostatic range in response to exposure to sub-toxic, non-damaging, signaling molecules or events, or the removal or cessation of such molecules or events". Alternatively, hormesis is a notion predicted by Southam and Ehrlich in 1943 that mainly differs from adaptive homeostasis by the involvement of a repair process: "the process by which sub-lethal damages caused by small doses of a toxin or poison would produce an

exaggerated repair response in which the organism actually becomes stronger than it was previously". These two concepts state that low intensity exposure to a stress factor induces an adaptation for further exposure. The acquired resistance capacities are then specific (Figure 4, straight line) or crossed (Figure 4, dotted line) depending on the agents against which adaptation is acquired. These two processes contribute to the variability of adaptive capacities in living organisms, determining if adaptation is possible, and against which range of stressors.

Interestingly, adaptation is not acquired permanently but rather for a given but variable period. This knowledge supports that adaptive homeostasis is permitted only temporary, and that prolonged exposure to a derivation of the homeostatic state is deleterious. This observation is in agreement with the notion of allostatic charge or load (see 1.2.4). Consequently, the kinetic of response is a key phenomenon in the process.

1.2.3) Mechanism of the Local Adaptive Syndrome

Contrary to GAS, LAS is characterized by a selective and specific response from organs or tissues that are locally exposed to a stressor. Its manifestations will be the same independently of the stressor nature, but it is limited to the exposed organ. The principal role of LAS is to create a barrier between the invaded or damaged region and the rest of the body. As during the general stress response, three steps characterize LAS:

- First, the invaded or injured zone sets up an inflammation response. This step corresponds to a local counter-attack to eliminate the stressor while it is still possible. Locally, tissue-resident macrophages and mast cells initiate recognition of an infection or aggression (Medzhitov, 2008). A wide range of inflammatory molecules are released such as chemokines or cytokines that will support host response, including recruitment of immune cells and plasma proteins, sent to the affected site to kill any eventual invader. To do so, immune cells release secretory vesicles containing highly reactive molecules including reactive oxygen species (ROS) and reactive nitrogen species (RNS) among others (Medzhitov, 2008). If the pathogenic agent is eliminated by this oxidative burst, a local repairing process occurs. Otherwise, a "fight-or-flight" strategy is applied: pursuing the counter-attack with the risk of major collateral damages due to reactive species production that do not discriminate host and invader cells, or withdrawing by repression of the local inflammation.
- When the local inflammation is not sufficient, the inflammatory process is followed by degeneration of proximal cells and formation of granulomas: a wall constituted by

macrophage layers surrounding invader's cells. It aims to block or at least minimize infiltration of the stressor within the infiltrated tissue.

- Finally, if stressor exposure is maintained, affected cells massively die by apoptosis or necrosis leading to important damages at the tissue level (Figure 5). Apoptosis is a cell death program playing important roles during cell damage and cell stress response, similar to the process setting during development and morphogenesis (Nikoletopoulou et al., 2013). Cell death is initiated by caspases (cysteine-aspartic proteases) and proteases activation that induce mitochondrial membrane permeabilization, chromatin condensation and DNA fragmentation. Conversely, necrosis is a caspase-independent cell death activated more specifically during damage or stress response, and also in some pathologies (Nikoletopoulou et al., 2013). A third cell death mechanism, autophagy, consists in self-cannibalization that can be activated in response to particular situations such as nutrient deprivation, hormonal depletion or hypoxia. This process involves the engulfment of cytoplasmic material and intracellular organelles within intracellular vesicles called autophagosomes (Nikoletopoulou et al., 2013).

Although local and general responses are two distinct phenomena, their activation mechanisms are linked together: a sufficiently intense local stressor can trigger a general response and a general stressor can exacerbate or repress local effects (Selye, 1976). Indeed, GAS involves in its initial and final steps secretion of adrenal hormones, such as corticoids, that are known to repress inflammation. These molecules belong to the steroid hormones family and are used by many organisms as anti-inflammatory agents (Perez-Castro et al., 2012). Corticoid secretion favors an easier and faster local healing process by repressing inflammation, and avoiding useless side effects and energetic wasting in case the invader can be easily countered. In contrary during the second step of GAS, adrenal hormones concentration decreases in blood, stimulating inflammation. This situation generally allows the elimination of the invader but can also results in unintentional activation of inflammatory processes through different places of the body. Conversely, local responses can also triggers a general one as in case of allergic reaction resulting from pollen inhalation and caused by an inappropriate dysregulation of the immune system against generally harmless stressors (Selye, 1976). Despite these observations, the functional relations between LAS and GAS are not well understood and only few indications on the nature of the first messenger involved in stress response are known.

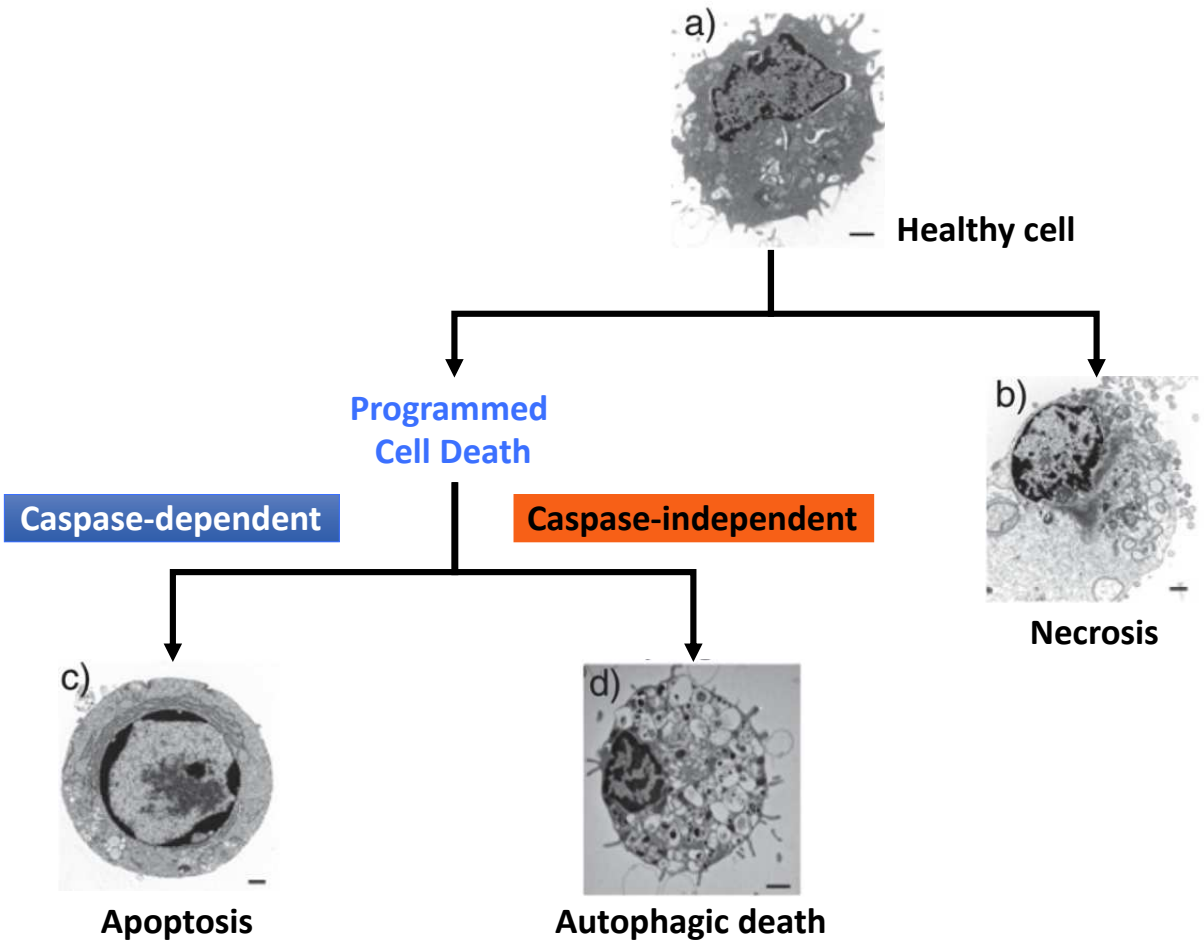


Figure 5 – Apoptosis, autophagy and necrosis. Apoptosis, autophagy and necrosis are the main pathways for programmed and non-programmed cell death. Necrosis is a non-programmed caspase-independent pathway activated in healthy cells (a) and leading to necrotic cells of particular morphology (b). Necrotic cells display an endoplasmic reticulum and mitochondrial swelling, rupture of the cell membrane, distension of the nucleus and cell lysis. In contrary, apoptosis (the caspase-dependent cell death) leads to rounded cells, chromatin condensation, nucleus fragmentation and the loss of apoptotic bodies (c). These bodies are vacuoles containing intact organelles and cytoplasm. Finally, autophagy is another caspase-independent but programmed cell death, contrary to necrosis (d). Cells under autophagic death display numerous intracellular vesicles: autophagosomes. Adapted from (Nikoletopoulou et al., 2013).

Although local and general responses are two distinct phenomena, their activation mechanisms are linked together: a sufficiently intense local stressor can trigger a general response and a general stressor can exacerbate or repress local effects (Selye, 1976). Indeed, GAS involves in its initial and final steps secretion of adrenal hormones, such as corticoids, that are known to repress inflammation. These molecules belong to the steroid hormones family and are used by many organisms as anti-inflammatory agents (Perez-Castro et al., 2012). Corticoid secretion favors an easier and faster local healing process by repressing inflammation, and avoiding useless side effects and energetic wasting in case the invader can be easily countered. In contrary during the second step of GAS, adrenal hormones concentration decreases in blood, stimulating inflammation. This situation generally allows the elimination of the invader but can also results in unintentional activation of

inflammatory processes through different places of the body. Conversely, local responses can also triggers a general one as in case of allergic reaction resulting from pollen inhalation and caused by an inappropriate dysregulation of the immune system against generally harmless stressors (Selye, 1976). Despite these observations, the functional relations between LAS and GAS are not well understood and only few indications on the nature of the first messenger involved in stress response are known.

1.2.4) Modulation of stress response

Response to general stress factors exposure has been clearly defined. However, several parameters are capable to modulate its manifestation. This is the origin of response variations observed between individuals, species and stressor natures.

Stress response is an ubiquitous process among species and individuals. GAS has been observed in different vertebrate organisms and always follows a stereotypical pattern. However, despite the conservation of this mechanism, individuals' response is variable in dynamic and intensity. Depending on individuals or stressor nature, stress response can stand for hours to days and be more or less pronounced among individuals as illustrated by exacerbated allergic response occurring in some people. These divergences are partly due to what Hans Selye referred as internal and external conditioning.

Internal conditioning is due to modifications performed by endogenic factors such as heritable processes (i.e. genetic predispositions), past experiences (i.e. epigenetic), sex or age. In addition, age also plays an important part in stress response conditioning as it has been shown that the capacity for re-establishment of homeostasis progressively declines with age (Ewald, 2018).

Conversely, external conditioning is due to modifications caused by exogenic and environmental factors, microbial and viral challenges, climate, diet, drugs, and pollution being some examples.

Complementary to conditioning, the other main modulators of stress response are stressor's specific effects. Being non-specifically induced, stress response is believed to be independent of the stressor nature. However, depending on stressor properties, specific effects can modulate the classical stress response profile or even completely hide some non-specific effects. If a stress response is induced by injection of insulin, the expected stereotypical effect of stress-induced increased glycaemia will be counteracted by the insulin specific effect that consists in decreasing blood sugar concentration.

The physiological flexibility of living organisms is an essential feature of their resilience. This term is increasingly used and mentioned in various fields from engineering (Lundberg and Johansson,

2015) to ecology and neurobiology (Feder et al., 2019). However, despite its common use, the word "resilience" represents a complex situation that is still poorly defined, sometimes explained by contradictory interpretations and lacking a universally accepted definition in the scientific literature (Aburn et al., 2016). In stress biology, resilience is generally employed to describe the ability of living organisms to respond appropriately to pertubations by resisting damages and setting a suitable biological response. This capacity to face a challenge usually consists in trying to maintain or at least to recover a pre-perturbation state, a strategy referred as homeostasis. In some complex situations, an alternative solution is to explore new set points to reach a novel and more resilient equilibrium called allostasis. This strategy is physiologically more costly and depends on the capacity to endure long-term consequences of the adaptation, referred as the allostatic load. Depending on the stressor nature, intensity, persistence and the general organism state, the allostatic load accumulates at a variable speed and can result in an allostatic overload: a failure to adaptation due to sustained activation of regulatory mechanisms (Baffy and Loscalzo, 2014). In humans, it has been demonstrated that such chronic exposure to stress promotes the development of diseases such as Alzheimer disease (Tönnies and Trushina, 2017), cancer (Moloney and Cotter, 2018), diabetes (Jha et al., 2016), and cardiovascular problems (Aldosari et al., 2018).

1.3) At the cellular level: the oxidative stress

Hans Selye's work had a major impact in the Biology field, not only in health Science or clinic (Rice, 2012). Following his publications, numerous works were undertaken to understand how these notions translate at the cellular and molecular levels and to identify molecular processes or pathways representative of the physiological descriptions made at the individual level.

Similarly to the physiologic phenomenon described by Selye, oxidative stress suffers from a semantic problem, as it is a term widely used, often misunderstand. It was initially defined by Helmut Sies as a "perturbation of the pro-oxidants to antioxidants balance in favor of pro-oxidants leading to, potentially, damages" (Sies, 1997). This concept relies on the maintenance of a dynamic equilibrium between oxidative and reductive reactions in cells, named the redox homeostasis, and oxidative stress is characterized by a shift towards an increased oxidant potential with toxic cellular effects. However, based on more recent findings, this definition has been updated by Dean Jones in 2006 and stated as "a disequilibrium between oxidants and antioxidants in favor to the firsts, leading to a perturbation of the redox signaling control and/or to molecular damages" (Jones, 2006). In this new concept, oxidative stress represents a disruption of redox signaling and control.

1.3.1) Oxygen, the initial problem

Since life appeared on Earth, the ecosystem has undergone significant environmental changes. Around 600 million years ago, the dioxygen (O_2) concentration in the atmosphere dramatically increased and finally reached today's level of 21% (Figure 6) (Eaton, 2006; Fischer et al., 2016). Dioxygen is a major driver of the equilibrium between oxidation and reduction reactions as it forms free radicals generated by electron acceptance. Due to the presence of one or two unpaired electrons, the superoxide radical ($O_2^{\bullet -}$) and peroxides (O_2^{2-}) react with water to produce hydroperoxyl ($HO_2^{\bullet}$), hydroxyl ($HO^{\bullet}$) or hydrogen peroxide (H_2O_2). According to their capacities to react with other organic molecules, these compounds are unified under the term of "reactive oxygen species" (ROS). In addition, the presence of O_2 favored the apparition of additional "reactive nitrogen species" (RNS), such as nitric oxide ($\cdot NO$) and peroxynitrite ($ONOO^-$).

Both ROS and RNS are generated by harmful environmental factors - UV light, ionizing radiations or toxics (smoke, chemicals, drugs, pollutants) - but also many cellular functions using O_2 molecules to perform biochemical reactions. One of these key reactions in aerobic organisms is the oxidative phosphorylation that takes place in mitochondria in eukaryotes and at the cellular membrane in prokaryotes. During this process, electrons are transferred from donor to acceptor proteins through a cascade of complex enzymatic redox reactions coupled with protons transport to the intermembrane space (Cadenas, 2018). Finally, electrons end up in O_2 molecules that are converted to water by the cytochrome c oxidase (also called complex IV), while ATP synthesis is permitted by the generated gradient of proton. However, leak in the electron transport chain can occur and transfer of electron to O_2 leads to production of the superoxide radical $O_2^{\bullet -}$. Other sources of cellular ROS are (i) the auto oxidation of haemoglobin ($Hb(Fe^{2+})\bullet O_2 \rightarrow metHb(Fe^{3+}) + O_2^{\bullet -}$), (ii) the Fenton reaction, involving redox-active labile iron, ($Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + HO^{\bullet} + OH^-$), and (iii) several reactions catalyzed by enzymes, such as NADPH oxidase, superoxide dismutase, myeloperoxidase, nitric oxide synthase.

Accumulation of O_2 and other highly reactive species derivatives leads to excessive oxidation and disruption of redox homeostasis. This in turn induces cellular damages, resulting from reaction with all major cellular components (Eaton, 2006; Halliwell and Gutteridge, 2015). Different ROS/RNS toxicity relies on their oxidative reactivity, together with their half-life within the cell (Table I). Therefore, in an O_2 containing environment, organisms had to evolve and develop new

strategies to integrate the properties of this factor. Some mechanisms appeared to counter harmful modifications induced by ROS, although others appeared to take advantages of these highly reactive species. These choices have led to the paradoxical situation in aerobic organisms: O_2 is an environmental poison indispensable for life (Davies et al., 2017).

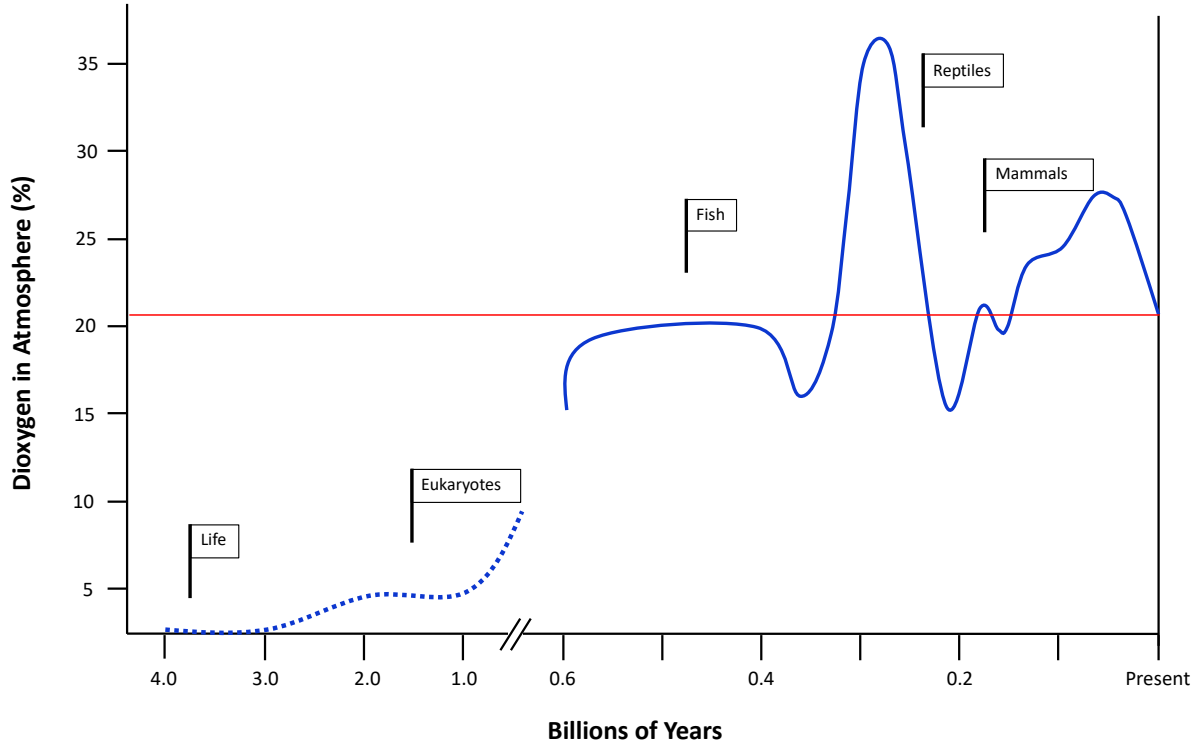


Figure 6 – Appearance of dioxygen in atmosphere is a major challenge for life. Evolution of dioxygen (O_2) concentration in Earth atmosphere. Life appeared on Earth when the dioxygen rate in the atmosphere was almost null. Around 600 million years ago, the dioxygen content in the Earth atmosphere dramatically increased to reach around 20%. Since that time, the dioxygen percentage in the air varied in a range from 15 to 35%. Red line depicts the actual dioxygen content in our atmosphere. Adapted from (Olson, 2012).

ROS	Symbol	Half-life	Properties
Superoxide radical	$O_2^{\cdot -}$	10^{-6} s	Poor oxidant
Hydroperoxyl radical	$HO_2^{\cdot}$	-	Stronger oxidant than $O_2^{\cdot -}$
Hydrogen peroxide	H_2O_2	minute	Oxidant, diffuses across membranes
Hydroxyl radical	$OH^{\cdot}$	10^{-9} s	Extremely reactive, diffuses only to very low distance
Alkoxyl radical	$LO^{\cdot}$	10^{-6} s	Less reactive than $OH^{\cdot}$, but more than $ROO^{\cdot}$
Peroxyl radical	$LOO^{\cdot}$	10^{-2} s	Weak oxidant, highly diffusible
Singlet oxygen	$^1O_2$	10^{-6} s	Powerful oxidizing agent

Table I – Characteristics of most abundant dioxygen radicals. Dioxygen derived free radicals and reactive oxygen species have a high potential to react with biological macromolecules according to their residing time in the cell and their oxidant properties. Alkoxyl and peroxyl radicals correspond to oxidized-lipid (L) entities.

1.3.2) Disrupting the redox homeostasis: the oxidative stress

In a context of high oxygen concentration, ROS have severe impacts on oxidation and reduction reactions possibly disrupting the redox homeostasis. When this equilibrium cannot be maintained and actual redox status overpasses a certain threshold, cells undergo oxidative-stress.

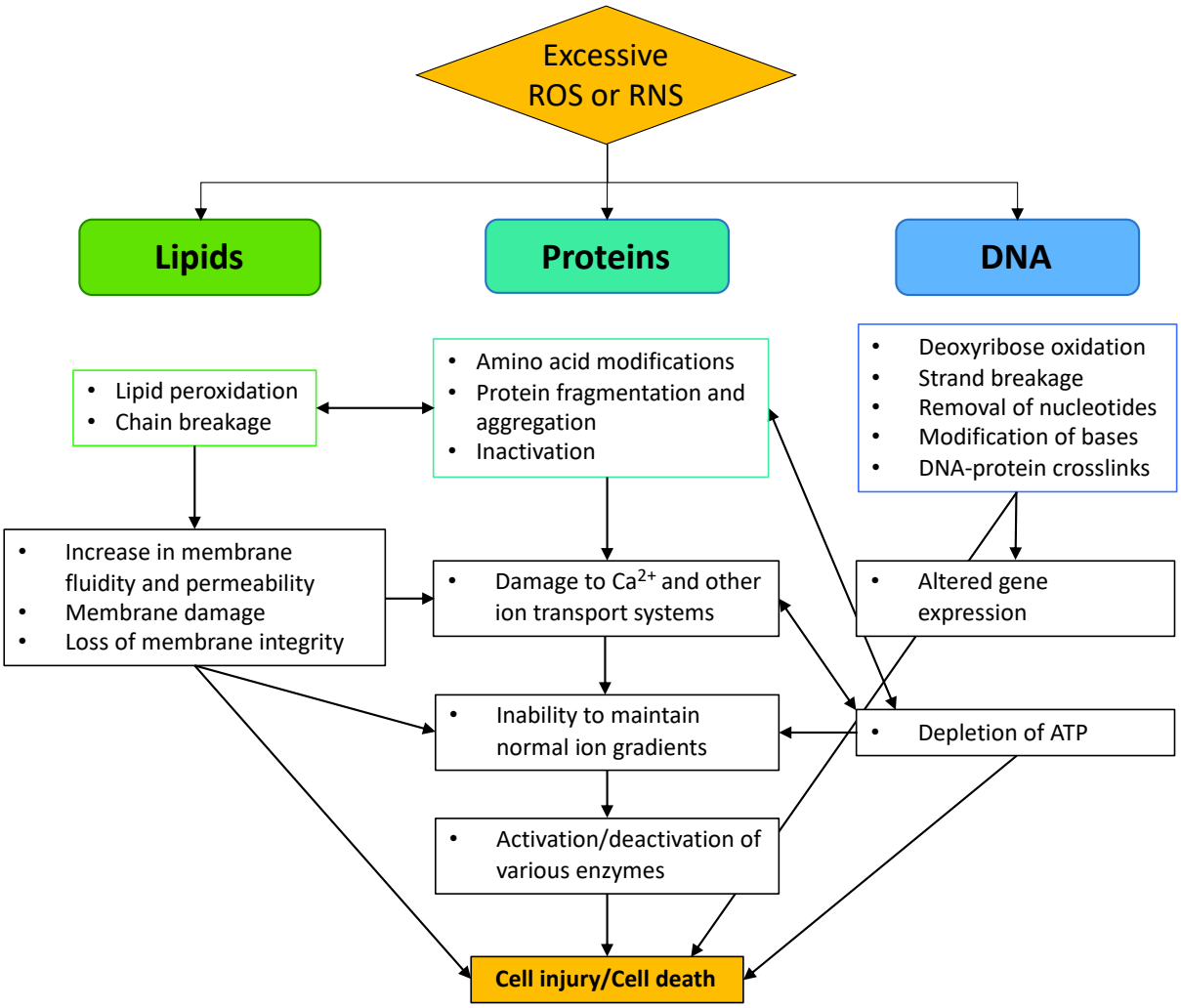


Figure 7 – Mechanisms of molecular and cellular injuries mediated by ROS and RNS. At high concentration, ROS can cause oxidative damages to all principal cellular components, thereby affecting cell functions. When ROS damages are too important, cells engage a suicide process: apoptosis or necrosis. Adapted from (Sharma et al., 2012).

Due to their high reactivity, ROS can modify DNA, proteins and lipids, thereby affecting their biological activities (Figure 7). DNA oxidation can lead to single or double strand breaks, and to nucleotide modifications (Cadet and Wagner, 2013). In one well-documented reaction, DNA oxidation targets the guanine nucleotide and leads to formation of 8-oxoguanine. Because of its alternative conformation, 8-oxoguanine induces an unusual pairing with an adenine nucleotide during DNA replication, consequently leading to its substitution by a thymine. Such mutations can be responsible for protein sequence alteration or modification of gene expression, by degradation of DNA motifs recognized by transcription factors within gene promoters. In addition to DNA, reactive

species also alter lipids. Main components of cell membranes, lipids play a key role in the maintenance of cell integrity. Because of the presence of several oxidizable double bonds in their lateral chain, polyunsaturated fatty-acids (PUFA) are the preferred targets of oxidation reactions. Referred as peroxidation, oxidation of PUFA molecules is a propagative process composed by three phases: initiation, propagation and termination (Gaschler and Stockwell, 2017). During initiation, hydroxyl and peroxyl radicals generated by metabolism or the Fenton reaction, oxidize PUFA, producing unstable fatty-acid radicals that immediately react with O_2, leading to the formation of alkoxyl- or peroxyl- fatty-acid radicals (Table I, $LO^{\cdot}$ and $LOO^{\cdot}$).

The propagation step consists in the reaction of this radical with the surrounding fatty-acids. Thereby, a unique oxidation event can cause propagation of lipid oxidation along a lipid membrane. Finally, the propagation can be interrupted by antioxidant molecules or enzymes (e. g. phospholipid glutathione peroxidase), when two radicals react together, forming a covalent bond. In addition to fatty acids, cell membranes contain proteins susceptible to oxidation as well. When free radicals react with proteins, they induce amino acids modifications such as carbonylation, intra- and inter-molecular crosslinks and formation of protein crosslink by dityrosine (Dalle-Donne et al., 2003). These reactions lead to proteins misfolding or cross-linking, disabling their biological function or at least modifying their catalytic capacities (Stadtman and Levine, 2003). Moreover, when proteins are unfolded, hydrophobic residues get exposed, promoting protein aggregation or increasing their susceptibility to be recognized and degraded by the proteasome. Altogether, oxidation of these cellular components has major impacts on central cellular activities resumed in Figure 7.

Similarly to the physiologic stress response, cells adapt their behaviour and metabolism during oxidative stress. These modifications have different outcomes depending on the cell type or the tissue affected, but a stereotypical response takes place that can be considered as a "cellular general adaptive response". In an optimal situation, cells can repair the damaged components to avoid the deleterious effects using the antioxidant system and other dedicated repair mechanisms (see 1.5.2). In case of a most severe situation, depending on the intensity and duration of oxidative stress, cells can go through five different states (Halliwell and Gutteridge, 2015):

- Stimulation of proliferation, only takes place at low oxidative stress intensity,
- Activation of adaptive pathways such as overexpression of defence systems,
- Alteration of cellular component functions due to oxidative damages,
- Survival but inactivation of dividing capacity (also called "senescence"),
- Activation of cell death by necrosis or apoptosis.

Similarly to the "fight-or-flight" strategy occurring during LAS, cells that initiate oxidative stress defence seem to follow a "sink-or-swim" method. First, cells try to replace their damaged components to maintain as much as possible the basic functions. When this solution is no longer possible, cells can still survive but with non-repairable damages. If the redox equilibrium is disrupted for a prolonged period, accumulation leads to general impairment of the whole cellular metabolism, severely affecting cell fate and activating a programmed suicide mechanism.

1.4) Cellular roles of redox reactions

Despite all their negative effects when present in excessive concentration, ROS, RNS and in more general free radicals are nevertheless essentials. When their production and activity are correctly regulated in cells, free radicals perform indispensable biological functions. Notably, they are central in defence mechanisms against pathogens (Lorenzen et al., 2017) as well as for signaling processes (Zhou et al., 2019). Therefore, contrary to what is now obvious in the public mind, if some antioxidants are good, more antioxidants are not necessarily better and can trigger reductive stress (Pérez-Torres et al., 2017): it has been shown that an excess of antioxidants is deleterious to the immune system, notably by impairing T-cell activation (Lorenzen et al., 2017), or to stimulate tumour progression (Hawk et al., 2016). To highlight and precise the essential role of redox reactions in biological systems, Dean Jones and Helmut Sies have proposed the notion of “Redox Code” (Jones and Sies, 2015). Similar to the genetic code, the redox code aimed to establish how oxidation and reduction reactions control metabolic and signaling pathways in living organisms. This concept has been stated as a list of four principles: (i) the metabolic organization, (ii) the linkage of metabolism to structure, (iii) the redox signaling and spatiotemporal differentiation, and (iv) adaptation to the environment.

1.4.1) Redox control in metabolism organization

Many vital metabolic pathways are intimately related to redox reactions involving electron transfers. Enzymes catalyzing these reactions are assisted by small molecules to capture released electrons: the so-called coenzymes. Among them, the nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP) are two systems based on the use of the NAD^+/NADH and $NADP^+$/NADPH couples. Maintained at a near thermodynamic equilibrium, both systems participate to metabolic processes, such as dehydrogenases coenzymes, including catabolic (i.e. the pentose phosphate) and anabolic (i.e. the gluconeogenesis) pathways. For example during glycolysis, glyceraldehyde-3-phosphate dehydrogenase (GAPDH) converts glyceraldehyde 3-

phosphate to 3-phospho-glyceroyl phosphate using NAD^+ as electron acceptor. Such system is also involved in the tricarboxylic acid (TCA) cycle in which three dehydrogenases depends on NAD^+: the isocitrate, alpha-ketoglutarate and malate dehydrogenases. In the pentose phosphate pathway, $NADP^+$ is preferred as electron acceptor for the 6-phosphogluconate dehydrogenase, responsible for ribulose-5-phosphate production. Production of NADH and NADPH during these reactions requires their regeneration as potential electron acceptor by other oxidative reactions. Therefore, both metabolic and oxidative activities are inter-dependent.

The $NADP^+$/NADPH couple is participating in many reductive reactions as co-factors for thioredoxin- or glutathione-reductases for example, but is also involved in ROS production. In humans, seven membrane-bound proteins convert NADPH to $NADP^+$: five NADPH oxidases (NOX1-5) and two dual oxidases (DUOX1 and DUOX2) (Lambeth, 2004; Bedard and Krause, 2007; Lambeth and Neish, 2014). These proteins produce superoxide radicals in different cell compartments participating in local ROS signaling, important for apoptosis or proteins modifications, as well as defence against pathogens by producing oxidative bursts. In addition, DUOX proteins possess a peroxidase domain that can directly transform $O_2^{\bullet-}$ to hydrogen peroxide (H_2O_2) (Figure 8). Interestingly, NOX and DUOX enzymes seem to be conserved in animals as suggested by the presence of homologous proteins in the nematode *Caenorhabditis elegans* (Ewald, 2018). The protein sequence is weakly conserved between human and nematode's DUOXs (around 30% of similarity), but their architectures and functions are strikingly conserved with functional domains sharing 90% of similarity. Consequently, the catalytic function of the worm's Duox is similar to its mammalian counterpart and is capable to perform signaling reactions, as well as defensive mechanism to kill pathogens.

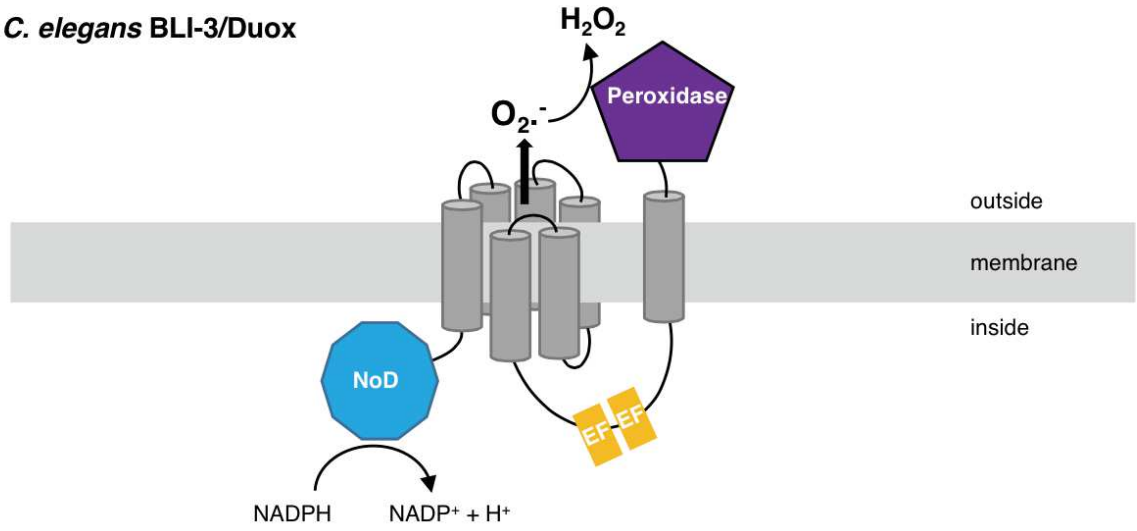


Figure 8 - DUOX protein topology. As in mammalian, *C. elegans* DUOX is a transmembrane protein. Conversion of NADPH to $NADP^+$ by NADPH oxidase domain (NoD) leads to $O_2^{\cdot-}$ production and its release through the membrane pore. In DUOX proteins, the additional peroxidase domain compared to NOX can then convert superoxide radicals to H_2O_2. Adapted from (Ewald, 2018).

1.4.2) Redox as a structural and functional switch

Other redox processes have an essential role for cells without being associated with a specific reaction or pathway. These processes act directly on proteins to modify their folding and consequently regulate their activity or their interactions with partners. Redox systems involved in this mechanism are considered as molecular switches and react with redox sensitive amino acids groups such as thiol group (R-SH), creating/breaking up intra- or inter-molecular disulfide bonds (R-S-S-R'). This activity is performed by thiol-dependent systems such as the cysteine/cystine (Cys/CySS), notably used by thioredoxin (TRX-SH/TRX-SS) and the tripeptide glutathione/glutathione disulfide (GSH/GSSG) among others couples (Figure 9) (Kemp et al., 2008). Four types of switch have been described: "on/off" switches that activate or inactivate a protein, allosteric switches responsible for the regulation of catalytic activities, thiolation switches orienting protein functions and interaction switches (Go and Jones, 2013). The paradigm of thiol switch controlling protein interactions is the case of the nuclear transcription factor NF-kappa-B (NFKB1) in which the DNA-binding capacity depends on modification of a cysteine residue (Toledano and Leonard, 1991). Interestingly, this thiol switch system has been identified also in prokaryotes, underlying its importance and conservation in living organisms (Hillion and Antelmann, 2015).

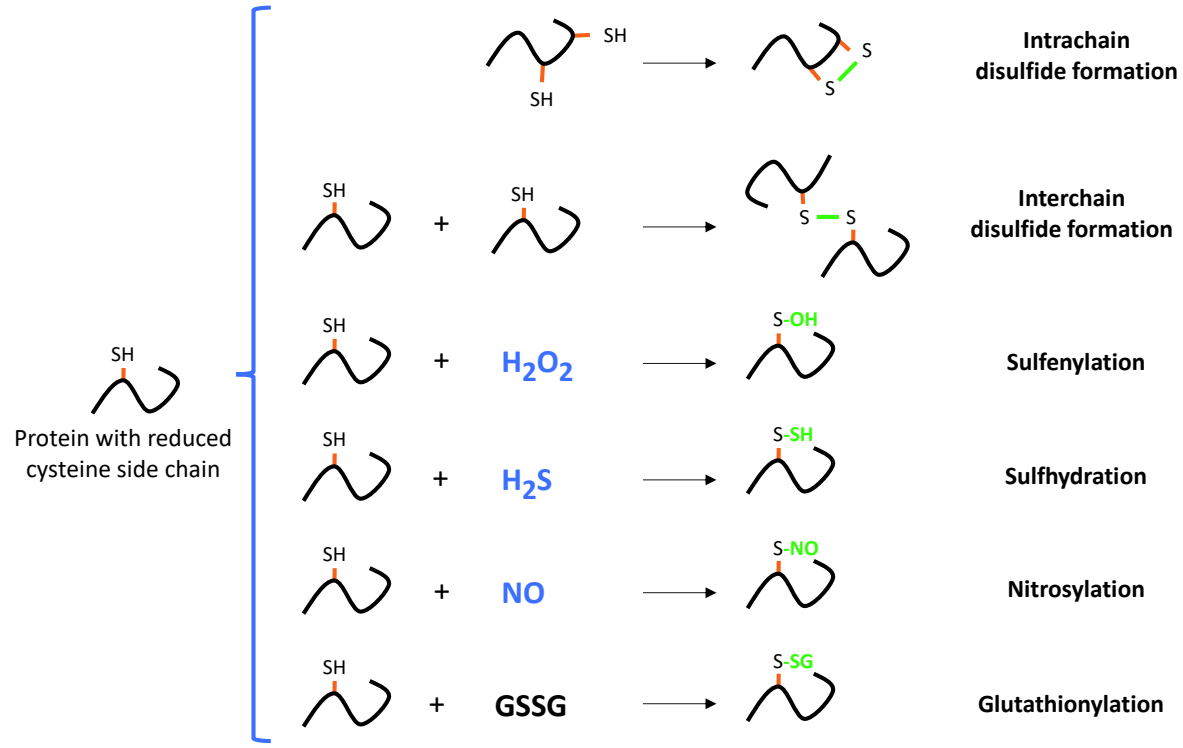


Figure 9 - Thiol systems. Thiol (R-SH) groups are highly reactive with their molecular environment. Cysteines can form internal disulfide bonds to modify a protein structure and external disulfide bonds to create proteins interactions. This reactivity of cysteine constitutes a switch system for activating/deactivating proteins or to create/break interactions. Reactive species are represented in blue. Adapted from (Ellgaard et al., 2017).

Of note, enzymes of the selenoprotein family – and among them thioredoxin reductase, glutathione peroxidase and methionine sulfoxireductase - play a central role in the regulation of the thiol system. Selenoproteins belong to a group of proteins that contain at least one selenocysteine amino acid, presenting a selenol group (R-SeH) in place of the thiol group in cysteine. This particular amino acid is co-translationally inserted into a specific set of proteins (25 in human), thanks to a dedicated translation machinery (Vindry et al., 2018). Presence of the selenium atom in selenocysteine confers specific catalytic properties to the enzyme that translate into increased reactivity compared to a cysteine homolog. In addition, presence of the selenocysteine was hypothesized to protect the enzyme of oxidation in case of oxidative stress, to preserve its catalytic activity (Reich and Hondal, 2016). Indeed, oxidation of selenocysteine is spontaneously reversible, while oxidized cysteine or methionine requires enzymatic-catalyzed reactions for their reduction. Therefore, selenoproteins are predicted to act as rescue enzymes, with preserved activity in condition of increased oxidative status.

1.4.3) Redox signaling and spatiotemporal differentiation

It exists another layer of complexity in the contribution of redox reactions to cell signaling, allowing space and time control. Cells components are subject to modifications that often result in a change of their properties or functions. Some of these modifications are permanent such as ubiquitination of proteins that triggers protein degradation. Conversely, other modifications such as phosphorylation are reversible and play a key role in signal transduction. Similarly, oxidation of cysteine thiols by endogenous metabolites and external substances is a reversible process. For example, oxidation/reduction of cysteine residues in actin controls its polymerisation/depolymerisation forms (Dalle-Donne et al., 2002; Farah et al., 2011). Moreover, thiol switches controlling cofilin oxidation level inhibits its interaction with actin and promotes its translocation to the mitochondria, where it stimulates apoptosis signaling (Klamt et al., 2009). As actin dynamics is essential for vital biological processes such as cell morphology, migration, growth or membrane trafficking, its direct and indirect regulation by redox reactions places thiol switch systems as a central signaling platform connecting intracellular with extracellular environments (Go et al., 2015). Similarly, activation/deactivation cycles of hydrogen peroxide (H_2O_2) metabolism was shown to support complex time-dependent processes controlling development of organisms.

Within the cell, redox reactions are spatially organized in two ways. Eukaryotic cells are divided in specialised compartments, containing a set of proteins catalyzing specific reactions. Membrane permeability being specific for different ROS, cellular repartition permits local independent activation of enzymes control within the cell. For example, oxidation state of TRX differs among major compartments within the cell, with oxidation status decreasing from endoplasmic reticulum, to cytoplasm, to nuclei, and being the lowest in mitochondria. A second way is the metabolic compartmentation, such as partitioning of H_2O_2 metabolism between different enzymatic systems. The knowledge of peroxide metabolizing systems has progressed with sequential discovery of catalase, selenium-dependent GSH peroxidase, selenium-independent GSH peroxidases (glutathione transferases), and peroxiredoxins (see 1.5.2). Several lines of evidence indicate that in mammalian cells, catalase has little contribution to peroxide metabolism outside the peroxisomes. Partitioning of metabolism between the three remaining systems was inferred from GSSG efflux after infusion of diamide in hepatocytes, resulting in irreversible oxidation of the thiol system. It showed that protein thiol oxidation constituted 73% of the net thiol oxidation while GSH oxidation was only 27% of the total. Thus substantial protein oxidation can occur without depletion of GSH, indicating that protein and GSH oxidation occur independently. This result showed that most of the reductive activity was independent of the two GSH peroxidases, mainly relying on the activity of peroxiredoxins (Tribble and Jones, 1990).

1.4.4) Redox as a structured network

The complex interplay between redox actors and their many different targets constitutes an intricate intracellular network controlling essential metabolic and signaling pathways in response to environmental signals. Thanks to the switch capacities of thiol systems, redox reactions are responsible for the fine-tuning of protein's activities, including enzymes and transcription factors (HIF-1 alpha, NRF2, see 1.5.3). Other redox actors, also participate to signal integration, for example by mediating insulin signaling as H_2O_2 production is increased in presence of extracellular insulin (Szypowska and Burgering, 2011). Also, during wound healing, cells activate a modification of their shape based on calcium transport, ATP and H_2O_2 production (Cordeiro and Jacinto, 2013). From the tissue viewpoint, local production of ROS by neutrophils sent to the invaded place to kill pathogens also assures signaling to the neighbouring cells (El-Benna et al., 2016; Glennon-Alty et al., 2018). Finally, oxidative stress is predicted to be an active condition in many diseases (Table II). For example, redox reactions are believed to play a key role during aging. Older organisms present a higher rate of carbonylated proteins, genomic and epigenomic alterations, inducing deregulation of metabolic processes, mitochondrial dysfunctions and even disturbed cellular communication (Go and Jones, 2017; Ewald, 2018).

Respiratory system	
	Inhalation of oxidants (SO_2, NO_2, O_3)
	Smoking
Brain	
	Alzheimer's disease
	Parkinson's disease
	Amyotrophic lateral sclerosis
	Down syndrome
	Traumatic injury
Cardiovascular system	
	Atherosclerosis
	Ischemia-reperfusion injury
	Myocardial infarction and heart failure
	Selenium deficiency (Keshan disease)
Skin	
	Ionizing radiation
	Thermal injury
	Porphyria
	Photosensitizers and other reagents
Muscle	
	Over exercise
	Muscular dystrophy
Others	
	Aging
	Cancer
	Cataracts
	Diabetes mellitus
	Inflammatory and autoimmune diseases
	Liver damage by endotoxins or halogen derivatives
	Kidney diseases/disorders
	Viral infections (AIDS)

Table II – Major diseases and disorders related to ROS/RNS. Because ROS and RNS are ubiquitous molecules, they interfere with various biological processes. They are consequently linked to many different diseases and disorders.

1.5) Maintaining redox homeostasis

This network constitutes an ideal biological structure sensing and responding to cell exposure to its environment and during its whole life, also defined as the exposome. To maintain redox homeostasis, cells deploy several strategies depending on the organism nature and on situations. Three main strategies have been established: flee the environment of high ROS concentration, decrease its endogenous production or increase antioxidant capacities.

1.5.1) Avoiding or reducing ROS exposure

To avoid perturbations of the redox homeostasis, particularly in situation of saturating oxidation, unicellular organisms such as bacteria, archaea or single-cell eukaryotes can just move away from external ROS production sources. Such solution still involves the use of ROS sensor systems to orient the organism. For multicellular eukaryotes, the situation is more complex, since oxidative stress generally constitutes an endogenous condition. In this situation, ROS exposure can be decreased by inhibition of its endogenous production reactions. However, because redox systems are interconnected to metabolism control, such inhibition can have deleterious effect.

1.5.2) The roles of antioxidants

Alternatively, cells can use antioxidant agents that are defined as “any substance capable to delay, avoid or repair molecular oxidative damages” (Halliwell and Gutteridge, 2015). These substances act by different ways: removing reactive species, protecting other biomolecules or preventing their oxidation. Depending on their location in cells and mode of action, these different activities are performed either by enzymatic or non-enzymatic agents (Eaton, 2006).

The non-enzymatic molecules generally consist in low-molecular weight entities scavenging ROS (Table III). Thanks to the reactive capacity of the thiol group present in cysteine, a consequent number of antioxidants consist in cysteine-containing molecules. Glutathione is one of these molecules, a tripeptide composed by a glutamate, cysteine and glycine amino acids. When oxidized, its cysteine can form a disulfide bridge with another glutathione molecule, thereby protecting other cell's components from oxidation. Other kind of antioxidants does not contain cysteine such as vitamin C (also referred as ascorbic acid), an enzymatic cofactor obtained from the diet in humans. This molecule is essential for the activity of antioxidant enzymes such as peroxidases, but also acts as an independent scavenger of free radicals. Being water-soluble, vitamin C is found in the cytosol and acts on ROS and RNS thanks to its two ionizable hydroxyl (R-OH) groups. Similarly, vitamin E refers to a group of fat-soluble compounds (tocopherols and tocotrienols) obtained from the diet and playing an antioxidant role. Thanks to its lipophilic properties, vitamin E plays a major role in cell

membrane integrity by blocking peroxidation propagation through its peroxyl-fatty-acid radical and its scavenging capacity.

In opposition to their non-enzymatic counterpart, antioxidant enzymes are high-molecular weight entities aimed to reduce reactive species (Table IV). Due to their enzymatic nature, these compounds are generally more efficient and specific catalysts of redox reactions (Figure 10). The two most often presented antioxidant enzymes are the superoxide dismutase (SOD) and the catalase (CAT). In humans, SOD exists in three forms that have a different location: SOD1 is localized in the cytoplasm and mitochondria, SOD2 is mitochondria-specific and SOD3 is found in the extracellular space. These enzymes catalyze the production of H_2O_2 and dioxygen from two $O_2^{\bullet-}$ radicals. As hydrogen peroxide is also a reactive specie, it must be reduced by a second enzyme: CAT.

Endogenous antioxidants
Bilirubin
Glutathione and other thio-compounds (thioredoxin)
Uric acid
Coenzyme Q (Ubiquinone-10/Ubiquinol-10)
Lipoic acid
Melatonin
Sex hormones
2-oxoacids (pyruvate, 2-oxoglutarate)
Dipeptides containing His (carnosine, anserine)
Albumin (-SH groups)
Dietary antioxidants
Ascorbic acid
Vitamin E
Carotenoids
Flavonoids – plant phenols (catechin, quercetin…)
Synthetic antioxidants
N-acetylcysteine (scavenger of ROS)
Deferoxamine (chelator)
Alopurinol (inhibitor of XO)
Acetyl salicylic acid (ferritin synthesis)

Table III – Major non-enzymatic antioxidants. Non-enzymatic antioxidants can be synthetized (endogenous antioxidants) or obtained from the diet. In addition, some synthetic molecules are also an antioxidant source used to treat patients.

Using four porphyrin heme groups, CAT converts two H_2O_2 molecules into harmless products: water and dioxygen. Other enzymatic systems occur in cells in addition to SOD and CAT, such as glutathione peroxidases (GPX), peroxiredoxins (PRX) and thioredoxins (TRX). The first one refers to a group of enzymes that use glutathione to reduce H_2O_2. Once oxidized, GPX enzymes are recycled by glutathione molecules, themselves reduced by glutathione reductase (GR) enzymes.

Peroxiredoxins and thioredoxins are two categories of enzymes that also work together. In a first step, peroxiredoxins reduce H_2O_2 molecules in water, leading to oxidation of their own cysteine. This step leads to the creation of disulfide bonds between two PRX, forming a PRX homodimer. Then, thioredoxins recycle peroxiredoxins by reducing their S-S bond. In a final step, thioredoxin reductase (TR) reduces oxidized TRX. Altogether, enzymatic and non-enzymatic compounds constitute the major mechanism of redox regulation in cells.

ENZYME	LOCATION
Superoxide dismutase	
Cu/Zn SOD (SOD1)	Primarily cytosol, mitochondria, nucleus
Mn SOD (SOD2)	Mitochondria
EC SOD (SOD3)	Extracellular fluid
Catalase CAT	Peroxisomes
Glutathione peroxidase GPX	Cytosol, mitochondria
Glutathione reductase GR	Cytosol, mitochondria

Table IV – Main enzymatic antioxidants. Enzymatic antioxidants display a specific repartition in different cellular compartments and allow the spatialization of redox responses.

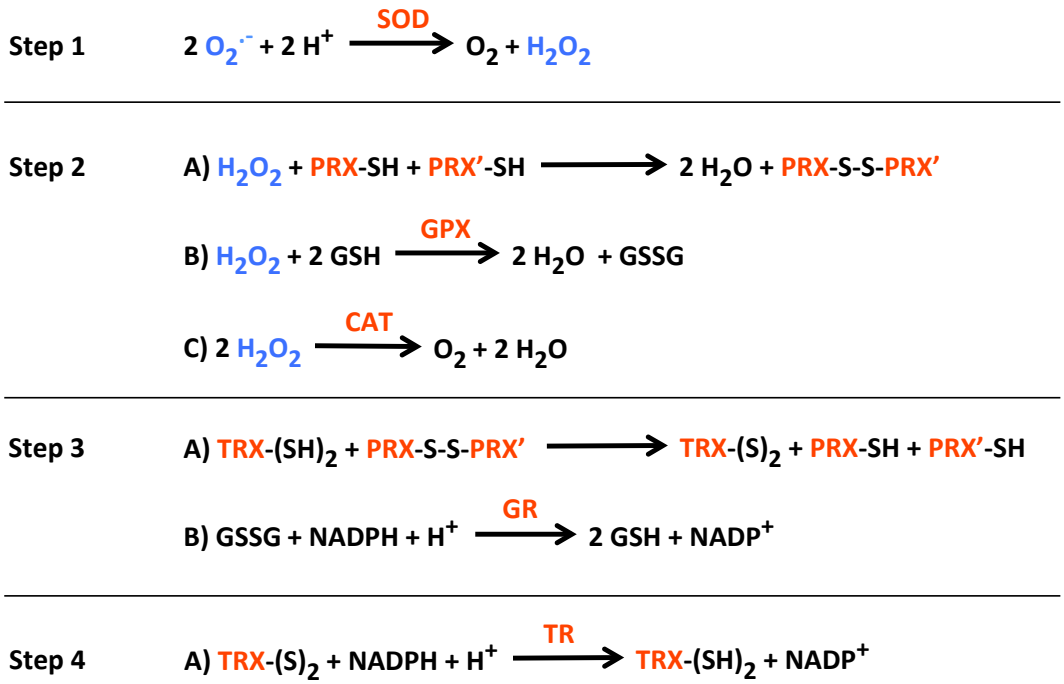


Figure 10 - Enzymatic antioxidant systems. In a first step, superoxide radical ($O_2^{\cdot-}$) is dismutated in hydrogen peroxide (H_2O_2) by superoxide dismutase (SOD) enzymes (represented in orange). Then, three enzymatic systems can manage this cellular ROS (represented in blue): A) the peroxiredoxin (PRX)/ thioredoxin (TRX)/ thioredoxin reductase (TR) system, B) the glutathione peroxidase (GPX)/ glutathione (GSH)/ glutathione reductase (GR) system and C) the catalase (CAT) system. For systems A) and B), the final step leads to $NADP^+$ production.

1.5.3) Oxidative stress sensing and control

In cells, oxidative stress defense is managed by transcription factors, proteins able both to sense oxidative signals and to regulate genes expression. Three transcription factors are known to be central components of the redox-disruption response: NF-kappa-B (discussed previously), HIF-1 (hypoxia-inducible factor 1) and NRF2 (NF-E2-related factor 2).

HIF-1, an obligatory heterodimer composed by two distinct subunits, respectively HIF-1 alpha and HIF-1 beta, was identified as an O_2 sensor. Under normoxia (a state of basic physiological oxygen concentration), the alpha subunit is modified by PHD enzymes (prolyl-hydroxylases), leading to its rapid ubiquitination and degradation (Movafagh et al., 2015). Conversely, under hypoxia (a state of reduced oxygen availability), PHD enzymes activity decreases, and HIF-1 alpha and beta subunits accumulate to form a functional heterodimer. It is still controversial, but it has been hypothesized that HIF-1 is also sensitive to high intracellular ROS levels induced by the hypoxic conditions. Oxidation of HIF-1 promotes its translocation to the nucleus where it binds DNA motifs called HREs for "hypoxia responsive elements", activating the transcription of related genes. Such activity mediates an adaptive metabolic response increasing the flux of enzymes of the glycolytic pathway, the serine synthesis and the folate cycle, while decreasing the TCA cycle turnover. These adaptations result in the production of antioxidant molecules and NADPH regeneration to counter the oxidative pressure (Semenza, 2017).

NRF2 is another system of cellular response to oxidative stress that is regulated by a protein heterodimer formation: KEAP1-CUL3 (kelch-like ECH-associated protein 1, cullin 3) (Yamamoto et al., 2018). In unstressed situations, the KEAP1-CUL3 complex interacts with NRF2, promoting its ubiquitin-dependent degradation through the proteasome. Upon a redox-disrupting stimulus, three cysteines of KEAP1, which is a thiol-rich redox sensor, are oxidized by ROS, thereby modifying its interaction with CUL3 and decreasing CUL3 ubiquitination activity. In this case, NRF2 is stabilized and gets translocated to the nucleus where it interacts with sMAF (small musculo-aponeurotic fibrosarcoma protein) to bind AREs (antioxidant responsive elements) motifs. Interaction of NRF2 with promoter regions activates the transcription of genes encoding cytoprotective enzymes, such as proteins involved in glutathione synthesis, thioredoxin reductase, peroxiredoxin, glutathione-S-transferase, NAD(P)H dehydrogenase and multidrug resistance-associated proteins. In addition to this regulation by ROS, NRF2 activity is also controlled in the nucleus through the PI3K-AKT signaling pathway. When the PI3K (phosphoinositide 3-kinase) phosphorylation activity is not stimulated, GSK-3ß (glycogen synthase kinase 3 beta) gets activated and phosphorylates NRF2. Consequently, NRF2 is recognized by the ß-TRCP/CUL1 (beta-transducin repeats-containing

protein/cullin 1) complex, resulting in its ubiquitination and subsequent degradation. These regulatory mechanisms based on thiol modifications and signal transduction are essential since it has been demonstrated that uncontrolled NRF2 activity promotes reductive stress, and is involved in many cancer types (Yamamoto et al., 2018).

1.5.4) The oxidative stress response through evolution

In a study of 2007 (Toledano et al., 2007), a genomic comparison was conducted to determine the degree of conservation of oxidative stress response components and mechanisms. For this purpose, the thiol redox system from the bacteria *Escherichia coli* and the yeast *Saccharomyces cerevisiae* were compared. Interestingly, two conserved response pathways were identified in these organisms: the GSH and the thioredoxin pathways. The GSH pathway consists in the glutathione peptide and the associated enzymes from the glutaredoxin family (Grx). In *E. coli*, four Grx enzymes exist: *GrxA, GrxB, GrxC* and *GrxD*. In yeast, this family contains five homologous enzymes: Grx1 to Grx5. Similarly for the thioredoxin pathway, the same enzymes exist both in the bacteria and yeast: *TrxA* and *TrxC* in *E. coli*, Trx1 and Trx2 in *S. cerevisiae*. This observation indicates that oxidative stress responsive proteins are conserved between distant organisms. However, the two pathways show a strong functional redundancy in bacteria while they display specialized activities in yeast. These data indicate that even if the factors are conserved, they can perform more or less specialized activities depending on the organism. This observation might explain species specificities in response to one stressor despite the involvement of the same general pathway. On the other hand, these results also support the idea that living organisms had to develop and to conserve stress response mechanisms to handle a highly oxidative environment.

1.6) Stress biomarkers

Stress is a state that can have severe impacts on the health of living organisms and even leads to death. Similarly, at the cellular level, oxidative stress can lead to cell disorganization and apoptosis, and consequently to tissue and organ damages. One of the main goals of stress research is to understand the biological mechanisms behind stress response and adaptation to propose better resilience solutions and treatments. However, to manage this state, in patients or animals, powerful markers are needed to detect stress issues and to respond with appropriated intervention.

A biomarker is any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction. In the medical field, biomarkers are of common use to anticipate or at least detect pathologies (Liu et al.,

2013). Biomarkers must be easy to measure, optimally using non-invasive techniques, for example through a blood test or urine collection. For disease diagnostic, these markers must be sensitive and robust enough to allow specific identification of a given disease. In comparison to observable symptoms, biomarkers have the advantages of being detectable before the manifestation of pathogenic signs. Indeed, as phenotypic (i.e. measurable) alterations have molecular origins, biomarkers detectability precedes the apparition of symptoms. This characteristic allows to act in anticipation and to facilitate the cure.

One major hurdle with the identification of stress biomarkers resides in the complexity of the biological process: (1) it involves a vast diversity of possible inductors and physiological targets (cells, organs and tissues); (2) It is a dynamic process that occurs through multiple phases with a wide degree of variability due to conditioning or specific effects of stressors, even if the response follows a stereotypical scheme.

1.6.1) Biomarkers of the physiologic stress

Hans Selye extensively described the physiologic stress response through its GAS and LAS manifestations. During GAS, three successive stages have been defined, each one with its own biological characteristics and its physiologic manifestations that are supposed to be ubiquitous and independent of the nature of the stressor. Hormones as blood circulating molecules, which are, by definition easily measurable, constitute ideal biomarker candidates. Some stress biomarkers have been previously proposed among circulating hormones, notably cortisol, ACTH (the adrenocorticotropic hormone), adrenaline, oxytocin or vasopressin (Covelli et al., 2005; Milivojevic and Sinha, 2018). In addition, local inflammation playing a major role in LAS, inflammation biomarkers such as local production of hydrogen peroxide or myeloperoxidase activity (the enzyme responsible of hypochlorous acid production to kill pathogens) constitute widely used markers (Marrocco et al., 2017).

However, these general markers show several limitations. Most of these molecules are only transiently detectable, ranging from few hours to several days depending on the stressor or the species considered. Consequently a fluctuating time window dictates their detection and general symptoms can be hidden even if the response is still on going. Moreover, many classical markers, such a high concentration of corticoids in blood, cannot differentiate if an organism is in a pre- or post-resistance state.

Regarding our current knowledge, it is still not understood why some organisms will respond better and faster than others, nor what determines adaptive capacities. To propose better biomarkers, it is needed to understand the biological mechanism behind stress response and adaptation. This could

help to understand the spatial and temporal specificities and to determine strategies for designing relevant markers.

1.6.2) Biomarkers of cellular oxidative stress

Several cellular biomarkers have been identified thanks to cell culture studies. They have the advantage of being easy to detect at different levels in cell cultures: expression of stress responsive genes, membrane lipids oxidation, proteins carbonylation or DNA damages. Indeed, as oxidative stress causes modifications and degradation of many cellular components, catabolism products are robust markers of an oxidative stress response: malondialdehyde (MDA), isoprostanes, hydroperoxides, oxidized low-density lipoproteins (LDL), hexanoyl-lysine are witnesses to the general accumulation of oxidized lipids; nitro-tyrosine and carbonylated proteins are indicative of oxidized proteins accumulation; and 8-hydroxy-2'-deoxyguanosine is a good indicator of DNA damages. Also, the activity of intracellular antioxidant systems are indicative of the setting of an oxidative stress response through superoxide dismutase, peroxiredoxins, thioredoxins, glutaredoxins, glutathione peroxidases and glutathione production rates (Frijhoff et al., 2015; Marrocco et al., 2017). Most of these enzymes correspond to genes which expression is controlled by the transcription factor NRF2 (Yamamoto et al., 2018).

However, the analysis of these markers cannot so easily be translated at the living organism's level, some of them requiring invasive methods such as biopsy and a special attention to avoid post-sampling oxidation. Moreover, even if lipid oxidation or protein carbonylation can be assessed by blood or urine test, they can fail to represent the oxidative damage state of a given tissue or organ. For example, oxidative damages in the brain could be undetectable by blood analysis. Finally, biological tissues are often constituted by an heterogeneous material composed of many different cell types, and the kinetic specificities of cell type dependent response are not well defined. Therefore, it is needed to understand the molecular basis of stress response in cells and its propagation to tissues, organs and whole organisms. With such a knowledge, it will be possible to identify more convenient biomarkers representative of stress response kinetic, stressor specificity, tissue specificity and ultimately extendable to a wide range of living species.

In conclusion, due to the inherent complexity of stress response mechanisms, it is predicted that no biomarker will, alone, be an indicator for all stressors and their effects. Some stress-specific biomarkers have already been defined at the physiological and cellular levels, but none of them is accepted as universal and current methods are neither informative about stress conditions, nor predictive of stress integration. A set of markers is likely to provide more exhaustive description of the condition and to cover most possible situations (Eline Slagboom et al., 2018) (Figure 11). To

identify a combination of markers, high-throughput technologies supported by bioinformatics tools are appropriated methodologies, as they allow conducting integrative investigations from intracellular molecules to whole organisms.

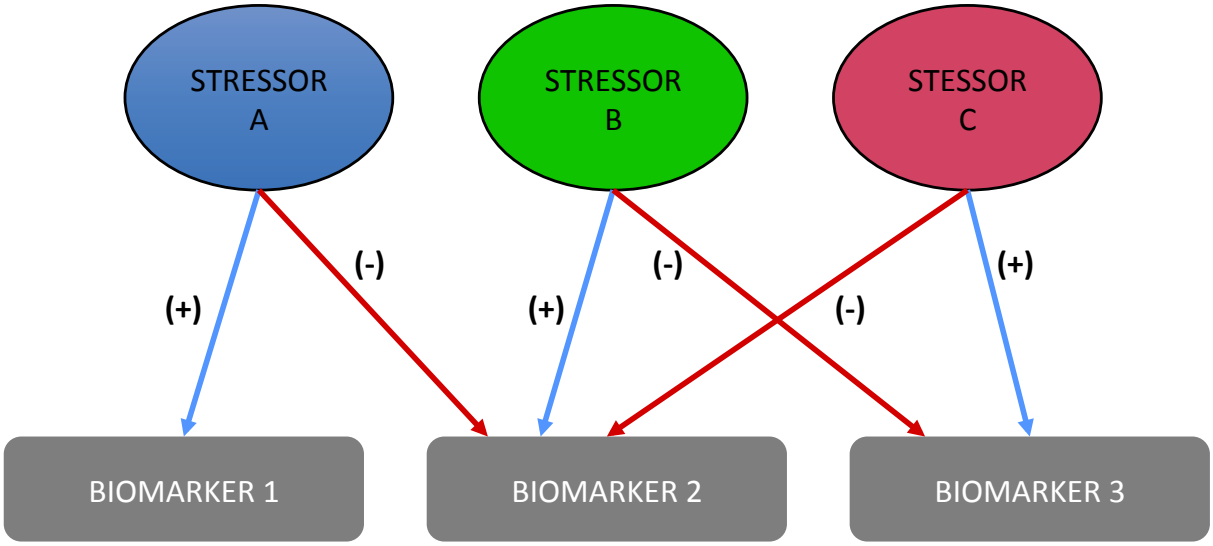


Figure 11 - Biomarkers for stress response. Stressor effects on biomarkers are complex. A given stressor can both up-regulate (+) or down-regulate (-) a biomarker that will become measurable or disappear. In such a complex situation, only the use of a set of several markers associated with a good knowledge of stress response can overcome this issue. Adapted from (Sanchez and Porcher, 2009).

1.7) Linking cellular to physiologic stress

One question that remains poorly addressed so far is how cellular stress response translates to a physiological response. Historically, scientists tried to define stress at the organism level because stress response was impacting the animal physiology. Then, with technological advances, it became possible to analyze stress response at the cellular and molecular levels. At that time, a fast shift occurred from generalist studies on whole organisms to detailed studies in cell cultures. These powerful analyses led to the identification of central factors implicated in stress response, such as the transcription factors HIF-1 alpha and NRF2, and provided detailed mechanistic insights at the cellular level. However, one difficulty in establishing a link between cellular and physiological stress originates from conceptual biases generated by experiments conducted on cell culture systems.

1.7.1) Cell culture biases

Most current cell cultures are composed of one single cell type grown in two-dimensions in contrary to tissues in living organisms. Accordingly, animal tissues are constituted of mixed cell types closely interconnected and communicating together.

Also, contrary to what happens in animals, cultured cells are not exposed to circulating hormones that establish a communication between distant organs. Consequently, phenomena observed during GAS such as temporal secretion of corticoids from adrenal glands to peripheral tissues does not occur.

Additionally to the gap between physiologic and cell cultures organization, cells in culture must also cope with a widely fluctuating O_2 concentration and a constant oxidative stress. Dioxygen concentration in atmosphere is 20.9% (140 mm mercury (Hg)). In laboratory condition, cell cultures are carried out in incubators containing of 18.5% O_2 and 5 % carbon dioxide concentrations. The only cells in the body that are exposed to an environment with 20% O_2 (110 mm Hg) are lung alveolar cells. Shortly after blood gets oxygenated, the O_2 level falls to 10.5-13% (80-100 mm Hg), and most organs function normally at O_2 levels ranging from 2-8% (19-70 mm Hg) (Figure 12). Moreover, within a given organ, substantial gradients can emerge and O_2 tensions within a given tissue can undergo temporal fluctuations after increased metabolic demand. Importance of this hypoxic gradient for tumour development, progression and resistance to therapy has been widely documented. Therefore, compared to its intra-tissue concentration, O_2 in the atmosphere causes a permanent oxidative stress for cultured cells (Ast and Mootha, 2019). This situation can drive natural selection in cells to develop or adapt metabolic and signaling pathways that are physiologically irrelevant. In addition, to compensate, cell culture media contain antioxidants in different concentration and composition compared to extracellular body fluids, corresponding also to an artificial environment. This problem is particular prominent in the study concerning oxidative stress defence, since O_2 concentration is a major component of ROS production.

These observations support the notion that, even if mammalian cell culture with supraphysiological O_2 tensions has led to transformative discoveries, cellular models are poorly representative of a physiological environment, especially concerning O_2 metabolism, and this could explain why the transposition of the data obtained from cell culture analysis to physiological models has failed in most cases (Ast and Mootha, 2019).

To minimize the biases of cell culture, it would be ideal to analyze molecular parameters at the full organism level. This is possible thanks to integrative approaches that allow to study a full system and without a priori. Using these approaches could lead to a precise understanding of stress response mechanisms occurring at the cellular level in a biologically relevant context.

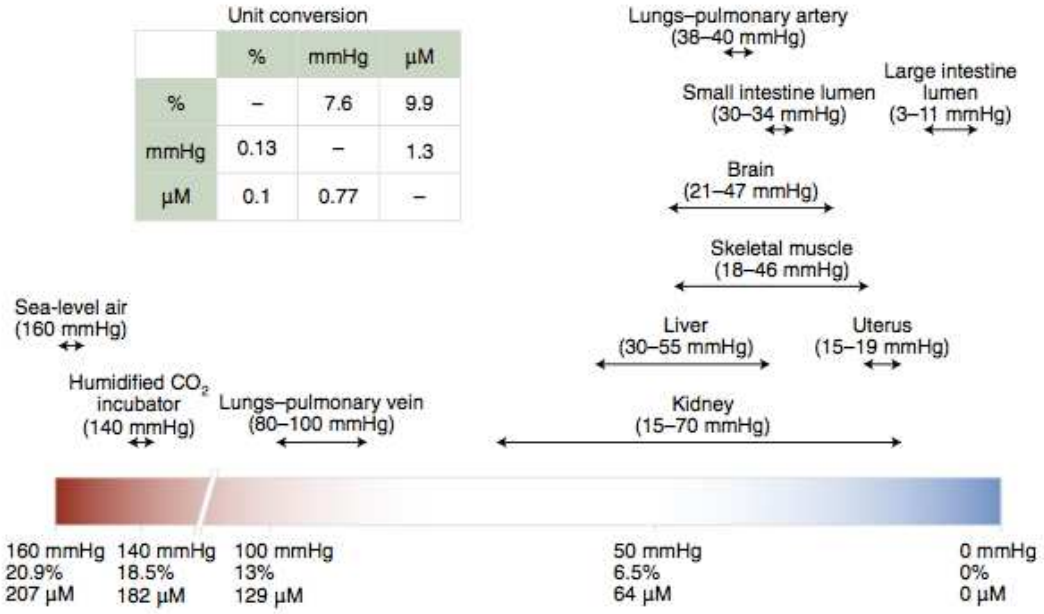


Figure 12 - Cell culture versus physiological system oxygen tensions. Cell culture is subject to many biologically irrelevant artefacts. In biological systems, organs are generally exposed to variable but generally lower dioxygen concentrations than in culture systems, depending on their biological activity and exposure to the environment. Taken from (Ast and Mootha, 2019).

1.7.2) Orthogonality of redox regulation

One possible explanation to how cellular redox processes translate an effect at higher levels of organization consists in considering the redox system as a whole rather than a sum of individual reactions. Indeed, as reported in a recent publication (Santolini et al., 2019), virtually all molecules can engage in electron transfer or redox processes, acting as electron donors or acceptors. Consequently, all these compounds may cross-react at any time and ,in response to multiple challenges, constituting a synchronized network referred as the "Redox Interactome" (Cortese-Krott et al., 2017). As all redox processes rely on a common ensemble of reactive species permanently interacting together, it results in the simultaneous superimposition of multiple activities occuring at any level of biological organisation: from molecular to physiological levels. Such regulation system orthogonal to many biological processes, including cell signaling, metabolism and bioenergetic pathways, offers the possibility for a constant re-adjustment of physiological parameters during an individual's life and in response to its environment. This regulatory mechanism is better described by the homeorhesis concept, a dynamic process involving the maintainance of an adjustable trend rather than the absolute need for set point conservation or homeostasis, by employing novel redox chemistry and biochemical processes in response to environmental perturbations (Waddington, 1968; Santolini et al., 2019).

1.7.3) Integrative approaches

Integrative approaches aimed to analyze living organisms as a whole and to integrate molecular and cellular information obtained on biological models such as *in vitro* experiments or cell cultures. It consequently intends to analyze complex but biologically relevant processes. One major limitation in this approach is the immeasurable complexity of biological systems under investigation, which requires to be organized and prioritized. Studying biological mechanisms from an evolutionary perspective is one possibility to focus on particularly interesting and relevant conserved functions. Anatomy and embryology are two fields of biology in which the comparison of similarities and differences between species have been applied with remarkable success. Comparative anatomy introduced the concepts of homologous and analogous structures, resulting from divergent (common ancestor) or convergent (similar environment) evolution.

Ageing is one example of a complex process. In a study of 2015 concerning ageing (Mansfeld et al., 2015), Ristow team searched for the impact of evolutionary conserved genes between the mouse *Mus musculus*, the fish *Danio rerio* and the worm *Caenorhabditis elegans* on life expectancy. Using a transcriptomic approach (RNA-seq), they investigated differential gene expression for these animals at three time points of their life. Doing so, they could identify 29 conserved genes up-regulated or down-regulated during life-time. To validate the importance of these genes during ageing, they used the RNA interference (RNAi) method in *C. elegans* to inactivate the candidate genes. Thanks to this technique, it has been shown that 12 genes are indeed able to extend the mean lifespan of *C. elegans* when their expression is repressed. One of these genes, *bcat-1* for "Branched-chain-amino-acid aminotransferase" (BCAT1 in human) has been particularly investigated as this gene was extending lifespan, also preserving the best vital aptitude, when inhibited. This gene encodes a protein important for the catabolism of branched-chain-amino-acids (BCAA), suggesting that BCAAs play an important role in lifespan regulation. Interestingly, overexpression of *bcat-1* was shown to decrease lifespan but also fertility in *C. elegans*. This behaviour suggests that a low expression of this gene is a selective advantage independent of its relation with ageing.

This example demonstrates that considering evolution is an efficient way to analyze the complexity of biological systems. Indeed, living organisms have evolved during billions of years, leading to the emergence of various life forms each one with their own specificities and complexity, but also preserving central functions. Phylogenetic comparisons provide crucial information to distinguish environmental adaptation from conserved mechanisms and allow the characterisation of the most relevant processes in the light of evolution.

1.7.4) Transcriptomics studies of stress response

Transcriptomics studies are based on the measurement and comparison of gene expression levels using microarrays or RNA sequencing (RNA-seq) technologies. While microarrays consist in chips on which are anchored sets of probes to detect targeted RNAs, RNA-seq allows gene expression measurement at the whole genome scale, without the need for *a priori* assumptions. Interestingly, as it is possible to sequence RNAs from a tissue or even from full organisms rather than using cell cultures, it is a well-suited protocol to move from reductionnist to integrative approaches. Many transcriptomics studies were performed to gain a better understanding of stress response using different species exposed to different stressors and investigating a variety of tissues including in time course experiments: transgenic versus wild-type mice submitted to chronic mild stress (Wassouf et al., 2019), salmons exposed to heat stress (Shi et al., 2019), zebrafish exposed to multiple chemicals (Schüttler et al., 2017) and corals exposed to thermal or cold stress (Lee et al., 2018) are some recent examples of such studies.

In the course of this Thesis project and during the writing of this manuscript, I tented to explore the literature of transcriptomics analyzes related to stress in different models, to compare the information obtained. However, I came to the conclusion that this study was impossible to conduct. Similarly to reductionnist approaches, integrative methods come with their own biases and limits, including the diversity of analytical methods available (RNA-seq or microarrays, bioinformatics tools, significance thresholds…) as well as the variety of tissues, species and stressors studied, preventing accurate comparison. In addition, in the large majority of these studies, only the most strongly differentially expressed genes are considered to perform functional analyzes. Otherwise, set of genes are filtered by their involvement in a particular biological process of interest such as inflammation, development or cancer. However, comparison of such restricted lists, especially those containing only the most differentially expressed genes between different models, rarely converge to a set of common genes, because they mainly focus on the specificities of the different processes studied. In addition, comparison of genes from different species is often a complicated task, as the gene name nomenclature is variable between distant species and because genomes are often uncomplete or poorly annotated. This observation raised the importance of evolutive and comparative approaches to determine gene and protein equivalence between species and to identify central mechanisms conserved during evolution.

2) The evolution principle

Estimations propose that around 10,000,000 different species exist on Earth, representing the so-called biodiversity (Mora et al., 2011). In the current scientific community, there is no doubt that these species have emerged from an evolutive process. This theory of evolution states that all actual living beings originate from a common ancestor that gave rise to all the diversity we know.

2.1) The theory of evolution

Before the establishment of the theory of evolution as proposed by Charles Darwin in 1859, several different explanations on the variety of living forms co-existed, some of them still supported today. In that time, the age of Earth was estimated at around few thousands to 100 million years, supporting the idea that life appeared early and already with a high level of complexity.

2.1.1) Context

In Darwin's time, the influence of religion was omnipresent in society. Most people believed that all living beings were created by a divine entity and that their forms were unchangeable over time. Although this vision of life's history was deeply rooted in society, some people over the world considered that these assumptions were wrong. Jean-Baptiste de Lamarck proposed that animals could adapt themselves to their environment during their life and transmit these adaptive changes to the progeny (Jablonka et al., 1998). In this vision, living beings are not immutable but rather highly adaptable and doomed to change from a generation to the next one.

2.1.2) Establishment of the theory of evolution

In 1831, Charles Darwin, student at the Cambridge University, was selected to participate in a maritime cartographic expedition along the South American coast (http://darwin-online.org.uk/). This journey let him observe the diversity of plants and animals and probably constituted the roots of the theory he will propose few years later. Notably, he observed fossils strikingly similar to animals still living in South America, leading him to consider that extinct species gave rise to the actual ones. In the Galapagos islands, he noted the resemblance between animals and plants he observed with those present in South America. Closer resemblance with these organisms rather than with animals and plants from other parts of the world suggested the existence of a certain continuity among them. Following these observations, Darwin started to write a book that laid the foundations of modern biology: "On the Origin of Species" (Darwin, 1859).

In this book published in 1859, Darwin presented his vision of evolution with a naturalist viewpoint based on comparison of observations, a disruptive view regarding the ideas in place at that time about the hierarchical organization of life (Figure 13). In contrary to Lamarck who was focused on changes occurring during the life of an organism, Darwin extended the notion of evolution to a time scale that includes all living and past organisms. This suggested that, at the root of actual species existed ancestral organisms, and ultimately converging to one single common ancestor that today we call LUCA for Last Universal Common Ancestor (Koskela and Annila, 2012). To explain how species could originate and evolve from common ancestors, Darwin proposed a mechanism central to his evolution theory: the notion of natural selection.

2.1.3) The mechanism of natural selection

A proposed explanation to how such a variety of life forms emerged from ancestor was that organisms are submitted to a constant modification process. For Lamarck, this process occurred at the level of an organism to permit a specific and transmissible adaptation required in a precise environmental situation. In contrast, Darwin proposed a mechanism based on chance. He argued that within a same species, a set of random variations pre-exist and do not correspond to a specific need for adaptation. Then, he stated that considering these variations, some organisms would experience success or difficulties to face a given environmental context. Individuals showing advantages are generally stronger, live longer and are consequently able to reproduce more easily to transmit their positive variations. In contrast, organisms that are disadvantaged by their variations will be less capable to reproduce. Such system should result to a selection of advantageous traits and to the disappearance of the deleterious ones. Charles Darwin considered that this selective process named "natural selection" was what favored the development of the actual diversity among living beings. At that time, the major counter argument was that such selection process was incompatible with the estimated age of Earth and could not give rise to the observed variability of living forms. However, 20 years after Darwin's death, the discovery of radioactivity led to a revise dating of Earth formation at around few billion years. This duration was then fully compatible with the theory of evolution (Figure 14).

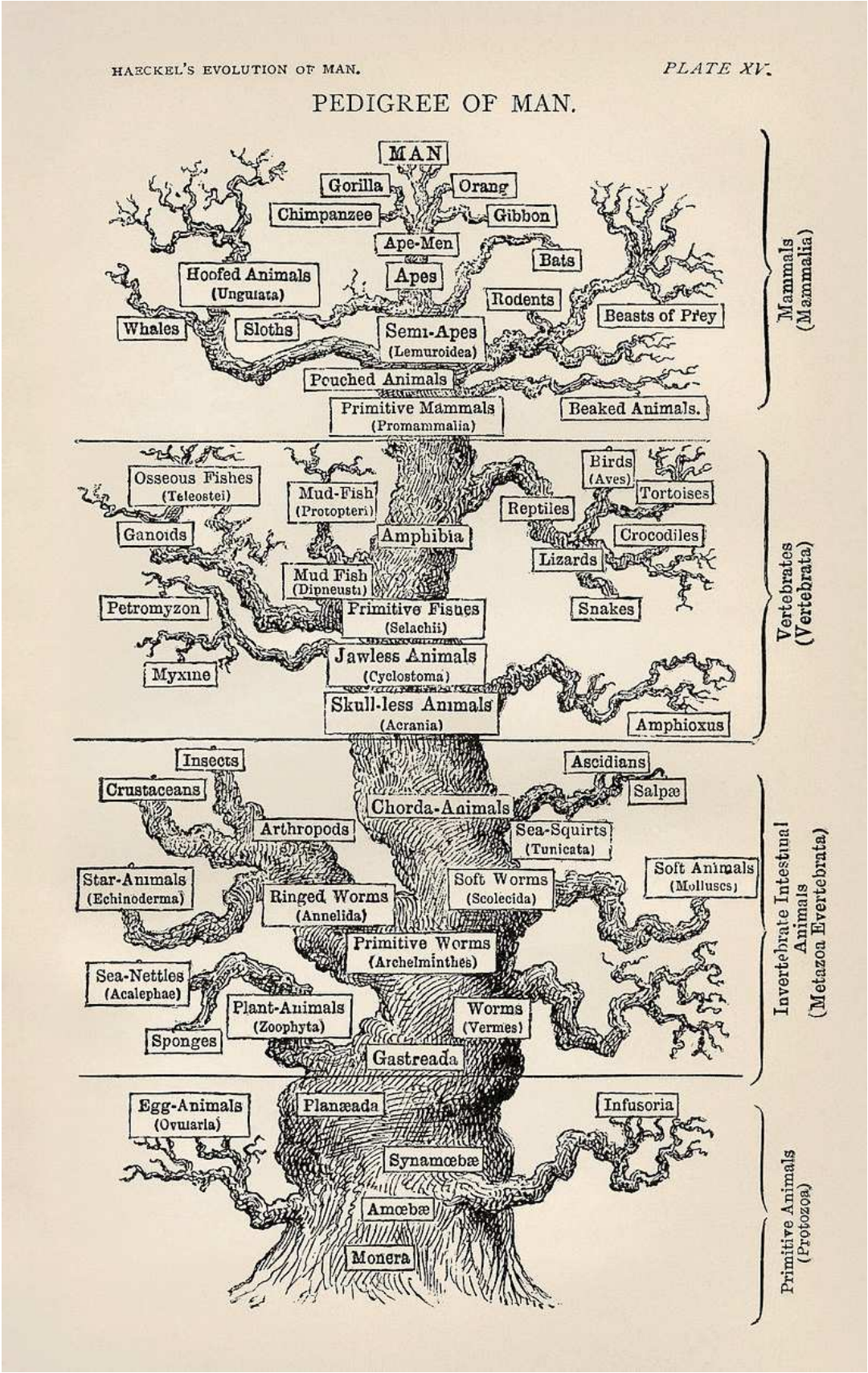


Figure 13 - Haeckel's tree of life. Even after publication of Darwin's evolution theory, many people were convinced that humans are the pinnacle of evolution. This representation from Ernst Haeckel (1879) places humans on the top of the animal evolutive tree. Adapted from (Gontier, 2011)

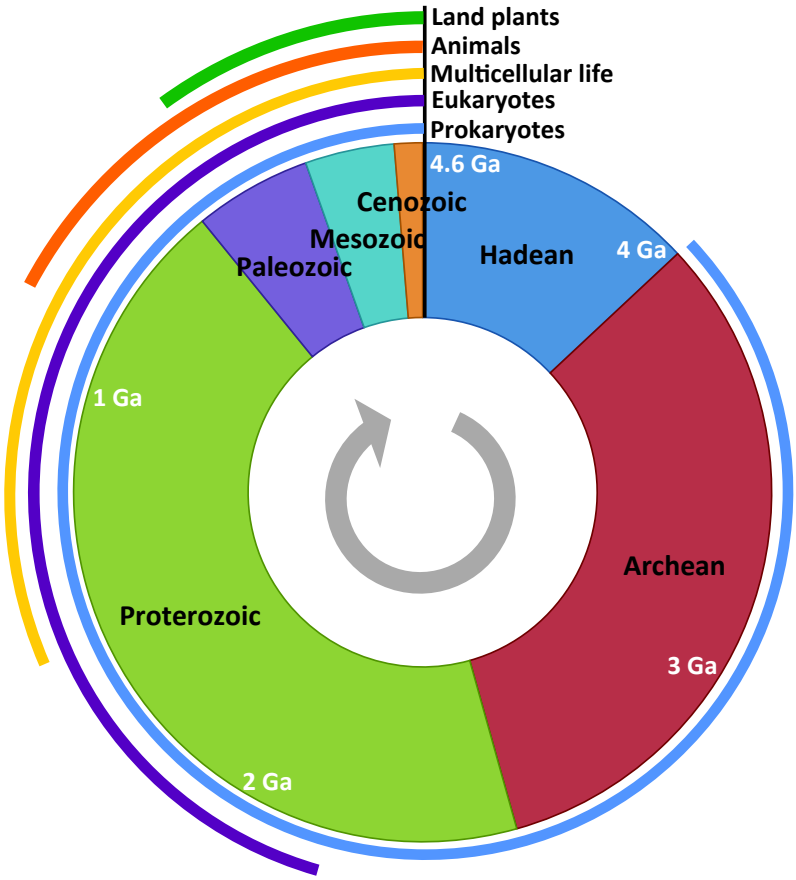


Figure 14 - Life at the geological timescale. Development of life through the evolutive process proposed by Darwin was difficult to believe because of the erroneous predicted age of Earth. Later, when geological timescale was specified (approximately 4.6 billion years), the theory of evolution became plausible. Coloured lines correspond to the apparition of some of the major life domains still present today. Years are represented in Ga: Giga years ago. Adapted from (Dias and Mattos, 2011).

2.1.4) Critics and development until today

At the publication of "On the Origin of Species", many people were disturbed by the idea that humans share a common ancestor with other species, including monkeys. Before the work of Darwin, humans were often considered as the pinnacle of evolution and systematically represented as the legitimate ruler of the living world (Figure 13). Considering that all living species share a common root, the theory of evolution was in total opposition with this worldview. Many criticisms appeared from people, even among scientists. One of the major objections was that the theory of evolution is based on transmission of variations. However, the precise mechanism of transmission and its biological support were not known. At that time, Hugo de Vries stated that random variations should led to the brutal apparition of new species, something inconsistent with the Darwinian gradualism exposed in his theory (De Vries, 1910a).

2.2) DNA, the molecular support of heredity

In 1889, Hugo de Vries introduced the notion of "pangene" (De Vries, 1910b) followed by its simplification proposed in 1905 by Wilhelm Johannsen (Johannsen, 1909), "gene", as the carrier of heredity. Inspired by the work of Gregor Mendel (1822-1884) on heredity, they postulated that inheritance of specific traits in living being is achieved by "particles". To establish the link between the ensemble of genes on one hand, and the ensemble of traits on the other hand, Wilhelm Johannsen coined the terms "genotype" and "phenotype" respectively (Johannsen, 1909). These concepts are still of main importance in today genetics, the science of gene variation and transmission in organisms. Concomitantly, in 1902 and 1903, Walter Sutton and Theodor Boveri discovered that transmission of genetic inheritance was achieved by a group of intracellular components, the chromosomes (Sutton, 1903). However at that time, their molecular nature was not clearly defined.

2.2.1) Discovery of the gene molecular support

During the 20s, progresses in the characterization of the cellular content led to the discovery of nucleic acids, a family of macromolecules subdivided in two types: one containing a ribose sugar, the other one containing a deoxyribose sugar. In 1919, Phoebus Levene specified the composition of these molecules constituted of polymers (believed to be short) of four repeated bases, linked by a sugar-phosphate backbone (Levene, 1919). The terms deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) were coined to differentiate the two types of molecules. However at that time, their roles were still unknown and no relation with the transfer of genetic information was made.

In 1928, Frederick Griffith conducted an experiment on mice infection that would lead to the molecular characterization of genes (Griffith, 1928). He worked with two strains of *Pneumococcus* bacteria: a pathogenic and a non-pathogenic one, inactivated by the removal of its lipopolysaccharide envelope, but containing an enzyme responsible for its toxicity. Surprisingly, Griffith showed that a mixture of dead pathogenic bacteria with alive non-pathogenic ones, which are separately harmless, led to mice death. After examination, Griffith could find pathogenic bacteria in the mixture and concluded that the information needed to synthetize the enzyme or the lipopolysaccharide envelope was transferred from the dead cells to the living ones. At that point, it was established that a genetic carrier, which nature was unknown, is able to carry information for protein synthesis.

In 1944, Oswald Avery, Colin MacLeod and Maclyn McCarty decided to determine the agent responsible for the genetic information transfer acting in the Griffith experiment (Avery et al., 1944). They mixed non-pathogenic *Pneumococcus* cells with different extracts of the pathogenic one containing either its envelope, its protein part or its DNA fraction. They could show that the non-pathogenic bacteria became infectious only when mixed with pure DNA. These observations led to

the conclusion that DNA is the molecular support of heredity and thereby, the molecular component of chromosomes and genes.

2.2.2) DNA composition and structure

After the demonstration that DNA is the molecular support of heredity, it was still not understood how this molecule could carry such information. Thanks to the work of Phoebus Levene in the 20s, the molecular composition of DNA was partially solved and was shown to be composed of the four bases: adenine (A), guanine (G), cytosine (C) and thymine (T) in DNA. In 50s, Erwin Chargaff studied in more detail the composition of DNA and established that this composition was highly divergent depending on the DNA origin, but systematically composed of equivalent amount of A and T, as well as C and G (Chargaff et al., 1952). Due to this property, Chargaff stated that DNA molecule always contain as much purines (big bases constituted by two cycles: A and G) than pyrimidines (small bases constituted by a unique cycle: C and T). These observations were of major importance for the characterization of DNA structure and to understand how it encodes the genetic information.

In 1953, based on the X-ray diffraction data collected by Rosalind Franklin and on the properties of DNA composition established by Chargaff, James Watson and Francis Crick proposed their model for DNA structure (Watson and Crick, 1953). This molecule is composed of two anti-parallel chains paired together and constituting a regular double helix. This pairing property implied that the nucleotide sequence of one strand is necessary and sufficient to determine the sequence of its complement and was at the origin of the idea that DNA is replicated in cells thanks to this singularity.

Five years later, Matthew Meselson and Franklin Stahl tested the three models proposed to explain replication of DNA: the conservative, semi-conservative and dispersive models (Meselson and Stahl, 1958). They concluded that the replication of DNA is performed in respect of the semi-conservative model involving the separation of each strand and the neo-synthesis of a complementary one thanks to the pairing rule. This replication is achieved by a specific enzyme discovered by Arthur Kornberg in 1956: the DNA polymerase (Kornberg et al., 1956).

2.2.3) Solving the genetic code

Due to its sequential composition and the conservative property of replication, many people became convinced that DNA could encode information based on its sequence order. Knowing that proteins are composed of an amino acids sequence coming from a set of 20 possible amino acids, DNA should be able to encode at least 20 different messages. Francis Crick and its colleagues concluded that, to be able to encode at least this number of amino acids, the encoding units, referred

as “codons”, should be a combination of more than two consecutive nucleotides. With three nucleotides, it is up to 64 possible combinations. In 1964, the work of Marshall Nirenberg and Philip Leder solved the genetic code (Nirenberg and Leder, 1964). From the 64 possible combinations, 61 encode amino acids with a certain degree of redundancy, and the remaining three ones correspond to stop signals aimed to stop protein synthesis. This discovery was a major advent in the understanding of evolutive mechanisms, for that for the first time, it became possible to read DNA and to deduce a protein composition encoded in the molecular support of genetic heredity.

2.2.4) The central dogma of molecular biology

Thanks to the knowledge acquired by studying DNA and its role in heredity, Francis Crick established in 1958 and clarified in 1970 the central dogma of molecular biology as such: “The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid”. Interestingly, the central dogma of molecular biology assumes that any modification in the DNA sequence will provoke a modification of the transcribed RNAs and with possible repercussion on proteins composition and function (Figure 15).

2.2.5) DNA organization in organisms

According to modern definitions, proteins are encoded by delimitated DNA regions, the so-called genes, and genes are part of long DNA strands organized linearly or circularly that form, in eukaryotes and prokaryotes respectively, a super-compact structure called chromosome (Kuzminov, 2014). In a cell, the overall content of DNA is referred as “genome” and can be constituted of several chromosomes. In prokaryotes, genomes are generally smaller and more compact. Genes are encoded close to each other and can sometimes be transcribed as one single unit, a structure called operon. In eukaryotes and more frequently in multicellular eukaryotes, genomes are larger and genes have a higher level of organization. They contain coding and non-coding regions respectively referred as exons and introns respectively. This mosaic structure of the gene allows a mechanism called alternative splicing and consisting in skipping of some exons under certain conditions to produce different proteins from one unique gene.

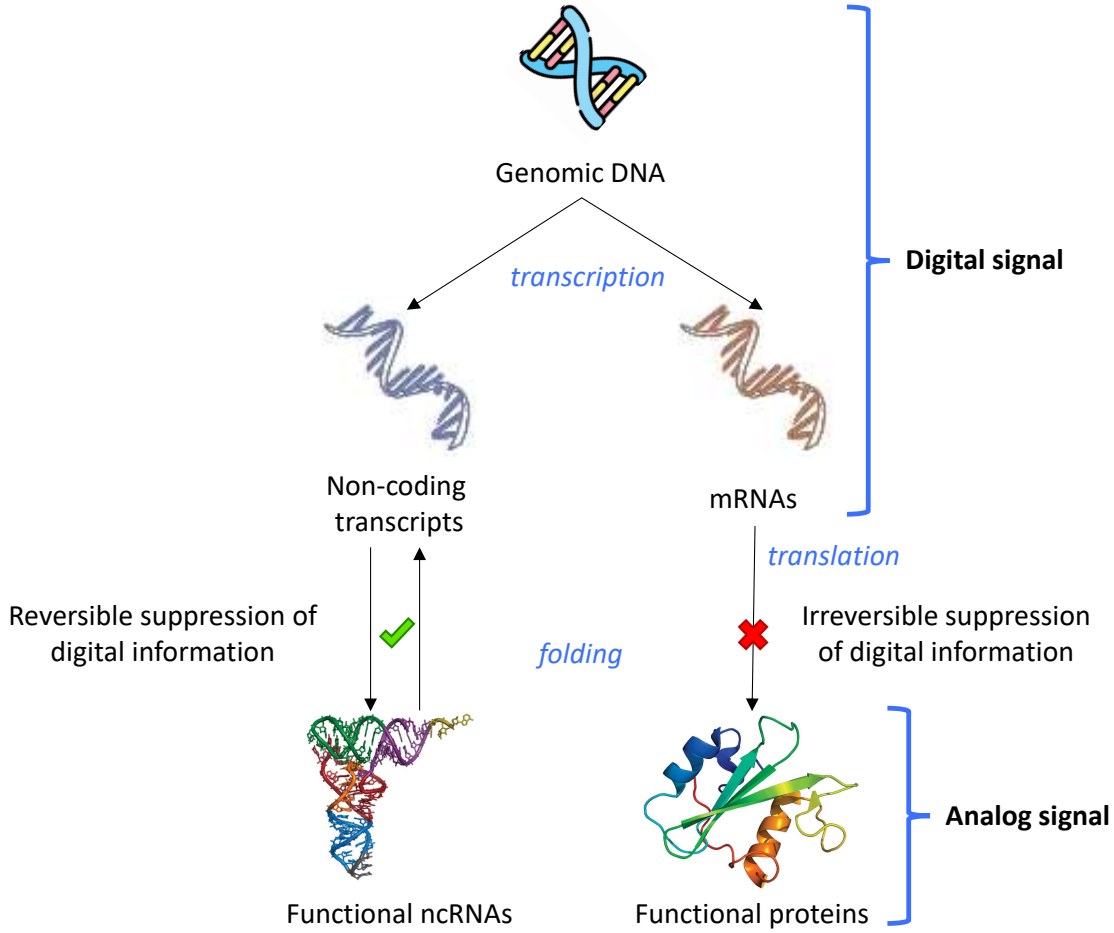


Figure 15 - The central dogma of molecular biology. DNA contains one-dimensional (digital) information copied in coding (mRNAs) or non-coding (ncRNAs) molecules through transcription. Then, the initial information is lost due to conversion into an analog or three-dimensional signal. For coding-genes, this process involves a translational step followed by a folding step. For ncRNAs, only a folding step is required for them to display their functions. Adapted from (Koonin, 2015)

2.2.6) The principle of biological relativity

Since its statement, the central dogma of molecular biology was considered as an essential pillar to understand the living world. However, the position of DNA at the top of the transcription-translation cascade was discussed in the light of recent discoveries about gene expression regulation, notably by epigenetic and environmental factors. In 2012, Denis Noble proposed the principle of biological relativity, arguing that there is, *a priori*, no privileged level of causation in biological systems, contrary to what is depicted by the central dogma of molecular biology (Noble, 2012). In this paper, Noble presented the DNA molecule as a passive entity that must be interpreted by a variety of other components to determine when and what to produce, similarly to a musician playing a music score. These essential components include hormones, transmitters, transcription factors as well as epigenetic marks such as methylation and histone modifications. Furthermore, it is also important to note that the integrity of the DNA molecule itself depends on protein machineries able to read and to continually correct mutations in the genome. The concept of biological relativity was then extended

to all components of living organisms, stating that "in multi-scale networks of interactions, as found everywhere in organisms, any parts of a network at any level might affect every other part" (Noble et al., 2019). For example, a variation in calcium ions concentration in the heart can lead to multiple molecular, cellular and tissular outcomes ultimately resulting in heart attack, a situation refered as upward causation. Conversely, the decision to practice physical exercise is a choice taken at the individual level that induces physiological modifications to finally result in adaptations at the molecular level, a situation refered as downward causation.

2.3) Evolution at the DNA scale

The central dogma of molecular biology positions DNA at the top of the transcription-translation cascade, and regarding the theory of evolution, this place makes DNA the most important effector of natural selection. Any change in the DNA sequence can lead to changes in proteins, consequently affecting the cell fate. When a change in DNA takes place in the germline, it become heritable, and it can cause diseases or decrease the organism's ability to face environmental challenges. In case, this change in DNA lead to a premature death, it will not be transmitted to the next generation and will not be conserved. Conversely, if a change favor adaptation to an environment or at least has no deleterious effect, it will have greater chances to be maintained and transmitted to the progeny. In DNA, different kinds of changes can occur ranging from single nucleotide to whole chromosome modification.

2.3.1) Point mutations

One of the main sources of DNA modifications is linked to the DNA replication step. Replication is an error-prone mechanism that sometimes generates nucleotide substitutions generally referred as mutations. The error rate of replication is a key concept of evolution at the DNA scale as a too low error rate would produce an insufficient amount of variations and consequently, limited evolution possibilities. Mutations are divided into three categories depending on their impact on the encoded protein: silent, missense and nonsense mutations. A silent mutation happens when a single nucleotide substitution does not alter the protein sequence thanks to the genetic code redundancy. Missense mutations correspond to nucleotide substitutions leading to a change in the protein sequence. This change can be conservative if the properties of the new amino acid are equivalent to the original one, or non-conservative otherwise. If the mutation is non-conservative, the protein function can be altered, improved or completely disrupted. Finally, nonsense mutations lead to four main issues: stop-gain, stop-loss, start-gain and start-loss. Stop-gain and stop-loss are due to the premature apparition of a stop-codon or the replacement of the true termination codon, leading to

shorter or longer proteins, and generally result in non-functional proteins. Similarly, start-gain and start-loss involve apparition of an additional ATG start codon or the replacement of a start codon respectively, leading to aberrant protein formation or to the complete loss of protein production.

2.3.2) Insertions and deletions

DNA replication, particularly of repeated regions, can also lead to nucleotides insertions or deletions grouped under the term "indels". Depending on the region where indels take place and the number of added or removed nucleotides, these anomalies cause more or less severe deleterious consequences. When the number of affected nucleotides is a multiple of three, codons are added which result in additional amino acids in the encoded sequence or in premature stop signal. Otherwise, the indel event induces a shift in the reading frame (frameshift) causing a consequent change of the protein sequence and function.

2.3.3) Transposable elements

Some changes in DNA can affect larger regions than indels and point mutations and are due to transposable elements, also called "transposons". They consist in DNA sequences of variable size able to change of positions within the genome. Two types of transposons exist: retrotransposons and DNA transposons. Retrotransposons originate from DNA regions transcribed in RNA and then reverse transcribed in DNA by a reverse transcriptase. The reverse transcribed DNA can then be inserted back at a different position in the genome. Conversely, DNA transposons are not transcribed. Their genomic transposition is catalyzed by transposase enzymes that bind DNA regions in a specific or non-specific manner. When transposons are inserted at a novel genomic locus, they can interrupt genes thereby disabling their function and sometime leading to diseases.

2.3.4) Duplication

DNA can undergo duplication at different level, from nucleotide to whole chromosomes or genomes. Five types of duplication are generally considered: partial gene duplication (nucleotide level), complete gene duplication, partial chromosomal duplication, complete chromosomal duplication (referred as aneuploidy) and full genome duplication (referred as polyploidy).

Gene duplication can occur through many different ways, notably by retrotransposons. In that case, DNA is transcribed, reverse transcribed and re-inserted within the genome, leading to two copies of the same gene. Gene duplication is believed to play a major role during evolution, participating to a neo-functionalization process. A duplicated gene is, indeed, free of selective pressure, as mutations will not lead to deleterious effects since the original gene is still functional.

Two types of gene duplications are observed: invariant and variant repeats. In the former case, the two sequences are almost identical and share identical functions. This case allows the increasing synthesis of one gene product. Transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) are a good example of invariant repeats. In the second case, one copy of the duplicated gene accumulates many mutations eventually leading either to the development of new functions or its inactivation. Indeed, mutations inducing a frameshift or a premature stop codon will turn the affected copy to a pseudogene.

When duplication happens internally due to exon duplication, the encoded protein can gain an additional domain, corresponding to a region of the protein performing a specific function (i.e. reaction catalysis) or constituting a structural unit. Serum albumin (ALB) is one example of a protein constituted by three repetitions of the same domain (Figure 16). Alternatively, exons encoding functional domains can also be transferred from one gene to another, creating mosaic proteins. The human tissue-type plasminogen activator (PLAT or TPA) is an example of mosaic protein evolution, combining five domains present in other proteins: one fibronectin type-I domain, one EGF-like domain, two Kringle domains and one peptidase S1 domain (Figure 16).

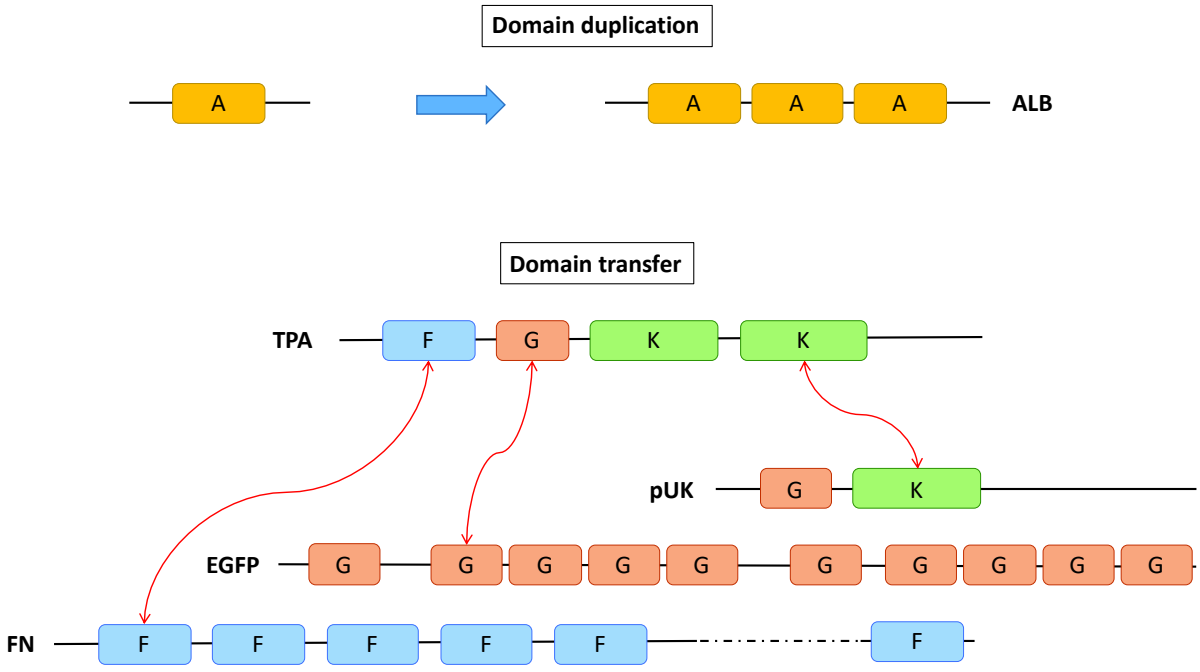


Figure 16 - Modularity of protein domains. Schematic representation of domain duplication in serum albumin (ALB) and mosaic protein organization of tissue-type plasminogen activator (TPA), prourokinase (pUK), fibronectin (FN) and epidermal growth factor precursor (EGFP) domains. ALB contains three repetitions of the albumin domain acquired from exon duplication. TPA contains diverse domains acquired through exon insertion from the three other proteins. Only non-proteinase regions are represented. A, albumin domain; K, kringle module; G, growth-factor module; F, finger module.

2.3.5) Horizontal gene transfer

While in eukaryotes the main way to acquire new functions is gene duplication followed by neo-functionalization, horizontal gene transfers (HGTs) are probably the favored way in bacteria and archaea. HGTs consist in transfer of genetic material between different species. This transfer can occur between representatives of very distant taxa of life: viruses, bacteria, archaea and eukaryotes. HGTs imply the need for a vehicle to transport the genetic information between cells, and a molecular machinery to insert DNA into another genome. Retroviruses are paradigm of this phenomenon. These entities insert their retro-transcribed genomic RNA into the host chromosomal DNA and are easily transmittable between individuals and species by nature. These two characteristics are what define HGTs. Genes acquired through retroviruses infection are referred as virogenes and are easily detectable (Todaro, 1975). Indeed, vertebrate genomes encode a non-negligible part of sequences homologous to retrovirus sequences. Interestingly, HGTs events have also taken place between eukaryotic host cells genome and mitochondrial or chloroplast genomes.

With the development of comparative genomics, HGTs, initially believe to occur only marginally, have appeared as an essential process for unicellular life and evolution. For example, bacterial cases of antibiotic resistance acquisition in hostile environments correspond to Lamarckian evolution by short-term adaptation. In bacteria, this key phenomenon takes place through plasmid exchange. Notably, antibiotic resistance can be acquired through horizontal transposon transfer (HTT). For example, a gene is transposed from a resistant bacterial genome to one of its plasmid, and then the plasmid is exchanged to another bacterium. The transferred gene is then referred as xenolog, an homologous gene originating from another species. In these organisms, gene transfer occurs notably thanks to gene transfer agents (GTAs), specialized HGT vectors made of defective derivatives of tailed bacteriophages (Lang et al., 2012). The ensemble of genetic information exchanged is often referred as “mobilome”, a term including genes and their vector system: bacteriophages, plasmids or transposons.

2.4) Studying evolution

Before the discovery of DNA and its role as carrier of the genetic information, evolution of organisms was assessed through morphological comparisons. In 1958, Francis Crick proposed that, if the phenotype relies on the genotype, morphological comparisons could also be achieved by gene comparisons. Indeed, thanks to its hereditary property and its mutation ability, DNA is the cellular unit that allows the evolution of species. Hence, analyzing DNA at all levels of organisation should be the best way to study evolution.

2.4.1) A science based on comparisons

Before proposing its now famous theory, Darwin spent quite some time to observe living organisms, and this is what led him to compare his observations between species. Similarly, in 1962, inspection of haemoglobin sequences led Emile Zuckerkandl and Linus Pauling to propose the concept of molecular clock based on the idea that if the evolution rate of a protein sequence is constant, the evolutive distance between two organisms can be inferred from sequence comparisons (Zuckerkandl and Pauling, 1962). However, when this concept was applied to specific cases, the molecular clock was not always identical and appeared to depend on the protein under consideration. Moreover, many genes are not represented in all living organisms, sometimes preventing the identification of the phylogenic relation between species. Carl Woese solved this problem in 1987 when he discovered that one ribosomal RNA (rRNA) present in all living species could constitute the ideal molecular clock: the 16S rRNA (Woese, 1987). For the first time, it was possible to compare and classify all living species using molecular criteria rather than morphology. These works resulted in the establishment of a new phylogeny and its representation by a new tree of life composed by three major domains: eukarya, bacteria and archaea.

Similarly, in 1990, Eric Westhof and François Michel applied a comparative method to RNA sequences investigations in order to decipher the three-dimensional architecture of group I catalytic introns (Michel and Westhof, 1990). Intron excision can be performed by a RNA-protein complex called the spliceosome, or thanks to an autocatalytic mechanism specific to some introns. Group I catalytic introns are part of this category and their function is closely related to their three-dimensional structural organization. The simple comparison of group I introns sequences between different although related species allowed Westhof and Michel to identify conservation or co-variation of nucleotides at specific position. As the sequence of the intron does not encode any protein, these nucleotides were apparently not important for the primary sequence of introns. Westhof and Michel proposed that these conservation/co-variations were linked to the three-dimensional structure and consequently essential to maintain the autocatalytic activity of the ribozyme. Based on the sequence comparison, they could model the 3D structure of group I catalytic introns and gain insights into the autocatalytic mechanism. This strategy was widely applied to solve multiple RNA structures of increasing complexity, up to the largest one of all, the ribosomal RNAs. Importantly, latter 3D structures of RNAs solved by X-ray crystallography or cryomicroscopy largely confirmed the results predicted from comparative sequence analyses (Miao et al., 2020). In parallel, Eugene V. Koonin and his group developed a similar strategy to establish the basis and concepts of protein sequence comparative analysis.

2.4.2) Genomic & bioinformatics

Since the mid 90s, thanks to massive genome sequencing, a plethora of genomes from various organisms has become available. The idea to compare these genomes then became obvious, but some biologists as Ernst Mayr believed that this could not be feasible. He argued that living organisms have so enormous phenotypic divergences that genes should also be too different to be compared, even between closely-related species. However, the comparison of genes and their encoded proteins led to another conclusion: genes and their products are strikingly conserved, even between organisms sharing a distant ancestor. Some genes are conserved from bacteria to humans, and their evolution does not preclude sequence comparison. This raised the conclusion that comparative genomic was not only feasible, but also supported a mine of information. With increasing number of genomic sequences available, manual comparison was no more conceivable. Computers appeared as the ideal tool for sequence comparison leading to the development of comparative genomics as one of the more important field in bioinformatics. Over time, central concepts such as homology, orthology and paralogy (see 3.1.1) have emerged and became important supports to study evolution at the molecular level (Sonnhammer and Koonin, 2002).

3) Bioinformatics and Big Data

Bioinformatics is the field of biology that aims to answer biological questions using *in silico* approaches that relies on mathematics, statistics, and data-mining methods. Computers are able to work at a high speed and to deal with massive data, two characteristics appropriated to extract biological meaning from all experiments generated by new technologies, the so-called "Big Data". Since the advent of the Big Data era, bioinformatics has gained unprecedented importance notably to analyze massively produced results from next-generation sequencing (NGS) technologies. However, this field of biology is not so recent, as it appeared long before the advent of NGS.

3.1) Origins of bioinformatics

Bioinformatics originates from the early 1960s and its first application was a program to reassemble partial protein sequences, established few years before the genetic code was solved. Initially designed to deal with proteins data, bioinformatics softwares from that time laid down the foundations for modern bioinformatics and many other fundamental fields in biology.

3.1.1) It all started with proteins

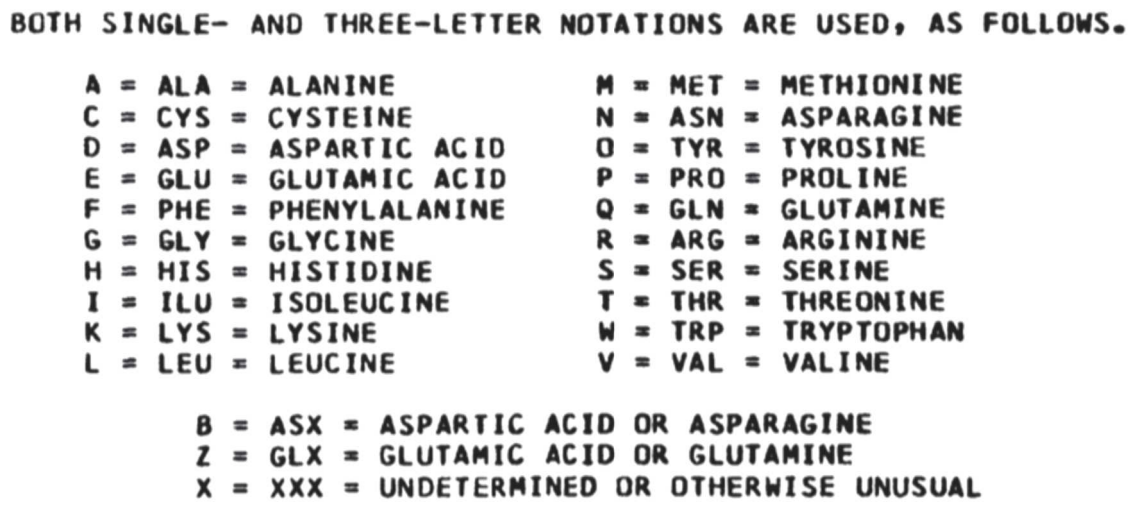
In the 1950s, the role of DNA as the molecular basis of heredity was still debated and its characteristic double helix structures not yet determined. At that time, most biological investigations concerned proteins and particularly enzymes. These macromolecules often considered as biological catalysts are of particular importance because they allow acceleration of chemical reactions in cells. To better understand their catalytic mechanisms, their content in amino acid could be investigated by a sequencing method developed by Pehr Edman: the Edman degradation method (Edman, 1950). This method consisted in labelling the amino-terminal acid of a protein with phenylisothiocyanate following by its specific cleavage at low pH. Once cleaved, the labelled amino acid was extracted and identified using chromatography or electrophoresis. By an iterative process, this technique allowed the sequencing of peptides one amino acid after the others. To speed up the process, it was possible to initially break a protein of interest into smaller peptides and to sequence them separately. Without surprise, the first bioinformatics program developed in this context was designed to help peptide's sequences reassembly emanating from the Edman degradation method. Developed in 1962 by Margaret Dayhoff, sometimes referred as the "mother and father of bioinformatics", together with Robert S. Ledley, "Comprotein" is since then considered as the first "*de novo* sequence assembler" (Dayhoff and Ledley, 1962).

Four years after, the combination of Comprotein with the Edman sequencing method led to the release of the first protein atlas constituting the first biological sequence database. This book contained 65 protein sequences represented with the one-letter amino acid code established by Dayhoff herself and that is still in use nowadays (Figure 17) (Eck and Dayhoff, 1966). The presence of same proteins from different species in this book stimulated the emergence of new ideas, notably to compare their amino acid sequences. This was the foundation of the hypothesis that species evolution can be assessed through comparison of their protein sequences (see 3.2.4). The bases of evolutive science were then established. Initially called "paleogenetics", a term coined by Emile Zuckerkandl and Linus Pauling, this field quickly became the major way to study life's history (Pauling and Zuckerkandl, 1963). In 1970, one of the key concepts in this field was proposed, the concept of "orthology". Similarity between two sequences is quantified thanks to several parameters, the most obvious one being the percentage of identity: when two sequences are similar enough, they can be considered as homologs. Homology between two proteins can derive from a gene duplication and the two resulting proteins are then paralogs. If the duplication process arose before a given speciation event, the relation is defined as outparalogy. Otherwise, duplication occuring after a speciation event is considered as inparalogy (Sonnhammer and Koonin, 2002). When homology between proteins from different species results from a speciation event rather than a duplication, they are considered as orthologs. Identification of orthologous proteins supported the theory of evolution and the notion of common ancestor.

As homologous proteins share a similar sequence, it was supposed that they also share the same function. This was demonstrated to be often true for orthologs (Gabaldón and Koonin, 2013), extending the observation of protein sequences conservation among living organisms to the conservation of biological functions. In the case of paralogs, duplication of a gene decrease the selection pressure on one of the two copies, resulting in its capacity to evolve independently to acquire a specialized activity (specialization), to develop a new function (neo-functionalization) or to become nonfunctional (pseudogene). In this situation, the assignment of biological functions by annotation transfer based on sequence similarity from known proteins is less reliable and precise, potentially rising to annotation mistakes.

At that time, sequence comparisons were performed manually and visually. Consequently, only short peptides or proteins from closely related species could be compared. This was a major limitation to study evolution at the protein level as it was particularly difficult and time consuming to compare proteins sharing a distant ancestor or of different lengths. The need for informatics stimulated conception of pairwise protein sequence alignment tools. The first algorithm dedicated to this task was created in 1970 by Needleman and Wunsch and facilitated pairwise alignment of protein sequences (Needleman and Wunsch, 1970). In 1987, tools for multiple sequence alignment (MSA)

appeared, which allowed to gain time and precision in protein alignments as to compare multiple sequences of divergent length and composition together. In 1988, one of the most popular of these MSA tools was developed: CLUSTAL, a program still maintained and widely used today (Higgins and Sharp, 1988).


BOTH SINGLE- AND THREE-LETTER NOTATIONS ARE USED, AS FOLLOWS.

A = ALA = ALANINE
C = CYS = CYSTEINE
D = ASP = ASPARTIC ACID
E = GLU = GLUTAMIC ACID
F = PHE = PHENYLALANINE
G = GLY = GLYCINE
H = HIS = HISTIDINE
I = ILU = ISOLEUCINE
K = LYS = LYSINE
L = LEU = LEUCINE
M = MET = METHIONINE
N = ASN = ASPARAGINE
O = TYR = TYROSINE
P = PRO = PROLINE
Q = GLN = GLUTAMINE
R = ARG = ARGININE
S = SER = SERINE
T = THR = THREONINE
W = TRP = TRYPTOPHAN
V = VAL = VALINE

B = ASX = ASPARTIC ACID OR ASPARAGINE
Z = GLX = GLUTAMIC ACID OR GLUTAMINE
X = XXX = UNDETERMINED OR OTHERWISE UNUSUAL

Figure 17 - Amino acid one-letter code. Table for the correspondence from one- to three-letters codes of amino acids as it was presented in the first protein atlas. Note that tyrosine is the only amino acid that its one-letter nomenclature was changed today: Y has replaced O. Taken from (Eck and Dayhoff, 1966).

3.1.2) From protein to DNA analysis

In 1968, as the bases of "paleogenetics" were established, the genetic code was completely solved. This was a main progress as it allowed biologists to decrypt the information encoded in DNA. In theory, it was then possible to determine any amino acid sequence encoded in any gene. This revolutionary discovery should have led to a major improvement in protein sequence prediction. However, in comparison to proteins, genes are less concentrated in cells. A unique gene is transcribed in multiple RNA molecules from which a larger amount of protein copies are translated. Therefore having access to a gene sequence was a major limitation and gene amplification was a pre-requisite to their investigation. In 1983, an amplification method was developed: the polymerase chain reaction (PCR) (Mullis and Faloona, 1987). This method now widely used in all molecular biology laboratories consists in a specific amplification of a selected DNA region using small complementary nucleic fragments called "primers". PCR is based on the iteration of DNA replication cycles performed by polymerase proteins working at a sufficiently high temperature to unwind DNA double helices. This techniqueal advance was necessary and sufficient for the emergence of the DNA sequencing era.

3.1.3) Sanger DNA sequencing

In 1977, the now famous Sanger DNA sequencing method - the classical chain-termination method - appeared, constituting a revolution in the world of biology (Sanger et al., 1977a).

This technological advent constituted a revolution for the study of genomes but also for the study of proteins. Before DNA sequencing, proteins needed to be individually expressed, concentrated and purified before sequencing. Thanks to DNA sequencing, it became theoretically possible to access to a whole genome from a single genomic DNA extraction. With the capacity to decrypt genes according to the genetic code, it was then possible to uncover the full proteome of an organism just by sequencing a single DNA molecule. However, the Sanger sequencing method was not so appropriated to determine the genomic sequence of eukaryotic or prokaryotic organisms. Indeed, this manual method was limited to small genomes such as viral ones. The first of them was sequenced in 1977: the bacteriophage ΦX174, a single-stranded DNA virus (Sanger et al., 1977b).

To facilitate the analysis of Sanger sequencing results, Roger Staden developed a set of multiple bioinformatics programs: the Staden Package (Staden, 1979). This bunch of tools allowed searching for overlaps between Sanger readings, assembly of small reads into bigger DNA sequences referred as "contigs", and annotation of the generated sequence files. However, this was insufficient to allow routine sequencing of eukaryotic or prokaryotic genomes that were still too long to be easily handled. Nevertheless, in 1995, the first prokaryotic genome was sequenced: the *Haemophilus influenzae* genome (Fleischmann et al., 1995).

3.1.4) Democratization of informatics

On the purely informatics side, techniqueal improvements also contributed to the advent of modern bioinformatics. As the years went by, computers became less and less expensive but also smaller, favoring their acquisition by laboratories, as well as private people. Concomitantly, their processing capacities continuously increased allowing faster data treatment, better data storage and improved computational power. These technological advents led to the establishment of national databases for biological data storage. Notably, three sequence databases where established during the 80s: the European Molecular Biology Laboratory (EMBL), its American counterpart GenBank and the DNA Data Bank of Japan (DDBJ). With an increasing number of users, the informatics as well as bioinformatics communities raised questions about the use and owning of data leading to the emergency of the free software philosophy. All these changes resulted into the union in 1987 of the EMBL, GenBank and DDBJ in a collaborative institution: the "International Nucleotide Sequence Database Collaboration" (INSDC).

In addition, new high-level programming languages appeared in the mid 90s: Perl in 1987 and Python in 1989. In comparison to their ancestors FORTRAN, C or BASIC, these languages were more flexible, using a simpler syntax. Consequently, they encouraged more people to have an interest into informatics and bioinformatics that favored tools development.

3.2) Bioinformatics to study evolution

Computers allow the efficient analysis of textual data. Consequently, bioinformatics appeared as an ideal solution to study the evolution of living organisms through DNA, RNA and protein's sequences that are easily converted into texts. Bioinformatics then became an essential method to study evolution.

3.2.1) Sequencing full genomes

In the mid 1990, biologists improved DNA sequencing technologies to allow whole genome sequencing. These combined improvements were instrumental to the full sequencing of the first bacterial genome in 1995: the *Haemophilus influenzae* genome (Fleischmann et al., 1995). Several genome-sequencing projects then succeeded, applied on more and more complex organisms: *Saccharomyces cerevisiae* (the first eukaryote) in 1996 (Goffeau et al., 1996), *Caenorhabditis elegans* (the first animal) in 1998 (The *C. elegans* Sequencing Consortium, 1998), *Arabidopsis thaliana* (the first plant) (The Arabiopsis Genome Initiative, 2000), and *Drosophila melanogaster* (Adams et al., 2000) in 2000 and finally the human genome released in 2003 (International Human Genome Sequencing Consortium, 2001). This success paved the way to the wave of massive genome sequencing and resulted in an impressive amount of DNA sequences. Due to the variety of sequenced organisms, these genomic data allowed the investigation of evolutionary events from living beings with a variable degree of divergence. Thanks to the unprecedented number of available sequences, it became for the first time possible to analyze evolution through statistically relevant observations. Concomitantly in 1990, the National Institutes of Health (NIH) released BLAST (Basic Local Alignment Search Tool), an algorithm allowing fast alignment of nucleotides and amino acids sequences (Altschul et al., 1990). This tool still widely used today, greatly helped the investigation of all the sequences generated by the different sequencing projects, making comparisons faster and easier.

3.2.2) Genomics, the mother of omics

With a high number of genomes available coming from various prokaryotes and eukaryotes, the 2000s were the cradle of genomics. From the analysis of genomic content, structure and organization, biologists have been able to characterize the main features differentiating prokaryotes and eukaryotes. Notably, genomes were then separated in two classes: those with a gene number proportional to the genome size, and those showing decoupling between these two parameters: prokaryotic and eukaryotic genomes respectively. To analyze all these data, dedicated bioinformatics methods to predict genes in genomes were developed. This was particularly challenging in eukaryotes due to their characteristic gene structure, alternating coding and non-coding regions: exons and introns. Nevertheless, algorithms based on the knowledge acquired with molecular biology could detect open reading frames (ORF) in genes with an ever-increasing precision. With these major progresses, it became theoretically possible to predict all proteins encoded in the genomic data from several species, from bacteria to mammals. To help the functional analysis of all these data, an initiative started in 1998 proposed a standardized system: the Gene Ontology initiative (Ashburner et al., 2000). The aim of this project was to represent and annotate gene functions using a controlled vocabulary usable across all species. A standardized annotation system, referred as GO Terms, was then divided in three categories: "cellular component" corresponding to the gene product location, "molecular function" corresponding to its catalytic activity and "biological process" corresponding to pathways or functions. This nomenclature is particularly relevant to compute ontological enrichments, allowing the identification of the main biological activity performed by the encoded proteins from a set of many genes.

3.2.3) Development of second- and third-generation sequencing technologies

In 2005, new DNA sequencing technologies appeared to counter the main drawbacks of Sanger sequencing method: its cost and speed. Based on the parallel sequencing of smaller reads (few hundreds base pairs versus up to one kilobase), the pyrosequencing technique from the 454 Life Sciences company opened the way for the second-generation sequencing (NGS) (Margulies et al., 2005), followed by the method developed by the Illumina company which is the actual leader in DNA sequencing (Bentley et al., 2008). The development of these high-throughput sequencing technologies caused a significant improvement in genome sequencing. Consequently, the number of sequenced genomes rapidly increased. To compile and analyze all these data, powerful automated strategies were needed. Generating smaller reads but at a higher rate, it was required to develop appropriated bioinformatics tools to reassemble DNA fragments into longer scaffolds: the assembly

process. However, this step is time-consuming even if performed by powerful computers because of the number and size of DNA reads, especially when no reference genome is available.

With the efforts made to improve sequencing efficiency and to continue decreasing its cost, recent new technologies have appeared. Pacific BioSciences (Rhoads and Au, 2015) and Oxford Nanopore (Stoddart et al., 2009) are developing third-generation sequencing (TGS) technologies that stand out from the previous one by allowing sequencing of longer reads (from 10 to 100 kilobases) from a single DNA molecule, countering the NGS drawbacks and facilitating *de novo* genome assemblies.

In parallel of technological improvements, ever more ambitious human sequencing projects appeared. In 2008 started the 1000 Genomes Project (1KGP), a project planned to sequence at least thousand human genomes (The 1000 Genomes Project Consortium, 2010). This goal was achieved in 2012. Similarly in 2016, a human sequencing project reached the number of 10,000 sequenced genomes (Telenti et al., 2016). More recently in 2018, an even more ambitious project initiated in the United-Kingdom reached his goal to sequence 100,000 human genomes. Similar projects centered on different species have also been initiated: the 1001 Genomes Project started in 2008 targeting the plant *Arabidopsis thaliana* (Weigel and Mott, 2009) or the 1002 Yeast Genomes Project started in 2013 targeting the yeast *Saccharomyces cerevisiae* (Peter et al., 2018).

These projects generating a high number of genomes coming from a large diversity of species opened the way to comparative genomics, a bioinformatics field that aimed to analyze evolution through comparison of protein or gene sequences originating from different species. More than allowing techniqueal comparisons of genomes and proteomes, these advances favored the emergence of new evolutionary hypotheses and concepts, new fields, and the discovery of new biological processes.

3.2.4) Comparative genomics

Comparing protein sequences between different taxa, it became obvious that some regions and even amino acids are more conserved than others. This observation raised the idea that these conserved residues are particularly important for protein function. Indeed, any modification of these residues should lead to a disruption of vital protein functions and are consequently under a high evolutionary pressure. As the protein function greatly depends on the three-dimensional structure, which can be altered by sequence modification, it became possible to establish a relation between a protein sequence and its structure. Comparison of protein sequences also allowed identification of functional and structural regions shared between different proteins: protein domains.

After a genome has been sequenced and its gene content predicted, it is possible to determine the encoded protein sequences. However, each protein from each living organism has not been experimentally studied to identify its biological role. Alternatively, sequence comparison allowed prediction of protein identity and function based on their similarity level with well-annotated and studied proteins. In this case, GO-terms annotations from these similar proteins are transferred and mentioned as "Inferred from Electronic Annotation" (IEA). To decide if a functional annotation can reasonably be transferred between two proteins, comparative genomics investigations rely on the characterization of the homologous proteins. Without any clue about gene duplication history, the most commonly used metrics to identify homology is sequence similarity. Many different methods were proposed to assess homology relations between proteins, the most straightforward being the identification of the best-hit result from a BLAST search and the assumption that this best-hit is the orthologous protein of an initial protein query. As this simple process does not consider the reciprocality of such relation, a more common approach is the bidirectional or reciprocal best-hit (RBH) method (Tatusov et al., 1997). In this method, the best-hit from an initial BLAST search is retrieved and used as query for a second BLAST analysis against the first species. If the best-hit found is the protein used as query in the initial search, and if the similarity score is significant (usually superior than 30% of identity), the two proteins are considered as orthologs. The RBH method is efficient to identify one-to-one orthology relations, but is not well suited to solve cases where a gene is duplicated in the target species, constituting a set of inparalogous sequences. To define these one-to-many (one protein is ortholog to many inparalogs) and many-to-many (many proteins are co-ortholog to many inparalogs) relations, an additionnal step is required to first identify these inparalogy relations. This step, implemented in Inparanoid (Sonnhammer and Ostlund, 2015) and OrthoInspector (Nevers et al., 2019) for examples, consists in identifying and clustering sequences from a given species that are more similar between them than with any protein from the target species, representative of their inparalogy. In addition to this method, determination of orthology relations are also performed based on alternative approaches involving phylogenetic trees, hybrid methods combining sequence similarity and evolutive trees, or accessible through integrative databases such as the Alliance for Genome References website (The Alliance of Genome Resources Consortium, 2020) that gather and process data coming from many different tools.

At the single-species level, facilitation of genome sequencing make it possible to consider the comparison of genes to detect single-nucleotide polymorphisms or variations (SNPs, SNVs). Combining these results with functional knowledge permitted to correlate these divergences with certain phenotypes or pathologies.

Genomics and comparative genomics concern the study of DNA history during evolution. However, comparing gene presence or absence between genomes is a relatively static way to analyze

DNA. To characterize the functional roles of genes in dynamical processes, sequencing technologies were applied to another target: RNA.

3.3) Bioinformatics to study the genome dynamic

Regulation of gene expression is the mechanism that allows the differentiation of specialized tissues in an organism constituted of cells that share the same genome. To permit this specialization, cells express specific genes with a variable rate depending on their type and function.

3.3.1) Transcriptomics

The transcriptome of a cell or a tissue corresponds to all the RNA molecules transcribed from the genome, namely all the genes that are effectively expressed. To study cellular RNAs, different techniques were developed. Initially performed on microarrays, the gene expression could be measured for only a restricted number of previously selected genes. To do so, specific probes were designed and anchored on a chip. Then, single stranded DNA or RNA molecules from a cell were added and expression of genes could be measured by the fluorescence emitted during the pairing process with the probe. With the development of high throughput sequencing technologies, it became possible to sequence a full transcriptome, without any *a priori* bias. This was the foundation for the advent of RNA sequencing or RNA-seq. Sequencing of RNA fragments is the first part of the RNA-seq process. Once all RNA molecules are sequenced as small fragments, bioinformatics tools are used to align RNA reads against a reference genome. This step allows the detection of genes that are effectively transcribed in cells, and presumably translated. Considering the number of read aligned to a gene, it is possible to determine its expression level, relatively to the total number of RNA reads sequenced. This technique helped to explain how a single and unique genome could produce different cellular phenotypes. Interestingly, RNA-seq is not limited to the analysis of the coding fraction of the genome, but is also applicable to the non-coding part of transcripts, including micro RNAs (miRNAs) and long non-coding RNAs (lncRNAs). These molecules are of particular interest as they were demonstrated to participate in post-transcriptional gene expression regulation and alternative splicing respectively (Guil and Esteller, 2015).

To understand in which extend transcriptomics could help to decipher the genome dynamics, an important issue was to establish whether transcript levels of a given gene are representative of the corresponding protein levels (Edfors et al., 2016). The presence or absence of such correlation between gene and protein level through RNA concentration has been largely debated and is still a question today. In 1997 before the omics era, based on biochemical experiments, Anderson and Seilhamer (Anderson and Seilhamer, 1997) showed that transcript and protein abundances strongly

varied between tissues and were poorly correlated together. More recent publications also pinpointed that proteome and transcriptome abundances were not sufficiently correlated to act as proxies for each other, precluding protein level prediction based on genome wide transcriptomics data (Payne, 2015). Such deviation between RNA and protein levels could be attributed to post-transcriptional regulation processes affecting mRNA stability, translation efficiency as well as the variable decay rate between proteins. However, controversial studies demonstrated that the protein–mRNA ratio in human cell lines (Lundberg et al, 2010) and tissues (Wilhelm et al, 2014) is constant. Recent results (Edfors et al., 2016) tend to nuance previous observations and indicate that prediction of the protein copy numbers from RNA levels is possible but significantly improved by the use of a gene-specific RNA-to-protein (RTP) conversion factor.

3.3.2) Differential expression analysis

Similarly to comparative genomics that consists in comparison of genome organization and content, transcriptomics studies also results in a comparison: the differential expression analysis. As genes transcribed in a condition or a tissue can be detected and their relative abundance determined, it is possible to compare gene expression between different cell types or environmental conditions. This kind of analysis shows genes that are overexpressed (or up-regulated) or repressed (or down-regulated) between two conditions, allowing to compare tissues between a group of patients and their controls, to determine the causes of a disease. More generally, it became possible to reconstruct all the complexity involved during a biological response at the gene resolution level and consequently, to understand the molecular mechanisms responsible of biological processes.

3.4) Integrative bioinformatics

The RNA-seq technology emerged with NGS and allowed a dynamic analysis of genome activity. Similarly to genomics and transcriptomics, the "-omics" suffix became then associated to all the fields impacted by computing analyses: proteomics, epigenomics, metabolomics, lipidomics, glycomics and interactomics being the most commonly discussed nowadays. To deal with these "Big Data", informatics is a helpful solution providing fast large-scale analytical capacities. High rate of data production associated with an efficient way to analyze them constitute what is called the "Big Data era".

3.4.1) A sea of data

Analyzing the available mass of data is the first step to extract knowledge. A lot of studies are focused on few molecules in a very particular context: a given cell type or developmental step, under

certain pathologies, or in response to particular environmental exposures. These kinds of approaches are usually considered as reductionist as they consist in testing multiple variables individually to identify an association with a phenomenon of interest. However, to really understand biological processes, it is mandatory to re-integrate at least partially these detailed mechanisms into a more general biological context. To this aim, omics studies play a significant role by providing clues at least about whole genome and whole transcriptome in given situations. Nevertheless, even these studies are focused on a unique problem, such as the analysis of tissue response to a drug, of cells behavior, or of organ physiology from a transgenic animal. Moreover, results are often only partially exploited, biased by the initial question of the study, while more knowledge could be extracted from such unique study or by comparing different ones. This is the reason why, at a time when so many data become available, it is possible and needed to integrate most of them into larger comprehensive models of life.

3.4.2) Systems biology

Systems biology is a branch of bioinformatics that aimed to recompose the biologic complexity to gain knowledge at the whole system level (Mast et al., 2014). To do so, it consists in modelling life by considering multiple genomes, transcriptomes, proteomes and metabolomes together rather than separately, to reconstitute life hierarchical organization at cell, organ or even population scale (Zierer et al., 2015). In cells, genes can be viewed as the upper part of this hierarchy, being the origin of RNA transcripts, themselves encoding proteins. This organization is partially regulated by a retro-control mechanism involving transcription factor proteins that will enhance or decrease expression of genes, which stand on the top of this cascade. Interestingly, a parallel can be made with another hierarchical organization observed at the animal organism level. The brain and secretory glands send signals to peripheral organs to regulate their activity; those in return, transmit retro-active signals to allow an appropriated response to environmental conditions. Systems biology aimed to reconstruct these hierarchical structures and to recompose a model as close as possible to the biological reality.

In such model, entities in various forms (molecules, proteins, genes, cells, organs) are interconnected based on their reciprocal activities and properties. At the genomic level, genes can be connected according to their organization in the genome (co-localization in a same chromosomal locus or synteny), their evolutive history (duplication-rearrangement…), or the function of the encoded protein. At the transcriptome level, encoded proteins can be deduced and their concentration can be estimated based on RNA expression levels. In addition, RNA-RNA interactions play important regulatory roles in controlling RNA translation or degradation. At the proteome level, proteins can

participate to common structural, metabolic or signaling processes, but they can also interact with DNA, RNA or other proteins contributing to regulation mechanisms. For example, in the case of a transcription factor protein, an interaction edge can represent the activity of the protein on DNA transcription. Compilation of all these data will ultimately lead to a complex physical and functional interactions network.

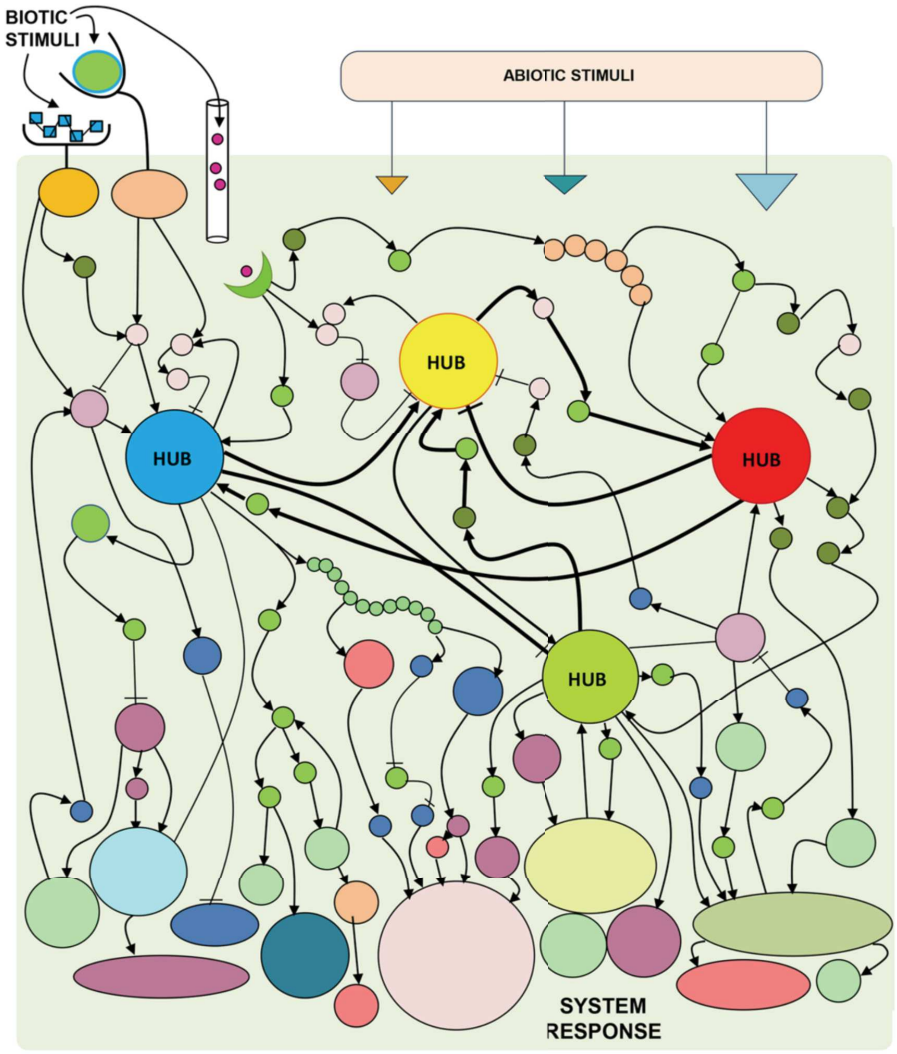


Figure 18 - Networks for complex systems. In cells, most components are interconnected making the system complex to understand. Indeed, biological networks are tuned to respond differently to different inputs but show a lot of redundancy allowing their regulation and increasing their adaptive capacities. Taken from (Hillmer, 2015).

Such an integrative way to study datasets is based on identification and analysis of interactions between all components of a complex system, which is possible thanks to network biology (Figure 18). This field was greatly inspired by Albert-László Barabási's works on purely informatics networks in the 90s. The foundation of his theory is that interactions between components and their intrinsic hierarchical structure are more important than the components themselves (Barabási and Albert, 1999).

3.4.3) Network biology

Biological interactions constituting a network can be divided in different categories. For example, the STRING interactions database (Szklarczyk et al., 2019) identifies four main categories: physical interactions, co-occurrence (in publications or within a pathway), co-expression and conservation (degree of homology). Integrating these different interaction modes, three main types of networks are currently represented in biology: regulatory networks, protein interactions networks and metabolic networks (Figure 19). Regulatory networks aims to represent all interactions able to affect proteins and RNAs expression such as the binding of a transcription factor to its gene target. Protein interactions networks aims to represent all physical binding between proteins, generally leading to complex formation and signaling events. These interactions are currently referred as PPI for "protein-protein interactions". Finally, metabolic networks are designed to represent the cascade of events occurring in a biological pathway. Usually, these networks connect enzymes together with their substrates and products.

By analyzing how biological components interact, it is possible to identify central elements, called hubs or groups of highly interconnected elements that are biologically relevant: clusters. Based on the foundation of network biology that supposes that interactions between components are more informative that the components themselves, the connectivity can give rise to emergent properties (Palsson, 2000). Indeed, proteins display functions that depend on their partners. Some transcription factor activities are repressed or enhanced after binding to other proteins and this is also true for several enzymes. Similarly, protein inhibition by its protein repressor demonstrated *in vitro* could be biologically irrelevant if the two proteins are in different compartments and unable to physically interact *in vivo*. Again, information on their interaction is at least as fundamental that information on these two independent components.

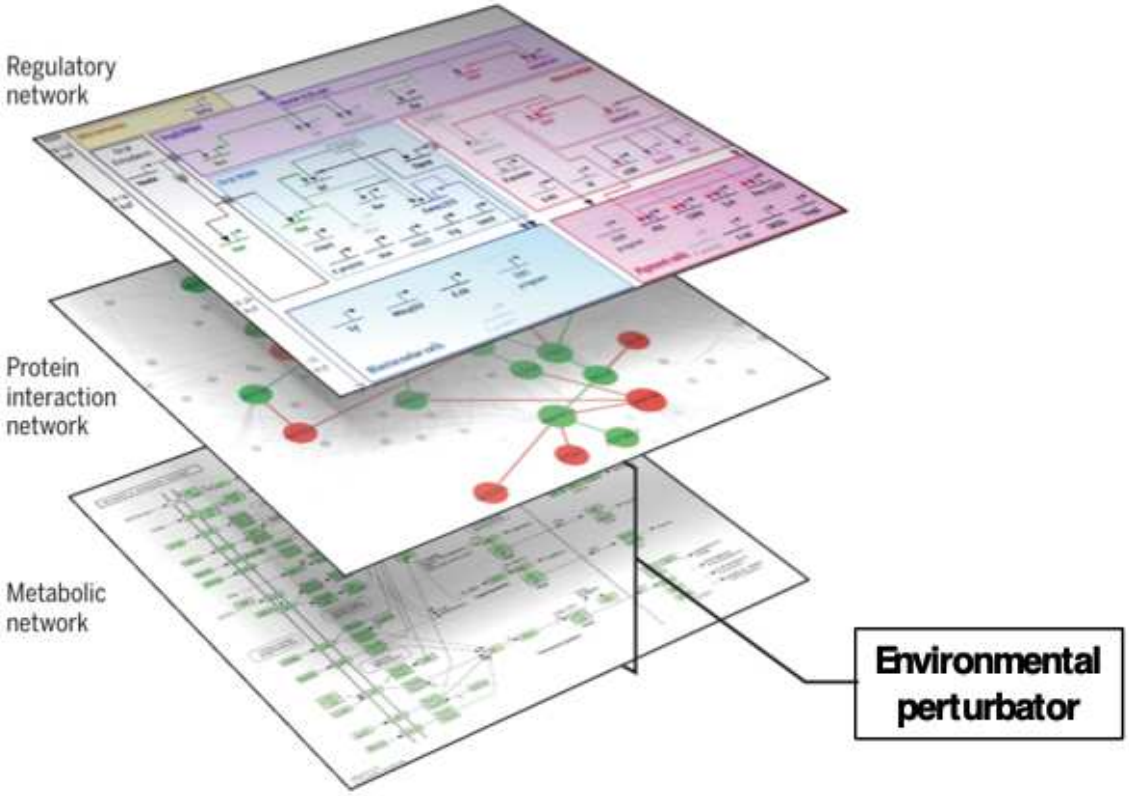


Figure 19 - Cells are multilayers networks. Cells are generally represented as three different network types: regulatory networks, protein interaction networks and metabolic networks. Each one is connected to the other ones and consequently, an environmental agent acting on at least one layer will impact all of them. Adapted from (Vermeulen et al., 2020).

3.4.4) From genotype to phenotype

As discussed previously, modifications of DNA lead to a cascade of effects: RNA molecules are altered, the encoded proteins are modified, cells behavior is affected, tissues function or inter-communications can be disturbed, finally leading to organs impairment and eventually to disease or death. In that way, genotype and phenotype are tightly connected, and genomics together with other omics data can help to predict the phenome of an individual (Figure 20). In theory, systems biology is able to predict disease appearance and its evolution in a particular organism or a given patient (Zierer et al., 2015). Similarly, such system is expected to predict stress response evolution regarding all the modulatory effects and the genomic background of a subject.

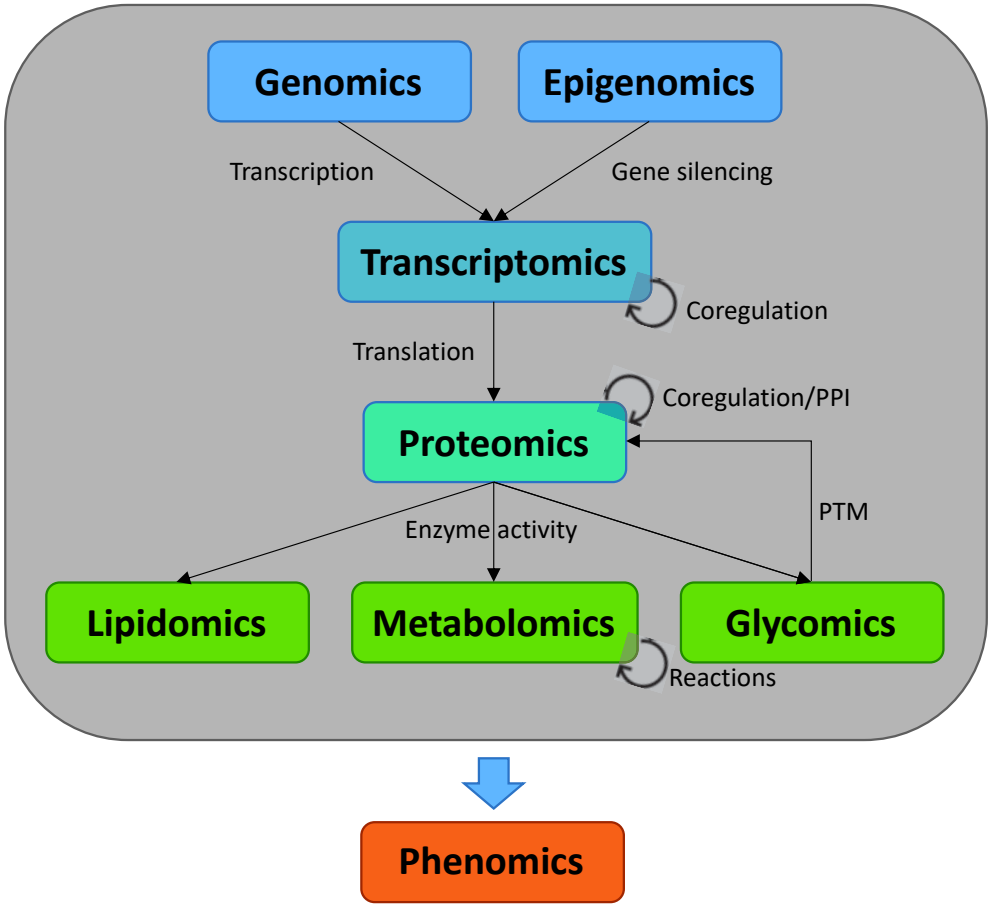


Figure 20 – Interdependencies of omics data in a hierarchical system. Because of the particular hierarchical organization of cellular processes, all related omics analyses are interdependent. Connecting these processes together is a clue to integrate biological complexity and to predict their impacts on phenotypes. However, predictions accuracy greatly depends on quality control, an essential step to avoid errors propagations from an omics layer to the other ones. Adapted from (Zierer et al., 2015).

3.4.5) Limitations in omics data integration

Similarly to the central role of DNA for living organisms, genomics investigations constitute an essential field impacting all the other omics (Conesa and Mortazavi, 2014). However, genomic data are rarely perfect and often of questionable quality.

The first problem with genomics is sequencing errors. Sequencing technologies are more and more precise but mis-identification of nucleotides still persists notably with the third generation sequencing systems that produce long reads but with a high error-rate. According to the central role of DNA, sequencing errors propagates at the RNA and protein levels. Similarly to true mutations, sequencing errors can results in premature stop codon and mis-prediction of the encoded amino acid peptide. Consequently, data resulting from the automatic transcription and translation from an erroneous DNA sequence are also wrong. Another important problem inherent to sequencing processes occurs during the assembly step. Sequencing project generally starts with the fragmentation of the genomic DNA into smaller pieces. Then, during the sequencing step, only small fragments are

determined and they must be re-assembled in order to reconstruct the whole genome. However, depending on the sequencing project, the assembly step can be incomplete. Instead of having access to a full genome, only an ensemble of long reads is accessible: contigs and scaffolds. This situation leads to a major problem, as a gene encoded at a contig extremity can be incomplete. Consequently, its automated identification and translation will produce erroneous data regarding the gene boundaries and protein sequence. Another example of a common misinterpretation in genomic data is the case of the selenocysteine amino acid. Indeed, selenocysteine (Sec) has the particularity of being encoded by a reprogrammed TGA codon (normally read as a stop) thanks to a specific re-encoding machinery. Automated algorithms for gene identification usually interpret this TGA(Sec) codon as a termination signal, resulting in a premature truncated protein sequence.

Genomic data quality also relies on steps performed after the sequencing process, notably during gene annotation. Automatic protein function inferred from sequence similarities between two proteins can be a source of errors. This annotation relies on thresholds and different methods to determine if a protein function can be transferred to another one. These automated processes are sources of error propagations, as a unique mis-annotated protein can serve as reference for annotating a bunch of other proteins. All these data taking together, it appears that a large part of the available genomic information is at least partially wrong, introducing noise in the source of information. This conclusion has a particularly dramatic impact for big data analyses and systems biology studies. As each layer of the biological complexity is integrated into a single model, bad quality data can generate a corrupted model, leading to erroneous predictions and irrelevant conclusions.

OBJECTIVES

During this thesis, I was involved in three different projects although sharing a common heuristic method, as they were conducted in an evolutive perspective. In the first one, I used comparative analyses to identify genes and mechanisms involved in stress response that are conserved between different vertebrate species. In the second one, I investigated the divergent evolution of bacterial and archaeal specific genes belonging to the phosphopantothenate biosynthetic pathway. To standardize and facilitate such investigations of evolutive relations between genes in different species, as part of a third project, I developed an integrative tool accessible online and combining different bioinformatics approaches: PROTEDEX.

I) Transcriptomic analysis to identify conserved genes and mechanisms involved in stress response and adaptation in vertebrate species

Stress is a major health issue both in Human and livestock. Indeed, stress factors are diverse in the environment and known for their potential to induce a stress state in living organisms (Vermeulen et al., 2020). To cope with such conditions, organisms developed a complex mechanism referred as stress response. This response was initially described as a stereotypical manifestation conserved between animals and independent of the stressor nature (Selye, 1956), suggesting the involvement of a conserved mechanism during this process. However, the dynamic and some specific effects of this response are variable and depends on individuals and species particularities, stressors nature as well as duration of exposure. Because of these variable parameters, stress response is a complex mechanism difficult to understand at the organism level. Several cellular stress models were developed to study this response. However, these reductionist approaches brought their own biases and specificities, and poorly translated at whole individual level to describe stress response and predict adaptive capacities.

Comparative biology approaches is another way to dissect such complex process by the identification of conserved mechanisms, with their associated genes, between different species. To this aim, we developed four animal stress models initially characterized using specific stress markers. We applied an unbiased approach based on transcriptomics methods to determine differentially expressed genes in each model. The comparison of these genes sets led to the identification of specific and, more importantly, common genes between the four stress models. Ontological analyses of these conserved genes allowed to highlight their associated biological processes and to dissect which of them are rather involved in stress response or adaptation to stress conditions. Interestingly, identification of conserved genes between species could constitute a particularly interesting application for stress state diagnostic. Indeed, such combination of candidate biomarkers could help to evaluate health state of different organisms for an accurate diagnostic, or allow testing the

pertinence of animals or patients therapeutic interventions, ultimately improving their resilience to stress situations.

II) Mosaic organization of metabolic pathways between bacteria and archaea species

In parallel to my main subject, I also worked on two emerging projects, the first of which concerning the mosaic organization of prokaryotic pathways for acetyl-CoA biosynthesis. To classify species, taxa-specific enzymes have been identified and are currently used as evolutive indicators in phylogenetic studies. Conducting analyses on a particular group of unclassified bacteria referred as *Candidatus poribacteria*, I serendipitously observed an interesting phylogenetic distribution of two enzymes involved in phosphopantothenate biosynthesis. This couple of enzymes is known to display a bacteria- or archaea-specific distribution and is used as a taxa-specific marker to classify these organisms. However, presence of the bacterial enzyme was previously identified in one archaeal group. In this project, I investigated the particular phylogenetic distribution of these enzymes within *Candidatus poribacteria* genomes and determined a symmetric situation in which bacteria encoding the archaeal enzymes were identified. This observation, first, demonstrate the mosaic organization of this metabolic pathway in prokaryotes, but also lead to a reconsideration of the use of these enzymes to classify bacteria and archaea species.

III) Development of an integrative tool for gene sets investigation

One common feature of bioinformatics studies and particularly omics data investigations is that they produce massive quantity of data, often difficult to interpret. Together with genomics, transcriptomics and proteomics are the three most represented omics categories and because of their intricate dependencies, outputs of these technologies (genes, RNAs, proteins) are affiliated to genes and convertible into gene sets. To analyze such sets, several approaches and associated tools have been developed among them ontological enrichment based on GO Terms is the most popular and practicable one. In addition, other approaches are possible, including for example extensive data-mining, homology inference or network construction. Each of these methods can be performed individually using different tools, generally proposing adjustable parameters and thresholds. However, such tools are sparse and their diversity contributes to reproducibility problems. In this methodological project, I worked on the establishment of a single analytic workflow combining investigation of genes sets through different but standardized approaches. This workflow, PROTEDEX, aimed to propose an integrative way to analyze genes ensembles relying on different but complementary methods: data-mining based on genomes annotations and GO Terms enrichments,

clustering based on differential expression, identification of transcription factor regulators, establishment of protein interaction networks and assessment of homology relations.

RESULTS

Chapter 1

1) Introduction to "Transcriptomic analysis to identify conserved genes and mechanisms involved in stress response and adaptation in vertebrate species"

The main project during this thesis was conducted in the framework of a collaborative program involving academic (CNRS, INRAe) and industrial (Adisseo) partners. Adisseo is a world-leader company developing nutritional solutions for farm animals and is interested in individual adaptive mechanisms in response to environmental changes. Adisseo research and development strategy aims to develop new tools for the investigation of animal stress state and to provide effective nutritional solutions to improve livestock adaptability, by designing preventive and curative strategies based on precise nutrition evaluation, nutrient health effect and farm practices. Indeed, challenging environmental factors impose biological constraints on living organisms. To manage such constraints, animals trigger a stress response. However, stress exposures have measurable impacts on livestock systems, similarly to human health. In the recent years, investigations were conducted on cellular stress models to characterize stress defense processes, but the conclusions were poorly translatable to whole animals and the mechanisms participating in stress response at the organism level remains only partially understood. Supported by the initial description of stress response as a stereotypical mechanism conserved between related species, we proposed that some important nodes of regulation must have been conserved during evolution and that these nodes could be involved in stress response independently of the stressor nature. To test this hypothesis and identify conserved genes involved in such response, we undertook a comparative analysis between different species exposed to different stress conditions. At this point, it came to the question what stress model to use, including animals and challenges. For the animal, we choosed to compare different species, although within one define evolution group, vertebrates, to facilitate the identification of homologous genes. Concerning the challenge, it needs to be well parameterized and controlled. We selected two farm animals, chicken and pig, with related classical rearing challenges and one laboratory animal model, a transgenic mouse with induced stress sensitivity. Considering that there is no ideal stress model, the requirement was to choose any previously established animal model in which physiological indicators have been identified to characterize their response to stress. Each collaborative partner designed one of these controlled and standardized stress models: chickens exposed to heat (INRAe Tours), pigs exposed to heat and inflammation (INRAe Rennes), chickens fed with an unbalanced diet (Adisseo) or transgenic mice presenting an increase susceptibility to oxidative stress submitted to physical exercises (CNRS Strasbourg). To investigate animal stress response, we performed transcriptomics analyzes to measure and to contrast gene expression between stress and control conditions for each model. Then, a comparison between models was conducted to determine a set of conserved genes activated or

repressed in response to challenging conditions and that are sensitive and quantitative enough to be used as biomarkers of the animal stress status. Indeed, the validity of biological indicators currently used for stress diagnostic is a subject of debate and so far none of these markers has been recognized as unequivocal. In addition, we also aimed to extend the actual knowledge of the molecular mechanisms underlying the stress response to the physiological level. Particularly, understanding stress response dynamics and determinant factors allowing adaptation are of central interest to manage animal well-being in rearing situations but also to provide better solutions for stress-related diseases in humans.

This work was possible thanks to a CIFRE (Conventions Industrielles de Formation par la REcherche) fellowship from the Association Nationale de la Recherche et de la Technologie.

The results presented in this study were the subject of an international application for patent, entitled “Process for identifying a stress state in a subject and/or for assessing the stress response level in a subject”, Submission number 8427705, PCT/EP2020/056869, Date of receipt 13 March 2020.

In this study, A.L., A.C., N.L.F., Y.M. and M.B. conceived and developed the animal stress models investigated. L.T. realized the bioinformatics analysis of these models including: developing a workflow to process and analyse transcriptomics data, performing comparative analyzes to identify orthology relations, investigating the stress-related genes expression profiles using parallel coordinates representation, performing data-mining analyzes, generating protein-protein networks and analyzing their architecture using clustering methods, assessing ontological analysis for regulatory pathways and developing TRACE module algorithm for transcription factor characterization. A.R. and M. B-T. performed the experimental validation of the bioinformatics results. All authors provided critical feedback and helped shape the research, analysis and manuscript. L.T. and A.L. drafted the manuscript.

2) Manuscript I

Identification of a set of genes involved in stress response and adaptation, conserved during evolution between vertebrate species.

*Luc Thomès[1], Ahmad Rida[1], Mélanie Braye-Thami[1], Christelle Hennequet-Antier[2], Aurélien Brionne[2], Vincent Coustham[2], Nathalie Le Floc’h[3], Anne Collin[2], Yves Mercier[4], Mickael Briens[4] and Alain Lescure[1]

[1]Architecture et Réactivité de l'ARN, CNRS, Université de Strasbourg, Strasbourg, France.

[2]INRAE, Université de Tours, BOA, 37380 Nouzilly, France.

[3]INRAE, AGROCAMPUS Ouest, PEGASE, 35590, Saint-Gilles, France

[4]ADISSEO FRANCE SAS, Commentry, France

Identification of a set of genes involved in stress response and adaptation, conserved during evolution between vertebrate species.

*Luc Thomès[1], Ahmad Rida[1,] Mélanie Braye-Thami[1], Christelle Hennequet-Antier[2], Aurélien Brionne[2], Vincent Coustham[2], Nathalie Le Floc'h[3], Anne Collin[2], Yves Mercier[4], Mickael Briens[4] and Alain Lescure[1]

Affiliation:
[1]Architecture et Réactivité de l'ARN, CNRS, Université de Strasbourg, Strasbourg, France.

[2]INRAE, Université de Tours, BOA, 37380 Nouzilly, France.

[3]INRAE, AGROCAMPUS Ouest, PEGASE, 35590, Saint-Gilles, France

[4]ADISSEO FRANCE SAS, Commentry, France

Corresponding author:
Alain Lescure
Architecture et Réactivité de l'ARN, CNRS, Université de Strasbourg,
IBMC-2 allée Konrad Roentgen, F-67000 Strasbourg, France
Tel: +33 388417106
E-mail: a.lescure@cnrs-ibmc.unistra.fr

Runing Title: Conservation of genes involved in stress response and adaptation

Key words: stress response, adaptation, evolutionary conservation, extracellular matrix, TGFbeta pathway

*Order of the authors is only indicative

SUMMARY

Stress has been defined as a specific syndrome induced by a large variety of challenging factors. Here, we identified using comparative transcriptomic analyzes a limited set of 26 conserved genes transcriptionally regulated in response to different stressors in the three vertebrate species *Gallus gallus* (chicken), *Sus scrofa* (pig) and *Mus musculus* (mouse). We observed that most of these genes are co-expressed during stress response, suggesting their involvement in one common pathway. In addition, it appeared that these genes encode secreted proteins from the matricellular family. By ontological and network analyzes, we showed that these proteins of the extracellular matrix (ECM) are largely interconnected and are part of a larger network of ECM-related proteins. Expression clustering of stress-related genes predicted these genes to be regulated by TGFß-, SRC- and CD44-mediated signaling pathways, which is consistent with the current knowledge about TGFß activity. Taken together, these results indicate that genes coding for extracellular proteins are highly responsive to stress exposure, participating to a common biological process conserved among vertebrate species.

INTRODUCTION

Stress induced by exposure to challenging environmental conditions has a major impact on individuals health and welfare, not only in human, but also in livestock for which global climate changes and anthropogenic actions impose additional stressful constrains. Interventions to reproducibly improve individual stress resilience or adaptive abilities have been searched for long, but so far the mechanistic of the stress defense system remains hindered by the complexity and the multiplicity of interacting conditions and intervening factors. Hence, it is required to develop increased knowledge about the molecular and physiological mechanisms underlying adaptive stress response, for guiding the optimal resources needed for preventing, controlling, or mitigating exposure to risks. At the physiological level, hormones of the hypothalamus pituitary adrenal (HPA) axis were shown to play a central role in vertebrates, by mediating the communication between the different organs and controlling or prioritizing survival over less essential physiological functions[1]. Studies conducted on cell culture models showed that most challenging environmental conditions translate into an increased oxidative stress, activating transcription factors such as HIF-1α or NRF2 that control expression of an antioxidant program[2]. However, how the antioxidant system and indeed the organism as a whole respond to elevated oxidative stress *in vivo* is not well understood. The inner workings of this system and its relationships with the network of hormonal communications *in vivo* could only be efficiently addressed using dedicated animal models.

The major challenge with stress studies resides in the multiplicity of stress factors and the large variability in response between species or individuals. However as pointed by Hans Selye, stress is a characteristic physiological state with stereotypical manifestations, including loss of weight, increased adrenocorticotropic hormone production, decreased circulating glucose concentration, hypotension, despite the diversity of the stress factors[3]. This situation indicates that diverse challenging conditions should be processed by different integrative pathways, which might converge to a limited number of conserved sequences[4]. Numerous studies have been conducted to unveil new mechanisms underlying stress response. In the post-genomic era, systemic and non-biased screening approaches based on deep-sequencing technologies and bioinformatics analysis were developed. These studies led to the identification of genes differentially expressed in different stress conditions or species, with particular interest on the more differentially expressed ones, conducting to the characterization of related biological processes. Comparison of the data obtained between different studies showed little overlap, pointing to the fact that these studies

focused more on specificities of the stress response according to stressors or species than common characteristics[5,6,7,8]. Moreover, time course analyzes of stress response showed that the list of the most differentially expressed genes largely varies at different time points, indicating a dynamic process with sequential expression patterns[9]. These specificities and expression dynamics reflect the plasticity of biological systems to adapt to various stress situations. However, according to the observation of stereotypical manifestations at the physiological level, it is predicted that conserved underlying mechanisms also contribute to stress defence. These mechanisms conserved during evolution among species could correspond to central nodes important for integration and regulation of a larger network of genes involved in stress response. Comparative analyzes should help to disclose such conserved mechanisms. These comparative analyzes should consider not only the most differentially expressed genes in each model but include all the genes differentially expressed between stress situations and different species.

In our project, we have conducted transcriptomic analyses to compare stress-related gene expression patterns to identify regulation programs in four standardized stress models including three different vertebrate species, namely chicken, pig and mouse, submitted to different stress conditions. This study was conducted on muscle, a dynamic tissue largely impacted by the metabolic perturbations generated by exposure to stressors. We identified a set of 26 differentially expressed genes common to the different species and stress models that displayed a common expression signature between controls and adapted or non-adapted stressed animals. Most of these genes encode proteins that are constituents of the extracellular matrix. In addition, RNA-seq-based pathway and network analyzes indicate that the identified set of genes is involved in a common biological process control by the TGFß signaling pathway, probably involved in cell to cell communication and intracellular signal transduction, in agreement with response to stress.

RESULTS

Identification of stress-related genes in four stress/species models

In an attempt to identify the models to study stress response, we had to choose which animal species to use and which stressors to apply. The species selected were limited to the vertebrate group to facilitate the identification of homologous genes and biological processes. As there is no optimal stressor or mode of exposure, we proposed that the most important factor is to dispose of previously studied stress models for which physiological indicators or molecular markers of the stress response have been characterized allowing to unambiguously distinguish stressed from unstressed animals. The models were selected to consider a combination of various vertebrate species and controlled stress situations, and the transcriptomic analyzes were conducted on a tissue that dynamically respond to multiple stress challenges. Loss of weight being a common feature observed in many stress situations and largely attributed to change in muscle mass, muscles were considered. Therefor we chose to study four well-characterized animal stress models, wherein three different species were submitted to four different stressors: (i) Chickens submitted to heat challenge[10]; (ii) Pigs submitted to heat and LPS-mediated inflammation challenges[11]; (iii) Chickens submitted to nutritional challenge consisting in low crude protein level supply; (iv) A mice depleted for the *SelenoN* gene submitted to physical exercise challenge. This transgenic mice strain was shown to present an increased sensitivity to oxidative stress leading to muscle dystrophy when submitted to forced exercise swimming-test[12]. These stress models were used to identify genes with increased or decreased transcriptional expression following exposure to the stressors.

Previous analyzes defined for each model a set of physiological indicators characteristic of challenged animals compared to controls, such as weight loss, blood hormone markers, measurements of oxidative level. In our experiments, analyzes of these markers indicated that the responses of stressor exposed animals were highly variable, some of them closer to control reference values, and the other ones significantly different (data not shown). To take into account this disparity among the stressed animals, we defined two subgroups: the adapted and the non-adapted animals, with parameters convergent or significantly different from control animals respectively. In the chickens submitted to heat challenge model, the adapted group consisted in animals with a pre-exposure to increased temperature during embryogenesis[10]. For the three remaining models, loss of weight in challenged animals was mostly considered in addition to other parameters as detailed in Material and Methods. For

each group - adapted, non-adapted and control – a set of four animals was selected (Figure 1). Transcriptomic analyzes were conducted on muscle, a dynamic tissue highly responsive to stressor exposure. Total RNA was extracted from muscle samples and subjected to Illumina RNA-sequencing (RNA-seq) technology. About 40-99 million reads were obtained for each individual sample (Supp Table1). Quality of data acquisition was validated using FastQC[13]. SARTool[14] was used to evaluate the data dispersion and the normalization procedure of the gene expression for each sample (Supp Table 1). Based on the transcriptomic data, a MDS plot representation verified the relative clustered distribution of the analyzed animals into the three identified groups – control, adapted and non-adapted, for each model (Figure 2A to D), despite their large inter-individual variability. However, for the chicken/heat stress model, one of the four control samples (a_ctrl) showed an atypical variability compared to other animals of his group (Figure 2D), and therefore was considered as an outlier and removed from the control set for the gene expression comparative analysis. For the pig/heat and inflammation stress model, we noticed an unusual clustering of the animals into two sets, independently of the stress context, that could be attributed to sex differences mainly (Figure 2B: animals clustered on the left side corresponded to females except j-adapt, and animals grouped on the right side corresponded to males). Therefore, a blocking factor "sex" was introduced for the comparative gene expression analysis to take into account the stress response differences between male and female pigs. Comparative analysis between the three categories of stressed and control animals identified a list of statistically significant differentially expressed (DE) genes in each stress model. The threshold of differential expression level was defined based on the chosen padj-value ≤ 0.05. Number of DE genes between the stressed and control animals is reported in Table 1. Remarkably, the comparison of non-adapted versus control showed the highest number of DE genes in most cases, but in one model (the chicken/heat stress model). This observation suggested that non-adapted animals are more divergent from controls than the adapted ones, based on the expression of their genomic program. This is not the case for the chicken/heat stress model, in which adapted animals are the more divergent from controls, but in this case adapted animals were exposed to a pretreatment during embryogenesis that induced a pre-adapted state, inducing the setting of a genomic program that stimulate acclimation.

To compare the models together, we first merged all DE genes in response to stress, or between adapted and non-adapted animals into one list for each model. To manage inter-species nomenclature heterogeneity, we searched the human ortholog for each gene; the

human gene name was chosen for this procedure as it permitted to use well-annotated gene databases for further analysis. Then, the lists of DE genes were compared between the four models. A Venn diagram representation (Figure 2E) showed that four genes were conserved between the four models. Because of the evolutionary distances between the animal species considered and to avoid orthologous misassignment biases, in particular in gene families composed of a large number and highly similar paralogous genes, we chose to extend the list to genes conserved between at least three of the models, giving rise to a list of 26 conserved genes differentially expressed during the stress response process (Table 2).

Differential expression of the stress-related genes was validated by RT-qPCR experiments for the two models chicken/nutritional and mouse/exercise stress models (Supp Figure1). Noticebly, the 26 conserved genes identified through the comparative analysis are not part of the list of the top differentially expressed genes for each model.

Analysis of differential expression of the 26 stress-related genes showed that they were not similarly regulated following stress exposure, between the four models of stress, but also for equivalent comparison between adapted or non-adapted and control animals (Figure 2F). Depending on the stress model, each gene can be differentially expressed, but not necessarily in the same orientation. As an example, the CHAC1 gene was down-regulated in non-adapted versus adapted animals in the chicken/nutritional and pig/heat and inflammation stress models, as it was up-regulated in the same comparison in the chicken/ heat stress model and not significantly deregulated in the mouse/exercise model; the KERA gene was up-regulated in the adapted and non-adapted conditions versus control in the pig model, as it was down-regulated in the equivalent conditions in the stressed mouse and down regulated in non-adapted versus both control and adapted animals in the chicken/nutritional stress model. This variability in stress-induced expression of the 26 conserved genes might reflect the differences between the modes of stress applied in each model, resulting in different sequences and progression in the stress response program.

Expression profile of the conserved differentially expressed genes during stress

To further investigate the transcription program of the 26 stress-related genes, we used a parallel coordinates visualisation to compare their expression profile in each model (Figure 3A). This representation showed that even if the differential expression level of those genes was not the same, most of them shared the same expression profile across stress models

being almost always up-regulated or down-regulated together. This convergent expression is clear for the two chicken and the pig models, although more variable in the stressed mouse model. The most divergent gene compared to this overall trait was CHAC1. This observation indicated that a major part of these genes are co-expressed, possibly controlled by one signaling pathway or transcription regulation mechanism. In addition, this observation suggested that these co-expressed genes are part of the same biological process, playing a functional role together.

Ontological analysis of the conserved stress-related genes

To define common features in the identified set of genes, we conducted an enrichment computational analysis using the STRING website[15] based on the gene ontology (GO) terms. This analysis pointed out that most of the genes are linked to the biological process terms "collagen fibril organization", "extracellular matrix organization", or different response pathways related to cellular signaling (Table 3). Combining literature and Uniprot information, we characterized a sub-list of 14 proteins out of 26 that localize outside of the cell, in the extracellular matrix (ECM) or being secreted (Table 2). This analysis also provided evidence that a part of them are members of the matricellular protein family. This family consists in extracellular proteins found in the ECM, but not only involved in its structural architecture. Matricellular proteins are known to participate in several processes like the regulation of cellular adhesion, differentiation and proliferation, cell-cell interactions and also signal transduction pathways to influence normal cell functions[16,17].

In addition, the STRING website provided a network analysis based on validated or predicted protein relations, that showed that 15 out of the 26 proteins of interest were interconnected (13 as part of a principal network, and two others) (Figure 3B). This network presented significantly more interactions than expected for a random set of proteins of same size (p-value 8.18e-13).

Together with the co-expression of the genes during stress response and the co-localization evidences for the encoded proteins, this connectivity suggests that the identified genes are involved in a common biological process occurring outside of the cell at the ECM level. Information obtained indicated that the stress-related genes code for proteins that are not only structural components of the ECM, but rather involved in cell fate regulation through cell-cell communication and intracellular signal transduction, in agreement with their involvment with response to stress.

Contribution of the identifed stress-related genes to other stress responses

The obtained list of genes and the link between stress and ECM was rather unexpected as the 26 genes identified in the present study are not part of the classical genes characterized in the majority of stress state or stress response studies, such NRF2 or HIF-1α related genes. To confirm their involvement in this biological process, we performed a data-mining search in the GEO database (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/) from NCBI. This publicly available database contains raw transcriptomic data from RNA-seq and microarray experiments and can be queried online or by command-line requests. We searched for studies that can be associated to each gene of the list. The number of genes associated with each experiment is displayed on Figure 4A. We found three experiments, each showing a maximum of 11 DE genes out of the 26 genes (but not the same 11 genes in each of the three experiments), one of them about skeletal muscle response to a physical exercise, and the other ones related to cancer diseases investigations (Supp Table 2). Interestingly, most of the studies identified using this method were related to stressors exposure, extracellular matrix-related diseases, or aging processes. The stress-related genes most often associated to transcriptomic studies are ANXA1 and 2, LGALS1, COL1A1 and THBS1 (Figure 4B). Of note, ANXA1 and 2 are also two of the four DE genes conserved between the four stress models that we analyzed.

To extend this analysis, the expression level of the set of conserved genes was investigated in another stress model, consisting in chicken exposed to a xenobiotic. Chickens were provided the reactive oxygen species-producing reagent paraquat in drinking water for one week. RNA was extracted from muscle and gene expression levels for the 26 genes were determined by quantitative PCR and compared to control animals (Supp Figure 2). The results showed that the expression of a majority of the studied genes was up- or down-regulated in this stress model as well, although displaying a unique pattern compared to the previous stress models. CHAC1 was the only strongly up-regulated gene, but eleven genes were down-regulated more than two folds.

Expression of the stress-related genes in the liver tissue during stress

Next, we investigated whether the change in expression of the set of the 26 genes was restricted to muscle tissue, or could be extended to other tissues, such as liver. Liver is an actively regulated organ to adjust body metabolic needs in response to stress. Quantitative PCR analyzed the level of expression of the genes from the chicken/nutritional and

mouse/exercise stress models in this tissue. Expression was compared between adapted or non-adapted and control animals as previously. Similarly to what was described in muscle, for each situation, it is a subset of genes significantly differentially expressed for each analysis. However, differential expression for the 26 genes was highly variable in the liver for the two models and between comparisons. In addition, the expression of the genes was different between liver and muscle tissues for most of the comparisons (Figure 5). This observation support the idea that a limited number of genes will not be sufficient for stress diagnostic at the animal level, but that accurate test will rather requires an ensemble of gene markers more likely to reflect the dynamic processes taking place during stress response.

Intra-model network structure and functional analyzes

To gain details into the RNA-seq data, we computed a network analysis for all DE genes in each model. STRING networks generated for each model contained 1259, 440, 295 and 265 nodes for the mouse/exercise, pig/heat and inflammation, chicken/heat and chicken/nutritional stress models respectively (Table 5). The obtained networks contained regions with high density of nodes, corresponding to clusters of highly interconnected proteins (Figure 6). These clusters were extracted from the four networks (Supp Table 3). For each model, the number of clusters and number of associated proteins is detailed in Table 5. Functional analyzes of the largest clusters showed that the proteins contained within these clusters were divergent between the four networks, but considering protein families and related functions instead of protein identity, revealed that clusters often contained proteins belonging to the same family or involved in the same biological pathway, such as collagen family, the ubiquitin family, thrombospondin (THBS)-related proteins or chemokine-related immune response (Supp Table 3).

Central regulatory nodes or hubs, corresponding to protein connected to a large number of partners that are not necessarily interconnected together, is another characteristic of networks. For the networks corresponding to the four stress models, the proteins with the highest node degree were characterized (represented with red dots on Figure 6) and this led to the identification of two proteins, CD44 and SRC, playing a central role in the chicken or mammalian models networks respectively. Interestingly, these two proteins are involved in the same signaling pathway: CD44 is a transmembrane receptor that binds hyaluronic acid and activates SRC, a modulator of several signaling pathways, including the focal adhesion kinase (FAK) and the PI3K-AKT pathways[18].

Alltogether, these observations suggested that biological functions are more conserved and important for stress response than the identity of individual proteins.

The stress-related genes are part of a larger network of co-expressed genes

The co-expression and co-localization of the 26 stress-related conserved genes, led us to the hypothesis that the transcription of these genes could be controlled by a unique signal or transcription regulation program. This observation raised the question of other less conserved genes, although co-expressed, being potential partners of the core set of 26 genes. Using an expression-based clustering method, we identified in our models genes following the same expression pattern than the core set. This clustering method allowed to extract four lists of co-expressed genes in each model. To avoid any bias introduced by one single model and to stay consistent with the idea of evolutionary conserved genes, we decided to consider only genes conserved between at least two models. A list of 93 of co-expressed genes, including 18 genes over the 26 of the intial core set was obtained. As previously, we computed an ontological enrichment for this new set of 93 genes. The list of "biological process" terms is largely convergent with the one obtained for the core set (Table 4), with "extracellular matrix organization" as the most significative term in both cases (false discovery rate of 5.80e-7 for the co-expressed genes and 4.40e-6 for the conserved stress-related genes). Then, a network was generated showing that 71 over the 93 proteins are inter-connected (Figure 7), including 16 over the 18 genes of the conserved protein set (green discs, Figure 7), and 14 genes related to extracellular matrix organization (red circles, Figure 7). This result showed that the genes co-expressed with the set of conserved genes during stress response are also largely interconnected and the new network extended the number of genes related to ECM. In addition, most proteins of the conserved or of the ECM-related sets occupied a central position within the network, being interconnected and connected to other nodes, suggesting their important role in the stress response. Remarkably, the two proteins CD44 and SRC, previoulsy identified as central regulatory nodes in the network analysis of the stress models taken individually, also occupy a central position in this extended network of co-expressed genes.

Then, we tested this extended list of 93 co-expressed genes to find a potential common transcriptional regulator. For this purpose, we used the TRACE method developped as part of the PROTEDEX protocol (see chapter 3.2). Basically, based on the three databases compiling transcription factors and their target genes ENCODE[19], TRANSFAC[20] and CHEA[21], this method predicts two metrics for each transcription factor: (i) its activity

corresponding to the involvement of the transcription factor in the regulation of a set of genes compared to the full-list of genes controlled by this factor; (ii) its influence that estimates the impact of the transcription factor regarding the set of genes of interest. Moreover, a statistical value is computed to evaluate the significance of each hit (p-value). A volcano plot positioning each transcription factor according to the activity and p-value metrics was drawn for the list of the 93 co-expressed genes (Figure 8). This predicted that five transcription factors have the highest probability to control the expression of the set of 93 genes (log2 of activity >0.66 and –log10 of p-value >6 and influence >20%): SOX2, SMARCA4, PPARG, SMAD3 and SMAD4. SOX2 is a transcription factor controlling the expression of a number of genes involved in embryonic and neuronal development[22]. SMARCA4, also known as BRG1, was shown to be a regulator of autophagy in response to oxidative stress and is also part of a chromatin remodeling complex recently identified as a regulator of ECM-related genes expression[23,24]. PPARG is a well-known transcription factor involved in the control of lipids/carbohydrates homeostasis[25]. Concerning SMAD3 and SMAD4, these two proteins of the same family are essential for the activity of the TGFß cytokine signaling pathway. TGFß is an extracellular protein that become activated upon cleavage by ECM residents proteases, such as matrix metalloproteinases (MMPs) or THBS1[26,27]. Noticebly, THBS1 is one of the 26 stress-related proteins identified as conserved between the stress models in this study. Once activated, TGFß displays two different functions: a canonical and a non-canonical one. The canonical function leads to the recruitment of SMAD2, SMAD3 and SMAD4 proteins to the nucleus where they regulate transcription of their target genes. The non-canonical function leads to activation of different signaling pathways, including the PI3K-AKT signaling cascade. Moreover, TGFß has been described to interact directly or indirectly with CD44 and SRC, two other modulators of these signaling pathways[27,28].
Altogether, the analyzes of the set of co-expressed genes pointed to the major importance of the TGFß pathway and several of its effectors in stress reponse.

DISCUSSION

In this study, we developed an evolutionary comparative approach to identify genes important for stress response and conserved in vertebrate species. Similar approaches have been successfully applied to address other complex biological processes, such as aging. In this case, it led to the identification of a conserved pathway involving branched-chain amino acid catabolism, controlling aging between distant species[29]. The main asset of this approach is that it allows us to hierarchize experimental observations and to characterize their importance based on conservation during evolution. It is assumed that processes that have been selected over a broad period are more likely to constitute the central nodes of elaborate biological networks controlling complex functions. By comparing stress response at the gene transcription level in three different species exposed to four independent stressors, we identified a limited number of 26 genes differentially expressed and conserved between at least three of these models.

Functional analysis of the 26 genes revealed that most of the encoded proteins localize in one cell compartment, the extracellular matrix (ECM). This observation was confirmed by ontological investigations, and network analyzes demonstrated the high connectivity of these co-localized proteins. Integration of the conserved stress-related genes into an extended set of co-expressed genes reinforced their affiliation to ECM-related processes based on ontological enrichment. ECM was considered for long as a structural architecture assuring cell physical cohesion in tissues. However, it is now recognized as a dynamic compartment and that many of its components, called matricellular proteins, constitute key signals controlling cell functions and cell to cell communication, therefore important for cell adhesion, migration, differentiation and proliferation[16,17,30]. In addition to its well-known role in organogenesis during development, the ECM components have been shown to be important for other normal or pathological biological processes, such as cancer and aging. During tumorigenesis, It has been established that cancer cells interact with neighbor stromal cells creating a microenvironment, named pre-metastatic niche, favorable for attachment, survival and growth of circulating tumor cells. Several soluble factors and cells involved in the pre-metastatic niche formation have been identified, acting through different mechanisms, which converge to modifications of ECM components, including collagen and matrix metalloproteinases[31]. ECM remodeling was also shown to play a central role in cellular senescence a process important to limit tumor progression and favor tissue repair. However, long-term presence of senescent cells in tissues may have detrimental role in promoting tissue damage and aging. Cellular senescence is associated with changes in

expression and secretion of cytokines and chemokines, together with ECM components and remodeling enzymes[32]. To the best of our knowledge, ECM-related genes were not previously characterized as part of the stress defense mechanism. However, in a study addressing the importance of the insulin regulatory pathway on aging control in the nematode *Caenorhabditis elegans*, Ewald et al.[33] demonstrated a decrease of collagen expression during aging, and conversely the extension of worms lifespan with increased production of collagen and other ECM-related proteins. It was shown that the regulation of ECM production is controlled by the PI3K-AKT signaling axis acting on the NRF2 transcription factor[33]. This transcription factor being known to play a central role in stress response, it indicated a coupling between stress defense and ECM remodeling mechanisms during aging. Altogether, these observations indicate that identification of ECM role in stress response was possible only because the studies were conducted at the organism level, highlighting processes that normally escape cell culture analysis This indicated that our work provide another layer of complexity to the study of stress response.

Investigations on the organization of the extended network aggregating all genes with similar expression pattern as the set of conserved stress-related genes pointed to the presence of highly interconnected groups or clusters of extracellular factors. Comparison of these ECM-related clusters showed limited conservation of proteins identity among models. However, it revealed that these clusters are mainly constituted by multigene families coding for many paralogous proteins, such as the collagen, chemokine or thrombospondin families. Compiling these data demonstrates that biological functions carried out by paralogous proteins are more conserved and probably more important for stress response, than the individual proteins themselves. One possible explanation is that families of paralogous proteins provide a functional adaptability to diverse situations where each member is likely to perform a specialized activity, some of them dedicated to stress response. Alternatively, specialized paralogs might be differentially recruited in stress defense according to the nature of the stressor.

Interestingly, the analysis of the expression profile for the set of 26 conserved genes revealed their co-expression in the different subgroups – adapted and non-adapted - of each model, leading to a convergent expression pattern. This behavior suggested the existence of a common signaling pathway or transcriptional control regulating the expression of the set of conserved genes. However, the comparison of their expression profiles between models showed a highly variable expression depending on stress models, species or tissues

considered. One should keep in mind that stress response is a dynamic process, the kinetic of which depends on multiple factors, such as nature and intensity of the stressor, period of exposure, pre-exposure to the stressor, environmental context, as well as sex, age and individual genetic or epigenetic predispositions. Because of this, it is impossible to predict the progression of the stress response programs in the different animal models at the time the tissue samples were collected. Kinetic study of the expression of the stress-related genes identified in this study following exposure to stressor should provide important information on the mechanistic and their contribution to early stress integration or later adaptive processes. This observation also sheds light on why looking for one universal marker for stress has failed so far. A combination of consistent markers is more likely to represent the perturbations state of an organism within its environment and the dynamic of the stress response. Transcriptomic is a relevant method to address this question since it provides unbiased analysis on a large set of biological parameters, representative of the whole gene expression program.

Taken together, co-expression of the genes during stress response and co-localization of their interconnected products stressed the importance of the ECM and suggests a contextual adaptation and remodeling of this compartment implicating multigenes families and induced through a common regulatory mechanism.

We searched for potential factors controlling a common regulatory program in response to stress by combining complementary investigations based on co-expression and network analyzes. First, among the proteins of the generated network, two factors, CD44 and SRC, were of particular interest: they are cellular partners involved in the control of the same signaling pathway. These two signaling molecules were identified because of their particular position in the network, central and highly connected to other nodes, suggesting their role as important regulators. Parallel investigations to characterize a transcriptional regulator controlling the set of co-expressed genes pointed to two transcription factors, SMAD3 and SMAD4. Interestingly, both SMAD3/4 and CD44/SRC are implicated in the TGFß signaling pathway. Collectively, these results converge to the importance of TGFß signaling during stress response, controlling the expression program through a canonical pathway involving SMAD3 and SMAD4, or a non-canonical pathway acting jointly with SRC and CD44 on common signaling axes (Figure 9).

Remarkably, the importance of the TGFß pathway was also observed during aging and in age-related pathologies in vertebrate species[34]. Multiple studies described an up-regulation of TGFß in elderly individuals in connection with the cellular senescence mechanism. Investigations of the relations between the production of TGFß in senescent cells and the development of aging-related pathologies have highlighted an alteration of TGFß levels during Alzheimer disease, sarcopenia, osteoarthritis, cardiovascular disease, as well as obesity. Because of its cell-type-, context- and age-dependent activity, the precise mechanism by which TGFß act on these processes has not been clearly identified yet, but the TGFß pathway was demonstrated to provoke cell degeneration and reduction of regenerative capacities after injury. Altogether, these observations revealed the tight interconnection between TGFß activity, aging, age-related pathologies and response to injuries. Our results are convergent with this knowledge and provide hints to dissect the relations between stress response capacities and aging. Further analyzes should help to understand how stress response dysfunctions could connect to age-related diseases.

The information obtained through this research program is expected to provide the core knowledge essential to the development of a multi-species generic model of the stress response, allowing a precise handling of stressor challenges and the adaptive processes. To validate the relevance of such set of 26 conserved genes during stress response, one could ask about their conservation in more distant species. So far, we identified fifteen homologous genes from this set in the *C. elegans* species. Remarkably, homologs of TGFß, that we predicted here to be a key regulator of stress response, were identified in distant organisms such as corals and it was demonstrated the capacity of the TGFß pathway to modulate immune response to adapt the symbiotic relations between the coral host and its unicellular algae in response to heat[35,36].

Importantly, a combination of the newly characterized stress-related genes could constitute relevant stress biomarkers. Such set is anticipated to provide useful effective tools for stress status diagnostic in individuals and to direct appropriated prophylactic bioindications.

MATERIAL AND METHODS

Stress model design

Four animal stress models were selected and designed based on previous studies, wherein three different species were submitted to four different stressors: (i) Chickens submitted to heat challenge[10]; (ii) Pigs submitted to heat and inflammatory challenges[11]; (iii) Chickens submitted to nutritional challenge; (iv) *SelenoN* transgenic mice submitted to physical exercise challenge[12].

Briefly, in the chicken/heat-stress model, eggs were maintained either at 37.8°C and 56% relative humidity during the whole incubation period or thermal-manipulated by incubation at 39.5°C and 65% relative humidity for 12h/24 from embryonic day E7 to E16 included. After hatching, male chicks were transferred to a single poultry house and reared from day 0 to day 34. The temperature was decreased from 33°C at day 0 to 21°C at day 25 and maintained at 21°C thereafter. On day 34, control or thermal manipulated chicken groups were exposed to 32°C for 5h. Animals without heat-challenge during embryogenesis and reared under standard conditions were used as controls and were characterized by body temperature of 41.2 +/- 0.1°C. For the gene expression analysis, animals better tolerating heat by means of embryo heat acclimation were selected for low body temperature (adapted group, 42.2 +/- 0.2°C) compared to the non-adapted and control groups that presented significantly higher body temperature (42.9 +/- 0.9°C; $P < 0.01$). They were slaughtered and breast muscles were recovered, snap-frozen and maintained at -80°C until further analysis.

Concerning the pigs/heat and inflammation model, 77-day old pigs were kept constantly at 24°C during a 14-day adaptation period, then divided into two groups, wherein animals were either maintained in thermo-neutral condition (24°C) or exposed to high temperature for 17 days. For the high-temperature group, the room was kept at 24°C during 5 days, then gradually increased to 30°C. Starting day 8 of the heat challenge period, pigs were administrated five injections of LPS from *E. coli* on days 8, 10, 12, 14, 16 of the heat stress period. Pigs were weighted individually at the beginning and at the end of the experimental period and rectal temperature was recorded. All animals were euthanized 24 hours after the final LPS injection. Tissues were collected and stored at -80°C. For the gene expression analysis, stressed animals presenting the largest deviations in average daily weight and feed efficiency compared to controls were assigned to the non-adapted group, and animals exposed to the experimental treatment but with body weight similar to controls were assigned to the adapted group. Qualification into these two stressor-exposed groups was

further validated based on plasmatic analyzes evaluating hormonal response and inflammatory status.

The chicken/nutrition-stress model consisted in chickens fed with two different diets supplying low (17%) or usual (22%) crude protein levels. Birds were put on standard corn-soybean based starter diet (22% CP/3000 kcal/kg) during the two first weeks of life to assure normal development. At day 15, chicken were treated with low or usual protein iso-energetic diet until 6 weeks old. For each condition and for 24 birds per treatments, blood and tissues samples were collected. Plasmatic corticosterol, iodotyronine T3-T4, TBA-RS and glutathione status were measured to evaluate the hormonal and oxidative status difference between dietary treatments during growth. In addition, animals weight was recorded before and after treatment. Based on gain of weight and oxidative parameters, the responses of stressed animals were highly variable, some of them closer to control reference values, the other one significantly divergent. To take in account this variability we defined two subgroups of stressed animals: adapted animals (feed conversion ratio similar to those of control animals) and non-adapted animals (feed conversion ratio significantly different to control animals). Breast muscle samples were collected at week 6 and stored at -80°C until further analysis.

For the mouse/physical exercise-stress model, 8 to 12 months old transgenic *SelenoN*$^{-/-}$ (KO/KO) mice or heterozygotes (KO/WT) mice were submitted to a forced swimming test. In this study, 15 KO/KO and 9 KO/WT mice were set to swim for six minutes each day during two months. Based on their ability to swim and body weight parameters, two subgroups were defined in the KO/KO cohort. The ones showing weight loss and difficulties to complete the swimming exercise were categorized as non-adapted animals, and the ones showing only subtle or no phenotypic alterations were categorized as adapted animals. Blood samples were collected and total oxidation-reduction potential capacity of the plasma was measured using the RedoxSYS® system (Luoxis, Englewood, USA). The values obtained for the stressed animals of both adapted and non-adapted subgroups compared to the KO/WT were in agreement with the loss of weight parameter. At the end of the two-month experimental period, animals were euthanized, paravertebral muscle tissues were collected and stored at -80°C until further analysis.

Model of chicken exposed to paraquat-induced stress ("chicken-paraquat model")

For this additional experimental model, a total of 144 one-day-old Ross 308 male broiler chicks, with an average body weight (BW) of 39 g, were reared from D1 to D21. They were allocated in 72 battery cages (0.5 × 0.42 m^2) with wire floors (6 chicks/cage) in environmentally controlled rooms. The birds were randomly assigned to treatment pens with similar starting weights. Each cage was equipped with one trough feeder and one drinker. Birds had *ad libitum* access to mash feeds and water during all study. Average temperature was 33°C at placement, being reduced by 1°C every 2 days until 23°C to provide comfort throughout the study. The lighting program was 18 hours light and 6 hours dark during each 24 hours period throughout the trial.

The experimental design was a completely randomized factorial design, consisting of a placebo or oxidative stress groups thus two experimental treatments, and 12 replicates per treatment with 6 birds for each replicate. The oxidative stress was applied only from D7 to D14, through the supplementation of a xenobiotic, i.e. paraquat dichloride hydrate (Sigma-Aldrich Company Ltd., Dorset, UK), through the water supply system at the dose of 110 µg/mL. This dosage was achieved by using water containers, individually located in each cage. The control group received standard water (placebo) on the same period using similar containers. A starter and grower diets were provided from D1 to D7, and D8 to D21, respectively. The basal diets were standard wheat/corn-soy-based broiler diets, and were formulated to meet or exceed the nutrient requirements of broilers, as recommended by the NRC (1994). Body weight (BW) was recorded on D1, D7, D14 and D21. On D14 and D21, one bird per pen replicate of each treatment was sacrificed for tissue collection. Briefly, 100 mg of tissue (breast, liver and ileum) were immerged in 2-mL Eppendorf tubes containing 1 mL RNAlater® (Sigma-Aldrich), and kept at -20°C until analysis.

All experimental procedures used in the current study were approved by the Ethics and Research Committee of the institutions conducting the study.

Total RNA extraction and purification

According to the biological indicators measured in each stress model, we determined four animals representative of the adapted and non-adapted subgroups respectively; four animals of the control group were selected as well. Using a FastPrep-24 5GTM (mpbio®) and 1.4 mm ceramic beads (6913-100, mpbio®), muscle samples were homogenized in Tri Reagent buffer (Sigma®) at 1.5 ml per 100 mg tissue, twice for 40 sec at 6 m/s speed. After centrifugation for 10 min at 12,000 g at 4°C, 1 mL of supernatant was collected. Two hundred

µL chloroform was added and after vortexing, the mix was centrifuged for 10 min at 12,000 g at 4°C. Five hundred µL of the upper aqueous phase was collected and RNAs were precipitated by addition of one volume isopropanol 100% at room temperature for 10 min. After centrifugation for 10 min at 10,000 g at 4°C, supernatant was removed and the pellet was dried at room temperature for 10 min. The RNA pellet was resolubilized in 50 µL of RNase-DNase free water and incubated for 10 min on ice. RNA concentration was determined using a NanoDrop 1000 spectrophotometer (Thermo Scientific).

Total RNAs, including mRNAs and long–non coding RNAs were then purified, using the RNA Clean & ConcentratorTM-5 kit (Zymo Research). Ten µg total RNA were purified according to manufacturer's instructions. The concentration and the purity of the RNA samples were measured using a NanoDrop, and their integrity (RIN) was evaluated using a Bio-Analyzer 2100 (Agilent Technologies). The RIN values ranged from 8.0 to 9.6.

RNA sequencing

Purified RNAs were reverse-transcribed into cDNAs and sequencing was conducted at the GenomEast Plateforme, IGBMC using the HiSeq Illumina® technology (HiSEQ 4000). FASTQ sequence files containing reads were retrieved.

RNA-seq data processing

HISAT2 tool version 2.0.4[37] with default parameters was used to perform alignment of reads against the genomes, according to genome annotations. The GTF annotations and FASTA genome files used in this step were as followed: (i) For chickens, Galgal5 genome with its associated NCBI annotations; (ii) For pigs, Sscrofa11.1 genome with its associated NCBI annotations and (iii) for mice, GRCm38.p5 (mm10.p5) genome with its associated Ensembl version M14 annotations

Gene expression level was measured by reads counting using the HTSeq tool version 0.6.1[38] with default parameters. Finally, the edgeR tool of the SARTools R package[14] was used to define differentially expressed genes between the three tested conditions: adapted versus control, non-adapted versus control and non-adapted versus adapted, respectively referred as AvsC, NAvsC and NAvsA. To manage samples variability, we used a modified version of the edgeR robust mode, which performs a different dispersion calculation according to the method developed by Zhou and Robinson[39].The same parameters were used for all models, applying the default Benjamini-Hochberg p-value adjustment method. Differentially expressed (DE) genes from each model were defined using a classical threshold for the adjusted p-value (padj) of 0.05.

Model comparisons

Once a list of DE genes was established for each model, we searched the human orthologous name of each of these genes using the BioMart tool accessible online from the Ensembl website (https://www.ensembl.org/). For several genes, no human ortholog could be inferred using this automatic procedure, therefore human homologs were identified by a manual curation process using BLASTp instead and keeping the human best hit name. This step let us manage inter-species nomenclature heterogeneity but also permitted further analyzes using well-annotated genes. We then merged all DE genes of each comparison in each model and compared the four lists to characterize conserved genes involved in stress response. We defined the conserved genes list by selecting all genes in common between at least three of our four models. This choice was made to be more permissive, comparing four strictly different models, and to avoid misassignment during the gene name conversion step mainly because of gene families composed of a large number and highly similar paralogous genes.

Conserved genes network and functional analysis

To investigate the connectivity degree of our genes of interest we used the STRING 11 website (https://string-db.org/) in which edges correspond to predicted and experimentally validated functional associations. The biological roles of these genes were identified and associated to biological functions based on ontological enrichment computed using again the STRING website that provides GO terms enrichments in addition to network representations.

Data-mining from the GEO database

At the time we performed this search (February 2017), the GEO database (https://www.ncbi.nlm.nih.gov/geo/) contained about 95096 experiments. To query this database with our 26 genes, we used a homemade Python script to retrieve all experiments in which at least one of these genes was differentially expressed. We then checked whether these genes were frequently differentially expressed in studies that we identified as stressor exposure experiments.

Intra-model network structural and functional analysis

To build stress response networks, we defined one list containing all genes differentially expressed in at least one comparison for each model: adapted versus control, not-adapted versus control or adapted versus non-adapted. The four lists obtained were used to generate

four protein networks using the STRING website. Network edition and analysis was performed using Cytoscape version 3.7[40] associated to the MCODE plugin providing an algorithm to extract groups of highly interconnected proteins referred as clusters[41]. Central proteins with numerous isolated partners (also referred as "hubs") were identified based on their elevated degree value. The top five proteins of each network were kept and compared between models to identify, common nodes.

Identification and analysis of additional conserved genes

To identify the functional partners of the 26 conserved genes we applied a gene expression-based hierarchical clustering method to group genes by their expression profile within each model. Genes within the same clusters that the 26 conserved ones were retrieved for each model. All the genes identified by this method were finally compared to identify those present at least in two stress models. These common genes were added to the list of the 26 previously identified genes. From this set, a larger protein network was generated and ontological enrichments were computed using the STRING services. In addition, we tested this extended list of co-expressed genes to identify potential transcriptional regulators using the TRACE module with default parameters implemented in the PROTEDEX protocol.

REFERENCES

1. Joseph, D. & Whirledge, S. Stress and the HPA Axis: Balancing Homeostasis and Fertility. *IJMS* **18**, 2224 (2017).
2. Halliwell, B. & Cross, C. E. Oxygen-derived species: their relation to human disease and environmental stress. *Environ. Health Perspect.* **102**, 8 (1994).
3. Selye, H. The Stress of life. United States of America: McGraw Hill Book Co. (1956).
4. Kültz D. Molecular and evolutionary basis of the cellular stress response. *Annu Rev Physiol.* **67**, 225-257 (2005).
5. Wassouf, Z. *et al.* Distinct Stress Response and Altered Striatal Transcriptome in Alpha-Synuclein Overexpressing Mice. *Front. Neurosci.*, **12**, 1033 (2019).
6. Shi, K.-P. *et al.* RNA-seq reveals temporal differences in the transcriptome response to acute heat stress in the Atlantic salmon (Salmo salar). *Comp. Biochem. Physiol. Part D Genomics Proteomics*, **30**, 169–178 (2019).
7. Schüttler, A. *et al.* The Transcriptome of the Zebrafish Embryo After Chemical Exposure: A Meta-Analysis. *Toxicol. Sci.*, **157**, 291–304 (2017).
8. Lee, S.T.M. *et al.* Transcriptomic response in Acropora muricata under acute temperature stress follows preconditioned seasonal temperature fluctuations. *BMC Res Notes*, **11**, 119 (2018).
9. Szustakowski, J.D. *et al.* Dynamic resolution of functionally related gene sets in response to acute heat stress. *BMC Mol Biol*, **8**, 46 (2007).
10. Loyau, T. *et al.* Thermal manipulation of the chicken embryo triggers differential gene expression in response to a later heat challenge. *BMC Genomics* **17**, 329 (2016).
11. Campos, P. H. R. F., Merlot, E., Damon, M., Noblet, J. & Le Floc'h, N. High ambient temperature alleviates the inflammatory response and growth depression in pigs challenged with Escherichia coli lipopolysaccharide. *Vet J.* **200**, 404–409 (2014).
12. Rederstorff, M. *et al.* Increased Muscle Stress-Sensitivity Induced by Selenoprotein N Inactivation in Mouse: A Mammalian Model for SEPN1-Related Myopathy. *PLoS One* **6**, e23094 (2011).
13. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010). Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
14. Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. & Dillies, M.-A. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLoS One* **11**, e0157022 (2016).

15. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
16. Murphy-Ullrich, J. E. & Sage, E. H. Revisiting the matricellular concept. *Matrix Biol.* **37**, 1–14 (2014).
17. Gerarduzzi, C., Hartmann, U., Leask, A. & Drobetsky, E. The Matrix Revolution: Matricellular Proteins and Restructuring of the Cancer Microenvironment. *Cancer Res* **80**, 2705–2717 (2020).
18. Chen, C., Zhao, S., Karnad, A. & Freeman, J. W. The biology and role of CD44 in cancer progression: therapeutic implications. *J Hematol Oncol* **11**, 64 (2018).
19. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
20. Wingender, E. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.***24**, 238–241 (1996).
21. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
22. Novak, D. *et al.* SOX2 in development and cancer biology. *Semin Cancer Biol.* (2019) doi:10.1016/j.semcancer.2019.08.007.
23. Barutcu, A. R. *et al.* SMARCA4 regulates gene expression and higher-order chromatin structure in proliferating mammary epithelial cells. *Genome Res.* **26**, 1188–1201 (2016).
24. Liu, M. *et al.* BRG1 attenuates colonic inflammation and tumorigenesis through autophagy-dependent oxidative stress sequestration. *Nat Commun* **10**, 4614 (2019).
25. Hong, F., Pan, S., Guo, Y., Xu, P. & Zhai, Y. PPARs as Nuclear Receptors for Nutrient and Energy Metabolism. *Molecules* **24**, 2545 (2019).
26. Derynck, R. & Budi, E. H. Specificity, versatility, and control of TGF-b family signaling. *Sci Signal* **12**, 570 (2019).
27. Nolte, M. & Margadant, C. Controlling Immunity and Inflammation through Integrin-Dependent Regulation of TGF-β. *Trends Cell Biol.* **30**, 49–59 (2020).
28. Zhang, H., Davies, K. J. A. & Forman, H. J. TGFβ1 rapidly activates Src through a non-canonical redox signaling mechanism. *Arch. Biochem. Biophys.* **568**, 1–7 (2015).
29. Mansfeld, J. *et al.* Branched-chain amino acid catabolism is a conserved regulator of physiological ageing. *Nat Commun* **6**, 10043 (2015).
30. Adams, J.C. Matricellular Proteins: Functional Insights From Non-mammalian Animal Models. *Curr. Top. Dev. Biol*. **130**, 39-105 (2018).

31. Paolillo & Schinelli. Extracellular Matrix Alterations in Metastatic Processes. *IJMS* **20**, 4947 (2019).
32. Levi, N., Papismadov, N., Solomonov, I., Sagi, I. & Krizhanovsky, V. The ECM path of senescence in aging: components and modifiers. *FEBS J* **287**, 2636–2646 (2020).
33. Ewald, C. Y., Landis, J. N., Abate, J. P., Murphy, C. T. & Blackwell, T. K. Dauer-independent insulin/IGF-1-signalling implicates collagen remodelling in longevity. *Nature* **519**, 97–101 (2015).
34. Tominaga, K. & Suzuki, H. I. TGF-β Signaling in Cellular Senescence and Aging-Related Pathology. *IJMS* **20**, 5002 (2019).
35. Detournay, O., Schnitzler, C. E., Poole, A. & Weis, V. M. Regulation of cnidarian–dinoflagellate mutualisms: Evidence that activation of a host TGFβ innate immune pathway promotes tolerance of the symbiont. *Dev Comp Immunol* **38**, 525–537 (2012).
36. Berthelier, J. *et al.* Implication of the host TGFβ pathway in the onset of symbiosis between larvae of the coral Fungia scutaria and the dinoflagellate Symbiodinium sp. (clade C1f). *Coral Reefs* **36**, 1263–1268 (2017).
37. Kim, D., Paggi, J.M., Park, C. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
38. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
39. Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **42**, e91–e91 (2014).
40. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
41. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 27 (2003).

Figure Legends

Figure 1. Sample collection and data processing scheme of the multi -species/-stress screening method. Total RNA was prepared from muscle samples and sequenced by RNA-seq. After passing quality controls and validation of sample clustering into predefine conditions, the sequences were mapped to the referring genomes. Quality of alignment was controlled. Number of reads for each annotated gene was used for statistical evaluation of significant differentially expressed (DE) genes between adapted, non-adapted and control conditions for each model. For models comparison, genes annotation was homogenized using human genes name as reference. Commonly differentially expressed genes over different stress models were further analyzed.

Figure 2. Comparative transcriptomic analysis of three different species exposed to four stressor conditions. (A-D) Multi-dimensional scaling (MDS) plot of the different RNA samples analyzed for each stress model: **(A)** chicken/nutritional stress; **(B)** pig/heat and inflammation stress; **(C)** mouse/exercise stress and **(D)** chicken/heat stress. Blue, red and purple dots correspond to the control (ctrl), adapted (adapt) and non-adapted (una) animals groups respectively. The distances between dots correspond to the leading log-fold-changes (logFC) between each pair of RNA samples. This plot was obtained using edgeR. **(E)** Venn diagram of all differentially expressed (DE) genes in the four stress models. This representation shows that four genes are differentially expressed in at least one comparison for each model. This list was extended to 26 genes by considering genes conserved in at least three of the four models. This diagram was generated using the Venny 2.1.0 website tool (http://bioinfogp.cnb.csic.es/tools/venny/). **(F)** Heatmap representation of differential gene expression in response to stress of the 26 conserved genes based on the transcriptomic data. Differential gene expression level between two conditions depicted on the top is represented by a white to red gradient ranging from 3.494 to -2.237 for the log2 of fold-change value (AvsC: adapted versus control; NAvsC: non-adapted versus control; NAvsA: non-adapted versus adapted). The different groups represented on the right correspond to: (Group 1) genes conserved between the four models; (Group 2) genes conserved between the chicken/nutritional, pig/heat and inflammation and mouse/exercise stress models; (Group 3) genes conserved between the chicken/heat, pig/heat and inflammation and mouse/exercise stress models; (Group 4) genes conserved between the chicken/heat, chicken/nutritional and mouse/exercise stress models; (Group 5) genes conserved between the chicken/heat, chicken/nutritional and pig/heat and inflammation models. The heatmap plot was generated using R.

Figure 3. Functional analysis of the stress-related conserved genes. (A) Differential expression profile for each of the 26 conserved genes in each comparison for the four models. Each vertical line represents the variation of gene expression, expressed as log2 of fold-change, between two conditions: adapted versus control (AvsC), non-adapted versus control (NAvsC) and non-adapted versus adapted (NAvsA). The black solid line indicates the null log2 fold-change level corresponding

to no differential expression. It was observed that most of the genes displayed a convergent expression profile across comparisons and models. **(B)** The list of human orthologs for the 26 stress-related encoded proteins were submitted to the STRING server (version 11.0) that provide network analyzes. The obtained network is composed of 26 nodes and 25 edges connecting 15 proteins, with an average node degree of 1.92, and a PPI enrichment p-value of 8.18e-13.

Figure 4. Data-mining the GEO database for occurrences of the 26 stress-related genes. (A) This graph displays the number of studies identifying a set of the stress-related genes; **(B)** Number of studies associated to each gene, reported in the GEO database, is shown.

Figure 5. Comparison of the differential expression of the 26 stress-related genes in muscle or liver tissues of stressed versus control animals. Measures realized on liver or muscle tissues are displayed in blue and red respectively. Graphs **(A, C and E)** correspond to tissues taken from the chicken/nutritional stress model and **(B, D and F)** correspond to tissues obtained from the mouse/exercise stress model. Differential expression compared non-adapted versus control **(A and B)**; adapted versus control **(C and D)**; or non-adapted versus adapted **(E and F)**. Change in expression is expressed as log2 of fold-change. Genes that were statistically significantly differentially expressed presented values over 0.66 or lower than -0.66.

Figure 6. Individual network for the four stress models. The list of all differentially expressed genes in each stress model, and the set of 26 conserved genes, were submitted to the STRING server (version 11.0) for analysis and obtained networks were edited using Cytoscape. Principal networks corresponding to the mouse/exercise **(A)** chicken/heat **(B)** chicken/nutritional **(C)** and pig/heat and inflammation **(D)** stress models are depicted. Networks parameters are detailed in Table 5. Nodes assembled into clusters are represented by green dots and central nodes with highest number of partners are represented by red dots. The two remarkable proteins CD44 and SRC identified in the chicken or mammals models respectively, are highlighted by dark circles: CD44 in panels B and C; SRC in panels A and D.

Figure 7. Network analysis of the list of genes co-expressed with stress-related genes. The list of human orthologs for the 93 encoded proteins co-expressed with the core of stress-related genes were submitted to the STRING server (version 11.0) that provide network analyzes. The obtained network is composed of 93 nodes and 217 edges connecting 71 proteins, with an average node degree of 4.67, and a PPI enrichment p-value of 1.0e-16. Nodes depicted in green correspond to the set of conserved stress-related genes identified in this study. Nodes with red circles display proteins annotated by biological process term “extracellular matrix organization”.

Figure 8. Prediction of transcription factors controlling the set of 93 co-expressed genes. Volcano plot representing TRACE results for the analysis of co-expressed genes in response to stress. Each dot corresponds to a transcription factor plotted according to its log2 of activity and its -log10 of p-value. The size of each dot corresponds to the transcription factor influence parameter on the set of co-expressed genes.

Figure 9. Model of transcriptional regulation of genes during stress response through the TGFß-, SRC- and CD44-mediated signaling pathways. TGFß is a cytokine secreted in the extracellular matrix (ECM) in an inactivated form called “latent complex”. **(A)** Latent TGFß is processed by THBS1 and other extracellular proteases from the matrix metalloproteinase (MMP) family, leading to its activation. **(B)** Binding of the active TGFß to one of the TGFß receptors (TGFßRs) promotes activation and phosphorylation of the receptor. Alternatively, TGFßRs can be activated by interaction with CD44 bound to hyaluronic acid (HA). Both HA and TGFß either alone or in combination can induce TGFßRs-mediated signaling. **(C)** The canonical signaling pathway promotes regulation of gene expression in a SMAD-dependent manner. Activation of TGFßRs induces phosphorylation of the SMAD2/3 complex, favoring its interaction with SMAD4. Once the SMAD2/3/4 complex is formed, it translocates to the nucleus to regulate gene expression. Our study proposes that this signaling pathway controls the expression of a stress-response program. **(D)** The non-canonical signaling pathway relies on phosphorylation cascades of different protein factors. This signal propagates through different parallel pathways, such as TNF receptor-associated factor 4/6 (TRAF4/6), Rho family of small GTPases (RhoA, Ras), phosphatidylinositide 3-kinase (PI3K) or the mitogen-activated protein kinases family (MAPKs). These phosphorylation cascades finally activate several transcription factors, not represented here, that control different gene expression programs, including regulation of stress-related genes. **(E)** In addition to the canonical and non-canonical pathways induced by TGFßRs, it has been proposed that binding of TGFß to its receptor also induces NOX-mediated production of reactive oxygen species (ROS), notably hydrogen peroxide (H_2O_2). This reactive molecule activates SRC through oxidation of a redox-sensitive cysteine residue. In its activated form, SRC mediates a cross-talk regulation of several signaling pathways together with TGFßRs, including the PI3K-AKT pathway. Arrows symbolize activation or signal transmission in each pathway. Phosphorylation is represented by a green symbol.

Figures

Figure 1

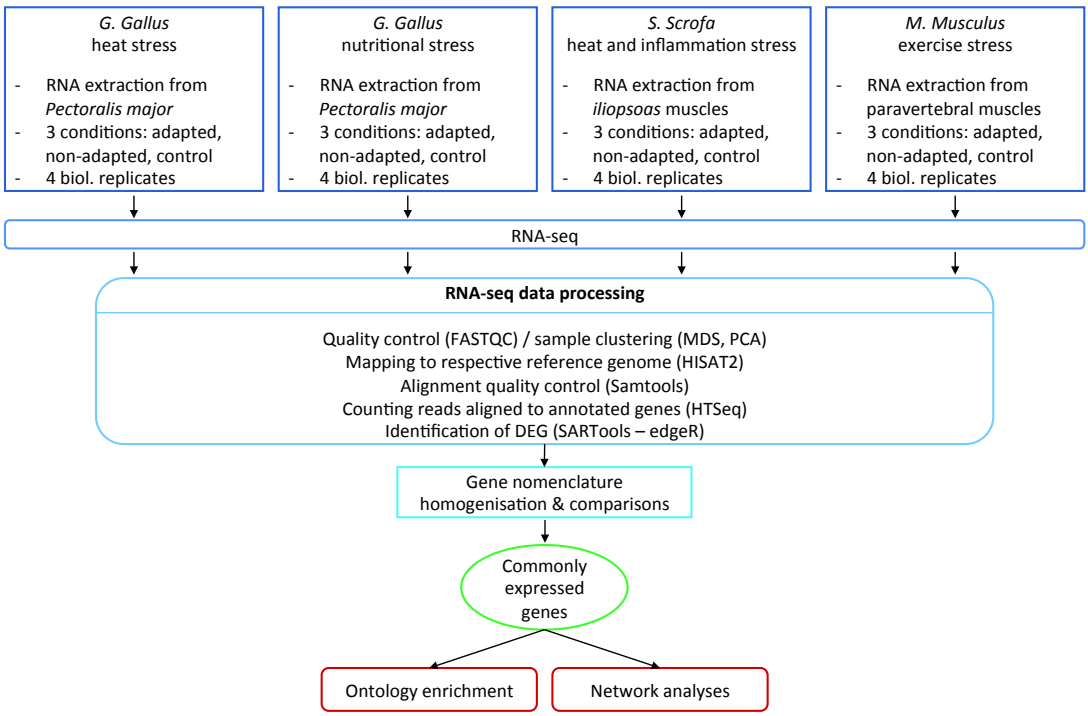


Figure 2

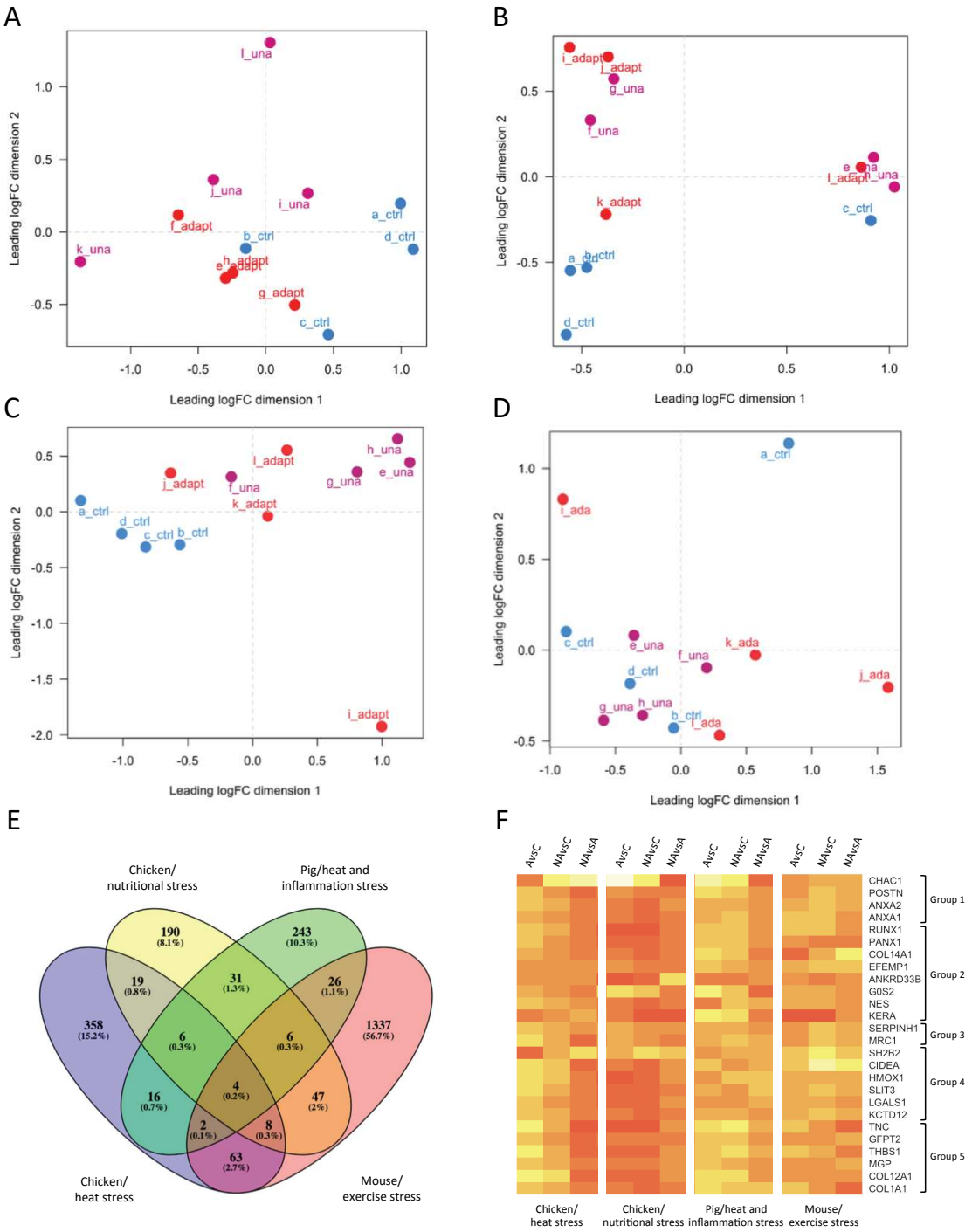
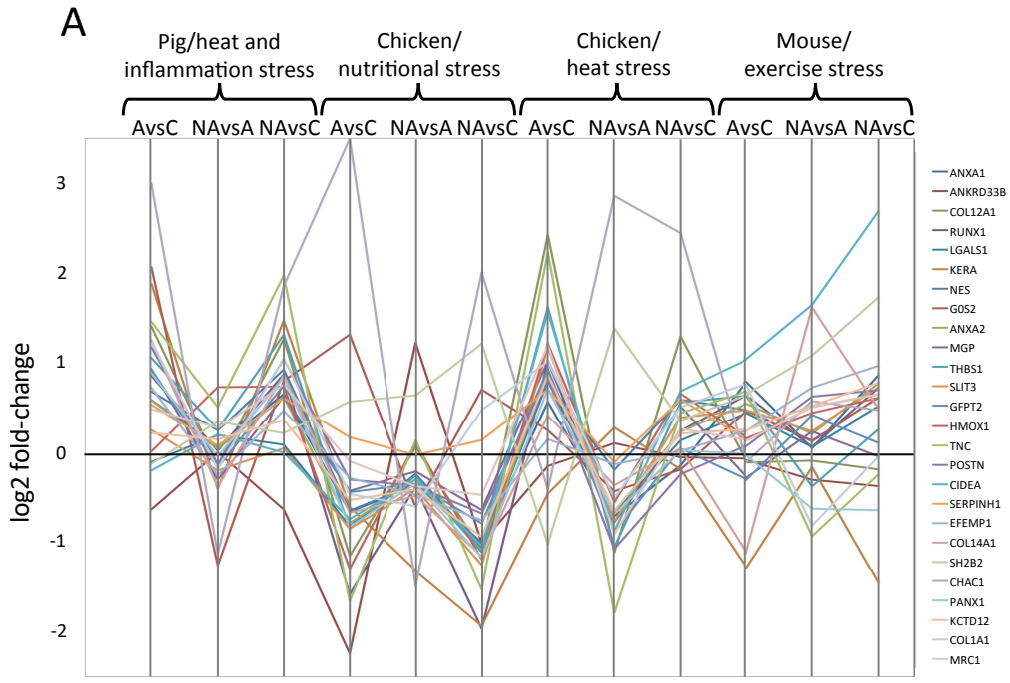

A
l_una
j_una
i_una
f_adapt
a_ctrl
b_ctrl
d_ctrl
k_una
h_adapt
e_adapt
g_adapt
c_ctrl
Leading logFC dimension 2
Leading logFC dimension 1
B
l_adapt
j_adapt
g_una
f_una
e_una
h_una
k_adapt
c_ctrl
a_ctrl
b_ctrl
d_ctrl
C
h_una
l_adapt
e_una
j_adapt
f_una
g_una
a_ctrl
k_adapt
d_ctrl
c_ctrl
b_ctrl
i_adapt
D
a_ctrl
i_ada
c_ctrl
e_una
k_ada
f_una
d_ctrl
j_ada
g_una
h_una
b_ctrl
l_ada
E
Chicken/
nutritional stress
Pig/heat and
inflammation stress
190
(8.1%)
243
(10.3%)
19
(0.8%)
31
(1.3%)
26
(1.1%)
358
(15.2%)
6
(0.3%)
6
(0.3%)
1337
(56.7%)
16
(0.7%)
4
(0.2%)
47
(2%)
2
(0.1%)
8
(0.3%)
63
(2.7%)
Chicken/
heat stress
Mouse/
exercise stress
F
AvsC
NAvsC
NAvsA
CHAC1
POSTN
ANXA2
ANXA1
RUNX1
PANX1
COL14A1
EFEMP1
ANKRD33B
G0S2
NES
KERA
SERPINH1
MRC1
SH2B2
CIDEA
HMOX1
SLIT3
LGALS1
KCTD12
TNC
GFPT2
THBS1
MGP
COL12A1
COL1A1
Group 1
Group 2
Group 3
Group 4
Group 5
Chicken/
heat stress
Chicken/
nutritional stress
Pig/heat and
inflammation stress
Mouse/
exercise stress

Figure 3


A
Pig/heat and inflammation stress
Chicken/ nutritional stress
Chicken/ heat stress
Mouse/ exercise stress
AvsC NAvsA NAvsC AvsC NAvsA NAvsC AvsC NAvsA NAvsC AvsC NAvsA NAvsC
log2 fold-change
3
2
1
0
-1
-2
ANXA1
ANKRD33B
COL12A1
RUNX1
LGALS1
KERA
NES
G0S2
ANXA2
MGP
THBS1
SLIT3
GFPT2
HMOX1
TNC
POSTN
CIDEA
SERPINH1
EFEMP1
COL14A1
SH2B2
CHAC1
PANX1
KCTD12
COL1A1
MRC1

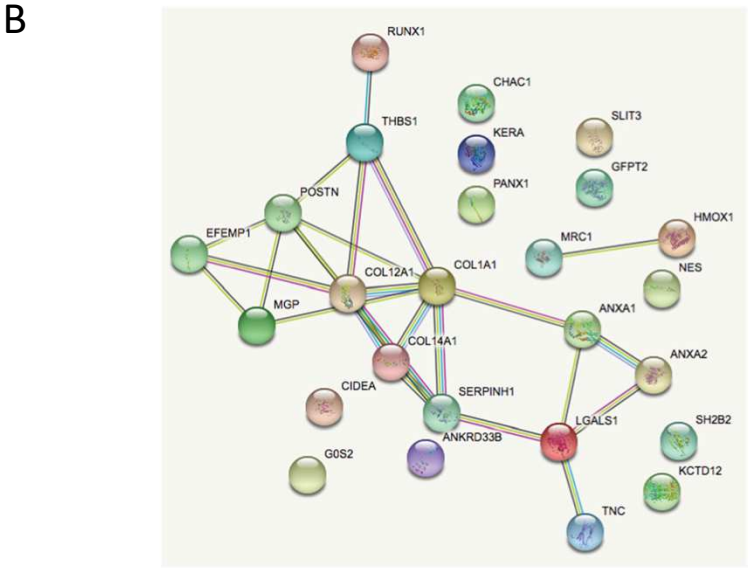
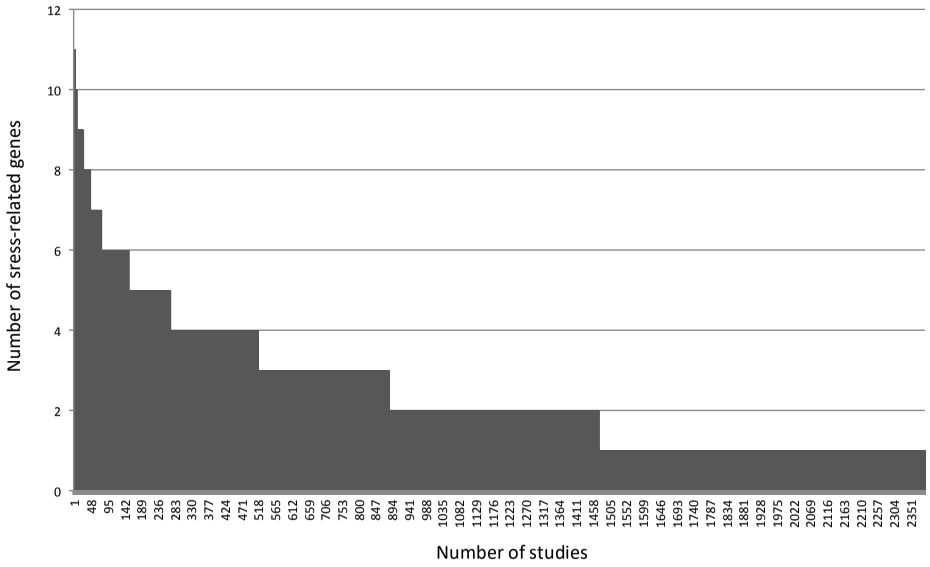

B
RUNX1
CHAC1
THBS1
SLIT3
KERA
GFPT2
PANX1
POSTN
HMOX1
EFEMP1
MRC1
COL12A1
COL1A1
NES
MGP
ANXA1
COL14A1
ANXA2
CIDEA
SERPINH1
LGALS1
SH2B2
ANKRD33B
G0S2
KCTD12
TNC

Figure 4

A


Number of sress-related genes
0
2
4
6
8
10
12
1
48
95
142
189
236
283
330
377
424
471
518
565
612
659
706
753
800
847
894
941
988
1035
1082
1129
1176
1223
1270
1317
1364
1411
1458
1505
1552
1599
1646
1693
1740
1787
1834
1881
1928
1975
2022
2069
2116
2163
2210
2257
2304
2351
Number of studies

B

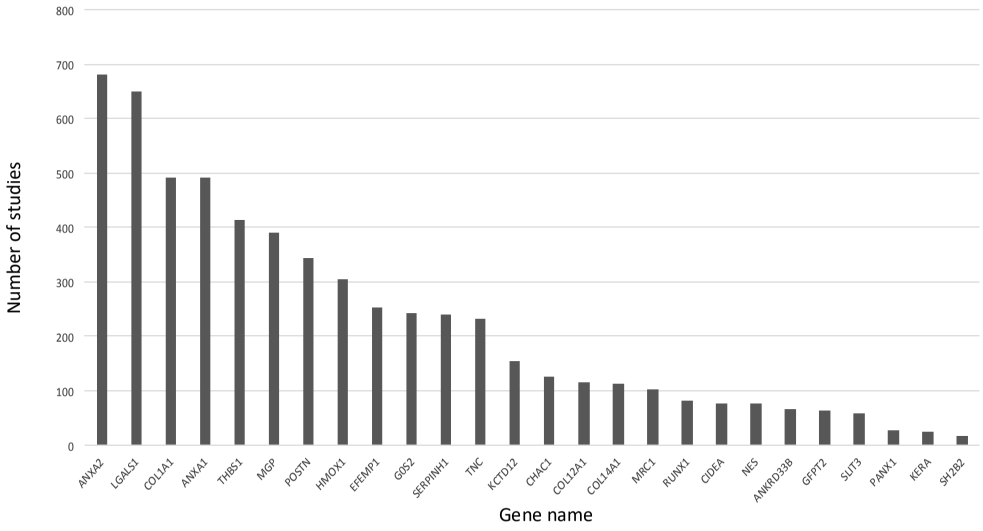

Number of studies
0
100
200
300
400
500
600
700
800
ANXA2
LGALS1
COL1A1
ANXA1
THBS1
MGP
POSTN
HMOX1
EFEMP1
GDS2
SERPINH1
TNC
KCTD12
CHAC1
COL12A1
COL14A1
MRC1
RUNX1
CIDEA
NES
ANKRD33B
GFPT2
SLIT3
PANX1
KERA
SH2B2
Gene name

Figure 5

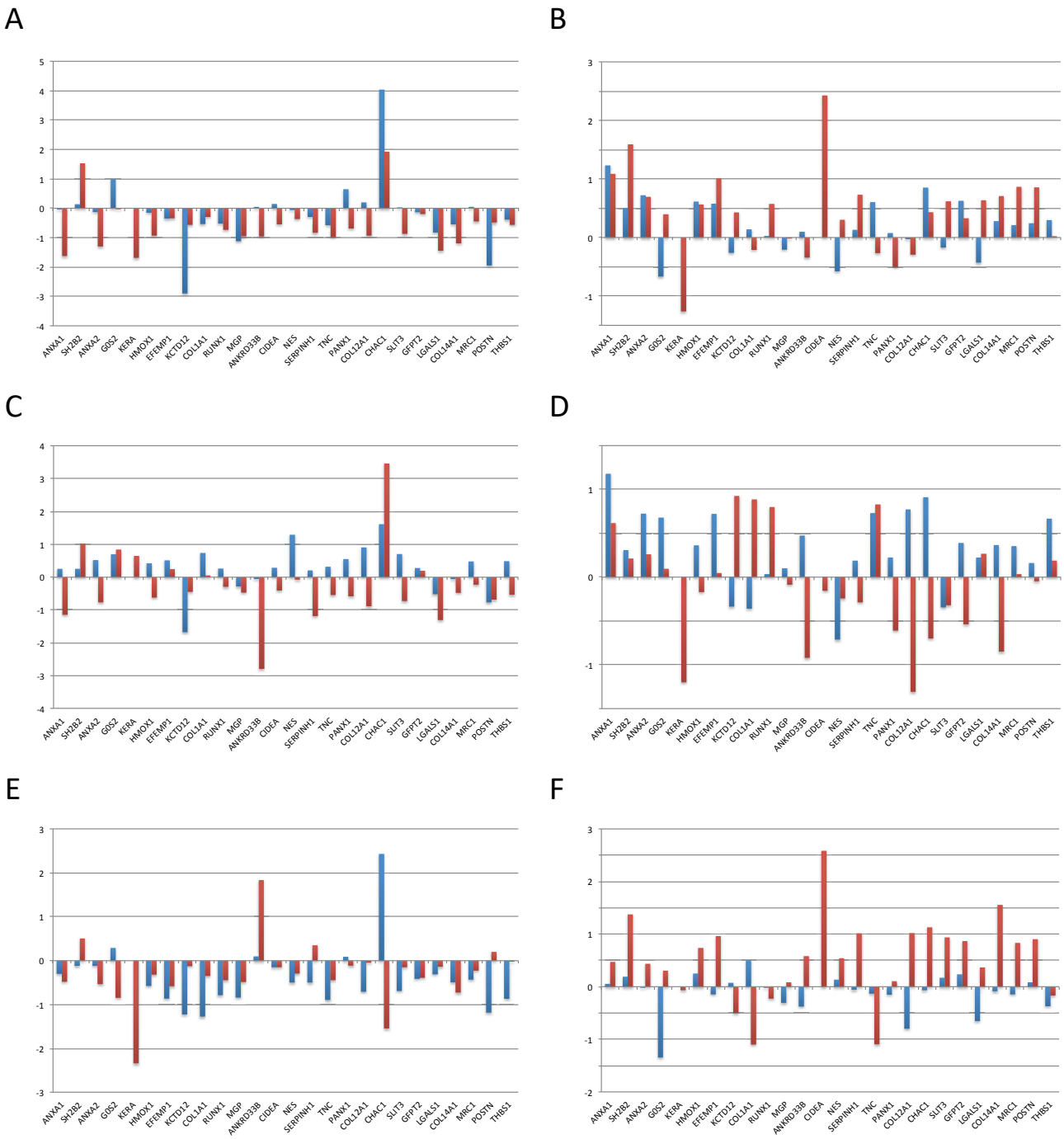

A
B
C
D
E
F
ANXA1
SH2B2
ANXA2
G0S2
KERA
HMOX1
EFEMP1
KCTD12
COL1A1
RUNX1
MGP
ANKRD33B
CIDEA
NES
SERPINH1
TNC
PANX1
COL12A1
CHAC1
SLIT3
GFPT2
LGALS1
COL14A1
MRC1
POSTN
THBS1
Liver
Muscle

Figure 6

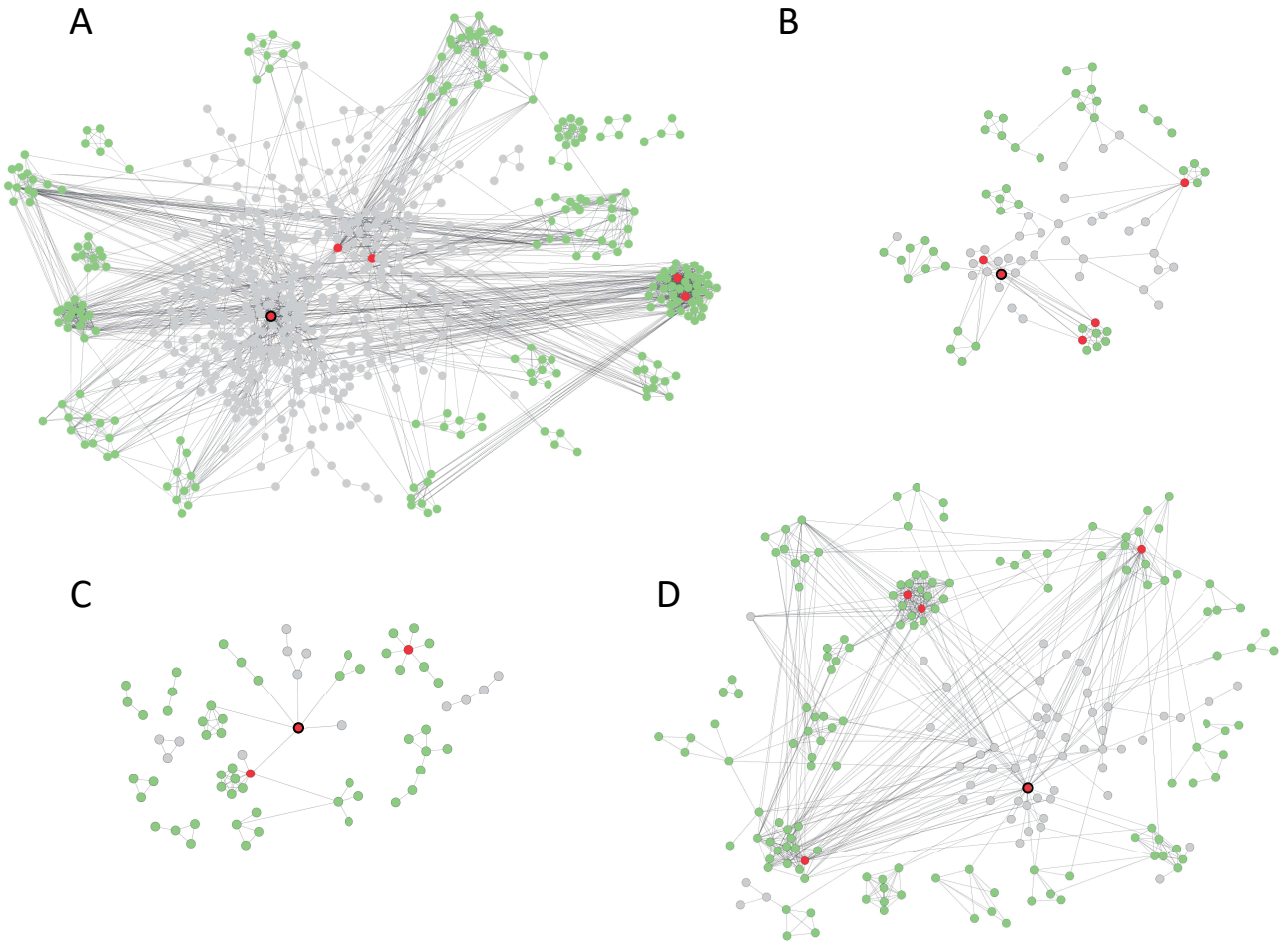


Figure 7

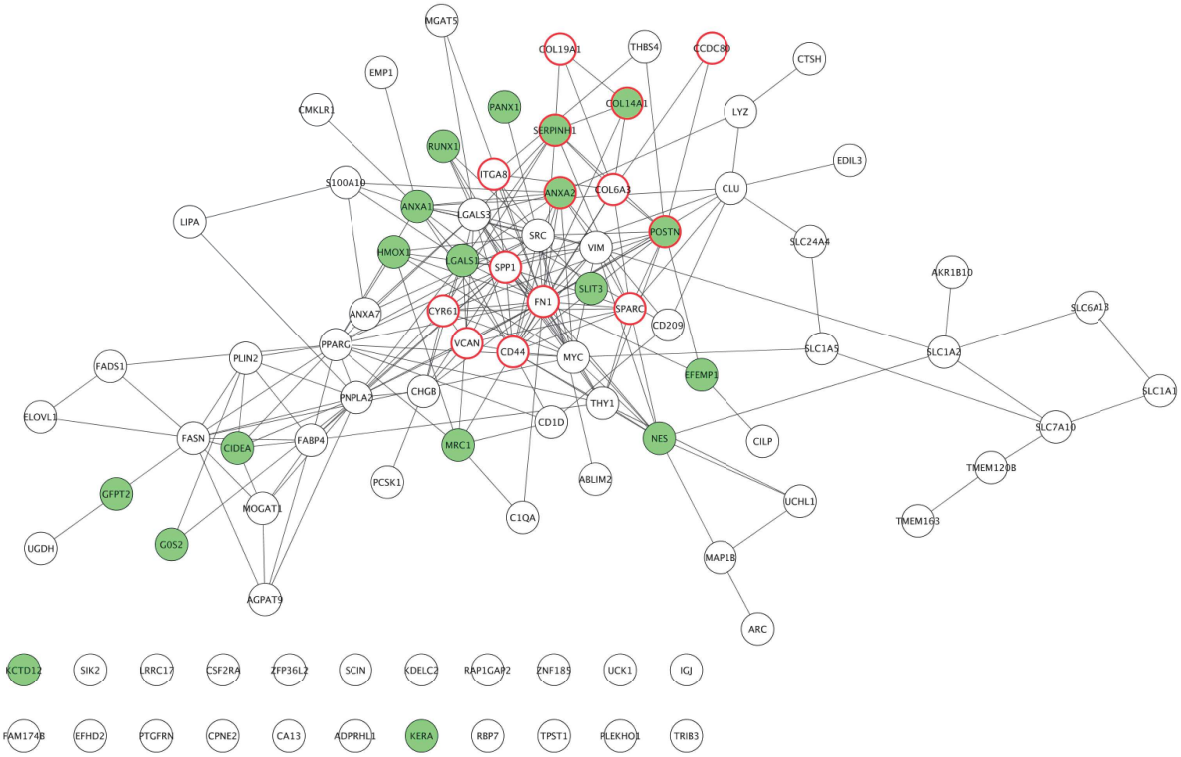

MGAT5
COL19A1
THBS4
CCDC80
CTSH
EMP1
CMKLR1
PANX1
COL14A1
LYZ
SERPINH1
RUNX1
EDIL3
ITGA8
S100A10
ANXA2
COL6A3
CLU
ANXA1
LGALS3
LIPA
SRC
POSTN
SLC24A4
VIM
HMOX1
LGALS1
SPP1
AKR1B10
SLIT3
FN1
SLC6A13
ANXA7
CYR61
SPARC
CD209
PPARG
VCAN
CD44
MYC
SLC1A5
SLC1A2
PLIN2
FADS1
EFEMP1
SLC1A1
CHGB
PNPLA2
THY1
ELOVL1
CD1D
SLC7A10
FASN
CIDEA
FABP4
MRC1
NES
CILP
TMEM120B
GFPT2
PCSK1
ABLIM2
UCHL1
MOGAT1
C1QA
TMEM163
UGDH
G0S2
MAP1B
AGPAT9
ARC
KCTD12
SIK2
LRRC17
CSF2RA
ZFP36L2
SCIN
KDELC2
RAP1GAP2
ZNF185
UCK1
IGJ
FAM174B
EFHD2
PTGFRN
CPNE2
CA13
ADPRHL1
KERA
RBP7
TPST1
PLEKHO1
TRIB3

Figure 8

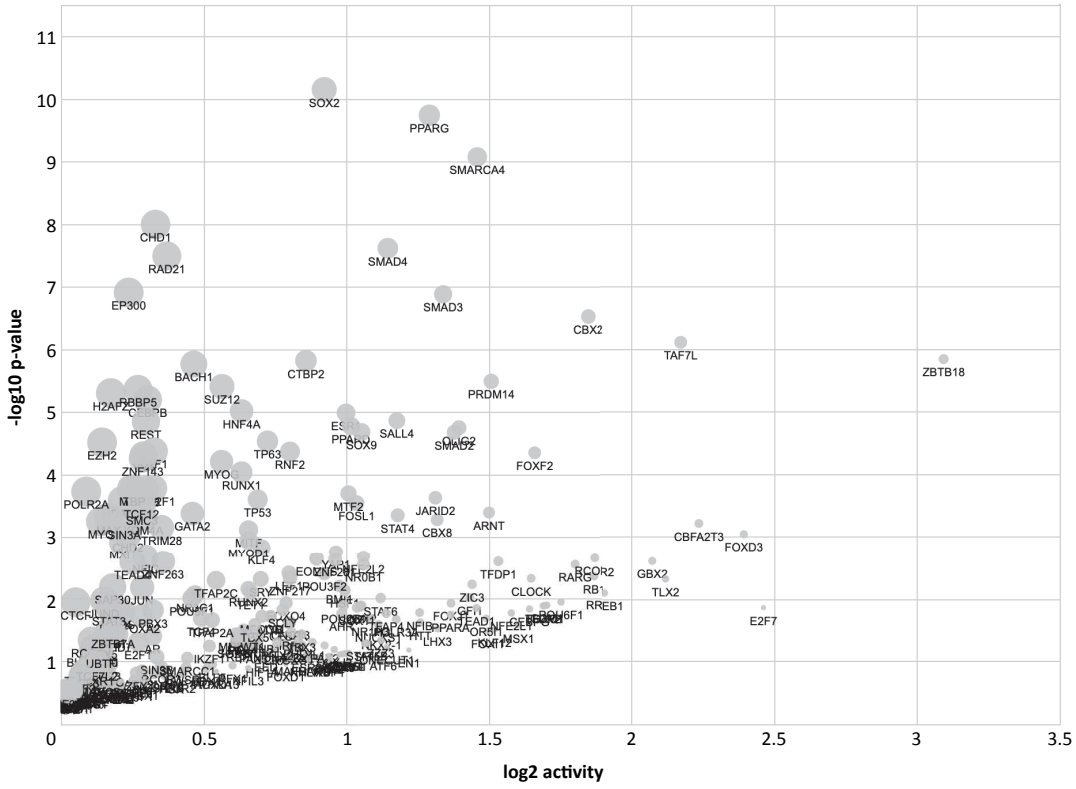
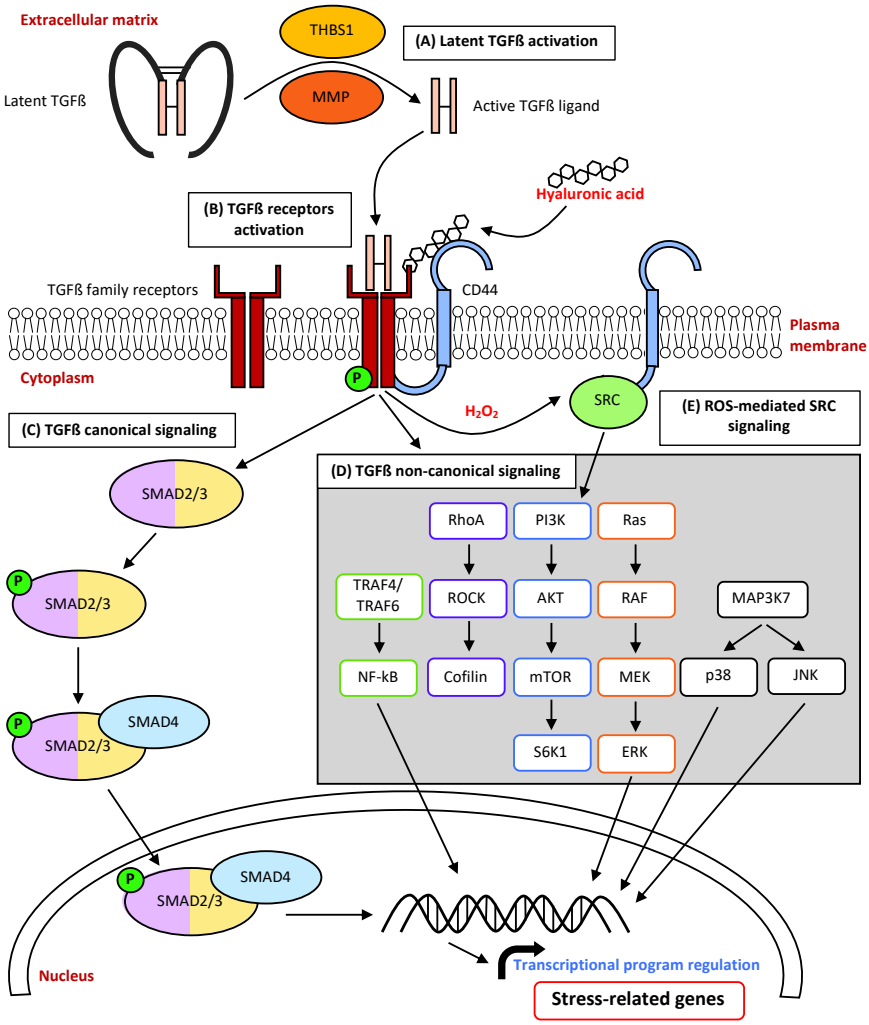

-log10 p-value
log2 activity
SOX2
PPARG
SMARCA4
CHD1
RAD21
SMAD4
EP300
SMAD3
CBX2
TAF7L
ZBTB18
BACH1
CTBP2
PRDM14
SUZ12
HNF4A
REST
SALL4
SOX9
EZH2
TP63
RNF2
FOXF2
RUNX1
POLR2A
TP53
MTF2
FOSL1
JARID2
GATA2
STAT4
CBX8
ARNT
CBFA2T3
FOXD3
KLF4
TFDP1
RCOR2
RARG
GBX2
CLOCK
RB1
TLX2
TFAP2C
ZIC3
RREB1
E2F7
LHX3
MSX1

Figure 9


Extracellular matrix
THBS1
(A) Latent TGFß activation
Latent TGFß
MMP
Active TGFß ligand
Hyaluronic acid
(B) TGFß receptors activation
TGFß family receptors
CD44
Plasma membrane
Cytoplasm
P
H_2O_2
SRC
(E) ROS-mediated SRC signaling
(C) TGFß canonical signaling
(D) TGFß non-canonical signaling
SMAD2/3
RhoA
PI3K
Ras
TRAF4/ TRAF6
ROCK
AKT
RAF
MAP3K7
NF-kB
Cofilin
mTOR
MEK
p38
JNK
SMAD4
S6K1
ERK
Nucleus
Transcriptional program regulation
Stress-related genes

Supplementary Figures

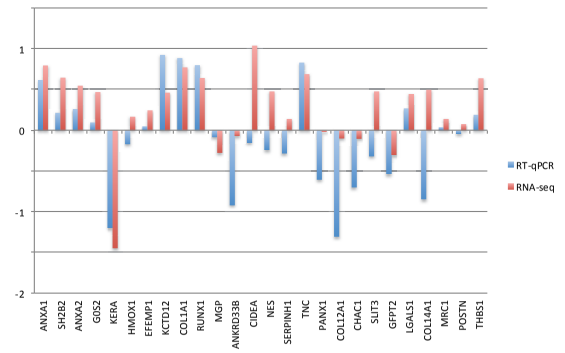
Supp Figure 1. Validation of transcriptomic data by quantitative PCR (RT-qPCR). RNA was extracted from muscles of adapted, non-adapted and control animals from the mouse/exercise **(A and B)** and chicken/nutritional **(C and D)** stress models, and expression level for each of the 26 stress-related genes was determined by classical RT-qPCR or Fluidigm. Comparison of the differential expression levels between adapted versus control **(A and C)** and non-adapted versus control **(B and D)** animals determined by RT-qPCR, Fluidigm and RNA-seq are displayed as log2 of fold-change. In most cases, differential expression levels are equivalent between the different methods (significative threshold values 0.66 or -0.66). The divergences obtained in **(A)** are mainly observed for genes not statistically significantly differentially expressed (log2 of fold-change ranging from -0.66 to 0.66).

Supp Figure 2. Differential expression of the stress-related genes between chicken exposed to the xenobiotic reagent paraquat and controls. RNA extracted from muscles of chicken submitted to paraquat exposure or controls and expression level for each of 26 stress-related genes was determined by RT-qPCR. Differential expression levels between paraquat-treated and control animals are displayed as log2 of fold-change. Most genes were statistically significantly differentially expressed (threshold values 0.66 or -0.66).

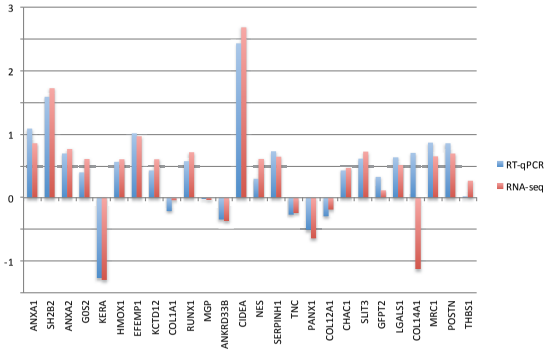
Supplementary Figures

Supp Figure 1

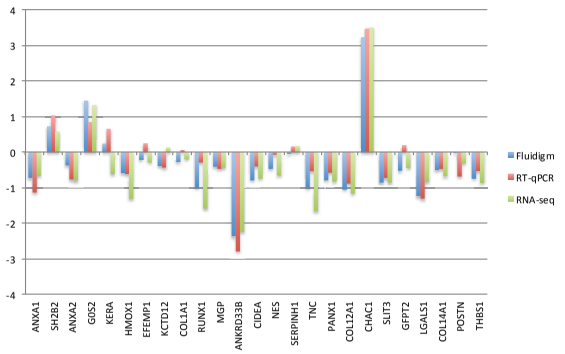
A



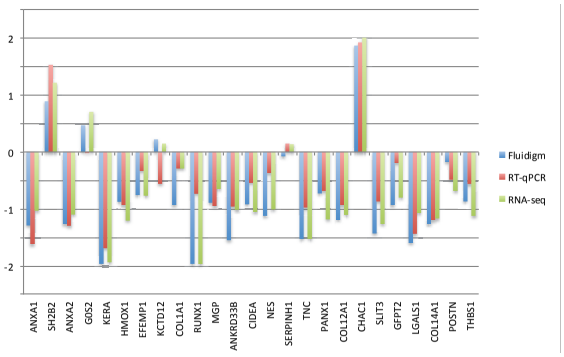
B



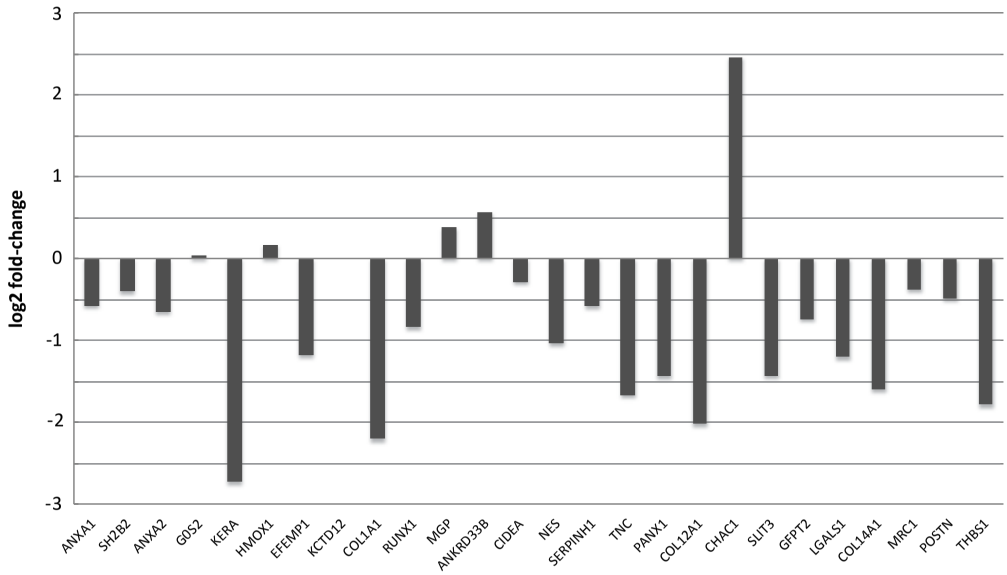
C



D



Supp Figure 2



Tables

Models / Comparison	Pig/heat and inflammation stress	Mouse/exercise stress	Chicken/heat stress	Chicken/nutritional stress
AvsC	282	356	278	65
NAvsC	339	1394	39	275
NAvsA	35	220	133	15

Table 1. Number of DE genes per comparison for each model, with a padj threshold of 0.05. AvsC: adapted versus control; NAvsC: non-adapted versus control; NAvsA: non-adapted versus adapted.

Gene	Identified function	Cell localization
Ankrd33b	Unknown	Unknown
Annexin A1 – Anxa1	Cell membrane reparation and inflammation	Memb, secreted, nucleus, cyto
Annexin A2 – Anxa2	Cell membrane reparation	Secreted
Chac1	Glutathion degradation	Cyto
Cidea	Apoptosis, energy metabolism	Nucleus, lipid droplets
Collagen 1 – Col1a1	ECM component	ECM
Collagen 12 – Col12a1	ECM component, fibril associated collagen	ECM
Collagen 14 – Col14a1	ECM component, fibril associated collagen	ECM
Fibulin 3 – Efemp1	Cell adhesion and differentiation	ECM
G0s2	Lipolysis and apoptosis control	Mitochondria
Gfpt2	glutamine-fructose-6-phosphate transaminase 2	Cyto
Hmox1	Heme oxygenase, forms biliverdin	ER
Kctd12	Auxiliary subunit GABA-B receptors	Memb
Kera	Keratan sulfate proteoglycane	ECM
Lgals1	Lectin binding galactoside, role in apotosis, adhesion and cell differentiation	ECM
Mgp	Calcium mineralization control	ECM
Mrc1	Macrophage mannose receptor	Memb
Nes	Vimentin intermediate filaments assembly dynamics	ECM, cytoskeleton, cyto
Pannexin 1 – Panx1	Structural component of gap junctions	Memb, ER
Periostin - Postn	Cell adhesion	Secreted, ECM
Runx1	Transcription factor important for muscle regeneration	Nucleus
Serpinh1	Collagen chaperone	ER
Sh2b2	Adapter protein for tyrosine kinase receptors, insulin response	Memb, Cyto
Slit3	Cellular migration	ECM
Thbs1	Adhesive glycoprotein, heparin and collagen binding	ECM, ER
Tenascin - Tnc	Cell adhesion and growth	ECM

Table 2. List of differentially expressed genes conserved between at least three different models. Identified function and cell localization were determined based on Uniprot database annotations and bibliographic analyzes. ECM=Extra Cellular Matrix; ER=Endoplasmic reticulum; Cyto=cytoplasm; Memb=membrane

Biological Process (GO)			
GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0030199	collagen fibril organization	5 of 39	4.40e-06
GO:0030198	extracellular matrix organization	8 of 296	4.40e-06
GO:0007162	negative regulation of cell adhesion	7 of 245	1.10e-05
GO:0010033	response to organic substance	16 of 2815	1.33e-05
GO:1901700	response to oxygen-containing compound	11 of 1427	0.00023
GO:0048731	system development	17 of 4144	0.00023
GO:0031670	cellular response to nutrient	4 of 59	0.00023
GO:0014070	response to organic cyclic compound	9 of 873	0.00023
GO:0010812	negative regulation of cell-substrate adhesion	4 of 55	0.00023
GO:0001501	skeletal system development	7 of 457	0.00025
GO:0071295	cellular response to vitamin	3 of 23	0.00068
GO:0009611	response to wounding	7 of 547	0.00068
GO:0007584	response to nutrient	5 of 208	0.00087
GO:0071310	cellular response to organic substance	12 of 2219	0.00098
GO:0009725	response to hormone	8 of 854	0.00098
GO:0001819	positive regulation of cytokine production	6 of 390	0.00098
GO:0001817	regulation of cytokine production	7 of 615	0.00098
GO:0042493	response to drug	8 of 900	0.0011
GO:0090049	regulation of cell migration involved in sprouting angiogene...	3 of 37	0.0013
GO:0071345	cellular response to cytokine stimulus	8 of 953	0.0014
GO:0009653	anatomical structure morphogenesis	11 of 1992	0.0014
GO:0048513	animal organ development	13 of 2926	0.0016
GO:0042060	wound healing	6 of 461	0.0016
GO:0032501	multicellular organismal process	19 of 6507	0.0021
GO:0051216	cartilage development	4 of 147	0.0022
GO:0031099	regeneration	4 of 151	0.0024
GO:0051241	negative regulation of multicellular organismal process	8 of 1098	0.0028
GO:0045766	positive regulation of angiogenesis	4 of 162	0.0028
GO:0043536	positive regulation of blood vessel endothelial cell migration	3 of 55	0.0028
GO:0050707	regulation of cytokine secretion	4 of 174	0.0031
GO:0033993	response to lipid	7 of 825	0.0031
GO:0032964	collagen biosynthetic process	2 of 8	0.0031
GO:0031340	positive regulation of vesicle fusion	2 of 8	0.0031
GO:0051239	regulation of multicellular organismal process	12 of 2788	0.0034
GO:0023051	regulation of signaling	13 of 3360	0.0042
GO:0051093	negative regulation of developmental process	7 of 910	0.0045
GO:1903587	regulation of blood vessel endothelial cell proliferation invol...	2 of 13	0.0052
GO:0009612	response to mechanical stimulus	4 of 210	0.0052
GO:0030335	positive regulation of cell migration	5 of 452	0.0080
GO:0001818	negative regulation of cytokine production	4 of 245	0.0080
GO:0002694	regulation of leukocyte activation	5 of 470	0.0088
GO:0043371	negative regulation of CD4-positive, alpha-beta T cell differe...	2 of 20	0.0094
GO:0090050	positive regulation of cell migration involved in sprouting a...	2 of 21	0.0098

Table 3. Ontological enrichment in the "biological process" category according to the STRING website for the 26 stress-related conserved genes. These data indicate that a majority of the 26 genes are involved in extracellular functions such as extracellular matrix organization and/or response to cell signaling. Only most significant results (false discovery rate ≤ 0.01) are displayed here.

Biological Process (GO)			
GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0030198	extracellular matrix organization	14 of 296	5.80e-07
GO:0050793	regulation of developmental process	32 of 2416	3.20e-05
GO:0045785	positive regulation of cell adhesion	13 of 375	3.20e-05
GO:0022603	regulation of anatomical structure morphogenesis	19 of 961	7.14e-05
GO:0022604	regulation of cell morphogenesis	13 of 442	0.00010
GO:0030155	regulation of cell adhesion	15 of 623	0.00012
GO:0010810	regulation of cell-substrate adhesion	9 of 189	0.00012
GO:0010033	response to organic substance	33 of 2815	0.00012
GO:0051128	regulation of cellular component organization	29 of 2306	0.00016
GO:0042221	response to chemical	41 of 4153	0.00018
GO:0051093	negative regulation of developmental process	17 of 910	0.00032
GO:0030335	positive regulation of cell migration	12 of 452	0.00040
GO:0019216	regulation of lipid metabolic process	11 of 373	0.00040
GO:0001501	skeletal system development	12 of 457	0.00040
GO:0032502	developmental process	47 of 5401	0.00042
GO:0007155	cell adhesion	16 of 843	0.00042
GO:0048856	anatomical structure development	45 of 5085	0.00043
GO:0045595	regulation of cell differentiation	23 of 1695	0.00047
GO:0051241	negative regulation of multicellular organismal process	18 of 1098	0.00048
GO:0046942	carboxylic acid transport	9 of 252	0.00048
GO:0045596	negative regulation of cell differentiation	14 of 683	0.00051
GO:0048731	system development	39 of 4144	0.00055
GO:1903039	positive regulation of leukocyte cell-cell adhesion	8 of 202	0.00065
GO:0034097	response to cytokine	17 of 1035	0.00071
GO:1903037	regulation of leukocyte cell-cell adhesion	9 of 278	0.00072
GO:0051239	regulation of multicellular organismal process	30 of 2788	0.00072
GO:0048513	animal organ development	31 of 2926	0.00072
GO:0042940	D-amino acid transport	3 of 6	0.00072
GO:0007275	multicellular organism development	42 of 4726	0.00072
GO:0010811	positive regulation of cell-substrate adhesion	6 of 109	0.0013
GO:2000026	regulation of multicellular organismal development	23 of 1876	0.0014
GO:0006952	defense response	18 of 1234	0.0014
GO:0009888	tissue development	21 of 1626	0.0015
GO:0042493	response to drug	15 of 900	0.0016
GO:0006950	response to stress	32 of 3267	0.0018
GO:0002682	regulation of immune system process	19 of 1391	0.0018
GO:0071310	cellular response to organic substance	25 of 2219	0.0019
GO:0015711	organic anion transport	10 of 414	0.0019
GO:0098609	cell-cell adhesion	10 of 416	0.0020
GO:0089718	amino acid import across plasma membrane	3 of 11	0.0021
GO:0051451	myoblast migration	3 of 11	0.0021
GO:0071345	cellular response to cytokine stimulus	15 of 953	0.0025
GO:0050870	positive regulation of T cell activation	7 of 193	0.0025
GO:0010889	regulation of sequestering of triglyceride	3 of 12	0.0025
GO:0003333	amino acid transmembrane transport	5 of 80	0.0027
GO:0048583	regulation of response to stimulus	35 of 3882	0.0029
GO:0030334	regulation of cell migration	13 of 753	0.0031
GO:0050896	response to stimulus	56 of 7824	0.0032
GO:0009611	response to wounding	11 of 547	0.0032
GO:0002684	positive regulation of immune system process	14 of 882	0.0036
GO:0070887	cellular response to chemical stimulus	27 of 2672	0.0041
GO:0045087	innate immune response	12 of 676	0.0041
GO:0002694	regulation of leukocyte activation	10 of 470	0.0041
GO:0002376	immune system process	25 of 2370	0.0041
GO:0050863	regulation of T cell activation	8 of 302	0.0047
GO:0009966	regulation of signal transduction	29 of 3033	0.0050
GO:0023051	regulation of signaling	31 of 3360	0.0051
GO:0048518	positive regulation of biological process	43 of 5459	0.0052
GO:0016477	cell migration	13 of 812	0.0052
GO:0051249	regulation of lymphocyte activation	9 of 401	0.0054
GO:0034389	lipid droplet organization	3 of 19	0.0054
GO:0009653	anatomical structure morphogenesis	22 of 1992	0.0054
GO:0002521	leukocyte differentiation	8 of 313	0.0054
GO:0015909	long-chain fatty acid transport	4 of 54	0.0059
GO:0016043	cellular component organization	41 of 5163	0.0063
GO:1902475	L-alpha-amino acid transmembrane transport	4 of 57	0.0068
GO:0052547	regulation of peptidase activity	9 of 420	0.0068
GO:0000902	cell morphogenesis	11 of 626	0.0070
GO:0050707	regulation of cytokine secretion	6 of 174	0.0071
GO:0002685	regulation of leukocyte migration	6 of 175	0.0071
GO:0030162	regulation of proteolysis	12 of 742	0.0072
GO:0032879	regulation of localization	25 of 2524	0.0077
GO:0010646	regulation of cell communication	30 of 3327	0.0078
GO:0014070	response to organic cyclic compound	13 of 873	0.0082
GO:0045862	positive regulation of proteolysis	8 of 347	0.0084
GO:0010769	regulation of cell morphogenesis involved in differentiation	7 of 263	0.0087
GO:0048869	cellular developmental process	31 of 3533	0.0089
GO:0070779	D-aspartate import across plasma membrane	2 of 4	0.0093
GO:0051050	positive regulation of transport	13 of 892	0.0093
GO:0040011	locomotion	15 of 1144	0.0093
GO:0051240	positive regulation of multicellular organismal process	18 of 1551	0.0097

Table 4. Ontological enrichment in the "biological process" category according to the STRING website for the 93 genes co-expressed with the core set of stress-related conserved genes. These data indicate that a majority of the 93 genes are involved in extracellular functions such as extracellular matrix organization and/or response to cell signaling. Only most significant results (false discovery rate ≤ 0.01) are displayed here.

Stress model	Number of proteins in the network	Number of edges in the network	Number of proteins in principal networks	Number of clusters	Number of proteins in clusters
Mouse/exercise	1259	2308	699	18	235
Pig/heat and inflammation	440	657	197	16	141
Chicken/heat	295	152	82	8	46
Chicken/nutritional	265	95	68	12	54

Table 5. Individual STRING network analyzes for the four stress models. Network connectivity parameters are indicated.

Supplementary Tables

Stress models	Average total number of reads	Aligned sequences (%)	Average aligned base number	Mis-aligned bases (%)	FastQC quality score
Chicken/heat stress	40 545 275	87.87	1 770 663 933	0.26	37.95
Chicken/nutritional stress	97 785 672	87.86	4 277 466 249	0.23	39.41
Pig/heat and inflammation stress	99 522 896	83.09	4 112 632 668	0.15	39.76
Mouse/exercise stress	92 750 256	96.02	4 434 834 419	0.08	39.3

Supp Table 1. Number of reads aligned to the reference genomes and quality controls. Numbers provided for each model correspond to the average value of the 12 sequenced samples distributed into three conditions.

GEO studies title	Number of stress-related genes	Genes involved in the stress-related studies
Skeletal muscle initial response to concentric resistance exercise training: time course	11	['ANXA1', 'ANXA2', 'COL14A1', 'COL1A1', 'EFEMP1', 'HMOX1', 'MRC1', 'NES', 'PANX1', 'SERPINH1', 'TNC']
Gliomas of grades III and IV (HG-U133A)	11	['ANXA1', 'ANXA2', 'COL1A1', 'EFEMP1', 'HMOX1', 'LGALS1', 'MGP', 'NES', 'POSTN', 'SERPINH1', 'TNC']
Anaplastic thyroid carcinomas: thyroid biopsies	11	['ANXA2', 'COL1A1', 'LGALS1', 'MGP', 'MRC1', 'POSTN', 'RUNX1', 'SERPINH1', 'SLIT3', 'THBS1', 'TNC']
Postinfarction heart failure model: left ventricle	10	['ANXA1', 'ANXA2', 'COL12A1', 'COL14A1', 'COL1A1', 'HMOX1', 'MGP', 'MRC1', 'POSTN', 'SLIT3']
Tendon development: embryonic limb tendon cells	10	['ANXA2', 'COL12A1', 'COL14A1', 'COL1A1', 'KCTD12', 'KERA', 'LGALS1', 'POSTN', 'SERPINH1', 'THBS1']
Type 1 interferon effect on primary neurons and fibroblasts from E14.5 embryos	10	['ANXA1', 'ANXA2', 'COL12A1', 'COL1A1', 'EFEMP1', 'LGALS1', 'POSTN', 'SERPINH1', 'THBS1', 'TNC']
Ductal carcinoma in situ: mammary gland	10	['ANXA1', 'ANXA2', 'COL1A1', 'EFEMP1', 'G0S2', 'KCTD12', 'LGALS1', 'MGP', 'MRC1', 'POSTN']
Ovarian endometriosis	10	['COL12A1', 'COL14A1', 'COL1A1', 'EFEMP1', 'KCTD12', 'LGALS1', 'MGP', 'SERPINH1', 'THBS1', 'TNC']
Gata4 heterozygous mutant heart response to pressure overload	9	['ANXA2', 'COL12A1', 'COL14A1', 'COL1A1', 'EFEMP1', 'POSTN', 'RUNX1', 'THBS1', 'TNC']
Pandemic and seasonal influenza A H1N1 infection of differentiated type I-like alveolar epithelial cells in vitro	9	['COL12A1', 'COL1A1', 'EFEMP1', 'GFPT2', 'HMOX1', 'LGALS1', 'POSTN', 'SERPINH1', 'TNC']
Articular and growth plate cartilage zones: 10-day old normal proximal tibia	9	['ANXA1', 'COL12A1', 'COL14A1', 'COL1A1', 'EFEMP1', 'KERA', 'LGALS1', 'MRC1', 'POSTN']
Articular cartilage zones: 1-week old normal proximal tibia	9	['ANXA1', 'COL12A1', 'COL14A1', 'COL1A1', 'EFEMP1', 'LGALS1', 'MGP', 'POSTN', 'TNC']
Turner syndrome amniocyte derived-induced pluripotent stem cells	9	['ANXA1', 'COL12A1', 'COL1A1', 'EFEMP1', 'LGALS1', 'POSTN', 'SLIT3', 'THBS1', 'TNC']
Nrf2-deficient type II lung epithelial cell response to antioxidant supplementation	9	['ANXA2', 'COL1A1', 'KCTD12', 'LGALS1', 'MGP', 'NES', 'SERPINH1', 'THBS1', 'TNC']
DNA demethylation effect on terminally differentiated cells	9	['ANXA1', 'ANXA2', 'COL12A1', 'COL1A1', 'HMOX1', 'MGP', 'POSTN', 'THBS1', 'TNC']
DNA demethylation effect on glioblastoma cultures	9	['ANXA1', 'ANXA2', 'COL1A1', 'HMOX1', 'LGALS1', 'NES', 'POSTN', 'SERPINH1', 'THBS1']
Osteoarthritic chondrocytes: monolayer and matrix cultures	9	['ANXA1', 'COL12A1', 'COL1A1', 'EFEMP1', 'MGP', 'POSTN', 'SERPINH1', 'THBS1', 'TNC']
Fibroblast-derived induced pluripotent stem cells harboring pathogenic leucine-rich repeat kinase 2 mutation	9	['ANXA1', 'COL1A1', 'EFEMP1', 'LGALS1', 'POSTN', 'RUNX1', 'SLIT3', 'THBS1', 'TNC']
Hutchinson Gilford Progeria Syndrome cell line response to oncogenic challenge	9	['ANXA2', 'COL12A1', 'COL1A1', 'EFEMP1', 'G0S2', 'HMOX1', 'POSTN', 'SERPINH1', 'THBS1']
Tibolone hormone effect on postmenopausal endometrium	9	['ANXA2', 'COL14A1', 'COL1A1', 'EFEMP1', 'LGALS1', 'MGP', 'POSTN', 'SERPINH1', 'TNC']
Adipose tissue response to dihydrotestosterone: time course	9	['ANXA2', 'CIDEA', 'COL1A1', 'HMOX1', 'LGALS1', 'MGP', 'POSTN', 'SERPINH1', 'TNC']

Chondrocyte differentiation: time course	9	['ANXA1', 'ANXA2', 'COL12A1', 'COL1A1', 'MGP', 'MRC1', 'POSTN', 'SERPINH1', 'THBS1']
Osteoblast differentiation (MG-U74A)	9	['ANXA1', 'KCTD12', 'LGALS1', 'MGP', 'MRC1', 'POSTN', 'SERPINH1', 'THBS1', 'TNC']
Death receptor knockout effect on NEMO-deficient model of chronic liver disease	9	['ANXA2', 'COL14A1', 'COL1A1', 'HMOX1', 'LGALS1', 'POSTN', 'RUNX1', 'THBS1', 'TNC']
Preadipocytes from anatomically separate fat depots (HG-U133A)	9	['ANXA2', 'COL1A1', 'EFEMP1', 'G0S2', 'LGALS1', 'MGP', 'POSTN', 'SERPINH1', 'THBS1']

Supp Table 2. Transcriptomics analyses including genes out of the 26 stress-related genes. The GEO database was searched for studies in which the identified stress-related genes were differentially expressed. Most significant experiments, including at least nine of the genes of interest, are displayed. The names of the genes identified in each study is reported.

Cluster Number	Mouse/exercise stress	Pig/heat and inflammation stress	Chicken/nutritional stress	Chicken/heat stress
1	ACSS2, ACSS3, ADH1, ALDH1A3, ALDH1B1, ALDH2, ALDH3B2, ALDH9A1, AOC3, APOL6, CAT, CNDP2, COMT, CYP1A1, CYP1B1, CYP2E1, ECHDC1, GGH, GGT5, GPX1, GPX8, GSTA3, GSTA4, GSTT1, MAOB, MGST1, MGST2, ODC1, SOD3	LYZ, **RPL3L**, **RPS9**, IFI30, IFI6, MTHFD1L, MID1, OAS2	NT5E, PDE7B, PDE8A	ATP6V0D2, ATP6V1G3, ATP5J2
2	**ASB14**, **ASB15**, CBX4, CUL7, **FBXL16**, **FBXO10**, **FBXO21**, KBTBD13, KLHL13, KLHL2, KLHL5, PHC1, **RNF114**	**ANXA1**, **ANXA2**, LGALS1, C3AR1, S100A11, S100A10, **ADCY1**, **CCL21**, S100A6, **CCR5**, **CXCR4**, C5AR1, **ACKR3**	**PPARG**, HMOX1, **FABP4**, TNFRSF1A, NFATC2, MAP4K5, **PLIN2**, **ADIPOR2**	**ASB14**, PARK2, **RNF217**, **ASB12**, UBB
3	PABPC1, MRPS6, **RPL17**, **RPL35**, **RPL36**, **RPL31**, **RPL3**, EEF1A1, EIF1AX	OBSCN, RHOF, RHOB, ARHGEF2, ARHGAP36, ARHGAP23	**ANXA1**, **ANXA2**, LGALS1, **ANXA5**, **P2RY2**, SSTR2, **ADRA2A**	**COL12A1**, POSTN, **COL14A1**, **COL1A1**, **SERPINH1**, **COL4A6**, **COL6A3**, LEPREL1
4	ASPA, FOLH1, ASNS, NAT8L	SPSB1, SOCS3, **FBXO32**, **RNF34**, **FBXL4**, **RNF144B**, **ASB15**, **ASB4**	TNC, ITGA8, THBS4, CHAD	LTBP1, **THBS2**, **ADAMTSL2**, **THBS1**
5	PDE1A, PDE1B, **AK4**, **AK2**, ENTPD5, CTPS2, NME4, PDE3B	**AK4**, AMPD1, AMPD3, NT5E	**THBS1**, **ADAMTS8**, F13A1, **CLU**	LGALS1, **ANXA1, ANXA2**, LPAR6, PROKR2, GATB, **CXCL12**, CNR1, APLN
6	STX1B, STX2, STX6, STXBP2	**ADAMTS12**, **ADAMTSL4**, **ADAMTSL3**, **ADAMTS2**, **THSD4**, SPON2	GK, AKR1B10, **AGPAT9**, GPAM	ATP1B4, PRKG1, ATP1A2, ATP1A1, PLN
7	AP1S1, FNBP1, AP1S2, TFRC, CTSZ, DAB2, OCRL, AP3S1, SNX2, ARPC5, TXNDC5, SORT1, CD74	F3, ITGB3, MMP14, PLAU, ECM1, **TGFB3**, F13A1, **IGFBP6**, **SPARC**, **SERPINE1**, FGL2, ISLR, TIMP2, VEGFA, TIMP1, SPP1, **THBS1**, FN1, **IGF2**	ATP5G1, SLC25A4, PAM16	GFPT2, PRKAB2, PRKAG3, PFKFB4, PFKFB3, CPS1
8	AFAP1, MYL12B, LCP1, MYO5C, MYO5A, ACTG2, CORO2B, EFCAB2, CORO1B, CORO1A	CD163, HBB, C1QB, C1QA, C1QC	UCHL1, PARK2, **FBXL5**	HSPA4L, HSPH1, HSP90B1, HSP90AA1, APOA1, TTR, DNAJA4
9	**COL5A3**, **COL18A1**, CRTAP, **COL14A1**, **COL12A1**, **COL15A1**, **COL4A2**, **COL7A1**, **SERPINH1**, **COL1A1**, **COL24A1**	EFEMP1, MFAP5, MFAP4, MFAP2, LOXL3, LOXL4, LOXL2, LOX, FBLN1, FBLN2	DNTTIP1, MBD3, BRPF3	
10	CFD, ORM1, RARRES2, **TGFB1**, QSOX1, **SERPINE1**, ALB, **IGF1**, **THBS1**, **SPARC**, ACTN4, MMRN1, **CLU**, PROS1, LGALS3BP, PCYOX1L	NR1D1, BHLHE41, NFIL3	FGF13, SH3GL3, KDR	
11	GNAI1, AGT, GNB1, C3, GNAQ, **ADCY9**, ADORA1, OPN3, **CXCL13**, **CCR2**, GRK5, **ADCY5**, HCAR1, GRK3, **ADCY1**, HCAR2, **ADCY3**, **CCL11**, PLCB1, SUCNR1, **ACKR3**, **ANXA1**, **CCR1**, TSHR, UTS2R, G G7, MC2R, GNG5, GNG2, VIPR1, OXTR, PTGER3, PTH1R, SLC9A3R1, CYSLTR1, KCNJ15, **ADRB3**, HEBP1, **P2RY2**, **CXCL9**, FFAR2	ACTC1, CFL1, EPHB2, TNNC2, EPHB3	VIM, TNNI1, TPM2, TNNT2, MYBPC1	

12	CHST15, CHST12, CHST11, DES	**COL1A1**, **COL14A1**, **COL1A2**, **COL3A1**, **COL17A1**, **COL8A2**, **COL8A1**, **COL16A1**, PCOLCE2, POSTN, **COL12A1**, **COL19A1**, **COL6A3**, **SERPINH1**, **COL6A6**, **COL5A1**, **COL5A2**, P4HA3	**COL1A1**, **COL12A1**, **SERPINH1**, **COL14A1**, **COL19A1**, **COL24A1**, **COL21A1**
13	LPCAT3, DEGS1, DGAT1, DGAT2, SGPL1, PLD2, PLPPR2, MGLL, LPL, MOGAT1, CERS2, THRSP, CHPT1, CDS1, PLPP2, PLPP3, CERS6, CERS5, ACSL5, ACSL6, CPT1A, **AGPAT2**, PNPLA2, PNPLA3, SMPD3, FASN	HS3ST6, SDC2, SDC3, DSE, DCN, TNC, CSPG4, VCAN, BGN	
14	AGMO, CYP51, IDI1, LSS, MSMO1, NSDHL, SC5D, SDR42E1, SIGMAR1	KERA, PRELP, FMOD, OGN, ECM2	
15	ACOT4, ACOT2, TKT, PGLS, PGD, FBP2, PFKP, TALDO1, PFKL, HK1, SCD2	FOS, CYR61, KDM6B, MGP, FOSB, ODC1, HMOX1, JUN, JUND, EGR1, JUNB, EGR2, DUSP1, EGR3	
16	**ADIPOQ**, ANGPTL4, CD36, CEBPA, EBF1, **FABP4**, LEP, MED13, PCK1, **PLIN1**, **PPARG**, RBP4, RETN, UCP2	CHRND, CHRNG, CHRNA1, CHRNB1	
17	PEX13, PEX16, PEX11A, PEX3, SLC25A17, ABCD2		
18	FUT4, CHST1, KERA, B3GALT2, ST3GAL4, FMOD, B4GALT2		

Supp Table 3. Identification of common proteins and protein families clustered in the four stress-model networks. All clusters components from each network described in (Figure 6) were extracted and listed in this table. Clusters identified in at least two stress-model networks are displayed with same colors; proteins or protein families common to equivalent clusters are represented in bold. Four clusters are composed of proteins and protein families differentially expressed in the four stress-models: the collagen family (depicted in green), the ubiquitin family (depicted in yellow), the chemokine-related immune response (depicted in cyan) and the thrombospondin-related proteins (depicted in pink).

Chapter 2

1) Introduction to "Mosaic organization of metabolic pathways between bacteria and archaea species"

During my thesis, I also worked to get insights into the selenoprotein N (SelenoN) function which is one of the main subject of the team. The selenoprotein N is one of the 25 selenium containing proteins in human. This protein is encoded by a gene known to cause different forms of central muscular dystrophies when mutated in humans, but its precise function still remains elusive. Initially, this gene was detected within nearly all animal genomes and its presence was believed to be restricted to this phylum. However, recently, we could identify that SelenoN orthologous proteins are also encoded within bacterial genomes from one single group of unclassified bacteria referred as *Candidatus poribacteria*. Surprisingly, SELENON orthologous genes were identified only within a fraction of the sequenced *C. poribacteria* genomes and their distribution were correlated to the Poribacteria lifestyle. Indeed, this taxa gathers bacteria living either as symbionts with sponges and corals, or as free-living organisms in the sea. Only the symbiotic Poribacteria contain the SELENON gene. Presence of other eukaryotic- or archaea-specific genes (xenologs) was also observed within *C. poribacteria* genomes, indicating abundant gene exchanges through horizontal gene transfers (HGTs) between the bacteria and its eukaryotic host, as well as with other organisms of the prokaryotic community constituting the host microbiome. These HGTs are predicted to play an important role for the establishment of the symbiotic relationship with eukaryotic hosts. To get clues about SelenoN possible function, I worked on genes specifically conserved in SELENON-containing *C. poribacteria*, and I serendipitously identified two enzymes from the phosphopantothenate biosynthetic pathway showing a singular distribution between these organisms. These enzymes participate in two consecutive reactions and are well-known to display a bacterial- or archaeal-specific phylogenetic profile. By comparing the gene distribution in the two *C. poribacteria* subgroups, we highlighted a mosaic organization of the two enzymes involved in phosphopantothenate biosynthesis, with some individuals using the bacterial enzymes while others relying on the archaeal proteins to achieve the production of this metabolite important to acetyl-Coenzyme A metabolism. In addition to improve gene annotations in this bacterial and other related groups, this observations highlights the dynamic evolution of the acetyl-Coenzyme A metabolic pathway during evolution.

This manuscript is submitted for publication as a short letter to *Molecular Biology and Evolution*.

In this study, A.L. and L.T. designed the experiments, and L.T. conducted the identification of orthologous proteins between the different prokaryotic groups, performed the multiple sequences alignment and computation of phylogenetic trees. Both A.L. and L.T. drafted the manuscript.

2) Manuscript II

Mosaic evolution of the phosphopantothenate biosynthesis pathway in bacteria and archaea.

Luc Thomès[1], Alain Lescure[1]

[1]Université de Strasbourg, CNRS, Architecture et Réactivité de l'ARN, UPR9002, Strasbourg, France.

Title: Mosaic evolution of the phosphopantothenate biosynthesis pathway in bacteria and archaea.

Authors: Luc Thomès[1], Alain Lescure[1]

Affiliation: [1]Université de Strasbourg, CNRS, Architecture et Réactivité de l'ARN, UPR9002, Strasbourg, France.

Corresponding author: a.lescure@cnrs-ibmc.unistra.fr

Abstract :

Phosphopantothenate is an essential precursor to synthesis of Coenzyme A (CoA), a metabolite central to many metabolic pathways. Organisms of the archaeal phyla were shown to utilize a different phosphopantothenate biosynthetic pathway from the eukaryotic and bacterial one. In this study, we report that symbiotic bacteria from the group *Candidatus poribacteria* present enzymes of the archaeal pathway, namely pantoate kinase (PoK) and phosphopantothenate synthetase (PPS), mirroring what was demonstrated for *Picrophilus torridus,* an archaea partially utilizing the bacterial pathway. Our results support the ancient origin of the CoA pathway in the three domains of life, but also highlight its complex and dynamic evolution. Importantly, this study helps to improve protein annotation for this pathway in the *Candidatus poribacteria* group and other related organisms.

Keywords: phosphopantothenate pathway, Coenzyme A, *Candidatus poribacteria*

Introduction

Coenzyme A (CoA) is an essential metabolite common to many biosynthetic pathways. More than 400 enzyme-catalyzed reactions are known to involve CoA as a substrate. In most bacteria and eukaryotes, synthesis of one of the first intermediates in this pathway, phosphopantothenate, is achieved in a two-step reaction: synthesis of pantothenate by condensation of pantoate with ß-alanine, followed by pantothenate phosphorylation. Interestingly, it was shown that archaea utilize an alternative pathway, where the two consecutive reactions are exchanged, with the phosphorylation step occurring first, followed by addition of ß-alanine (see Fig. 1) (Yokooji et al., 2009; Ishibashi et al., 2012; Tomita et al., 2012; Katoh et al., 2013). This difference was proposed as an intrinsic characteristic that distinguishs bacterial and archaeal phyla. In comparative genome analyses of a group of bacteria, *Candidatus poribacteria*, we made the striking observation that the enzymes corresponding to the phosphopantothenate synthesis pathway were not correctly annotated. *Candidatus poribacteria* refers to an unclassified group of marine bacteria, evolutionarily related to the superphylum *Planctomycetes-Verrucomicrobia-Chlamydia* (Fieseler et al., 2004; Kamke et al., 2014). These bacteria were originally identified as members of the bacterial community living in symbiosis with diverse sponge species, including *Aplysina aerophoba*. *Candidatus poribacteria* present the peculiarity of sharing several eukaryotic-like features, such as complex inner membrane structures similar to eukaryotic intracellular compartments and a nucleoid-like structure. Interestingly, a recent metagenomic study identified additional strains of the *Candidatus poribacteria* group living as free-living organisms present in seawater, defining two distinct subgroups characterized according to their lifestyle, and designated Entoporibacteria for the sponge-associated and Pelagiporibacteria for the free-living ones. Genomic analyses revealed a high level of inner divergence between the two groups, indicating a different evolutionary history (Podell et al., 2019). Ontological analysis of the gene sets specific to each subgroup predicted that a large part of the genes specific to the Entoporibacteria group contribute to the host-symbiont interaction.

Results

In bacteria, pantothenate synthetase (PS) is an enzyme responsible for condensation of ß-alanine and D-pantoate resulting in D-pantothenate. Subsequently, pantothenate kinase (PanK) phosphorylates D-pantothenate to D-4'-phosphopantothenate (Fig. 1). D-4'-phosphopantothenate enzymes can be classified into three different types based on their sequences: PanKs of type I and type III are found in a wide range of bacteria, while type II is mostly present in eukaryotes, but has also been identified in *Staphylococci* (see Fig. 2A). Intriguingly, a search for PanK and PS genes in the symbiotic *Candidatus poribacteria* genomes failed to identify homologs for these enzymes. Based on multiple sequence alignments and phylogenetic tree construction, we determined that the proteins annotated as GHMP kinase (GHMPK) and phosphopantothenate/pantothenate synthetase (PP/PS) in these genomes are similar to the archaeal enzymes pantoate kinase (PoK) that phosphorylates D-pantoate, and phosphopantothenate synthetase (PPS), responsible for condensation of D-4-phosphopantoate with ß-alanine, respectively (Fig. 2A and 2B, and Supp Fig. 1 and 2). The *Candidatus poribacteria* proteins displayed 31% or 44% identity with *Methanospirillum hungatei* PoK and PPS respectively (Supp Table 2). Important PPS residues for substrates binding, deduced form the 3D structure (Kim et al., 2013), appeared to be conserved (Supp Fig 2). This observation suggests the presence of the archaeal pathway in the symbiotic Poribacteria. Consequently, some ambiguous protein annotations can be resolved, since GHMPK and PP/PS are orthologs of PoK and PPS enzymes respectively, based on reciprocal best-hit BLAST searches. The use of the archaeal rather than the bacterial reaction order for the synthesis of the CoA intermediate D-4'-phosphopantothenate in the symbiotic Entoporibacteria group mirrors what has been shown in the archaea *Picrophilus torridus,* in which an enzyme much closer to the bacterial PanK than to the canonical archaeal PoK was found and has been annotated as "archaeal PanK" (Takagi et al., 2010; Shimosaka et al., 2016) (Fig. 2A).

Strikingly, the archaeal enzymes of phosphopantothenate pathway present in Entoporibacteria appeared to be absent from Pelagiporibacteria that possess the classical bacterial genes coding for type-III PanK and PS (Fig. 2A and 2B, and Supp Fig. 3 and 4). The *Candidatus poribacteria* proteins presented 26% or 47% identity with *Pseudomonas aeruginosa* type-III PanK and *Escherichia coli* PS respectively (Supp Table 2). Importantly, PanK and PS residues for substrates binding, deduced form the 3D structure (von Delft et al., 2001; Yang et al., 2006), appeared also to be conserved (Supp Fig 3 and 4). This observation implies that utilization of the alternative phosphopantothenate pathways is dictated by the bacterial interaction with its environment and likely contributes to the holobiont interaction.

Discussion

This study revealed that the origin of phosphopantothenate biosynthesis is more complex than anticipated and that what was initially defined as an archaeal pathway is also used in some bacterial groups. It also suggests a high degree of evolutive and functional plasticity in the biosynthesis of the metabolic intermediates of CoA. Interestingly, a similar mosaic evolution utilizing alternative routes in different bacteria and archaea were identified for the mevalonate pathway, a biosynthetic process downstream of phosphopantothenate that converts CoA into isoprenoid precursors (Lombard and Moreira, 2011; Hoshino and Gaucher, 2018). Of note, despite the similarities in the catalyzed reactions, multiple sequence alignment showed no common domain between PanK and PoK on one side, nor PS and PPS on the other side. This observation clearly indicated that these enzymes originate from different ancestral genes.

Despite the ubiquity of the CoA pathway, the uneven taxonomic distribution of the two routes for phosphopantothenate synthesis raises several questions about their evolutionary origin. Two alternative but non-exclusive explanations for this phylogenic plasticity can be proposed. On the one hand, the exceptions to the phyla-specific synthesis pathways were acquired by distinct archaeal or bacterial groups through horizontal gene transfers. Our results support this hypothesis, since we showed that, from the pool of sequenced *Candidatus poribacteria* genomes, only symbiotic Entoporibacteria use the archaeal pathway, as the free-living Pelagiporibacteria use the bacterial pathway. Entoporibacteria are part of a large microbial community that colonizes the sponge mesohyl, constituting an ancestral form of microbiota (Webster and Thomas, 2016; Pita et al., 2018). This community includes both bacteria and archaea in close proximity, a condition favorable for interindividual gene transfer. Indeed, analysis of *Candidatus poribacteria* genomes revealed the presence of many genes coding for eukaryote-like proteins, which were predicted to be involved in mediating host-microbe interactions (Kamke et al., 2014; Podell et al., 2019). Alternatively, the two pathways may have originally coexisted in a common ancestor and one or another of the two branches were then positively or negatively selected according to metabolic properties or toxic effects of the possible metabolic intermediates. However, so far coexistence of the two pathways has not been described in any organism, suggesting that they are exclusive to each other. Our phylogenetic profiling study reinforces the notion of a mosaic of orthologous relationships of CoA biosynthetic genes between bacteria and archaea as originally proposed by Genschel (2004).

Materials and methods

Sequence retrieval and multiple alignment construction

Reference protein sequences for each enzyme of interest (type I/II/III PanKs, PoK, PS, PPS) were retrieved from the UniprotKB database(https://www.uniprot.org/). Accession numbers for the selected enzymes are shown in Supp Table 1. For these sequences, conserved protein domains were extracted from the CDD database (https://www.ncbi.nlm.nih.gov/Structure/cdd) and conserved regions were identified in the multiple alignment representative of each domain (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Then, three additional sequences from other species were retrieved for each enzyme of interest based on their reviewed Uniprot annotations (see Supp Table 1). For each enzyme, a multiple alignment of the retrieved sequences was constructed using Muscle and conserved regions were manually compared to the expected regions identified in the conserved protein domains.

We then used each reference protein sequence as a query for a tBLASTn search of the *Candidatus poribacteria* genomes in the WGS database (https://www.ncbi.nlm.nih.gov/genbank/wgs/), since there are more genomic assemblies (64) than proteomes available (22). This allowed us to retrieve four protein sequences from pelagic *C. poribacteria* using the reference type-III PanK protein as a query, and four protein sequences from symbiotic *C. poribacteria* using the reference PoK protein as a query (Supp Table 2). Genome assemblies and contigs used are shown in Supp Table 3. Finally, multiple alignments of the retrieved PoK enzymes including Entoporibacteria mis-annotated sequences and of the retrieved type-III PanK enzymes including Pelagiporibacteria sequences were constructed using Muscle 3.8.31 (Edgar, 2004) (Supp Fig. 1 and 3).

A similar approach was used to obtain PS and PPS sequences. First, we retrieved PS sequences for the three reference bacteria previously determined and four PPS sequences for the four PoK-coding archaea from UniprotKB. Accession numbers for the selected enzymes are shown in Supp Table 1. For these sequences, conserved protein domains were identified in CDD. For each enzyme, a multiple alignment was constructed and conserved regions were compared to the expected regions from the conserved protein domains. Using the *E. coli* and the *M. hungatei* sequences, we then performed a tBLASTn search of the WGS database to retrieve the four protein sequences from the previously considered Pelagiporibacteria and Entoporibacteria. Genome assemblies and contigs used are presented in Supp Table 2. Finally, multiple alignments of the retrieved PPS enzymes (including Entoporibacteria mis-annotated sequences)

and of the retrieved PS enzymes (including Pelagiporibacteria sequences) were constructed using Muscle (Fig. Supp. 2 and 4).

Phylogenetic analyses

The tree construction was carried out using PhyloBayes v.4.1 (Lartillot et al. 2009) for (i) the set of type-III PanKs or (ii) PoK proteins, and (iii) the set of PS or (iv) PPS proteins (Supp Figures 1 and 3 or 2 and 4 respectively). For each set, two Bayesian analyses were performed using either the single substitution model (LG) or the profile mixture model (CAT-GTR). Each analysis was performed in duplicate, and the convergence was assessed using the bpcomp function provided by PhyloBayes. For each analysis, 100 sampled points were removed as burn-in. The tree topologies obtained for each set using both models are almost identical, and therefore only the CAT-GTR trees are shown here (Fig. 2).

Supplementary materials

Supp Figure 1. Multiple protein alignment of reference pantoate kinase (PoK) and mis-annotated Entoporibacterial sequences.

Supp Figure 2. Multiple protein alignment of reference phosphopantothenate synthetase (PPS) and mis-annotated Entoporibacterial sequences.

Supp Figure 3. Multiple protein alignment of reference and Pelagiporibacterial type-III pantothenate kinase (PanK) sequences.

Supp Figure 4. Multiple protein alignment of reference and Pelagiporibacterial pantothenate synthetase (PS) sequences.

Supp Table 1. Conserved domains and Uniprot accession numbers of bacterial and archaeal phosphopantothenate biosynthetic enzymes.

Supp Table 2. Identitfiers of Candidatus poribacteria contigs encoding bacterial or archaeal phosphopantothenate biosynthetic enzymes.

References

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792-1797.

Fieseler L, Horn M, Wagner M, Hentschel U. 2004. Discovery of the Novel Candidate Phylum "Poribacteria" in Marine Sponges. *Appl. Environ. Microbiol*. 70(6):3724-3732.

Genschel U. 2004. Coenzyme A Biosynthesis: Reconstruction of the Pathway in Archaea and an Evolutionary Scenario Based on Comparative Genomics. *Mol Biol Evol*. 21:1242–1251.

Ishibashi T, Tomita H, Yokooji Y, Morikita T, Watanabe B, Hiratake J, Kishimoto A, Kita A, Miki K, Imanaka T, et al. 2012. A detailed biochemical characterization of phosphopantothenate synthetase, a novel enzyme involved in coenzyme A biosynthesis in the Archaea. *Extremophiles* 16:819–828.

Kamke J, Rinke C, Schwientek P, Mavromatis K, Ivanova N, Sczyrba A, Woyke T, Hentschel U. 2014. The Candidate Phylum Poribacteria by Single-Cell Genomics: New Insights into Phylogeny, Cell-Compartmentation, Eukaryote-Like Repeat Proteins, and Other Genomic Features. *PLoS ONE.* 9:e87353.

Katoh H, Tamaki H, Tokutake Y, Hanada S, and Chohnan S. 2013. Identification of pantoate kinase and phosphopantothenate synthetase from Methanospirillum hungatei. *J Biosci Bioeng*. 115:372–376.

Kim M-K, An YJ, Cha S-S. 2013. The crystal structure of a novel phosphopantothenate synthetase from the hyperthermophilic archaea, Thermococcus onnurineus NA1. *Biochem Bioph Res Co.* 439:533–538.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17):2286-2288

Pita L, Rix L, Slaby BM, Franke A, Hentschel U. 2018. The sponge holobiont in a changing ocean: from microbes to ecosystems. *Microbiome*. 6:46.

Podell S, Blanton JM, Neu A, Agarwal V, Biggs JS, Moore BS, Allen EE. 2019. Pangenomic comparison of globally distributed Poribacteria associated with sponge hosts and marine particles. *ISME J*. 13:468–481.

Shimosaka T, Tomita H, Atomi H. 2016. Regulation of Coenzyme A Biosynthesis in the Hyperthermophilic Bacterium Thermotoga maritima. *J. Bacteriol.* 198:1993–2000.

Takagi M, Tamaki H, Miyamoto Y, Leonardi R, Hanada S, Jackowski S, Chohnan S. 2010. Pantothenate Kinase from the Thermoacidophilic Archaeon Picrophilus torridus. *J. Bacteriol*. 192:233–241.

Tomita H, Yokooji Y, Ishibashi T, Imanaka T, Atomi H. 2012. Biochemical Characterization of Pantoate Kinase, a Novel Enzyme Necessary for Coenzyme A Biosynthesis in the Archaea. *J. Bacteriol.* 194:5434–5443.

von Delft F, Lewendon A, Dhanaraj V, Blundell TL, Abell C, Smith AG. 2001. The Crystal Structure of E. coli Pantothenate Synthetase Confirms It as a Member of the Cytidylyltransferase Superfamily. *Structure.* 9:439–450.

Webster NS, Thomas T. 2016. The Sponge Hologenome. *mBio*. 7:e00135-16.

Yang K, Eyobo Y, Brand LA, Martynowski D, Tomchick D, Strauss E, Zhang H. 2006. Crystal Structure of a Type III Pantothenate Kinase: Insight into the Mechanism of an Essential Coenzyme A Biosynthetic Enzyme Universally Distributed in Bacteria. *JB* 188:5532–5540.

Yokooji Y, Tomita H, Atomi H, Imanaka T. 2009. Pantoate Kinase and Phosphopantothenate Synthetase, Two Novel Enzymes Necessary for CoA Biosynthesis in the Archaea. *J. Biol. Chem*. 284:28137–28145.

Acknowledgements

We are indebted to Julie Thompson for fruitful discussions and comments about the results of this study. This study was supported by funding of CNRS, Fondation Meyer and ADISSEO FRANCE SAS –Laboratoire Commun DiagnOxi - to AL, and a CIFRE fellowship from the Association Nationale Recherche et Technologie to LT.

Figures legends

Figure 1: Bacterial and archaeal phosphopantothenate biosynthetic pathways. Most bacteria, like eukaryotes, use PS and PanKs to synthetize 4'-phosphopantothenate from pantoate. The alternative pathway utilized by most archaea involve PoK and PPS enzymes that catalyze similar reactions but in the reverse reaction order.

Figure 2: Bayesian phylogenetic trees of enzymes involved in phosphopantothenate pathway in archaeal and bacterial groups. The obtained trees show the distribution of the symbiotic (Entoporibacteria*)* and free-living (Pelagiporibacteria) *Candidatus poribacteria* groups according to the use of (**A**) pantothenate kinase (PanK) and pantoate kinase (PoK) enzymes and (**B**) pantothenate synthetase (PS) and phosphopantothenate synthetase (PPS) enzymes (upper and lower panels respectively). Branch lengths are shown for major nodes. Scale bar represents 0.2 and 0.1 amino acid replacements per site per unit evolutionary time on panels **A** and **B** respectively. Abbreviation: CPO: *Candidatus poribacteria*; MHU: *Methanospirillum hungatei*; TKO: *Thermococcus kodakarensis*; MJA: *Methanocaldococcus jannaschii*; MMA: *Methanosarcina mazei*; ECO: *Escherichia coli*; SAU: *Staphylococcus aureus*; PAR: *Psychrobacter arcticus*; ABA: *Acinetobacter baumannii*; PAE: *Pseudomonas aeruginosa*; BSU: *Bacillus subtilis*.

Figure 1

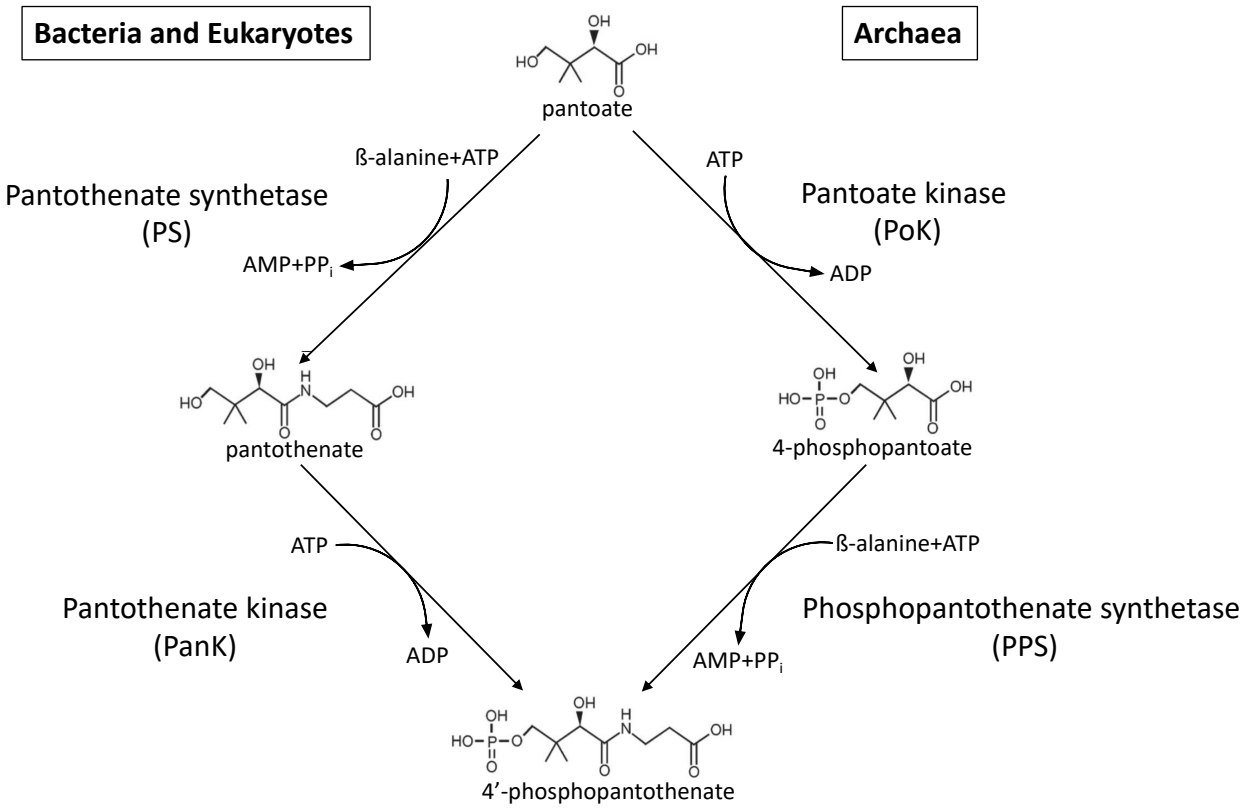
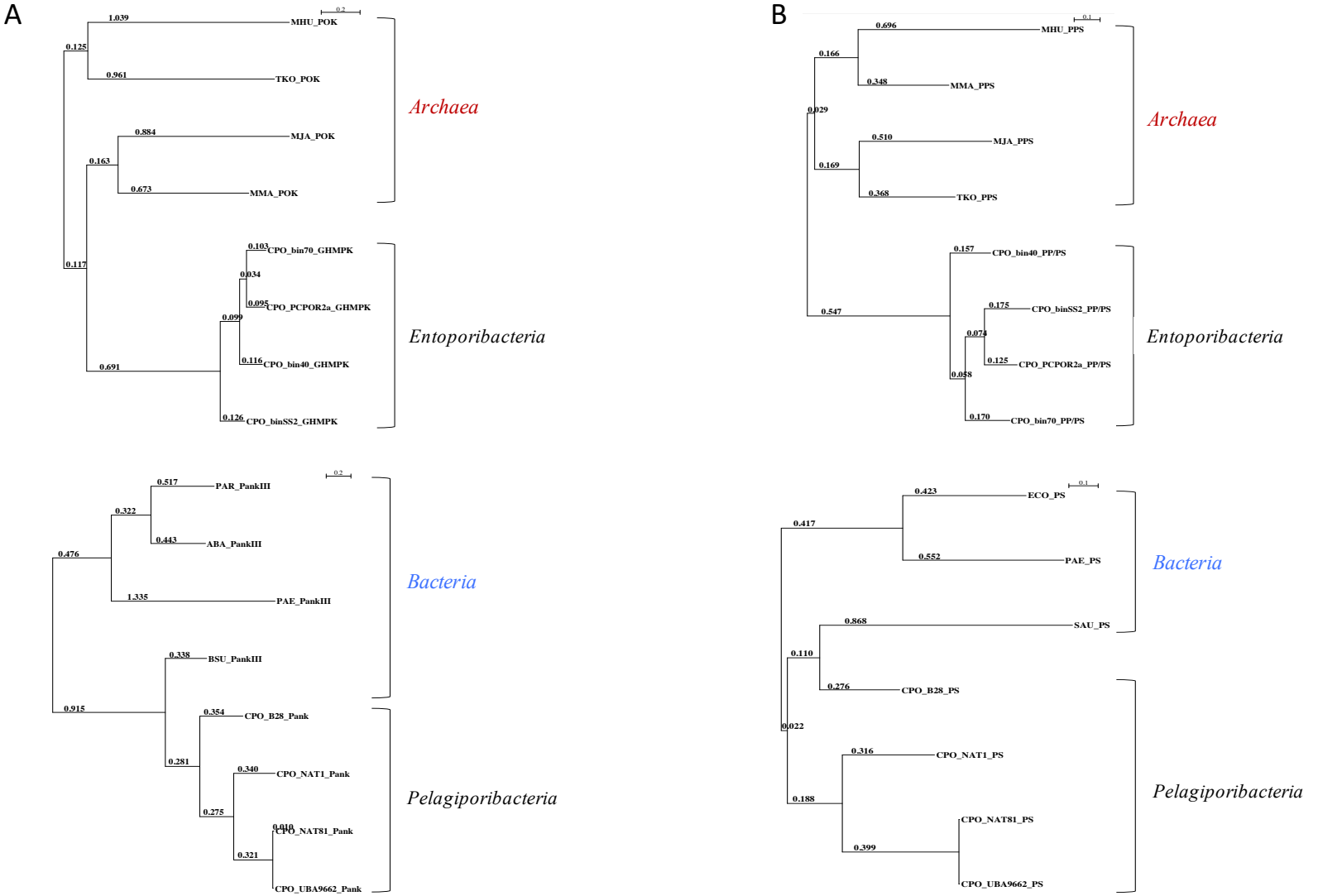
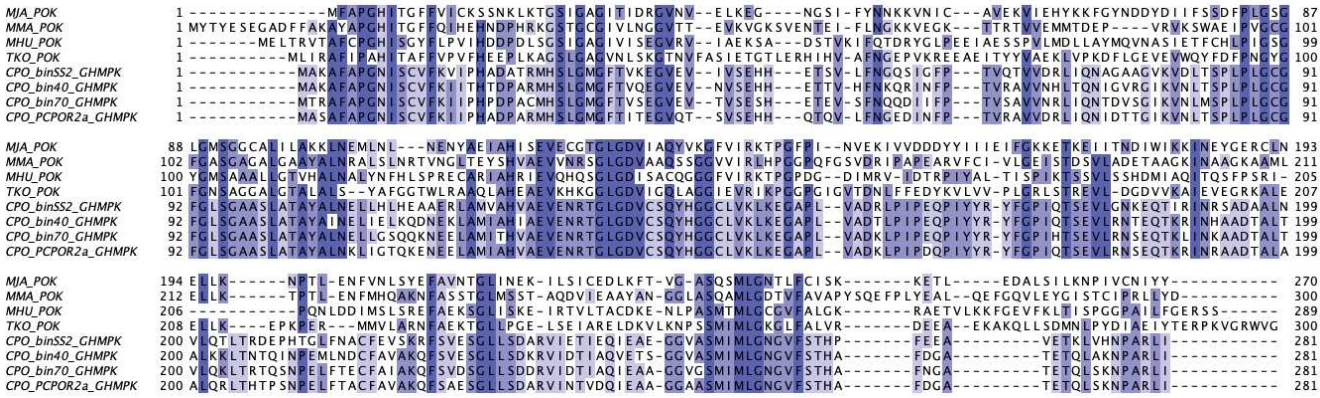
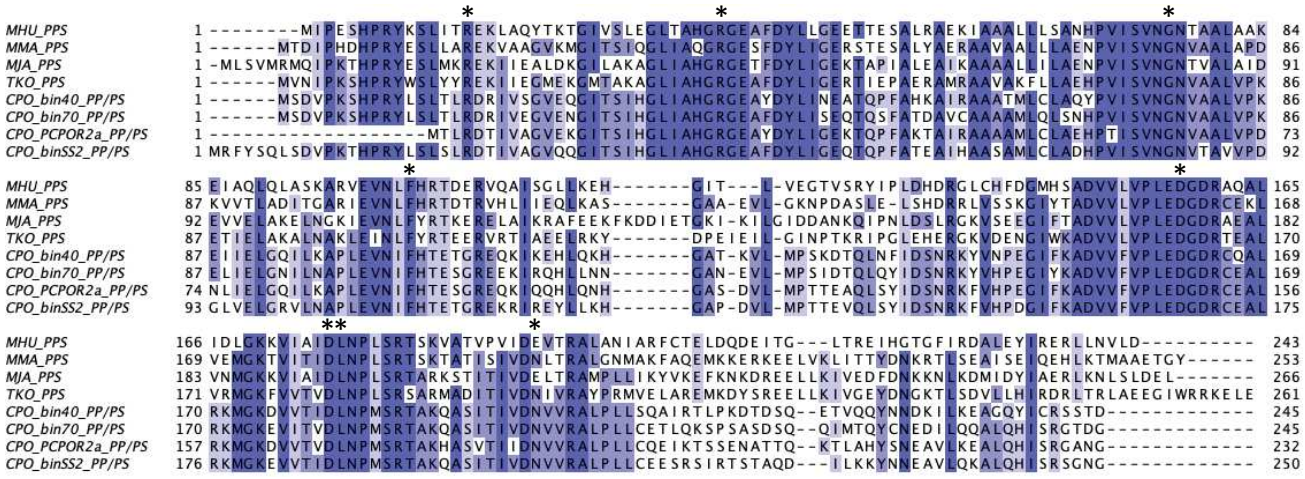

Bacteria and Eukaryotes
Archaea
pantoate
ß-alanine+ATP
ATP
Pantothenate synthetase
(PS)
Pantoate kinase
(PoK)
AMP+PPi
ADP
pantothenate
4-phosphopantoate
ATP
ß-alanine+ATP
Pantothenate kinase
(PanK)
Phosphopantothenate synthetase
(PPS)
ADP
AMP+PPi
4'-phosphopantothenate

Figure 2

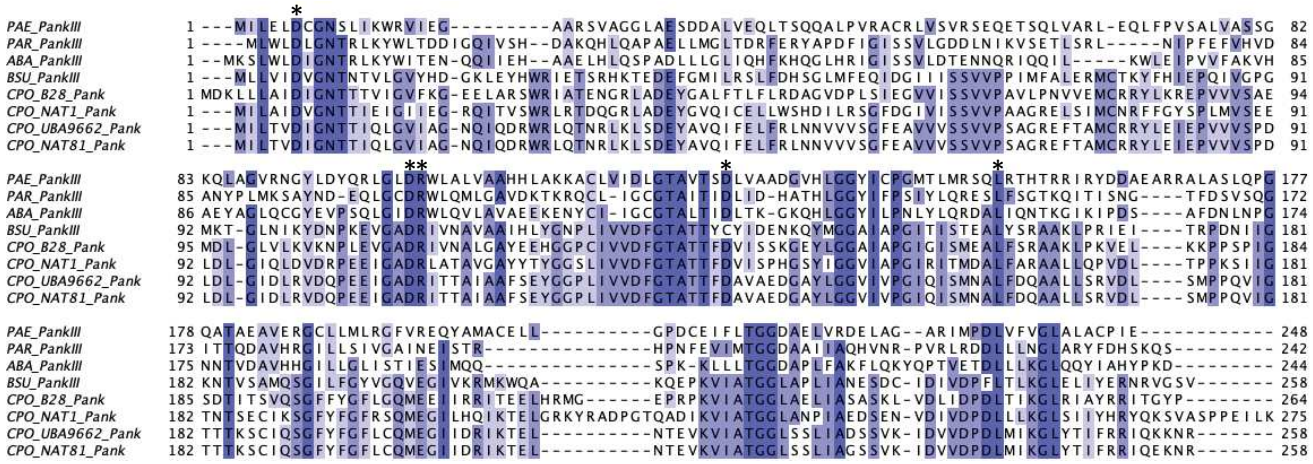

A
0.2
1.039
MHU_POK
0.125
0.961
TKO_POK
0.884
MJA_POK
0.163
0.673
MMA_POK
Archaea
0.103
CPO_bin70_GHMPK
0.034
0.095
CPO_PCPOR2a_GHMPK
0.099
0.117
0.116
CPO_bin40_GHMPK
0.691
0.126
CPO_binSS2_GHMPK
Entoporibacteria
0.2
0.517
PAR_PankIII
0.322
0.443
ABA_PankIII
0.476
1.335
PAE_PankIII
Bacteria
0.338
BSU_PankIII
0.915
0.354
CPO_B28_Pank
0.281
0.340
CPO_NAT1_Pank
0.275
0.010
CPO_NAT81_Pank
0.321
CPO_UBA9662_Pank
Pelagiporibacteria
B
0.1
0.696
MHU_PPS
0.166
0.348
MMA_PPS
0.029
0.510
MJA_PPS
0.169
0.368
TKO_PPS
Archaea
0.157
CPO_bin40_PP/PS
0.175
CPO_binSS2_PP/PS
0.547
0.074
0.125
CPO_PCPOR2a_PP/PS
0.058
0.170
CPO_bin70_PP/PS
Entoporibacteria
0.1
0.423
ECO_PS
0.417
0.552
PAE_PS
Bacteria
0.868
SAU_PS
0.110
0.276
CPO_B28_PS
0.022
0.316
CPO_NAT1_PS
0.188
CPO_NAT81_PS
0.399
CPO_UBA9662_PS
Pelagiporibacteria



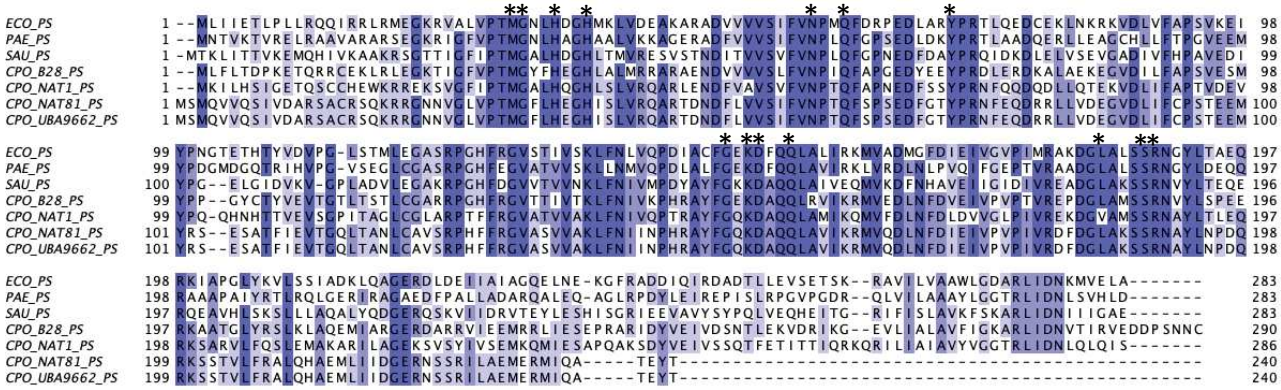
Supp Fig. 1: Multiple protein alignment of reference archaeal pantoate kinase (PoK) sequences with Entoporibacterial sequences annotated as GHMP kinase. Alignment was constructed using Muscle and visualized using Jalview (Waterhouse et al., 2009). Blue boxes correspond to the percent identity with the consensus sequence in the alignment: dark blue > 80%, medium blue > 60%, light blue > 40%, white < 40%. Abbreviation: MJA: *Methanocaldococcus jannaschii*; MMA: *Methanosarcina mazei*; MHU: *Methanospirillum hungatei*; TKO: *Thermococcus kodakarensis*; CPO: *Candidatus poribacteria*.



Supp Fig. 2: Multiple protein alignment of reference archaeal phosphopantothenate synthetase (PPS) sequences with Entoporibacterial sequences annotated as phosphopantothenate/pantothenate synthetase. Alignment was performed using Muscle and visualized using Jalview (Waterhouse et al., 2009). Conserved residues identified as important for substrate binding are indicated by a star (Kim et al., 2013). Blue boxes correspond to the percentage identity with the consensus sequence in the alignment: dark blue > 80%, medium blue > 60%, light blue > 40%, white < 40%. Abbreviation: MHU: *Methanospirillum hungatei*; MMA: *Methanosarcina mazei*; MJA: *Methanocaldococcus jannaschii*; TKO: *Thermococcus kodakarensis*; CPO: *Candidatus poribacteria*.



Supp Fig. 3: Multiple protein alignment of reference and correctly annotated Pelagiporibacterial type-III pantothenate kinase (PanK) sequences. Alignment was performed using Muscle and visualized using Jalview (Waterhouse et al., 2009). Conserved residues identified as important for catalysis are indicated by a star (Yang et al., 2006). Blue boxes correspond to the percentage identity with the consensus sequence in the alignment: dark blue > 80%, medium blue > 60%, light blue > 40%, white < 40%. Abbreviation: PAE: *Pseudomonas aeruginosa*; PAR: *Psychrobacter arcticus*; ABA: *Acinetobacter baumannii*; BSU: *Bacillus subtilis*; CPO: *Candidatus poribacteria*.



Supp Fig. 4: Multiple protein alignment of reference and correctly annotated Pelagiporibacterial pantothenate synthetase (PS) sequences. Alignment was performed using Muscle and visualized using Jalview (Waterhouse et al., 2009). Conserved residues identified as important for substrate binding are indicated by a star (von Delft et al., 2001). Blue boxes correspond to the percentage identity with the consensus sequence in the alignment: dark blue > 80%, medium blue > 60%, light blue > 40%, white < 40%. Abbreviation: ECO: *Escherichia coli*; PAE: *Pseudomonas aeruginosa*; SAU: *Staphylococcus aureus*; CPO: *Candidatus poribacteria*.

References

Kim M-K, An YJ, Cha S-S. 2013. The crystal structure of a novel phosphopantothenate synthetase from the hyperthermophilic archaea, Thermococcus onnurineus NA1. *Biochem Bioph Res Co.* 439:533–538.

von Delft F, Lewendon A, Dhanaraj V, Blundell TL, Abell C, Smith AG. 2001. The Crystal Structure of E. coli Pantothenate Synthetase Confirms It as a Member of the Cytidylyltransferase Superfamily. *Structure.* 9:439–450.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189–1191.

Yang K, Eyobo Y, Brand LA, Martynowski D, Tomchick D, Strauss E, Zhang H. 2006. Crystal Structure of a Type III Pantothenate Kinase: Insight into the Mechanism of an Essential Coenzyme A Biosynthetic Enzyme Universally Distributed in Bacteria. *JB* 188:5532–5540.

PanK/PoK			
Species	**Enzymes**	**Conserved domain**	**Accession number**
Escherichia coli	**type-I PanK**	**PRK05439**	**P0A6I3**
Yersinia pestis	type-I PanK		A9R361
Lactobacillus paracasei	type-I PanK		Q036Y4
Picrophilus torridus	type-I PanK		Q6L2I5
Staphylococcus aureus	**type-II PanK**	**PRK13317**	**Q6G7I0**
Bacillus cereus	type-II PanK		B7JSQ5
Oceanobacillus iheyensis	type-II PanK		Q8EN08
Bacillus thuringiensis subsp. Konkukian	type-II PanK		Q6HHK0
Pseudomonas aeruginosa	**type-III PanK**	**PRK13322**	**Q9HWC1**
Psychrobacter arcticus	type-III PanK		Q4FUX4
Bacillus subtilis	type-III PanK		P37564
Acinetobacter baumannii	type-III PanK		B0VU08
Methanospirillum hungatei	**PoK**	**COG1829**	**Q2FUB2**
Thermococcus kodakarensis	PoK		Q5JHF1
Methanocaldococcus jannaschii	PoK		Q58379
Methanosarcina mazei	PoK		A0A0E3RDM2
PS/PPS			
Species	**Enzymes**	**Conserved domain**	**Accession number**
Escherichia coli	PS	PRK00380	Q8X930
Staphylococcus aureus	PS	PRK00380	P65658
Pseudomonas aeruginosa	PS	PRK00380	A6VCI6
Methanospirillum hungatei	PPS	PRK13761	Q2FUA9
Methanocaldococcus jannaschii	PPS	PRK13761	Q57662
Thermococcus kodakarensis	PPS	PRK13761	Q5JIZ8
Methanosarcina mazei	PPS	PRK13761	Q8PUQ1

Supp Table 1. Conserved domains and Uniprot accession numbers of bacterial and archaeal phosphopantothenate biosynthetic enzymes. Bacterial type-I, type-II, type-III PanKs and archaeal PoK enzymes were retrieved from four reference organisms depicted in bold. Conserved domains were identified to validate protein annotations and the sequences were used to retrieve homologs from three other species for each enzyme based on reviewed Uniprot annotations. PS and PPS sequences were retrieved for the three reference bacteria and the four archaea encoding the PoK enzymes respectively. Archaeal species are depicted in red.

Query protein (Organism: UniprotKB ID)	*C. poribacteria* contig ID (strain)	E-value	Identity (%)	Query cover (%)
Type-III PanK (*P. aeruginosa:* Q9HWC1)	PACG01000088.1 (NAT1)	2E-05	23.2	97
	DPVI01000463.1 (UBA9662)	5E-04	26.6	58
	NZUQ01000044.1 (NAT81)	5E-03	26	58
	QNBQ01000081.1 (B28)	5E-04	28.5	50
PoK (*M. hungatei*: Q2FUB2)	VXXJ01000247.1 (bin40)	6E-27	31.7	89
	PYJA01000004.1 (PCPOR2a)	4E-23	30.7	89
	RKRR01000028.1 (binSS2)	4E-25	30.8	89
	MPMY01000015.1 (bin70)	2E-24	32	91
PS (*E. coli*: Q8X930)	PACG01000104.1 (NAT1)	9E-70	44.7	92
	DPVI01000290.1 (UBA9962)	4E-61	43.7	92
	NZUQ01000037.1 (NAT81)	4E-61	43.7	92
	QNBQ01000082.1 (B28)	3E-72	50.2	91
PPS (*M. hungatei*: Q2FUA9)	VXXJ01000393.1 (bin40)	2E-54	44.1	95
	PYJA01000017.1 (PCPOR2a)	3E-50	43.5	95
	RKRR01000051.1 (binSS2)	5E-51	41.3	95
	MPMY01000029.1 (bin70)	8E-50	48	81

Supp Table 2. Identifiers of Candidatus poribacteria contigs encoding bacterial or archaeal phosphopantothenate biosynthetic enzymes. Presence of PanK, PoK, PS or PPS orthologs within Candidatus poribacteria genomes was established using tBLASTn searches against the WGS database. BLAST values representing the degree of similarity are indicated for each protein. Proteins were extracted and translated from eight different Poribacteria assemblies: four Pelagiporibacteria (NAT1, UBA9662, NAT81, B28) and four Entoporibacteria (bin40, PCPOR2a, binSS2, bin70). Archaeal species are depicted in red.

Chapter 3

1) Introduction to "Development of an integrative tool for gene sets investigation"

To test hypotheses and to answer biological questions rose by comparative transcriptomic or genomic analyzes conducted during my thesis, I was brought to investigate and manage large sets of genes originating from different species. In the recent years, many informatics tools have been developed to analyze and compare such sets resulting from omics studies. Searching and using these tools during my thesis allowed me to realize that it is not always as straightforward. Indeed, such tools are accessible through different platforms (websites, softwares or command-lines), using different methods, and intended to solve different problems. To face such a complex situation, one possible approach is to design new programs adapted to the specific biological questions, and providing better control, understanding and accessibility. However by doing this, each new tool is added to the pool of those already existing, making the situation even more complicated. To solve this conundrum, we wanted to propose an alternative approach based on a combination of pre-existing well-known tools and databases (Uniprot, STRING, and BLAST), implemented with complementary analytical methods. This led us to the design of an original workflow providing a systematic and standardized process to analyze sets of genes, through an intuitive interface accessible online: PROTEDEX. In addition to its application in the following manuscript, PROTEDEX was also successfully applied in the study entitled "Transcriptomic analysis to identify conserved genes and mechanisms involved in stress response and adaptation in vertebrate species" (see Chapter 1).

This project was conducted in collaboration with Victor Loegler, an undergraduate student, who worked on this workflow as part of an internship under my supervision.

The computational protocol designed during this project is submitted for publication as an application note to *Bioinformatics*.

V.L. and L.T. conceived and developed the bioinformatics workflow. V.L. implemented the EVOBLAST module and L.T. developed the remaining part of the bioinformatics protocol, including the TRACE module. This project was possible thanks to the advice of A.L. and F.J. and the work was conducted under their supervision. The manuscript was drafted by V.L., L.T. and A.L.

2) Manuscript III

PROTEDEX: a modular workflow allowing standardized and integrative investigation of gene sets

Victor Loegler[1], Fabrice Jossinet[1], Alain Lescure[1] and Luc Thomès[1]

[1]Architecture et Réactivité de l'ARN, Université de Strasbourg, CNRS, Strasbourg, France.

PROTEDEX: a modular workflow allowing standardized and integrative investigation of gene sets

Victor Loegler[1], Fabrice Jossinet[1], Alain Lescure[1] and Luc Thomès[1]

[1]Architecture et Réactivité de l'ARN, University of Strasbourg, CNRS, Strasbourg, France.

Correspondance:
Luc Thomès,
University of Strasbourg, UPR ARN du CNRS,
IBMC, 2 allée Konrad Roentgen,
67084 Strasbourg, France.
Tel +33 388 41 71 06
l.thomes@ibmc-cnrs.unistra.fr

Abstract

Summary: In the omics era, an increasing number of high-throughput studies result in large lists of proteins or genes and their associated quantitative data. Here we present PROTEDEX, an online workflow aimed at extracting pertinent biological information from these lists. Using a combination of five modules including both new and pre-existing tools, it allows standardized and integrative investigation of gene sets from a unique request. It was designed to perform biological enrichment analyses, network construction, cluster extraction and identification of regulators, as well as prediction of homology relations. Moreover, PROTEDEX can also generate and analyze a generic list of genes by querying Gene Ontology and Reactome nomenclatures from the Uniprot database. As an illustrative example, the usefulness of PROTEDEX is demonstrated by investigating the use of human cholesterol-related genes in *Drosophila melanogaster*.

Availability and implementation: https://protedex.fedcloud.fr

Contact: l.thomes@ibmc-cnrs.unistra.fr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Introduction

With the advent of high-throughput technologies in biology, an increasing number of studies, referred to as "omics" studies, lead to large lists of genes or proteins of interest. Compilation and interpretation of such lists requires dedicated bioinformatics tools, a challenging and time-consuming task considering the profusion of available methods. In many instances, extraction of biological meaning from a gene list starts with computation of Gene Ontology Terms (GO-terms) enrichments (Ashburner *et al.*, 2000). However, even for this simple step, there is no unified approach and the absence of a consensus method leads to a variety of interpretations depending on the algorithms and databases used. Subsequent data mining is even more complex and requires a combination of different analyses, including network analyses, identification of gene regulators and comparative or evolutionary studies. For each of these individual steps, many tools are available such as GOrilla (Eden *et al.*, 2009), ShinyGO (Ge and Jung, 2018) or DAVID (Jiao *et al.*, 2012) for ontological enrichment analysis; BioGRID (Oughtred *et al.*, 2019) or IntAct (Orchard *et al.*, 2014) for network investigations; iRegulon (Janky *et al.*, 2014) or ChEA3 (Keenan *et al.*, 2019) for identification of gene regulators and Inparanoid (Sonnhammer and Östlund, 2015), OrthoMCL (Fischer *et al.*, 2011), OrthoInspector (Nevers *et al.*, 2019) or Ensembl Compara (Vilella *et al.*, 2009) for investigation of homology relations, to cite few examples. Such tools are essential to facilitate the analysis and guide the interpretation of large datasets in a concrete biological framework. Here we present PROTEDEX, a web-based standardized workflow to easily investigate gene lists in an integrative manner following such biological framework. PROTEDEX consists of five independent modules that combine on the same webpage well-known services, such as String (Szklarczyk *et al.*, 2019) protein interaction networks and BLAST sequence database searches (Altschul *et al.*, 1997), together with new tools and methods. Starting from a gene list or a functional query, PROTEDEX allows detection of potential biological functions based on (i) biological enrichment computations; (ii) analysis of the connectivity between the encoded proteins by network constructions and (iii) extraction of protein clusters, as well as identification of key regulators such as transcription factors. Additionally, the data can be interpreted in an evolutionary context by prediction of homology relations. PROTEDEX can also exploit quantitative data, such as differential gene expression datasets, to classify and prioritize the gene list.

Methods

PROTEDEX is a Tornado-based web-service written in Python 3. It allows investigation of gene lists from ten well-annotated reference species covering most studied branches of the tree of life: *Homo sapiens, Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, Gallus gallus, Arabidopsis thaliana, Saccharomyces cerevisiae, Escherichia coli, Staphylococcus aureus* and *Pyrococcus furiosus*. PROTEDEX proposes a modular workflow combining five autonomous and selectable modules: UNIDEX, ENRICH, BIONET, TRACE and EVOBLAST.

UNIDEX is a MySQL database that contains protein annotations retrieved from the reviewed proteins available in Uniprot. This module converts the provided gene names into Uniprot accession identifiers required by EVOBLAST. Alternatively, UNIDEX is also able to retrieve or to filter a gene list based on GO-terms "Biological Process" or Reactome annotations (Jassal *et al.*, 2019).

ENRICH calculates functional enrichments from the gene set using the services provided by the String website API (Szklarczyk *et al.*, 2019). PROTEDEX permits standardized submission to this website using customized parameters.

BIONET reconstructs protein-protein interaction (PPI) networks, again using the services provided by the String website API. In addition, BIONET allows extraction of interaction clusters using the MCODE algorithm (Bader and Hogue, 2003). Assuming that clusters are formed by proteins involved in common biological pathway, functional enrichment is computed for each cluster individually using String.

TRACE is a new method to query data from TRANSFAC (Wingender, 1996), ChEA (Lachmann *et al.*, 2010) and ENCODE (The ENCODE Project Consortium, 2012) databases to models the degree of involvement of transcription factors (TFs) in the control of gene expression within a target list. TRACE is designed to propose an accessible and complementary approach to other tools such as iRegulon (Janky *et al.*, 2014), which predicts TFs involved in gene expression regulation based on the presence of DNA-binding motifs within the promotors of these genes. Briefly, TRACE assigns three values to each known TF corresponding to: “Influence” (ratio of the number of target genes compared to the total number of genes in the list), “Activity” (ratio of the number of target genes in the list compared to the total number of genes controlled by TF), and an associated *P*-value (see Supplementary File S1). ENRICH, BIONET and TRACE can also integrate quantitative data, since they analyze

separately genes associated with positive and negative values, in addition to the overall set.

EVOBLAST is a module relying on the BLAST tool provided by the NCBI. It uses the "Bidirectional Best Hit" method (Tatusov *et al.*, 1997) in real time to identify up-to-date homology relations between the gene list of reference organisms and genes encoded by any target species available at the NCBI. The gene list retrieved from the second BLASTp is scanned using an E-value Fluctuation Analysis (EFA) method (see Supplementary File S2). EVOBLAST annotates genes as (i) "Ortholog" if the results list includes the initial gene in the first position with an E-value below 10^{-10} and an identity score greater than 25%, (ii) "Homolog" if the results list includes the initial gene but not in the first position, with an E-value of the best hit below 10^{-10} and identity score greater than 25% suggesting a paralogous or co-orthologous relation and (iii) "Domain-level" in the remaining cases. "Ortholog" and "Homolog" annotations together indicate the presence of an equivalent protein within the target proteome while the "Domain-level" homology only highlight conservation of short amino-acid regions. Finally, genes are annotated as "None" if the encoded protein has no equivalent in the target, which means that BLASTp results contain only hits with either a BLAST score below 40 or an expect value (E-value) above 10^{-2}, or no hit at all. Running tasks in real time, EVOBLAST does not rely on pre-computed data and allows homology relations investigation in any target species accessible from the NCBI. This strategy solves the limitation of available species and the problem of maintaining databases updated that are encountered by many current tools such as OrthoInspector (Nevers *et al.*, 2019) or OrthoMCL (Fischer *et al.*, 2011), but consequently reduces EVOBLAST processing speed, making it completely dependent on the NCBI server status.

For each module, results are retrieved as text files that can be visualized with dedicated software, *e.g.* Cytoscape (Shannon, 2003) for network results visualisation.

Usage

As a case study, we investigated the presence of cholesterol-related human genes equivalent in *Drosophila melanogaster*. This study is designed to analyze the possible function of cholesterol-related genes in flies, organisms that cannot synthesize cholesterol (see Supplementary File S3). The results obtained using PROTEDEX confirmed previously published information showing that cholesterol-related genes are either absent or reoriented to new metabolic pathways (Seegmiller *et al.,* 2002; Rawson, 2003).

Conclusion

PROTEDEX is a web-based workflow implemented to provide a standardized and intuitive protocol to analyze gene sets in a biologically relevant framework. It allows investigation of biological processes, protein networks, gene regulators and homology relations from a single website, optionally including quantitative data. Each module can be independently selected with adjustable parameters: statistical thresholds, PPI confidence level or target group for the homology analysis.

Acknowledgements

We are indebted to Odile Lecompte for fruitful discussions about the results of this study. We are grateful to Julie Thompson for critical reading of the manuscript.

Funding: This work was supported by funding of CNRS and ADISSEO FRANCE SAS to AL, and a CIFRE fellowship from the Association Nationale Recherche et Technologie to LT.

Conflict of Interest: none declared.

References

Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389-3402.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**, 25–29.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 27.

Eden,E. et al. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Fischer,S. *et al.* (2011) Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. In, Goodsell,D.S. (ed), *Curr. Protoc. in Bioinform.*, 35, 6.12.1-6.12.19

Ge,S.X. and Jung,D. (2018) ShinyGO: a graphical enrichment tool for animals and plants. *Bioinformatics*, **36**, 2628-2629

Jassal,B. *et al.* (2019) The reactome pathway knowledgebase. *Nucl. Acids Res.*, gkz1031.

Jiao,X. *et al.* (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.

Keenan,A.B. *et al.* (2019) ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucl. Acids Res.*, **47**, W212–W224.

Lachmann,A. *et al.* (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.

Orchard,S. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucl. Acids Res.*, **42**, D358–D363.

Oughtred,R. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, **47**, D529–D541.

Rawson,R.B. (2003) The SREBP pathway — insights from insigs and insects. *Nat Rev Mol Cell Biol*, **4**, 631–640.

Seegmiller,A.C. *et al.* (2002) The SREBP Pathway in Drosophila: Regulation by Palmitate, Not Sterols. *Developmental Cell*, **2**, 229-238.

Shannon,P. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, **13**, 2498–2504.

Sonnhammer,E.L.L. and Östlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucl. Acids Res.*, **43**, D234–D239.

Szklarczyk,D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucl. Acids Res.*, **47**, D607–D613.

Tatusov,R.L. (1997) A Genomic Perspective on Protein Families. *Science*, **278**, 631–637.

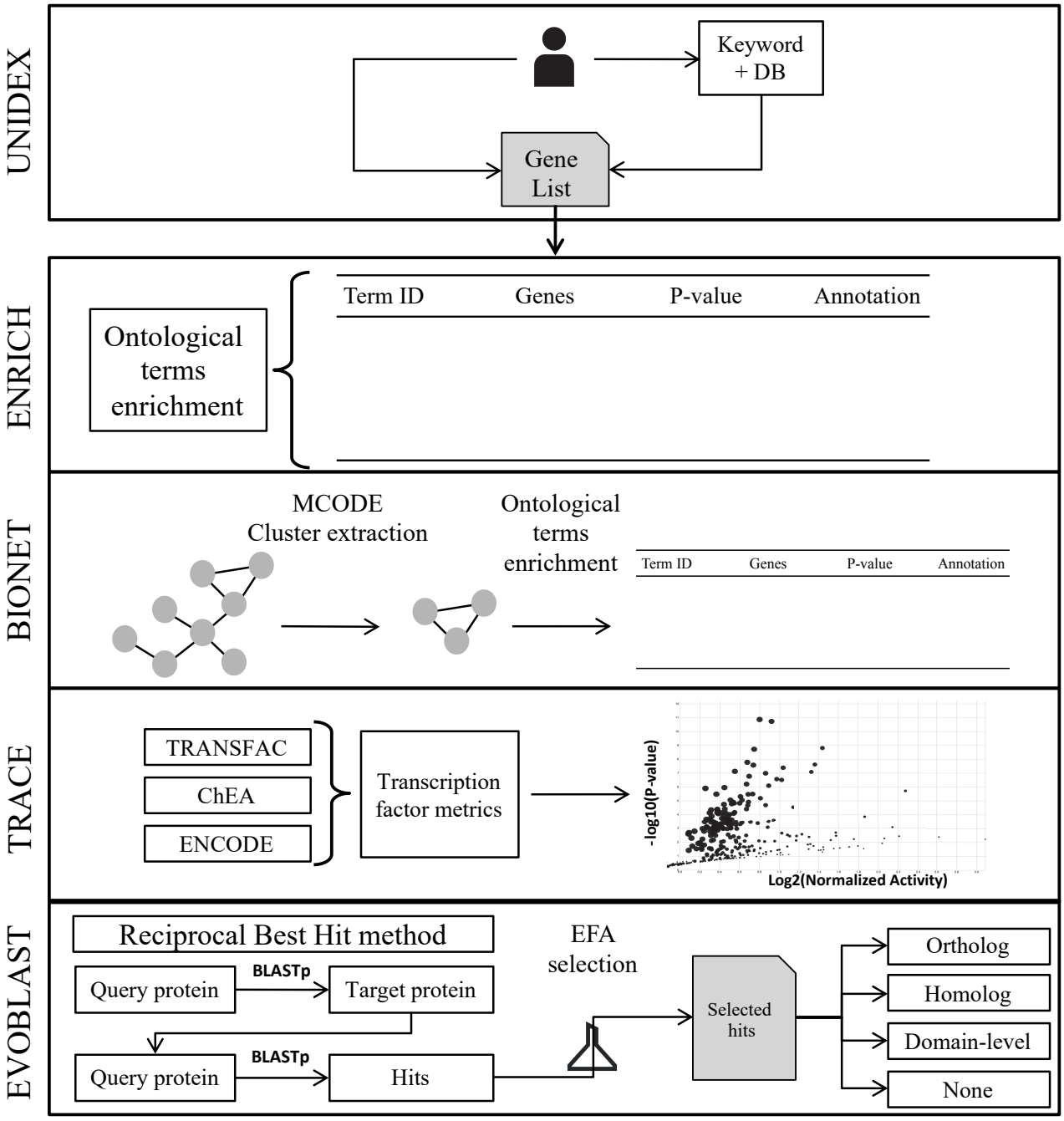
The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Vilella,A.J. *et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, **19**, 327–335.

Wingender,E. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucl. Acids Res.*, **24**, 238–241.

Figure Legend

Fig. 1. General workflow of PROTEDEX. Starting from a gene list, the PROTEDEX protocol includes five modules that compute and integrate ontological terms enrichment, a protein interaction network and its highly interconnected nodes, prediction of important gene regulators, and homology annotations.


UNIDEX
Keyword
+ DB
Gene
List
ENRICH
Ontological
terms
enrichment
Term ID
Genes
P-value
Annotation
BIONET
MCODE
Cluster extraction
Ontological
terms
enrichment
Term ID
Genes
P-value
Annotation
TRACE
TRANSFAC
ChEA
ENCODE
Transcription
factor metrics
- log10(P-value)
Log2(Normalized Activity)
EVOBLAST
Reciprocal Best Hit method
Query protein
BLASTp
Target protein
Query protein
BLASTp
Hits
EFA
selection
Selected
hits
Ortholog
Homolog
Domain-level
None

Supplementary Information

Suppl. File S1: TRACE, a method to identify transcription factors involved in the expression control of a gene set.

Suppl. File S2: E-value Fluctuation Analysis (EFA), a method used by EVOBLAST to annotate homology relations between a tested and a reference organism.

Suppl. File S3: Case study in *Drosophila melanogaster*

Suppl. File S1: TRACE, a method to identify transcription factors involved in the expression control of a gene set.

TRACE scores computation

To identify transcription factors (TFs) potentially controlling the expression of genes from a list, we developed a method called TRACE. The central idea of the TRACE module is to compute two values for each TF, namely "Activity" and "Influence", in addition to a statistical value.

Activity:

"Activity" describes the contribution of a TF in the regulation of a set of genes compared to the full list of reported genes controlled by this factor within the databases TRANSFAC (Wingender, 1996), ChEA (Lachmann *et al.*, 2010) and ENCODE (The ENCODE Project Consortium, 2012). Raw "Activity" is computed as follows:

$$\boldsymbol{Activity} = \left(\frac{\boldsymbol{Number\ of\ TF\ target\ genes\ in\ the\ list}}{\boldsymbol{Total\ number\ of\ target\ genes\ for\ this\ TF}}\right) \times \mathbf{100}$$

Influence:

"Influence" estimates the impact of a defined TF regarding the set of genes of interest. It is computed as follows:

$$\boldsymbol{Influence} = \left(\frac{\boldsymbol{Number\ of\ TF\ target\ genes\ in\ the\ list}}{\boldsymbol{Full\ number\ of\ genes\ in\ the\ list}}\right) \times \mathbf{100}$$

Normalization factor:

The significance of a TF's "Activity" is highly dependent on the total number of target genes for this TF. Indeed, a TF with few target genes will more easily reach maximum "Activity" for a large list of genes of interest without necessarily having biological meaning. To evaluate this bias, it is necessary to compute a "Normalized Activity" corresponding to the ratio between the "Calculated Activity" and an "Expected Activity". To calculate this "Expected Activity", we determined a normalization factor.

Taken together, the TRANSFAC (Wingender, 1996), ChEA (Lachmann *et al.*, 2010) and ENCODE (The ENCODE Project Consortium, 2012) databases cover 23,595 genes, which is considered to be the full human genome capacity. For each experiment, the normalization factor is then a constant computed as follows:

$$\boldsymbol{Normalization\ factor} = \frac{\boldsymbol{Number\ of\ genes\ in\ the\ list\ of\ interest}}{\boldsymbol{Number\ of\ genes\ in\ the\ database(s)}}$$

The factor corresponds to the fraction of the genome that is represented in the list of genes. If this list is equal to the full genome, all TF "Activities" are then expected to reach 100%. Accordingly, if the list only corresponds to a fraction of the genome, the "Expected Activity" is determined as follows, with the "Theoretical Maximal Activity" equal to 100:

$$\boldsymbol{Expected\ Activity} = \boldsymbol{Normalization\ factor} \times \boldsymbol{Theoretical\ Maximal\ Activity}$$

From the "Expected Activity", a "Normalized Activity" is calculated for each TF by computing the following ratio:

$$\boldsymbol{Normalized\ Activity} = \frac{\boldsymbol{Activity}}{\boldsymbol{Expected\ Activity}}$$

Note that PROTEDEX allows users to select data from TRANSFAC, ChEA and ENCODE databases together or independently. The normalization factor is modified accordingly.

Statistical value:

For each TF, a *P*-value is computed according to the hypergeometric law using a Python command line:

_P_-value = hypergeom.sf(gene_activated, Full_Controllable_Genes, gene_activable , n)

Here, "gene_activated" corresponds to the number of target genes for one TF in a given list. The "Full_Controllable_Genes" variable corresponds to the total number

of target genes controlled by any TFs within the database(s). The "gene_activable" variable corresponds to the number of all reported genes controlled by one TF within the databases. Finally, the variable "n" corresponds to the number of genes within the list of interest that is part of the "Full_Controllable_Genes" list. The *P*-value estimates the probability, for each TF, to regulate "gene_activated" over "gene_activable" genes after drawing "n" genes without replacement in a set containing "Full_Controllable_Genes" genes. If the *P*-value equals 0, it is replaced by the lowest *P*-value observed to the power of 1.2, as the output file calculates the -log10(*P*-value). The value of 1.2 was selected to facilitate graphical representation.

Evaluation of TRACE prediction for transcription regulators of cholesterol-related genes and comparison with the iRegulon package.

TRACE generates multiple metrics used to visualize the contribution of candidate regulators controlling a set of genes (see above). A volcano plot is generated representing the log2(Normalized Activity) as a function of the -log10(*P*-value) for each factor according to selected thresholds (for example, see Fig. Supp. 1). Here we conducted a TRACE analysis to predict key regulators involved in the control of 66 human genes related to cholesterol metabolism (for detailed analysis see Supplementary file S3). The results were compared with the predictions of the iRegulon application (Janky *et al.*, 2014) available in the Cytoscape software (Shannon, 2003). iRegulon predicts TFs involved in the expression control of a set of genes, based on the presence of related DNA-binding motifs within the promotors of the genes. It generates a table of the top TF candidates from the most relevant group of TFs associated with a DNA-binding motif.

To identify top candidate factors using TRACE, we set the -log10(*P*-value) threshold to 5, and the log2(Normalized Activity) threshold to 1. With these thresholds, the most likely TF candidates predicted by TRACE are: SREBF2, SREBF1, IRF3, HNF4G, NFYA, MBD4 and NR1I2 (Fig. Supp. 1). In parallel, iRegulon identified 26 top TF candidates, each representative of a group, and sorted them by an enrichment score, called "NES", as shown in Fig. Supp. 2.

Two TFs were predicted by both approaches: SREBF1 and SREBF2. Interestingly, these TFs corresponded to the two best candidates detected by the two methods. In addition, SREBF1 and SREBF2 are known to be the main regulators of cholesterol metabolism in humans (Espenshade and Hughes, 2007). NFYA was identified by TRACE as a potential TF, while iRegulon predicted two related TFs, NFYB and NFYC. To determine the relevance of the results obtained by the two methods, we performed bibliographical investigations using Uniprot and Pubmed to search for direct or indirect relations between the predicted TFs and cholesterol metabolism regulation. Based on these investigations, all seven TFs predicted by TRACE were known to be directly or indirectly related to cholesterol homeostasis regulation. Concerning iRegulon, nine of the 26 candidates predicted were previously shown to be directly or indirectly involved in this process (34.6%): SREBF1, SREBF2, NFYB, YBX1, SP1, GLI2, DBP, HSF1 and NFYC.

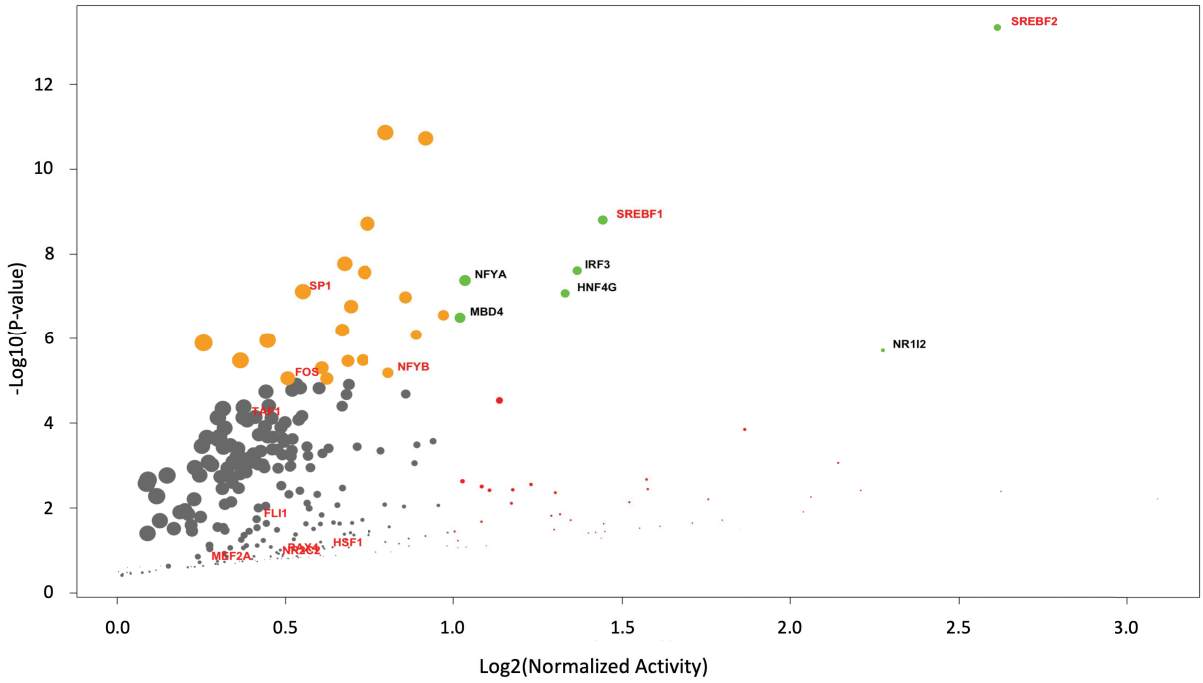


Fig. Supp. 1. Volcano plot representing TRACE results for the analysis of 66 human genes related to cholesterol metabolism. Each dot corresponds to one transcription factor (TF) and is plotted according to its log2(Normalized Activity) and its -log10(P-value) values. The size of the dots corresponds to their respective TF "Influence" value. TFs with a log2(Normalized Activity) >=1 are depicted in red. TFs having a -log10(P-value) >=5 are represented in orange. TFs corresponding to most significant candidates (log2(Normalized Activity) >=1 and a -log10(P-value) >=5) are represented

in green and tagged with their respective gene names. Gene names of the significant TF candidates identified by iRegulon are highlighted in red.

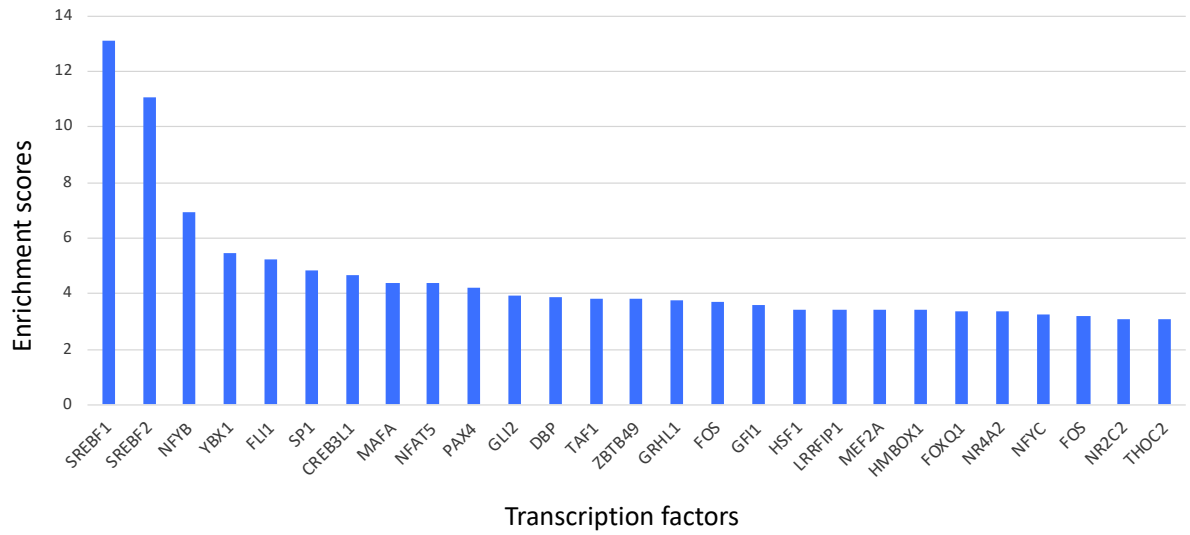


Fig. Supp. 2. Candidate transcription factors predicted by iRegulon for the analysis of 66 human genes related to cholesterol metabolism. The 26 transcription factors (TFs) detected are sorted according to their "NES" enrichment scores. SREBF1 and SREBF2, the two major regulators of cholesterol synthesis in cells, are the two most enriched TFs predicted using this method.

Suppl. File S2: E-value Fluctuation Analysis (EFA), a method used by EVOBLAST to annotate homology relations between a tested and a reference organism.

Computation of E-value Fluctuation Analysis (EFA)

In EVOBLAST, a protein is annotated as "Ortholog" when the best hit found in a bidirectional BLASTp search (Altschul *et al.*, 1997) is identical to the protein query. In contrast, the homology relation is described as “None” when results of the initial BLASTp are not statistically significant. However it exists other homology relations without reciprocal identity of the BLASTp best hits, such as co-orthologous or non-orthologous proteins arising from gene duplication, and block of conserved regions due to divergent evolution and genetic rearrangements. To annotate these homolog groups, existing methods have been developed (Tatusov *et al.*, 1997; Linard *et al.*, 2011) that require multiple BLAST queries, a precise but time-consuming protocol difficult to apply for large datasets. To circumvent this limitation, we developed the E-value Fluctuation Analysis (EFA) method to identify a potential homolog group from a single bidirectional BLASTp, allowing to perform this investigation in real-time at each EVOBLAST request.

Homolog proteins or proteins containing resembling domains should share sequence similarities and thus have comparable E-values in a BLASTp result. The EFA method was designed to discriminate different homolog groups or domains based on the E-value score. The method takes as input the results of the reciprocal BLASTp performed by EVOBLAST for the Bidirectional Best Hit method. The EFA method identifies a most significant E-value fluctuation between two consecutive hits within the BLASTp results. To quantify this increase, an R factor is computed for each hit "n" as:

$$\boldsymbol{R(n) = log2\left(\frac{E-value(n)}{E-value(n-1)}\right)}$$

As the R factor cannot be computed for the first hit (n=1) or for a null E-value at the denominator, it is assigned a null value. Using the R(n) values, EVOBLAST extracts all protein hits above the highest R factor, if the E-value is less than 10^{-2}. If the protein initially queried is found in the extracted protein list, but is not the best hit of the

reciprocal BLASTp, the homology relation will be annotated as either "Homolog" or "Domain-level" given E-value and identity scores.

To demonstrate the predictive capacity of the EFA method, we used the case of the KCTD family proteins using KCTD12 as query. This protein is known to be part of a closer subgroup sharing a higher inner similarity within this family and consisting of KCTD8, KCTD12 and KCTD16 (Liu *et al.*, 2013). Table Supp. 1 and Fig. Supp. 3 show results of the EFA method applied on a BLASTp search output against *Homo sapiens* using the human KCTD12 protein (KCD12_HUMAN) as query. Out of the 24 proteins in the BLASTp results, the EFA method allowed to identify a group of three proteins sharing significantly high similarity: KCTD12, KCTD16 and KCTD8. These results are in agreement with the current knowledge about this family, confirming the efficiency of the EFA method.

Hit number	Protein name	E-value	R factor
1	BTB/POZ domain-containing protein KCTD12	0E+00	0.00
2	BTB/POZ domain-containing protein KCTD16	2E-107	0.00
3	BTB/POZ domain-containing protein KCTD8	6E-105	5.70
4	BTB/POZ domain-containing protein KCTD21	1E-19	196.23
5	BTB/POZ domain-containing protein KCTD15	1E-16	6.91
6	Potassium channel regulatory protein	4E-16	1.39
7	BTB/POZ domain-containing protein KCTD6	3E-15	2.01
8	BTB/POZ domain-containing protein KCTD4	7E-15	0.85
9	BTB/POZ domain-containing protein KCTD1	3E-14	1.46
10	BTB/POZ domain-containing protein KCTD18	9E-14	1.10
11	BTB/POZ domain-containing protein KCTD2	8E-11	6.79
12	BTB/POZ domain-containing protein KCTD14	9E-11	0.12
13	BTB/POZ domain-containing protein KCTD7	5E-10	1.71
14	BTB/POZ domain-containing protein KCTD9	6E-10	0.18
15	BTB/POZ domain-containing protein KCTD3	3E-09	1.61
16	BTB/POZ domain-containing protein KCTD5	1E-08	1.20
17	SH3KBP1-binding protein 1	3E-08	1.10
18	BTB/POZ domain-containing protein KCTD19	7E-06	5.45
19	BTB/POZ domain-containing protein KCTD11	3E-05	1.46
20	Potassium voltage-gated channel subfamily D member 1	3E-03	4.61
21	Potassium voltage-gated channel subfamily C member 3	0.04	2.64
22	Potassium voltage-gated channel subfamily V member 1	0.1	1.05
23	Potassium voltage-gated channel subfamily D member 2	0.2	0.41
24	Signal-induced proliferation-associated 1-like protein 1	9.5	3.97

Table Supp. 1. Results of the BLASTp request on *Homo sapiens* using KCD12_HUMAN as query. Performing the EFA method on these results allowed to obtain a maximum R factor value for hit 4, identifying the three first proteins as a homolog group (highlighted in green).

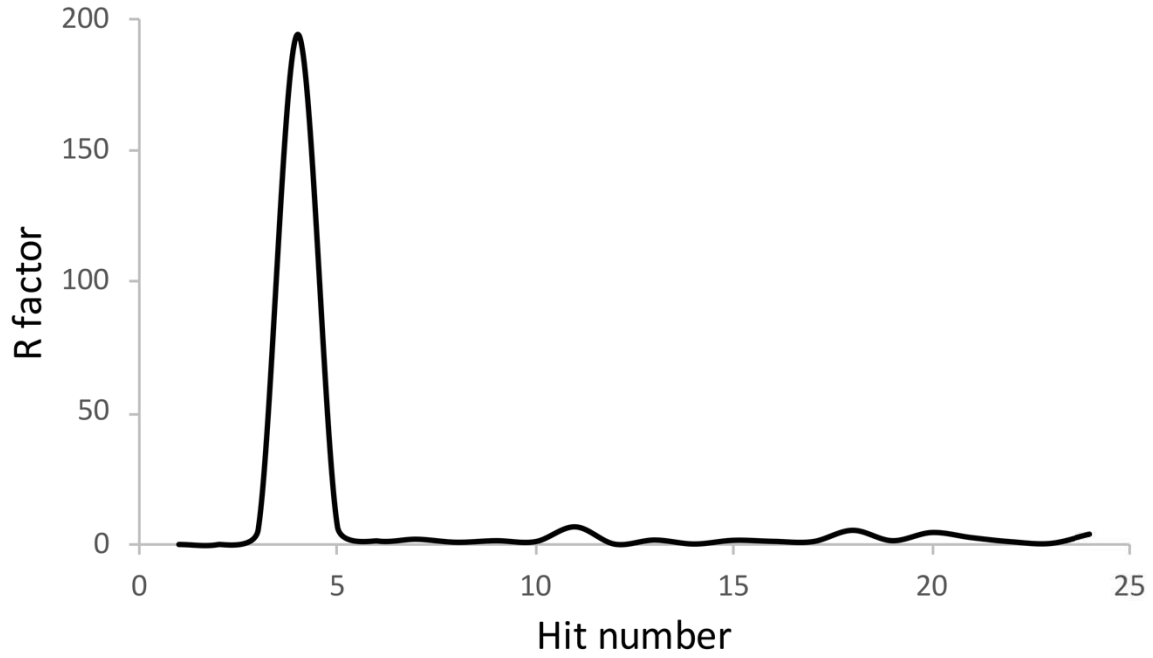


Fig. Supp. 3. Graphical representation of the computed R factor values corresponding to each KCTD12 BLASTp hit. Maximum R factor value is reached for hit 4. Hits located above the maximum R factor peak correspond to a known group of highly similar proteins within the KCTD family. Hits located below the maximum R factor peak correspond to other KCTD groups and unrelated proteins.

As an example of the use of the EFA method within the EVOBLAST pipeline, Table Supp. 2 and Fig. Supp. 4 show results obtained from a request using ANXA1_HUMAN as the gene query and *Caenorhabditis elegans* as the target organism. ANXA1 codes for a protein from the annexin family, a large group consisting of twelve genes within the human genome (Rescher and Gerke, 2004). As expected, the EFA method allowed to discriminate annexins and annexin-related proteins from the other ones. Since ANXA1 is part of the identified annexin group but is not the best hit, EVOBLAST will annotate this gene as "Homolog".

Hit number	Protein name	E-value	R factor
1	Annexin A7	7E-111	0.00
2	Annexin A11	1E-105	11.87
3	Annexin A4	1E-95	23.03
4	Annexin A5	7E-92	8.85
5	Annexin A6	8E-92	0.13
6	Annexin A13	2E-88	7.82
7	Annexin A8-like protein 1	1E-87	1.61
8	Annexin A8	4E-87	1.39
9	Annexin A3	4E-86	2.30
10	Annexin A1	3E-81	11.23
11	Annexin A10	4E-73	18.71
12	Annexin A2	5E-72	2.53
13	Putative annexin	4E-69	6.68
14	Annexin A9	3E-48	48.07
15	Protein lifeguard 1	3E-05	99.01
16	Homeobox protein Hox-B7	0.18	8.70
17	Galectin-3	1.9	2.36
18	Programmed cell death 6-interacting protein	3.1	0.49
19	M-phase-specific PLK1-interacting protein	7.9	0.94

Table Supp. 2. Results of the reciprocal BLASTp of an EVOBLAST request using ANXA1_HUMAN as gene query and *Caenorhabditis elegans* as target organism. Hits 1 to 14 correspond to proteins from the annexin family or related proteins. Hits 15 to 19 correspond to proteins belonging to unrelated families. Maximum R factor value is achieved for hit 15, and therefore EVOBLAST retains proteins corresponding to hits 1 to 14. Since ANXA1 is not the first hit in this list, but it is present in hits 1 to 14, the homology relation will be annotated as "Homolog" by EVOBLAST.

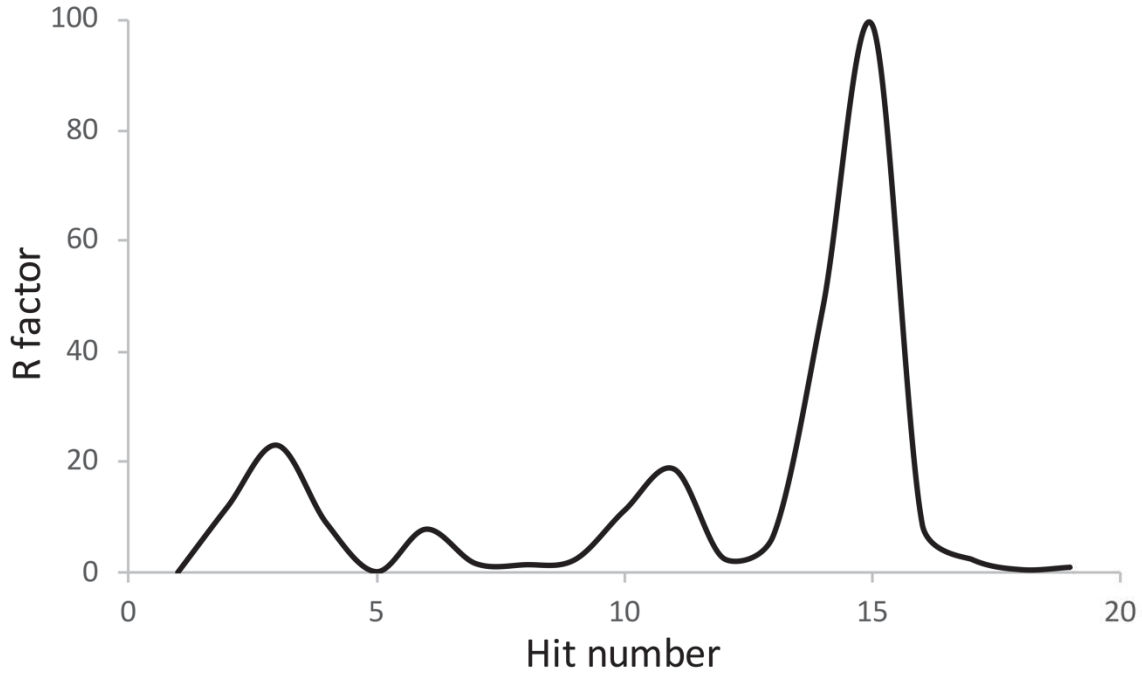


Fig. Supp. 4. Graphical representation of the computed R factor values corresponding to each BLASTp hit. Maximum R factor value is reached for hit 15. Hits located above the maximum R factor peak correspond to the annexin family proteins or related proteins. Hits located below the maximum R factor peak correspond to unrelated proteins.

Evaluation of EVOBLAST prediction of homology relations for a set of human cholesterol-related proteins and comparison with DIOPT.

DIOPT (Hu *et al.*, 2011) is a tool available on the Alliance for Genome References website (The Alliance of Genome Resources Consortium, 2020) that allows orthologs inference based on several reference approaches mostly benchmarked by the Quest for Orthologs (QfO) consortium (Gabaldón *et al.*, 2009). We performed a comparative analysis using a set of 66 human proteins related to cholesterol metabolism to search for *Drosophila melanogaster* homologous proteins (for further analysis see Supplementary file S3). EVOBLAST annotates homology relations as "Ortholog", "Homolog" (excluding orthology relations), "Domain-level" or "None". In DIOPT, proteins from the target organism are associated with a weighted score representing the number of tools that identified this protein as ortholog of the query protein (between 0 and 14.75 for *D. melanogaster*) and a rank related to the orthology prediction confidence level: "None", "low", "moderate" or "high".

Using EVOBLAST to annotate the 66 human cholesterol-related proteins versus the *Drosophila melanogaster* proteome, we identified 26 "Ortholog", 17 "Homolog", 14 "Domain-level" and 9 "None" proteins. DIOPT generated a list of potential *D. melanogaster* orthologs for each human protein. Considering the most significative predicted orthologs based on their weighted score values, DIOPT identified 31 "high", 19 "moderate", 11 "low" and 5 "None" proteins. Results are detailed in Table Supp. 3.

1. To compare DIOPT and EVOBLAST annotations, we assumed that EVOBLAST "Ortholog" and "Homolog" annotations were equivalent to DIOPT "high" and "moderate" annotations respectively, indicative of the presence of an homologous protein. Interestingly, all human proteins annotated as "moderate" or "high" and associated with a weighted value greater than five by DIOPT are also predicted as "Homolog" or "Ortholog" of *D. melanogaster* proteins by EVOBLAST. After establishing this correspondence, DIOPT predicted 50 potential homology relations, including the 43 proteins annotated by EVOBLAST as "Ortholog" or "Homolog". From this set, the potential homologous proteins identified by EVOBLAST and DIOPT are identical in 90.7% of cases. However, for some proteins, EVOBLAST and DIOPT predictions appeared more divergent. These divergences can be classified in three general cases: EVOBLAST predicts “Domain-level” restricted homology while DIOPT results suggest protein presence in the target organism (CYP46A1, HSD17B7, MSMO1, NR1H4, DHCR7, NCOA1 and NCOA2). In most of these cases, the *D. melanogaster* protein identified by DIOPT was with a low weighted score value, indicating that only few of the orthology prediction tools implemented in DIOPT effectively detected this protein as ortholog. Finally, concerning the NCOA1 and NCOA2 proteins, EVOBLAST and DIOPT made divergent predictions while associated with better weighted score values. This observation suggests that EVOBLAST failed to detect the presence of NCOA1 and NCOA2 homologs in *D. melanogaster* based on sequences similarity. However, the sequence similarity of NCOA1/NCOA2 with their respective *D. melanogaster* homolog taiman (predicted by DIOPT) or their "Domain-level" equivalents tango and spineless respectively (predicted by EVOBLAST) is limited to the N-terminal domains bHLH

(basic helix-loop-helix) and PAS (Per-Arnt-Sim) only, in agreement with the EVOBLAST "Domain-level" annotation.

2. EVOBLAST and DIOPT annotations are convergent for homology detection, but the protein predicted in EVOBLAST was not the best one identified by DIOPT (SLC27A5, AKR1C1, AKR1C4 and ACOX2). Of note, the proteins predicted as homolog by EVOBLAST were also identified by DIOPT but with slightly lower weighted values: 3.93 for SLC27A5, 4.92 for AKR1C1 and AKR1C4, 6.84 for ACOX2.
3. In the last cases, EVOBLAST predicted a different target protein as potential equivalent of the human protein compared to DIOPT in cases of "Domain-level" and "None" annotations. However, since the DIOPT score value obtained for these genes was below one, we assumed that the divergence of the potential orthologous protein predicted can be neglected.

Altogether, these comparisons indicate that EVOBLAST predictions are mostly consistent with reference orthology inference tools as divergent predictions mainly concern "None" or "Domain-level" annotations.

Human Gene	PROTEDEX - EVOBLAST Annotations	DIOPT Rank	DIOPT Weighted Score	Identical protein predicted ?
FDPS	Ortholog	high	14.75	Yes
HMGCS1	Ortholog	high	14.75	Yes
HSD17B4	Ortholog	high	14.75	Yes
KPNB1	Ortholog	high	14.75	Yes
MBTPS2	Ortholog	high	14.75	Yes
PMVK	Ortholog	high	13.85	Yes
PLPP6	Ortholog	high	13.8	Yes
SCP2	Ortholog	high	13.8	Yes
SEC23A	Ortholog	high	13.8	Yes
ACAT2	Ortholog	high	13.74	Yes
GGPS1	Ortholog	high	13.72	Yes
MBTPS1	Ortholog	high	13.72	Yes
MVD	Ortholog	high	13.72	Yes
SCAP	Ortholog	high	12.89	Yes
RAN	Ortholog	high	12.87	Yes
SREBF2	Ortholog	high	12.82	Yes
SEC24C	Ortholog	high	12.81	Yes
LBR	Ortholog	high	12.79	Yes
HMGCR	Ortholog	high	12.77	Yes
RXRA	Ortholog	high	12.77	Yes
AMACR	Ortholog	high	12.7	Yes
SAR1B	Ortholog	high	11.96	Yes
SEC24A	Homolog	high	11.84	Yes
SREBF1	Homolog	moderate	11.84	Yes
SEC24B	Ortholog	high	11.79	Yes
SEC24D	Homolog	moderate	11.79	Yes
IDI1	Ortholog	high	11.76	Yes
IDI2	Homolog	moderate	10.87	Yes
ARV1	Ortholog	high	10.79	Yes
ABCB11	Homolog	moderate	9.91	Yes
MVK	Ortholog	high	9.89	Yes
TM7SF2	Homolog	moderate	9.78	Yes
ACOX2	Homolog	high	8.8	No
HSD3B7	Homolog	moderate	7.88	Yes
AKR1D1	Homolog	moderate	7.85	Yes
SLC27A2	Homolog	moderate	5.84	Yes
CYP27A1	Homolog	moderate	5.8	Yes
NCOA1	Domain-level	high	4.98	No
NCOA2	Domain-level	high	4.98	No
AKR1C1	Homolog	moderate	4.96	No
AKR1C4	Homolog	moderate	4.96	No
SLC27A5	Homolog	moderate	4.94	No
AKR1C2	Homolog	moderate	4.92	Yes
AKR1C3	Homolog	moderate	4.92	Yes
DHCR7	Domain-level	moderate	4.91	Yes
NR1H4	Domain-level	moderate	3.02	Yes
NSDHL	Homolog	moderate	2.88	Yes
MSMO1	Domain-level	moderate	2.83	No
HSD17B7	Domain-level	moderate	2.04	Yes
CYP46A1	Domain-level	high	1.91	No
CYP39A1	Domain-level	low	0.95	Yes
CYP51A1	Domain-level	low	0.95	No
CYP7A1	Domain-level	low	0.95	Yes
CYP7B1	Domain-level	low	0.95	Yes
CYP8B1	Domain-level	low	0.95	No
PTGIS	Domain-level	low	0.95	No
BAAT	None	low	0.9	No
DHCR24	None	low	0.9	No
FDFT1	None	low	0.9	No
SC5D	Domain-level	low	0.9	Yes
SQLE	None	low	0.9	No
ACOT8	None	None	0	NA
EBP	None	None	0	NA
INSIG1	None	None	0	NA
INSIG2	None	None	0	NA
LSS	None	None	0	NA

Table Supp. 3. Homology relations between 66 human proteins related to cholesterol metabolism and *D. melanogaster* equivalents. Results are sorted by "DIOPT Weighted Score" values. Corresponding EVOBLAST and DIOPT annotations are highlighted using the same color: orthology or

high confidence are blue; other homology relations or moderate confidence are green; domain-level homology or low confidence are orange; protein absence or absence of prediction are grey. Identity of the EVOBLAST and the DIOPT top predicted orthologs is specified by "Yes", "No" or by "NA" in absence of ortholog predictions. In 53 cases out of 66, EVOBLAST and DIOPT share equivalent annotations.

Suppl. File S3: Case study in Drosophila melanogaster

Introduction

To evaluate the usefulness of the PROTEDEX workflow, we investigated the presence of cholesterol-related human gene homologs in *Drosophila melanogaster.* It is well-known that, in contrast to most animals, insects are auxotroph for cholesterol. However, they use this molecule, relying on cholesterol supply from food. On the basis of this knowledge, presence of genes related to cholesterol synthesis in flies is unexpected. We used PROTEDEX and a set of human protein-coding genes related to cholesterol metabolism in order to assess their conservation and to examine their possible biological functions in the insect *Drosophila melanogaster*.

Material and method

To perform this analysis, a list of 66 genes of interest was extracted from the human genome by textual query of the Reactome annotations (Jassal *et al*. 2019) using the UNIDEX module and the keyword "cholesterol". To validate that the genes of the list are indeed involved in cholesterol-related pathways, we performed an enrichment analysis using the PROTEDEX ENRICH module with a statistical threshold of 0.001. To identify functionally interconnected proteins, the PROTEDEX BIONET module was used to build a network with the 66 proteins of the list, keeping default parameters. Next we used the PROTEDEX TRACE module with default method and parameters to identify key regulators, including factors controlling the synthesis of sterol-related molecules. Finally, we used the PROTEDEX EVOBLAST module to identify protein homologs in the insect *Drosophila melanogaster* proteome and to map their ontological relations.

Results

The PROTEDEX workflow was performed by applying the different modules successively. Using the UNIDEX module, we retrieved a list of 66 human genes related to "cholesterol" according to Reactome annotations. These genes and their products are thereafter referred as "cholesterol-related proteins". As expected, computing biological enrichments with the ENRICH module generated a list of ontological annotations referred as "biological process" mostly related to cholesterol metabolism.

The top 10 enrichment terms retrieved are shown on Supp. Table 4. Using the BIONET module, we generated an interaction network including 64 interconnected proteins out of the 66 candidates. Inspection of this network highlighted its organization into six clusters containing 3 to 23 proteins (Fig. Supp. 5). For each cluster, an ontological-terms enrichment analysis was performed and the results are summarized in Table Supp. 5. This approach defined three subgroups within the network: clusters 1 and 2 were associated to sterol biosynthesis or metabolism; clusters 3, 4 and 6 were linked to bile acids, bile salts and oxysterols synthesis; and cluster 5 was related to regulation of cholesterol synthesis annotations. The TRACE module allowed the identification of a set of candidate factors involved in expression control of the genes coding for the cholesterol-related proteins of our list. Obtained results were visualized on a volcano plot representation displayed on Fig. Supp. 6. In agreement with the current knowledge, these results stress the importance of the transcription factors SREBF2 and SREBF1 as two key regulators of cholesterol-related genes expression. Based on our analysis, other transcription factors, such as HNF4G or IRF3 were also predicted to be involved in sterol synthesis control. Interestingly, a publication from Castrillo *et al.* (2003) established a connection between the IRF3 factor activity and the control of cholesterol homeostasis mediated by the transcription factor LXR. Indeed, LXR is involved in regulation of cholesterol absorption, transport and elimination and was shown to be inhibited by IRF3 through activation of the TLR (toll-like receptor) signalling pathway and competition for a common transcriptional co-activator: p300/CBP. Finally, EVOBLAST results showed that among the 66 proteins used as query, 43 homology relations were identified in the target species and classified as 26 orthologs and 17 other homologs (Table Supp. 6). For the 23 remaining proteins, no direct homologs were inferred as the proteins annotations were annotated "Domain-level" only (14) or "None" (9) and therefore, they were considered as absent from the *D. melanogaster* proteome.

Combining the clusters generated by the BIONET module and the homology relations analyses, the percentage of proteins conserved between *D. melanogaster* and *Homo sapiens* within each cluster was calculated (represented on Fig. Supp. 7). By integrating the enrichment data, we also computed this percentage according to the ontological annotations referred as "biological process": sterol biosynthesis (51.7%);

bile acids, bile salts and oxysterols synthesis (61.1%); regulation of cholesterol synthesis (88.9%). This result indicated that even if proteins linked to sterol metabolism are partially represented in *Drosophila melanogaster* proteome (51.7 and 61.1%), proteins related to regulation of sterol synthesis are more conserved in the insect (88.9%).

Discussion

Using the PROTEDEX protocol, we showed that cholesterol-related proteins are only partially conserved in *Drosophila melanogaster* compared to human. Indeed, it is known that insects, such as the fly, are unable to synthetize sterols and they rely on cholesterol from nutrition. However, although enzymes involved in sterol biosynthesis (cluster 1 and 2) are largely absent compared to human, proteins involved in cholesterol synthesis control (cluster 5) are mostly conserved in the fly, including the ortholog of the two human paralogous proteins SREBF1 and SREBF2. In mammalian cells, these transcription factors are activated by a mechanism based on sterol sensing within the endoplasmic reticulum membrane. When sterol concentration drops, the active part of SREBF1 and SREBF2 is translocated inside the nucleus to promote expression of genes coding the enzymes involved in cholesterol synthesis (Espenshade and Hughes, 2007). In *D. melanogaster*, one possible explanation for SREBF conservation is that, even if insects do not need a control mechanism to regulate sterol synthesis, a sterol sensor is still required to regulate food absorption or synthesis of steroid hormones, such as ecdysteroids, which control insect moulting. The case of SREBF conservation in this organism sheds light on this matter, as it was demonstrated that *D. melanogaster* SREBF responds to the intracellular fatty-acid content instead of cholesterol, indicating an evolution of its function and an adaptation to the absence of cholesterol synthesis (Seegmiller *et al.,* 2002; Rawson, 2003).

Conclusion

In conclusion, we developed a workflow combining different tools for analysing a set of genes, PROTEDEX. Conducting a PROTEDEX analysis on human cholesterol-related proteins, we showed their selective conservation in the *Drosophila*

melanogaster proteome, suggesting an adaptive mechanism, consistent with current knowledge.

Term ID	Statistical value	Enriched term
HSA-8957322	1.09E-122	*Metabolism of steroids*
GO.0008202	1.91E-87	*steroid metabolic process*
HSA-556833	3.85E-86	*Metabolism of lipids*
GO.0006694	8.56E-79	*steroid biosynthetic process*
GO.1901615	2.27E-76	*organic hydroxy compound metabolic process*
GO.1901617	7.02E-75	*organic hydroxy compound biosynthetic process*
GO.0016125	4.20E-65	*sterol metabolic process*
GO.0008203	6.74E-63	*cholesterol metabolic process*
GO.0006066	1.07E-58	*alcohol metabolic process*
HSA-1655829	1.41E-56	*Regulation of cholesterol biosynthesis by SREBP (SREBF)*

Table Supp. 4. Ontological terms enrichment computed by the ENRICH module using the 66 human cholesterol-related genes. Only the ten most significant terms are displayed. The statistical value corresponds to the false discovery rate computed by the String website.

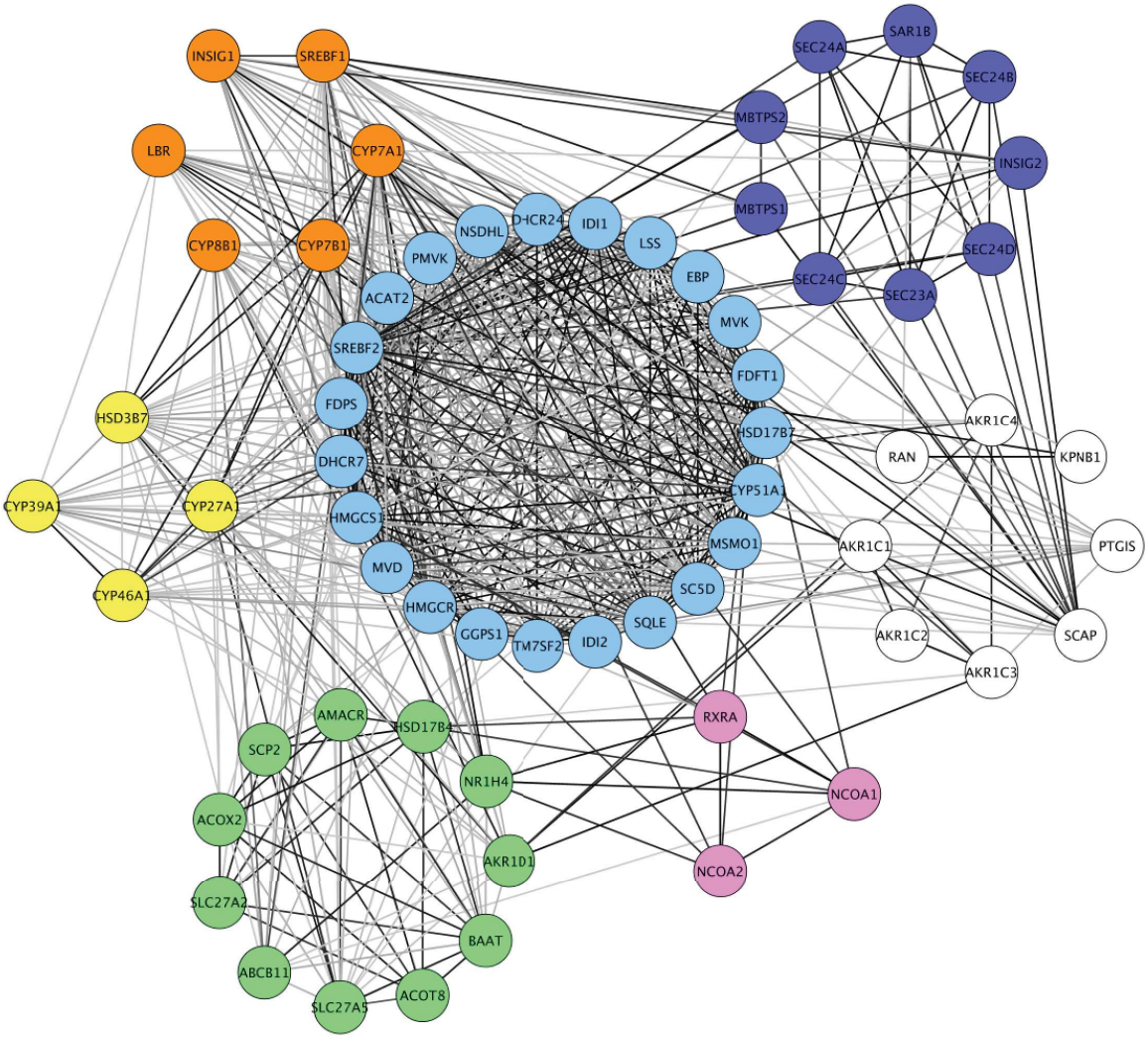


Fig. Supp. 5. Network graph of the 64 interconnected human proteins related to cholesterol. Nodes correspond to proteins, edges correspond to interactions. Proteins belonging to the same cluster are shown in the same color and defined in the text as: Cluster 1 (light blue), cluster 2 (orange), cluster 3 (yellow), cluster 4 (green), cluster 5 (dark blue), cluster 6 (pink). Edges are represented with different grey intensity according to the interaction confidence score available on the String website. The graph was generated using Cytoscape.

Cluster N°	N° of proteins	Enriched terms
1	23	*Cholesterol biosynthesis, cholesterol biosynthetic process, cholesterol metabolic process, Steroid biosynthesis, Metabolism of steroids*
2	6	*Cholesterol metabolism, Cytochrome P450 cholesterol 7-alpha-monooxygenase-type, Cytochrome P450 E-class group IV, Primary bile acid biosynthesis, Endoplasmic reticulum*
3	4	*Primary bile acid synthesis, Synthesis of bile acids and bile salts via 24-hydroxycholesterol, bile acid biosynthesis process, Endogenous sterols, steroid hydroxylase activity*
4	11	*Bile acid metabolic process, Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol, bile acid biosynthesis process, Primary bile acid biosynthesis, Metabolism of lipids*
5	9	*Regulation of cholesterol biosynthesis by SREBP (SREBF), Protein processing in endoplasmic reticulum, Antigen presentation: Folding, assembly and peptide loading of class I MHC, Cargo concentration in the ER, Sec23/Sec24 zinc finger*
6	3	*Endogenous sterols, Synthesis of bile acids and bile salts via 27-hydroxycholesterol, Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol, Recycling of bile acids and salts, BMAL1: CLOCK, NPAS2 activates circadian gene expression*

Table Supp. 5. Ontological terms enrichment calculated for each cluster of the BIONET network. Only the five most significant terms are displayed. Terms corresponding to Pubmed publications (PMID) were not included.

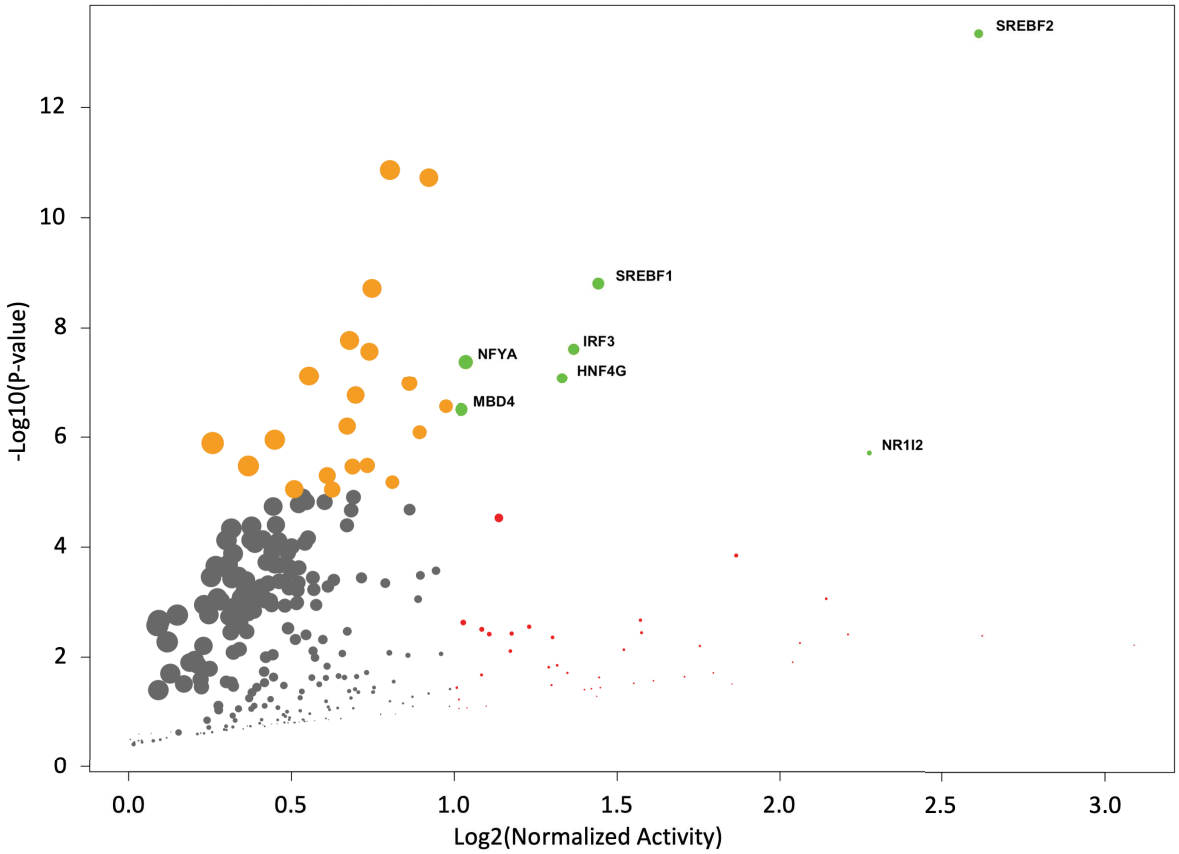


Fig. Supp. 6. Volcano plot displaying transcription factors involvement in the expression control of the 66 human cholesterol-related genes. Only factors with "Normalized Activity" >1 are represented. Each dot corresponds to a transcription factor (TF) and is represented according to its log2(Normalized Activity) and its -log10(P-value). The size of the dots corresponds to the TF “Influence” value. TFs with a log2(Normalized Activity) >=1 are represented in red. TFs having a -log10(P-value) >=5 are represented in orange. TFs corresponding to most significant candidates (log2(Normalized Activity) >=1 and a -log10(P-value) >=5) are represented in green and labelled with their respective gene names. The SREBF2 TF appears as the best candidate for cholesterol-related gene regulation. In addition, its paralog SREBF1 is also one of the most involved transcription factors.

Cluster N°	Nb of proteins	Nb of "Ortholog"	Nb of "Homolog"	Nb of "Domain-level"	Nb of "None"
1	23	10	3	5	5
2	6	1	1	3	1
3	4	0	2	2	0
4	11	3	5	1	2
5	9	6	2	0	1
6	3	1	0	2	0
Not clustered	8	3	4	1	0
Not in network	2	2	0	0	0
Total	66	26	17	14	9

Table Supp. 6. Protein homology-relation distribution by cluster. Protein homology relation was calculated by EVOBLAST as described in Supplementary File S2. For each cluster or unclustered protein set, total number of proteins, number of proteins annotated as "Ortholog", "Homolog", "Domain-level" and "None" by EVOBLAST are shown. Degree of conservation between human and D. melanogaster is variable between clusters with some containing mainly conserved proteins (e.g. cluster 5) and others composed of proteins with no equivalent in the fly (e.g. cluster 2).

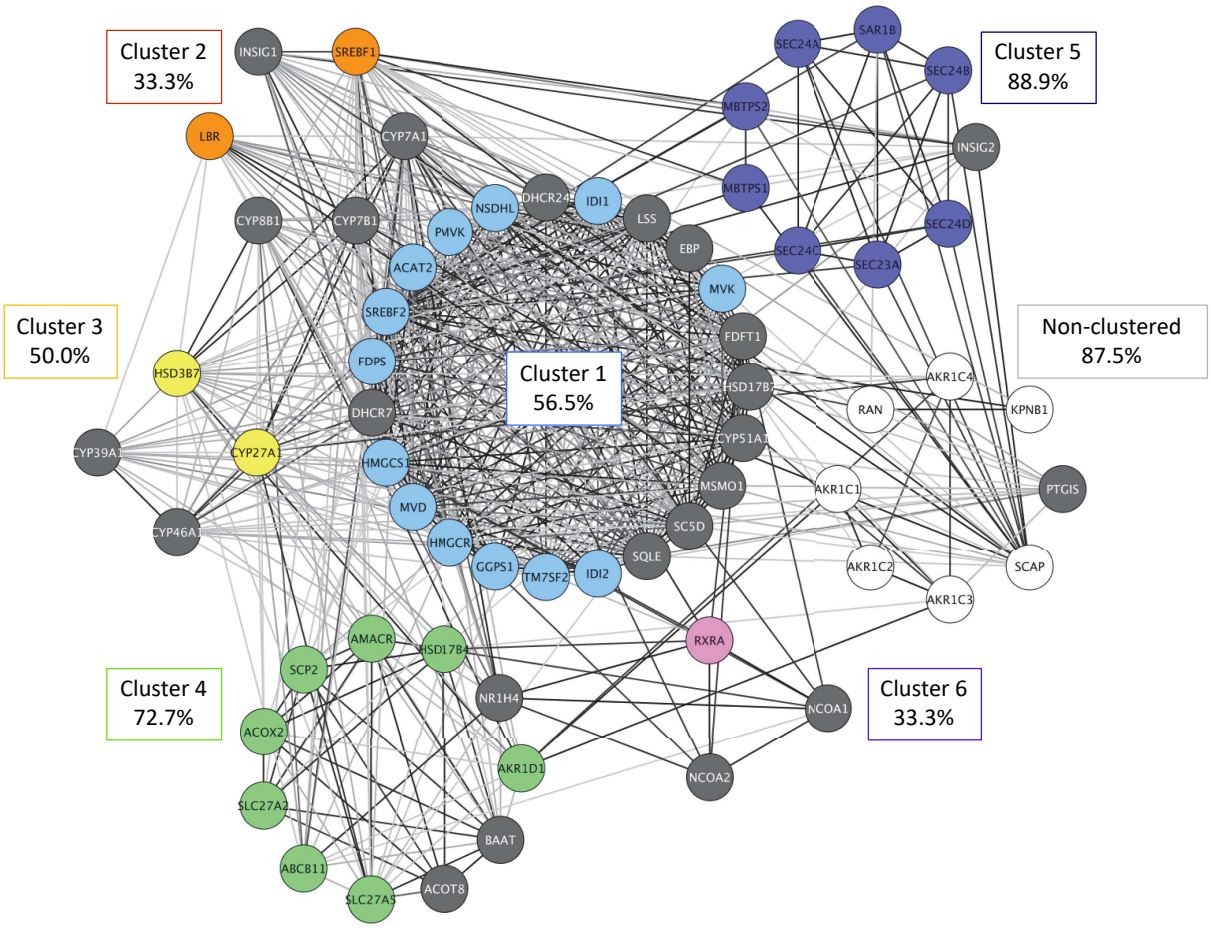


Fig. Supp. 7. EVOBLAST annotations in the network of the 64 interconnected human proteins related to cholesterol. Nodes correspond to proteins, edges correspond to interactions. Proteins conserved between human and *D. melanogaster* proteomes are highlighted by the respective cluster colour. Nodes filled with grey represent proteins annotated as "Domain-level" or "None" by EVOBLAST and therefore considered as absent in the fly proteome. Percentages displayed on the figure correspond to the frequencies of the *Homo sapiens* proteins conserved in *D. melanogaster* within each cluster. The graph was generated using Cytoscape.

References

Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet*, **25**, 25–29.

Castrillo,A. *et al.* (2003) Crosstalk between LXR and Toll-like Receptor Signaling Mediates Bacterial and Viral Antagonism of Cholesterol Metabolism. *Molecular Cell*, **12**, 805–816.

Espenshade,P.J. and Hughes,A.L. (2007) Regulation of Sterol Synthesis in Eukaryotes. *Annu. Rev. Genet.*, **41**, 401–427.

Jassal,B. *et al.* (2019) The reactome pathway knowledgebase. *Nucleic Acids Research*, gkz1031.

Janky,R. *et al.* (2014) iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Comput Biol*, **10**, e1003731.

Lachmann,A. *et al.* (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.

Rawson,R.B. (2003) The SREBP pathway — insights from insigs and insects. *Nat Rev Mol Cell Biol*, **4**, 631–640.

Rescher,U. and Volker,G. (2004) Annexins - unique membrane binding proteins with diverse functions. *Journal of Cell Science*, **117**, 2631–2639.

Seegmiller,A.C. *et al.* (2002) The SREBP Pathway in Drosophila: Regulation by Palmitate, Not Sterols. *Developmental Cell*, **2**, 229-238.

Shannon,P. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, **13**, 2498–2504.

Szklarczyk,D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, **47**, D607–D613.

Tatusov,R.L. (1997) A Genomic Perspective on Protein Families. *Science*, **278**, 631–637.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Wingender,E. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, **24**, 238–241.

Zhang,H. and Reilly,M.P. (2015) IRF2BP2: A New Player at the Crossroads of Inflammation and Lipid Metabolism. *Circ Res*, **117**, 656–658.

CONCLUSIONS AND PERSPECTIVES

Comparative approaches led to many fundamental advances in Biology. One of the most impressive achievements is probably comparative embryology, which constituted an important support to the theory of evolution (Richardson and Keuck, 2002*)*. By demonstrating that even distant species emanate from nearly identical embryological structures formed at early stages, it appeared clearly that ontogenesis corroborate a continuous evolutionary relation between species (Figure 21). The power of comparative approaches resides in the fact that it allows the dissection of biological complexity into similar and dissimilar items and highlight priority traits based on conservation criteria during evolution on large time scales, characteristics that are not accessible in another way.

Figure 21 – Plates of eight species compared at three stages of development. Left to right: fish, salamander, turtle, chicken, pig, cow, rabbit and human. This comparative picture shows that even distant animals share identical embryonic stages, supporting their evolutive relations. This draw by Haeckel is reported in Richardson and Keuck, 2002.

During my thesis, I followed this common thread and applied comparative approaches at the level of genes, based on transcriptomic or genomic analyses. Thanks to this method, I

identified a set of genes conserved among vertebrate species involved in stress response and adaptation. Moreover, I highlighted the presence of archaeal and eukaryotic genes in bacterial genomes from the *Candidatus poribacteria* phylum and demonstrated the mosaic organization of the Coenzyme A biosynthetic pathway in this group. Ultimately, I also designed an analytical workflow to investigate sets of genes, mainly based on comparisons of genes expression and distribution among species. These works paved the way for further studies about stress response and to the extension of observations made from *C. poribacteria* to other related taxa. Finally, the establishment of the PROTEDEX workflow also initiated the design of an integrative platform that could be further extended and improved.

Conclusion and perspectives to "Transcriptomic analysis to identify conserved genes and mechanisms involved in stress response and adaptation in vertebrate species"

This work was based on the provocative assumption of the existence of mechanism(s) participating in stress response that are conserved between species submitted to different stressors. This hypothesis is based on the idea that evolution will have selected an important functional architecture, which constitutes the support for an elaborate and complex biological network allowing the response to a multitude of situations or challenges to which individuals will be exposed. In this way, we identified 26 genes differentially expressed during stress response, conserved between at least three of the models that we analyzed. This number of genes was almost ideal, restricted enough to be experimentally validated and sufficient enough to initiate bioinformatics investigations, such as ontological enrichment and network analyzes. However, it is likely that our comparative analysis only disclosed the emerged part of a larger set of genes effectively involved in stress response. Indeed, the approach used in this study introduced unavoidable biases, resulting from the choices of species considered, stressor tested, the tissue investigated or inherent to the selected bioinformatics tools, notably from the process for orthology-based conversion of gene names in case of highly duplicated gene families. Despite these limitations, the analyzes of the obtained set of genes unveiled different convergent trends, such as extracellular matrix (ECM) ontological term, co-expression, network connectivity. Moreover, thanks to the access to the full transcriptional program involved in our four models, it was possible to extend this set using additional criteria than conservation, as discussed below.

The similarity of the expression profiles of the stress-related genes, despite the diversity of the experimental conditions, reinforces the relevance of this set of genes and suggested their involvement in one common biological process. Network analyses permitted to extend this group of initial genes to other co-localized or co-expressed proteins, and highlighted the importance of one signal transduction pathway controlled by TGFß. The main question that remains to be answer yet is to what extend these results could be generalized: is it limited to the species/stressor models analyze in this work; to the muscle tissue that was considered; to the vertebrate group studied? How robust is this conservation through evolution?

Concerning the tissue specificity, in the manuscript presented in the Results section, we showed preliminary investigations on liver tissues obtained from the animals used in the

mouse/exercise and chicken/nutritional stress models. The results obtained confirmed the significative differential expression of most of the 26 genes in this tissue as well, validating their involvement as systemic indicators in animals. However, the detailed investigations of the stress-related genes expression also highlighted a variable expression profile between muscles and livers, suggesting that stress response displays a tissue-specific pattern. We proposed that this variability reflect the course and kinetic of stress response process that might differ between different organs. To complete this approach and provide additional hints for the contribution of these genes to the response of different stressors, we also developed an alternative stress model consisting in chickens exposed to a xenobiotic challenge induced by paraquat, an herbicide known to provoke oxidative damages in cells. This additional model, also described in the manuscript, validated the differential expression of the stress-related genes in response to this new stressor. These results reinforced the conclusions proposed in the study, but also rose another question: how many different models will be necessary to validate the set of 26 stress-related genes? Indeed, considering the variety of possible stressors from a given environment and the number of living species, all possible situations cannot be reasonably addressed and there is probably no ideal sufficient number of models to test and definitively validate this set. Consequently, each individual validation in an other model can just comfort our results and increase their statistical significance and robustness. Alternatively, conducting a meta-analysis over a broad number of stress situations, analyzed using a dedicated algorithm, could allow us to integrate the data in form of characteristic profiles or signatures. These characteristic signatures would be beneficial to either a better diagnostic of stress stages or to increase fundamental knowledge about the biological processes involved in stress response.

Another possible validation concerning the relevance of the set of 26 stress-related genes, will be to ask about their conservation in more distant species. To answer this question, we considered *Caenorhabditis elegans*, a well-studied animal model. This laboratory model present several assets for our studies: it is easy to grow in relative large quantities at low cost; It has a short life-cycle (three weeks); although it is a relatively simple organism, most genes and signaling pathways are conserved with more complex animals. For these reasons, *C. elegans* is a widely used model to study diseases or to screen active molecules that can be used on other species (Rodrigez et al., 2013). This nematode also presents the advantage that it can be easily genetically manipulated using the RNA interference technique. In addition, bibliographic studies showed that several experimental stress models were already developed and studied in *C. elegans*, such as heat, paraquat or starving stresses (Castro et al., 2012; Zevian

et Yanowitz, 2014; Possik and Pause, 2015; Crombie et al., 2016; Dilberger et al., 2019). So far, we identified in the worm a list of fifteen genes homologous to the 26 vertebrate set, including seven validated orthologs and eight paralogs. It is plan to conduct analyzes on the expression of these fifteen genes during heat- and paraquat-induced stresses. In addition, this model will help to address the dynamic of expression of each gene by time course experiments (Jovic et al., 2017). It is expected to characterize genes with early and late expression profiles, suggesting their involvement in different sequences of the response. Early genes are predicted to participate in stress signal integration and the setting of defense mechanisms, as the ones expressed latter are more likely to correspond to repair mechanisms and the setting of adaptive processes. For this, it remains to be defined whether the same group of genes will participate to the same sequence, independently of the stressor. In second step, taking advantage of the RNA interference, it will be possible to inactivate the expression of each genes of interest individually and to test the impact on stress response, using the stress models previously established. First, it will necessary to verify that the inactivation of each gene is not lethal in basal conditions. Instead, in case the RNA interference approach will not be successful, a library of mutants for most *C. elegans* genes is also available; these mutants could be used for physiological tests as well.

Based on our transcriptomic study, we determined that proteins located in the ECM are important actors of stress response and adaptation. The importance of this observation was discussed previously. What remains to be understand is how remodeling of cell environment into tissues relates to stress response. This remodeling could constitute a signaling pathway induced by the release of signaling molecules sequestered in the ECM, that get available to bind to membrane cellular receptors. Alternatively, ECM remodeling could contribute to a repair mechanism, mediated for example through collagen turnover, an active mechanism in wounding (Kjaer et al., 2006; Fisher et al., 2009). Another interesting point to address will be to analyze the interplay between alteration of ECM and other stress-related processes, such as inflammatory response and infiltration of immune cells in the tissue (Robert et al., 2016).

We also proposed TGFß and the PI3K-AKT signaling pathways to be key regulators of gene expression program during stress response. These predictions emanated from different convergent results: SMAD3 and SMAD4 were predicted as transcription factors involved in the regulation of the stress-related genes; THBS1, one of 26 stress-related genes, is one of the factors that activate TGFß in the extracellular matrix; SRC et CD44, two signal transducing

factors of the TGFß pathway were identified to occupy a central position in the networks specific to the different stress models, as well as in the extended network of genes co-expressed. Moreover, an experiment designed to identify TGFß target genes in embryo dermal fibroblast, by comparing gene expression profiles in cells treated with MPPN (2-(3-(6-Methylpyridin-2-yl)-1H-pyrazol-4-yl)-1,5-naphthyridine), an inhibitor of the TGFß receptor kinase, demonstrated the differential expression of many ECM protein coding genes, including several genes of the stress-related list (TNC, COL12A1, POSTN, RUNX1) or paralog genes corresponding to multigene families (THBS2, COL15A1, COL8A2, COL11A1, COL23A1, ANKRD1) (Kosla et al., 2013).

However, these transcriptomic data remains to be validated at the protein level. For this, we initiated experiments aiming to explore the involvement of the TGFß signaling pathway by measuring phosphorylation level of different reporter proteins. Since the TGFß cytokine is acting through two different pathways (Figure 9 manuscript I), we decided to assess the phosphorylation status of SMAD2 and SMAD3, two, effectors of the canonical pathway, as well as phosphorylation of AKT, representing the activation of the non-canonical pathway. In addition, knowing that SRC can act either alone or in combination with TGFß, we are also planning to measure SRC phosphorylation status. Altogether, these results are expected to validate the activation of the TGFß signaling in the stress models and considering the redundancy of the signaling pathways, to characterize the contribution of each effector on transcriptional control of stress-related genes. Interestingly, preliminary results tend to validate the specific activation of the PI3K-AKT signaling pathway in adapted animals. Again, *C. elegans* might be the model of providence to address this question. Indeed, inhibitors of the TGFß pathway could be tested for their interference with stress response. Conversely, the effect of hyaluronic acid (HA), a ligand of the CD44 receptor involved in the activation of the TGFß pathway, acting as a promoter of stress resistance and adaptive capacity could be tested. Of note, no homolog for CD44 was identified in *C. elegans*, but another hyaluronic receptor called RHAMM was characterized as part of an ancient family of HA binding proteins (Csoka and Stern, 2013).

Although we choose to focus our attention on mRNAs encoding proteins, we have also access to complementary expression data for long non-coding RNAs (lncRNAs) and micro RNAs (miRNAs). Information about these components may be of particular interest as these RNAs are predicted to regulate mRNAs and proteins expression on a shorter time scale than

transcription regulation, consistent with rapid adaptation to environmental challenges. Therefore, non-coding RNAs constitute potential regulatory nodes but also eventual biomarkers of the stress animal status. Our transcriptomics analyzes included characterization of differentially expressed lncRNAs and miRNAs, and ten to twenty such genes have been identified for each model. However, at the time we performed these studies, functional information about these RNAs was not sufficient to improve our understanding of stress response, especially in non-model animals such as pig and chicken. In addition, validation of the ncRNA candidates in our biological models could not be considered; most ncRNA activities are assessed using genetically modified cellular models. Such approach may be poorly representative of what we observed in this study at the tissue level and cell culture models are not appropriate to study the ECM remodeling process. Consequently, it would be difficult to correlate these results with the observations from whole animal studies in absence of relevant biological models.

To gain insight on the functional relations between proteins encoded by the DE genes of each model, I generated four protein-protein interaction networks. Such method allowed me to highlight groups of highly interconnected proteins and also to propose central regulators of these networks. However, the biggest flaw of such methodology is that these networks condensed all three comparisons for each model (adapted versus control, non-adapted versus control and non-adapted versus adapted), a complicated situation to characterize adapted and non-adapted animals specificities. To decompose the complexity of these stress-related networks, one possibility is to hierarchize the DE genes involved in adapted and non-adapted individuals during stress response. An option consists in the identification of common and specific DE genes between adapted versus control and non-adapted versus control animals respectively and to compute biological enrichments. Eventually, a preliminary step of network could be generated to put away encoded proteins not interconnected. Alternatively, DE genes could also be classified according to their expression variation in the two comparisons to determine group of co-up-regulated, co-down-regulated or with an opposite differential expression between comparisons. To this aim, hierarchical clustering methods could provide an appropriated solution. Such methodology would help to analyze the specificities of adapted and non-adapted organisms and to complete our results mainly focused on genes conserved between models and their co-expressed partners.

Concerning the diagnostic for animal stress state, we propose that the set of 26 stress-related genes identified here could be used as candidate biomarkers. Measuring the expression of genes in tissues from farm animals is possible particularly in chickens, where feathers can constitute a sufficient source of RNAs, easily accessible. However, most of the common diagnostic protocols are rather based on methods to measure proteins or metabolites in a non-invasive way, directly from blood samples or urine. Investigation of the biological processes in which the 26 gene products are involved could help to define other markers such as metabolite produced by enzymes or peptides modifications, directly accessible in blood. In practice, once the biomarkers are measured, it remains to establish its relation with the animal status. To correlate gene expression with the stress status in animals, I developed a simple algorithm based on the data observed in our animal models. This program uses the expression of all or part of the 26 conserved genes to determine if (i) an animal is enduring stress conditions and (ii) shows signs of adaptation or maladaptation, and (iii) to quantify the similarity with one of our reference models. This tool also provides useful outputs to visualize the stress state and similarity with the other models using a radar plot representation. The current version of the algorithm has been tested using the expression profile of the 26 conserved genes from non-adapted chickens submitted to paraquat-induced stress. Obtained diagram anticipated their stressed non-adaptation status, but also predicted this profile to be highly similar to the one of the non-adapted chickens from the nutritional stress model (Figure 22). To improve these predictions, it will require an iterative process based on additional results from new stress models. As another possible improvement, we could consider model-specific differentially expressed genes, in addition to the common ones, to propose a method assessing more accurately specific components of stress response.

The set of genes or biomarkers identified in this study being conserved between species, it is expected that it will be applicable to Human as well. Indeed, it will of interest to explore whether these markers could be useful indicators of stress status in patients with different pathological conditions. It remains to be defined which diseases will be the most relevant to carry these investigations, but cancer appears as a promising candidate as ECM remodeling has already been shown to constitute a key factor for tumorigenesis. In case it will be applicable, diagnostic using these genes could be helpful to direct targeted therapeutic strategies.

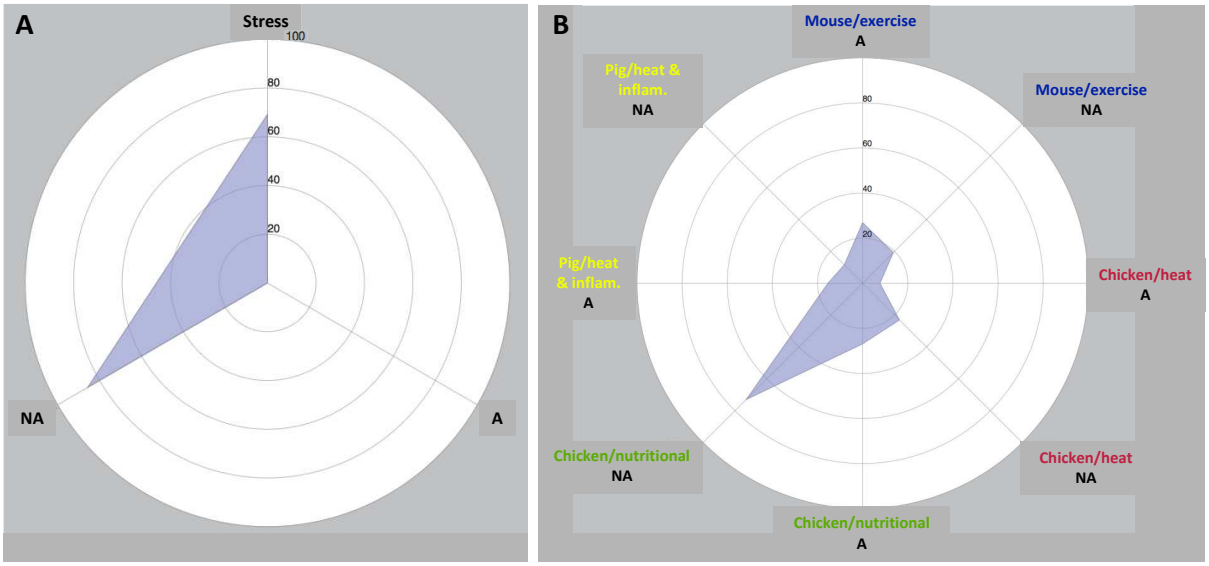


Figure 22 – Radar plots generated by the algorithm based on differential expression of genes from chickens exposed to a xenobiotic stress (paraquat). (A) According to the genes expression profile from the paraquat model, the algorithm predicted a stress state in chickens exposed to the stressor (Stress score > 60%) and anticipated that the animal response is characteristic of non-adapted animals (Non-adaptation score > 80%). NA= non-adapted; A= adapted. **(B)** The comparison of the 26 genes expression profile from the animals exposed to paraquat with our reference models showed the highest similarity with the non-adapted chickens from the chicken/nutritional model (similarity with chicken/nutritional-NA > 70%).

Conclusion and perspectives to "Mosaic organization of metabolic pathways between bacteria and archaea species"

Despite the absence of a clear classification, the *Candidatus poribacteria* group is of particular interest because it gathers organisms with different lifestyles that radically differ at the genomic level as confirmed by comparative genomics studies (Podell et al., 2019). Indeed, the percentage of genes conserved between the free-living and symbiotic Poribacteria is about 50%, indicating an important specialization depending on the lifestyle. In our study, we demonstrated that phosphopantothenate biosynthesis, one of the initial steps in Coenzyme A synthesis, is part of this specialization, with the bacterial-specific enzymes PanK and PS in free-living Poribacteria (Pelagiporibacteria) genomes and the archaeal-specific enzymes PoK and PPS in symbiotic Poribacteria (Entoporibacteria) genomes. It was concluded that the presence of the archaea genes in this bacterial group could be the result of an horizontal gene transfer favored by the close proximity of bacteria and archaea organisms as part of the symbiotic community within the sponge. Alternatively, the archaeal pathway was selected because it provides an evolutionary advantage for the symbiosis. Because of the impossibility to grow these organisms outside of the sponge host, it is difficult to solve this question. It is also possible that this observation of bacteria containing the archaea-specific genes for pantothenate biosynthesis could be extended to other organisms. Indeed, during this work, I observed the presence of the archaeal enzymes in a few members of the bacterial superphylum Planctomycetes-Verrucomicrobia-Chlamydiae (PVC). Interestingly, it was proposed that *Candidatus poribacteria* constitute a group phylogenetically related to the PVC clade. It would be of interest to disclose if these organisms are also symbiotic bacteria, but the genomic sequences retrieved from metagenomic project analyzes were poorly annotated. As mentioned before, Entoporibacteria are impossible to grow in laboratory conditions because it needs the unique environment provided by the eukaryotic host. By default, we could engineer an *Escherichia coli* mutant strain expressing the archaeal enzymes instead of the endogenous bacterial ones, and to test for its viability. This will tell whether the two pathways are functionally redundant. Another important aspect of this question concerns the regulation of these pathways. Indeed, it has been shown that the PanK enzyme is regulated through a feedback inhibition mechanism by Coenzyme A (Shimosaka et al., 2016). In the case of the archaeal PoK, this enzyme was shown to be moderately regulated by its substrate pantoate to prevent accumulation of phosphopantoate within the cell (Tomita et al. 2012). During this study, I investigated in more details the case of these couples of two enzymatic, but the question

of other genes showing such atypical distribution can be asked. It would be of interest to search for other archaeal genes specific to the Entoporibacteria sub-group, indicative of their characteristic link with the symbiotic lifestyle.

Conclusion and perspectives to "Development of an integrative tool for gene sets investigation"

PROTEDEX is a workflow developed to propose a straightforward way to manage large sets of genes and to investigate their roles. With this project, we aimed to design a standardized protocol combining well-known programs and to increase reproducibility. However, despite our initial intentions, we made the choice to develop two additional original modules, TRACE and EVOBLAST, while other solutions already existed. TRACE was proposed instead of iRegulon. The iRegulon analysis is based on the identification of consensus DNA binding motifs recognized by transcription factors, as TRACE relies on publicly available databases of transcription factors and their related gene targets. One advantage of TRACE, in contrast to iRegulon, is that it does not require any local software installation. Some specialized tools to perform such online identification of transcription factor involved in gene expression regulation have been recently proposed and could constitute interesting alternatives to TRACE, including BART (Wang et al., 2018) and ChEA3 (Keenan et al., 2019). EVOBLAST was designed to offer a complementary approach to other well-known tools for orthology search, such as OrthoInspector (Nevers et al., 2019), with no limit for target organisms but conversely, a slower computational capacity, because it relies on no precomputed data. In return, EVOBLAST is also completely dependent on the NCBI server status, a necessary condition to maintain it up-to-date. Another possible criticism to PROTEDEX is that it offers a standardized approach, but it is based on arbitrary choices that we made, such as the use of certain annotation systems (GO Biological Process and Reactome), certain databases (Swissprot, Transfac, ChEA and ENCODE) and certain tools (STRING, BLAST). Here is a dilemma between adding more possible tools and parameters, and keeping a reduced standardized workflow to increase reproducibility. One way to improve PROTEDEX impact would be to ensure that the options we choose for each module are the most relevant. For example, using the STRING API to build networks is a good choice, but is not the best way to compute ontological enrichments because it does not offer the possibility to integrate a user-provided background list.

In addition, several improvements are possible and planed for each module:

First, we plan to make PROTEDEX usable with unique IDs in addition to gene names and to implement a more robust gene converter than UNIDEX, such as BioMart (Durinck et al., 2005).

Concerning the network construction and clustering by the BIONET module, we want to test other clustering algorithms instead of MCODE, such as ClusterONE and to propose the possibility to manage cluster property parameters through the PROTEDEX interface directly. Moreover, the output files generated could provide more quantitative information about the network basic parameters such as the number of nodes, number of edges and mean degree. Finally, BIONET and ENRICH results are currently generated as text files, but it would be useful to automatically generate pictures of the networks and graphical representations of the computed enriched terms.

One possibility that PROTEDEX offers to the user is to group genes based on quantitative values if they are provided. To go further than just generating independent networks and computing specific enrichments, we would like to implement TRACE in order to integrate these values for the transcription factor (TF) scoring by coupling it with functional data. Currently, if a list of genes contains many targets for a TF, this regulator will be pointed by TRACE as a good candidate for transcription activation, independently of the up- or down–regulation of the gene. However, TRACE should consider if the TF is an activator or repressor of transcription, regarding the values associated to its target genes.

EVOBLAST could be improved in many ways notably to perform faster analyzes. First, EVOBLAST could be associated to other services like those provided by the Alliance of Genome Resources (The Alliance of Genome Resources Consortium, 2020) that propose pre-computed homology relations for some model organisms. By this way, our module could run faster and rely on well-annotated homology relations. Another way to decrease the computational time of our method would be to pool the best hits from protein requests. In its actual version, EVOBLAST considers each protein successively to use it as query for a BLASTp search first, and then retains the best hit for a reciprocal BLASTp. However, in some situations, different proteins from the user list can point to the same best hit in a target organism, a relevant situation in case of large family members. In this situation, EVOBLAST will launch redundant reciprocal BLASTp requests. This could be avoided by: first, launching all BLASTp; then, establishing the best hit redundancy and, finally, running only unique reciprocal BLASTp searches. We also consider the possibility to improve the EVOBLAST computation speed by launching several requests in parallel without exceeding what is permitted by NCBI.

Finally, we plan to extend the possibilities of PROTEDEX by implementing additional databases and services. For example, we would like to add databases specialized on signaling pathways such as SIGNOR 2.0 (Licata et al., 2019) to complete the network generated by

BIONET with oriented edges based on functional data. We are also interested to extend the possibilities of EVOBLAST to allow it to use the tBLASTn service in addition to BLASTp. Indeed, a high amount of data is exclusively accessible through the WGS database from NCBI, and can be requested with protein sequences queries only by tBLASTn.

To conclude, we are currently working on an offline version of PROTEDEX accessible from GitHub and to be used locally.

BIBLIOGRAPHY

Aburn G, Gott M, Hoare K. 2016. What is resilience? An Integrative Review of the empirical literature. *J Adv Nurs* 72:980–1000.

Adams MD. 2000. The Genome Sequence of Drosophila melanogaster. *Science* 287:2185–2195.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403-410.

Anderson L, Seilhamer J. 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18:533–537.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet* 25:25–29.

Ast T, Mootha VK. 2019. Oxygen and mammalian cell culture: are we repeating the experiment of Dr. Ox? *Nat Metab* 1:858–860.

Avery OT, MacLeod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med* 79(2):137-158.

Baffy G, Loscalzo J. 2014. Complexity and network dynamics in physiological adaptation: An integrated view. *Physiol Behav* 131:49–56.

Barabasi AL, Albert R. 1999. Emergence of Scaling in Random Networks. *Science* 286:509-512.

Bedard K, Krause K-H. 2007. The NOX Family of ROS-Generating NADPH Oxidases: Physiology and Pathophysiology. *Physiol Rev* 87:245–313.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.

Cadenas S. 2018. Mitochondrial uncoupling, ROS generation and cardioprotection. *Biochim. Biophys. Acta* 1859:940–950.

Cadet J, Wagner JR. 2013. DNA Base Damage by Reactive Oxygen Species, Oxidizing Agents, and UV Radiation. *Csh Perspect Biol* 5:a012559–a012559.

Castro PV, Khare S, Young BD, Clarke SG. 2012. Caenorhabditis elegans Battling Starvation Stress: Low Levels of Ethanol Prolong Lifespan in L1 Larvae. *PLoS ONE* 7:e29984.

Chargaff E, Lipshitz R, Green C. 1952. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J Biol Chem* 195:155-160.

Conesa A, Mortazavi A. 2014. The common ground of genomics and systems biology. *BMC Syst Biol* 8:S1.

Cordeiro JV, Jacinto A. 2013. The role of transcription-independent damage signals in the initiation of epithelial wound healing. *Nat Rev Mol Cell Biol* 14:249–262.

Cortese-Krott MM et al. 2017. The Reactive Species Interactome: Evolutionary Emergence, Biological Significance, and Opportunities for Redox Metabolomics and Personalized Medicine. *Antioxid Redox Signal*. 27:684–712.

Covelli V, Passeri M, Leogrande D, Jirillo E, Amati L. 2005. Drug Targets in Stress-Related Disorders. *CMC* 12:1801–1809.

Crombie TA, Tang L, Choe KP, Julian D. 2016. Inhibition of the oxidative stress response by heat stress in *Caenorhabditis elegans*. *J Exp Biol* 219:2201–2211.

Csoka AB, Stern R. 2013. Hypotheses on the evolution of hyaluronan: A highly ironic acid. *Glycobiology* 23:398–411.

Dalle-Donne I, Giustarini D, Colombo R, Rossi R, Milzani A. 2003. Protein carbonylation in human diseases. *Trends. Mol. Med*. 9:169–176.

Dalle-Donne I, Rossi R, Giustarini D, Gagliano N, Di Simplicio P, Colombo R, Milzani A. 2002. Methionine oxidation as a major cause of the functional impairment of oxidized actin. *Free Radical Bio Med* 32:927–937.

Darwin C. 1859. On the Origin of Species by Means of Natural Selection Or the Preservation of Favoured Races in the Struggle for Life. H. Milford; Oxford University Press.

Davies JMS, Cillard J, Friguet B, Cadenas E, Cadet J, Cayce R, Fishmann A, Liao D, Bulteau A-L, Derbré F, et al. 2017. The Oxygen Paradox, the French Paradox, and age-related diseases. *GeroScience* 39:499–550.

Davies KJA. 2016. Adaptive homeostasis. *Mol Aspects Med* 49:1–7.

Dayhoff MO, Ledley RS. 1962. Comprotein: a computer program to aid primary protein structure determination. In: Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall). Philadelphia, Pennsylvania: ACM Press. p. 262–274. Available from: http://portal.acm.org/citation.cfm?doid=1461518.1461546

De Vries H. 1910. The Mutation Theory: The origin of varieties by mutation (Vol. 2). Open Court Publishing Company.

De Vries H. 1910. Intracellular pangenesis. Chicago: Open Court.

Dias MS, Mattos JRLD. 2011. HUMAN RACE ACTIONS VERSUS THE BREAKING OF THE CO2 EQUILIBRIUM LIMIT. Available from: http://rgdoi.net/10.13140/RG.2.1.3853.6723

Dilberger B, Baumanns S, Schmitt F, Schmiedl T, Hardt M, Wenzel U, Eckert GP. 2019. Mitochondrial Oxidative Stress Impairs Energy Metabolism and Reduces Stress Resistance and Longevity of C. elegans. *Oxid Med Cell Longev* 2019:1–14.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21:3439–3440.

Eaton S. 2006. The biochemical basis of antioxidant therapy in critical illness. *Proc. Nutr. Soc.* 65:242–249.

Eck RV, Dayhoff MO. 1966. Atlas of Protein Sequence and Structure. *Natl. Biomed. Res. Found.*, Washington, DC.

Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, Forsström B, Uhlén M. 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* 12:883.

Edman P. 1950. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chemica Scandinavica* 4:283-293.

El-Benna J, Hurtado-Nedelec M, Marzaioli V, Marie J-C, Gougerot-Pocidalo M-A, Dang PM-C. 2016. Priming of the neutrophil respiratory burst: role in host defense and inflammation. *Immunol Rev* 273:180–193.

Eline Slagboom P, van den Berg N, Deelen J. 2018. Phenome and genome based studies into human ageing and longevity: An overview. *Biochim. Biophys. Acta* 1864:2742–2751.

Ellgaard L, Sevier CS, Bulleid NJ. 2018. How Are Proteins Reduced in the Endoplasmic Reticulum? *Trends Biochem. Sci.* 43:32–43.

Ewald C. 2018. Redox Signaling of NADPH Oxidases Regulates Oxidative Stress Responses, Immunity and Aging. *Antioxidants* 7:130.

Farah ME, Sirotkin V, Haarer B, Kakhniashvili D, Amberg DC. 2011. Diverse protective roles of the actin cytoskeleton during oxidative stress. *Cytoskeleton* 68:340–354.

Feder A, Fred-Torres S, Southwick SM, Charney DS. 2019. The Biology of Human Resilience: Opportunities for Enhancing Resilience Across the Life Span. *Biol. Psychiatry* 86:443–453.

Fischer WW, Hemp J, Valentine JS. 2016. How did life survive Earth's great oxygenation? *Curr Opin Chem Biol* 31:166–178.

Fisher GJ, Quan T, Purohit T, Shao Y, Cho MK, He T, Varani J, Kang S, Voorhees JJ. 2009. Collagen Fragmentation Promotes Oxidative Stress and Elevates Matrix Metalloproteinase-1 in Fibroblasts in Aged Human Skin. *Am. J. Clin. Pathol.* 174:101–114.

Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, Bult C, Tomb J, Dougherty B, Merrick J, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269:496–512.

Frijhoff J, Winyard PG, Zarkovic N, Davies SS, Stocker R, Cheng D, Knight AR, Taylor EL, Oettrich J, Ruskovska T, et al. 2015. Clinical Relevance of Biomarkers of Oxidative Stress. *Antioxid Redox Sign* 23:1144–1170.

Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* 14:360–366.

Gaschler MM, Stockwell BR. 2017. Lipid peroxidation in cell death. *Biochem Bioph Res Co* 482:419–425.

Glennon-Alty L, Hackett AP, Chapman EA, Wright HL. 2018. Neutrophils and redox stress in the pathogenesis of autoimmune disease. *Free Radical Bio Med* 125:25–35.

Go Y-M, Chandler JD, Jones DP. 2015. The cysteine proteome. *Free Radical Bio Med* 84:227–245.

Go Y-M, Jones DP. 2013. The Redox Proteome. *J. Biol. Chem.* 288:26512–26520.

Go Y-M, Jones DP. 2017. Redox theory of aging: implications for health and disease. *Clin. Sci.* 131:1669–1688.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 Genes. *Science* 274:546–567.

Gontier N. 2011. Depicting the Tree of Life: the Philosophical and Historical Roots of Evolutionary Tree Diagrams. *Evo Edu Outreach* 4:515–538.

Griffith F. 1928. The Significance of Pneumococcal Types. *J. Hyg.* 27:113–159.

Guil S, Esteller M. 2015. RNA–RNA interactions in gene regulation: the coding and noncoding players. *Trends in Biochem. Sci.* 40:248–256.

Halliwell B, Gutteridge JMC. Free radicals in biology and medicine. United States of America: Oxford University Press, 2015, 905 p.

Hawk M, McCallister C, Schafer Z. 2016. Antioxidant Activity during Tumor Progression: A Necessity for the Survival of Cancer Cells? *Cancers* 8:92.

Higgins DG, Sharp PM. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73:237–244.

Hillion M, Antelmann H. 2015. Thiol-based redox switches in prokaryotes. *Biol. Chem.* 396:415–444.

Hillmer RA. 2015. Systems Biology for Biologists.True-Krob HL, editor. *PLoS Pathog* 11:e1004786.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Jablonka E, Lamb MJ, Avital E. 1998. 'Lamarckian' mechanisms in darwinian evolution. *Trends Ecol. Evol.* 13:206–210.

Jha JC, Banal C, Chow BSM, Cooper ME, Jandeleit-Dahm K. 2016. Diabetes and Kidney Disease: Role of Oxidative Stress. *Antioxid Redox Sign* 25:657–684.

Johannsen W. 1909. Elements of an exact theory of heredity. Jena: Gustav Fischer.

Jones DP. 2006. Redefining Oxidative Stress. *Antioxid Redox Sign* 8:1865–1879.

Jones DP, Sies H. 2015. The Redox Code. *Antioxid Redox Sign* 23:734–746.

Jovic K, Sterken MG, Grilli J, Bevers RPJ, Rodriguez M, Riksen JAG, Allesina S, Kammenga JE, Snoek LB. 2017. Temporal dynamics of gene expression in heat-stressed Caenorhabditis elegans. *PLoS ONE* 12:e0189445.

Keenan AB et al. 2019. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucl. Acids Res.* 47:W212–W224.

Kjaer M, Magnusson P, Krogsgaard M, Moller JB, Olesen J, Heinemeier K, Hansen M, Haraldsson B, Koskinen S, Esmarck B, et al. 2006. Extracellular matrix adaptation of tendon and skeletal muscle to exercise. *J Anat.* 208:445–450.

Klamt F, Zdanov S, Levine RL, Pariser A, Zhang Y, Zhang B, Yu L-R, Veenstra TD, Shacter E. 2009. Oxidant-induced apoptosis is mediated by oxidation of the actin-regulatory protein cofilin. *Nat Cell Biol* 11:1241–1246.

Koonin EV. 2015. Why the Central Dogma: on the nature of the great biological exclusion principle. *Biol Direct* 10:52.

Kornberg A, Kornberg SR, Simms ES. 1956. Metaphosphate synthesis by an enzyme from Escherichia coli. *Biochim. Biophys. Acta* 20:215-227.

Koskela M, Annila A. 2012. Looking for the Last Universal Common Ancestor (LUCA). *Genes* 3:81–87.

Kosla J, Dvorak M, Cermak V. 2013. Molecular Analysis of the TGF-beta Controlled Gene Expression Program in Chicken Embryo Dermal Myofibroblasts. *Gene.* 513(1):90-100.

Kuzminov A. 2014. The Precarious Prokaryotic Chromosome. *J. Bacteriol.* 196:1793–1806.

Lambeth JD. 2004. NOX enzymes and the biology of reactive oxygen. *Nat Rev Immunol* 4:181–189.

Lambeth JD, Neish AS. 2014. Nox Enzymes and New Thinking on Reactive Oxygen: A Double-Edged Sword Revisited. *Annu. Rev. Pathol. Mech. Dis.* 9:119–145.

Lang AS, Zhaxybayeva O, Beatty JT. 2012. Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol* 10:472–482.

Lee STM, Keshavmurthy S, Fontana S, Takuma M, Chou W-H, Chen CA. 2018. Transcriptomic response in Acropora muricata under acute temperature stress follows preconditioned seasonal temperature fluctuations. *BMC Res Notes* 11:119.

Levene PA. 1919. The structure of yeast nucleic acid. *J. Biol. Chem.* 40:415-424.

Licata L, Lo Surdo P, Iannuccelli M, Palma A, Micarelli E, Perfetto L, Peluso D, Calderone A, Castagnoli L, Cesareni G. 2019. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Res* 48:D504–D510.

Liu X, Liu R, Zhao X-M, Chen L. 2013. Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. *BMC Med Genomics* 6:S8.

Lorenzen I, Mullen L, Bekeschus S, Hanschmann E-M. 2017. Redox Regulation of Inflammatory Processes Is Enzymatically Controlled. *Oxid Med Cell Longev* 2017:1–23.

Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Älgenäs C, Lundeberg J, Mann M, Uhlen M. 2010. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol* 6:450.

Lundberg J, Johansson BJE. 2015. Systemic resilience model. *Reliab. Eng. Syst. Saf.* Available from: http://dx.doi.org/10.1016/j. ress.2015.03.013

Mansfeld J, Urban N, Priebe S, Groth M, Frahm C, Hartmann N, Gebauer J, Ravichandran M, Dommaschk A, Schmeisser S, et al. 2015. Branched-chain amino acid catabolism is a conserved regulator of physiological ageing. *Nat Commun* 6:10043.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.

Marrocco I, Altieri F, Peluso I. 2017. Measurement and Clinical Significance of Biomarkers of Oxidative Stress in Humans. *Oxid Med Cell Longev* 2017:1–32.

Mast FD, Ratushny AV, Aitchison JD. 2014. Systems cell biology. *J. Cell Biol.* 206:695–706.

Medzhitov R. 2008. Origin and physiological roles of inflammation. *Nature* 454:428–435.

Meselson M, Stahl FW. 1958. The replication of DNA in Escherichia coli. *Proc Natl Acad Sci USA* 44:671-682.

Miao Z, Adamiak RW, Antczak M, Boniecki MJ, Bujnicki J, Chen S-J, Cheng CY, Cheng Y, Chou F-C, Das R, et al. 2020. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA* 26:982–995.

Michel F, Westhof E. 1990. Modelling of the Three-dimensional Architecture of Group I Catalytic Introns Based on Comparative Sequence Analysis. *J Mol Biol* 216:585-610.

Milivojevic V, Sinha R. 2018. Central and Peripheral Biomarkers of Stress Response for Addiction Risk and Relapse Vulnerability. *Trends. Mol. Med.* 24:173–186.

Moloney JN, Cotter TG. 2018. ROS signalling in the biology of cancer. *Semin. Cell Dev. Biol.* 80:50–64.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. 2011. How Many Species Are There on Earth and in the Ocean?Mace GM, editor. *PLoS Biol* 9:e1001127.

Movafagh S, Crook S, Vo K. 2015. Regulation of Hypoxia-Inducible Factor-1a by Reactive Oxygen Species : New Developments in an Old Debate: Regulation of Hypoxia-Inducible Factor-1a. *J. Cell. Biochem.* 116:696–703.

Mullis KB, Faloona FA. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Method Enzymol* 155:335-350.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–453.

Nevers Y, Kress A, Defosset A, Ripp R, Linard B, Thompson JD, Poch O, Lecompte O. 2019. OrthoInspector 3.0: open portal for comparative genomics. *Nucleic Acids Res.* 47:D411–D418.

Nikoletopoulou V, Markaki M, Palikaras K, Tavernarakis N. 2013. Crosstalk between apoptosis, necrosis and autophagy. *Biochim. Biophys. Acta* 1833:3448–3459.

Nirenberg M, Leder P. 1964. RNA Codewords and Protein Synthesis: The Effect of Trinucleotides upon the Binding of sRNA to Ribosomes. *Science* 145:1399–1407.

Noble D. 2012. A theory of biological relativity: no privileged level of causation. *Interface Focus.* 2:55–64.

Noble R, Tasaki K, Noble PJ, Noble D. 2019. Biological Relativity Requires Circular Causality but Not Symmetry of Causation: So, Where, What and When Are the Boundaries? *Front. Physiol.* 10:827.

Olson KR. 2012. Mitochondrial adaptations to utilize hydrogen sulfide for energy and signaling. *J Comp Physiol B* 182:881–897.

Palsson B. 2000. The challenges of in silico biology. *Nat Biotechnol* 18:1147–1150.

Pauling L, Zuckerkandl E. 1963. Chemical Paleogenetics. *Acta Chem. Scand.* 17:S9-S16.

Payne SH. 2015. The utility of protein and mRNA correlation. *Trends Biochem. Sci.* 40:1–3.

Perez-Castro C, Renner U, Haedo MR, Stalla GK, Arzt E. 2012. Cellular and Molecular Specificity of Pituitary Gland Physiology. *Physiol Rev* 92:1–38.

Pérez-Torres I, Guarner-Lans V, Rubio-Ruiz ME. 2017. Reductive Stress in Inflammation-Associated Diseases and the Pro-Oxidant Effect of Antioxidant Agents. *IJMS* 18:2098.

Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, et al. 2018. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* 556:339–344.

Podell S, Blanton JM, Neu A, Agarwal V, Biggs JS, Moore BS, Allen EE. 2019. Pangenomic comparison of globally distributed Poribacteria associated with sponge hosts and marine particles. *ISME J* 13:468–481.

Possik E, Pause A. 2015. Measuring Oxidative Stress Resistance of Caenorhabditis elegans in 96-well Microtiter Plates. *JoVE*:52746.

Reich HJ, Hondal RJ. 2016. Why Nature Chose Selenium. *ACS Chem. Biol.* 11:821–841.

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *GPB* 13:278–289.

Richardson MK, Keuck G. 2002. Haeckel's ABC of evolution and development. *Biol. Rev.* 77:495–528.

Rice VH. 2012. Theories of stress and its relationship to health. In V. H. Rice (Ed.), Handbook of stress, coping, and health: Implications for nursing research, theory, and practice (p. 22–42). Sage Publications, Inc.

Robert S, Gicquel T, Victoni T, Valença S, Barreto E, Bailly-Maître B, Boichot E, Lagente V. 2016. Involvement of matrix metalloproteinases (MMPs) and inflammasome pathway in molecular mechanisms of fibrosis. *Bioscience Reports* 36:e00360.

Rodriguez M, Snoek LB, De Bono M, Kammenga JE. 2013. Worms under stress: C. elegans stress response and its relevance to complex human disease and aging. *Trends Genet.* 29(6):367-374.

Sanchez W, Porcher J-M. 2009. Utilisation des biomarqueurs pour la caractérisation de l'état écotoxicologique des masses d'eau. *TSM*:29–38.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463-5467.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison III CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of the bacteriophage ϕX174 DNA. *Nature* 265:687-695.

Santolini J, Wootton SA, Jackson AA, Feelisch M. 2019. The Redox architecture of physiological function. *Curr Opin Physiol* 9:34–47.

Selye H. 1976. Forty years of stress research: principal remaining problems and misconceptions. *Can Med Assoc J.* 115(1):53-56.

Selye H. The Stress of life. United States of America: McGraw Hill Book Co, 1956.

Semenza GL. 2017. Hypoxia-inducible factors: coupling glucose metabolism and redox regulation with induction of the breast cancer stem cell phenotype. *EMBO J* 36:252–259.

Sharma P, Jha AB, Dubey RS, Pessarakli M. 2012. Reactive Oxygen Species, Oxidative Damage, and Antioxidative Defense Mechanism in Plants under Stressful Conditions. *J. Bot.* 2012:1–26.

Shi K-P, Dong S-L, Zhou Y-G, Li Y, Gao Q-F, Sun D-J. 2019. RNA-seq reveals temporal differences in the transcriptome response to acute heat stress in the Atlantic salmon (Salmo salar). *Comp. Biochem. Physiol. Part D Genomics Proteomics* 30:169–178.

Shimosaka T, Tomita H, Atomi H. 2016. Regulation of Coenzyme A Biosynthesis in the Hyperthermophilic Bacterium Thermotoga maritima. Becker A, editor. *J. Bacteriol.* 198:1993–2000.

Schüttler A, Reiche K, Altenburger R, Busch W. 2017. The Transcriptome of the Zebrafish Embryo After Chemical Exposure: A Meta-Analysis. *Toxicol. Sci* 157:291–304.

Sies H. 1997. Oxidative stress: oxidants and antioxidants. *Exp Physiol* 82:291–295.

Sonnhammer ELL, Koonin EV. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18:619–620.

Sonnhammer ELL, Östlund G. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* 43:D234–D239.

Southam CM, Erhlich J. 1943. Effects of extracts of western red-cedar heartwood on certain wood-decaying fungi in culture. *Phytopathology* 33:517-524.

Staden R. 1979. A strategy of DNA sequencing employing computer programs. *Nucl Acids Res* 6:2601–2610.

Stadtman ER, Levine RL. 2003. Free radical-mediated oxidation of free amino acids and amino acid residues in proteins. *Amino Acids* 25:207–218.

Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H. 2009. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *PNAS* 106:7702–7707.

Sutton WS. 1903. THE CHROMOSOMES IN HEREDITY. *Biol. Bull.* 4:231–250.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47:D607–D613.

Szypowska AA, Burgering BMT. 2011. The Peroxide Dilemma: Opposing and Mediating Insulin Action. *Antioxid Redox Sign* 15:219–232.

Tatusov RL. 1997. A Genomic Perspective on Protein Families. *Science* 278:631–637.

Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C, et al. 2016. Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci USA* 113:11901–11906.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

The Alliance of Genome Resources Consortium, Agapite J, Albou L-P, Aleksander S, Argasinska J, Arnaboldi V, Attrill H, Bello SM, Blake JA, Blodgett O, et al. 2020. Alliance of Genome Resources Portal: unified model organism research platform. *Nucleic Acids Res*. 48:D650–D658.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408:796-815.

The C. elegans Sequencing Consortium. 1998. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science* 282:2012–2018.

Todaro GJ. 1975. Evolution and Modes of Transmission of RNA Tumor Viruses. *Am. J. Clin. Pathol*. 81:590-606.

Toledano MB, Kumar C, Le Moan N, Spector D, Tacnet F. 2007. The system biology of thiol redox system in *Escherichia coli* and yeast: Differential functions in oxidative stress, iron metabolism and DNA synthesis. *FEBS Letters* 581:3598–3607.

Toledano MB, Leonard WJ. 1991. Modulation of transcription factor NF-kappa B binding activity by oxidation-reduction in vitro. *Proc. Natl. Acad Sci. USA* 88:4328-4332.

Tomita H, Yokooji Y, Ishibashi T, Imanaka T, Atomi H. 2012. Biochemical Characterization of Pantoate Kinase, a Novel Enzyme Necessary for Coenzyme A Biosynthesis in the Archaea. *J Bacteriol* 194:5434–5443.

Tribble DL, Jones DP. 1990. Oxygen dependence of oxidative stress. *Biochem. Pharmacol*. 39:729–736.

Vermeulen R, Schymanski EL, Barabási A-L, Miller GW. 2020. The exposome and health: Where chemistry meets biology. *Science* 367:392–396.

Vindry C, Ohlmann T, Chavatte L. 2018. Translation regulation of mammalian selenoproteins. *Biochim. Biophys. Acta* 1862:2480–2492.

Waddington CH. 1968. Towards a Theoretical Biology. *Nature* 218:525–527.

Wang Z et al. 2018. BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics*. 34:2867–2869.

Wassouf Z, Hentrich T, Casadei N, Jaumann M, Knipper M, Riess O, Schulze-Hentrich JM. 2019. Distinct Stress Response and Altered Striatal Transcriptome in Alpha-Synuclein Overexpressing Mice. *Front. Neurosci*. 12:1033.

Watson JD, Crick FHC. 1953. Molecular structure of nucleic acids. *Nature* 171:737-738.

Weigel D, Mott R. 2009. The 1001 Genomes Project for Arabidopsis thaliana. *Genome Biol* 10:107.

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509:582–587.

Woese CR. 1987. Bacterial evolution. *Microbiol Rev* 51:221-271.

Wyhe, John van ed., 2002- The Complete Work of Charles Darwin Online (http://darwin-online.org.uk/)

Yamamoto M, Kensler TW, Motohashi H. 2018. The KEAP1-NRF2 System: a Thiol-Based Sensor-Effector Apparatus for Maintaining Redox Homeostasis. *Physiol Rev* 98:1169–1203.

Zevian SC, Yanowitz JL. 2014. Methodological considerations for heat shock of the nematode Caenorhabditis elegans. *Methods* 68:450–457.

Zhou DR, Eid R, Miller KA, Boucher E, Mandato CA, Greenwood MT. 2019. Intracellular second messengers mediate stress inducible hormesis and Programmed Cell Death: A review. *Biochim. Biophys. Acta* 1866:773–792.

Zierer J, Menni C, Kastenmüller G, Spector TD. 2015. Integration of 'omics' data in aging research: from biomarkers to systems biology. *Aging Cell* 14:933–944.

Zuckerkandl E, Pauling LB. 1962. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New York, pp 189–225

Université de Strasbourg

Luc THOMES

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Étude génomique multi-espèces de la réponse adaptative au stress oxydatif

Résumé en français

Durant cette thèse, j'ai employé des approches de transcriptomique comparative pour identifier des processus biologiques impliqués dans la réponse au stress et l'adaptation. Cette étude a mis en évidence l'existence d'un jeu de gènes conservés entre des espèces de vertébrés. L'étude de ces gènes a révélé qu'ils codent majoritairement pour des protéines de la matrice extracellulaire et qu'ils sont co-exprimés, suggérant leur implication dans un même processus de régulation en réponse à un stress. Des études en réseau ont proposé que ces gènes seraient régulés par la voie de signalisation TGFß associée aux facteurs de signalisation CD44 et SRC. En parallèle, j'ai mené une étude de génomique comparative portant sur l'évolution de la voie de biosynthèse du pantothénate chez des bactéries du groupe *Candidatus poribacteria*. Cette étude a démontré une distribution en mosaïque de deux voies jusqu'alors considérées comme caractéristiques des bactéries et des archées. Par ailleurs, j'ai également développé un outil informatique, PROTEDEX, pour l'analyse intégrative et standardisée de listes de gènes, utilisant des données évolutives et fonctionnelles.

Mots-clé : réponse au stress, évolution, approche comparative, régulation par TGFß, pantothénate

English abstract

During this thesis project, I used comparative transcriptomic approaches to identify biological processes involved in stress response and adaptation. This study demonstrated the existence of a set of conserved genes between vertebrate species. The study of these genes revealed that they predominantly encode proteins of the extracellular matrix and that they are co-expressed, suggesting their involvement in a common regulatory pathway in response to stress. Network analyzes proposed that these genes are regulated by the TGFβ signaling pathway and associated with the signaling factors CD44 and SRC. In parallel, I conducted a comparative genomics study on the evolution of the pantothenate biosynthetic pathway in bacteria of the group *Candidatus poribacteria*. This study demonstrated a mosaic distribution of two pathways previously identified as characteristic of bacteria and archaea. In addition, I also developed a bioinformatic tool, PROTEDEX, for integrative and standardized analysis of gene lists, using evolutionary and functional data.

Keywords: stress response, evolution, comparative approach, TGFß regulation, pantothenate