Bioinformatic Tools for Exposome Data Analysis: Application to Human Molecular Signatures of Ultraviolet Light Effects

Carles Hernandez-Ferrer

THESIS DIRECTOR
Dr. Juan R. González Ruiz
Bioinformatic Research Group in Epidemiology
Barcelona Institute for Global Health

THESIS TUTOR
Dr. Luís. A. Pérez Jurado
Department of Experimental and Health Sciences
Universitat Pompeu Fabra

UPF Doctoral Thesis
– YEAR 2017 –

ISGlobal
Barcelona
Institute for
Global Health

upf. Universitat
Pompeu Fabra
Barcelona

A tu, Ricard

you don’t make mistakes,
mistakes make you

The Last Word (2017)
by Mark Pellington

Acknowledgements

I would like to start by expressing my most sincere gratitude to my director, Dr. Juan R. Gonzalez, who has trusted me from the very beginning of this journey, which started the first day I set foot in the old CREAL. Your guidance has greatly enriched me both as a person and as a scientist. I will never forget what you have done for me during this time.

I am also deeply grateful to my non-official mentor, Dr. Mariona Bustamante, for her valuable support, advice, and efforts in bringing forward the multiple projects in which we have worked together, including – of course – this thesis.

I want to thank without a doubt Carlos Ruiz for sharing his knowledge with me in the many chats in the corridor as well as during breakfast, conferences, and trips. Because you make it easy and pleasant to work alongside you. Special thanks to Marcos López, who has been a professional fellow since we started our master studies together. I hope this is not our last step together.

Three names are a “must” in this list: Ibon Tamayo (now in Boston), José Vargas (back to Brazil) and José Urquiza. Thanks Ibon for all the lunches and evenings we have spent together. Thanks for sharing your experiences about work, life, and love with me. Thanks José (Vargas) for keeping me in your mind and sharing your mobility experiences with me. Thanks José (Urquiza) for the breakfasts at the PRBB’s cafeteria and for the sporadic lunches; your opinion was, and is, very valuable for me. Many thanks to you three for sharing your research experience with me, for considering me in your ideas and projects, and for listening to me when I needed it.

Never forgotten, I have a place in this list to all the co-authors of the scientific articles listed in this document, who patiently read the documents and helped me improving the manuscripts.

I want to give a deep thank you to my parents; without you, I could never have become the person I am today. Thanks for trusting me even when I did not. To my grandfather Josep, for letting me play with all the machines in your workshop, and to my grandmother Carme, for never ceasing to try to make me pay attention to the "here-and-now" instead of losing myself in the imagination. To my cousin Marc Berenguer and to Shinju Park for being a model of integrity, and for seeding in me the path of curiosity that led me to complete this thesis. Also to the other members of the family, for they tireless support along this adventure.

Un racó especial d'aquesta llista és per a la iaia Teresa i el iaio Joan. Sense el seu afecte incondicional, les seves ganes d'escoltar-me i les el seu esforç per incentivar-me a donar sempre el millor de mi ara no estaria aquí.

A piece of me is going with my two sisters-in-law Maria and Carla Giner. Because, with your personal motivation and your commitment, you are a role model for all of us. And another piece goes with "els meus nens", Maria and David, because during the months I have been guiding your *treball de recerca* you showed an enthusiasm I have never seen before.

Heartfelt thanks also to "my two brothers" Marc Girons and Joan Puigserver. You have been always on my side and you supported me every time I have been full of doubts. To Mireia Vidal and Xavi Clotet, for making me discover once and again the magic of friendship.

Finally, thanks to the love of my life. The most genuine thanks. For your endless love, patience, and unbreakable faith in me. You may not know it, but you are a true example of constant self-improvement and my true inspiration.

Barcelona, September 2017

Preface

This thesis was written at the Campus Mar of the Global Health Institute Barcelona (ISGlobal), former Centre for Research in Environmental Epidemiology (CREAL), between 2013 and 2017, and it was supervised by Dr. Juan R. González. This work consists of a compilation of the scientific publications co-authored by the PhD candidate according to the procedures of the Biomedicine PhD program of the Department of Experimental and Health Sciences of Universitat Pompeu Fabra.

The present document includes: 1) its abstract, 2) a general introduction, 3) two blocks of results with their own rationale, methods and discussions, 4) final conclusions, and 5) a description of future work.

This thesis focuses on the analysis of the *exposome* - understood as the complete set of exposures a human being is in contact from conception to death - and its molecular signatures.

The first part of the thesis aims to study the effects of artificial ultraviolet radiation on transcription in humans to disentangle the molecular mechanisms underlying the last effects of ultraviolet radiation on heath. In particular, I display two original scientific articles about its acute effects on blood and skin human transcriptome, at level of gene and micro-RNA expression. The second part of the thesis is based on the tools required to analyses the *exposome*. In this part I present three original scientific papers that corresponds to a set of four tools. The tools were developed as R packages and cover the topics of: *exposome* data management, *exposome* data characterization and analysis, and *exposome* enrichment. Two of this tools are already in Bioconductor and all them for are available through GitHub platform (`http://github.com/isglobal-brge`).

At the end of this document, in the “future work” chapter, I introduce the HELIX project. The aim of this project is to characterize the early-life *exposome* to advance our knowledge into the causal relationship between the *exposome* and human health. The tools presented as result of this thesis are currently being used in the HELIX project.

The printed version of this document is accompanied by a CD that includes: A) a PDF copy of it, B) the supplementary materials of the articles presented in chapter 3, C) the supplementary materials of the articles included in chapter 4, and D) a registry of the copyright of each one of the figures - from third-party scientific articles - used in this thesis.

Abstract

Most common diseases are caused by a combination of genetic, environmental and lifestyle factors. These diseases are referred to as complex diseases. Examples of this type of diseases are obesity, asthma, hypertension or diabetes. Several empirical evidence suggest that exposures are necessary determinants of complex disease operating in a causal background of genetic diversity. Moreover, environmental factors have long been implicated as major contributors to the global disease burden. This leads to the formulation of the *exposome*, that contains any exposure to which an individual is subjected from conception to death. The study of the underlying mechanics that links the *exposome* with human health is an emerging research field with a strong potential to provide new insights into disease etiology.

The first part of this thesis is focused on ultraviolet radiation (UVR) exposure. UVR exposure occurs from both natural and artificial sources. UVR includes three subtypes of radiation according to its wavelength (UVA 315-400 nm, UVB 315-295 nm, and UVC 295-200 nm). While the main natural source of UVR is the Sun, UVC radiation does not reach Earth's surface because of its absorption by the stratospheric ozone layer. Then, exposures to UVR typically consist of a mixture of UVA (95%) and UVB (5%). Effects of UVR on human can be both beneficial and detrimental, depending on the amount and form of UVR. Detrimental and acute effects of UVR include erythema, pigment darkening, delayed tanning and thickening of the epidermis. Repeated UVR-induced injury to the skin, may ultimately predispose one to the chronic effects photoaging, immunosuppression, and photocarcinogenesis. The beneficial effect of UVR is the

cutaneous synthesis of vitamin D. Vitamin D is necessary to maintain physiologic calcium and phosphorous for normal bone mineralization and to prevent rickets, osteomalacia, and osteoporosis.

But the *exposome* paradigm is to work with multiple exposures at a time and with one or more health outcomes rather focus in a single exposures analysis. This approach tends to be a more accurate snapshot of the reality that we live in complex environments. Then, the second part is focused on the tools to explore how to characterize and analyze the *exposome* and how to test its effects in multiple intermediate biological layers to provide insights into the underlying molecular mechanisms linking environmental exposures to health outcomes.

Resumen

Las enfermedades complejas se encuentran entre las mas comunes y son causadas por una combinación de factores genéticos y ambientales (contaminación ambiental, estilo de vida, etc). Entre las enfermedades complejas que se pueden destacar se encuentran la obesidad, el asma, la hipertensión o la diabetes. Diversos estudios científicos sugieren que el hecho de padecer enfermedades complejas esta condicionado a la aparición o acumulación de determinados factores ambientales. Asimismo, se ha descrito que los factores ambientales son unos de los principales contribuyentes a la carga mundial de morbilidad. Todo esto nos lleva a definir el término *exposoma* como el conjunto de factores ambientales a los que un individuo se ve expuesto desde la concepción hasta la muerte. El estudio de la mecánica subyacente que vincula el *exposoma* con la salud es un campo de investigación emergente con un fuerte potencial para proporcionar nuevos conocimientos sobre la etiología de las enfermedades.

La primera parte de esta tesis se centra en la exposición a la radiación ultravioleta. La exposición a la radiación ultravioleta proviene de fuentes tanto naturales como artificiales. La radiación ultravioleta incluye tres subtipos de radiación según su longitud de onda (UVA 315-400 nm, UVB 315-295 nm y UVC 295-200 nm). Si bien la principal fuente natural de radiación ultravioleta es el Sol, la UVC no llega a la superficie de la Tierra debido a su absorción por la capa estratosférica de ozono. En consecuencia, la exposición a radiación ultravioleta a la que estamos usualmente sometidos consisten en una mezcla de UVA (95 %) y UVB (5 %). Los efectos de la radiación ultravioleta en humanos pueden ser beneficiosos o perjudiciales dependiendo de su cantidad y forma. Los efectos perjudiciales

y agudos de la radiación ultravioleta incluyen eritema, oscurecimiento del pigmento, retraso en el bronceado y engrosamiento de la epidermis. Repetidas lesiones en la piel producidas por radiación ultravioleta pueden predisponer, en última instancia, a efectos crónicos de fotoenvejecimiento, inmunosupresión y fotocarcinogénesis. El mayor efecto beneficioso de la radiación ultravioleta es la síntesis cutánea de la vitamina D. La vitamina D es necesaria para mantener el calcio fisiológico y del fósforo para la mineralización ósea y para prevenir el raquitismo, la osteomalacia y la osteoporosis.

El paradigma del *exposoma* es trabajar con múltiples exposiciones a la vez en vez centrarse en una sola exposición. Este enfoque permite tener una visión más parecido a la realidad que vivimos. Luego, la segunda parte se centra en las herramientas para explorar cómo caracterizar y analizar el *exposoma* y cómo probar sus efectos en múltiples capas biológicas intermedias para proporcionar información sobre los mecanismos moleculares subyacentes que vinculan las exposiciones ambientales a los resultados de salud.

Resum

Les malalties complexes es troben entre les més comuns i són causades per una combinació de factors genètics i ambientals (contaminació ambiental, estil de vida, etc.). Entre les malalties complexes que es poden destacar es troben l'obesitat, l'asma, la hipertensió o la diabetis. Diversos estudis científics suggereixen que el fet de desenvolupar malalties complexes està condicionat a l'aparició o l'acumulació de determinats factors ambientals. Seguint amb aquesta línia, s'ha descrit que els factors ambientals són uns dels principals contribuents a la càrrega mundial de morbiditat. Tot això ens porta a definir el terme *exposoma* com el conjunt de factors ambientals als quals un individu es veu exposat des de la seva concepció fins a la mort. L'estudi de la mecànica subjacent que vincula el *exposoma* amb la salut és un camp de recerca emergent amb un fort potencial per proporcionar nous coneixements sobre l'etiologia de les malalties.

La primera part d'aquesta tesi es centra en l'exposició a la radiació ultraviolada. L'exposició a la radiació ultraviolada prové de fonts tant naturals com artificials. La radiació ultraviolada inclou tres subtipus de radiació segons la seva longitud d'ona (UVA 315-400 nm, UVB 315-295 nm i UVC 295-200 nm). Si bé la principal font natural de radiació ultraviolada és el Sol, la UVC no arriba a la superfície de la Terra a causa de la seva absorció per la capa estratosfèrica d'ozó. En conseqüència, l'exposició a radiació ultraviolada a la qual estem sotmesos usualment consisteixen en una barreja d'UVA (95 %) i UVB (5 %). Els efectes de la radiació ultraviolada en humans poden ser beneficiosos o perjudicials depenent de la seva quantitat i forma. Els efectes perjudicials i aguts de la radiació ultraviolada inclouen eritema, enfosquiment del pigment, retard en el bronzejat i engrossiment

de l'epidermis. Repetides lesions a la pell produïdes per radiació ultraviolada poden predisposar, en última instància, a efectes crònics de fotoenvelliment, immunosupressió i fotocarcinogènesi. El major efecte beneficiós de la radiació ultraviolada és la síntesi cutània de la vitamina D. La vitamina D és necessària per mantenir el calci fisiològic i del fòsfor per a la mineralització òssia i per prevenir el raquitisme, l'osteomalàcia i l'osteoporosi.

El paradigma de l'*exposoma* és treballar amb múltiples exposicions al mateix temps en comptes de focalitzar-se en una sola exposició. Aquest enfocament permet tenir una visió més semblant a la realitat que vivim. Després, la segona part del document se centra en les eines per explorar com caracteritzar i analitzar l'*exposoma* i com provar els seus efectes en múltiples capes biològiques intermèdies per proporcionar informació sobre els mecanismes moleculars subjacents que vinculen les exposicions ambientals als resultats de salut .

Thesis Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Complex Diseases

Many of the most common diseases such as obesity, asthma, hypertension or diabetes are caused by a combination of genetic, environmental and lifestyle factors. These diseases are called complex diseases.

A central question in biology is whether observed variation in a particular disease is due to environmental or to genetic factors. Heritability is a concept that summarizes how much of the variation in a disease is due to variation in genetic factors [1]. Other causes of measured variation in a trait are characterized as environmental factors. In human studies of heritability, these are often apportioned into factors from “shared environment” and “non-shared environment” based on whether they tend to result in persons brought up in the same household more or less similar to persons who were not.

Since heritability is estimated by comparing individual disease variation among related individuals (in a population), heritability is specific to a particular population in a particular environment.

Genome-wide association studies (GWAS) have been proven a powerful tool for investigating the genetic architecture of complex diseases [2], [3], in which several hundred thousand to more than a million single nucleotide polymorphisms (SNPs) are assayed in thousands of individuals.

The underlying rationale for GWAS is that they look for common variants for common diseases. So on, common diseases are highly attributable to allelic variants present in more than 1–5% of the population [4], [5]. The allelic architecture of some conditions reflects the contributions of several variants of great effect. In spite of this, the most common variants individually or in combination confer relatively small increases in risk and explain only a small proportion of heritability (a portion of phenotypic variance in a population attributable to additive genetic factors) [6].

Table 1.1, from Manolio et. al. [7], summarizes the estimated heritability for several complex traits. Age-related macular degeneration may provide the best example of a common disease in which heritability is substantially explained by a small number of common variants of large effect, but for other conditions, such as Crohn's disease, the proportion of heritability explained is not nearly so large despite a much larger number of identified variants.

Table 1.1: Estimates of heritability and number of loci for several complex traits, from Manolio et. al. [7].

Disease	# Loci	Herit. Explained	Herit. Measure	Ref.
Age-related macular degeneration	5	50%	Sibling recurrence risk	[8]
Crohn's disease	32	20%	Genetic risk (liability)	[9]
Systemic lupus erythematosus	6	15%	Sibling recurrence risk	[10]
Type 2 diabetes	18	6%	Sibling recurrence risk	[11]
HDL cholesterol	7	5.2%	Residual-phenotypic variance*	[12]
Height	40	5%	Phenotypic variance	[13]
Early onset myocardial infarction	9	2.8%	Phenotypic variance	[14]
Fasting glucose	4	1.5%	Phenotypic variance	[15]

*Residual is after adjustment for age, gender, diabetes.

After 10 years of GWAS [16], it has been shown:

- Complex Diseases Are Highly Polygenic. More than 10,000 associations have been reported between genetic variants and one or more complex diseases through GWA studies. GWAS associations have proven highly replicable, both within and between populations under the assumption of adequate sample sizes. A conclusion from GWASs is that for almost any complex disease, many loci contribute to standing genetic variation. This means that for most diseases polymorphisms in many genes contribute to genetic variation in the population.

- Pleiotropy Is Pervasive. This means that the paradigm of “one gene, one function, one trait” is the wrong way to view genetic variation in humans. This conclusion appears after: 1) mendelian mutations for a specific diseases are frequently associated with other traits; 2) pedigree studies have reported genetic correlations between traits and diseases (same variants affects two or more diseases at the same time); and 3) analytical methods that estimate genetic correlations from GWAS data have provided evidence for widespread pleiotropy. Then, the study of diseases in isolation might lead to the wrong inference.

- The Importance of Sample Size to Detect Association. The number of discoverable loci associated with a specific disease depends on the disease and on sample size. But all show a sharp increase at a critical sample size. This observation implies that larger experimental sample sizes will lead to new discoveries, and that is exactly what has occurred over the last decade. To date, there has been no trait with evidence of a plateau of the number of risk loci discovered with increasing sample size.

- First Steps of Personalized Medicine. A long the last decade, the experimental design of GWAS led to remarkable discoveries in human genetics: understanding of the genetic architecture of some complex diseases; the discovery of variants and genes that play a relevant role in biological pathways for complex diseases; and providing sets of candidates of therapeutics and molecular targets. Into the future, “personalized medicine” will use knowledge and strategies (as well as prevention interventions or risk stratification) that, directly or indirectly, will be built on information obtained in first step by GWAS results.

The questions arise as to why so much of the heritability is apparently unexplained by GWAS findings. It is important to find an answer to this question because a substantial proportion of individual differences facing disease susceptibility is known to be due to genetic factors. Then, understanding these genetic variations may contribute to prevention, diagnosis and treatment of diseases. GWAS have identified hundreds of variants in many dozens of traits, but for many traits they have explained only a small proportion of estimated heritability [17].

While any explanations for this missing heritability have been suggested consensus is lacking. Proposed explanations includes much larger numbers of variants of smaller effect yet to be found; rarer variants with possibly larger effects that are poorly detected by available and commercial genotyping arrays (that focus on variants present in 5% or more of the population); structural variants poorly captured by existing arrays; low power to detect gene-gene interactions; and inadequate accounting for shared environment among relatives [7].

It is reasonable to assume that allelic architecture (number, type, effect size and frequency of susceptibility variants) may differ across traits. Also, that missing heritability may take a different form for different diseases [18]. Unfortunately, current knowledge is too limited to distinguish these possibilities [7].

Immune and infectious agents have been recognized as among the strongest selection pressures in human evolution [19], and immune-related genes have been strongly implicated in Crohn's disease and other immune-mediated diseases [6], suggesting either that pleiotropic effects of these variants reduce the efficiency of negative selection or that strong environmental perturbation in modern societies might expose the disease risk associated with these variants.

In this line, several empirical evidence suggest that exposures are necessary determinants of complex disease and that operate in a causal background of genetic diversity [20].

For instance, genetically-stable populations experience profound alterations in cancer incidence across generations and in migrations which have been subjected to different environments [21]–[23].

1.2 Global Burden Disease of Environmental Exposures

The Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) is the largest and most comprehensive effort to date to measure epidemiological levels and trends worldwide [24]. Led by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington, the GBD quantifies the comparative magnitude of health loss to diseases, injuries, and risk factors by age, sex, and geography over time. One of their last update published in 2015, evaluated 300 diseases and injuries in 195 countries (by age and sex), from 1990 to 2013 [25].

In this study, they focused on three groups of risk: “behavioral”, “environmental and occupational”, “and metabolic”. Risks factors were organized into a four level hierarchy with first level blocks of *environmental and occupational*, *behavioural* and *metabolic*. the next level in the hierarchy included nine clusters of relative risks with more detail in the third and four levels.

The *environmental and occupational*, risks included “unsafe water, sanitation, and hand-washing” (with 3 nested risk factors), “air pollution” (with 3 nested risk factors), “other environmental risks” (with 2 nested risk factors) and “occupational risks” (with 6 nested risk factors; one of them, “occupational carcinogens” included another 14 nested risk factors). The *behavioural* risks contained “child and maternal malnutrition” (with 5 nested risk factors from that “suboptimal breastfeeding” has another 2 nested risk factors and “childhood under-nutrition” has 3 nested risk factors), “tobacco smoke” (with 2 nested risk factors), “alcohol and drug use” (with 2 nested risk factors), “dietary risks” (with 14 nested risk factors), “sexual abuse and violence” (with 2 nested risk factors), “unsafe sex” and “low physical activity”. Finally, *metabolic* risks included 6 nested risk factors (with body-mass index and blood pressure among others).

They adopted the World Cancer Research Fund grading the evidences assessing an association between a risk factor exposures and an outcome in four levels: 1) convincing, 2) probable, 3) possible, and 4) insufficient. Only convincing and probable risk outcome pairs were taken into account.

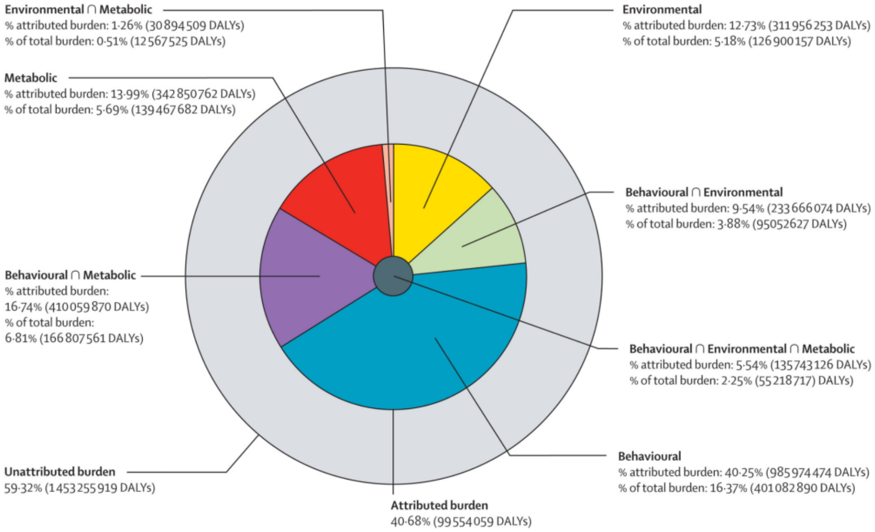


Figure 1.1: Proportion of all-cause DALYs attributable to behavioural, environmental and occupational, and metabolic risk factors and their overlaps for all ages in 2013, from GBD 2013 [25].

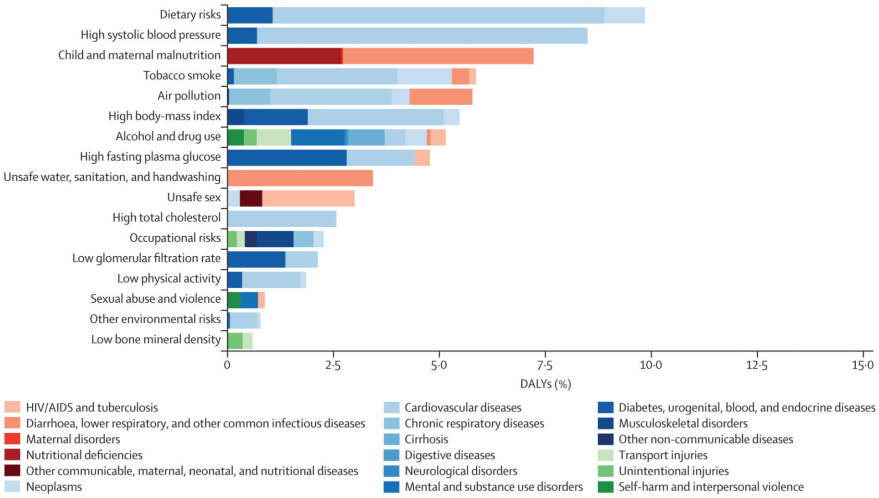


Figure 1.2: Global DALYs attributed to level 2 risk factors in 2013, adapted from GBD 2013 [25].

All risks combined account for 57.2% (95% uncertainty interval 55.8–58.5) of deaths and 41.6% (40.1–43.0) of disability-adjusted life-years (DALYs). As shown in Figure 1.1, 17.73% DALYs attributable to environmental risk factors (for all ages in 2013). Figure 1.2 shows effects of different categories of environmental risk factors by disease.

Two exposures studied by the GBD 2013 and GBD 2015 are relevant in terms of public health: *tobacco smoke* (from behavioural) and *air pollution* (from environmental and occupational). The first due to its large effects effects [26], air pollution due to the wide range of affected populations [27].

The case of "tobacco smoke"

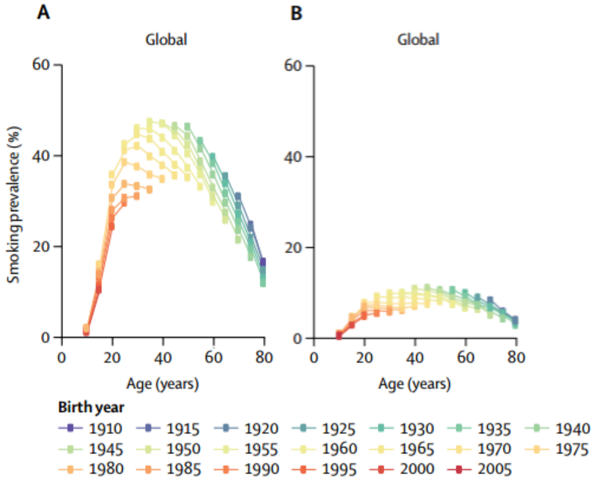


Figure 1.3: Prevalence of daily smoking over time at the global level by men (A) and women (B), adapted from GBD Smoking 2015 [26].

In 2015 smoking was the second leading risk factor for early death and disability worldwide. Claiming more than 5 million lives every year since 1990 its contribution to overall disease burden stills growing.

The GBD Smoking 2015 [26] investigated differences in smoking prevalence and attributable burden according to the Socio-demographic Index (SDI). Figure 1.3 shows the prevalence patterns by year according to age. Additionally, they assess age and sex patterns by birth cohort across levels of development and performed a decomposition analysis of potential drivers of smoking attributable disease burden over time.

Worldwide in 2015, the age-standardized prevalence of daily smoking was 25% (95% UI 24.2–25.0,) in men and 5.4% (5.1–5.7) in women. 51 countries and territories had significantly higher prevalence of smoking than the global average for men, and these countries were located mainly in central and eastern Europe and south-east Asia. For women, 70 countries, mainly in western and central Europe, significantly exceeded the global average [26]. Figure 1.4 shows the ranking of smoking as a risk factor worldwide.

Despite a global decrease, several countries still had a high prevalence of smoking among individuals aged between 15 and 19 years.

In 2015, 6.4 million deaths (95% UI 5.7–7.0) were attributable to smoking worldwide, representing a 4.7% (1.2–8.5) increase in smoking-attributable deaths since 2005. In both 2005 and 2015, smoking was the second-leading risk factor for attributable mortality among both sexes, following high-systolic blood pressure.

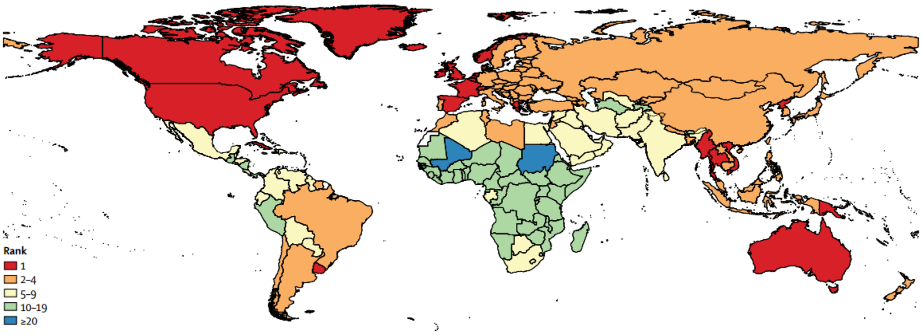


Figure 1.4: Rankings of smoking as a risk factor for all-cause, all-age attributable DALYs for both sexes combined in 2015, fragment from GBD Smoking 2015 [26].

There were 148.6 million (95% UI 134.2–163.1) smoking-attributable DALYs worldwide in 2015. Moreover, as can be seen in Figure 1.4, smoking was the leading risk factor for attributable disease burden in 24 countries.

Overall, in 2015, cardiovascular diseases (41.2%), cancers (27.6%), and chronic respiratory diseases (20.5%) were the three leading causes of smoking-attributable age-standardized DALYs for both sexes. Of all risk factors, smoking was the leading risk factor for cancers and chronic respiratory diseases [26].

The case of "air pollution"

Exposure to ambient air pollution increases mortality and morbidity and shortens life expectancy [28], [29]. GBD 2015 estimated the burden of disease attributable to 79 risk factors in 195 countries from 1990 to 2015. GBD 2015 identified air pollution as a leading cause (in top 10 leading causes from 2005 to 2015) of global disease burden, especially in low-income and middle-income countries [25].

A comparison of the percentage change in risk exposure from 1990 to 2015 with the level of attributable DALYs in 2015 helps to identify large risks for which a long-term increase in global exposure has occurred.

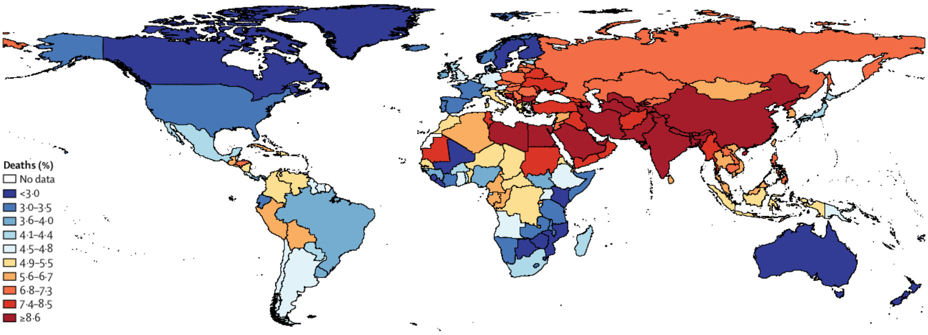


Figure 1.5: Deaths attributable to ambient particulate matter pollution in 2015, fragment from GBD Air Pollution 2015 [27].

Air pollution is a complex mixture of gases and particles whose sources and composition vary spatially and temporally. Population-weighted annual mean concentrations of particle mass with the aerodynamic diameter less than 2.5 μm (PM 2.5) and tropospheric ozone are the two indicators used to quantify exposure to air pollution. PM 2.5 is the most consistent and robust predictor of mortality in studies of long-term exposure [30], [31]. Ozone, a gas produced via atmospheric reactions of precursor emissions, is associated with respiratory disease independent of PM 2.5 exposure [32], [33].

Deaths attributable to long-term exposure to PM 2.5 in 2015 varied substantially among countries (Figure 1.5). also along time were PM 2,5 increased by 11.2% from 1990 (39.7 μg/m^3) to 2015 (44.2 μg/m^3), increasing most rapidly from 2010 to 2015.

In high-income countries, exposure to ambient PM 2.5 contributed to 4.3% of total deaths in 2015 versus 9.0% in upper-middle income, 8.7% in lower-middle-income, and 4,9% in low-income countries. These differences in attributable mortality mostly reflect the fraction of total deaths from cardiovascular disease among countries.

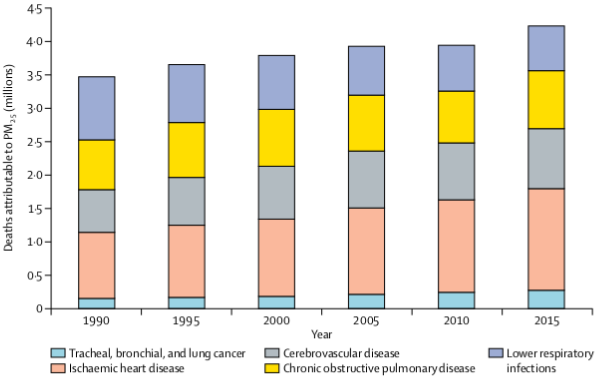


Figure 1.6: Deaths attributable to ambient particulate matter pollution by year and disease, from GBD Air Pollution 2015 [27].

Cohen et. al. estimated the burden attributable to PM 2.5 for ischaemic heart disease (IHD), cerebrovascular disease (ischaemic stroke and hemorrhagic stroke), lung cancer, chronic obstructive pulmonary disease

(COPD), and lower respiratory infections (LRI) [34]. Evidence linking these diseases with exposure to ambient air pollution was judged to be consistent with a causal relationship on the basis of criteria specified for GBD risk factors [25].

Finally, they conclude that long-term exposure to PM 2.5 contributed to 4.2 million deaths in 2015, representing 7,6% of total global deaths. Household air pollution from solid fuel use was responsible for 2.8 million deaths (Figure 1.6).

1.3 The Exposome Concept

As previously illustrated, environmental factors have long been implicated as major contributors to the global disease burden [25]. This lead to the formulation of the *exposome*, concept first described by Wild on 2005.

The *exposome* is composed of every exposure to which an individual is subjected from conception to death. Therefore, it requires consideration of both the nature of those exposures and their changes over time [35]. The consideration of the nature of the exposures generates three domains (Figure 1.7): internal environment, specific external environment and general external environment [36].

First, the general external exposures domain includes the wider social, economic and psychological influences on the individual, for example: social capital, education, financial status, psychological and mental stress, urban-rural environment and climate. Second, the specific external exposures is an extensive range which includes infectious agents, chemical contaminants, diet, lifestyle factors (e.g. tobacco, alcohol...), occupation and medical interventions. Last, the exposures in the internal domain include all the internal biological processes in response to the external exposures domain, to maintain homeostasis and which are influenced by the genome (further in this thesis called molecular signatures).

Measures in one domain or another may reflect to differing degrees one component of the *exposome*, e.g. the urban environment (general external), pesticides (specific external) and inflammation (internal) [36].

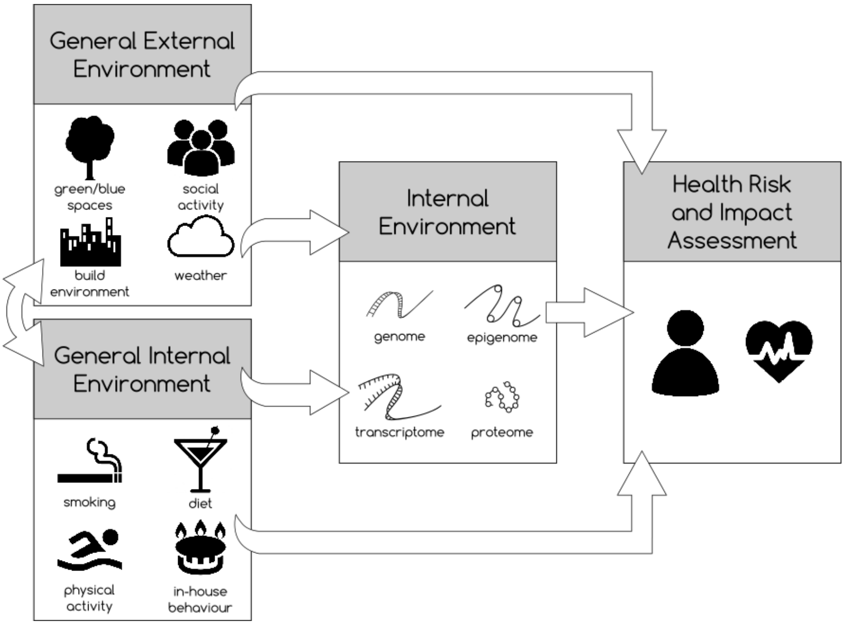


Figure 1.7: The effects and interactions between the different domains constituting the exposome (specific and general external environments and internal environment) and health risk. Figure inspired by Vrijheid [37].

This original concept proposed by Wild was further expanded by Rappaport and Smith [38], who functionalized the *exposome* in terms of circulating chemicals in the body that reflect both exogenous and endogenous exposures. In other words, the *exposome* represents the combined exposures from all sources that reach the internal chemical environment. Subsequently, Miller and Jones refined the concept.

There are 3 distinct differences between Wild (*old*) definition and Rappaport and Smith plus Miller & Jones (*new*) definition:

1. The concept of the cumulative biological responses, representing body's response to external forces and chemicals.
2. The inclusion of behavior, including lifestyle as a dynamic interaction with our surroundings, our relationships, our interactions, and physical and emotional stressors.

3. The addition of "endogenous processes" that are affected by complex exposures. Our bodies are complex biochemical reaction vessels with countless reactions occurring at any time. The lingering damage seen as DNA mutations, epigenetic alterations, protein modifications... is the evidence of a real effect and may be present decades after exposure.

This last version of the definition of the *exposome* led to understand that the levels of endogenous molecules (internal *exposome*) and specific external exposures do not need to be the same.

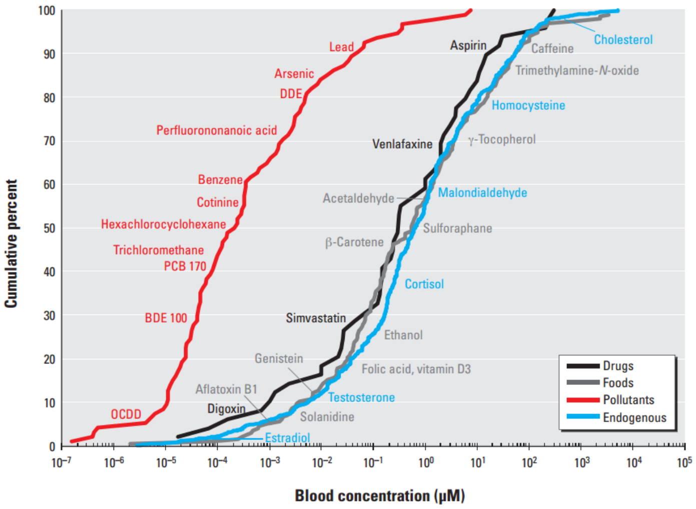


Figure 1.8: Each curve represents the cumulative distribution of chemical concentrations from a particular source category, from Rappaport et. al. [22].

So, as proposed by Rappaport et. al. 2014, measuring the *exposome* in human blood offers am an interesting approach for interrogating biologically relevant exposure-associated processes, because blood transports chemicals to and from tissues and represents a reservoir of all endogenous and exogenous chemicals in the body at a given time [39]. Figure 1.8 shows the cumulative distributions of blood concentrations for the four sources of chemicals studied by Rappaport et. al. 2014 (endogenous chemical - with 1,223 elements -, food chemical - with 195 elements -, pollutant -

with 94 elements - and drug - with 49 elements).

The dynamic nature of the *exposome* presents one of the most challenging features of its characterization. To fully characterize an individual's *exposome* would require either sequential measures that spanned a lifetime (Figure 1.9). Therefore, innumerable cross-sectional measures of the exposure profile building to a continuous real-time monitoring will be required, which cumulatively would represent the *exposome* of the individual.

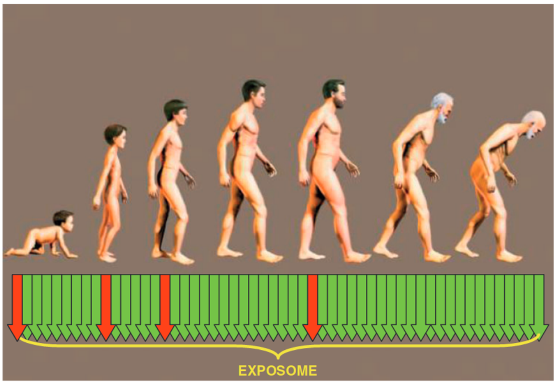


Figure 1.9: Exposome requires multiple measurements over human life including in utero exposures (prenatal *exposome* - not included in the schema), from Wild [36].

The *exposome* captures the essence of nurture; it is the summation and integration of external forces acting upon our genome throughout our lifespan [40]. This measurable quantity of the *exposome* represents a biological index of our nurture and is the context in which specific exposures have an impact on health [41].

Exposure during fetal or early life to environmental chemicals has been associated with adverse fetal growth and with developmental neurotoxic and immunotoxic effects in children [42], [43]. This clears up with two situations: 1) up to now the environment and child health field focused on single exposure health effect relationships [44], and 2) environmental exposures during fetal stages linked with structural and functional changes in later life stages, predisposing to disease [45].

The role of the impact of prenatal *exposome* in human health has been highly explored. There is evidence that shows that manipulation of the environment in the prenatal and infancy stages can be associated with permanent changes in physiology and/or structure:

- The link between the nature of infant feeding to later health consequences [46].
- The relationship between birth size and later risk of disease [47]–[49].
- Prematurity independent of growth retardation is associated with long-term metabolic consequences [50].
- Offspring of women subjected to severe undernutrition in early pregnancy did not have reduced birth weight but do have an increased risk of obesity [51].

Many of these environmental changes are associated with permanent alterations in gene expression regulated by epigenetic factors such as DNA methylation and histone alteration. Gluckman et. al proposed the "developmental origins of health and disease" (DOHaD) paradigm that leads to the recognition that early life influences can alter later disease risk. DOHaD phenomenon can be considered as a subset of the broader processes of developmental plasticity by which organisms adapt to their environment. The adaptive processes allow genotypic variation to be preserved through transient environmental changes and they may affect a single organ or system, but generally, they induce integrated adjustments in the mature phenotype, a process underpinned by epigenetic mechanisms and influenced by prediction of the mature environment [45].

The study of the underlying mechanics that links the *exposome* with human health is an emerging research field with a strong potential to provide new insights into disease etiology [52].

One example of a successful effort is the National Health and Nutrition Examination Survey (NHANES). Measured factors include environmental exposures such as chemicals, nutrients, and infectious agents. It also includes other indicators of environmental exposures such as self-reported nutrient consumption, physical activity, and prescribed pharmaceutical drugs [53]. With systematic information on exposures, environment-wide association studies (ExWAS) could become much more powerful and complement GWAS and deep sequencing studies [54].

1.4 Molecular Signatures of Exposome

Research has clearly established that the environment plays a significant role in our health and in the development of diseases. At the same time studies of genetic variants and disease have been conducted to reveal links between environmental exposures and health outcomes. Other studies have identified environmental factors as significant contributors to disease, yet the specific exposures of concern are poorly defined [55], [56].

The main goal in *exposome* analysis is to understand how chemicals are altering human biology to explain its association with human health outcomes [57]. Such effects could include binding to macromolecules, inducing structural changes and disruption of biological pathways. The need remains for a systematical evaluation of the environmental contributors to health and disease [58], [59].

Although the term *biomarker* refers to any measurable state in a living organism, a useful biomarker can differentiate between biological states, particularly those represented by diseased and healthy populations. Discovery of new biomarkers is important for epidemiology, which seeks causes of diseases (biomarkers of exposure), as well as for diagnosis and treatment of diseases (biomarkers of disease). Moreover, biomarkers of exposure can also take into account a given specific *exposome* period (prenatal, early life, adulthood...). To this end, the different layer of molecular signatures, as well as their relations must be taken into account. Figure 1.10 illustrates a simplification of the hierarchy between the different molecular signatures.

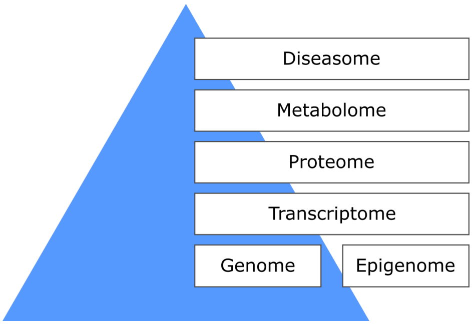


Figure 1.10: Simplification of the hierarchy of the molecular signatures (different omic data types).

At the bottom of the pyramid both the *genome* and *epigenome* are located. While the *genome* includes the codification of any function any cell type can do, the *epigenome* is in charge to cover and uncover those genomic regions needed by each one of the cell types. For this, any aberrations in the *genome*, such as copy number variations (CNVs) or chromosomal inversions, and any epigenetic perturbations, as changes in methylation patterns or histone modifications, may have an impact in the *transcriptome*. Modifications in *transcriptome* levels affects the levels and types of *proteome*, that at its time effects, joint with external risk factors, the proportions and patterns in *metabolome*. The joint effect of the cascade modifications in each level of the pyramid triggers the alterations in *diseaseome* (also called *phenome*).

Example of defining biomarkers of exposure is the case of Reese et. al. They looked for biomarkers in newborns of sustained smoking by the mother during pregnancy. They result in a set of methylation probes (CpGs) that can be easily applied to other newborn studies having methylation data and lacking cotinine levels [60].

Disease biomarker discovery has grown over the last years guiding the development of drugs and diagnostic products. Moreover, it has vigorously embraced the omics revolution and thereby offers hope that whole new classes of biomarkers of disease will be found. Parallel developments of biomarkers of exposure have been more modest, not only because these biomarkers lack clear commercial interests, but also because knowledgedriven designs are still favored over omic tools for characterizing exposures.

1.4.1 The Three Elements of Epigenetics

Epigenetics is defined as the study of heritable changes in gene expression that occur without changes in DNA sequence [61]. Epigenetic mechanisms are flexible genomic parameters that can change genome function under exogenous and endogenous influences. There are three (see Figure 1.11): DNA methylation, histone modifications and miRNA expression.

This layer of regulatory information is essential for proper development of cellular function. The genome is static and present in all cells, the epigenome is variable by cell, tissue and developmental stage. then, it determines of cellular functions and identity. These mechanisms also represent an adaptive intermediary that interprets and responds to environmental factors, resulting in alterations in the transcriptome.

Epigenome patterns have been characterized in different tissues and time points in international projects such as ROADMAP [62] and ENCODE [63]. They do not act alone but in combination determining chromatin states which have specific regulatory functions (i.e. enhancers vs repressors) [64].

5-methylcytosine (5MeC) represents 2-5% of all cytosines in mammalian genomes and is found primarily on CpG dinucleotides. Methylation is involved in regulating many cellular processes, including chromatin structure and remodeling, X-chromosome inactivation, genomic imprinting, chromosome stability, and gene transcription [65], [66]. Generally, gene promoter hypermethylation is associated with decreased expression of the gene [67]. However, more than 90% of all genomic 5MeC are not directly related to gene function as they lie on CpG dinucleotides located in transposable repetitive elements [68]. Then, global hypomethylation, as well as hypomethylation of transposable repetitive elements, have been associated with reduced chromosomal stability and altered genome function [69], [70].

DNA methylation is a covalent modification by which methyl groups are to the DNA molecules. Methylation can change the activity of a DNA segment without changing the sequence and it is heritable by somatic cells after cell division. Perturbations in methylation during fetal growth and early lifetime can lead to irreversible changes in structure and function.

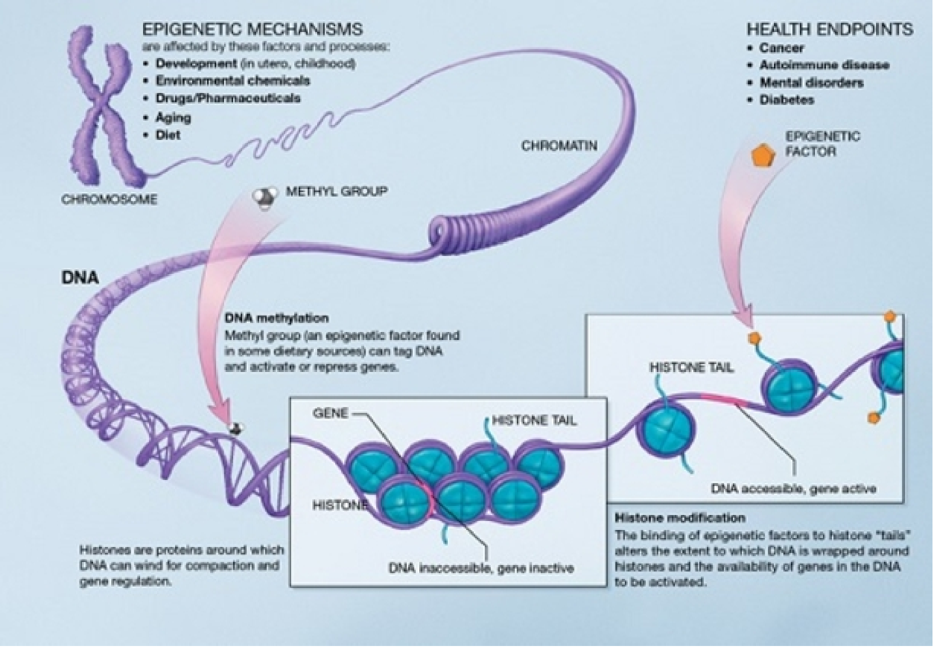

EPIGENETIC MECHANISMS
are affected by these factors and processes:
• Development (in utero, childhood)
• Environmental chemicals
• Drugs/Pharmaceuticals
• Aging
• Diet
HEALTH ENDPOINTS
• Cancer
• Autoimmune disease
• Mental disorders
• Diabetes
CHROMOSOME
CHROMATIN
METHYL GROUP
EPIGENETIC FACTOR
DNA
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.
GENE
HISTONE TAIL
HISTONE
HISTONE TAIL
DNA inaccessible, gene inactive
DNA accessible, gene active
Histones are proteins around which DNA can wind for compaction and gene regulation.
Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

Figure 1.11: Epigenetic mechanisms, by National Institutes of Health.

DNA methylation is the main epigenetic biomarker investigated in molecular epidemiology studies in relation to environmental exposures. A common finding in this studies is the small epigenetic effect that is associated with exposures. A reasonable answer for this situation is proposed by Breton et. al. when referring to the consequences of these small effects. Small effects result to be magnified over time, raising the risk for developing diseases, so we do not find larger effects just because they are incompatible with continued development [71].

microRNAs (miRNA) are single-stranded RNAs of 21-23 nucleotides in length that are transcribed from DNA but not translated into proteins (non-coding RNAs). Mature miRNAs are partially complementary to one or more messenger RNA (mRNA) molecules. miRNA main function is to down-regulate gene expression by interfering with mRNA functions [72].

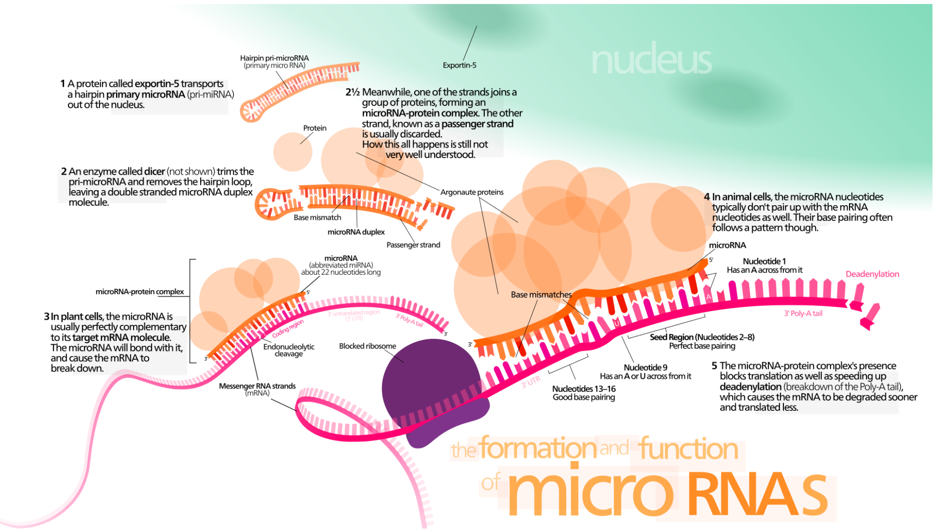

nucleus
Exportin-5
Hairpin pri-microRNA (primary micro RNA)
1 A protein called exportin-5 transports a hairpin primary microRNA (pri-miRNA) out of the nucleus.
2½ Meanwhile, one of the strands joins a group of proteins, forming an microRNA-protein complex. The other strand, known as a passenger strand is usually discarded. How this all happens is still not very well understood.
Protein
2 An enzyme called dicer (not shown) trims the pri-microRNA and removes the hairpin loop, leaving a double stranded microRNA duplex molecule.
Argonaute proteins
Base mismatch
microRNA duplex
Passenger strand
microRNA (abbreviated miRNA) about 22 nucleotides long
microRNA-protein complex
5'
3'
3 In plant cells, the microRNA is usually perfectly complementary to its target mRNA molecule. The microRNA will bond with it, and cause the mRNA to break down.
Coding region
3' untranslated region (3' UTR)
3' Poly-A tail
Endonucleolytic cleavage
Blocked ribosome
Messenger RNA strands (mRNA)
4 In animal cells, the microRNA nucleotides typically don't pair up with the mRNA nucleotides as well. Their base pairing often follows a pattern though.
microRNA
Nucleotide 1
Has an A across from it
Deadenylation
Base mismatches
A
U
Seed Region (Nucleotides 2–8)
Perfect base pairing
Nucleotide 9
Has an A or U across from it
Nucleotides 13–16
Good base pairing
3' UTR
3' Poly-A tail
5 The microRNA-protein complex's presence blocks translation as well as speeding up deadenylation (breakdown of the Poly-A tail), which causes the mRNA to be degraded sooner and translated less.
the formation and function of micro RNAs

Figure 1.12: Formation and function of micro-RNA, by Wikipedia.

Research provided pieces of evidence to support two distinct modes of miRNA-mediated translational repression, one acting at the initiation phase of protein synthesis [73] and the other at a stage post-initiation [74]. While the majority of miRNAs are located within the cell, some miRNAs, commonly known as circulating miRNAs or extracellular miRNAs, have also been found in the extracellular environment, including various biological fluids and cell culture media [75]. miRNAs derive from regions of RNA transcripts that fold back on themselves to form short hairpins [76].

miRNAs function via base-pairing with complementary sequences within mRNA molecules. As a result, these mRNA molecules are silenced, by one or more of the following processes [77], [78]:

- Cleavage of the mRNA strand into two pieces.
- Destabilization of the mRNA through shortening of its poly(A) tail.
- Less efficient translation of the mRNA into proteins by ribosomes.

As already seen, the most studied environmental factor in relation to methylation is smoking, also in miRNA [79]. Vrijens et. al. conducted a systematic review looking for miRNA as potential signatures of environmental exposure. Table 1.2, adapted from the systematic review, shows the miRNA related to smoking from "in vivo" studies.

Table 1.2: In vitro studies on the effects of smoking on differential miRNA expression, from Vrijens et. al. [80].

miRNA	**miRNA function**	**Regulation**	**Tissue/cell type**
miR-15a	Tumor suppressor	↓	Primary bronchial epithelial cells
miR-125b	Targets *p53*, stress response		
miR-199b	Oncogene activation		
miR-218	Tumor suppressor		
miR-31	Apoptosis, tumor suppressor	↑	Normal and cancer lung cells
miR-21	Fatty acid synthesis, apoptosis	↑	Human squamous carcinoma cells
miR-452	Targets *CDK1*	↓	Human alveolar macrophages

1.4.2 Transcriptomics and Beyond

Transcriptome analysis in molecular epidemiology studies has become a promising tool in order to evaluate the impact of environmental exposures. These analyses aim to help in establishing the *exposome* both by identifying the chemical nature of the exposures and the induced molecular responses. Transcriptomic signatures can be regarded as a biomarker of exposure as well as markers of effect which reflect the interaction between individual genetic background and exposure levels [81], [82].

Recent research has shown the usefulness of measuring transcriptomics responses induced by environmental factors in order to:

- Define new biomarkers of exposure and effect at gene expression level.
- Identify relevant gene–environment interactions.
- Establish mechanistic pathways involved in both initiation and prevention of disease.

An already published study about the effects of the blueberry-apple juice in human transcriptome revealed that most of the gene expression changes

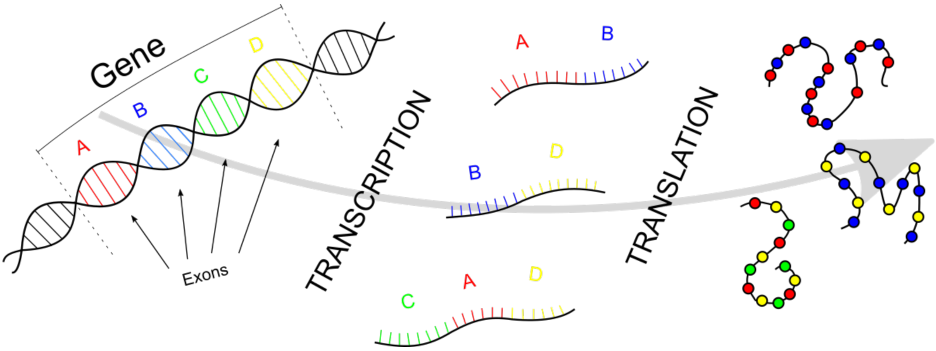


Figure 1.13: Gene-protein codification process.

were produced in biological pathways involved in the immune system. Cell adhesion and lipid metabolism pathways were also perturbed.

Transcriptome was also a target for the study of smoking effects on human health. Example of this is the study presented by Paul and Amundson where they targeted 300 genes with significantly different expression, of which 170 genes were up-regulated and 130 genes were down-regulated in smokers [83].

The proteome is the set of proteins expressed in a given type of cell, at a given time, under defined conditions. As seen in Figure 1.13 the proteins results from gene codification and more than one protein can be produced from one gene due to alternative splicing events. Then, exposure to environmental exposures often elicits a change in cellular signaling which is carried out in part by changes in the post-translational state of proteins, for instance changing their abundance levels.

Alterations in the proteome, as seen in the transcriptome, have an impact on human health. Following the smoking topic, research has seen that many of the substances included in tobacco smoke readily pass through the placental barrier [84]. So on, maternal smoking significantly affected 72 protein out of 392 protein analyzed by Huuskonen et. al.: 27 protein levels were increased and 45 protein decreased their volumes. The protein affected included constructs of hemoglobin subunits, protective proteases, and proteins directly involved in cellular structure or carbon dioxide metabolism [85].

Other studies linked aberrations in proteome with the *exposome* and human diseases. For instance, the immune response is trigged intermediate when air pollutants enter the body. Such a response is observable by assessing levels of small proteins such interleukin-1β (IL-1β), interleukin-6 (IL-6), among others. These cytokines are prominent inflammatory signaling mediators that contribute to widespread neuroinflammation in the children's brain [86].

Metabolic profiling (metabolomics) is now used routinely as a tool to provide information-rich data-sets for biomarker discovery, promoting and augmenting detailed mechanistic studies. Hence it can be used to explore the integrated response of an organism to environmental changes. Numerous metabolic phenotyping studies have investigated the impact of anthropometric factors such as age, sex, and obesity in an attempt to understand the human metabolome [87], [88].

Gu et. al. identified 25 metabolites associated with smoking status. In their findings, they identified associations with metabolites involved in the benzoate, caffeine, vitamin, steroid, amino acid and carbohydrate pathways, which potentially underscore smoking associated aetiological mechanisms [89]. These findings may have implications regarding the etiology of smoking-related diseases. In the same direction, Rolle-Kampczyk et. al. performed a metabolic profiling of serum from cord blood discovering that the effects of environmental tobacco smoke on the fetal metabolome are affected in a different way than the maternal metabolome [90].

Chapter 2

Objective

The aim of this Ph.D. thesis is to study the role of environmental exposures on human molecular signatures. The thesis begins by studying the role of a single exposure, while the second part aims to provide tools for extending such analyses to the exposome paradigm. Both scopes are related to the analysis of the impact of the exposome on human molecular signatures. The specific objectives are:

1. To study the effect of ultraviolet radiation (UVR) on the blood and skin transcriptome, including both mRNA and miRNA in an experimental design.
   (a) Scientific Article I: Analysis of the effect of solar fluorescent simulated radiation on human blood transcriptome.
   (b) Scientific Article II: Analysis of the effect of UVB on human skin transcriptome.
2. To develop bioinformatics tools under R programming language for the analysis of the exposome and multiple omic data-sets.
   (a) Scientific Article III: Development of a coordinated data organization system for multiple omic data-sets.
   (b) Scientific Article VI: Development of a framework to perform:

- exposome data characterization; including exposures standardization, transformation and description.
- univariate association analysis between exposome and diseasome.
- univariate and multivariate association analysis between exposome and multiple omic data-sets.

(c) Scientific Article V: Development of a tool to perform queries to *Comparative Toxicogenomics Database* (CTD$^{\text{TM}}$) for exposome enrichment analysis.

Chapter 3

Effect of Ultraviolet Radiation in Human Transcriptome

3.1 Rationale

Exposure to ultraviolet radiation (UVR) occurs from both natural and artificial sources. UVR can be classified into three regions according to its wavelength (see Figure 3.1): UVA (315-400 nm), UVB (315-295 nm) and UVC (295-200 nm). The main natural source is the Sun. However, radiation under 295 nm in wavelength does not reach Earth's surface due to the absorption by the stratospheric ozone layer. As a result, UVR from Sun typically consist in 95% of UVA and 5% of UVB. On the other hand, artificial UVR sources are widely used in industry and health care organizations for their germicidal properties.

The health effects of UVR on humans can be beneficial or detrimental, depending on the amount and form of UVR, as well as on the skin type of the individual exposed. Detrimental and acute effects of UVR include erythema, pigment darkening, delayed tanning and thickening of the epidermis. Erythema, redness of the skin that occurs with sunburn, is a

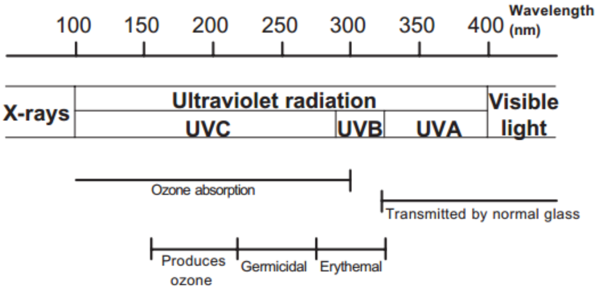


Figure 3.1: UVR spectrum and the chemical physical and biological effects, from Matsumura and Ananthaswamy [91].

cutaneous inflammatory reaction that can be accompanied by warmth and tenderness. In fair skin types, sunlight may induce a transient flush of erythema during or immediately after exposure. A delayed erythema response is common in all skin types, and peaks between 6–24h [92]. Repeated UV-induced injury to the skin, may ultimately predispose one to the chronic effects photoaging (the development of deep wrinkles, leathery skin, dilatation of blood vessels, and multiple dark spots on the Sun exposed skin), immunosuppression, and photocarcinogenesis [93], [94].

The main established beneficial effect of UVR is the cutaneous synthesis of vitamin D_3. Vitamin D is necessary to maintain physiologic calcium and phosphorous for normal bone mineralization and to prevent rickets, osteomalacia and osteoporosis [95].

Vitamin D_3 is synthesized endogenously in human skin following exposure to UVB radiation in sunlight, which spontaneously photoisomerizes 7-dehydrocholesterol to pre-vitamin D_3 [96], [97]. Pre-vitamin D_3 is subsequently converted to vitamin D_3 by thermal isomerisation, which then enters the circulation and is hydroxylated in the liver to long-lived 25-hydroxyvitamin D (25OHD3) [98]. This process is seen in Figure 3.2.

Since foods are naturally low in vitamin D main source for most people is by solar exposure. The solar zenith angle, which varies by latitude, season and time of day, determines the amount of absorption and scattering of solar UVB radiation and thus the intensity of sunlight at ground-level [100]. The association between solar UVB and vitamin D is not straight-

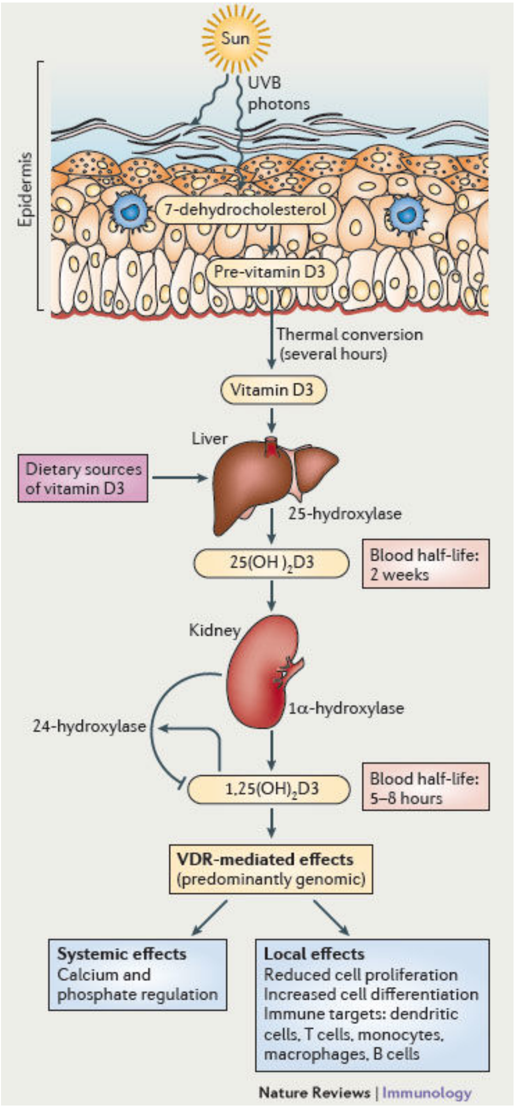

Sun
UVB photons
Epidermis
7-dehydrocholesterol
Pre-vitamin D3
Thermal conversion (several hours)
Vitamin D3
Liver
Dietary sources of vitamin D3
25-hydroxylase
25(OH)$_2$D3
Blood half-life: 2 weeks
Kidney
1α-hydroxylase
24-hydroxylase
1,25(OH)$_2$D3
Blood half-life: 5–8 hours
VDR-mediated effects (predominantly genomic)
Systemic effects
Calcium and phosphate regulation
Local effects
Reduced cell proliferation
Increased cell differentiation
Immune targets: dendritic cells, T cells, monocytes, macrophages, B cells
Nature Reviews | Immunology

Figure 3.2: From UVB to Vitamin D$_3$ synthesis, from Hart el. al. [99].

forward, since living in a sunny climate does not ensure sufficient vitamin D status [101], [102]. Vitamin D availability also depends on personal and lifestyle factors including skin pigmentation (increased melanin in darker skin naturally blocks cutaneous synthesis of vitamin D3) [103], age (the amount of 7-dehydrocholesterol in the skin decreases with age) [104], dietary and supplemental intake [105] and sunlight exposure (when and how long unprotected skin is exposed) [106].

Besides vitamin D production, UVR has also been related to other beneficial effects. Ecological and epidemiological studies have suggested that UVB exposure and vitamin D protects against several cancers [107], [108].

For all this, we planned to study the effects of UVR on transcription in humans to disentangle the molecular mechanisms underlying the last effects of UVR on heath. In particular, we investigated the acute effects of UVR exposure on both blood and skin human transcriptome, at level of gene expression and micro-RNA expression.

3.2 Methods

3.2.1 Blood Analysis Design

Nine healthy males from UK and with similar anthropometric and sun-sensitive skin type II were selected for the study. The whole body of the participants was exposed to 3 standard erythemal dose (SED) of fluorescent solar simulated radiation (FSSR). FSSR imitates real solar exposures and compresses UVA and UVB at similar proportions.

Five participants were exposed in spring (March-April), and 4 in summer (July-September). Blood samples for the nine volunteers were obtained pre-exposure and 6h, 24h and 48h post-exposure. RNA was extracted and quantified with a Nanodrop spectrophotometer (Thermo Fisher Scientific), all RNA samples had a RNA Integrity Number (RIN) > 7. Plasma vitamin D_2 and vitamin D_3 levels were measured in duplicate with liquid ultra high pressure chromatography tandem mass spectrometry (Waters, Milford, Massachusetts, USA) before detection by a TQD Mass Spectrometer

(Waters, Milford, Massachusetts, USA) with electrospray ionization using multiple reaction monitoring.

The description of the study can be seen in Figure 3.3.

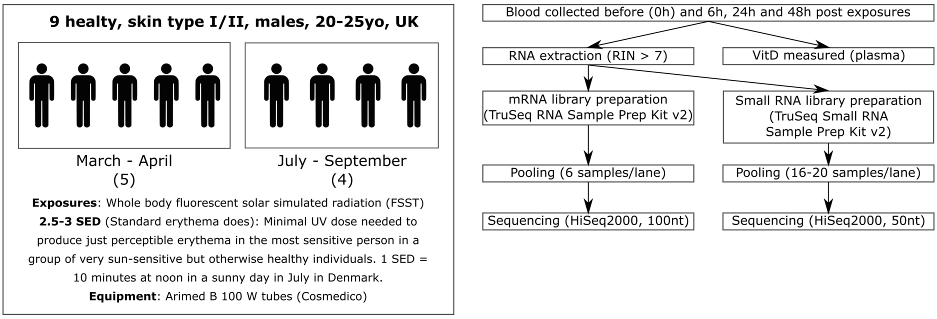

9 healty, skin type I/II, males, 20-25yo, UK
March - April
(5)
July - September
(4)
Exposures: Whole body fluorescent solar simulated radiation (FSST)
2.5-3 SED (Standard erythema does): Minimal UV dose needed to produce just perceptible erythema in the most sensitive person in a group of very sun-sensitive but otherwise healthy individuals. 1 SED = 10 minutes at noon in a sunny day in July in Denmark.
Equipment: Arimed B 100 W tubes (Cosmedico)
Blood collected before (0h) and 6h, 24h and 48h post exposures
RNA extraction (RIN > 7)
VitD measured (plasma)
mRNA library preparation (TruSeq RNA Sample Prep Kit v2)
Small RNA library preparation (TruSeq Small RNA Sample Prep Kit v2)
Pooling (6 samples/lane)
Pooling (16-20 samples/lane)
Sequencing (HiSeq2000, 100nt)
Sequencing (HiSeq2000, 50nt)

Figure 3.3: Experimental design for the study of the UVR in human blood transcriptome.

3.2.2 Skin Analysis Design

Seven of the nine volunteers had a region of the gluteus exposed to 3 SED, a region exposed to 6 SED and a region kept covered as control. Each exposed region was exposed to fluorescent solar simulated radiation (FSSR). FSSR imitates real solar exposures and compresses UVA and UVB at similar proportions. For these seven volunteers biopsies of the control region, 3 SED exposed region and 6 SED exposed region were obtained at 6h and at 24h post-exposures. The description of the study can be seen in Figure 3.4.

3.2.3 Bioinformatic Pipeline

The total number of samples was 84, 40 blood samples and 44 skin samples. Seven technical replicates were included (different library preparation).

mRNA initial quality control was done by visual inspect of the reports generated by Lappalainen et. al. `FastQC` [109]. Alignment (with a maximum of 5 mishmashes) and mapping (to NNCBI hg19 - EntrezId - allowing

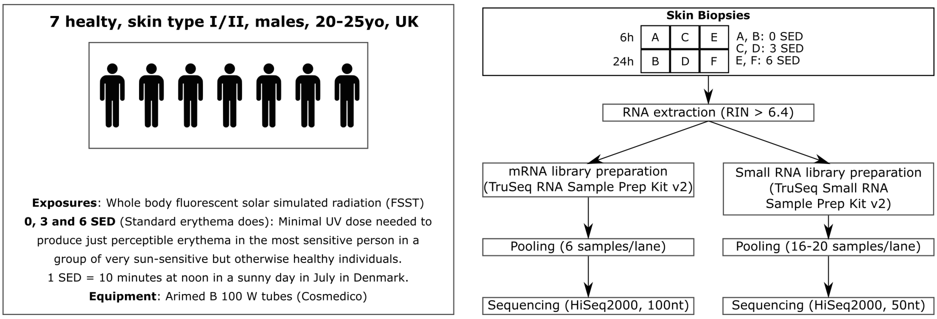

7 healty, skin type I/II, males, 20-25yo, UK
Exposures: Whole body fluorescent solar simulated radiation (FSST)
0, 3 and 6 SED (Standard erythema does): Minimal UV dose needed to produce just perceptible erythema in the most sensitive person in a group of very sun-sensitive but otherwise healthy individuals.
1 SED = 10 minutes at noon in a sunny day in July in Denmark.
Equipment: Arimed B 100 W tubes (Cosmedico)
Skin Biopsies
6h
24h
A
C
E
B
D
F
A, B: 0 SED
C, D: 3 SED
E, F: 6 SED
RNA extraction (RIN > 6.4)
mRNA library preparation (TruSeq RNA Sample Prep Kit v2)
Small RNA library preparation (TruSeq Small RNA Sample Prep Kit v2)
Pooling (6 samples/lane)
Pooling (16-20 samples/lane)
Sequencing (HiSeq2000, 100nt)
Sequencing (HiSeq2000, 50nt)

Figure 3.4: Experimental design for the study of the UVR in human skin transcriptome.

multiple overlap and discarded if matched in more than one locus) was done using `Rsubread` R package [110]. Three samples were filtered (due to low number of reads) after checking GC content per sample, number of read per gene and number of read in *globin* and *collagen* genes. Same samples were marked to be discarded after 5'-3' degradation inspection. Minimum correlation between technical duplicates was of 0.983 and maximum of 1. Visual inspection of PCAs by tissue, exposure, technical and anthropometric variables were done and no sample was discarded.

micro-RNA sequencing data were analysed as previously described [111]. At least $\sim$ 4M reads mapped to micro-RNAs in each sample. Despite this, we detected a highly abundant micro-RNA, *hsa-miR-486-5p* that represented between 78% and 90% of the reads. Principal component analysis did not show any technical bias. One sample was removed due to a completely different behaviour, possibly due to contamination. The correlation between replicates was > 0.98 and only one of them was kept in the analysis.

Normalization of the counts and differential expression analysis were done using `DESeq2` R package [112], including *library batch*, *flowcell* and *id* as covariates and taking as reference the unexposed or first time. Pathway analysis was performed on the results using DAVID (6.7 Jan 2010) [113].

27 blood genes and 55 skin genes were validates though qPCR (quantitative polymerase chain reaction) performed using the TaqMan Real-Time

PCR system (Thermo Fisher Scientific). Four housekeeping genes were selected from the mRNA data experiment [114]. Samples were run in triplicate and their correlations were > 0.7 for blood genes and > 0.6 for skin genes, except for three and two genes. Outlier replicates, defined by a standard deviation > 0.25 and a qPCR threshold cycle > 33 were excluded.

For three participants a second unexposed blood sample, collected between 1 and 22 days before UVR exposure, was also analysed and correlations between the two unexposed samples were > 0.97.

3.3 Results

Two manuscripts corresponds to the results of this project, one with the results obtained from the transcriptom analysis in blood and a second one with the results obtained from skin analysis.

3.3.1 Effect of FSSR on Human Blood Transcriptome

Bustamante M, Hernandez-Ferrer C, Sarria Y, Harrison GI, Nonell L, Kang W, et al. The acute effects of ultraviolet radiation on the blood transcriptome are independent of plasma 25OHD 3. Environ Res. 2017 Nov;159:239–48. DOI: 10.1016/j.envres.2017.07.045

3.3.2 Effect of FSSR on Human Skin Transcriptome

Journal	Environmental Research
Title	Ultraviolet radiation induced changes on skin transcriptome in humans: effect of dose and time
Authors	Mariona Bustamante*, Carles Hernandez-Ferrer*, Yaris Sarria, Graham I Harrison, Lara Nonell, Wenjing Kang, Marc R Friedländer, Xavier Estivill, Juan R González, Mark Nieuwenhuijsen, Antony Young
Submitted	Draft

* Both authors contributed equally to this work.

Ultraviolet radiation induced changes on skin transcriptome in humans: effect of dose and time

Mariona Bustamante[a,b,c,d*], Carles Hernandez-Ferrer[a,c,d], Angela Tewari[e], Yaris Sarria[a,c,d], Graham I. Harrison[e], Eulalia Puigdecanet[f], Wenjing Kang[g], Marc R. Friedländer[g], Xavier Estivill[b,c,d], Juan R. González[a,c,d], Mark Nieuwenhuijsen[a,c,d], Antony R. Young[e*]

[a]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
[b]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
[c]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[d]CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain
[e]King's College London (KCL), St John's Institute of Dermatology, London, UK
[f]Servei d'Anàlisi de Microarrays, IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain
[g]Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden

The authors wish it to be known that, in their opinion, the two first and last two authors should be regarded as joint First and joint Last Authors.

*Corresponding authors:

Mariona Bustamante, Tel +34 93 214 73 00, Fax +34 93 214 73 02, mariona.bustamante@isglobal.org, Av. Dr Aiguader 88, 08003, Barcelona

Antony Young, Tel +44 (0) 20 7188 6412, Fax +44 (0) 20 7188 8050, antony.young@kcl.ac.uk, St John's Institute of Dermatology, 9th Floor, Tower Wing, Guy's Hospital, Great Maze Pond, SE1 9RT, London, UK

Abstract

Ultraviolet radiation (UVR) produces erythema, oxidative stress, DNA mutations, and finally skin cancer. Skin can adapt to UVR damaging effects by increasing keratinisation, tanning and apoptosis. In this study we investigated the skin transcriptional response to different doses of fluorescent solar simulated radiation (FSSR) in humans.

Seven healthy male volunteers were exposed to 3 different doses of FSSR [0 (unexposed), 3 and 6 standard erythemal doses (SED)]. Skin biopsies were obtained at 6h and 24h post-exposure for each SED. Gene and miRNA expression was assessed with next generation sequencing, and a set of differently expressed genes were validated by quantitative PCR (qPCR).

The number of differentially expressed genes increased with higher FSSR dose and shorter time post-exposure. At 6h and at 6 SED, 4,071 genes were differently expressed with an average absolute fold change of 1.5 (range: 1.2 - 2.5). A slight increase of the number of upregulated vs. downregulated genes was observed. At each time points, differently expressed genes were involved in main cellular functions such as transcription and translation. Moreover, keratinization and apoptosis were found at 6h, while inflammation and immune response was more prominent at 24h. Only four miRNAs (*hsa-miR-146b-5p*, *hsa-miR-223-3p*, *hsa-miR-204-5p* and *hsa-miR-142-5p*) were differently expressed, suggesting less strong effects or an earlier response. None time*dose interaction was validated.

The high FSSR dose used in this study is similar to the daily UVR dose experimented by holiday makers, suggesting that their skin transcriptional profile is markedly altered.

Keywords: ultraviolet radiation (UVR), ultraviolet radiation type B (UVB), fluorescent solar simulated radiation (FSSR), erythemal dose (SED), skin, dermis, epidermis, biopsy, transcription, gene expression, miRNA expression

Abbreviations

UVR: ultraviolet radiation

UVA: ultraviolet radiation type A

UVB: ultraviolet radiation type B

FSSR: fluorescent solar simulated radiation

SED: standard erythemal dose

MED: minimal erythemal dose

ROS: reactive oxygen species

miRNA: micro RNA

RIN: RNA Integrity Number

qPCR: quantitative polymerase chain reaction

FC: fold change

Log2FC: Log_2 fold change

FDR: False Discovery Rate

Introduction

Skin is the largest and most external organ in the body and forms a physical barrier to environment. It is organized in two layers: epidermis and dermis made of epithelial, mesenchymal, glandular and neurovascular components. Exposure to solar ultraviolet radiation (UVR) is one of the most important environmental factors affecting skin physiology. It is the main cause of the three most common types of skin cancer: basal cell carcinoma, squamous cell carcinoma and malignant melanoma (Greinert et al., 2015).

Terrestrial solar ultraviolet radiation typically comprises ≤5% of UVB (~295-315 nm) and ≥95% of UVA (315-400 nm). The impact of UVA and UVB radiation on the skin is dependent on their energy which is absorbed by cellular chromophores such as melanin, DNA, aromatic amino acids (ie. tyrosine and tryptophan), and urocanic acid, among others (Damiani and Ullrich, 2016)(Ramasamy et al., 2017)(Schuch et al., 2017).

UVB radiation is more energetic than UVA. It can directly damage the epidermal cells producing erythema, associated with an inflammatory response (D’Orazio et al., 2013)(Schuch et al., 2017). UVB also produces DNA mutations, which, if not eliminated via apoptosis or DNA repair mechanisms, can lead to photo-carcinogenesis (D’Orazio et al., 2013)(Ramasamy et al., 2017)(Schuch et al., 2017). On the other hand, UVB participates in the synthesis of vitamin D in skin, which has many beneficial effects on health.

UVA rays penetrate deeper within the skin and are mostly responsible for the generation of reactive oxygen species (ROS), and therefore they can also generate DNA damage, but at lesser extent than UVB (Tewari et al., 2012)(Ramasamy et al., 2017)(Schuch et al., 2017). UVA can reach the deep dermis producing skin photo-aging (Krutmann, 2000). Both UVR types are also involved in skin immunosupression (Hart et al., 2011).

Skin can adapt to UVR exposure by increasing keratinocyte cell division (epidermal hyperkeratosis) and by increasing tanning, melanization of the skin (D'Orazio et al., 2013). The first step of melanization consists in a redistribution of existing epidermal melanin pigments. This is mainly lead by UVA. UVB is responsible of the de novo melanin synthesis. Its transfer to keratinocyes starts after several hours or days of UVR exposure. However, tanning response is generally insufficient to prevent UVR mutagenic effects (Greinert et al., 2000)(Ridley et al., 2009) and erythema in lighter skin types (Sheehan et al., 1998).

The molecular consequences of exposure to UVR, both the adverse effects as well as the adaptive responses, can be reflected in the skin transcriptional profile (D'Orazio et al., 2013)(Ramasamy et al., 2017)(Schuch et al., 2017). UVR-induced gene expression changes in skin have extensively been investigated, but the response of miRNAs, small non-coding RNAs that regulate gene expression, is less well characterized (Syed et al., 2013). Moreover, not many of these studies performed *in vivo* and involving skin biopsies from exposed human volunteers (Enk et al., 2006)(Ramasamy et al., 2017). Usually they only evaluate one UVR dose and one time point post-exposure (Daniell, 2012)(Dawes et al., 2014). The study of skin transcriptional response to the whole terrestrial UVR spectrum (UVA + UVB), as in real life settings, may provide relevant information for public health.

In the present study, we investigated the effects of different doses of fluorescent solar simulated radiation (FSSR) over time on the skin transcriptional profile in humans. In particular, seven healthy male volunteers were exposed to two different doses of FSSR (3 SED and 6 SED). Unexposed and exposed skin biopsies from these volunteers were collected at 6h and 24h post-exposure, and gene and miRNA expression was assessed using next generation sequencing.

Material and methods

Participants, biological samples and exposure to fluorescent solar simulated radiation

The study was conducted according to the Declaration of Helsinki after approval was obtained from the Ethics Committee of St Thomas's Hospital, London, UK. All participants gave written informed consent. The study design, library preparation and bioinformatic analysis can be found in Supplementary Figure 1.

Seven healthy males from UK and with similar anthropometric and sun-sensitive skin type II were selected for the study (Table 1). Previously unexposed buttock skin was exposed to FSSR using Arimed B in a two Waldmann UV 100L W tubes (Waldmann GmbH & Co, Villingen-Schwenningen, Germany). The emission spectrum comprises 5.3% UVB (280-315 nm) which accounts for 79.6% of the erythemally effective energy. Two different regions were irradiated to 3 and 6 standard erythemal doses (SED). An unexposed region was taken as control. At 6h and 24h the following punch biopsies (4mm including epidermis and dermis) were taken under local anaesthesia by a dermatologist: 6h – 0 SED (A), 6h – 3 SED (C); and 6h – 6 SED (E); and at 24h: 24h – 0 SED (B), 24h – 3 SED (D), and 24h – 6 SED (F). All samples were immediately frozen. Total RNA from skin biopsies was extracted using the RNeasy Mini Kit (Qiagen, Hilden, Germany). RNA was quantified with a Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA), and quality evaluated with a RNA 6000 Nano Kit in a Bioanalyzer equipment (Agilent, Santa Clara, California, USA). All RNA samples had a RNA Integrity Number (RIN) >6.4, expect ICE_004_E (6h – 6 SED), which was eliminated from the study. Additional details on the samples can be found in Supplementary Table 1.

Gene and miRNA expression (next generation sequencing)

RNA libraries were prepared with the TruSeq RNA Sample Prep Kit v2 (mRNA) and the TruSeq Small RNA Sample Prep Kit (small RNA) (Illumina, San Diego, California, USA). Small RNA library size selection was done with acrylamide gels and quality was assessed with the DNA 1000 or High Sensitivity Kit in a Bioanalyzer (Agilent, Santa Clara, California, USA). Pools of 6 (mRNA) or 16-20 (small RNA) cDNA libraries were prepared at 30 nM after quantification with KAPA SYBR® FAST qPCR (Kapa Biosystems, Hoffmann-La Roche, Basel, Switzerland). Libraries were single-end sequenced (100 nt and 50 nt for mRNA and small RNA, respectively) on a HiSeq2000 platform (Illumina, San Diego, California, USA). Samples were randomized during library preparation and during the sequencing. Three mRNA and one miRNA samples were analyzed in duplicate. Correlation between replicates was >0.98 and only one of them was kept in the analysis.

For mRNA, reads were mapped against the genome using the R package Rsubread (Liao et al., 2013), allowing a maximum of 5 mismatches and using the hs37d5 as reference. Gene annotation was performed with NCBI hg19 (Entrez Gene) database. Three samples had <10M reads and were excluded from the analysis (ICE_003_A, ICE_004_A, and ICE_004_F) (Supplementary Table 2). None of the samples showed degradation (Wang et al., 2012). 19,877 genes were detected in >80% of the samples, 17,612 with an average of >19 reads.

The small RNA sequencing data were analyzed as previously described (Lappalainen et al., 2013). At least 1.79 M reads mapped to miRNAs in each sample (Supplementary Table 3) and 1,427 miRNAs were annotated.

Validation of the expression levels of top genes (qPCR)

Validation of 55 genes was performed using the TaqMan Real-Time PCR system (Thermo Fisher Scientific, Waltham, Massachusetts, USA). The selection criteria is specified in Supplementary Table 4. Four housekeeping genes were selected from the mRNAseq experiment using the RefFinder web-based tool (Xie et al., 2012). Unexposed and paired samples collected at 6h (A) and at 24h (B) showed a correlation > 0.97. Samples were run in triplicate and their correlations were >0.6, except for two genes (*POMC*, and *CYP27B1*). According to the NormFinder method (Andersen et al., 2004), *UBE2D2* and *TBP* were found to be the most stable housekeeping genes. ΔCt was calculated subtracting mean (*UBE2D2* + *TBP*) Ct to the candidate gene Ct. Sample size in the mRNAseq and qPCR experiments was slightly different (Supplementary Table 1).

Statistical analysis

All analyses were done in R3.1.0 and R3.2.3 environment ('R: A language and environment for statistical computing.', 2017, http://www.R-project.org).

Differential expression: next generation sequencing

mRNA and miRNA differential expression was analyzed using the R package DESeq2 v.1.14.1 (Love et al., 2014).

All samples were normalized together using the scaling factor method implemented in DESeq2. Then, a negative binomial generalized model to test the association between gene expression and a dummy variable that combined time and dose, was fitted. The model was adjusted for participant ID, library preparation batch, and flowcell (mRNA). Log2 Fold Changes (Log2FC) were estimated with an automatic shrinkage function that uses empirical Bayes priors. The following contrasts were tested: 6h - 0 SED vs. 6h – 3 SED; 6h - 0 SED vs. 6h – 6 SED; 24h - 0 SED vs. 24h – 3 SED; 24h - 0 SED vs. 24h – 6 SED, and 6h - 0 SED vs. 24h – 0 SED. Linear FSSR dose effects were tested at 6h and 24h, separately. Seventy-five % and 90% of the genes detected at dose 6 SED in the main analysis were also significant in the linear dose response analysis at 6h and 24h, respectively (data not shown). Note that the differences between both analyses are the assumption of linear effects and the normalization process. In the linear dose response analysis, samples collected at 6h were normalized together and separately from samples collected at 24h, and vice versa.

To search for genes whose expression level deviated from the additive effects of dose and time, a negative binomial generalized model with a time*dose interaction term was fitted. The following time*dose interaction parameters were retrieved: 6h – 0 SED vs. 24h – 3 SED; and 6h - 0 SED vs. 24h – 6 SED.

Multipletesting was addressed with DESeq2 by filtering genes that had little chance of showing significant evidence and by calculating adjusted p values with the False Discovery Rate (FDR) method within each comparison. Moreover, in the small RNA analysis, Bonferroni correction was applied by dividing the nominal significance (p value = 0.05) by the number of miRNAs detected with a minimum of 10 normalized counts (Bonferroni adjusted p value: 0.05 / 389 = 1.29E-04).

Differential expression: qPCR

The association between gene expression assessed by qPCR (ΔCt) and FSSR dose at different times was tested with linear mixed models adjusting for participant ID as a random effect. The interaction between time and dose was tested by introducing a time*dose interaction term in the models that contained all samples together.

Note that sample size in the mRNAseq, miRNAseq and qPCR validation studies was slightly different.

Functional enrichment analysis

Genes with an arbitrary significance of p value <1E-03 were selected to perform functional enrichment analysis. No filtering of the genes based on Log2FC was done.

Gene-set enrichment analysis

Gene-set enrichment analysis was performed with the Database for Annotation, Visualization and Integrated Discovery (DAVID) v.6.7 (Huang da et al., 2009a)(Huang da et al., 2009b), using GO-BP, KEGG, Biocarta, and Reactome databases. To reduce redundancy, we used the Functional Annotation Clustering option that displays similar annotations together. Results were filtered for enrichment scores ≥1.3 (equivalent to 0.05 in the non-log scale).

miRNA – gene regulatory networks

The regulatory networks of miRNAs and genes, including transcription factors, were analyzed using MAGIA2 (Bisognin et al., 2012). This tool combines expression profiles analysis with in silico regulatory interaction predictions. To create the miRNA-gene networks, all samples with available gene and miRNA expression data were analyzed together, and not stratified by condition. Briefly, miRNA - gene (including transcription factors) interactions were predicted using DIANA-microT (Maragkakis et al., 2009) with mean stringency. Transcription factor - miRNA interactions are retrieved from mirGen2.0 (Friard et al., 2010) and TransmiR (Wang et al., 2010), whereas transcription factors - gene interactions were obtained from 'TFBS conserved' track of the UCSC genome annotation for human (version hg19). The correlation of the expression profiles was calculated with the Spearman test.

Validated miRNA-targeted gene pairs were retrieved from 'miRWalk 2.0: a comprehensive atlas of predicted and validated miRNA-target interactions' (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/).

Results

Study design

Seven healthy males with similar anthropometric and dermatological parameters were enrolled in the study (Table 1). Three skin biopsies were collected at 6h and three at 24h after exposure to different FSSR doses: unexposed (0 SED), exposed to 3 SED and exposed to 6 SED.

Gene expression after exposure to FSSR

Discovery phase: mRNAseq

The number of differentially expressed genes at 5% FDR increased with higher FSSR dose (6 SED) and shorter time post-exposure (6h), suggesting a dose dependent short term effect on the majority of genes (Table 2). Therefore, the highest number of differently expressed genes was found at 6h and at 6 SED. Absolute mean fold change (FC) at 6h and 6 SED was 1.5 (ranging from 1.2 to 2.5). In general, a faint increase of the number of upregulated vs. downregulated genes was observed. At 6 SED, the effect size of upregulated genes was slightly more pronounced than the effect of downregulated, as shown in the Volcano plots (Figure 1). After multiple-testing correction, no statistically significant differences were observed among expression levels in unexposed (0 SED) skin biopsies collected at 6h and at 24h (data not shown).

At 6h, 132 genes were differently expressed (5% FDR) after exposure to 3 SED (Supplementary Table 5). The number increased to 4,071 after exposure to 6 SED (Supplementary Table 6). Ninety-seven % of the genes detected at low dose (3 SED) were also altered at high dose (6 SED). At 24h, 16 genes were differently expressed (5% FDR) after exposure to 3 SED and 1,583 after exposure to 6 SED (Supplementary Table 7 and Supplementary Table 8). Again, 94% of the genes detected at low dose (3 SED) were also altered at high dose (6 SED).

Figure 2 summarizes the overlap of genes differently expressed at 5% FDR at different conditions. At low dose (3 SED), 7.5% of the genes with FSSR induced changes at 6h were also found at 24h; while at high dose (6 SED), this increased to 29.7%. Less than 25% of the genes were exclusively identified at 24h. Ten genes (7 downregulated and 3 upregulated) showed altered expression patterns after FSSR at all times points and doses. Their absolute FC ranged from 1.3 to 3.8 with a mean of 2.1. In general, their differential expression was more pronounced at 6 SED and 6h.

Only two genes showed an interaction between time and dose: *IGSF9B* and *TMEM127* (at 5 % FDR and with >10 averaged normalized counts). However, their baseline levels at 6h and 24h were different and the time*dose interaction was not validated by qPCR (Supplementary Figure 4 and Supplementary Table 9).

When the analysis was restricted to 5 volunteers with available samples in all conditions, gene expression patterns were similar to the ones observed in the main analysis (Supplementary Table 10 and Supplementary Figure 5).

Gene-set enrichment analysis

At 6 SED and 6h, the following biological processes were identified: keratinization, apoptosis, transcription and translation, splicing, tRNA modifications, and cytoskeleton organization (Supplementary Table 11). At 24, we detected the same pathways found at 6h except for keratinization and apoptosis, and plus inflammation, immune response (IL1, IL6, IL10, TNF, NFKB, INF gamma, TLR), and hyaluronan biosynthesis (Supplementary Table 12). At low dose, we did not detect any enrichment, except for signaling through different factors at 6h (Supplementary Tables 13).

Validation of candidate genes: qPCR

The expression of 55 genes was validated by qPCR (Supplementary Table 4 and Supplementary Table 9). These included genes which were at least nominally associated with exposure to FSSR in the discovery phase or were in pathways with genes fulfilling this criteria: 9 genes found at 5% FDR at all doses and time conditions, 2 genes that showed a potential time*dose interaction, and 34 genes involved in candidate pathways [immunity and inflammation (N=4), DNA repair (N=3), pigmentation (N=20) and vitamin D (N=7)].

All tested pigmentation genes were repressed after FSSR, except for *EDNRB*, which was upregulated at 24h (Supplementary Figure 6). Some of the pigmentation genes showed a transient repression only at 6h post-exposure (*EDN3*, *KIT*, *MITF*, *SLC24A5*, *PMEL* and *ASIP*); whereas others had a more sustained downregulation (*ATRN*, *LYST*, *TYRP1*, *OCA2*, and *FGFR2)*. Three of these genes, *FGFR2*, *LYST* and *EDNRB*, survived Bonferroni correction.

Regarding vitamin D genes, only the expression of *CYP2R1* was validated. *CYP2R1* is involved in 25-hydroxyvitamin D production from vitamin D, mainly in liver. In the skin, its expression was decreased at high doses of FSSR and particularly at 6h (Supplementary Figure 7).

In addition, 10 genes with differential expression in blood after whole body exposure to FSSR and with nominal associations in the discovery phase in skin were selected for validation (Bustamante et al 2017). Similarly, to what was observed in blood, the expression of *CD83* was increased, but only at high doses. *SCPEP1*, *FKBP5*, *FLT3* and *ITSN1* were downregulated, especially at 6 SED and 6h, in agreement with the acute pattern observed in blood. In contrast to what was observed in blood, *PLA2G7* that participates in PAF (platelet-activating factor) degradation was not downregulated in skin, while the expression of *PTGS2* (*COX2*) regulated by PAF and involved in prostaglandin PGE2 production was increased. PAF transmits the UVR immunosuppressive signal from the skin to the immune system (Damiani and Ullrich, 2016).

miRNA expression after exposure to FSSR in skin

Discovery phase: miRNAseq

We also explored the effect of FSSR exposure on skin miRNA expression over time and at different doses. Sixteen miRNAs had an abundance >1%. *Hsa-miR-10b-5p* was the most common miRNA with a mean abundance of 26.3% of the reads.

The FSSR-induced effect on the miRNA expression was of smaller magnitude compared with the gene expression analysis as shown in the Volcano plots (Figure 3). Maximum absolute FC was 2.1 (*hsa-miR-223-3p*). Only 4 miRNAs survived multipletesting correction (Figure 4, Supplementary Table 15). *Hsa-miR-146b-5p* and *hsa-miR-223-3p* were upregulated at 6h and high dose (6 SED). At 24h, the levels of *hsa-miR-223-3p* were still high, while the levels of *hsa-miR-146b-5p* were close to baseline. The expression patterns of *hsa-miR-204-5p* and *hsa-miR-142-5p* were more complex, with potential time*dose interactions. *Hsa-miR-204-5p* was upregulated at 24h and low dose. *Hsa-miR-142-5p* was upregulated at 6h and 6 SED and downregulated at 24h and 3 SED.

miRNA – gene regulatory networks

Then, we investigated the miRNA-gene regulatory networks combining information from expression profiles and by in silico predictions. All miRNAs, expect for *hsa-miR-142-5p*, showed significant correlations with some predicted target genes or transcription factors (Figure 5). The following miRNA-gene correlations were negative and significant after multipletesting (q-value <0.05): *hsa-miR-146b-5p* and *TMEM237*, *TMEM132E*, *LANCL1*, *SLC6A4*; *hsa-miR-204-5p* and *IL1B*; *hsa-miR-223-3p* and *HLF* (transcription factor).

A list of validated miRNA-gene interactions for these 4 miRNAs is shown in Supplementary Table 16. Some of them are: *IL6 (hsa-miR-223-3p)* in myeloid cells (Dorhoi et al., 2013)*, KIT (hsa-miR-146b-5p)* in papillary thyroid carcinoma (He et al., 2005) and *PRKCB (hsa-miR-142-5p)* in breast cancer (Pillai et al., 2014).

Discussion

In the present study we investigated the transcriptional profile from skin biopsies from human volunteers exposed to FSSR. We tested two FSSR doses (3 SED and 6 SED) and two post-exposure time points (6h and 24h).

FSSR dose was the main driver of gene transcriptional changes in skin. The number of differently expressed genes at 3 SED was 1-4% of the genes detected at 6 SED, and only <5% were specific to low dose. At 6h and at 6 SED, 4,071 genes were differently expressed, which represents around 20% of the human transcriptome. Volunteers were exposed to 3 and 6 SED that is about 1 or 2 minimal erythemal dose (MED) for skin type II, respectively (Harrison and Young, 2002). Our maximal exposure dose of 6 SED was similar to Danish holiday makers who received a daily average of 9.4 ± 7.0 SED during a holiday in Tenerife in March, when at least 50% body surface was exposed to greatest UVB intensity (Petersen et al., 2013).

The number of differently expressed genes was reduced over time after FSSR exposure. Although the vast majority of biological functions detected at 6h and 24h were the same, a few of them were time specific. At 6h, we identified pathways related to apoptosis and keratinization; whereas at 24h, pathways related to inflammation, immune response and hyaluronan biosynthesis. Indeed, this expression pattern over time was confirmed by qPCR for some particular genes related to DNA repair (*AEN*, *POLH*) and immunity (*IL1A*, *IL20*, *IL6* and *TNF*). Hyaluronan, one of the main extracellular matrix molecules of epidermal keratinocytes, has previously been reported to increase after low dose UVB exposure (Rauhala et al., 2013), however results are not consistent among studies.

Ten genes were differently expressed in all conditions. Nine of them were validated by qPCR. *AEN* (upregulated) encodes a nuclear exonuclease required for P53-dependent apoptosis (Kawase et al., 2008). *CDKAL1* (downregulated) is expressed in immune cells and becomes downregulated when immune cells are activated with proliferating signals. Genetic variants in *CDKAL1* have been associated with several diseases: psoriasis, Crohn's disease and type 2 diabetes (Quaranta et al., 2009). *EPHB1* (downregulated) encodes an ephrin receptor tyrosine kinase that mediates cell-cell communication by interacting with ephrin ligands residing on adjacent cell surfaces. They participate in development, maintenance, and repair processes in cutaneous biology (Surawska et al., 2004)(Lin et al., 2012). *GRIP1* (downregulated) is a scaffolding protein required for the formation and integrity of the dermo-epidermal junction (Bladt et al., 2002). *PRKCB* (*protein kinase C beta*, downregulated) activates *TYR*, the key and rate-limiting enzyme in pigmentation. Topical application of a *PRKCB* inhibitor reduces skin and hair pigmentation (Park et al., 2004). *SLC24A3* (downregulated), also known as *NCKX3*, is a Na+/Ca2+ exchanger. Another family member (*SLC24A5*) has a role in the development of pigmentation in skin and retinal epithelia (Giot et al., 2003). Moreover, we also validated by qPCR the expression of other key genes in the pigmentation processes. The expression of selected pigmentation genes was repressed at 6h, expect for *EDNRB* which showed increased levels. Consistently, *EDNRB* levels have been shown to increase after UVB exposure in cultured melanocytes. Downregulated genes promote skin pigmentation, except for *ASIP*. *ASIP* (agouti signaling protein) binds to *MC1R* and produces a switch in the melanin production, from eumelanogenesis to pheomelanin. This expression pattern suggests that in individuals of skin type II and at 24h the delayed tanning response that involves increases in the number and activity of functional melanocytes with increased activity *TYR*, has not still started (Brenner and Hearing, 2008).

In contrast to genes, the expression of miRNAs was not massively affected by FSSR in the present time and dose conditions. Whether this is because transcription of miRNA is less influenced by FSSR or because effects take place in different time point (earlier response) deserves further investigation. It also could be that the precision of the RNAseq was lower for miRNAs, difficulty the identification of differently expressed miRNAs. In any case, due to their regulatory role, subtle differences in miRNA levels might be relevant in skin biology.

Four miRNAs were differently expressed. *Hsa-miR-146b-5p* and *hsa-miR-223-3p* were upregulated after FSSR exposure. In agreement with our findings, *hsa-miR-146b-5p* and *hsa-miR-223-3p* have been found to be upregulated in cellular models irradiated with a UV lamp (254 nm) (Al-Khalaf et al., 2013) and UVB irradiated mice (Xu et al., 2012), respectively. Moreover they both seem to be upregulated in skin of psoriasis patients (Løvendorf et al., 2015)(Hermann et al., 2017). The miRNA-gene network analysis identified the following genes as potential targets for *hsa-miR-146b-5p: TMEM237, TMEM132E, LANCL1,* and *SLC6A4. SLC6A4* encodes a serotonin transporter, which terminates the action of serotonin and recycles it in a sodium-dependent manner. Activation of the serotonin pathway has been suggested to mediate UVB-induced immune suppression (Wolf et al., 2016). *hsa-miR-223-3p* expression was inversely correlated with the expression of *HLF* (*PAR BZIP transcription factor*). *Hsa-miR-204-5p* and *hsa-miR-142-5p* exhibited a more complex pattern, making interpretation more cautious. *Hsa-miR-204-5p* was upregulated at 24h and low dose. *hsa-miR-204-5p* may control the signaling towards the MAPK and STAT3 pathway in the progression of actinic keratosis to squamous cell carcinoma (Toll et al., 2016), and is involved in skin wound (Etich et al., 2016). In our data, *hsa-miR-204-5p* levels were inversely correlated with *IL1B* levels. Another study has validated the interaction between *hsa-miR-204-5p* and *IL1B* (Li et al., 2011). *Hsa-miR-142-5p* has been found to be upregulated in chronically UVR treated mice skin (Singh et al., 2016). The role of miRNAs in UVR effects in human skin has been reviewed elsewhere (Syed et al., 2013). None of the miRNAs described there was deregulated in our study.

The study has several strengths. First, the effects FSSR on transcription were investigated in biopsies obtained from volunteers locally exposed to FSSR, in contrast to the more artificial in vitro cellular models. Selected volunteers had a similar skin, type II according to the Fitzpatrick's scale. Since we did not separate dermis from epidermis, present findings reflect the transcriptional pattern of all cell types in the skin biopsy. Secondly, the transcriptional profile was investigated comprehensively, including genes and miRNAs. Moreover, the expression of some genes was validated by qPCR. Finally, our study used a broad-spectrum UVR source, which has the advantage of simulating natural UVR.

The study also has some limitations. On one hand, the sample size is still limited to detect small effect sizes, which is of special importance for miRNAs, whose FSSR-induced expression change seems to be subtler. On the other hand, the study investigates the acute effects (up to 24h) of one unique FSSR exposure. The chronic effects on transcription of multiple FSSR exposures deserve further investigation.

Conclusions

FSSR induced changes on gene expression were dose dependent, with the highest number of differently expressed genes at 6h and high dose (6 SED). The FSSR effect on transcription decreased over time, with a fewer number of genes differently expressed at 24h compared with 6h. Gene-set enrichment analysis suggested a first response involving apoptosis and keratinization, followed by activation of inflammation and immune pathways. No time*dose interactions were detected. At these dose and time conditions, subtle effects of FSSR on miRNA expression were observed.

Ethical declaration

The study was conducted according to the Declaration of Helsinki after approval was obtained from the Ethics Committee of St Thomas's Hospital, London, UK. All participants gave written informed consent.

Acknowledgments

We thank Karl Lawrence, Kylie Morgan and Jessica Greenaway for carrying out some of the irradiations and all the participants for their collaboration.

Funding

This study was supported by CERCA Programme / Generalitat de Catalunya and it was funded by the Instituto de Salud Carlos III FEDER (PI10/02235), the Spanish Ministry of Economy and Competitiveness (MTM2015-68140-R), the European Commission, under the Framework 7 Programme Environment Theme [Contract No. 227020: The Impact of Climate and Environmental Factors on Personal Ultraviolet Radiation Exposure and Human Health (ICEPURE)], and the UK National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London, UK. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the UK Department of Health. Wenjing Kang and Marc R. Friedländer acknowledge funding from the Strategic Research Area program of the Swedish Research Council through Stockholm University.

Author contributions

MB, ARY, MN designed the study. GIH and AT were responsible for ethics, recruitment, logistics, irradiations and dosimetry. YS, EP, MB performed or supervised the laboratory experiments (gene expression and miRNA). CH-F, MB, JRG, WK, and MRF performed bioinformatic and statistical analyses. MB, CH-F, XE, JRG, MN, AT, and ARY interpreted the results. MB and ARY wrote the manuscript.

References

Andersen, C.L., Jensen, J.L., Orntoft, T.F., 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res 64, 5245–5250.

Bisognin, A., Sales, G., Coppe, A., Bortoluzzi, S., Romualdi, C., 2012. MAGIA2: From miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). Nucleic Acids Res. 40, 13–21. doi:10.1093/nar/gks460

Bladt, F., Tafuri, A., Gelkop, S., Langille, L., Pawson, T., 2002. Epidermolysis bullosa and embryonic lethality in mice lacking the multi-PDZ domain protein GRIP1. Proc Natl Acad Sci U S A 99, 6816–21. doi:10.1073/pnas.092130099

Brenner, M., Hearing, V.J., 2008. The protective role of melanin against UV damage in human skin. Photochem. Photobiol. doi:10.1111/j.1751-1097.2007.00226.x

D’Orazio, J., Jarrett, S., Amaro-Ortiz, A., Scott, T., 2013. UV radiation and the skin. Int. J. Mol. Sci. 14, 12222–12248. doi:10.3390/ijms140612222

Damiani, E., Ullrich, S.E., 2016. Understanding the connection between platelet-activating factor, a UV-induced lipid mediator of inflammation, immune suppression and skin cancer. Prog Lipid Res 63, 14–27.

Daniell, H., 2012. NIH Public Access 76, 211–220. doi:10.1007/s11103-011-9767-z.Plastid

Dawes, J.M., Antunes-Martins, A., Perkins, J.R., Paterson, K.J., Sisignano, M., Schmid, R., Rust, W., Hildebrandt, T., Geisslinger, G., Orengo, C., Bennett, D.L., McMahon, S.B., 2014. Genome-wide transcriptional profiling of skin and dorsal root ganglia after ultraviolet-B-induced inflammation. PLoS One 9. doi:10.1371/journal.pone.0093338

Dorhoi, A., Iannaccone, M., Farinacci, M., Faé, K.C., Schreiber, J., Moura-Alves, P., Nouailles, G., Mollenkopf, H.J., Oberbeck-Müller, D., Jörg, S., Heinemann, E., Hahnke, K., Löwe, D., Del Nonno, F., Goletti, D., Capparelli, R., Kaufmann, S.H.E., 2013. MicroRNA-223 controls susceptibility to tuberculosis by regulating lung neutrophil recruitment. J. Clin. Invest. 123, 4836–4848. doi:10.1172/JCI67604

Enk, C.D., Jacob-Hirsch, J., Gal, H., Verbovetski, I., Amariglio, N., Mevorach, D., Ingber, A., Givol, D., Rechavi, G., Hochberg, M., 2006. The UVB-induced gene expression profile of human epidermis in vivo is different from that of cultured keratinocytes. Oncogene 25, 2601–2614. doi:1209292 [pii]10.1038/sj.onc.1209292

Etich, J., Bergmeier, V., Pitzler, L., Brachvogel, B., 2016. Identification of a reference gene for the quantification of mRNA and miRNA expression during skin wound healing. Connect. Tissue Res. 8207, 03008207.2016.1210606. doi:10.1080/03008207.2016.1210606

Friard, O., Re, A., Taverna, D., De Bortoli, M., Cora, D., 2010. CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. BMC Bioinformatics 11, 435. doi:1471-2105-11-435 [pii]\r10.1186/1471-2105-11-435

Giot, L., Bader, J.S., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., Mcdaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., Dasilva, A., Zhong, J., Stanyon, C.A., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., Mckenna, M.P., Chant, J., Lamason, R.L., Mohideen, M.-A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., Sinha, S., Moore, J.L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P.M., O'Donnell, D., Kittles, R., Parra, E.J., Mangini, N.J., Grunwald, D.J., Shriver, M.D., Canfield, V.A., Cheng, K.C., 2003. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science (80-.). 302, 1727–1736. doi:10.1126/science.1116238

Greinert, R., Boguhn, O., Harder, D., Breitbart, E.W., Mitchell, D.L., Volkmer, B., 2000. The dose dependence of cyclobutane dimer induction and repair in UVB-irradiated human keratinocytes. Photochem Photobiol 72, 701–708.

Greinert, R., de Vries, E., Erdmann, F., Espina, C., Auvinen, A., Kesminiene, A., Schüz, J., Schuz, J., 2015. European Code against Cancer 4th Edition: Ultraviolet radiation and cancer. Cancer Epidemiol. 39 Suppl 1, S75–S83. doi:10.1016/j.canep.2014.12.014

Harrison, G.I., Young, A.R., 2002. Ultraviolet radiation-induced erythema in human skin. Methods 28, 14–19.

Hart, P.H., Gorman, S., Finlay-Jones, J.J., 2011. Modulation of the immune system by UV radiation: more than just the effects of vitamin D? Nat Rev Immunol 11, 584–596.

He, H., Jazdzewski, K., Li, W., Liyanarachchi, S., Nagy, R., Volinia, S., Calin, G.A., Liu, C.G., Franssila, K., Suster, S., Kloos, R.T., Croce, C.M., de la Chapelle, A., 2005. The role of microRNA genes in papillary thyroid carcinoma. Proc Natl Acad Sci U S A 102, 19075–19080. doi:10.1073/pnas.0509603102

Hermann, H., Runnel, T., Aab, A., Baurecht, H., Rodriguez, E., Magilnick, N., Urgard, E., Šahmatova, L., Prans, E., Maslovskaja, J., Abram, K., Karelson, M., Kaldvee, B., Reemann, P., Haljasorg, U., Rückert, B., Wawrzyniak, P., Weichenthal, M., Mrowietz, U., Franke, A., Gieger, C., Barker, J., Trembath, R., Tsoi, L.C., Elder, J.T., Tkaczyk, E.R., Kisand, K., Peterson, P., Kingo, K., Boldin, M., Weidinger, S., Akdis, C.A., Rebane, A., 2017. miR-146b Probably Assists miRNA-146a in the Suppression of Keratinocyte Proliferation and Inflammatory Responses in Psoriasis. J. Invest. Dermatol. doi:10.1016/j.jid.2017.05.012

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009b. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37, 1–13.

Kawase, T., Ichikawa, H., Ohta, T., Nozaki, N., Tashiro, F., Ohki, R., Taya, Y., 2008. p53 target gene AEN is a nuclear exonuclease required for p53-dependent apoptosis. Oncogene 27, 3797–810. doi:10.1038/onc.2008.32

Krutmann, J., 2000. Ultraviolet A radiation-induced biological effects in human skin: Relevance for photoaging and photodermatosis. J. Dermatol. Sci. doi:10.1016/S0923-1811(99)00077-8

Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D.G., Lek, M., Lizano, E., Buermans, H.P., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S.B., Donnelly, P., McCarthy, M.I., Flicek, P., Strom, T.M., Geuvadis, C., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S.E., Hasler, R., Syvanen, A.C., van Ommen, G.J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I.G., Estivill, X., Dermitzakis, E.T., 2013. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511. doi:10.1038/nature12531

Li, G., Luna, C., Qiu, J., Epstein, D.L., Gonzalez, P., 2011. Role of miR-204 in the regulation of apoptosis, endoplasmic reticulum stress response, and inflammation in human trabecular meshwork cells. Invest. Ophthalmol. Vis. Sci. 52, 2999–3007. doi:10.1167/iovs.10-6708

Liao, Y., Smyth, G.K., Shi, W., 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41, e108.

Lin, S., Wang, B., Getsios, S., 2012. Eph/ephrin signaling in epidermal differentiation and disease. Semin. Cell Dev. Biol. doi:10.1016/j.semcdb.2011.10.017

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550.

Løvendorf, M.B., Mitsui, H., Zibert, J.R., Røpke, M.A., Hafner, M., Dyring-Andersen, B., Bonefeld, C.M., Krueger, J.G., Skov, L., 2015. Laser capture microdissection followed by next-generation sequencing identifies disease-related microRNAs in psoriatic skin that reflect systemic microRNA changes in psoriasis. Exp. Dermatol. 24, 187–193. doi:10.1111/exd.12604

Maragkakis, M., Alexiou, P., Papadopoulos, G.L., Reczko, M., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Simossis, V. a, Sethupathy, P., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., Hatzigeorgiou, A.G., 2009. Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinformatics 10, 295. doi:10.1186/1471-2105-10-295

Park, H.-Y., Lee, J., González, S., Middelkamp-Hup, M. a, Kapasi, S., Peterson, S., Gilchrest, B. a, 2004. Topical application of a protein kinase C inhibitor reduces skin and hair pigmentation. J. Invest. Dermatol. 122, 159–66. doi:10.1046/j.0022-202X.2003.22134.x

Petersen, B., Thieden, E., Philipsen, P.A., Heydenreich, J., Wulf, H.C., Young, A.R., 2013. Determinants of personal ultraviolet-radiation exposure doses on a sun holiday. Br J Dermatol 168, 1073–1079. doi:10.1111/bjd.12211

Pillai, M.M., Gillen, A.E., Yamamoto, T.M., Kline, E., Brown, J., Flory, K., Hesselberth, J.R., Kabos, P., 2014. HITS-CLIP reveals key regulators of nuclear receptor signaling in breast cancer. Breast Cancer Res. Treat. 146, 85–97. doi:10.1007/s10549-014-3004-9

Quaranta, M., Burden, a D., Griffiths, C.E.M., Worthington, J., Barker, J.N., Trembath, R.C., Capon, F., 2009. Differential contribution of CDKAL1 variants to psoriasis, Crohn's disease and type II diabetes. Genes Immun. 10, 654–8. doi:10.1038/gene.2009.51

R: A language and environment for statistical computing. , 2008. . R Found. Stat. Comput. Vienna, Austria. ISBN 3-900051-07-0, URL http//www.R-project.org.

Ramasamy, K., Shanmugam, M., Balupillai, A., Govindhasamy, K., Gunaseelan, S., Muthusamy, G., Robert, B., Nagarajan, R., 2017. Ultraviolet radiation-induced carcinogenesis: Mechanisms and experimental models. J. Radiat. Cancer Res. 8, 4. doi:10.4103/0973-0168.199301

Rauhala, L., H??m??l??inen, L., Salonen, P., Bart, G., Tammi, M., Pasonen-Sepp??nen, S., Tammi, R., 2013. Low dose ultraviolet b irradiation increases hyaluronan synthesis in epidermal keratinocytes via sequential induction of hyaluronan synthases has1-3 mediated by P38 and Ca2+/calmodulin-dependent protein kinase II (CaMKII) signaling. J. Biol. Chem. 288, 17999–18012. doi:10.1074/jbc.M113.472530

Ridley, A.J., Whiteside, J.R., McMillan, T.J., Allinson, S.L., 2009. Cellular and sub-cellular responses to UVA in relation to carcinogenesis. Int J Radiat Biol 85, 177–195.

Schuch, A.P., Moreno, N.C., Schuch, N.J., Menck, C.F.M., Garcia, C.C.M., 2017. Sunlight damage to cellular DNA: Focus on oxidatively generated lesions. Free Radic. Biol. Med. 107, 110–124. doi:10.1016/j.freeradbiomed.2017.01.029

Sheehan, J.M., Potten, C.S., Young, A.R., 1998. Tanning in human skin types II and III offers modest photoprotection against erythema. Photochem Photobiol 68, 588–592.

Singh, A., Willems, E., Singh, A., Ong, I.M., Verma, A.K., 2016. Ultraviolet radiation-induced differential microRNA expression in the skin of hairless SKH1 mice, a widely used mouse model for dermatology research. Oncotarget 7, 84924–84937. doi:10.18632/oncotarget.12913

Surawska, H., Ma, P.C., Salgia, R., 2004. The role of ephrins and Eph receptors in cancer. Cytokine Growth Factor Rev. doi:10.1016/j.cytogfr.2004.09.002

Syed, D.N., Khan, M.I., Shabbir, M., Mukhtar, H., 2013. MicroRNAs in skin response to UV radiation. Curr. Drug Targets 14, 1128–34. doi:10.1016/j.biotechadv.2011.08.021.Secreted

Tewari, A., Sarkany, R.P., Young, A.R., 2012. UVA1 induces cyclobutane pyrimidine dimers but not 6-4 photoproducts in human skin in vivo. J Invest Dermatol 132, 394–400. doi:10.1038/jid.2011.283

Toll, A., Salgado, R., Espinet, B., Díaz-Lagares, A., Hernández-Ruiz, E., Andrades, E., Sandoval, J., Esteller, M., Pujol, R.M., Hernández-Muñoz, I., 2016. MiR-204 silencing in intraepithelial to invasive cutaneous squamous cell carcinoma progression. Mol. Cancer 15, 53. doi:10.1186/s12943-016-0537-z

Wang, J., Lu, M., Qiu, C., Cui, Q., 2010. TransmiR: a transcription factor-microRNA regulation database. Nucleic Acids Res. 38, D119-22. doi:10.1093/nar/gkp803

Wang, L., Wang, S., Li, W., 2012. RSeQC: quality control of RNA-seq experiments. Bioinformatics 28, 2184–2185.

Wolf, P., Byrne, S.N., Limon-Flores, A.Y., Hoefler, G., Ullrich, S.E., 2016. Serotonin signalling is crucial in the induction of PUVA-induced systemic suppression of delayed-type hypersensitivity but not local apoptosis or inflammation of the skin. Exp. Dermatol. 25, 537–543. doi:10.1111/exd.12990

Xie, F., Xiao, P., Chen, D., Xu, L., Zhang, B., 2012. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. Plant Mol Biol.

Xu, Y., Zhou, B., Wu, D., Yin, Z., Luo, D., 2012. Baicalin modulates microRNA expression in UVB irradiated mouse skin. J. Biomed. Res. 26, 125–134. doi:10.1016/S1674-8301(12)60022-0

3.4 Discussion

Effect of FSSR on Human Blood Transcriptome

Seventeen genes showed decreased expression at 6h, but recovered to their baseline status at 24h or 48h. The direction of the effect of these genes was validated by qPCR. Some are notable because of their relationship with the reported effects of UVR on health. Table 3.1 shows the differentially expressed genes involved in immune regulation, Table 3.2 lists the genes related to cancer and Table 3.3 the genes associated with cardiovascular diseases.

Table 3.1: Differently expressed genes in human blood after FSSR exposure involved in immune regulation processes.

Gene	Regulation	Description
FKBP5	↓	Member of the immunophilin protein family that particpates in immunoregulation. Response to psychological stress. Genetic polymorphisms have been associated with mood disorders. Mechanisms of tumouriogenesis and chemoresistance.
IL13RA1	↓	Subunit of the *IL13* and the *IL4* receptors. High levels of *IL4* and *IL13* have been reported in asthmatic patients after allergic sensitization.
EMR1	↓	Orphan receptor expressed on mature eosinophils, involved in cell adhesion. An anti-*EMR1* antibody induces depletion of eosinophils: potential therapy for eosinophilic diseases.
TLR2	↓	Cooperates with *LY96* to mediate the innate immune response to bacterial lipoproteins and other microbial cell wall components. It has been implicated in the pathogenesis of several autoimmune diseases.
CLECE4	↓	Cell-surface receptor for several ligands and induces the secretion of inflammatory cytokines. Increased levels in bone marrow derived mononuclear cells in rheumatoid arthritis
STAB1	↓	Binds to both Gram-positive and Gram-negative bacteria and may play a role in defense against bacterial infection. First line of defense against infections.
CD83	↑	Cell surface marker of mature dendritic cells that stimulates the immune system. In contrast, the soluble form of *CD83* has been shown to be a potent immunosuppressor.

Other studies have explored the effect of UVB on gene expression in whole blood or in specific blood cell types. Jung et al exposed healthy male volunteers to narrow-band UVB (311 nm) and identified, at 24h, 9 down-

Table 3.2: Differently expressed genes in human blood after FSSR exposure involved in cancer.

Gene	Regulation	Description
FLT3	↓	Class III receptor tyrosine kinase that regulates hematopoiesis [38]. Activating mutations cause acute myeloid leukemia (AML), acute lymphoblastic leukaemia (ALL) and myelodysplasia.
CACNA2D3	↓	Acts as a suppressor gene in nasopharyngeal and esophageal squamous carcinomas. Hypermethylation is a poor prognostic factor in gastric cancer and is associated with breast cancer relapses.
MARVELD1	↓	Microtubule-associated protein that is down-regulated in multiple cancers and silenced by DNA methylation.

Table 3.3: Significant differently expressed genes in human blood after FSSR exposure involved in cardiovascular diseases.

Gene	Regulation	Description
PLA2G7	↓	Phosphilipase that hydrolyzes phospholipids into fatty acids and other lipophilic molecules. Degradation of the platelet-activating factor (PAF): cell proliferation, angiogenesis, immflamation, immunity, and vasodilatation]. PAF transmits the UVR immunosuppressive signal from the skin to the immune system: UVB absorbed in the epidermis → keratinocytes stimulate PAF production → up-regulation of *CXCR4* in mast cell surface → migration of mast cells to the lymph nodes → secretion of *IL10* → suppress contact hypersensitivity, activation of T follicular helper cells, germinal centre formation and antibody formation.
ADORA3	↓	Regulation of several homeostatic processes. Cardioprotective and anti-ischemic effects, and promotes vasoconstriction through mast cell degranulation. Regulates the immune system and is over-expressed in cancer cells and in rheumatoid arthritis, psoriasis, and Crohn's disease.
CPM	↓	Participates in the differentiation of monocytes to macrophages. Part of the kinin-kallikrein system: inflammation, coagulation, pain and blood pressure control (through nitric oxide (NO) production). In humans, whole body UVA lowers blood pressure by release of NO. In rabbits, acute and chronic exposures to UVR increases kinin levels and decreases activity of kininases (CPM).

regulated genes out of the 2,000 candidates tested using custom microarrays followed by qPCR [120]. All photobiological responses show spectral dependence, and monochromatic UVB is not environmentally or physiologically relevant. In human monocyte-derived dendritic cells, exposure to solar simulated UVA/UVB, up-regulated expression of genes involved in cellular stress and inflammation, and down-regulated genes involved in chemotaxis, vesicular transport and RNA processing [121]. These genes do not overlap with those detected in the present study. To the best of our knowledge, our study, which evaluates the blood transcriptome through next generation sequencing, is the most comprehensive of its type to date. Other studies have investigated seasonal effects on gene expression and on blood cell types proportions [122]. Whether seasonal transcriptional patterns are, in part, consequence of differential exposure to UVR through the year deserves further research.

Only *hsa-miR-3940-3p* was differently expressed (5% FDR). A modest change in the effect of FSSR on *hsa-miR-3940-3p* expression was observed after adjustment by 25(OH)D3 levels, but not after adjustment for monocyte proportions. No miRNA showed differential expression after UVR exposure in any of the pair-wise comparisons. Little is known of the role of *hsa-miR-3940-3p*, found at high levels in plasma and serum. There are at least two possible reasons for the lack of changes in other miRNAs. Firstly, our initial sampling time of 6h may be too late because miRNAs are thought to be the first line of defence against stressors. Secondly, we found that *hsa-miR-486-5p*, highly expressed in red blood cells [123], represented 90% of whole blood reads, reducing the chances of detecting other differentially expressed miRNAs at intermediate or low levels.

A bias towards *hsa-miR-486-5p* in samples prepared with the Illumina kit has been reported [124].

Effect of FSSR on Human Skin Transcriptome

The number of differentially expressed genes increased with higher FSSR dose (6 SED) and shorter time post-exposure (6h). At 6h, 132 genes were differently expressed (5% FDR) after exposure to 3 SED. The number increased to 4,071 after exposure to 6 SED (5% FDR), which represents around 20% of the human transcriptome. 97% of the genes detected at low dose (3 SED) were also altered at high dose (6 SED). At 24h, 16 genes were differently expressed (5% FDR) after exposure to 3 SED and 1,583 after exposure to 6 SED. 94% of the genes detected at low dose (3 SED) were also altered at high dose (6 SED). At low dose (3 SED), 7.5% of the genes with FSSR induced changes at 6h were also found at 24h; while at high dose (6 SED), this increased to 29.7%.

Less than 25% of the genes were exclusively identified at 24h. Ten genes (7 down-regulated and 3 up-regulated) showed altered expression patterns after FSSR at all times points and doses.

In general, differentially expressed genes patterns were more pronounced at 6 SED and 6h, suggesting a dose dependent short term effect on the majority of genes.

At 6 SED and 6h, the following biological processes were identified:

- Keratinization
- Apoptosis
- Transcription and translation
- Splicing
- tRNA modifications
- Cytoskeleton organization

At 24h, we detected the same pathways found at 6h except for keratinization and apoptosis. Three new pathways were also detected:

- Inflammation
- Immune response
- Hyaluronan biosynthesis

At low dose, we did not detect any enrichment, except for signalling through different factors at 6h.

Rauhala et al., 2013 already suggested that hyaluronan, also known as hyaluronic acid and one of the main extracellular matrix molecules of

Table 3.4: Differently expressed genes in human skin after FSSR exposure (across both extraction time-points and both FSSR doses).

Gene	Status	Description
AEN	↑	Encodes a nuclear exonuclease required for P53-dependent apoptosis.
CDKAL1	↓	It is expressed in immune cells and becomes downregulated when immune cells are activated with proliferating signals.
EPHB1	↓	It encodes an ephrin receptor tyrosine kinase that mediates cell-cell communication by interacting with ephrin ligands residing on adjacent cell surfaces. They participate in development, maintenance, and repair processes in cutaneous biology.
GRIP1	↓	It is a scaffolding protein required for the formation and integrity of the dermo-epidermal junction
PRKCB	↓	Protein kinase C beta activates *TYR*, the key and rate-limiting enzyme in pigmentation. Topical application of a *PRKCB* inhibitor reduces skin and hair pigmentation
SLC24A3	↓	It is also known as *NCKX3*. It is a Na+/Ca2+ exchanger. *SLC24A5*, another member of the same family, has a role in the development of pigmentation in skin and retinal epithelia.

Table 3.5: Four miRNA were found differently expressed (up-regulated) in human skin after FSSR exposure.

Gene	Status	Target Genes*
Hsa-miR-146b-5p	↑	TMEM237, TMEM132E, LANCL1, SLC6A4
Hsa-miR-223-3p	↑	HLF
Hsa-miR-204-5p	↑	MAPK and STAT3
Hsa-miR-142-5p	↑	–

*Experimentally validated target gene.

epidermal keratinocytes, reported to increase after low dose UVB exposure [125].

Ten genes were differently expressed in all conditions. Nine of them were validated by qPCR. Table 3.4 summarize the most relevant detected genes and their functions, apparat of key genes in the pigmentation processes.

In contrast to genes, the expression of miRNAs was not massively affected by FSSR in the present time and dose conditions. Four miRNAs were differently expressed. Table 3.5 shows them with their possible targeted genes.

Whether this low number of differentially expressed miRNA is because their transcription is less influenced by FSSR or because effects take place in different time point - understood as earlier response process - deserves further investigation. In any case, due to their regulatory role, subtle differences in miRNA levels might be relevant in skin biology.

In agreement with our findings, *hsa-miR-146b-5p* and *hsa-miR-223-3p* have been found to be upregulated in cellular models irradiated with a UV lamp [126] and UVB irradiated mice [127], respectively. Moreover they both seem to be upregulated in skin of psoriasis patients [128], [129]. *SLC6A4* encodes a serotonin transporter, which terminates the action of serotonin and recycles it in a sodium-dependent manner. Activation of the serotonin pathway has been suggested to mediate UVB-induced immune suppression [130].

hsa-miR-204-5p may control the signalling pathway in the progression of actinic keratosis to squamous cell carcinoma [131], and is involved in skin wound [132]. In our data, *hsa-miR-204-5p* levels were inversely correlated with *IL1B* levels [133]. *Hsa-miR-142-5p* has been found to be upregulated in chronically UVR treated mice skin [134].

Strengths & Limitations

The main strength of our study is the evaluation of physiologically and environmentally relevant UVR exposure in humans in vivo. Selected volunteers had a similar skin, type II according to the Fitzpatrick's scale.

The exposure of $\sim$ 3 SED on blood analysis was low compared with Danish holiday makers who received a daily average of 9.4 ± 7.0 SED during a holiday in Tenerife during March, when at least 50% body surface was exposed when the UVB intensity was greatest [135]. The study population was relatively small, but was phenotypically homogenous for erythemal response after similar UVR exposures. Current maximal expression changes were $\sim$1.5-fold reduction (for *FLT3* and *FKBP5*), but the presented FSSR exposure doses were limited by ethical considerations.

The effects FSSR on skin transcription were investigated in biopsies obtained from volunteers locally exposed to FSSR, in contrast to the more artificial in vitro cellular models. Since we did not separate dermis from epidermis, present findings reflect the transcriptional pattern of all cell types in the skin biopsy. Moreover, the transcriptional profile was investigated comprehensively, including genes and miRNAs.

Another strengths of both studies is the validation of results with an independent method (qPCR).

The study also has some limitations. On one hand, the sample size is still limited to detect small effect sizes, which is of special importance for miRNAs, whose FSSR-induced expression change seems to be subtler. On the other hand we have that, due to their design, we can only evaluate acute effects from 6h to 48h post exposure. Then, the link between acute effects on gene expression and chronic effects on health outcomes can only be hypothesized, and should be assessed in future studies.

Chapter 4

Tools Development for Exposome Data Analysis

4.1 Rationale

As the reduction of costs of genomic assays drops, projects in medicine and biotechnology continue generating large volume and diversity of data. While advances in knowledge of the molecular mechanics of human diseases are expected, scientists are continuously challenged on data management and analysis [136], [137].

Major public projects have performed experiments to a group of individuals generating different types of datasets [138]. For instance, the Cancer Genome Atlas (TCGA) [139], is the largest resource available for multi-assay cancer genomics data and the International Cancer Genome Consortium (ICGC) [140] coordinates 55 research projects to characterize the genome, transcriptome and epigenome of multiple tumors.

In addition, large repositories collect data of several smaller projects allowing unified storage and stimulating data sharing. Gene Expression Omnibus (GEO) [141], [142] is the primary database where data from multi-assay experiments is shared publicly. But there are other reference databases like dbSNP [143], a deposit for short genetic variations.

In parallel to this increase of *omic* data and to the improvement of the understanding of the molecular origins of certain human genetic diseases through genomics data, the environment also has a strong influence on our health.

The term *exposome* was coined to describe the totality of human environmental (i.e. non-genetic) exposures from conception onward, complementing the genome [35]. The study of the underlying mechanics that links the *exposome* with human health is an emerging research field with a strong potential to provide new insights into disease etiology [52]. The *exposome* paradigm is to work with multiple exposures at a time and one or more health outcomes rather focus in a single exposures analysis. This approach tends to be a more accurate snapshot of the reality that we live in complex environments with between-exposure confounding, complex interactions, and potential for multi-dimensional joint effects [144]. Exploring how the *exposome* affects multiple intermediate biological layers (e.g. transcriptome, methylome, proteome, metabolome, etc.) will provide insights into the underlying molecular mechanisms linking environmental exposures to health outcomes.

Rapid developments in technology and rapidly declining costs have led to a virtual explosion in the amount of data regarding individuals' exposures over time. Current health studies are able to measure hundreds of exposures simultaneously in the same individuals using combinations of questionnaires, an array of sensors and biochemical assays [145]. Commonly assessed exposures include chemicals in the air, water, food, or household products, as well as information about individual behaviors, activities, and surrounding physical environments.

Some international projects have started to investigate the exposome systematically and including the different layers of complexity [146]. These projects (Table 4.1) have provided a large amount of data on exposures, health outcomes and omics that need to be analyzed and interpreted.

While the conceptual parallel to genomic analyses is attractive, a number of unique methodological challenges need to be overcome. Unlike genomic data, which has an underlying ordering with variables having the same statistical distribution across the genome, *exposome* data have some char-

Table 4.1: Short list of the most relevant research project on *exposome* effects in human health.

Project	**Web Site**
The Human Early-Life Exposome (HELIX Project)	http://www.projecthelix.eu/
The EXPOsOMICS project	http://www.exposomicsproject.eu/
HEALS	http://www.heals-eu.eu/
The Human Exposome Project	http://humanexposomeproject.com/

acteristics that prevent the direct application of methods designed for other omics. Both Minari et. al. [53] and Robinson et. al. [44] disused about them:

- The *exposome* measurement is a study of a heterogeneous constellation of variables that have specific characteristics. This should be understood that the `exposome` is composed by several biomarkers obtained from plasma and blood test having a nature extremely different than the exposures obtained from individual sensors of external sensors.

- Due to the nature of the exposures measured in the *exposome* analysis, the involved statistical test will need to be available for both continuous and categorical (including diatomic and multi-categoric) variables.

- Usually, *exposome* studies group the exposures in families (aka. PCBs, air pollutants, heavy metals...) [147]. This leads to a dense correlation structure within variables in the same family. Hence, the presence of high correlation between exposures within families makes difficult to disentangle the effect of individual exposures [148].

- The conception of *exposome* is linked to time dependency. Due to the nature of certain exposures, the *expoomse* is also linked to spatial dependency.

Bioconductor initiative raised in 2001 with the aim to provide a portal for free software - written in R - to centralize and standardize methods to analyze high-throughput biological data [149], [150]. Both the core members of the project and the community has made a great effort to provide a standard infrastructure to represent biological data.

Several R/Bioconductor packages implement methods to analyze, integrate and visualize biological data. Each of these packages implements a different strategy to manage input biological data and to perform the analysis. Therefore a standard structure to manage in a coordinated multiple data-sets of the different *omic* types, allowing to include *exposome* data, obtained from the same individuals is required. At the same time, the development of a basic structure to contain the *exposome* data that could be transferred to the previous multi-set structure is mandatory. Then, this will allows operating at exposures level for *exposome*-health analysis and exposome-omic analysis. Moreover, following the tendency of current genetic analysis, a tool to perform enrichment analysis based on exposures instead of genes will be the next step.

4.2 Methods

4.2.1 Coordinated Data Organization System for Multiple Omic Data-Sets

`MultiDataSet` is a new R class based on Bioconductor standards developed to encapsulate multiple data-sets. `MultiDataSet` deals with the usual difficulties of managing multiple and non-complete data-sets while offering a simple way of subsetting features and selecting individuals.

Its structure is an extension of the abstract `eSet` class. `MultiDataSet` is therefore a data-storage class that comprises data-sets of different omic data (assay data), feature data and phenotypic data. Despite its general form, `MultiDataSet` maintains the specific characteristics of the datasets (e.g. it preserves matrices of calls and probabilities of a `SnpSet`).

The internal structure of `MultiDataSet` comprises five fields that are R standard `lists`. Their names match other Bioconductor classes: `assayData` containing the measurement values; `phenoData` that stores the description of the samples; `featureData` and `rowRanges` that have the description of the features; and `return_method` that allows recovering the original dataset.

Six accessors are available to retrieve information from each `MultiDataSet`'s fields: `assayData`, `pData`, `fData`, `rowRanges`, `rowRangesElements` and `sampleNames`. The first four retrieve the content of `assayData`, `phenoData`, `featureData` and `rowRanges`. `rowRangesElements` returns the names of datasets with a genomic coordinates in a `GenomicRanges`. The accessor `sampleNames` returns a named list with the samples names of each data set.

Following Bioconductor guidelines, `MultiDataSet` objects are created empty through its constructor. Once the object is created, data-sets of class `eSet` can be added with `add_eset` and data-sets of class `SummarizedExperiment` can be added using `add_rse`. `MultiDataSet` package incorporates three *specific functions* to include specific omic data sets: `ExpressionSet` and `SnpSet` from `Biobase` package, `MethylationSet` from `MultiDataSet` package and `GenomicRatioSet` from `minfi` package.

Adding a new type of data to `MultiDataSet` objects (data-sets with a type of data that is not natively supported in `MultiDataSet` package) is easy by implementing a new *specific function* that validates the data and then uses `add_eset` or `add_rse` to incorporate the data-set to the `MultiDataSet` objects.

4.2.2 Framework for Exposome Data Analysis

In order to enable comprehensive analyses of the *exposome* and its connections to human health the *rexposome project* was created. *rexposome project* is a compendium of two R packages, `rexposome` and `omicRexposome`, that provide a freely available and open-source framework for robust, scalable and reproducible of state-of-the-art methods to perform a comprehensive analysis of the exposome and its relationship with health outcomes and molecular intermediates.

rexposome project envisions a typical sequence of analyses for *exposome* data:

1. Detailed characterization of the *exposome*.

2. Linear and non-linear association between exposure and health outcome.

3. Univariate and multivariate association between exposures and molecular intermediates (*omic* data).

To this end, `rexposome` R package implements a new R class, `ExposomeSet`, that is based in standard `eSet` Bioconductor and designed to incorporate exposome data into R/Bioconductor analysis framework. Objects of class `ExposomeSet` can be created from data files using the function `readExposome` or from standard `data.frame`s using `loadExposome`.

In some occasions *exposome* data can contains missing information. In that case, multiple imputation process is suggested to be applied to estimate the values of the missing data. To properly use the multiple-imputed *exposome* data, `rexposome` implements the new R class `imExposomeSet`, that can be created using `loadImputed`. Methods for downstream association analyses using `imExposomeSet` objects are already implemented in `rexposome`.

The internal structure of `ExposomeSet` is homologous to standard `eSet`-like objects. Then, three internal fields stores the exposome data: 1) `assayData`, having the matrix of exposures; 2) `phenotypeData`, having the table of phenotypes, covariates and health outcomes; and 3) `featureData`, having the exposure' description table. All these objects are coordinated by: 1) the names of the rows of the `assayData` and the `phenotypeData` are the same and in the same order; 2) the names of the columns of the `assayData` and the names of the rows of the `featureData` are the same and in the same order.

Internal organization of `imExposomeSet` objects is the same as `ExposomeSet`. But their tables must contains two special columns, labelled as ".imp" and ".id". ".imp" column must contain the number of imputation set (starting for 1, since 0 is for raw data) and column ".id" must contain the real samples' ID.

Accomplishing with the steps of the exposome data analysis pipeline, the characterization process of the *exposome* has the aim to better understand the underlying structure of the exposure data, prior to investigating their associations with markers of health and disease. This process includes: 1) data pre-processing; 2) descriptive analyses; and 3) association analyses.

For *exposome* pre-processing step, `rexposome` implements the method `trans`, which allows to apply a transformation function to `ExposomeSet`'s exposures. The transformation aims to guarantee normality assumption, required in downstream analyses. The method `standardize` allows the user to standardize the data by using "normal standardization" $\left(\frac{x-\bar{x}}{s_x}\right)$, "robust standardization" $\left(\frac{x-\widetilde{x}}{MAD_x}\right)$ or "interquartile range standardization" $\left(\frac{x}{p_{75}(x)-p_{25}(x)}\right)$. Once *exposome* data is transformed, its can be categorized into variables coded as low/high exposed values by using tertiles, quartiles or any other criteria by using the function `highAndLow`.

`rexposome` package contains functions to perform basic description of both exposures and phenotypes. There `plotFamily` method allow the user to get box-plots (continuous variables) or accumulated-bar plots (categorical variables) by family of exposures. Correlation between exposures from an `ExposomeSet` can be computed (method `correlation`). The nature of the two involved exposures is taken into account: continuous vs. continuous uses `cor` function from R `base`; categorical vs. categorical uses `cramerV` function from `lsr` R package; and categorical vs. continuous exposures correlation is calculated as the square root of the adjusted r-square obtained from fitting a lineal model with the categorical exposures as dependent variable and the continuous exposure as independent variable.

Principal Component Analysis (PCA) can be applied to *exposome* data (method `pca`). PCA on *exposome* data is a way to explore the relationship between exposures through the principal components. The correlation of the exposures with the principal components can be obtained and plotted, indicating the "meaning" of each principal component in terms of exposures. The association between the phenotypes, covariates and health outcome with the same principal components can also be obtained, and, therefore, will help in deciphering the possible relationship between exposures and phenotypes.

Finally, and after the descriptive of the *exposome*, its correlation and the

PCA, in some occasions it may be useful to analyze the *exposome* as a whole. If so, a clustering analysis, based on samples, can be performed to cluster individuals having similar exposure profiles. This can be done in `rexposome` package by using method `clustering` that has been designed to accept any implemented clustering method.

`rexposome` provides two different approaches to analyze the association between *exposome* data and health outcomes. The first of them is called Exposome-Wide Association Study (ExWAS) that is equivalent to a Genome-Wide Association Study (GWAS) in genomics or to Epigenetic-Wide Association Study (EWAS) in epigenomics. The ExWAS was first described by Patel et al. [54]. This type of analysis is performed using the method `exwas` for `ExposomeSet` objects. The equivalent analysis was designed for `imExposomeSet` where an analysis is done for each imputed set and P-Values are pooled to obtain a global association score. Results are encapsulated in `ExWAS` objects than can be plotted and exported to standard tables and files. Multiple comparisons in the ExWAS analysis is addressed by computing the number of effective ($\mathrm{N_{eff}}$) tests as described by Li and Ju [151]. The method estimates $\mathrm{N_{eff}}$ by using the exposure correlation matrix that is corrected when it is not positive definite by using `nearPD` from `Matrix` package. The significant threshold is computed as $1 - (1 - 0.05)^{\mathrm{N_{eff}}}$.

There are some authors that proposed to perform association analysis in a multivariate fashion, just to take into account the correlation across exposures [152]. To this end, `rexposome` allows to perform a multivariate association analysis between the *exposome* and health outcomes using Elastic-Net regularized generalized linear models, from `glmnet` R package. The procedure is encapsulated in the `mexwas` method.

`omicRexposome` provides the function association that perform association analyses between exposures and molecular signatures (*omic* data). The function, `association` takes as input a `MultiDataSet` object [153]. This object must contain an `ExposomeSet` and an *omic container* (`ExpressionSet`, `SummarizedExperiment`, `MethylationSet`, `GenomicRatioSet`...) and fits linear models as described in the `limma` R package [154], [155]. The pipeline implemented in association allows performing surrogate variable analysis in order to correct for unwanted variability. This adjustment is provided by SmartSVA, isva [156] and SVA R packages.

For multivariate association studies involving *exposome* and molecular signatures omicRexposome provides the method crossomics, allows performing integration analysis using different approaches: multiple co-inertia analysis (MCIA), that is implemented in the omicade4 R package; and Multi-Canonical Correlation Analysis (MultiCCA) that is implemented in the PMA R package.

MCIA is an analysis method that identifies co-relationships between multiple data-sets by projecting the features of the multiple data-sets into a single dimensional space. By using this approach the most relevant features from each data-set can be obtained.

MultiCCA is an extension of Canonical Correlation Analysis - which has gained popularity as a method for the integration of several omic data - and provides a sparse version of it. The main advantage regarding MCIA is that this method generates a list of features whose loadings are statistically significantly different from zero.

4.2.3 Post-exposome Data Analysis: Enrichment Based in Exposures

The *Comparative Toxicogenomics Database* (CTD$^{\mathrm{TM}}$; http://ctdbase.org) is a public resource for toxicogenomic information manually curated from the peer-reviewed scientific literature, providing key information about the interactions of environmental chemicals with gene products and their effect on human disease [157]. CTD provides information of a triad of core interactions describing chemical-gene, chemical-disease and gene-disease relationships. It includes more than 30.5 million toxicogenomic connections relating chemicals/drugs, genes/proteins, diseases, exposures, Gene Ontology (GO) annotations, pathways (KEGG/Reactome), and gene interaction modules [158].

The CTDquerier R package has been developed to allow R/Bioconductor user to query CTD from an R session. It facilitates the inclusion of CTD data in downstream statistical and enrichment analyses in

R/Bioconductor pipelines. CTDquerier R package allows querying CTD by genes (query_ctd_gene), by chemicals or exposures (query_ctd_chem) and by diseases (query_ctd_dise). The queries can be performed using a single or multiple terms (gene names in gene symbol format, chemical names or disease names).

The three functions to query CTD return an object of class CTDquery, that encapsulates the data retrieved from the data base. This class ensures compatibility with R/Bioconductor third packages by implement 4 methods: get_terms allows to see the terms that were validated in CTD; extract returns the data retrieved from CTD, enrich performs a Fisher's exact test for testing the enrichment between two CTDquery objects. Moreover, the plot functions allows to obtain different representations of the obtained data.

4.3 Results

Three manuscripts corresponds to the results of this project. The first manuscript corresponds to the design of the MultiDataSet R/Bioconductor package. The second manuscript corresponds to the implementation of the *rexposome project* (rexposome and omicRexposome R/Bioconductor packages). Lastly, the third manuscript corresponds to CTDquerier R package.

4.3.1 `MultiDataSet` R/Bioconductor package

Hernandez-Ferrer C, Ruiz-Arenas C, Beltran-Gomila A, González JR. MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration. BMC Bioinformatics. 2017 Jan 17;18(1):36. DOI: 10.1186/s12859-016-1455-1

4.3.2 *rexposome project* – `rexposome` and `omicRexposome` R/Bioconductor packages

Journal	Nat. Methods
Title	Comprehensive analysis of the exposome, exposome-health associations and omics intermediates
Authors	Carles Hernandez-Ferrer, Gregory A. Wellenius, Ibon Tamayo, Xavier Basagaña, Jordi Sunyer, Martine Vrijheid, Juan R. González
Status	Submitted (August 1, 2017)

Comprehensive analysis of the exposome, exposome-health associations and omics intermediates

Carles Hernandez-Ferrer[1,2,3], Gregory A. Wellenius[5], Ibon Tamayo[1,2,3], Xavier Basagaña[1,2,3], Jordi Sunyer[1,4,2,3], Martine Vrijheid[1,2,3], Juan R. Gonzalez*[,1,2,3]

Affiliations

1. ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

2. Universitat Pompeu Fabra (UPF), Barcelona, Spain.

3. CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.

4. IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain.

5. Department of Epidemiology, Brown University School of Public Health, Rhode Island, USA.

* Correspondece to Juan R. Gonzalez <juanr.gonzalez@isglobal.org>

ABSTRACT

Genomics has dramatically improved our understanding of the molecular origins of certain human diseases. However, our health is also influenced by the cumulative impact of exposures experienced across the life course (termed “exposome”). The study of the high-dimensional exposome offers a new paradigm for investigating environmental contributions to disease etiology. However, there is a lack of advanced bioinformatics tools for managing these data, characterizing the exposome and associating the exposome to health outcomes and different omic layers. We provide a generic framework in the R/Bioconductor architecture that includes object oriented classes and methods to leverage high-dimensional exposome data in disease association studies including its integration with a variety of high-throughput data types. The *rexposome* project offers a free infrastructure for robust, scalable, reproducible and open-source development of state-of-the-art methods to perform a comprehensive analysis of the exposome and its consequences on human health.

MAIN TEXT

Although genomics has dramatically improved our understanding of the molecular origins of certain human genetic diseases, the environment also has a strong influence on our health. As an illustration, the World Health Organization’s Global Burden of Disease collaboration estimates that 79 behavioral, environmental, occupational and metabolic risks factors account for half of global mortality and more than a third of global disability-adjusted life-years[1]. The term “exposome” was coined by Wild to describe the totality of human environmental (i.e. non-genetic) exposures from conception onward, complementing the genome[2]. The study of the links between the exposome and health is an emerging research field with a strong potential to provide new insights into disease etiology[3,4,5,6]. The exposome paradigm breaks with the common practice in observational epidemiology of analyzing and reporting the association between a single health outcome and a single exposure at a time. Studies of the exposome can provide a more holistic view of the complex relationships between constellations of exposures and one or more health outcomes, accounting for the reality that we live in complex environments with between-exposure confounding, complex interactions, and potential for multi-dimensional joint effects[7]. Moreover, the exposome approach may reduce publication bias since findings of positive, negative, and null associations are equally likely to be published if found and the rate of false positive results is decreased by properly controlling for multiple testing. In addition, exploring how the exposome affects multiple intermediate biological layers (e.g. transcriptome, methylome, proteome, metabolome etc.) can provide insights on the underlying molecular mechanisms linking environmental exposures to health effects.

Rapid developments in technology and rapidly declining costs have led to a virtual explosion in the amount of data regarding individuals’ exposures over time. For example, current health studies are able to measure hundreds of exposures simultaneously in the same individuals[3] using combinations of questionnaires, an array of sensors and biochemical assays. Commonly assessed exposures include chemicals in the air, water, food, or household products, as well as information about individual behaviors, activities, and surrounding physical environments. Given the growing numbers of variables commonly collected, there is a pressing and palpable need for the development of novel methods, pipelines and platforms that allow systematic and semi-automatic analyses such as those conducted in the genomics or other omic analyses.

While the conceptual parallel to genomic analyses is attractive, a number of unique methodologic challenges need to be overcome. Unlike genomic data, which has an underlying ordering with variables having the same statistical distribution across the genome, exposome data have some characteristics that prevent the direct application of methods designed for other omics. These include having multiple markers for the same underlying exposure (e.g. questionnaire-based data on exposure to certain chemical, data from personal monitors, or data from biomarkers in blood or urine); different degrees of measurement error; diverse patters of missing data; and heterogeneity of distributions, including censored variables (e.g. values below the limit of detection).

To enable comprehensive analyses of the exposome and its connections to human health, we have created the *rexposome* project. *rexposome* provides a freely available framework for robust, scalable, reproducible and open-source development of state-of-the-art methods to perform a comprehensive analysis of the exposome and its relationship with health outcomes and molecular intermediates. The project includes complementary R/Bioconductor packages capable of addressing a variety of scientific questions (**see On-Line Methods**).

We envision the following typical sequence of analyses: (i) detailed characterization of the exposome, (ii) exposure signature discovery, (iii) linear and non-linear exposure-disease association studies and (iv) omic-exposure association and integration.

The aim of characterizing the exposome is to better understand the underlying structure of the exposure data prior to investigating their associations with markers of health and disease. This process includes: 1) data preprocessing (understanding missing data, imputation of missing data, and data transformation); 2) descriptive analyses; and 3) association analyses.

Missing data is a common problem in epidemiological studies[8], which is aggravated in studies of the exposome due to the high dimensionality. Tools to describe the missing data patterns have been implemented in *rexposome*. Therefore, the characterization of the exposome begins by describing the patterns of missingness and then performing imputation to enhance downstream analyses. A comprehensive approach to imputation of missing values in the setting of high dimensional data is beyond the scope of *rexposome*. Nonetheless, *rexposome* incorporates a naïve method for missing data imputation with the goal of allowing users to perform a complete analysis without the loss of statistical power typical of complete case analyses (see On-line Methods). We recommend the user to properly perform multiple imputation (for instance, by using *mice* R package) and then incorporate the multiple imputed data into the *rexposome* framework to perform downstream analyses (**see On-line Methods and Supplementary Material 3**).

Once the exposome is free of missing data, several descriptive analyses can be performed to characterize the exposome. Box-plots for continuous exposures and cumulative bar-plots for categorical exposures can be obtained for any single exposure, family of exposures or for the whole exposome (**Figure 1A and Supplementary Material 2 - Section 2.2.4**). The exposome can be further characterized by exploring patterns of correlation within and between groups of exposures or exposure families (**Supplementary Material 2 - Section 2.2.6**). Principal component analysis (PCA) can be used to gain insights into how exposures are grouped and whether any quantitative trait may explain differences in exposure patterns (**Supplementary Material 2 - Section 2.2.5**). PCA will also facilitate identification of constellations of exposures that explain much or most of the variability in the exposome[9] (**Figure 1B and Supplementary Material 2 - Section 2.3.1)**. Clustering analysis can be used to identify signatures of exposures (**Supplementary Material 2 - Section 2.2.7)**. These signatures will help when associating the whole exposome with omics data-sets representing molecular intermediates between exposure and health.

The association between the exposome and health can be assessed by univariate or multivariate analysis of the exposures[10,11]. Univariate analysis is performed by performing Exposome-Wide Association Study (ExWAS) which is conceptually analogous to a Genome-Wide Association Study (GWAS) based on a high dimensional set of exposure metrics rather than gene loci[12]. A P-value of the association between each exposure and a given health outcome is computed from a generalized linear model using likelihood ratio or score tests. Some exposures do not present a linear dose-response relationship. Therefore, the association analysis can be performed using splines or any other non-linear method. The function to perform ExWAS includes an argument to define any type of modeling that includes the possibility of adjusting for confounding variables, interactions among variables, or splines to capture any possible non-linear exposure-outcome relationship (**Supplementary Material 2 - Section 2.3.2**).

Importantly, *rexposome* provides several methods for visualizing the results of analyses. Manhattan-like plots provide a visual representation of the statistical significance of each association with p-values grouped by family of exposure (**Figure 1C)**. The effects of different models to test the proper set of covariates can be visually inspected by using forest plots and volcano plots. The multivariate association analysis of the full exposome and health outcomes is performed by using elastic net, a penalized regression method that outperforms other multivariate methods in terms of controlling the false discovery rate (FDR) while maximizing statistical power[13]. Results can be visualized in a heat-map showing the level of the coefficients assigned to each exposure.

An important aim of exposomics is the identification of molecular signatures (methylome, transcriptome, proteome) that can represent molecular biomarkers of exposure. Currently there is interest on testing the association between the exposome and different high-throughput data types. These include genome[14], transcriptome[15] and epigenome[16]. The outcome of exposome-omic association studies will provide a list of molecular signatures (genes, CpGs...) associated with a single exposure or exposure signatures. The association analyses can be performed exposure by exposure or using exposure signature level by using state-of-the-art methods implemented in widely used Bioconductor packages (**Figure 1D and Figure 1E. Also see On-Line Methods**).

A step further in exposomics is to integrate both the measurement of environmental factors with molecular signatures. The aim is to investigate the complex interactions between exposures and multiple omic features[17]. To achieve this goal, we implement two multivariate methods that are able to integrate multiple tables: sparse multiple canonical correlation analysis and multiple co-inertia analysis (**see On-Line Methods)**.

We demonstrate the usefulness of this platform through two applied examples. First, we investigate the association between PCBs, PDCs and PBDEs and markers of obesity using cross-sectional data from the US National Health and Nutrition Examination Survey (NHANES). **Supplementary Material 1** illustrates how to perform this analysis of 41 exposures and 3 obesity-related phenotypes among 41,473 individuals (**see Online Methods**). The analyses performed using *rexposome* showed as PCBs, PDCs and BDEs are highly correlated with obesity (**Supplementary Material 1**). In a second applied example, we used *rexposome* and *omicRexposome* to characterize the exposome and its association with omics data (transcriptome and methylome) from a prospective pregnancy cohort in Spain, the INMA-Sabadell study (**Supplementary Material 1**). The analysis revealed that exposure to PCBs can alter different pathways related to brain and neurodevelopmental regulation processes (**Figure 1E and Supplementary Material 1 – section 3.5**). This would be highly relevant in this study since one of the main objectives in INMA is to established mechanisms linking exposome, biomarkers and neurodevelopmental disorders in children.

To conclude, we have developed *rexposome* as an open-source platform to facilitate the application of novel methods to the analysis of the exposome and its connections to markers of human health, disease, and molecular signatures. In a time in which modern exposure assessment tools are providing ever-increasing amounts of detailed exposure data for epidemiologic studies, our tool can facilitate their analysis and the discovery of new health determinants.

Figure 1. [top-left] Box-plots indicating the level of maternal serum concentration of each member of the PCBs family (INMA-Sabadell). [top-middle] First and second principal components showing the exposures space (INMA-Sabadell). [top-right] Manhattan plot showing the results obtained from the Exposome-Wide Association Study on asthma (INMA-Sabadell). [bottom-left] QQ plot showing the P-Values of the association between the cluster of exposures and transcriptome (INMA-Sabadell). [bottom-middle] Volcano plot indicating the P-Value (Y axis) and the strength of the association (beta coefficient) between PCBs and each loci (X axis) of the association between the cluster of exposures and transcriptome (INMA-Sabadell). [bottom-right] Heat-map with the P-Value obtained from the pathway enrichment analysis from PCBs family and transcriptome (INMA-Sabadell).

ON-LINE METHODS

rexposome Project Overview

The aim of *rexposome* project is to provide a free framework for robust, scalable, reproducible and open-source development of statistical methodologies of exposome. It includes data characterization and analysis and integration with other omic data. To this end, the project includes *rexposome* and *omicRexposome* that are written in the Rprogramming language (http://www.r-project.org) and use Bioconductor infrastructure (https://www.bioconductor.org). They are available under the MIT open source license and they have been submitted for inclusion in Bioconductor. Development version is available in the GitHub repository of our group (https://github.com/isglobal-brge). The main package is *rexposome* that provides a set of functions and methods to load exposome data into R environment and prepare it to perform exposome analysis. This allows users for full flexibility on adapting the exposome data to their goals. It contains functions to perform required exposome data processing steps such as: exposure categorization (low/high exposure), data transformation and standardization.The *omicRexposome* package focuses on performing exposome-omic data association (based on *limma* R package - https://bioconductor.org/packages/limma) and exposome-omic data integration (based on both *omicade4* - https://bioconductor.org/packages/omicade4 - and *PMA* - https://cran.r-project.org/package=PMA - R packages).

Data Description

NHANES

We download 41 exposures and 5 phenotypes (mean diastolic, mean systolic, weight, BMI and maximal calf circumference) from *nhanes-prod* database for a total of 41473 individuals (See Supplementary Table 1 and Supplementary Table 2).

INMA-Sabadell Cohort

We subset original INMA-Sabadell Cohort selecting 32 exposures, 3 health outcomes (asthma, rhinitis and whistles) and 3 phenotypes (sex, age, cbmi) for 657 individuals from full pregnancy period or third semester if full pregnancy period not available. Exposures included matrix source and limit-of-detection (See Supplementary Table 3).

Exposome Data Loading

rexposome R package is designed to load exposome data from files in external format (.csv, .tsv…) or standard R *data.frame* objects. The function readExposome requires three external files (.csv, .tsv…) in specific format (**Supplementary Material 2 – Section 2.1.1**). One file must contain a table with the exposures, having the samples as rows and the exposures as columns. The second file must contain the phenotypic data (e.g. it includes co-variates and outcomes of interests), having the samples as rows and the phenotypes as columns. The third file must contain the description of the exposures, with the exposures as rows and the descriptors as columns (**Supplementary Material 2 - Section 1.3**). The minimum requirement of this file is that it hast to provide two columns, one with the name of the exposure and another with the family each exposure belongs to. If all these tables are already loaded into R, the function loadExposome can be used to encapsulate all this information into a single object (**Supplementary Material 2 – Section 2.1.2**). In both cases, the functions create an object of class *ExposomeSet*. In some occasions, exposome contains missing information. In that case, multiple imputation process can be applied to generate multiple tables of imputed exposures. These imputed datasets are then used to properly perform association studies by using multiple imputation techniques (paper). *rexposome*, can deal with these multiple imputed tables by encapsulating them into an *imExposomeSet* object using the function loadImputed. Methods for downstream association analyses using *imExposomeSet* objects are already implemented in *rexposome* (**Supplementary Material 3 – Section from 1.1 to 1.3**).

The internal structure of *ExposomeSet* is homologous to standard *ExpressionSet* objects. This means that three elements are stored in different fields: 1) “assayData” having the matrix of exposures as an *assayData* object; 2) “phenotypeData” having the table of phenotypes and outcomes as an *AnnotatedDataFrame* object; and 3) “featureData” having the exposure description table as an *AnnotatedDataFrame*. All these objects are coordinated by: 1) the names of the rows of the “assayData” and the “phenotypeData” are the same and in the same order; 2) the names of the columns of the “assayData” and the names of the rows of the “featureData” are the same and in the same order (**Supplementary Material 2 – Figure 2**). This is a bit different for *imExposomeSet* objects that having the same elements with the same names it stores the information in *DataFrame*s objects. Moreover, these *DataFrame*s needs to have two columns labeled as “.imp” and “.id”, having in “.imp” the number of imputation set (starting for 1, since 0 is for raw data) and in the real samples’ ID in “.id”.

<u>Omic Data Loading</u>

The main input data for *omicRexposome* are *ExposomeSet* and *ExpressionSet* objects (also *SummarizedExperiment*), encapsulated in a *MultiDataSet* object.Both sets are used to test the association between the exposome in the *ExposomeSet* and any type of omic measure (transcriptome, methylome, proteome, metabolome…). The results are encapsulated in *ResultSet* objects, than can be used to visualize the results.

Exposome Data Pre-processing

<u>Data Transformation</u>

rexposome offers full flexibility to the user to pre-process its exposome data. The package has several methods that perform the main pre-processing steps that are required when dealing with the exposome. The method trans allows the user to apply any function to transform the exposures encapsulated in an *ExposomeSet*. The transformation aims to guarantee normality assumption that is required in downstream analyses (**Supplementary Material 2 – Section 2.2.2**). The method standardize allows the user to standardize the data by using three different procedures (**Supplementary Material 2 – Section 2.2.5**):

1. normal standardization:
2. robust standardization: $\left({}^{x-\tilde{x}}/_{MAD_x}\right)$
3. interquartile range (IQR): $\left({}^{x}/_{p_{75}(x)-p_{25}(x)}\right)$

Once data has been transformed the user can categorize the exposures into variables coded as low/high exposed values by using tertiles, quartiles or any other criteria by using the function highAndLow.

Missing Data

As previously stated, exposome data can be loaded having missing data. In our settings two different types of missingness patterns can be observed: missing at random and missing due to limit of detection (LOD). *rexposome* contains functions to create tables and plots to study missing patterns. Although performing multiple imputation of exposome is beyond the scope of *rexposome* project we have described in the Supplementary Material 1 and 3 how to perform either a single imputation method - *Hmisc* (https://cran.r-project.org/package=Hmisc) - or a more sophisticated one based on multiple imputation - *mice* (https://cran.r-project.org/package=mice). Therefore, we recommend the user to perform imputation in a separate framework and then incorporate imputed data (i.e. multiple tables) into *rexposome* as an *imExposomeSet* object as described in **Supplementary Material 1** and 3.

Exposome Characterization

The *rexposome* package contains functions to perform basic description of both exposures and phenotypes. There *Summary* and *plotFamily* methods allow the user to get descriptive statistics and box-plots (continuous variables) or accumulated-bar plots (categorical variables) by family of exposures (**Supplementary Material 2 – Section 2.2.4**). Additionally, *rexposome* is able to compute the correlation between exposures. The correlation is computed using the method correlation. The method takes into account the nature of each pair of exposures: continuous vs. continuous uses *cor* function from R *base*, categorical vs. categorical uses *cramerV* function from *lsr* R package (https://cran.r-project.org/package=lsr) and categorical vs. continuous exposures correlation is calculated as the square root of the adjusted r-square obtained from fitting a lineal model with the categorical exposures as dependent variable and the continuous exposure as independent variable (**Supplementary Material 2 – Section 2.2.6**).

Principal Component Analysis (PCA) can be applied to exposome data (with method *pca*) as an alternative way to explore the relation between exposures (through the principal components) (**Supplementary Material 2 – Section 2.2.5**). Additionally it allows exploring the correlation of the exposures with the principal components and also the association of the phenotypes with the same principal components (**Supplementary Material 2 – Section 2.3.1**). This type of analysis will help in deciphering possible relationship between exposures and phenotypes.

Finally, in some occasions the user may want to analyze the exposome as a whole. If so, a clustering analysis on samples can be performed to cluseter individuals having similar exposure profiles. This can be done in the rexposome package by using method *clustering* that has been designed to accept any clustering method the user can implement. The **Supplementary Material 2 – Section 2.2.6** illustrates how to perform this by using hierarchical clustering. The exposure signature that is generated by using this method can be used later in *omicRexposome* to test the association of the exposome signatures with any omic data-set.

Exposome Association Analysis

Single Association Analysis

rexposome provides to different approaches to analyze the association between exposome data and health outcomes. The first of them is called Exposome-Wide Association Study (ExWAS) that is equivalent to a Genome-Wide Association Study (GWAS) in genomics or to Epigenetic-Wide Association Study (EWAS) in epigenomics. The ExWAS was first described by Patel et al.[12]. This type of analysis is performed using the method *exwas* for *ExposomeSet* objects (**Supplementary Material 2 – Section 2.3.2**). Equivalent analysis was designed (in the homologous exwas method) for *imExposomeSet* where an analysis is done for each imputed set and P-Values are pooled to obtain a global association score (**Supplementary Material 3 – Section 2**). Results are encapsulated in an *ExWAS* objects than can be plotted and exported to standard tables and files (*plotExwas*, extract methods that perform these operations). The statistical analyses behind ExWAS are based on generalized linear models. The function *exwas* allows the user to indicate any formula describing the model that should be adjusted for. This follows standard formula options in R. That is, continuous or factor variables can be incorporated in the design, as well as interaction or splines using standard R functions and formulas. Multiple comparisons in the ExWAS analysis is addressed by computing the number of effective (N_{eff}) tests as described by Li and Ju[18]. The method estimates N_{eff} by using the exposure correlation matrix that is corrected when it is not positive definite by using *nearPD* R function. The significant threshold is computed as $1-(1-0.05)^{Meff}$. This threshold is added to the Manhattan-like plots obtained from *ExWAS* objects. *ExWAS* objects also offers *plotEffect* that allows to plot the effect of each coefficient of a given model or to compare the effects of the same coefficient given to models (two *ExWAS* objects).

Multivariate Association Analysis

There are some authors that proposed to perform association analysis in a multivariate fashion, just to take into account the correlation across exposures[13]. *rexposome* has implemented a method to perform association analysis between the exposome and health outcomes using Elastic-Net regularized generalized linear models implemented in *glmnet* R package (https://cran.r-project.org/package=glmnet). The procedure is encapsulated in the *mexwas* method (**Supplementary Material – Section 2.3.3**). Results are encapsulated in a *mExWAS* object than can be plot and exported to standard R *data.frame*s (*plotExwas* and extract methods can deal with *mExWAS* objects too).

Exposome-Omic Association Analysis

omicRexposome provides the function *association* that perform association analyses between exposures and omic data. This function takes as input a *MultiDataSet* object containing an *ExposomeSet* and an omic container (or *ExpressionSet*, *SummarizedExperiment*, *MethylationSet*, *GenomicRatioSet*...) and fits linear models as described in the *limma* R package[19]. The pipeline implemented in association allows performing surrogate variable analysis in order to correct for unwanted variability. This adjustment is provided by *SmartSVA* (https://cran.r-project.org/web/packages/SmartSVA) and *SVA* R package (https://bioconductor.org/packages/SVA). Function association allows analyze a single exposure, the full exposome or a subset of exposures. association returns an object of class *ResultSet*, that contains the results obtained from the *limma* pipeline joint with other information that can be latter use to enrich the results (the exact number of individuals used, both *featureData* from *ExposomeSet* and *ExpressionSet* and the arguments passed to association). *omicRexposome* implements a series of methods for *ResultsSet* that can be useful for the user:

- *tableLambda* provides the lambda value, that accounts for inflation, of each association analysis (e.g. each exposure),
- *tableHits* returns the number of hits per association analysis (e.g. each exposure) given a threshold of a given P-Value,
- *plotAssociation* allows to create a QQ-plot, Manhattan or Volcano plot of a given association analysis (e.g. selected exposure).

Examples about how to obtain these results are described in **Supplementary Material 4.**

Exposome-Omic Integration Analysis

The function *crossomics* allows performing integration analysis using different type of *rexposome* or Bioconductor objects. Any of them are added into a *MultiDataSet*[20] object (*ExposomeSet*, *ExpressionSet*, *SummarizedExperiment* and *SnpSet*) that is given to the function. It contains two different pipelines for integration analysis.

The first one is based on multiple co-inertia analysis (MCIA) that is implemented in the *omicade4* Bioconductor package (https://bioconductor.org/packages/omicade4). MCIA is an analysis method that identifies co-relationships between multiple data-sets by projecting the features of the multiple data-sets into a single dimensional space, transforming diverse sets of features onto the same scale. By using this approach the most relevant features from each data-set can be obtained in a multivariate fashion (**Supplementary Material 4 – Section 2.3.1**).

The second approach is an extension of Canonical Correlation Analysis (CCA), which has gained popularity as a method for the integration of several omic data (**Supplementary Material 4 – Section 2.3.2**). The method is called Multi-Canonical Correlation Analysis (MultiCCA) that is implemented in the *PMA* R package (https://cran.r-project.org/web/packages/PMA/). MultiCCA extends CCA by providing a sparse version of canonical correlation analysis. The main advantage with regard MCIA is that this method provide a list of features whose loadings are statistically significant different from cero.

Author's Contribution

JRG had the original idea of developing an R package for exposome data analysis. CH-F improved original idea and implemented rexposome and omicRexposome R packages. CH-F developed the web-pages and manages the publication processes of the packages in Biocondcutor. JRG and CH-F conducted the analysis presented in the supplementary material. GAW carefully tested the implementation of ExWAS and provided useful feedback to improve the packages. IT and XB worked on data imputation methods and tested the performance of *Hmic*, *mice* and other R packages designed to address single and multiple imputation. IT provided useful feedback by testing several parts of the *rexposome* package. JS and MV provided data used in supplementary material and helped with data analysis interpretation. JRG and XB provided guidance in statistical methods. CH-F and JRG drafted the manuscript with critical revisions provided by GW, XB and MV. JRG supervised the project. All authors read and approved the final manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Acknowledgments

The author thankfully acknowledges the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Malaga.

Funding

This work has been partly funded by the Spanish Ministry of Economy and Competitiveness (MTM2015-68140-R) by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no308333 – the HELIX project, and by grant R21-ES023073 from the National Institute of Environmental Health Sciences (NIEHS), NIH. The sponsoring agencies had no role in the design of the study, the collection, analysis, and interpretation of data or in writing the manuscript.

Availability

rexposome project is available at Bioinformatic Research Group GitHub's page (http://github.com/isglobal-brge). A specific repository is available to each one of the R packages:

rexposome: https://github.com/isglobal-brge/rexposome

omicRexposome: https://github.com/isglobal-brge/omicRexposome

Current version of both packages are being reviewed in Bioconductor.

REFERENCES

[1] Forouzanfar M. H., et al., Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet (2015).

[2] Wild C. P. Complementing the Genome with an 'Exposome': The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiology Biomarkers & Prevention 14, 1847–1850 (2005).

[3] HELIX: Building the early life exposome. Available at: http://www.projecthelix.eu/. (Accessed: 16th January 2017).

[4] EXPOsOMICS Available at: http://www.exposomicsproject.eu/. (Accessed: 16th January 2017).

[5] HEALS. Available at: http://www.heals-eu.eu/. (Accessed: 16th January 2017).

[6] The Human Exposome Project. Available at: http://humanexposomeproject.com/. (Accessed: 16th January 2017).

[7] Svingen T., Vinggaard A.M., The risk of chemical cocktail effects and how to deal with the issue. J Epidemiol Community Health (2016).

[8] Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R. & Simpson, J. A. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Medical Research Methodology 12, 96 (2012).

[9] Robinson, O. et al. The Pregnancy Exposome: Multiple Environmental Exposures in the INMA-Sabadell Birth Cohort. Environ. Sci. Technol. 49, 10632–10641 (2015).

[10] Lind, P. M., Risérus, U., Salihovic, S., Bavel, B. van & Lind, L. An environmental wide association study (EWAS) approach to the metabolic syndrome. Environment International 55, 1–8 (2013).

[11] Patel, C. J., Chen, R., Kodama, K., Ioannidis, J. P. A. & Butte, A. J. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. Hum. Genet. 132, 495–508 (2013).

[12] Patel, C. J., Bhattacharya, J. & Butte, A. J. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS One 5, (2010).

[13] Agier, L. et al. A Systematic Comparison of Linear Regression–Based Statistical Methods to Assess Exposome-Health Associations. Environmental Health Perspectives 124, (2016).

[14] Liu, C. et al. Characterization of genome-wide H3K27ac profiles reveals a distinct PM2.5-associated histone modification signature. Environmental Health 14, 65 (2015).

[15] Wittkopp, S. et al. Nrf2-related gene expression and exposure to traffic-related air pollution in elderly subjects with cardiovascular disease: An exploratory panel study. J Expo Sci Environ Epidemiol 26, 141–149 (2016).

[16] Joubert, B. R. et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. The American Journal of Human Genetics 98, 680–696 (2016).

[17] Gomez-Cabrero, D. et al. Data integration in the era of omics: current and future challenges. BMC Systems Biology 8, I1 (2014).

[18] Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (2005).

[19] Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., Smyth G.K.. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. (2015).

[20] Hernandez-Ferrer C., Ruiz-Arenas C., Beltran-Gomila A., González JR, MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration. BMC Bioinformatics (2017)

4.3.3 `CTDquerier` R package

Hernandez-Ferrer C, Gonzalez JR. CTDquerier: a bioconductor R package for Comparative Toxicogenomics DatabaseTM data extraction, visualization and enrichment of environmental and toxicological studies. Bioinformatics. 2018 Apr 24; 1-3. DOI: 10.1093/bioinformatics/bty326

4.4 Discussion

MultiDataSet in Bioconductor Ecosystem

MultiDataSet R package was the de first piece of software developed during the first year of this Ph.D. thesis. MultiDataSet meets the objectives that encouraged its development allowing coordinated management of multiple data-sets; extraction of common samples along the multiple data-sets; allowing filtering by data-set, sample, feature, phenotype (covariates and/or health outcomes) and genomic region; and to re-obtain the original data-sets.

Parallel to the development of MultiDataSet, a Bioconductor initiative to provide a tool in multiple data-set structure management was started by Prof. Dr. Levi Waldron (from City University of New York) in mid-2014. While MultiDataSet was make public available through Bioconductor 3.3, MultiAssayExperiment was included in Bioconductor 3.4. From then, both initiatives have been coexisting in Bioconductor ecosystem.

Homologous to MultiDataSet, MultiAssayExperiment introduces a new Bioconductor class with three key components: 1) colData is the primary data-set containing sample data, understood as characteristics such as pathology and histology; 2) ExperimentList is a list of results of data-sets; and 3) sampleMap is a table that relates elements from ExperimentList with those in colData. Consequently, a set of methods are provided to work with MultiAssayExperiment:a constructor function, subsetting and methods (by data-set, genomic ranges, features and samples), extraction methods, and functions for sample selection with complete cases and merging replicates among others.

Ramos et. al. wrote a comparison between MultiDataSet and MultiAssayExperiment in the supplementary material of its publication from *bioRxiv* (June 1st, 2017) [159]. In summary, MultiDataSet provides wrappers for several analysis packages while MultiAssayExperiment provides generic extraction and reshaping functions capable of generating output suitable for these and other packages. MultiAssayExperiment does not provide a slot for feature annotations, however MultiDataSet includes the annotation of each added data-set allowing, for example, to subset non ranged based data-set (like *exposome* data).

The main limitation of `MultiDataSet`, not highlighted by Ramos et. al., must be highlight: memory management is not optimized it `MultiDataSet`. As data-sets grow in size, further developments on the storage need to be devised [153]. This limitation is overcome by `MultiAssayExperiment`, allowing to include on-disk, remote and delayed data representation of very big data-sets through `DelayedMatrix` classes via an `HDF5` back-end.

`rexposome` and `omicRexposome` for Exposome Data Analyses

As well as `MultiDataSet`, `rexposome` provides with a new class, `ExposomeSet`, to incorporate *exposome* data in R/Bioconductor in order to use already existing tools for its analysis. The join efforts for both `rexposome` and `omicRexposome` allowed the development of a robust implementation of a pipeline for *exposome* data analysis (excluding missing data imputation).

The development of *rexposome project* was motivated by the several analysis that were carried out - and that steel being carried out - at the Barcelona Global Health Institute and its projects (INMA [160], MeDALL [161], BREATH [162]...). The development of the `ExposomeSet` provides a data structure for intra-center *exposome* data sharing, also for exposome data publication in data repositories (for instance, the in-development Bioconductor's `ExperimentHub`). At the same time, *rexposome project* standardized the methods to be used in standard *exposome* data analysis pipelines. In a time in which modern exposure assessment tools are providing ever-increasing amounts of detailed exposure data for epidemiologic studies, `rexposome` and `omicRexposome` can facilitate their analysis and the discovery of new health determinants.

The usage and usefulness of the *rexposome projects* were proved on analyzing two data-sets. The first data-set was obtained from the public data repository of the the US the National Health and Nutrition Examination Survey (NHANES) [163], [164] - cross-sectional data. The second data-set was obtained from a prospective pregnancy cohort in Spain, Infancia y Medio Ambiente (INMA), corresponding to a single (INMA-Sabadell) of the several cohorts in the project.

First, an investigation on the associations between *polychlorinated biphenyls* (*PCB*s), *polychlorinated dibenzofurans* (*PDC*s) and *polybrominated diphenyl ethers* (*PBDE*s) and markers of obesity using the NHANES cross-sectional data. The analyses showed *PCB*s, *PDC*s and *organochlorines* (*BDE*s) are highly correlated with obesity. The second illustrative analysis performed the characterization of the *exposome* and its association with molecular signatures. It revealed that exposure to *PCB*s can alter different pathways related to brain and neurodevelopmental regulation processes.

Incorporating CTD into Exposome Data Analysis with `CTDquerier`

In the same direction that `rexposome` and `omicRexposome`, `CTDquerier` was built to meet the lack of tools to provide information about existing chemicals-gene or chemicals-disease associations. The information provided by CTD linking chemicals to genes and diseases can provide significant biological insight when joint with results from *exposome* analysis.

`CTDquerier` helps to improve the results of *exposome* analysis by: 1) providing a list of genes associated with the studied chemicals; and 2) provide a list of diseases associated with the chemicals used. If the *exposome* analysis is based on the study of the relationship between exposures and molecular signatures, it may provide with: 1) verify whether resulting features linked to genes are associated with specific diseases; 2) provide a list of chemicals that change linked genes expression, and 3)literature evidence to disentangle the relationship between genes, exposures and diseases.

A Complete Framework for Exposome Data Analysis

Four software pieces have been presented in this chapter as part of the results of this Ph.D. thesis. These packages have been built to provides a full infrastructure for a complete pipeline to analyze *exposome* data. This pipeline is operational for analysis including or leaking the study of the molecular intermediates for *exposome* molecular signatures.

Figure 4.1 illustrates how the presented packages are four puzzles pieces than can be interconnected to perform a study were the underlying causes of diseases are studied through investigating the link between the *exposome* and its molecular signatures.

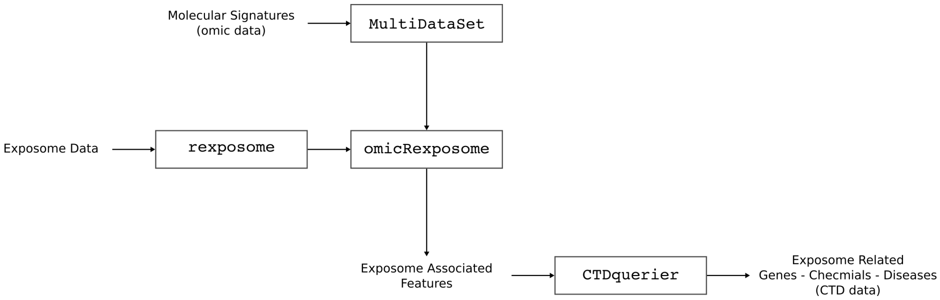


Figure 4.1: A Complete Framework for Exposome Data Analysis.

Therefore, this pipeline perfectly matches with the purposes of the European HELIX (The Human Early-Life Exposome) project [52]. Hence, `rexposome` and `omicRexosome` are being used, joint to `MultiDataSet`, to perform the first hight-thought analysis were $\sim$ 300 exposures are tested for association with $\sim$ 500,000 omic features (including transcriptome, methylome, proteome and metabolome).

The aim of this screening is to end up with a “Molecular Signature Catalogue for Early Life Exposome”, a data base having the degree of association between all the exposures and all the omic features in the HELIX project and a web portal for public access to the association data.

Chapter 5

Conclusion

UV Radiation Effect

Blood

- Exposure to UVR induced transient changes in gene expression in blood cells.
- Affected genes are related to immune regulation, cancer and blood pressure.
- UVR increased vitamin D_3 over time, especially in participants with low baseline levels.
- The acute effects of UVR on the blood transcriptome are independent of vitamin D_3.

Skin

- Effect to UVR induced was dose dependent, with a decreased over time.
- Vast majority of the detected biological functions in skin transcriptome at both 6h and 24h were similar.
- However apoptosis and keratinization were more prominent at 6h while inflammation, immune response and hyaluronan biosynthesis

at 24h.

- Four miRNA where found differentially expressed and they were related to activation of skin wound and to serotonin pathway – that has been suggested to mediate UVB-induced immune suppression.

Software and Tools

- `MultiDataSet` allows to manage data from different omics experiment from the same or different individuals. Its implementation eases standardization of multiple omic sets.
- `rexposome` implements `ExposomeSet` class, the first structure for *exposome* data management under R and Bioconductor framework.
- `rexposome` implements exposome-wide association study, principal component analysis, correlation analysis and clustering analysis. Then, it is the first tool for *exposome* characterization and analysis.
- Combining of `rexposome` and `omicRexposome` allows to look for *exposome* biomarkers.
- `CTDquerier` allows to obtain CTD data under R and Bioconductor framework allowing enrichment analysis from the *exposome* (chemical) side.
- All the five packages can be stand-alone used, providing individual functionality that can be used in multiple downstream analysis. But they can also be combined to provide a full framework for *exposome* data analysis.

Chapter 6

Future Work

While the results shown in chapter 3 are auto-conclusive the results from chapter 4 can be extended to other projects: tools designed for *exposome* data analysis. Then, this chapter focuses on the application of these tools, their possible improvement and the dissemination of the own tools and the results obtained from their application.

6.1 Application

The HELIX (Human Early-Life Exposome) project [52] has as its general aim to implement tools and methods to characterize early-life exposure to a wide range of chemical and physical environmental factors and associate these with data on major child health outcomes, thus developing an "early-life exposome" approach.

Then, the project aims to find new *exposure biomarkers* through the development and testing of new personal exposure devices and the development and application of new statistical tools for *exposome* and molecular signatures analysis.

The project takes pregnancy and childhood periods as the starting point for developing the life-course *exposome*.

Then, both ExposomeSet and imExposomeSet classes, implemented in rexposome, are used as main structures to store, transfer and analysed the *exposome*. At the same time, the implementation of the Exposome-Wide Association Study (ExWAS) [54] is being used by the members of the project to study the *exposome* association with growth and obesity, neurodevelopment and respiratory health among others.

In parallel to the strict *exposome* analysis, HELIX project is testing the association of the exposuers with six sets of molecular signatures with omicRexposome, through MultiDataSet objects to coordinated manage *exposome* data with transcriptome data (from micro RNA and gene expression), methylation data, proteome data and two sets of metabolome data (urine and plasma).

6.2 Improvement

MultiDataSet R/Bioconductor package

While the class implemented in MultiDataSet R/Biocondcutor package is useful and completely operational has a lack in terms of memory management. This limitation has made that most of the Biocondcutor users moved to the MultiAssayExperiment R/Bioconductor package [159], as we observed during "*BioC 2017: Where Software and Biology Connect*" celebrated in Boston.

The team in charge of developing MultiAssayExperiment improved the package once MultiDataSet was published and hired a programmer for full time development and maintenance. This has moved MultiAssayExperiment to a superior position compared to MultiDataSet.

rexposome R/Bioconductor package

Current version of exwas method, implemented in rexposome, allows to test the association between the exposures in a ExposomeSet or imExposomeSet versus an outcome of interest. This method relies in the standard glm from base package so the outcomes of interests must follow one of the standard distributions:

```
family(object, ...)

binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

On of the lines for future work is to improve the exwas method to allow to use a more extensive family of outcomes. Current ideas are:

- Cox models (proportional hazards models) are a class of survival models in statistics. Survival models relate the time that passes before some event occurs to one or more covariates that may be associated with that quantity of time.
- Longitudinal data analysis including repeated measures, mixed models analysis, and multilevel modeling.
- Time series analysis by integration the ts() function from base package and using ARIMA models.

Another improvement for rexposome is about the *exposome* containers. Both ExposomeSet and imExposomeSet are based on the eSet class. Bioconductor is encouraging new development to be based on SumarizedExperiment instead to eSet-like classes. Hence, rexposome needs an update to re-implement both ExposomeSet and imExposomeSet based on SumarizedExperiment.

omicRexposome R package

In order to improve omicRexposome, a data-package called BRGEdata was developed. The package including a series of synthetic data-sets that can be used for *exposome* data analysis, for omic data analysis and for exposome-omic data integration. The data-sets included in `BRGEdata` are listed in Table 6.1.

Table 6.1: Five data-sets were encapsulated in `BRGEdata` to be used for illustrative purposes in `omicRexposome` vignette.

Data Type	# Samples	# Features	Technology	Object Name	Class
Exposome	110	15		brge_expo	ExposomeSet
Transcriptome	75	67528	Affymetrix HTA 2.0	brge_gexp	ExpressionSet
Methylome	115	476946	Illumina Human Methylation 450K	brge_methy	GenomicRatioSet
Proteome	90	47		brge_prot	ExpressionSet

After the direction Biocondcutor is taking, the data-sets needs to be updated to `SumarizedExperiment`-like objects instead of the `eSet`-like objects that are currently being used.

Current version of `omicRexposome` accepts `MultiDataSet` objects as main input. Since `MultiAssayExperiment` objects are becoming popular in Bioconductor ecosystem `omicRexposome` needs to be updated to also accept this type of objects.

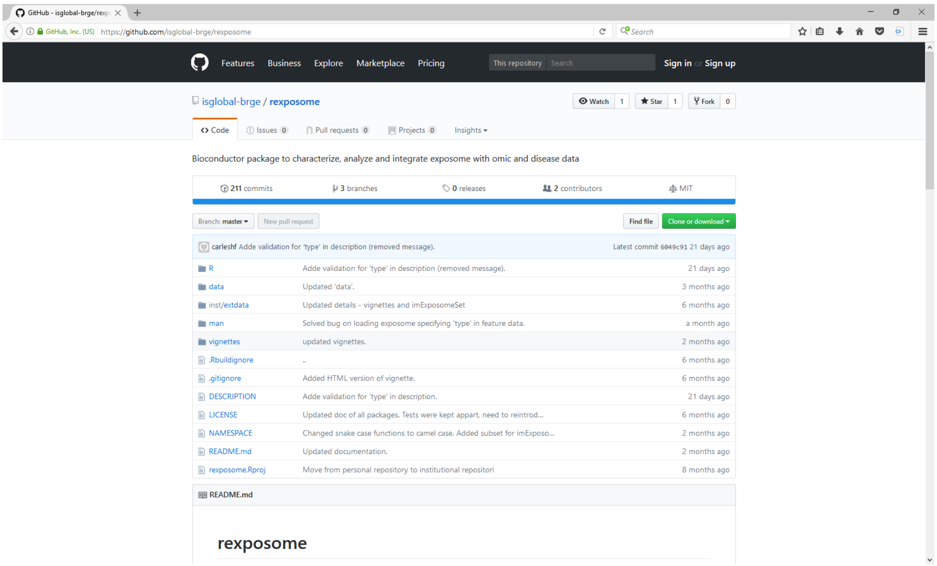
`CTDquerier` R package

`CTDquerier` package was develop to query the *Comparative Toxicogenomics Database* (CTD^{TM}) using genes, chemicals and diseases. During the development of the package, CTD got improved and a new “compartment” was added: exposures. This “compartment” allows performing “exposure studies” allowing to perform queries to CTD selecting a chemical, a gene, a disease and/or a phenotype (described with Gene Ontology terms - biological processes). The “compartment” also allows to select the “receptor”, that describes the category of human subjects that are being acted upon by a chemical exposure stressor; the study factor, that describes any main circumstance influencing the overall outcome of the exposure study (such as age, genetic predispositions, etc.). At least one of the fields must be filled. The result is a table listing all the studies that match the queried terms, including the reference, the study title, the summary, the study factors (age, sex, BMI, etc.), the chemical, the receptors and the outcome among others.

Then, CTDquerier could be extended to allow a user to perform "exposure studies" from R environment and obtain the literature results, with the possibility to obtain the genes, chemicals and/or diseases from the results.

R Package Maintenance

To properly maintain and distribute the four R packages included as results of this Ph.D. thesis, they were developed using a version control system (VCS). The selected VSC was Git and the system used was through the GitHub web service (http://github.com).



Bioconductor package to characterize, analyze and integrate exposome with omic and disease data

carleshf Adde validation for 'type' in description (removed message). Latest commit 6049c91 21 days ago

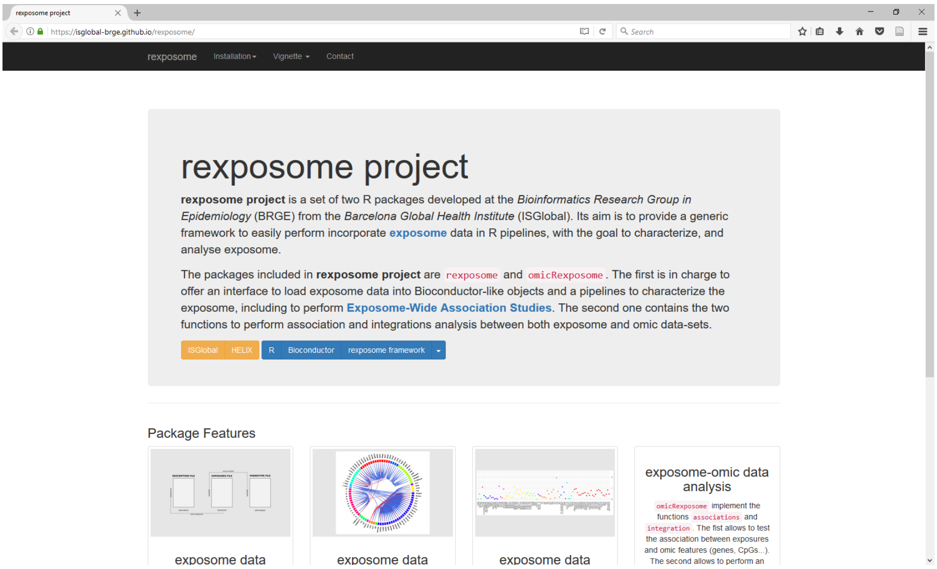
R	Adde validation for 'type' in description (removed message).	21 days ago
data	Updated 'data'.	3 months ago
inst/extdata	Updated details - vignettes and imExposomeSet	6 months ago
man	Solved bug on loading exposome specifying 'type' in feature data.	a month ago
vignettes	updated vignettes.	2 months ago
.Rbuildignore	..	6 months ago
.gitignore	Added HTML version of vignette.	6 months ago
DESCRIPTION	Adde validation for 'type' in description.	21 days ago
LICENSE	Updated doc of all packages. Tests were kept appart, need to reintrod...	6 months ago
NAMESPACE	Changed snake case functions to camel case. Added subset for imExposo...	2 months ago
README.md	Updated documentation.	2 months ago
rexposome.Rproj	Move from personal repository to institutional repositori	8 months ago

Figure 6.1: All the four presented packages (MultiDataSet, rexposome, omicRexposome and CTDquerier) have a repository in the BRGE's GitHub account for version control and easy installation process.

Figure 6.1 shows a screen-shoot of the rexposome 's GitHub repository. Each package uses their repository for an up-to-date distribution using the R package devtools and its method install_github. The idea is to start using the *issue tab* for issue tracking and new features development.

6.3 Dissemination

Communicating science to the public is a must in any scientific career. So on, Figure 6.2 shows the web-page created for *rexposome project* using the *GitHub pages* framework (`https://isglobal-brge.github.io/rexposome/`). The goal of this web-page is to be a place to centralize the source-code of both `rexposome` and `omicRexposome` packages as well as installation guides and their vignettes.



rexposome project

rexposome project is a set of two R packages developed at the *Bioinformatics Research Group in Epidemiology* (BRGE) from the *Barcelona Global Health Institute* (ISGlobal). Its aim is to provide a generic framework to easily perform incorporate **exposome** data in R pipelines, with the goal to characterize, and analyse exposome.

The packages included in **rexposome project** are `rexposome` and `omicRexposome`. The first is in charge to offer an interface to load exposome data into Bioconductor-like objects and a pipelines to characterize the exposome, including to perform **Exposome-Wide Association Studies**. The second one contains the two functions to perform association and integrations analysis between both exposome and omic data-sets.

Package Features

exposome data

exposome data

exposome data

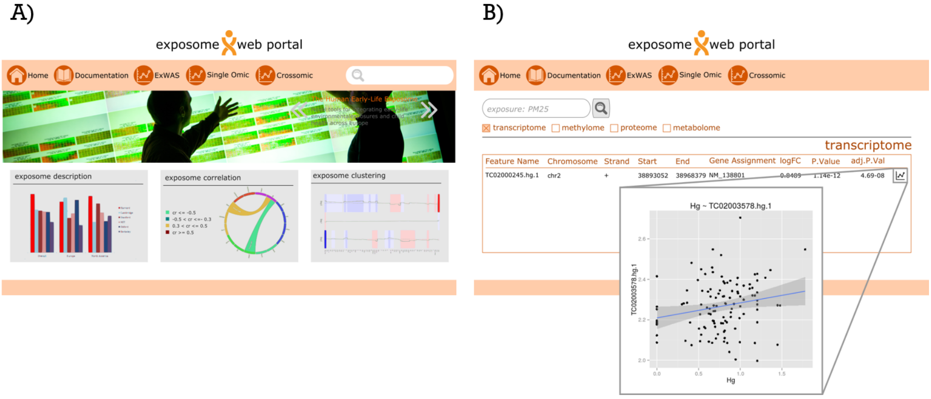
exposome-omic data analysis

`omicRexposome` implement the functions `associations` and `integration`. The fist allows to test the association between exposures and omic features (genes, CpGs...). The second allows to perform an

Figure 6.2: *rexposome project* web page - `https://isglobal-brge.github.io/rexposome/`

While the `rexposome project`'s web page is for tool dissemination, this Ph.D. thesis has the future goal to develop a web portal for disseminating the results obtained from the screening performed in the HELIX project. The screen, more described before (Application section), includes the association test between more than 200 exposures (distributed into prenatal and post-natal time points) and six molecular signature sets.

In order to make the results of this screening public available a web portal (The "Exposome Web Portal") will be developed. Figure 6.3 shows a mockup of the possible design of the portal.



The Human Early-Life Exposome

Novel tools for integrating early-life environmentalexposures and child health across Europe

Feature Name	Chromosome	Strand	Start	End	Gene Assignment	logFC	P.Value	adj.P.Val
TC02000245.hg.1	chr2	+	38893052	38968379	NM_138801	0.8409	1.14e-12	4.69-08

Figure 6.3: Tentative mockup of the "Exposome Web Portal" that will be built for HELIX project.

Figure 6.3 - A shows the main page of the portal. This section will show a basic summary of the analyzed data (number of exposures, number of features per molecular signature, etc.). From this screen, the user may be able to go to a series of pages describing the data (correlation between exposures, description of each molecular signatures, etc.) and to the query page. The query page, seen in Figure 6.3 - B, will allow the user to query giving a criteria. The criteria may involve any exposure, any feature from any molecular signature set, an annotation of a feature and/or a threshold for the association. The result of the query given user's criteria will be displayed in a table (that would be able to be exported as CSV file) and will offer toms visualization per results like QQ-plot and Volcano-plot of individual analysis.

Chapter 7

Scientific Work Related to this Thesis

7.1 Peer-reviewed Publications Not Presented as Results

The Pregnancy Exposome: Multiple Environmental Exposures in the INMA-Sabadell Birth Cohort

Journal	Environmental Science & Technology
Authors	Oliver Robinson, Xavier Basagaña, Lydiane Agier, Montserrat de Castro, Carles Hernandez-Ferrer, Juan R. Gonzalez, Joan O. Grimalt, Mark Nieuwenhuijsen, Jordi Sunyer, Rémy Slama, and Martine Vrijheid
Status	Published (July 13, 2015)

Abstract

The “exposome” is defined as “the totality of human environmental exposures from conception onward, complementing the genome” and its holistic approach may advance understanding of disease etiology. We aimed to describe the correlation structure of the exposome during pregnancy to better understand the relationships between and within families of exposure and to develop analytical tools appropriate to exposome data. Estimates on 81 environmental exposures of current health concern were obtained for 728 women enrolled in The INMA (INfancia y Medio Ambiente) birth cohort, in Sabadell, Spain, using biomonitoring, geospatial modeling, remote sensors, and questionnaires. Pair-wise Pearson's and polychoric correlations were calculated and principal components were derived. The median absolute correlation across all exposures was 0.06 (5th–95th centiles, 0.01–0.54). There were strong levels of correlation within families of exposure (median = 0.45, 5th-95th centiles, 0.07-0.85). Nine exposures (11%) had a correlation higher than 0.5 with at least one exposure outside their exposure family. Effectively all the variance in the data set (99.5%) was explained by 40 principal components. Future exposome studies should interpret exposure effects in light of their correlations to other exposures. The weak to moderate correlation observed between exposure families will permit adjustment for confounding in future exposome studies.

psygenet2r: a R/Bioconductor package for the analysis of psychiatric disease genes

Journal	Bioinformatics
Authors	Alba Gutiérrez-Sacristán*, Carles Hernández-Ferrer*, Juan R. González, Laura I. Furlong
Status	Published (August 17, 2017)

Abstract

Motivation: Psychiatric disorders have a great impact on morbidity and mortality. Genotype-phenotype resources for psychiatric diseases are key to enable the translation of research findings to a better care of patients. PsyGeNET is a knowledge resource on psychiatric diseases and their genes, developed by text mining and curated by domain experts.

Results: We present psygenet2r, an R package that contains a variety of functions for leveraging PsyGeNET database and facilitating its analysis and interpretation. The package offers different types of queries to the database along with variety of analysis and visualization tools, including the study of the anatomical structures in which the genes are expressed and gaining insight of gene's molecular function. +sygenet2r is especially suited for network medicine analysis of psychiatric disorders.

Availability: The package is implemented in R and is available under MIT license from Bioconductor.

A systemic approach to identify signalling pathways activated during short term exposure to traffic related-urban air pollution from human blood

Journal	Environmental Health Perspectives
Authors	José Eduardo Vargas, Nadine Kubesch, Carles Hernandez-Ferrer, Glória Carrasco-Turigas, Mariona Bustamante, Mark Nieuwenhuijsen, Juan R González
Status	Submitted (January 30, 2017)

Abstract

Background: The molecular mechanisms that promote pathologic alterations in human physiology mediated by short-term exposure to traffic pollutants remains not well understood.

Objetive: In this work was to develop mechanistic networks to determine which specific pathways are activated by real world exposures of traffic air pollution (TRAP) during rest and moderate physical activity (PA).

Methods: A controlled crossover study to compare whole blood gene expression pre and post short-term exposure to high and low of TRAP was performed together with systems biology analysis. Twenty eight healthy volunteers aged between 21-53 years were recruited. These subjects were exposed during 2 hours to different pollution levels (high and low TRAP levels) while either cycling or resting. Global transcriptome profile of each condition was performed from human whole blood samples. Microarrays analysis was performed to obtain differential expressed genes (DEG) to be used as initial input for GeneMania software to obtain protein-protein (PPI) networks.

Results: Two networks were found reflecting high or low TRAP levels, which shared only 5.6% and 15.5% of its nodes, suggesting specific cell signalling pathways being activated in each environmental condition. However, gene ontology analysis of each PPI network suggests that each level of TRAP regulate common members of NFK-B signalling pathway.

Conclusion: Our work provides the first approach describing mechanistic networks to understand TRAP effects on a system level.

Circulating miRNAs, isomiRs and small RNA clusters in human plasma and breast milk

Journal	PLoS ONE
Authors	Mercedes Rubio, Mariona Bustamante, Carles Hernandez-Ferrer, Dietmar Fernandez-Orth, Lorena Pantano, Yaris Sarria, Maria Piqué-Borras, Killian Vellve, Silvia Agramunt, Ramon Carreras, Xavier Estivill, Juan R Gonzalez, Alfredo Mayor
Status	Submitted (August 2, 2017)

Abstract

Circulating small RNAs, including miRNAs but also isomiRs and other small RNA species, have the potential to be used as non-invasive biomarkers for communicable and non-communicable diseases. In this study, 1) feasibility and pitfalls of analysing circulating small RNA in plasma and breast milk using next generation sequencing have been evaluated, and 2) small RNAs patterns in these biofluids have been compared. RNA from plasma and breast milk samples from 15 healthy postpartum mothers, was

extracted. Small RNA libraries were prepared with the NEBNext small RNA library preparation kit and sequenced in an Illumina HiSeq2000 platform. After an initial quality control, miRNAs, isomiRs and clusters of small RNAs were annotated using seqBuster/seqCluster framework. The average amount of extracted RNA was 81 ng/mL [standard deviation, (SD), 41] and 3985 ng/mL (SD 3767) for plasma and breast milk, respectively. In plasma, the mean number of good quality reads were 4.04 million (M) (40.01% of the reads), while 12.5M (89.6%) in breast milk. 1,182 miRNAs, 74,317 isomiRs and 1,053 small RNA clusters that included piwi-interfering RNAs (piRNAs), tRNAs, small nucleolar RNAs (snoRNA) and small nuclear RNAs (snRNAs) were detected. Samples grouped by biofluid, with 308 miRNAs, 4,737 isomiRs and 778 small RNA clusters differentially detected. Plasma and milk showed a completely different small RNA profile. In both, miRNAs, piRNAs, tRNAs, snRNAs, and snoRNAs were identified, in line with previous studies showing the presence of non-miRNA species in biofluids.

7.2 Conference Presentations

Molecular signature time variability in HELIX panel paired-samples
Carles Hernandez-Ferrer, Carlos Ruiz-Arenas, Martine Vrijheid, Juan R. González.

- Event: New Horizons for Early Life Exposome Research - Final HELIX Symposium (Oct 2017)
- Location: Barcelona Biomedical Research Park, Barcelona, Spain
- Type: Oral Communication

Linkage between methylation probes and expression transcripts
Carlos Ruiz-Arenas, Carles Hernandez-Ferrer, Martine Vrijheid, Juan R. González.

- Event: New Horizons for Early Life Exposome Research - Final HELIX Symposium (Oct 2017)
- Location: Barcelona Biomedical Research Park, Barcelona, Spain

- Type: Poster

`rexposome`: A bioinformatic tool for characterizing multiple environmental factors and its association with different omics biomarkers and diseases Carles Hernandez-Ferrer, Martine Vrijheid, Juan R. González.

- Event: Workshop in Environmental Omics, Integration and Modelling (Oct 2017)
- Location: CosmoCaixa, Barcelona, Spain
- Type: Oral Communication

Extending Bioconductor To Exposome Data Analysis Carles Hernandez-Ferrer, Juan R. González

- Event: BioC 2017: Where Software and Biology Connect (Jul 2017)
- Location: Dana Farber Cancer Institute, Boston (MA), United States of America
- Type: Oral Communication

`rexposome`: A Bioconductor package for characterizing multiple environmental factors and its association with different omics biomarkers and diseases
Carles Hernandez-Ferrer, Martine Vrijheid, Juan R. González

- Event: European Bioconductor Developers' Meeting (Dec 2016)
- Location: University of Basel (ZLF Building), Basel, Switzerland
- Type: Oral Communication

`rexposome`: a bioinformatic tool for charactering multiple environmental factors and its association with different omic biomarkers and diseases
Carles Hernandez-Ferrer, Martine Vrijheid, Juan R. González

- Event: The Emory Exposome Course (Jun 2016)
- Location: Emory University, Atlanta, United States of America
- Type: Oral Communication, Hands-on-lab

rexposome: a bioinformatic tool for charactering multiple environmental factors and its association with different omic biomarkers and diseases
Carles Hernandez-Ferrer, Martine Vrijheid, Juan R. González

- Event: XIII Symposium on Bioinformatics (May 2016)
- Location: Universidad de Valencia, Valencia, Spain
- Type: Poster

Analyzing SNPs, CNVs, inversions and mosaicisms association studies using Affymetrix CytoScan
Carles Hernandez-Ferrer, Ines Quintela Garcia, Katharina Danielski, Angel Carracedo Alvarez, Luis A. Perez-Jurado, Juan R. Gonzalez

- Event: Molecular Targets for Predictive and Personalized Medicine of Cancer (Apr 2015)
- Location: Hospital St. Pau i de la Santa Creu, Barcelona, Spain
- Type: Poster

Analyzing SNPs, CNVs, inversions and mosaicisms association studies using Affymetrix CytoScan
Carles Hernandez-Ferrer, Ines Quintela Garcia, Katharina Danielski, Angel Carracedo Alvarez, Luis A. Perez-Jurado, Juan R. Gonzalez

- Event: XII Symposium on Bioinformatics (Sep 2014)
- Location: Centro de Investigaciones Cientficas Isla de la Cartuja, Sevilla, Spain
- Type: Poster

Bibliography

[1] T. J. C. Polderman, B. Benyamin, C. A. de Leeuw, P. F. Sullivan, A. van Bochoven, P. M. Visscher, and D. Posthuma, "Meta-analysis of the heritability of human traits based on fifty years of twin studies," *Nature Genetics*, vol. 47, no. 7, pp. 702–709, Jul. 2015, ISSN: 1546-1718. DOI: 10.1038/ng.3285.

[2] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 7164, pp. 851–861, Oct. 18, 2007, ISSN: 0028-0836. DOI: 10.1038/nature06258. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689609/ (visited on 08/11/2017).

[3] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *The New England journal of medicine*, vol. 360, no. 17, pp. 1759–1768, Apr. 23, 2009, ISSN: 0028-4793. DOI: 10.1056/NEJMra0808700. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3422859/ (visited on 08/11/2017).

[4] D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *Trends in Genetics*, vol. 17, no. 9, pp. 502–510, Sep. 1, 2001, ISSN: 0168-9525. DOI: 10.1016/S0168-9525(01)02410-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168952501024106 (visited on 08/11/2017).

[5] J. K. Pritchard, "Are rare variants responsible for susceptibility to complex diseases?" *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 124–137, Jul. 1, 2001, ISSN: 0002-9297, 1537-6605. DOI: 10.1086/321272. [Online]. Available: http://www.

cell.com/ajhg/abstract/S0002-9297(07)61452-9 (visited on 07/09/2017).

[6] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 23, pp. 9362–9367, Jun. 9, 2009, ISSN: 1091-6490. DOI: 10.1073/pnas.0903103106.

[7] T. A. Manolio, F. S. Collins, N. J. Cox, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 8, 2009, ISSN: 0028-0836. DOI: 10.1038/nature08494. [Online]. Available: http://www.nature.com/nature/journal/v461/n7265/full/nature08494.html (visited on 08/11/2017).

[8] J. Maller, S. George, S. Purcell, J. Fagerness, D. Altshuler, M. J. Daly, and J. M. Seddon, "Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration," *Nature Genetics*, vol. 38, no. 9, pp. 1055–1059, Sep. 2006, ISSN: 1061-4036. DOI: 10.1038/ng1873.

[9] J. C. Barrett, S. Hansoul, D. L. Nicolae, *et al.*, "Genome-wide association defines more than thirty distinct susceptibility loci for crohn's disease," *Nature genetics*, vol. 40, no. 8, pp. 955–962, Aug. 2008, ISSN: 1061-4036. DOI: 10.1038/NG.175. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2574810/ (visited on 08/13/2017).

[10] J. B. Harley, M. E. Alarcón-Riquelme, L. A. Criswell, *et al.*, "Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci," *Nature genetics*, vol. 40, no. 2, pp. 204–210, Feb. 2008, ISSN: 1061-4036. DOI: 10.1038/ng.81. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3712260/ (visited on 08/13/2017).

[11] E. Zeggini, L. J. Scott, R. Saxena, *et al.*, "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes," *Nature genetics*, vol. 40, no. 5, pp. 638–645, May 2008, ISSN: 1061-4036. DOI: 10.1038/ng.

120. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2672416/ (visited on 08/13/2017).

[12] S. Kathiresan, O. Melander, C. Guiducci, *et al.*, "Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans," *Nature genetics*, vol. 40, no. 2, pp. 189–197, Feb. 2008, ISSN: 1061-4036. DOI: 10.1038/ng.75. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2682493/ (visited on 08/13/2017).

[13] P. M. Visscher, "Sizing up human height variation," *Nature Genetics*, vol. 40, no. 5, pp. 489–490, May 2008, ISSN: 1061-4036. DOI: 10.1038/ng0508-489. [Online]. Available: http://www.nature.com/ng/journal/v40/n5/full/ng0508-489.html.

[14] Myocardial Infarction Genetics Consortium, "Genome-wide association of early-onset myocardial infarction with common single nucleotide polymorphisms, common copy number variants, and rare copy number variants," *Nature genetics*, vol. 41, no. 3, pp. 334–341, Mar. 2009, ISSN: 1061-4036. DOI: 10.1038/ng.327. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2681011/ (visited on 08/13/2017).

[15] I. Prokopenko, C. Langenberg, J. C. Florez, *et al.*, "Variants in MTNR1b influence fasting glucose levels," *Nature genetics*, vol. 41, no. 1, pp. 77–81, Jan. 2009, ISSN: 1061-4036. DOI: 10.1038/ng.290. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2682768/ (visited on 08/13/2017).

[16] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of GWAS discovery: Biology, function, and translation," *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, Jul. 6, 2017, ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2017.06.005. [Online]. Available: http://www.cell.com/ajhg/abstract/S0002-9297(17)30240-9 (visited on 08/27/2017).

[17] B. Maher, "Personal genomes: The case of the missing heritability," *Nature News*, vol. 456, no. 7218, pp. 18–21, Nov. 5, 2008, ISSN: 0028-0836. DOI: 10.1038/456018a. [Online]. Available: http://www.nature.com/news/2008/081105/full/456018a.html.

[18] J. K. Pritchard and N. J. Cox, “The allelic architecture of human disease genes: Common disease-common variant...or not?” *Human Molecular Genetics*, vol. 11, no. 20, pp. 2417–2423, Oct. 1, 2002, ISSN: 0964-6906.

[19] M. C. Campbell and S. A. Tishkoff, “AFRICAN GENETIC DIVERSITY: Implications for human demographic history, modern human origins, and complex disease mapping,” *Annual review of genomics and human genetics*, vol. 9, pp. 403–433, 2008, ISSN: 1527-8204. DOI: 10.1146/annurev.genom.9.081307.164258. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2953791/ (visited on 08/13/2017).

[20] S. M. Rappaport, “Genetic factors are not the major causes of chronic diseases,” *PLoS ONE*, vol. 11, no. 4, Apr. 22, 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0154387. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4841510/ (visited on 07/09/2017).

[21] K. Hemminki, J. Lorenzo Bermejo, and A. Försti, “The balance between heritable and environmental aetiology of human disease,” *Nature Reviews. Genetics*, vol. 7, no. 12, pp. 958–965, Dec. 2006, ISSN: 1471-0056. DOI: 10.1038/nrg2009.

[22] S. M. Rappaport, D. K. Barupal, D. Wishart, P. Vineis, and A. Scalbert, “The blood exposome and its role in discovering causes of disease,” *Environmental Health Perspectives*, vol. 122, no. 8, pp. 769–774, Aug. 2014, ISSN: 1552-9924. DOI: 10.1289/ehp.1308015.

[23] S. Wu, S. Powers, W. Zhu, and Y. A. Hannun, “Substantial contribution of extrinsic risk factors to cancer development,” *Nature*, vol. 529, no. 7584, pp. 43–47, Jan. 7, 2016, ISSN: 0028-0836. DOI: 10.1038/nature16166. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836858/ (visited on 08/13/2017).

[24] S. S. Lim, T. Vos, A. D. Flaxman, *et al.*, “A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010,” *Lancet*, vol. 380, no. 9859, pp. 2224–2260, Dec. 15, 2012, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(12)61766-8. [Online]. Available:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4156511/ (visited on 08/13/2017).

[25] GBD 2013 Risk Factors Collaborators, "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: A systematic analysis for the global burden of disease study 2013," *Lancet (London, England)*, vol. 386, no. 10010, pp. 2287–2323, Dec. 5, 2015, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(15)00128-2. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4685753/ (visited on 08/14/2017).

[26] GBD 2015 Tobacco Collaborators, "Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: A systematic analysis from the global burden of disease study 2015," *The Lancet*, vol. 389, no. 10082, pp. 1885–1906, May 13, 2017, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(17)30819-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S014067361730819X (visited on 09/02/2017).

[27] A. J. Cohen, M. Brauer, R. Burnett, *et al.*, "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the global burden of diseases study 2015," *Lancet (London, England)*, vol. 389, no. 10082, pp. 1907–1918, May 13, 2017, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(17)30505-6. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5439030/ (visited on 08/14/2017).

[28] WHO, "Air quality guidelines - global update 2005," 2005. [Online]. Available: http://www.who.int/phe/health_topics/outdoorair/outdoorair_aqg/en/ (visited on 08/14/2017).

[29] U.S. Environmental Protection Agency., "Integrated science assessment (ISA) for particulate matter (final report, dec 2009).," U.S. Environmental Protection Agency, Washington, 2009. [Online]. Available: https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=216546 (visited on 08/14/2017).

[30] H. Chen, M. S. Goldberg, and P. J. Villeneuve, "A systematic review of the relation between long-term exposure to ambient air

pollution and chronic diseases," *Reviews on Environmental Health*, vol. 23, no. 4, pp. 243–297, Dec. 2008, ISSN: 0048-7554.

[31] G. Hoek, R. M. Krishnan, R. Beelen, A. Peters, B. Ostro, B. Brunekreef, and J. D. Kaufman, "Long-term air pollution exposure and cardio- respiratory mortality: A review," *Environmental Health*, vol. 12, p. 43, May 28, 2013, ISSN: 1476-069X. DOI: 10.1186/1476-069X-12-43. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679821/ (visited on 08/14/2017).

[32] M. Jerrett, R. T. Burnett, C. A. I. Pope, K. Ito, G. Thurston, D. Krewski, Y. Shi, E. Calle, and M. Thun, "Long-term ozone exposure and mortality," *New England Journal of Medicine*, vol. 360, no. 11, pp. 1085–1095, Mar. 12, 2009, ISSN: 0028-4793. DOI: 10.1056/NEJMoa0803894. [Online]. Available: http://dx.doi.org/10.1056/NEJMoa0803894 (visited on 08/14/2017).

[33] U.S. Environmental Protection Agency., "Integrated science assessment (ISA) of ozone and related photochemical oxidants (final report, feb 2013)," U.S. Environmental Protection Agency, Washington, 2013. [Online]. Available: https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=247492 (visited on 08/14/2017).

[34] H. H. Shin, A. J. Cohen, C. A. Pope, M. Ezzati, S. S. Lim, B. J. Hubbell, and R. T. Burnett, "Meta-analysis methods to estimate the shape and uncertainty in the association between long-term exposure to ambient fine particulate matter and cause-specific mortality over the global concentration range," *Risk Analysis: An Official Publication of the Society for Risk Analysis*, Jun. 4, 2015, ISSN: 1539-6924. DOI: 10.1111/risa.12421.

[35] C. P. Wild, "Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology," *Cancer Epidemiology Biomarkers & Prevention*, vol. 14, no. 8, pp. 1847–1850, Aug. 1, 2005, ISSN: 1055-9965, 1538-7755. DOI: 10.1158/1055-9965.EPI-05-0456. [Online]. Available: http://cebp.aacrjournals.org/cgi/doi/10.1158/1055-9965.EPI-05-0456 (visited on 01/16/2017).

[36] C. P. Wild, "The exposome: From concept to utility," *International Journal of Epidemiology*, vol. 41, no. 1, pp. 24–32, Feb. 2012, ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyr236. [Online]. Avail-

able: http://www.ije.oxfordjournals.org/lookup/doi/10.1093/ije/dyr236 (visited on 01/16/2017).

[37] M. Vrijheid, "The exposome: A new paradigm to study the impact of environment on health," *Thorax*, vol. 69, no. 9, pp. 876–878, Sep. 1, 2014, ISSN: 0040-6376, 1468-3296. DOI: 10.1136/thoraxjnl-2013-204949. [Online]. Available: http://thorax.bmj.com/content/69/9/876 (visited on 08/14/2017).

[38] S. M. Rappaport and M. T. Smith, "Environment and disease risks," *Science (New York, N.Y.)*, vol. 330, no. 6003, pp. 460–461, Oct. 22, 2010, ISSN: 0036-8075. DOI: 10.1126/science.1192603. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4841276/ (visited on 09/02/2017).

[39] J. K. Nicholson, E. Holmes, J. M. Kinross, A. W. Darzi, Z. Takats, and J. C. Lindon, "Metabolic phenotyping in clinical and surgical environments," *Nature*, vol. 491, no. 7424, pp. 384–392, Nov. 15, 2012, ISSN: 0028-0836. DOI: 10.1038/nature11708. [Online]. Available: http://www.nature.com/nature/journal/v491/n7424/full/nature11708.html?foxtrotcallback=true (visited on 09/02/2017).

[40] G. Miller, *The Exposome.* Elsevier, 2013, 118 pp., ISBN: 978-0-12-417217-3.

[41] G. W. Miller and D. P. Jones, "The nature of nurture: Refining the definition of the exposome," *Toxicological Sciences: An Official Journal of the Society of Toxicology*, vol. 137, no. 1, pp. 1–2, Jan. 2014, ISSN: 1096-0929. DOI: 10.1093/toxsci/kft251.

[42] M. Gascon, E. Morales, J. Sunyer, and M. Vrijheid, "Effects of persistent organic pollutants on the developing respiratory and immune systems: A systematic review," *Environment International*, vol. 52, pp. 51–65, Feb. 1, 2013, ISSN: 0160-4120. DOI: 10.1016/j.envint.2012.11.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0160412012002425 (visited on 09/03/2017).

[43] D. T. Wigle, T. E. Arbuckle, M. C. Turner, A. Bérubé, Q. Yang, S. Liu, and D. Krewski, "Epidemiologic evidence of relationships between reproductive and child health outcomes and environmental chemical contaminants," *Journal of Toxicology and Environmental*

Health. Part B, Critical Reviews, vol. 11, no. 5, pp. 373–517, May 2008, ISSN: 1521-6950. DOI: 10.1080/10937400801921320.

[44] O. Robinson, X. Basagaña, L. Agier, *et al.*, "The pregnancy exposome: Multiple environmental exposures in the INMA-sabadell birth cohort," *Environmental Science & Technology*, vol. 49, no. 17, pp. 10 632–10 641, Sep. 1, 2015, ISSN: 0013-936X. DOI: 10.1021/acs.est.5b01782. [Online]. Available: http://dx.doi.org/10.1021/acs.est.5b01782 (visited on 01/18/2017).

[45] P. D. Gluckman, M. A. Hanson, and A. S. Beedle, "Early life events and their consequences for later disease: A life history and evolutionary perspective," *American Journal of Human Biology: The Official Journal of the Human Biology Council*, vol. 19, no. 1, pp. 1–19, Feb. 2007, ISSN: 1042-0533. DOI: 10.1002/ajhb.20590.

[46] A. Lucas, "Programming by early nutrition in man," *Ciba Foundation Symposium*, vol. 156, 38–50, discussion 50–55, 1991, ISSN: 0300-5208.

[47] D. J. Barker and C. Osmond, "Infant mortality, childhood nutrition, and ischaemic heart disease in england and wales," *Lancet (London, England)*, vol. 1, no. 8489, pp. 1077–1081, May 10, 1986, ISSN: 0140-6736.

[48] C. N. Hales, D. J. Barker, P. M. Clark, L. J. Cox, C. Fall, C. Osmond, and P. D. Winter, "Fetal and infant growth and impaired glucose tolerance at age 64.," *BMJ : British Medical Journal*, vol. 303, no. 6809, pp. 1019–1022, Oct. 26, 1991, ISSN: 0959-8138. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1671766/ (visited on 09/03/2017).

[49] H. E. Syddall, A. A. Sayer, S. J. Simmonds, C. Osmond, V. Cox, E. M. Dennison, D. J. P. Barker, and C. Cooper, "Birth weight, infant weight gain, and cause-specific mortality: The hertfordshire cohort study," *American Journal of Epidemiology*, vol. 161, no. 11, pp. 1074–1080, Jun. 1, 2005, ISSN: 0002-9262. DOI: 10.1093/aje/kwi137.

[50] P. L. Hofman, F. Regan, W. E. Jackson, C. Jefferies, D. B. Knight, E. M. Robinson, and W. S. Cutfield, "Premature birth and later insulin resistance," *New England Journal of Medicine*, vol. 351, no. 21, pp. 2179–2186, Nov. 18, 2004, ISSN: 0028-4793. DOI: 10.

1056/NEJMoa042275. [Online]. Available: http://dx.doi.org/10.1056/NEJMoa042275 (visited on 09/03/2017).

[51] R. C. Painter, T. J. Roseboom, and O. P. Bleker, "Prenatal exposure to the dutch famine and disease in later life: An overview," *Reproductive Toxicology (Elmsford, N.Y.)*, vol. 20, no. 3, pp. 345–352, Oct. 2005, ISSN: 0890-6238. DOI: 10.1016/j.reprotox.2005.04.005.

[52] M. Vrijheid, R. Slama, O. Robinson, *et al.*, "The human early-life exposome (HELIX): Project rationale and design," *Environmental Health Perspectives*, vol. 122, no. 6, pp. 535–544, Jun. 2014, ISSN: 0091-6765. DOI: 10.1289/ehp.1307204. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4048258/ (visited on 08/10/2017).

[53] A. K. Manrai, Y. Cui, P. R. Bushel, *et al.*, "Informatics and data analytics to support exposome-based discovery for public health," Apr. 2017. [Online]. Available: http://www.annualreviews.org.sare.upf.edu/doi/10.1146/annurev-publhealth-082516-012737 (visited on 08/01/2017).

[54] C. J. Patel, J. Bhattacharya, and A. J. Butte, "An environment-wide association study (EWAS) on type 2 diabetes mellitus," *PLoS ONE*, vol. 5, no. 5, May 20, 2010, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0010746. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2873978/ (visited on 01/18/2017).

[55] R. Doll and R. Peto, "The causes of cancer: Quantitative estimates of avoidable risks of cancer in the united states today," *Journal of the National Cancer Institute*, vol. 66, no. 6, pp. 1191–1308, Jun. 1981, ISSN: 0027-8874.

[56] P. L. Remington, R. C. Brownson, and Centers for Disease Control and Prevention (CDC), "Fifty years of progress in chronic disease epidemiology and control," *MMWR supplements*, vol. 60, no. 4, pp. 70–77, Oct. 7, 2011, ISSN: 2380-8942.

[57] K. K. Dennis, S. S. Auerbach, D. M. Balshaw, Y. Cui, M. D. Fallin, M. T. Smith, A. Spira, S. Sumner, and G. W. Miller, "The importance of the biological impact of exposure to the concept of the exposome," *Environmental Health Perspectives*, vol. 124, no. 10, pp. 1504–1510, Oct. 2016, ISSN: 0091-6765. DOI: 10.1289/EHP140.

[Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5047763/ (visited on 09/03/2017).

[58] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki, "Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from sweden, denmark, and finland," *The New England Journal of Medicine*, vol. 343, no. 2, pp. 78–85, Jul. 13, 2000, ISSN: 0028-4793. DOI: 10.1056/NEJM200007133430201.

[59] W. C. Willett, "Balancing life-style and genomics research for disease prevention," *Science (New York, N.Y.)*, vol. 296, no. 5568, pp. 695–698, Apr. 26, 2002, ISSN: 1095-9203. DOI: 10.1126/science.1071055.

[60] S. E. Reese, S. Zhao, M. C. Wu, *et al.*, "DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy," *Environmental Health Perspectives*, vol. 125, no. 4, pp. 760–766, Apr. 2017, ISSN: 1552-9924. DOI: 10.1289/EHP333.

[61] A. P. Wolffe and M. A. Matzke, "Epigenetics: Regulation through repression," *Science (New York, N.Y.)*, vol. 286, no. 5439, pp. 481–486, Oct. 15, 1999, ISSN: 0036-8075.

[62] M. J. Ziller, H. Gu, F. Müller, *et al.*, "Charting a dynamic DNA methylation landscape of the human genome," *Nature*, vol. 500, no. 7463, pp. 477–481, Aug. 22, 2013, ISSN: 0028-0836. DOI: 10.1038/nature12433. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3821869/ (visited on 09/05/2017).

[63] ENCODE Project Consortium, "The ENCODE (ENCyclopedia of DNA elements) project," *Science (New York, N.Y.)*, vol. 306, no. 5696, pp. 636–640, Oct. 22, 2004, ISSN: 1095-9203. DOI: 10.1126/science.1105136.

[64] A. Yen and M. Kellis, "Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type," *Nature Communications*, vol. 6, Aug. 18, 2015, ISSN: 2041-1723. DOI: 10.1038/ncomms8973. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4557131/ (visited on 09/05/2017).

[65] W. Reik, W. Dean, and J. Walter, "Epigenetic reprogramming in mammalian development," *Science*, vol. 293, no. 5532, pp. 1089–

1093, Aug. 10, 2001, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1063443. [Online]. Available: http://science.sciencemag.org/content/293/5532/1089 (visited on 08/05/2017).

[66] S. I. S. Grewal and D. Moazed, "Heterochromatin and epigenetic control of gene expression," *Science (New York, N.Y.)*, vol. 301, no. 5634, pp. 798–802, Aug. 8, 2003, ISSN: 1095-9203. DOI: 10.1126/science.1086887.

[67] G. Orphanides and D. Reinberg, "A unified theory of gene expression," *Cell*, vol. 108, no. 4, pp. 439–451, Feb. 22, 2002, ISSN: 0092-8674.

[68] A. S. Yang, M. R. H. Estécio, K. Doshi, Y. Kondo, E. H. Tajara, and J.-P. J. Issa, "A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements," *Nucleic Acids Research*, vol. 32, no. 3, e38, 2004, ISSN: 0305-1048. DOI: 10.1093/nar/gnh032. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC373427/ (visited on 08/05/2017).

[69] W. A. Schulz, "L1 retrotransposons in human cancers," *Journal of Biomedicine and Biotechnology*, vol. 2006, 2006, ISSN: 1110-7243. DOI: 10.1155/JBB/2006/83672. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1559935/ (visited on 08/05/2017).

[70] R. K. Slotkin and R. Martienssen, "Transposable elements and the epigenetic regulation of the genome," *Nature Reviews Genetics*, vol. 8, no. 4, pp. 272–285, Apr. 2007, ISSN: 1471-0056. DOI: 10.1038/nrg2072. [Online]. Available: http://www.nature.com.sare.upf.edu/nrg/journal/v8/n4/full/nrg2072.html?foxtrotcallback=true (visited on 08/05/2017).

[71] C. V. Breton, C. J. Marsit, E. Faustman, *et al.*, "Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: The children's environmental health and disease prevention research center's epigenetics working group," *Environmental Health Perspectives*, vol. 125, no. 4, pp. 511–526, Apr. 2017, ISSN: 0091-6765. DOI: 10.1289/EHP595. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5382002/ (visited on 09/05/2017).

[72] R. J. Jackson and N. Standart, "How do MicroRNAs regulate gene expression?" *Sci. STKE*, vol. 2007, no. 367, re1–re1, Jan. 2, 2007, ISSN: 1525-8882. DOI: 10.1126/stke.3672007re1. [Online]. Available: http://stke.sciencemag.org/content/2007/367/re1 (visited on 08/05/2017).

[73] G. Mathonnet, M. R. Fabian, Y. V. Svitkin, *et al.*, "MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4f," *Science*, vol. 317, no. 5845, pp. 1764–1767, Sep. 21, 2007, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1146067. [Online]. Available: http://science.sciencemag.org/content/317/5845/1764 (visited on 09/05/2017).

[74] S. Nottrott, M. J. Simard, and J. D. Richter, "Human let-7a miRNA blocks protein production on actively translating polyribosomes," *Nature Structural & Molecular Biology*, vol. 13, no. 12, pp. 1108–1114, Dec. 2006, ISSN: 1545-9993. DOI: 10.1038/nsmb1173.

[75] M. H. Sohel, "Extracellular/circulating MicroRNAs: Release mechanisms, functions and challenges," *Achievements in the Life Sciences*, vol. 10, no. 2, pp. 175–186, Dec. 1, 2016, ISSN: 2078-1520. DOI: 10.1016/j.als.2016.11.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2078152016300797 (visited on 09/05/2017).

[76] D. P. Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 23, 2004, ISSN: 0092-8674.

[77] ——, "MicroRNA target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, Jan. 23, 2009, ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.01.002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3794896/ (visited on 09/05/2017).

[78] M. R. Fabian, N. Sonenberg, and W. Filipowicz, "Regulation of mRNA translation and stability by microRNAs," *Annual Review of Biochemistry*, vol. 79, pp. 351–379, 2010, ISSN: 1545-4509. DOI: 10.1146/annurev-biochem-060308-103103.

[79] E. A. Vucic, K. L. Thu, L. A. Pikor, *et al.*, "Smoking status impacts microRNA mediated prognosis and lung adenocarcinoma biology," *BMC Cancer*, vol. 14, Oct. 24, 2014, ISSN: 1471-2407. DOI:

10.1186/1471-2407-14-778. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216369/ (visited on 09/05/2017).

[80] K. Vrijens, V. Bollati, and T. S. Nawrot, "MicroRNAs as potential signatures of environmental exposure or effect: A systematic review," *Environmental Health Perspectives*, vol. 123, no. 5, pp. 399–411, May 2015, ISSN: 1552-9924. DOI: 10.1289/ehp.1408459.

[81] P. Vineis, A. E. Khan, J. Vlaanderen, and R. Vermeulen, "The impact of new research technologies on our understanding of environmental causes of disease: The concept of clinical vulnerability," *Environmental Health*, vol. 8, p. 54, Nov. 30, 2009, ISSN: 1476-069X. DOI: 10.1186/1476-069X-8-54. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2793242/ (visited on 09/06/2017).

[82] V. Breda, S. G. J, L. C. Wilms, S. Gaj, D. G. J. Jennen, J. J. Briedé, J. C. S. Kleinjans, D. Kok, and T. M. C. M, "The exposome concept in a human nutrigenomics study: Evaluating the impact of exposure to a complex mixture of phytochemicals using transcriptomics signatures," *Mutagenesis*, vol. 30, no. 6, pp. 723–731, Nov. 1, 2015, ISSN: 0267-8357. DOI: 10.1093/mutage/gev008. [Online]. Available: https://academic.oup.com/mutage/article/30/6/723/2622750/The-exposome-concept-in-a-human-nutrigenomics (visited on 09/06/2017).

[83] S. Paul and S. A. Amundson, "Differential effect of active smoking on gene expression in male and female smokers," *Journal of carcinogenesis & mutagenesis*, vol. 5, 2014, ISSN: 2157-2518. DOI: 10.4172/2157-2518.1000198. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4303254/ (visited on 09/06/2017).

[84] V. Karttunen, P. Myllynen, G. Prochazka, O. Pelkonen, D. Segerbäck, and K. Vähäkangas, "Placental transfer and DNA binding of benzo(a)pyrene in human placental perfusion," *Toxicology Letters*, vol. 197, no. 2, pp. 75–81, Aug. 16, 2010, ISSN: 1879-3169. DOI: 10.1016/j.toxlet.2010.04.028.

[85] P. Huuskonen, M. R. Amezaga, M. Bellingham, *et al.*, "The human placental proteome is affected by maternal smoking," *Reproductive*

Toxicology (Elmsford, N.y.), vol. 63, pp. 22–31, Aug. 2016, ISSN: 0890-6238. DOI: 10.1016/j.reprotox.2016.05.009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4991937/ (visited on 09/06/2017).

[86] S. Brockmeyer and A. D'Angiulli, "How air pollution alters brain development: The role of neuroinflammation," *Translational Neuroscience*, vol. 7, no. 1, pp. 24–30, Mar. 21, 2016, ISSN: 2081-3856. DOI: 10.1515/tnsci-2016-0005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017593/ (visited on 09/06/2017).

[87] C. Menni, G. Kastenmüller, A. K. Petersen, *et al.*, "Metabolomic markers reveal novel pathways of ageing and early development in human populations," *International Journal of Epidemiology*, vol. 42, no. 4, pp. 1111–1119, Aug. 2013, ISSN: 0300-5771. DOI: 10.1093/ije/dyt094. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3781000/ (visited on 09/06/2017).

[88] P. Elliott, J. M. Posma, Q. Chan, *et al.*, "Urinary metabolic signatures of human adiposity," *Science Translational Medicine*, vol. 7, no. 285, 285ra62, Apr. 29, 2015, ISSN: 1946-6242. DOI: 10.1126/scitranslmed.aaa5680.

[89] F. Gu, A. Derkach, N. D. Freedman, *et al.*, "Cigarette smoking behaviour and blood metabolomics," *International Journal of Epidemiology*, vol. 45, no. 5, pp. 1421–1432, Oct. 2016, ISSN: 1464-3685. DOI: 10.1093/ije/dyv330.

[90] U. E. Rolle-Kampczyk, J. Krumsiek, W. Otto, S. W. Röder, T. Kohajda, M. Borte, F. Theis, I. Lehmann, and M. von Bergen, "Metabolomics reveals effects of maternal smoking on endogenous metabolites from lipid metabolism in cord blood of newborns," *Metabolomics*, vol. 12, 2016, ISSN: 1573-3882. DOI: 10.1007/s11306-016-0983-z. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4783445/ (visited on 09/06/2017).

[91] Y. Matsumura and H. N. Ananthaswamy, "Toxic effects of ultraviolet radiation on the skin," *Toxicology and Applied Pharmacology*, Toxicology of the Skin, vol. 195, no. 3, pp. 298–308, Mar. 15, 2004, ISSN: 0041-008X. DOI: 10.1016/j.taap.2003.08.019. [Online].

Available: http://www.sciencedirect.com/science/article/pii/S0041008X03004952 (visited on 08/26/2017).

[92] L. R. Sklar, F. Almutawa, H. W. Lim, and I. Hamzavi, "Effects of ultraviolet radiation, visible light, and infrared radiation on erythema and pigmentation: A review," *Photochemical & Photobiological Sciences*, vol. 12, no. 1, pp. 54–64, Dec. 13, 2012, ISSN: 1474-9092. DOI: 10.1039/C2PP25152C. [Online]. Available: http://pubs.rsc.org/en/content/articlelanding/2013/pp/c2pp25152c (visited on 08/26/2017).

[93] J. H. Epstein, "Photocarcinogenesis: A review," *National Cancer Institute Monograph*, no. 50, pp. 13–25, Dec. 1978, ISSN: 0083-1921.

[94] Y. Matsumura and H. N. Ananthaswamy, "Short-term and long-term cellular and molecular events following UV irradiation of skin: Implications for molecular medicine," *Expert Reviews in Molecular Medicine*, vol. 4, no. 26, pp. 1–22, Dec. 2, 2002, ISSN: 1462-3994. DOI: doi:10.1017/S146239940200532X.

[95] M. F. Holick, "Vitamin d: A d-lightful health perspective," *Nutrition Reviews*, vol. 66, S182–S194, suppl_2 Oct. 1, 2008, ISSN: 0029-6643. DOI: 10.1111/j.1753-4887.2008.00104.x. [Online]. Available: https://academic.oup.com/nutritionreviews/article/66/suppl_2/S182/1856240/Vitamin-D-a-D-Lightful-health-perspective (visited on 08/10/2017).

[96] J. A. MacLaughlin, R. R. Anderson, and M. F. Holick, "Spectral character of sunlight modulates photosynthesis of previtamin d3 and its photoisomers in human skin," *Science*, vol. 216, no. 4549, pp. 1001–1003, May 28, 1982, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.6281884. [Online]. Available: http://science.sciencemag.org/content/216/4549/1001 (visited on 08/10/2017).

[97] E. M. Hume, N. S. Lucas, and H. H. Smith, "On the absorption of vitamin d from the skin," *Biochemical Journal*, vol. 21, no. 2, pp. 362–367, 1927, ISSN: 0264-6021. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1251921/ (visited on 08/10/2017).

[98] M. F. Holick, "Vitamin d deficiency," *New England Journal of Medicine*, vol. 357, no. 3, pp. 266–281, Jul. 19, 2007, ISSN: 0028-

4793. DOI: 10.1056/NEJMra070553. [Online]. Available: http://dx.doi.org/10.1056/NEJMra070553 (visited on 08/10/2017).

[99] P. H. Hart, S. Gorman, and J. J. Finlay-Jones, "Modulation of the immune system by UV radiation: More than just the effects of vitamin d?" *Nature Reviews Immunology*, vol. 11, no. 9, pp. 584–596, Sep. 2011, ISSN: 1474-1733. DOI: 10.1038/nri3045. [Online]. Available: http://www.nature.com/nri/journal/v11/n9/full/nri3045.html?foxtrotcallback=true (visited on 08/26/2017).

[100] A. R. Webb, "Who, what, where and when-influences on cutaneous vitamin d synthesis," *Progress in Biophysics and Molecular Biology*, vol. 92, no. 1, pp. 17–25, Sep. 2006, ISSN: 0079-6107. DOI: 10.1016/j.pbiomolbio.2006.02.004.

[101] N. Binkley, R. Novotny, D. Krueger, T. Kawahara, Y. G. Daida, G. Lensmeyer, B. W. Hollis, and M. K. Drezner, "Low vitamin d status despite abundant sun exposure," *The Journal of Clinical Endocrinology & Metabolism*, vol. 92, no. 6, pp. 2130–2135, Jun. 1, 2007, ISSN: 0021-972X. DOI: 10.1210/jc.2006-2250. [Online]. Available: https://academic.oup.com/jcem/article/92/6/2130/2597445/Low-Vitamin-D-Status-despite-Abundant-Sun-Exposure (visited on 08/10/2017).

[102] F. M. Moy, "Vitamin d status and its associated factors of free living malay adults in a tropical country, malaysia," *Journal of Photochemistry and Photobiology. B, Biology*, vol. 104, no. 3, pp. 444–448, Sep. 2, 2011, ISSN: 1873-2682. DOI: 10.1016/j.jphotobiol.2011.05.002.

[103] T. L. Clemens, J. S. Adams, S. L. Henderson, and M. F. Holick, "Increased skin pigment reduces the capacity of skin to synthesise vitamin d3," *Lancet (London, England)*, vol. 1, no. 8263, pp. 74–76, Jan. 9, 1982, ISSN: 0140-6736.

[104] J. MacLaughlin and M. F. Holick, "Aging decreases the capacity of human skin to produce vitamin d3.," *Journal of Clinical Investigation*, vol. 76, no. 4, pp. 1536–1538, Oct. 1985, ISSN: 0021-9738. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC424123/ (visited on 08/10/2017).

[105] T. Chen, F. Chimeh, Z. Lu, *et al.*, "Factors that influence the cutaneous synthesis and dietary sources of vitamin d," *Archives of*

biochemistry and biophysics, vol. 460, no. 2, pp. 213–217, Apr. 15, 2007, ISSN: 0003-9861. DOI: `10.1016/j.abb.2006.12.017`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698590/` (visited on 08/10/2017).

[106] J. Wortsman, L. Y. Matsuoka, T. C. Chen, Z. Lu, and M. F. Holick, "Decreased bioavailability of vitamin d in obesity," *The American Journal of Clinical Nutrition*, vol. 72, no. 3, pp. 690–693, Sep. 1, 2000, ISSN: 0002-9165, 1938-3207. [Online]. Available: `http://ajcn.nutrition.org/content/72/3/690` (visited on 08/10/2017).

[107] S.-W. Lin, D. C. Wheeler, Y. Park, M. Spriggs, A. R. Hollenbeck, D. M. Freedman, and C. C. Abnet, "Prospective study of ultraviolet radiation exposure and mortality risk in the united states," *American Journal of Epidemiology*, vol. 178, no. 4, pp. 521–533, Aug. 15, 2013, ISSN: 0002-9262. DOI: `10.1093/aje/kws589`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3736750/` (visited on 08/10/2017).

[108] W. B. Grant, "Roles of solar UVB and vitamin d in reducing cancer risk and increasing survival," *Anticancer Research*, vol. 36, no. 3, pp. 1357–1370, Mar. 1, 2016, ISSN: 0250-7005, 1791-7530. [Online]. Available: `http://ar.iiarjournals.org/content/36/3/1357` (visited on 08/10/2017).

[109] S. Andrews. (2010). FastQC a quality control tool for high throughput sequence data, [Online]. Available: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[110] Y. Liao, G. K. Smyth, and W. Shi, "The subread aligner: Fast, accurate and scalable read mapping by seed-and-vote," *Nucleic Acids Research*, vol. 41, no. 10, e108, May 2013, ISSN: 0305-1048. DOI: `10.1093/nar/gkt214`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3664803/` (visited on 08/10/2017).

[111] T. Lappalainen, M. Sammeth, M. R. Friedländer, *et al.*, "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, vol. 501, no. 7468, pp. 506–511, Sep. 26, 2013, ISSN: 0028-0836. DOI: `10.1038/nature12531`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3918453/` (visited on 08/26/2017).

[112] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, Dec. 5, 2014, ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. [Online]. Available: https://doi.org/10.1186/s13059-014-0550-8 (visited on 08/10/2017).

[113] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009, ISSN: 1750-2799. DOI: 10.1038/nprot.2008.211.

[114] F. Xie, P. Xiao, D. Chen, L. Xu, and B. Zhang, "miRDeepFinder: A miRNA analysis tool for deep sequencing of plant small RNAs," *Plant Molecular Biology*, Jan. 31, 2012, ISSN: 1573-5028. DOI: 10.1007/s11103-012-9885-2.

[115] M. Moukayed and W. B. Grant, "Molecular link between vitamin d and cancer prevention," *Nutrients*, vol. 5, no. 10, pp. 3993–4021, Sep. 30, 2013, ISSN: 2072-6643. DOI: 10.3390/nu5103993. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820056/ (visited on 09/17/2017).

[116] D. L. Stirewalt and J. P. Radich, "The role of FLT3 in haematopoietic malignancies," *Nature Reviews. Cancer*, vol. 3, no. 9, pp. 650–665, Sep. 2003, ISSN: 1474-175X. DOI: 10.1038/nrc1169.

[117] M. Levis and D. Small, "FLT3: ITDoes matter in leukemia," *Leukemia*, vol. 17, no. 9, pp. 1738–1752, 2003, ISSN: 0887-6924. DOI: 10.1038/sj.leu.2403099. [Online]. Available: https://www.nature.com/leu/journal/v17/n9/full/2403099a.html (visited on 09/17/2017).

[118] E. K. Cahoon, R. M. Pfeiffer, D. C. Wheeler, J. Arhancet, S.-W. Lin, B. H. Alexander, M. S. Linet, and D. M. Freedman, "Relationship between ambient ultraviolet radiation and non-hodgkin lymphoma subtypes: A u.s. population-based study of racial and ethnic groups," *International journal of cancer. Journal international du cancer*, vol. 136, no. 5, E432–E441, Mar. 1, 2015, ISSN: 0020-7136. DOI: 10.1002/ijc.29237. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4268147/ (visited on 09/17/2017).

[119] E. T. Chang, A. J. Canchola, M. Cockburn, Y. Lu, S. S. Wang, L. Bernstein, C. A. Clarke, and P. L. Horn-Ross, "Adulthood residential ultraviolet radiation, sun sensitivity, dietary vitamin d, and risk of lymphoid malignancies in the california teachers study," *Blood*, vol. 118, no. 6, pp. 1591–1599, Aug. 11, 2011, ISSN: 0006-4971. DOI: 10.1182/blood-2011-02-336065. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3156047/ (visited on 09/17/2017).

[120] E. J. Jung, T. Kawai, H. K. Park, Y. Kubo, K. Rokutan, and S. Arase, "Identification of ultraviolet b-sensitive genes in human peripheral blood cells," *The journal of medical investigation: JMI*, vol. 55, no. 3, pp. 204–210, Aug. 2008, ISSN: 1349-6867.

[121] H. d. l. Fuente, A. Lamana, M. Mittelbrunn, S. Perez-Gala, S. Gonzalez, A. García-Diez, M. Vega, and F. Sanchez-Madrid, "Identification of genes responsive to solar simulated UV radiation in human monocyte-derived dendritic cells," *PLOS ONE*, vol. 4, no. 8, e6735, Aug. 26, 2009, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006735. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006735 (visited on 09/17/2017).

[122] X. C. Dopico, M. Evangelou, R. C. Ferreira, *et al.*, "Widespread seasonal gene expression reveals annual differences in human immunity and physiology," *Nature Communications*, vol. 6, ncomms8000, May 12, 2015, ISSN: 2041-1723. DOI: 10.1038/ncomms8000. [Online]. Available: https://www.nature.com/articles/ncomms8000 (visited on 09/17/2017).

[123] C. C. Pritchard, E. Kroh, B. Wood, J. D. Arroyo, K. J. Dougherty, M. M. Miyaji, J. F. Tait, and M. Tewari, "Blood cell origin of circulating microRNAs: A cautionary note for cancer biomarker studies," *Cancer prevention research (Philadelphia, Pa.)*, vol. 5, no. 3, pp. 492–497, Mar. 2012, ISSN: 1940-6207. DOI: 10.1158/1940-6207.CAPR-11-0370. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4186243/ (visited on 09/17/2017).

[124] X. Huang, T. Yuan, M. Tschannen, *et al.*, "Characterization of human plasma-derived exosomal RNAs by deep sequencing," *BMC Genomics*, vol. 14, p. 319, May 10, 2013, ISSN: 1471-2164. DOI:

10.1186/1471-2164-14-319. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3653748/ (visited on 09/17/2017).

[125] L. Rauhala, L. Hämäläinen, P. Salonen, G. Bart, M. Tammi, S. Pasonen-Seppänen, and R. Tammi, "Low dose ultraviolet b irradiation increases hyaluronan synthesis in epidermal keratinocytes via sequential induction of hyaluronan synthases has1–3 mediated by p38 and ca2+/calmodulin-dependent protein kinase II (CaMKII) signaling," *The Journal of Biological Chemistry*, vol. 288, no. 25, pp. 17 999–18 012, Jun. 21, 2013, ISSN: 0021-9258. DOI: 10.1074/jbc.M113.472530. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3689945/ (visited on 09/17/2017).

[126] H. H. Al-Khalaf, P. Mohideen, S. C. Nallar, D. V. Kalvakolanu, and A. Aboussekhra, "The cyclin-dependent kinase inhibitor p16ink4a physically interacts with transcription factor sp1 and cyclin-dependent kinase 4 to transactivate MicroRNA-141 and MicroRNA-146b-5p spontaneously and in response to ultraviolet light-induced DNA damage," *The Journal of Biological Chemistry*, vol. 288, no. 49, pp. 35 511–35 525, Dec. 6, 2013, ISSN: 0021-9258. DOI: 10.1074/jbc.M113.512640. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3853297/ (visited on 09/17/2017).

[127] Y. Xu, B. Zhou, D. Wu, Z. Yin, and D. Luo, "Baicalin modulates microRNA expression in UVB irradiated mouse skin," *Journal of Biomedical Research*, vol. 26, no. 2, pp. 125–134, Mar. 2012, ISSN: 1674-8301. DOI: 10.1016/S1674-8301(12)60022-0. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3597329/ (visited on 09/17/2017).

[128] M. B. Løvendorf, H. Mitsui, J. R. Zibert, M. A. Røpke, M. Hafner, B. Dyring-Andersen, C. M. Bonefeld, J. G. Krueger, and L. Skov, "Laser capture microdissection followed by next-generation sequencing identifies disease-related microRNAs in psoriatic skin that reflect systemic microRNA changes in psoriasis," *Experimental Dermatology*, vol. 24, no. 3, pp. 187–193, Mar. 2015, ISSN: 1600-0625. DOI: 10.1111/exd.12604.

[129] H. Hermann, T. Runnel, A. Aab, *et al.*, "miR-146b probably assists miRNA-146a in the suppression of keratinocyte proliferation and

inflammatory responses in psoriasis," *The Journal of Investigative Dermatology*, vol. 137, no. 9, pp. 1945–1954, Sep. 2017, ISSN: 1523-1747. DOI: 10.1016/j.jid.2017.05.012.

[130] P. Wolf, S. N. Byrne, A. Y. Limon-Flores, G. Hoefler, and S. E. Ullrich, "Serotonin signaling is crucial in the induction of PUVA-induced systemic suppression of delayed type hypersensitivity but not local apoptosis or inflammation of the skin," *Experimental dermatology*, vol. 25, no. 7, pp. 537–543, Jul. 2016, ISSN: 0906-6705. DOI: 10.1111/exd.12990. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4927393/ (visited on 09/17/2017).

[131] A. Toll, R. Salgado, B. Espinet, *et al.*, "MiR-204 silencing in intraepithelial to invasive cutaneous squamous cell carcinoma progression," *Molecular Cancer*, vol. 15, Jul. 25, 2016, ISSN: 1476-4598. DOI: 10.1186/s12943-016-0537-z. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4960761/ (visited on 09/17/2017).

[132] J. Etich, V. Bergmeier, L. Pitzler, and B. Brachvogel, "Identification of a reference gene for the quantification of mRNA and miRNA expression during skin wound healing," *Connective Tissue Research*, vol. 58, no. 2, pp. 196–207, Mar. 2017, ISSN: 1607-8438. DOI: 10.1080/03008207.2016.1210606.

[133] G. Li, C. Luna, J. Qiu, D. L. Epstein, and P. Gonzalez, "Role of miR-204 in the regulation of apoptosis, endoplasmic reticulum stress response, and inflammation in human trabecular meshwork cells," *Investigative Ophthalmology & Visual Science*, vol. 52, no. 6, pp. 2999–3007, May 2011, ISSN: 0146-0404. DOI: 10.1167/iovs.10-6708. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3109013/ (visited on 09/17/2017).

[134] A. Singh, E. Willems, A. Singh, I. M. Ong, and A. K. Verma, "Ultraviolet radiation-induced differential microRNA expression in the skin of hairless SKH1 mice, a widely used mouse model for dermatology research," *Oncotarget*, vol. 7, no. 51, pp. 84 924–84 937, Oct. 26, 2016, ISSN: 1949-2553. DOI: 10.18632/oncotarget.12913. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5356709/ (visited on 09/17/2017).

[135] B. Petersen, E. Thieden, P. A. Philipsen, J. Heydenreich, H. C. Wulf, and A. R. Young, "Determinants of personal ultraviolet-radiation exposure doses on a sun holiday," *The British Journal of Dermatology*, vol. 168, no. 5, pp. 1073–1079, May 2013, ISSN: 1365-2133. DOI: `10.1111/bjd.12211`.

[136] S. Pineda, P. Gomez-Rubio, A. Picornell, K. Bessonov, M. Márquez, M. Kogevinas, F. X. Real, K. Van Steen, and N. Malats, "Framework for the integration of genomics, epigenomics and transcriptomics in complex diseases," *Human Heredity*, vol. 79, no. 3, pp. 124–136, 2015, ISSN: 1423-0062. DOI: `10.1159/000381184`.

[137] P. Suravajhala, L. J. A. Kogelman, and H. N. Kadarmideen, "Multiomic data integration and analysis using systems genomics approaches: Methods and applications in animal production, health and welfare," *Genetics, Selection, Evolution : GSE*, vol. 48, Apr. 29, 2016, ISSN: 0999-193X. DOI: `10.1186/s12711-016-0217-x`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4850674/` (visited on 08/10/2017).

[138] L. Kannan, M. Ramos, A. Re, *et al.*, "Public data and open source tools for multi-assay genomic investigation of disease," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 603–615, Jul. 2016, ISSN: 1467-5463. DOI: `10.1093/bib/bbv080`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4945830/` (visited on 08/10/2017).

[139] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, ISSN: 1061-4036. DOI: `10.1038/ng.2764`. [Online]. Available: `http://www.nature.com/ng/journal/v45/n10/full/ng.2764.html?foxtrotcallback=true` (visited on 08/10/2017).

[140] T. J. Hudson, W. Anderson, A. Aretz, *et al.*, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, pp. 993–998, Apr. 15, 2010, ISSN: 0028-0836. DOI: `10.1038/nature08987`. [Online]. Available: `https://www.nature.com/nature/journal/v464/n7291/full/nature08987.html` (visited on 08/10/2017).

[141] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data reposi-

tory," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, Jan. 1, 2002, ISSN: 0305-1048. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC99122/ (visited on 08/10/2017).

[142] T. Barrett, S. E. Wilhite, P. Ledoux, *et al.*, "NCBI GEO: Archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, pp. D991–D995, D1 Jan. 1, 2013, ISSN: 0305-1048. DOI: 10.1093/nar/gks1193. [Online]. Available: https://academic.oup.com/nar/article/41/D1/D991/1067995/NCBI-GEO-archive-for-functional-genomics-data-sets (visited on 08/10/2017).

[143] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: The NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, Jan. 1, 2001, ISSN: 0305-1048. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC29783/ (visited on 08/10/2017).

[144] T. Svingen and A. M. Vinggaard, "The risk of chemical cocktail effects and how to deal with the issue," *J Epidemiol Community Health*, vol. 70, no. 4, pp. 322–323, Apr. 1, 2016, ISSN: 0143-005X, 1470-2738. DOI: 10.1136/jech-2015-206268. [Online]. Available: http://jech.bmj.com/content/70/4/322 (visited on 08/10/2017).

[145] M. C. Turner, M. Nieuwenhuijsen, K. Anderson, D. Balshaw, Y. Cui, G. Dunton, J. A. Hoppin, P. Koutrakis, and M. Jerrett, "Assessing the exposome with external measures: Commentary on the state of the science and research recommendations," *Annual Review of Public Health*, vol. 38, no. 1, pp. 215–239, Mar. 20, 2017, ISSN: 0163-7525. DOI: 10.1146/annurev-publhealth-082516-012802. [Online]. Available: http://www.annualreviews.org.sare.upf.edu/doi/10.1146/annurev-publhealth-082516-012802 (visited on 08/24/2017).

[146] J. A. Stingone, G. M. Buck Louis, S. F. Nakayama, R. C. Vermeulen, R. K. Kwok, Y. Cui, D. M. Balshaw, and S. L. Teitelbaum, "Toward greater implementation of the exposome research paradigm within environmental epidemiology," *Annual Review of Public Health*, vol. 38, no. 1, pp. 315–327, Mar. 20, 2017, ISSN: 0163-7525. DOI: 10.1146/annurev-publhealth-082516-012750.

[Online]. Available: http://www.annualreviews.org.sare.upf.edu/doi/10.1146/annurev-publhealth-082516-012750 (visited on 08/25/2017).

[147] O. Robinson and M. Vrijheid, "The pregnancy exposome," *Current Environmental Health Reports*, vol. 2, no. 2, pp. 204–213, Jun. 1, 2015, ISSN: 2196-5412. DOI: 10.1007/s40572-015-0043-2. [Online]. Available: https://link.springer.com/article/10.1007/s40572-015-0043-2 (visited on 08/26/2017).

[148] F. Dominici, R. D. Peng, C. D. Barr, and M. L. Bell, "Protecting human health from air pollution: Shifting from a single-pollutant to a multipollutant approach," *Epidemiology*, vol. 21, no. 2, pp. 187–194, Mar. 2010, ISSN: 1044-3983. DOI: 10.1097/EDE.0b013e3181cc86e8. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001648-201003000-00005 (visited on 08/26/2017).

[149] R. C. Gentleman, V. J. Carey, D. M. Bates, *et al.*, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, R80, 2004, ISSN: 1474-760X. DOI: 10.1186/gb-2004-5-10-r80. [Online]. Available: http://dx.doi.org/10.1186/gb-2004-5-10-r80 (visited on 01/16/2017).

[150] W. Huber, V. J. Carey, R. Gentleman, *et al.*, "Orchestrating high-throughput genomic analysis with bioconductor," *Nature Methods*, vol. 12, no. 2, pp. 115–121, Feb. 2015, ISSN: 1548-7091. DOI: 10.1038/nmeth.3252. [Online]. Available: https://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html (visited on 08/10/2017).

[151] J. Li and L. Ji, "Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix," *Heredity*, vol. 95, no. 3, pp. 221–227, Sep. 2005, ISSN: 0018-067X. DOI: 10.1038/sj.hdy.6800717.

[152] L. Agier, L. Portengen, M. Chadeau-Hyam, *et al.*, "A systematic comparison of linear regression–based statistical methods to assess exposome-health associations," *Environmental Health Perspectives*, vol. 124, no. 12, May 24, 2016, ISSN: 0091-6765. DOI: 10.1289/EHP172. [Online]. Available: http://ehp.niehs.nih.gov/EHP172 (visited on 01/18/2017).

[153] C. Hernandez-Ferrer, C. Ruiz-Arenas, A. Beltran-Gomila, and J. R. González, "MultiDataSet: An r package for encapsulating multiple data sets with application to omic data integration," *BMC Bioinformatics*, vol. 18, p. 36, Jan. 17, 2017, ISSN: 1471-2105. DOI: 10.1186/s12859-016-1455-1. [Online]. Available: https://doi.org/10.1186/s12859-016-1455-1 (visited on 08/11/2017).

[154] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, e47, Apr. 20, 2015, ISSN: 1362-4962. DOI: 10.1093/nar/gkv007.

[155] B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth, "Robust hyperparameters estimation protects agains hypervariable genes and improves power to detect differential expression," *The annals of applied statistics*, vol. 10, no. 2, pp. 946–963, Jun. 2016, ISSN: 1932-6157. DOI: 10.1214/16-AOAS920. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5373812/ (visited on 08/11/2017).

[156] A. E. Teschendorff, J. Zhuang, and M. Widschwendter, "Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies," *Bioinformatics (Oxford, England)*, vol. 27, no. 11, pp. 1496–1505, Jun. 1, 2011, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr171.

[157] M. CJ, R. MC, C. GT, F. JN, and B. JL, "The comparative toxicogenomics database (CTD): A resource for comparative toxicological studies," *Journal of experimental zoology. Part A, Comparative experimental biology*, vol. 305, no. 9, pp. 689–692, Sep. 1, 2006, ISSN: 1548-8969. DOI: 10.1002/jez.a.307. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1586110/ (visited on 08/10/2017).

[158] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The comparative toxicogenomics database: Update 2017," *Nucleic Acids Research*, vol. 45, pp. D972–D978, Database issue Jan. 4, 2017, ISSN: 0305-1048. DOI: 10.1093/nar/gkw838. [Online]. Avail-

able: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210612/ (visited on 08/10/2017).

[159] M. Ramos, L. Schiffer, A. Re, *et al.*, "Software for the integration of multi-omics experiments in bioconductor," *bioRxiv*, p. 144 774, Jun. 1, 2017. DOI: 10.1101/144774. [Online]. Available: http://www.biorxiv.org/content/early/2017/06/01/144774 (visited on 08/12/2017).

[160] M. Guxens, F. Ballester, M. Espada, *et al.*, "Cohort profile: The INMA–INfancia y medio ambiente–(environment and childhood) project," *International Journal of Epidemiology*, vol. 41, no. 4, pp. 930–940, Aug. 2012, ISSN: 1464-3685. DOI: 10.1093/ije/dyr054.

[161] J. Bousquet, J. Anto, C. Auffray, *et al.*, "MeDALL (mechanisms of the development of ALLergy): An integrated approach from phenotypes to systems medicine," *Allergy*, vol. 66, no. 5, pp. 596–604, May 1, 2011, ISSN: 1398-9995. DOI: 10.1111/j.1398-9995.2010.02534.x. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1398-9995.2010.02534.x/abstract (visited on 08/12/2017).

[162] J. Sunyer, M. Esnaola, M. Alvarez-Pedrerol, *et al.*, "Association between traffic-related air pollution in schools and cognitive development in primary school children: A prospective cohort study," *PLOS Medicine*, vol. 12, no. 3, e1001792, Mar. 3, 2015, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001792. [Online]. Available: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001792 (visited on 08/12/2017).

[163] C. J. Patel, N. Pho, M. McDuffie, J. Easton-Marks, C. Kothari, I. S. Kohane, and P. Avillach, "A database of human exposomes and phenomes from the US national health and nutrition examination survey," *Scientific Data*, vol. 3, p. 160 096, Oct. 25, 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.96. [Online]. Available: http://www.nature.com/articles/sdata201696 (visited on 01/16/2017).

[164] C. for Disease Control {and} Prevention (CDC). (2016). National health and nutrition examination survey data, [Online]. Available: https://nhanes.hms.harvard.edu/ (visited on 08/26/2017).