Classifications transcriptomiques des dermatoses inflammatoires : application à la dermatite atopique

Alain Lefèvre-Utile

Thèse de doctorat

NNT : 2021UPASL041

université PARIS-SACLAY

Inflammatory dermatitis transcriptomic classifications: application to atopic dermatitis

Classifications transcriptomiques des dermatoses inflammatoires : application à la dermatite atopique

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°582 : Cancérologie, Biologie, Médecine, Santé (CBMS)
Spécialité de doctorat : Aspects moléculaires et cellulaires de la biologie
Unité de recherche : Immunologie humaine, physiopathologie and immunothérapie (U976)
Référent : Faculté de médecine

Thèse présentée et soutenue à Paris-Saclay le 1er juin 2021 par

Alain LEFEVRE-UTILE

Composition du Jury

Christine BODEMER Professeur des universités, praticien hospitalier. INSERM, APHP, Université de Paris.	Présidente
Sylvie CHEVRET Professeur des universités, praticien hospitalier. INSERM, APHP, Université de Paris.	Rapportrice et examinatrice
Audrey NOSBAUM Professeur des universités, praticien hospitalier. INSERM, HCL, Université de Lyon	Rapportrice et examinatrice
Gaëlle LELANDAIS Professeur des universités. Université Paris Saclay, CNRS.	Examinatrice
Rodolphe THIEBAUT Professeur des universités, praticien hospitalier. INSERM, Université de Bordeaux.	Examinateur

Directeur de thèse

Vassili SOUMELIS
Professeur des universités, praticien hospitalier.
INSERM, APHP, Université de Paris.

Mon problème,
avec les classements,
c'est qu'ils ne durent pas ;
à peine ai-je fini de mettre de l'ordre
que cet ordre est déjà caduc.

Comme tout le monde,
je suppose,
je suis pris parfois de frénésie de rangement ;
l'abondance des choses à ranger,
la quasi-impossibilité de les distribuer
selon des critères vraiment satisfaisants
font que je n'en viens jamais à bout,
que je m'arrête à des rangements provisoires et flous,
à peine plus efficace que l'anarchie initiale.

In *Penser/classer*, Georges Perec, 1985

ACKNOWLEDGMENTS

Excuse my French but…
Commençons cette thèse par les remerciements comme le veut la coutume.

Merci au **Professeure Christine Bodemer** pour avoir accepté d'être la Présidente de mon jury de thèse. Il y a quelques années, j'étais dans votre service, accueilli comme un des votre : cela a laissé des traces. Cette thèse s'intéresse de très près à la dermatologie, discipline qui bien qu'elle ne soit pas la mienne, reste ma danseuse.

Merci au **Professeures Sylvie Chevret** et **Audrey Nosbaum** pour avoir accepté d'être Rapporteures (ou Rapporteuses ou Rapportrices, le dictionnaire n'a pas encore tranché) de cette thèse. Merci pour votre temps et pour l'attention toute particulière que vous avez porté à ce travail. J'espère vous avoir intéressé plus que désespéré , et questionné d'avantage qu'énervé.
En souvenir, **Professeure Chevret**, de ces heures passées sur les bancs des amphithéâtres de la Faculté Villemin, à écouter (attentivement bien sûr) vos cours de biostatistiques. C'était dans une autre vie.
En souvenir, **Professeure Nosbaum**, de cette réunion GREAT à la veille d'un certain confinement. Dernier rendez-vous entre collègues non masqués qui me restent en mémoire. C'était dans un autre monde.

Merci au **Professeure Gaëlle Lelandais** pour avoir accepté d'être dans mon jury et surtout pour m'avoir tant transmis. Assister à ton DU a été un des bons moments de la thèse. Je me suis senti étudiant, me suis vu progresser, cette sensation m'avais manqué.

Merci au **Professeur Rodolphe Thiébaut** pour avoir accepté d'être dans mon jury. Merci aussi à vous et à Boris Hejblum de m'avoir fait confiance en me proposant d'encadrer un projet de vos étudiants de Master 2.

Merci au **Professeur Jean-David Bouaziz** et à **Philippe Hupé** pour leur participation à mon Comité de thèse et tous les conseils et encouragements qu'ils m'ont prodigués.

Merci au **Professeur Vassili Soumelis**, mon directeur de thèse. Tu m'as fait confiance, moi le débutant. Merci pour tes explications, ta pédagogie, face à mes questions parfois hors cadre. Tu insuffles dans ton équipe la bienveillance et l'hédonisme qui t'animent, même quand l'époque s'obscurcie. J'espère que tu garderas cette passion et cette curiosité infinie qui sont en toi. J'espère que l'on continuera à travailler ensemble.

Merci à **Melissa Saichi**, my unsupervised partner. Ce fut bref, mais j'ai beaucoup aimé notre binôme. Désolé s'il y a eu des malentendus parfois. Je te souhaite plein de projets et de belles idées pour la suite.

Merci à **Enzo Battistella**, toujours patient avec mes question naive. J'ai été tout de suite admiratif de ta maturité technique et scientifique. Notre collaboration s'est faite en toute confiance et en toute fluidité. Au plaisir de retravailler ensemble.

Passe le merci à **Nikos Paragios** et **Maria Vakalopoulou**, pour leur maitrise de la direction, leur rigueur et la qualité des discussions qu'ils ont animé.

Merci à **Daniel Herrero Saboya**, camarade de bureau, mais pas seulement. Malgré ton arrivée dans ce Paris chaotique, tu as su t'y faire ta place. J'espère que tu t'épanouiras suffisamment pour continuer à réfléchir et l'apprendre. Un jour peut-être nous aurons un projet commun, cela me ravirait ! Que par toi survive la tradition du pique-nique sous la pluie dans le Carré Saint Louis.

Merci à **Justine Poirot**, ma co-thésarde, sœur de galère, et qui cartonne sa thèse comme une reine (*mais si, mais si*). Souvenons-nous du 1er étage du bâtiment Hayem, désaffecté, à attendre les copains. Nous aurions pu y tourner un film de zombies, et avons préféré boire beaucoup de café, mais trop peu d'apéro.
Bientôt le Champagne.

Merci à **Élise Amblard**, médecin en devenir, mais pas seulement. Merci pour avoir ralenti ton débit de parole et de réflexion pour que je puisse suivre, et d'avoir relu si attentivement mon article. Si tu veux je te lègue mon bureau.

Merci à toute ma belle équipe #9 avec en premier sur votre droite en entrant : le bureau bio-info : **Floriane**, **Iris**, **Faezeh** et bien sûr les copains des autres bureaux : **Salima**, **Arturo**, **Jasna**, **Pierre**, **Grégoire**, **Alba**, **Lucile**, **Maéva**, **Lilith**. Et aussi nos ancien(nes) : **Sarah**, **Camille**, **Sarantis**, **Caroline**, **Léa**, **Coline**, **Charlotte** et **Fanny**.

Merci au **Docteure Nuala Mooney** pour son travail de relecture, sa bienveillance. C'était toujours très bénéfique et encourageant de discuter avec toi.

Merci à **Camille Braun**, camarade d'un Semiderm infini. J'espère qu'on se croisera dans le grand Nord.

Merci à la **Fondation pour la recherche médicale**, qui m'a fait confiance et accordé un financement qui m'a permis de vivre sereinement mes trois années.

Merci au **Professeur Loïc De Pontual** de m'avoir permis de continuer à exercer la Pédiatrie dans son équipe. Pas évident à concilier avec une vie de doctorant, mais cela a eu le mérite de me remonter le moral quand je me sentais utile en apprenant des choses aux internes. En continuant d'exercer, j'ai pu garder quelques reflexes et rafraichir mes connaissances cliniques. Je n'ai qu'une peur, c'est de devenir mauvais médecin en plus d'être mauvais chercheur. A suivre donc….
Merci donc à toutes **l'équipe médicale et paramédicale de la Pédiatrie de Jean Verdier** pour m'avoir si bien accueilli et accepté de me ménager quand c'était nécessaire.

Merci à **l'équipe de Data for good** de s'être intéressée à mon projet. Et surtout merci à ceux qui y ont directement participé, et ce jusqu'au bout : **Sara Kazdaghli, Aoife Fogarty, Skander Kasdaghli, Vincent Gargasson, et Pierre-Louis Patenotre**.

Merci à **David Boutboul** (l'astronaute, pas le footballeur). Amigo, tu m'as bien aidé à rester sain d'esprit dans cet étrange état qui est d'*être en train de chercher*. A nos 459 litres de café, 182 sandwichs, bahos, pizze et nos trop rares plats du jours. Aux milliers de musiques écoutées, discutées, et partagées. A nos discussions à bâtons rompus sur tout et rien, et surtout sur rien. Tu n'oublieras pas de me rendre ma BD tout de même.

Merci à mes parents : **Louis** et **Maria**, pour leur présence constante et discrète. Ils ont su affronter quelques chantiers ces dernières années, chapeau. Malgré cela vous avez été là pour moi, pour les filles. Sans votre aide je n'aurai pas pu me sortir de toutes ces injonctions.

Merci à **Jean**, mon frère, qui sera docteur en Éthique dans quelques semaines. Reçois toute mon admiration pour ces années de labeur. Je crois que tu es un vrai chercheur, garde cette hargne face à l'institution. Reste, tout de même, à redécouvrir le goût du repos, de la bière fraiche, et de la contemplation.

Merci à mes filles : **Mathilde** et **Salomé** (dans l'ordre d'apparition), garantes de mon élan vital et de mon réveil matinal. Ce n'était pas évident de bien m'occuper de vous alors que je devais jongler avec toutes ces responsabilités et que l'époque s'obscurcissait. J'espère que vous garderez quelques souvenir de ce papa qui vous accompagnait à l'école et la crèche le matin. Vous êtes mes lumières indispensables.

Merci à **Fleur**, ma personne préférée de tous les temps. Tu n'as pas arrêté ces dernières années et la pandémie n'a pas aidé. Malgré cela tu étais là, près de moi. Notre maison (appartement… il ne faut pas exagérer), notre famille, nos projets me portent et m'inspirent. J'espère avoir le privilège d'être à tes côtés tant que possible. Par pudeur, ma déclaration s'arrête là.

Ça va aller, ça va aller, ça va aller la vie.
Ça va aller, ça va aller, bien

A la mémoire de Luc, Audrey et Mahmoud

INDEX

ABBREVIATION LIST

89ADGES	89 Atopic Dermatitis Gene Expression Signatures
AD	Atopic Dermatitis
AKR1B	Aldo-Keto Reductase Family 1 Member
BTC	Betacellulin
CASK	Calcium/Calmodulin Dependent Serine Protein Kinase
CLEC2A	C-Type Lectin Domain Family 2 Member A
CREB	cAMP Responsive Element Binding Protein
CRTC	CREB Regulated Transcription Coactivator
DEFB4A	Defensin Beta 4A
DGAT2L6	Diacylglycerol O-Acyltransferase 2 Like 6
EPGN	Epithelial Mitogen
ER	Estrogen Receptor
ERK	Extracellular Signal-regulated Kinase
FLG	Filaggrin
GO	Gene Ontology
GJB4	Gap Junction Protein Beta 4
GSTA3	Glutathione S-Transferase Alpha 3
HER2	Human Epidermal Growth Factor Receptor-2
HMOX1	Heme Oxygenase 1
IGH	Immunoglobulin Heavy Locus
IGK	Immunoglobulin Kappa Locus
IGL	Immunoglobulin Lambda Locus
IL	Interleukin
k-NN	k-Nearest Neighbors
KRT	Keratin
KRTAP	Keratin Associated Protein
LOR	Loricrin
MAE	Mean Absolute Error
MAARS	Microbes in Allergy and Autoimmunity Related to the Skin
MADAD	Meta-Analyses Derived Atopic Dermatitis
MAN2A1	Mannosidase Alpha Class 2A Member 1
MG	Metagene
ML	Machine Learning
MLP	Multilayer perceptron
MST	Minimal Spanning Tree
NFXL1	Nuclear Transcription Factor/X-Box Binding Like 1
NRS	Numeric Rating Scale
OGN	Osteoglycin
PARP1	Poly-ADP-ribose-polymerase-1
PCA	Principal Component Analysis
PI	Peptidase Inhibitor
PILAR	Proliferation-Induced Lymphocyte-Associated Receptor
PLA2G4D	Phospholipase A2 Group IVD

PM20D1	Peptidase M20 Domain Containing 1
PR	Progesterone Receptor
PRSS22	Serine Protease 22
QDA	Quadratic discriminant analysis
RBF	Radial Basis Function
RNA	Ribo Nucleic Acid
S100A	S100 Calcium Binding Protein A
SA	*Staphylococcus aureus*
SCORAD	Score Atopic Dermatitis
SLC46A2	Solute Carrier Family 46 Member 2
sPLS	Sparse Partial Least Square
SNAP23	Synaptosome Associated Protein 23
SPRR	Small Proline-Rich Protein
SVM	Support Vector Machine
TCFL5	Transcription Factor Like 5
TNM	Tumors Nodes Metastases
TOX2	TOX High Mobility Group Box Family Member 2
UGT3A2	UDP Glycosyltransferase Family 3 Member A2
VGLL2	Vestigial Like Family Member 2
VRS	Visual Rating Scale
XIST	X Inactive Specific Transcript

SYNTHESE

Introduction

Une classification est une organisation de concepts en groupes, selon un objectif donné, d'après des critères définis. Il s'agit de trouver un compromis entre l'hétérogène et l'homogène, de telle sorte que les classes, comme produits de classifications, regroupent des entités selon leurs points communs et se distinguent selon leurs différences. L'exercice de classification dépend ainsi fortement de la population étudiée et du niveau de résolution des variables utilisées. Ce qui nous pousse vers cet effort est la prise de conscience d'une complexité qui nécessite, pour être appréhendée, d'être simplifiée. Cette prise de conscience est la souvent la conséquence de l'utilisation de nouvelles approches ou d'outils, qui apportent leurs points de vue originaux (Figure 2). Les classifications étaient initialement d'avantage basées sur l'intuition et la croyance, comme l'illustre la *Doctrine des signatures* qui estimait la fonction d'un végétal selon des critères anthropomorphiques. Elles devinrent plus méthodiques avec l'arrivée de la *Systématique*, ou science de classer, théorisée par Carl Linée (1701–1777) puis Georges Cuvier (1769–1832). Les rapports entre classes sont alors régentés par des liens hiérarchiques, d'inclusion et d'exclusion, leur permettant de définir précisément le Vivant. C'est enfin Emil Hans Willi Hennig (1913–1976), un biologiste allemand qui modernisa ces approches en y apportant d'avantage de rigueur. Il introduisit en particulier, le principe de parcimonie qui place la relation la plus directe comme la plus vraisemblable, favorisant ainsi la généralisation et la reproductibilité. Les classifications en Biologie comparée cherchent à établir un lien entre les espèces, et illustrent bien le phénomène[1]. Elles ont évolué parallèlement à celles de techniques d'observation et méthode d'organisation. Initialement, les liens de proximité entre les espèces étaient déterminés sur critères anatomiques, puis embryologiques, et enfin génétiques sous la forme des arbres phylogénétiques que l'on utilise encore (Figure 5).

Tout comme en biologie, la médecine a été le terrain de nombreuses classifications. La première classification des maladies humaines remonte à Hippocrate (-460– -377) et Galien (129-201). Ils ont supposé que la physiologie et la pathologie humaine dépendaient de l'équilibre de quatre humeurs : le sang, la lymphe, les biles noire et jaune (Figure 6). Cette *Théorie des humeurs* permettait de donner une explication mécanistique des maladies (défaut

ou excès de l'humeur) ainsi qu'une orientation thérapeutique (e.g. saignée, purgatif). Ces interprétations ont fait référence jusqu'à la fin du Moyen-Âge et c'est avec l'arrivée de sciences descriptives comme l'anatomie et la physiologie que les maladies humaines ont commencé à être définies comme nous les concevons actuellement. L'Organisation Mondiale de la Santé actualise régulièrement son *International statistical Classification of Diseases and related health problems* (ICD : actuellement dans sa onzième version)[2]. Ce support offre une vue exhaustive des troubles de la santé humaine, classée par organes atteints, mécanismes impliqués, et caractérisés selon des variables cliniques, biologiques et radiologiques. Il permet de situer les pathologies, les unes par rapports aux autres, et est particulièrement utilisé pour le codage administratif des diagnostiques. Par contre, il n'a pas la nuance suffisante pour être utilisé dans la prise en charge des malades, où les tableaux cliniques ne répondent que partiellement aux définitions. Cela ouvre plusieurs réflexions comme celle de l'objectif de la classification (diagnostique, pronostique, prédictif) et de la nécessité de combiner les variables pour augmenter la précision du classement. L'intégration de multiples natures de données fait partie du raisonnement médical. Pour obtenir un diagnostic, choisir un traitement, prévoir une évolution, le médecin prend en compte des informations provenant de son interrogatoire, son examen clinique, des imageries, tests sanguins et urinaires. Nous sommes entrés dans l'ère des données omiques, dite de haut débit, et l'enjeu de classifications futures est de réussir à intégrer ces données complexes à nos critères habituels pour améliorer la compréhension des maladies humaines[3].

L'évolution des classifications des cancers du sein est un modèle à suivre. Il s'agit d'une maladie fréquente est grave parmi les plus étudiée au monde. Elle a depuis longtemps été le terrain de découverte de biomarqueurs et d'innovation thérapeutiques ciblées. C'est un chirurgien britannique, Georges Beatson (1849–1933), qui a démontré le rôle des hormones sexuelles dans la carcinogenèse en induisant la fonte tumorale après un ovariectomie[4]. L'hormonothérapie a depuis été proposée aux malades dont les tumeurs sur expriment les récepteurs de le oestrogènes (ER+) ou de la progesterone (PR+)[5]. Les classes de tumeurs ER/PR+ et ER/PR- constituent ainsi deux maladies distinctes, avec un traitement, et un pronostic qui leur est propre. Par la suite, le marqueur tumoral HER2 a permis de caractériser une nouvelle classe de cancer du sein avec la possibilité d'un traitement ciblé[6]. Ces différents marqueurs de biologies moléculaires sont encore, en combinaison avec la classification TNM, les variables utilisées en routine pour la prise en charge des malades. Plus récemment, a été

proposée une classification transcriptomique avec des classes de cancer du sein aux pronostics distincts (Figure 7)[7,8]. Ces approches ont permis d'orienter la recherche sur de nouvelles thérapeutiques[9] et d'élaborer des signatures génétiques permettant d'orienter la prise en charge en pratique courante[10,11].

La dermatite atopique (DA) est une maladie inflammatoire cutanée fréquente (Figure 8) caractérisée par des poussées d'eczéma (Figure 9) et de prurit[12–14]. Elle est marquée par une forte hétérogénéité tant sur les plans de l'âge de survenue, de l'intensité des symptômes, des comorbidités, et des thérapeutiques[15–18]. Dans ce travail, nous avons émis l'hypothèse que les différences interindividuelles entre les patients atteintes de DA pouvaient être le reflet de différents mécanismes moléculaires identifiables sur le transcriptome cutané. La physiopathologie de la DA est complexe, elle implique l'interaction entre le *monde extérieur* et son exposome[19] (Figure 10), le monde intérieur et les équilibres immunologiques[20,21], neuro-métaboliques qui y résident, en passant par l'interface cutanée superficielle, garante de l'homéostasie[22] (Figure 11). La DA s'intègre dans l'atopie, soit la tendance d'origine familiale, à l'hypersensibilité en réponse à une exposition antigénique, s'associant fréquemment à l'asthme, la rhinoconjonctivite allergique et les allergies alimentaires[23,24]. Ainsi la DA constitue la facette cutanée de cette maladie systémique qu'est l'atopie (Figure 12). Mais la DA semble d'avantage complexe que son nom le laisse paraître. Après avoir de nombreuses fois changé de noms (Table 1), l'usage veut qu'on la nomme « atopique ». Alors que toutes les DA ne sont pas IgE médiées, il serait d'avantage correcte de faire référence à l'eczéma atopique et non-atopique[25]. La DA pourrait d'avantage être considérée comme un syndrome avec une physiopathologie commune, à laquelle s'ajouterai des mécanismes spécifiques de chaque endotypes (Figure 11). La nécessité de classer la DA et ses sous-entités s'est avérée nécessaire. La première catégorisation prenait en compte le taux d'IgE circulantes et les allergies associées définissant la DA intrinsèque (non allergique à IgE normales) et extrinsèque[26,27]. Seulement, cette dichotomie révélait peu les différences interindividuelles et n'aidait pas à la prise en charge des malades. Depuis, différents endotypes de DA ont été définis en fonction de l'âge de survenue, de l'origine ethnique, ou encore du mode évolutif révélant des mécanismes sous-jacents spécifiques. Ainsi, la DA du sujet asiatique semble d'avantage liée une polarisation Th17, ce qui suggère un possible rôle bénéfique des anti-Th17[28] chez ces malades, pourtant réservés au psoriasis (Figure 14). Le recours aux données omiques devrait permettre d'aller plus en profondeur dans la compréhension de

l'hétérogénéité de la DA (i.e. au sein de l'AD du patient asiatique). Les premières thérapeutiques ciblées, le Dupilumab (anti-IL4rα)[29] et Baricitinib (anti JAK1-2)[30], montrent des résultats encourageants, mais persistent respectivement 31-52% et 50-65% de non-répondeurs. A mesure que les pistes thérapeutiques s'élargissent (Table 2), il est de plus en plus nécessaire d'identifier les biomarqueurs permettant d'orienter le choix du traitement. Pour cela, plusieurs méthodes s'offrent à nous. « Celui qui ne sais pas ce qu'il cherche ne comprendra pas ce qu'il trouve » disait Claude Bernard (1813–1878). A l'opposé, appréhender la question de la classification avec le moins d'a priori possible amène à des découvertes plus originales et inattendues. Ces méthodes supervisées et non supervisées peuvent s'opposer comme se compléter. Elles permettent d'exploiter pleinement les données omiques complexes. L'approche supervisée se base sur des annotations cliniques ou biologiques, sélectionnées selon une hypothèse médiée par la connaissance. Elle constitue l'approche privilégiée dans la DA car nécessite de plus faibles effectifs. C'est le cas des signatures transcriptomiques MADAD[31] et 89ADGES[32] qui permettent de classer les patients AD par opposition à des sujets sains ou atteints d'autres dermatoses. La supervision peut aussi porter sur les gènes, ainsi Thijs and Bakker et al[33,34] ont élaborée une méthode mixte avec une sélection restreinte de biomarqueurs, suivie d'une classification non supervisée. Jusqu'à présent, une approche purement non supervisée : du choix des variables, au clustering, n'existe pas encore. C'est ce que nous avons voulu développer dans la première partie de ce travail qui visera à définir des classes de malades selon leurs mécanismes communs. Le deuxième projet, au contraire, consistera en une approche supervisée sur un symptôme complexe. En effet, il se concentrera sur l'identification d'une signature moléculaire minimaliste permettant de classer les patients selon l'intensité de leur prurit.

Résultats

Présentation de la cohorte

En tant que membre du Consortium Microbes in Allergy and Autoimmunity Related to the Skin (MAARS), nous avons eu accès à ses données cliniques, transcriptomiques, génétiques et métagénomiques. Ce projet financé par l'Union Européen, pour 7,8 millions d'euros, entre 2011 et 2015, est à ce jour le plus grand jeu de données complexes concernant l'AD, le

psoriasis, et des sujets sains. Il n'a à ce jour, fait l'objet que de deux publications[35,36]. Nous avons décidé de nous concentrer sur la question de l'hétérogénéité de la DA, et n'avons donc pas utilisé la cohorte de psoriasis. Les patients atteints de DA répondaient aux critères de Hanifin et Rajkan[37], ne présentaient pas de comorbidité auto-immune et n'avaient pas reçus récemment de traitement locaux ou systémiques. Les contrôles ne présentaient pas d'antécédent de dermatose inflammatoire. Tous ont été biopsiés en deux sites (lésionnels et non-lésionnels pour les DA) pour les analyses transcriptomiques (Affymetrix GeneChip Whole Transcript Expression Array®) et métagénomiques, et de nombreux critères cliniques, dont l'activité de la maladie, ont été recueillis (Annexe 1). Les données omiques ont été générées de manière centralisée, selon un protocole standardisé pour éviter les biais techniques. Une phase importante d'exploration des données cliniques a permis de caractériser les populations (Table 3), de sélectionner seulement les gènes codants (Figure 5) et d'affiner les questions de recherches.

Projet 1 : Identification de nouveaux endotypes de DA à partir de données transcriptomiques

Attention : le numéro des figures fait référence à celui de l'article (partie Results).

L'apport des données omiques a permis de caractériser plus profondément les mécanismes de la DA. Pourtant, il n'existe pas encore de classification purement non supervisée de la DA à partir de transcriptome de peau. Pour répondre à cette question, nous avons utilisé les biopsies cutanées faites en peau lésée de 82 patients atteints de DA et de 117 sujets sains. Nous avons pris en compte les différences qui peuvent exister entre situations physiologiques en sélectionnant les gènes selon leurs variances. Les gènes hyper-variants spécifiques de la DA avaient une variance > 0.5 et étaient > 2 fois la variance en situation physiologique (Figure 1b). Ces 222 gènes avaient des fonctions intéressantes, proches de la physiopathologie de la DA : différenciation kératinocytaires, activité des métalloprotéases et de multiples voies immunologiques (Figure 1c). Nous avons utilisé ces gènes pour classer les échantillons de malades selon différentes méthodes de clustering non supervisées et avec nombre optimal de clusters. Nous avons choisi la méthode des *k-means,* qui divisait les échantillons en quatre groupes (Figure 2b), et montrait les meilleurs scores de Dunn, Rand et

Jaccard reflet de la robustesse du clustering (Supplementary 4). L'interprétation des différences cliniques et biologiques entre les groupes (Figure 2c) selon les tests de Kruskal-Wallis et Fisher révélaient qu'ils différaient en termes de sévérité de la DA (p=0,003) et d'importance de colonisation à Staphylococcus Aureus (SA), (p = 0,004). De manière à identifier les gènes qui était les plus importants dans le clustering nous avons effectué une analyse différentielle d'expression (*t test, p<0,05*) entre les clusters (Figure 3b). Nous avons regroupé les gènes caractéristiques d'un groupe (Supplementary 6) en 3 métagènes (Figure 3c) qui se sont révélés avoir des fonctions biologiques intéressantes (Figure 3d). En effet, le métagènes 1 (MG I), 19 gènes, caractéristique du cluster 1, était composé de cytokines de la famille de l'IL-1, comme IL-36A et IL-36G) ainsi que de gènes de destruction tissulaire. MG II, 23 gènes, caractéristique du cluster 2 avaient une activité immunorégulatrice négative avec les cytokines IL-34 et IL-37 ainsi que de reconstruction cutanée (FLG2, LOR). MG III, caractéristique du cluster 4 était composé entièrement de gènes de l'immunité lymphocytaire B. Le cluster 3 était quant à lui caractérisé par une expression moyenne de tous les métagènes. En comparant la corrélation entre l'expression des métagènes et les paramètres cliniques d'intérêt nous avons retrouvé que le l'expression du MG I, pro-inflammatoire, était fortement corrélé avec le SCORAD (R=0.48, p=4.7e-06) et la colonisation à SA (R=0.45, p=3.1e-05), alors que le MG II, anti-inflammatoire, leur était inversement corrélé (respectivement R=-0.88, p=3.5e-06 and R=-0.46, p=1.8e-05). Enfin, nous avons pu valider ces observations sur une cohorte indépendante que nous avons sélectionnée car il s'agissait de la deuxième plus importante (après la nôtre), utilisant la même technologie et avec des informations sur le SCORAD pour chaque échantillon[38].

Dans cette première partie, nous avons élaboré une classification de la DA en utilisant au maximum la structure des données, sans prendre en compte les informations cliniques et biologiques avant l'étape d'interprétation. La sélection de gènes a été élaborée de manière logique, à partir de la variance des gènes en situation physiologique et pathologique, ce qui nous a été permis par l'importance de la cohorte de contrôle. Ainsi, l'équilibre entre les métagènes pro-inflammatoires destructeurs de l'architecture cutanée, et anti-inflammatoires reconstructeurs est un phénomène jusqu'ici jamais reporté qui pourrait aider à mieux comprendre cette pathologie complexe et orienter la thérapeutique.

Projet 2: Élaboration d'une signature transcriptomique du prurit dans la dermatite atopique, en combinant les modèles statistiques et d'apprentissage machine.

Attention : le numéro des figures fait référence à celui de l'article (partie Results).

Le prurit, est la sensation désagréable qui mène au grattage. C'est un symptôme central dans la physiopathologie et le retentissement de la DA[39–41]. Ses mécanismes sont complexes et partiellement compris rendant son traitement compliqué[42,43] (Supplementary 1c). Nous avons posé l'hypothèse que les données transcriptomiques étaient par leur importante précision, à même d'aider à la compréhension du prurit. Ainsi, nous avons utilisé les transcriptomes des 82 biopsies cutanées pratiquées en zone lésées chez les patients atteints de DA, et le score visuel analogique (1-10, médian = 7), côté pour tous les patients (Figure 1). Dans un premier temps nous avons utilisé des modèles statistiques comme l'analyse différentielle, ou les corrélations de Spearman et Pearson (Figure 2). La comparaison avec un méthode mixte, alliant statistique et apprentissage machine, comme le sparse PLS[44], a pointé la nécessité de réduire préalablement la dimensionnalité pour obtenir la puissance nécessaire (Figure 3). Ainsi, en collaboration avec une équipe de Mine Paris Tech, nous avons utilisé un modèle combinant les outils statistiques et d'apprentissage machine[45,46] capable à la fois de réduire le nombre de gènes tout en optimisant la prédiction du prurit. Nous avons défini deux classes de prurit (bas < 7, haut ≥ 7) et la cohorte a été divisée en cohorte d'entrainement et de test (80%-20%). La sélection de variables a été effectuée sur la cohorte d'entrainement en combinant plusieurs techniques : l'arbre de décision, la machine a vecteur de support linéaire, XGBoosting, AdaBoost, et Lasso. Ces classifieurs ont été entrainés et validés pour distinguer les classes de prurit. Parallèlement, nous avons utilisé des modèles statistiques comme l'information mutuelle, le Chi^2, et le modèle de régression linéaire univariée. Les gènes ont été ordonner selon l'importance que leurs donnaient chaque technique et nous avons retenu ceux qui avait été retenus le plus souvent (>40%). La signature minimaliste finale a été ensuite obtenue par l'ablation des gènes dont l'exclusion ne modifiait pas la précision de la prédiction (Figure 4). Sept gènes faisaient partie de la signature finale : Heme Oxygenase 1 (HMOX1), Calcium/Calmodulin Dependent Serine Protein Kinase (CASK),

Vestigial Like Family Member 2 (VGLL2), Mannosidase Alpha Class 2A Member 1 (MAN2A1), un ARN non-codant (GPRC5D-AS1) et deux nouveaux transcrits (AC113382.1 and AL031123.1). Les classes étaient prédites sur la cohorte de test avec une balanced accuracy = 0.77, une précision = 0.86, une sensibilité = 0.67, et une spécificité = 0.88. Nous n'avons pas pu valider notre signature sur une cohorte externe indépendante car quelques gènes n'étaient pas communs aux différentes technologies. Par contre nous avons pu valider la robustesse de la méthode, en groupant deux cohortes indépendantes incluant n=70 biopsies de peaux lésées de patients avec DA. Une nouvelle signature a pu être élaborée avec des performances comparables (balanced accuracy = 0.90, une précision = 0.90, une sensibilité = 1,00, et une spécificité = 0,80), mais sans gêne en commun avec la précédente (Figure 5). Par contre, l'interprétation fonctionnelle de deux signatures a révélé des fonctions communes.

A ce jour, ce travail est le premier à utiliser une combinaison de modèle statistiques et d'apprentissage machine pour mieux comprendre le prurit dans la dermatite atopique. Il permet d'extraire les gènes les plus importants dans la prédiction, mais rend difficile l'interprétation fonctionnelle du fait de leur faible nombre. La reproductibilité de nos résultats n'a pas été vérifiée, mais l'utilisation de la même méthode sur une cohorte indépendante a montré des résultats très encourageants. Cela peut être dû à plusieurs facteurs comme les différences de recrutements, de localisation anatomique des biopsies, la différence de technologies utilisées entre les cohortes ou encore la difficulté d'évaluer par ce simple score, un symptôme aussi subjectif et multifactoriel qu'est le prurit (Supplementary 2). Les gènes sélectionnés ne faisaient pas partie de la physiopathologie connue de la DA. Au contraire, ils laissent entrevoir des perspectives nouvelles, en particulier via des thérapeutiques ciblées sur le trafic intracellulaire de vésicules.

Discussions et perspectives

Au-delà des discussions relatives aux résultats des projets, nous avons voulu développer trois axes de réflexions soulevés par cette thèse et relativement aux difficultés: 1) l'élaborer une classification, 2) d'utiliser de données déjà publiées, et 3) d'appliquer *au lit du malade* les découvertes basées sur des données complexes.

Qu'est ce qui fait une classification, comment est-elle construite et mise en pratique ? Dans le domaine des maladies humaine, les classifications actuelles tendent à corréler des données observées avec des états pathologiques pour définir des symptômes, puis des syndromes caractérisant un diagnostic. Les données omiques peinent encore à être intégrées à la clinique et la biologie standard due à l'évolution des techniques, les disparités de qualité, et la difficulté à les générer. Classer revient à simplifier un tableau complexe en concepts intelligibles. Ce *principe de parcimonie*, qui structure le rationalisme occidental, nous permet de réduire la quantité d'informations et de définir les classes (i.e. un diagnostic). Mais, alors que les données omiques nous informent sur de multiples mécanismes à la fois, cette simplification peut paraitre simpliste. La biologie des systèmes modélise les maladies comme des réseaux d'interactions complexes entre différentes states d'informations (cliniques, génétiques, environnementales…). Elle annonce ainsi comment seront classer les maladies quand nous seront capable d'intégrer toutes ces données. Une des difficultés rencontrées est la validation des classifications basées sur des données omiques. Le coût des technologies restreint souvent le nombre d'échantillons, alors que la quantité d'information par échantillon est importante. Cela implique de trouver un compromis entre la puissance statistique (donc l'atténuation du bruit de fond) et la résolution biologique. L'idéal est de pouvoir vérifier ses découvertes sur une cohorte indépendante et prospective utilisant une technologie similaire qui puisse être comparer avec une technologie de référence (e.g. *in silico* confirmée *in situ*). Le renouvellement des classifications passera par une sensibilisation des différentes communautés qui sont impliqué. Cela devra commencer par le choix d'une sémantique commune. En effet, la terminologie peut sembler floue aux non-initiés (e.g. *predire* en apprentissage machine n'a pas de valeur prospective, et se rapproche d'*estimer*) . Ces ambiguïtés sont en parties lié à la richesse des initiatives et un défaut de consensus entre les multiples disciplines impliquées.

La *Science ouverte* (*open Science*) désigne le cercle vertueux de la recherche basée sur le partage des données scientifiques. Ce projet en est un exemple. En effet, la cohorte MAARS a été constitué dans le courant des années 2010 pour aboutir, à ce jour, à deux publications[35,36]. Ces données, qui ont coutées environ 7 millions d'euros, sont devenues publiques, mais restent encore sous-exploitées. Selon les principes FAIR (Findable, Accessible, Interoperale and Reusable)[47], les informations utilisées dans un projets doivent être partagées à la communauté à leur publication. Toutefois, la réalité ne suit pas encore la théorie, et il est parfois difficile d'avoir accès à certains types de données publiées. Dans le domaine de la DA, une signature diagnostique a pu être développée[31] en utilisant la base de données GEO[48]. Certains disciplines ont créé des bases de données consultables et utilisables par tous, comme en vaccinologie[49], ou en biologie forestière[50]. Cette pratique montre toutefois certaines limites. L'utilisation de données dont on ne connait pas exactement les circonstances de génération peut entrainer des biais importants. Ce fut le cas dans notre deuxième projet où, les disparités de recrutement et de technologie, même minimes, ne nous ont pas permis de valider nos résultats. L'on doit trouver un équilibre entre la réutilisation de données existantes et la génération de nouvelles données pour ne pas que cette pratique freine l'innovation. Mais étant donné que le volume d'information biologique ne fait que croitre, ce n'est pour l'instant pas d'actualité.

Comment faire en sorte que les données complexes puissent être utiles aux malades. De nombreuses découvertes ont été faite grâce à l'apport des données génomiques. Pour ce qui est de la DA, l'étude du génome de milliers de patients a mis en évidence le rôle important du gène (FLG) de la filagrine dans le mécanisme de la maladie, définissant ainsi une classe de malade plus précoce, plus graves et plus chronique. Toutefois, le génotypage FLG n'est pas fait en routine et son résultat ne modifie pas la prise en charge du malade. L'identification de groupes, au sein d'un même diagnostique devrait avoir un impact sur le malade, mais c'est rarement le cas en général, et n'existe pas dans la DA. A ce jour, le cancer du sein est la seule maladie où, dans certains pays, une signature transcriptomique est proposée pour orienter le traitement[11]. Il est possible que la grande diversité de nature d'informations nuise à leurs interprétations et que de nouvelles approches d'analyse soient nécessaires pour aider à leur interprétation. L'apport de l'intelligence artificielle, pour l'instant inutilisée en pratique courante de soins, devrait à terme débloquer cette situation. Cela impliquera de dépasser le choc des cultures entre les disciplines. En effet, médecins et biologistes nécessitent de

comprendre pour apprécier leur décision alors que le choix opéré par l'intelligence artificielle peut parâtre obscure. Ces nouvelles approches peinent encore à s'imposer et à montrer leur supériorité face aux modèles statistiques sur lesquels notre médecine moderne est fondée[51]. De plus, un certain nombre des membres de la communauté médicale devra être initié à ces techniques, permettant de garder un regard critique et de participer à des recherches dans le domaine. En parallèle, les *data scientists* devront faire preuve de curiosité sur la discipline et discuter constamment avec les biologistes et médecins pour aboutir à un outil proche de la réalité du terrain. Dans tous les cas, pour favoriser l'applicabilité du modèle, le choix final devra rester entre les mains de l'expert.

Cette thèse, nous l'espérons, semble présenter de multiples perspectives. Elle regroupe deux travaux qui ont permis de mieux comprendre les mécanismes de la DA et de son prurit à l'échelle de l'individus. Ces méthodes pourraient être utilisées dans la partie non exploité de la cohorte concernant le psoriasis, ainsi que dans d'autres maladies complexes. A titre personnel, elle m'a permis d'appréhender une nouvelle discipline, celle de la *Science des données*, et de la mêler à mes connaissances médicales. J'espère avoir ainsi acquis un regard neuf et élargi qui me sera utile dans mes projets et ma pratique future.

INTRODUCTION

Classifying: an evolving definition of a moving concept applied to the Living and other fields

What is classifying?

A compromise between finding differences and commonalities

Classifying is interpreting data, whatever their nature may be, in order to simplify apparent complexity but without denying it. If we focus on extreme situations: at the lowest degree of resolution, everything seems homogeneous. People belong to a unique category: human beings. On the other side, if we zoom to the maximum, every person is singular and represents a full-fledged category. Strictly speaking, these two examples are not classification. To build a classification is to find a compromise between the observation scale and the desired level of information.

Classifications produce classes regrouping samples with commonalities related to an addressed question. It may refer to the origin, the present, or the future of the studied object. Thus, classifying implies defining: a *non-too-homogeneous-non-too-heterogeneous* population of samples, a scale of observation, one or several classification criteria, and a question. As in all Science, completeness is an illusion, and the ambition of the classification work must be balanced by the fact that we only ever classify samples, and current classifications are meant to be redefined by future ones.

Description vs applicability: do classifications have to be useful?

Guillaume Lecointre (1964-), a French zoologist and systematician (i.e. Science classification specialist) speaks about classification this way:

> *"Classifying is clustering objects in an ensemble. This ensemble is an argued concept containing objects with common properties. […] Classification tells something about the world while determination tree purpose is to be practical. […] It creates concepts, that do not have to be practical. […] But it could be simplified for a pedagogical reason."(ref Lecointre ENS)*

He thus opposes the purely descriptive approach, whose aim is to bring out concepts, to the finalist approach, whose aim is to answer a practical question. In other words, he announces the duality between unsupervised and supervised strategies that will frame our problematic. In human biology, and especially in the human diseases field, a choice should be made between discovering original concepts with an unbiased method and addressing pragmatic questions related to patient care issues. What is sure is that the perspective of improving patient management should not be forgotten.

A universal need for classifying

Is there any field that does not have classification needs? Is organizing knowledge through hierarchized concepts inherent to the way we think? It appears that none makes an exception. Whether it is willfully or not, our reasoning is based on concepts organized together, sometimes exclusively, sometimes not. And the manner we place the objects depends on the question's angle of attack.

An illustrative example of infinite and moving classification concerns musical genres. As musical genres influence each other, we can study their historical and geographical relations. Musical notation begun during the 14th century BC in Middle East and adopted progressively its contemporary form with the musical partition since Middle Ages. It has allowed bypassing classical evolutionary thinking, as written music have influenced the music of other places and times. And this is even more true with music recording and the possibility for musicians to travel along each with their own musical universe. Thus, many musical genres result from the evolution of others, but external influence can disrupt the chronological relationships. As an example, jazz music derived from blues, that derived from Afro-American Black slaves' music, that derives from traditional African music. But Afro-Caribbean jazz was born when Cuban percussionists arrived in the United States of America to play with jazz bands (Figure 1).

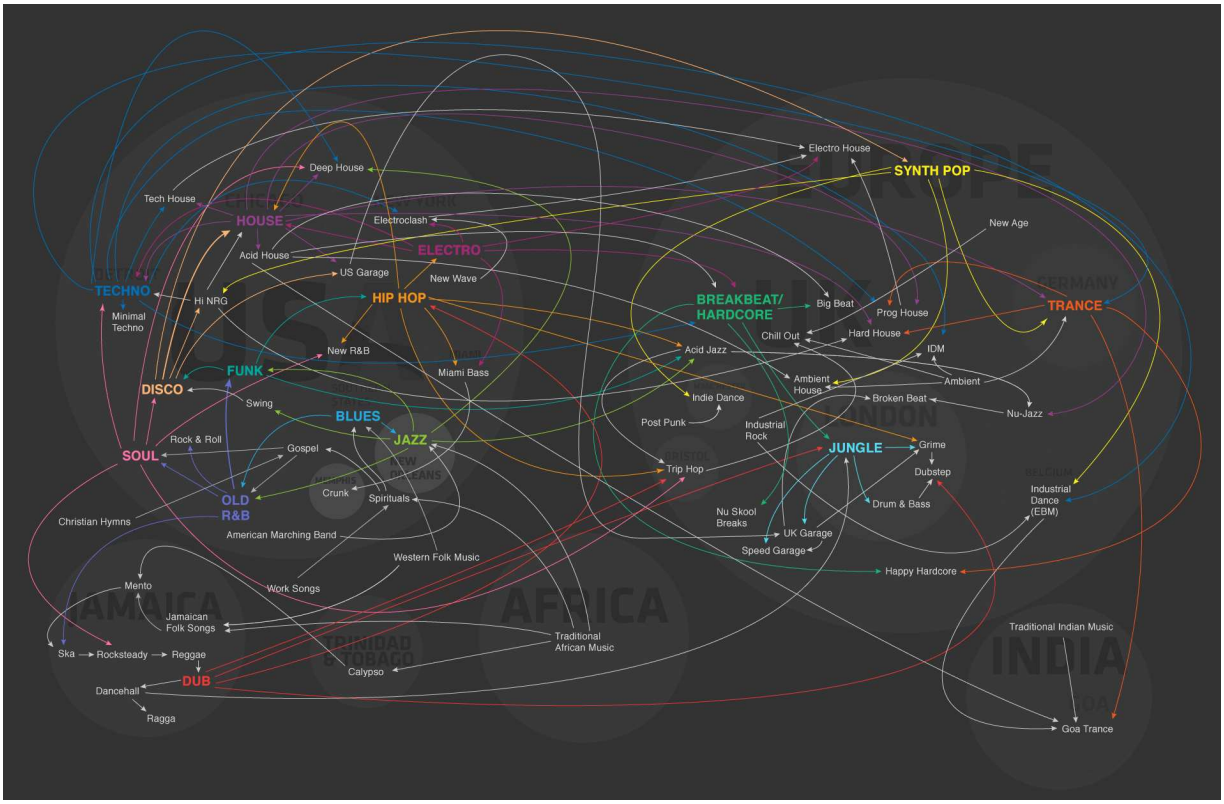


***Figure 1* How music travels: a geographical classification of musical genres**. *Arrows symbolize the influences and evolutions of musical genres on each other.*

Jazz music could have been also sub-classified according to the instrument types (e.g. with or without piano), the amplification method (e.g. acoustic, amplified), or the number of musicians (e.g. solo, trio, big band). Possibilities are infinite, especially since the criteria can be combined. The musical industry recent needs for automatic music recognition promotes artificial intelligence approaches. Nowadays, machine learning and deep learning can identify the precise musical genre and even the name of the piece without using previous criteria but dissecting the physical music structure (e.g. timbral texture, rhythmic, pitch)[52].

Why do we classify and for what purpose?

Originally: when light reveals an intriguing complexity

It is legitimate to question the beginning of our classification needs. It may be the consequence of a revelation: the one of heterogeneity. As soon as Humans took heed of the diversity surrounding them, as soon the classification need began. It started with the observation and description of the Living, and it became progressively enriched by new lights

on previously shaded areas. In biology, classifications evolved with technologies and each new tool brings its share of fresh data (Figure 2).

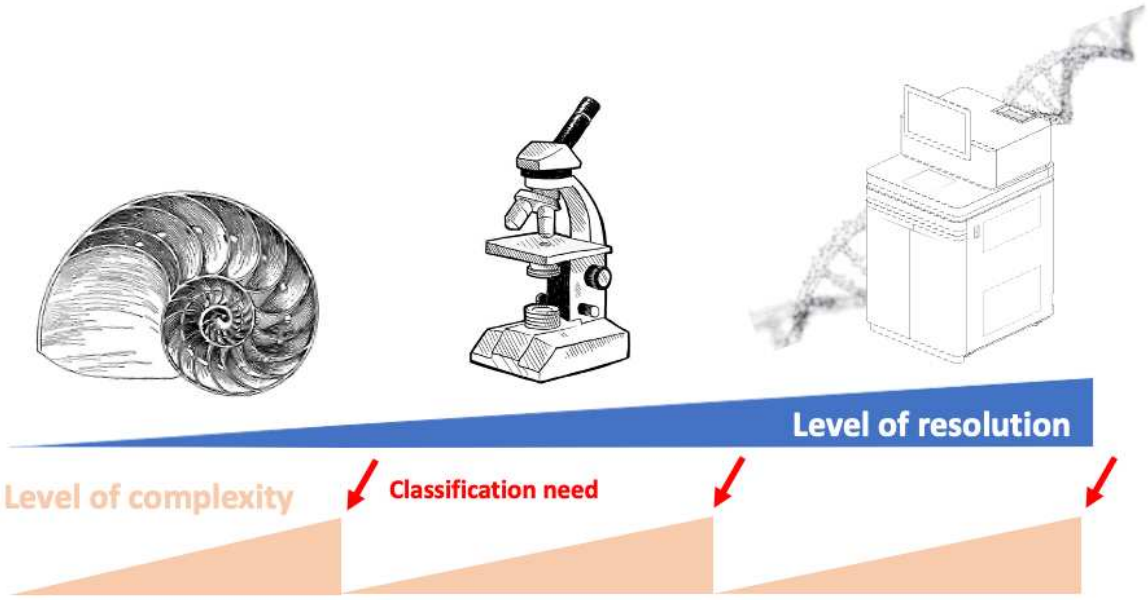


***Figure 2** **Increasing level of resolution brings forth complexity and provokes the need to classify**. Anatomical drawing of a nautilus, a microscope, and an Illumina® sequencer illustrate the evolution of technology to describe the Living. This highlighted complexity gives rise to questioning.*

An illustrative example is Comparative biology (Figure 5) which we will see more in detail later. This science uses natural variations and disparities to understand the patterns of life. It hypothesizes that conservation among species greatly assists the detection and characterization of functional elements, whereas inter-species differences are probably the best indicators of biological adaptation[1]. Its classification has been updated progressively with new tools and disciplines such as Descriptive Anatomy, Embryology, and more recently Genetics showing that new classifications can both reinforce and contradict old ones.

Aim: highlighting and simplifying diversity

It appears that being aware of the complexity of your environment bring to the need for its simplification. Our way of thinking forces us, unconsciously or actively, to define classes to organize our knowledge. Data are grouped into understandable and named concepts. Regardless of the scale, the simplification work begins with a selection of relevant features called dimensionality reduction. Selected features should be strongly informative. The ultimate goal is answering one or several specific questions from a certain attack angle (e.g.

Ordering the emergence of animal species in the course of evolution through the description of their embryogenesis or *Endotyping a disease according to treatment response and biomarker levels*).

How do we classify: is it essential to have a method?

A brief history of classification methods

Classifications have changed with technologies but also with culture and method evolution. Early stages of classification should date to the Paleolithic period (1.8 million years ago), where the hunter-gatherer *Homo erectus* probably already distinguished gender and species. These informal classification premises had an operative role concerning necessity to use plants and animals according to their properties (i.e. feeding without being poisoned)[53]. The intuitive and interpretative classification approach was formalized in the *Doctrine of signatures*. It was theorized by the Greeks Pedanius Dioscorides (40 – 90 AD) and Aelius Galenus (129 – 201 AD). It was based on an anthropomorphic interpretation of flora so that plants were classified according to their resemblances with human body parts and their therapeutic functions for those body parts. As an example, walnuts were used for head ailments because of their similarities with brain morphology.

It is still tempting to rely on intuition to simplify the complexity of the world. Regularly, we classify the new concepts coming to us without proper rules, by more or less passive aggregation to previous ones. But when it appears too confusing, an active reset is needed and a method should be used to hierarchize concepts afresh (Figure 2). This is how our Western and contemporary way of classifying developed along with Systematics classification during the XIIXth century.

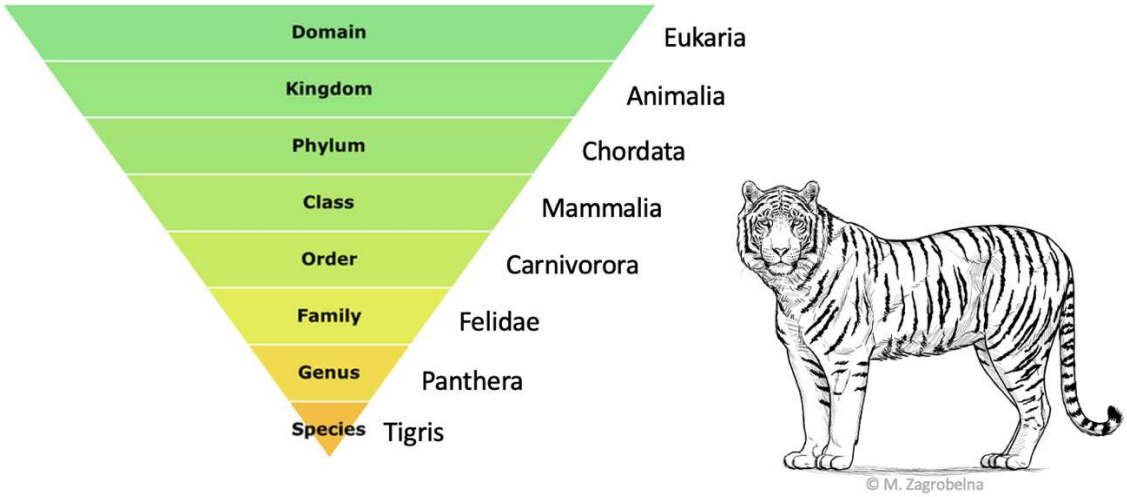


Figure 3 Tiger's position in G.Cuvier hierarchical classification of Living.

Systematic classification was theorized by Carl Linnaeus (1707 – 1778) whose ambition was to illustrate the Creator's map. He implemented a standard naming system, dividing the Living in concepts with hierarchical relationships between them. The hierarchical connections between the concepts were based on the Bernard de Jussieu (1699 – 1777 AD) principle of *Character subordination* so that a unique and constant feature is equal or even superior to inconstant ones. It was then formalized by Georges Cuvier (1769 – 1832), a French naturalist, who defined the different layers of Living as kingdoms, phyla, classes, orders, families, genera, then species (Figure 3). These rules unknowingly defined here what would be the basis of the hierarchical clustering method.

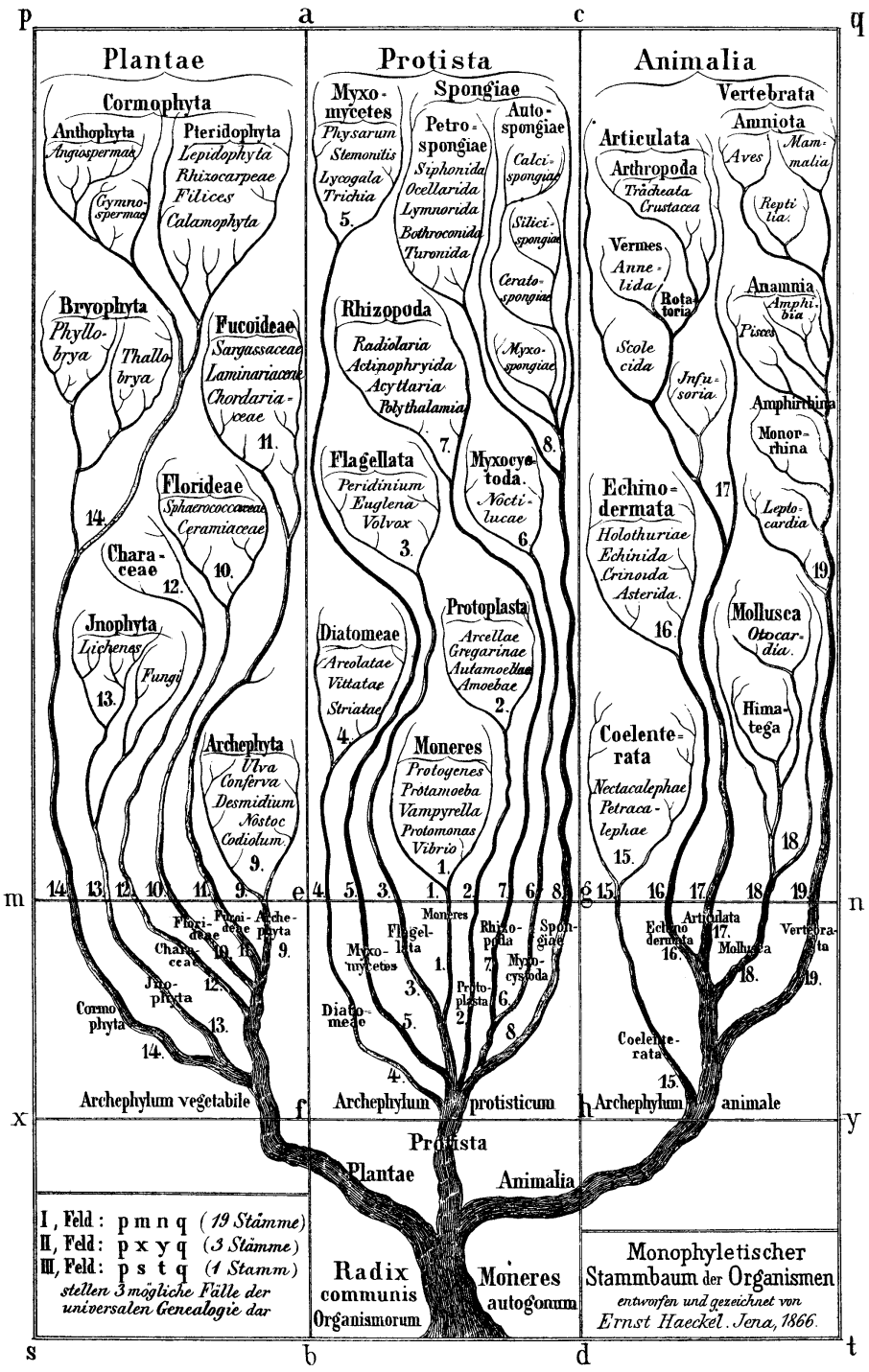


***Figure 4** Haeckel's original conception of the Tree of life (1866).*
The branch junctions symbolize the common origins between the species that are regrouped in kingdoms.

The formal framework defined by Systemics allowed Charles Darwin (1809–1882) to model his Theory of evolution. He used the shared nature of a character as a marker of evolution and introduced chronological relationships between species. These relationships are figured in a universal Tree of life (Figure 4) the ancestor of the phylogenetic tree that we

still widely use and can be stated as follow: relationships between concepts (e.g. species) are represented with two components: the branching order (illustrating group proximity) and the branch length (illustrating evolution course).

> *"Therefore, a man should examine for himself the great piles of superimposed strata, and watch the rivulets bringing down mud, and the waves wearing away the sea-cliffs, in order to comprehend something about the duration of past time, the monuments of which we see all around us"*
> Charles Darwin*, The Origin of Species (1859).*

The last disruption in classification method evolution was the Systematic phylogenetic classification published in 1950 by Emil Hans Willi Hennig (1913–1976) a German biologist. He offered another version of Systematics tenets. Less mystical, his intention was not to reflect God's vision, but to "*classify a representative sample of life"*. He was also more rigorous; he introduced the *parsimony principle* favoring the more direct relationship as the most likely. He aimed to gain reproducibility and reduce the possible hypothesis. As a concept derives from another, the first cannot help to make groups in the second. Thus, he paved the way to recent classification based on the computerization of high throughput data such as omics data (genomics, transcriptomics, *etc.*).

The example of the classification of living organisms in light of Comparative biology

The Living has always been intriguing to the eyes of Science so that classification efforts have been historically significant on this topic. The classification of living organisms based on Comparative biology is the reflection of these changes. First, it was based on anatomical homologies and differences. Pierre Belon (1517–1564) a French naturalist, discussed the comparisons between the skeleton of birds and humans. This reasoning was then adopted by anatomists and surgeons that enriched it with their observations and led to modern Comparative anatomy (Figure 5A).

Later, Karl Ernst von Baer (1792–1876) introduced Ontogenesis to go deeper than anatomical description. He hypothesized that embryos start from one or a few basic forms that are similar in different animals, as the more general characters of a large group appear

earlier in the embryo than the more special characters (e.g. embryological similarities between lizard's mandible and mammal's ear) (Figure 5B).

Recent advances in genomics have revealed even more subtle links and disparities and have led to a total reclassification of the Living. Principles are the same as previously except that gene sequences have replaced characters, defining that the less a gene is shared between individuals, the more recent it is (Figure 5C).

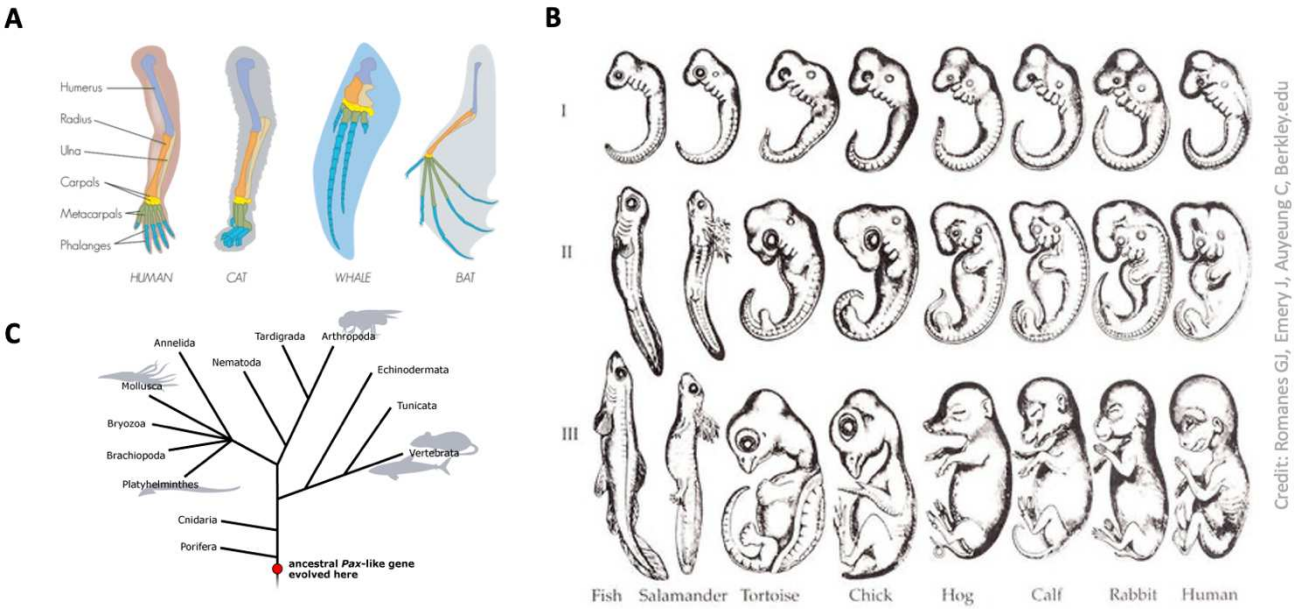


***Figure 5** Comparative biology classification has changed with technologies and disciplines evolution. A: Anatomical homologies between vertebrae. B: Embryological homologies between several living beings. C: Genetical proximity between species using Pax-like gene evolution.*

In this part, we have seen how the history of Science has shaped classification methodology toward systematism and thoroughness carried by the Classification of the Living and the Theory of evolution. It leads us to a clearer definition of what classification is. It answers a specific question (e.g. relationships between living beings), with one or several types of data (e.g. anatomical data), to define classes (e.g. species). We will see afterward that this frame is common to human disease classification issues and that contribution of high throughout data will challenge the methodological reasoning introducing unsupervised classification to generate original hypotheses.

Human disease classification: a matter of era and scale

Classifying is naming and defining new diseases based on common characteristics and mechanisms. Without being aware of it, doctors manipulate classes in patient everyday care. In making disease diagnosis, or predicting its evolution, they thus validate, through use, the existence of these concepts. But as for Living, disease classifications have moved over times.

A brief history of human disease classification

From Hippocrates to Galen: a certain sense of humors

The first formal attempt to classify human diseases can be attributed to Hippocrates (-460 – -377 BC), the Greek philosopher and physician, who has founded modern medicine. He theorized that four humors acted as vital body fluids: blood, yellow bile, phlegm, and black bile (in reference to blood in internal hemorrhage). Good health was defined as a balance of these humors while a pathological condition could be due to excess or deficiency of one of them. Hippocrates classified diseases according to these variables to categorize, name, understand the diseases, and guide their care (e.g. bloodletting for an excess of blood humor). It was a diagnostic and therapeutic classification.

Several centuries later another Greek physician, Galen (129 – 201 AD) used these humors to define the concepts of *Temperaments*, which he associated with natural elements, seasons, age, organs (Figure 6).

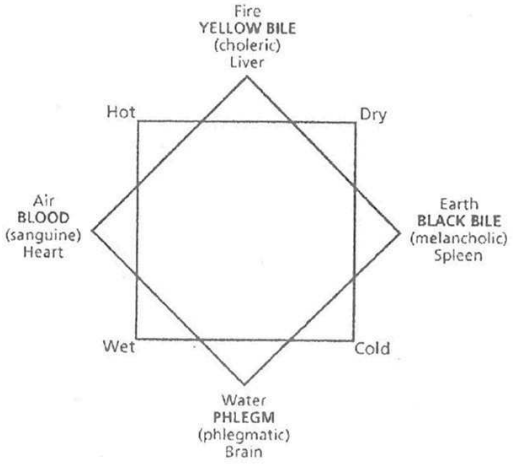


Figure 6 Galen's Humoral theory. *From Strathern et al.*

This classification of human disease using mechanisms that oriented treatment, inspired Avicenna (980 – 1037) to write The Canon of Medicine (1025), thus influencing Islamic medicine. In Western medicine, the legacy of Hippocrates and Galen had dominated medical thinking until the end of the Middle Ages. In this period Andreas Vesalius (1514 – 1564), a Flemish physician and anatomist founded modern Anatomy while the British physician and physiologist William Harvey (1578 – 1657) described for the first time the cardio-vascular system physiology. These innovative disciplines will change radically our understanding of human diseases.

Describe dispassionately to cluster better: the rise of modern classifications

It is therefore with the arrival of disciplines based on the objectification of descriptive data that classifications we still use today were established. They derived from observed correlations between pathological states and clinical syndromes. They characterized disease, establishing a nosology depending on observational skills and simple laboratory tools to define a syndromic phenotype. Like in Systematics evolutionary classification of Living follows the parsimony principle in which the simplest explanation is favored over the others. The aim is to facilitate the applicability of classification for establishing syndromic patterns that streamline the number of phenotypes to consider for the clinician. As clinicians are often dealing with atypical situations due to variable phenotypic expression, classifications might suffer from a lack of specificity when challenged in real life[3].

Large classification of human diseases typically uses clinical and standard biological symptoms. The World Health Organization updates regularly the International Statistical Classification of Diseases and Related Health Problems (ICD-11) that attempts to classify comprehensively human diseases[2]. It is used to code and classify morbidity data from the patient records for a public health survey. On a similar scale, the American Psychiatric Association wrote the Diagnostic and Statistical Manual of Mental Disorders: DSM-5[54]. It aims to describe psychiatric clinical symptoms in the most atheoretical way. Even if the way of elaborating its lists of symptoms and of weighting them in a given disease is controversial and arbitrary and use in daily practice is not the rule[55], it remains a base for mental disorder training and student formation. As we will develop later, classification applications reveal their weakness, and lead to their loss, and finally serve the reconstruction of future ones.

Data integration in the pre- and post-genomic era

As the nature of data is changing, so is the one of disease classification. Classifications with different types of data have been designed for decades. Because variables do not have the same importance, this could be a challenge for their normalization and weighting. That is what physicians do when they diagnose or predict while dealing with a variety of data.

First example, a diagnostic classification integrating data of different natures: the EULAR/PRINTO/PRES classification of inflammatory vasculitis in children[56]. Its objective was to diagnose the sub-type of vasculitis (between Henoch–Schönlein purpura, childhood polyarteritis nodosa, childhood Wegener granulomatosis, and childhood Takayasu arteritis). A group of experts first defined the scope of application (i.e. context in which classification criteria can be applied): children with suspicion of inflammatory vasculitis. Then they selected the classification variables which are biologic, radiologic, and histologic. Lastly, they tested the statistical robustness of the classifier to select the most specific and sensitive variables for diagnoses. Second example, a prognosis classification that all medical students have learned: the clinic-biological Ranson score for acute pancreatitis evolution[57]. It classes patients according to their risk of severe pancreatitis. Although still used, it has never been validated on prospective independent studies. These two examples illustrate the strengths and limits of current classification that future ones should take into consideration.

We are now at the beginning of a *post-genomic era* where molecular big data classification at the individual level puts us in a unique position to redefine human diseases with optimal accuracy, sensitivity, and specificity. The ambition is now to define precisely each patient's endophenotype offering the perspective of personalized diagnosis, prognosis, and therapeutics[3]. As we will see further with breast cancer, particular issues will arise notably dimensionality reduction, or how to deal with a huge amount of data, and biological interpretation, or how to apply these discoveries in real life.

Breast cancer as a figurehead of human disease classification

Why is breast cancer classification a major concern for the past decades?

Malignant diseases are heterogeneous entities, even within a specific organ. The most illustrative might be breast cancer. Indeed, breast cancer is a very frequent disease, with a disparity in terms of diagnosis of sub-types, therapeutic strategies, and prognosis[58]. A large research effort has been made in this field. In the PubMed database, more than 292,000 publications have already been indexed on the topic, to compare with lymphoma (approx. 256,000), prostate cancer (approx. 128,000), or colon cancer (approx. 50,000). For decades, breast cancer has been at the forefront of innovation in terms of biomarkers and treatment thus illustrate well the need for aggregating new approaches to classical ones. This leads to updated classifications that fit, at least partially, to old and validated ones while providing original information[59]. To clarify these evolutions that reinforce knowledge on breast cancer, we will detail the sequence of these discoveries.

History of breast cancer classification

First breast cancer classification occurred when the estrogen role was suggested in 1896 while Georges Beatson reduced the malignancy with bilateral oophorectomies[4]. Hormonotherapy has since been studied even before the estrogen and/or progesterone receptors (ER/PR) could be detected in situ in approximately 70% of patients[5]. In 1987 Slamon et al. discovered a new subtype where expression of Human Epidermal Growth Factor Receptor-2 (HER2) was associated with prognosis[6]. At this point breast cancer was divided into three major in-situ molecular groups: ER/PR+, HER2 amplified group, and the triple-negative group (ER-/PR-/HER2-) with distinct presentations, evolution, and prognosis. This came as a complement to the clinical and radiological Tumor-Node-Metastases (TNM) classification whose aim is to harmonize criteria between all types of cancers.

Omics data increased the level of resolution of disease pathophysiology but especially allowed to conceive it as multifactorial. Perou *et al* and Sorlie *et al* defined in 2000 and 2001 the first transcriptomic diagnosis and prognosis classification from breast cancer biopsies (Figure 5)[7,8]. This led to various molecular clusters that partially fitted with previous in situ

classes[60], progressively evolved to their specific sub-classes division[61], and were finally integrated into vaster ones[62].

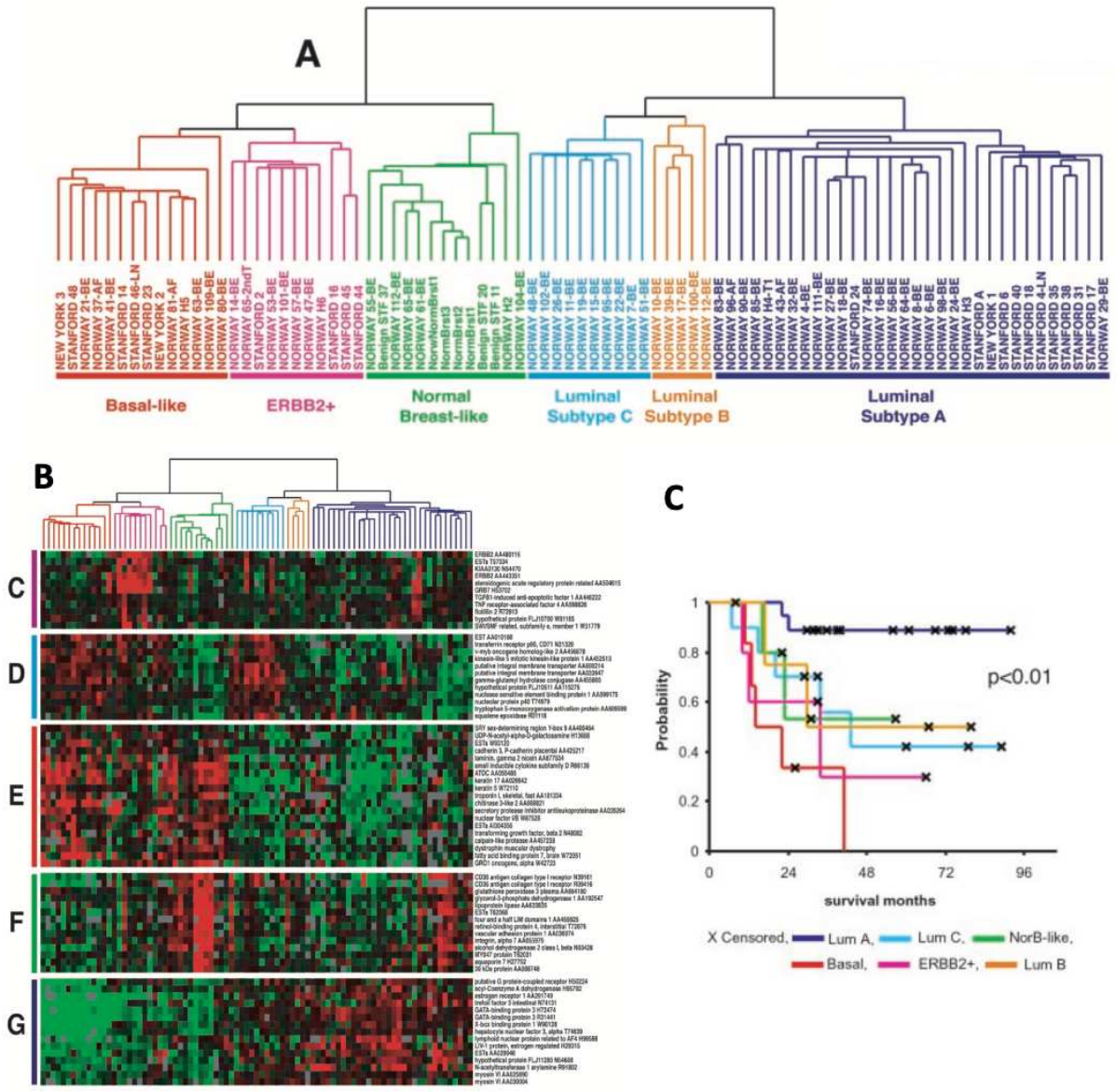


Figure 7 ***First transcriptomic classification of breast cancer****. Six breast cancer classes (A) have been identified from gene expression array (B). This classification highlights new disease mechanisms and prognosis disparity between classes (C). From Sørlie et al.*

Biological and therapeutical impact of breast cancer classification

The understanding and management of breast cancer have been revolutionized by these successive discoveries with a direct impact on patients. ER was at the same time an important pathophysiological actor and a prognosis marker. When studies of endocrine therapies were

analyzed, it became apparent that patients with tumors expressing ER could benefit from such therapy[63]. However, even if tumors do express ER or PR, not all women benefit from hormonotherapy. HER-2 amplification in tumors had a comparable value than hormonal receptors or TNM. Initially seen as a marker for bad prognosis, it became the marker of a good response to anti-HER2 therapy. The first monoclonal therapy targeting HER2 in association to chemotherapy began in 1998[64] but resulted in 40-50% of non-responder among HER2+ metastatic breast cancers.

The more we know about breast cancer, the more it appears complex, and remain women whose disease mechanisms are still poorly understood. Omics classifications offered mechanistic explanations but also prognosis and therapeutic value. As an example, poly-ADP-ribose-polymerase-1 (PARP1) expression on basal-like breast cancer suggests possible benefit for PARP-1 inhibitor therapy[9]. Taking advantage of multiple gene expression signatures, MammaPrint®, contains 70 genes that are associated with outcome. Applied to women under sixty-one years old, with lymph node-negative breast cancer and tumors smaller than 5 cm. It identified women that do not need adjuvant chemotherapy despite high clinical risk and classifies them as low and high transcriptomic risks which represent approximately half of women in each class[10]. It was designed in 2002, validated prospectively in 2016[11], and it is beginning to be used in daily practice in the United States of America, and the United Kingdom. But the recognition of its benefit is not consensual so that it has not received approval for the French market. Despite this, Mammaprint® is one of the rare transcriptomic-based prognoses classifications that has a direct impact on patients.

In the field of breast cancer, researchers and clinicians have been constantly striving to capitalize on innovation. This has made it possible to considerably increase the understanding of the disease and the outcome of the patients. It illustrates the fact that the more you understand the disease heterogeneity, the more you realize that it is broader than expected.

The need for classification in atopic dermatitis

Generalities about atopic dermatitis

Epidemiology: a worldwide frequent disease

Atopic dermatitis (AD) is one of the most frequent inflammatory skin conditions. Its lifetime prevalence has shown a worldwide increase in the past 30 years. In developed countries, it seems to plateau now at 10–20%[12–14] and continues to increase in many developing countries (Figure 8)[65].

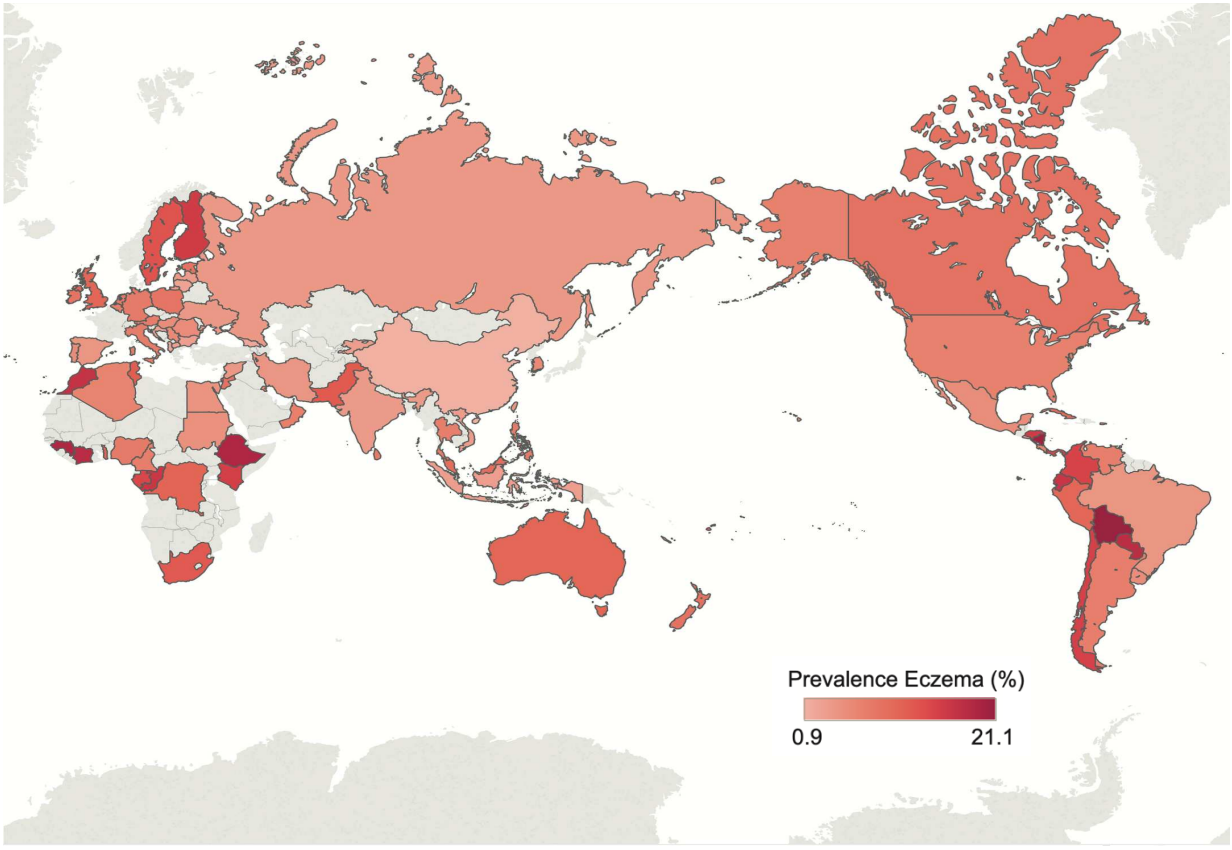


***Figure 8* AD geographical prevalence disparity.** *From Brunner et al (ref)*

Most AD begins during childhood, but persistent or adult-onset forms are not rare. In France, it touches 3,6% of the adult population (2.1 million) with 40 % of moderate (1 million) and 12.5 % of severe forms. Among them, 100 000 to 150 000 require systemic treatment and 26 500 to 42 500 are eligible for targeted biologics[30].

A rich semiology likely to blur the vision

AD is marked by chronic and recurrent episodes of pruriginous, and erythematous lesions of eczema of varying severity. Non-lesional skin is not intact, often dry and sensitive[14]. Nonetheless, AD is a multifaceted disease that can appear in multiple ways and varies according to its lesion morphology and distribution, its age of onset (Figure 7), its course, its associated symptoms, and comorbidities[15–17] . The bio-clinical heterogeneity of a disease can be hard to apprehend, while 78 clinical variables could be associated with AD depending on the geographical region, and the age[18]. In this work, we hypothesized that clinical and biological heterogeneity of AD could reflect different underlying mechanisms.

***Figure 9* Typical clinical appearance and localization of AD at different ages.** *From Weidinger et al.*

Complex pathophysiology that combines multiple mechanisms

The skin facing the external environment

AD pathophysiology is complex. It includes basal mechanisms common to all AD patients, and individual particularities. Different layers are implicated. First, the *outer world* which is the skin exposome, that combines the sum of all external factors the skin is exposed to[19], such as temperature, humidity, ultraviolet radiation, diet, pollution, water hardness, and microbiome (Figure 8).

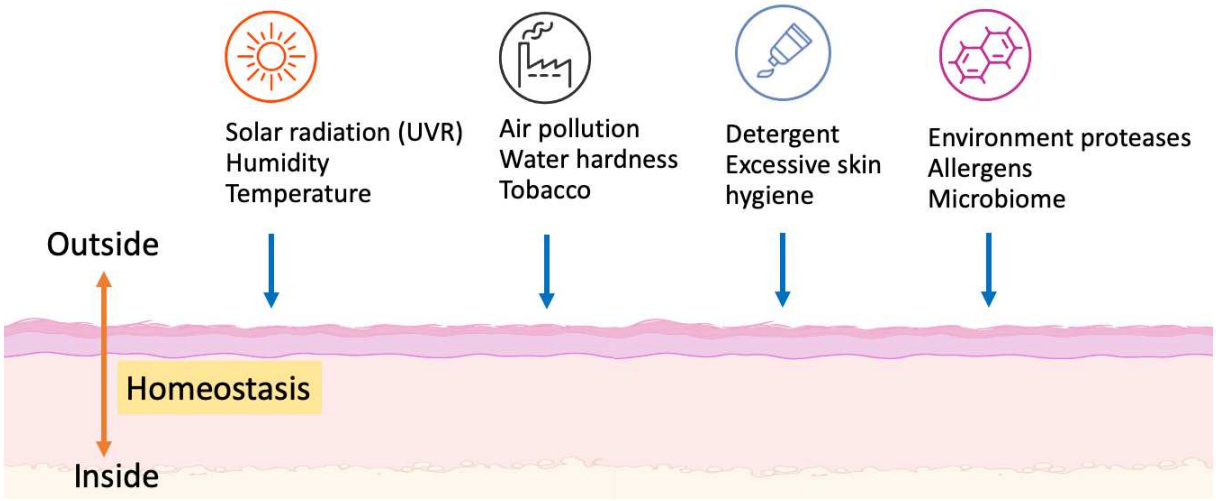


***Figure 10** The skin facing the outside world.*
Adapted from Passeron et al and Stefanovic et al. Designed with Biorender®

The semi-permeable barrier ensuring homeostasis between the inside and the outside

Second, the *interface* which is the epidermis that acts as the cutaneous barrier itself. It plays a critical role in preventing allergen and microbial penetration into the human body, skin water loss, and skin homeostasis[22]. The filaggrin, coded by the FLG gene, is one of the most important proteins of the skin architecture. Mutated in 9% of European people[66] it leads to an increased risk of developing AD of 3.12 to 4.78 for heterozygous FLG -/+ loss-of-function mutations[67].

The skin micro-environment involving immunological balance and nervous system

Third, the *inner world* which is the dermis and in particular the immune micro-environment. AD has been typically described as a Th2 disease[26], but the underlying cytokine networks appear more complex with the influences of Th1, Th17, Th22, or Tfh polarization[20,21] (Figure 11). Primary immunodeficiency associated with eczema, such as STAT3, TYK2, or DOCK8 loss-of-function mutations reinforce the immune-mediated character of the disease[68]. This immune unbalance produces a variety of mediators that activate pruritus signal conduction along nerves to be interpreted by the central nervous system.

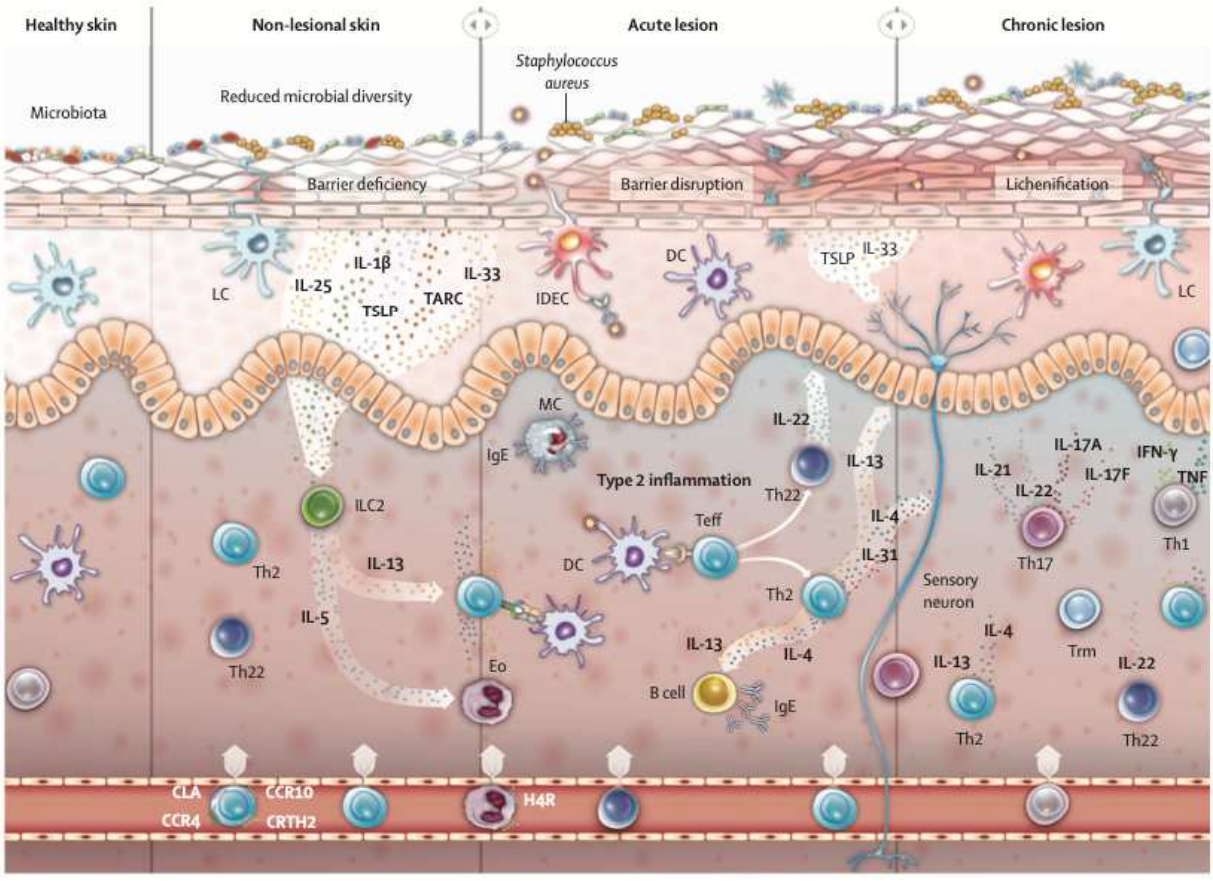


***Figure 11* Key pathophysiological changes in AD.** *From Langan et al.*

Moreover, it is the interdependencies of these 3 layers that makes the complexity of the disease pathophysiology.

Placing AD within atopy as a cutaneous facet of a systemic condition

Definition of atopy

Atopy is a personal or familial tendency due to polygenic predisposition, usually in childhood or adolescence, to become sensitized and produce IgE antibodies in response to ordinary exposures to allergens. As a consequence, these persons can develop asthma, rhinoconjunctivitis, food allergies, or atopic dermatitis[69].

The atopic march and the association of AD and other atopic comorbidities

The sequence of these diseases forms the atopic march which illustrates the systemic nature of atopy where AD would be the cutaneous facet (Figure 10). Due to defective skin barrier function, AD is usually the first step before developing other atopic diseases. This defective skin barrier is thought to allow both epidermal water loss and penetration by high molecular weight structures such as allergens, bacteria, and viruses causing hypersensitivity reactions[23]. Cross-sectional studies showed up to 62% of asthma and 29% of rhinitis in a Thai AD children cohort[24]. These values were completed in a longitudinal study showing that early-onset (<2 years old) persistent eczema was associated with an increased risk of asthma (OR 7.48), allergic rhinitis (OR 3.47), and food allergy (OR 13.4) at year 7. This strong tendency was not confirmed in late-onset AD, suggesting the possibility of alternative mechanisms in older children and adults[70].

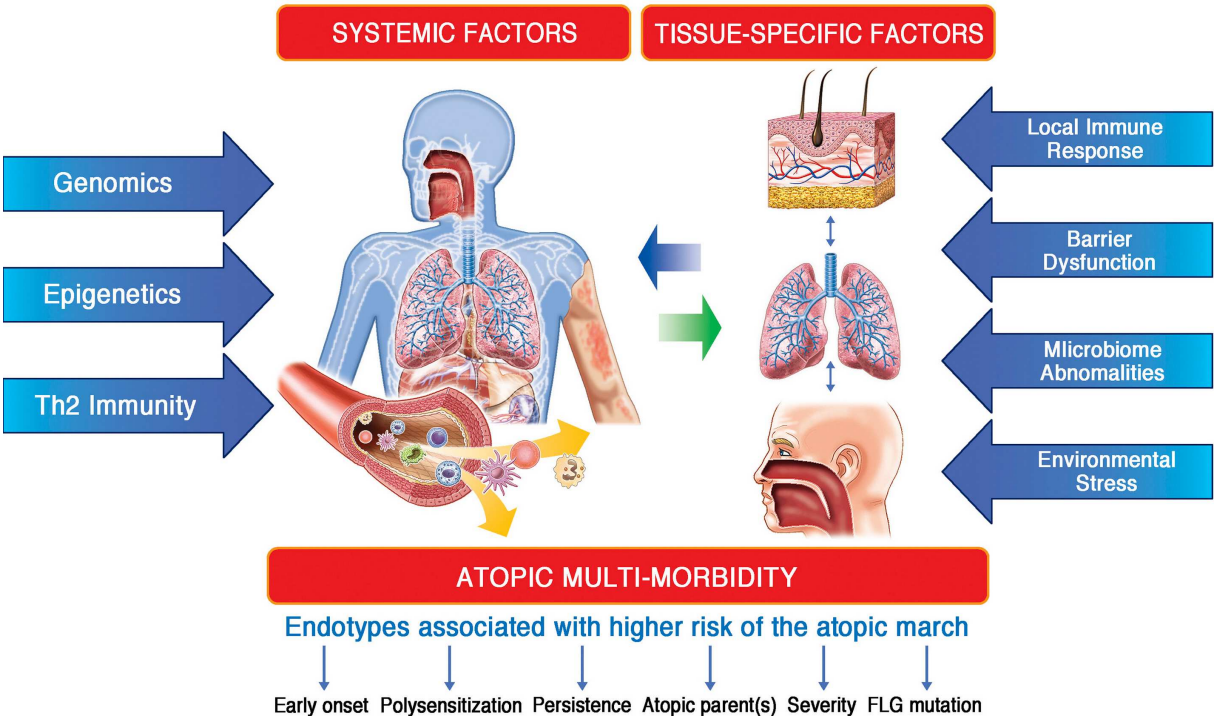


***Figure 12** The systemic nature of AD as the dermatological manifestation of atopy. From Paller et al.*

AD is not only atopic, in its semantic as in its mechanisms

One of the first questions that strikes the young researcher who begins to take an interest in AD is its name. Over time, AD has changed its name many times (Table 1). Thus, it would be more correct to refer to eczema to introduce the concepts of atopic and non-atopic eczema[25]. While eczema is not always associated with elevated circulating IgE, it is still customary to name it *atopic dermatitis* whether it is IgE-mediated or associated with asthma and rhinoconjunctivitis. Despite constant effort to define allergic disease nomenclatures, use prevails. Thus, in our study *AD* will refer both to atopic (or extrinsic AD) or non-atopic eczema (or intrinsic AD).

To a much lesser extent, AD can be associated with non-atopic comorbidities without knowing whether it is a cause, a consequence, or a shared mechanism. Studies have described a higher risk of depression and anxiety, cutaneous and extra-cutaneous infections, cardiovascular disease, inflammatory diseases (e.g. rheumatoid arthritis, inflammatory bowel disease)[17,71].

Classification challenges in AD

AD endotyping history

It is important to design well the shape of AD so that it contains entities sharing important common points while representing its heterogeneity. AD must follow the way of asthma where endotyping strategies have been applied[72]. Recently has been identified an IL-17 immunity asthma endotype that shared pathophysiological mechanism with psoriasis and then suggest original therapeutic approaches[73]. Thus, AD should be considered more as a syndrome with a base of mechanisms and manifestations common to all patients, and endophenotype specificities (Figure 11).

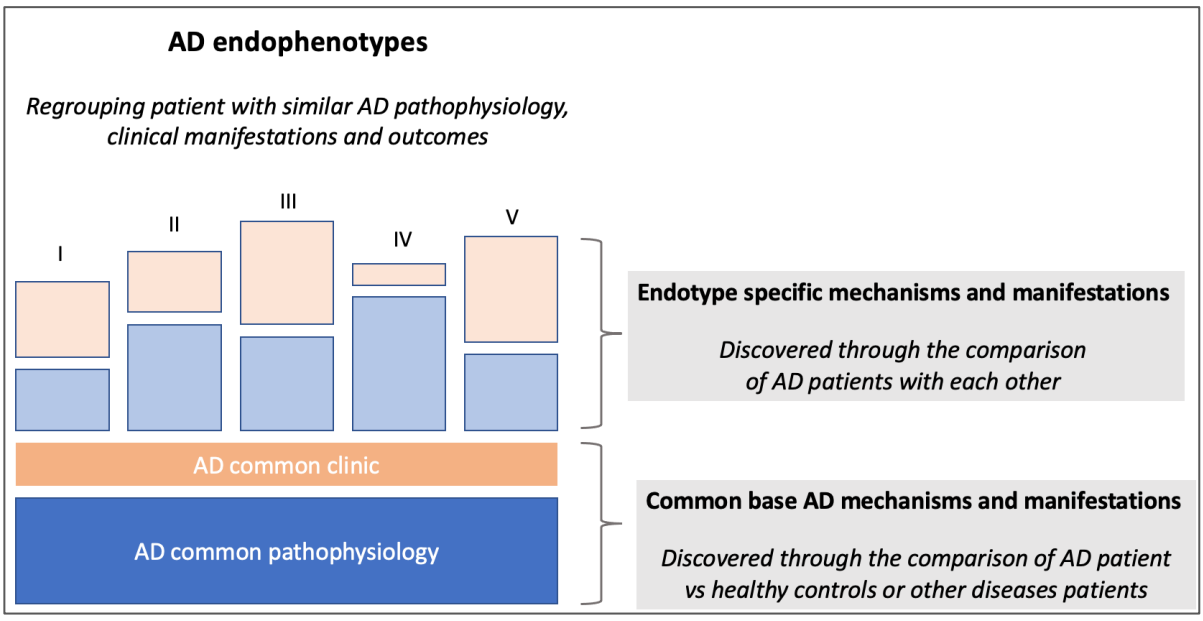


***Figure 13* Disease endophenotyping illustrates with the example of AD.** *A single disease is defined by shared mechanisms and symptoms between patients. To identity endophenotypes, patients with a similar diagnosis must be compared with each other. Adapted from Lotval et al.*

AD is a moving concept that has been renamed almost twenty times in the two past centuries (Table 1). Likely, the constant attempts by different authors to redefine and reclassify AD reflect a recognition that the clinical syndrome encompasses more than one disease entity. Opposingly, during the same period, psoriasis has remained a well-defined, non-controversial diagnosis with multiple well-recognized subtypes[74].

Prurigo diathésique	Besnier (1892)
Névrodermite diffuse	Brocq (1902)
Prurigo Besnier	Rasch (1903)
Eczéma constitutionel	Brocq (1927)
Early and late exsudative eczematoid	Rost (1928)
Eczema infantum	
Eczema flexurarum	
Neurodermatitis disseminata and pruriginosa	
Hay fever eczema–asthmatic eczema	
Neurodermitits	Rost and Marchionini (1932)
Atopic dermatitis	Wise and Sulzberger (1935)
Endogenous eczema	Korting (1954)
Neurodermitis constitution-alis sive atopice	Schnyder and Borelli (1967)
Neurodermitis atopica sive constitutionalis	Wuthrich (1983)
Pure-mixed atopic eczema	Wuthrich (1989)
Atopiforme dermatitis	Bos (1998) (2002)
Extrinsic-intrinsic atopic eczema	Wuthrich (1989)
IgE-mediated–non-IgE-mediated atopic eczema	
Allergic-nonallergic atopi eczema dermatitis syndrome	Johansson (2001)

***Table 1** Different names that have been used for AD over time.*
From Novak et al.

The first endotyping taking into account a biological signature defines the dual extrinsic (with elevated IgE level and associated atopic diseases) and intrinsic AD (with normal IgE level and no atopic comorbidities)[26,27]. These two concepts are less and less used because further discoveries revealed their overlap and did not show a significant impact on patient management. But this biologically and clinically intuitive classification showed the way to other stratification of the AD syndrome. Recent findings in AD endophenotyping, based on very diverse categories such as ethnicity, age of onset, disease chronicity, have revealing their underlying specific mechanism (Figure 12). Thus, Asian AD phenotype can be considered as a predominant Th-17 disease that could eventually benefit from Th-17 targeted treatment[28]. As treatment efficiency is still debated[75], ethnicity-based endotypes appear not to be the only contributor to individual specificities. The arrival of high throughput data allows to interpret more closely the complex pathophysiology of AD and to understand AD heterogeneity inside a specific entity (e.g. AD heterogeneity inside Asian population).

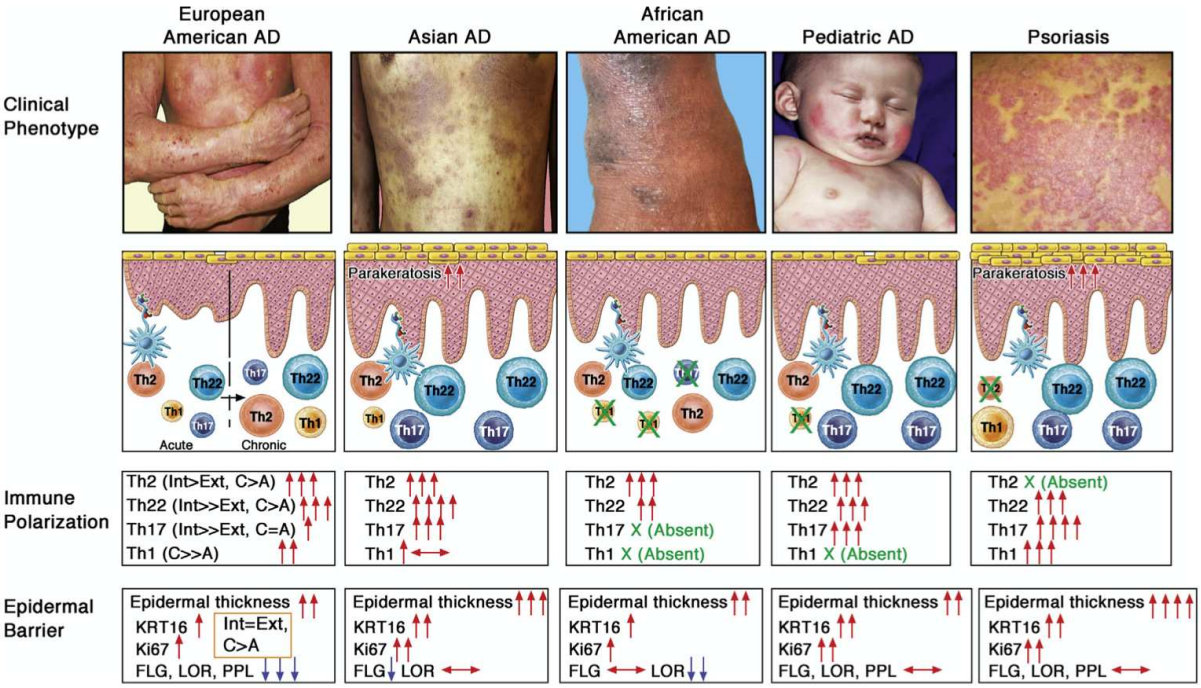


Figure 14 ***The multiplicity of AD endophenotypes****. Clinical phenotypes are associated with immune polarization and epidermal barrier characterization. Pathophysiology could vary between acute or chronic status. From Czarnowicki et al.*

AD endotyping heralds the consideration of inter-individual specificities and personalized medicine

Previous approaches have helped to distinguish between important entities. As an example, Simpson et al compared AD *colonized with SA* vs *non-colonized with SA* and distinguished two distinct endotypes based on clinical and biological. Studies that focus on deeper heterogeneity, like intra-Caucasian heterogeneity, are rarer. The more illustrative example is the AD endotyping by Thijs et al, which has been partially prospectively validated by Bakker *et al*[33,34]. In these studies, AD patients are divided into 4 classes based on a knowledge-driven list of 278 blood biomarkers revealing interesting associations. This could be the first step to identify inter-individual specificities and go beyond the *one size fit all* strategy that prevails in AD therapy.

In official recommendations, no biomarkers are used for treatment choice, and the only variables that are taken into account to guide patient care are patient age, therapeutic history, and disease severity[76]. In France, 37% of severe AD and 5,9 % of moderate AD patients have received systemic immunosuppressive drugs, mainly ciclosporin, in the past 12 months[30]. Among them, 13% have stopped systemic immunosuppressive therapies each year, for inefficiency or intolerance and will require alternative targeted therapeutics. At this step, 31 to 52% will be non-responsive to Dupilumab (anti-IL4rα)[29] and 50-65% to Baricitinib (anti-JAK1,2)[30]. Understanding individual specificities is needed to guide therapeutic orientation.

Biologic	Target	Family	Developmental Phase
Dupilumab	IL-4Ra	Th2	Approved
Lebrikizumab	IL-13	Th2	II, completed
Tralokinumab	IL-13	Th2	III, active not recruiting
Nemolizumab	IL-31Ra	Th2	III, recruiting
GBR830	OX40	Th2	II, recruiting
KHK4083	OX40	Th2	II, active not recruiting
Tezepelumab	TSLP	Th2	II, recruiting
Etokimab	IL-33	Th2	II, recruiting
PF-06817024	IL-33	Th2	I, active not recruiting
REGN3500	IL-33	Th2	II, recruiting
Fezakinumab	IL-22	Th22	II, completed
Bermekimab	IL-1a	Innate Immunity	II, completed
Ustekinumab	IL-12/23p40	Th17/IL-23	II, completed
Risankizumab	IL-23p19/IL-23A	Th17/IL-23	II, recruiting
Secukinumab	IL-17A	Th17/IL-23	II, completed
Omalizumab	IgE	IgE	IV, completed
Ligelizumab	IgE	IgE	II, completed

***Table 2** Current and ongoing biologics in clinical development for atopic dermatitis. From Wu et al.*

The past years have brought significant progress in the current treatment of AD in the form of biological treatment. Cytokines and other mediators that play an important role in the pathogenesis of skin inflammation have become a target for new forms of therapy[77,78].

Supervised vs unsupervised strategy

"Who does not know what he is looking for, does not understand what he finds", Claude Bernard (1813-1878) said. On the other way, exploration with the lesser *a priori* could lead to original discoveries. This echoes *Meno's paradox*, a Socratic dialogue by Plato (428-347 BC), and questioned our openness to likely consider original discoveries. Our ability to discover AD classes could depend on a knowledge-driven hypothesis (e.g. racial or age disparities). This so-called supervised approach is complementary to an unsupervised approach which relies on

data structure independently to annotations and thus generates original hypothesis. Due to their size and high level of resolution, high throughput data can be approached in these two ways.

The supervised approach needs clinical or biological annotations and knowledge expertise to formulate a question. Although some analysis steps can use unsupervised technics, so far, the AD classification effort has been mainly supervised. Using skin transcriptomic data, MADAD[31] and 89ADGES[32] allowed to classify AD patients, psoriasis patients, and healthy controls in their pre-definite diagnostic class. This highlighted AD mechanisms that acted as molecular signatures of AD, in comparison to non-AD mechanisms. It thus defined the common pathophysiology of AD. To zoom on AD heterogeneity, Thijs and Bakker et al followed a mixed approach made with a supervised step of feature selection by choosing known relevant blood biomarkers. Then they stratified patients with unsupervised clustering to identify original associations between their proteins of interest. So far, a purely data-driven AD classification is still lacking.

In this work, we wanted to develop both ways of classifying. In the first part of our results, we designed an unsupervised analysis from the initial step of feature selection to the final clustering. We finally interpreted clinically and biologically the different clusters with the ambition to discover previously unknown mechanisms. On the other hand, in the second part of our results, we conducted a supervised analysis to better understand pruritus, combining statistical and machine learning models, thus identifying exciting novel actors of this complex syndrome.

RESULTS

Cohort presentation

MAARS consortium

As a total beginner in complex data analysis, I spent a lot of time getting to know my datasets at the early stages of my thesis. The aim was to identify their strengths and weaknesses to be able to ask a research question that they could effectively respond to. High throughput data generation is a long process that requires several steps that cannot be done by a unique actor. That is why sequential quality control is needed, preferably by technical experts.

The Microbes in Allergy and Autoimmunity Related to the Skin (MAARS) project has been funded by the European Union to the value of 7.8 million euros between April 2011 and March 2015. This collaborative project implied ten different clinical and research teams from seven distinct countries. The overall goal was to unravel the inflammatory pathways during host-pathogen interactions which may trigger allergic or autoimmune inflammation using atopic dermatitis (a surrogate for allergic diseases) and psoriasis (a surrogate for autoimmune diseases) as disease models. To this end, multiscale characterisations have been generated such as clinical, metagenomic, transcriptomic, and genomic data. Quality and homogeneity were assessed and confirmed in a pilot study before starting the major sampling collection.

So far, this effort has resulted in only two published papers:

1) *Microbe-host interplay in atopic dermatitis and psoriasis*, Fyhrquist et al. Nature communications, Oct. 2019[35]
2) *Microbial and transcriptional differences elucidate atopic dermatitis heterogeneity across skin sites*, Ottman et al. *Allergy* Oct. 2020[36]

Here, we chose to assess an original problematic from these underused high-quality data. We did not consider psoriasis samples in order to focus on AD lesional skin and healthy controls non-lesional skin. As we were dealing with transcriptomic classification issues, we

only considered *Staphylococcus aureus* colonization, which are the main bacteria implicated in AD, and excluded the whole of metagenomic data.

Method

Methods of data generation and quality control steps are presented synthetically in results parts. We take the time here to develop the different aspects that may have influenced our questions and the ways of answering them

Subject recruitment

Patients were recruited in 3 European dermatological centers: London, Dusseldorf, Helsinki. The MAARS cohort was composed of 91 AD adult patients. The diagnosis was made by a dermatologist according to the Hanifin and Rajka criteria (Figure 13A)[37].

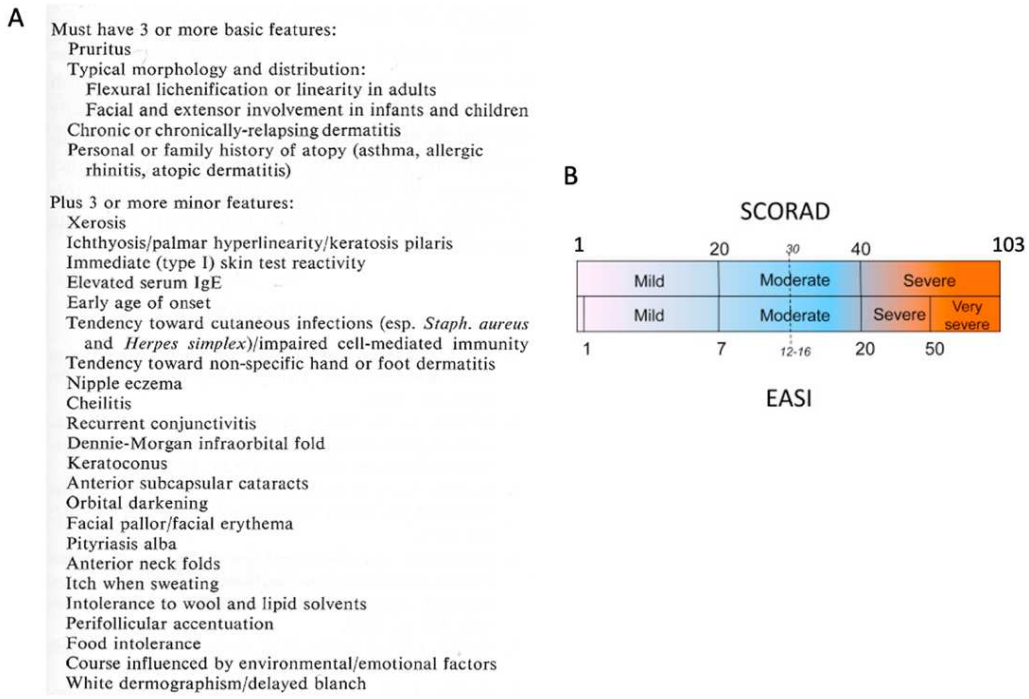


***Figure 3* A: Hanifin and Rajka crieria in the original article**. *From Hanifin et al.*

B: SCORAD-based the severity level. *Equivalence on EASI score.*

AD patients had moderate-to-severe chronic disease with a SCORAD score greater or equal to 25[79] (Figure 13B, Annex 1). In the same 3 centers, were recruited 126 controls that did not have significant comorbidity or treatment.

Exclusion criteria for the AD cohort were as followed:

- Concomitant autoimmune disease
- Personal or familial history of psoriasis
- Use of the following therapeutics
    - within 2 weeks: systemic antibiotics, topical steroids on the biopsy site
    - within 12 weeks: systemic immunosuppressive drugs, systemic biologic agent, phototherapy

Ethical aspect

The study was approved by the appropriate local Institutional Review Boards (University of Helsinki, Dnro 91/13/03/00/2011; Heinrich Heine University Düsseldorf, 3647/2011; King's College London, 11/H0802/6. All subjects provided written informed consent before participation.

Sampling

As a standardized sampling procedure was applied, skin biopsy sites were left untreated, and cleaned with antibacterial Dove soap for 2 weeks before. Washing was avoided for the last 24 hours before sampling. Active disease parts of the skin were selected for a skin swab and a 6 mm biopsy punch with local anesthesia. All were localized in the upper back and tight posterior area (areas II and III on Fig. 14A) and matched for AD lesional and healthy controls.

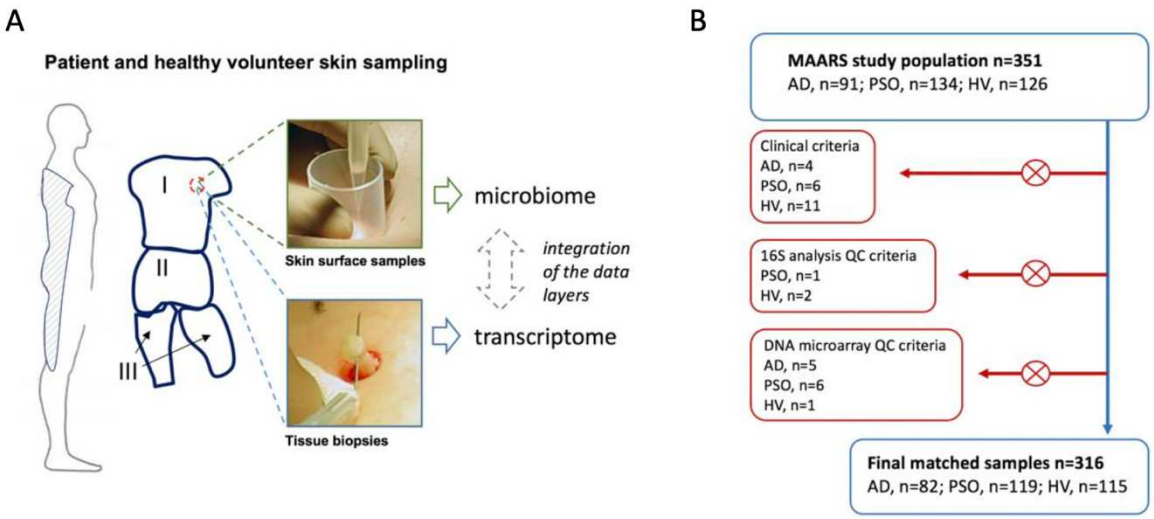


Figure 4 A : Standardized sampling procedure. B : Clinical and technical quality control steps.

Biological data generation and quality control

Biological data generation involved multiples standardized steps implying different members of the consortium. Each step had its pipeline of data generation and quality controls (Figure 14B), and can be divided as follow.

Metagenomics were generated and analyzed following these steps

1) DNA extraction Pathogen lysis and QIAamp UCP pathogen Mini Kit (Qiagen®) according to the manufacturer's instructions.
2) 16 rRNA gene amplification with RT-PCR GradeWater (Life technology®)
3) The output from each sample was further processed in QIIME (Quantitative Insights Into Microbial Ecology) and OTU taxonomies were assigned from the Greengenes Database Consortium.

Microarray transcriptional profiles were generated and pre-processed following these steps

1) RNA extraction with RNeasy Fibrous Tissue Mini kit (Qiagen®)
2) 100 ng of total RNA were amplified according to Affymetrix protocols (Affymetrix GeneChip Whole Transcript Expression Array®)
3) RNA hybridization in Affymetrix Gene ST 2.1® 96 plates, after several quality control steps

- Multichannel Nanodrop (Thermofisher®) was used to normalize RNA quantity
- QUIxcel DNA electrophoresis (Quiagen®) was used to ensure RNA quality
- Universal RNA and bacterial spikes were added to total RNA to check the quality of the hybridization procedure

Microarrays were then normalized using the Robust Multi-Array Average (RMA) approach and technical batch effects were removed.

Exploration of clinical data

Quality control and variable preprocessing

Clinical data were collected from a standard form completed by the examinator, and containing several dozens of close-ended questions and semi-open-ended questions (Annex 1). Thus, we had several clinical datasets at our disposal. One was common to all patients (AD and HV) and contained only 33 variables. One was only for AD patients and contained 296 variables, and one was only for HV and contained 166 variables. The major part of these variables concerned technical and identification information. It was important to consider them for labeling trackability but they were not clinically relevant. We then focused on the clinical and biological information.

Even if the quality control step of the clinical data is less standardized than for its biological counterpart, it remains an important milestone. We chose not to include in our analyses the variables with many missing values (e.g. “AD treatment history”). When necessary, we created variables by combining several others (e.g. “All kinds of allergy” combined the different allergies-related variables for one patient).

Identifying relevant variables for analysis and interpretation

Prior to analysis and interpretation, it was important to identify potential bias especially comparing AD patients to healthy controls (Table 3)

	AD patients	Controls
Individuals (nb)	82	113
Samples (nb)	82	213
Age mean (yo) [min-max]	44 [20-83]	35 [19-77]
Gender (F/M)	36/46	44/69

Table 3 Main demographical characteristic of AD and controls.

Due to the significant overrepresentation of the white skin population (> 90%), we would not be able to show differences due to ethnic endotypes. In like manner, our methods should take into account the gender disparity between AD and controls. Gender was also unbalanced between groups so we designed an analysis pipeline to neutralize the gender bias by excluding genes located on the sex chromosomes in our unsupervised feature selection process. For the supervised part of the classification, we chose to focus on the clinically relevant pruritus score because it was well annotated in the whole cohort, using a good precision score from 1 to 10.

Exploration of transcriptomic data

Expression values

Unlike RNA sequencing, another transcriptomics technology, which yields very sparse data (i.e. that contains many zeros), microarray data are continuous values strictly greater than 0. They have been log2 normalized so that their distributions range from 0 to 13. As expected, expression values were distributed according to a bi-modal gaussian. The first gaussian distribution, on the left, was the result of background noise, while the second gaussian distribution represented the biological signal (Figure 15).

Selection of coding genes

To limit the influence of non-biologically relevant signals on our analyses we chose to exclude non-coding genes. To do so, we used the ExpFilter function from the EMA R package (version 1.4.7) and we defined a low expressed gene as being inferior to 4 in all samples. From

32633 probes, we obtained a 22635 coding genes matrix, corresponding approximatively to the order of magnitude of known biologically relevant genes in *Homo sapiens*. Thereafter, we used this smaller matrix in all our analyses (Figure 15).

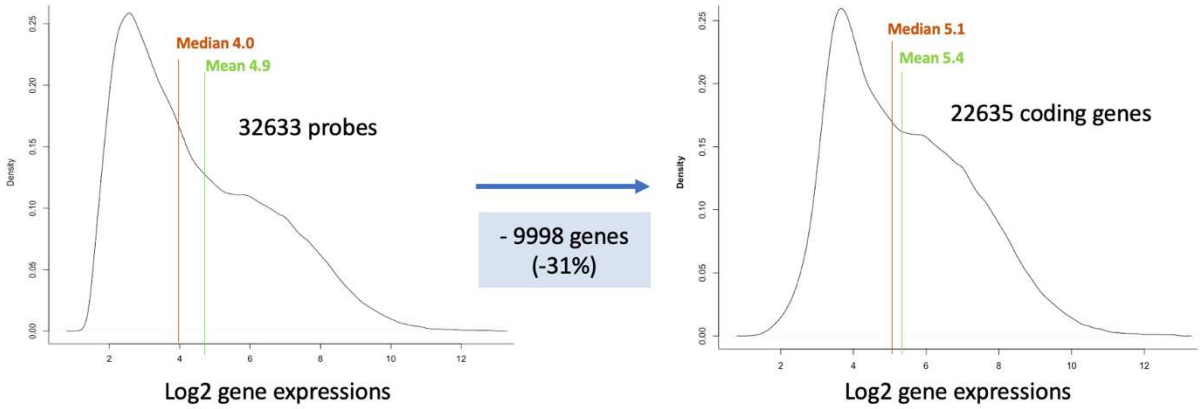


***Figure 5** Gene expressions distribution before and after coding gene filtering.*

PART 1: Unsupervised classification

What does skin transcriptome tell about AD heterogeneity?

Approach rational

For the first part of the thesis, we wanted to take advantage of the cohort size which is so far the largest skin transcriptomic cohort published for AD patients and healthy controls. It provided us with enough statistical power for making possible an original unsupervised approach. To be less influenced by clinical and biological annotations, we started from the formal mathematical structure of the data. Then we advanced as far as possible in the analysis, remaining blind to a possible biological significance. We have therefore selected our genes, determined the optimal number of clusters and the optimal clustering method following this way. In the last part of the study, we began to use the data annotation to define how our AD clusters differed biologically and clinically.

Results announcement

Our initial hypothesis was that disease heterogeneity could be revealed using the AD-specific hyper variables genes. We designed an intuitive and logical way to select genes using physiological and pathological conditions. This helped us to reduce dramatically the data dimensionality and concentrate the biological information. We discovered four AD clusters that had distinct biological and clinical properties and we succeeded to validate this finding on an independent cohort.

Graphical abstract

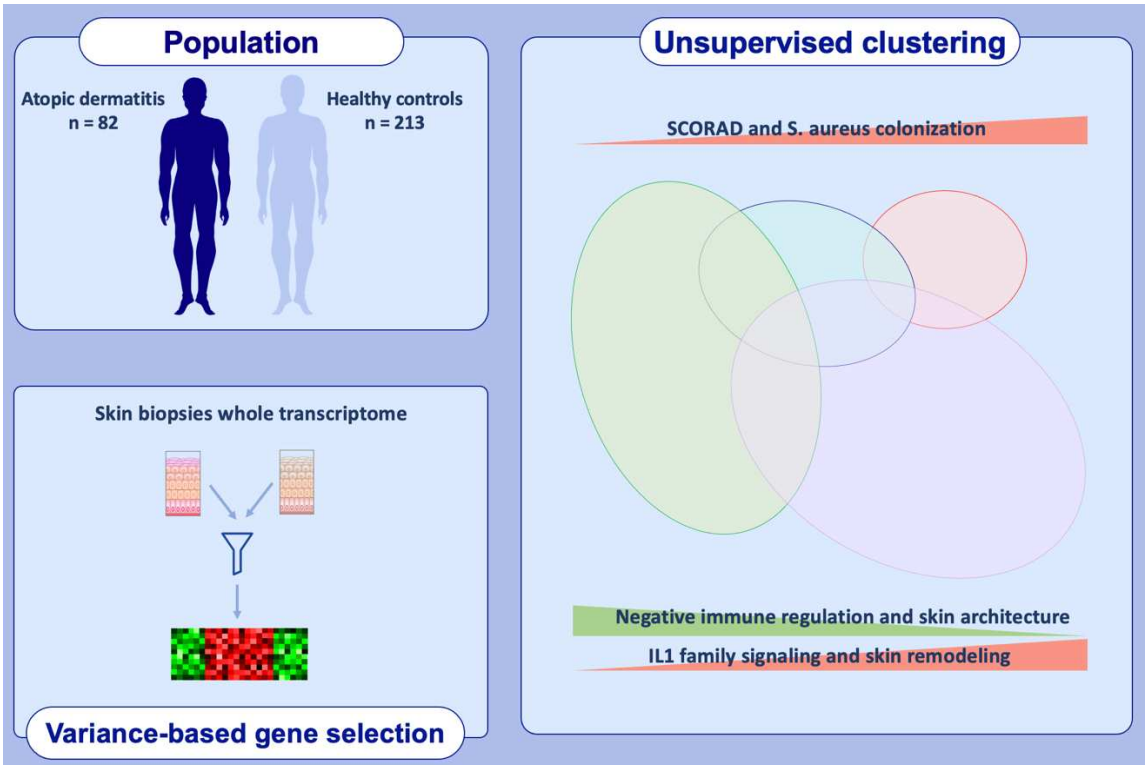


Figure 6 Graphical abstract of study design and results

Transcriptome-based identification of novel endotypes in adult atopic dermatitis

Authors

Alain Lefèvre-Utile[1,2], Melissa Saichi[1], Péter Oláh[3,4], Marc Delord[5], Bernhard Homey[3], Vassili Soumelis[1,6] ; MAARS consortium

Affiliations

[1] Université de Paris, Inserm, U976 HIPI Unit, Institut de Recherche Saint-Louis, F-75010, Paris, France

[2] Assistance Publique-Hôpitaux de Paris (APHP), General Pediatrics and Pediatric Emergency Department, Jean Verdier Hospital, Bondy, France

[3] Department of Dermatology, University of Duesseldorf, Duesseldorf, Germany

[4] Department of Dermatology, Venereology, and Oncodermatology, Medical Faculty, University of Pécs, Hungary.

[5] Clinical Research Center, Centre Hospitalier de Versailles, Le Chesnay, France

[6] Assistance Publique-Hôpitaux de Paris (AP-HP), Laboratoire d'Immunologie, Hôpital Saint-Louis, F-75010, Paris, France

MAARS Consortium

Juha Kere[a,b], Francesca Levi-Schaffer[c], Dario Greco[d,e], Noora Ottman[f], Jonathan Baker[g], Björn Andersson[h], Mauricio Barrientos-Somarribas[h], Stefanie Prast-Nielsen[i], Lukas Wisgrill[j], Sophia Tsoka[k], Nanna Fyhrquist[fl], Harri Alenius[fl], Helen Alexander[g], Jens M. Schröder[m], Frank O. Nestle[g], Antti Lauerma[n], Philippe Hupé[o], Annamari Ranki[n]

MAARS Consortium affiliations

[a] Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

[b] Stem Cells and Metabolism Research Program, Folkhälsan Research Institute, University of Helsinki, Helsinki, Finland

[c] Pharmacology and Experimental Therapeutics Unit, Institute for Drug Research, School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

[d] Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[e] Institute of Biotechnology, University of Helsinki, Helsinki, Finland
[f] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[g] St John's Institute of Dermatology, Faculty of Medicine and Life Sciences, Kings College London, London, United Kingdom
[h] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden
[i] Department of Microbiology, Tumour and Cell Biology, Karolinska Institutet, Stockholm, Sweden
[j] Division of Neonatology, Pediatric Intensive Care and Neuropediatrics, Comprehensive Center for Pediatrics, Department of Pediatrics and Adolescent Medicine; Medical University of Vienna, Austria
[k] Department of Informatics, Faculty of Natural and Mathematical Sciences, Kings College London, London, United Kingdom
[l] Department of Bacteriology and Immunology, Medicum, University of Helsinki, Helsinki, Finland
[m] Department of Dermatology, University Hospital Schleswig-Holstein, Kiel, Germany
[n] Department of Dermatology, Allergology, and Venerology, University of Helsinki, Helsinki University Hospital, Inflammation Centre, Helsinki, Finland
[o] INSERM U900, CNRS UMR144, Institut Curie, Mine Paris Tech, Paris, France

Disclosure statement

We declare no competing interests.

Keywords

Atopic dermatitis, adult, skin, transcriptome, endophenotype, clustering, IL-34, IL-36, IL-37

Acknowledgments

This work was supported by the *Fondation pour la Recherche Médicale*, grant FDM201806006187, to Alain Lefèvre-Utile. The MAARS Consortium has received funding from the FP7/2007–2013 (Grant 261366). Thanks to Nuala Mooney, Andrei Zinovyev, and Elise Amblard for their critical reviews.

Abstract

Atopic dermatitis (AD) is a frequent and heterogeneous inflammatory skin disease, for which personalized medicine remains a challenge. High-throughput approaches have recently improved understanding of the complex pathophysiology of AD. However, a purely data-driven AD classification is still lacking. To address this question, we applied an original unsupervised approach on the largest available AD transcriptome dataset (n=82) and healthy (n=213) skin samples (MAARS dataset). Taking into account pathological and physiological state, variance-based filtering revealed 222 AD-specific hyper-variable genes that efficiently classified the AD samples into 4 clinically and biologically distinct clusters. Comparison of gene expression between clusters identified 3 sets of upregulated genes used to derive metagenes (MG): MG-I (19 genes) was associated with IL-1 family signaling (including IL-36A and 36G) & skin remodeling, MG-II (23 genes) with negative immune regulation (including IL-34 and 37) & skin architecture, and MG-III (17 genes) with B lymphocyte immunity. Sample clusters differed in disease severity (p=0.003) and *S. aureus* (SA) colonization (p=0.004). Cluster 1 contained the most severe AD samples, highest SA colonization, and overexpressed MG-I. Cluster 2 was characterized by less severe AD samples, low SA colonization, and high MG-II expression. Cluster 3 included mild AD samples, mild SA colonization, and low expression of all MGs. Cluster 4 had the same clinical features as cluster 3 but had hyper-expression of MG-3. Last, we successfully validated our method and results in an independent cohort of samples. Our study revealed unrecognized AD endotypes with specific underlying biological pathways, which highlight novel pathophysiological mechanisms. These data provide new insights to establish novel personalized treatment strategies.

Introduction

Atopic dermatitis (AD), i.e. atopic eczema, is an inflammatory skin condition characterized by chronic and recurrent episodes of pruriginous, and erythematous lesions of varying severity[12]. Depending on the geographical region, AD can affect up to 20% of children and 5% of adults[80,81]. AD is clinically very heterogeneous in terms of symptoms, severity, comorbidities, and response to therapy.

AD pathophysiology is multifaceted, involving several biological and organizational levels, such as dermo-epidermal architecture, water and lipid metabolism, and skin immunity[13]. These interact with the patient's genetic background as well as multiple environmental factors like the microbiome, radiations, or allergens[19]. AD has been typically described as a Th2 disease[26], but the underlying cytokine networks appear more complex with inter-relationship of Th1, Th17, Th22, or Tfh polarization[20,21]. Notably, IL-36 which is known to be involved in psoriasis and Th17 immunity[82,83], and negative immunoregulatory IL-34 and IL-37[84–86], may also contribute to AD pathophysiology.

The marked clinical and biological heterogeneity of AD has suggested the existence of different disease endotypes, defined as patient subtypes with shared pathobiological mechanisms[87]. Since the description of extrinsic and intrinsic AD[26,27], several other endophenotypes have been identified, mainly based on the age of disease onset, duration of symptoms, and patient ethnicity[28]. Selected biomarkers, including serum IgE levels, blood eosinophils, or panels of pre-defined cytokines, have been applied to design supervised approaches allowing stratification of patient groups[34,88] without reaching a consensus regarding their clinical applications and utility.

Large-scale and omics approaches are increasingly used because they offer the possibility of characterizing complex diseases based on a diversity of biological variables and pathways. Previous AD studies have identified diagnostic molecular signatures from skin transcriptome[31,32], and endotypes from a knowledge-driven set of blood biomarkers[33,34]. However, a purely data-driven skin transcriptomic classification of AD is still lacking.

In this study, we sought to establish a systematic, unbiased strategy for patient classification and characterization of AD heterogeneity. We exploited the largest available AD and normal skin transcriptomic datasets, and validate our results on an independent cohort.

Our work provides new insights into AD heterogeneity and suggests how unprecedented potential strategies for personalized treatments could be designed.

Material and methods

Study cohort

The data were obtained from the MAARS Consortium[35], whose dataset is publicly available on the Array Express interface (E-MTAB-8149). Patient recruitment and data generation methodologies are comprehensively described in Fyhrquist et al publication[35]. Briefly, AD patients and healthy controls have been recruited in three European Dermatology departments, after provided written informed consent under institutional review board-approved protocols. All AD patients met the Hanifin and Rajka criteria[37]. Sampling and data generation occurred between 2012 and 2013. A vast amount of clinical features was collected, including disease severity AD scoring (SCORAD)[79] (Annex 1). A 6 mm punch biopsy was performed in lesional skin of AD patients and at two different sites for healthy controls. Bulk transcriptomic analysis was performed after mRNA extraction with Affymetrix GeneChip® Whole Transcript Expression Arrays. SA colonization was determined by 16s rRNA amplification and sequencing from a non-invasive skin swab made on the lesional skin biopsy site.

Validation cohort

To assess the reproducibility and robustness of both our unsupervised approach and the main findings of our study, we applied the same analysis protocol on an independent dataset. To reduce technological and technical biases, we sourced independent cohorts using a comparable transcriptomic technology and with available annotation on disease severity. Among the total number of pre-selected transcriptomic cohorts (n=48), only six studies met the above criteria (Suppl 8). We selected the largest one, with n=51 lesional skin sample, and whose methodology is described in Guttman *et al* study[38]. Bulk transcriptomic data were generated using Affymetrix Human U133Plus 2.0® gene arrays. The expression matrix was

downloaded through the Gene expression omnibus (GEO) interface (GSE130588) using *GEOquery* package[89] (version 2.51.1).

Expression array preprocessing

The MAARS dataset was loaded and analyzed with *R* language (version 3.6.0) on the *R Studio* interface (version 1.2.1335). As an exploratory step, we projected the dataset using Principal Component Analysis and performed several clustering methods. We discarded the sample *MAARS_3_070_03* as a potential mislabeled outlier. After quality control steps, 213 samples from the 113 healthy controls were pooled with AD samples. Thus, 82 AD lesional and 213 healthy skin samples were used for further analyses

Variance-based feature selection

Expression variance was computed for each gene across AD lesional and healthy control samples. Subsequently, a ratio of the corresponding variance between the two populations (AD lesional variance / healthy controls variance) was calculated for each gene. We included genes that were highly variable and specific of AD lesional skin: variance >0.5 (corresponding approximatively to the 5th percentile of all variances), and ratio >2. We further term the selected genes from this step as *AD-specific hyper-variable genes* and use them for subsequent steps.

Clustering method and optimal number of clusters estimation

Unsupervised clustering algorithms, such as *k*-means, hierarchical clustering, k-Nearest Neighbors (*k*-NN) followed by Minimal Spanning Tree (MST) partitioning, were applied on the expression matrix, including only AD samples (n=82). The optimal number of clusters to consider for downstream analysis was determined by the elbow method, average silhouette width, and MST. To assess the clustering efficiency, we calculated the clustering purity with the Dunn index, Rand index, and Jaccard similarity coefficient, and chose the method which scored the best in the different metrics.

Metagene construction

Differential analysis, using Student test and Bonferroni correction, was computed between patient clusters to identify the set of genes characteristic of each sample group. Unique and upregulated genes were further pooled to construct *Metagenes* (MG) using pairwise correlation. MG expressions correspond to the expression's mean of their respective genes. Clinical and biological parameters were compared using the Kruskal-Wallis test for continuous variables and the Fisher test for binary variables. Sample clusters were then characterized using the most significant features and metagene expressions. Correlation analyses between MG, features, and AD clusters were performed with non-parametric Spearman or parametric Pearson statistics, depending on the nature of the data.

Functional enrichment

Pathway enrichment analysis and functional annotation of the metagenes were performed first using g:Profiler web server (https://biit.cs.ut.ee/gprofiler/gost) with Kyoto Encyclopedia of Genes and Genomes database (KEGG)[90] and then Cytoscape (version 3.8.0)[91] with ClueGO extension (version 2.5.6)[92] using gene ontology (GO) biological process. An extensive manual curation of the literature through PubMed search was carried out to ensure a complete functional annotation of genes identified by non-supervised analysis.

Statistics and data visualization

Statistics were conducted with *stats* (version 3.6.0) package and *Rquery* function. Figures were generated with *ggplot2* package (version 3.3.1). Statistical differences in clinical data were considered significant at an adjusted P-value of less than 0.05 (Benjamini and Hochberg correction). At the transcriptomic level, the adjusted P-value below 0.05 (Bonferroni correction) was required. For correlation analysis, $r > 0.3$ and $p < 0.05$ were considered significant.

Results

Variance-based gene selection revealed pathways relevant to AD pathophysiology

To identify clinically and biologically relevant AD endotypes, we exploited a large and multicentric transcriptomic dataset (Suppl 1a) containing 82 AD samples and 213 healthy skin control samples from 113 healthy volunteers[35]. The control cohort was composed of 69 females and 44 males with a mean age of 35 years [min-max: 19-77]. We have chosen to keep all healthy samples to maximize statistical power. AD cohort contained different levels of AD severity since it included mild (n=6), moderate (n=39), and severe (n=37) AD, with a mean SCORAD of 52 [min-max: 17-89]. It included 36 females and 46 males. The mean age was 44 years [min-max: 20-83]. A more than 90% of AD patients and controls were Caucasian. All patients were recruited after a therapeutic wash-out of 3 months for systemic immunosuppressive drugs, and 2 weeks for topical steroids.

Transcriptomic profiles were obtained using Affymetrix protocols, which enabled the detection of 22,635 coding genes. We designed a feature selection strategy based on the premise that hyper-variable genes in AD samples would contain the most discriminatory information for the subsequent definition of patient clusters (Fig 1a). We applied a cut-off on variance values (> 0.5) to filter out low variable genes that resulted in a selection of 575 “hyper-variable genes” (Fig 1b). Next, we considered that hyper-variable genes in both AD and healthy states could represent physiologically hyper-variable genes. Such genes would also vary across healthy skin and their corresponding ratio (Variance in AD/ Variance in controls) would oscillate around 1, such as gender-related genes. Hence, we applied an additional filter based on a variance ratio >2, to select highly variable genes solely in the AD state, that would represent “AD-specific hyper-variable genes”. As expected, XIST and DGAT2L6, which are involved in gender representation, are two of the most variable genes in AD lesional and control skin with respective variances of 11.6 vs 11.3, and 7.2 vs 6.7. Genes not associated with sex chromosomes may also vary among healthy skin samples for physiological reasons. This was the case for KRTAP4-12 and PM20D1 with respective variances of 5.4 vs 8.6, and 6.7 vs 6.6. We reasoned that such genes would not inform on the disease-specific variability that should form the basis for characterizing putative AD clusters. Overall, the application of these

two filters on the initial dataset identified 222 AD-specific hyper-variable genes (Fig 1b, Suppl 2).

We exploited several databases such as the GO biological process (Fig 1c) and the KEGG pathways (Suppl 1b) for functional interpretation and biological relevance of these 222 genes. AD-specific hyper-variable genes were enriched in functions such as "keratinocyte differentiation", "type-I interferon signaling pathways", "lymphocyte-mediated immunity" (Fig 1c). Most of these pathways were consistent with previously described mechanisms of AD onset and development, confirming that the selected genes were representative of AD pathophysiology. This was a first step that revealed the biological relevance of the AD-specific hyper-variable genes.

AD-specific hyper-variable genes identify AD clusters with distinct clinical features

The 222 AD-specific hyper-variable genes were used to cluster the AD samples (Fig 2a). First, we estimated the optimal number of clusters and performed clustering through different methods such as the elbow method coupled with *k*-means (Fig 2b), MST partitioning coupled with *k*-NN (Suppl 3a), and the average silhouette width coupled with hierarchical clustering (Suppl 3b-c). All approaches indicated four clusters. We compared the different clustering methods based on the intra-cluster metric Dunn index, and the inter-cluster metrics Rand index and Jaccard similarity coefficient. We selected *k*-means clustering because of its highest score in all metrics (Suppl 4). *k*-means clustering classified samples in four AD clusters of n=12, 23, 29, and 18 samples, respectively. Multidimensional scaling (MDS) showed the embedding of previously defined clusters (Fig 2b).

Second, we asked whether clinical, biological, bacteriological, or genetic features were different between those four AD clusters (Fig 2c). We found that the SCORAD[79] and the 16s rRNA quantification of skin SA colonization differed in at least one cluster compared to the others (p=0.003, and p=0.004, respectively) (Suppl 5a). Other variables, such as association with other allergic diseases (p=0.85), gender (p=0.80), pediatric age at disease onset (p=0.20), age (p=0.98) and raised IgE levels (p=0.56) (Fig 2c, Suppl 5b) did not show significant variation among clusters.

Construction of metagenes and identification of their specific biological functions

Considering the 222 AD-specific hyper-variable genes, and the 4 AD clusters, we sought to identify the genes with the best ability to discriminate between clusters (Fig 3a).

We performed a sequential differential expression analysis of one AD cluster versus all the others. It revealed 3 sets of genes since one cluster did not contain any specific gene (Suppl 6a). In total, 62 differentially expressed genes were either up- or down-regulated, and some were shared between several gene sets (Fig 3b). To get the strongest positive signal, we considered only the 59 unique and upregulated genes. We then carried out a correlation analysis, and grouped genes that had the same behavior into metagenes (Fig 3c). We thereby defined three metagenes: MG-I, MG-II, and MG-III containing respectively 19, 23, and 17 genes (Suppl 6b).

Next, we wanted to functionally characterize those three metagenes. GO biological process analysis showed enrichment in *keratinization* and *innate immune pathways* for MG-I, *complement activation,* and *humoral response* for MG-III. Probably due to the low number of genes and their apparent heterogeneity, functional enrichment analysis did not identify a significant association for MG-II. Hence, we turned to manual curation and functional annotation through an extensive PubMed search for each gene in each metagene (Fig 3d). This approach revealed more specific information about each gene and allowed grouping genes into relevant biological families. MG-I had a strong enrichment in IL-1 family signaling (DEFB4A, IL36A, IL36G, PI3, PLA2G4D, S100A7A, S100A12), and epidermal proliferation and differentiation (EPGN, KRT6A, KRT6C, AKR1B10, SPRR2B). MG-II was enriched in negative immune regulation (CLEC2A (=PILAR), IL34, IL37, SLC46A2), skin architecture (FLG2, LOR, OGN), and epidermal homeostasis (BTC, GJB4, GSTA3, KRT77, UGT3A2). Finally, MG-III was entirely composed of B lymphocyte immunity-related genes (IGLs, IGHs, IGKs). In total, we could characterize the predominant molecular functions of the 59 most important genes in our AD clusters.

Multilayer characterization of AD clusters

We compared the metagene expression distribution across AD clusters (Fig 3b, Suppl 7). Hyper-expression of MG-I was associated with AD cluster 1 (p=6.0e-14), MG-II with AD cluster

2 (p=1.4e-11), and MG-III with AD cluster 4 (p=1.2e-08). AD cluster 3 was the only one with a mild expression of all metagenes. Differential expression analysis between metagenes confirmed that they could act as discriminating variables allowing the classification of AD samples and characterization of AD clusters at the biological level.

Then we analyzed the correlation of metagene expression levels with the clinical and biological features that differed between groups (Fig 4a). This revealed a positive correlation of MG-I with the SCORAD (R=0.48, p=4.7e-06), and with SA colonization (R=0.45, p=3.1e-05), and a strong inverse-correlation between MG-II and those 2 parameters (R=-0.88, p=3.5e-06 and R=-0.46, p=1.8e-05, respectively). In contrast, MG-III was not correlated with any of these parameters (Fig 4b). In conclusion, the expression level of *IL-1 family signaling & skin remodeling* MG-I was associated with greater clinical severity and SA colonization, whereas the *Negative immune regulation & skin architecture* MG-II was associated with milder forms of the disease.

Validation on an independent cohort

We set out to validate the main results of this study, which is the contrasting association of the *IL-1 family signaling & skin remodeling* MG-I, and the *Negative immune regulation and skin architecture* MG-II with disease severity. To do so, we screened independent datasets from Affymetrix® microarray AD skin studies that included annotations about disease severity. Six cohorts with 12 to 51 lesional AD skin samples met these criteria in the GEO database (Suppl 8). We selected the largest dataset[38] containing 51 mild-to-severe adult Caucasian AD samples. We used our reference AD-specific hyper-variable genes (Suppl 2) with the same analysis pipeline by clustering the patient samples in four groups using the *k*-means method. We then built metagenes with the same genes as in our study (Suppl 6b) to compare the metagene expression and the AD severity among clusters (Fig 4c).

SCORAD significantly differed between AD clusters (p = 0,04), as well as for the 3 metagene expressions (with respectively p = 5.9e-06, 1.1e-06, 0.004). As in our cohort, the most severe AD cluster was associated with MG-I hyper-expression, and the less severe AD cluster was associated with the hyper-expression of MG-II. Finally, we calculated the correlation of metagene expression and disease severity (Fig 4d) and confirmed the significant positive association of *IL-1 family signaling & skin remodeling* MG-I with disease severity (R=0.35,

p=0.013) and the significant negative association of *Negative immune regulation and skin architecture* MG-II with disease severity (R=-0.42, p=0.0019).

This analysis, based on the largest independent dataset available, validates the existence of four AD clusters characterized by different disease severity and underlying mechanisms.

Discussion

The question of *how to evaluate and interpret human disease heterogeneity?* is difficult to assess. Since the beginning of the high-throughput data era, the duality between supervised and unsupervised approaches prevails. Practically, both strategies are complementary. As AD is a frequent and complex disease, multiple approaches to data collection and interpretation have been developed. Because they are less difficult to design, supervised approaches are mainly used. The idea is to compare different patients' groups with pre-defined clinical or biological characteristics with the scope of identifying specific pathological mechanisms. Thus, AD disease signatures as MADAD (Meta-Analyses Derived Atopic Dermatitis)[31] and 89ADGES (89 Atopic Dermatitis Gene Expression Signature)[32], have been robustly designed comparing AD patients to healthy skin transcriptomes. In AD and other complex diseases, such as systemic lupus erythematosus, supervised approaches on disease severity score were able to highlight particular pathways[93–95]. Unsupervised approaches are more rarely used and consist in using data structure to generate new hypotheses independently of clinical and biological annotations. They require a higher number of samples compared to supervised methods to maximize their statistical robustness. We decided to design a purely data-driven unsupervised strategy using the largest AD and controls microarray cohort. Our method is novel and therefore able to reveal original findings. It also consolidates previous discoveries because of its complementarity to published studies[84,96,97].

Feature selection is an essential step in complex data analysis. It consists in reducing data dimension and complexity by removing redundant and irrelevant features to reduce noise and improve classification performance. Differential expression and machine learning wrapping are the most commonly used method in the supervised approach[98]. As genes with high expression variance could drive phenotypical diversity, we hypothesized that their study is important for the success of unsupervised approaches. The variance could be used directly to

select genes[7,99] or through variance-based methods like principal component analysis (PCA)[100]. Considering that important gene expression differences have been reported between healthy individual[101] we elaborate an original strategy of gene selection using their variabilities in AD versus physiological context.

From the description of its extrinsic and intrinsic forms[26,27], AD has been extensively sub-divided and many endophenotypes have been identified, based on the age of disease onset, duration of symptoms, or patient ethnicity[28]. Recent studies discovered[33] and validated[34] prospectively the existence of four AD endotypes using unsupervised clustering based on a knowledge-driven set of blood biomarkers. These are efficient strategies for endotyping patients using low invasive and cost-effective procedures. Our study differs from the above ones on two main points. First, we used skin transcriptomic data considering that the skin microenvironment would be the more representative of AD development mechanisms. Second, we did not follow preconceived ideas for selecting the genes of interest to be able to identify original genes and new mechanisms.

While down-regulation of IL-34 and IL-37 expressions have already been described in AD in comparison to controls[84,97], the role of IL-37 as a negative immune regulator[102], especially in a Th2 context has just been reported[85]. Also, IL-34 can lead to a pro- or anti-inflammatory state, depending on the context[86], and has been recently described as inversely-correlated with AD severity[94]. FLG2 and LOR are part of the epidermal differentiation complex, they act as important mediators of skin architecture. As a reflection of skin reconstruction defects, they are also downregulated in AD in comparison to healthy skin[103]. In our study, IL-34 and IL-37 are part of MG-II that groups skin architecture genes such as FLG2, LOR, OGN, and skin homeostasis genes. We showed that MG-II expression is strongly inversely-correlated with AD severity and is the functional signature of the less severe AD endotype. This supports the notion that their co-expression, as anti-inflammatory and skin reconstruction signals may be responsible for a protective effect.

The more severe SCORAD and SA colonized AD endotype was characterized by the up-regulation of MG-I that was enriched in IL-1 family signaling, epidermal proliferation, and differentiation. This metagene shows a high level of overlap with the molecular signature of

psoriasis, in particular Th17 immunity[104]. Although Th17 polarization is often described as downregulated in AD compared with psoriasis[105], it could be upregulated according to the AD endotype[106] and participate in AD development mechanisms[107]. As part of Th17 polarization, upregulated IL-36 pathway is responsible for a mendelian monogenic psoriasis[108], besides IL-36G and IL-36A have been described as upregulated in AD[96] where it seemed to play a role in SA induced hypersensitivity[109]. In contrast with MG-II, pro-inflammatory IL-1 signaling is grouped with genes of keratinocyte proliferation (such as KRT6A and C) and protease activity (such as PRSS22 and PI15) which could reveal a disorganized tissue response. Similarly to asthma, where a Th17 endotype has been recently described[73], our findings could suggest the introduction of IL-17 and IL-36 blockers into the AD therapeutic arsenal, representing an important step towards personalized medicine.

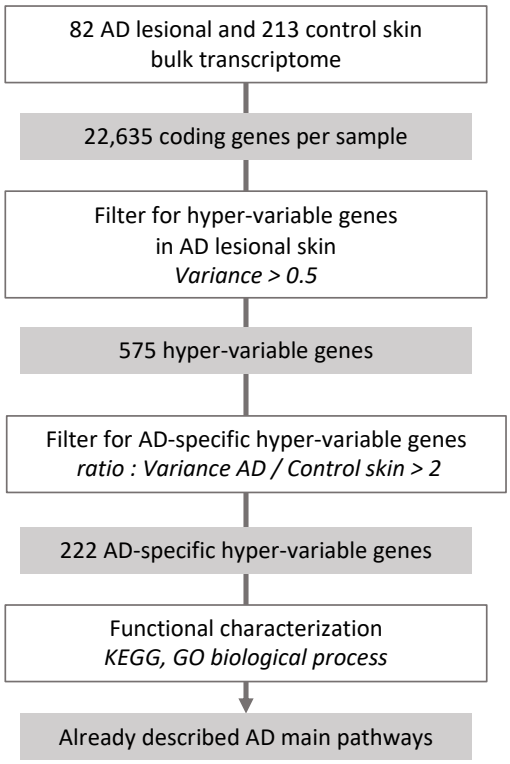
The use of patient stratification based on transcriptomic data in daily care is still very rare. The validation step of those data is critical and should be carried out in each study, at least on independent sample groups and ideally on prospective cohorts. However, prospective validation of a transcriptomic signature could be a long process. Breast cancer is currently the only situation in which a transcriptomic signature is routinely used in patient management[10,11]. In our study, we have decided to validate our main biological findings taking advantage of an independent dataset that was generated with a comparable transcriptomic technology. The questions of the method and the observation scale remain. Although RNA sequencing offers a better resolution than microarray, the latter is the most used in the AD field and remains robust for bulk transcriptomic analyses[110]. In addition, classification studies require large bulk transcriptome cohorts. Those are limited by the fact that sampling AD patients and control volunteers involve invasive skin biopsies that are not common practice in AD management, especially in children. This could be addressed by implementing recent non-invasive and validated methods such as skin tape strippings[84,94,111]. In addition to innovative sampling techniques, newer high-throughput approaches could allow the detection of gene expression with higher resolution. Indeed, the first studies using single-cell RNA sequencing have been recently published[112–115] and have begun dissecting AD heterogeneity at the cellular level.

Perspectives

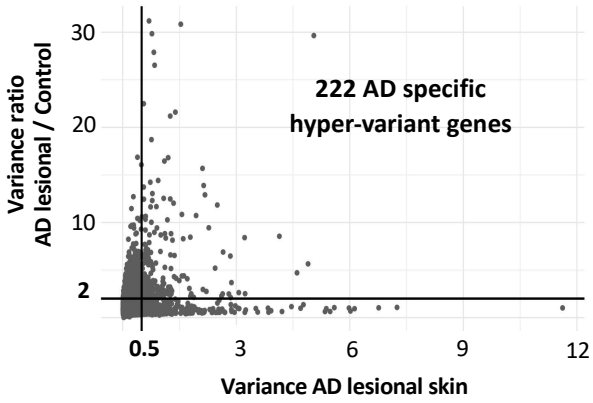
Our logical and intuitive method selected biologically relevant genes from a large amount of data. Combined with a data-driven unsupervised classification, it enabled reproducible AD endotype identification. This pipeline could be applied in other complex diseases for which classification remains challenging. Regarding AD, endotypes were characterized by distinct disease severities and conflicting biological mechanisms. This could help to better understand AD heterogeneity as well as develop new and personalized therapeutic approaches.

Figure 1

a



b



c **Functional enrichment of all AD specific hyper-variable genes with GO biological process**

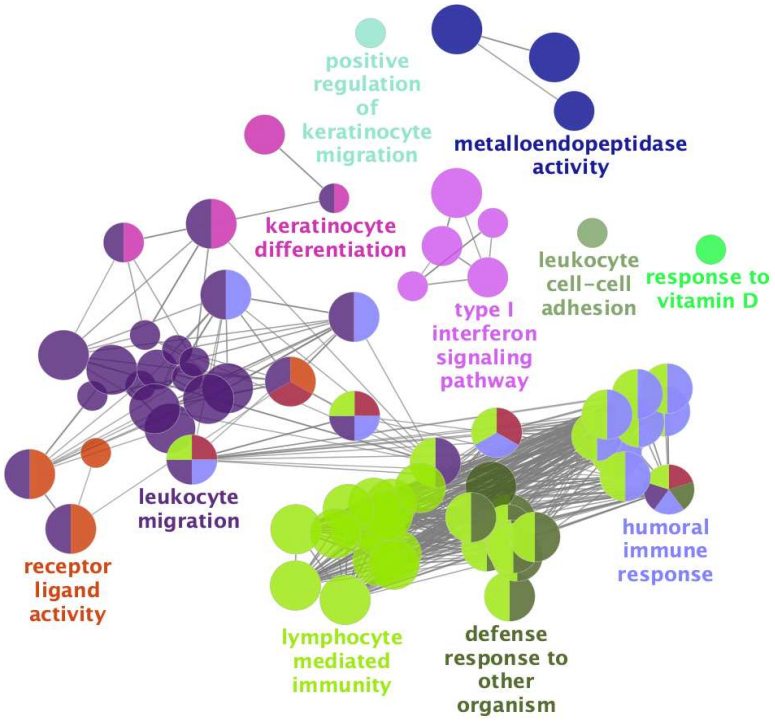


Figure 1: **AD-specific hyper-variable genes are representative of known AD mechanisms.**

a: Workflow of feature selection and functional annotation. First, we selected only hyper-variable genes in AD lesional skin (variance > 0.5). Second, we kept genes that were twice less variable in healthy skin (ratio: variance AD / Control skin > 2) (Fig 1b). From the 222 AD-specific hyper-variable genes, we performed several functional enrichments that highlighted known AD mechanisms (Fig 1c, Suppl 1b).

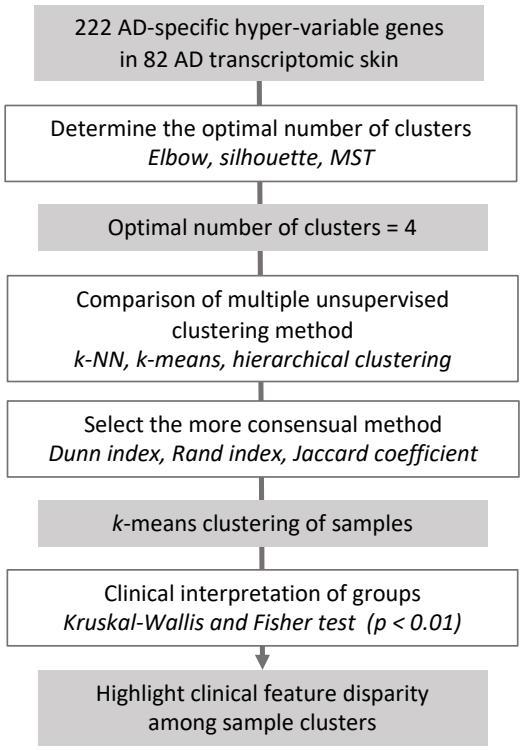
b: AD-specific hyper-variable gene selection. Among AD lesional skin, n = 575 genes were highly variable (variance > 0.5). Among them, n = 222 were twice more variable than healthy skin and called AD-specific hyper-variable genes.

c: Functional enrichment of all AD-specific hyper-variable genes with GO biological process. It confirmed the AD-specific hyper-variable gene's role as representative of AD known pathophysiology. Network representation of significative terms (p.adj > 0.05). Nodes represent main GO terms. Edges symbolize interactions between nodes.

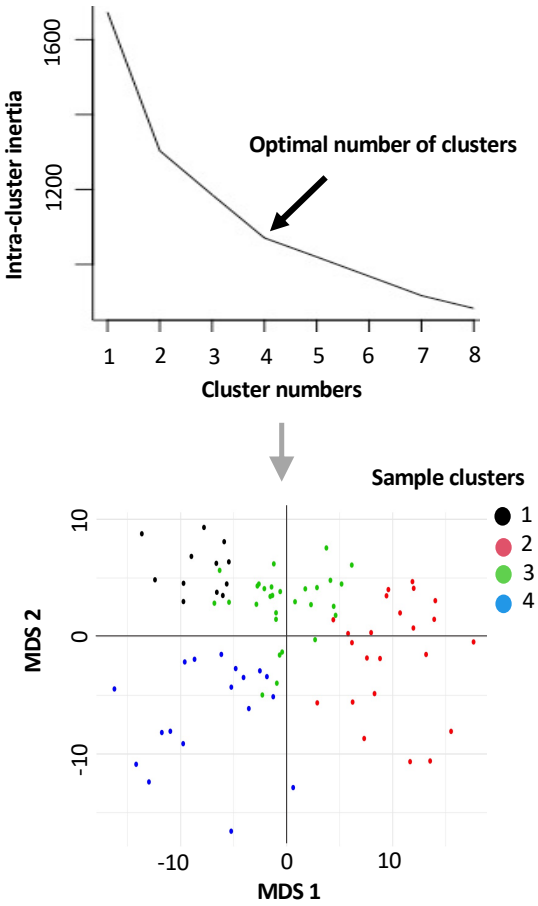
AD: atopic dermatitis, GO: gene ontology

Figure 2

a



b



c

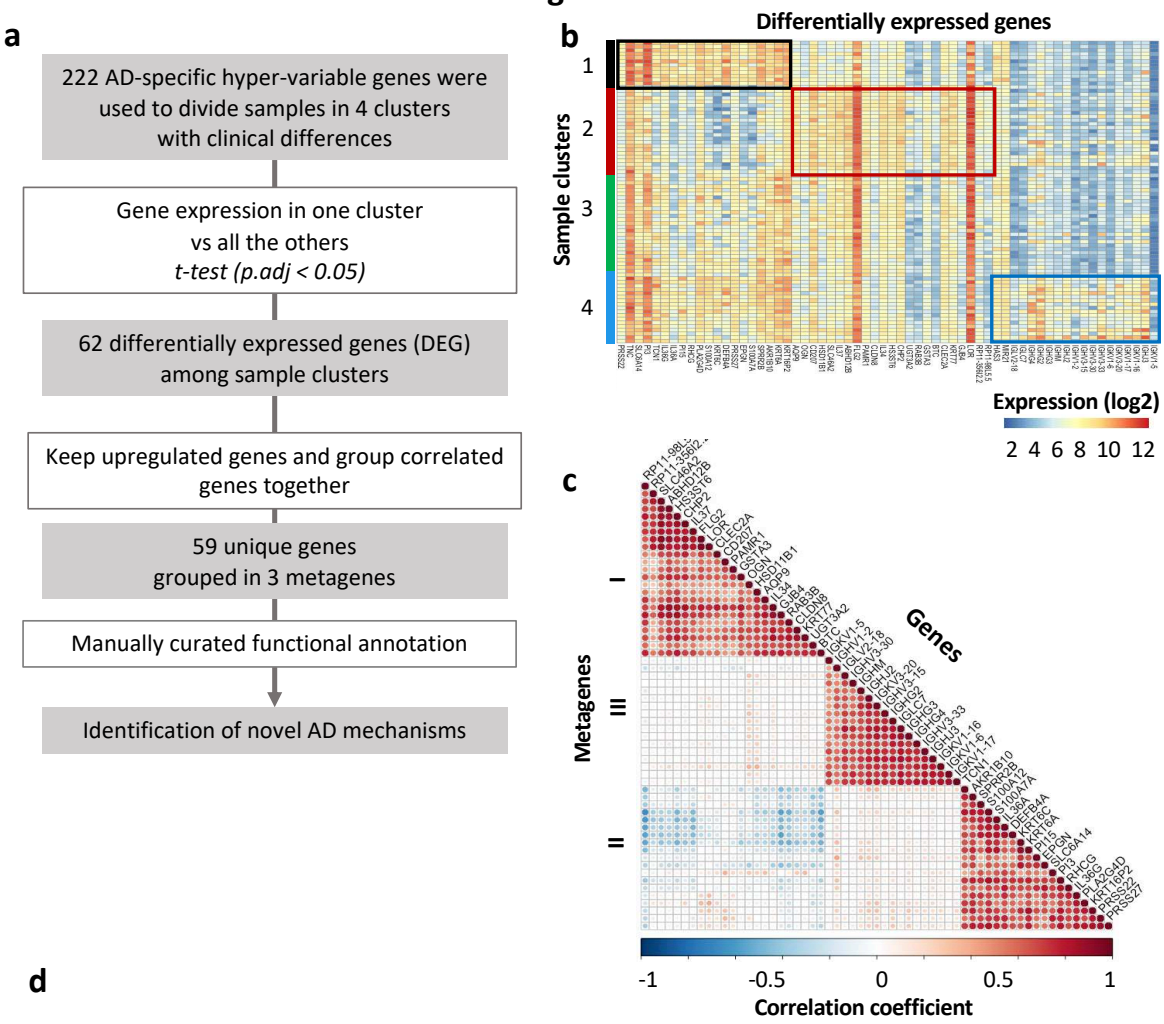
Relevant clinical variables among clusters (adjusted p-value < 0.01)

SCORAD **SA colonization**	**< 0.01**
Personal history	
Age	
Gender	
Early age of AD onset	
Allergies	
Asthma	
Cardiovascular diseases	
Familial history	
Allergies	
Clinical features	> 0.05
Cheilitis	
Lichenification	
Orbital darkening	
Palmar hyperlinearity	
Perifollicular accentuation	
Dennie Morgan infraorbital fold	
Pruritus score	
Sleep loss score	
Biological features	
Raised IgE serum level	

Figure 2: **Unsupervised clustering of four sample clusters with distinct clinical features.** **a:** Workflow of sample clustering and clinical characterization. The 222 AD-specific hyper-variable genes have been used to cluster samples. The optimal number of clusters was estimated at four by all the different methods (Fig 2b and Suppl 3). The different clustering methods have been compared to choose the most consensual and consistent (Suppl 4). **b:** *k*-means clustering representation using the optimal number of clusters according to the elbow method. The optimal number of clusters has been evaluated to 4 by the elbow method. *k*-means clusters have been represented in the dimension reduction MDS graph. **c:** Relevant clinical variables among clusters. Variables are ordered by statistical significance. Kruskal Wallis test has been used for continuous, and Fisher test for Boolean variables. The most significant clinical parameters were: SCORAD, and *SA* colonization (adjusted $p<0.01$). All other variables had an adjusted $p>0.05$.

AD: atopic dermatitis, MDS: multi-dimensional scale, MST: Minimal Spanning Tree, SA: Staphylococcus aureus, *SCORAD: score AD*

Figure 3



Metagene I (19 genes) – IL-1 family signaling & skin remodeling
Th17 : DEFB4A, IL36A, IL36G, PI3, PLA2G4D, S100A7A, S100A12 Epidermal proliferation and differentiation : EPGN, KRT6A, KRT6C, AKR1B10, SPRR2B Protein metabolism : PRSS22, SLC6A14, PI15, TCN1 Poorly described : KRT16P2, PRSS27, RHCG
Metagene II (23 genes) – Negative immune regulation & skin architecture
Negative immune regulation: CLEC2A (=PILAR), IL34, IL37 Skin architecture: FLG2, LOR, OGN Epidermal homeostasis: BTC, GSTA3, GJB4, KRT77, UGT3A2, Water metabolism: AQP9, CLDN8 Langerhans cell marker: CD207 Lipase: ABHD12B Endogenous steroid activation: HSD11B1 Poorly described : CHP2, HS3ST6, PAMR1, RAB3B, RP11-356I2.2, RP11-98L5.5, SLC46A2
Metagene III (17 genes) – B lymphocyte immunity
Immunoglobulin recombination: IGLV2-18, IGLC7, IGHG4, IGHG2, IGHG3, IGHM, IGHJ2, IGHV1-2, IGHV3-15, IGHV3-30, IGHV3-33, IGKV1-6, IGKV3-20, IGKV1-17, IGKV1-16, IGHJ3, IGKV1-5

Figure 3: Metagene expression in sample clusters reveal exciting molecular and functional signatures.

a: Workflow of discriminant genes selection and grouping, and functional annotation to build molecular signature of sample clusters. We used the four sample clusters made from the 222 AD-specific hyper-variable genes. The second phase of gene selection was performed based on their differential analysis (Fig 3b).

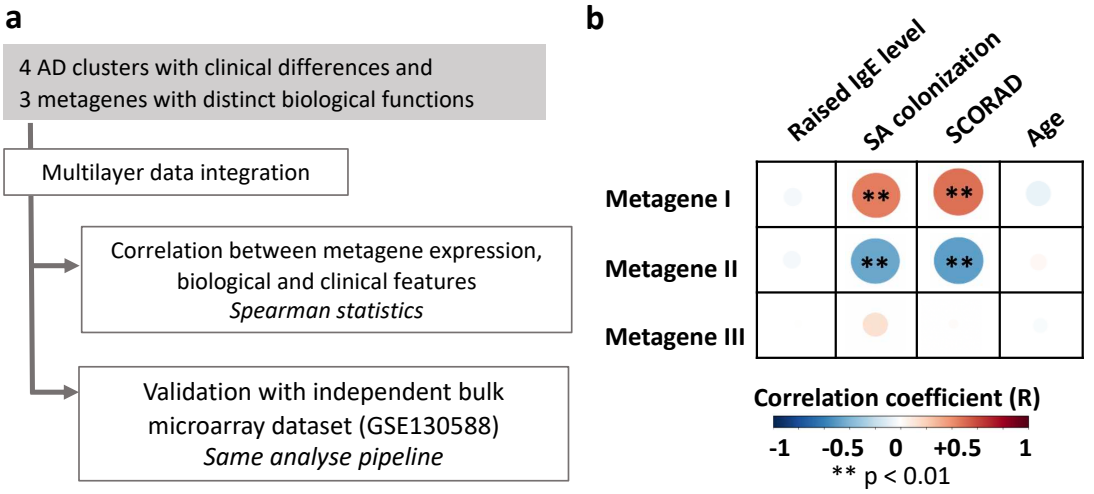
b: Heatmap representation of gene expression disparity among samples and clusters. Co-expressed genes had the same behavior according to sample clusters They are surrounded by their specific colored boxes.

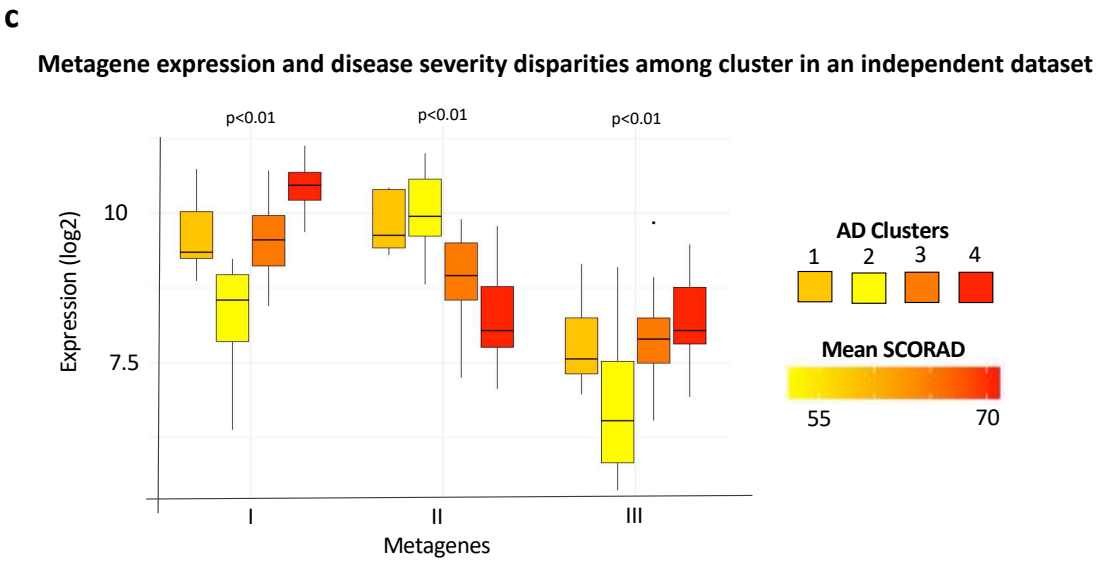
c: Metagenes construction by correlation analysis using Pearson statistic. We select unique and up-regulated genes from the differential expression analysis and grouped correlated genes into their respective metagenes.

d: Literature mining manually curated signature revealed interesting and defined metagenes functions. Genes belonging to each metagene shared common biological functions. Computational approaches were not suitable for this analysis due to the low number of genes per metagene. Thus, a manually curated approach was favored.

AD: atopic dermatitis

Figure 4


a
4 AD clusters with clinical differences and
3 metagenes with distinct biological functions
Multilayer data integration
Correlation between metagene expression, biological and clinical features
Spearman statistics
Validation with independent bulk microarray dataset (GSE130588)
Same analyse pipeline
b
Raised IgE level
SA colonization
SCORAD
Age
Metagene I
Metagene II
Metagene III
**
Correlation coefficient (R)
-1 -0.5 0 +0.5 1
** p < 0.01


c
Metagene expression and disease severity disparities among cluster in an independent dataset
p<0.01
p<0.01
p<0.01
Expression (log2)
10
7.5
I
II
III
Metagenes
AD Clusters
1 2 3 4
Mean SCORAD
55
70

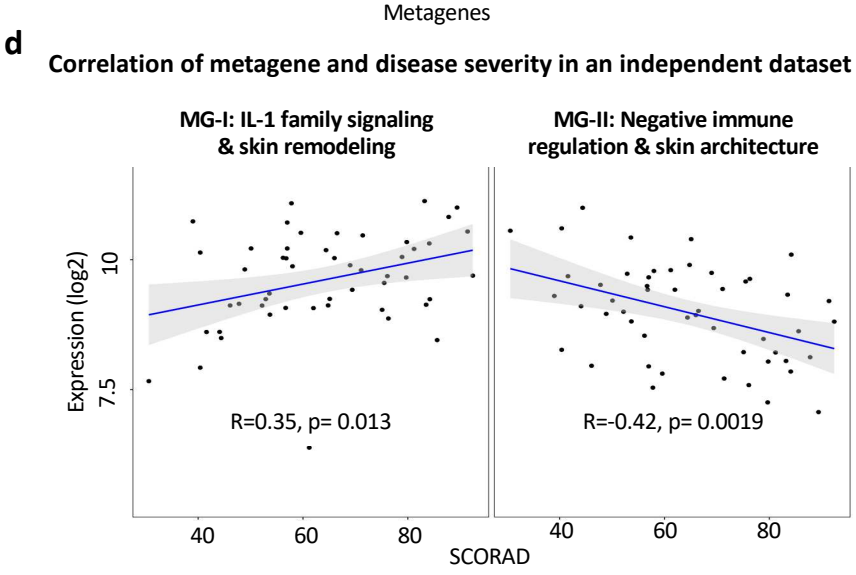

d
Correlation of metagene and disease severity in an independent dataset
MG-I: IL-1 family signaling & skin remodeling
MG-II: Negative immune regulation & skin architecture
Expression (log2)
10
7.5
R=0.35, p= 0.013
R=-0.42, p= 0.0019
40
60
80
SCORAD

Figure 4: Data integration for multiscale cluster characterization and external validation. **a:** Workflow of the data integration step. The different phases were independent of each other.

b: Correlation between metagene expressions and clinical features using Spearman statistics. Significant positive correlations were shown between MG-I, SCORAD, and SA colonization. Opposingly, significant negative correlations were found between MG-II, SCORAD, SA colonization, and MG-I. No significant results were found between metagenes expression and raised IgE levels (binary variable) or age (continuous variable).

c: Metagene expression and disease severity disparities among clusters in an independent dataset (GSE130588). AD cluster colors have been chosen relatively to their mean severity score. ANOVA test showed metagenes as differentially expressed between clusters. As in our cohort, the cluster with the more severe AD hyper-expressed MG-I, and the cluster with the less severe AD hyper-expressed MG-II.

d: Correlation of metagene and disease severity in an independent dataset (GSE130588). Spearman statistics confirmed the positive correlation between MG-I expression and disease severity as well as the negative correlation between MG-II expression and disease severity.

AD: atopic dermatitis, MG: metagene, SA: Staphylococcus aureus, *SCORAD: score AD*

Supplementary 1

a

Demographical data of AD and control cohort

	AD patients	Controls
Individuals (nb)	82	113
Samples (nb)	82	213
Age mean (yo) [min-max]	44 [20-83]	35 [19-77]
Gender (F/M)	36/46	44/69

b

Functional enrichment of all AD-specific hyper-variable genes with KEGG database

Term name	Terme ID	p.adj
Cytokine-cytokine receptor interaction	KEGG:04060	3.4 e-9
Viral protein interaction with cytokine	KEGG:04061	1.4 e-8
IL-17 signaling pathway	KEGG:04657	2.8 e-4
Influenza A	KEGG:05164	3.3 e-3
AGE-RAGE signalling pathway in diabetic complications	KEGG:04933	3.9 e-3

Supplementary 1:

a: Demographical data of AD and control cohort.

b: Functional enrichment of all AD-specific hyper-variable genes
based on the KEGG database. Statistically significant biological processes are ranked by adjusted p-value.

Supplementary 2

	GeneSymbol	Variance in AD	Variance in controls	Ratio
ENSG00000175426	PCSK1	0.695	0.022	31.2
ENSG00000197915	HRNR	1.534	0.05	30.8
ENSG00000127318	IL22	0.772	0.026	29.8
ENSG00000196611	MMP1	5.05	0.17	29.7
ENSG00000162892	IL24	0.822	0.029	27.9
ENSG00000142224	IL19	0.842	0.032	26.5
ENSG00000254651	RP11-430H10.3	0.554	0.025	22.5
ENSG00000244057	LCE3C	1.391	0.064	21.6
ENSG00000108702	CCL1	1.255	0.059	21.2
ENSG00000105641	SLC5A5	0.76	0.041	18.7
ENSG00000169429	IL8	1.201	0.071	16.8
ENSG00000006074	CCL18	1.102	0.067	16.5
ENSG00000149968	MMP3	2.109	0.134	15.7
ENSG00000211936	IGHV4-4	0.936	0.065	14.4
ENSG00000143520	FLG2	0.729	0.051	14.2
ENSG00000185962	LCE3A	2.139	0.154	13.9
ENSG00000162891	IL20	0.554	0.04	13.7
ENSG00000123496	IL13RA2	0.784	0.06	13.0
ENSG00000182585	EPGN	2.173	0.168	12.9
ENSG00000103569	AQP9	1.082	0.086	12.5
ENSG00000134028	ADAMDEC1	1.256	0.101	12.5
ENSG00000166509	CLEC3A	0.559	0.045	12.4
ENSG00000212556	Y_RNA	0.781	0.063	12.3
ENSG00000153802	TMPRSS11D	1.343	0.112	12.0
ENSG00000136694	IL36A	2.499	0.211	11.8
ENSG00000163661	PTX3	0.89	0.076	11.7
ENSG00000182566	CLEC4G	0.736	0.063	11.6
ENSG00000175592	FOSL1	1.564	0.144	10.8
ENSG00000188293	IGFL1	1.939	0.181	10.7
ENSG00000181617	FDCSP	0.606	0.057	10.6
ENSG00000236481	AC002331.1	0.51	0.048	10.6
ENSG00000158859	ADAMTS4	0.6	0.058	10.3
ENSG00000206384	COL6A6	1.181	0.115	10.3
ENSG00000203782	LOR	0.757	0.076	10.0
ENSG00000110347	MMP12	2.275	0.241	9.4
ENSG00000171889	MIR31HG	0.904	0.096	9.4
ENSG00000223872	AC006372.5	0.601	0.065	9.2
ENSG00000189433	GJB4	0.836	0.093	9.0
ENSG00000248329	RP11-366M4.3	1.114	0.126	8.8
ENSG00000198074	AKR1B10	1.277	0.145	8.8
ENSG00000162040	HS3ST6	0.706	0.081	8.7
ENSG00000204936	CD177	1.006	0.116	8.7
ENSG00000170465	KRT6C	4.139	0.484	8.6
ENSG00000102837	OLFM4	1.787	0.211	8.5
ENSG00000184330	S100A7A	3.221	0.383	8.4
ENSG00000205362	MT1A	1.137	0.136	8.4
ENSG00000105205	CLC	1.598	0.193	8.3
ENSG00000213886	UBD	1.315	0.161	8.1
ENSG00000227471	AKR1B15	0.699	0.089	7.9
ENSG00000115758	ODC1	0.584	0.076	7.7
ENSG00000166736	HTR3A	0.603	0.079	7.6
ENSG00000104368	PLAT	0.663	0.093	7.2
ENSG00000103044	HAS3	0.605	0.085	7.1
ENSG00000211966	IGHV5-51	0.997	0.143	7.0
ENSG00000163221	S100A12	2.652	0.385	6.9
ENSG00000138135	CH25H	0.743	0.108	6.9
ENSG00000163600	ICOS	0.57	0.086	6.6
ENSG00000211791	TRAV13-2	0.638	0.096	6.6
ENSG00000157368	IL34	0.543	0.082	6.6
ENSG00000243466	IGKV1-5	1.263	0.191	6.6
ENSG00000211653	IGLV1-40	1.367	0.209	6.5
ENSG00000134827	TCN1	2.844	0.44	6.5
ENSG00000163638	ADAMTS9	0.525	0.081	6.5
ENSG00000103888	KIAA1199	0.661	0.102	6.5
ENSG00000102970	CCL17	0.509	0.08	6.4
ENSG00000227300	KRT16P2	0.638	0.102	6.3
ENSG00000166869	CHP2	0.681	0.109	6.3
ENSG00000169385	RNASE2	0.681	0.109	6.3
ENSG00000188404	SELL	0.598	0.099	6.1
ENSG00000119457	SLC46A2	0.744	0.126	5.9
ENSG00000124731	TREM1	0.669	0.116	5.8
ENSG00000211598	IGKV4-1	0.565	0.098	5.8
ENSG00000226145	AC022596.6	0.876	0.153	5.7
ENSG00000171711	DEFB4A	4.897	0.867	5.7
ENSG00000131969	ABHD12B	0.755	0.134	5.6
ENSG00000119508	NR4A3	0.514	0.092	5.6
ENSG00000159337	PLA2G4D	0.679	0.122	5.6
ENSG00000137440	FGFBP1	0.755	0.136	5.5
ENSG00000169213	RAB3B	0.788	0.148	5.3
ENSG00000169248	CXCL11	1.113	0.211	5.3
ENSG00000200972	RNU5A-8P	2.442	0.47	5.2
ENSG00000094804	CDC6	0.523	0.101	5.2
ENSG00000125571	IL37	1.276	0.247	5.2
ENSG00000137648	TMPRSS4	0.8	0.159	5.0
ENSG00000211900	IGHJ6	4.605	0.979	4.7
ENSG00000189182	KRT77	0.743	0.16	4.7
ENSG00000211838	TRAJ52	0.578	0.126	4.6
ENSG00000112984	KIF20A	0.527	0.117	4.5
ENSG00000168671	UGT3A2	1.614	0.367	4.4
ENSG00000239855	IGKV1-6	1.566	0.358	4.4
ENSG00000149090	PAMR1	0.516	0.119	4.3
ENSG00000117594	HSD11B1	1.358	0.316	4.3
ENSG00000138642	HERC6	0.65	0.153	4.2
ENSG00000157601	MX1	0.669	0.159	4.2
ENSG00000126787	DLGAP5	0.56	0.133	4.2
ENSG00000172382	PRSS27	0.673	0.16	4.2
ENSG00000231475	IGHV4-31	1.065	0.257	4.2
ENSG00000177257	DEFB4B	1.41	0.341	4.1
ENSG00000187498	COL4A1	0.542	0.132	4.1
ENSG00000105664	COMP	1.706	0.418	4.1
ENSG00000137558	PI15	1.494	0.384	3.9
ENSG00000106366	SERPINE1	0.803	0.206	3.9
ENSG00000159167	STC1	0.856	0.223	3.8
ENSG00000006327	TNFRSF12A	0.7	0.183	3.8
ENSG00000236543	RP11-98L5.5	1.059	0.278	3.8
ENSG00000211753	TRBV28	0.618	0.163	3.8
ENSG00000134321	RSAD2	0.572	0.152	3.8
ENSG00000012223	LTF	2.873	0.78	3.7
ENSG00000151006	PRSS53	0.617	0.17	3.6
ENSG00000184613	NELL2	0.834	0.232	3.6
ENSG00000039987	BEST2	0.643	0.179	3.6
ENSG00000211904	IGHJ2	1.09	0.304	3.6
ENSG00000005001	PRSS22	0.512	0.143	3.6
ENSG00000187116	LILRA5	0.688	0.193	3.6
ENSG00000231331	AC103563.2	0.74	0.209	3.5
ENSG00000111335	OAS2	0.711	0.203	3.5
ENSG00000211866	TRAJ23	0.721	0.206	3.5
ENSG00000211952	IGHV4-28	0.754	0.216	3.5
ENSG00000211880	TRAJ9	0.508	0.148	3.4
ENSG00000137959	IFI44L	1.052	0.314	3.4
ENSG00000161905	ALOX15	1.047	0.314	3.3
ENSG00000152268	SPON1	0.552	0.166	3.3
ENSG00000211950	IGHV1-24	0.616	0.186	3.3
ENSG00000211956	IGHV4-34	0.683	0.207	3.3
ENSG00000166948	TGM6	0.563	0.171	3.3
ENSG00000158485	CD1B	0.923	0.282	3.3
ENSG00000184557	SOCS3	0.656	0.201	3.3
ENSG00000183813	CCR4	0.515	0.158	3.3
ENSG00000162739	SLAMF6	0.62	0.191	3.2
ENSG00000211949	IGHV3-23	0.718	0.223	3.2
ENSG00000111012	CYP27B1	0.612	0.19	3.2
ENSG00000140519	RHCG	1.252	0.389	3.2
ENSG00000041982	TNC	0.654	0.204	3.2
ENSG00000205678	TECRL	0.763	0.238	3.2
ENSG00000119125	GDA	1.207	0.377	3.2
ENSG00000162692	VCAM1	0.527	0.167	3.1
ENSG00000116031	CD207	0.63	0.2	3.1
ENSG00000136688	IL36G	0.73	0.232	3.1
ENSG00000211875	TRAJ14	0.542	0.173	3.1
ENSG00000205364	MT1M	0.581	0.186	3.1
ENSG00000211853	TRAJ36	0.539	0.175	3.1
ENSG00000240382	IGKV1-17	1.761	0.574	3.1
ENSG00000241244	IGKV1D-16	0.546	0.179	3.0

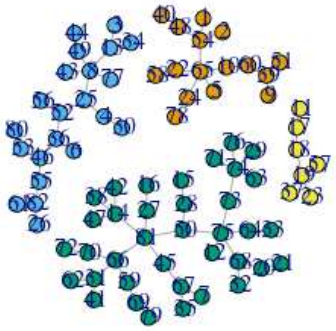
ENSG00000239951	IGKV3-20	0.922	0.305	3.0
ENSG00000117228	GBP1	0.524	0.173	3.0
ENSG00000211945	IGHV1-18	0.838	0.279	3.0
ENSG00000102962	CCL22	0.961	0.32	3.0
ENSG00000174502	SLC26A9	0.74	0.248	3.0
ENSG00000007908	SELE	1.278	0.429	3.0
ENSG00000235842	RP11-356I2.2	1.052	0.354	3.0
ENSG00000211953	IGHV3-30	2.119	0.715	3.0
ENSG00000140534	TICRR	0.514	0.175	2.9
ENSG00000200648	U6	0.517	0.176	2.9
ENSG00000164687	FABP5	0.507	0.174	2.9
ENSG00000211860	TRAJ29	0.503	0.173	2.9
ENSG00000199004	MIR21	0.696	0.24	2.9
ENSG00000185745	IFIT1	0.515	0.178	2.9
ENSG00000119938	PPP1R3C	0.525	0.184	2.9
ENSG00000107984	DKK1	0.681	0.24	2.8
ENSG00000240834	IGKV1D-12	1.198	0.423	2.8
ENSG00000211934	IGHV1-2	1.102	0.39	2.8
ENSG00000240864	IGKV1-16	1.301	0.464	2.8
ENSG00000156284	CLDN8	0.558	0.199	2.8
ENSG00000172752	COL6A5	0.92	0.33	2.8
ENSG00000127954	STEAP4	0.524	0.188	2.8
ENSG00000168685	IL7R	0.606	0.22	2.8
ENSG00000196805	SPRR2B	2.209	0.802	2.8
ENSG00000183760	PAPL	0.541	0.198	2.7
ENSG00000158488	CD1E	0.601	0.221	2.7
ENSG00000174156	GSTA3	0.579	0.213	2.7
ENSG00000140285	FGF7	0.564	0.208	2.7
ENSG00000211899	IGHM	0.985	0.371	2.7
ENSG00000188393	CLEC2A	0.527	0.199	2.6
ENSG00000211892	IGHG4	3.067	1.167	2.6
ENSG00000251616	RP11-485M7.3	0.583	0.223	2.6
ENSG00000119917	IFIT3	0.547	0.211	2.6
ENSG00000147138	GPR174	0.625	0.242	2.6
ENSG00000211859	TRAJ30	0.521	0.202	2.6
ENSG00000169313	P2RY12	0.524	0.204	2.6
ENSG00000184348	HIST1H2AK	0.676	0.266	2.5
ENSG00000211867	TRAJ22	0.797	0.317	2.5
ENSG00000242887	IGHJ3	3.233	1.285	2.5
ENSG00000199377	RNU5F-1	1.107	0.44	2.5
ENSG00000211943	IGHV3-15	1.81	0.723	2.5
ENSG00000188257	PLA2G2A	1.284	0.513	2.5
ENSG00000211673	IGLV3-1	1.189	0.475	2.5
ENSG00000124102	PI3	2.64	1.058	2.5
ENSG00000211955	IGHV3-33	2.804	1.127	2.5
ENSG00000087916	SLC6A14	0.927	0.374	2.5
ENSG00000211858	TRAJ31	1.01	0.411	2.5
ENSG00000211940	IGHV3-9	0.504	0.206	2.4
ENSG00000137965	IFI44	0.739	0.307	2.4
ENSG00000162654	GBP4	0.518	0.216	2.4
ENSG00000211664	IGLV2-18	1.256	0.529	2.4
ENSG00000174808	BTC	0.6	0.253	2.4
ENSG00000211855	TRAJ34	0.571	0.242	2.4
ENSG00000113070	HBEGF	0.795	0.34	2.3
ENSG00000153234	NR4A2	0.528	0.227	2.3
ENSG00000100985	MMP9	0.503	0.216	2.3
ENSG00000243290	IGKV1-12	0.892	0.383	2.3
ENSG00000211893	IGHG2	2.611	1.147	2.3
ENSG00000211685	IGLC7	1.791	0.793	2.3
ENSG00000211765	TRBJ2-2	0.548	0.245	2.2
ENSG00000170801	HTRA3	0.564	0.254	2.2
ENSG00000211884	TRAJ5	0.545	0.25	2.2
ENSG00000171246	NPTX1	0.585	0.268	2.2
ENSG00000249437	NAIP	0.595	0.274	2.2
ENSG00000211897	IGHG3	1.077	0.498	2.2
ENSG00000211856	TRAJ33	0.62	0.289	2.1
ENSG00000177575	CD163	0.699	0.327	2.1
ENSG00000205420	KRT6A	1.876	0.878	2.1
ENSG00000211873	TRAJ16	0.653	0.308	2.1
ENSG00000090659	CD209	0.618	0.293	2.1
ENSG00000169245	CXCL10	2.853	1.355	2.1
ENSG00000115008	IL1A	0.98	0.466	2.1
ENSG00000211699	TRGV3	0.603	0.288	2.1

Supplementary 2: **The 222 AD specific hyper-variable genes with their respective variance in the pathological and physiological states are ranked by ratio.**

Supplementary 3

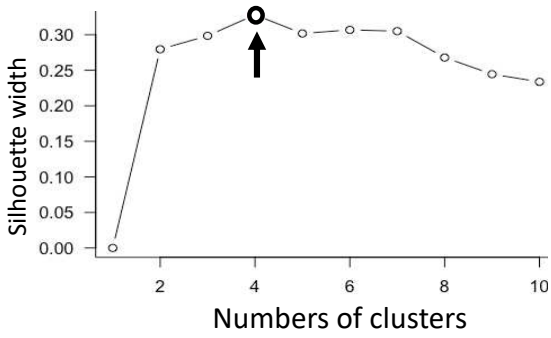
a

***k*-NN and MST sample clustering in four clusters**



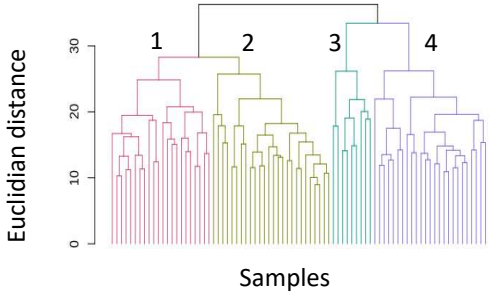
b

Average silhouette width in all samples



c

Hierarchical clustering of samples in four clusters



Supplementary 3: **Cluster number optimization**

a: Sample k-NN and MST clustering in four clusters. The number written within each circle refers to the sample identity.

b: The highest average silhouette width corresponded to four sample clusters.

c: Hierarchical clustering: Euclidian distance has been used to constitute four clusters

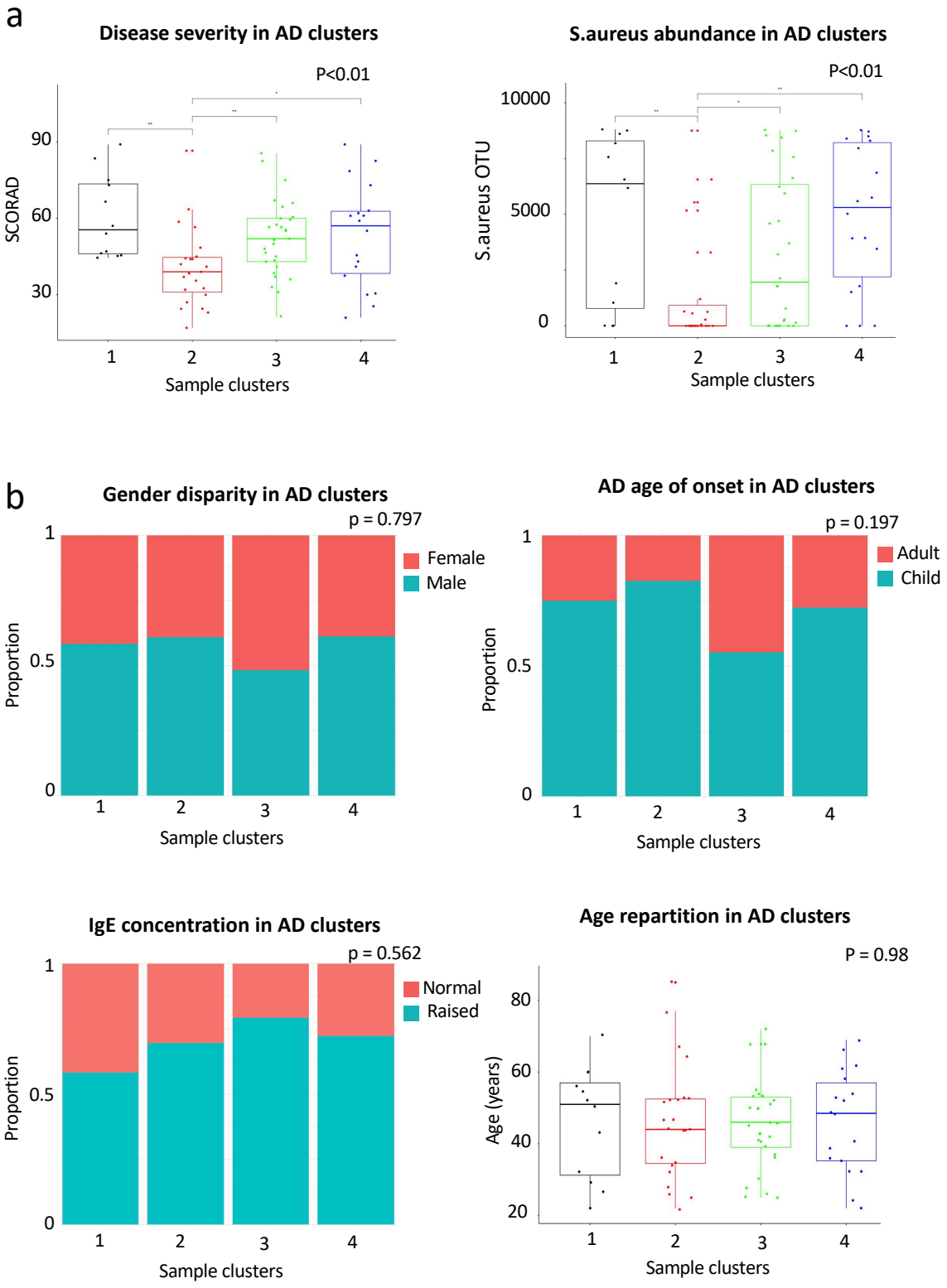
Supplementary 4

Intra and inter-cluster metrics used to select the more consensual method

Methods	**Hierarchical clustering**	***k*-NN clustering**	***k*-means clustering**
Cluster numbers	4	4	4
Intra-cluster metric			
Dunn Index	0,42	0,34	0,43
Inter-cluster metrics			
Rand Index			
Hierarchical clustering		0,68	0,79
***k*-NN clustering**	0,68		0,73
***k*-means clustering**	0,79	0,73	
Jaccard similarity coefficient			
Hierarchical clustering		0,30	0,44
***k*-NN clustering**	0,30		0,38
***k*-means clustering**	0,44	0,38	

Supplementary 4: **Clustering methods comparison**. *k*-means clustering method has been chosen for sample clustering because of its highest intra-cluster (Dunn index) and inter-cluster (Rand index and Jaccard similarity coefficient) metrics.

Supplementary 5



Supplementary 5: **Clinical features representation in sample clusters.**

a: Statistically significant clinical features: SCORAD and SA colonization

b: Important non-significant clinical features: early age of disease onset, IgE concentration, gender, age

SA: Staphylococcus aureus, *SCORAD: score AD*

Supplementary 6

a **Differentially expressed genes in one cluster vs the others**

Cluster 1 vs 2, 3 and 4	Cluster 2 vs 1, 3 and 4	Cluster 3 vs 1, 3 and 4	Cluster 4 vs 1, 2, and 3
AKR1B10	ABHD12B	No gene	AKR1B10
DEFB4A	AQP9		HAS3
IL36A	BTC		IGLV2-18
IL36G	CD207		IGLC7
EPGN	CLDN8		IGHG4
KRT6A	CLEC2A		IGHG2
KRT6C	CHP2		IGHG3
KRT16P2	HS3ST6		IGHM
PI3	HSD11B1		IGHJ2
PI15	FLG2		IGHV1-2
PLA2G4D	GJB4		IGHV3-15
PRSS22	GSTA3		IGHV3-30
PRSS27	IL34		IGHV3-33
RHCG	IL37		IGKV1-6
S100A7A	KRT77		IGKV3-20
S100A12	LOR		IGKV1-17
SLC6A14	OGN		IGKV1-16
SPRR2B	PAMR1		IGHJ3
TCN1	RAB3B		IGKV1-5
TNC	RP11-356I2.2		KRT6A
	RP11-98L5.5		MIR21
	SLC46A2		PI3
	UGT3A2		PRSS27
			S100A12
			TCN1

b **Upregulated differentially expressed genes grouped in corresponding metagenes**

Metagenes		
1	2	3
AKR1B10	ABHD12B	IGLV2-18
DEFB4A	AQP9	IGLC7
EPGN	BTC	IGHG4
IL36A	CD207	IGHG2
IL36G	CHP2	IGHG3
KRT16P2	CLDN8	IGHM
KRT6A	CLEC2A	IGHJ2
KRT6C	FLG2	IGHV1-2
PI15	GJB4	IGHV3-15
PI3	GSTA3	IGHV3-30
PLA2G4D	HS3ST6	IGHV3-33
PRSS22	HSD11B1	IGKV1-6
PRSS27	IL34	IGKV3-20
RHCG	IL37	IGKV1-17
S100A12	KRT77	IGKV1-16
S100A7A	LOR	IGHJ3
SLC6A14	OGN	IGKV1-5
SPRR2B	PAMR1	
TCN1	RAB3B	
	RP11-356I2.2	
	RP11-98L5.5	
	SLC46A2	
	UGT3A2	

Supplementary 6: **AD clusters transcriptomic signature.**

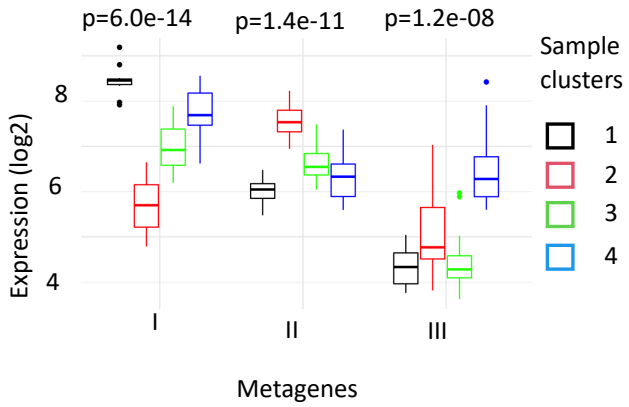
a: Differentially expressed genes in one cluster versus the others

b: Upregulated differentially expressed genes grouped in their metagenes

AD: atopic dermatitis

Supplementary 7

Metagene expression among sample clusters



Supplementary 7: Metagene expression among sample clusters. ANOVA test showed differential expression of metagenes between clusters.

Supplementary 8

TITLE	DOI	GEO	ARRAY EXPRESS	AUTHOR	YEAR	TECH_MICROARRAY	SAMPLES	AGE	SCORAD
Dupilumab progressively improves systemic and cutaneous abnormalities in patients with atopic dermatitis	10.1016/j.jaci.2018.08.022	GSE130588	NA	E. Guttman-Yasski et al	2018	Affymetrix Human U133Plus 2.0 gene arrays	51	Adult	YES
Cyclosporine in patients with atopic dermatitis modulates activated inflammatory pathways and reverses epidermal pathology	10.1016/j.jaci.2014.03.003	GSE58558	E-GEOD-58558	S. Khattri et al	2014	Affymetrix Human U133Plus 2.0 gene arrays	12	Adult	YES
Dupilumab improves the molecular signature in skin of patients with moderate-to-severe atopic dermatitis	10.1016/j.jaci.2014.10.013	GSE59294	E-GEOD-59294	JD. Hamilton et al	2014	Affymetrix Human U133Plus 2.0 gene arrays	18	Adult	YES
Progressive Activation of Th2/Th22 characterizes acute and chronic atopic dermatitis	10.1016/j.jaci.2012.07.012	GSE36842	NA	JK. Gittler et al	2012	Affymetrix Human U133Plus 2.0 gene arrays	17	Adult	YES
Reversal of atopic dermatitis with narrow-band UVB phototherapy and biomarkers for therapeutic response	10.1016/j.jaci.2011.05.042	GSE27887	E-GEOD-27887	S Tintle et al	2011	Affymetrix Human U133Plus 2.0 gene arrays	12	Adult	YES
Nonlesional atopic dermatitis skin is characterized by broad terminal differentiation defects and variable immune abnormalities	10.1016/j.jaci.2010.12.1124	GSE32924	NA	M Suárez-Fariña et al	2011	Affymetrix Human U133Plus 2.0 gene arrays	12	Adult	YES

Supplementary 8: Comprehensive list of independent skin AD transcriptome with clinical annotations on disease severity. The largest expression matrix (GSE130588) was used to validate our main findings.

PART 2: Supervised approach

Could a combination of statistical and machine learning models highlight pruritus multiple mechanisms?

Rational of the approach

For the second part of the thesis, we wanted to start from a clinical issue as we had an important and well characterized clinical dataset. We have chosen to focus on pruritus for several reasons. First, this is a very debilitating symptom for which targeted treatments development has just began. Second, the pruritus intensity score was annotated for all patients. Finally, we thought that pruritus is a complex symptom involving a wide range of mechanisms that could be well captured by transcriptomic data.

Result announcement

We hypothesized that relationships between pruritus intensity and genetic expression could be diverse. Thus, we used different kinds of approaches that assessed different natures of data. We used classical statistical models to first find out that they were not the most pertinent. We then used a combination of statistical and machine learning approaches to extract the most minimalistic pruritus signature. We optimized our pipeline to be able to highlight a distinct pruritus signature on an independent cohort. Both signatures showed great accuracy for pruritus predictions. Although, no gene was common to both signatures, they shared interesting functions, had never been described in AD and showed interesting therapeutic potential.

Graphical abstract

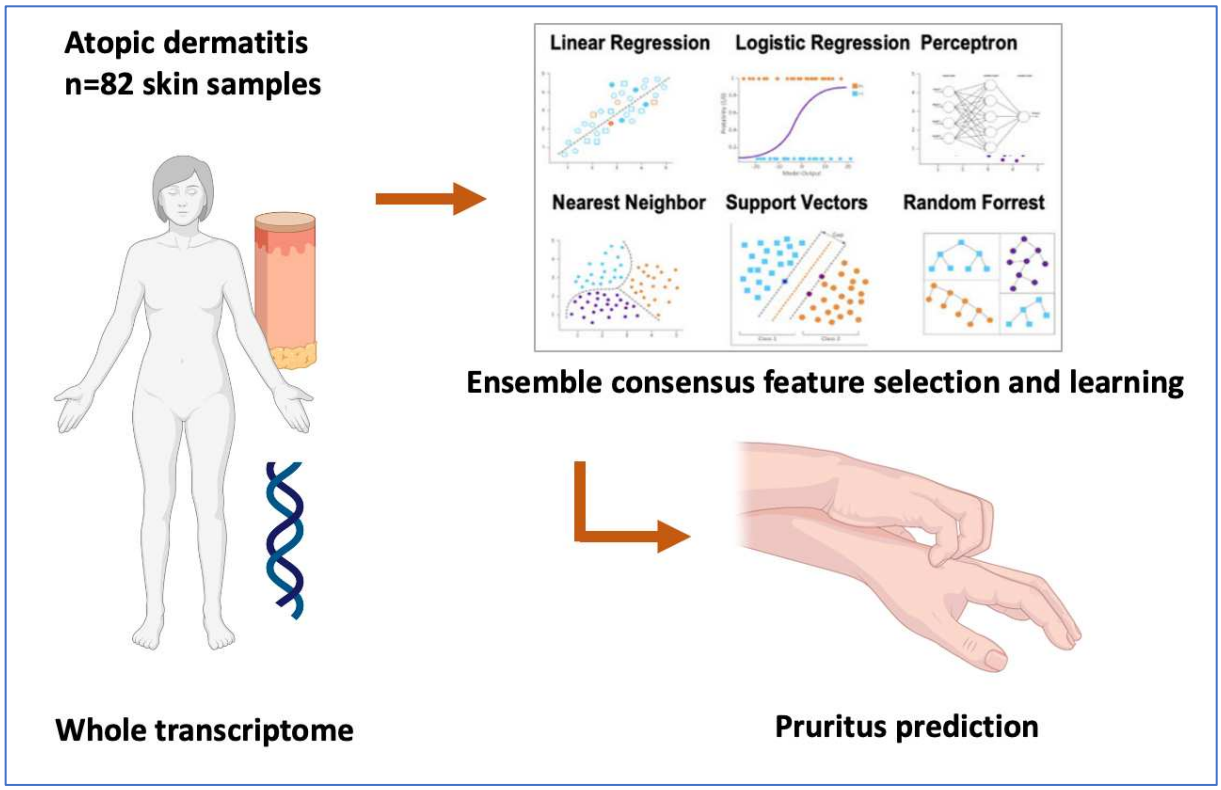


Figure 7 Graphical abstract of study design and results. Designed with Biorender©.

Statistic and machine learning model combination for minimal gene signature identification of atopic dermatitis pruritus

Authors

Alain Lefèvre-Utile[1,2], Enzo Battistella[3,4], Ana B Pavel[5], Maria Vakalopoulou[4], Emma Guttman-Yasky[5], Nikos Paragios[4,6], Vassili Soumelis[1,7], MAARS consortium

Affiliations

[1] Université de Paris, Inserm, U976 HIPI Unit, Institut de Recherche Saint-Louis, F-75010, Paris, France

[2] Assistance Publique-Hôpitaux de Paris (APHP), General Pediatrics and Pediatric Emergency Department, Jean Verdier Hospital, Bondy, France

[3] Molecular Radiotherapy and Innovative Therapeutics, INSERM UMR1030, Gustave Roussy Cancer Campus, Paris-Saclay University, Villejuif, France

[4] Mathematics and Informatics for Complex, CentraleSupélec, Paris-Saclay University, 91190, Gif-sur-Yvette, France

[5] Laboratory of Inflammatory Skin Diseases, Department of Dermatology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[6] TheraPanacea, Paris, France

[7] Assistance Publique-Hôpitaux de Paris (AP-HP), Laboratoire d'Immunologie, Hôpital Saint-Louis, F-75010, Paris, France

MAARS Consortium

Juha Kere[a,b], Francesca Levi-Schaffer[c], Dario Greco[d,e], Noora Ottman[f], Jonathan Baker[g], Björn Andersson[h], Mauricio Barrientos-Somarribas[h], Stefanie Prast-Nielsen[i], Lukas Wisgrill[j], Sophia Tsoka[k], Nanna Fyhrquist[fl], Harri Alenius[fl], Helen Alexander[g], Jens M. Schröder[m], Frank O. Nestle[g], Antti Lauerma[n], Philippe Hupé[o], Annamari Ranki[n], Bernhard Homey[p]

MAARS Consortium affiliations

[a] Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

[b] Stem Cells and Metabolism Research Program, Folkhälsan Research Institute, University of Helsinki, Helsinki, Finland

[c] Pharmacology and Experimental Therapeutics Unit, Institute for Drug Research, School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel
[d] Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[e] Institute of Biotechnology, University of Helsinki, Helsinki, Finland
[f] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[g] St John's Institute of Dermatology, Faculty of Medicine and Life Sciences, Kings College London, London, United Kingdom
[h] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden
[i] Department of Microbiology, Tumour and Cell Biology, Karolinska Institutet, Stockholm, Sweden
[j] Division of Neonatology, Pediatric Intensive Care and Neuropediatrics, Comprehensive Center for Pediatrics, Department of Pediatrics and Adolescent Medicine; Medical University of Vienna, Austria
[k] Department of Informatics, Faculty of Natural and Mathematical Sciences, Kings College London, London, United Kingdom
[l] Department of Bacteriology and Immunology, Medicum, University of Helsinki, Helsinki, Finland
[m] Department of Dermatology, University Hospital Schleswig-Holstein, Kiel, Germany
[n] Department of Dermatology, Allergology, and Venerology, University of Helsinki, Helsinki University Hospital, Inflammation Centre, Helsinki, Finland
[o] INSERM U900, CNRS UMR144, Institut Curie, Mine Paris Tech, Paris, France
[p] Department of Dermatology, University of Duesseldorf, Duesseldorf, Germany

Keywords
Atopic dermatitis, Adults, Pruritus, Transcriptome, Machine Learning, Trafficking

Abstract

Pruritus is a major symptom of atopic dermatitis (AD) and causes an important burden for patients and society. Its mechanisms are complex and partly understood, making therapeutic perspectives promising. To address this question, we used the largest (n=82) available AD transcriptome lesional skin dataset (MAARS dataset). All patients auto-evaluated pruritus intensity using a visual scale going from 1 to 10. The median score was 7.

We first explore our data using correlation, differential analysis, and sparse PLS to conclude that more innovative approaches should be favored. We applied an automatic deep learning and statistical-based model using an ensemble of architectures, and a data-driven consensus for the gene selection and the pruritus prediction. Final minimalist signatures were obtained using an ablation study. Its application on our data revealed interesting genes for pruritus prediction: Heme Oxygenase 1 (HMOX1), Calcium/Calmodulin Dependent Serine Protein Kinase (CASK), Vestigial Like Family Member 2 (VGLL2), Mannosidase Alpha Class 2A Member 1 (MAN2A1), one long non-coding RNA (GPRC5D-AS1) and two novel transcripts (AC113382.1 and AL031123.1). It predicted pruritus classes with 0.77 balanced accuracy, 0.86 precision, 0.67 sensitivity, and 0.88 specificity. We validated our ensemble approach on two merged external cohorts, with n=70 samples in total. A new signature was designed, without gene in common with the previous one, but with similar prediction performance. Functional interpretation including both signatures showed interesting shared function and potential therapeutic targets. Our study is so far the first to apply ML to pruritus understanding, and encourage the use of innovative approaches for complex data comprehension.

Introduction

Atopic dermatitis (AD), i.e atopic eczema, is one of the most common inflammatory dermatitis worldwide. Depending on the geographical region, AD can affect up to 20% of children and 5% of adults[80,81]. Pruritus, defined as an unpleasant urge to scratch, is one of its major symptoms according to Hanifin and Rajka criteria[37]. It concerns 87-100% of patients[39] and places AD as one of the main causes of pruritus from dermatological origins[116] (Suppl. 1a).

Pruritus is a very debilitating symptom for AD patients, it alters the quality of life, sleeping, and concentration, it causes psychiatric disorders such as depression, anxiety, or helplessness, reduces self-esteem, and provokes pruritus-induced skin lesions[40,41] (Suppl. 1b).

But *why pruritus is so hard to treat?* A part of the answer lies in its complex pathophysiology involving:

1) The *exterior world* regrouping all itch-causing agents such as mechanical or chemical irritants, skin microbiome or infection, allergens.
2) The *interface* constituted by the epidermis, mainly composed of keratinocytes, lipids, and structural proteins.
3) The *inner world* which begins with the dermis and its rich immune microenvironment continues through nerves and spinal cord then ends with the pruritus signal interpretation in several cortex areas.

Each layer implies specific pruritogenic molecular actors e.g. cytokines, proteases, histamines, leukotrienes, or neuropeptides that constitute general and disease-specific mechanisms.

AD shares common mechanisms with other pruriginous diseases but has its specificities. Indeed, AD patients, have an unbalanced skin microbiome with an overrepresentation of *Staphylococcus aureus* that can secrete itching proteases[42]. The skin barrier is altered due to the downregulation of essential skin architecture protein (e.g. FLG, LOR)[117,118]. AD dermis immune environment is Th2, Th22, Th17 or TfH polarized with important concentrations of prurigenic cytokines such (e.g. IL4, IL13, or IL31)[20]. Finally, AD skin shows a higher density of nerves[119,120] that expresses receptors for all these prurigenic mediators[42] (Suppl. 1c).

All these molecular and cellular actors play a role in an itch-scratch circle[43] that antipruritic therapeutics try to break (Suppl. 1d). So far only anti-IL31 has been validated specifically for its anti-pruritic effect[121,122], but as a Th2 blocker, it is hard to determine the anti-pruritic effect independently. Other targeted treatments more focused on the symptom are emerging[123,124].

In this study, we took advantage of the largest skin microarray AD cohort. We hypothesize that machine learning (ML) will be an innovative way to understand better pruritus mechanism and identify new targeted treatments. While the benefit of ML over statistical models is still debated[51] we used an analysis pipeline that combines ML and statistical approaches to select the more consensual pruritus signature. We then applied the same method on an independent cohort and highlighted the common mechanism shared in both signatures.

Method

Learning cohort

The data were obtained from the MAARS Consortium[35,36] whose dataset is publicly available on the Array Express interface (E-MTAB-8149). Patient recruitment and data generation methodologies have been comprehensively described in Fyhrquist et al publication[35]. Briefly, AD patients have been recruited in three European Dermatology departments, after provided written informed consent under institutional review board-approved protocols. All AD patients met the Hanifin and Rajka criteria[37]. Sampling and data generation occurred between 2012 and 2013. Numbers of clinical features were collected, including visual auto-evaluation of the pruritus scale[41] (Fig. 1a). A 6 mm punch biopsy was performed in the lesional skin of AD patients. Bulk transcriptomic analysis was performed after mRNA extraction with Affymetrix GeneChip® Whole Transcript Expression Arrays.

External and independent cohort

To assess the reproducibility and robustness of our classifier we applied our prediction on two independent datasets. To reduce technological and technical biases, we sourced independent cohorts using a comparable transcriptomic technology and with available annotation on pruritus severity. Among the total number of pre-selected transcriptomic cohorts (n=48), only two studies met the above criteria (Suppl. 2a) with n = 30 and n = 40 AD lesional skin samples[125,126]. They were generated by the same team, using homogeneous protocols described in Bissonnette *et al* and *Pavel et al* studies. Pruritus intensity was evaluated by the patient using NRS (Numeric rating scale) (Suppl. 2b) Bulk transcriptomic data were generated using Affymetrix Human U133Plus 2.0® gene arrays. Expression matrices GSE133385 and GSE133477 were downloaded through the Gene expression omnibus (GEO) interface using *GEOquery* package[89] (ver. 2.51.1). Annotations on pruritus severity were shared by collaborators.

Expression array preprocessing

The MAARS dataset was loaded and analyzed with *R* language (version 3.6.0) on the *R Studio* interface (version 1.2.1335). As an exploratory step, we projected the dataset using Principal Component Analysis and performed several clustering methods. We discarded the sample *MAARS_3_070_03* as a potential mislabeled outlier. Thus, 82 AD lesional skin samples were used for further analyses on the training cohort. The two external cohorts have been merged so that 70 AD lesional skin samples were included in the testing cohort. Coding genes have been filtered in both cohorts, reducing the expression array respectively from 32,633 probes to 22,637 genes, and 30,409 to 18,588.

No outlier was excluded in both validation cohorts. Due to their perfect homogeneity, the two validation cohorts have been merged. We then selected probes at the intersection that were covered by all technologies. Thus, we used for validation analyses 12837 gene expression matrices.

Statistical models

To identify important genes in the pruritus mechanism we first used statistical models such as differential expression and correlation analyses.

Differential expressions were computed using *limma* R packages (3.42.2). Correlation analyses were conducted using the *psych* R package (ver. 2.0.9). For statistical significance, r > 0.3 and adjusted p-values < 0.05 were considered significant, Benjamini-Hochberg correction was applied for multiple testing. Graphical representations were designed using the *ggplot2* R package (ver. 3.3.1).

sPLS was performed using the *mixomics* R package (ver. 6.13.3)[127]. sPLS regression mode was used to identify a combination of variables able to explain relationships between expression array and pruritus score. Repeated k-fold cross-validation was performed to optimize the number of dimensions using the Q2 value (calculated as 1 – (Predictive residual Error Sum of Squares/ Total Sums of Squares)), and Mean Absolute Error (MAE) were computed for final gene selection.

Functional enrichment has been done using the *g:Profiler* interface (https://biit.cs.ut.ee/gprofiler/gost) using GO: Biological process terms. Adjusted p-values < 0.01 (Benjamini-Hochberg correction) were considered as significant.

Predictive genes' selection

As classical statistical models did not show a definitive conclusion, we then used machine learning approaches. Using all the coding genes considered, we built a high-dimension space of size 22 637. A min-max normalization of the attributes was performed for the training and validation cohorts. The same values were also applied to the test set.

To tackle the dimensionality curse and discover significant and robust predictive genes for the pruritus score, we adapted our in-house feature selection pipeline[45]. First, applying a space dimension reduction step, which is of prime importance especially in genomics studies[46].

We separated the samples' cohort into two classes: low (< median i.e. 7) and high pruritus (≥ 7). The cohort was subdivided into training and test on the principle of 80%-20% maintaining the observed distribution of classes between the two subsets. Then, on this basis, the training set was further divided into 5 subdivisions to perform feature selection. We evaluated a variety of classical machine learning classifiers - using the entire feature space and 4 subdivisions for training - such as Decision Tree Classifier, Linear Support Vector Machine, XGBoosting, AdaBoost, and Lasso. These classifiers were trained and validated to distinguish between pruritus classes. Besides, we considered statistics-based approaches based on Mutual Information, Chi-squared statistics, and Univariate linear regression tests.

Each of these methods was used to assess the importance of the features regarding pruritus. Features were ranked according to their selection prevalence for each method. Our experiments indicated that different classifiers highlight different attributes as important. We adopted a consensus approach choosing features with the highest sum of prevalence (>40%) over all methods.

Classification task

The classification was addressed using an ensemble learning approach on the gene signature designed in the feature selection step. The same training/test sets as the ones for feature selection were used. We performed 5-fold cross-validation and evaluated the average performance of the following supervised classification methods: Nearest Neighbor, Linear, Sigmoid, Radial Basis Function (RBF), Polynomial Kernel Support Vector Machines (SVM), Gaussian Process, Decision Trees, Random Forests, AdaBoost, XGBoosting, Gaussian Naive Bayes, Bernoulli Naive Bayes, Multi-Layer Perceptron (MLP) & Quadratic Discriminant Analysis. For each binary classification task, a consensus model was designed selecting the top 5 classifiers. The selected models were trained and combined through a winner takes all approach to determine the optimal outcome. The final prediction was performed thanks to a majority voting scheme.

Signature refinement

A step of signature refinement was then performed thanks to ablation. In the same cross-validation settings as before, we iteratively trained the ensemble classifier on the training set using the signature genes except one. Then, we removed the gene which ablation incurred the best-averaged results on validation. This process was repeated until no gene remained. The final retained signature was designed by considering all the genes after the inflection point, according to the elbow method.

Then, the selected classifiers were retrained using the entire training set and the refined signature, and their performance was reported on the test set.

Implementation details

Concerning the implementation details, for the majority voting classifier, the top 5 classifiers consist of RBF SVM, Linear SVM, Polynomial SVM, QDA, and MLP. The RBF SVM had a penalty parameter of 0.7 and a kernel coefficient gamma of 1. The Linear kernel had a penalty parameter of 3. The Polynomial SVM was granted a kernel degree of 2. The QDA classifier was considered without any prior or regularization parameter and with an absolute threshold of 10. The MLP classifier was trained with a lbfgs solver, an alpha of 0.1, a ReLU

activation a maximal number of iterations of 1000, a batch size of 500, and an invscaling learning rate. To prevent overfitting, early stopping was used.

Results

Explore clinical and transcriptomic data

The learning cross-sectional cohort contained 82 AD lesional skin samples. All AD patients had auto-evaluated their pruritus intensity using the pruritus Visual Rating Scale (VRS) ranging from 1 to 10 (Fig. 1a). Pruritus intensity distribution was characterized by, mean = 6.4, median = 7, interquartile = 5-8, a minimum = 1, and a maximum = 10 (Fig. 1b).

All patients had lesional skin transcriptome arrays, including 22637 coding genes per sample. We wondered if a relationship between gene expressions and pruritus score could be established. We first projected genetic data using Principal Component Analysis (PCA). Pruritus score did not seem associated with the two first principal components (Fig. 1c). Pearson correlations were computed for any of the 30 first dimensions, and none was significantly correlated with pruritus score (r^2 were between -0.3 <> 0.3, and adjusted p-value > 0.05) (data not shown). This led us to use statistical models to identify interesting genes in the pruritus mechanism.

Statistical models: differential expression and correlation analyses

We first computed differential expression analyses. For well-balanced comparisons, we separated the cohort into two groups centered on the median pruritus score, with n= 42 with pruritus score < 7 (*pruritus low*) and n = 40 with pruritus score ≥ 7 (*pruritus high*). Differential expression analyses showed no significant differences between the two populations with all adjusted p-values >0.05. Similar results were obtained with different group comparisons, like first vs last third (data not shown). Differential expression comparison was not adapted to assess the pruritus questioning our cohort.

We then wanted to take advantage of the pruritus score's continuous nature. We computed correlation analyses with Spearman and Pearson statistics to establish what would be the best to highlight relationships between pruritus intensity and gene expressions. Both technics showed strengths and weaknesses. Spearman statistics revealed a restricted list of pruritus-correlated genes (with r coefficient -0.3 < or > 0.3), but without statistical significance

(all adjusted p-values were above 0.05) (Fig. 2a,b).On the opposite, Pearson statistics showed an extensive gene list positively (n=879 genes) and negatively correlated (n=808 genes) with pruritus intensity (Fig. 2b). A functional annotation based on *GO: Biological process* terms was applied on genes correlated with pruritus score. It revealed that positively correlated genes were highly enriched in immune function while negatively correlated genes were enriched in neuronal function (Suppl. 3a, b).

Pearson and Spearman's statistics were complementary to reveal interesting biological functions about pruritus. But with both approaches, the too-large dimension of our dataset provoked the selection of inappropriate gene numbers for biological interpretation (Fig. 2c).

sPLS: a mixed approach between statistical models and ML models

As dimensionality reduction was required, we used sparse partial least square (sPLS) to apply a gene's number reduction supervised on pruritus intensity score. It consists of achieving variable selection by introducing Lasso penalization on the pair of loading matrices, here expression array, and pruritus score. We first applied the sPLS algorithm on all genes' expression dataset (Fig. 3a). The two first dimensions explained 5 and 4% of the total variance and well-segregated samples according to their pruritus score. The next crucial step was about tuning the sPLS model to reduce the number of variables. To do so we repeated 20 times 10-fold cross-validation. We selected the optimal number of sPLS components that was defined as the lowest dimension with Q2 level ≥0.0975[44](Fig. 3b). Based on the lowest MAE obtained on both components, the optimal number of variables was respectively 15 and 5 genes (for MAE=0.400 and 0.385) (Fig. 3c). The sPLS algorithm was then rerun with optimized parameters (2 components and 20 selected genes) supervised on the pruritus score. Variance per component was comparable according to the first run (4 and 2%), meaning that little information has been lost, and the pruritus score was well distributed (Fig. 3d). Finally, we ordered the gene of interest according to their weight in the prediction of pruritus score (Fig. 3e).

Predictive genes' signature: selection and performances

To add robustness to previous approaches we decided to use an analysis pipeline combining ML and statistical model. This can reduce gene number, and predict pruritus categories in a consensual manner, taking advantage of multiple approaches. To do so we categorize our in-house cohort into two categories: low pruritus (score <7) and high pruritus (score ≥7) with respectively n= 40 and 42 samples. Relying on the aforementioned selection method, we extracted 22 genes and obtained after ablation a minimalist signature composed of the 7 following genes (Fig. 4a): Heme Oxygenase 1 (HMOX1), Calcium/Calmodulin Dependent Serine Protein Kinase (CASK), Vestigial Like Family Member 2 (VGLL2), Mannosidase Alpha Class 2A Member 1 (MAN2A1), one long non-coding RNA (GPRC5D-AS1) and two novel transcripts (AC113382.1 and AL031123.1).

Our proposed ensemble approach reported high performance overall considered evaluation metrics in intra-cohort validation (Fig. 4b). With only a 7 genes signature, we reached on test 0.77 balanced accuracy, 0.86 precision, 0.67 sensitivity, 0.88 specificity. Also, we manage to correctly classify 87.5% of the low pruritus class' samples and 66.67% for the high pruritus class (Fig. 4c).

Ensemble learning approach applied on an external and independent cohort

The genes included in the external cohorts massively varying from the ones of the MAARS dataset, the previously identified gene signature could not be tested in those new settings. Thus, the same pipeline was repeated on our external cohort to demonstrate the generalizability of our approach despite the microarray technology disparity. A very different signature was obtained, again including 7 genes: Nuclear Transcription Factor/X-Box Binding Like 1(NFXL1), TOX High Mobility Group Box Family Member 2 (TOX2), Transcription Factor Like 5 (TCFL5), Synaptosome Associated Protein 23 (SNAP23), one long non-coding RNA (ENSG00000279064), and one novel transcript (AC011815.3) (Fig. 5a).

Notwithstanding, the differences between the cohorts, we managed excellent results on the test. With only a 7 genes signature, we reached on test 0.90 balanced accuracy, 0.90 precision, 1.00 sensitivity, 0.80 specificity (Fig. 5b). Also, we manage to correctly classify 80.0% of the low pruritus class' samples and 100% for the high pruritus class(Fig. 5c).

Discussion

Our molecular signature may seem destabilizing for several reasons. Our pipeline was designed so that the minimal number of genes would be selected. This had the advantage of highlighting the essential substance of pruritus prediction. But it makes functional interpretation harder. In our case, our predictive genes could orientate AD pruritus treatment development. The recourse to the drug repositioning databases did not show convincing results but manually PubMed research revealed interesting tracks to follow.

sPLS signature and the Spearman statistics top 10 positively correlated genes (Fig. 3e, Suppl. 3c), contained candidates for pruritus comprehension and treatment. A special interest has to be brought to CRTC (CREB Regulated Transcription Coactivator), 2 and 3. These genes are part of the CREB complex (cAMP Responsive Element Binding Protein) which acts as an activator of ERK (Extracellular Signal-regulated Kinase) cascade inflammatory skin pruritus and pain[128,129] and could be a potential therapeutic target[130].

Our ensemble approach showed different gene signatures within in-house and external cohorts. Nevertheless, we excluded novel transcript and long non-coding RNA to extracted common functions from better-characterized genes. Interestingly, CASK, SNAP23, and IFT46 were implied in vesicle trafficking[131,132]. This could be central in pruritus mechanisms in particular considering mast cell degranulation, cytokine emission by lymphocytes, and cell-nerve interaction at the synapse level. Interesting treatments are developed targeting prurigenic meditators trafficking, like botulin toxin[133], and others could follow, as CASK inhibitors. HMOX1 appeared in both ML and Pearson correlation gene selection. Positively associated with pruritus intensity, its expression could be a response to tissue injury and act as a protective factor[134], and it can be upregulated upon specific therapeutics[135].

In this study, we discovered that combining ML and statistical models led to robust pruritus predictions. Our minimalistic gene signatures highlight potential molecular actors of pruritus mechanisms. Surprisingly, our gene of interest did not overlap with current knowledge about AD pruritus. This recall us that predicting is not explaining. Indeed, none of our genes belongs to classical AD pathways such as Th2,17,22 polarization nor skin architecture[14]. But variables selected to predict are not always indicative of a mechanism logic. Mamaprint®, which is the only transcriptomic prognosis signature with a direct impact

on the patient, includes 70 genes, and a large majority is not already described in the breast cancer pathophysiology[10,11].

To allow ML model application, we categorized our pruritus score and thus lost its continuous nature. We divide our cohort into two groups by whether they were above or below the median pruritus score. But due to the inclusion criteria that excluded low severity AD, low pruritus scores were then underrepresented and the median was high (=7/10). This could have attenuated the expression contrast between our two groups and made the work of the classifier harder. Although independent validation cohorts were selected because of their similarities with our learning cohort (Suppl. 2c). Unfortunately, our results were not validated on an external cohort. In general, this can be due to disparities in patient recruitment (age, gender, race disparities), sampling procedure (anatomical localization, skin preparation), microarray technologies, and platform protocols. In our case, differences between learning and validation cohorts are subtle and could be due to various sample anatomical localizations and the diverse Affymetrics® microarray generations. A recent study using the same data as ours showed that gene expressions differed according to the anatomical localization in the AD context[36]. It highlights the importance of standardized sampling procedures within skin transcriptomic studies. The technical bias might be even strongest in our case. As genome coverage is not homogeneous among Affymetrix® technologies[136], focusing on common genes led to biological information loss.

Beyond technology disparities that make validation steps more difficult. One potential issue is the lack of a consensual and routinely used pruritus scale. In this study, we used two simple ways to evaluate pruritus intensity: the VRS and the NRS (Fig. 1a and Suppl. 2b). These two auto-evaluation scales ranging from 1 to 10 have their subtilities and can be discordant[137]. But beyond that, they point out the difficulty of measuring a symptom with a strong subjective component. Many ways to explore and quantify this symptom exist (Suppl. 4a)[41] and a consensual research-oriented pruritus scale still needs to be designed and used. As we develop a pruritus classifier based on complex data, it might be time to use emerging approaches to objectively assess pruritus in atopic dermatitis by multidimensionality scale[138].

ML models have advantages over statistical ones as they can identify multicollinear and complex relationships in multidimensional data. Yet, these approaches have not been used extensively in AD fields. ML has been applied for histological-based AD diagnosis[139]. Recently, AD severity prediction, using a knowledge-based set of biomarkers and intern cross-validation

has been performed in a cross-sectional manner[140,141]. Personalized prediction of disease severity based on prospective clinical followings[142] showed interesting results. ML superiority on statistical models is still debated[51]. When compared in previous examples, ML did not outperform statistical models[140]. Moreover, studies are tempted to overuse the term ML to designate more "classical" clustering approaches, without learning step. To be convinced of ML contribution in AD understanding, purely data-driven studies with external validation are still missing. In our study, we conducted a feature selection capable of accurate predictions in our cohort and an independent one.

Pearson and Spearman correlations can be used complementarily[143]. In our case, both approaches showed interesting results but did not assess the pruritus issue. With the increasing use of high throughput data, analysis methods need to adapt. While traditional statistical models still dominate the field, even in omics data. Omics data began to exceed the capabilities of the conventional statistical model[144]. Many machine learning methods can derive models for pattern recognition, classification, and prediction from complex data[145]. It appears nevertheless less effective in producing explicit models with biological significance. In our study, we chose to compare different approaches, and even if ML superiority was not clear, its use should be more generalized to update progressively the way we analyze complex data.

Perspectives

We aimed to understand deeper pruritus mechanism in the AD context. To do so we used statistical and machine learning models in a complementary manner. It led to identifying important transcriptomic signatures that bring a new light on pruritus and suggest targeted therapeutic development.

Figures and legends

Figure 1

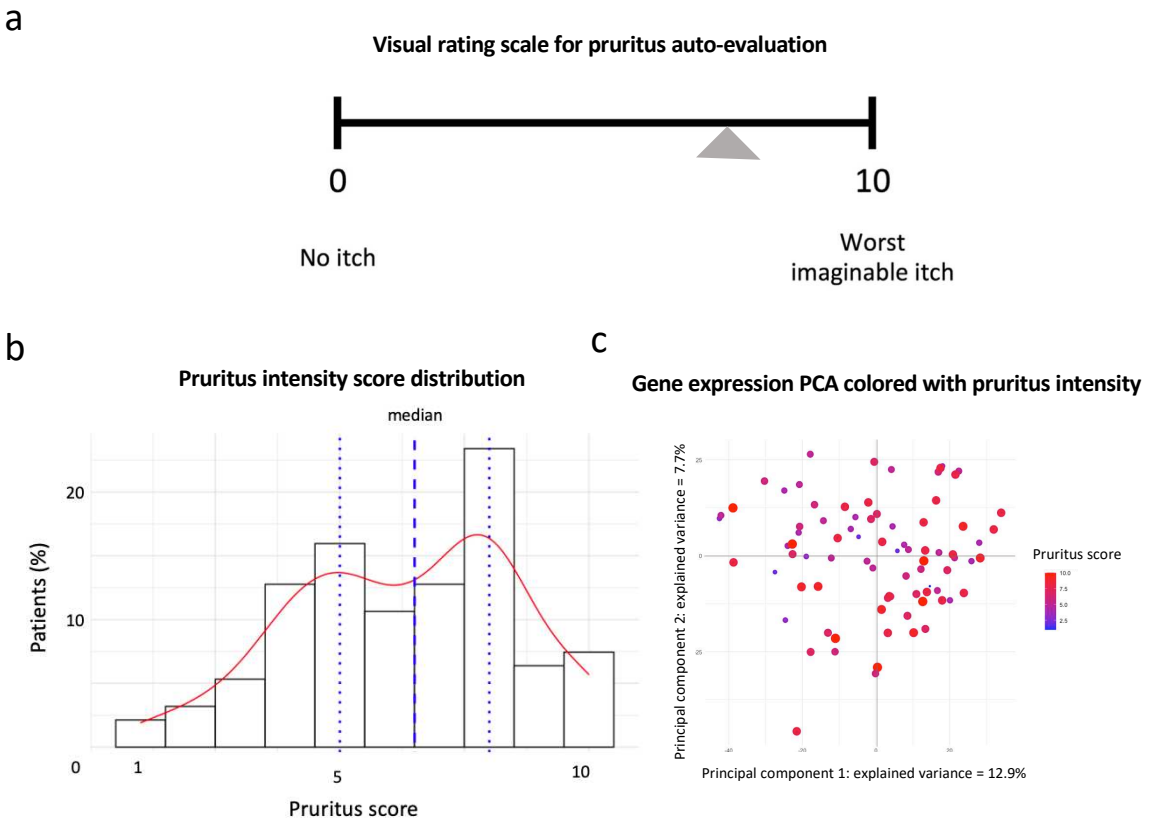


Figure 1: Exploration of pruritus score in the learning in-house cohort

a: Visual scale for pruritus auto-evaluation.

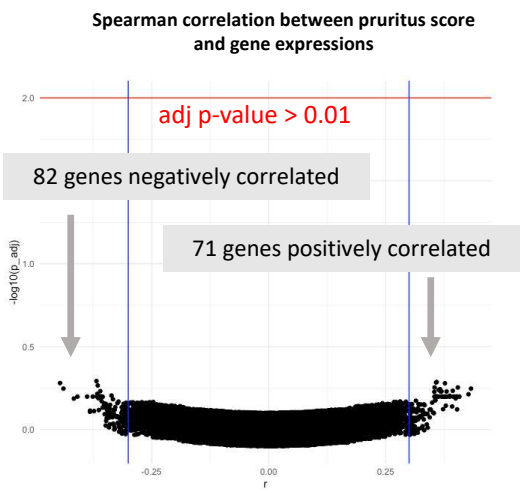
b: Pruritus distribution among learning cohort patients.

c: PCA representation of whole cohort genetic data colored with pruritus intensity. PCA two first dimensions poorly illustrate pruritus' intensity distribution.

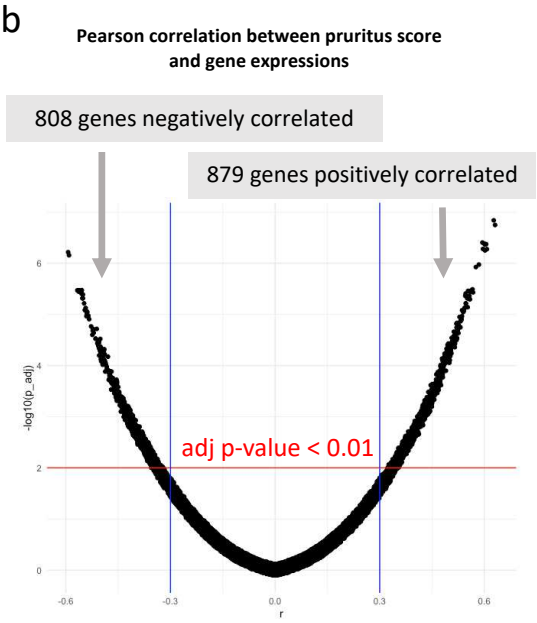
PCA: Principal component analysis.

Figure 2

a


Spearman correlation between pruritus score and gene expressions
adj p-value > 0.01
82 genes negatively correlated
71 genes positively correlated
-log10(p_adj)
r

b


Pearson correlation between pruritus score and gene expressions
808 genes negatively correlated
879 genes positively correlated
adj p-value < 0.01
-log10(p_adj)
r

c

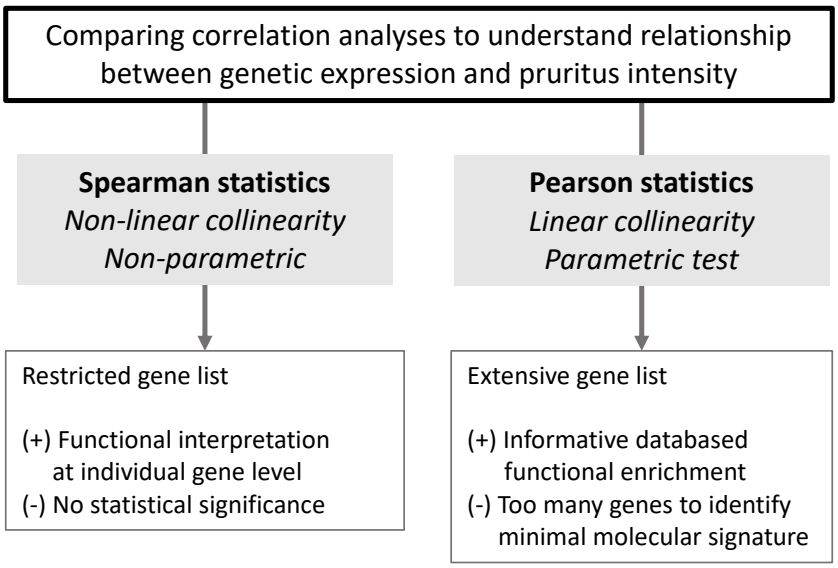

Comparing correlation analyses to understand relationship between genetic expression and pruritus intensity
Spearman statistics
Non-linear collinearity
Non-parametric
Pearson statistics
Linear collinearity
Parametric test
Restricted gene list
(+) Functional interpretation at individual gene level
(-) No statistical significance
Extensive gene list
(+) Informative databased functional enrichment
(-) Too many genes to identify minimal molecular signature

Figure 2: Comparison of Spearman and Pearson's correlations between pruritus intensity and gene expression.

a: No correlation between pruritus intensity and gene expression according to Spearman statistics. Few genes have correlation coefficients above 0.3 but none showed significance.

b: Correlation between pruritus intensity and gene expression according to Pearson statistics. A large number of genes have correlation coefficients above 0.3 and high significance.

c: Both correlation models' contribution points to the need for dimension reduction.

Figure 3

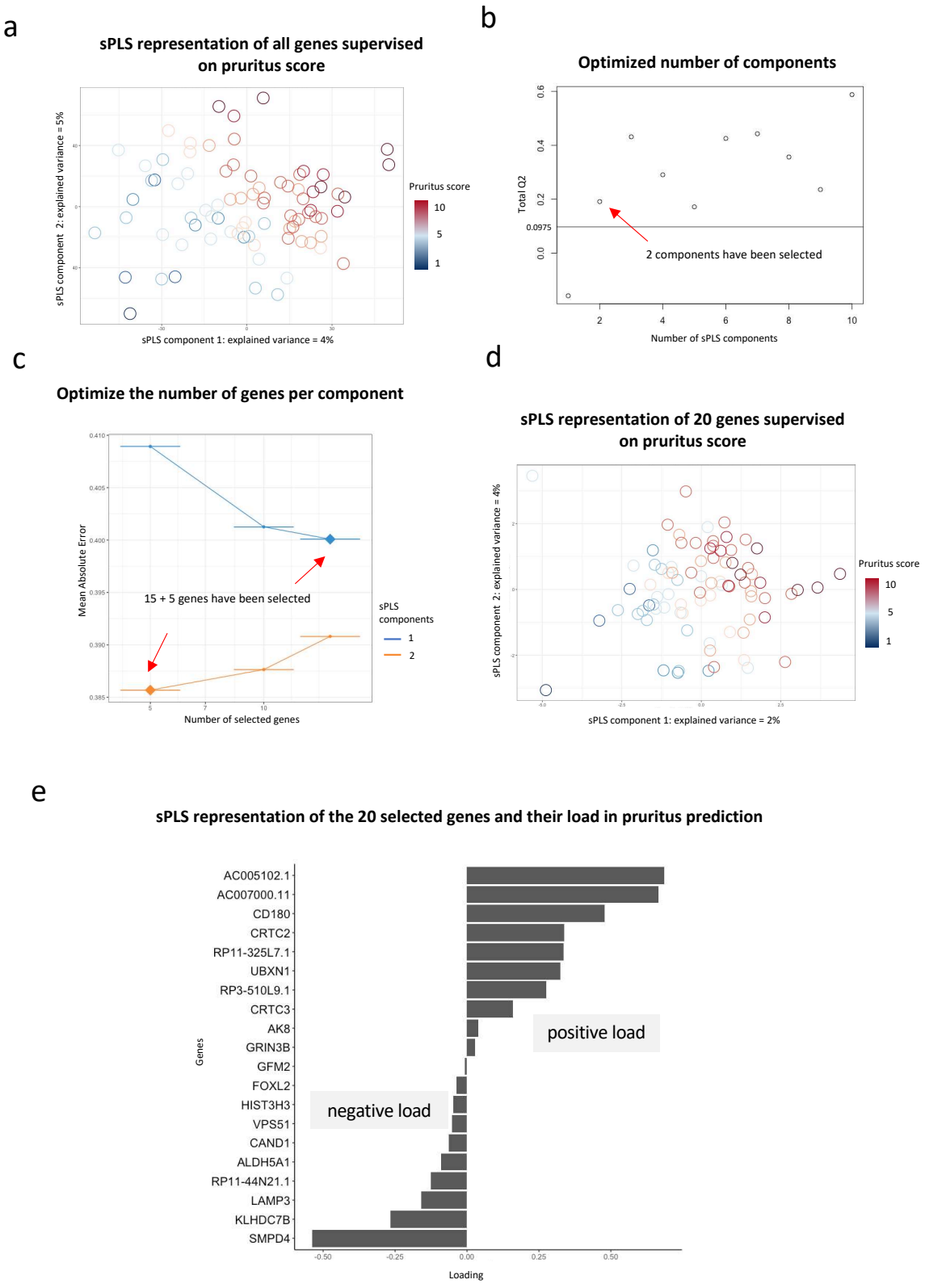

a
sPLS representation of all genes supervised on pruritus score
sPLS component 2: explained variance = 5%
sPLS component 1: explained variance = 4%
Pruritus score
10
5
1
b
Optimized number of components
Total Q2
0.0975
2 components have been selected
Number of sPLS components
c
Optimize the number of genes per component
Mean Absolute Error
15 + 5 genes have been selected
sPLS components
1
2
Number of selected genes
d
sPLS representation of 20 genes supervised on pruritus score
sPLS component 2: explained variance = 4%
sPLS component 1: explained variance = 2%
Pruritus score
10
5
1
e
sPLS representation of the 20 selected genes and their load in pruritus prediction
AC005102.1
AC007000.11
CD180
CRTC2
RP11-325L7.1
UBXN1
RP3-510L9.1
CRTC3
AK8
GRIN3B
GFM2
FOXL2
HIST3H3
VPS51
CAND1
ALDH5A1
RP11-44N21.1
LAMP3
KLHDC7B
SMPD4
Genes
positive load
negative load
Loading

Figure 3: sPLS model optimize dimension and gene reduction for pruritus estimation

a: Representation of all genes supervised on pruritus score, on sPLS first two dimensions.

b: Optimized number of components according to Q2 value. 10-fold cross-validation repeated 20 times allow reduction to 2 components.

c: Optimized number of genes according to MAE. 15 genes from 1st component and 5 from 2nd component was selected.

d: New run of sPLS model optimized with 2 components and 20 genes. Pruritus score is still well segregated.

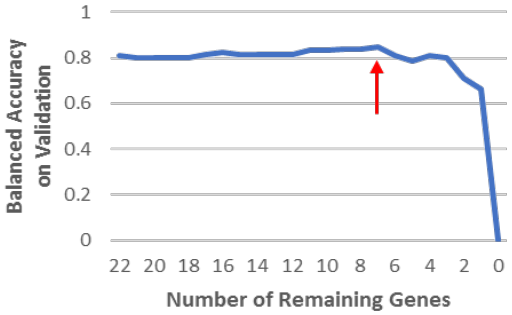
e: Ranked genes by load values in pruritus estimation.

MAE: Mean Average Error, sPSL: sparse Partial Least Square

Figure 4

a

Number of Genes	Removed Feature	Balanced accuracy
22	ENSG00000138944 – SHISAL1	0,812
21	ENSG00000184210 – DGAT2L6	0,8
20	ENSG00000167969 – ECI1	0,8
19	ENSG00000089916 - GPATCH2L	0,8
18	ENSG00000063761 - ADCK1	0,8
17	ENSG00000215405 - GOLGA6L6	0,815
16	ENSG00000129646 - QRICH2	0,826
15	ENSG00000188171 – ZNF626	0,814
14	ENSG00000143502 – SUSD4	0,815
13	ENSG00000005238 – FAM214B	0,815
12	ENSG00000120708 - TGFBI	0,814
11	ENSG00000102910 – LONP2	0,835
10	ENSG00000162949 - CAPN13	0,835
9	ENSG00000059377 – TBXAS1	0,836
8	ENSG00000162877 – PM20D1	0,836
7	ENSG00000052802 – MSMO1	0,849
6	**ENSG00000100292 – HMOX1**	**0,811**
5	**ENSG00000246323 - AC113382.1**	**0,788**
4	**ENSG00000147044 - CASK**	**0,811**
3	**ENSG00000170162 – VGLL2**	**0,799**
2	**ENSG00000226281 - AL031123.1**	**0,708**
1	**ENSG00000247498 - GPRC5D-AS1**	**0,664**
0	**ENSG00000112893 – MAN2A1**	**0**



b

Classifier	Balanced Accuracy		Precision		Sensitivity		Specificity	
	Training	Test	Training	Test	Training	Test	Training	Test
Linear SVM	0,91	0,65	0,89	0,67	0,94	0,67	0,88	0,62
poly SVM	0,94	0,77	0,94	0,86	0,94	0,67	0,94	0,88
RBF SVM	0,89	0,65	0,89	0,67	0,91	0,67	0,88	0,62
Neural Net	0,99	0,77	1	0,86	0,97	0,67	1	0,88
QDA	0,88	0,65	0,86	0,67	0,91	0,67	0,84	0,62
Ensemble Classifier	**0,94**	**0,77**	**0,94**	**0,86**	**0,94**	**0,67**	**0,94**	**0,88**

c

		Predicted Pruritus	
		< 7	**≥ 7**
Actual Pruritus	**< 7**	7 (87.5%)	1 (12.5%)
	≥ 7	3 (33.3%)	6 (66.7%)

Figure 4: Feature ablation and test result details on the MAARS

a: Gene signature before and after ablation study on MAARS cohort. The table presents the balanced accuracy on validation for each removed gene. Genes are ordered from the least predictive to the most predictive. The curve represents balanced accuracy according to the number of genes. The inflection point is evidenced by the red arrow.

b: Test results on MAARS cohort. Predictive performance over the different metrics considered for the classifiers selected on cross-validation to design the ensemble classifier. Classifiers have been retrained on the full training set using the signature obtained by selection and ablation.

c: Confusion matrix. Prediction of pruritus classes (columns) according to the actual classes (rows) performed on the MAARS cohort part that has not been used for ablation study nor classification.

a

Figure 5

Number of Genes	Removed Feature	Balanced accuracy
19	ENSG00000185483 - ROR1	0,775
18	ENSG00000118514 - ALDH8A1	0,79
17	ENSG00000105497 - ZNF175	0,805
16	ENSG00000114770 - ABCC5	0,836
15	ENSG00000186532 - SMYD4	0,874
14	ENSG00000279967 - FP671120.5	0,874
13	ENSG00000279186 - FP236315.2	0,845
12	ENSG00000118473 - SGIP1	0,831
11	ENSG00000279303 - CU634019.3	0,865
10	ENSG00000258634 - AL160006.1	0,865
9	ENSG00000264720 - MIR3117	0,879
8	ENSG00000115827 - DCAF17	0,874
7	ENSG00000011143 - MKS1	0,874
6	**ENSG00000170448 - NFXL1**	**0,861**
5	**ENSG00000124191 - TOX2**	**0,827**
4	**ENSG00000101190 - TCFL5**	**0,829**
3	**ENSG00000279064 - FP236315.1**	**0,765**
2	**ENSG00000092531 - SNAP23**	**0,714**
1	**ENSG00000279989 - AC011815.3**	**0,675**
0	**ENSG00000118096 - IFT46**	**0**

b

Classifier	Balanced Accuracy		Precision		Sensitivity		Specificity	
	Training	Test	Training	Test	Training	Test	Training	Test
Linear SVM	0,85	0,9	0,92	0,9	0,81	1	0,88	0,8
poly SVM	0,9	0,9	0,97	0,9	0,84	1	0,96	0,8
RBF SVM	0,85	0,9	0,9	0,9	0,86	1	0,85	0,8
Neural Net	0,96	0,9	0,96	0,9	1	1	0,92	0,8
QDA	0,87	0,9	0,92	0,9	0,86	1	0,88	0,8
Ensemble Classifier	**0,91**	**0,9**	**0,93**	**0,9**	**0,93**	**1**	**0,88**	**0,8**

d

		Predicted Pruritus	
		< 7	**≥ 7**
Actual Pruritus	**< 7**	4 (80.0%)	1 (20.00%)
	≥ 7	0 (0%)	9 (100.00%)

<u>Figure 5:</u> Feature ablation and test result details on the independent and external cohort

a: Gene signature before and after ablation study on the external cohort. The table presents the balanced accuracy on validation for each removed gene. The genes are ordered from the least predictive to the most predictive.

b: Test results on the external cohort. Predictive performance over the different metrics considered for the classifiers selected on cross-validation to design the ensemble classifier. Classifiers have been retrained on the full training set using the signature obtained by selection and ablation.

c: Confusion matrix. Prediction of pruritus classes (columns) according to the actual classes (rows) performed on the external cohort part that has not been used for ablation study nor classification.

Supplementary figure and legends

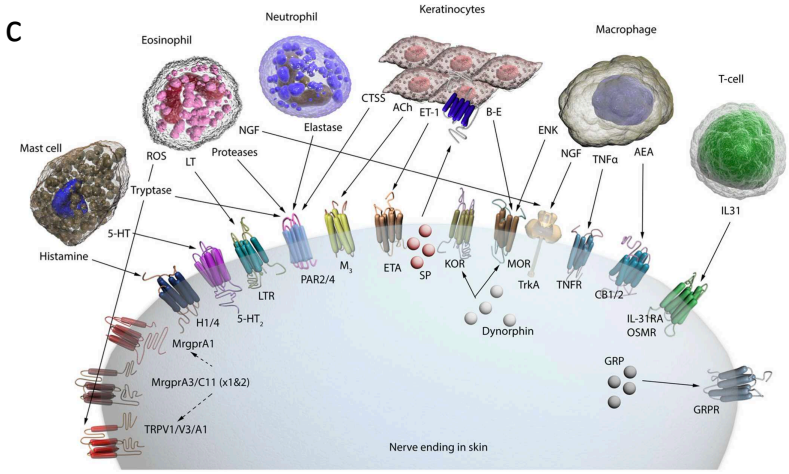
Supplementary 1

a

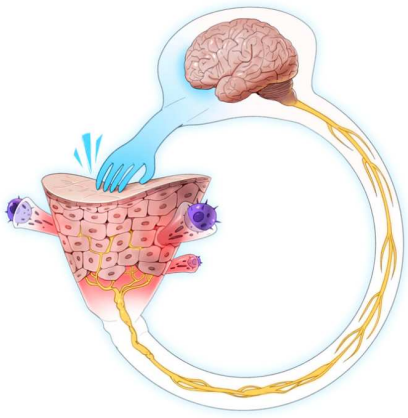
Dermatological diseases	Examples of diagnoses
Inflammatory dermatoses	Atopic dermatitis, psoriasis, contact dermatitis, dry skin, drug reactions, scars, ”invisible dermatoses”
Infectious dermatoses	Mycotic, bacterial and viral infections and folliculitis, scabies, pediculosis, arthropod reactions, insect bites
Autoimmune dermatoses	Bullous dermatoses, especially dermatitis herpetiformis Duhring, bullous pemphigoid, dermatomyositis
Genodermatoses	Darier’s disease, Hailey-Hailey disease, ichthyoses, Sjögren-Larsson syndrome, EB pruriginosa
Dermatoses of pregnancy	Polymorphic eruption of pregnancy, pemphigoid gestationis, prurigo gestationis
Neoplasms	Cutaneous T-cell-lymphoma (especially erythrodermic variants), cutaneous B-cell-lymphoma, leukaemic infiltrates of the skin

b

c



d



<u>Supplementary 1</u>: Pruritus overview in dermatoses and atopic dermatitis

a: Main causes of pruritus among dermatoses. From *Pereira et al*.

b: Photographs of pruritus lesion in AD patient hands. From *Pereira et al*.

c: Cellular and molecular actors of pruritus signal and their neuronal receptors. From *Mollanazar et al*.

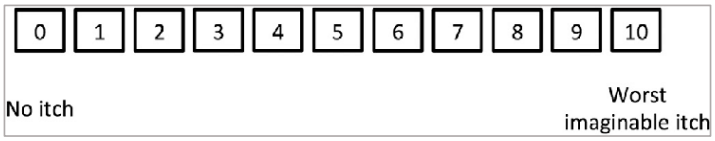
d: Pruritus vicious circle implying skin, immunity, and nervous system. From *Cevikbas and Lerner.*

Supplementary 2

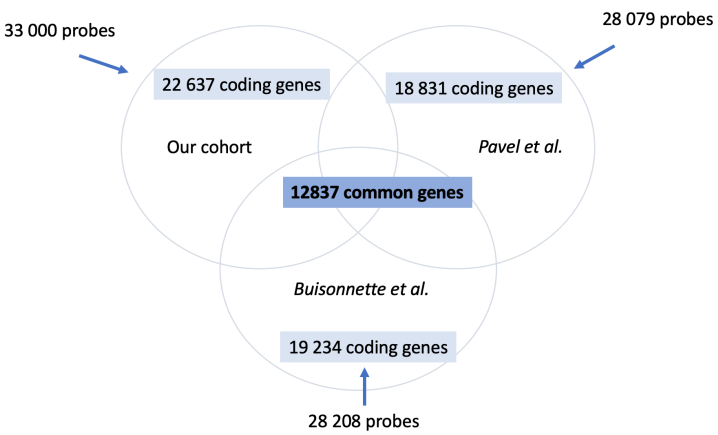
a

	Pavel et al.	Bissonnette et al.	Our cohort
DOI	10.1016/j.jaci.2019.07.013	10.1016/j.jaci.2019.06.047	
Journal	JACI	JACI	
Year	2018	2019	
Patient recruitment			
Centers	Multicentric (North America)	Monocentric (North America)	Multicentric (North Europe)
Age	> 18 yo	> 18 yo	>18 yo
Majoritary skin color	White	White	White
Hanifin and Rajka criteria	NA	Yes	Yes
Active AD	Yes	Yes	Yes
Topical washout	2 w	2 w	2 w
Systemic washout	12 w	4 w	12 w
AD severity	Moderate to severe	Mild to moderate	Moderate to severe
Pruritus severity (/10) med[min-max]	7 [1-10]	6 [2-10]	7[1-10]
Technical aspect			
number of samples	30	40	82
Anatomical site	Various	Various	Standardized
Microarray technology	Affymetrix Human U133Plus 2.0	Affymetrix Human U133Plus 2.0	Affymetrix® GeneChip® Whole Transcript
Data availability	GSE133385	GSE133477	E-MTAB-8149

b



c



<u>Supplementary 2</u>: In-house and external cohort comparability

a: Clinical and demographical comparison of our in-house cohort and two external cohorts.

b: Numeric Rating Scale (NRS) for pruritus evaluation in both external cohorts.

c: Gene intersection of in-house and external cohorts. Due to technology disparities, only 12837 coding genes are covered by all microarrays.

Supplementary 3

a

Top 10 functional enrichment of significant positively correlated genes

GO:BP		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
myeloid leukocyte activation	GO:0002274	1.224×10^{-16}	
myeloid cell activation involved in immune response	GO:0002275	7.108×10^{-16}	
leukocyte degranulation	GO:0043299	8.212×10^{-15}	
cell activation	GO:0001775	1.054×10^{-14}	
myeloid leukocyte mediated immunity	GO:0002444	2.906×10^{-14}	
neutrophil activation involved in immune response	GO:0002283	3.765×10^{-14}	
leukocyte activation involved in immune response	GO:0002366	4.783×10^{-14}	
granulocyte activation	GO:0036230	5.491×10^{-14}	
neutrophil degranulation	GO:0043312	5.491×10^{-14}	
leukocyte activation	GO:0045321	5.491×10^{-14}	

b

Top 10 functional enrichment of significant negatively correlated genes

GO:BP		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
cell junction organization	GO:0034330	1.053×10^{-4}	
synapse organization	GO:0050808	1.494×10^{-4}	
cellular component morphogenesis	GO:0032989	2.113×10^{-4}	
cell morphogenesis	GO:0000902	6.004×10^{-4}	
plasma membrane bounded cell projection morphogene...	GO:0120039	6.004×10^{-4}	
cell projection morphogenesis	GO:0048858	6.004×10^{-4}	
neuron projection morphogenesis	GO:0048812	6.004×10^{-4}	
cell part morphogenesis	GO:0032990	6.004×10^{-4}	
cell morphogenesis involved in neuron differentiation	GO:0048667	6.660×10^{-4}	
generation of neurons	GO:0048699	1.167×10^{-3}	

c

Top 20 positively correlated genes with pruritus score

Ensembl	Symbol	r	p_adj
ENSG00000160741	CRTC2	0,43	0,57
ENSG00000246323	AC113382.1	0,41	0,63
ENSG00000247925	AL139807.1	0,40	0,63
ENSG00000069702	TGFBR3	0,40	0,63
ENSG00000234199	LINC01191	0,40	0,63
ENSG00000176658	MYO1D	0,39	0,63
ENSG00000197191	CYSRT1	0,39	0,63
ENSG00000237815	Unknow	0,39	0,63
ENSG00000140577	CRTC3	0,38	0,63
ENSG00000112214	FHL5	0,38	0,63
ENSG00000152377	SPOCK1	0,38	0,63
ENSG00000100292	HMOX1	0,37	0,63
ENSG00000116032	GRIN3B	0,37	0,63
ENSG00000240032	LNCSRLR	0,37	0,63
ENSG00000180818	HOXC10	0,36	0,63
ENSG00000258435	AC048337.1	0,36	0,63
ENSG00000165695	AK8	0,36	0,63
ENSG00000215305	VPS16	0,36	0,63
ENSG00000103942	HOMER2	0,36	0,63
ENSG00000213612	FAM220CP	0,36	0,63

d

Top 20 negatively correlated genes with pruritus score

Ensembl	Symbol	r	p_adj
ENSG00000178445	GLDC	-0,34	0,76
ENSG00000170162	VGLL2	-0,35	0,76
ENSG00000164347	GFM2	-0,35	0,73
ENSG00000086696	HSD17B2	-0,35	0,70
ENSG00000130487	KLHDC7B	-0,35	0,70
ENSG00000187554	TLR5	-0,36	0,63
ENSG00000185875	THNSL1	-0,36	0,63
ENSG00000174327	SLC16A13	-0,36	0,63
ENSG00000234614	C2CD4D-AS1	-0,36	0,63
ENSG00000143502	SUSD4	-0,36	0,63
ENSG00000147573	TRIM55	-0,36	0,63
ENSG00000186510	CLCNKA	-0,36	0,63
ENSG00000111530	CAND1	-0,36	0,63
ENSG00000231584	FAHD2CP	-0,36	0,63
ENSG00000205037	AC134312.1	-0,37	0,63
ENSG00000252980	RNU6-367P	-0,37	0,63
ENSG00000183770	FOXL2	-0,38	0,63
ENSG00000247498	GPRC5D-AS1	-0,39	0,63
ENSG00000257556	LINC02298	-0,41	0,63
ENSG00000223648	IGHV3-64	-0,44	0,57

Supplementary 3: Functional interpretation of correlated genes with pruritus. Even with too low or too much significance, correlated genes showed interesting functions.

a: Top 10 functional enrichment according to GO: Biological process using all positively correlated genes with pruritus according to Pearson correlation. Immune functions are overrepresented.

b: Top 10 functional enrichment according to GO: Biological process using all inversely correlated genes with pruritus according to Pearson correlation. Neuronal functions are overrepresented.

c: Detailed list of top 20 genes positively correlated with pruritus score according to Spearman correlation.

d: Detailed list of top 20 genes negatively correlated with pruritus score according to Spearman correlation.

Supplementary 4

a An overview of tools to measure pruritus. From Pereira *et al*

	Tool
Intensity	*Monodimensional:* • Visual analogue scale • Numerical rating scale • Verbal rating scale *Multidimensional:* • Itch Severity Scale • Pruritus Grading System
Scratch lesions	• Scratch Symptom Score • Prurigo Activity Scale
Scratching activity	• Actigraphy • Accelerometer
Course of pruritus	• Dynamic Pruritus Score • Itch-Free Days • 5-D Scale • Patient Benefit Index • ItchApp®
Psychiatric comorbidities	• Hospital Anxiety and Depression Scale • Beck Depression Inventory • Hamilton Rating Scale for Depression
Sleep impairment	• Stanford Sleepiness Scale • Epworth Sleepiness Scale • Athens Insomnia Scale
Quality of life	• ItchyQol • Dermatological Life Quality Index • 36-item short form

Supplementary 4: Overview of tools used to measure pruritus. From Pereira et al.

GENERAL DISCUSSION & PERSPECTIVES

GENERAL DISCUSSION

The trans-disciplinarity of this thesis calls for multiple discussions. I chose to develop three questions that have been raised by this work. First will be the constraints and expectations of classification work. Second will be the need for open data science, with its pros and cons. Last will be the difficult application of complex data-based discoveries to patient management, and the physician position in the understanding and decision process.

Classification issues

Create a temporary and unperfect object

Contemporary classification approaches rely on the ability to correlate observed features with pathological states to define syndromes. Throughout the last century, this approach became more objective, as the molecular underpinnings of many disorders were identified and definitive laboratory tests became an essential part of the overall diagnostic paradigm[3].

In many aspects, the ambition of molecular classifying still faces obstacles such as data quality, method performances, results reproducibility, and overall real-life interpretation. It is difficult being definitive while knowing that current classifications will sooner or later be broken and replaced by newer ones. What makes this field so unstable is the lack of a community-wide, consensus-based, human- and machine-interpretable language for describing phenotypes, in their genomic and environmental contexts. This could be the most pressing scientific bottleneck to develop robust classification integrating many biological key fields[146].

From reductionism's pitfall to comprehensiveness illusion

Current disease classifications and medical diagnosis are the direct consequence of inductive generalization predicated on Occam's razor. This parsimony principle reduces

variable number to reach a limited number of possible classes making diagnosis more feasible. But doctors' reliance on Cartesian reductionism in establishing diagnoses may appear insufficient for out-of-square situations[3].

As the quality and quantity of biological data are growing, this simplification may now appear simplistic. Even in a monogenic disorder such as sickle cell disease, whose mechanisms seem well defined, clinical presentations are heterogeneous and intermediate phenotypes frequent[147]. This is due to polygenic interferent mechanisms and host-environment interactions that current diagnosis classification hardly considers. To be able to identify new mechanisms, Systems biology proposes to model disease as a complex network integrating data's modules of different nature (clinic, environmental, omics, *etc…*), connected according to a certain probabilistic strength (Figure 15). Although more comprehensive, these dense graphical representations are often hard to understand for clinicians and barely applicable.

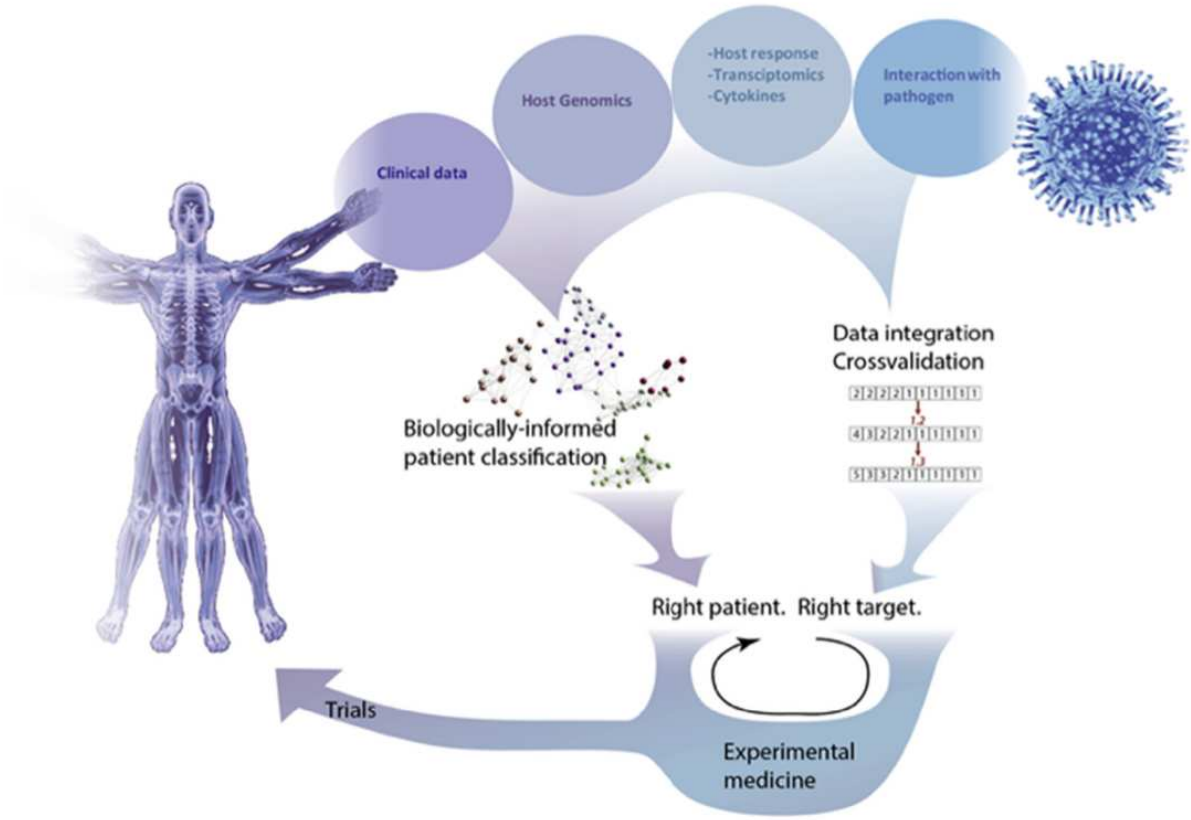


***Figure 15* Summary of a systems medicine approach to infection**. *Data sources can be combined to obtain clinically-informative phenotyping of patients, and identification of therapeutic targets. From Russel et al.*

Classification biological interpretation and validation

In a biologic research project, the validation step is the point that strengthens the discovery, and classifying is not an exception. In omics-based classification, this step is rarely reached as opposed to more classical experimental biology where triplicates can robustly underly a strong biological effect between two conditions. With omics data, the large scale allows to identify more subtle mechanisms, but the noise background is also more important. Added to this potential smaller cohort size due to the cost issue and the statistical robustness is difficult to maintain during the validation steps.

A first possibility could be to go back to the bench and to validate with more routine experiments a selection of relevant biomarkers. Recently, has been designed an AD skin transcriptome classification supervised on patient age and validated biomarker existence, at the protein level, using *in situ* assays[148]. This is a secure way to convince the reader of the replicability of results but, due to variables biased selection, this removes a part of the picture, the complex one.

Moreover, this is often done at the expense of replicating the results on an independent cohort which is a far too rare step in the omics field, for some reasons. Due to rapid technical evolution, experimental protocol or computational scripts cannot be applied to all publicly available cohorts. Besides, the recruitment and sampling procedure can differ from the original and challenge the result verification. In our project, finding an independent cohort with comparable patients and technologies was hard and only one, for the unsupervised classification, and two cohorts, for the supervised classification, met these criteria and had enough patients [38,125,126].

The ideal validation is hard to design and requires the constitution of an independent prospective cohort. And it has been done only once in AD with serum-biomarkers-based classification[33,34]. In breast cancer, this took a decade but opened the unique opportunity for the first transcriptome-based classification signature to be applied at the patient bedside and recognize by public health authorities[10,11].

Seeking a semantic consensus

Last but not least, the scientific community should begin classification issues clarification with semantic harmonization. Over time, all disciplines have added up their terminology, bringing semantic richness but also confusion. This issue has been asked for a long time[149] however, in current research, no effort has been made to define the terms they use. The literature overview that was done for this work showed that the classification objects could be: clustered / classified / subclassified / divided /sub-divided / stratified, into: classes / sub-classes / groups / sub-groups / entities / clusters / types / endophenotypes … This wealth of terminology reflects a willingness of rigor but could also make the message less accessible for non-expert readers. So far attempts in terminology harmonization are field-specific (e.g. hierarchical classification of species, Figure). But due to the multidisciplinary influences, a global harmonization, if desired, would be a long process.

Added to this is the ambiguity that results from certain statistical terms, which can confuse the clinician. In the same way that *association* does not mean *causality*. Machine learning makes use of the term *predict* which does not have the meaning of projection into the future that it implies. In this context, *predict* is used as *estimate*. To make a true prediction, the estimation should be confirmed prospectively.

To make classification articles more understandable for biologists and clinicians, the semantic harmonization effort must be initiated simultaneously at the biological and technical levels.

Recycling data, a dilemma of consciousness

We surely can try to assess original questions with already used data. Provided that its quality allows it, it relies on the inspiration of the researcher to recycle them for new purposes.

The moral obligation of sharing published data

Open science refers to the virtuous circle of making research based on shared data while publishing results to the community. This increases transparency in the research process, confidence in findings, and facilitates reproducibility[150]. This could take different forms such as give access to published datasets or actively collect them about a specific topic in a collaborative manner. Many initiatives are growing this way, such as Global Forest Biodiversity Initiative, an international research collaboration, that contains information about more than 1 million locations. They are publicly available, stored in CSV plain-text files, therefore accessible to all, and have already been used for high-impact research projects[50] .

Good practice data management is now defined as the FAIR (Findable, Accessible, Interoperable, and Reusable) principles[47]. And many scientific journals have adopted policies that encourage or require data sharing. But still, many collaborators support open data in principle but have a specific reason for keeping their collections private. It reflects the current state of Science: partly open and partly closed, still driven by competitive emulation more than collaboration[151]. Associating as co-authors those who share their data could be a solution to encourage generous comportments.

All fields are not equal in the sharing process. Especially in medical research, data generation could be a long process, dependent on ethical aspects. Thus, the volume of data can be limited in comparison to other fields. Practically, anonymous clinical outputs are rarely publicly available as they can be reused at the research team level in other projects. But omics field sets the example, as datasets are increasingly shared[48,152].

Example of great medical discoveries based on shared data

About AD classifications, significant discoveries have been done thanks to publicly available microarray pooling, such as the MADAD transcriptomic signature[31]. In this project,

the merging of several microarray datasets allowed the authors to obtain sufficient statistical power to design a disease molecular signature. Unfortunately, no international open database exists for AD clinical information. Therefore, important AD translational studies using clinical and omics features are rare and remain at the exclusive disposal of the team that generated them.

About the vaccine follow-up survey, there is the Vaccine Adverse Event Reporting System (VAERS) which is a publicly available dataset (https://vaers.hhs.gov/) where United States' health practitioners can declare any adverse effects related to a vaccine. Motivated by an ideal of transparency and sharing, reports are at the disposal of the scientific community, thus serving as a high dimensional reference dataset for research projects. A four years phase IV survey of quadrivalent influenza vaccine has been published recently, with a comprehensive adverse event overview. It underlined the poor allergenic nature and the benefits-risks ratio of the vaccine[49].

The Covid-19 crisis, raised the obvious need for global collaboration[153] especially in areas suffering from a lack of resources[154]. Open data has become a reality, even for the public, with live epidemiological statistics diffusion on platforms such as www.covidtracker.fr or www.data.gouv.fr. These efforts saw the birth of many international data sharing collaborative initiatives and allowed the publication of biological[155] and clinical[156] disease phenotype that increased rapidly our knowledge on this emerging disease.

Limits of data sharing

Practicing Science from underused data has many advantages. It is an ecological and almost free way to make discoveries. In our project, the MAARS cohort, data generation cost more than 7 million euros. Only one person, with one computer, is needed to reuse them. Also, it allows increasing the sample size in studies merging several datasets[157]. This approach is notably adapted to high throughput data whose generation is often not driven by a closed-ended question so they can be used for many other problematics. But even if original data generation has a cost, we must not deny the financial and ecological impact of storage and indexation for them to be reuse as *second-hand* data.

As technologies are moving fast, old data could be out of date, they became hard to use and to validate discoveries based on recent technologies. In our project, the validation of the machine learning pruritus predictive signature was not easy because it required comparable transcriptomic technologies. And even with almost similar technologies, subtle differences can make validation impossible. While validating results on other types of data could strengthen them, it can also add technological biases hard to identify and correct. It is generally recommended, when possible, to start from the raw data to avoid biases related to the pre-processing stages.

Above technological constraints, data quality increases when they are generated and analyzed in an identical environment. *Can it be assumed that the differences in study populations, data collection and analysis, and treatments, both protocol-specified and unspecified, can be ignored?*[151] Thus, researchers planning to use publicly available data should have skills in data management, curation, and quality control to avoid using poor quality data. Lastly, recycling data could have the perverse effect of slowing down the generation of original data, putting a brake on innovation. But as we generate more and more data, this is currently not the point.

Researchers have to be conscious of the strengths and weaknesses of their data. To do so, the sharing spirit should concern clinical and technical meta-data as well as omics data. Thus, an informed researcher will be able to judge and criticize its results, with the final objective of applying these discoveries in practice.

Are complex-data-based discoveries lost in translation?

High throughput data take more and more part of research discoveries. They can be analyzed as unique layers or as multiple and integrated. Ironically, they remain underused in real-life clinical practice. We will develop here the increasing implication of complex data-based-discoveries on patient bedsides.

Impact of diagnostic, prognostic, and therapeutic classifications for patients

Diagnoses, prognoses, and therapeutics are often interrelated so classifications could address several of these aspects. Omics data contributed to discovering major disease subtypes. Next genome sequencing applied on AD population revealed the important role of FLG protein in skin barrier integrity. Indeed, heterozygous loss of function mutations of *FLG* gene, carried by 9% of the European population, increases the disease risk of 3 to 5 times. It defined a mutated patient class with earlier age of onset, chronic and more severe evolution[22,66]. While this classification strongly highlighted one of the main disease mechanisms, it does not lead to therapeutical intervention. Thus, FLG characterization at the DNA level is not a part of standard care. In the cancer field, several blood-based combinatorial proteomic biomarker assays have been recently developed to assess the breast biopsy indication in women routine screening. While it has not been approved by public health instances, this kind of approach might have a high impact on breast cancer diagnosis[158].

Endotyping complex diseases in biological-based classes support the revelation of actionable therapeutic targets. It is susceptible to give information about treatment efficacy and tolerance. AD targeted treatment, is currently indicated to patients with moderate-to-severe AD (exclusively based on clinical criteria). This is a revolution in patient therapeutic strategies, although there remains a significant proportion of non-responders. Omics-based studies are designed to demonstrate treatment biological efficacity[38,125,126,159] or to identify good-response biomarkers[160,161]. However, no robust longitudinal study for treatment response prediction has been done in AD. Prediction of treatment responses is not easy and has to deal with various problems such as how to define a response or responder, what are clinically relevant outcome measures, and what should be the timing of response evaluation.

Classifications are often based on a unique level of biological information stratified using statistical models. With the increasing ability to generate even more high throughput data, multi-omics data integration using machine learning strategies seems promising.

When artificial intelligence integrates data: predict instead of understand

Complex data require complex methods. Data layer addition forces analyses strategies to update constantly and to grow in complexity. That is why machine learning is increasingly used in medical research: because the machine has access to a precision that humans cannot see. As an example, drug repurposing (or repositioning) is a cost-effective approach for revealing drugs that can be used to treat diseases for which they are currently not prescribed. In inflammatory skin conditions, a machine learning algorithm has been designed to model drug-disease relationships taking into account drugs whose effects were already known. Then, was attributed a supposed effectiveness to drugs and identified potential therapeutics[162]. This generated original hypotheses to facilitate future treatment development, reducing cost and time expense.

While machine learning shows promising perspectives, its implementation in medical research has two main paradoxes. The first would be the culture shock between both disciplines. Biologists and clinicians are seeking biological meaning and interpretability. To do so they need a certain amount of information to be able to group variables, as genes, in common molecular pathways or functions. On the opposite, mathematicians and statisticians aim to design the more minimalistic signature for more accurate prediction. Collaboration between both sides requires dialogue and compromise ability. The second paradox would be the necessity of a minimal sample quantity concerning the variable number. Indeed, the machine learning models we used, estimated their accuracies using intra-cohort cross-validation. So, as for statistical models, its robustness will strongly depend on cohort size and the sample/variable ratio. Thus, while complex data generation is costly, machine learning analyses might require an increased number of samples, with increased expends.

Complexification has its detractors. Simple methods might perform almost as well as more sophisticated ones[163]. Why a smart black box that produces hardly interpretable results should be better than more classical tools that Statistics made three centuries to develop. A recent systematic review showed no global performance benefit of machine learning over logistic

regression for binary clinical prediction models[51]. As statistics and machine learning are part of a unique entity: data science, future analysts should learn the different facets of this discipline and combine statistical and machine learning approaches as we did in our supervised analysis.

Good old clinic for real-life application, until the advent of *data physician*

Clinical decision support system has been intended to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information[164]. It can be knowledge-based or not, thus requiring IA tools. Its aims are as diverse as diagnosis and treatment management, cost containment, administrative task automation, *etc*. All medical aspects can be concerned by this assistance, but so far, it has not been used with a direct impact on patient care. Indeed, ethical issue about responsibility and transparency of the decisions made by such systems remains[165] .

As the final choice stands in the physicians' hands, they have to be educated on Data Science, its strengths, and weaknesses. They should overcome their inferiority complex which leads them to a blind validation or a total rejection when they deal with scientific papers based on complex data. Thus, these innovative approaches should help them refine their clinical intuition rather than vanishing it.

On the other hand, data scientists should show curiosity towards clinical data, share their thinking with the project referring clinician, and simplify their discovery in the most interpretable concepts. To do so, study protocols should strive to validate their findings with routine gold-standard approaches such as rtPCR for transcriptomic studies or immunohistochemistry for proteomics.

Interaction between Data and Medical Sciences has become a staple. It has announced a revolution probably as important the contribution of Statistic contribution to evidence-based medicine. A peaceful and balanced collaboration between data scientists and medical doctors shall be a prerequisite to finally hope for personalized medicine.

PERSPECTIVES

In the field of complex diseases

The two parts of thesis results could be informative in other complex disease contexts for their methodological aspects. Feature selection based on variance disparities between physiological and pathological states appeared as a simple and logical way to reduce dimensionality while keeping the maximum of biological information. It could be used to assess disease heterogeneity issues using unsupervised clustering. After being used in the Covid context, our analysis pipeline combining statistical and machine learning models showed interesting results in identifying symptom mechanisms. The complementarity of the different tools it includes makes it a more flexible pipeline, more easily applicable in another complex disease context.

In the field of AD

Our findings revealed unknown facets of atopic dermatitis. Our four skin-transcriptomic-based endotypes were related to mechanisms which were, until now, poorly considered. These non-canonical pathways should be taken into account to develop new therapeutics that could be used alone or in addition to standard treatments. These endotypes should be searched in independent cohorts with the gene signature we designed using targeted gene expression arrays. The role of the IL-36 pathway in disease severity suggests that anti-IL36 biologics should be reconsidered in the AD therapeutic arsenal, at least for patients belonging to the IL-36 dependent endotype.

Pruritus remains a complex symptom. Our results revealed the unsuspected role of vesicle trafficking and suggest that therapeutic development should be oriented in this direction. As pruritus is shared in variety of diseases, these new mechanisms should be screened in other

itching conditions such as *prurigo nodularis*, or others, to determine whether they are AD-specific or not.

In the team

Although they are publicly available, the MAARS data (transcriptomics, metagenomics and clinics) are still stored in the team server and could be inspiring for novel research projects. Given the wealth of clinical data, original supervised questions can be addressed in AD cohort. Moreover, data about psoriasis are important and of good quality while similar analyses should be applied to decipher psoriasis heterogeneity. That is probably what I would have done if I have had encouraging feedback on the AD side, as an early first publication.

Moreover, classifying is now a new pilar of the team activity. We are now part of an important European consensus whose aim is to classify inflammatory disorders. I hope that this work has started a good momentum in this direction.

At the personal level

Independently to the global context, that has been hard for every human being in the past year, and has darkened this second half of the thesis, the main challenge for me was to take possession of the omics culture while bringing it to my medical culture.

I do not pretend to become an expert in data analysis. But I have lost a part of my inferiority complex when I apprehend a medical research paper that deals with omics data. I discovered an open-minded research spirit, more collaborative and transparent than clinical research. I collaborated with passionate and rigorous data scientists that shared their expertise with me for this thesis but also in parallel projects (such as Data for Good). I did not become a *data physician*, I don't know if this even exists, but I feel I can be at the interface between both cultures, to be able to participate in medical research projects based on complex data.

BIBLIOGRAPHY

1. Tirosh, I., Bilu, Y. & Barkai, N. Comparative biology: beyond sequence analysis. *Current Opinion in Biotechnology* **18**, 371–377 (2007).
2. World Health Organization. International Classification of Diseases 11th Revision.
3. Loscalzo, J., Kohane, I. & Barabasi, A. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. *Mol Syst Biol* **3**, 124 (2007).
4. Beatson, G. Tn the Treatment of Inoperable Cases of Carcinoma of the Mamma: Suggestions for a New Method of Treatment, with Illustrative Cases. *Trans Med Chir Soc Edinb* **15**, 153–179 (1896).
5. McGuire, W. L. & Chamness, G. C. Studies on the estrogen receptor in breast cancer. *Adv Exp Med Biol* **36**, 113–136 (1973).
6. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177–182 (1987).
7. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98**, 10869–10874 (2001).
8. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
9. Domagala, P., Huzarski, T., Lubinski, J., Gugala, K. & Domagala, W. PARP-1 expression in breast cancer including BRCA1-associated, triple negative and basal-like tumors: possible implications for PARP-1 inhibitor therapy. *Breast Cancer Res Treat* **127**, 861–869 (2011).
10. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
11. Cardoso, F. *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).
12. Langan, S. M., Irvine, A. D. & Weidinger, S. Atopic dermatitis. *The Lancet* **396**, 345–360 (2020).
13. Weidinger, S., Beck, L. A., Bieber, T., Kabashima, K. & Irvine, A. D. Atopic dermatitis. *Nat Rev Dis Primers* **4**, 1 (2018).
14. Weidinger, S. & Novak, N. Atopic dermatitis. *The Lancet* **387**, 1109–1122 (2016).
15. Silverberg, N. B. Typical and atypical clinical appearance of atopic dermatitis. *Clin. Dermatol.* **35**, 354–359 (2017).
16. Silverberg, J. I. *et al.* Phenotypical Differences of Childhood- and Adult-Onset Atopic Dermatitis. *The Journal of Allergy and Clinical Immunology: In Practice* **6**, 1306–1312 (2018).
17. Silverberg, J. I. Comorbidities and the impact of atopic dermatitis. *Annals of Allergy, Asthma & Immunology* **123**, 144–151 (2019).
18. Yew, Y. W., Thyssen, J. P. & Silverberg, J. I. A systematic review and meta-analysis of the regional and age-related differences in atopic dermatitis clinical characteristics. *J Am Acad Dermatol* **80**, 390–401 (2019).
19. Stefanovic, N., Flohr, C. & Irvine, A. D. The exposome in atopic dermatitis. *Allergy* **75**, 63–74 (2020).
20. Carmi-Levy, I., Homey, B. & Soumelis, V. A modular view of cytokine networks in atopic dermatitis. *Clin Rev Allergy Immunol* **41**, 245–253 (2011).
21. Szabó, K. *et al.* Expansion of circulating follicular T helper cells associates with disease severity in childhood atopic dermatitis. *Immunology Letters* **189**, 101–108 (2017).
22. Kim, B. E. & Leung, D. Y. M. Significance of Skin Barrier Dysfunction in Atopic Dermatitis. *Allergy Asthma Immunol Res* **10**, 207–215 (2018).
23. Hogan, M. B., Peele, K. & Wilson, N. W. Skin barrier function and its importance at

the start of the atopic march. *J Allergy (Cairo)* **2012**, 901940 (2012).
24. Somanunt, S., Chinratanapisit, S., Pacharn, P., Visitsunthorn, N. & Jirapongsananuruk, O. The natural history of atopic dermatitis and its association with Atopic March. *Asian Pac J Allergy Immunol* **35**, 137–143 (2017).
25. Johansson, S. G. *et al.* A revised nomenclature for allergy. An EAACI position statement from the EAACI nomenclature task force. *Allergy* **56**, 813–824 (2001).
26. Reinhold, U., Kukel, S., Goeden, B., Neumann, U. & Kreysel, H. W. Functional characterization of skin-infiltrating lymphocytes in atopic dermatitis. *Clinical & Experimental Immunology* **86**, 444–448 (1991).
27. Wüthrich, B. Atopic neurodermatitis. *Wien Med Wochenschr* **139**, 156–165 (1989).
28. Czarnowicki, T., He, H., Krueger, J. G. & Guttman-Yassky, E. Atopic dermatitis endotypes and implications for targeted therapeutics. *Journal of Allergy and Clinical Immunology* **143**, 1–11 (2019).
29. Gooderham, M. J., Hong, H. C., Eshtiaghi, P. & Papp, K. A. Dupilumab: A review of its use in the treatment of atopic dermatitis. *Journal of the American Academy of Dermatology* **78**, S28–S36 (2018).
30. Haute Autorité de Santé. Comission de transparence avant acceptation de mise sur le marché du Baricitinim. (2021).
31. Ewald, D. A. *et al.* Meta-analysis derived atopic dermatitis (MADAD) transcriptome defines a robust AD signature highlighting the involvement of atherosclerosis and lipid metabolism pathways. *BMC Medical Genomics* **8**, (2015).
32. Ghosh, D. *et al.* Multiple Transcriptome Data Analysis Reveals Biologically Relevant Atopic Dermatitis Signature Genes and Pathways. *PLOS ONE* **10**, e0144316 (2015).
33. Thijs, J. L. *et al.* Moving toward endotypes in atopic dermatitis: Identification of patient clusters based on serum biomarker analysis. *Journal of Allergy and Clinical Immunology* **140**, 730–737 (2017).
34. Bakker, D. S. *et al.* Confirmation of multiple endotypes in atopic dermatitis based on serum biomarkers. *Journal of Allergy and Clinical Immunology* S0091674920308010 (2020) doi:10.1016/j.jaci.2020.04.062.
35. Fyhrquist, N. *et al.* Microbe-host interplay in atopic dermatitis and psoriasis. *Nat Commun* **10**, 4703 (2019).
36. Ottman, N. *et al.* Microbial and transcriptional differences elucidate atopic dermatitis heterogeneity across skin sites. *Allergy* all.14606 (2020) doi:10.1111/all.14606.
37. Hanifin, J.M., R., G. Diagnostic Features of Atopic Dermatitis. *Acta Derm. Venereol.* (1980).
38. Guttman-Yassky, E. *et al.* Dupilumab progressively improves systemic and cutaneous abnormalities in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology* **143**, 155–172 (2019).
39. Dawn, A. *et al.* Itch characteristics in atopic dermatitis: results of a web-based questionnaire. *Br J Dermatol* **160**, 642–644 (2009).
40. Schmitt, J. *et al.* Usage and effectiveness of systemic treatments in adults with severe atopic eczema: First results of the German Atopic Eczema Registry TREATgermany. *J Dtsch Dermatol Ges* **15**, 49–59 (2017).
41. Pereira, M. P. & Ständer, S. Assessment of severity and burden of pruritus. *Allergol Int* **66**, 3–7 (2017).
42. Mollanazar, N. K., Smith, P. K. & Yosipovitch, G. Mediators of Chronic Pruritus in Atopic Dermatitis: Getting the Itch Out? *Clinic Rev Allerg Immunol* **51**, 263–292 (2016).

43. Cevikbas, F. & Lerner, E. A. Physiology and Pathophysiology of Itch. *Physiol Rev* **100**, 945–982 (2020).
44. Tenenhaus, Michel. *La régression PLS: théorie et pratique.* (1998).
45. Chassagnon, G. *et al.* AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. *Med Image Anal* **67**, 101860 (2021).
46. Battistella, E. *et al.* Cancer Gene Profiling through Unsupervised Discovery. *arXiv:2102.07713 [cs, q-bio]* (2021).
47. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
48. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).
49. Woo, E. J. & Moro, P. L. Postmarketing safety surveillance of quadrivalent recombinant influenza vaccine: Reports to the vaccine adverse event reporting system. *Vaccine* **39**, 1812–1817 (2021).
50. Liang, J. *et al.* Positive biodiversity-productivity relationship predominant in global forests. *Science* **354**, aaf8957–aaf8957 (2016).
51. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* **110**, 12–22 (2019).
52. Oramas, S., Barbieri, F., Nieto, O. & Serra, X. Multimodal Deep Learning for Music Genre Classification. *Transactions of the International Society for Music Information Retrieval* **1**, 4–21 (2018).
53. Rouhan, G. & Gaudeul, M. Plant Taxonomy: A Historical Perspective, Current Challenges, and Perspectives. *Methods Mol Biol* **2222**, 1–38 (2021).
54. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). (2013).
55. Wakefield, J. C. Diagnostic Issues and Controversies in DSM-5: Return of the False Positives Problem. *Annu Rev Clin Psychol* **12**, 105–132 (2016).
56. Ozen, S. *et al.* EULAR/PRINTO/PRES criteria for Henoch-Schonlein purpura, childhood polyarteritis nodosa, childhood Wegener granulomatosis and childhood Takayasu arteritis: Ankara 2008. Part II: Final classification criteria. *Annals of the Rheumatic Diseases* **69**, 798–806 (2010).
57. Ranson, J. H. Etiological and prognostic factors in human acute pancreatitis: a review. *Am J Gastroenterol* **77**, 633–638 (1982).
58. Harbeck, N. & Gnant, M. Breast cancer. *Lancet* **389**, 1134–1150 (2017).
59. Szymiczek, A., Lone, A. & Akbari, M. R. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. *Clin Genet* **99**, 613–637 (2021).
60. Kristensen, V. N. *et al.* Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* **14**, 299–313 (2014).
61. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).
62. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
63. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365**, 1687–1717 (2005).
64. Pegram, M. D. *et al.* Phase II study of receptor-enhanced chemosensitivity using

recombinant humanized anti-p185HER2/neu monoclonal antibody plus cisplatin in patients with HER2/neu-overexpressing metastatic breast cancer refractory to chemotherapy treatment. *JCO* **16**, 2659–2671 (1998).
65. Brunner, P. M. & Guttman-Yassky, E. Racial differences in atopic dermatitis. *Annals of Allergy, Asthma & Immunology* **122**, 449–455 (2019).
66. Palmer, C. N. A. *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nature Genetics* **38**, 441–446 (2006).
67. Liang, Y., Chang, C. & Lu, Q. The Genetics and Epigenetics of Atopic Dermatitis—Filaggrin and Other Polymorphisms. *Clinical Reviews in Allergy & Immunology* **51**, 315–328 (2016).
68. Lyons, J. J. & Milner, J. D. Primary atopic disorders. *Journal of Experimental Medicine* **215**, 1009–1022 (2018).
69. Paller, A. S., Spergel, J. M., Mina-Osorio, P. & Irvine, A. D. The atopic march and atopic multimorbidity: Many trajectories, many pathways. *J Allergy Clin Immunol* **143**, 46–55 (2019).
70. Carlsten, C. *et al.* Atopic dermatitis in a high-risk cohort: natural history, associated allergic outcomes, and risk factors. *Ann Allergy Asthma Immunol* **110**, 24–28 (2013).
71. Schmitt, J. *et al.* Atopic dermatitis is associated with an increased risk for rheumatoid arthritis and inflammatory bowel disease, and a decreased risk for type 1 diabetes. *J. Allergy Clin. Immunol.* **137**, 130–136 (2016).
72. Lötvall, J. *et al.* Asthma endotypes: A new approach to classification of disease entities within the asthma syndrome. *Journal of Allergy and Clinical Immunology* **127**, 355–360 (2011).
73. Östling, J. *et al.* IL-17–high asthma with features of a psoriasis immunophenotype. *Journal of Allergy and Clinical Immunology* **144**, 1198–1213 (2019).
74. Ardern-Jones, M. R. Characterisation of atopic dermatitis (AD) endotypes and novel treatment targets: towards a molecular classification. *Experimental Dermatology* **27**, 433–434 (2018).
75. Ungar, B. *et al.* Phase 2 randomized, double-blind study of IL-17 targeting with secukinumab in atopic dermatitis. *Journal of Allergy and Clinical Immunology* **147**, 394–397 (2021).
76. Agache, I. *et al.* EAACI Biologicals Guidelines - dupilumab for children and adults with moderate-to-severe atopic dermatitis. *Allergy* (2020) doi:10.1111/all.14690.
77. Klasa, B. & Cichocka-Jarosz, E. Atopic Dermatitis - Current State of Research on Biological Treatment. *J Mother Child* **24**, 53–66 (2020).
78. Wu, J. & Guttman-Yassky, E. Efficacy of biologics in atopic dermatitis. *Expert Opin Biol Ther* **20**, 525–538 (2020).
79. European Task Force on Atopic Dermatitis. Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology (Basel)* **186**, 23–31 (1993).
80. Deckers, I. A. G. *et al.* Investigating International Time Trends in the Incidence and Prevalence of Atopic Eczema 1990–2010: A Systematic Review of Epidemiological Studies. *PLoS ONE* **7**, 28 (2012).
81. Barbarot, S. *et al.* Epidemiology of atopic dermatitis in adults: Results from an international survey. *Allergy* (2018) doi:10.1111/all.13401.
82. D'Erme, A. M. *et al.* IL-36γ (IL-1F9) is a biomarker for psoriasis skin lesions. *J. Invest.*

Dermatol. **135**, 1025–1032 (2015).
83. Miura, S. *et al.* IL-36 and IL-17A Cooperatively Induce a Psoriasis-like Gene Expression Response in Human Keratinocytes. *J Invest Dermatol* (2021) doi:10.1016/j.jid.2021.01.019.
84. Guttman-Yassky, E. *et al.* Use of Tape Strips to Detect Immune and Barrier Abnormalities in the Skin of Children With Early-Onset Atopic Dermatitis. *JAMA Dermatol* **155**, 1358 (2019).
85. Li, W. *et al.* IL-37 is protective in allergic contact dermatitis through mast cell inhibition. *International Immunopharmacology* **83**, 106476 (2020).
86. Guillonneau, C., Bézie, S. & Anegon, I. Immunoregulatory properties of the cytokine IL-34. *Cell. Mol. Life Sci.* **74**, 2569–2586 (2017).
87. Russell, C. D. & Baillie, J. K. Treatable traits and therapeutic targets: Goals for systems biology in infectious disease. *Current Opinion in Systems Biology* **2**, 140–146 (2017).
88. Thijs, J. L., de Bruin-Weller, M. S. & Hijnen, D. Current and Future Biomarkers in Atopic Dermatitis. *Immunology and Allergy Clinics of North America* **37**, 51–61 (2017).
89. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).
90. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
91. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003).
92. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
93. Alarcón-Riquelme, M. E. New Attempts to Define and Clarify Lupus. *Curr Rheumatol Rep* **21**, 11 (2019).
94. He, H. *et al.* Tape strips detect distinct immune and barrier profiles in atopic dermatitis and psoriasis. *Journal of Allergy and Clinical Immunology* S0091674920308241 (2020) doi:10.1016/j.jaci.2020.05.048.
95. Toro-Dom, D. & Carmona-S, P. Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. 11.
96. Tsoi, L. C. *et al.* Atopic Dermatitis Is an IL-13–Dominant Disease with Greater Molecular Heterogeneity Compared to Psoriasis. *Journal of Investigative Dermatology* **139**, 1480–1489 (2019).
97. Esaki, H. *et al.* Identification of novel immune and barrier genes in atopic dermatitis by means of laser capture microdissection. *Journal of Allergy and Clinical Immunology* **135**, 153–163 (2015).
98. Hira, Z. M. & Gillies, D. F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics* 13.
99. Dozmorov, I. Hypervariable genes--experimental error or hidden dynamics. *Nucleic Acids Research* **32**, e147–e147 (2004).
100. Thijs, J. L. *et al.* Serum biomarker profiles suggest that atopic dermatitis is a systemic disease. *Journal of Allergy and Clinical Immunology* **141**, 1523–1526 (2018).
101. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat Genet* **46**, 430–437 (2014).
102. Wang, L., Quan, Y., Yue, Y., Heng, X. & Che, F. Interleukin-37: A crucial cytokine with multiple roles in disease and potentially clinical therapy (Review). *Oncol Lett* (2018) doi:10.3892/ol.2018.7982.

103. Guttman-Yassky, E. *et al.* Broad defects in epidermal cornification in atopic dermatitis identified through genomic analysis. *Journal of Allergy and Clinical Immunology* **124**, 1235-1244.e58 (2009).
104. Li, J., Chen, X., Liu, Z., Yue, Q. & Liu, H. Expression of Th17 cytokines in skin lesions of patients with psoriasis. *J Huazhong Univ Sci Technolog Med Sci* **27**, 330–332 (2007).
105. Guttman-Yassky, E. *et al.* Low Expression of the IL-23/Th17 Pathway in Atopic Dermatitis Compared to Psoriasis. *J Immunol* **181**, 7420–7427 (2008).
106. Suárez-Fariñas, M. *et al.* Intrinsic atopic dermatitis shows similar TH2 and higher TH17 immune activation compared with extrinsic atopic dermatitis. *Journal of Allergy and Clinical Immunology* **132**, 361–370 (2013).
107. Sugaya, M. The Role of Th17-Related Cytokines in Atopic Dermatitis. *IJMS* **21**, 1314 (2020).
108. Akiyama, M., Takeichi, T., McGrath, J. A. & Sugiura, K. Autoinflammatory keratinization diseases. *J. Allergy Clin. Immunol.* **140**, 1545–1547 (2017).
109. Patrick, G. J. *et al.* Epicutaneous Staphylococcus aureus induces IL-36 to enhance IgE production and ensuing allergic disease. *J Clin Invest* **131**, (2021).
110. Suárez-Fariñas, M. *et al.* RNA sequencing atopic dermatitis transcriptome profiling provides insights into novel disease mechanisms with potential therapeutic implications. *Journal of Allergy and Clinical Immunology* **135**, 1218–1227 (2015).
111. Hulshof, L. *et al.* A minimally invasive tool to study immune response and skin barrier in children with atopic dermatitis. *Br J Dermatol* **180**, 621–630 (2019).
112. He, H. *et al.* Single-cell transcriptome analysis of human skin identifies novel fibroblast subpopulation and enrichment of immune subsets in atopic dermatitis. *Journal of Allergy and Clinical Immunology* **145**, 1615–1628 (2020).
113. Rojahn, T. B. *et al.* Single-cell transcriptomics combined with interstitial fluid proteomics defines cell type–specific immune regulation in atopic dermatitis. *Journal of Allergy and Clinical Immunology* S009167492030556X (2020) doi:10.1016/j.jaci.2020.03.041.
114. Bangert, C. *et al.* Persistence of mature dendritic cells, TH2A, and Tc2 cells characterize clinically resolved atopic dermatitis under IL-4Rα blockade. *Sci Immunol* **6**, (2021).
115. Kalinina, P. *et al.* The Whey Acidic Protein WFDC12 Is Specifically Expressed in Terminally Differentiated Keratinocytes and Regulates Epidermal Serine Protease Activity. *J Invest Dermatol* (2020) doi:10.1016/j.jid.2020.09.025.
116. Ständer, S. *et al.* Clinical classification of itch: a position paper of the International Forum for the Study of Itch. *Acta Derm Venereol* **87**, 291–294 (2007).
117. Bao, L. *et al.* A molecular mechanism for IL-4 suppression of loricrin transcription in epidermal keratinocytes: implication for atopic dermatitis pathogenesis. *Innate Immun* **23**, 641–647 (2017).
118. Gutowska-Owsiak, D., Schaupp, A. L., Salimi, M., Taylor, S. & Ogg, G. S. Interleukin-22 downregulates filaggrin expression and affects expression of profilaggrin processing enzymes. *Br J Dermatol* **165**, 492–498 (2011).
119. Pincelli, C. *et al.* Neuropeptides in skin from patients with atopic dermatitis: an immunohistochemical study. *Br J Dermatol* **122**, 745–750 (1990).
120. Andersen, H. H., Elberling, J., Sølvsten, H., Yosipovitch, G. & Arendt-Nielsen, L. Nonhistaminergic and mechanical itch sensitization in atopic dermatitis. *Pain* **158**, 1780–1791 (2017).
121. Ruzicka, T. *et al.* Anti–Interleukin-31 Receptor A Antibody for Atopic Dermatitis. *New*

England Journal of Medicine **376**, 826–835 (2017).
122. Kabashima, K., Matsumura, T., Komazaki, H. & Kawashima, M. Trial of Nemolizumab and Topical Agents for Atopic Dermatitis with Pruritus. *N Engl J Med* **383**, 141–150 (2020).
123. Cowan, A., Kehner, G. B. & Inan, S. Targeting Itch with Ligands Selective for κ Opioid Receptors. *Handb Exp Pharmacol* **226**, 291–314 (2015).
124. Golpanian, R. S. & Yosipovitch, G. Current and emerging systemic treatments targeting the neural system for chronic pruritus. *Expert Opin Pharmacother* **21**, 1629–1636 (2020).
125. Pavel, A. B. *et al.* Oral Janus kinase/SYK inhibition (ASN002) suppresses inflammation and improves epidermal barrier markers in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology* **144**, 1011–1024 (2019).
126. Bissonnette, R. *et al.* Crisaborole and atopic dermatitis skin biomarkers: An intrapatient randomized trial. *Journal of Allergy and Clinical Immunology* **144**, 1274–1289 (2019).
127. Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).
128. Chen, Y. *et al.* Role of ERK1/2 activation on itch sensation induced by bradykinin B1 activation in inflamed skin. *Exp Ther Med* **12**, 627–632 (2016).
129. Yeom, M. *et al.* Atopic dermatitis induces anxiety- and depressive-like behaviors with concomitant neuronal adaptations in brain reward circuits in mice. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **98**, 109818 (2020).
130. Vollono, L. *et al.* Potential of Curcumin in Skin Disorders. *Nutrients* **11**, (2019).
131. Blue, R. E., Curry, E. G., Engels, N. M., Lee, E. Y. & Giudice, J. How alternative splicing affects membrane-trafficking dynamics. *J Cell Sci* **131**, (2018).
132. Hou, Y. & Witman, G. B. The N-terminus of IFT46 mediates intraflagellar transport of outer arm dynein and its cargo-adaptor ODA16. *Mol Biol Cell* **28**, 2420–2433 (2017).
133. Gharib, K., Mostafa, A. & Elsayed, A. Evaluation of Botulinum Toxin Type A Injection in the Treatment of Localized Chronic Pruritus. *J Clin Aesthet Dermatol* **13**, 12–17 (2020).
134. Chen, L. & Zhong, J. L. MicroRNA and heme oxygenase-1 in allergic disease. *Int Immunopharmacol* **80**, 106132 (2020).
135. Casares, L. *et al.* Cannabidiol induces antioxidant pathways in keratinocytes by targeting BACH1. *Redox Biol* **28**, 101321 (2020).
136. Robinson, M. D. & Speed, T. P. A comparison of Affymetrix gene expression arrays. *BMC Bioinformatics* **8**, 449 (2007).
137. Storck, M. *et al.* Pruritus Intensity Scales across Europe: a prospective validation study. *J Eur Acad Dermatol Venereol* jdv.17111 (2021) doi:10.1111/jdv.17111.
138. Smith, M. P. *et al.* Emerging Methods to Objectively Assess Pruritus in Atopic Dermatitis. *Dermatol Ther (Heidelb)* **9**, 407–420 (2019).
139. Guimarães, P., Batista, A., Zieger, M., Kaatz, M. & Koenig, K. Artificial Intelligence in Multiphoton Tomography: Atopic Dermatitis Diagnosis. *Sci Rep* **10**, 7968 (2020).
140. Holm, J. G. *et al.* Immunoinflammatory Biomarkers in Serum Are Associated with Disease Severity in Atopic Dermatitis. *Dermatology* 1–8 (2021) doi:10.1159/000514503.
141. Jurakic Toncic, R. *et al.* Stratum corneum markers of innate and T helper cell-related immunity and their relation to the disease severity in Croatian patients with atopic dermatitis. *J Eur Acad Dermatol Venereol* (2021) doi:10.1111/jdv.17132.
142. Hurault, G., Domínguez-Hüttinger, E., Langan, S. M., Williams, H. C. & Tanaka, R. J.

Personalized prediction of daily eczema severity scores using a mechanistic machine learning model. *Clin Exp Allergy* **50**, 1258–1266 (2020).
143. Liesecke, F. *et al.* Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci Rep* **8**, 10885 (2018).
144. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N. & Fanos, V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina* **56**, 455 (2020).
145. Arcadu, F. *et al.* Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digit. Med.* **2**, 92 (2019).
146. Deans, A. R. *et al.* Finding our way through phenotypes. *PLoS Biol* **13**, e1002033 (2015).
147. Kato, G. J., Gladwin, M. T. & Steinberg, M. H. Deconstructing sickle cell disease: reappraisal of the role of hemolysis in the development of clinical subphenotypes. *Blood Rev* **21**, 37–47 (2007).
148. Renert-Yuval, Y. *et al.* The molecular features of normal and atopic dermatitis skin in infants, children, adolescents, and adults. *J Allergy Clin Immunol* (2021) doi:10.1016/j.jaci.2021.01.001.
149. Kosolapoff, G. M. Basis for scientific terminology and classification. *Science* **101**, 89–90 (1945).
150. Popkin, G. Data sharing and how it can benefit your scientific career. *Nature* **569**, 445–447 (2019).
151. Longo, D. L. & Drazen, J. M. Data Sharing. *N Engl J Med* **374**, 276–277 (2016).
152. Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Sci Data* **6**, 251 (2019).
153. Cosgriff, C. V., Ebner, D. K. & Celi, L. A. Data sharing in the era of COVID-19. *Lancet Digit Health* **2**, e224 (2020).
154. Wehbe, S. *et al.* COVID-19 in the Middle East and North Africa region: an urgent call for reliable, disaggregated and openly shared data. *BMJ Glob Health* **6**, (2021).
155. Weber, G. M. *et al.* International Comparisons of Harmonized Laboratory Value Trajectories to Predict Severe COVID-19: Leveraging the 4CE Collaborative Across 342 Hospitals and 6 Countries: A Retrospective Cohort Study. *medRxiv* (2021) doi:10.1101/2020.12.16.20247684.
156. Ning, W. *et al.* Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. *Nat Biomed Eng* **4**, 1197–1207 (2020).
157. Milham, M. P. *et al.* Assessment of the impact of shared brain imaging data on the scientific literature. *Nat Commun* **9**, 2818 (2018).
158. Loke, S. Y. & Lee, A. S. G. The future of blood-based biomarkers for the early detection of breast cancer. *Eur J Cancer* **92**, 54–68 (2018).
159. Guttman-Yassky, E. *et al.* Molecular signatures order the potency of topically applied anti-inflammatory drugs in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology* **140**, 1032-1042.e13 (2017).
160. Brunner, P. M. *et al.* Baseline IL-22 expression in patients with atopic dermatitis stratifies tissue responses to fezakinumab. *Journal of Allergy and Clinical Immunology* **143**, 142–154 (2019).
161. Nakahara, T. *et al.* Exploration of biomarkers to predict clinical improvement of atopic dermatitis in patients treated with dupilumab: A study protocol. *Medicine (Baltimore)*

99, e22043 (2020).
162. Patrick, M. T. *et al.* Drug Repurposing Prediction for Immune-Mediated Cutaneous Diseases using a Word-Embedding–Based Machine Learning Approach. *Journal of Investigative Dermatology* **139**, 683–691 (2019).
163. Hand, D. J. Classifier Technology and the Illusion of Progress. *Statist. Sci.* **21**, (2006).
164. Sutton, R. T. *et al.* An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit. Med.* **3**, 17 (2020).
165. Jayatilake, S. M. D. A. C. & Ganegoda, G. U. Involvement of Machine Learning Tools in Healthcare Decision Making. *Journal of Healthcare Engineering* **2021**, 1–20 (2021).

APPENDICES

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Date of Visit
☐☐ ☐☐ ☐☐☐☐

CASE REPORT FORM (Atopic Eczema)

MAARS: Microbes in Allergy and Autoimmunity Related to the Skin

Principal Investigator

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

INCLUSION CRITERIA
If the answer is NO to any of these questions then the patient is to be excluded from the study.

	YES	NO
Written informed consent	☐	☐
≥ 18 years old	☐	☐
Patients with Atopic Dermatitis, diagnosed using Hanifin-Rajka Criteria	☐	☐

EXCLUSION CRITERIA
If the answer is YES to any of these questions then the patient is to be excluded from the study.

	Yes	**No**
Patient unable to give written informed consent	☐	☐
Patient has no allergen-specific IgE and no allergen-specific immediate type reactions	☐	☐
Patients who have received treatment at the biopsy site at least 2 weeks prior to screening	☐	☐
Patients who have received systemic antibiotics within the previous 4 weeks prior to screening	☐	☐
Patients who have received systemic immunosupressive therapy within the previous 12 weeks prior to screening	☐	☐
Patients who have received systemic biologic agents within the previous 12 weeks prior to screening	☐	☐

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Diagnostic & Phenotypic Data

ATOPIC DERMATITIS
Anamnesis sheet

*The identification number contains:
- Two digit code for the disease (AD=atopic dermatitis)
- The date of the biopsy in YYMMDD-format
- A two digit consecutive number
- Two digits for the initials of the patient
- One digit for indentification of the Center (D: Düsseldorf; H: Helsinki; L: London)
- One digit for additional biopsies. Should be 0 for a patient with one biopsy and leaves room for up to 10 biopsies per patient.

Identification numbers must be unique

Name .....
Date of birth .....
Address .....
ID-No.* AD .....
Male ☐ Female ☐

Ethnicity/Family History

Ethnicity: ■ White ■ Black - African ■ Black - Caribbean ■ Black - Other ■ Indian ■ Bangladeshi ■ Chinese ■ Asian - Other ■ Other If Other, Please Specify: _____

Family History of atopic diseases: ❑ Yes ❑ No ❑ Unknown If Yes, Please Specify (1st relative only)_____

Known Allergies: ❑ Yes ❑ No ❑ Unknown If Yes, Please Specify:

.....

.....

Hanifin and Rajka diagnostic criteria

	No	Yes
1. Pruritus (major)	☐	☐
2. Typical morphology and distribution	☐	☐
3. Chronic or chronically-relapsing dermatitis	☐	☐
4. White dermographismus	☐	☐
5. Xerosis	☐	☐
6.Palmar hyperlinearity/ Keratosis pilaris	☐	☐
7. Immediate (type 1) skin-test reactivity	☐	☐
(specify if possible)	☐	☐
Grass polls	☐	☐
Birch polls	☐	☐
Ragweed	☐	☐
Dermatophagoides pteronissimus	☐	☐
Dermatophagoides farinae	☐	☐
Dog	☐	☐
Cat	☐	☐
Total IgE (IU/ml)		
a) 150 IU/ml < x < 400 IU/ml	☐	☐
b) >400 IU/ml	☐	☐
8. Raised serum IgE	☐	☐
9. Early age of onset		
10. ECP measurement [ug/ml]		
11. Tendency toward cutaneous infections or impaired cell-mediated immunity	☐	☐
Staphylococcus aureus	☐	☐
Viral: HSV	☐	☐
HPV	☐	☐
12. Tendency toward non-specific hand or foot dermatitis	☐	☐
13. Nipple eczema	☐	☐
14. Intolerance to:	☐	☐
wool	☐	☐
lipid solvents	☐	☐
15. Cheilitis	☐	☐
16. Food intolerance	☐	☐

17. Recurrent conjunctivitis	☐	☐
18. Dennie-Morgan infraorbital fold	☐	☐
19. Keratoconus	☐	☐
20. Anterior subcapsular cataracts	☐	☐
21. Orbital darkening	☐	☐
22. Facial pallor/erythema	☐	☐
23. Pityriasis alba	☐	☐
24. Anterior neck folds	☐	☐
25. Itch when sweating	☐	☐
26. Perifollicular accentuation	☐	☐
27. Course influence by enviromental or emotional factors	☐	☐
28. Responsiveness to		
Glucocorticosteroids	☐	☐
Topical steroids	☐	☐
Systemic cyclosporin	☐	☐
29.) Concurrent diseases	☐	☐
30.) Specify if yes:		
31.) Additional medication	☐	☐
32.) Specify if yes:		

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

33. Personal or family history of atopy ☐ ☐
a) Personal history of: ☐ ☐
Allergic rhinitis ☐ ☐
Allergic conjunctivitis ☐ ☐
Bronchial asthma ☐ ☐
Urticaria ☐ ☐

food ☐ ☐
drug ☐ ☐
other ☐ ☐
Contact allergy (nickel) ☐ ☐

33 b) Family history

	Parents				Children				Siblings			
	Mother		Father		Boy		Girl		Brother		Sister	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Atopic dermatitis												
Allergic rhinitis												
Allergic conjunctivitis												
Asthma												
Urticaria												

Atopic Dermatitis Treatment –

Past treatment- including UV	Dose	Freq	Start date	Ongoing	End Date	Responder
			DD/MM/YY	Y/N	DD/MM/YY	Y/N
			DD/MM/YY	Y/N	DD/MM/YY	Y/N
			DD/MM/YY	Y/N	DD/MM/YY	Y/N
			DD/MM/YY	Y/N	DD/MM/YY	Y/N
			DD/MM/YY	Y/N	DD/MM/YY	Y/N
			DD/MM/YY	Y/N	DD/MM/YY	Y/N
			DD/MM/YY	Y/N	DD/MM/YY	Y/N
			DD/MM/YY	Y/N	DD/MM/YY	Y/N

Concomitant Medication-

Medication	Dose	Freq	Start date	Ongoing	End Date
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY

Other concurrent chronic diseases

			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY
			DD/MM/YY	Y/N	DD/MM/YY

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Smoking: ❑ Yes ❑ No

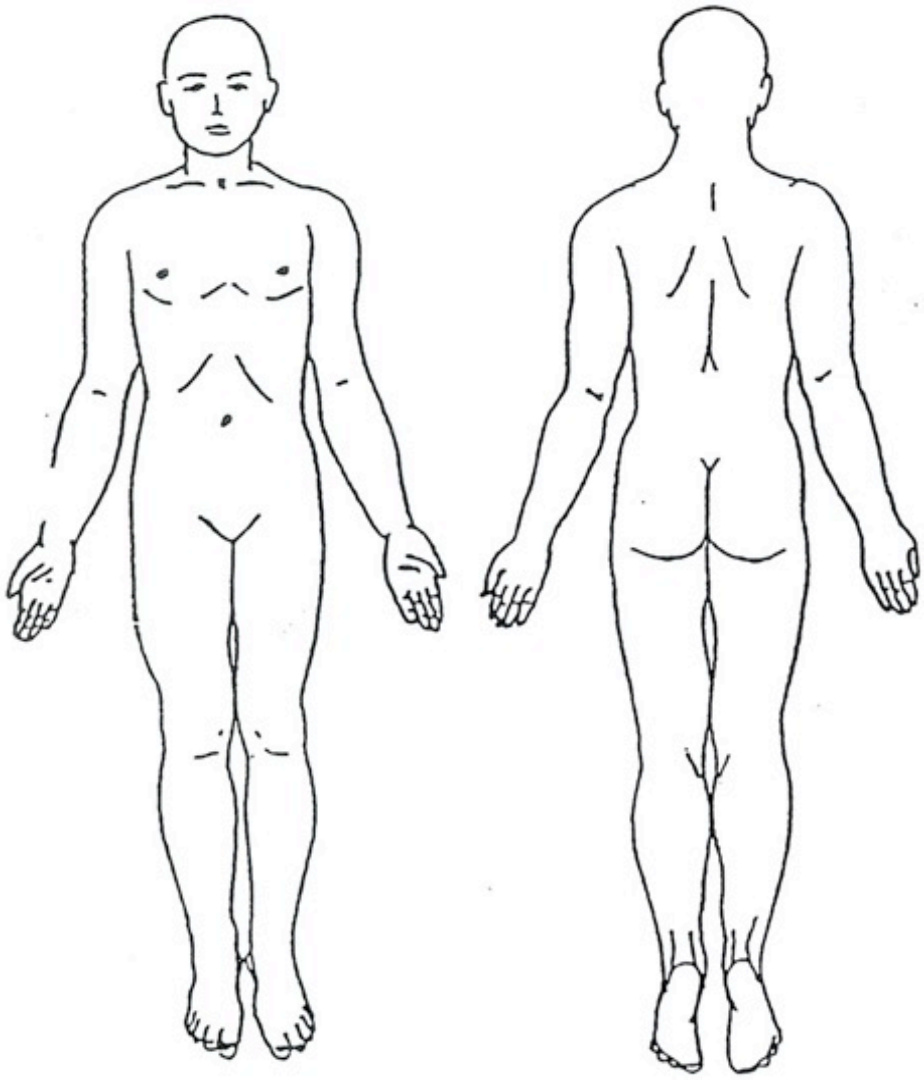
Malignancies

Skin cancers	treated	not treated
	DD/MM/YY	DD/MM/YY
	DD/MM/YY	DD/MM/YY
	DD/MM/YY	DD/MM/YY
Other	DD/MM/YY	DD/MM/YY
	DD/MM/YY	DD/MM/YY
	DD/MM/YY	DD/MM/YY

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

SCORAD index

Institution: .....
Physician: .....



A)Extend: Please state the areas involved

.....

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

B) Intensity

Erythema	
Edema/Papulation	
Oozing/crusts	
Excoriation	
Lichenification	
Dryness**	

**Dryness is evaluated on uninvolved areas

Means of calculation
0 = absence
1 = mild
2 = moderate
3 = severe

C) Subjective symptoms: Pruritus and sleep loss

Pruritus (1-10): ___

1.....10

Sleep loss (1-10): ___

(Visual analog scale average for the last 3 days or nights)

Objective SCORAD A/5 + 7B/2 / 83

SCORAD A/5 + 7B/2 + C / 103

Remarks:

.....

.....

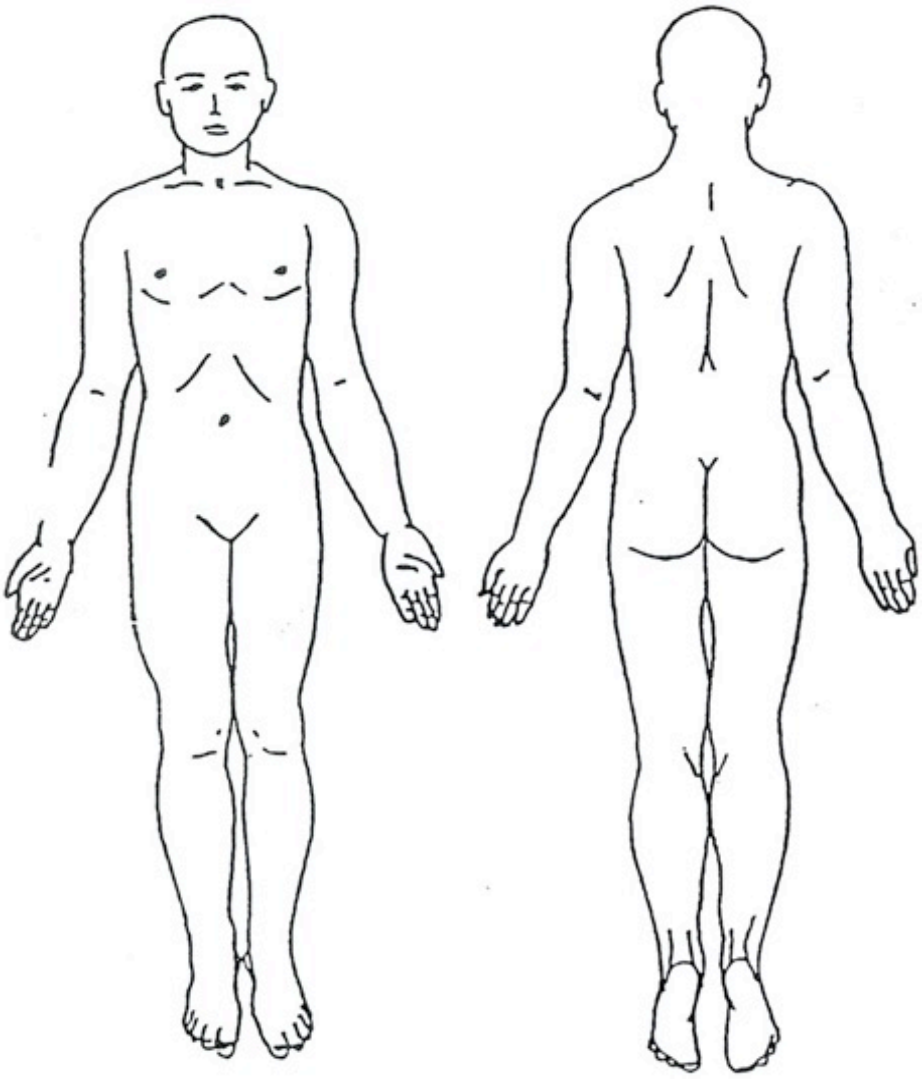
.....

.....

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

LOCAL SCORAD

Institution: .....
Physician: .....



A)Please state the areas used for sampling

Lesional skin: _____

❑ **Acute lesion** (< 1 week) ❑ **Chronic lesion** (> 1 week) ❑ **Unknown**

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Evidence of superinfection:
a) clinical _____
b) microbiological (bacterial)_____
c) microbiological (viral)_____

Non-lesional skin: _____

B) Intensity

Erythema	
Edema/Papulation	
Oozing/crusts	
Excoriation	
Lichenification	
Dryness**	

**Dryness is evaluated on uninvolved areas

Means of calculation
0 = absence
1 = mild
2 = moderate
3 = severe

Remarks:

.....

.....

.....

.....

Clinical Assessment:

Date of Assessment (DD/MM/YY**) _____**

Atopic Dermatitis Patients:

SCORAD Score: _____

LOCAL SCORAD Score: _____

Global Assessment Score	Severe/ moderate-severe/ moderate/ mild/ almost clear/ clear

Sample Collection:

Swipe sample

	Site- Flexor		
Disease Type	**Lesional**	**Non-lesional**	**Date** DD/MM/YY
AD			

Biopsy

	Site- Flexor		
Disease Type	**Lesional**	**Non-lesional**	**Date** DD/MM/YY
AD			

Patient Initials ☐☐☐ Patient ID ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Blood sample collection:		**Date:** DD/MM/YY
Heparinised blood sample for PBMC:	**YES / NO**	
Blood sample for genomic DNA:	**YES / NO**	

Photographs taken	**YES / NO**	**Date:** DD/MM/YY

COLLABORATIVE WORKS

During my PhD I participate to several collaborative works, whether for scientifical and clinical aspects.

Scientifical main contributions

- **Influence of FLG loss-of-function mutations in host–microbe interactions during atopic skin inflammation**
  Peter Oláh, (…) Alain Lefèvre-Utile, (…),Bernhard Homey
  Rejected from Allergy after reviewing, currently under review in the Journal of Investigative Dermatology
  I made a contribution in the analyses design and the draft redaction

- **Characterization and functional interrogation of SARS-CoV-2 RNA interactome**
  Athéna Labeau, Alain Lefèvre-Utile, (…) Ali Amara, Laurent Meertens
  Available here: https://doi.org/10.1101/2021.03.23.436611
  Currently under review in Cell Host and Microbes
  I made a contribution in the functional enrichment, the figure design, and the draft redaction

Clinical main contributions

- **Coronavirus Disease 2019 Pandemic: Impact Caused by School Closure and National Lockdown on Pediatric Visits and Admissions for Viral and Nonviral Infections-a Time Series Analysis**
  François Angoulvan, (…), Alain Lefèvre-Utile, (…) David Skurnik.
  Published in the Clinical Infectious Disease Journal, since January 2021
  Available here: https://doi.org/10.1093/cid/ciaa710
  &
  COVID-19 pandemic: Impact caused by school closure and national lockdown on pediatric visits and admissions for viral and non-viral infections, a time series analysis.
  François Angoulvan, (…), Alain Lefèvre-Utile, (…) David Skurnik.
  Published in the Clinical Infectious Disease Journal, since June 2020
  Available here: https://doi.org/10.1093/cid/ciaa710
  I made a contribution in data generation and research question design

- **Different Clinical Presentations and Outcomes of Disseminated Varicella in Children With Primary and Acquired Immunodeficiencies I made a contribution in performing the statistics**
  Paul Bastard, Aurelien Galerne, Alain Lefèvre-Utile, (…), Benedicte Neven
  Published in the Frontiers of Immunology, since November 2020
  Available here: https://doi.org/10.3389/fimmu.2020.595478
  I made a contribution in the statistical analyses

université PARIS-SACLAY

ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Classifications transcriptomiques des dermatites inflammatoires : application à la dermatite atopique

Mots clés : dermatite atopique, transcriptome, endotype, prurit, statistique, apprentissage machine

Résumé : La dermatite atopique (AD) est une maladie inflammatoire de la peau qui touche selon l'origine géographique jusqu'à 20% des enfants et 5% des adultes. Elle est marquée par une forte hétérogénéité aussi bien clinique, biologique, qu'en terme de conséquence pour le patient. Cette variabilité interindividuelle est probablement liée à des mécanismes distincts qui pourraient être la cible de traitements personnalisés. Dans cette thèse, nous avons appliqué deux stratégies distinctes sur une importante cohorte transcriptomique issue du consortium MAARS, d'échantillons cutanés AD (n=82) et sains (n=213).
1) L'approche non-supervisée, a permis l'identification de gènes hypervariables spécifiques de l'AD. A partir de ces gènes, quatre groupes de patients ont été constitués, caractérisés par des mécanismes biologiques et des présentations cliniques distinctes. L'existence de ces quatre endotypes a été validé sur une cohorte indépendante.
2) L'approche supervisée s'est intéressée à l'exploration du prurit, un symptôme de l'AD très invalidant. En utilisant des approches innovantes comme l'apprentissage machine, nous avons identifié une signature génétique du prurit. Elle a permis de prédire l'intensité du prurit avec une précision importante et d'identifier de nouveaux mécanismes.
Ces deux stratégies complémentaires ont tiré profit de l'importance de la cohorte et de la technologie utilisé permettant des découvertes originales, aussi bien méthodologiques et biologiques.

Title : Inflammatory dermatitis transcriptomic classifications: application to atopic dermatitis

Keywords : atopic dermatitis, transcriptomics, endotype, pruritus, statistics, machine learning

Abstract : Atopic dermatitis (AD) is a frequent inflammatory dermatitis that concerns up to 5% of adults and 20% of children depending on their origins. It is a highly heterogeneous disease at the clinical and biological levels. Interindividual variabilities should underline distinct pathophysiological mechanisms that could be targeted by personalized therapeutics. In this thesis work, we applied two strategies to the largest transcriptomics cohort (MAARS Consortium). We used n=82 AD and n=213 healthy controls skin samples.
1) The unsupervised approach identified AD hyper-variable genes in comparison to healthy controls. These genes were used to cluster patients into four groups with distinct underlying mechanisms and clinical characterization. The exitance of four endotypes was then validated on an external cohort.
2) The supervised approach focused on the understanding of pruritus. This symptom takes an important part in the AD burden. Combining machine learning and statistical models, we identified a pruritus genetic signature able to accurately predict pruritus intensity and revealed new mechanisms.
These complimentary strategies took advantage of our large cohort and good quality data to reveal original findings at the methodological and biological levels.