nature communications

Article

https://doi.org/10.1038/s41467-023-37025-7

L'analyse unicellulaire révèle une inflammation interactions à l'origine de la dégénérescence maculaire

Reçu : 12 mai 2022

Accepté : 27 février 2023

Published online: 05 May 2023

Vérifier les mises à jour

Manik Kuchroo1,19, Marcello DiStasio2,19, Eric Song3,19, Eda Calapkulu3 , Le Zhang4, Maryam Ige5 , Amar H. Sheth 5, Abdelilah Majdoubi3 , Madhvi Menon6 , Alexandre Tong 7, Abhinav Godavarthi8, Yu Xing3 , Scott Gigante 9, Holly Steach10, Jessie Huang7 ,Guillaume Huguet 11,12, Janhavi Narain13, Kisung You14, George Mourgkos3 , Rahul M. Dhodapkar5 , Matthew J. Hirn 15,16, Bastian Rieck17, Guy Wolf 11,12, Smita Krishnaswamy 7,14,20 ✉ & Brian P. Hafler 3,18,20 ✉

En raison de points communs physiopathologiques, la dégénérescence maculaire liée à l'âge (AMD) représente un modèle accessible de manière unique pour étudier les thérapies pour maladies neurodégénératives, ce qui nous amène à examiner si les voies de progression de la maladie sont communes à toutes les maladies neurodégénératives. Ici nous utiliser le séquençage d'ARN mononucléaire pour profiler les lésions de 11 post-mortem rétines humaines atteintes de dégénérescence maculaire liée à l'âge et 6 rétines témoins sans antécédent de maladie rétinienne. Nous créons un pipeline d'apprentissage automatique basé sur les progrès récents en matière de géométrie et de topologie des données et d'identifier les cellules gliales activées populations enrichies au début de la maladie. Examen des données unicellulaires de la maladie d'Alzheimer et de la sclérose en plaques progressive grâce à notre pipeline, nous retrouvons un profil d'activation gliale similaire enrichi dans la phase précoce de ces maladies neurodégénératives. Dans la dégénérescence maculaire liée à l'âge à un stade avancé, nous identifions un axe de signalisation microglie-astrocytes médié par l'interleukine-1β qui pilote l'angiogenèse caractéristique de la pathogenèse de la maladie. Nous avons validé ce mécanisme en utilisant des tests in vitro et in vivo chez la souris, identifiant une nouvelle cible thérapeutique possible pour la DMLA et éventuellement d'autres maladies neurodégénératives. conditions. Ainsi, en raison des états gliaux partagés, la rétine offre un potentiel système d'investigation d'approches thérapeutiques dans les maladies neurodégénératives maladies.

La DMLA est une maladie neurodégénérative de la rétine qui touche 196 millions de personnes dans le monde et a un impact significatif sur la santé des patients qualité de vie1 . À l'instar d'autres maladies neurodégénératives du système nerveux central (SNC), comme la maladie d'Alzheimer (MA) et sclérose en plaques progressive (SEP), la DMLA peut être classée en étapes. Initialement, au stade précoce et « sec » de la DMLA, des dépôts extracellulaires contenant de l'amyloïde bêta, appelés drusen, s'accumulent dans la rétine, conduisant à l'activation de glia2 . Dans la DMLA avancée « néovasculaire », angiogenèse et fibrose induites par le facteur de croissance endothélial vasculaire (VEGF) provoquent une perte de photorécepteur et de vision3· Dans la SEP et la MA, la dysfonction gliale la régulation est associée à des lésions neuronales et à une déficience neurologique progressive4,5 . Cela soulève la question de savoir si les états d'activation des cellules gliales sont partagés dans la neurodégénérescence et la question de savoir si la rétine humaine peut être utilisée comme modèle pour des interventions ciblant les cellules gliales pour des maladies neurodégénératives similaires.

Alors que la transcriptomique unicellulaire a donné un aperçu des perturbations cellulaires dans la MA et la MS4-7 , une transcriptomique unicellulaire l'analyse de la DMLA n'a pas été réalisée. Pour identifier les types de cellules et

Une liste complète des affiliations apparaît à la fin du document. ✉courrier électronique : smita.krishnaswamy@yale.edu ; brian.hafler@yale.edu

états enrichis à travers les stades de la DMLA, nous avons effectué un séquençage d'ARN à noyau unique basé sur la microfluidique massivement parallèle (snRNA-seq) créer un ensemble de données transcriptomiques unicellulaires sur la pathologie de la DMLA, comprenant 70 973 cellules à plusieurs stades de la maladie. Dans un tel de grands ensembles de données, identifiant les populations cellulaires à l'origine des maladies et pourraient être ciblés pour un bénéfice thérapeutique reste un défi avec approches actuelles. Cela se produit souvent parce que les populations pathogènes peuvent constituer un petit sous-ensemble d'un compartiment reconnu du tissu. Il peut donc être difficile d'identifier de telles populations parmi le bruit et la complexité présents dans les données unicellulaires. À Pour résoudre ce problème, nous avons développé un outil d'apprentissage automatique d'inspiration topologique suite d'outils appelée Analyse Cellulaire avec Topologie et Homologie de Condensation (CATCH). Au centre de ce cadre se trouve un pipeline de découverte de populations pathogènes dont l'élément clé est un méthode appelée condensation par diffusion8 . Condensation par diffusion identifie systématiquement des groupes de cellules similaires à travers les échelles pour découvrir des sous-populations d'intérêt dans un cadre de diffusion de données. Dans cette approche, les cellules sont tirées de manière itérative vers la valeur pondérée. moyenne de leurs voisins dans l'espace génétique de grande dimension, lentement éliminant les variations. Lorsque les cellules se rapprochent les unes des autres, la diffusion la condensation les fusionne, créant un nouveau cluster. Quand combiné avec une méthode d'abondance différentielle unicellulaire MELD9 , la condensation de diffusion peut identifier des sous-populations distinctes associées à la progression de la maladie. Cela représente une amélioration sur les outils de clustering qui partitionnent les données en fonction des métriques de interconnectivité des clusters. Puisque cette approche identifie des populations enrichies en maladies, des signatures spécifiques à une condition peuvent être construit et comparé à travers des conditions neurodégénératives, aidant construire une compréhension commune des mécanismes communs des maladies.

À l'aide du pipeline CATCH, nous avons identifié deux populations de des cellules gliales activées, un sous-ensemble microglial et un sous-ensemble d'astrocytes, enrichis dans la phase précoce de la DMLA sèche. Ces sous-ensembles ont été caractérisés par les signatures de la phagocytose, du métabolisme lipidique et des fonctions lysosomales. En réappliquant notre pipeline aux ensembles de données unicellulaires AD4 et MS5 , nous avons identifié les mêmes signatures dans les premières phases de plusieurs maladies neurodégénératives, indiquant un mécanisme commun pour les maladies gliales activation dans la phase précoce de la neurodégénérescence. La microglie et les signatures d'expression des astrocytes ont été validées dans la rétine humaine et tissu cérébral. Dans la DMLA néovasculaire à un stade avancé, CATCH a identifié un signature d'expression de l'inflammasome dans les microglies ainsi qu'une signature pro-angiogénique dans les astrocytes. Grâce à une analyse informatique des interactions récepteur-ligand, nous avons identifié un axe de signalisation clé entre IL-1β dérivée des microglies et astrocytes pro-angiogéniques, le moteur de néovascularisation et perte de photorécepteurs dans les maladies avancées AMD3 . Grâce à une combinaison de cellules souches pluripotentes induites par l'homme (iPSC), des tests de stimulation des astrocytes dérivés, des expériences in vivo sur des souris et des analyses d'échantillons rétiniens de DMLA humaine post-mortem, nous validé cet axe microglial-astrocytaire pro-angiogénique médié par l'IL-1β dans la DMLA néovasculaire à un stade avancé. Comme IL-1β inflammatoire et gliale la signalisation sont importantes dans la MA et la MS10-12, ces voies représentent signatures moléculaires gliales partagées entre les affections neurodégénératives qui affectent la rétine et le cerveau. Cette étude offre à la fois une cadre pour identifier les populations cellulaires affectées par la maladie et signatures de maladies à partir de données unicellulaires complexes ainsi que d'informations clés dans les moteurs communs de la neurodégénérescence.

Résultats

CATCH identifie, caractérise et compare efficacement populations enrichies en maladies dans des données transcriptomiques unicellulaires complexes

Faisant partie du système nerveux central (SNC), la rétine contient de nombreux différentes couches fonctionnelles et strates distinctes occupées par un ensemble très diversifié de types et d'états de cellules (Fig. 1A). De plus, en tant que composant du SNC, la rétine partage des caractéristiques avec le cerveau au niveau niveau de biologie cellulaire et de pathologie dégénérative (Fig. 1B). Semblable à La DMLA, la SEP et la MA ont des phases définies de la maladie, chacune avec un début ou un stade aigu actif et un stade tardif ou chronique inactif de la maladie13–15 . Pour identifier états cellulaires pathogènes enrichis en DMLA, et les relier aux états trouvés dans la MA et la SEP, nous avons effectué un séquençage de snRNA basé sur la microfluidique massivement parallèle pour profiler les lésions de la macula de 11 rétines avec différents degrés de pathologie DMLA et 6 échantillons de contrôle, créant ainsi un vue unicellulaire de la pathologie DMLA. Nous avons ensuite appliqué un pipeline, CATCH, pour analyser cet ensemble de données en groupements significatifs de types de cellules et d'états identifier les mécanismes pathogènes de la maladie, qui peuvent être partagés dans les maladies neurodégénératives. Nous avons utilisé snRNA-seq pour notre analyse, qui s'est avérée performante en termes de sensibilité et de classification des types de cellules par rapport au scRNA-seq16. snRNA-seq a le avantages supplémentaires: il minimise les changements d'expression génique résultant de la dissociation des tissus et minimise les défis de dissociation pour les tissus tels que la rétine et le cerveau

Les cellules peuvent exister dans divers états transcriptionnels, qui naturellement tomber dans une hiérarchie ou une organisation. Au sein de cette hiérarchie, les cellules d'un niche fonctionnelle plus similaire, par exemplemicroglie et astrocytes, sont plus étroitement liées les unes aux autres que les cellules d'un amour plus disparates niche, par exemple les microglies et les cellules endothéliales. Apprendre ceci La hiérarchie à partir des données est importante pour le développement d'une approche systématique compréhension de la fonction biologique et peut fournir un aperçu de mécanismes de pathogenèse de la maladie. Comme les types de cellules peuvent être différemment affectés par la maladie, l'identification et l'identification simultanées caractérisation de classes abondantes de cellules à granularité grossière comme ainsi que des types de cellules rares ou des états à granularité fine, fournit un cadre complet pour la définition, la modélisation et la compréhension. voies cellulaires spécifiques dans la maladie. Même si les données biologiques ont structure à de nombreux niveaux de granularité différents, la plupart étant groupés les méthodes offrent un ou quelques niveaux de granularité. Ces quelques niveaux de granularité peut créer des identifications inexactes des états cellulaires associés à la maladie. Pour résoudre ce problème, nous avons développé CATCH, un cadre qui combine les principes de la géométrie des collecteurs de données avec la topologie informatique pour créer une meilleure compréhension de états cellulaires à travers les granularités. Alors que la composante essentielle de CAPTEUR, condensation par diffusion8 , et ses propriétés mathématiques17 ont été établis et utilisés pour identifier la structure multigranulaire dans les ensembles de données biomédicales18, il n'a pas été appliqué aux données transcriptomiques unicellulaires. Ici, nous avons adapté et construit un pipeline autour de la condensation par diffusion pour balayer systématiquement tous les possibles granularités de la hiérarchie cellulaire pour identifier les populations pathogènes et déduire les mécanismes de neurodégénérescence.

Pour apprendre la hiérarchie cellulaire à partir de données transcriptomiques complexes unicellulaires, nous avons adapté la condensation par diffusion pour efficacement déplacer les cellules vers leurs voisines les plus similaires en termes de profil transcriptomique à travers les itérations successives. Quand les cellules s'effondrent les uns dans les autres, la condensation par diffusion les fusionne, les regroupant ainsi à un niveau de granularité spécifique (Fig. 1C). Par En se condensant lentement puis en fusionnant des cellules similaires, la condensation par diffusion apprend efficacement comment les cellules interagissent les unes avec les autres sur des centaines de niveaux de granularité. Puisque la condensation par diffusion ne forcer les cellules à fusionner à n'importe quelle itération donnée, comme le font d'autres approches de regroupement hiérarchique, la durée pendant laquelle une cellule ou un groupe de cellules cellules fusionnées, reste persistant dénote non seulement leur transcriptomique interdépendance mais aussi leur caractère unique par rapport aux autres cellules. Des cellules qui ne prennent que quelques itérations pour fusionner sont très similaires les unes aux autres, tandis que les cellules qui nécessitent un nombre important d'itérations pour fusionner sont plus différents dans leur profil transcriptomique global. Cette approche est fondamentalement distinct du clustering de détection communautaire populaire des méthodes basées sur des métriques telles que la modularité et le score de silhouette, qui optimiser les étiquettes de cluster en fonction de l'interconnectivité du réseau. La condensation par diffusion est une approche à gros grains qui fusionne lentement des populations similaires ensemble à toutes les échelles. Cette fonctionnalité de l'algorithme nous permet d'effectuer une analyse en aval et d'identifier des populations enrichies en états pathologiques.

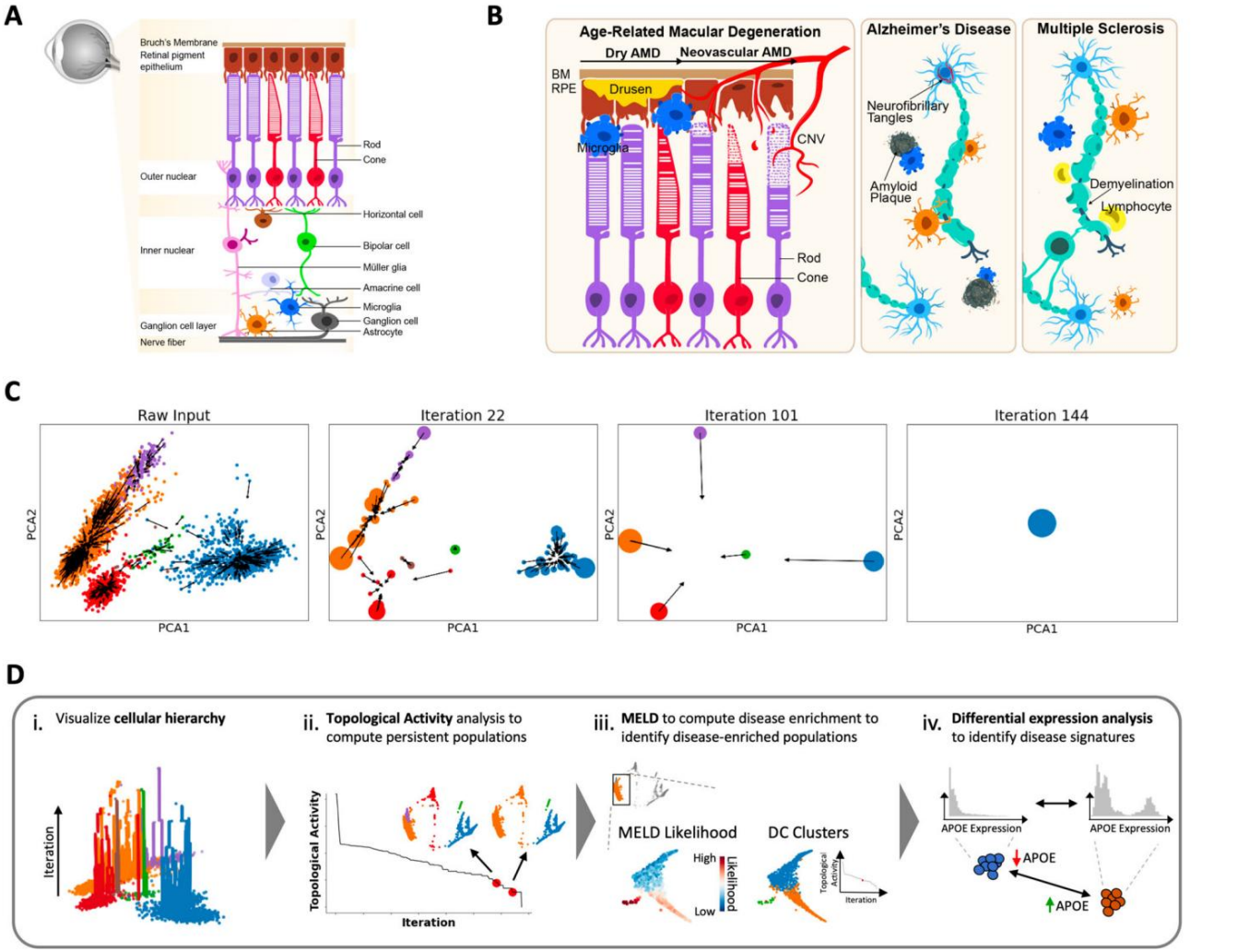

Figure 1 | Aperçu des processus des maladies neurodégénératives et de la topologie approche de condensation par diffusion. Un croquis de la coupe transversale de la rétine montrant les couches et les principaux types de cellules. B Illustration du rôle des cellules immunitaires innées dans la pathogenèse des maladies neurodégénératives. Au stade sec de la DMLA, il y a accumulation de débris de drusen extracellulaires entre la membrane de Bruch (BM) et l'épithélium porcin rétinien (RPE), conduisant à l'activation de la glie. Au stade néovasculaire avancé de Une DMLA et une néovascularisation choroïdienne (NVC) médiée par le VEGF se développent, ce qui peut entraîner à la perte de vision due à la mort des cellules photoréceptrices en bâtonnets et en cônes. Accumulation de plaques extracellulaires et enchevêtrements neurofibrillaires intracellulaires dans la maladie d'Alzheimer et les dommages à la myéline dans la sclérose en plaques progressive sont tous deux accompagnés de activation des microglies (bleu) et des astrocytes (orange). C Description visuelle du cellulaire processus de condensation entrepris par condensation par diffusion sur quatre granularités. Les points sont déplacés et fusionnés avec leurs voisins les plus proches au fur et à mesure déterminé par une marche aléatoire pondérée sur le graphique de données. Au fil de nombreuses successions itérations, les cellules s'effondrent, dénotant l'identité du cluster à diverses itérations. D Le le processus de granularité grossière décrit dans Ccrée des centaines de granularités de clusters, qui peuvent être analysés de manière significative : (i) nous pouvons visualiser la hiérarchie des clusters calculés par condensation de diffusion, pour identifier le comportement de fusion à travers les granularités ; (ii) nous pouvons identifier des partitions significatives et persistantes des données en effectuant une analyse d'activité topologique ; (iii) en conjonction avec MELD9 , nous pouvons analysez ces granularités significatives pour identifier les résolutions qui conviennent de manière optimale séparer les populations de cellules enrichies en maladies des populations de cellules saines et enfin; (iv) nous pouvons calculer des gènes différentiellement enrichis entre des populations de intérêt.

Le framework CATCH utilise la caractéristique de persistance de condensation de diffusion pour apprendre et analyser la hiérarchie cellulaire afin identifier les états transcriptomiques pathogènes et créer des signatures robustes de la maladie à partir de données unicellulaires. La hiérarchie cellulaire est visualisé pour identifier la structure hiérarchique et de persistance du données (Fig. 1D-i). Des granularités significatives de la hiérarchie cellulaire sont identifié grâce à l'analyse de l'activité topologique, une analyse qui identifie des granularités hautement persistantes et stables pour la caractérisation en aval (Fig. 1D-ii). Grâce à cette analyse, nous identifions des clusters qui isoler les cellules trouvées de manière disproportionnée dans des échantillons pathogènes ou sains à l'aide de la méthode d'analyse d'enrichissement unicellulaire MELD9 (Fig. 1D-iii). Enfin, nous créons de riches signatures de maladies en identifiant des gènes exprimés différemment dans des populations de cellules pathogènes à l'aide d'une approche rapide. modification de la distance du Earth Mover (EMD) qui exploite la hiérarchie cellulaire (Fig. 1D-iv).

Pour plus de détails sur chaque composante du CATCH pipeline, y compris les adaptations à la condensation par diffusion, visualisation de la hiérarchie cellulaire, analyse de l'activité topologique et notre implémentation de l'analyse d'expression différentielle, voir rubrique méthodes.

Comparaison avec d'autres algorithmes de clustering sur synthétique et réel données unicellulaires. Nous avons comparé notre approche CATCH à stratégies de clustering existantes appliquées aux données unicellulaires. Utilisation d'une combinaison de 40 ensembles de données synthétiques unicellulaires ainsi que de données réelles unicellulaires et les données de cytométrie en flux, nous avons comparé les performances de regroupement de notre implémentation adaptée de la condensation par diffusion contre Lou-vain et Leiden, techniques de clustering multigranulaire souvent appliquées à données unicellulaires dans les packages de Monocle 3, ainsi que Shared de Seurat Algorithme de clustering des voisins les plus proches et FlowSOM, état de l'art

méthodes de regroupement de transcriptomiques et de cytométrie en flux unicellulaires
données, respectivement.

Splatter est un simulateur de données réalistes sur une seule cellule où
les étiquettes des clusters de vérité terrain sont connues19. Utiliser ces vérités terrain
étiquettes, nous avons généré des ensembles de données unicellulaires de plus en plus bruyants avec
deux types différents de bruit biologique : variation et décrochage
(Fig. 1A supplémentaire). Avec chacun de ces ensembles de données, nous suivons
le cadre CATCH : nous calculons et visualisons d'abord l'homologie de condensation (Fig.
1B supplémentaire) avant d'effectuer
analyse de l'activité topologique pour identifier les quatre granularités les plus persistantes
(Fig. 1C supplémentaire), puis enfin calculer l'indice rand ajusté, une mesure courante
pour déterminer
précision du regroupement par rapport à un ensemble d'étiquettes de regroupement de vérité terrain
(Fig. 1D supplémentaire), en conservant le score le plus élevé de notre
comparaisons. Curieusement, la population la plus persistante (iv),
avait presque toujours le score d'indice Rand ajusté le plus élevé. En utilisant
cette approche de comparaison, nous avons comparé la condensation par diffusion
aux algorithmes de regroupement des voisins les plus proches partagés de Louvain,
Leiden et Seurat sur 40 ensembles de données synthétiques unicellulaires. Pour
Louvain et Leiden de l'approche comparative, quatre différentes
les résolutions des clusters ont été calculées et comparées, en gardant
seule la comparaison, qui a produit le rand ajusté le plus élevé
indice. Qu'il s'agisse de niveaux croissants d'abandon scolaire ou d'augmentation
degrés de variation, CATCH a mieux performé que Louvain,
Leiden et Shared Nearest Neighbours de Seurat regroupant des algorithmes dans 10
simulations différentes. Alors que le bruit augmentait à 0,7
et 0,9 abandon et variation de 0,3 et 0,4, CATCH a surpassé les autres approches de
manière statistiquement significative
($p < 0,05$, ANOVA avec tests t de Student bilatéraux post hoc avec
correction de comparaisons multiples) (Fig. 1E supplémentaire).

Ensuite, nous avons comparé CATCH au clustering de Louvain et de Leiden
approches sur des données réelles unicellulaires où les clusters multigranulaires avaient
été identifié par un expert biologique20,21. Tout d'abord, nous avons analysé le réel
données transcriptomiques unicellulaires générées à partir d'un poisson zèbre en développement
avec des vérités fondamentales connues sur les clusters de types de cellules20. Nous avons organisé ces
étiquettes de cluster en étiquettes de cluster multigranulaires en agrégeant d'abord 18
types de cellules trouvés dans quatre types de tissus avant de les agréger en
trois couches germinales. De cette manière, nous avons produit un cluster de vérité terrain
étiquettes à travers les granularités. Nous avons ensuite comparé les quatre granularités
CATCH les plus persistantes avec des clusters multigranulaires calculés
en utilisant Louvain et Leiden, en réglant à nouveau le paramètre de résolution sur
produire dix étiquettes de cluster différentes. À toutes les granularités de la vérité terrain
labels de cluster, CATCH a surperformé Louvain et Leiden malgré
plus de granularités sont calculées pour les approches de comparaison
(Fig. 3B supplémentaire).

Enfin, comme l'analyse de déclenchement par cytométrie en flux est pratiquée depuis longtemps
comme référence en matière d'identification et de comparaison des types de cellules,
nous avons comparé CATCH à d'autres approches de clustering sur le flux
données de cytométrie. En utilisant 1,3 millions de cellules générées à partir de 30
patients, nous avons comparé les performances de CATCH à celles de Louvain,
Leiden et l'étalon-or du clustering par cytométrie en flux
FluxSOM21. Dans les 30 comparaisons, CATCH a surpassé les autres comparaisons de
manière statistiquement significative.
(test t bilatéral entre CATCH et chacun des autres clustering
approches, valeur p <0,01) (Fig. 3A supplémentaire). Tous ces
les comparaisons établissent que CATCH identifie les populations connues
des cellules dans les données synthétiques et réelles des cellules de signal mieux que les
techniques établies, en particulier lorsqu'il y a un degré élevé de
bruit biologique et variation. De plus, CATCH calcule un
hiérarchie complète des états cellulaires lors de l'identification des populations, permettant
de regrouper rapidement des groupes de cellules pour les identifier
états d'activation d'intérêt. Cette sous-population de cellules constitue un
sous-regroupement direct du groupe de grains les plus grossiers d'intérêt,
permettant la comparaison des états d'activation cellulaire. Tandis qu'un
peut modifier à plusieurs reprises les paramètres d'autres techniques pour acquérir
des grappes de grains plus fins ou plus grossiers, ces regroupements seraient déconnectés
les uns des autres, ce qui signifie qu'une hiérarchie complète est
non capturés et les groupes cellulaires d'une course à l'autre peuvent changer
considérablement. CATCH résout ce problème en identifiant les regroupements à travers
granularités dans un cadre unique.

Pour valider davantage l'analyse informatique, nous effectuons
études d'ablation sur chaque composant du pipeline CATCH (Fig. 2 supplémentaire).
Enfin, nous montrons la capacité de ce pipeline à
identifier les types de cellules rares (Fig. 5 supplémentaire) et les signatures de
populations de maladies dans des données unicellulaires réelles (Fig. 10 supplémentaire).
Pour un aperçu de l'analyse informatique et des comparaisons supplémentaires, voir la
section méthodes.

Analyse de séquençage d'ARN mononucléaire de la macula chez des individus humains
atteints d'une pathologie DMLA. Nous avons appliqué CATCH à l'ensemble de données
AMD snRNA-seq pour identifier les principaux types de cellules présents dans le contrôle.
et des échantillons AMD. Nous avons effectué une analyse d'activité topologique et
identifié trois granularités de la hiérarchie cellulaire pour une analyse en aval (granularités
à faible activité et forte persistance). Nous avons visualisé l'ensemble de données snRNA-
seq en utilisant PHATE et le
Clusters définis par CATCH au niveau des deux granulations identifiées les plus grossières
(Fig. 2A). En visualisant la troisième granularité, nous
observé un certain nombre de clusters, que nous avons classés en types de cellules
sur la base de l'expression de gènes marqueurs spécifiques au type de cellule
précédemment établis22 (Fig. 4A supplémentaire) (voir Méthodes).
Grâce à cette approche, nous avons identifié des types de cellules neuronales, notamment
cellules ganglionnaires rétiniennes, cellules horizontales, cellules bipolaires, photorécepteurs
à bâtonnets, photorécepteurs à cônes et cellules amacrines, ainsi que
types rares de cellules non neuronales, notamment les microglies, les astrocytes,
Glie de Müller et cellules vasculaires (Fig. 2B, C). Pour déterminer si ces
populations pourraient être trouvées avec des approches établies, nous
appliqué le clustering Louvain23 aux données monocellulaires AMD. Louvain
a révélé 22 populations à granularité grossière et 40 populations
à granularité fine (Fig. Supplémentaire 5A, B). Cependant, dans les deux résolutions, de
rares types de cellules immunitaires innées, comme les microglies,
les astrocytes et les gliales de Müller, n'ont pas été identifiés avec le Louvain
méthode, avec des marqueurs spécifiques à ces types de cellules ne localisant pas
à n'importe quel cluster. Enfin, démontrer la capacité de CATCH à
identifier des populations significatives de cellules à travers les granularités, nous
exploré plus en détail les sous-types de cellules bipolaires, un ensemble diversifié
d'interneurones qui transmettent des signaux à partir de photorécepteurs en bâtonnets et en cônes
aux cellules ganglionnaires de la rétine24-26. En analysant une granularité grossière de
Dans les cellules bipolaires, nous avons identifié les deux premiers sous-types principaux,
ON-center et OFF-center (Fig. 4B supplémentaire). En analysant un
une granularité plus fine, nous avons identifié les 12 principaux sous-types de cellules sur la base
sur l'expression de gènes marqueurs spécifiques au sous-type cellulaire (Fig. 4C – E
supplémentaire).

Identifier les types de cellules impliqués dans la pathogenèse de la DMLA dans un
de manière impartiale, nous avons appliqué un différentiel basé sur la condensation
analyse de l'expression des types de cellules identifiés par CATCH. En comparant
les cellules provenant de rétines sèches ou néovasculaires
DMLA aux cellules des rétines témoins, nous avons identifié différentiellement
gènes exprimés en utilisant la distance de Earth Mover dans chaque type de cellule
(définir la valeur p corrigée du FDR <0,1 dans toutes les comparaisons)27. En analysant
le nombre de gènes exprimés différentiellement dans toutes les cellules
types, nous avons constaté que les cellules vasculaires, les microglies et les astrocytes avaient le
le plus grand nombre de gènes différentiellement exprimés à travers les étapes de
AMD comparée aux échantillons témoins (Fig. 2D). De plus, nous avons effectué une
analyse d'abondance pour identifier si certains types de cellules étaient significativement
plus enrichis en DMLA sèche ou néovasculaire. Ce
l'analyse a révélé une augmentation statistiquement significative de la proportion
de noyaux de microglies et d'astrocytes provenant de donneurs présentant à la fois des
DMLA néovasculaire par rapport aux échantillons témoins (test multinomial bilatéral,
valeur $p < 0,01$) (Fig. 2E). De plus, il y avait un enrichissement statistiquement significatif
en cellules vasculaires dans la DMLA néovasculaire,

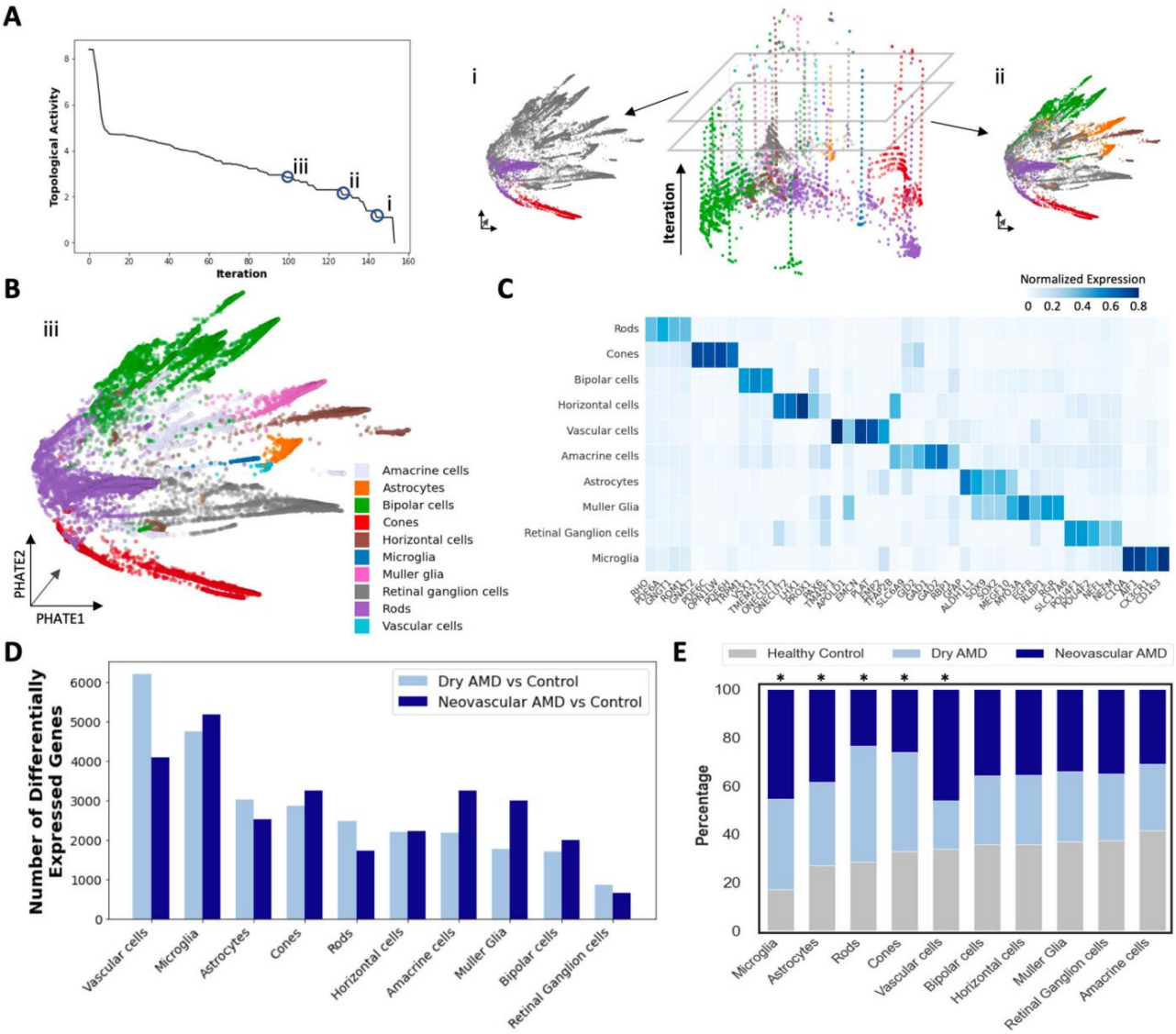

Figure 2 | Profilage d'ARN mononucléaire de la macula d'individus humains présentant différents stades de pathologie de la DMLA. Une analyse (à gauche) de l'activité topologique des données unicellulaires de la rétine humaine sur toutes les itérations de condensation. En calculant les gradients sur l'activité topologique (voir Méthodes), nous identifions trois granularités auxquelles se produisent des partitions persistantes des données (représentées par les résolutions i, ii et iii) et les sélectionnons pour une analyse en aval. (à droite) Processus de condensation des données monocellulaires AMD visualisés à travers les itérations (de bas en haut) avec les clusters de granularité les plus grossiers visualisés sur l'intégration PHATE : résolution i. représente les clusters et la résolution les plus grossiers ii. représente le deuxième cluster le plus grossier. B Les populations identifiées avec la granularité la plus fine identifiées par l'analyse d'activité topologique (résolution iii.) ont été visualisées et toutes les populations se sont vu attribuer un type de cellule en fonction de la signature génétique du type de cellule dont elles présentaient l'expression la plus élevée. C Gènes spécifiques au type de cellule visualisés ainsi que l'expression normalisée moyenne de gènes marqueurs spécifiques au type de cellule connu. Tous les principaux types de cellules rétiniennes ont été identifiés par le processus CATCH décrit dans A, B. D Gènes exprimés différentiellement et identifiés par la distance de Wasserstein Earth Mover (EMD) entre les cellules provenant de lésions de DMLA néovasculaire sèche à un stade précoce et à un stade avancé et les cellules provenant de rétines témoins sur une base spécifique au type de cellule. Nombre de gènes exprimés de manière significativement différentielle entre les cellules témoins et les cellules AMD, rapportés d'une manière spécifique au type de cellule et au stade (valeur p corrigée par le FDR < 0,1). Types de cellules triés selon la plupart des gènes différentiels entre la DMLA sèche et la comparaison témoin. Les cellules vasculaires, les microglies et les astrocytes possèdent les gènes exprimés de la manière la plus différentielle dans la DMLA sèche par rapport aux échantillons témoins. Le graphique à barres E indique la contribution des types de cellules dans chaque cluster à partir des échantillons de contrôle, de DMLA sèche et de DMLA néovasculaire. Les microglies et les astrocytes sont les types de cellules statistiquement les plus enrichis dans la DMLA, tandis que les bâtonnets et les cônes sont les types de cellules les plus appauvris dans la DMLA néovasculaire. Les cellules vasculaires constituent le type de cellule le plus enrichi dans la DMLA néovasculaire. Toutes les statistiques ont été calculées à l'aide de tests multinomiaux bilatéraux avec correction de comparaisons multiples (* p < 0,1).

mettant en évidence l'importance des cellules vasculaires dans le développement de l'angiogenèse pathologique présente à ce stade de la maladie (test multinomial bilatéral, valeur p < 0,01). Il y avait une diminution relative de l'abondance des photorécepteurs en bâtonnets et en cônes dans la DMLA néovasculaire avancée, ce qui concorde avec la perte connue de photorécepteurs au stade avancé de la maladie (test multinomial bilatéral, valeur p < 0,01) (Fig. 2E) .). Ces résultats suggèrent que les types de cellules non neuronales, notamment les microglies, les astrocytes et les cellules vasculaires, sont des types de cellules importants dans la pathogenèse de la DMLA, avec non seulement les plus altérations transcriptionnelles mais également changements d'abondance au cours de la progression de la DMLA.

La signature d'activation microgliale identifiée dans la DMLA sèche est partagée au cours de la phase précoce de plusieurs maladies neurodégénératives. Bien que les états d'activation des microglies et leur dynamique aient été identifiés dans des modèles murins d'AD7 et des états d'expression associés trouvés chez l'homme28, on ne comprend pas bien dans quelle mesure ces états et dynamiques sont communs à toutes les maladies neurodégénératives humaines. Le

L'étude des microglies dans le SNC a été difficile en raison de leur rareté, nécessitant des stratégies d'enrichissement ciblées7,28. Grâce à la capacité de CATCH à parcourir toutes les hiérarchies de clusters, nous pouvons identifier des sous-populations de types de cellules rares avec une granularité fine pour effectuer une analyse rigoureuse et approfondie des états cellulaires. Pour identifier les sous-populations microgliales enrichies dans des phases spécifiques de la DMLA et construire des signatures transcriptomiques de la maladie, nous avons identifié des granularités CATCH qui isolaient des scores de probabilité MELD élevés calculés pour les conditions de DMLA témoin, sèche et néovasculaire. Nous avons calculé les scores de vraisemblance MELD pour chaque condition sur toutes les microglies de la DMLA (Fig. 3A). Ensuite, nous avons identifié une granularité mise en évidence par l'analyse de l'activité topologique qui séparait les régions à forte probabilité de maladie des régions à faible probabilité de maladie (voir Méthodes). Avec cette approche, nous avons identifié trois clusters, chacun enrichi pour une condition différente : un cluster enrichi pour les cellules provenant d'échantillons de contrôle, un cluster enrichi pour les cellules provenant d'échantillons précoces et secs de DMLA, et un cluster enrichi pour les cellules d'un stade avancé. échantillons de DMLA néovasculaire (Fig. 3A).

Pour identifier les signatures de la DMLA présentes dans les microglies au début de la pathogenèse de la maladie sèche, phase dans laquelle les microglies ont déjà été impliquées2 , nous avons effectué une analyse d'expression différentielle entre les clusters enrichis en contrôle et les clusters enrichis en DMLA sèche. En analysant les gènes les plus différentiellement exprimés (valeur p corrigée par FDR <0,1) entre ces sous-populations, une signature d'activation claire est apparue dans le groupe précoce et sec enrichi en DMLA, comprenant APOE, TYROBP et SPP1 (Fig. 3D) , gènes connus pour jouer un rôle dans la neurodégénérescence7 . L'association de TYROBP et d'APOE a été validée sur des coupes de macula rétinienne humaine par immunofluorescence simultanée pour IBA1, un gène associé aux microglies, et hybridation in situ pour TYROBP et APOE. Sur des coupes de macula rétinienne humaine, les cellules IBA1-positives provenant de patients atteints de DMLA sèche ont montré un enrichissement par rapport aux contrôles des transcrits de gènes de TYROBP et APOE, indiquant la polarisation d'un sous-ensemble de microglies vers le phénotype microglial neurodégénératif au début de la maladie (Fig. 3G) .

Une expression accrue de TYROBP et d'APOE dans la microglie a également été identifiée par hybridation in situ sur des lésions de tissu cérébral humain présentant une MA à un stade précoce et une SEP progressive précoce par rapport aux témoins (Fig. 7C supplémentaire).

En raison de la similitude entre cet état d'activation et un état microglial associé à une maladie précédemment défini et décrit chez la souris7,29, nous avons effectué une analyse complète des états microgliaux dans deux autres maladies neurodégénératives, la MA et la SEP progressive.

En appliquant l'approche CATCH aux données snRNA-seq d'AD4 et MS5 , nous avons identifié tous les principaux types de cellules sur la base de l'expression de gènes marqueurs spécifiques au type de cellule (Fig. Supplémentaire 6A – D). Comme dans la DMLA, l'analyse d'enrichissement a révélé que les microglies étaient significativement enrichies dans la MA et la SEP par rapport au tissu cérébral témoin (Fig. 6E, F supplémentaires). Semblable à notre analyse de la DMLA identifiant les états transcriptomiques spécifiques à la phase de la maladie, nous avons appliqué l'analyse MELD et l'activité topologique aux microglies dans les ensembles de données AD et MS et avons identifié trois groupes de microglies dans chaque maladie : un groupe enrichi en cellules provenant du tissu cérébral témoin ; un groupe enrichi en cellules provenant de tissus atteints de MA à un stade précoce ou de lésions actives aiguës de SEP ; et un cluster enrichi en cellules provenant de tissus atteints de MA à un stade avancé ou de lésions chroniques inactives de SEP (Fig. 3B, C). L'analyse de l'expression différentielle entre les groupes enrichis en contrôle et enrichis en début de maladie a donné un profil d'activation commun partagé dans les trois maladies lors de l'analyse des gènes les plus exprimés différentiellement (Fig. 3D, panneaux du milieu et de droite) (valeur p corrigée par FDR < 0,1)).

Pour comprendre les populations microgliales enrichies en début de maladie, nous avons visualisé la signature d'activation microgliale (CD74, SPP1, VIM, FTL, B2M) (APOE, TYROBP, CTSB) (C1QB et C1QC) ainsi qu'une signature homéostatique (P2RY12, P2RY13, et OLFML3) sur des clusters de maladies neurodégénératives enrichis en contrôle et enrichis en début de maladie (Fig. 3E). Une nette divergence est observée entre le modèle d'expression de la signature homéostatique dans les populations enrichies en contrôle et les populations précoces. populations enrichies en maladies dans toutes les conditions. Avec une expression plus élevée des gènes d'activation et une expression plus faible des gènes homéostatiques, la population précoce de microglies activées présente un état de polarisation divergent. Nous avons construit une signature d'activation microgliale composite et l'avons cartographiée sur les grappes avec une signature microgliale associée à la maladie décrite précédemment trouvée dans un modèle de souris AD7 . Le stade précoce des clusters enrichis en maladies neurodégénératives présentait une expression plus élevée des deux signatures par rapport aux clusters enrichis en contrôle (Fig. 3F avec des valeurs d'expression allant de 5 à 25 pour notre signature d'activation et de 7 à 26 pour la signature DAM).

Ce phénotype microglial neurodégénératif commun à la DMLA, à la SEP et à la MA implique une régulation positive de plusieurs gènes impliqués dans les études sur le risque de maladie neurodégénérative. Ceux-ci incluent l'APOE, un régulateur clé de la transition entre les états homéostatiques et neurotoxiques dans la microglie30, fortement impliqué dans le risque de MA31,32 et de DMLA33 ; TYROBP qui code pour la protéine adaptatrice TREM2 DAP12, dont les mutations sont impliquées dans un syndrome du lobe frontal avec une pathologie de type MA34 et dont l'expression est régulée positivement dans les microglies de la substance blanche dans les lésions de SEP ; SPP1 (ostéopontine), impliquée dans l'activation microgliale dans les cerveaux affectés par MS35 et AD36 ; et CTSB, codant pour la protéase majeure de la cathepsine-B des lysosomes, qui est régulée positivement dans les microglies répondant aux plaques β-amyloïdes dans AD36. L'initiation de l'accumulation pathologique de matériel extracellulaire se produit par différents moyens dans ces trois maladies neurodégénératives. Cependant, la découverte selon laquelle les voies d'activation microgliales phagocytaires, lipidiques et lysosomales sont régulées positivement au stade actif précoce ou aigu des trois maladies suggère un rôle convergent pour la dérégulation de la microglie orientée vers l'élimination des dépôts extracellulaires de débris.

La signature d'activation des astrocytes identifiée dans la DMLA sèche est partagée au cours de la phase précoce de plusieurs maladies neurodégénératives. Bien que les états et la dynamique transcriptomiques des astrocytes aient été établis dans des modèles murins de MA, les profils des astrocytes n'ont pas été profilés dans les lésions de la DMLA humaine à un moment donné. résolution cellulaire6 . Comme notre analyse initiale impliquait les astrocytes dans la pathogenèse de la maladie (Fig. 2D, E), nous avons effectué une analyse croisée similaire au sein des populations d'astrocytes en utilisant la méthode CATCH. En utilisant MELD et l'analyse d'activité topologique, nous avons identifié quatre groupes d'astrocytes à granularité fine dans la hiérarchie de condensation de diffusion : un groupe enrichi en cellules provenant d'échantillons témoins, un groupe enrichi en cellules provenant de patients atteints de DMLA sèche et précoce, un groupe enrichi en cellules provenant d'échantillons de contrôle. des patients atteints de DMLA néovasculaire à un stade avancé et un groupe contenant un nombre égal de cellules provenant des trois affections (Fig. 4A). Lors de la comparaison des profils transcriptomiques des cellules au sein des populations d'astrocytes sèches enrichies en DMLA et enrichies en contrôle, les gènes clés associés à l'activation et à la dégénérescence, tels que GFAP, VIM et B2M, ont été régulés positivement (Fig. 4D) .

À l'aide de MELD et de l'analyse de l'activité topologique, nous avons identifié des groupes qui isolaient des populations spécifiques à un stade donné au sein des astrocytes MS et AD. Dans les deux maladies, nous avons identifié trois groupes : un groupe enrichi en cellules provenant de tissus cérébraux témoins, un groupe enrichi en cellules provenant de tissus atteints de MA à un stade précoce ou de lésions actives aiguës de SEP, et un groupe enrichi en cellules provenant de tissus atteints de MA à un stade avancé ou de lésions chroniques inactives de SEP (Fig. 4B, C). En comparant les clusters enrichis en contrôle et en maladies précoces au sein de chaque ensemble de données à l'aide d'un transport condensé, nous avons identifié une signature génétique partagée enrichie dans le sous-groupe de maladies neurodégénératives à un stade précoce pour les trois maladies (Fig. 4E) . La signature génétique intégrée comprenait des marqueurs d'astrocytes activés, notamment VIM, GFAP, CRYAB et CD8137,38, un complexe majeur d'histocompatibilité (CMH) de classe I (B2M) 39,40, le métabolisme du fer (FTH1 et FTL), un composant des canaux hydriques impliqué dans l'élimination des débris (AQP4) 41, ainsi que dans l'activation lysosomale et la phagocytose lipidique et amyloïde (CTSB, APOE). Il est intéressant de noter que de nombreux gènes régulés positivement étaient partagés entre les signatures d'activation microgliales et astrocytes à un stade précoce, ce qui suggère que les voies communes du stress glial sont activées dans la neurodégénérescence.

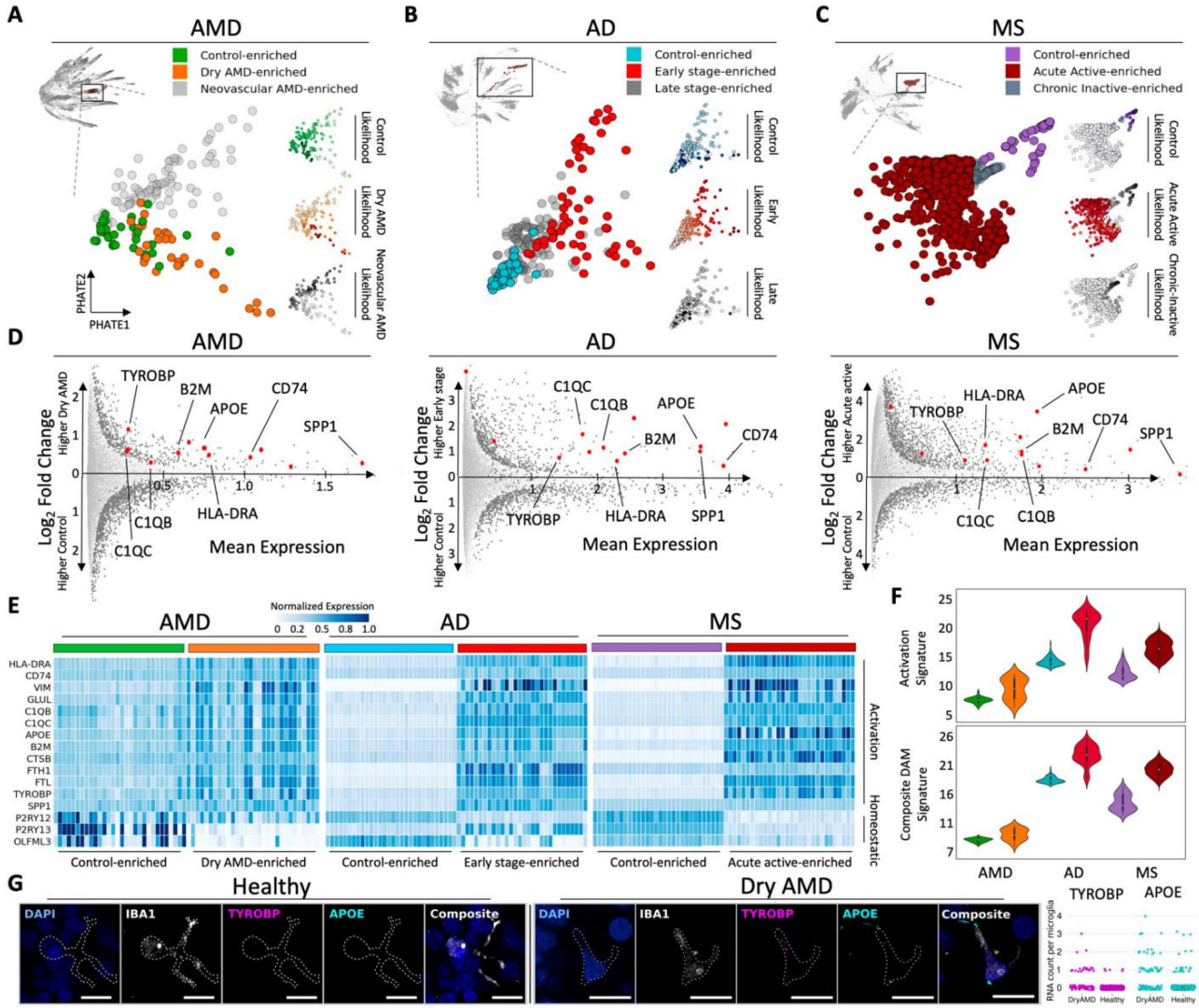

Figure 3 | L'analyse des grains fins des microglies révèle une signature d'activation commune enrichie au début de trois maladies neurodégénératives différentes. 141 microglies identifiées par condensation de diffusion à granularité grossière (en haut à gauche) peuvent être subdivisées en trois groupes à granularité fine, chacun enrichi en cellules provenant d'un état pathologique différent. L'enrichissement de l'état pathologique a été calculé à l'aide de MELD (à droite) pour chaque condition : contrôle (en haut), DMLA sèche (au milieu) et DMLA néovasculaire (en bas), avec des probabilités MELD plus élevées affichées avec des couleurs plus foncées. Une résolution de l'homologie de condensation, qui isole de manière optimale les scores de vraisemblance MELD de chaque condition, a été identifiée à l'aide d'une analyse d'activité topologique. Les microglies sont revisualisées à l'aide de PHATE. B Comme dans le panneau A, trois sous-ensembles de 288 microglies se trouvent dans la MA avec condensation de diffusion et analyse de l'activité topologique, chacun enrichi en cellules d'un stade différent de la pathologie, tel que calculé par MELD (à droite). Les cellules sont revisualisées avec PHATE. C Comme dans le panneau A, trois sous-ensembles de 1 263 microglies sont trouvés dans la SEP avec condensation par diffusion et analyse de l'activité topologique, chacun enrichi en cellules d'un stade différent de la maladie, tel que calculé par MELD (à droite). Les cellules sont revisualisées avec PHATE. D L'analyse de l'expression différentielle entre les microglies enrichies en contrôle et celles enrichies en phase active précoce ou aiguë dans les maladies neurodégénératives révèle un modèle d'activation partagé au début de la maladie (expression accrue de TYROBP, B2M, APOE, CD74, SPP1, HLA-DR, C1QB, C1QC). Les gènes différentiellement exprimés significatifs sont visualisés en gris foncé (test EMD bilatéral avec valeur p corrigée FDR <0,1 comme décrit dans les méthodes). E Heatmap démontrant les différences dans l'expression du modèle d'activation partagé neurodégénératif et un signature homéostatique entre les microglies enrichies en contrôle et les microglies enrichies en maladies actives précoces ou aiguës dans les maladies neurodégénératives. Les conventions de couleur sont celles des panneaux A à C. Les lignes correspondent aux gènes et les colonnes représentent les cellules individuelles. Nous avons tracé 40 cellules de chaque ensemble de données sélectionnées par échantillonnage aléatoire pour révéler la différence entre les états cellulaires de type contrôle et ceux de type maladie précoce. (F, en haut) Signature d'activation microgliale composite pour le modèle d'activation partagé neurodégénératif dans les microglies enrichies en contrôle et enrichies en maladies actives précoces ou aiguës dans les maladies neurodégénératives (axe y - expression génique de la signature). (F, inférieur) Signature de microglies associées à la maladie (DAM) (de la réf. 7) pour les microglies enrichies en contrôle et enrichies en maladies actives précoces ou aiguës dans les maladies neurodégénératives. Les conventions de couleur sont celles des panneaux A à C (axe y – expression génique de la signature). Les détails sur les statistiques sont disponibles dans la G Micrographies d'hybridation d'ARN in situ combinée et d'immunofluorescence IBA1 démontrant une expression élevée de composants clés du modèle d'activation partagé neurodégénératif (TYROBP et APOE) dans des cellules IBA1-positives, un marqueur de la microglie, provenant de rétines atteintes de DMLA sèche (à droite groupe) par rapport aux rétines témoins (groupe de gauche). Toutes les barres d'échelle = 10 µm. Le nombre moyen de points lacrymaux identifiés par cellule IBA1-positive pour TYROBP était de 0,28 ± 0,05 dans la DMLA sèche (n = 191) contre 0,02 ± 0,01 pour le contrôle (n = 464 ; p < 1e-10 ; test du Chi carré pour 0 vs .>0). Le nombre moyen de points lacrymaux identifiés par cellule IBA1-positive pour l'APOE était de 0,57 ± 0,09 dans la DMLA sèche contre 0,14 ± 0,03 pour le contrôle (p < 1e-08 ; test du Chi carré pour 0 contre >0).

Semblable à la microglie, nous avons cartographié les signatures d'activation homéostatiques (GPC5, LSAMP, TRPM3) et composites (B2M, CRYAB, VIM, GFAP, AQP4, APOE, ITM2B, CD81, FTL) sur des clusters d'astrocytes enrichis en début de maladie et enrichis en contrôle à travers maladies neurodégénératives. Similaire par rapport à la microglie, la signature d'activation composite et les signatures homéostatiques étaient exprimées de manière divergente par les premiers groupes enrichis (Fig. 4E, F supérieur avec des valeurs d'expression allant de 0 à 17). L'utilisation d'une signature d'astrocyte associée à une maladie récemment publiée et établie

https://doi.org/10.1038/s41467-023-37025-7

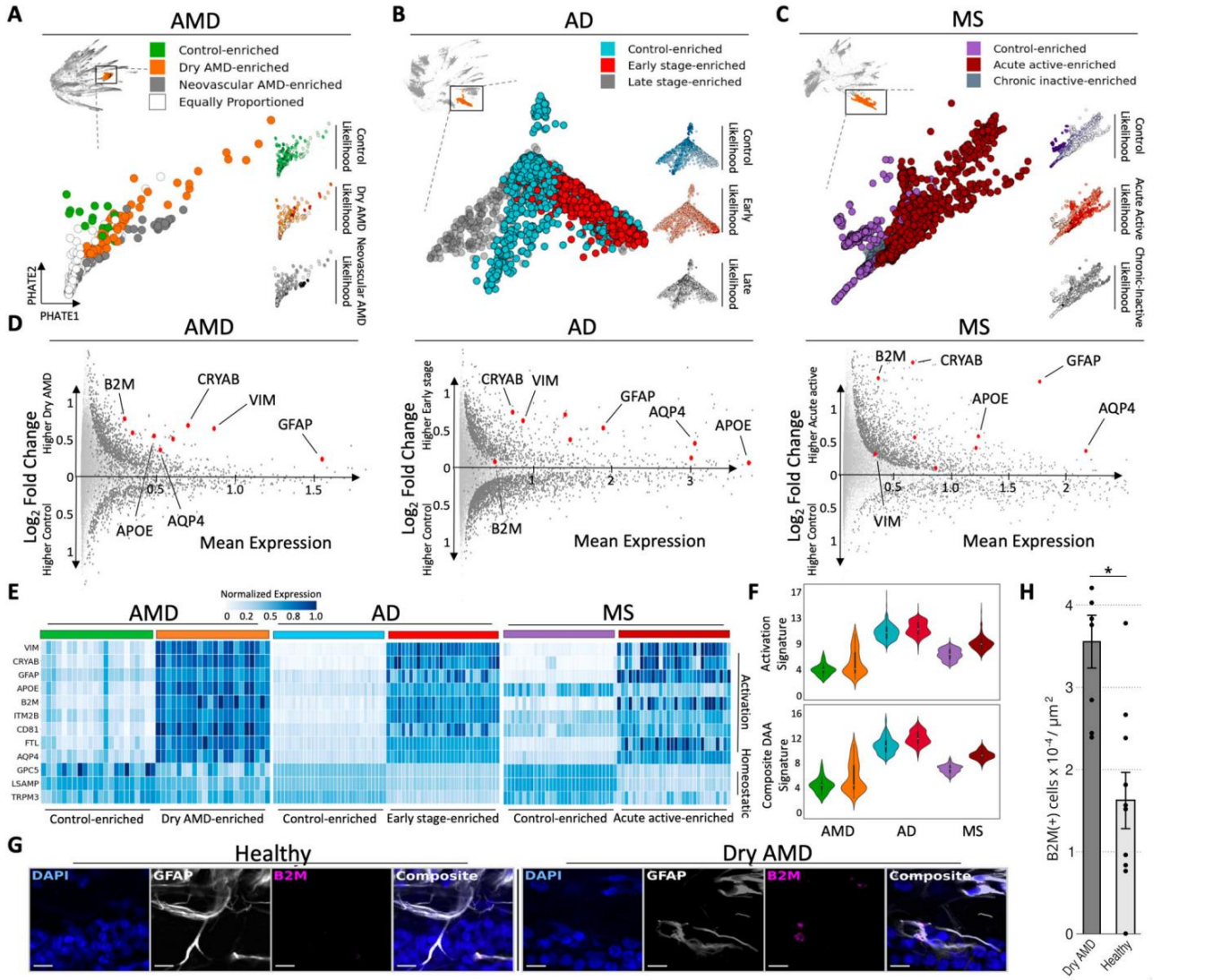

Figure 4 | L'analyse des grains fins des astrocytes révèle une signature d'activation partagée enrichi dans la phase précoce des maladies neurodégénératives. A 474 astrocytes identifié par condensation de diffusion à granularité grossière (en haut à gauche) peut être subdivisé en trois groupes à granularité fine, chacun enrichi pour les cellules d'un différents stades de la maladie neurodégénérative. L'enrichissement de l'état pathologique a été calculé à l'aide de MELD (à droite) pour chaque condition : contrôle (en haut), DMLA sèche (au milieu) et DMLA néovasculaire (en bas), avec des probabilités de MELD plus élevées affichées avec des couleurs plus foncées. Une résolution de l'homologie de condensation, qui isole de manière optimale les scores de vraisemblance MELD de chaque condition, a été identifiée à l'aide de l'activité topologique. analyse. Les astrocytes sont revisualisés à l'aide de PHATE. B Comme dans le panel A, trois sous-ensembles de 2361 astrocytes sont trouvés dans la MA avec condensation de diffusion et topologie analyse d'activité, chacune enrichie en cellules provenant d'un stade différent de la maladie d'Alzheimer, ainsi que calculé par MELD (à droite). Les astrocytes sont revisualisés avec PHATE. C Comme dans le panneau A, trois sous-ensembles de 5469 astrocytes sont trouvés dans la SEP avec condensation par diffusion et analyse de l'activité topologique, chacune enrichie pour les cellules d'un stade différent de MS tel que calculé par MELD (à droite). Les astrocytes sont revisualisés avec PHATE. D Analyse d'expression différentielle entre le stade enrichi de contrôle et le stade précoce de clusters enrichis en maladies neurodégénératives parmi les maladies neurodégénératives révèle un modèle d'activation partagé au stade précoce de la maladie. Cette signature inclut B2M, CRYAB, VIM, GFAP, AQP4, APOE, ITM2B, CD81, FTL. Significatif gènes différentiellement exprimés visualisés en gris foncé (test EMD bilatéral avec FDR valeur p corrigée <0,1 comme décrit dans les méthodes). E Heatmap démontrant les différences dans l'expression des astrocytes du modèle d'activation neurodégénérative partagée et une signature homéostatique entre enrichi par le contrôle et précoce ou aigu astrocytes actifs enrichis en maladies dans les maladies neurodégénératives. Les conventions de couleurs sont celles des panneaux A à C. Les lignes correspondent aux gènes et les colonnes représentent cellules individuelles. Nous avons tracé 40 cellules de chaque ensemble de données sélectionnées par échantillonnage aléatoire pour révéler la différence entre un type de contrôle et un type de maladie précoce. états cellulaires. F Signature composite d'activation des astrocytes (en haut) et signature des astrocytes associée à la maladie (DAA) pour l'activation neurodégénérative partagée modèle dans les clusters enrichis en contrôle et les clusters enrichis en maladies précoces dans les maladies neurodégénératives. Les conventions de couleur sont comme dans les panneaux A à C (axe y - gène expression de signature). Les détails sur les statistiques sont disponibles dans la section méthodes. G Micrographies d'hybridation d'ARN in situ combinée et d'immunofluorescence GFAP montrant une expression plus abondante de B2M dans la rétine riche en astrocytes. couches de rétine sèche de DMLA par rapport au contrôle. Toutes les barres d'échelle = 10 µm. Barre H tracé montrant la densité des transcrits B2M dans la couche plexiforme interne riche en astrocytes, couche de cellules ganglionnaires rétiniennes et couches de fibres nerveuses dans les échantillons de rétine affectés par la sécheresse AMD (n = 8 cellules) et contrôle (n = 10 cellules). Les données sont présentées sous forme de valeurs moyennes ± SEM ; *p < 1e-03 ; Test t de Welch à deux échantillons.

dans un modèle de souris AD6, nous avons construit une signature d'activation composite et cartographié cela sur les clusters de maladie précoce et enrichis en contrôle dans toutes les conditions. Les clusters enrichis en maladies précoces affichés expression plus élevée de la nature de la signature du gène des astrocytes associés à la maladie (DAA) en plus de la signature d'activation composite (Fig. 4F, inférieure avec des valeurs d'expression allant de 0 à 16).

Pour valider la signature des astrocytes dans les tissus, nous avons effectué immunofluorescence GFAP simultanée et hybridation in situ d'ARN pour B2M, un composant du CMH-I et membre du signature génétique partagée sur des sections de la macula humaine. Le Les couches rétiniennes occupées par les astrocytes GFAP-positifs (couche plexiforme interne jusqu'à la membrane limitante interne) contenaient une concentration plus élevée

densité des transcriptions B2M dans les rétines affectées par la DMLA sèche par rapport à la rétine témoin (valeur p < 1e-03, test t de Student bilatéral) (Fig. 4G, H).

Les microglies présentent une signature d'activation de l'inflammasome et les astrocytes affichent une signature pro-angiogénique dans la DMLA néovasculaire à un stade avancé

Bien que les signatures d'activation gliale soient partagées au cours de la phase précoce des maladies neurodégénératives multiples, il est intéressant de comprendre si elles persistent ou évoluent au stade avancé des maladies neurodégénératives. Pour comprendre cette dynamique d'activation gliale à travers les stades de la DMLA, de la MA et de la SEP, nous avons effectué une analyse d'expression différentielle entre le stade précoce des grappes enrichies en maladies neurodégénératives et le stade avancé des grappes d'astrocytes et de microglies enrichies en maladies neurodégénératives. Dans les deux comparaisons, les signatures moléculaires présentes au stade précoce de la DMLA, de la SEP et de la MA ne sont pas détectées dans les microglies et les astrocytes au stade avancé de la neurodégénérescence (Fig. 9A, B supplémentaires), ce qui indique des modifications transcriptionnelles dans la glie au cours de la progression de la maladie.

Pour examiner les changements transcriptionnels dans les cellules gliales au cours de la progression d'une pathologie de DMLA néovasculaire sèche à un stade avancé, nous avons effectué un snRNAseq sur trois rétines supplémentaires provenant de rétines de donneurs humains atteintes de DMLA néovasculaire et appliqué l'analyse CATCH à 46 783 noyaux lorsqu'elle était combinée avec les échantillons séquencés précédemment. Nous avons identifié une granularité de la hiérarchie CATCH avec une faible activité topologique et attribué des étiquettes de type cellulaire basées sur l'expression de signatures génétiques spécifiques au type cellulaire (Fig. 5A , B). Suite à l'analyse CATCH à grain fin, nous avons identifié deux groupes de microglies : un groupe enrichi en cellules provenant de rétines témoins et un groupe enrichi en cellules provenant de rétines de DMLA néovasculaire à un stade avancé (Fig. 5C) . Pour identifier les changements transcriptionnels spécifiques à un type de cellule dans la sous-population de microglies enrichies en pathologie de DMLA néovasculaire à un stade avancé, nous avons effectué une analyse d'expression différentielle basée sur la condensation entre les clusters enrichis en contrôle et enrichis en DMLA néovasculaire. L'analyse des gènes les plus différentiellement exprimés entre ces sous-populations (valeur p corrigée par le FDR <0,1) a révélé une signature liée à l'inflammasome, notamment IL1B, NOD2 et NFKB1. La protéine pro-IL-1β nécessite à la fois un clivage et une libération via l'activation de la caspase médiée par l'inflammasome et la pyroptose pour la bioactivité . Ici, l'activation des capteurs de l'inflammasome et l'oligomérisation en complexes protéolytiquement actifs peuvent se produire en réponse à une baisse significative et durable de la tension en oxygène ou à une exposition chronique aux lipides42,43, toutes deux connues pour piloter l'activation de l'inflammasome via NLRP3 (NOD-, LRR- et pyrine contenant le domaine 3) (Fig. 5D). Aux stades avancés de la MA et de la SEP, d'autres voies cellulaires associées au stress ont été régulées, notamment les régulateurs transcriptionnels de la réponse au stress du RE (XBP1) et leurs gènes cibles impliqués dans le repliement et le transport des protéines (HSPA1A, HSPA1B, HSP90AA1) et la glycosylation (ST6GAL1 et ST6GALNAC3), ainsi que les régulateurs de l'autophagie. et protéostasie (ATG7, MARCH1, USP53). Ces signatures mettent en évidence une induction de stress cellulaire partagée.

À l'aide du flux de travail CATCH à grain fin, nous avons identifié deux sous-populations d'astrocytes : un cluster enrichi en cellules provenant d'échantillons rétiniens témoins et un cluster enrichi en cellules provenant d'échantillons rétiniens de DMLA néovasculaire à un stade avancé (Fig. 5E) . Pour identifier les signatures de la DMLA présentes dans les astrocytes au stade avancé de la pathogenèse de la maladie, nous avons effectué une analyse d'expression différentielle basée sur la condensation entre les clusters enrichis en contrôle et les clusters enrichis en DMLA néovasculaire. Analyse des gènes les plus exprimés différentiellement (valeur p corrigée par FDR < 0,1) entre ces sous-populations a révélé une élévation de l'expression de VEGFA, NR2E1 et HIF1A (Fig. 5F), qui sont tous des régulateurs des réponses cellulaires à une faible tension d'oxygène44-46. Alors que le VEGFA est connu pour être un médiateur important de la croissance anormale des vaisseaux sanguins qui caractérise la DMLA néovasculaire à un stade avancé et qu'il est la cible des thérapies actuelles pour le traitement de la maladie33,47,48, nos données démontrent chez l'homme une sous-population spécifique d'astrocytes rétiniens . qui sont à l'origine de ce signal.

L'IL-1β dérivée des microglies entraîne une néovascularisation pathologique via les astrocytes.

Comme on sait que les microglies influencent les états fonctionnels des astrocytes par la sécrétion de facteurs solubles, nous avons voulu déterminer si les cytokines dérivées des microglies pouvaient piloter l'expression de VEGFA à partir des astrocytes rétiniens . Étant donné que CATCH était capable d'isoler les états astrocytes et microgliaux, nous avons utilisé l'analyse d'interaction CellPhoneDB52 pour créer une liste putative de cytokines possibles dérivées des microglies qui peuvent interagir avec les astrocytes pour piloter l'expression de VEGFA (Fig. 6A). À partir de cette analyse, le groupe de microglies enrichies en néovasculaires a interagi de manière plus significative avec les astrocytes via l'IL-1β et l'IL-6, tandis que chez les témoins, l'interaction microglie-astrocyte était principalement médiée par l'IL-4. De plus, l'IL-1β a interagi de manière plus significative avec la sous-population d'astrocytes enrichies en néovasculaires. En utilisant l'estimation de l'information mutuelle rééchantillonnée à densité conditionnelle (DREMI), une méthode permettant d'identifier des associations non linéaires dans les données53, nous constatons que la signalisation de l'IL-1β sur les astrocytes était associée de la manière la plus significative à la production de VEGFA par les astrocytes. Pendant ce temps, la signalisation de l'IL-4 était associée de manière plus significative à une diminution de la production de VEGFA par les astrocytes (Fig. 6B). Nous avons ensuite entrepris de valider de manière impartiale les régulateurs de cytokines de la production de VEGFA par les astrocytes.

Les cytokines font partie d'un réseau complexe de protéines pouvant produire des effets additifs, synergiques ou antagonistes. Pour démontrer cette relation, nous avons utilisé deux méthodes de sélection. Nous avons d'abord utilisé une approche de criblage combinatoire utilisant toutes les cytokines identifiées dans notre ensemble de données snRNAseq, en supprimant une à la fois pour tester sa nécessité dans la création d'un astrocytes exprimant VEGFA. Le dépistage avec des astrocytes humains dérivés d'iPSC a démontré que l'IL-1β, l'IL-10 et l'IL-17 sont des régulateurs positifs de la production de VEGFA dans ces cellules, car leur soustraction provoque une diminution des VEGFA par rapport aux astrocytes humains dérivés d'iPSC stimulés par toutes les cytokines (Fig. 6C). Nous avons ensuite testé la capacité de certaines de ces cytokines à réguler la production de VEGFA en effectuant une stimulation protéique unique et avons noté que seule l'IL-1β provoquait la sécrétion de VEGFA par les astrocytes (Fig. 6D) . Dans les deux analyses, l'IL-1β a régulé positivement l'induction de VEGFA à partir des astrocytes à la fois in vitro (Fig. 6C, D) et in silico (Fig. 6B). Notre analyse de la régulation du VEGFA a validé la prédiction informatique selon laquelle l'IL-4 serait un régulateur négatif de la production de VEGF-A (Fig. 6B, C), montrant l'utilité de notre approche pour identifier les interactions de signalisation entre les sous-ensembles cellulaires identifiés avec CATCH.

Avec l'identification de médiateurs cytokines de la production de VEGFA astrocytes, nous avons validé nos résultats in vivo en injectant de l'IL-1β par voie intravitréenne à une souris. Cela a entraîné une régulation positive du VEGFA (Fig. 6E, F). Non seulement il y a eu une augmentation de la quantité de VEGFA (Fig. 6G, à droite), mais il y a également eu une augmentation des signaux superposés de GFAP et de VEGFA, indiquant l'activation et la sécrétion de VEGFA des astrocytes (Fig. 6G, à gauche), ainsi que Expression de VEGFA s'étendant de la localisation de la couche de cellules ganglionnaires jusqu'aux autres couches de la rétine. Une tendance similaire a également été observée dans l'épithélium pigmentaire rétinien (RPE) adjacent, mais n'a pas atteint une signification statistique (Fig. 8 supplémentaire), probablement en raison de la variation de l'autofluorescence intrinsèque parmi les cellules RPE. Au total, cela a démontré la suffisance des cytokines telles que l'IL-1β pour induire la sécrétion de VEGFA dans les astrocytes in vitro et in vivo. Les cytokines telles que l'IL-1β sont augmentées dans le corps vitré des patients atteints de DMLA néovasculaire54, mais la source et le rôle de ces cytokines dans l'angiogenèse n'ont pas été explorés. Nous avons entrepris une coloration immunohistochimique pour l'IL-1β dans des échantillons rétiniens provenant de la macula de patients atteints de DMLA et de témoins sains, observant une augmentation de l'intensité de l'IL-1β dans les couches internes de la rétine, où résident les astrocytes (Fig. 6I) . . De plus, une régulation positive du VEGFA a été observée dans ces zones (Fig. 6G), indiquant que le phénomène que nous observons in vitro et chez la souris se produit probablement également dans la DMLA néovasculaire humaine (Fig. 6G – I).

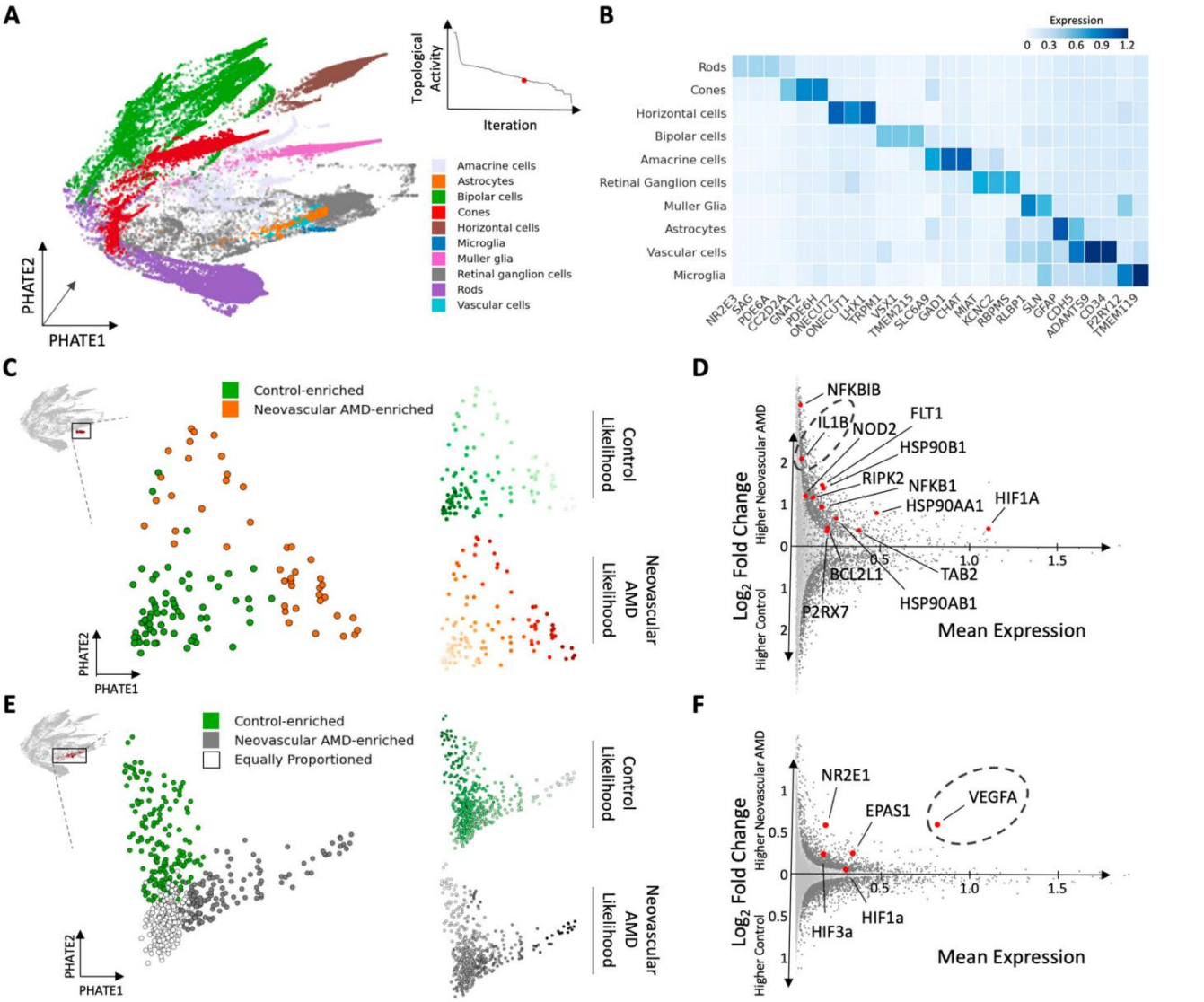

Figure 5 | Modifications spécifiques à chaque type de cellule dans l'expression des gènes au cours de la progression de la DMLA. Une visualisation PHATE de 46 783 noyaux isolés de la DMLA néovasculaire et des rétines témoins65. L'analyse CATCH a identifié une résolution de l'homologie de condensation, qui a isolé les types de cellules. Comme dans la figure 3, chaque groupe cellulaire s'est vu attribuer une identité de type cellulaire en fonction de la signature génétique qu'il exprimait au niveau le plus élevé. Types de cellules identifiés par B CATCH, comme le montre l'expression normalisée moyenne de gènes marqueurs spécifiques au type de cellule connu. C L'enrichissement de l'état de la maladie a été calculé à l'aide de MELD (à droite) pour chaque condition : contrôle (en haut) et DMLA néovasculaire (en bas), avec des probabilités MELD plus élevées affichées avec des couleurs plus foncées. Une résolution de l'homologie de condensation, qui isole de manière optimale les scores de vraisemblance MELD de chaque condition, a été identifiée à l'aide d'une analyse d'activité topologique. Les microglies sont revisualisées à l'aide de PHATE. Deux sous-ensembles de cellules microgliales, un enrichi en microglies provenant de rétines atteintes de DMLA néovasculaire et un autre provenant de rétines témoins. D L'analyse de l'expression différentielle entre les amas microgliaux enrichis en contrôle et enrichis en maladie néovasculaire a révélé un modèle d'activation différent en fin de maladie. Gènes significativement exprimés de manière différentielle, visualisés dans l'obscurité gris (test EMD bilatéral avec valeur p corrigée FDR <0,1 comme décrit dans les méthodes). Cette signature inclut NFKBIB, IL1B, NOD2, FLT1, HSP90B1, RIPK2, NFKB1, HSP90AA1, HIF1A, BCL2L1, P2RX7, TAB2, HSP90AB1. E L'enrichissement de l'état de la maladie a été calculé à l'aide de MELD (à droite) pour chaque condition : contrôle (en haut) et DMLA néovasculaire (en bas), avec des probabilités MELD plus élevées affichées avec des couleurs plus foncées. Une résolution de l'homologie de condensation, qui isole de manière optimale les scores de vraisemblance MELD de chaque condition, a été identifiée à l'aide d'une analyse d'activité topologique. Les astrocytes sont revisualisés à l'aide de PHATE. CATCH a identifié trois sous-ensembles de cellules astrocytes, un enrichi en astrocytes provenant de rétines néovasculaires, un autre provenant de rétines témoins et un troisième également réparti entre les conditions. F L'analyse de l'expression différentielle entre les groupes d'astrocytes enrichis en contrôle et enrichis en maladies néovasculaires révèle un modèle d'activation différent dans la maladie néovasculaire à un stade avancé. Gènes exprimés de manière différentielle significative, visualisés en gris foncé (test EMD bilatéral avec valeur p corrigée FDR <0,1 comme décrit dans la section Méthodes). Cette signature comprend NR2E1, EPAS1, VEGFA, HIF1A, HIF3A.

Discussion

Ici, nous avons utilisé snRNA-seq pour générer un atlas transcriptomique unicellulaire de la DMLA au cours de la progression pathologique, ainsi que pour développer un pipeline d'apprentissage automatique qui permet des comparaisons significatives entre les types et les états de cellules à travers les maladies et les phases. Afin de générer de riches signatures pour la comparaison de maladies croisées parmi des sous-populations cellulaires rares, nous avons développé une suite d'outils d'apprentissage automatique inspirés de la topologie pour l'analyse d'une seule cellule, « CATCH », un outil qui identifie les spécificités des sous-populations enrichies dans une condition spécifique en calculant la hiérarchie complète des états cellulaires à l'aide de la « condensation de diffusion ». Ce pipeline a identifié des états cellulaires enrichis en maladies, caractérisé les signatures d'expression pathogène et prédit les interactions cellulaires entre les populations pathogènes, découvrant ainsi des cibles thérapeutiques potentielles.

Grâce à CATCH, nous avons identifié et caractérisé des sous-populations spécifiques de microglies et d'astrocytes enrichies au stade précoce de

https://doi.org/10.1038/s41467-023-37025-7

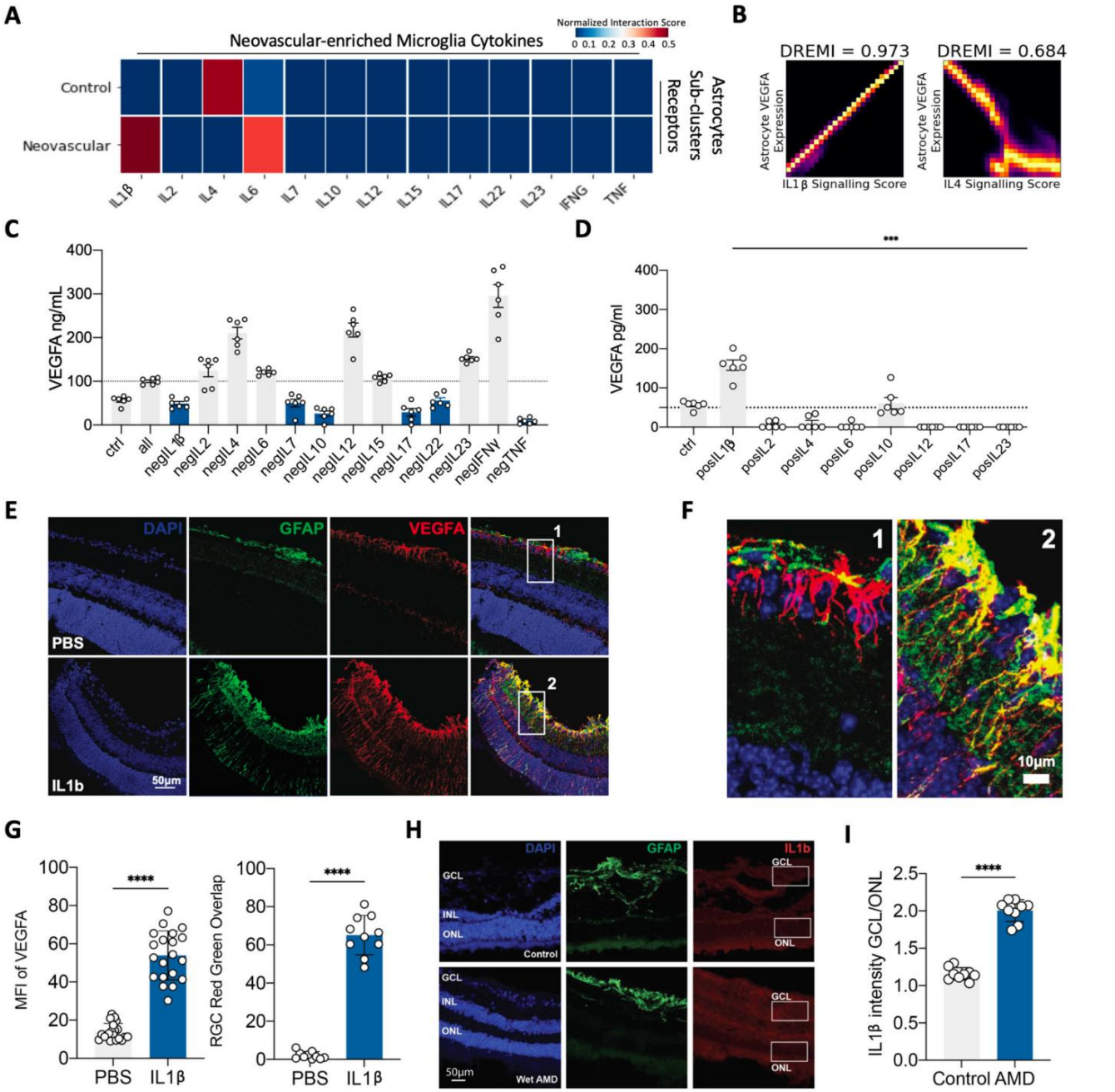

Figure 6 | Identification des régulateurs de cytokines de la sécrétion de VEGFA des astrocytes. Une analyse d'interaction entre la condensation par diffusion a identifié des sous-types de astrocytes et microglies enrichies en néovasculaires (détaillées sur la Fig. 5) calculées avec Téléphone portableDB52. Les interactions entre les cytokines produites à partir de microglies enrichies en néovasculaires ont été calculées par rapport aux récepteurs de cytokines des sous-types d'astrocytes. Des interactions entre des paires spécifiques de cytokines et de récepteurs ont été ajoutées pour produire une valeur d'interaction de cytokine unique pour le contrôle et les astrocytes néovasculaires. sous-types. Analyse d'association B DREMI entre l'expression des astrocytes VEGFA et l'IL-1β score de signalisation et score de signalisation IL-4. Les scores de signalisation pour l'IL-1β et l'IL-4 étaient calcul en ajoutant l'expression du récepteur de l'IL-1β et de l'IL-4, respectivement, astrocytes enrichis en néovasculaires de la Fig. 5. C Dépistage négatif réalisé dans astrocytes humains dérivés d'iPSC 24 h après la stimulation, en soustrayant une cytokine (par exemple, 'negIL2') du pool combinatoire pour tester sa nécessité pour générer un Astrocyte producteur de VEGFA par rapport au contrôle du véhicule (ctrl). Tout représente une stimulation avec un mélange de cytokines (IL-1β, IL-2, IL-4, IL-6, IL-7, IL-10, IL-12, IL-15, IL-17, IL-22 , IL-23, IFNγ, TNF). La protéine VEGFA est mesurée à l'aide d'enzymes test immunosorbant (ELISA). Les données ont été évaluées en utilisant une ANOVA unidirectionnelle avec correction de comparaisons multiples à l'aide de Dunnetts. D Conduite cytokine unique dépistage positif dans les astrocytes humains dérivés d'iPSC pour tester la suffisance de chacun cytokine pour stimuler la production de VEGFA des astrocytes. Les niveaux de protéine VEGFA sont mesurés par ELISA 24 h après la stimulation avec chaque cytokine par rapport au véhicule. contrôle (ctrl). Les données ont été évaluées à l'aide d'une ANOVA unidirectionnelle avec correction de comparaisons multiples à l'aide de Dunnetts. E IL-1β ou PBS a été injecté par voie intravitréenne dans un oeil de souris. Les rétines ont été collectées 72 heures plus tard pour une imagerie immunofluorescente. GCL : couche de cellules ganglionnaires ; IPL : couche plexiforme interne ; INL : couche nucléaire interne ; OPL : extérieur couche plexiforme ; ONL : couche nucléaire externe. Solution saline tamponnée au phosphate PBS (témoin). Les expériences ont été répétées au moins trois fois indépendamment avec des résultats similaires. F Zoom sur les images des régions indiquées en E. G Quantification de l'intensité moyenne de fluorescence (MFI) du VEGFA après injection d'IL-1β ou de PBS dans les yeux de souris après 72 h (à gauche) et quantification de la quantité de chevauchement de VEGFA et de GFAP dans le ganglion couche cellulaire de la rétine de souris après injection d'IL-1β ou de PBS (à droite). Le centre du les barres d'erreur sont la moyenne. Un test t de Student bilatéral a été réalisé. **** représente $p < 0,0005$. H Imagerie par immunofluorescence du contrôle post-mortem humain et rétines néovasculaires de DMLA. Les expériences ont été répétées au moins trois fois indépendamment fois avec des résultats similaires. I Quantification de l'intensité de l'IL-1β dans la couche de cellules ganglionnaires (GCL) sur la couche nucléaire externe (ONL) de la rétine de F. Les données sont présentées sous forme valeurs moyennes ± SEM ; ****$p < 0,0005$; Test t de Student bilatéral non apparié.

DMLA sèche présentant des signatures d'activation liées à la phagocytose, au métabolisme lipidique et à la fonction lysosomale. Nous avons trouvé des populations similaires de microglies et d'astrocytes dans les analyses de données unicellulaires sur la MA et la SEP précédemment publiées. Bien que les événements déclencheurs initiaux diffèrent probablement selon les affections neurodégénératives, les plaques extracellulaires riches en lipides jouent un rôle de premier plan dans chaque affection. Il est probable que les cellules gliales coordonnent l'élimination des débris extracellulaires et, à leur tour, s'activent. Bien que la clairance phagocytaire initiale puisse être bénéfique, il a été démontré que l'activation gliale joue un rôle dans la dégénérescence dans la DMLA, la MA et la SEP. Aux stades ultérieurs de la maladie, ce paysage d'activation partagée évolue. Dans la DMLA néovasculaire avancée, notre analyse a identifié une signature liée à l'inflammasome des microglies qui détermine la polarisation pro-angiogénique des astrocytes et la néovascularisation pathologique. L'activation de l'inflammasome microglial et la libération ultérieure d'IL-1β pourraient être médiées par divers capteurs de signalisation. Le capteur NLRP3 peut être activé en réponse à divers signaux de stress, notamment une exposition prolongée aux lipides ou une hypoxie prolongée, et a déjà été impliqué en tant que moteur microglial de l'immunopathologie neurodégénérative, ce qui en fait un candidat probable55. Les microglies sont des cellules très mobiles et sensibles à une grande variété de stimuli. Bien que le traçage de la lignée qui différencie définitivement l'origine des phagocytes mononucléaires en macrophages circulants, en macrophages résidant dans les tissus et en microglies reste difficile, on pense que les phagocytes mononucléés trouvés sur le côté apical de l'EPR à proximité des drusen, qui induisent l'activation de l'inflammasome, proviennent des trois populations56. En outre, de nouvelles données suggèrent que l'inflammasome et l'IL-1β jouent un rôle essentiel dans la promotion de la dégénérescence dans la SEP et l'AD10-12. Le traitement par l'IL-1β des cellules RPE in vitro entraîne une régulation positive de l'expression de VEGFA57. Ainsi, nos résultats impliquent un cadre dans lequel

Cet ensemble d'analyses a des implications claires pour les thérapies potentielles contre la DMLA et d'autres maladies neurodégénératives. Actuellement, le traitement anti-VEGF est la principale intervention approuvée pour traiter la DMLA et n'est efficace qu'au stade le plus avancé de la maladie. Notre analyse impartiale a non seulement identifié la spécificité du type cellulaire de l'expression du VEGFA, mais a également identifié les interactions de signalisation pathogènes qui favorisent la progression de la DMLA. Étant donné que le VEGFA est une glycoprotéine librement diffusable, sa production à partir des astrocytes rétiniens peut induire une angiogenèse à partir de la choroïde. Actuellement, des thérapies qui inhibent l'IL-1β sont disponibles et utilisées en pratique clinique pour le traitement d'autres maladies. L'inhibition de l'IL-1β dérivée des microglies dans la DMLA néovasculaire pourrait apporter un bénéfice thérapeutique, en empêchant une néovascularisation ultérieure chez les patients avancés, ou même en empêchant la néovascularisation avant qu'elle ne commence chez les patients présentant des stades précoces de la maladie. Étant donné que ces mécanismes sont communs à la SEP et à la MA, il est plausible que ces interventions puissent également bénéficier aux patients souffrant d'autres maladies neurodégénératives. L'identification de candidats thérapeutiques prometteurs à tester dans le cadre d'essais cliniques sur les maladies neurodégénératives reste importante, et nos données suggèrent que les approches ciblant les cellules gliales pourraient être largement applicables à plusieurs maladies neurodégénératives.

Méthodes

Déclaration éthique

Cette étude, acquisition et utilisation d'échantillons de rétine humaine post-mortem ont été approuvées par le comité d'examen institutionnel du Yale Human Research Protection Program (numéro de protocole Yale 2000028616). Nous avons respecté toutes les règles éthiques pertinentes pour le travail avec des participants humains. Tous les échantillons de tissus humains ont été obtenus avec le consentement éclairé avant le prélèvement de tissus auprès des participants s'ils étaient inscrits ante mortem ou des tuteurs légaux s'ils étaient post mortem. Les protocoles expérimentaux sur les souris ont été approuvés par le comité institutionnel de protection et d'utilisation des animaux de l'université de Yale (numéro de protocole Yale 2022-20275). Toutes les expériences ont été réalisées conformément aux directives décrites par le comité institutionnel de protection et d'utilisation des animaux de l'université de Yale.

Détails de l'analyse CATCH

Le cadre CATCH constitue un groupe d'outils d'apprentissage automatique d'inspiration topologique pour identifier, caractériser et comparer des populations de cellules enrichies en conditions à travers la hiérarchie cellulaire. Ce cadre est centré sur le processus de condensation par diffusion, qui apprend la structure des données à travers les granularités. Au-delà d'adaptations significatives à la condensation de diffusion, nous avons introduit des outils pour aider à analyser la riche quantité d'informations multigranulaires produites par la condensation de diffusion : visualisation de la hiérarchie cellulaire, analyse de l'activité topologique, caractérisation automatisée des clusters et analyse de l'expression différentielle.

Dans les sections suivantes, nous fournissons une description détaillée de chaque aspect de CATCH. Cela comprend des descriptions détaillées du processus de condensation par diffusion ainsi que sa relation avec MELD, la distance de terrassement de Wasserstein (EMD) et l'analyse de l'activité topologique. Nous complétons cette section avec un ensemble rigoureux de comparaisons pour comparer notre méthode.

Contexte des filtres d'apprentissage et de diffusion multiples. Bon nombre des concepts fondamentaux de la condensation par diffusion et de ses adaptations présentés ici sont basés sur les progrès de la théorie des variétés et des filtres graphiques. Typiquement, les données à n dimensions X = {x1, …, xN} peuvent être modélisées comme provenant d'une variété à d dimensions Md collectée via une fonction non linéaire xi = f(zi). En effet, les stratégies de collecte de données (telles que le séquençage d'ARN unicellulaire) créent des observations de grande dimension même lorsque la dimensionnalité intrinsèque est relativement faible. Les algorithmes qui utilisent cette hypothèse de variété58-61 exploitent la géométrie intrinsèque de faible dimension de la variété pour explorer les relations dans les d

Les auteurs de diffusion59 ont présenté un cadre qui capture la géométrie multiple intrinsèque à l'aide de marches aléatoires qui agrègent les relations locales entre les points de données pour révéler des géométries non linéaires. Ces relations locales, appelées affinités, sont construites à l'aide d'une fonction noyau gaussienne :

$$K(x_i, x_j) = \exp\left(\frac{k\,\|x_i - x_j\|_2}{\varepsilon}\right) \quad (1)$$

où K est une matrice de Gram N × N et un paramètre de bande passante ε, qui contrôle la localité. Un opérateur de diffusion est défini comme la normalisation des lignes de la matrice N × N Gram K :

$$P = D^{-1}K \quad (2)$$

où D(xi, xi) = ∑jK(xi, xj). La matrice d'opérateur de diffusion P représente les probabilités de transition en une seule étape pour une marche aléatoire ou un processus de diffusion markovien. De plus, comme le montre la figure 59, les puissances de cet opérateur de diffusion P (représentées par Pt où t > 0) représentent une marche aléatoire en t étapes.

Des travaux récents sur la diffusion de données27,62–64 ont montré que ce cadre proposé par59 peut être utilisé comme filtre passe-bas lorsque l'opérateur P est directement appliqué aux caractéristiques des données, déplaçant efficacement les points de données à proximité de leurs voisins de diffusion sur la variété. . Ce processus de filtrage passe-bas élimine efficacement les variations de haute fréquence, ou bruit, et conserve uniquement la géométrie principale de faible dimension du collecteur de données.

Aperçu de la condensation par diffusion et de ses limites. La condensation de diffusion est un processus dynamique qui s'appuie sur des concepts précédemment établis en matière de filtres de diffusion, de géométrie de diffusion et d'analyse de données topologiques. L'algorithme déplace lentement et de manière itérative les points ensemble de manière à révéler la topologie de la géométrie sous-jacente. L'approche de condensation par diffusion implique deux étapes qui sont répétées de manière itérative jusqu'à ce que tous les points convergent :

1. Calculer un opérateur de diffusion de Markov inhomogène dans le temps à partir de les données;
2. Appliquez cet opérateur aux données comme filtre de diffusion passe-bas, déplacer les points vers les centres de gravité locaux.

Comme établi dans des travaux antérieurs8,17,27, l'application du l'opérateur P sur un vecteur v fait la moyenne des valeurs de v sur de petits quartiers dans les données. Lorsqu'il est appliqué directement à une coordonnée fonction, cette application condense les points vers les centres locaux de la gravité tel que déterminé par le paramètre de bande passante ε, créant un ensemble filtré de coordonnées. Dans ce processus, si X(0) = X est l'ensemble de données d'origine avec l'opérateur de diffusion P0 = P, alors Xð1Þ = X = Xð0Þ*P0. Alors que les applications précédentes des filtres de diffusion appliquez simplement une itération de ce processus de filtrage par diffusion à données, nous pouvons répéter ce processus pour réduire davantage la variabilité dans les données en calculant la matrice de Markov P1 en utilisant le X(1) filtré par coordonnées. Une nouvelle représentation de coordonnées filtrée X(2) est obtenu en appliquant P1 aux fonctions de coordonnées de X(1). Les premières applications de l'opérateur de diffusion P à X atténuent les variations haute fréquence de la fonction de coordonnées, rapprochant efficacement les points similaires les uns des autres. Les applications ultérieures atténuent les variations de basse fréquence, déplaçant des groupes de points similaires les uns envers les autres. Une explication plus complète de la diffusion la condensation et ses propriétés mathématiques peuvent être trouvées dans réf. 8 et réf. 17.

Dans sa forme originale, le processus de condensation par diffusion ne peut être appliqué aux données scRNAseq. Bien qu'utile pour les tâches générales d'analyse de données, ce processus a des limites :

1. l'approche ne fonctionne pas dans l'espace non linéaire de la variété transcriptomique unicellulaire ;
2. ne s'adapte même pas à des milliers de points de données ;
3. n'identifie pas les granularités de la topologie, ce qui signifie- partitionner complètement l'espace d'état cellulaire et
4. n'identifie pas les populations pathogènes impliquées dans la maladie processus.

Dans ce travail, nous abordons chacune de ces limites et approfondissons étendre le cadre pour effectuer efficacement des analyses monocellulaires clés tâches telles que la caractérisation des clusters et l'expression différentielle analyse.

Pour répondre à ces préoccupations, nous avons apporté les adaptations importantes suivantes pour une application aux données unicellulaires :

1. Apprenez dynamiquement la géométrie du collecteur unicellulaire avec chaque filtre de diffusion utilisant des marches aléatoires en t optimisées avec entropie spectrale ;
2. Visualisez la hiérarchie apprise en intégrant l'arbre de condensation ;
3. Utiliser l'activité topologique pour identifier des granularités significatives pour analyse en aval ;
4. Mettre en œuvre le repérage des opérateurs de diffusion, les parcours aléatoires pondérés et la fusion des données pour atteindre efficacement des milliers de personnes de cellules ;
5. Implémenter la condensation par diffusion avec un noyau de désintégration alpha pour caractérisation automatisée des clusters et calcul efficace de gènes d'expression différentielle.

La condensation de diffusion intrinsèque au collecteur apprend la hiérarchie cellulaire à partir de données transcriptomiques unicellulaires. Encadré 1

Algorithme 1. Condensation par diffusion intrinsèque au collecteur

Entrée : matrice de données X cellule par PC, paramètre de bande passante initiale du noyau. compteur ε0 et seuil de fusion ζ

Résultat : étiquettes de cluster par itération

1 : X0← X, je ← 0

2 : alors que nombre de points dans Xi > 1

3 : Fusionner les points de données a, b si Xi(a) − Xi(b) 2 < ζ, où Xi(a) est la énième rangée de Xi

4 : Mettre à jour l'affectation du cluster pour chaque point de données d'origine basé sur la fusion

Di← calcule la matrice de distance par paire à partir de Xi

5 : 6 : Affinité du noyau de désintégration alpha Ki← (Di, εi)

7: La ligne Pi← normalise Ki pour obtenir une matrice de transition de Markov (opérateur de diffusion)

ti← entropie spectrale de Pi

8 : 9 : Xi + 1 Pti XI

dix: εi+1← mise à jour(εi)

11 : je = je + 1

12 : fin pendant

Notre implémentation de l'algorithme de condensation par diffusion prend un matrice X cellule par composant principal (généralement les 50 premiers composants) et calcule un opérateur de diffusion P, représentant la probabilité distribution de la transition d'une cellule à une autre en une seule étape en utilisant une fonction de noyau de désintégration α avec une bande passante fixe ε (Alg. 1 : étapes 5-7). Alors que d'autres techniques d'apprentissage multiple résument les données à un point où les caractéristiques intrinsèques de la variété dérivées ont une relation peu claire avec l'expression des gènes, notre approche apprend la variété tout en travaillant sur des composants principaux, qui ont une relation claire avec gènes. En utilisant les composants principaux comme substrat pour la condensation, nous pouvons facilement caractériser les clusters et effectuer des analyses différentielles. analyse de l'expression dans l'espace d'expression génique dans l'analyse en aval.

Une autre amélioration clé que nous apportons à l'algorithme de condensation consiste à élever P à la puissance t (plutôt que 1 comme dans 8), simulant une marche aléatoire en t étapes sur les données. Cette approche débruit de manière adaptative et affine ces probabilités de transition à travers les itérations de telle sorte que des transitions se produisent sur la variété unicellulaire non linéaire27,59,65 . Ce L'opérateur de diffusion en t-étapes Pt est appliqué aux données d'entrée, agissant comme un filtre de diffusion intrinsèque au collecteur, remplaçant efficacement les coordonnées d'un point avec la moyenne pondérée de ses voisins de diffusion en t-step. Nous suivons les valeurs de t calculées au fil des itérations et effectuons une étude d'ablation pour montrer la nécessité d'ajuster de manière adaptative t dans chaque itération de la condensation par diffusion intrinsèque du collecteur (Fig. 2A, B supplémentaires). Voir Alg. 1 pour le pseudocode de cet algorithme. Quand la distance entre deux cellules tombe en dessous d'un seuil de distance ζ, les cellules sont fusionnés, les désignant comme appartenant au même cluster à l'avenir (Alg. 1 : étapes 3,4). Il est important de noter que dans le œuvre originale8, n'a pas fusionné les points. Ce processus est ensuite répété itérativement jusqu'à ce que toutes les cellules se soient réduites à un seul cluster. Cette fusion étape, mise en œuvre dans notre condensation par diffusion intrinsèque au collecteur approche, permet le calcul rapide de la hiérarchie cellulaire lors du gros grain. Lors de l'application de cette diffusion intrinsèque multiple processus de condensation en données transcriptomiques unicellulaires, nous pouvons voir les cellules se condensent pour regrouper les centroïdes au fil des itérations, de manière efficace et apprendre rigoureusement la hiérarchie des cellules uniques (Fig. 1C). Enfin, grâce à des astuces de mise en œuvre évolutives, telles que l'opérateur de diffusion le repérage66 et les marches aléatoires pondérées, nous avons permis la diffusion condensation pour s'adapter à des milliers de cellules individuelles (Supplémentaire Figure 2F). Détails supplémentaires sur la sélection de t ainsi que évolutif des astuces de mise en œuvre peuvent être trouvées ci-dessous.

Apprentissage dynamique de la géométrie des variétés avec entropie spectrale et des filtres de diffusion à étapes en T. Alors que la mise en œuvre initiale de la condensation par diffusion a été créée pour comprendre la structure multigranulaire de données linéaires, les cellules individuelles occupent un espace hautement non linéaire nécessitant stratégies d'apprentissage multiple27,59,65. Dans les données unicellulaires, le bruit technique, tels que les abandons et les variations, créent des artefacts de mesure. Quand en construisant des probabilités de diffusion sur ce type de données bruyantes, des probabilités de transition élevées peuvent être calculées de manière inappropriée entre des cellules non liées. Ainsi, travailler directement avec P ne parvient pas à reconnaître les non-linéarités et les artefacts techniques présents dans les données unicellulaires.

Des travaux antérieurs sur la diffusion de données ont montré qu'élever l'opérateur de diffusion P à la puissance t affine ces probabilités de transition, augmentant ainsi les chances de transition vers des cellules plus apparentées . Cette étape d'alimentation permet d'apprendre la géométrie non linéaire pertinente de la variété de données, ce qui nous permet d'ignorer les voisins parasites trouvés dans l'espace de mesure ambiant des cellules et de trouver à la place des voisins de diffusion situés sur la variété unicellulaire.

Comme les ensembles de données unicellulaires peuvent souvent souffrir de différents types et échelles de bruit, les approches précédentes ont montré que le nombre correct de pas à effectuer doit être calculé de manière adaptative en fonction des données27,67 . Cependant, les stratégies proposées précédemment pour sélectionner t sont souvent lentes, car elles nécessitent une approche par essais et erreurs, qui repose sur la structure de l'ensemble de données sous-jacent. Cependant, dans la condensation de diffusion, la structure de l'ensemble de données sous-jacent se déplace continuellement entre les granularités en raison de l'application répétée de filtres de diffusion, ce qui rend le calcul répété de t nécessaire et, grâce à ces techniques, compliqué en termes de calcul. Par conséquent, nous proposons de sélectionner t de manière adaptative à chaque itération de condensation en utilisant une approche basée sur l'entropie spectrale. Précédemment, il a été montré que l'alimentation de l'opérateur de diffusion P affecte différentiellement les vecteurs propres de la matrice alimentée. Alors que les vecteurs propres bruyants à haute fréquence se réduisent rapidement à zéro, les vecteurs propres à basse fréquence, plus informatifs, diminuent beaucoup moins rapidement27. Nous pensons qu'il existe une valeur de t, qui réduit de manière optimale les informations bruitées provenant des vecteurs propres haute fréquence tout en conservant le maximum d'informations provenant des vecteurs propres informatifs basse fréquence. Pour identifier ce point, nous calculons l'entropie spectrale des probabilités de diffusion P lorsqu'elles sont alimentées à différents niveaux de t.

L'entropie spectrale est définie comme l'entropie de Shannon des valeurs propres normalisées, c'est-à-dire

$$SðP,tÞ = X \psi t \log ðψt\,iÞ, \qquad ð3Þ$$

Comme il existe un certain degré de perte d'informations à chaque valeur croissante de t, nous essayons d'identifier le point auquel cette courbe de perte d'informations se stabilise. Alors que l'alimentation à de faibles valeurs de t diminue rapidement l'entropie spectrale à mesure qu'une grande quantité de bruit diminue, l'alimentation à des valeurs plus élevées de t ne réduit que lentement l'entropie en raison de la suppression plus lente des informations des vecteurs propres informatifs à basse fréquence. En prenant le point auquel cette stabilisation se produit comme indiqué dans la réf. 65, nous permet de manière optimale de sélectionner de manière adaptative une valeur de t à chaque itération de condensation de diffusion, nous permettant ainsi de produire un filtre de diffusion qui a appris la variété unicellulaire.

En fait, la dérivation adaptative basée sur les données est essentielle à l'apprentissage de la structure de cluster multigranulaire des données. Afin d'illustrer ce point, nous avons généré des données synthétiques unicellulaires à l'aide de Splatter19.
Comme on peut le voir, pour différentes quantités de bruit variationnel et de suppression, la sélection optimale de t via l'entropie spectrale produit un meilleur ensemble d'étiquettes de cluster que lors de la définition de tin d'une manière fixe et déterminée par l'utilisateur (Fig. 3B supplémentaire). En fait, nous pouvons voir que le réglage de t sur 1 n'apprend pas la variété de données ou la structure de cluster de données unicellulaires, même assez silencieuses, révélant la nécessité de sélectionner un niveau élevé de t de manière adaptable et basée sur les données. Enfin, nous voyons qu'au fil des étapes de condensation successives, la complexité des données diminue et nécessite donc des niveaux de t inférieurs pour l'apprentissage (Fig. 3A supplémentaire).

Améliorer l'évolutivité avec des marches aléatoires pondérées, des opérateurs de diffusion repérés et des points de données fusionnés. Le calcul répété d'un opérateur de diffusion à partir de données unicellulaires de haute dimension, permettant à cet opérateur de diffusion d'identifier la valeur optimale de t suivie par l'application d'un filtre de diffusion via une multiplication matricielle, est coûteux en calcul. Répéter ces calculs, potentiellement des centaines de fois, comme le fait la condensation par diffusion, est fastidieux. En fait, cette approche, dans sa mise en œuvre la plus élémentaire, s'adapte très mal aux données unicellulaires de grande dimension comportant des dizaines de milliers de caractéristiques et potentiellement des centaines de milliers de cellules. Pour améliorer l'efficacité du calcul, nous effectuons les étapes suivantes :

1. Fusionner les points qui tombent en dessous d'un seuil de distance prédéfini ζ pour créer un cluster et pondérer les marches aléatoires pour maintenir l'effet de la densité des données ;
2. Calculez l'opérateur de diffusion compressé via le marquage66 pour calculer efficacement l'entropie spectrale, comme indiqué dans la réf. 65.

Collectivement, ces avancées améliorent considérablement la vitesse de calcul de la condensation par diffusion (Fig. 2F supplémentaire). En pratique, une hiérarchie cellulaire complète d'un ensemble de données de 13 000 cellules peut être analysée en 6 minutes dans un ordinateur portable Google Colaboratory (un service qui fournit gratuitement un processeur à 4 cœurs de 2 GHz et 20 Go de RAM).

Visualisation et analyse d'un arbre de condensation avec analyse d'activité topologique pour identifier des granularités significatives pour l'analyse en aval. L'analyse des données topologiques (TDA) est un cadre puissant qui apprend et analyse les données à travers des granularités. Dans TDA, on identifie les points de données associés en identifiant toutes les paires dont la distance tombe en dessous d'un seuil de distance δ dans une matrice de distance D. Toute paire de points qui tombe en dessous de ce seuil est considérée comme faisant partie du même composant ou cluster connecté. À mesure que δ augmente, davantage de paires de cellules seront connectées, créant rapidement moins de composants connectés, ou moins de clusters plus grands, à des granularités plus grossières. Dans l'analyse des données topologiques, l'homologie persistante est une approche de principe pour suivre les composants connectés qui sont créés et détruits à travers une gamme de granularités. Alors que la condensation par diffusion apprend la structure multigranulaire des données grâce à une cascade d'approches de filtration par diffusion non linéaire au lieu d'un seuil de distance croissant, ces approches sont intuitivement liées.

Nous pouvons étudier ce processus de condensation par diffusion soit de manière holistique, en évaluant toutes les granularités simultanément, soit de manière détaillée, en évaluant indépendamment les granularités significatives. À un niveau élevé, la hiérarchie cellulaire peut être étudiée en visualisant la hiérarchie cellulaire, contenant toutes les fusions dans toutes les granularités. Comme la condensation de diffusion intrinsèque au collecteur opère dans les dimensions PCA, nous implémentons pratiquement cette visualisation en empilant les deux premiers axes de Xi → Xi + 1 XI , créant ainsi un arbre hiérarchique qui résume la structure de cluster des données à travers les granularités (Fig. 1D- je).

Pour une analyse plus détaillée, nous pouvons découper cet arbre hiérarchique à des niveaux significatifs pour identifier les granularités des clusters qui divisent de manière optimale les cellules en clusters significatifs en fonction de la géométrie des données.
En utilisant l'homologie persistante, nous définissons une analyse d'activité topologique, une technique pour analyser la création et la destruction de clusters au cours d'itérations consécutives (Xi → Xi + 1) du processus de condensation par diffusion intrinsèque du collecteur. L'analyse de l'activité topologique est une variante de la statistique récapitulative de persistance totale souvent utilisée pour caractériser l'activité topologique dans l'analyse classique des données topologiques68. Dans ce cadre d'analyse, nous résumons la fusion de points au cours du processus de condensation et attribuons à chaque cluster une valeur de « proéminence » topologique connue sous le nom de persistance. Les composants hautement persistants sont considérés comme représentant des groupes de cellules dont le profil transcriptionnel est similaire et distincts des autres cellules. Ces clusters, et leurs valeurs de persistance associées, sont mieux représentés à l'aide d'un « code-barres de persistance ». Il s'agit d'une visualisation69 constituée de barres horizontales de différentes longueurs ; chaque barre correspond à une caractéristique topologique – un sous-groupe de cellules dans notre cas – tandis que la longueur de chaque barre représente la persistance de cette caractéristique, indiquant directement dans quelle mesure la caractéristique est importante. En supposant que le code-barres de persistance est constitué d'un ensemble de barres de coordonnées finales B : = fb1, ... ,bk g, nous calculons une courbe d'activité A : R → N définie par AðiÞ : = fb 2 B b ≤ ig , c'est-à-dire le nombre de caractéristiques topologiques (clusters de cellules) actives et indépendantes à une itération i donnée. Cette courbe d'activité, proposée pour la première fois par70 et mise en œuvre

by71, nous permet d'identifier des itérations de condensation rapide ainsi que des itérations d'inactivité relative à travers le gradient de A. Plus précisément, nous nous intéressons aux segments contigus dans la préimage de ∂A/∂i = 0, que nous appelons i- segments. La longueur d'un i-segment est le nombre d'itérations pour lesquelles il n'y a aucun changement dans l'activité topologique. Ainsi, le nombre d'itérations pour lesquelles ∂A/∂i = 0 fournit un moyen fondamental de sélectionner des granularités de condensation significatives calculées par le processus de condensation par diffusion. Inspirés par la nomenclature de l'homologie persistante, nous appelons la longueur d'un i-segment sans activité topologique sa persistance, ce qui signifie que nous recherchons le plus persistant de ces segments d'activité topologique.

Identification de populations enrichies en maladies en collaboration avec MELD. Bien que l'analyse de la hiérarchie cellulaire identifie les populations de cellules apparentées de manière impartiale et multigranulaire, elle n'utilise pas les informations sur les conditions d'origine pour identifier les populations cellulaires enrichies en maladies d'intérêt. Bien que nous puissions intégrer

Dans notre analyse, dans notre analyse, les cellules d'un certain état transcriptomique pathogène peuvent être surreprésentées dans une sous-variété d'un type cellulaire donné. En comparant directement les cellules d'un type particulier les unes aux autres en fonction de leur condition d'origine, nous diluons ces informations d'enrichissement et perdons un signal important. En fait, il a été démontré que l'identification de ces états pathogènes et leur comparaison directe avec des outils de regroupement et d'expression différentielle constituent une méthode plus puissante pour identifier les états cellulaires enrichis en conditions et les signatures d'expression9,72. Nous explorons ce point plus loin dans cette section.

Pour prendre en compte les informations spécifiques à une condition, nous utilisons MELD pour identifier les populations cellulaires enrichies ou épuisées au cours de différentes phases de la maladie. MELD est une méthode basée sur la géométrie multiple pour calculer un score de vraisemblance pour chaque cellule, indiquant si elle est plus susceptible d'être vue dans l'échantillon normal ou malade. Trouver une méthode de regroupement qui sépare ces groupes enrichis en conditions est un problème difficile qui doit être résolu pour identifier des populations cellulaires discrètes, qui peuvent être décrites de manière approfondie. Pour identifier rigoureusement les populations cellulaires présentant de forts signaux d'enrichissement spécifiques à la maladie, nous combinons ce score MELD au niveau cellulaire avec les informations de notre analyse d'activité topologique pour identifier les résolutions qui produisent des clusters stables. Ensuite, au sein de ce regroupement stable, nous identifions des populations enrichies en différentes pathologies.

Caractérisation automatisée des clusters via une diffusion multiple intrinsèque condensation. Même si l'identification des états cellulaires pathogènes est essentielle, les biologistes s'intéressent davantage à ce qui définit ces populations. La plupart des méthodes d'apprentissage multiple visualisent ou regroupent des populations d'intérêt, ce qui nécessite des calculs supplémentaires coûteux pour caractériser les populations cellulaires et découvrir des gènes différentiellement exprimés. Comme notre approche condense continuellement les profils transcriptomiques de cellules individuelles en centroïdes de cluster locaux dans un espace multiple, à toute itération, les états transcriptomiques des données condensées peuvent être extraits sans coût de calcul supplémentaire. Pour améliorer cette convergence vers les centroïdes, nous mettons en œuvre notre processus de condensation par diffusion avec un noyau de désintégration α (Fig. 2C supplémentaire). Ce noyau limite plus fortement la conversion des distances en affinités, ressemblant beaucoup au noyau boîte, qui calcule avec précision les centroïdes de cluster au cours des fusions de points principaux. Lorsque la condensation par diffusion fusionne deux cellules lors d'une itération particulière, le point nouvellement formé se trouve proche du centre de gravité des deux cellules d'origine dans l'espace transcriptomique.

Dans des conditions spécifiques, le nouveau point est exactement le centre de gravité du cluster tel que défini dans la proposition ci-dessous. Tout d'abord, nous définissons le noyau de désintégration α comme :

$$K\alpha(x_i, x_j) = \exp\left(\frac{k\, x_i x_j k\alpha}{\varepsilon\alpha}\right), \quad je, j = 1, \ldots, N : \qquad \text{ð4Þ}$$

La fonction du noyau gaussien standard, comme indiqué dans l'équation (1), a un α de 2. Le noyau de désintégration α par défaut utilise quant à lui une valeur beaucoup plus élevée (la valeur par défaut dans notre implémentation est 40), qui convertit les distances proches en affinités de manière beaucoup plus stricte (Supplémentaire Figure 2C). À mesure que α augmente vers l'infini, cette fonction noyau converge presque complètement vers le noyau boîte. Avec ce noyau, nous sommes prêts à énoncer un ensemble de conditions dans lesquelles la condensation par diffusion le processus peut être facilement caractérisé.

Proposition 1. Supposons qu'il existe une unique distance minimale globale non nulle δi entre les points xa, xb à chaque itération i, avec la paire de points suivante à une distance d'au moins δi + τi avec 0 < τi. Notez que xa, xb pourraient avoir une multiplicité supérieure à 1, représentant des clusters de taille > 1. Définissez ensuite la bande passante sur i : = δi + τi/2 à chaque itération du processus de condensation. Pour un α suffisamment grand, le processus de condensation par diffusion maintiendra deux invariants pour les N − 1 premières étapes :

1. Le nombre de points sera N − i ; 2. Les points uniques seront situés au centre de gravité de leur cluster.

Preuve. Il est facile de vérifier que (1) et (2) sont valables pour le pas zéro. Pour tout i < N et pour α suffisamment grand, Kα(xk, xj) devient arbitrairement proche de 1 pour (k, j) {(a, a), (a, b), (b, a), (b , b)} et 0 sinon. Exactement une fusion se produit à chaque pas de temps entre les points en xa et xb. Étant donné Pi comme décrit ci-dessus, ils fusionnent jusqu'au point xa + xb , c'est-à-dire le centre de gravité du cluster. Par récurrence (1) et (2) sont valables pour tout i < N. □

Dans ce contexte, le processus de condensation converge toujours en exactement N − 1 étapes. En pratique, nous visons des temps de convergence beaucoup plus courts car il y a beaucoup moins de N − 1 niveaux de clustering intéressants.
Pour 50 498 cellules, on retrouve un ensemble de paramètres permettant une convergence en 150 pas. Pour cette raison, nous utilisons une bande passante i plus grande, ce qui conduit à une convergence beaucoup plus rapide et donne des centres de cluster à chaque niveau qui sont proches, mais pas exactement, des centroïdes de cluster des points qu'ils représentent. Un autre facteur est le réglage du paramètre α. Étant donné que la condensation de diffusion intrinsèque au collecteur opère dans les dimensions PC, le profil complet d'expression génique du centroïde du cluster xab peut facilement être extrait en inversant les dimensions PC. Nous montrons que ce point est non seulement vrai mathématiquement mais aussi empiriquement vrai dans la pratique (Fig. 3C supplémentaire).

Analyse de l'expression différentielle via approximation de la distance gène Wasserstein. Au-delà de la caractérisation des clusters, l'analyse de l'expression différentielle est une méthode essentielle pour identifier les signatures des populations pathogènes. La distance de Earth Mover (EMD), également connue sous le nom de « transport optimal », qui se manifeste généralement par la distance 1D-Wasserstein, est une méthode populaire et établie pour extraire des gènes différentiellement exprimés entre des grappes . EMD, cependant, est coûteux en termes de calcul, car il calcule une cartographie optimale entre les points, s'exécutant en un temps O~ðn3Þ. Auparavant, des implémentations basées sur des arbres telles que FlowTree76 et QuadTree77 ont été capables de se rapprocher étroitement de la distance de Wasserstein de la vérité terrain tout en améliorant considérablement le temps d'exécution en limitant le transport des points à travers les branches d'un arbre hiérarchique78. Étant donné que la condensation de diffusion produit également une intégration arborescente des données, nous utilisons le transport basé sur les arbres pour l'expression différentielle.

EMD, ou distance de Wasserstein 1-D, est une mesure de la distance entre deux distributions. Pour une distance au sol donnée, la distance de Wasser-stein entre les distributions peut être considérée comme la distance totale minimale nécessaire pour déplacer une distribution vers l'autre. Soient μ, ν deux distributions sur un espace mesurable Ω de métrique d(,), et Π(μ, ν) l'ensemble des distributions conjointes π sur l'espace Ω × Ω, tel que pour tout sous-ensemble ω Ω , π(ω × Ω) = μ(ω) et π(Ω × ω) = ν(ω). La distance 1-Wasserstein Wd également connue sous le nom de distance du terrassement

(EMD) est défini comme :

$$Wdðμ, vÞ : = \inf_{π2Πðμ, vÞZ\ Ω × Ω} dðx, yÞπðdx, dyÞ : \quad ð5Þ$$

Lorsque μ, v sont des distributions discrètes sur des points de Rd, de taille m, n respectivement, cela peut être exprimé de manière équivalente en notation matricielle comme :

$$Wdðμ, vÞ : = \min_{Π ≥ 0} \sum_{je=1}^{PM} \sum_{j=1}^{Pn} Πijdðxi, xjÞ$$

$$\sum_{je=1}^{PM} Πij = vj,\ 8j\ 2\ f1, ... ,ng \quad ð6Þ$$

$$\text{sous réserve de : } \sum_{j=1}^{Pn} Πij = μj,\ 8i\ 2\ f1, ... ,mg$$

Pour les distances au sol générales, cela peut être calculé à l'aide de l'algorithme hongrois en temps O~ðn3Þ. Intuitivement, la difficulté du calcul du transport optimal est de trouver la carte Π, qui optimise le coût dans les limites des contraintes. Cependant, pour une métrique arborescente, cette carte optimale est facile à calculer sous forme fermée car il n'y a qu'un seul chemin (à travers l'arbre) entre les paires de points. Ce chemin unique entre des paires de points entraîne une complexité de calcul réduite de O~ðnÞ. Ceci est mieux compris en utilisant la forme double Kantorovich-Rubinstein de la distance de Wasserstein :

$$Wdðμ, vÞ = \sup_{kL ≤ 1} f : kf \quad Z\ Ω f\ ðxÞdμ\ Z \quad Ω f\ ðxÞdv \quad ð7Þ$$

où la fonction témoin f : Ω ! R et L désignent la norme de Lipschitz. Cette double forme est valable sous quelques conditions mineures, qui s'appliquent aux espaces considérés ici. Pour plus d'informations, voir 79.

Étant donné un arbre enraciné T avec des longueurs d'arêtes strictement non négatives, nous définissons la métrique de l'arbre naturel dT(x, y) comme la longueur du chemin unique entre les nœuds x, y. Nous notons la masse d'une distribution sur un sous-arbre Tr enraciné au nœud r comme μðTrÞ = P μðxÞ. Pour chaque nœud v T x2Tr , nous désignons son arête parent associée par ev avec un poids wv. Dans ce contexte, il est facile de construire la fonction témoin optimale dans l'équation. (7). Sans perte de généralité, on part de la racine et construit f tel que f(r) = 0 et pour chaque arête e(u, v) où u est parent de v, f(v) = f(u) + we signe(μ(Tv) − v(Tv)). Compte tenu de cette construction, il est facile de voir que la distance de Wasserstein avec la distance entre les arbres et le sol a la forme fermée suivante :

$$WdT ðμ, vÞ = X_{v2T} wv \quad μðTvÞ vðTvÞ \quad : \quad ð8Þ$$

La question se pose alors : quelles sont les métriques utiles des arbres ? Une métrique d'arbre idéale qui présente une faible distorsion de l'espace euclidien et qui est évolutive à des dimensions élevées. QuadTree77 est un algorithme de métrique arborescente conçu pour approximer la distance de transport optimale entre des mesures discrètes avec la distance au sol euclidienne en divisant de manière récursive l'espace en hypercubes, mais ne s'adapte pas bien à la dimension. Plus précisément, supposons, sans perte de généralité, que les données se trouvent dans l'hypercube [0, 1]d , puis à chaque niveau h [0, H) divisez l'espace en hypercubes 2dh de longueur de côté 2−h . Cela forme un arbre de niveau H avec chaque nœud représentant un hypercube.

Si le centre de l'hypercube est déplacé de manière aléatoire, alors la distance QuadTree WdQT présente une distorsion au plus Oðd log 1=τÞ où τ est la distance minimale entre les points de données, c'est-à-dire

$$c\ ðd \log τÞWdQT ðμ,vÞ ≤ Wkk2 ðμ,vÞ ≤ C\ ðd \log τÞWdQT ðμ,vÞ \quad ð9Þ$$

pour certaines constantes c, C en attente77.

Cependant, la distance QuadTree évolue mal car elle est calculée dans OðNd logðd1=τÞÞ. Dans le cadre de haute dimension, comme snRNAseq données, la mauvaise mise à l'échelle par rapport à d à la fois sur le plan informatique et dans l'approximation n'est pas souhaitable. Dans ce contexte78, il est suggéré d'échantillonner les arbres en utilisant le regroupement des points les plus éloignés80. De plus,76 implémente FlowTree, une petite modification de QuadTree qui rend les distances Wasser-stein des arbres beaucoup plus précises avec l'ajout d'un faible coût de calcul supplémentaire.

S'appuyant à la fois sur FlowTree et QuadTree, CATCH implémente une nouvelle formulation d'EMD sur l'arbre de condensation de diffusion. Pour deux amas de condensation par diffusion a, b situés respectivement à Ca, Cb , nous définissons la distance d'approximation de Wasserstein basée sur la condensation entre eux comme :

$$WCT ða, b,TÞ = k\ Ca\ Cbk2 + X_{eðu,vÞ2Ta} nous\ aðTvÞ + X_{eðu,vÞ2Tb} nous\ bðTvÞ \quad ð10Þ$$

où nous :=2−h Cv − Cu 2 pour le bord e(u, v) à la profondeur h et a(x), b(x) sont définis comme fonctions indicatrices de leurs clusters respectifs.

Cela conduit à la proposition suivante indiquant que peu importe à quel point nous sommes proches des paramètres de la proposition 1, WCT représente toujours une distance de Wasserstein d'arbre valide entre les clusters.

Proposition 2. La distance d'approximation de la distance de Wasserstein basée sur la condensation WCT, pour tout arbre de condensation de diffusion T, définit une distance de Wasserstein valide sur une distance au sol de l'arbre pour deux clusters quelconques de cet arbre.

Preuve. Nous montrons cela en construisant la métrique d'arbre associée dCT sur un arbre de condensation arbitraire TCT et concluons en montrant que WdTCT est équivalent à WCT. Commencez par enraciner l'arbre à un nœud représentant Ca avec deux enfants, la racine de Ta nommée ra et Cb. L'arête e(Ca,ra) a un poids 0 et l'arête (Ca, Cb) a un poids Ca − Cb 2.

Le nœud Cb aura un seul nœud enfant la racine de Ta nommé rb, et est connecté par une arête de longueur nulle. Tous les autres nœuds seront définis comme dans Ta et Tb avec les poids de bord associés.

Il est facile de vérifier que la mesure du chemin sur la construction TCT représente une distance dCT valide. Enfin, nous vérifions que la distance de Wasserstein avec une distance au sol de dCT est équivalente à WCT telle que définie dans l'éq. (dix). En effet, parce que nous avons ajouté une connexion par saut dans l'arbre pour connecter directement les nœuds a, b avec une arête de longueur Ca − Cb 2 et puisque a(Tv) pour v Tb est toujours nul et vice versa, nous avons

$$\begin{aligned} WdCT ða, bÞ &= X_{eðu,vÞ2TCT} nous\ \ aðTvÞ\ bðTvÞ \\ &= weðCa,CbÞ\ \ aðTCbÞ\ bðTCbÞ + X_{eðu,vÞ2Ta} we\ \ aðTvÞ\ bðTvÞ + X_{eðu,vÞ2To} nous\ \ aðTvÞ\ bðTvÞ \\ &= k\ Ca\ Cbk2\ \ 0\ 1 + X_{eðu,vÞ2Ta} nous\ \ aðTvÞ\ 0 + X_{eðu,vÞ2Tb} we\ \ 0\ bðTvÞ \\ &= k\ Ca\ Cbk2 + X_{eðu,vÞ2Ta} nous\ aðTvÞ + X_{eðu,vÞ2To} nous\ bðTvÞ \\ &= WCT ða,b,TÞ : \end{aligned} \quad ð11Þ$$

Notez que WCT ne calcule pas la distance de Wasserstein sur le même arbre pour chaque ensemble de clusters, et comme le montre la figure 76 , cela améliore souvent la précision par rapport à la comparaison. De plus, il est utile conceptuellement mais pas essentiel que les centres de cluster Ca, Cb soient proches des centroïdes de cluster. Dans la proposition 1, nous avons délimité le cadre dans lequel cela se vérifie exactement, mais ces paramètres ne sont pas pratiques pour notre calcul efficace nécessitant n − 1 étapes de diffusion. Au lieu de cela, nous nous contentons de centres proches des centroïdes mais pouvant être calculés efficacement en beaucoup moins d'étapes de diffusion. Notre formulation est similaire à la distance standard de Wasserstein avec la distance entre les arbres et le sol comme dans l'équation. (8), mais simplifié et optimisé pour le cas de comparaison de clusters, qui sont des éléments de la métrique arborescente. Nous effectuons deux changements. Tout d'abord, nous ajoutons une connexion sautée dans l'arbre pour connecter directement les nœuds a, b avec une arête de longueur Ca − Cb 2 comme dans la réf. 76, qui est empiriquement plus fidèle à leurs expériences et aux nôtres. Ensuite, on remarque que a(Tv) pour v Tb est toujours nul et vice versa, simplifiant ainsi les deuxième e

deux optimisations nous donnent un algorithme efficace en haute
dimensions et est efficace empiriquement (Fig. 1E et Supplémentaires
(Fig. 2D supplémentaire) à travers les granularités (Fig. 2E supplémentaire).

Grâce à cette intuition, CATCH est capable d'effectuer rapidement des calculs différentiels
analyse d'expression en approximant la métrique de Wasserstein sur une base par gène le long
des hiérarchies générées par la condensation de diffusion-fusion intrinsèque multiple. Tirer parti
de la capacité de notre approche à résumer
paysages transcriptomiques avec le noyau de désintégration α, nous utilisons plusieurs
granularités de la hiérarchie cellulaire pour se rapprocher avec précision
vérité terrain, distance de Wasserstein entre les gènes et identification des signatures
d'expression spécifiques au cluster78 (Fig. 1D-iv). Nous montrons que c'est
empiriquement vrai avec nos comparaisons (Fig. 2D supplémentaire et
Fig. 1F supplémentaire).

Inspiré par des méthodes antérieures statistiquement solides d'identification
gènes différentiellement exprimés, nous implémentons une méthode basée sur le rééchantillonnage
approche pour identifier les vrais gènes différentiellement exprimés73,81. Dans ce
approche, nous estimons le taux de fausses découvertes (FDR), qui est la proportion attendue
d'hypothèses nulles rejetées faussement pour le test de chaque gène.
statistique à un niveau de signification donné73,81. Pour calculer les FDR à partir de notre
valeurs de Wasserstein, nous générons une distribution nulle en permutant les
regroupez les étiquettes (en pratique 1000 fois) et calculez à chaque fois la distance de
Wasserstein entre les classes permutées. En utilisant la médiane de
En permutant les distances de Wasserstein pour chaque gène, nous créons une distribution
nulle à partir de laquelle nous pouvons calculer les valeurs p par gène. L'atteint
Les valeurs p sont corrigées à l'aide de la procédure Benjamini – Hochberg82.

Caractérisation automatisée des clusters et Earth Mover's Distance
entre les gènes dans les données synthétiques et réelles d'une seule cellule. Alors que
condensation par diffusion intrinsèque au collecteur mise en œuvre avec une désintégration α
Le noyau peut théoriquement se rapprocher des caractérisations de clusters de vérité terrain et
calculer des gènes exprimés différentiellement, nous voulions
démontrer ce raisonnement dans des données unicellulaires synthétiques et réelles. À
montrent empiriquement que notre approche basée sur la condensation se rapproche
EMD entre deux clusters, nous calculons les valeurs EMD entre gènes
en utilisant le transport optimal de Wasserstein ainsi que nos approximations
approche sur données synthétiques et réelles utilisant un noyau gaussien et de désintégration α
implémentations de condensation par diffusion. À l'aide de données unicellulaires générées à
partir d'éclaboussures, nous calculons la condensation de diffusion et identifions
la granularité avec la persistance topologique la plus élevée en utilisant l'analyse d'activité
topologique. Nous avons ensuite calculé la vérité terrain et approximé
valeurs d'expression différentielle en comparant chaque cluster à cette granularité avec tous les
autres clusters. Dans notre analyse, un total de 12 130 200 et
4 535 640 comparaisons de gènes ont été calculées respectivement à l'aide des approches
gaussiennes et de désintégration α. Comparaison des désintégrations gaussienne et α
Distances approximatives de Wasserstein par rapport à la vérité terrain par gène
Valeurs de Wasserstein, nous pouvons voir la valeur dans notre approche de désintégration α
(Fig. 2D supplémentaire) car il se rapproche de la vérité terrain Wasserstein
distance avec un coefficient de corrélation de 0,979. De plus, notre
L'approche a calculé la comparaison des 4 535 640 gènes en 63 s au sol
les valeurs de vérité ont été calculées en 43 125 s, ce qui équivaut à une multiplication par 684
en vitesse de calcul.

Nous avons répété notre comparaison avec des données réelles sur une seule cellule,
en comparant à nouveau les deux approches des valeurs de vérité terrain de Wasserstein EMD.
temps sur 10 granularités identifiées par analyse d'activité topologique.
Comme précédemment réalisé, à chaque granularité, tous les clusters ont été comparés à tous
les autres clusters en utilisant chaque approche. Parmi toutes les comparaisons, un total de
10 166 640 et 2 541 660 comparaisons ont été calculées.
pour les implémentations gaussiennes et α-décroissance, respectivement. Encore une fois, nous
voir que la désintégration α est essentielle pour capturer avec précision la vérité terrain EMD
valeurs, notre approche de désintégration α étant fortement corrélée à la vérité terrain
EMD alors que l'approche gaussienne était moins corrélée (complémentaire
Figure 1F). De plus, nous constatons à nouveau une augmentation de la vitesse de calcul
avec notre approche basée sur la condensation. Dans notre implémentation pondérée, nous
sommes capables de calculer les 2 541 660 comparaisons en 32 s,
tandis que les valeurs EMD de vérité terrain ont été calculées en 27 517 s, ce qui équivaut à
une augmentation similaire de 860 fois de la vitesse de calcul. Ensuite, nous montrons que
cette corrélation entre la vérité terrain EMD et la condensation
L'approximation de la distance de Wasserstein n'est pas une caractéristique de la granularité
des clusters telle que définie par le nombre de clusters (Fig. 3D supplémentaire). Enfin,
nous utilisons également des implémentations de désintégration α et gaussienne pour calculer et
comparer les caractérisations de clusters à la vérité terrain dans une cellule unique réelle
données. Utiliser le même ensemble de clusters et de granularités que précédemment
calculé, nous voyons que le noyau de désintégration α caractérise à nouveau les clusters avec
plus de précision qu'un noyau gaussien (Fig. 3C supplémentaire).

CATCH identifie les gènes différentiellement exprimés à partir de données unicellulaires
bruyantes. Auparavant, les signatures de maladies au sein d'un type cellulaire étaient
déterminé en comparant les profils d'expression génique des cellules basés sur
leur état d'origine. Par exemple, les microglies seraient séparées
en deux groupes en fonction de la condition d'origine, soit malade, soit en bonne santé,
qui serait ensuite comparé. Nous pensons que CATCH améliore
ce cadre en identifiant d'abord les états enrichis en maladies, puis
identifier des gènes différentiellement exprimés entre ces états. C'est
car notre procédure prend en compte un bruit important qui peut apparaître
dans les données unicellulaires pour identifier de manière plus pure les états cellulaires enrichis
dans des contextes pathologiques particuliers. En fait, des études antérieures ont validé que
cette approche identifie mieux les processus biologiques que la précédente
approches de comparaison des « conditions d'origine »9 .

Pour illustrer ce point dans des données réelles sur une seule cellule, nous avons effectué
analyse de l'expression différentielle entre les microglies en fonction de leur
condition d'origine des trois maladies neurodégénératives
ensembles de données. Nous pensons que si notre approche est plus sensible pour identifier
des gènes différentiellement exprimés, une approche moins sensible serait
pas trouvé une signature partagée aussi forte. Après avoir défini la signification
seuils basés sur nos taux de fausses découvertes par gène, nous avons identifié
gènes significativement enrichis dans la phase active précoce ou aiguë de
chaque maladie (Fig. 10a supplémentaire). Cependant, dans toutes les comparaisons, nous
avons identifié beaucoup moins de
gènes dans cette analyse de type cellulaire (135, 68 et 416) qu'avec notre
pipeline (618, 795 et 1551 pour AMD, AD et MS, respectivement),
indiquant que l'identification de sous-types cellulaires pathogènes
avec CATCH avant la comparaison augmente notre capacité à détecter les gènes exprimés de
manière différentielle. Dans les comparaisons entre maladies entre
microglies neurodégénératives à un stade précoce, seulement 17 gènes communs
ont été trouvés, nettement moins que les 168 gènes communs trouvés
avec notre pipeline. Parmi les gènes communs, seulement la moitié de l'activation
une signature a été trouvée (APOE, B2M, FTH1, FTL, SPP1). Semblable à notre
comparaison microgliale à gros grains, nous avons comparé la force
de notre approche dans les astrocytes. Après avoir défini des seuils de signification
sur la base de nos valeurs q par gène, nous avons identifié beaucoup moins
gènes enrichis (221, 271 et 886) que ceux que nous avons trouvés avec notre analyse
(1 444, 680 et 2 278 gènes pour la DMLA, la MA et la SEP, respectivement)
(Fig. 10b supplémentaire). Dans notre analyse au niveau du type de cellule, seulement 28
des gènes communs ont été trouvés, nettement moins que les 630 gènes communs
gènes trouvés avec notre pipeline. Parmi les gènes communs, seulement la moitié
la signature d'activation a été trouvée (AQP4, CD81, CRYAB, GFAP).

Collectivement, ces comparaisons révèlent la sensibilité de ce pipeline de découverte
pour trouver des signatures génétiques et des relations biologiquement significatives dans des
données bruyantes d'expression génique unicellulaire.

Autres détails sur les méthodes de calcul

Séquençage et prétraitement de l'ARN AMD mononucléaire. Les données snRNA-seq provenant
d'échantillons maculaires ont été traitées selon les étapes suivantes. Exemple de démultiplexage
et alignement de lecture sur le NCBI
la référence pré-ARNm GRCh38 a été complétée pour mapper les lectures aux deux
transcrits de pré-ARNm et d'ARNm matures non épissés à l'aide de CellRanger
version 3.1.0. Matrices génétiques et cellulaires de rétines atteintes de DMLA sèche (n = 4),
DMLA néovasculaire (n = 7) ou témoins sans maladie rétinienne connue
(n = 6) ont ensuite été regroupés en un seul fichier. Nous avons préfiltré en utilisant
paramètres dans scprep (v1.0.3, https://github.com/KrishnaswamyLab/

scprep). Les cellules contenant au moins 1 400 transcrits uniques ont été conservées pour une analyse plus approfondie afin de générer une matrice cellule par gène contenant 70 973 cellules. La normalisation a été effectuée en utilisant les paramètres par défaut avec la normalisation L1, en ajustant le côté total de la bibliothèque de chaque cellule à 1 000. Toute cellule présentant plus de 200 comptes normalisés d'ARNm mitochondrial a été supprimée. La correction par lots a été effectuée à l'aide d'Harmony (https://github.com/immunogenomics/harmony) pour aligner les effets de lot introduits par le lot de séquençage, l'intervalle post-mortem, le lieu d'acquisition de l'échantillon et la chimie du séquençage 10X83. Les fichiers de données brutes pour les données snRNA-seq humaines seront disponibles en téléchargement via GEO sous un numéro d'accès à attribuer sans aucune restriction sur la disponibilité des données.

Prétraitement du séquençage d'ARN mononucléaire AD et MS. Les données snRNA-seq pour la MA et la SEP ont été acquises à partir de sources publiées4,5 .
Les cellules contenant au moins 1 000 transcrits uniques ont été conservées pour une analyse plus approfondie afin de générer une matrice cellule par gène pour chaque maladie. La normalisation a été effectuée en utilisant les paramètres par défaut de scprep avec la normalisation L1, en ajustant le côté total de la bibliothèque de chaque cellule à 1 000. Toute cellule présentant plus de 200 comptes normalisés d'ARNm mitochondrial a été supprimée. Une correction par lots a été effectuée sur les données MS à l'aide d'Harmony (https://github.com/immunogenomics/harmony) pour aligner les effets de lot introduits par le lot de séquençage, le lot de capture et le sexe.

Identification du type de cellule avec CATCH. Tous les types de cellules ont été identifiés en effectuant une analyse d'activité topologique sur l'homologie de condensation calculée par condensation de diffusion. Afin d'identifier les types de cellules, nous avons identifié une résolution sans activité topologique, qui divise bien l'espace d'état cellulaire et attribue à chaque groupe un type de cellule basé sur des gènes marqueurs spécifiques au type de cellule.

Analyse des interactions. L'analyse cellule-cellule ligand-récepteur a été réalisée sur des données d'expression de snRNA prétraitées à l'aide du package python CellPhoneDB (// github.com/Teichlab/cellphonedb, v2.1.4)52.
Avant de procéder à l'analyse, la base de données de 834 combinaisons ligand-récepteur et complexes protéiques à unités multiples a été complétée par 2557 interactions ligand-récepteur trouvées dans la base de données celltalker (https://github.com/arc85/celltalker) 84. La fonction intégrée de génération de base de données a été utilisée pour mettre à jour la base de données existante. Notre base de données complète générée par l'utilisateur a été invoquée à chaque exécution de la fonction de commande d'analyse statistique CellPhoneDB.

Les cartes d'interaction CellPhoneDB ont été calculées sur différentes entrées. Tout d'abord, des microglies et des astrocytes enrichis en phase de maladie avec une identité de sous-groupe ont été analysés pour identifier les interactions de signalisation entre les états d'activation des astrocytes et des microglies (Fig. 6B).
Le nombre de permutations a été fixé à 2 000 et le seuil de valeur p à 0,01.

Les méthodes biologiques détaillent
les tissus humains. Les yeux post-mortem pour le test Chromium Single Cell 3' (n = 17) et les dossiers médicaux contenant le stade de la DMLA ont été obtenus auprès d'Advancing Sight Network (Alabama), de la Lions Gift of Sight Eye Bank (Minnesota) ou du Département de pathologie de Yale avec un intervalle post mortem maximum de 13 h. Les globes ont été examinés pour détecter une maladie rétinienne par un ophtalmologiste (HBP) avant la dissection et la dissociation des échantillons. La rétine pour snRNA-seq a été obtenue auprès de donneurs post-mortem humains non apparentés, comprenant des stades de DMLA normaux, intermédiaires secs sur AREDS2 et néovasculaires (Tableau supplémentaire 1). Pour chaque échantillon, nous avons profilé la macula, qui est la région de la rétine responsable de la vision centrale et la plus gravement touchée par la pathologie DMLA. Nous avons identifié quatre échantillons de DMLA intermédiaire provenant de patients prenant le supplément oculaire en vitamines et minéraux AREDS2 avec des drusen, un signe pathologique associé au stade sec intermédiaire de la maladie. Sept échantillons de DMLA post-mortem présentaient une néovascularisation à un stade avancé de la maladie. Normale
les donneurs n'avaient aucun antécédent de maladie rétinienne. Des données cliniques supplémentaires sur les sujets sont présentées dans le tableau supplémentaire 2.

Dissection rétinienne et isolement des noyaux du tissu rétinien congelé.
Les globes ont été placés dans du RNAlater (ThermoFisher) et transportés sur la glace. Des poinçons trépan (diamètre 6 mm) ont été utilisés pour isoler des échantillons de la macula de la rétine centrale, située à l'écart de la papille optique et des artérioles principales. Pour chaque morceau de tissu, la rétine a été séparée mécaniquement de l'épithélium-choroïde pigmentaire rétinien sous-jacent, congelée sur de la neige carbonique et conservée à –80 ° C. Les noyaux ont été isolés et purifiés à l'aide du kit d'isolation Nuclei EZ Prep Nuclei (Sigma), en suivant le protocole du fabricant, avec quelques modifications. Toutes les procédures ont été effectuées sur de la glace ou à 4°C. En bref, le tissu rétinien congelé a été soumis à une homogénéisation rapide (25 fois avec le pilon A suivi de 25 fois avec le pilon serré B) à l'aide du kit de broyeur de tissus KIMBLE Dounce (Sigma) dans 2 ml de tampon de lyse EZ. L'échantillon a été transféré dans un tube de 15 ml avec 2 ml supplémentaires de tampon de lyse EZ et incubé sur de la glace pendant 5 min. Après l'incubation, l'échantillon a été centrifugé à 500 xg, 5 min à 4°C. Les surnageants ont été jetés et les noyaux isolés ont été remis en suspension dans 4 ml de tampon de lyse EZ, incubés pendant 5 minutes sur de la glace et centrifugés à 500 xg pendant 5 minutes à 4 ° C. Ensuite, les noyaux ont été lavés avec 4 ml de tampon de suspension Nuclei glacé (1x solution saline tamponnée au phosphate (PBS) contenant 0,01% de BSA et 0,1% d'inhibiteur de RNase), remis en suspension dans 1 ml de tampon de stockage Nuclei EZ et passés à travers un nylon de 40 µM. tamis cellulaire. Les suspensions de noyaux ont été comptées avec du bleu trypan avant leur chargement sur la plateforme microfluidique.

SnRNA-seq microfluidique basé sur des gouttelettes. Les noyaux isolés de chaque échantillon maculaire ont été traités par séquençage d'ARN nucléaire unique basé sur la microfluidique. Des bibliothèques unicellulaires ont été préparées à l'aide des plates-formes Chro-mium 3' v2 et v3 (10x Genomics) en suivant le protocole du fabricant. En bref, les noyaux uniques ont été divisés en billes de gel dans une émulsion dans l'instrument 10x Chromium Controller, suivis d'une lyse et d'une transcription inverse avec code à barres de l'ARN, d'une amplification, d'un cisaillement et d'un adaptateur 5' et d'une fixation d'index d'échantillon. En moyenne, 7 000 noyaux ont été chargés sur chaque canal, ce qui a permis de récupérer 4 000 noyaux. Les bibliothèques ont été séquencées sur la plateforme Illumina NextSeq 500. Les données de séquence brutes ont été alignées sur le génome humain GRCh38-3.0.0 à l'aide de l'aligneur STAR, et le logiciel Cell Ranger (v3.1.0, 10x Genomics) a été utilisé pour démultiplexer les lectures et attribuer le nombre de lectures à des cellules individuelles. (Après le prétraitement du contrôle qualité, les profils snRNA-seq ont été utilisés dans les analyses ultérieures. Cet ensemble de données a été corrigé pour tenir compte des effets de lot sur les échantillons à l'aide de l'algorithme Harmony83.

Hybridation d'ARN in situ et immunofluorescence. Pour valider les différences d'expression génique, une hybridation in situ a été réalisée à l'aide du test RNAscope Multiplex Fluorescent V2 (Advanced Cell Diagnostics, Hayward, CA, USA). Des macules disséquées de globes humains entiers ont été fixées dans du paraformaldéhyde à 4 % (PFA) à 4 °C pendant la nuit. Les tissus ont été séquentiellement déshydratés avec 15 % de saccharose, puis 30 % de saccharose avant d'être incorporés dans de l'OCT, et congelés sur de la neige carbonique. Les moules OCT ont été sectionnés à 10 µm d'épaisseur. L'hybridation in situ de l'ARN a été réalisée selon le protocole du fabricant. En bref, les coupes congelées fixes ont été cuites à 60 ° C pendant 1 heure avant l'incubation dans du PFA à 4% pendant 10 minutes et le prétraitement par digestion par la protéase. Les sondes cibles ont été hybridées avec un système d'amplification de signal sensible à la température basé sur HRP, suivi d'un développement de couleur. Les gènes de ménage POLR2A, PPIB et UBC ont été utilisés comme ARNm de contrôle interne (Fig. 7 supplémentaire); si les sondes de ces ARNm n'étaient pas visualisées, l'échantillon était considéré comme non disponible pour l'étude de l'expression génique. Les sondes utilisées incluent APOE, TYROBP, B2M, VEGFA et HIF1A (Advanced Cell Diagnostics, Hayward, CA, USA). Les lames ont été contre-colorées au DAPI pendant le protocole d'immunofluorescence (voir ci-dessous). La coloration positive a été déterminée par des points ponctués fluorescents dans les canaux appropriés

dans le noyau et/ou le cytoplasme. Après le protocole d'hybridation in situ de l'ARN, les coupes congelées fixes ont été bloquées avec du sérum animal et incubées pendant une nuit à 4 ° C avec des anticorps primaires (voir le segment d'anticorps ci-dessous). L'incubation des anticorps secondaires a duré 1 heure à température ambiante et les noyaux cellulaires ont été contre-colorés avec du DAPI. Les images ont été capturées immédiatement à l'aide d'un microscope confocal (Zeiss LSM800, Jena, Allemagne). Les anticorps suivants contre les antigènes humains ont été utilisés : GFAP (1 : 500, MA5-12023, Invitrogen) et Iba1 (1 : 500, 019-19741, Fujifilm). Les anticorps ont été visualisés avec Alexa Fluor 488 (1: 200, A-11001/A-21208, Invitrogen).

Souris. Des souris mixtes C57BL/6 âgées de quatre à huit semaines ont été achetées auprès du National Cancer Institute, puis élevées et hébergées à l'Université de Yale. Toutes les procédures utilisées dans cette étude (appariées selon le sexe et l'âge) étaient conformes aux directives fédérales et aux politiques institutionnelles du Comité de protection et d'utilisation des animaux de la Yale School of Medicine (protocole approuvé par l'IACUC n° 2022-20275) régissant le bien-être et l'éthique des animaux. traitement.

Cellules. Les cellules astrocytes dérivées d'IPSC ont été achetées sur Brainxell.com (numéro de catalogue BX-0600; Brainxell, Madison Wisconsin). Les cellules ont été cultivées conformément aux directives du fournisseur en utilisant un milieu DMEM/F12 1:1 et Neurobasal avec un supplément de N2 (1x), du Glutamax (0,5 mM), un supplément d'Astrocytes (1x), du sérum de veau fœtal (1 %).

Culture de cellules. Les cellules d'astrocytes dérivées d'IPSC ont été cultivées jusqu'à un état complètement différencié avant la stimulation par les cytokines. Les cytokines (IL-1β, IL2, IL4, IL6, IL7, IL10, IL12, IL15, IL17, IL22, IL23, IFNG, TNF) ont toutes été achetées auprès de PeproTech.com (Peprotech, Cranbury, NJ). Pour la stimulation d'une seule cytokine, les cellules ont été stimulées avec chaque cytokine à une concentration de 100 ng/mL pendant 24 h. Pour la stimulation combinatoire des cytokines, un cocktail de toutes les cytokines moins la cytokine d'intérêt a été réalisé avec une concentration de chaque cytokine à 50 ng/mL. Les cellules ont été stimulées pendant 24 heures avant la collecte du milieu. Le milieu collecté a été centrifugé à 1 000 xg pour éliminer les cellules et débris avant d'effectuer un ELISA.

Dosage immuno-enzymatique. Le test ELISA (Enzyme-linked immunosor-bent) a été réalisé à l'aide d'un kit ELISA VEGF-A pour souris (Cusabio LLC) en suivant les instructions du fabricant. En bref, deux puits d'une plaque de microtitration en PVC ont été recouverts de 100 µL d'antigène (10 µg/mL dans du PBS), après quoi la plaque a été scellée et incubée pendant 2 h à température ambiante. Après trois lavages avec du PBS et l'application d'un tampon de blocage (5 % de lait sec dans du PBS), la plaque a été refermée et incubée pendant 2 ha à température ambiante. La plaque a été lavée deux fois avec du PBS et un anticorps anti-VEGF-A dans un tampon de blocage a été ajouté aux puits. Après une nouvelle incubation de 2 h à température ambiante, la plaque a été lavée 5 fois avec du PBS et 100 µL de solution de substrat ont été ajoutés aux puits. Une solution d'arrêt a été ajoutée aux puits et l'absorbance à 450 nm a été enregistrée dans un lecteur de plaques.

Injection intravitréenne. Les souris ont été anesthésiées à l'aide d'un mélange de kétamine (50 mg/kg) et de xylazine (5 mg/kg), injecté par voie intrapéritonéale. Les yeux des souris ont été stérilisés à la bétadine. Un petit trou a été réalisé sur la face latérale du limbe à l'aide d'une seringue à insuline de calibre 33. À l'aide d'une seringue Hamilton à extrémité émoussée, 1 µL de PBS ou d'IL-1β (100 ng) a été injecté à un angle de 45 degrés au niveau du limbe par voie intravitréenne.
Une fois la perfusion terminée, la seringue a été laissée en place pendant une minute avant de la retirer. Le site d'injection a été lavé avec du PBS stérile et une pommade vétérinaire Puralube a été appliquée sur les yeux. Les souris ont été surveillées jusqu'à leur guérison complète.

Traitement des tissus de souris et microscopie. Les rétines ont été disséquées, fixées dans 2 % de PFA pendant 1 h et immédiatement traitées dans une solution de blocage (10 % de sérum d'âne normal, 1 % d'albumine sérique bovine, 0,3 % de PBS-Triton X-100) pour une incubation d'une nuit à 4 C. Après incubation, un

Un sous-ensemble de rétines pour l'imagerie RPE a été blanchi avec un traitement avec 2 ml de H2O2 à 30 % + 8 ml de PBS + 2 pastilles de NaOH jusqu'à ce qu'elles soient optiquement claires (30 min). Des anticorps primaires (VEGFA ; Invitrogen cat# MA5-13812) ont été appliqués et les coupes ont été incubées pendant une nuit à 4 ° C, puis lavées cinq fois à température ambiante dans du PBS et du Triton X-100 à 0,5 %, avant d'être incubées avec un anticorps secondaire fluoroconjugué. dilué dans du PBS et 0,5% de Triton X-100 pendant 2 h à température ambiante. Les coupes ont été lavées cinq fois à température ambiante, colorées au DAPI et montées avant l'imagerie. Les images confocales ont été prises sur un microscope Leica SP8. L'analyse quantitative a été réalisée à l'aide du logiciel de traitement d'image FIJI ou ImageJ (NIH ou Bethesda) ou du logiciel Imaris 8 (Oxford Instruments).

Statistiques et reproductibilité. Lorsque deux groupes indépendants ont été comparés, le test t de Welch a été utilisé lorsque des variances inégales étaient supposées, et le test t de Student pour une variance présumée égale. Toutes les comparaisons ont été effectuées à l'aide de tests bilatéraux. Des tests du chi carré ont été utilisés pour comparer les proportions entre deux groupes. Les expériences d'hybridation in situ (telles que représentées sur les figures 3G et 4G) ont été répétées deux fois dans chaque cas. Lorsque trois groupes indépendants ou plus étaient comparés, des tests multinomiaux bilatéraux avec correction de comparaisons multiples étaient utilisés, le cas échéant. Les barres d'erreur tracées sur les visualisations des moyennes représentent l'erreur standard de la moyenne. L'analyse de l'expression différentielle dans le cadre de l'algorithme CATCH comprend une distance bilatérale de Earth Mover, c'est-à-dire une distance de Wasserstein unidimensionnelle, avec des seuils de signification établis sur la base des taux de fausses découvertes par gène (test EMD bilatéral avec FDR corrigé). valeur p < 0,1). Sur la figure 3A, 141 microglies ont été identifiées, dont 30 dans une population enrichie en bonne santé, 32 dans la population sèche enrichie en DMLA et 79 dans la population humide enrichie en DMLA néovasculaire. Sur la figure 4A, 474 astrocytes ont été identifiés, dont 301 dans la population également proportionnée, 22 dans la population enrichie en bonne santé, 96 dans la population sèche enrichie en DMLA et 55 dans la population humide néovasculaire enrichie en DMLA.

Dans les Fig. 3f et 4f, les diagrammes en boîte et en moustaches sont définis comme suit : les moustaches contiennent l'intervalle de confiance interne de 95 %, la limite inférieure de la boîte est les 25 % et la limite supérieure les 75 % des valeurs. Enfin, la médiane au centre de la case désigne les 50 %. Dans cette figure et ci-dessous, toutes les valeurs sont rapportées en valeurs totales d'expression génique normalisées. Dans la figure 3, les signatures d'activation microgliale sont présentées dans des groupes de microglies dans trois maladies neurodégénératives. Les microglies enrichies en contrôle AMD ont une signature minimale de 7,3, une moustache inférieure de 7,5, une limite inférieure de 7,9, une médiane de 8,4, une limite supérieure de 8,8, une moustache supérieure de 9,0 et un maximum de 9,4. Les microglies sèches enrichies en DMLA ont un minimum de 6,6, une moustache inférieure de 7,0, une limite inférieure de 8,3, une médiane de 9,2, une limite supérieure de 10,2, une moustache supérieure de 10,8 et un maximum de 11,9. Dans le cluster enrichi pour le contrôle de la maladie d'Alzheimer, cette signature a un minimum de 16,7, une moustache inférieure de 17,0, une limite inférieure de 17,2, une médiane de 17,6, une limite supérieure de 18,2, une moustache supérieure de 19,0 et un maximum de 20.1. Dans le groupe enrichi en maladies précoces, cette signature a un minimum de 16,3, une moustache inférieure de 17,7, une limite inférieure de 20,5, une médiane de 21,4, une limite supérieure de 23,2, une moustache supérieure de 24,2 et un maximum de 25.6. Dans le premier cluster MS enrichi en contrôle progressif, cette signature a un minimum de 11,7, une moustache inférieure de 12,5, une limite inférieure de 13,2, une médiane de 13,8, une limite supérieure de 15,4, une moustache supérieure de 17,3 et un maximum de 17.7. Dans le groupe enrichi en maladie progressive précoce, cette signature a un minimum de 15,4, une moustache inférieure de 15,5, une limite inférieure de 17,9, une médiane de 19,0, une limite supérieure de 19,9, une moustache supérieure de 21,4 et un maximum de 21,6. . Sur la figure 4, les signatures d'activation des astrocytes sont présentées pour les groupes d'astrocytes dans les trois maladies. Les astrocytes enrichis en contrôle AMD ont une signature minimale de 1,4, une moustache inférieure de 1,9, une limite inférieure de 2,5, une médiane de 3,0, une limite supérieure de 3,4, une moustache supérieure de 4,3 et un maximum de 6,8. Les astrocytes secs

limite inférieure de 2,5, médiane de 3,1, limite supérieure de 6,7, moustache supérieure de 9,2 et maximum de 10,8. Dans le cluster enrichi pour le contrôle de la maladie d'Alzheimer, cette signature a un minimum de 6,6, une moustache inférieure de 6,7, une limite inférieure de 8,8, une médiane de 9,3, une limite supérieure de 10,1, une moustache supérieure de 12,5 et un maximum de 16,8. . Dans le groupe enrichi en maladies précoces, cette signature a un minimum de 6,4, une moustache inférieure de 6,5, une limite inférieure de 9,1, une médiane de 9,5, une limite supérieure de 11,3, une moustache supérieure de 12,4 et un maximum de 14.1. Dans le premier cluster enrichi en contrôle MS progressif, cette signature a un minimum de 3,2, une moustache inférieure de 3,5, une limite inférieure de 4,6, une médiane de 5,3, une limite supérieure de 6,4, une moustache supérieure de 7,9 et un maximum de 8,1. Dans le groupe enrichi en maladie progressive précoce, cette signature a un minimum de 4,3, une moustache inférieure de 4,4, une limite inférieure de 6,4, une médiane de 7,0, une limite supérieure de 7,8, une moustache supérieure de 9,4 et un maximum de 14,6. .

Résumé du rapport De plus
amples informations sur la conception de la recherche sont disponibles dans le résumé du rapport du portefeuille Nature lié à cet article.

Disponibilité des données
Les données sources sont fournies sous forme de fichier de données sources. Les fichiers de données brutes et traitées pour les données snRNA-seq utilisées dans cette étude sont disponibles en téléchargement via GEO sous le numéro d'accès GSE221042. Les données utilisées dans cette étude proviennent de la réf. 4 est disponible sur le centre de partage de ressources de recherche du Rush Alzheimer's Disease Center à l'adresse https://www.radc. rush.edu/docs/omics.htm ou sur Synapse (https://www.synapse.org/#!Synapse : syn18485175) sous https://doi.org/10.7303/syn18485175.
Les données utilisées dans cette étude proviennent de la réf. 5 sont disponibles dans Sequence Read Archive (SRA) sous le numéro d'accès PRJNA544731 (NCBI Bioproject ID : 544731) ou sur https://ms.cells.ucsc.edu. Les données sources sont fournies avec ce document.

Disponibilité du code Le
package CATCH, tel qu'implémenté en python, est disponible en téléchargement avec un didacticiel guidé sur la page Github du Krishnaswamy Lab : https://github.com/KrishnaswamyLab/CATCH.

Références 1.

Wong, WL et al. Prévalence mondiale de la maladie maculaire liée à l'âge
projection de la dégénérescence et de la charge de morbidité pour 2020 et 2040 : une revue systématique et une méta-analyse. Lancette Glob. Santé 2, e106-e116 (2014).
2. Mitchell, P., Liew, G., Gopinath, B. et Wong, TY
dégénérescence maculaire. Lancette 392, 1147-1159 (2018).
3. Bird, AC et coll. Un système international de classification et de notation de la maculopathie liée à l'âge et de la dégénérescence maculaire liée à l'âge. Le Groupe international d'étude épidémiologique ARM. Survivre. Ophtalmol. 39, 367-374 (1995).
4. Mathys, H. et al. Analyse transcriptomique unicellulaire de la maladie d'Alzheimer maladie. Nature 570, 332-337 (2019).
5. Schirmer, L. et al. Vulnérabilité neuronale et diversité multilignée dans la sclérose en plaques. Nature 573, 75-82 (2019).
6. Habib, N. et al. Astrocytes associés à la maladie d'Alzheimer et le vieillissement. Nat. Neurosci. 23, 701-706 (2020).
7. Keren-Shaul, H. et al. Un type de microglie unique associé au développement restreint de la maladie d'Alzheimer. Cellule 169, 1276-1290 (2017).
8. Brugnone, N. et al. Granulation grossière des données via des données inhomogènes condensation par diffusion. Lors de la conférence internationale de l'IEEE 2019 sur Big Data (Big Data), 2624-2633 (IEEE, 2019).
9. Burkhardt, DB et al. Quantification de l'effet des perturbations expérimentales dans les données de séquençage d'ARN unicellulaire à l'aide du traitement du signal graphique. Nat. Biotechnologie. 39, 619-629 (2020).
10. Lemprière, S. Activité inflammatoire du NLRP3 comme biomarqueur du pri-Mary sclérose en plaques progressive. Nat. Révérend Neurol. 16, 350-350 (2020).
11. Zhang, Y., Dong, Z. & Song, W. Inflammasome NLRP3 en tant que roman Cible thérapeutique pour la maladie d'Alzheimer. Transduction de signal. Cible. Là. 5, 37 (2020).
12. White, CS, Lawrence, CB, Brough, D. & Rivers-Auty, J. Inflam-masomes comme cibles thérapeutiques pour la maladie d'Alzheimer. Pathologie cérébrale. 27, 223-234 (2017).
13. Faissner, S., Plemel, JR, Gold, R. & Yong, VW Sclérose en plaques progressive : de la physiopathologie aux stratégies thérapeutiques. Nat. Rév. 18, 905-922 (2019).
14. Huang, W.-J., Chen, W.-W. & Zhang, X. Sclérose en plaques : pathologie, diagnostic et traitements. Exp. Là. Méd. 13, 3163-3166 (2017).
15. Braak, H. & Braak, E. Stadification neuropathologique des changements liés à la maladie d'Alzheimer. Acta Neuropathol. 82, 239-259 (1991).
16. Ding, J. et coll. Comparaison systématique des méthodes de séquençage d'ARN unicellulaire et mononucléaire. Nat. Biotechnologie. 38, 737-746 (2020).
17. Huguet, G. et al. Géométrie et topologie de diffusion inhomogènes dans le temps. https://arxiv.org/abs/2203.14860 (2022).
18. Moyle, MW et coll. Principes structurels et développementaux de l'assemblage des neuropiles en c. élégans. Nature 591, 99-104 (2021).
19. Zappia, L., Phipson, B. & Oshlack, A. Splatter : simulation de données de séquençage d'ARN unicellulaire. Génome Biol. 18, 174 (2017).
20. Wagner, DE et al. Cartographie unicellulaire des paysages d'expression génique et de la lignée dans l'embryon de poisson zèbre. Science 360, 981-987 (2018).
21. Aghaeepour, N. et al. Évaluation critique des techniques automatisées d'analyse des données de cytométrie en flux. Nat. méthodes 10, 228-238 (2013).
22. Menon, M. et al. L'atlas transcriptomique unicellulaire de la rétine humaine identifie les types de cellules associés à la dégénérescence maculaire liée à l'âge. Nat. Commun. 10, 4902 (2019).
23. Blondel, VD, Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast déploiement des communautés en grands réseaux. J.Stat. Mécanique. Théorie Exp. 2008, P10008 (2008).
24. Shekhar, K. et al. Classification complète des neurones bipolaires rétiniens par transcriptomique (unicellulaire). Cellule 166, 1308-1323.e30 (2016).
25. Peng, Y.-R. et coll. Classification moléculaire et taxonomie comparative des cellules fovéales et périphériques de la rétine des primates. Cellule 176, 1222-1237.e22 (2019).
26. Yan, W. et coll. Atlas cellulaire de la fovéa humaine et de la rétine périphérique. Sci. Rep.10, 9802 (2020). 27. van
Dijk, D. et coll. Récupération des interactions génétiques à partir de données unicellulaires par diffusion de données. Cellule 174, 716-729.e27 (2018).
28. Srinivasan, K. et coll. Les microglies des patients atteints de la maladie d'Alzheimer présentent un vieillissement accru et une activation transcriptionnelle unique. Cell Rep.31, 107843 (2020).
29. Friedman, BA et coll. Divers profils d'expression myéloïde cérébrale révèlent des états d'activation microgliale distincts et des aspects de la maladie d'Alzheimer qui ne sont pas évidents dans les modèles murins. Cell Rep.22, 832-847 (2018).
30. Krasemann, S. et coll. La voie TREM2-APOE pilote la transition phénotype descriptif des microglies dysfonctionnelles dans les maladies neurodégénératives. Immunité 47, 566-581 (2017).
31. Corder, EH et coll. Dose génétique de l'allèle de l'apolipoprotéine E de type 4 et risque de maladie d'Alzheimer dans les familles à apparition tardive. Science 261, 921-923 (1993).
32. Lambert, JC et coll. Une méta-analyse de 74 046 individus identifie 11 nouveaux locus de susceptibilité à la maladie d'Alzheimer. Nat. Genet. 45, 1452-1458 (2013).

33. Fritsche, LG et coll. Une vaste étude d'association à l'échelle du génome sur la dégénérescence maculaire liée à l'âge met en évidence les contributions de variantes rares et courantes. Nat. Genet. 48, 134-143 (2016).
34. Satoh, JI, Kino, Y., Yanaizu, M. et Saito, Y. Maladie d'Alzheimer pathologie dans les cerveaux atteints de la maladie de Nasu-Hakola. Dis rare intraitable. Rés. 7, 32-36 (2018).
35. van der Poel, M. et coll. Le profilage transcriptionnel des microglies humaines révèle une hétérogénéité de la matière gris-blanche et des changements associés à la sclérose en plaques. Nat. Commun. 10, 1139 (2019).
36. Sala Frigerio, C. et al. Les principaux facteurs de risque de la maladie d'Alzheimer : l'âge, le sexe et les gènes modulent la réponse des microglies aux plaques Aβ. Cell Rep.27, 1293-1306 (2019).
37. Giovannoni, F. & Quintana, FJ Le rôle des astrocytes dans l'inflammation du SNC. Tendances Immunol. 41, 805-819 (2020).
38. Zamanian, JL et al. Analyse génomique de l'astrogliose réactive. J. Neurosci. 32, 6391-6410 (2012).
39. Bombeiro, AL, Hell, RC, Simões, GF, Castro, MV et Oliveira, A. L. Importance de l'expression du complexe majeur d'histocompatibilité de classe I (CMH-I) pour la réactivité astrogliale et la stabilité des circuits neuronaux in vitro. Neurosci. Lett. 647, 97-103 (2017).
40. Ransohoff, RM & Estes, ML Expression des astrocytes des histogrammes majeurs produits génétiques complexes de compatibilité dans le tissu cérébral de la sclérose en plaques obtenus par biopsie stéréotaxique. Cambre. Neurol. 48, 1244-1246 (1991).
41. Xie, L. et coll. Le sommeil entraîne la clairance des métabolites du cerveau adulte. Sciences 342, 373-377 (2013).
42. Latz, E., Xiao, TS & Stutz, A. Activation et régulation des inflammasomes. Nat. Révérend Immunol. 13, 397-411 (2013).
43. Cantuti-Castelvetri, L. et al. Une clairance défectueuse du cholestérol limite la remyélinisation dans le système nerveux central âgé. Sciences 359, 684-688 (2018).
44. Shweiki, D., Itin, A., Soffer, D. & Keshet, E. Endothélial vasculaire le facteur de croissance induit par l'hypoxie peut médier l'angiogenèse initiée par l'hypoxie. Nature 359, 843-845 (1992).
45. Zeng, ZJ et coll. TLX contrôle l'angiogenèse par interaction avec la protéine von Hippel-Lindau. Biol. Ouvert 1, 527-535 (2012).
46. Wang, GL, Jiang, BH, Rue, EA & Semenza, GL Le facteur 1 inductible par l'hypoxie est un hétérodimère basique-hélice-boucle-hélice-PAS régulé par la tension cellulaire d'O2. Proc. Natl Acad. Sci. États-Unis 92, 5510-5514 (1995).
47. Kliffen, M., Sharma, HS, Mooy, CM, Kerkvliet, S. et de Jong, PT Expression accrue des facteurs de croissance angiogéniques dans la maculopathie liée à l'âge. Frère. J. Ophthalmol. 81, 154-162 (1997).
48. Wong, TY, Liew, G. et Mitchell, P. Mise à jour clinique : nouveaux traitements pour la dégénérescence maculaire liée à l'âge. Lancet 370, 204-206 (2007).
49. Escartin, C. et al. Nomenclature des astrocytes réactifs, définitions et orientations futures. Nat. Neurosci. 24, 312-325 (2021).
50. Guttenplan, KA et al. Les astrocytes réactifs neurotoxiques entraînent la mort neuronale après une lésion rétinienne. Cell Rep.31, 107776 (2020).
51. Liddelow, SA et al. Les astrocytes réactifs neurotoxiques sont induits par les microglies activées. Nature 541, 481-487 (2017).
52. Efremova, M., Vento-Tormo, M., Teichmann, SA et Vento-Tormo, R. CellPhoneDB : déduire la communication cellule-cellule à partir de l'expression combinée de complexes ligand-récepteur multi-sous-unités. Nat. Protocoles 15, 1484-1506 (2020).
53. Krishnaswamy, S. et coll. Analyse conditionnelle basée sur la densité de la signalisation des cellules T dans les données unicellulaires. Sciences 346, 1250689-1250689 (2014).
54. Zhao, le niveau de métal.interleukine-1β est augmenté dans le corps vitré des patients atteints de dégénérescence maculaire néovasculaire liée à l'âge (DMLA) et de vasculopathie choroïdienne polypoïdale (PCV). PLoS ONE 10, e0125150 (2015).
55. Heneka, MT, McManus, RM & Latz, E. Signalisation de l'inflammation dans la fonction cérébrale et les maladies neurodégénératives. Nat. Révérend Neurosci. 19, 610-621 (2018).
56. Guillonneau, X. et al. Sur les phagocytes et la dégénérescence maculaire. Programme. Rétine. Résolution des yeux. 61, 98-128 (2017).
57. Nagineni, CN, Kommineni, VK, William, A., Detrick, B. et Hooks, J. J. Régulation de l'expression du VEGF dans les cellules rétiniennes humaines par les cytokines : implications sur le rôle de l'inflammation dans la dégénérescence maculaire liée à l'âge. J. Cell. Physiol. 227, 116-126 (2012).
58. Moon, KR et coll. De nombreuses méthodes d'analyse basées sur l'apprentissage données de séquençage d'ARN à cellule unique.Curr.Opin.Syst.Biol.7,36-46 (2018).
59. Coifman, RR & Lafon, S. Cartes de diffusion. Appl. Calculer. Harmonie. Anal. 21, 5–30 (2006).
60. Van Der Maaten, L., Postma, E. & Van den Herik, J. Réduction de dimensionnalité : un comparatif. J. Mach. Apprenez la résolution. 10, 66-71 (2009).
61. Izenman, AJ Introduction à l'apprentissage multiple. Wiley interdisciplinaire. Rév. Calcul. Stat. 4, 439-446 (2012).
62. Lindenbaum, O., Stanley, J., Wolf, G. et Krishnaswamy, S. dans Avancées des systèmes de traitement de l'information neuronale, 1400-1411 (MIT Press, 2018).
63. Gama, F., Ribeiro, A. et Bruna, J. Transformations par diffusion par diffusion sur des graphiques. Dans Conférence internationale sur les représentations d'apprentissage (ICLR, 2019).
64. Gao, F., Wolf, G. & Hirn, M. Diffusion géométrique pour l'analyse de données graphiques. À paraître dans les actes de la 36e Conférence internationale sur l'apprentissage automatique (PMLR, 2019).
65. Moon, KR et coll. Visualiser la structure et les transitions en haute résolution données biologiques dimensionnelles. Nat. Biotechnologie. 37, 1482-1492 (2019).
66. Gigante, S. et coll. Diffusion compressée. En 2019, 13e conférence internationale sur la théorie et les applications de l'échantillonnage (SampTA) (IEEE, 2019).
67. Batson, J., Royer, L. & Webber, J. Validation croisée moléculaire pour ARN-seq unicellulaire. https://www.biorxiv.org/content/early/2019/09/30/786269 . https://www.biorxiv.org/content/early/2019/09/ 30/786269.full.pdf bioRxiv (2019).
68. Chen, C. & Edelsbrunner, H. La diffusion manque rapidement de persistance. Dans Actes de la Conférence internationale IEEE sur la vision par ordinateur (ICCV) 423-430 (Curran Associates, Inc., Red Hook, NY, États-Unis, 2011).
69. Ghrist, R. Barcodes : La topologie persistante des données. Taureau. Suis. Mathématiques. Soc. 45, 61-75 (2008).
70. Rieck, B., Sadlo, F. et Leitte, H. dans Méthodes topologiques dans les données Analyse et visualisation. (éd. Carr, H., Fujishiro, I., Sadlo, F. et Takahashi, S.) 87-101 (Springer, Cham, Suisse, 2020).
71. O'Bray, L., Rieck, B. & Borgwardt, K. Courbes de filtration pour le graphique représentation. Dans Actes de la 27e Conférence internationale ACM SIGKDD sur la découverte des connaissances et l'exploration de données (KDD).1267-1275 (Association for Computing Machinery, New York, NY, États-Unis, 2021).
72. Dann, E., Henderson, NC, Teichmann, SA, Morgan, MD et Marioni, JC Tests d'abondance différentielle sur des données unicellulaires à l'aide de graphiques des k voisins les plus proches. Nat. Biotechnologie. https://doi.org/10.1038/s41587-021-01033-z _ (2021).
73. Nabavi, S., Schmolze, D., Maitituoheti, M., Malladi, S. et Beck, AH EMDomics : une méthode robuste et puissante pour l'identification de gènes différentiellement exprimés entre classes hétérogènes. Bioinformatique 32, 533-541 (2015).
74. Wang, T. & Nabavi, S. Analyse différentielle de l'expression génique dans les données de séquençage d'ARN unicellulaire. En 2017, Conférence internationale de l'IEEE sur la bioinformatique et la biomédecine (BIBM) 202-207 (IEEE, 2017).
75. Orlova, DY et coll. Earth Mover's Distance (EMD) : une véritable mesure pour comparer les niveaux d'expression de biomarqueurs dans les populations cellulaires. PLoS ONE 11, e0151859 (2016).
76. Backurs, A., Dong, Y., Indyk, P., Razenshteyn, I. et Wagner, T. Recherche évolutive du voisin le plus proche pour un transport optimal. https://arxiv. org/abs/1910.04126 (2020).
77. Indyk, P. & Thaper, N. Récupération rapide d'images via des intégrations. Lors du 3e atelier international sur les théories statistiques et computationnelles de la vision (IEEE Computer Society Press, 2003).

78. Le, T., Yamada, M., Fukumizu, K. et Cuturi, M. dans Advances in neural information Processing Systems, 12304-12315 (Neural Information Processing Systems Foundation, 2019).
79. Peyré, G. & Cuturi, M. Transport optimal informatique. https://arxiv.org/abs/1803.00567 (2019).
80. Gonzalez, TF Clustering pour minimiser la distance intercluster maximale. Théorique. Calculer. Sci. 38, 293-306 (1985).
81. Storey, JD Une approche directe des taux de fausses découvertes. JR Stat. Soc. Ser. B (Stat. Methodol.) 64, 479-498 (2002).
82. Benjamini, Y. & Hochberg, Y. Contrôler le taux de fausses découvertes : une approche pratique et puissante des tests multiples. JR Stat. Soc. Ser. B (méthodologique) 57, 289-300 (1995).
83. Korsunsky, I. et al. Intégration rapide, sensible et précise des données unicellulaires avec harmonie. Nat. Méthodes 16, 1289-1296 (2019).
84. Ramilowski, JA et al. Un projet de réseau de signalisation multicellulaire médiée par ligand-récepteur chez l'homme. Nat. Commun. 6, 7866 (2015).

Remerciements

Nous tenons à remercier les donneurs de rétine et leurs familles pour leur contribution à ce travail. Sans leur sacrifice, notre étude n'aurait pas été possible. L'HBP reçoit un financement de recherche de NEI K08-EY026652, NEI R01-EY034234, de la Thome Memorial Foundation, de la Doris Duke Charitable Foundation, de la H. Eric Cushing Foundation, de la Nancy Lurie Marks Family Foundation, de la CJL Charitable Foundation, du Reynold and Michiko Spector. Prix en neurosciences et Hoffmann-La Roche Pharmaceuticals. MK reçoit un soutien à la recherche grâce à la subvention de formation NIAID 1F30-AI157270. MD reçoit le soutien à la recherche de la subvention de formation NCI K12CA215110 et du Robert E. Leet et Clara Guthrie Patterson Trust. SK reçoit le soutien à la recherche du NIAID 5U19-AI089992-08. SK et GW reçoivent le soutien de recherche du NIGMS 1RO1-1355929. GW reçoit un financement du Canada CIFAR AI (CCAI)

Subvention à la découverte du CRSNG 03267. LZ reçoit un financement de recherche du NIA R56-AG074015 et du NIDA DP2-DA056169. AHS reçoit un financement du Bureau de recherche étudiante de la Yale School of Medicine. Nous remercions le réseau Advancing Sight et la banque d'yeux Lions Gift of Sight pour la récupération rapide des yeux des donneurs.

Contributions des auteurs Conception :

MK, MM, SK, BPH ; Conception des travaux : MK, MD, EC, SKBPH ; Acquisition de données : MD, EC, MI, LZ, MM, YX, BPH, ES, AM, GM ; Analyse des données : MK, MD, EC, AHS, RMD, BPH ; Interprétation des données : MK, MD, EC, AS, BR ; GW ; Sask.; HBP ; Création de nouveaux logiciels : MK, SG, JH, AT, AG, HS, GH, JN, KY, MH, BR, GW ; Rédaction—rédaction : MK, MD, EC, BR, BPH, SK ;

Intérêts concurrents Le Dr Krishnaswamy est membre du conseil consultatif scientifique de KovaDx et d'AI Therapeutics. Le Dr Hafler reçoit un financement de recherche de Nayan Ther-apeutics et de Hoffmann-La Roche Pharmaceutical. Le Dr Hafler est membre du conseil consultatif scientifique de Carmine Therapeutics. Tous les autres auteurs ne déclarent aucun intérêt concurrent.

Informations supplémentaires Informations supplémentaires La version en ligne contient du matériel supplémentaire disponible sur https://doi.org/10.1038/s41467-023-37025-7.

La correspondance et les demandes de matériel doivent être adressées à Smita Krishnaswamy ou Brian P. Hafler.

[1]Département de neurosciences, Université de Yale, New Haven, CT, États-Unis. [2]Département de pathologie, Université de Yale, New Haven, CT, États-Unis. [3]Département de Ophtalmologie et sciences visuelles, Université de Yale, New Haven, CT, États-Unis. [4]Département de neurologie, Université de Yale, New Haven, CT, États-Unis. [5]École de Yale de Médecine, New Haven, CT, États-Unis. [6]Division de médecine infectieuse, immunitaire et respiratoire, Université de Manchester, Manchester, Royaume-Uni. [7]Département de Informatique, Université de Yale, New Haven, CT, États-Unis. [8]Département de mathématiques appliquées, Université de Yale, New Haven, CT, États-Unis. [9]Biologie computationnelle, Programme de bioinformatique, Université de Yale, New Haven, CT, États-Unis. 10Département d'immunobiologie, Faculté de médecine de l'Université de Yale, New Haven, CT, États-Unis. 11Mila—Institut québécois d'IA, Montréal, QC, Canada. 12Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada. 13Département d'informatique, Rutgers University, Nouveau-Brunswick, NJ, États-Unis. 14Département de génétique, Université de Yale, New Haven, CT, États-Unis. 15Département de mathématiques computationnelles, sciences et ingénierie, Michigan State University, East Lansing, MI, États-Unis. 16Département de mathématiques, Michigan State University, East Lansing, MI, États-Unis. 17Département de science et d'ingénierie des biosystèmes, ETH Zurich, Zurich, Suisse. 18Broad Institute du MIT et Harvard, Cambridge, MA, États-Unis. 19Ces auteurs ont contribué à parts égales : Manik Kuchroo, Marcello DiStasio, Eric Song. 20Ces auteurs ont co-dirigé ces travaux : Smita Krishnaswamy, Brian P. Hafler.
courrier électronique : smita.krishnaswamy@yale.edu ; brian.hafler@yale.edu