Méthodes statistiques pour l'analyse différentielle de données RNA-seq en masse et en cellule unique appliquées en immunologie

Marine Gauthier

université de BORDEAUX

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SOCIÉTÉ, POLITIQUE,

SANTÉ PUBLIQUE

SPÉCIALITÉ SANTÉ PUBLIQUE, OPTION BIOSTATISTIQUE

Par **Marine GAUTHIER**

Méthodes statistiques pour l'analyse différentielle de données RNA-seq en masse et en cellule unique appliquées en immunologie

Sous la direction de : Rodolphe THIÉBAUT
Co-directeur : Boris P. HEJBLUM

Soutenue le 2 Décembre 2021

Membres du jury :

PICARD Franck	Directeur de recherche	CNRS Lyon	Rapporteur
NEUVIAL Pierre	Directeur de recherche	CNRS Toulouse	Rapporteur
VIALANEIX Nathalie	Directrice de recherche	INRAE Toulouse	Présidente du jury
THIÉBAUT Rodolphe	PU-PH	Université de Bordeaux	Directeur de thèse
HEJBLUM Boris P.	Chargé de recherche	INSERM Bordeaux	Co-directeur de thèse

Remerciements

À Boris. Je n'aurais pas pu avoir un autre directeur de thèse que toi. Tu as su être présent en toutes circonstances, que ce soit pour un problème de code, pour l'écriture des articles ou pour la préparation des conférences. Tu as essayé de soigner mon syndrome de l'imposteur. Tu m'as laissé le temps de trouver ma place dans ce monde impressionnant de la recherche. Tu as eu à coeur de me transmettre ton expérience et tes connaissances. Tu ne m'as jamais laissée seule face à mes appréhensions. Tu as sans cesse eu l'énergie de me rassurer jusqu'au rendu de la thèse. Tu as justement dosé l'accompagnement humain et l'accompagnement professionnel. J'ai beaucoup appris à tes côtés. Durant ces 3 années et demie, j'ai rencontré un chercheur attentionné, pédagogue et d'une rare gentillesse. Ta bienveillance naturelle et ton optimisme à toute épreuve ont compensé mes inquiétudes. Je te remercie pour ton temps, pour ta patience et pour m'avoir soutenue. Mes remises en question ne t'ont jamais découragé et j'espère sincèrement avoir été à la hauteur de ta confiance. Je me souviens encore d'une doctorante qui me disait que le plus important dans une thèse c'était le sujet. L'un des premiers enseignements de cette thèse est qu'elle a tort. Le plus important c'est le directeur de thèse. Le plus important c'est toi.

À Rodolphe. Je retiendrai ton dynamisme et ton cerveau qui fonctionne à 100 à l'heure. Malgré ton emploi du temps de ministre, tu as pu te rendre pleinement disponible lors de nos rendez-vous. À chaque fois, j'en suis ressortie motivée et déterminée. Tu as également su valoriser mes idées et me donner une vision sur la portée de mes travaux. Tes multiples compétences en médecine et en statistiques ont apporté toute la profondeur et l'originalité dont l'équipe SISTM bénéficie aujourd'hui et je suis honorée d'avoir pu en faire partie.

To Denis. Thank you for your availability on the other side of the world, for helping me more than necessary in my wandering in front of the asymptotic test, for encouraging me, for bringing your brilliant ideas to my work which would not be what it is without you and for allowing me to progress in English. I was very lucky

to be able to collaborate with such a talented researcher and a kind person.

À tous les membres de l'équipe SISTM. Merci pour votre accueil chaleureux, votre gentillesse et votre bienveillance. Merci d'avoir toujours eu un petit mot gentil lors de nos rencontres. Je pense en particulier à Solenne, Mélany et Laura avec qui j'ai eu de multiples échanges sur les données d'expression génique.

À Sandrine et Audrey. Merci d'avoir géré mes problèmes administratifs d'une main de maître et d'avoir défendu ma cause pour ces fameux billets d'avion.

À Benjamin. Tu as été d'abord mon élève, puis mon stagiaire et maintenant tu es doctorant dans l'équipe. Je suis sincèrement heureuse d'avoir croisé ton chemin. Malgré toi, tu m'as apporté la légitimité et la confiance dont j'avais besoin à un moment creux de ma thèse. Je me suis beaucoup reconnue en toi et j'espère avoir pu te donner les clés pour affronter ton doctorat.

Aux membres du VRI. Merci de m'avoir donné l'opportunité d'être au coeur des recherches sur le COVID-19 et d'avoir écouté avec beaucoup de curiosité les prémisses de ma dernière méthode.

Aux membres du jury. Merci d'avoir accepté d'évaluer ma thèse. Franck Picard, merci d'avoir accepté de rapporter cette thèse et de m'avoir conviée à votre groupe de travail durant lequel j'ai pu présenter dans une ambiance détendue et bienveillante. Cela a été un réel tournant dans l'approche de ma thèse. Pierre Neuvial, merci d'avoir accepté d'être rapporteur et d'avoir participé activement depuis le début à mes comités. Vos retours et vos connaissances ont été d'une grande aide. Marie-Agnès Dillies, merci d'avoir pris part à mon comité en cours de route et d'avoir apporté votre gentillesse ainsi que votre expertise. Nathalie Vialaneix, merci de votre présence. Votre intérêt pour mes recherches lors d'une conférence a renforcé ma confiance.

À Mamy Odette. Tu es une Mamy incroyable. Depuis toute petite, tu t'es occupée de moi et tu sais que notre complicité est chère à mon coeur. Mes séjours dans ta maison de campagne ont toujours été ressourçants et apaisants. Je me souviens du

premier confinement chez toi, un moment hors du temps, au calme, pendant lequel tu as pris soin de moi. Tu connais les hauts et les bas par lesquels je suis passée. Tu as su me réconforter et me redonner confiance avec ta sagesse de Mamy. Je ne te remercierai jamais assez... Dans quelques semaines, nous allons réaliser ton rêve d'aller en Égypte et nous créer de nouveaux souvenirs.

À Maman. Merci de m'avoir toujours encouragée peu importe la voie que je voulais prendre, de mes tergiversations en première année de fac jusqu'à l'après-thèse. Tu as eu confiance en moi et tu m'as soutenue dans mes choix. Cela a énormément compté pour moi.

À Papa. Partager quelques années à l'Université avec toi a été une expérience pour le moins atypique. J'ai hérité de toi la persévérance mais aussi l'entêtement. Tes mots rassurants et motivants m'ont portée pendant toute cette thèse. Je suis très heureuse qu'on ait pu accomplir nos projets de vie en quasi synchronisation. À Gaëlle, merci pour tes encouragements et ta bonne humeur.

À Papy Jean-Pierre. Tu as eu à coeur de suivre mon avancement de doctorante même si je te parlais d'un monde qui t'était inconnu. Tu n'as jamais douté de moi et je t'en remercie.

A Laurent et Murielle. Merci pour toutes ces années durant lesquelles vous m'avez accueillie à bras ouverts dans ce cocon familial d'une générosité sans pareille.

À Emma et Mathieu. Vous avez joué vos rôles de psychologue et philosophe à la perfection. Partager cette aventure avec vous a été d'une richesse incomparable. Vivement que nous puissions fêter avec Louis, nos 4 titres ! Je pense à toutes nos soirées où nous ne nous pouvions plus nous quitter, trop occupés à réinventer la recherche, la France et le monde. Mathieu, tu en as fait du chemin depuis ta première année. Je suis certaine que peu importe la voie que tu choisiras, tu y brilleras. Emma, j'aurais aimé avoir une amie comme toi plus tôt dans ma vie. Ton écoute et tes conseils m'ont été d'une aide que tu ne peux imaginer. Notre amitié est devenue essentielle. Ton esprit cortiqué te mènera loin et je serai toujours présente pour t'épauler.

À Louis, mon amour de toujours. Voilà plus de 13 ans que tu partages ma vie et que je ne m'imagine pas sans toi. Ton amour est ce que j'ai de plus précieux. Cette thèse, c'est aussi la tienne. Je ne compte plus le nombre de fois que tu m'as aidée, écoutée et rassurée. Tu as supporté mes joies et mes peines. Ton esprit brillant, ton réalisme avec ton éternel flegme, et ta volonté, n'ont cessé de m'impressioner durant toutes ces années. Tu m'apportes au quotidien une force infinie. Tu m'as tracé le chemin avec ta thèse et je suis fière d'avoir été témoin de ta réussite. Vivre nos doctorats ensemble a été une expérience unique, j'ai à présent hâte que nous puissons partager ce titre de Docteur et partir main dans la main pour de nouvelles aventures. J'ai une tendre pensée pour notre grand chat, Athéna, qui est avec nous depuis le début de nos études universitaires et qui est d'une certaine façon le symbole de cette traversée, et pour notre petit Minos, qui a amené ce brin de folie à cette fin de thèse.

Valorisations scientifiques et enseignements

Publications scientifiques

- **Gauthier Marine**, Agniel Denis, Thiébaut Rodolphe, Hejblum Boris P. dearseq : a variance component score test for RNA-Seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics*. 2(4), 2020.
https://doi:10.1093/nargab/lqaa093

- Yves Lévy, Aurélie Wiedemann, Boris P. Hejblum, Mélany Durand, Cécile Lefebvre, Mathieu Surénaud, Christine Lacabaratz, Matthieu Perreau, Emile Foucat, Marie Déchenaud, Pascaline Tisserand, Fabiola Blengio, Benjamin Hivert, **Marine Gauthier**, Minerva Cervantes-Gonzalez, Delphine Bachelet, Cédric Laouénan, Lila Bouadma, Jean François Timsit, Yazdan Yazdanpana, Giuseppe Pantaleo, Hakim Hocini, Rodolphe Thiébaut. CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. *iScience*. 24(7) :102711 (2021).
https://doi.org/10.1016/j.isci.2021.102711

- **Gauthier Marine**, Agniel Denis, Thiébaut Rodolphe, Hejblum Boris P. Distribution-free complex hypothesis testing for single-cell RNA-seq dif-

ferential expression analysis. *Submitted.*

Communications orales

- **Gauthier M**, Agniel D, Thiébaut R, Hejblum BP. Conditional cumulative distribution function estimation for differential gene expression analysis in single-cell RNA-seq data. *StatOmique*, Paris, France, Novembre 2019.

- **Gauthier M**, Agniel D, Thiébaut R, Hejblum BP. Test d'hypothèse complexe appliqué à l'analyse de l'expression différentielle pour données RNA-seq en cellule unique. *51èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Nice, France, Mai 2020.

- **Gauthier M**, Agniel D, Thiébaut R, Hejblum BP. Conditional cumulative distribution function estimation for differential gene expression analysis in single-cell RNA-seq data. *ISCB*, Krakovie, Pologne, Août 2020.

- Gauthier M, Agniel D, Thiébaut R, **Hejblum BP**. Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis. *Statistical Methods for Post-Genomic Data 2021*, France, Janvier 2021.

- **Gauthier M**, Agniel D, Thiébaut R, Hejblum BP. Test d'hypothèse complexe appliqué à l'analyse de l'expression différentielle pour données RNA-seq en cellule unique. *SingleStatomics*, France, Mars 2021.

- **Gauthier M**, Agniel D, Thiébaut R, Hejblum BP. Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis. *8th Channel Network Conference*, Paris, France, Avril 2021.

- **Gauthier M**, Agniel D, Thiébaut R, Hejblum BP. Test d'hypothèse complexe appliqué à l'analyse de l'expression différentielle pour données RNA-seq en cellule unique. *52èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Nice, France, Juin 2021.

Paquets R

- `dearseq` : un paquet R pour l'analyse de données *bulk* RNA-seq. Disponible sur Bioconductor, version de développement sur GitHub (`github.com/borishejblum/dearseq`).

- `ccdf` : un paquet R pour test d'hypothèses complexes adapaté à l'analyse différentielle de données *single-cell* RNA-seq. Disponible sur le CRAN, version de développement sur GitHub (`github.com/Mgauth/ccdf`).

Enseignements

Master 2 Mathématiques appliquées et Statistiques, Parcours Modélisation Statistique et Stochastique, Université de Bordeaux, France :

- 2019-2020 : **Statistiques en grande dimension**, TD, 10h
- 2020-2021 : **Statistiques en grande dimension**, TD, 10h
- 2020-2021 : **Chaînes de Markov**, TD, 14h

Co-encadrement

Benjamin Hivert, stage de 6 mois, 24 février 2020 - 21 août 2020. « Réduction de dimension pour la visualisation de données RNA-seq en cellule unique appliquée en immunologie »

Table des matières

Chapitre 1

Introduction

L'expression des gènes est le processus par lequel l'information inscrite dans un gène est utilisée pour la production d'une protéine. A partir d'une simple prise de sang, nous sommes aujourd'hui capables de mesurer cette expression génique dans un échantillon biologique à l'échelle du génome entier. Ce type de données permet d'approfondir notre compréhension des mécanismes biologiques dans différents contextes, et notamment concernant la réponse immunitaire dans de nombreuses études vaccinales [Querec et al., 2009; Furman et al., 2014; Hejblum et al., 2015]. Par exemple, il a été montré que les modifications de l'expression génique des sujets vaccinés précèdent de plusieurs jours, voire semaines, d'autres biomarqueurs de l'immunité induite [Rechtien et al., 2017] et ont donc un fort potentiel de prédiction et de compréhension de la réponse au vaccin. Il est alors question de proposer une médecine personnalisée, c'est-à-dire avec des protocoles de soins adaptés à chaque patient, grâce au séquençage et l'analyse de son ADN [De Lecea and Rossbach, 2012; Frese et al., 2013; Chen and Snyder, 2013; Medaglini et al., 2018; Cotugno et al., 2019].

Les mécanismes d'expression des gènes sont plus que jamais au coeur des recherches scientifiques du XXI*ème* siècle [Scoazec, 2006; Geall et al., 2013; Costa et al., 2013; Saliba et al., 2014; Wickramasinghe et al., 2014].

1.1 L'expression génique et sa mesure

1.1.1 L'expression génique

Le génome est l'ensemble du matériel génétique d'un être vivant. Il est contenu dans le noyau des cellules qui sont toutes composées de ce même génome. Ce dernier comprend l'ensemble des gènes codant la synthèse des protéines. Le gène est considéré comme l'unité moléculaire de base de l'hérédité [Crick, 1958, 1970] et constitue un fragment d'ADN (Acide DésoxyriboNucléique). L'ADN est une molécule à double brin en forme de double hélice, composée de quatre nucléotides de base : adénine (A), cytosine (C), guanine (G) et thymine (T). Un projet de recherche international appelé « Projet Génome Humain » [DeLisi, 1988] initié en 1988 et achevé en 2003, qui avait pour but de déterminer le séquençage complet de l'ADN du génome humain et d'identifier les gènes qu'il contient, a estimé que les humains ont entre 20 000 et 25 000 gènes.

La production de protéines est contrôlée par les gènes via la transcription et la traduction. C'est le "dogme central" de la biologie moléculaire, représenté figure 1-1. La transcription fait référence au processus pendant lequel un fragment d'ADN codant le gène est copié en ARN messager par l'intermédiaire de l'ARN polymérase, une enzyme. C'est la première étape de l'expression d'un gène. L'ARN messager (ARNm) est une molécule à simple brin, complémentaire de la séquence de nucléotides de la molécule d'ADN où le nucléotide uracile (U), se substitue à la thymine. L'ensemble des molécules d'ARN présentes dans la cellule forme le transcriptome. La traduction quant à elle représente l'assemblage des acides aminés à partir de l'ARN messager pour former la protéine. L'immunologiste Steve Pascolo, chercheur à l'Hôpital universitaire de Zurich (Suisse), utilise l'analogie suivante pour expliquer ce qu'est l'ARN messager [Demey, 2020] : *« Le noyau de la cellule, c'est une bibliothèque. L'ADN, c'est un livre. L'ARN, c'est une photocopie de quelques pages. La spécificité d'une cellule fait que la photocopieuse sélectionne certaines pages et pas d'autres. Ainsi, toutes nos cellules ont le gène de l'insuline ; mais c'est seulement dans celles du pancréas que celui-ci est "photocopié". C'est un message qui est lu par la machinerie cellulaire pour lui instruire la fabrication d'une protéine. Une fois lue, la photocopie est détruite. L'ARNm permet*

d'exprimer aussi bien un anticorps qu'une protéine intervenant dans des cancers ou des pathologies génétiques. ».

Les cellules de tout organisme vivant sont donc régies par les instructions envoyées par le génome. Leur capacité à réguler l'expression des gènes leur permet de créer la protéine nécessaire à leur fonctionnement normal ou à leur survie à un moment donné. Quand un gène est transcrit, on dit qu'il s'exprime, c'est à dire qu'il transmet un message afin de créer la protéine associée. L'expression génique renvoie alors à l'ensemble des processus qui aboutissent à la formation d'un ARN et d'une protéine à partir d'un gène.

En observant l'ADN, nous avons accès aux gènes présents mais il est impossible de savoir lesquels sont exprimés et donc impliqués dans les nombreux phénomènes biologiques d'une cellule. Examiner les molécules qui évoluent de façon dynamique au sein de l'organisme comme les protéines serait une alternative étant donné que nous connaissons leur fonction biologique. Néanmoins, le codage des protéines est complexe (20 acides aminés contre quatre nucléotides dans l'ADN et l'ARN) et leur dégradation rapide dans la cellule rend leur quantification très délicate. L'ARN semble alors être le matériel génétique idéal à explorer. L'abondance de transcrits est considérée comme équivalente au niveau d'expression du gène et par conséquent équivalente à l'abondance des protéines crées. En effet, mesurer la quantité d'ARN messager permet d'évaluer le niveau de transcription d'un gène et ainsi, de déterminer si celui-ci est affecté par certaines conditions biologiques ou expérimentales. Il devient alors possible d'identifier les gènes qui s'expriment différentiellement, c'est-à-dire qui présentent des niveaux d'expression différents. Pour un chercheur, l'analyse de l'expression génique permet de comprendre la relation entre les profils d'expression des gènes et le phénotype (ensemble des traits observables) des organismes.

1.1.2 Le séquençage de l'ADN

$\Phi X174$ est le nom du premier être vivant (un bactériophage) dont le génome fut séquencé en 1977 par la méthode de Sanger [Sanger et al., 1977], scientifique ayant reçu le Prix Nobel de chimie en 1958 et 1980. La séquence comportait "seulement"

FIGURE 1-1 : Dogme central de la biologie moléculaire. Source : molecool.ch

5000 nucléotides pour un coût de séquençage très conséquent et de longs mois de travail. En 1998, le premier génome d'un animal (un ver de terre Caernohabditis elegans) fut publiée [C. Elegans Sequencing Consortium, 1998]. La méthode Sanger a rapidement montré ses limites quand il a été question de séquencer le génome humain [Schuster, 2008]. Depuis les biotechnologies n'ont cessé d'évoluer. La grande révolution en génomique fut amorcée par l'arrivée des technologies de séquençage à haut débit. Alors que les premiers séquençeurs historiques ne pouvaient traiter qu'environ 1000 nucléotides en l'espace de quelques jours, il est aujourd'hui possible de séquencer des milliards de nucléotides en parallèle [Zhang, Chiodini, Badr and Zhang, 2011]. En outre, c'est plusieurs milliers de brins d'ADN qui sont lus en même temps. C'est là toute la puissance de la technologie à haut débit dite RNA-*sequencing*.

D'après Stephens et al. [2015], en 2025, 100 millions à 2 milliards de génomes humains auront été séquencés. Le stockage des séquences devrait rapidement devenir problématique puisqu'il pourrait atteindre 40 exaoctets, c'est-à-dire 40 millards de millards d'octets. En comparaison, le plus gros centre de données au monde est l'Utah Data Center qui possède une capacité de stockage estimée entre 3 et 12 exaoctets [Hill, 2013]. En revanche, le coût de séquençage d'un génome complet humain a connu une baisse fulgurante, passant de 100 000$ en 2001 à environ 1 000$ 2020.

Les puces à ADN

Les puces à ADN, aussi appelées *microarrays*, sont une technique de séquençage inventée dans les années 90 et basée sur le principe d'hybridation [Southern et al., 1999], *i.e.* la propriété selon laquelle l'ADN à simple brin s'associe naturellement à

son brin complémentaire. D'une part, des séquences d'ADN synthétique sont placées sur un support : on les appelle des sondes. Ces séquences sont préalablement choisies afin de s'hybrider aux transcriptions des gènes ciblés pour l'expérience biologique. D'autre part, on extrait des ARNm de l'échantillon puis une amplification est appliquée afin d'accroître la quantité de matériel génétique. Ce sont les cibles. Par rétrotranscription, les ARNm sont convertis en brin d'ADN complémentaires (ADNc). Ils sont ensuite imprégnés d'une substance fluorescente ou radioactive. En mettant en contact les sondes avec les cibles, les nucléotides complémentaires vont s'associer. Enfin, la puce est décryptée par un scanner à très haute résolution et l'image scannée est analysée par ordinateur. Par la détection des signaux fluorescents ou radioactifs sous la forme de valeur d'intensité, on mesure les différents niveaux d'expression des gènes contenus dans la puce. Cette quantité est continue. L'atout majeur des puces à ADN est la capacité à analyser l'expression des dizaines de milliers de gènes simultanément, générant ainsi des données de grande dimension étant donné le nombre de gènes grandement supérieur au nombre d'individus ($n \ll p$).

Le RNA-seq en masse (*bulk* RNA-seq)

Le RNA-seq est une technique relativement récente et fait partie de ce qu'on appelle le « séquençage de nouvelle génération » (next-generation sequencing) ou « séquençage à haut débit » (high-throughput sequencing) ou encore « séquençage massif en parallèle ». Il constitue une réelle révolution technologique [Wang et al., 2009]. Le séquençage à ARN permet de récolter toutes les séquences d'ARN observé dans un échantillon. Pour cela il faut isoler les séquences d'ARN. Ensuite, elles sont converties en ADN complémentaire (ou cDNA) via une transcriptase-inverse. Les chaînes de nucléotides étant extrêmement longues, elles doivent être séparées en plus petits morceaux de taille identiques. On les appelle les *reads*. Le séquenceur Illumina HiSeq2000 peut traiter aujourd'hui plus de 200 paires de bases. La longueur des *reads* a un impact non-négligeable dans le séquençage : si les *reads* sont de taille trop faible, le risque de mauvais alignement par rapport au génome de référence augmente. Cette étape d'alignement s'appelle le *mapping* et consiste à rechercher dans le génome la

localisation d'une sous-séquence identique à celle du *read*. Théoriquement, la position est unique pour un *read* suffisamment long. Si un gène a été séquencé entièrement, il sera couvert par des *reads*, et ce jusqu'à plusieurs fois. Le nombre de *reads* alignés pour un échantillon est appelé la profondeur de séquençage (*sequencing depth*). C'est pour garantir une profondeur suffisante qu'on doit séquencer des centaines de millions de reads et s'assurer de leur qualité. Notons également que la profondeur a un impact sur la détection des gènes faiblement exprimés définissant alors la limite de détection. Enfin, les *reads* alignés sont dénombrés par l'intermédiaire de logiciels spécifiques. Il existe une base de référence regroupant des données génomiques publiques afin d'identifier les séquences et les gènes [Pruitt and Maglott, 2001]. Contrairement aux puces à ADN, si une séquence encore inconnue est détectée, elle peut être ajoutée à la base initiale. Un tableau de comptage est ainsi obtenu et peut être directement exploité pour des analyses comme par exemple l'analyse différentielle qui nous intéresse particulièrement dans cette thèse.

Le RNA-seq en cellule unique (*single-cell RNA-seq*)

Le RNA-seq en masse permet d'obtenir l'expression moyenne d'un gène dans un tissu potentiellement composé de centaines voire milliers de cellules. Avec le RNA-seq en cellule unique, il est devenu possible de réaliser un séquençage de l'ARN de chaque cellule (*single-cell* RNA-seq) issue d'une population cellulaire et d'étudier le transcriptome à la résolution de cellules individuelles. Par exemple, les analyses RNA-seq en masse peuvent être affectées par les proportions inconnues et variables des différents types de cellules présents dans un échantillon (interprétable en tant que variable confondante). Par ce séquençage novateur, nous pouvons mieux appréhender la distinction entre les populations cellulaires et découvrir des types de cellules auparavant inconnus.

Le premier article sur le scRNA-seq a été publié en 2009 [Tang et al., 2009]. Depuis, le nombre de cellules dans les expériences scRNA-seq a augmenté drastiquement [Svensson et al., 2018] avec l'usage de la microfluidique. La microfluidique est une technique impliquant la manipulation des fluides dans des microcanaux. La

plateforme de capture cellulaire la plus connue est Fluidigm C1 (voir DeLaughter [2018] pour les détails de son utilisation). Ce système utilise la microfluidique pour séparer les cellules dans des puits individuels (par exemple 96 puits) d'une plaque où elles sont lysées, transcrites en sens inverse et l'ADNc collecté est enfin amplifiées par PCR. Après cette étape, le produit est extrait de la plaque et les bibliothèques sont préparées pour le séquençage Illumina ce qui augmente considérablement le nombre de cellules pouvant être capturées à la fois.

Isoler les cellules en microfluidique (*Droplet-based cell capture*)

Par la technologie *droplet-based*, le séquençage en cellule unique est devenu encore plus rapide. Chaque cellule d'un échantillon est isolée dans une gouttelette d'huile contenant des réactifs, comme la rétrotranscriptase, et une bille appelée GEM (Gel bead in EMulsion) (voir Salomon et al. [2019] pour un guide pratique). En résumé, le but est d'obtenir un ensemble de fragments d'ADN pour le séquenceur où chaque *read* possède un *barcode* identifiant sa cellule d'origine. Ainsi, après la phase de séquençage, les *reads* pourront être associés à une unique cellule. La matrice des comptes aura en ligne les gènes et en colonne les cellules. A chaque bille GEM est associée un barcode et un *Unique Molecular Identifiers* (UMI). Le barcode joue le rôle d'identifiant et est unique à la bille GEM et à la cellule. L'UMI est une courte séquence aléatoire et unique à chaque fragment. Il enrobe la bille. Plusieurs UMI peuvent être associés à une bille. Les cellules sont ensuite lysées à l'intérieur des gouttelettes et l'ARNm est inversement transcrit pour produire de l'ADNc avec l'ensemble *barcode* et UMI. Les gouttelettes sont enfin brisées et l'ADNc est collecté pour le séquençage. A la sortie du séquençage et de l'alignement, les *reads* provenant d'une même cellule sont repérés par leur barcode.

Unique Molecular Identifiers (UMI)

Comme vu dans la sous-partie précédente, chaque bille se voit attribuer des UMI qui sert à éliminer les biais d'amplification [Mullis et al., 1986; Mullis, 1990]. En effet, les *reads* sont le résultat de la PCR pour *Polymerase Chain Reaction* qui amplifie

une petite quantité d'ADN afin d'en avoir une quantité suffisante pour le séquençage. Etant donné que nous travaillons à l'échelle d'une cellule, il est normal de ne récupérer qu'une faible quantité d'ARN. La PCR duplique plusieurs fois une même molécule afin d'intensifier le signal. Si une molécule est trop amplifiée, elle va être comptée plusieurs fois. Grâce aux UMI, elle sera détectée car le même UMI sera représenté à plusieurs reprises. Les cellules individuelles contiennent de très petites quantités d'ARN et pour obtenir suffisamment d'ADNc pour le séquençage, une étape d'amplification par PCR (*Polymerase Chain Reaction*) est nécessaire. En fonction de leur séquence nucléotidique, les différents transcrits peuvent être amplifiés à des vitesses différentes, ce qui peut fausser leurs proportions relatives dans l'échantillon. Les UMI servent à améliorer la quantification de l'expression des gènes en éliminant les doublons produits pendant l'amplification. La PCR distord les données notamment à cause de cette potentielle multiplication. Après la transcription inverse, l'amplification, le séquençage et l'alignement, la déduplication peut être effectuée en identifiant les *reads*, avec le même UMI, qui s'alignent à la même position sur le génome de référence et qui devraient donc être des duplications PCR plutôt que des copies réellement exprimées d'un transcrit.

1.2 Les données RNA-seq

1.2.1 Les données RNA-seq en masse

Cette dernière décennie, les données RNA-seq ont permis d'approfondir notre compréhension des mécanismes biologiques à l'oeuvre dans divers contextes, et en particulier concernant la réponse immunitaire dans de nombreuses études vaccinales. La vaccination est une pierre angulaire de la médecine moderne. Elle constitue l'intervention sanitaire la plus efficace dans la prévention et le contrôle de maladies infectieuses [Pulendran and Ahmed, 2011]. La mesure de l'expression génique ouvre une porte inédite sur les mécanismes moléculaires mis en jeu dans la réponse vaccinale. Une question centrale dans la recherche vaccinale est de déterminer s'il existe des prédic-

teurs de l'immunité induite. Les vaccins stimulent l'immunité du patient par le biais de réponses humorales (anticorps) et celullaires, l'intérêt est alors de détecter des signaux avant-coureurs qui peuvent être utilisés pour prédire la réponse au vaccin administré et son efficacité. La prise en compte des données RNA-seq dans les schémas expérimentaux a alors permis de décrire des signatures transcriptomiques hautement prédictives de la réponse ultérieure en anticorps quelques jours après la vaccination, par exemple contre la grippe ou la fièvre jaune [Nakaya et al., 2011; Furman et al., 2013; Li, Rouphael, Duraisingham, Romero-Steiner, Presnell, Davis, Schmidt, Johnson, Milton, Rajam et al., 2014]. De plus, l'élaboration d'un vaccin contre le virus EBOLA suite à la recrudescence des cas s'est également enrichie de l'inclusion des données RNA-seq dans les études. L'expression génique a été un outil primordial pour déterminer si les réponses immunitaires innées précoces sont corrélées à la réponse au vaccin. [Rechtien et al., 2017; Menicucci et al., 2017; Medaglini et al., 2018]. La mesure de l'expression génique est donc devenu un outil clé dans la recherche médicale. Elle permet de mieux comprendre les processus cellulaires normaux, les réponses à des traitements médicaux mais aussi les processus pathologiques tels que l'infection par des bactéries ou des virus. Par exemple, Singhania et al. [2018] ont découvert une signature de gènes qui permet de distinguer la Tuberculose dite active de l'infection dite latente et des individus sains. Dans le Chapitre 3, nous proposons une ré-analyse de leur jeu de données. Globalement, l'utilisation des données RNA-seq dans le diagnostic, l'orientation du traitement et la prédiction de l'évolution de la maladie par des méthode d'apprentissage statistique constituent un domaine prometteur pour la médécine moderne [Ching et al., 2018; Cheerla and Gevaert, 2019; Huang et al., 2020].

Modélisation des données de comptage

Les données RNA-seq sont des données de comptage se caractérisant par un faible nombre de comptes associés à une grande proportion de gènes et par une queue de distribution lourde à droite, qui s'explique par l'absence de limite supérieure des comptes. Typiquement, les données de comptage sont modélisées par une loi de Poisson. Cette dernière donne la probabilité d'obtenir k événements dans une population

et la moyenne de la distribution est toujours égale à sa variance. Elle ne peut ainsi décrire que le bruit liée aux comptes. Cependant, ce ne sera pas la seule source de variation. Dans la plupart des schémas expérimentaux, on peut être en présence de réplicats biologiques. Ces derniers désignent des échantillons distincts, par exemple issus d'individus différents mais soumis aux mêmes conditions expérimentales. Malgré des similarités entre le matériel biologique, il y aura inévitablement de nombreuses dissimilarités. Ces variations biologiques sont à prendre en compte. En effet, si les proportions d'ARNm étaient constantes entre les réplicats biologiques sous une même condition, la distribution de Poisson conviendrait, mais il y a toujours une variabilité naturelle entre les réplicats. Lors de la mesure l'expression génique dans plusieurs échantillons, la variance de ces mesures sera alors supérieure à la moyenne. On parle de surdispersion. C'est la raison pour laquelle la distribution de Poisson est insuffisante. La source de cette variabilité peut être attribuée à de nombreux facteurs biologiques mais aussi techniques, identifiables ou non [Oshlack and Wakefield, 2009; McIntyre et al., 2011]. C'est là que la distribution Binomiale Négative fait son entrée. Comme expliqué ci-dessus, pour les données RNA-seq, il est logique d'observer un paramètre de la loi de Poisson différent pour chaque échantillon i. Ce phénomène peut être modélisé en posant que le paramètre de Poisson suit une loi Gamma pour chaque gène g : $Y_{ig}|\lambda_{ig} \sim \mathcal{P}(\lambda_{ig}), \lambda_{ig} \sim Gamma(\beta, \tau)$ et donc $Y_{ig} \sim \mathcal{NB}(\tau, \frac{\beta}{\beta+1})$. La relation entre la moyenne et la variance dans les données RNA-seq entrevue ici mérite de s'y attarder un peu plus longtemps.

Relation moyenne-variance

Les gènes dont l'expression moyenne est élevée ont tendance à avoir une variance supérieure à la moyenne. Si nous nous focalisons sur les gènes dont l'expression moyenne est faible, nous observons une dispersion des valeurs de variance assez importante comparativement aux gènes avec une expression moyenne plus importante. Ce graphique met en évidence une caractéristique importante des données RNA-seq en masse et en cellule unique (en l'occurence données RNA-seq en masse dans cette figure) : la variance n'est pas la même pour toutes les observations et il existe

visiblement une relation entre la moyenne et la variance. On parle dans ce cas d'hétéroscédasticité. Pour comprendre ce que représente l'hétéroscédasticité en pratique, prenons l'exemple suivant. Au décollage d'une fusée et en se plaçant à une distance raisonnable de la zone de tir, on peut avec des outils adaptés mesurer la distance qu'elle parcourt en une seconde. L'estimation sera alors assez précise, au centimètre près. Quelques minutes plus tard, la fusée quitte l'atmosphère terrestre. La précision des mesures devient de plus en plus faible, à 100 mètres près, en raison de la distance accrue, de la distorsion atmosphérique et de divers autres facteurs. Pour de petites valeurs la variance est faible tandis que les grandes valeurs possèdent une grande variance. C'est ce même phénomène qui apparaît dans les données RNA-seq.

La relation moyenne-variance est un problème statistique qui a été soulevé et débattu lors des premières collectes de données RNA-seq en masse, d'où une littérature plus abondante pour le RNA-seq en masse. Dans un premier temps, la log-transformation a été naturellement utilisée pour contrer l'hétéroscédasticité. Néanmoins, comme décrit par Law et al. [2014], les comptes élevés en échelle logarithmique ont dorénavant des écarts-types beaucoup plus importants que les comptes faibles.

Les modèles linéaires supposent que la variance est constante et ne dépend pas de la moyenne. Cela signifie que les modèles linéaires ne fonctionnent qu'avec des données homoscédastiques. C'est pourquoi il a fallu se tourner vers les modèles linéaires généralisés qui permettent d'inclure la relation moyenne-variance comme ceux faisant appel à une hypothèse distributionnelle de Binomiale Négative.

Les méthodes `edgeR` [Robinson et al., 2010] et `DESeq2` [Love et al., 2014] utilisent la distribution binomiale négative alors que `voom` [Law et al., 2014] se base sur un modèle linéaire avec une hypothèse de normalité sur les données. Dans la partie 1.3.2 sur l'état de l'art, nous résumerons brièvement les procédures employées par ces trois méthodes pour tenir compte du lien qui existe entre la moyenne et la variance.

1.2.2 Les données RNA-seq en cellule unique

Le séquençage de l'ARN en cellule unique [Nawy, 2014] permet de changer d'échelle biologique et de discerner l'hétérogénéité du tissu [Patel et al., 2014] (voir figure

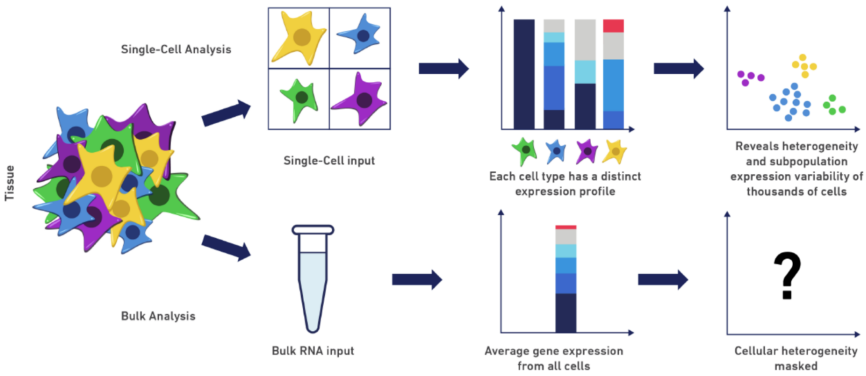


FIGURE 1-2 : Différences entre les techniques de séquençage, *community.10xgenomics.com*

1-2). L'une des principales applications est la découverte de nouvelles populations cellulaires mais aussi la caractérisation de types cellulaires rares à partir de tissus plus ou moins complexes [Shalek et al., 2014; Buettner et al., 2015; Jiang et al., 2016; Villani et al., 2017]. En particulier, la recherche en immunologie a profité de l'avénement du RNA-seq en cellule unique pour mieux comprendre l'hétérogénéité cellulaire du système immunitaire [Proserpio and Mahata, 2016; Stubbington et al., 2017; Papalexi and Satija, 2018; Neu et al., 2017; Bacher et al., 2020]. Le système immunitaire est responsable de la défense de l'organisme contre les infections diverses (*e.g.* par les virus, par les bactéries). Dans ce processus, une grande quantité de cellules se coordonnent pour activer la réponse immunitaire. Par exemple, la recherche sur le VIH s'est aussi tournée vers ce nouveau séquençage pour mettre en évidence des signatures de gènes impliqués dans certains phénomènes biologiques observables seulement à l'échelle celullaire [Golumbeanu et al., 2018; Bradley et al., 2018; Szabo et al., 2019]. Plus récemment, les scientifiques l'ont utilisé pour mieux appréhender les mécanismes d'infection du COVID-19 [Wen et al., 2020; Bost et al., 2020; Zhang et al., 2020; Xu et al., 2020; Kusnadi et al., 2021].

Dans les données RNA-seq en cellule unique, en raison du nombre peu important des molécules d'ARNm au sein d'une seule cellule (de l'ordre du picogramme) [Brennecke et al., 2013] et du seuil de capture du séquençage relativement faible, les transcrits ont tendance à ne pas être mesurés lors la transcription inverse. Par consé-

quent, certains gènes sont très exprimés dans une cellule mais ne le sont pas dans une autre. Cette absence de détection des comptes est appelée *dropout*. Pourtant, il est tout à fait possible qu'un gène ne s'exprime pas dans une cellule donnée et que l'absence d'expression traduite par un compte égal à zéro soit la réalité biologique. De plus, du fait de la variété des types cellulaires et de l'état dans lequel se trouve la cellule (le gène peut s'exprimer plus ou moins au cours du temps et ce, sans synchronisation entre les cellules), la distribution de l'expression génique pour un gène dans les cellules est généralement hétérogène et multimodale [Korthauer et al., 2016].

Les zéros dans les données RNA-seq en cellule unique

Pour donner un ordre d'idées, les données RNA-seq en cellule unique sont généralement composées à plus de 90% de zéros [Townes et al., 2019]. A titre d'exemple, le jeu de données réelles [Kusnadi et al., 2021] utilisé dans le Chapitre 4 contient 85% de comptages nuls. Les valeurs nulles de l'expression génique dans les données RNA-seq en cellule unique peuvent être soit d'origine biologique soit d'origine technique. La figure 1-3 par Jiang et al. [2020] résume les différentes étapes d'apparition des zéros. Les zéros biologiques peuvent résulter de l'absence de transcription (gène 1) ou bien de l'absence d'ARNm dûe à la dégradation de l'ARNm (gène 2). En effet, les gènes ne sont pas transcrits de manière constante mais plutôt par intermittence Hicks et al. [2018]; Suter et al. [2011]. D'ailleurs, les scientifiques ont décrit les cellules comme pouvant être dans des états actifs et des états inactifs. Combiné à la dégradation de l'ARNm, la distribution de l'expression d'un gène peut présenter une distribution avec une inflation en zéro [Paszek, 2007]. Si les transcriptions d'ARNm d'un gène dans une cellule ne sont pas converties en molécules d'ADNc alors le gène apparaîtra comme non-exprimé à tort, ce qui entraînera un zéro technique (gène 3). L'étape de la transcription inverse de l'ARNm en ADNc voit son efficacité varier considérablement selon le protocole utilisé [Bustin et al., 2015], ce qui en fait une étape critique dans la génération de zéros. Les zéros d'échantillonnage sont renvoyés dans deux cas de figure : les ADNc ne sont pas amplifiés par PCR générant peu de copies (gène 4) ou le nombre de molécules d'ARNm d'un gène est trop faible ce qui se répercute par

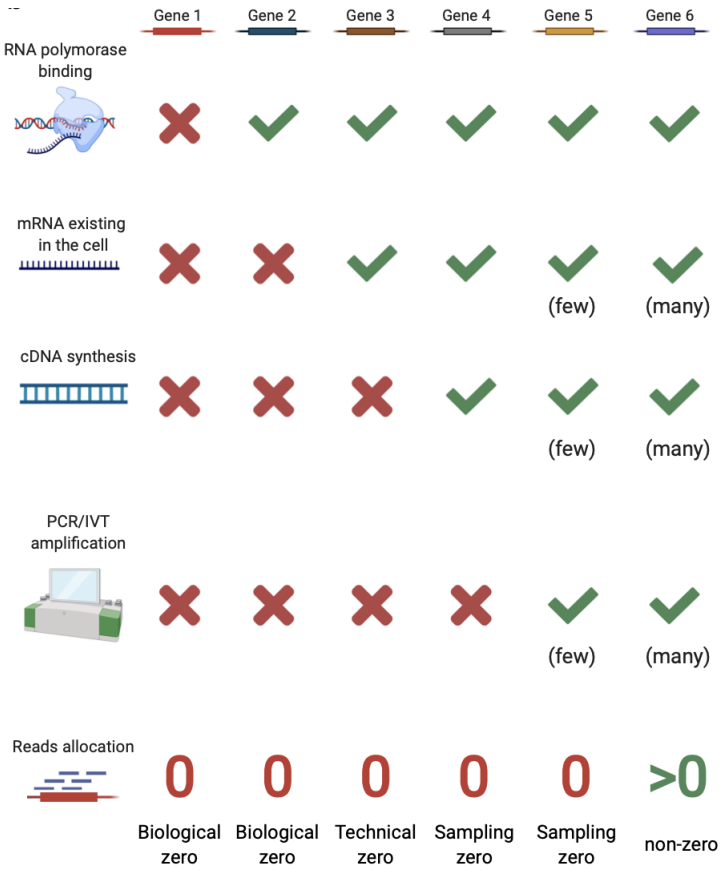


FIGURE 1-3 : Origines des comptes nuls dans les données RNA-seq en cellule unique. Les croix rouges indiquent les occurrences des zéros tandis que les coches vertes indiquent le contraire. [Jiang et al., 2020]

un faible nombre d'ADNc même après amplification (gène 5). Si un gène possède des ADNc dans la bibliothèque de séquençage, mais que ses ADNc sont en trop faible quantité pour être détectés lors du séquençage, le zéro renvoyé sera un zéro d'échantillonnage, directement lié à la contrainte sur le nombre total de *reads* séquencés. En outre, pendant le séquençage, les ADNc sont capturés de manière aléatoire, donc, un gène comportant moins d'ADNc a plus de chances de ne pas être détecté en raison de cet échantillonnage aléatoire. Enfin, si le nombre d'ADNc est suffisant pour un gène, le gène aura une mesure d'expression non nulle (gène 6).

De ce fait, les modèles enflés en zéro ont été largement utilisés [Finak et al., 2015; Kharchenko et al., 2014; Risso et al., 2018]. Ces modèles probabilistes sont composés d'une masse en zéro et d'une distribution pour les valeurs positives. La

loi de Poisson ou la loi binomiale négative sont employées pour les comptes type UMI et la loi normale pour les comptes transformés/normalisés (valeurs continues). Récemment, en se basant sur un jeu de données de contrôle négatif, Svensson [2020] a montré que le nombre de zéros dans les UMI est conforme à ce qui est attendu pour une distribution binomiale négative. Ceci indique donc que le grand nombre de zéros est plutôt dû à une variation biologique plutôt qu'à des effets techniques. L'absence d'inflation en zéro dans les UMI a également été décrite dans Townes et al. [2019]. Ces derniers illustrent ques les comptes seraient bien enflés en zéro et aurait une distribution multimodale alors que les UMI suivraient une distribution discrète enflée en zéro. On ne doit donc pas traiter les comptes comme les UMI puisque leurs caractéristiques mathématiques diffèrent. Townes et al. [2019] suggèrent de modéliser les UMI par une distribution multinomiale.

Jiang et al. [2020] et Sarkar and Stephens [2021] soulignent qu'il existe une grande ambiguïté dans l'utilisation de termes tels que *dropouts*, excès de zéros, inflation en zéro pour décrire la grande quantité de zéros dans les données RNA-seq en cellule unique. *Dropouts* est le terme le plus largement utilisé concernant les données scRNA-seq [Kharchenko et al., 2014] et décrit une expression génique nulle. Son utilisation dans les articles est devenue discordante. Certains font référence aux zéros non biologiques, d'autres à tous les zéros . *Dropouts* peut aussi désigner la présence de "nombreux" zéros. Jiang et al. [2020] ajoutent que les termes "excès de zéros" est aussi trompeur. Ils peuvent renvoyer à la plus grande proportion de zéros dans les données RNA-seq en cellule unique comparativement aux données RNA-seq en masse, ou bien aux zéros non-biologiques ou encore aux zéros supplémentaires que ne peut pas modéliser une loi binomiale négative. Il est donc préférable de parler d'inflation en zéro qui est un concept statistique qui traduit la proportion de zéros qui ne peut être modélisée par la distribution choisie (comme celle de la distribution de Poisson ou la distribution Binomiale négative).

Distribution complexe

Après avoir mis en évidence la multimodalité des données RNA-seq en cellule

unique (on peut observer jusqu'à plus de trois modes), Korthauer et al. [2016] (et repris par Wang et al. [2019]) décrivent quatre types distincts de différences entre deux conditions dans l'expression génique mesurés par RNA-seq en cellule unique :

- *differential expression* (DE) : deux distributions unimodales avec une moyenne différente dans chaque condition.

- *differential proportion* (DP) : deux distributions bimodales avec des moyennes dans les modes identiques dans les deux conditions ; la proportion dans le mode le plus faible est de 0,3 pour la condition 1 et de 0,7 pour la condition 2.

- *differential modality* (DM) : une distribution unimodale pour la condition 1 et une distribution bimodale pour la condition 2 avec un mode coïncidant avec la distribution unimodale. Pour la condition 2, la moitié des observations appartient à chaque mode.

- *both differential modality and different component means within each condition* (DB) : une distribution unimodale pour la condition 1 et une distribution bimodale pour la condition 2. Les distributions n'ont pas de composantes qui coïncident. La moyenne de la condition 1 est à mi-chemin entre les moyennes de chaque mode de la condition 2. La moitié des observations appartient à chaque mode pour la condition 2.

Korthauer et al. [2016] mettent l'accent sur la nuance entre la différence DM et la différence DP. La première suggère la présence d'un type de cellule distinct dans une condition, mais pas dans l'autre alors que que le second suggère des réponses cellulaires spécifiques [Tay et al., 2010]. La multimodalité des distributions ainsi que les multiples différences en distribution décrites précédemment suggèrent que la modélisation de l'expression génique doit être plus versatile et que les méthodes d'analyse différentielle (voir section 1.3) basées seulement sur la différence en moyenne ne sont pas adéquates. Cela appelle donc à la création d'outils plus flexibles dans la découverte de différences en distribution. Dans le Chapitre 4, nous nous efforçons d'élargir cette différence en

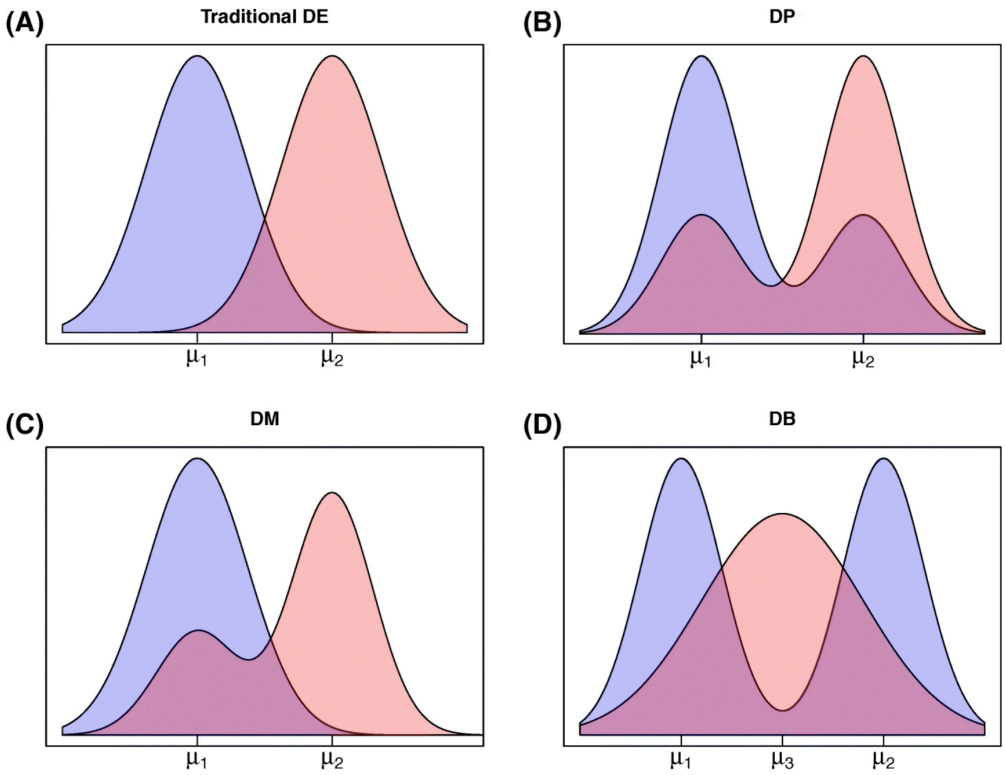


FIGURE 1-4 : Schéma des différents types de différence en distribution dans les données RNA-seq en cellule unique. (A) différence en moyenne (DE), (B) différence en proportions (DP), (C) différence en modalités (DM), (D) différence en modalités et en moyenne des composantes (DB). Source : Korthauer et al. [2016]

distribution entre un nombre arbitraire de conditions, en particulier plus de deux conditions.

1.2.3 Normalisation

Les données d'expression génique peuvent présenter des différences inhérentes aux conditions expérimentales dans lesquelles elles ont été générées et qui sont en général incontrôlables [Li, Łabaj, Zumbo, Sykacek, Shi, Shi, Phan, Wu, Wang, Wang, Thierry-Mieg, Thierry-Mieg, Kreil and Mason, 2014]. En effet, les technologies de séquençage reprennent involontairement des biais expérimentaux dans les données de séquençage de l'ARN. Les données brutes doivent alors être ajustées afin que les comparaisons entre les échantillons soient comparables pour la variabilité biologique. Ce recalibrage s'appelle normalisation. C'est une étape primordiale au traitement des données issues

de technologies à haut débit telles que le RNA-seq. Bullard et al. [2010] considèrent même que la normalisation est encore plus décisive que le choix de la statistique de test dans les tests d'hypothèse pour l'expression différentielle.

Nous pouvons citer comme deux sources de biais technique : la longueur des transcripts et la profondeur de séquençage. En effet, plus un gène est long, plus il aura de *reads* alignés (figure 1-5) et plus le séquençage est profond, plus il y aura de *reads* alignés sur chaque gène (figure 1-6). D'une part, la longueur d'un gène est un effet intra-échantillon, c'est-à-dire qu'elle affecte la comparaison du nombre de *reads* entre les différents gènes au sein d'un échantillon. D'autre part, la profondeur de séquençage est un effet inter-échantillon qui modifie la comparaison des *reads* entre le même gène dans différents échantillons. La profondeur de séquençage, aussi appelée taille de librairie, correspond au nombre total de *reads* séquencés pour un échantillon donné. Si l'échantillon B a été séquencé avec une profondeur deux fois supérieure à l'échantillon A alors l'échantillon B présentera deux fois plus de *reads*. Un gène non-DE entre les deux échantillons apparaîtra comme DE sans normalisation puisqu'il aura deux fois plus de *reads* alignés dans l'échantillon B. La variabilité du nombre total de molécules séquencées peut conduire à un nombre total de *reads* différents dans les échantillons. L'objectif de la normalisation est de faire en sorte que les différences entre les *reads* normalisés représentent de réelles différences biologique dans l'expression génique. En termes d'expression différentielle, les gènes non-DE devraient en moyenne avoir les mêmes *reads* normalisés dans toutes les conditions, tandis que les gènes DE devraient avoir des *reads* normalisés dont les différences entre les conditions représentent les véritables différences en ARNm.

Evans et al. [2018] décrivent la nécessité de normaliser les données RNA-seq brutes en raison des différentes proportions d'ARNm dans les cellules. Nous reprenons ici leur illustration en figure 1-7 et leurs explications qui mettent en lumière la nécessité de normaliser. Dans un échantillon donné, le nombre de molécules et par suite le nombre de *reads* associé à un gène est lié à la part de ce gène dans la population de molécules séquencées. Si des gènes sont fortement exprimés dans une seule des conditions, ils représenteront une plus grande proportion du total des molécules et

des *reads* comparativement aux autres (car la profondeur de séquençage est fixe et limitée). La figure 1-7 illustre ce cas de figure. Trois gènes sont mesurés pour deux conditions. Le gène 3 présente une quantité d'ARNm plus grande dans la condition B que dans la condition A. Le gène 1 et le gène 2 ont la même quantité d'ARNm dans les deux conditions (A). Le gène 3 étant fortement exprimé, il occupe la plus grande part de ARNm et affaiblit celles des gènes 1 et 2 (B). Cela se répercute sur la proportion des *reads* alignés (C) : le gène 3 se voit attribuer un grand nombre de *reads* alignés dans la condition B qui est également plus élevé que dans la condition A, ce qui est attendu. Cependant, puisque le gène 3 provoque une réduction du nombre de *reads* des gènes 1 et 2 dans la condition B, ces derniers se retrouvent sous-exprimés par rapport à la condition A. C'est ce qui apparait en (D), sans normalisation. Etant donné que le nombre total de *reads* est le même dans chaque condition, si une normalisation qui consiste à diviser par la taille de la librairie est appliquée, tous les gènes seront DE. En revanche, une normalisation appropriée, comme nous le verrons ci-après, équilibre le nombre de *reads* pour les gènes 1 et 2 et les rend non-DE. De plus, l'expression du gène 3 est corrigée et rend bien compte de l'expression réelle. On définit le *fold-change* (FC) comme le ratio du nombre de *reads* dans la condition A1 sur le nombre de *reads* dans la condition B pour chaque gène. Sans normalisation et avec la normalisation par la taille de librairie, les FC sont erronés dépassant 1 pour les gènes 1 et 2 et dépasse 0.5 pour le gène 3. Lorsque la bonne normalisation est appliquée, les deux premiers gènes ne sont pas DE et le gène 3, seul gène DE en réalité, obtient bien un FC de 0.5 (il y a deux fois plus de *reads* dans la condition B que la condition A ce qui est représentatif de la quantité d'ARNm/cellule).

Etat de l'art des méthodes de normalisation

Bien que le type de normalisation peut avoir un grand impact sur les analyses qui suivent, il n'existe pas de consensus quant à la meilleure méthode de normalisation à utiliser. De nombreuses approches ont vu le jour dans la littérature que l'on nommera par leurs termes anglais. Les principales sont les suivantes : *Total Count* (TC), *Upper Quartile* (UQ) [Bullard et al., 2010], *Median* (Med), la normalisation du

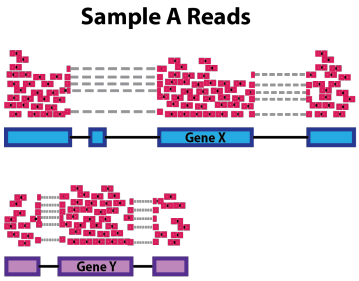


FIGURE 1-5 : L'expression de chaque gène apparaît comme deux plus importante dans l'échantillon A par rapport à l'échantillon B, mais cela est dû au fait que l'échantillon A ait une profondeur de séquençage deux fois plus élevée. [Harvard Chan Bioinformatics Core]

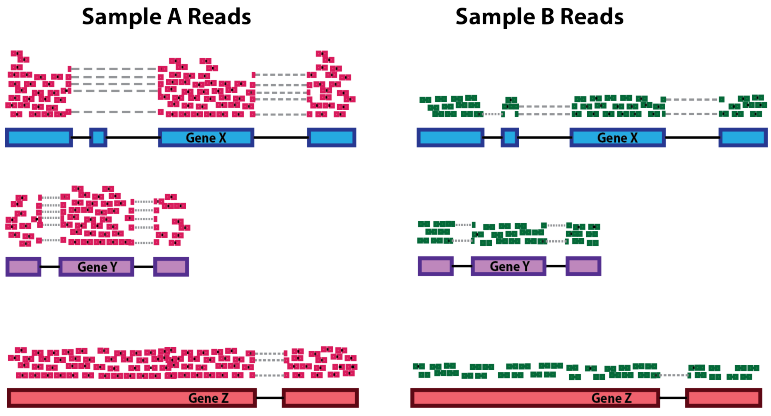


FIGURE 1-6 : Le gène X et le gène Y ont des niveaux d'expression identiques, mais le nombre de *reads* alignés au gène X apparaît comme plus important que le nombre de *reads* alignés au gène Y simplement parce que le gène X est plus long. [Harvard Chan Bioinformatics Core]

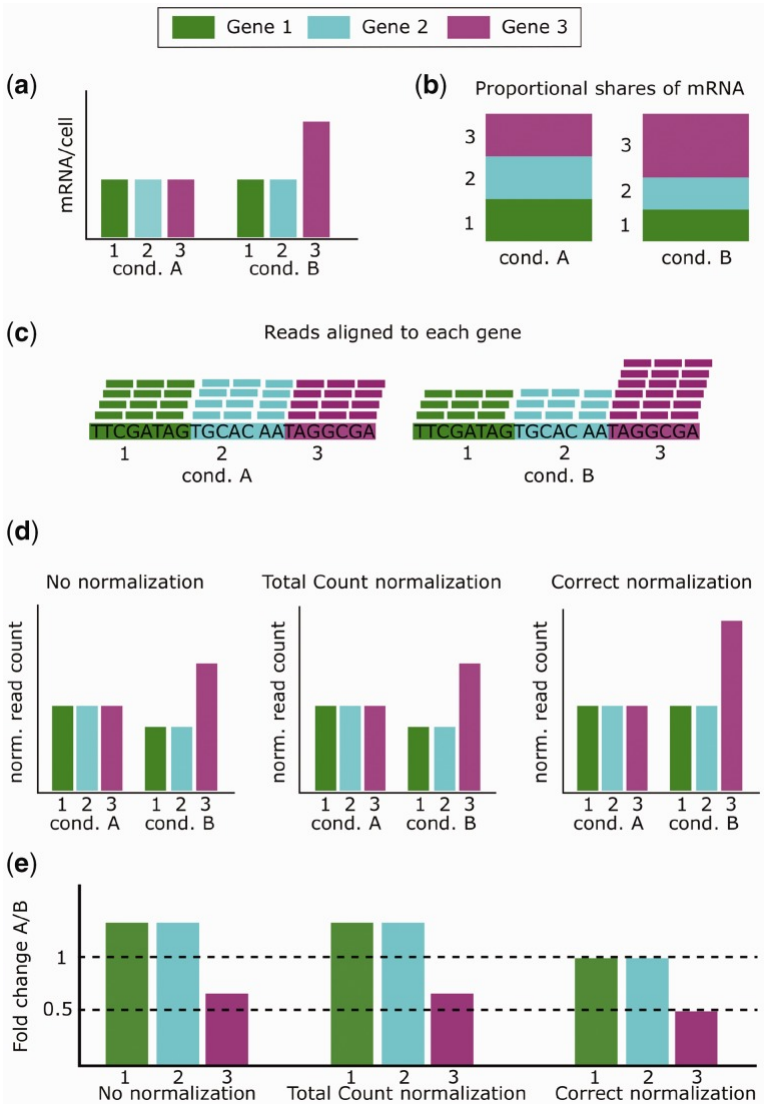


FIGURE 1-7 : Différentes proportions d'ARNm et normalisation. **(a)** Quantité d'ARNm/cellule dans la conditon A et B pour les gènes 1, 2 et 3. **(b)** Proportions d'ARNm/cellule dans la conditon A et B. **(c)** *reads* alignés pour chaque gène dans chaque condition pour une profondeur de séquençage donnée. **(d)** Quantité de *reads* sans normalisation, avec la normalisation TC et avec la normalisation adaptée. **(e)** *Fold-changes* associé à chaque gène pour chaque normalisation. [Evans et al., 2018]

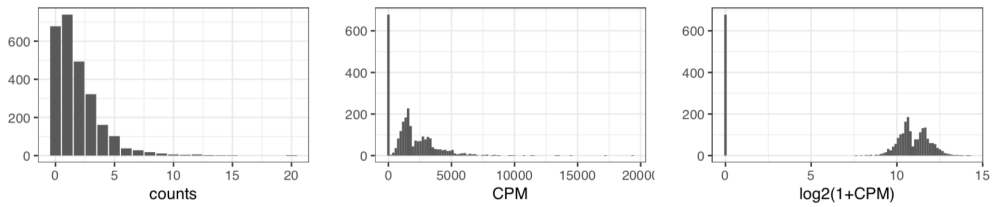
package `DESeq` [Anders and Huber, 2010], *Median Ratio* (MRN) [Maza et al., 2013], PoissonSeq (PS) [Li et al., 2012], *Trimmed Mean of M values* (TMM) du package `edgeR` [Robinson and Oshlack, 2010], *Quantile* (Q) [Bolstad et al., 2003; Yang and Thorne, 2003], *Counts per Millions* (CPM) [Robinson and Oshlack, 2010], transcripts per million mapped reads (TPM) [Wagner et al., 2012] et la normalisation *Reads Per Kilobase per Million mapped reads* (RPKM) [Mortazavi et al., 2008]. Au fur et à mesure de leur parution, des articles comparatifs [Dillies et al., 2013; Bullard et al., 2010; Maza et al., 2013; Lin et al., 2016; Li et al., 2015; Zyprych-Walczak et al., 2015; Rapaport et al., 2013; Kadota et al., 2012; Lovén et al., 2012; Risso et al., 2014; Evans et al., 2018] ont entrepris des comparaisons plus ou moins poussées afin que la communauté scientifique puisse avoir une vision plus claire sur la normalisation adaptée à leurs données. Plusieurs critères sont évalués comme la corrélation entre les comptes normalisés et les données qRT-PCR (considéré comme *gold standard*), la concordance des gènes DE sur données réelles, le taux de faux positifs, la puissance statistique, la sensibilité ou encore la spécificité via des courbes ROC. Les auteurs ont émis des recommadations s'agissant des méthodes de normalisation à éviter ou à privilégier. Dillies et al. [2013], Bullard et al. [2010] et Evans et al. [2018] s'accordent par exemple sur le fait que le *Total Count* et les RPKM sont des méthodes à éviter tandis que les approches `DESeq` et TMM sont les seules méthodes qui donnent de bons résultats à la fois en ce qui concerne la capacité à détecter les gènes DE et le contrôle des faux positifs.

Bien que les méthodes de normalisation pour données RNA-seq en masse soient encore utilisées, elles ne sont pas adaptées aux spécificités statistiques des données RNA-seq en cellule unique. Vallejos et al. [2017] encouragent à changer les habitudes des scientifiques et à utiliser des méthodes de normalisation spécifiques aux données RNA-seq en cellule unique plutôt que d'utiliser naïvement celles pour données RNA-seq en masse. Nous dressons une liste non-exhaustive des différentes normalisations pour données RNA-seq en cellule unique trouvées dans la littérature : *Single-Cell Tagged Reverse Transcription* (`SAMstrt`) [Katayama et al., 2013], *Bayesian Analysis of Single-Cell Sequencing Data* (`BASiCS`) [Vallejos et al., 2015], *Gamma Regression*

Model (`GRM`) [Ding et al., 2015], `scran` [Lun et al., 2016], un paquet pour l'analyse des données scRNA-seq, *Robust Normalization of Single-cell RNA-seq Data* (`SCnorm`) [Bacher et al., 2017], `Linnorm` [Yip et al., 2017], une méthode de transformation basée sur un modèle linéaire et la normalité pour les données scRNA-seq et récemment `Sanity` [Breda et al., 2021] basée sur de l'inférence bayésienne. De nombreux comparatifs ont été publiés [Lytal et al., 2020; Vallejos et al., 2017; Cole et al., 2019; Ding et al., 2015] et malgré la comparaison des performances à travers de multiples simulations et jeux de données réelles, aucune méthode ne se détache dans l'absolu.

Récemment, Townes et al. [2019] ont dénoncé, dans un papier très pédagogique, les effets délétères de la normalisation CPM suivie d'une transformation logarithmique sur des UMI (données RNA-seq en cellule unique), approche largement employée. En effet, dans le cas des UMI, la duplication due à la PCR n'est plus de mise et les comptes prennent de plus faibles valeurs comparativement aux *reads* après PCR. La transformation logarithmique est communément utilisée sur données RNA-seq en masse. L'expression des gènes étant mesurée dans un ensemble de cellules et renvoyée ensuite comme une moyenne, les comptes ont une plus grande magnitude. De plus, une particularité des données de comptage est qu'elles ont une variance fonction de leur moyenne, le passage au log permet alors d'éviter que les gènes fortement exprimés (qui ont donc des comptes élevés, et également une grande variance) prennent le dessus sur d'autres ayant des valeurs d'expression plus faibles. La transformation par la fonction log permet ainsi de faire une mise à l'échelle. Townes et al. [2019] insiste sur le fait que la log-transformation exagère artificiellement la différence entre les comptes nuls, qui sont déjà en grande quantité, et les comptes non nuls. C'est la raison pour laquelle ils décrivent l'inflation en 0 comme un artefact du à la log-transformation. Pour comprendre leurs arguments, prenons un exemple simple. La taille de la librairie d'une cellule varie en règle général entre 1000 et 3000 [Townes et al., 2019]. La normalisation CPM requière l'ajout d'un pseudo-compte, par exemple 1. Si un compte est nul, après la normalisation, il sera toujours nul car $\log\left(1 + \frac{0}{1000} \times 10^6\right) = 0$. Cependant, les comptes positifs vaudront $\log\left(1 + \frac{1}{1000} \times 10^6\right) \simeq 7$. Il est clair avec cet exemple que la log-transformation et la multiplication par le facteur 10^6 exagèrent artificiellement la

différence entre les comptes nuls (qui ne sont pas affectés par la normalisation) et les comptes non-nuls. La figure 1-8 illustre ce résultat. Les données brutes ne paraissent pas enflées en 0 alors que les CPM où chaque compte est divisé par la taille de la librairie et multiplié par 10^6 semblent enflés en 0. Enfin, les log-CPM sont très enflés en 0 et l'écart entre les comptes nuls et les comptes non-nuls apparaît comme démesuré.



(a) UMI counts (b) counts per million (CPM) (c) log of CPM

FIGURE 1-8 : Effets délétères de la normalisation CPM suivie d'une transformation logarithimique sur les UMI. Source : Townes et al. [2019]

Dans cette thèse, nous n'avons pas eu pour ambition de comparer les méthodes de normalisation dans nos simulations et les analyses de données réelles. Nous avons préféré employer les méthodes les plus couramment utilisées dans la littérature tout en avertissant sur les potentielles limites que le choix d'une approche peut générer. Dans les comparatifs sur données réelles, nous avons logiquement choisi de prendre la même normalisation que celle utilisée dans les articles dont sont issues les données (quand les matrices disponibles ne sont pas déjà normalisées) afin de pouvoir comparer nos résultats. En outre, les méthodes d'analyse différentielle que nous avons élaborées donnent la possibilité d'utiliser une matrice de comptes normalisés au préalable et ce, peu importe l'approche de normalisation retenue, contrairement aux méthodes les plus populaires qui se révèlent être limitantes. L'utilisateur peut choisir la normalisation qui lui convient en fonction de son design expérimental, de la nature des données et de ses propres recherches.

1.3 L'analyse d'expression différentielle

1.3.1 Principe et design expérimental

L'objectif de l'analyse d'expression différentielle est de déterminer quels gènes sont exprimés de manière différente, on dit qu'ils sont différentiellement exprimés, selon des conditions expérimentales. Pour cela, on doit se placer dans le cadre des tests d'hypothèse dont on donnera quelques généralités dans le Chapitre 2. Il s'agit alors de choisir l'outil statistique approprié, par exemple un modèle dont on testera les paramètres, et de définir l'hypothèse que l'on souhaite mettre à l'épreuve. En RNA-seq en masse, il s'agit de détecter la différence en moyenne alors qu'en RNA-seq en cellule unique, un plus grand intéret est porté à la différence en distribution. Typiquement, le cadre classique est la comparaison deux à deux entre des groupes expérimentaux. Par exemple, il est commun de vouloir comparer l'expression génique de patiens atteints d'une pathologie avec l'expression génique de patients sains, ou alors comparer l'expression génique de patients ayant reçus un vaccin avec ceux n'ayant pas été vaccinés. On peut envisager la comparaison entre deux populations cellulaires ou deux tissus cellulaires. Plus largement, de nouveaux designs expérimentaux voient le jour dans lesquels plusieurs conditions sont intéressantes à tester [Lévy et al., 2021; Singhania et al., 2018; Kuksin et al., 2021]. Lorsqu'un vaccin est administré, les scientifiques peuvent tester différentes doses et différents types de vaccins, il y aura donc autant de conditions que de dosages. Les injections peuvent également être réalisées à des intervalles de temps variés. Dans le Chapitre 4, nous analysons des données issus de patients atteints du COVID-19. Nous différencions alors les patients sains, les patients avec un COVID-19 modéré et les patients avec un COVID-19 sévère. Ces trois conditions constituent alors un design particulier que toutes les méthodes pour DEA ne peuvent pas prendre en compte. On peut aussi vouloir déterminer si l'expression génique évolue différentiellement au cours du temps [Hejblum et al., 2015; Agniel and Hejblum, 2017]. Les essais cliniques impliquant le suivi des patients au cours du temps sont nombreux [Dorr et al., 2015; Rechtien et al., 2017; Thiébaut et al., 2019]. Dès lors, il faut prendre en compte l'aspect temporel qui implique de gérer des données

groupées. L'analyse différentielle peut également consister à identifier les gènes qui s'expriment différentiellement selon la quantité d'une protéine ou de n'importe quelle molécule. La condition étant ici généralement continue, elle représente déjà une difficulté pour la majorité des méthodes. Pour aller plus loin, il est parfois necéssaire de vouloir prendre en compte une ou plusieurs variables confondantes, aussi dites variables de confusion. Une variable de confusion est un facteur autre que celui qui est étudié, qui est associé à la fois à la variable à expliquer (*e.g.* l'expression génique) et à la ou les variables explicatives. Une variable de confusion peut altérer ou masquer les effets de la variable à tester sur la mesure du phénomène en question. Elles peuvent être contrôlées statistiquement au cours de l'analyse, ce qui permet une mesure plus directe de la relation entre les variables d'intérêt.

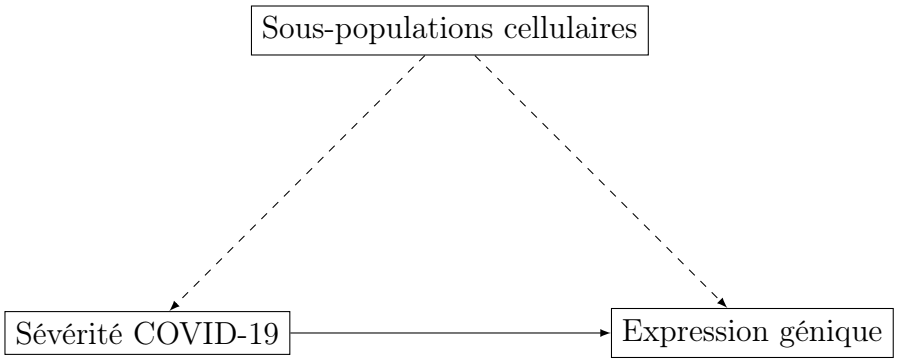


FIGURE 1-9 : Schéma de dépendance conditionnelle.

Dans le Chapitre 4, nous proposons l'étude d'un jeu de données RNA-seq en cellule unique [Kusnadi et al., 2021] dans lequel l'association de l'expression génique avec la sévérité de la maladie COVID-19 peut être influencée par la différence d'abondance dans les groupes de cellules T CD8+. Cette dernière peut potentiellement interargir en tant que variable confondante et masquer ou créer des liens entre la maladie et l'expression des gènes. Pour cela, nous analysons l'expression des gènes en fonction de la gravité de la maladie en ajustant sur les sous-groupes cellulaires afin d'exclure ce facteur de confusion. Aux multiples facettes de l'analyse différentielle s'ajoutent les défis méthodologiques intrinsèques aux données RNA-seq en masse et en cellule unique décrits dans la section 1.2.

1.3.2 Etat de l'art des méthodes d'analyse différentielle

Nous proposons une liste non-exhaustive des différentes méthodes d'analyse différentielle pour d'une part, les données RNA-seq en masse, en choisissant les trois approches les plus citées et utilisées auxquelles nous comparerons notre méthode dans le Chapitre 3, et d'autre part, pour les données RNA-seq en cellule unique dont certaines seront utilisées dans notre comparatif exposé dans le Chapitre 4.

Méthodes pour données RNA-seq en masse

Les méthodes d'analyse différentielle pour données RNA-seq en masse qui reviennent le plus dans la littérature sont `edgeR`, `DESeq2` et `limma-voom`. Nous résumons la méthodologie et les outils statistiques que chacune d'elles emploie.

edgeR [Robinson et al., 2010]

Pour un gène donné, `edgeR` modélise les données RNA-seq par une distribution binomiale négative pour l'échantillon i de la condition j :

$$Y_{ij} \sim NB(\mu_{ij}, \phi) \quad i = 1, ..., n_j \text{ et } j = 1, 2$$

avec $\mu_{ij} = m_i \lambda_j$, m_i étant la taille de la librairie pour l'échantillon i, λ_j l'abondance relative du gène dans le groupe expérimental j, ϕ la dispersion. Par suite, $\mathbb{E}(Y_{ij}) = \mu_{ij}$ et $Var(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}\phi)$. A noter que si $\phi = 0$ alors Y_{ij} suit une loi de Poisson. Le test porte sur le paramètre λ_j en posant l'hypothèse nulle "$\lambda_1 = \lambda_2$", et ce pour chaque gène. Deux tests sont possibles : un test de Wald en prenant comme statistique de test $\lambda_1 - \lambda_2$ divisé par l'écart-type estimé des résidus et un test dit exact, adapté aux données surdispersées, décrit dans [Robinson and Smyth, 2008].

L'estimation du paramètre de dispersion ϕ se fait en plusieurs étapes. D'abord, pour le gène g, on définit la log-vraisemblance conditionnelle pour ϕ sachant $z_j = \sum_{i=1}^{n_j} Y_{ij}$ par :

$$l_g(\phi) = \sum_{j=1}^{2} \left[\sum_{i=1}^{n_j} \log\Gamma(y_{ij} + \phi^{-1}) + \log\Gamma(n_j\phi^{-1}) - \log\Gamma(z_j + n_j\phi^{-1}) \right.$$
$$\left. - n_j \log\Gamma(\phi^{-1}) \right]$$

Ensuite, on obtient l'estimateur de la dispersion commune comme étant la valeur qui maximise la vraisemblance commune $l_C(\phi) = \sum_{g=1}^{G} l_g(\phi)$ où G est le nombre total de gènes. Cependant, il n'est pas toujours vrai que que chaque gène ait un paramètre de dispersion identique. A l'aide d'un estimateur de vraisemblance conditionnelle pondérée [Robinson and Smyth, 2007], les dispersions individuelles vont être comprimées vers la dispersion commune afin d'ajuster le paramètre de dispersion à chaque gène (*shrinkage*) :

$$WL(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g).$$

Enfin, l'estimateur de la dispersion individuelle ϕ_g est calculé.

Le paquet `edgeR` est implémenté en R et est disponible sur Bioconductor.

DESeq2 [Love et al., 2014]

Pour le gène g et l'échantillon i, Les comptes sont modélisés par un modèle linéaire généralisé de la famille binomiale négative avec un lien logarithmique : $y_{ij} \sim NB(\mu_{ij}, \phi_g)$ tels que $\mu_{gi} = s_g q_{gi}$ et $log(q_{gi}) = \sum_r x_{ir}\beta_{gr}$. Le paramètre de dispersion ϕ_g doit alors être estimé pour chaque gène. Dans un premier temps, l'estimation de la dispersion de chaque gène est obtenue par maximum de vraisemblance. Ensuite, comme illustré figure 1-10, pour estimer la relation moyenne-variance, la tendance de ces estimateurs est ajustée par une courbe paramétrée. Enfin, l'estimation finale du paramètre de dispersion du gène est ramenée à une valeur située entre l'estimation individuelle obtenue par maximum de vraisemblance et la valeur ajustée par la courbe, en calculant le maximum a posteriori (MAP) dans un modèle hiérarchique. On a donc un phénomène d'attraction de l'estimation pour un gène donné vers la tendance moyenne en empruntant de l'information à tous les gènes (*shrinkage*). La figure 1-10

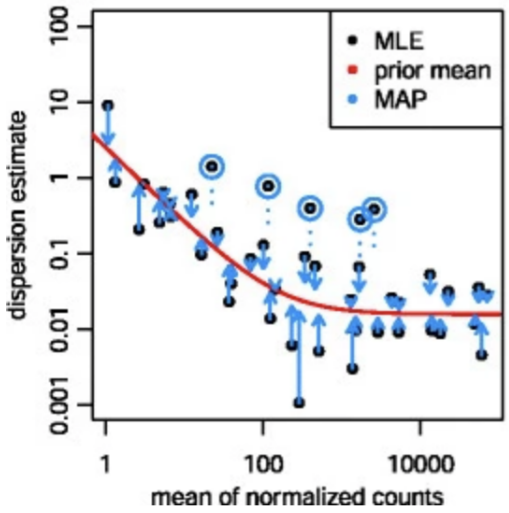


FIGURE 1-10 : Estimation de la dispersion dans `DESeq2`. Les points noirs représentent les estimateurs du maximum de vraisemblance (MLE) de la dispersion pour les différents gènes. La courbe rouge est ajustée aux MLE pour capturer la tendance générale. Les têtes des flèches correspondent aux estimations du maximum a posteriori, calculés à partir de la moyenne *a priori* (la courbe rouge). Les corps des flèches bleues matérialisent le « rétrécissement » des valeurs initiales vers un consensus. Les points noirs encerclés en bleu sont détectés comme des valeurs aberrantes et ne subissent pas le *shrinkage*. Source : Love et al. [2014]

illustre cette technique. Un test de Wald va permettre de tester les coefficients β_{ir} qui sont en réalité les estimations des *shrunken* log *fold-changes*.

Le paquet `DESeq2` est implémenté en R et est disponible sur Bioconductor.

limma-voom [Law et al., 2014]

La méthode `limma` commence par estimer un modèle linéaire avec hypothèse gaussienne pour chaque gène : $\mathbb{E}(y_g) = X\boldsymbol{\beta}_g$ où X est la matrice de design et β_g les coefficients à estimer. Puis, elle emploie également des méthodes bayésiennes empiriques pour obtenir des estimateurs de variance *a posteriori* décrites dans le papier de Ritchie et al. [2015]. Le modèle linéaire est estimé à partir des valeurs log-CPM y_{gi} pour chaque gène. On pose alors $\widehat{\beta}_g$ le coefficient estimé, $\widehat{\mu}_{gi} = x_i^T\widehat{\boldsymbol{\beta}}_g$ la valeur estimée de y_{gi} et s_g l'écart-type des résidus. Ensuite, la relation moyenne-variance est estimée de manière non-paramétrique, via une courbe de régression lissée appelée LOWESS (*LOcally Weighted Scatterplot Smoother*) notée $lo()$. Les écarts-types $s_g^{1/2}$ sont ajustés en

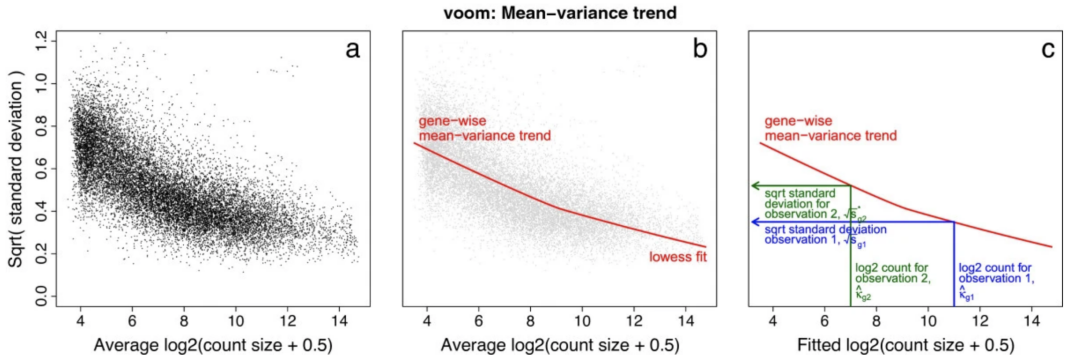


FIGURE 1-11 : Modélisation de la relation moyenne-variance par `voom` . a) Pour chaque gène, la racine carré de l'écart-type résiduel est représentée en fonction de la moyenne des *log-counts*. b) Ajustement de la relation moyenne-variance par une courbe LOWESS. c) La courbe LOWESS permet à chaque observation d'être associée à la racine carré d'un écart-type. Source : Law et al. [2014]

tant que fonction des moyennes ordonnées des comptes en échelle logarithmique (une transformation permet au préalable de passer des log-CPM aux log-comptes) pour chaque gène. Puis, les valeurs estimées $\widehat{\mu}_{gi}$ sont également converties en comptes et notés $\hat{\lambda}_{gi}$. Pour finir, $lo(\hat{\lambda}_{gi})$ donne la prédiction de la racine carré de l'écart-type des résidus de y_{gi}. Les poids $w_{gi} = lo(\hat{\lambda}_{gi})^{-4}$ ainsi que les log-CPM y_{gi} sont ensuite greffés à la procédure bayésienne empirique de `limma`. Les poids étant incorporés au modèle linéaire, la relation moyenne-variance est alors compensée. Une difficulté technique réside dans le fait qu'il faille estimer les variances des observations individuelles bien qu'il n'y ait aucune réplication au niveau de l'observation à partir de laquelle les variances pourraient être estimées. La relation moyenne-variance est estimée au niveau du gène, puis en interpolant cette tendance il est possible de prédire les variances des observations individuelles (voir figure 1-11). Finalement, les auteurs utilisent une *moderated t-statistic* pour tester $\boldsymbol{\beta}_g = 0$, détaillée dans McCarthy and Smyth [2009] et Smyth [2004].

Le paquet `limma` et le paquet `voom` sont implémentés en R et sont disponibles sur Bioconductor.

Méthodes pour données RNA-seq en cellule unique

Bien que les méthodes développées pour RNA-seq en masse puissent être utilisées

sur données RNA-seq en cellule unique tout en ne performant pas significativement moins bien (à condition de pré-filtrer les gènes au préalable) [Soneson and Robinson, 2018], elles ne sont pas capables de modéliser correctement l'inflation en zéro ni de détecter des différences plus complexes comme les différences en distribution.

MAST [Finak et al., 2015]

`MAST` propose un modèle linéaire généralisé en deux parties (*hurdle model*). La première partie modélise le niveau d'expression Y_{ig} en tant que variable binaire (compte nul ou non-nul) en fonction d'une matrice de design X contenant la condition à tester et le taux de détection cellulaire (optionnel) à l'aide d'une régression logistique, puis, la seconde partie modélise l'expression positive (excluant donc les zéros) par un modèle linéaire avec hypothèse gaussienne :

$$\begin{aligned} &logit\left(P\left(Z_{ig}=1\right)\right) = X_i\beta_g^D \\ &P\left(Y_{ig}=y \mid Z_{ig}=1\right) = \mathrm{N}\left(X_i\beta_g^C, \sigma_g\right) \end{aligned}$$

où β_g^D est le coefficient associé à la modélisation de la partie discrète, β_g^C est le coefficient associé à la partie continue et $\mathrm{Z} = [z_{ig}]$ indique si le gène g est exprimé dans la cellule i, c'est-à-dire $z_{ig} = 0$ si $y_{ig} = 0$ et $z_{ig} = 1$ si $y_{ig} > 0$. L'estimation de la variance σ_g se fait par une méthode empirique de Bayes qui consiste à réduire les estimations de la variance vers une variance globale. Nous ne la détaillerons pas ici. Les auteurs suggèrent d'inclure le taux de détection cellulaire pour chaque cellule i dans la matrice de design, défini comme $CDR_i = \frac{1}{N}\sum_g Z_{ig}$ où N est le nombre total de gènes et Z_{ig} est l'indicatrice spécifiant si le gène g s'exprime dans la cellule i. Le test statistique s'effectuant sur la nullité de β_g^D et β_g^C, un test de Wald pour chaque paramètre est fait. Z_g et Y_g étant définies comme conditionnellement indépendants pour chaque gène, les deux tests de Wald peuvent être additionnés tout en conservant leur distribution asymptotique, ici celle du χ^2 (et en sommant les degrés de liberté). La correction de Benjamini-Hochberg [Benjamini and Hochberg, 1995] pour la multiplicité des tests est appliquée. Notons que `MAST` permet de prendre en compte une

ou plusieurs covariables en plus du CDR, ce qui rend la méthode particulièrement attractive malgré l'approche modéliste.

SCDD [Korthauer et al., 2016]

Soit $Y_g = (y_{g1}, \ldots, y_{gJ})$ les mesures d'expression non nulles log-transformées du gène g dans J cellules issues de deux conditions. Sous l'hypothèse nulle de distributions équivalentes, Y_g l'expression du gène g est modélisée par un modèle de mélange à processus de Dirichlet (DPM) de gaussiennes. Les zéros sont modélisés à part et testés également séparément à la manière de `MAST` [Finak et al., 2015].

Un facteur de Bayes est défini pour déterminer si les données proviennent bien de deux distributions différentes traduisant l'impact significatif de la condition ou bien d'un modèle global qui ignore la condition (distributions équivalentes ou ED). On note $\mathcal{M}_{\text{DD}}$ l'hypothèse DD, et $\mathcal{M}_{\text{ED}}$ l'hypothèse de ED. Un facteur de Bayes pour le gène g s'écrit alors :

$$\text{BF}_g = \frac{f\left(Y_g \mid \mathcal{M}_{\text{DD}}\right)}{f\left(Y_g \mid \mathcal{M}_{\text{ED}}\right)}$$

où $f\left(Y_g \mid \mathcal{M}\right)$ désigne la distribution prédictive des observations du gène g sous l'hypothèse donnée. En général, il n'existe pas de solution analytique pour cette distribution dans le cadre du DPM.

Les auteurs proposent alors une approximation du facteur de Bayes et définissent ce dernier en tant que score. Ce score permet en outre de classer chaque gène en quatre catégories (optionnel) que nous avons défini figure 1-4. S'en suit un test par permutations qui permet d'estimer la distribution nulle du score et d'en déduire la p-valeur associée. Les observations sont permutées aléatoirement entre les deux conditions. La correction de Benjamin-Hochberg [Benjamini and Hochberg, 1995] est appliquée. Cette procédure passant par le facteur de Bayes engendre des temps de calculs particulièrement élevés, les auteurs proposent alors d'utiliser à la place le test non-paramétrique de Kolmogorov-Smirnov [Massey Jr, 1951] (ce qui ne permet pas de classifier les gènes en quatre catégories).

EMDomics [Nabavi et al., 2016]

`EMDomics` est une approche non-paramétrique basée sur la comparaison de deux histogrammes (un pour chaque condition testée). La distance de Wasserstein, aussi connue sous le nom de *Earth's Mover Distance* (EMD) va permettre d'identifier les différences en distribution pour un gène donné. Nous décrivons ce dernier procédé. Reprenons les notations de [Nabavi et al., 2016] et supposons que les gènes soient observés sous deux conditions différentes P et Q, avec M_1 cellules et M_2 cellules respectivement. Pour un gène, les histogrammes normalisés (*i.e.* la somme des hauteurs des intervalles est égale à 1) sont alors définis par : $P = \{(p_1, w_{p1}), \dots (p_i, w_{pi}), \dots (P_{M_1}, w_{pM_1})\}$ et $Q = \{(q_1, w_{q1}), \dots (q_j, w_{qj}), \dots \ (q_{M_2}, w_{qM_2})\}$, où p_i (respectivement q_j) est le centre de chaque intervalle de l'histogramme et le poids w_{pi} (respectivement w_{qj}) est la fréquence de chaque intervalle dans la condition P (respectivement Q). Le coût de transformation de la valeur p_i à q_j est défini comme la distance euclidienne $d_{ij} = \|p_i - q_j\|$. Le flux f_{ij} représente la proportion du poids associée à l'intervalle i qu'il faut "déplacer" pour qu'elle soit égale à celle de l'intervalle j. Le coût total s'écrit :

$$COST(P, Q, F) = \sum_{i=1}^{M_1} \sum_{j=1}^{M_2} f_{ij} d_{ij}$$

Il s'agit alors d'optimiser $F = [f_{ij}]$ afin de minimiser $COST$. Le score associé à l'EMD est défini comme suit :

$$EMD(P, Q) = \frac{\min COST}{\sum_{i=1}^{M_1} \sum_{j=1}^{M_2} f_{ij}}$$

Pour un gène, et sous l'hypothèse nulle qu'il n'y a pas de différence entre les deux conditions, un test par permutations, consistant à permuter aléatoirement les cellules entre elles, est appliqué pour approximer la distribution de ce coût sous H_0 et d'en déduire les p-valeurs correspondantes. Posons K le nombre total de permutations. A chaque permutation k, le score EMD est calculé selon la formule ci-dessus et noté emd_k. La p-valeur est donc exprimée par la formule suivante :

$$p\text{-valeur} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{\{emd_k \geq emd_{obs}\}}$$

avec emd_{obs}, le score associé à l'échantillon observé. La correction de Benjamini-

Hochberg [Benjamini and Hochberg, 1995] est enfin appliquée à l'ensemble des p-valeurs de tous les gènes. Notons que cette stratégie de calcul de l'EMD ajouté aux permutations est computationnellement très intensive et dépend intrinséquement du nombre de cellules observées et du nombre de permutations.

SigEMD [Wang and Nabavi, 2018]

`SigEMD` utilise le même test que `EMDomics` mais donne la possibilité de passer par une étape préalable d'imputation qui est décrite ci-après. Afin de réduire l'impact des zéros sur la distribution des gènes, `SigEMD` inclut une étape d'imputation des comptes nuls. Soit $\mathbf{Y} = [y_{gm}]$ la matrice d'expression génique de taille $N \times M$ telle que y_{gm} est l'expression du gène g dans la cellule m. On définit z_{gm} tels que $z_{gm} = 1$ si $y_{gm} > 0$ et $z_{gm} = 0$ si $y_{gm} = 0$. L'expression binaire du gène g est ensuite modélisée par une régression logistique :

$$logit\left(P\left(\mathbf{z_g} = 1\right)\right) = \mathbf{X}\mathbf{w_g}$$

où $\mathbf{X}$ est la matrice de design incluant la condition à tester et le taux de détection cellulaire, et $\mathbf{w}_g$ est le vecteur des coefficients. Pour un gène donné, si la condition n'est pas significative par le test de Wald alors cela signifie, d'après les auteurs, qu'elle ne peut pas expliquer si le gène est en effet exprimé. Pour ce gène, les zéros seront alors retirés. Si la condition est significative alors les zéros seront remplacés par des valeurs positives lors d'une étape d'imputation. Les gènes concernés par ce dernier cas de figure sont regroupés dans l'ensemble S_{IP}. Posons GDR_g le taux de détection du gène g (*Gene Detection Rate*) tel que $GDR_g = \frac{1}{M}\sum_{m=1}^{M} z_{gm}$ où M représente le nombre de cellules et $z_{gm} = 1$ si $y_{gm} > 0$, $z_{gm} = 0$ sinon. On calcule GDR_g pour l'ensemble des gènes nécessitant une imputation. Les gènes avec un $GDR > 80\%$ constituent l'ensemble S_G. Les s gènes les plus corrélés avec le gène g dans l'ensemble S_G au sens de la corrélation de Pearson forment l'ensemble S_{G_g}. Pour le gène g, une régression LASSO [Tibshirani, 1996] est appliquée tel que :

$$\min\left\{\frac{1}{|S_{Gg}|}\left\|\mathbf{y}_{G_g,m} - \mathbf{Y}_{G_g,-m}\beta_{gm}\right\|_2^2 + \lambda\left\|\beta_{gm}\right\|_1\right\}$$

avec $\left|S_{G_g}\right|$ le nombre de gènes dans S_{G_g}, $\mathbf{y}_{G_g,m}$ le vecteur des G_g observations dans la cellule m, $\mathbf{Y}_{G_g,-m}$ est la matrice de design avec G_g lignes et $(M-1)$ colonnes, λ est le paramètre de régularisation de la régression LASSO et $\boldsymbol{\beta}_{gm}$ est le vecteur des $(M-1)$ coefficients de régression. Les auteurs proposent ensuite d'estimer l'expression y_{gm} du gène g dans la cellule m, à l'aide des coefficients précédemment estimés $\boldsymbol{\beta}_{gm}$:

$$\begin{cases} \hat{y}_{gm} = \mathbf{y}_{g,-m} \cdot \beta_{gm} & \text{if } y_{gm} = 0 \text{ and } g \in S_{IP} \\ \hat{y}_{gm} = y_{gm} & \text{if } y_{gm} \neq 0 \text{ or } g \notin S_{IP} \end{cases}$$

où $\mathbf{y}_{\mathrm{g,-m}}$ est le vecteur d'expression du gène g privé de la cellule m. Finalement, seules les valeurs nulles des gènes de l'ensemble S_{IP} sont imputées et les autres valeurs restent inchangées. L'analyse différentielle s'effectuera sur une nouvelle matrice post-imputation en suivant la méthodologie de `EMDomics`.

D3E [Delmans and Hemberg, 2016]

La spécificité de `D3E` est de proposer un test du rapport de vraisemblance passant par l'ajustement d'un modèle Poisson-Bêta. Il est aussi possible d'utiliser le test non-paramétrique de Cramer-von Mises [Anderson, 1962] ou le test de Kolmogorov-Smirnov [Smirnov, 1939].

distinct [Tiberi et al., 2020]

`distinct` se base sur les fonctions de répartition empiriques pour identifier les gènes différentiellement exprimés entre deux conditions. Elle est uniquement utilisable dans le cas de designs expérimentaux avec réplicats biologiques puisqu'elle s'appuie sur la comparaison entre groupes de distributions (en utilisant la moyenne des fonctions de répartition). Le papier décrit la méthodologie pour prendre en compte des covariables mais cela n'a pas encore été implémenté. On se place dans le contexte de la comparaison entre deux groupes A et B avec réplicats biologiques, c'est-à-dire plusieurs échantillons, avec N_A et N_B échantillons. Pour chaque échantillon, la fonction de répartition empirique des observations (ECDF pour *empirical cumulative distribution function*) le i-ème échantillon dans le j-ème groupe est désignée par $ecdf_i^{(j)}(.)$,

pour $j \in \{A, B\}$ et $i = 1, \ldots, N_j$. On définit $b_1, \ldots, b_K$, les seuils auxquels les ECDF sont calculées. [Tiberi et al., 2020] proposent de calculer la statistique de test comme la différence absolue entre l'ECDF moyen des deux groupes (moyenne faite sur les réplicats dans une condition) sommée sur tous les seuils :

$$s^{obs} = \sum_{k=1}^{K} \left| \frac{\sum_{i=1}^{N_A} ecdf_i^{(A)}(b_k)}{N_A} - \frac{\sum_{i=1}^{N_B} ecdf_i^{(B)}(b_k)}{N_B} \right|$$

La distribution nulle est estimée via une approche par permutations.

DEsingle [Miao et al., 2018]

`DEsingle` utilise un modèle de Binomiale Négative avec inflation en zéro (ZINB pour *Zero Inflated Negative Binomial*). La distribution ZINB est un mélange d'une fonction de masse en zéro et d'une distribution binomiale négative (NB) :

$$\begin{aligned} &P(N_g = n \mid \theta, r, p) = \theta.\mathbb{1}(n = 0) + (1 - \theta).f_{\text{NB}}(r, p) \\ &= \theta \cdot \mathbb{1}(n = 0) + (1 - \theta) \cdot \begin{pmatrix} n + r - 1 \\ n \end{pmatrix} p^n (1 - p)^r, \quad n = 0, 1, 2, \ldots \end{aligned}$$

où θ est la proportion de zéros du gène g, $\mathbb{1}(n = 0)$ est une fonction indicatrice qui est égale à 1 si $n = 0$ et 0 $n \neq 0$, f_{NB} est la densité de la distribution discrète Négative Binomiale, r est le paramètre de taille de la loi NB et p est le paramètre de probabilité de la loi NB. Les auteurs font remarquer que la loi NB peut également générer des valeurs nulles. Les valeurs nulles observées sont le mélange de zéros constants et de zéros de la distribution NB. `DEsingle` utilise un test de rapport de vraisemblance comparant deux modèles ZINB pour les deux conditions biologiques afin de détecter les gènes DE et de les subdiviser en trois types : DEs pour les gènes qui montrent une proportion significativement différente de zéros mais l'expression de ces gènes dans les autres cellules ne présente aucune différence significative, DEa pour les gènes dont l'expression n'est pas significative dans la proportion de zéros, mais qui présentent une expression différentielle significative dans les autres cellules et DEg pour les gènes qui

présentent une différence significative dans la proportion de zéros, mais sont également différentiellement exprimés de dans les valeurs positives de l'expression.

1.4 Objectifs et plan de la thèse

L'objectif de cette thèse a été de proposer des méthodes d'analyse différentielle de l'expression génique qui reposent sur la volonté de créer des outils les plus flexibles possible tant par les différents types de variables pris en compte selon le schéma expérimental que par l'absence d'hypothèses distributionnelles.

À cet objectif principal s'ajoutent, d'une part, les problématiques intrinsèques aux données RNA-seq en masse :

- la modélisation de la relation moyenne-variance,
- le contrôle du taux de fausses découvertes, négligé dans les méthodes les plus connues,

et d'autre part, les spécificités des données RNA-seq en cellule unique :

- la prépondérance des valeurs nulles,
- la distribution de l'expression génique généralement hétérogène et multimodale.

Dans le Chapitre 2, nous proposons des rappels statistiques et méthodologiques nécessaires à la bonne compréhension de la suite du manuscrit. Le Chapitre 3 présente `dearseq`, une approche assurant un bon contrôle du FDR, basée sur un test du score en composante de variance qui permet d'identifier les transcrits dont le niveau d'expression est significativement associé à un ensemble de variables, conditionnellement à des covariables et sans poser d'hypothèses distributionnelles sur les données RNA-seq en masse. `dearseq` se différencie par sa grande flexibilité à l'aide d'un modèle mixte. En effet, il est possible d'analyser des données groupées ou non (e.g. longitudinales)

grâce à l'ajout d'un effet aléatoire. Au moyen d'une régression linéaire locale, l'information est empruntée à tous les gènes pour estimer la relation moyenne-variance. La méthode développée a été publiée dans *Nuclear Acid Research Genomics and Bioinformatics* (doi : `https://doi.org/10.1093/nargab/lqaa093`) et a été intégralement implémentée dans un paquet R disponible sur Bioconductor. Dans le Chapitre 4, nous introduisons `ccdf`, une nouvelle méthode d'analyse différentielle pour données RNA-seq en cellule unique. En généralisant le concept d'analyse différentielle, nous nous ramenons à un test d'indépendance conditionnelle. Afin de caractériser la distribution du gène, nous utilisons les fonctions de répartition conditionnelles. Nous proposons une estimation originale de celles-ci en passant par plusieurs régressions linéaires. Bien que cette méthode ait été pensée pour des données RNA-seq en cellule unique, elle dépasse largement le cadre de l'analyse différentielle et de la génomique en se voulant générale dans son implémentation et dans les types de variables en entrée. Ainsi, il est tout à fait raisonnable de vouloir l'appliquer à d'autres types de données dans des pans différents de la santé ou même de l'économie, tant que le nombre d'observations est suffisant pour que la puissance statistique soit satisfaisante. La méthode développée a fait l'objet d'un article soumis et a été intégralement implémentée dans un paquet R disponible sur le CRAN. Dans le Chapitre 5, nous discutons de nos développements en considérant leurs forces et leurs faiblesses tout en apportant des pistes de réflexions sur de possibles extensions et améliorations.

Chapitre 2

Rappels statistiques

2.1 Modèles mixtes

Comme vu dans le chapitre 1, il arrive fréquemment en immunologie d'être confronté à des schémas impliquant des mesures groupées ou répétées, longitudinales ou non. Par exemple, on retrouve des observations groupées lorsque des tissus distincts sont prélevés chez un patient. Les deux prélèvements appartiennent donc au même individu et présentent donc une certaine structure de corrélation. En multipliant cette opération pour plusieurs patients, on peut formuler que les échantillons provenant d'un même patient sont corrélés alors que ceux issus d'individus différents sont indépendants. En ce qui concerne les données longitudinales, le schéma expérimental dans Rechtien et al. [2017] comprend le transcriptome observé chez 20 sujets à 4 temps de mesure. Pour un patient donné, les 4 mesures de l'expression génique dites intra-sujets, sont donc intrinséquement corrélées alors que les mesures temporelles inter-sujets sont indépendantes.

Dans cette optique, la méthode d'analyse différentielle décrite au Chapitre 3 s'appuie sur un modèle linéaire mixte. Nous rappelons ici quelques généralités.

Le modèle linéaire classique, aussi appelé modèle à effets fixes, s'écrit :

$$Y_{ij} = X_{ij}\boldsymbol{\alpha}^{\intercal} + \varepsilon_{ij} \tag{2.1}$$

où Y_{ij} est de $p \times 1$ vecteur de la variable à expliquer, X_{ij} est le $p \times 1$ vecteur des variables explicatives pour la mesure j du sujet i, $\boldsymbol{\alpha}$ est le vecteur des effets fixes, $\varepsilon_{ij} \sim N(0, \sigma^2)$ l'erreur résiduelle. Une des limites du modèle à effets fixes est qu'il suppose une indépendance entre les observations y_{ij}. Or, les y_{ij} issues d'un même individu présentent un certain degré de corrélation et l'hypothèse d'indépendance est alors violée. Dans ce cas le modèle à effets fixes n'est plus adapté. Pour prendre en compte la dépendance entre certains individus, on rajoute au modèle à effets fixes un effet aléatoire, le modèle ainsi obtenu est appelé modèle à effets mixtes.

Le modèle linéaire à effets mixtes introduit dans Laird and Ware [1982] adapte le modèle de régression linéaire au cadre des données groupées et longitudinales. Supposons qu'on dispose de n sujets (ou patients), les variables du patient i sont observées n_i fois au cours du temps (structure longitudinale) ou sont simplement observées n_i fois (structure groupée mais non-longitudinale). On note Y_{ij} la j-ième mesure de l'individu i. Y_{ij} est modélisée par :

$$Y_{ij} = X_{ij}^{\intercal}\boldsymbol{\alpha} + \Phi_{ij}^{\intercal}\boldsymbol{\xi}_i + \varepsilon_{ij} \tag{2.2}$$

où X_{ij} est le $p \times 1$ vecteur des variables explicatives pour la mesure j du sujet i, $\boldsymbol{\alpha}$ est le $p \times 1$ vecteur des effets fixes, $\boldsymbol{\xi}_i$ est le $q \times 1$ vecteur des effets aléatoires, Φ_{ij} le $q \times 1$ vecteur de covariables, et ε_{ij} est l'erreur résiduelle. De plus, $\varepsilon_i = (\varepsilon_{ij})_{j=1,\ldots,n_i} \sim N(0, \Sigma_i)$. Pour tout $i = 1, \ldots, n$, les $\boldsymbol{\xi}_i$ sont indépendantes. Les ε_{ij} sont également indépendantes pour tout $i = 1, \ldots, n,\ j = 1, \ldots, n_i$. De plus les $\boldsymbol{\xi}_i$ et ε_{ij} sont mutuellement indépendantes. Les $\boldsymbol{\xi}_i$ sont distribuées selon une normale $\mathcal{N}(0, \Sigma_\xi)$ où Σ_ξ est une matrice définie positive de taille $q \times q$.

On définit le modèle linéaire mixte sous sa forme matricielle :

$$Y_i = X_i\boldsymbol{\alpha} + \Phi_i\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i \tag{2.3}$$

où X_i est la matrice de taille $n_i \times p$ des variables explicatives de l'individu i, $\boldsymbol{\alpha}$ est le $p \times 1$ vecteur des effets fixes, Φ_i est la matrice des covariables de taille $n_i \times q$, $\boldsymbol{\xi}_i$ est le $q \times 1$ vecteur des effets aléatoires de l'individu i, et $\boldsymbol{\epsilon}_i$ est le $n_i \times 1$ vecteur des

erreurs résiduelles. Dans le modèle (2.3) et en présence de données longitudinales, il est considéré que l'évolution au cours du temps de la variable réponse Y_i pour le sujet i varie autour d'un comportement moyen $X_i\alpha$. Ce comportement moyen est commun à tous les sujets, il est supposé être linéaire en les variables qui composent la matrice X_i. Les variations individuelles autour de ce comportement moyen sont caractérisées par le vecteur des effets aléatoires $\boldsymbol{\xi}_i$ qui est spécifique à chaque individu. Du modèle (2.3), on peut en déduire la distribution marginale de Y_i :

$$Y_i \sim N\left(X_i\alpha, V_i\right), \text{ telle que } V_i = \Phi_i \Sigma_\xi \Phi_i^\intercal + \Sigma_i.$$

On note γ les paramètres de variance et de covariance inclus dans V_i et $\theta = (\gamma^\intercal, \alpha^\intercal)^\top$. La log-vraisemblance s'écrit

$$\ell_n(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\left\{n_i \log(2\pi) + \log det(V_i(\gamma)) + \left(Y_i - X_i\alpha\right)^\top V_i^{-1}(\gamma)\left(Y_i - X_i\alpha\right)\right\}.$$

On peut obtenir l'estimateur du maximum de vraisemblance (ML) θ^{ML} en maximisant de manière itérative la log-vraisemblance. Si γ est connu, alors en résolvant l'équation du score suivante

$$\frac{\partial \ell_n(\theta)}{\partial \alpha} = 0,$$

on obtient un estimateur de α qui dépend de γ :

$$\widehat{\alpha}(\gamma) = \left(\sum_{i=1}^{n} X_i^\top V_i^{-1}(\gamma) X_i\right)^{-1} \sum_{i=1}^{n} X_i^\intercal V_i^{-1}(\gamma) \boldsymbol{y}_i$$

avec $\boldsymbol{y}_i$ l'observation de Y_i. Si γ est inconnue, on peut utiliser son estimateur du maximum de vraisemblance $\hat{\gamma}^{ML}$ calculé en maximisant $\ell_n(\gamma, \hat{\beta}(\gamma))$ en γ.

Dans l'estimation d'un modèle linéaire mixte, il est préférable d'utiliser le maximum de vraisemblance restreint (REML) plutôt que le maximum de vraisemblance classique afin d'éviter les estimations biaisées des paramètres de variance [Harville, 1977]. L'estimation par REML n'inclut que les paramètres de variance, les effets fixes sont estimés dans une seconde étape. De plus, dans un but de prédiction par

exemple, il est intéressant de calculer une estimation des effets aléatoires. La distribution postérieure $f(\boldsymbol{\xi}_i \mid \boldsymbol{y}_i)$ suit une densité normale multivariée et l'effet aléatoire $\boldsymbol{\xi}_i$ est généralement estimé par la moyenne de cette distribution postérieure : $\boldsymbol{\xi}_i(\theta) = \mathbb{E}(\boldsymbol{\xi}_i|\boldsymbol{Y}_i = \boldsymbol{y}_i) = \Sigma_\xi \Phi_i^\intercal V_i^{-1}(\gamma)(\boldsymbol{y}_i - X_i\boldsymbol{\alpha})$. Cet estimateur des effets aléatoires individuels est le meilleur prédicteur linéaire sans biais (BLUP) [Verbeke, 2000]. Pour plus de détails sur l'estimation des différents paramètres, nous renvoyons au livre "Modèles biostatistiques pour l'épidémiologie" [Commenges and Jacqmin-Gadda, 2015].

Dans le Chapitre 3, nous estimons seulement les paramètres sous H_0, qui de ce fait nous évitera d'avoir à utiliser l'algorithme EM qui peut être computationnellement intensif de par son caractère itératif. De plus, le paramètre de variance résiduelle dépendra de i et de j ce qui demandera une effort supplémentaire en plus de l'hétéroscédasticité inhérente aux données RNA-seq.

2.2 Généralités sur les tests d'hypothèse

L'analyse différentielle consiste en réalité en un test d'hypothèse. Même si les tests sont multiples et variés, les étapes de leur constructions sont communes. Nous proposons ici un rappel sur la théorie des tests d'hypothèse.

Un test d'hypothèse se formalise en plusieurs étapes :

1. poser l'hypothèse nulle H_0 et l'hypothèse alternative H_1
2. choisir une statistique de test pour tester H_0
3. définir la distribution de la statistique de test sous l'hypothèse nulle
4. calculer la valeur de la statistique de test observée
5. calculer la p-valeur associée à la statistique de test

La p-valeur est la probabilité sous l'hypothèse nulle d'obtenir une valeur au moins aussi extrême que celle observée.

Un test d'hypothèse consiste dans un premier temps à formuler deux hypothèses définies comme l'hypothèse nulle et l'hypothèse alternative. L'hypothèse nulle est ce que nous souhaitons tester.

L'hypothèse nulle

Typiquement, l'hypothèse nulle renvoie à l'absence d'effet et doit être définie avec attention. Le rejet ou non de l'hypothèse nulle se tranche à partir de règle de décision statistique. Il est important de rappeler que H_0 n'est jamais acceptée en tant que tel. En effet, il n'y a pas de certitude mathématique à statuer sur le fait que H_0 soit vraie, l'hypothèse est simplement rejetée ou non-rejetée au seuil de significativité choisi α.

Selon la structure de H_0, le test sera soit unilatéral, soit bilatéral. Soit μ le paramètre d'intêrét et μ_0 la valeur de comparaison. Un test est bilatéral lorsque l'hypothèse nulle est de la forme $H_0 : \mu = \mu_0$ et l'hypothèse alternative $H_1 : \mu \neq \mu_0$. Un test est unilatéral lorsque l'hypothèse nulle est de la forme $H_0 : \mu = \mu_0$ et l'hypothèse alternative $H_1 : \mu < \mu_0$ ou bien $H_1 : \mu > \mu_0$ (table 2.1) car le signe de la différence potentielle est connu.

Statistique de test et niveau de significativité

Une statistique de test est une fonction des variables aléatoires et est propre à chaque test. Le modèle mathématique, les outils statistiques préalablement utilisés, le design expérimental et les hypothèses distributionnelles faites sur les données sont autant d'élements déterminant la statistique de test. La prise de décision quant au rejet de H_0 se fait par la valeur observée de la statistique de test. Une fois la loi de probabilité de la statistique S sous l'hypothèse H_0 posée (la loi de probabilité n'est pas toujours triviale et nécessite souvent un travail approfondi), il est possible d'établir une valeur seuil, S_{seuil} de la statistique pour une probabilité donnée appelée le niveau de significativité α du test. La région critique R_c correspond à l'ensemble des valeurs telles que : $\mathbb{P}(S \in R_c) = \alpha$. Selon la nature unilatérale ou bilatérale du test, la définition de la région critique varie comme décrit table 2.1.

Le choix du niveau de significativité α dépend intrinséquement de l'impact qu'il

Test	Unilatéral $H_0 : \mu = \mu_0$		Bilatéral $H_0 : \mu = \mu_0$
Hypothèse alternative	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$
Niveau de significativité	$\mathbb{P}(S > S_{\text{seuil}}) = \alpha$	$\mathbb{P}(S < S_{\text{seuil}}) = \alpha$	$\mathbb{P}(\lvert S\rvert > S_{\text{seuil}}) = \alpha$

TABLE 2.1 : Test d'hypothèse unilatéral et bilatéral

aura en pratique. α est un compromis entre le risque que l'on est prêt à prendre de rejeter à tort et la capacité à faire des découvertes. A vouloir être trop conservateur, on peut aussi passer à côté de potentiels résultats. α peut être très faible, par exemple en astrophysique où les calculs doivent être d'une extrême précision, en physique nucléaire où une infime erreur peut avoir des conséquences dramatiques ou encore dans les essais cliniques où la santé des patients est en jeu. En biostatistiques, il est commun de choisir $\alpha = 0.05, 0.01, 0.001$.

Il existe deux stratégies équivalentes pour prendre la décision de rejeter l'hypothèse nulle :

- Sous l'hypothèse "H_0 est vraie" et pour un seuil de significativité α fixé, si la valeur de la statistique S_{obs} calculée à partir des données appartient à la région critique alors l'hypothèse H_0 est rejetée au risque d'erreur α. A contrario, si la valeur de la statistique S_{obs} n'appartient pas à la région critique alors l'hypothèse H_0 ne peut être rejetée.

- La probabilité critique α_{obs} telle que $P(S \geq S_{obs}) = \alpha_{obs}$ est évaluée. Si $\alpha_{obs} \geq \alpha$ l'hypothèse H_0 n'est pas rejetée car la probabilité de rejeter H_0 à tort est supérieure au niveau spécifié. Si $\alpha_{obs} \leq \alpha$, l'hypothèse H_0 est rejetée car la probabilité de rejeter H_0 à tort est considérée comme négligeable.

Risques d'erreur

On appelle risque d'erreur de première espèce la probabilité de rejeter H_0 et d'accepter H_1 alors que H_0 est vraie. Ceci se produit si la valeur de la statistique de test tombe dans la région de rejet alors que l'hypothèse H_0 est vraie. La probabilité de cet évènement est le niveau de significativité α. On dit aussi que le niveau de significativité est la probabilité de rejeter l'hypothèse nulle à tort.

On appelle risque de seconde espèce, noté β la probabilité de ne pas rejeter H_0 alors que H_1 est vraie. Ceci se produit si la valeur de la statistique de test ne tombe pas dans la région de rejet alors que l'hypothèse H_1 est vraie. Pour quantifier le risque β, il faut connaître la loi de probabilité de la statistique sous l'hypothèse H_1, ce qui est rarement le cas en pratique.

Puissance d'un test

L'aptitude d'un test à rejeter H_0 alors que celle-ci est fausse constitue la puissance du test. On appelle donc puissance d'un test, la probabilité de rejeter H_0 alors que H_1 est vraie. Sa valeur est $1 - \beta$. La puissance augmente aussi avec la taille d'échantillon et diminue lorsque α diminue. De plus, un test unilatéral aura tendance à être plus puissant qu'un test bilatéral dans le sens où il requiert une taille d'échantillon plus faible pour atteindre au moins la même puissance. Les différentes situations que l'on peut rencontrer dans le cadre des tests d'hypothèse sont résumées dans la table 2.2.

Décision / Réalité	H_0 vraie	H_1 vraie
H_0 non rejetée	correct	risque de seconde espèce β
H_0 rejetée	rejet à tort risque de première espèce α	puissance du test $1 - \beta$

TABLE 2.2 : Résultats possible d'un test d'hypothèse

2.3 Correction pour la multiplicité des tests

Dans le contexte de l'analyse différentielle en gène par gène (par opposition aux groupes de gènes), il y a autant de tests qu'il n'y a de gènes. En général, il est nécessaire d'analyser entre 10 000 et 20 000 gènes. Lorsque plusieurs tests sont effectués simultanément sur les mêmes données, on observe une augmentation de l'erreur de type I. La table 2.3 récapitule les différents résultats possibles lors de la réalisation de tests multiples. Dans de nombreuses applications, notamment en santé, on souhaite contrôler le risque de première espèce α lors de la réalisation d'un test. Lors de la réalisation de tests multiples, il existe différentes généralisations de la probabilité

d'erreur de type I. Si le *Family Wise Error Rate* (FWER) semble être l'extension naturelle de l'erreur de type I, plusieurs autres quantités, telles que le taux de fausses découvertes *False Discovery Rate* (FDR), sont de plus en plus utilisées.

Réalité \Décision	H_0 non rejeté	H_0 rejeté	Total
H_0 vraie	U	V	m_0
H_0 fausse	T	S	m_1
Total	W	R	m

TABLE 2.3 : Table croisée de la réalité et de la décision à partir de m hypothèses nulles H_0

Family Wise Error Rate

Le FWER est défini comme la probabilité qu'au moins un des m tests considérés rejette à tort H_0 : $FWER = \mathbb{P}(V > 0)$. Plusieurs procédures existent pour contrôler le FWER dans une expérience [Dudoit et al., 2008]. Une méthode simple et largement utilisée est la méthode de Bonferroni. Si chaque test est contrôlé au niveau α^*, alors le FWER est borné :

$$FWER \leq 1 - (1 - \alpha^*)^m$$

Contrôler chaque test au niveau :

$$\alpha^* = \frac{\alpha}{m}$$

assure que le FWER est contrôlé au niveau α.

Taux de fausses découvertes

Une autre quantité qui a émergé pour la correction des tests multiples est l'espérance du taux de faux positifs, appelé taux de fausse découverte (FDR) :

$$FDR = \mathbb{E}\left(\frac{V}{R} \mid R\right) P(R > 0)$$

Le contrôle du FDR à un niveau α est moins conservatif que le contrôle du FWER au même niveau α puisque $FDR \leq FWER$ [Dudoit et al., 2008], le cas limite étant

celui où toutes les hypothèses nulles sont vraies ($m_1 = 0$). Ainsi, toute procédure contrôlant le FWER contrôle par suite le FDR. Il existe plusieurs procédures pour contrôler le FDR. En particulier, la procédure de Benjamini et Hochberg [Benjamini and Hochberg, 1995] pour corriger les p-valeurs afin de contrôler le FDR est largement utilisée. Considérons un vecteur de m p-valeurs :

1. Les p-valeurs sont classées par ordre croissant : $p_{(1)} \dots p_{(m)}$, où $p_{(1)}$ est la plus petite p-valeur obtenue parmi les m tests effectués, et $p_{(m)}$ la plus grande.
2. Pour le niveau α, on cherche le plus grand k tel que : $p_{(k)} \leq \frac{k}{m}\alpha$
3. Toutes les hypothèses nulles $H_{0;(i)}$ sont rejetées pour $i = 1 \dots k$.

Les travaux présentés dans cette thèse font usage de la correction de Benjamini-Hochberg pour l'obtention des p-valeurs ajustées. Néanmoins, d'autres méthodes existent comme celles proposée par Storey [2002]. Au lieu de fixer α puis d'estimer la région de rejet à partir de k, c'est la région de rejet qui est choisie pour ensuite estimer α.

Chapitre 3

dearseq : a variance component score test for RNA-Seq differential analysis that effectively controls the false discovery rate

Marine Gauthier [1,2], **Denis Agniel** [3,4]
Rodolphe Thiébaut [1,2,5], **Boris P. Hejblum** [1,2]

[1]University of Bordeaux, INSERM Bordeaux Population Health Research Center, INRIA SISTM, F-33000 Bordeaux, France. [2]Vaccine Research Institute, F-94000 Créteil, France. [3]Rand Corporation, Santa Monica (CA), USA. [4]Harvard Medical School, Boston (MA), USA. [5]CHU de Bordeaux, Bordeaux, F-33000 France.

Published in *Nuclear Acid Research Genomics and Bioinformatics*
DOI: https://doi.org/10.1093/nargab/lqaa093

Sections of the article have been rearranged for the purpose of this thesis

Abstract

RNA-seq studies are growing in size and popularity. We provide evidence that the most commonly used methods for differential expression analysis (DEA) may yield too many false positive results in some situations. We present `dearseq`, a new method for DEA which controls the FDR without making any assumption about the true distribution of RNA-seq data. We show that `dearseq` controls the FDR while maintaining strong statistical power compared to the most popular methods. We demonstrate this behavior with mathematical proofs, simulations, and a real data set from a study of Tuberculosis, where our method produces fewer apparent false positives.

3.1 Introduction

With the rise of next generation sequencing technologies that measure gene expression on the scale of the entire genome, RNA-seq differential expression analysis (DEA) has become ubiquitous in many research fields. While numerous approaches have been proposed to perform DEA of RNA-seq data, there is no clear consensus on which method is the most efficient. Three methods stand out as the most commonly used in practice: `edgeR` [Robinson et al., 2010], `DESeq2` [Love et al., 2014], and `limma-voom` [Law et al., 2014] (respectively 6,999, 6,856, and 1,004 citations in PubMed as of December 11$^{\text{th}}$, 2019). `edgeR` and `DESeq2` both rely on the assumption that gene counts from RNA-seq measurements follow a Negative Binomial distribution, and `limma-voom` is based on a weighted linear model and assumes resulting test statistics follow a normal distribution.

Following long-standing statistical practice, researchers typically attempt to control the probability of finding a gene to be differentially expressed (DE) when the opposite is true in reality (i.e. the Type-I error) at a pre-specified level (conventionally 5%). In a high-dimensional context such as gene expression data, the false positive rate or False Discovery Rate (FDR) [Benjamini and Hochberg, 1995] has been largely adopted as the target probability to be controlled in exploratory studies. The FDR is the expected proportion of features identified as significant that are actually false positives: for instance, an FDR of 5% implies that among all the genes declared DE, 5% are not DE in reality. Controlling this error rate results in fewer false pos-

itives than controlling the per-gene Type-I error, while not being as restrictive as controlling the probability of any false positive (the family-wise error rate) among all of the potentially thousands of genes.

This control is usually taken for granted and often left out from the benchmarks of DEA methods, while in fact, an excessive FDR can be quite problematic. Not controlling the FDR means getting more false positives than expected, which limits the reproducibility of study results. Whole genome DEAs are usually exploratory steps prior to subsequent studies to confirm a gene signature is associated with a particular biological condition. If a majority of the selected genes turn out to be false positives, results may fail to replicate and any downstream health benefits may remain elusive, not to mention the waste of research resources.

When comparing DEA methods, the evaluation of their empirical FDR with respect to the targeted (nominal) level is often overlooked [Zhang et al., 2014; Tang et al., 2015; Seyednasrollah et al., 2015; Costa-Silva et al., 2017; Lamarre et al., 2018; Łabaj and Kreil, 2016]. Nonetheless, some issues with inflated FDR in DEA have been previously reported in the literature [Mazzoni et al., 2015; Rocke et al., 2015; Germain et al., 2016; Rigaill et al., 2018; Assefa et al., 2018], but those warnings have made little apparent impact on DEA practices.

Inflation of the empirical FDR in DEA can have numerous causes, from inadequate preprocessing of the data to violations of the DEA method's underlying assumptions. In particular, `edgeR`, `DESeq2` and `limma-voom` make potentially strong distributional assumptions on RNA-seq data. This type of model-based inference may be required when RNA-seq studies include only a small number of observations. However, these methods' parametric assumptions are not typically verifiable in practice. Any deviation from the hypothesized distribution of test statistics will translate into ill-behaved p-values and therefore uncontrolled FDR. FDR control rests upon the entire distribution of p-values being uniform under the null hypothesis H_0 (i.e. for genes that are truly not DE).

So even a slight deviation from strict Type-I error control can have dramatic consequences on the empirical FDR. In addition, even if Type-I error were controlled

at say 5%, non-uniformity in the p-value distribution under the null hypothesis could lead to failure to control the Type-I error at lower levels (such as 1% or lower) and/or failure to control the FDR. Larger sample sizes do not always solve issues with p-value distributions and FDR control arising from violation of modeling assumptions, and can sometimes even exacerbate the problem of misspecification and its consequences. As sequencing costs keep falling and study sample sizes are increasing, this issue needs to be addressed in large sample sizes as well.

Here, we propose `dearseq`, a new method to perform DEA that effectively controls the FDR, regardless of the distribution of the underlying data. `dearseq` is a robust approach that uses a variance component score test and relies on nonparametric regression to account for the intrinsic heteroscedasticity of RNA-seq data.

In the Results section we compare the performance of `dearseq` to the three most popular state-of-the-art methods for DEA: `edgeR`, `DESeq2` and `limma-voom`. We demonstrate that `dearseq` enforces strict control of Type-I error and FDR while maintaining good statistical power in a realistic and extensive simulation study where knowing the ground truth facilitates benchmarking the properties of the different methods. We also present a comparative re-analysis of a real-world Tuberculosis data set from Singhania et al. [2018] studying apparent false positives identified by the leading DEA methods compared to `dearseq`. `dearseq` can efficiently identify the genes whose expression is significantly associated with one or several factors of interest in complex experimental designs (including longitudinal observations) from RNA-seq data while providing robust control of FDR. `dearseq` is freely available as an R package on the Bioconductor library.

3.2 Methods

The general objective of DEA is to identify genes whose expression is significantly associated with a set of clinically relevant characteristics. `dearseq` is a new DEA framework based on a variance component score test [Lin, 1997; Huang and Lin, n.d.; Agniel and Hejblum, 2017], a flexible and powerful test that requires few assumptions

to guarantee rigorous control of Type-I and false discovery error rates. The method can be adapted to various experimental designs (comparisons of multiple biological conditions, repeated or longitudinal measurements, integrated supervision by several biomarkers at once). It builds upon recent methodological developments for the analysis of genomic data [Hejblum et al., 2015; Agniel and Hejblum, 2017; Agniel et al., 2018]. Variance component tests offer the speed and simplicity of classical score tests, but potentially gain statistical power by using many fewer degrees of freedom and have been shown to have locally optimal power in some situations [Goeman et al., 2006].

The `dearseq` method comprises 3 steps (with an optional initial normalization):

0. (optional) **normalize** gene expression across samples
1. **Estimate the mean-variance relationship** through a local linear regression borrowing information across all genes
2. **Test** each gene
3. Apply a **multiple-testing correction** controlling the FDR, such as the Benjamini-Hochberg procedure

3.2.1 Model specification

`dearseq` can be used to analyze longitudinal, grouped, or repeated measurements. Simplifications for a single observation per individual are given in 3.2.4. Let G be the total number of observed genes. Let r_i^g be the raw count and y_{ij}^g be the normalized gene expression of the g^{th} gene for the i^{th} sample at the j^{th} measure (any normalization can be used such as log-counts per million values, see Supplementary Materials for more details), for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Further, let X_{ij} be the p covariates to take into account and $\boldsymbol{\phi}_{ij}$ be the K variables we are interested in testing. $\boldsymbol{\phi}_{ij}$ contains the variables for DEA, such as disease status, treatment arm, other clinical characteristics which are to be associated with gene expression or any combination of continuous or binary measures that are under study.

To build a variance component score test statistic, we rely on the following working model for each gene g:

$$y_{ij}^g = \alpha_0^g + X_{ij}^T \boldsymbol{\alpha}^g + \boldsymbol{\phi}_{ij}^T \boldsymbol{\beta}^g + \boldsymbol{\phi}_{ij}^T \boldsymbol{\xi}_i^g + \epsilon_{ij}^g, \tag{3.1}$$

which can be factorized into matrix form as:

$$\boldsymbol{y}_i^g = \boldsymbol{\alpha}_0^g + X_i \boldsymbol{\alpha}^g + \Phi_i \boldsymbol{\beta}^g + \Phi_i \boldsymbol{\xi}_i^g + \boldsymbol{\epsilon}_i^g, \tag{3.2}$$

where, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ is a $n_i \times 1$ vector of normalized gene expression measurements, $\boldsymbol{\epsilon}_i \sim N(0, \Sigma_i)$ is a $n_i \times 1$ vector of measurement error, $\boldsymbol{\alpha}_0$ is a $n_i \times 1$ vector of intercepts, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_i \sim N(0, \Sigma_\xi)$ are respectively a $m \times 1$ vector of fixed effects and a $m \times 1$ vector of individual-level random effects of the variables of interest, X_i and Φ_i are the associated $n_i \times p$ matrix of covariates and $n_i \times m$ matrix of the variables of interest. Σ_ξ is the $m \times m$ covariance matrix of $\boldsymbol{\xi}_i$. Σ_i is the $n_i \times n_i$ covariance matrix of measurement errors. $\boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_i$ are assumed to be independent. Note that, to take into account the correlation between the different measurements of the same individual we use a random effect in the model. In addition, it is important to note that the variance of the residuals depends on i and j to model the heteroscedasticity of the data. This means that each measure of each individual has a different variance. Note that the method does not require a contrast matrix and can perform DEA across multiple conditions, as well as test the association of gene expression with continuous variables, or even a group of variables (continuous or categorical) at once.

3.2.2 Estimation of the mean-variance relationship

A key step in our method is the estimation of $\Sigma_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{in_i}^2)\ \forall i$, and for each gene, which will be useful for calculating the test statistic. Because of the intrinsic heteroscedasticity of the data, the variance of the residuals depends on i and j. To estimate the mean-variance relationship in $\mathbf{y}^g$, we borrow information across all G

genes to be able to estimate observation-specific variances. Let $v_{ij}^g = Var(y_{ij}^g \mid X_{ij}, \boldsymbol{\xi}_i^g)$ and $m_{ij}^g = E(y_{ij}^g \mid X_{ij}, \boldsymbol{\xi}_i^g)$ respectively the variance and the mean of gene g for sample i and measure j given the covariates and the random effects. We assume that v_{ij}^g may be modeled as a function of its mean m_{ij}^g. To save computational time and reduce the number of points used in the nonparametric fit, one could follow Law et al. [2014] and model the mean-variance relationship at the gene level. Specifically, $v^g = \omega(m^g) + e^g$ for some unknown function $\omega(\cdot)$ and errors which follow the moment conditions $E(e^g) = 0, V(e^g) = \tau^2$, $\tau > 0$. Thus, we used a local linear regression proposed by Wasserman [2006] which offers good asymptotic convergence. For practical reasons, we further added the two following steps:

1. Because we use the same window bandwidth h for all observations in kernel estimation, and in order to avoid over-fitting at rare expression levels (usually extremely high or low expression levels are encountered less often), we first perform a transformation of the data so that all observation neighborhoods are properly populated: $\widetilde{m}^g = f(m^g) = \Phi\left(\frac{m^g - \overline{m}}{s_m}\right)$ where $\overline{m}$ is the average observed expression level and s_m is the standard deviation of the gene average expression, $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution $\mathcal{N}(0, 1)$.

2. In order to remove the possibility of negative weights, we smooth over the *log*-transformed squared-errors $\widetilde{s}^g = \log(v^g)$ rather than the natural variances.

The full local linear regression for weight estimation performed is then:

$$\overline{m} = \frac{1}{G}\sum_g m^g\,, \quad \text{and} \quad s_m = \sqrt{\frac{1}{G-1}\sum_g (m^g - \overline{m})^2}$$

$$\widetilde{s}^g = \log(v^g)\,, \quad \text{and} \quad \widetilde{m}^g = f(m^g) = \Phi\left(\frac{m^g - \overline{m}}{s_m}\right)$$

$$\widetilde{S}_{nd}(x) = \sum_g K\left(\frac{\widetilde{m}^g - x}{h}\right)(\widetilde{m}^g - x)^d, \text{ for } d = 1, 2$$

$$\widetilde{b}^g(x) = K\left(\frac{\widetilde{m}^g - x}{h}\right)(\widetilde{S}_{n2}(x) - (\widetilde{m}^g - x)\widetilde{S}_{n1}(x)),$$

$$\widetilde{l}^g(x) = \frac{\widetilde{b}^g(x)}{\sum_g \widetilde{b}^g(x)}, \quad \widetilde{\omega}_n(x) = \exp\left(\sum_g \widetilde{l}^g(x)\widetilde{s}^g\right) \tag{3.3}$$

for some kernel function $K(\cdot)$ and bandwidth $h > 0$. Standard cross-validation techniques can be used to select h in practice.

Because the mixed effects model (3.1) may be computationally costly, we restrict ourselves to the fixed effect part of the model for estimating the mean-variance relationship:

$$\boldsymbol{y}_i^g = \alpha_0^g + X_i^T\boldsymbol{\alpha}^g + \Phi_i^T\boldsymbol{\beta}^g + \widetilde{\varepsilon}_i^g. \tag{3.4}$$

Based on this model, the mean-variance relationship could be estimated by $\widehat{\omega}_n(x) = \widetilde{\omega}_n(x)_{|m^g=\widehat{m}^g, v^g=\widehat{v}^g}$ with the estimate of the mean
$\widehat{m}^g = n^{-1}\sum_{i=1}^n n_i^{-1}\sum_{j=1}^{n_i}\widehat{\alpha}_0^g + X_{ij}^T\widehat{\boldsymbol{\alpha}}^g + \Phi_{ij}^T\widehat{\boldsymbol{\beta}}^g$ and the estimate of the variance $\widehat{v}^g = n^{-1}\sum_{i=1}^n n_i^{-1}\sum_{j=1}^{n_i}(y_{ij}^g - \widehat{\alpha}_0^g - X_{ij}^T\widehat{\boldsymbol{\alpha}}^g - \Phi_{ij}^T\widehat{\boldsymbol{\beta}}^g)^2$ where $\widehat{\boldsymbol{\alpha}}^g$ and $\widehat{\boldsymbol{\beta}}^g$ are estimated with Ordinary Least Squares.

Now that we have the estimate of ω_n, we can calculate the variance estimate of y_{ij}^g for all P genes as: $(\widehat{\sigma}_{ij}^g)^2 = \widehat{\omega}_n\left(\widehat{f}\left(\widehat{m}_{ij}^g\right)\right)$ with $\widehat{m}_{ij}^g = \widehat{\alpha}_0^g + X_{ij}^T\widehat{\boldsymbol{\alpha}}^g + \Phi_{ij}^T\widehat{\boldsymbol{\beta}}^g$.

3.2.3 Variance component score test statistic estimation

In this section, we derive a variance component score test statistic for the effects of interest. For the sake of simplicity, we omit the gene index g in the following, bear in mind that a test is carried out for each gene g.

According to the model (3.2), the null hypothesis of no effect of interest is:

$$H_0 : \text{“}\boldsymbol{\beta} = 0 \text{ and } \Sigma_\xi = 0\text{”} \tag{3.5}$$

If the variance-covariance matrix of the random effects is identically zero then the random effects $\boldsymbol{\xi}_i$ are also identically zero for all i. If at the same time, $\boldsymbol{\beta} = 0$ then

the expression of the gene will not be significantly associated with the variables of interest Φ_i.

Let Q be the variance component score test statistic such as $Q = \boldsymbol{q}^T\boldsymbol{q}$ with

$$\boldsymbol{q}^T = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\Phi_i = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \mu_i)^T\Sigma_i^{-1}\Phi_i$$

Considering that we have a consistent estimator of Σ_i, we still must provide estimates of α_0 and $\boldsymbol{\alpha}$. A natural way to estimate these quantities, given the heteroscedasticity in $\boldsymbol{y}$, is to fit a weighted mixed effects model. The weights are taken to be $\boldsymbol{w}_i = \text{diag}(\widehat{\Sigma}_i)^{-1}$. However, to avoid excessive computation time, instead of estimating the full mixed effects model from (3.2), we may fit a simpler fixed effects model (3.4), from which we can obtain estimates of α_0 and $\boldsymbol{\alpha}$.

3.2.4 Simplification when the measurements are not repeated

In this section, we detail how the generic formulation of the variance component score test simplifies into the form when the data are not repeated. When there is only one observation per individual and only one variable of interest (i.e., ϕ_{ij} is a scalar), the variance component score test simplifies to a standard score test. When there are multiple variables of interest and Φ_{ij} is a vector, the variance component score test may gain additional statistical power thanks to its exploitation of potential correlation among the tested variables. Here, we assume that the data are not grouped (e.g. repeated or longitudinal) and therefore the index j has to be removed, as used in the main manuscript. Thus, let y_i^g be the normalized gene expression of the g^{th} gene for the i^{th} sample. The working model is written as follows:

$$y_i^g = \alpha_0^g + X_i\boldsymbol{\alpha}^g + \Phi_i\boldsymbol{\beta}^g + \varepsilon_i^g \tag{3.6}$$

where $\varepsilon_i^g \sim N(0, (\sigma_i^g)^2)$, α_0^g is the intercept, X_i is a vector of p observations from covariates that needs to be adjusted, $\boldsymbol{\alpha}^g$ is the corresponding vector of p fixed effects, Φ_i is a vector of K observations from the variables of interest (with whom the ex-

pression association is tested), and $\boldsymbol{\beta}^g$ is the corresponding K vector of fixed effects associated to those variables of interest. The variance of the residuals depends on i to model the heteroscedasticity of the observations $\boldsymbol{y}$. The parameter of interest is $\boldsymbol{\beta}^g$: if $\boldsymbol{\beta}^g \neq \mathbf{0}$, then the gene is differentially expressed. The variance of the residuals ε_i^g depends on i to model the heteroscedasticity inherent to RNA-seq data.

According to the working model (3.6), a gene has its expression associated with the variable(s) of interest in Φ if $\boldsymbol{\beta}^g \neq 0$. `dearseq` thus tests the following null hypothesis:

$$H_0^g : \boldsymbol{\beta}^g = 0$$

The associated variance component score test statistic can be written as $Q^g = \boldsymbol{q}^{gT}\boldsymbol{q}^g$ with

$$\boldsymbol{q}^{gT} = n^{-1/2} \sum_{i=1}^{n} (y_i^g - \mu_i^g)(\sigma_i^g)^{-1}\Phi_i,$$

where μ_i is the conditional mean normalized expression given the covariates X_i.

Because of their count nature, RNA-seq data are intrinsically heteroscedastic. We model this mean-variance relationship through σ_i^g. But obviously, this individual variance cannot be estimated from a single observation. Instead, we adopt the same strategy described section 3.2.2, but omitting j index, to estimate $\widehat{\sigma_i}^g$.

3.2.5 Asymptotic and permutation tests

The asymptotic distribution of the test statistic Q can be shown to be a mixture of χ_1^2 random variables

$$Q \to \sum_{l=1}^{n_i} a_l \chi_1^2 \tag{3.7}$$

where the mixing coefficients a_l depend on the covariance of $\boldsymbol{q}$ (see supplementary materials for details). This asymptotic result rests solely upon the Central Limit Theorem, and this is why `dearseq` is particularly robust to misspecification: the distribution of Q is the same whether the model (3.2) holds or not. Therefore, the Type-I error and the FDR are controlled as long as the Central Limit Theorem is in action (meaning n is large enough).

One advantage of using a variance component score test over a regular score test is the gain in statistical power, that comes from exploiting the correlation among $\boldsymbol{\beta}^g$ coefficients to potentially reduce the degrees of freedom of the test. Another advantage is its flexibility that can accommodate random effects in the model to test mixed hypotheses (see supplementary materials for details).

To overcome the shortcomings of this asymptotic test in small samples, we propose to use a permutation test using the same statistic Q. Since we are in multiple testing setting it is of the utmost importance to carefully compute the associated p-values [Phipson and Smyth, 2010] before applying the Benjamini-Hochberg correction. Finally, in order to preserve statistical power, we use Phipson & Smyth's correction to account for random permutations (see supplementary materials).

Availability of data and materials

`dearseq` is freely available on Bioconductor at `https://bioconductor.org/packages/release/bioc/html/dearseq.html`. The sequence data set from the Singhania *et al.* Tuberculosis study analyzed in this article is accessible from the NCBI GEO database with the primary accession code GSE107991. The code used to analyze the data set and the results are available from the GitHub repository (https://github.com/borishejblum/dearseq)

Software versions

All computations were run under R v3.6.1 using `DESeq2` package v1.25.11, `edgeR` package v3.27.13, `limma` package v3.41.16, and `dearseq` package v0.99.8.

3.3 Results

3.3.1 Synthetic simulation study

As highlighted by both Conesa et al. [2016] and Assefa et al. [2018], engaging in realistic yet clear simulations of gene expression is difficult. One has to find the right balance between the controlled settings necessary to know the ground truth,

and the realism necessary to be convincing that the results would translate in real-world analyses. In an attempt to cover as broad a spectrum as possible, we present a performance evaluation of our methods under four data-generating scenarios: a) a Negative Binomial parametric assumption for RNA-seq data, b) a highly non-linear model designed to violate most modeling assumptions, c) a resampling from SEQC data [SEQC/MAQC-III Consortium, 2014] with truncated Gaussian noise, and d) a data-driven Negative Binomial parametric assumption. Simulations a) and b) were designed to drastically depart from the models used for all methods whereas simulations c) and d) aim to be more realistic. Scenario a) may be more favorable to `edgeR` and `DESeq2` as it relies on their parametric assumption of a Negative Binomial distribution for RNA-seq count data. Scenario b) may be unfavorable for all three compared methods (`edgeR`, `DESeq2` and `limma-voom`) since it features a high degree of non-linearity, deviating from any assumed model. Scenario c) is likely the most realistic of the three because it relies only on resampling real RNA-seq samples from the SEQC study [SEQC/MAQC-III Consortium, 2014], similarly to what was done in Germain et al. [2016]. A multivariate truncated Gaussian noise (using the estimated covariance stucture across the observed genes) was added to enable the generation of larger sample sizes while preserving the count nature of the data. Scenario d) relies on a Negative Binomial distribution whose parameters are estimated using data from Singhania et al. [2018] in order to avoid using arbitrary settings. Like scenario a), it favors both `edgeR` and `DESeq2` as they both rely on the Negative Binomial distribution assumption for the counts.

We simulated 1,000 synthetic data sets at different sample sizes using each one of these four scenarios. For scenarios a) and b), 0.5% of genes were generated as truly DE while the remaining 99.5% were not DE, among 10,000 genes. For the scenario c), since it is based on resampling from homogeneous samples, it was impossible to induce truly DE genes without making further parametric assumptions (which would have made the scenario less realistic). For this reason, in scenario c), FDR corresponded to the probability of finding any genes to be DE. In scenario d), 5% of the genes were generated as truly DE while the remaining 95% were not DE, among 10,000 genes.

Details of the data-generating mechanisms are provided in supplementary materials.

We evaluated the four methods (the leading methods and `dearseq`) in terms of Type-I error control and statistical power, as well as in terms of FDR and True Discovery Rate (TDR) after Benjamini-Hochberg correction [Benjamini and Hochberg, 1995] for multiple testing. Throughout, we used a targeted control rate for the FDR at a nominal level of 5%. Fig. 3-1 presents the Monte-Carlo estimation over the 1,000 simulations in each of the three scenarios for both the Type-I error and the FDR according to increasing samples sizes (from 4 to 200 samples). Fig. 3-2 presents the results of the first two scenarios and the data-driven Negative Binomial one for both the statistical power and the TDR. For all `edgeR`, `DESeq2` and `limma-voom` analyses we used the default values and followed the guidelines from their respective online user-guides.

In Fig. 3-1, `dearseq` exhibited good control of both Type-I error and FDR in all four scenarios, as soon as asymptotic convergence was reached (between 16 and 50 samples depending on the scenarios). To accommodate small sample sizes, we have also developed a permutation-based version of `dearseq`, which always controlled Type-I error and FDR, regardless of the sample size. `edgeR` appeared to control the Type-I error in scenarios a), c) and d) onwards, but exhibits slightly inflated Type-I error for large sample sizes in scenario a) (from 100 samples). This was much more visible for the FDR, which `edgeR` failed to control as soon as the sample size rose above 50. Under the non-linear model (b), neither the Type-I error nor the FDR are controlled by `edgeR`. `limma-voom` exhibited good control of both the Type-I error and FDR as long as its linear hypothesis was not violated (i.e., not scenario b)) and the sample size was large enough (between 8 to 50 samples depending on the scenario). Finally, `DESeq2` failed to control either the Type-I error or the FDR in scenarios a), b) and c), with its problems worsening as the sample size increased. In scenario d), `DESeq2`, `limma-voom` and `edgeR` exhibited a high FDR at small sample sizes while the permutation-based version of `dearseq` controlled the FDR for both small and large sample size.

Fig. 3-2 shows that this robust control of Type-I error and FDR from `dearseq` does

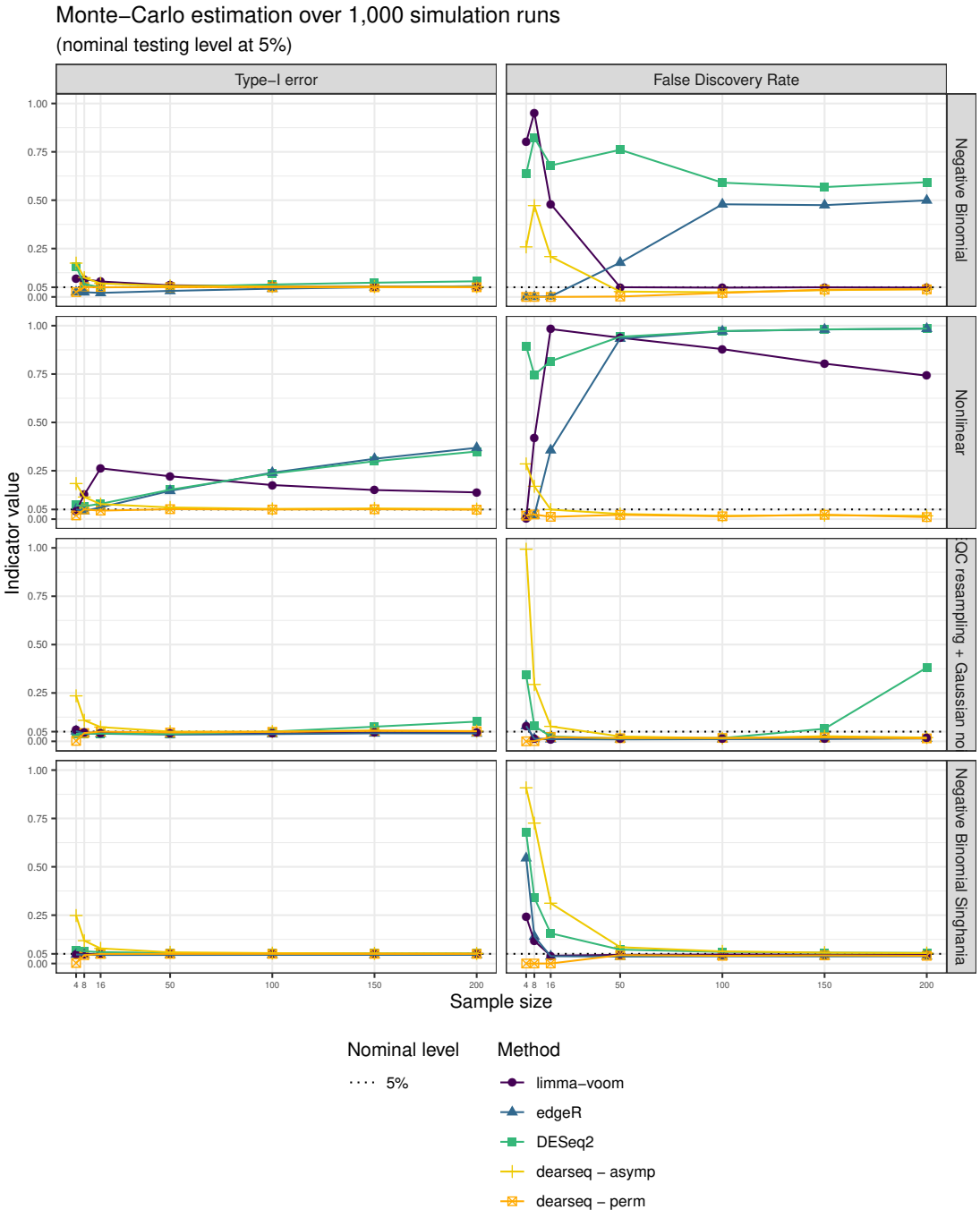


Figure 3-1: **Type-I error and FDR curves for each DEA method with increasing sample sizes** In each setting (Negative Binomial, Non-linear, SEQC data resampling and data-driven Negative Binomial), the Type-I error is computed as the number of significant genes among the true negative, and the FDR as the average number of false positives among the genes declared DE.

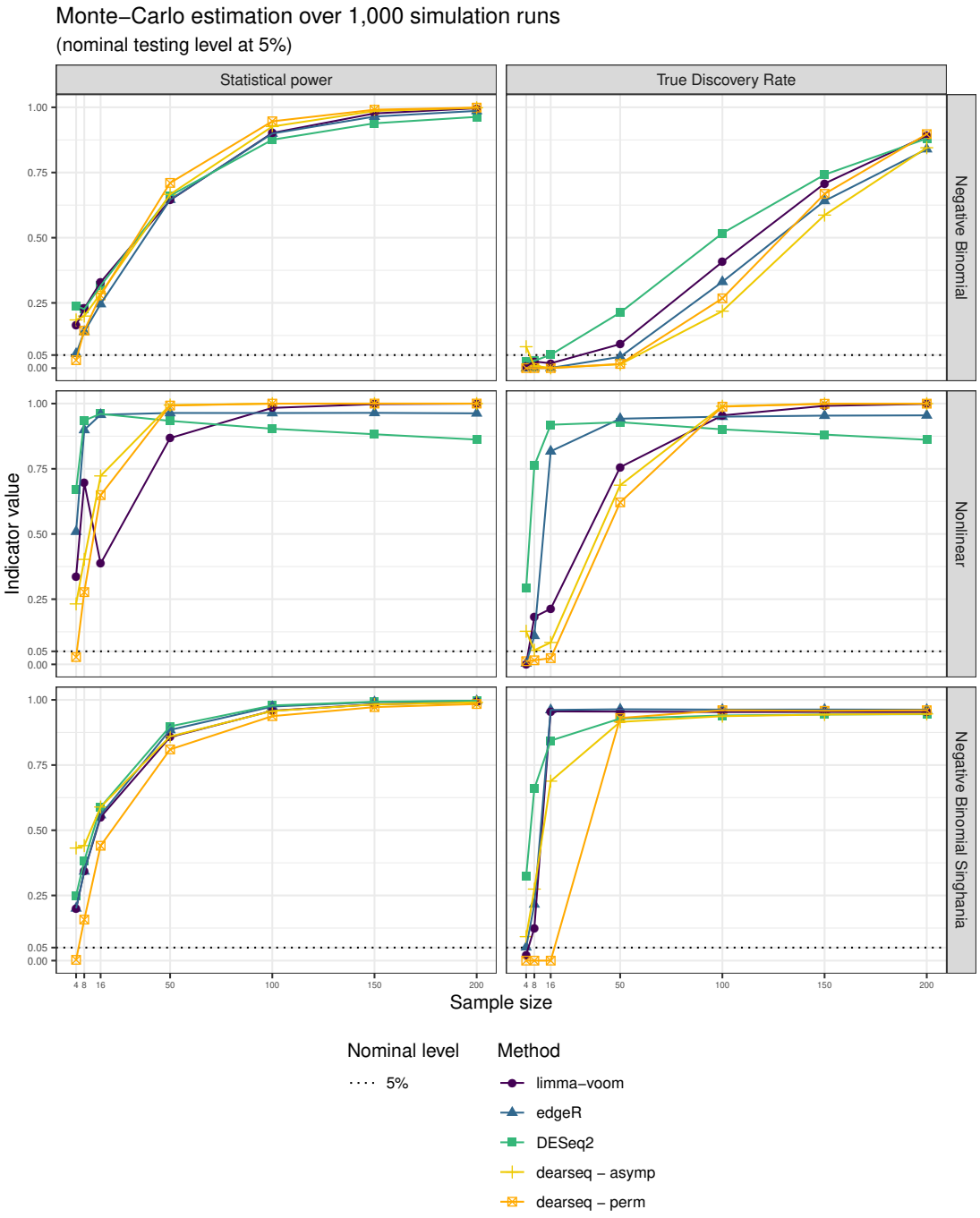


Figure 3-2: **Power and True Discovery Rate curves for each DEA method with increasing sample size** Because SEQC data resampling only generates non-significant genes, this setting does not allow to estimate statistical power or TDR.

not come at a price of reduced statistical power (or True Discovery Rate, its multiple-testing correction equivalent). Interestingly, the permutation approach also exhibits good statistical power (regarding competing approaches, interpreting statistical power when the Type-I error is not controlled would be dubious).

3.3.2 Real data set

In a recent paper, Singhania *et al.* identified a 373-genes signature of active tuberculosis from RNA-seq data [Singhania et al., 2018]. Tuberculosis (TB) is a disease caused by a bacterium called Mycobacterium tuberculosis. Bacteria typically infect the lungs, but they can also affect other parts of the body. Tuberculosis can remain in a quiescent state called latent tuberculosis infection (LTBI), where the patient is infected but has no clinical, bacteriological or radiological signs of the disease. Participants to this study were recruited from several medical institutes in London, UK (see Berry et al. [2010] for a detailed description). All participants were aged over 18 years old. Active TB patients were confirmed by laboratory isolation of M. tuberculosis on mycobacterial culture of a respiratory specimen, while Latent TB patients were characterized by a positive tuberculin-skin test (TST) together with a positive result using a M. tuberculosis antigen specific IFN-γ release assay (IGRA). Healthy control participants were recruited from volunteers at the National Institute for Medical Research (NIMR, Mill Hill, London, UK) and were negative to both TST and IGRA. In total, 54 participants were included, of whom 21 were active TB patients, 21 were LTBI patients, and 12 were healthy controls.

The signature was derived by contrasting active tuberculosis (TB) patients on the one hand against healthy individuals (Control) or those with a latent infection (LTBI) on the other hand (see Fig. 3-3). Their original analysis applied `edgeR` to their Berry London RNA-seq data, which included 14,150 normalized-gene counts measured across 54 samples after preprocessing (see Singhania et al. [2018] or supplementary materials for more information on this preprocessing) available from GEO (GSE107991). In light of our simulation results, we sought to explore if the signature Singhania *et al.* found using `edgeR` was likely to contain false positives. We therefore

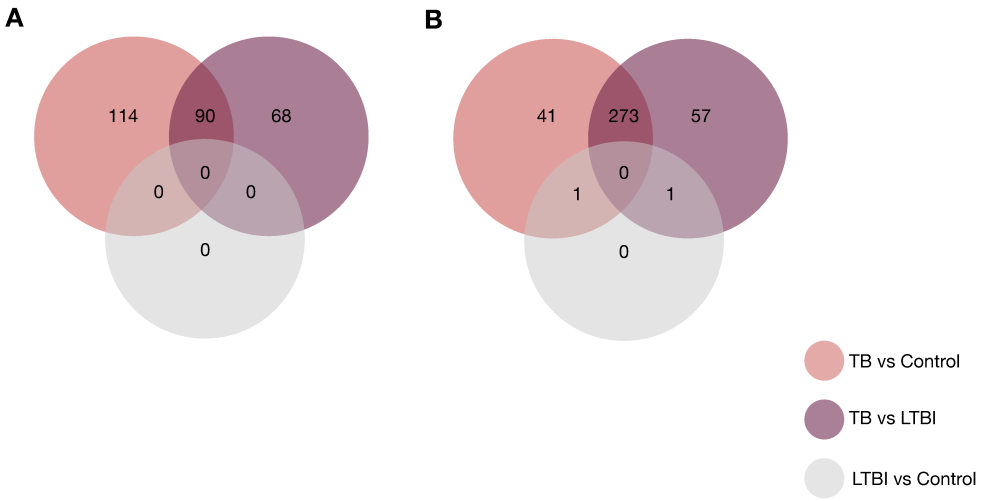


Figure 3-3: **Venn diagram showing overlap of DE genes using `dearseq` and the original `edgeR` signature among the three comparisons performed** a) Venn diagram showing the results of the three DEA using `dearseq`. Note that no gene differentially expressed was found with our method comparing the LTBI group and the control group, unlike `edgeR` which found two such genes to be DE. b) Venn diagram showing the results of the DEA using `edgeR` (Singhania *et al.*).

conducted a comparative re-analysis of these data, first comparing DE genes found by `dearseq` to the original signature of Singhania *et al.*. Secondly, we further compared the results obtained from the other leading methods, `DESeq2` and `limma-voom`.

Following Singhania *et al.*, to be included in the signature a DE gene g must have had both: i) an absolute $\log_2$(fold change) > 1, and ii) an FDR adjusted p-value < 0.05 (after correction for multiple testing with the Benjamini-Hochberg procedure). To ensure reproducibility of the numerical values from Singhania *et al.*, the $\log_2$ fold changes were calculated using `edgeR`. The signature was then evaluated by its capacity to distinguish between active TB versus all others. In order to quantify the relevance of each selected gene for distinguishing active TB from control and LTBI, we computed two measures of association, the leave-one-out cross-validated Brier score [Brier, 1950] and the marginal p-value for the association between the gene and TB status. The Brier score was computed as $BS_g = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\pi}_{gi}^{TB} - \mathbb{1}_{i\in TB})^2$. It compares each patient's TB status to $\widehat{\pi}_{gi}^{TB}$, their predicted probability of TB based on the selected gene g estimated using leave-one-out cross validation. A gene with Brier score BS_g close to 0 is a good predictor of TB, while a gene with Brier score far away from 0 is a poor predictor and potential false positive. Similarly, we compute the marginal p-value for each gene from a logistic regression predicting TB status from the gene expression. We estimate the Brier score and p-value for each gene separately. We do this rather than a multivariate model including all genes because the presence of a single predictive gene in the multivariate signature would be enough to yield accurate predictions, thus masking the potential false positive genes included in the model.

Applying the `dearseq` permutation test (see Methods) to the three comparisons originally performed in Singhania *et al.* (TB vs Control, LTBI vs Control, and TB vs LTBI) yields a global signature of 272 DE genes (see Fig. 3-3) of which 231 are in common with those found by the original `edgeR` analysis (see Fig. 3-5). We isolated the genes only identified by `dearseq` from the genes only identified by `edgeR` and from the genes in common between the two signatures to further assess the differences between the two results. Comparing the gene specific Brier scores BS_g between the two

signatures clearly shows that the overwhelming majority of the highest scores (i.e. the lowest predictive abilities) is due to `edgeR`-private genes (see Fig. 3-4 b)). Indeed, the univariate Brier scores of the `dearseq`-private genes have significantly smaller values on average than the `edgeR`-private genes (according to a t-test – see Fig. 3-4). This is further confirmed by the marginal association p-values, for which all of the highest values are again from `edgeR` private, notably all the values above 0.05. Thus, `edgeR`-private genes are likely false positives whereas the `dearseq`-private sound more relevant. From a biological point of view, the main pathways concerned by the 142 `edgeR`-private genes, that are "Inhibition of matrix metalloproteinases", "Granulocyte Adhesion and Diapedesis", "Inhibition of Angiogenesis by TSP1" using Ingenuity Pathway Analysis (IPA) were not directly related to the main pathways observed in the retained 373-gene signature (IFN-inducible genes, B- and T-cell genes) although interferon signalling pathway was represented by two genes and some upstream regulators. In contrary, among the 41 `dearseq`-private genes, HERC5 is upregulated by regulators belonging to IFN signaling pathways including IFN$\alpha\beta$ and IFNϵ (known to be regulated by Mycobacterium tuberculosis). Those results emphasize the better predictive ability of the genes identified by `dearseq`, and highlights the potential false positives arising from `edgeR`.

In addition, we performed the same analysis using `limma-voom` and `DESeq2` to further benchmark the performance of `dearseq`. For `DESeq2` and `limma-voom` we used the default values following the guidelines from their respective online user-guides. For `edgeR`, we rely on the results directly provided by Singhania et al. [2018] – see supplementary materials for details. Fig. 3-5 displays the Venn diagram of significantly DE genes across these four analyses. There are 228 genes common across all these tools. Interestingly, all of the 272 genes identified by `dearseq` are also identified by at least one of the three competing methods (and only 2 genes are identified by less than two other methods – namely only by `DESeq2`), illustrating that `dearseq` is less prone to generate false positives. `DESeq2` identifies the largest signature comprising of 471 genes, including all of the 272 genes identified by `dearseq` and 360 out of the 373 originally identified by `edgeR`, while `limma-voom` identify 402

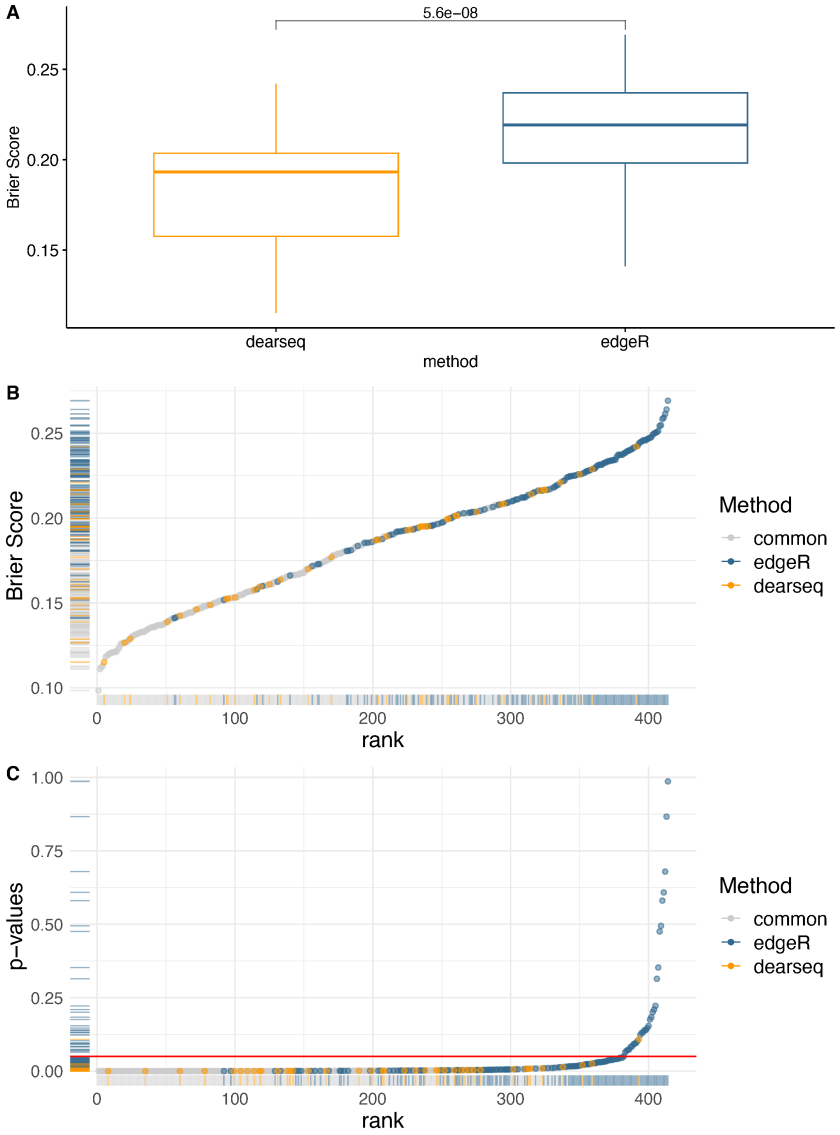


Figure 3-4: **Comparing `edgeR`-based signature to the signature derived by `dearseq`** a) Boxplots of the Brier scores of the 41 genes private to `dearseq` (i.e., not also declared DE by `edger`) and the 142 genes private to the original `edgeR` analysis. b) Univariate Brier scores. The blue points correspond to genes found only in the original `edgeR` signature, the yellow points found only in the `dearseq` signature, and the grey points found in both signatures. c) Marginal p-values from a univariate logistic regression combined with a leave-one-out cross validation for the 40 `dearseq`-private and the 142 `edgeR`-private genes. The red line indicates the common 5% p-value threshold.

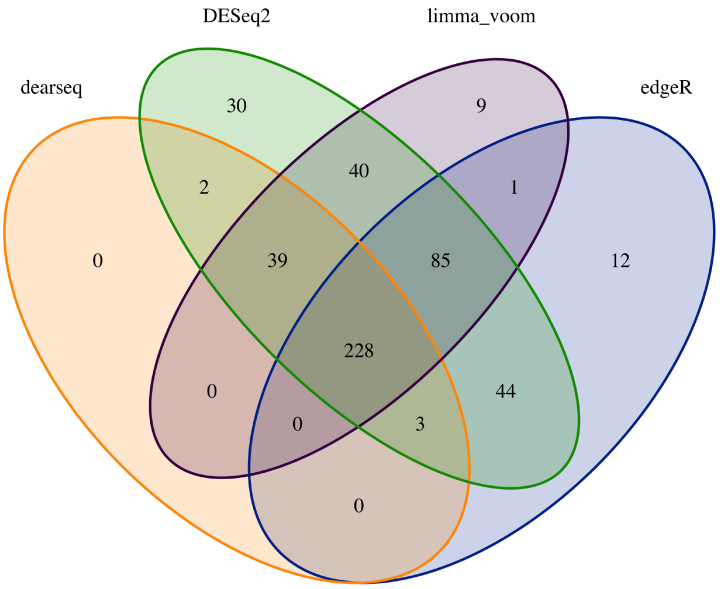


Figure 3-5: **Venn diagram summarizing the different signatures from the four methods.** Venn diagram of the genes declared DE by `dearseq`, `DESeq2`, `limma-voom` and `edgeR` (Singhania *et al.*) under an FDR-adjusted p-value of 0.05. None of the genes is found with `dearseq` only.

genes, among which 267 are in common with `dearseq` and 314 are in common with `edgeR`. As can be seen on Fig. 3-6 the `dearseq` signature has the lowest average Brier score, meaning that most of the additional genes identified by the three competing methods are less predictive of active TB status. Fig. 3-7 strengthens this conclusion by showing again that the `limma-voom`-private are largely over-represented among the highest Brier scores and the highest marginal p-values. The same conclusion can be drawn for the `DESeq2`-private genes.

3.4 Discussion

The proposed method `dearseq` represents an innovative and flexible approach for performing gene-wise analysis of RNA-seq gene expression measurements with complex design. As demonstrated in our simulation study, `edgeR`, `DESeq2` or `limma-voom` can

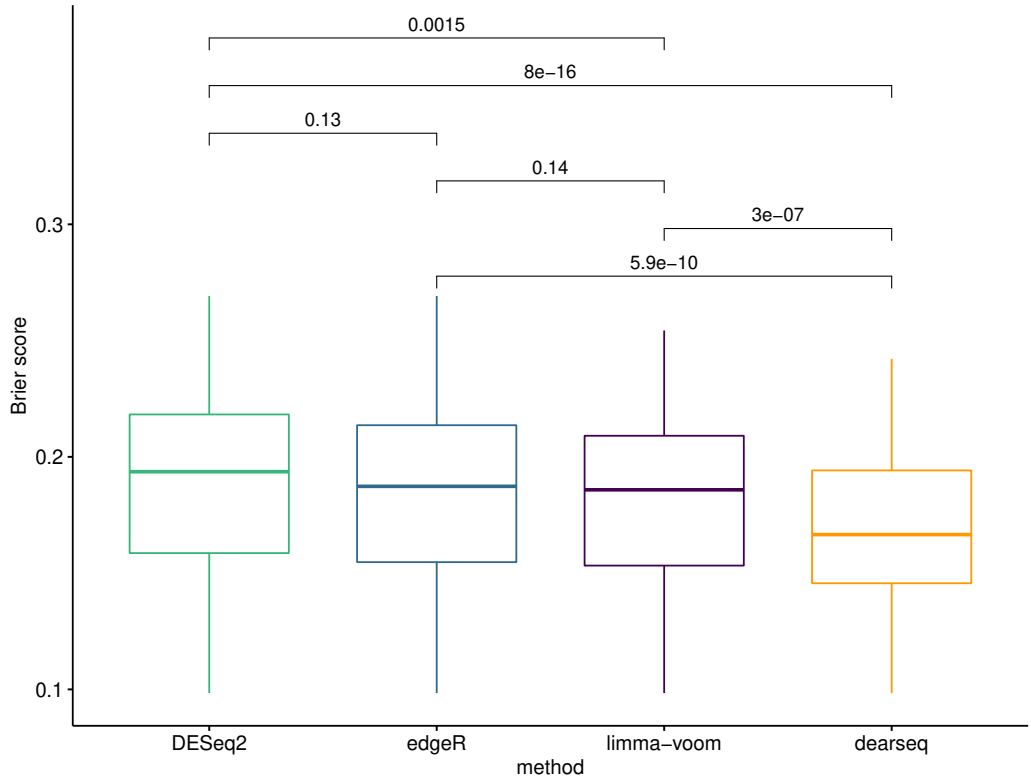


Figure 3-6: **Boxplots of the Brier scores of all the genes declared DE by the four methods.** Boxplots of the Brier scores of all the DE genes called by `dearseq`, `DESeq2`, `limma-voom` and `edgeR` (Singhania *et al.*). The predictions are derived from a logistic regression combined with a leave-one-out cross validation. Smaller Brier scores are better.

all fail to control the Type-I error and the FDR when the sample size increases, while our method behaves correctly. Moreover, the re-analysis of the London Berry Tuberculosis data set revealed that the differentially expressed genes identified by `dearseq` are highly predictive of active Tuberculosis status, while results from the three state-of-the-art methods (including the original `edgeR` analysis) likely include numerous false positives. While Agniel and Hejblum [2017] focused on the analysis of longitudinal data and only considered gene set analysis, here we introduce a much broader framework which generalizes the previous mathematical results beyond longitudinal studies and gene set analysis – most notably `dearseq` allows gene-by-gene analysis and can accommodate many different experimental designs including the usual two (or more) group(s) comparison.

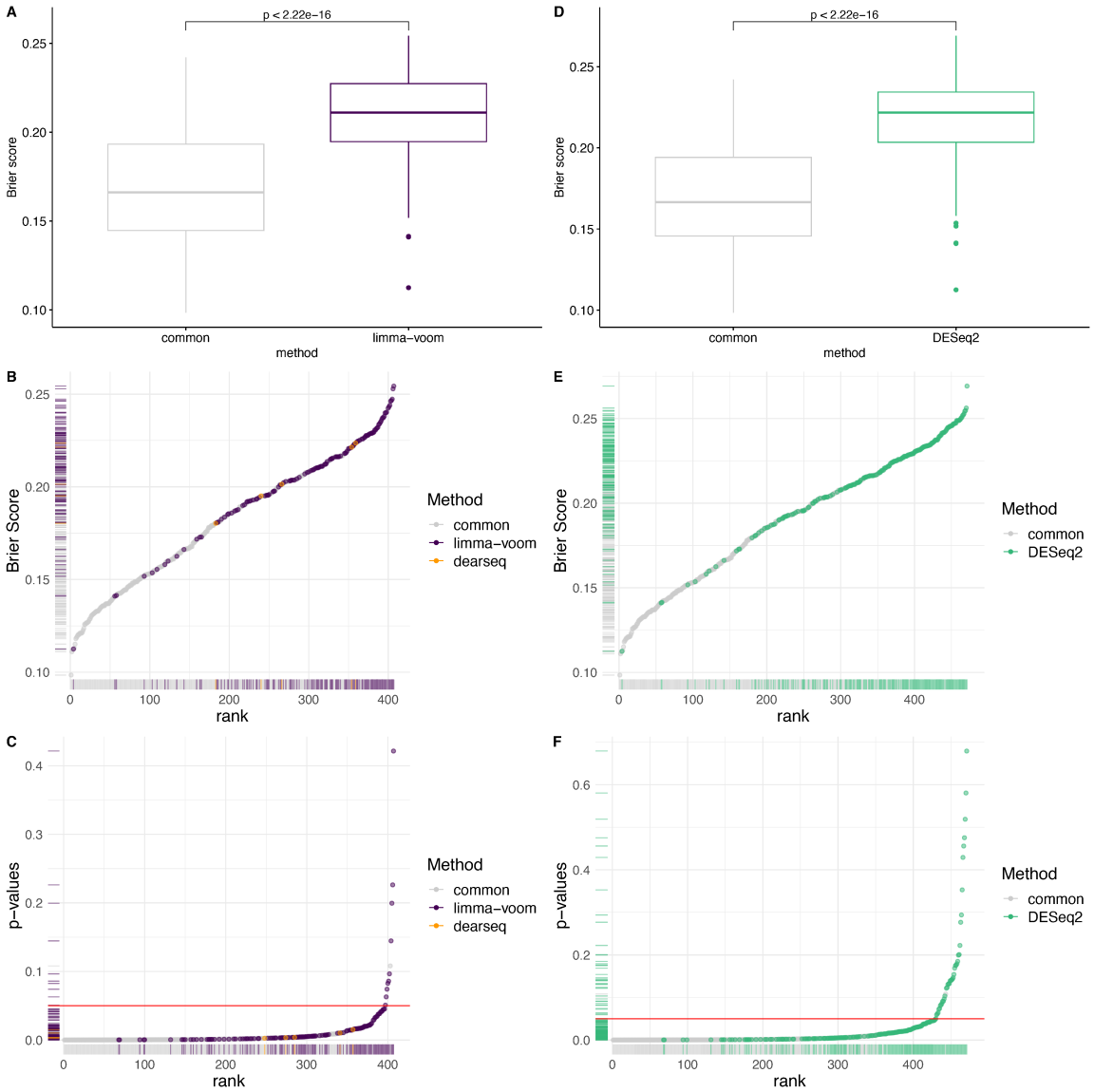


Figure 3-7: **Comparison of the `dearseq` derived signature to both the `DESeq2` and `limma-voom` derived signatures** a) Boxplots of the Brier scores of the DE genes private to `limma-voom` and the DE genes common to both `dearseq` and `limma-voom`. Note that only 5 genes are identified only by `dearseq` and not `limma-voom`. Therefore we exclude the associated boxplot. b) Univariate Brier scores. The purple points correspond to the DE genes called by `limma-voom` and the grey points to the genes common with `dearseq`. c) Marginal p-values. d) Boxplots of the Brier scores of the DE genes private to `dearseq` and the DE genes common to both `dearseq` and `DESeq2`. All genes declared DE by `dearseq` were also declared DE by `DESeq2`. e) Univariate Brier scores. The green points correspond to the DE genes called by `DESeq2` and the grey points to the genes common with `dearseq`. All genes declared DE by `dearseq` were also declared DE by `DESeq2`. f) Marginal p-values.

It is important to note that `edgeR`, `DESeq2` or `limma-voom` will not systematically have inflated FDR. As illustrated by our simulation studies, there are some scenarios in which, for some given sample sizes, they control the FDR adequately. However, we have shown here that this is no guarantee, and in practice it is very difficult to know under which circumstances a data analysis is taking place.

Because `dearseq` solely relies on the Central Limit Theorem convergence for its asymptotic test to work, it guarantees a control of the FDR without needing any model to hold as long as the sample size is large enough. This contrasts with `limma-voom` which hinges on weighted least squares and can yield incorrect inference if the heteroskedasticity in the variances are not modeled correctly [Romano and Wolf, 2017]. For lower sample sizes, where convergence is not reached, a robust permutation test can be used instead. Using Phipson and Smyth [2010]'s correction, it adequately controls the FDR regardless of the sample size and exhibits good statistical power in our simulation study. An alternative approach to permutation tests would be to use a Bayesian estimate of the posterior p-value with either a uniform prior or a Jeffreys' prior for instance. Considering m the number of permutations, and b the number of successes, the p-value is then equal to $\frac{b+1}{m+2}$ or $\frac{b+1/2}{m+1}$, which is also never equal to 0. Notice when K is large, this ends up very close to the unbiased estimator from Phipson and Smyth [2010]. Besides, this permutation test introduces a trade-off between the numerical precision (i.e. the number of permutations performed) and the associated computational time. We have undertaken substantial efforts to speed up the implementation of the `dearseq` R package on Bioconductor, which allows for parallel computing. This leads to reasonable computation times (from a few seconds to no more than a couple of minutes on a laptop) depending on the sample size and the computing power available. If more permutations were deemed essential, one option would be to selectively increase the number of permutations only for the genes where numerical precision is not sufficient to confidently estimate their adjusted p-value, thus limiting the computational burden.

Among the three state-of-art methods compared here, `DESeq2` seems to fail to control the FDR most often. In particular, even under its model assumption of

a Negative Binomial distribution for the data, it can suffer from inflated FDR. This seems counter-intuitive as our synthetic data were generated under the Negative Binomial distribution, and this should advantage `DESeq2` and `edgeR` – since both methods assume this model. As has been noted previously, this behavior can be caused by non-uniformity in the distribution of the p-values arising from `DESeq2` or `edgeR` (especially when combined with a multiple testing correction such as the Benjamini-Hochberg procedure) [Burden et al., 2014; Rocke et al., 2015; Yang et al., 2016; León-Novelo et al., 2017]. In addition, it should be noted that this behavior is not expected to be universal and other parameterizations of the Negative Binomial generative model could lead to better performance for these methods.

DEA can have numerous preprocessing steps, and the various possibilities can complicate the fair comparison of different methods. Since here our primary goal was to compare to the original `edgeR` analysis, we used the `edgeR`-preprocessed data as input to `dearseq` (`dearseq` does not rely on a specific preprocessing step and only requires that gene expression measurements are comparable across samples – all preprocessing regarding systematic bias or batch effect must be performed beforehand with any procedure deemed appropriate). For `DESeq2` and `limma-voom` we used the raw counts. Indeed both `edgeR` and `DESeq2` assume the input data to be strictly counts (i.e. integers), due to their Negative Binomial distribution assumption, though `edgeR` also has some support for so-called "non-integer counts". While this seems sensible given the nature of RNA-seq data, recent innovations in RNA-seq alignment methods such as `salmon` [Patro et al., 2017] or `kallisto` [Bray et al., 2016] return pseudocounts that are not integers. If the loss of precision is likely not severe when rounding up pseudo-counts, this same limitation prevents the use of already pre-processed (i.e. normalized or transformed data) and forces the DEA practitioner to stick to the specific processing of the methods. In that regard, `dearseq` is extremely flexible and offers to use either raw or transformed data (the default applies a log-cpm transformation similarly to `limma-voom`).RNA-seq is subject to various technical biases, and in particular the library size has an important and well-known impact on down-stream analysis if ignored. Therefore, it is important to account for library-size

one way or another, and several RNA-seq data normalization have been proposed to do so (e.g. TMM, RPKM, FPKM, CPM [Dillies et al., 2013]). The choice of which normalization to use is often linked to the biological context of an analysis, and thanks to its distribution-free assumptions, `dearseq` can be paired with any normalization method that is deemed appropriate. Following Law et al. [2014], we implemented the log2-CPM by default.

In addition, these methods have been designed to compare two (or multiple) conditions (several treatment regimen), and are not specifically oriented towards grouped or longitudinal data. Therefore there is a need in the broader DEA community for a more flexible method. `dearseq` relies on a general methodology that can easily accommodate more complex designs including gene set analysis while correctly controlling the false discovery rate [Agniel and Hejblum, 2017].

We have demonstrated that the three most popular RNA-seq DEA methods may not guarantee control of the number of false positive in their results, even when the sample size increases. To exemplify this problematic behavior, we present extensive simulation studies ranging from realistic resampling of real data to synthetic data generation under the models' assumptions, as well as a re-analysis of a real world data-set. To offer an alternative solution to DEA practitioners, we have developed `dearseq`, a new DEA method that uses a variance component score test to provide a robust, powerful and versatile approach to DEA while avoiding the pitfall of FDR inflation exhibited by the current state-of-the-art methods in certain situations. We also benchmarked this new method alongside the three established methods on both the simulations and the real data analysis to illustrate its excellent performance, both in terms of FDR control and of statistical power.

These results have important implications for the field, as DEA of RNA-seq data has become widespread. The distributional assumptions and model-based inference inherent to `DESeq2`, `edgeR` and `limma-voom` can underestimate the number of false positives in realistic settings. Users should be aware of the possibility of inflated FDR when using these procedures and should consider the use of `dearseq` which gives theoretical and empirical control of the FDR without sacrificing its statistical

power. Given the results of both our simulations and our real-world data re-analysis, we thus formulate the following recommendations: i) do not rely on a single DEA method and compare the results across several tools, as this strategy may likely eliminate the majority of false positives ; ii) for your main analysis, we recommend using `dearseq` or `limma-voom` over `DESeq2` or `edgeR`. Indeed, `limma-voom` appears to control the FDR adequately as long as your sample size is large enough and the model assumptions (in particular the linearity) are reasonable. On the other hand, `dearseq` ensures an effective control of the FDR regardless of the sample-size (thanks to its permutation test for small sample sizes) and demonstrates good statistical power.

Funding

MG is supported within the Digital Public Health Graduate's school, funded by the PIA 3 (Investments for the Future - Project reference: 17-EURE-0019). The project is supported through SWAGR Inria Associate-Team from the Inria@SiliconValley program. *Conflict of interest statement.* None declared.

Acknowledgements

Computer time for this study was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

Author's contributions: MG was a major contributor in writing the manuscript and performed the real data re-analysis. DA was a major contributor in writing the manuscript, wrote the original R code for `dearseq`, performed the simulation study under both Negative Binomial and Nonlinear settings and participated in the real data re-analysis. RT was a major contributor in writing the manuscript, participated in the real data analysis and performed the biological interpretation of the results. BH was a major contributor in writing the manuscript, implemented `dearseq` in an R package, performed the resampling simulations and directed the real data re-analysis.

All authors read and approved the final manuscript.

3.5 Supplementary materials

3.5.1 Singhania *et al.* re-analysis

Input data

RNA-seq data from Singhania *et al.* are publicly available on GEO with the primary accession code GSE107991 for the Berry London cohort. Two files are available:

- raw data : Raw_counts_Berry_London
- `edgeR` preprocessed data : edgeR_normalized_Berry_London

In our re-analysis, we used the `edgeR` preprocessed data to run `limma-voom`, `DESeq2` and `dearseq`. The log fold changes are calculated using the raw data.

Preprocessing The matrix of raw counts contains $58,051$ genes and 54 samples. As described in Singhania *et al.*, only genes expressed with counts per million (CPM) > 2 in at least five samples were considered and normalized using trimmed mean of M-values (TMM) to remove the library-specific artefacts. The filtering is carried out with `edgeR`. It results in a matrix of normalized counts containing $14,150$ genes and 54 samples (edgeR_normalized_Berry_London).

Analyses settings

`dearseq` Due to the low sample size for each DEA, the permutation test was used with 1000 permutations. The variable to be tested is the TB group for each of the three comparisons (i.e. TB versus Control, TB versus LTBI and LTBI versus Control). In the absence of covariates, we simply use an intercept.

`DESeq2` We performed the Wald test. The design matrix required was composed of an intercept and the group variable. The other parameters are those given by default in the user guide.

`limma-voom` A linear model is fitted to the log2 CPM for each gene. The `voom` step allows to obtain weights for each gene and sample that are passed into `limma`. The design matrix is the same as the two previous methods. The other parameters are those given by default in the user guide without the `contrasts.fit` step.

`edgeR` We used the genes signature from Singhania *et al.* supplementary file.

The code to reproduce the results is provided as a supplementary file.

3.5.2 Detailed simulation settings

The settings for `limma-voom` and `DESeq2` are the same as those given in section 1.2. Regarding `edgeR`, we followed the quick start section of the user guide, using the default parameters. The associated code is provided as a supplementary file.

Negative Binomial scenario a)

In this scenario, gene expression is generated from the following Negative Binomial distribution $NB(\mu_{ij}, \tau_{ij})$, such that:

$$E[y_{ij}] = \mu_{ij} \qquad \text{Var}(y_{ij}) = \mu_{ij} + \frac{\mu_{ij}^2}{\tau_{ij}}$$

$$\mu_{ij} = \max\{0, \widetilde{\mu}_{ij}\}$$

$$\widetilde{\mu}_{ij} = \begin{cases} \alpha + b_{i0} + x_i, & j = 1, ..., p_1 \\ \alpha + b_{i0} + x_i + (\beta + b_j) x_i z_i, & j = p_1 + 1, ..., p \end{cases}$$

$$\tau_{ij} \sim \text{exponential}(1), \quad b_j \sim N(0, \sigma_g^2), \quad \alpha = 1000, \quad b_{i0} \sim N(0, 1),$$

$$z_i \sim N(0, 1), \quad x_i \sim N(\mu_x, 1), \quad \mu_x \sim \text{exponential}(1/10)$$

Non-linear scenario b)

In this scenario, gene expression is generated according to the following model:

$$
\begin{aligned}
y_{ij} &= \min\left\{\max\left\{\frac{\mu_{ij}+\epsilon_{ij}}{10}, 0\right\}, 10^9\right\} \\
\mu_{ij} &= \begin{cases} \eta_{ij} + (\beta + b_j + b_{i1})z_i, & j = 1, ..., p_1 \\ \eta_{ij}, & j = p_1+1, ..., p \end{cases} \\
\eta_{ij} &= \frac{\gamma_{ij}\sum_{i=1}^{n}\gamma_{ij}}{1000n} \\
\gamma_{ij} &= \begin{cases} \nu_{ij} + (\beta + b_j + b_{i1})z_i, & j = 1, ..., p_1 \\ \nu_{ij}, & j = p_1+1, ..., p \end{cases} \\
\nu_{ij} &= \xi_{ij} + \delta_{ij} + x_i \\
\delta_{ij} &\sim N(0, \tau_{ij}^2) \\
\xi_{ij} &= \iota_j \zeta_{ij} + \iota_j \\
\zeta_{ij} &= N(0, 1) \\
\iota_j &\sim \text{exponential}(0.01) \\
b_j &\sim N(0, \sigma_g^2) \\
\log(\epsilon_{ij}) &\sim N(0, \sigma_{ij}^2) \\
\sigma_{ij} &\sim \text{exponential}(1)
\end{aligned}
$$

and τ_j is 0.01 times the standard deviation of $\{\xi_{1j}, ..., \xi_{nj}\}$

SEQC resampling scenario c)

In this scenario, gene expression was generated by randomly sampling among the five "A" replicate samples from the SEQC data. In practice, we used the data provided as supplementary data by Rapaport et al. [2013]. For a given sample size (4, 8, 16, 20, 50, 100, 150, 200), each simulated sample was first drawn from the original five real ones and arbitrarily assigned to one of the two mock comparison groups. Then,

random noise was added (using a multivariate Gaussian distribution centered on 0 with a covariance matrix for all the genes estimated from the five real original samples) in order to obtain different values for each simulated samples. Finally, values were rounded to the nearest integer and truncated at 0 (included), in order to emulate count data. Since the five A samples are replicates, such simulated samples were homogeneous and did not feature any truly DE gene.

Data-driven Negative Binomial scenario d)

Gene expression is generated from a Negative Binomial distribution of which the parameters have been estimated from Singhania *et al* real data set. For each gene of the real data set, the couple of parameters (μ and size) defining the Negative Binomial are estimated through Maximum Likelihood method. For a given sample size (4, 8, 16, 20, 50, 100, 150, 200), we simulated 10,000 genes of which 500 were differentially expressed. Each non-DE gene is sampled from a Negative Binomial given a couple of estimated parameters. Each DE gene is sampled as a non-DE gene and a random noise (using a Negative Binomial distribution) is added or subtracted to half of the samples.

3.5.3 `dearseq` method

Normalized gene expression

The `dearseq` methodology assumes that the gene expression measurement are comparable across samples. As this is not always the case with raw RNA-seq counts (due to technical effects for instance), a normalization step is often required. `dearseq` does not assume any specific normalization and can work with any kind of quantitative variables.

Toy example

We propose the following toy example to better understand the mixed model notations for `dearseq`. Consider 20 genes observed in 8 patients. 4 subjects received a vaccine and 4 received a placebo. For each patient, gene expression has been

derived from two different tissues (*in vivo* whole blood and *in vitro* stimulated Peripheral Blood Mononuclear Cell, respectively denoted WB and PBMC). Thus, gene expression of each patient has been measured twice, resulting in 16 measurements (2 measurements per subject). In this case, we have to deal with grouped data. We want to test which genes are differentially expressed according to the condition vaccine vs. placebo. We write:

- $G = 20$ the number of genes
- y_{ij}^g the gene expression of the g^{th} gene for the i^{th} patient at the time measurement t_{ij}, for $i = 1, \ldots, 10$, $j = 1, 2$. Thus, for $i = 1, \ldots, 10$, $\boldsymbol{y}_i = (y_{i1}, y_{i2})^T$ is the gene expression vector.
- $\boldsymbol{\phi}_i$ the vector of condition for the patient i. This is the condition to be tested. If the patient i has been vaccinated, we can write the following vector of factors: $\boldsymbol{\phi}_i = (\text{"}vaccine\text{"}, \text{"}vaccine\text{"})^T$. There is only one variable to be tested, so $K = 1$.
- $X_i = (1, 1)^T$, as there is no variable to take into account (i.e. which is not tested). Therefore, $p = 1$

Patient	Condition	Tissue type
1	vaccine	WB
1	vaccine	PBMC
2	vaccine	WB
2	vaccine	PBMC
3	vaccine	WB
3	vaccine	PBMC
4	vaccine	WB
4	vaccine	PBMC
5	placebo	WB
5	placebo	PBMC
6	placebo	WB
6	placebo	PBMC
7	placebo	WB
7	placebo	PBMC
8	placebo	WB
8	placebo	PBMC

Since we have measurements made on the same subject, we have to deal with grouped data. The function `dear_seq` of the R package `dearseq` takes this group structure as an argument to group according to the patients and so, to add a random effect $\boldsymbol{\xi}_i$.

Test statistic

In this section, we derive a variance component score test statistic for the effects of interest. For the sake of simplicity, we omit the gene index g in the following, bear in mind that a test is carried out for each gene g.

According to the model (3.2), the null hypothesis of no effect of interest is:

$$H_0 : \text{“}\boldsymbol{\beta} = 0 \text{ and } \Sigma_\xi = 0\text{”} \tag{3.8}$$

If the variance-covariance matrix of the random effects is identically zero then the random effects $\boldsymbol{\xi}_i$ are also identically zero for all i. If at the same time, $\boldsymbol{\beta} = 0$ then the expression of the gene will not be significantly associated with the variables of interest Φ_i.

Under the working model (3.2), for all i, we will distinguish the effects of covariates and the effects of variables of interest on gene expression by posing $\boldsymbol{\mu}_i = \boldsymbol{\alpha}_0 + X_i\boldsymbol{\alpha}$ and $\boldsymbol{\theta}_i = \boldsymbol{\beta}+\boldsymbol{\xi}_i$. We write $\boldsymbol{\theta}_i = \eta\boldsymbol{\nu}_i = \eta(\boldsymbol{\gamma}+\boldsymbol{\zeta}_i)$ with $\boldsymbol{\nu}_i \sim N(0, \Sigma_\nu), \boldsymbol{\gamma} \sim N(0, I), \boldsymbol{\zeta}_i \sim N(0, \Sigma_\zeta), \Sigma_\nu = I + \Sigma_\zeta$. $\boldsymbol{\nu}_i$ is the nuisance parameter. We can rewrite the null hypothesis as H_0 : "$\eta = 0$" and the model as $\boldsymbol{y}_{\mu_i} = \eta\Phi_i\boldsymbol{\nu}_i + \boldsymbol{\varepsilon}_i$ with $\boldsymbol{y}_{\mu_i} = \boldsymbol{y}_i - \boldsymbol{\mu}_i$ the centered outcome. Then, $\boldsymbol{y}_{\mu_i}|\boldsymbol{\nu}_i \sim N(\eta\Phi_i\boldsymbol{\nu}_i, \Sigma_i)$. We write the likelihood of $\boldsymbol{y}_{\mu_1}, \ldots, \boldsymbol{y}_{\mu_n}|\boldsymbol{\nu}_i$:

$$\begin{aligned}
\mathcal{L}(\eta) &= \mathcal{L}(\boldsymbol{y}_{\mu_1}, \ldots, \boldsymbol{y}_{\mu_n}, \eta|\boldsymbol{\nu}_i) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi)^{n_i/2} \mid \Sigma_i \mid^{1/2}} \\
&\exp\left(-\frac{1}{2}(\boldsymbol{y}_{\mu_i} - \eta\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}(\boldsymbol{y}_{\mu_i} - \eta\Phi_i\boldsymbol{\nu}_i)\right)
\end{aligned}$$

Then, we derive a variance component score test. It has the advantage of avoiding the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_i$ because it only requires estimating the model under the null.

The score being null, we follow the argument of Commenges and Andersen [1995] by considering $\lim_{\eta \to 0} \frac{\partial}{\partial(\eta^2)} \log(\mathcal{L}^*(\eta))$ to obtain the expression of the test statistic. Let $\mathcal{L}^*(\eta)$ be the likelihood of $\boldsymbol{y}_{\mu_1}, ..., \boldsymbol{y}_{\mu_n}$ such as:

$$\mathcal{L}^*(\eta) = \mathcal{L}(\boldsymbol{y}_{\mu_1}, ..., \boldsymbol{y}_{\mu_n}; \eta) = \mathbb{E}[\mathcal{L}(\boldsymbol{y}_{\mu_1}, ..., \boldsymbol{y}_{\mu_n}; \eta | \boldsymbol{\nu}_i) \mid \mathbb{V}],$$

with $\mathbb{V} = \{\boldsymbol{V}_i = (\boldsymbol{y}_{\mu_i}^T, X_i^T, \Phi_i^T)^T\}_{i=1}^n$

Then,

$$\begin{aligned}
\lim_{\eta \to 0} \frac{\partial \log(\mathcal{L}^*(\eta))}{\partial(\eta^2)} &= \lim_{\eta \to 0} \frac{1}{2\eta \mathcal{L}^*(\eta)} \frac{\partial \mathcal{L}^*(\eta)}{\partial \eta} \\
&= \lim_{\eta \to 0} \frac{1}{2\mathcal{L}^*(\eta)} \left[\eta^{-1} \frac{\partial \mathcal{L}^*(0)}{\partial \eta} + \frac{\partial^2 \mathcal{L}^*(0)}{\partial \eta^2} + o(1) \right] \\
&= \frac{1}{2\mathcal{L}^*(0)} \left[\frac{\partial^2 \mathcal{L}^*(0)}{\partial \eta^2} + o(1) \right]
\end{aligned}$$

Thus, removing $1/2$ for the sake of simplicity, we have :

$$
\begin{aligned}
\mathcal{L}^*(0)^{-1}\frac{\partial^2\mathcal{L}^*(0)}{\partial\eta^2}+o(1) &= \mathcal{L}^*(0)^{-1}\frac{\partial}{\partial\eta}\left(\frac{\partial\mathcal{L}^*(0)}{\partial\eta}\right)+o(1)\\
&= \frac{\partial^2\log\mathcal{L}^*(0)}{\partial\eta^2}+o(1)\\
&= \mathbb{E}\left[\frac{\partial^2\log\mathcal{L}(0)}{\partial\eta^2}|\mathbb{V}\right]+o(1)\\
&= \mathbb{E}\left[\frac{\partial}{\partial\eta}\left(\frac{\partial\log\mathcal{L}(0)}{\partial\eta}\right)|\mathbb{V}\right]+o(1)\\
&= \mathbb{E}\left[\frac{\frac{\partial^2\mathcal{L}(0)}{\partial\eta^2}\mathcal{L}(0)-\frac{\partial\mathcal{L}(0)}{\partial\eta}\frac{\partial\mathcal{L}(0)}{\partial\eta}}{\mathcal{L}(0)^2}|\mathbb{V}\right]+o(1)\\
&= \mathbb{E}\left[\frac{\frac{\partial^2\mathcal{L}(0)}{\partial\eta^2}}{\mathcal{L}(0)}-\left(\frac{\frac{\partial\mathcal{L}(0)}{\partial\eta}}{\mathcal{L}(0)}\right)^2|\mathbb{V}\right]+o(1)\\
&= \mathbb{E}\left[\frac{\partial^2\log\mathcal{L}(0)}{\partial\eta^2}-\left(\frac{\partial\log\mathcal{L}(0)}{\partial\eta}\right)^2|\mathbb{V}\right]+o(1)\\
&= \mathbb{E}\left[\frac{\partial^2\log\mathcal{L}(0)}{\partial\eta^2}|\mathbb{V}\right]+\mathbb{E}\left[\left(\frac{\partial\log\mathcal{L}(0)}{\partial\eta}\right)^2|\mathbb{V}\right]+o(1)
\end{aligned}
$$

Then standardizing by n,

$$
\lim_{\eta\to 0} n^{-1}\frac{\partial\log(\mathcal{L}^*(\eta))}{\partial(\eta^2)} = n^{-1}\mathbb{E}\left[\left(\frac{\partial\log\mathcal{L}(0)}{\partial\eta}\right)^2|\mathbb{V}\right]+\text{constant}+o(1)
$$

because

$$
n^{-1}\frac{\partial^2\log(\mathcal{L}(\eta))}{\partial\eta^2} = n^{-1}\sum_{i=1}^{n}-(\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i = \text{constant}
$$

Yet,

$$\begin{aligned}\frac{\partial \log(\mathcal{L}(\eta))}{\partial \eta} &= \sum_{i=1}^{n} \frac{\partial}{\partial \eta} - \frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log(\mid \Sigma_i \mid) - \frac{1}{2}(\boldsymbol{y}_{\mu_i} - \eta\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}(\boldsymbol{y}_{\mu_i} - \eta\Phi_i\boldsymbol{\nu}_i) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \eta} - \frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log(\mid \Sigma_i \mid) - \frac{1}{2}(\boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\boldsymbol{y}_{\mu_i} - \boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\eta\Phi_i\boldsymbol{\nu}_i \\ &- \eta(\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}\boldsymbol{y}_{\mu_i} + \eta^2(\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i) \\ &= \sum_{i=1}^{n} -\frac{1}{2}\left(-\boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i - (\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}\boldsymbol{y}_{\mu_i} + 2\eta(\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i\right)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \log(\mathcal{L}(\eta))}{\partial \eta} \mid_{\eta=0} &= \sum_{i=1}^{n} -\frac{1}{2}(-\boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i - (\Phi_i\boldsymbol{\nu}_i)^T\Sigma_i^{-1}\boldsymbol{y}_{\mu_i}) \\ &= \sum_{i=1}^{n} \frac{1}{2}(\boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i + \boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i) \\ &= \sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T\Sigma_i^{-1}\Phi_i\boldsymbol{\nu}_i\end{aligned}$$

So,

$$
\begin{aligned}
n^{-1}\mathbb{E}\left[\left(\frac{\partial \log \mathcal{L}(0)}{\partial \eta}\right)^2 | \mathbb{V}\right] &= n^{-1} Var\left[\frac{\partial \log \mathcal{L}(0)}{\partial \eta} \mid \mathbb{V}\right] \\
&= n^{-1} Var\left[\sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i \boldsymbol{\nu}_i \mid \mathbb{V}\right] \\
&= n^{-1} Var\left[\sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i (\boldsymbol{\gamma} + \boldsymbol{\zeta}_i) \mid \mathbb{V}\right] \\
&= n^{-1} Var\left[\left(\sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i\right) \boldsymbol{\gamma} \mid \mathbb{V}\right] + n^{-1} Var\left[\sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i \boldsymbol{\zeta}_i \mid \mathbb{V}\right] \\
&= n^{-1}\left(\sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i\right)\left(\sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i\right)^T + n^{-1} \sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i \Sigma_\zeta \Phi_i^T \Sigma_i^{-1} \boldsymbol{y}_{\mu_i} \\
&= \left(n^{-1/2} \sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i\right)\left(n^{-1/2} \sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i\right)^T + \text{constant} + o(1) \\
&= \boldsymbol{q}^T \boldsymbol{q}
\end{aligned}
$$

Let Q be the variance component score test statistic such as $Q = \boldsymbol{q}^T \boldsymbol{q}$ with

$$
\boldsymbol{q}^T = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i = n^{-1/2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \mu_i)^T \Sigma_i^{-1} \Phi_i
$$

Considering that we have a consistent estimator of Σ_i, we still must provide estimates of α_0 and $\boldsymbol{\alpha}$. A natural way to estimate these quantities, given the heteroscedasticity in $\boldsymbol{y}$, is to fit a weighted mixed effects model. The weights are taken to be $\boldsymbol{w}_i = \text{diag}(\widehat{\Sigma}_i)^{-1}$. However, to avoid excessive computation time, instead of estimating the full mixed effects model from (3.2), we may fit a simpler fixed effects model (3.4), from which we can obtain estimates of α_0 and $\boldsymbol{\alpha}$.

Test statistic limiting distribution

We have $Q = \boldsymbol{q}^T \boldsymbol{q}$. We note $\Gamma = \text{cov}(\boldsymbol{q})$. Then, we can write:

$$
Q = \boldsymbol{q}^T \Gamma^{-1/2} \Gamma \Gamma^{-1/2} \boldsymbol{q}
$$

The matrix Γ being square and diagonal, we carry out a singular value decomposition of Γ:

$$Q = \boldsymbol{q}^T \Gamma^{-1/2} U A U^T \Gamma^{-1/2} \boldsymbol{q},$$

where U is an orthogonal matrix of eigen vectors of Γ, A is a diagonal matrix of eigen values of Γ.
We take $u = \Gamma^{-1/2}\boldsymbol{q}$, with $\boldsymbol{q}^T = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}\Phi_i$, which give us: $Q = u^T U A U^T u$. Under the normal residual hypothesis, u immediately follows a standard normal distribution, in which case it is not necessary to apply the central limit theorem. Nevertheless, the variance component test is intended to be robust to the mispecification of the model, i.e., if the normal residual assumption is not verified. Therefore, we propose an asymptotic test to ensure its robustness against any data distribution, for example a negative binomial. This is one of the reasons why we propose the use of permutations when the number of individuals is considered too small or simply to ensure the reliability of the test. Then,

$$\begin{aligned}
\mathbb{E}(u) = \Gamma^{-1/2}\mathbb{E}(\boldsymbol{q}^T) &= \Gamma^{-1/2}\mathbb{E}(n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}\Phi_i) \\
&= \Gamma^{-1/2} n^{-1/2} \sum_{i=1}^{n} \underbrace{\mathbb{E}(\boldsymbol{y}_i^T - \boldsymbol{\mu}_i^T)}_{=0} \Sigma_i^{-1}\Phi_i \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{cov}(u) &= \mathrm{cov}(\Gamma^{-1/2} q^T) \\
&= (\Gamma^{-1/2})^T \mathrm{cov}(q) \Gamma^{-1/2} \\
&= \Gamma^{-1/2}\Gamma\Gamma^{-1/2} \\
&= \Gamma^{-1/2}\Gamma^{1/2} I_K \Gamma^{1/2}\Gamma^{-1/2} \\
&= I_K
\end{aligned}$$

By the central limit theorem, u asymptotically follows a multivariate standard normal distribution. U being orthonormal, $U^T u$ also asymptotically follows a multivariate standard normal distribution. So, $u^T U A U^T u = \sum_{k=1}^{K} a_k (u_k^*)^2$ where u_k^* is an element of the asymptotic multivariate standard normal distribution of $U^T u$ and a_k is an eigen value of Γ. So, it follows that $Q \underset{+\infty}{\sim} \sum_{k=1}^{K} a_k \chi_1^2$. Let $\widehat{Q}$ be the estimate of Q. Because Q and $\widehat{Q}$ are asymptotically equivalent, $\widehat{Q} \underset{+\infty}{\sim} \sum_{k=1}^{K} \widehat{a}_k \chi_1^2$. (See Agniel and Hejblum (2017) for the proof.)

Asymptotic and permutation tests

When n is sufficiently large, we propose an asymptotic test. The asymptotic distribution of the test statistic Q is a mixture of χ_1^2 random variables, i.e. $Q \to \sum_{k=1}^{K} a_k \chi_1^2$ where the mixing coefficients a_k depend on the covariance of $\boldsymbol{q}$. When n is very small, relying on the limiting distribution may not be adequate. To overcome this difficulty, we provide a permutation alternative to our asymptotic test. Permutations can be used to estimate the empirical distribution of $\widehat{Q}$ under the null hypothesis. Indeed, permutation tests are attractive because the only assumption we make is that the observations are independent and identically distributed under the null. As explained Phipson and Smyth [2010], it is essential to notice that permutation p-values that are really estimates of p-values, i.e., $\widehat{p}$-values, can lead to $\widehat{p}$-values exactly equal to zero. However, it is senseless to obtain $\widehat{p}$-values equal to zero when all permutations were enumerated, therefore it is not accurate to assume that the $\widehat{p}$-value can be reduced to zero by taking a smaller subset of all the permutations. So, estimating the p-value by B/m where B is the number of permutations for which the associated test statistics are at least as extreme as the observed one can be misleading.

Considering our model, the observations of a given individual i are exchangeable under the null, regardless of sampling measure. We assume that an independent random sample of S permutations is drawn with replacement such as $\boldsymbol{y}_i^* \in \mathbb{R}^{n_i}, y_{ij}^* = y_{i\sigma(j)}$ with $\sigma \in Perm\{1, \ldots, n_i\}$. We generate m test statistics which can contain repeat values, including the original observed value t_{obs}. Let B be the number of permutations for which the m test statistics are at least as extreme as t_{obs}, m_t be all

possible distinct permutations, B_t be the unknown total number of possible distinct test statistics exceeding t_{obs}, and $p_t = (B_t + 1)/(m_t + 1)$ be the ideal permutation p-value which is obviously unknown. If the null hypothesis is true, then B_t follows a discrete uniform distribution on the integers $0, \ldots, m_t$. Conditional on $B_t = b_t$, B follows a binomial distribution $\mathcal{B}(m, p_t)$. An approximation of this quantity can be calculated by:

$$p_e = \frac{b+1}{m+1} - \int_0^{0.5/m_t+1} F(b; m, p_t), \tag{3.9}$$

F is the cumulative probability function of the binomial distribution.

3.6 Analyse d'un jeu de données réelles sur le COVID-19 par `dearseq`

Cette section est un court résumé de l'analyse de données effectuée à l'aide de `dearseq` et incluse dans l'article publié suivant et également fourni en Annexe B :

Yves Lévy, Aurélie Wiedemann, Boris P. Hejblum, Mélany Durand, Cécile Lefebvre, Mathieu Surénaud, Christine Lacabaratz, Matthieu Perreau, Emile Foucat, Marie Déchenaud, Pascaline Tisserand, Fabiola Blengio, Benjamin Hivert,
Marine Gauthier, Minerva Cervantes-Gonzalez, Delphine Bachelet, Cédric Laouénan, Lila Bouadma, Jean François Timsit, Yazdan Yazdanpana, Giuseppe Pantaleo, Hakim Hocini, Rodolphe Thiébaut. CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. *iScience*. 24(7):102711 (2021).
DOI: 10.1016/j.isci.2021.102711

En pleine pandémie mondiale de COVID-19, j'ai été amenée à collaborer avec le *Vaccine Research Institute* (VRI), basé à l'hôpital Henri Mondor (APHP). Le VRI travaille sur la mise au point de vaccins contre le VIH et diverses maladies infectieuses. Les données sur lesquelles se base l'article précédemment cité sont issues d'une cohorte française, regroupant des individus contaminés par le COVID-19 : *French COVID* [Yazdanpanah et al., 2020]. Tous les patients ont été classés comme étant atteints

d'une forme sévère. 53 patients ont été hospitalisés dans une unité de soins intensifs depuis le début des symptômes, ou après une aggravation clinique, et 8 n'ont pas eu besoin d'une hospitalisation, amenant à un total de 61 individus inclus dans l'étude. L'âge médian était de 60 ans et 80% étaient des hommes. Le prélèvement d'échantillons pour les analyses immunologiques a été effectué dans les trois jours suivant l'admission et après une durée médiane de 11 jours après l'apparition des symptômes, à l'hôpital Paris-Bichat. 10 donneurs sains ont également été inclus afin de former un groupe contrôle, dont les échantillons issus des donneurs sains ont été collectés auprès de l'Etablissement Français du sang avant l'épidémie de COVID-19.

L'expression génique a été mesurée par RNA-seq en masse chez 44 patients infectés. C'est également le cas pour les 10 donneurs sains, portant ainsi le total des individus à 54. Les gènes ayant des comptes nuls pour tous les patients ont été retirés. 29 302 gènes ont alors été retenus pour l'analyse différentielle. Les données sont accessibles publiquement sur *Gene Expression Omnibus repository* avec le code GSE171110. Les comptes bruts sont normalisés par $\log_2$CPM. L'analyse différentielle proposée ici vise à détecter quels gènes s'expriment différentiellement entre les patients atteints du COVID-19 et les donneurs sains. Nous sommes dans le cas classique de l'analyse différentielle de la comparaison entre deux conditions. De par son bon comportement en termes de puissance statistique et de contrôle du FDR, `dearseq` a été la méthode privilégiée. Le nombre d'échantillons étant relativement faible, le test par permutations a été appliqué.

Entre la publication de `dearseq` et cette analyse de données, un test par permutations adaptatives a pu être implémenté (l'idée sera reprise dans la méthode `ccdf` détaillée dans le Chapitre 4). Le test par permutations adaptatives consiste à calculer le même nombre de permutations pour tous les gènes, puis à accroître le nombre de permutations seulement pour les gènes ayant une p-valeur inférieure à un seuil choisi. On répète cette dernière étape en augmentant progressivement le nombre de permutations et en abaissant le seuil pour les p-valeurs. La procédure s'arrête au nombre de permutations maximum choisi par l'utilisateur. Cela permet d'obtenir des p-valeurs avec une plus grande précision pour les plus petites p-valeurs tout en limitant l'aug-

mentation des temps de calcul. Dans cette analyse, nous avons donc choisi un nombre initial de 1 000 permutations pour aller jusqu'à 64 000 permutations.

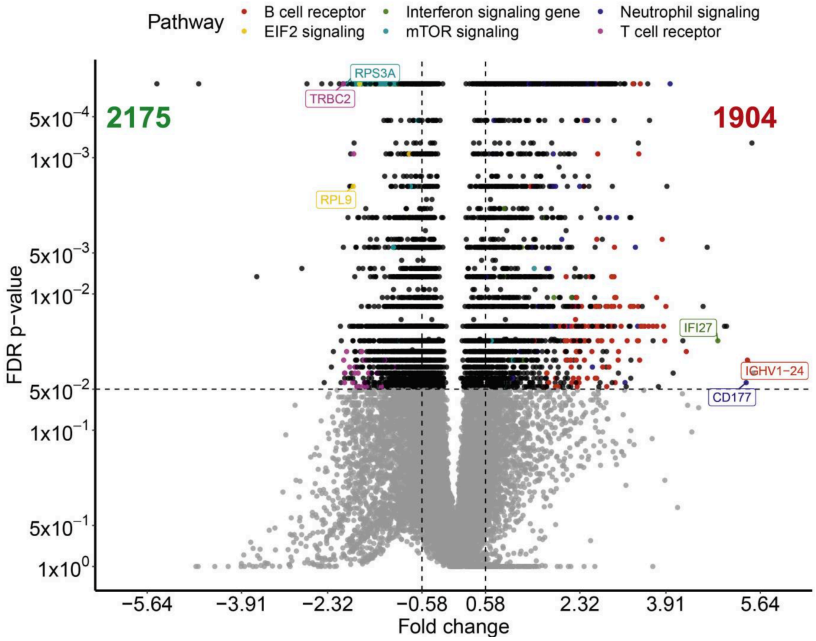


FIGURE 3-8 : Graphique de type *Volcano plot* représentant les gènes différentiellement exprimés identifiés par `dearseq` selon le FC et la p-valeur ajustée après la correction de Benjamini and Hochberg [1995] (appelée FDR *p-value*). La ligne horizontale en pointillés désigne le seuil de significativité des p-valeurs égal à 0.05. Les deux lignes verticales et parallèles en pointillés désignent les seuils tels que $\log_2(|FC|) \geq \log_2(1.5)$ Les gènes et les voies biologiques (*Pathways*) d'intérêt sont mis en évidence par différentes couleurs. Source : Lévy et al. [2021]

Le *fold-change* (FC) est une notion communément utilisée en analyse différentielle qu'il est nécessaire d'expliciter. Pour un gène, on appelle *fold-change* le rapport entre la valeur moyenne de l'expression génique sous la condition 1 et la valeur moyenne de l'expression génique sous la condition 2. Un *fold-change* égal à 2 signifie que l'expression génique moyenne sous la condition 1 est deux fois plus grande que l'expression génique moyenne sous la condition 2. Dans notre analyse, cela signifie que l'expression moyenne des patients atteints du COVID-19 est 2 fois plus grande que l'expression moyenne des individus sains, pour un gène donné. Un *fold-change* égal à 0,5 signifie que l'expression génique moyenne sous la condition 1 représente la moitié de l'expression génique moyenne sous la condition 2. Typiquement, les $\log_2$FC sont davantage utilisés pour leur symétrie lors d'une visualisation graphique. En effet, si $FC = 2$

alors $\log_2(FC) = 1$. Le gène sera considéré comme sur-exprimé. Si $FC = 0.5$ alors $\log_2(FC) = -1$. Le gène sera considéré comme sous-exprimé.

A l'aide de `dearseq`, 4 079 gènes ont été trouvés comme étant différentiellement exprimés tels que $\log_2(|FC|) \geq \log_2(1.5)$, dont 1 904 sur-exprimés et 2 175 sous-exprimés (voir figure 3-8). De ces résultats, le VRI a noté la présence de plusieurs voies biologiques associées aux gènes différentiellement exprimés qui correspondent à la réponse immunitaire, notamment la signalisation des neutrophiles et des interférons ainsi que les réponses des récepteurs des lymphocytes T et B. Parmi les gènes les plus sur-exprimés, on retrouve le gène CD177, un marqueur d'activation des neutrophiles. Les analyses statistiques et biologiques complémentaires présentées dans la publication ont pu mettre en évidence que le gène CD177 était non seulement un marqueur de la progression de la maladie et de sa sévérité mais aussi du décès.

Chapitre 4

Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis

Marine Gauthier [1,2], **Denis Agniel** [3,4]
Rodolphe Thiébaut [1,2,5]**Boris P. Hejblum** [1,2]

[1]University of Bordeaux, INSERM Bordeaux Population Health Research Center, INRIA SISTM, F-33000 Bordeaux, France. [2]Vaccine Research Institute, F-94000 Créteil, France. [3]Rand Corporation, Santa Monica (CA), USA. [4]Harvard Medical School, Boston (MA), USA. [5]CHU de Bordeaux, Bordeaux, F-33000 France.

Submitted

Abstract

State-of-the-art methods for single-cell RNA sequencing (scRNA-seq) Differential Expression Analysis (DEA) often rely on strong distributional assumptions that are difficult to verify in practice. Furthermore, while the increasing complexity of clinical and biological single-cell studies calls for greater tool versatility, the majority of existing methods only tackle the comparison between two conditions. We propose a novel, distribution-free, and flexible approach to DEA for single-cell RNA-seq data. This new method, called `ccdf`, tests the association of each gene expression with one

or many variables of interest (that can be either continuous or discrete), while potentially adjusting for additional covariates. To test such complex hypotheses, `ccdf` uses a conditional independence test relying on the conditional cumulative distribution function, estimated through multiple regressions. We provide the asymptotic distribution of the `ccdf` test statistic as well as a permutation test (when the number of observed cells is not sufficiently large). `ccdf` substantially expands the possibilities for scRNA-seq DEA studies: it obtains good statistical performance in various simulation scenarios considering complex experimental designs (*i.e.* beyond the two condition comparison), while retaining competitive performance with state-of-the-art methods in a two-condition benchmark. We apply `ccdf` to a large publicly available scRNA-seq dataset of 84,140 SARS-CoV-2 reactive CD8+ T cells, in order to identify the differentially expressed genes across 3 groups of COVID-19 severity (mild, hospitalized, and ICU) while accounting for seven different cellular subpopulations

4.1 Introduction

Single-cell RNA sequencing (scRNA-seq) makes it possible to simultaneously measure gene expression levels at the resolution of single cells, allowing a refined definition of cell types and states across hundreds or even thousands of cells at once. Single-cell technology significantly improves on bulk RNA-sequencing, which measures the average expression of a set of cells, mixing the information in the composition of cell types with different expression profiles. New biological questions such as detection of different cell types or cellular response heterogeneity can be explored thanks to scRNA-seq, broadening our comprehension of the features of a cell within its microenvironment [Eberwine et al., 2014].

Several challenges arise from the sequencing of the genetic material of individual cells like in transcriptomics (see Lähnemann et al. [2020] for a thorough and detailed review). Differential expression analysis (DEA) is a major field of exploration to better understand the mechanisms of action involved in cellular behavior. A gene is called differentially expressed (DE) if its expression is significantly associated with the variations of a factor of interest. Single-cell data have different features from bulk RNA-seq data that require special consideration for developing DEA tools. Indeed, scRNA-seq data display large proportions of observed zeros (i.e. "dropouts"), due either to biological processes or technical limitations [Lähnemann et al., 2020].

The large amount of single-cell measurements provides an opportunity to estimate and characterize the distribution of each gene expression and to compare it under different conditions in order to identify DE genes. In fact, scRNA-seq distributions usually show complex patterns. Therefore, the `scDD` method [Korthauer et al., 2016] condenses the difference in distribution between two conditions into four categories: the usual difference in mean, the difference in modality, the difference in proportions and the difference in both mean and modality. Since scRNA-seq data analysis lay unique challenges, new statistical methodologies are needed.

Several strategies making strong distributional assumptions on the data have been proposed to perform single-cell DEA. MAST [Finak et al., 2015] and SCDE [Kharchenko et al., 2014] are two well-know differential methods, the former using a two-part generalized linear model to take into account both the dropouts and the non-zero values by making a Gaussian assumption of each gene and the latter relying on a Bayesian framework combined with a mixture of Poisson and negative binomial distributions. `scDD` [Korthauer et al., 2016] makes use of a Bayesian modeling framework to detect differential distributions and then to classify the gene into four differential patterns using Gaussian mixtures. `DEsingle` [Miao et al., 2018] proposes a zero-inflated negative binomial (`ZINB`) regression followed by likelihood-ratio test to compare two samples. `D`^{3}`E` [Delmans and Hemberg, 2016] also applies a likelihood-ratio test after fitting a Poisson-Beta distribution.

Yet, Risso et al. [2018] have advanced that scRNA-seq data are zero-inflated and have proposed to use zero-inflated negative binomial models, while Svensson [2020] have argued that the number of zero values is consistent with usual count distributions. Then, Choi et al. [2020] illustrated that scRNA-seq data are zero-inflated for some genes but "this does not necessarily imply the existence of an independent zero-generating process such as technical dropout". In fact, biological information (e.g. cell type and sex) may explain it. The authors also discourage imputation as zeros can contain relevant information about the genes. In addition, Townes et al. [2019] argue that single-cell zero inflation actually comes from normalization and log-transformation. The distribution and the sparsity of scRNA-seq data remains difficult

to model, and – as there is no consensus on which model is the best one – it is of utmost importance to develop general and flexible methods for analyzing scRNA-seq data which do not require strong parametric assumptions.

Fewer distribution-free tools have been developed to model single-cell complex distributions without making any parametric assumption. EMDomics [Nabavi et al., 2016] and more recently SigEMD [Wang and Nabavi, 2018] are two non-parametric methods based on the Wassertein distance between two histograms, the latter including data imputation to handle the problem of the great number of zero counts. D^3E offers in addition the possibility to perform either the Cramer-von Mises test or the Kolmogorov-Smirnov test to compare the expression values' distributions of each gene, thus delivering a distribution-free option. In a comparative review, Wang et al. [2019] illustrated that non-parametric methods, i.e. distribution-free, perform better in distinguishing the four differential distributions. Recently, Tiberi et al. [2020] presented `distinct`, a hierarchical non-parametric permutation approach using empirical cumulative distribution functions comparisons. The method requires biological replicates (at least 2 samples per group) and allows adjustment for covariates but only tackles the comparison between two conditions to our knowledge.

However, the limitations of these state-of-the-art methods for scRNA-seq DEA are many. The approaches based on a distributional assumption face methodological issues, as they rely on strong distributional assumptions that are difficult to test in practice. In fact, a deviation from the hypothesized distribution will translate into erroneous p-values and may lead to inaccurate results. While the increasing complexity of clinical and biological studies calls for greater tools versatility, the majority of existing methods, whether parametric or non-parametric, cannot handle data sets with a complex design, making them very restrictive. In fact, the most commonly used methods remain in the traditional framework of DEA and only tackle the comparison between two conditions. One might be interested in the genes differentially expressed across several conditions (e.g. more than two cell groups, multi-arm...) or in testing the genes differentially expressed according to a continuous variable (e.g. cell surface markers measured by flow cytometry...). In particular, the identification of surrogate

biomarkers from transcriptomic measurements is becoming an emerging field of interest, especially in cancer therapy [Wang et al., 2007] or in new immunotherapeutic vaccines [Cliff et al., 2004]. Gene expression could be used to compare treatments in observational settings. Yet in such cases, adjusting for some technical covariates or some confounding factors is paramount to ensure the validity of an analysis, as this external influence can impact the outcome as well as the dependent variables and thus generates spurious results by suggesting a non-existent link between variables.

Overall, the need of testing the association between gene expression and the variables of interest, potentially adjusted for covariates, makes an additional motivation for developing suitable tools. The complex hypothesis we aim to test consists in performing a conditional independence test (CIT). A CIT broadens the classical independence test by testing for independence between two variables given a third one, or a set of additional variables (see Figure 4.1). Two random variables X and Y are conditionally independent given a third variable Z if, and only if, $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$. As described in Li and Fan [2020], many CIT have been developed previously and are readily available such as discretization-based tests [Huang et al., 2010], metric-based tests [Runge, 2018; Su and White, 2007; Huang et al., 2016], permutation-based two-sample tests [Doran et al., 2014; Gretton et al., 2012; Sen et al., 2017], kernel-based tests [Muandet et al., 2016; Li et al., 2009] and regression-based tests (see Li and Fan [2020] for a short review). Yet, these CIT either suffer from the curse of dimensionality or are hardly applicable to a large number of observations [Muandet et al., 2017; Zhang, Peters, Janzing and Schölkopf, 2011]. Zhou et al. [2020] have converted the conditional independence test into an unconditional independence test and then used the Blum–Kiefer–Rosenblatt correlation [Blum et al., 1961] to develop an asymptotic test. Yet, the latter cannot be applied when X is discrete. Those limitations make these tests impractical in our context of scRNA-seq DEA and thus require adaptation.

Performing DEA necessarily involves performing as many independent tests as there are genes. The variables of interest may be either discrete or continuous, while the number of covariates to condition upon may also increase. Consequently there is

an urgent need for a CIT that is both flexible and fast. Here, we propose a novel, distribution-free, and flexible approach, called `ccdf`, to test the association of gene expression to one or several variables of interest (continuous or discrete) potentially adjusted for additional covariates. Because of the current limitations of existing CIT and the growing interest in testing differences in distribution, we make use of a CIT based on the conditional cumulative distribution function (CCDF), estimated by a regression technique. We derive the asymptotic distribution of the `ccdf` test statistic, which does not rely on the underlying distribution of the data, as well as a permutation test to ensure a good control of type I error and FDR, even with a limited sample size.

Section 2 describes our proposed method with both asymptotic and permutation tests. Section 3 presents several simulation scenarios to illustrate the good performances of `ccdf` when we consider complex designs (*i.e.* beyond the two condition comparison), while a benchmark in the two condition case shows our method retains similar performance in terms of statistical power compared to competitive state-of-the-art methods. Section 4 compares the performances of our method with several methods on a "positive data set" that included differentially expressed genes as well as a "negative data set" without any differentially expressed gene. Section 5 illustrates the application of `ccdf` to a scRNA-seq study in COVID-19 patients. Section 6 discusses the strengths and limitations of our proposed approach.

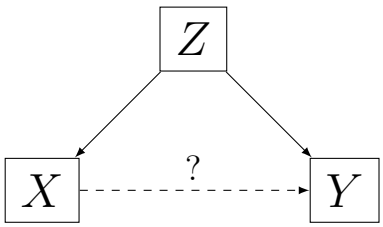


Figure 4-1: Conditional dependence graph [Li and Fan, 2020]

4.2 Method

In this section, we propose a new, easy-to-use, fast, and flexible test for scRNA-seq DEA. We give both its asymptotic distribution as well as a permutation approach to obtain valid p-values in small samples.

4.2.1 Conditional independence test

Null hypothesis.

Testing the association of Y, namely the expression of a gene, with a factor or a group of factors of interest X either discrete (multiple comparisons) or continuous given covariates Z is equivalent to test conditional independence between Y and X knowing Z: " $H_0 : Y \perp X \mid Z$ ". Several statistical tools can be used to characterize the probability law of a random variable, such as the characteristic function, the probability density distribution or the cumulative distribution function. The characteristic function is not often used in practice, due to its relative analytical complexity. The probability density distribution, while more popular, remains difficult to estimate in practice when the number of variables increases due to the increasing number of bandwidths to be optimized. This curse of dimensionality quickly leads to classical computational problems because of complexity growth. As for the cumulative distribution function, its estimation does not require any parameter akin to these bandwidths, making it an efficient tool in high-dimensional non-parametric statistics. From this point, we built a general DEA method including an estimation of the CCDF based on regression technique.

The conditional independence test we propose is based on CCDFs. Indeed, if a group of factors is associated with the expression of a gene, the immediate consequence is that the CCDF of the gene expression would be significantly different from the marginal cumulative distribution, which overlooks this conditioning. Thus, the null hypothesis can be written as: " $H_0 : F_{Y|X,Z}(y, x, z) = F_{Y|Z}(y, z)$ ", where the CCDF of Y given X and Z is defined as $F_{Y|X,Z}(y, x, z) = \mathbb{P}(Y \leq y \mid X = x, Z = z)$. If there are no covariates, the conditional independence test turns into a traditional

independence test as we test the null hypothesis $Y \perp X$ which is equivalent to test $F_{Y|X}(y,x) = F_Y(y)$.

Test statistic.

In this section, we propose a general test statistic for testing the null hypothesis of conditional independence that is easy to compute. We denote $\boldsymbol{Y}^g = (Y_1^g, ..., Y_n^g)$ an outcome vector (i.e. normalized read counts for gene g in n cells) and $X^g = (\boldsymbol{X}_1^g, ..., \boldsymbol{X}_n^g)$ a $s \times n$ matrix encoding the condition(s) to be tested that can be either continuous or discrete. One may want to add exogenous variables, which are not to be tested but upon which it is necessary to adjust the model. Let $Z^g = (\boldsymbol{Z}_1^g, ..., \boldsymbol{Z}_n^g))$ be a $r \times n$ matrix for continuous or discrete covariates to take into account. For the sake of simplicity, we drop the notation g in the remainder as we refer to gene-wise DEA.

We have $\boldsymbol{Y} \in [\zeta_{\min}, \zeta_{\max}]$ for some known constants ζ_{min}, ζ_{max}. Let $\zeta_{\min} \leq \omega_1 < \omega_2 < ... < \omega_p < \zeta_{\max}$ is a sequence of p ordered and regular thresholds. For each ω_j with $j = 1, ..., p$, the CCDF $F_{Y|X,Z}(\omega_j \mid x, z)$ may be written as a conditional expectation $F_{Y|X,Z}(\omega_j \mid x, z) = \mathbb{E}\left[\mathbb{1}_{\{Y \leq \omega_j\}} | X = x, Z = z\right] = \mathbb{E}\left[\widetilde{Y}_{ij} | X_i = x, Z_i = z\right]$ where $\widetilde{Y}_{ij} = \mathbb{1}_{\{Y_i \leq \omega_j\}}$ is a binary random variable that is 1 if $Y_i \leq \omega_j$ and 0 otherwise. We propose to estimate these conditional expectations through a sequence of p working models:

$$g\left(\mathbb{E}\left[\widetilde{Y}_{ij} | X_i, Z_i\right]\right) = \beta_{0j} + \boldsymbol{\beta}_{1j}\boldsymbol{X}_i + \boldsymbol{\beta}_{2j}\boldsymbol{Z}_i, \quad \forall i = 1, ..., n \tag{4.1}$$

where $\boldsymbol{\beta}_{1j} = (\beta_{1j1}, ..., \beta_{1js})$ is the vector of size s referring to the regression of $\widetilde{Y}_{ij}$ onto $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_{2j}$ is the vector of size r referring to the regression of $\widetilde{Y}_{ij}$ onto $\boldsymbol{Z}_i$, for the fixed thresholds $\omega_1, \omega_2, ..., \omega_p$. If X has no link with Y given Z, then we expect that $\boldsymbol{\beta}_{1j}$ will be null. So, we aim to test:

$$H_0 : \boldsymbol{\beta}_{1j} = \mathbf{0}, \quad j = 1, ..., p \tag{4.2}$$

Although, there are many different test statistics associated with this null hypothesis ,

we propose to use the following test statistic that can be written as $D = n\sum_{j=1}^{p}\sum_{k=1}^{s}\boldsymbol{\beta}_{1jk}^2$.

Estimation and asymptotic distribution.

In this section, we describe how to estimate β_{1j}, which allows the computation of the test statistic D.

While in principle any link function $g(\cdot)$ could be selected for the models (4.1), we select the identity link $g(y) = y$ for its computational simplicity, and we compute coefficient estimates using ordinary least squares (OLS). Because our approach requires p models for each of possibly thousands of genes, speed is of utmost importance, so we use OLS. Other selections of $g(\cdot)$ are of course possible and could be explored at the cost of additional computation time.

We show in the Appendix that using OLS to estimate (4.1), $\widehat{\boldsymbol{\beta}}_{1j}$ can be expressed by $\widehat{\boldsymbol{\beta}}_{1j} = n^{-1}\sum_{i=1}^{n}\boldsymbol{h}_i\tilde{Y}_{ij}$, where $\boldsymbol{h}_i$ is a function of the design matrix W with i^{th} row $\boldsymbol{W}_i = (1, \boldsymbol{X}_i, \boldsymbol{Z}_i)$. These estimates may be plugged in to obtain the estimated test statistic $\widehat{D}_n = n\sum_{j=1}^{p}\sum_{k=1}^{s}\widehat{\boldsymbol{\beta}}_{1jk}^2$. We furthermore show in the Appendix that the asymptotic distribution of the test statistic may be approximated by a mixture of χ_1^2 random variables:

$$\widehat{D} = \tilde{\boldsymbol{u}}^{\intercal}A\tilde{\boldsymbol{u}} + o_p(1) = \sum_{j=1}^{ps} a_j\tilde{u}_j^2 + o_p(1) \tag{4.3}$$

where $\tilde{\boldsymbol{u}} = (\tilde{u}_1, ..., \tilde{u}_{ps}) \sim N(0, I)$ are standard multivariate normal random variables and a_j are the eigenvalues of $\Sigma = \text{cov}(\sqrt{n}\widehat{\boldsymbol{\gamma}}_1)$ where $\widehat{\boldsymbol{\gamma}}_1$ is a vectorized version of $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\boldsymbol{\beta}}_{11}, \widehat{\boldsymbol{\beta}}_{12}, ..., \widehat{\boldsymbol{\beta}}_{1p})$ (the $s \times p$ matrix) concatenating the s rows of $\widehat{\boldsymbol{\beta}}_1$ one after another.

We may then compute p-values by comparing the observed test statistic $\widehat{D}_n$ to the distribution of $\sum_{j=1}^{ps}\widehat{a}_j\chi_1^2$, where $\widehat{a}_j$ is an estimate of a_j based on a consistent estimator for Σ (see Appendix for details). Therefore, this allows us to derive a p-value for the significance of a gene with regards to the variable(s) to be tested. In practice, we make use of saddlepoint approximation for distributions of quadratic forms Kuonen [1999]; Chen and Lumley [2019] to compute p-values for the mixture

of χ^2s, implemented in the `survey` R package [Lumley, 2004]. Note that we obtain a simple limiting distribution without relying on any distributional assumptions on the gene expression. In fact, based on the results in Li and Duan [1989], our test will be asymptotically valid as long as there exist any $g(\cdot)$ and any $\beta_{0j}, \beta_{1j}, \beta_{2j}$ such that (4.1) holds. Lastly, the great number of tests requires the Benjamini–Hochberg [Benjamini and Hochberg, 1995] correction afterwards, which is automatically applied to the raw p-values in the R package of `ccdf`.

4.2.2 Permutation test

Permutation tests are a simple way to obtain the sampling distribution for any test statistic, under the null hypothesis that there is no link between the outcome Y and the variable X. The observations of X can then be shuffled. Permutation tests are recommended when the number of observations is too small, so that the asymptotic distribution can not be assumed to hold. When the sample size n is low, we propose to perform permutations to estimate the empirical distribution of $\widehat{D}_n$ under the null hypothesis. We distinguish two cases: i) testing the association between Y and X without any covariate and ii) testing the association between Y and X given a covariate Z.

i) In the absence of covariates. Under the null hypothesis, Y and X are independent, so the observations of X are exchangeable. Hence, we can randomly permute the observations of X.

ii) In the presence of covariates. When we need to perform a conditional independence testing with a covariate Z, the observations of X are not exchangeable without conditioning on Z. Indeed, if we randomly permute the observations of X, we break not only the link between X and Y but also the link between X and Z. To preserve the dependency between X and Z, we are facing two cases: (a) if Z is a categorical variable and (b) if Z is continuous. In case (a), we randomly switch X within the groups defined by the categories of Z. Under scenario (b), the permutations become tricky. The idea is to permute two observations of X only if the two corresponding observations of Z are close. To do so, a conditional permutation algorithm based on

the distance between the observations of Z is proposed in supplementary materials.

Following the appropriate method, we can permute the observations of X. Under i), we are able to compute the test statistic D from the observations of Y and the permuted observations of X while under ii), the test statistic is obtained from the observations of Y and Z as well as the permuted observations of X. Then, under the null hypothesis, B permutation-based test statistics $\{D_1^*, ..., D_B^*\}$ are calculated and p-values are computed as $\widehat{p} = \frac{1}{1+B}\left(1 + \sum_{b=1}^{B} \mathbb{1}_{\{\widehat{D}_n \leq D_b^*\}}\right)$, to avoid getting zero p-values [Phipson and Smyth, 2010], where D_b^* is the test statistic obtained in the b-th permutation. Finally, we apply the Benjamini–Hochberg correction to the raw p-values to obtain FDR-adjusted p-values.

4.3 Simulation study

4.3.1 Comparisons with state-of-the-art methods in the two conditions case

We compared the performance of our method `ccdf` with three state-of-the-art methods, `MAST`, `scDD` and `SigEMD`, to find differentially distributed (DD) genes. We have selected methods implemented on R that have shown good results in Wang et al. [2019] benchmarks and specially designed for single-cell data (excluding methods for bulk RNA-seq like `edgeR` and `DESeq2`). Also, we considered methods that use normalized (and therefore continuous) counts as input, so that the results of our simulations are comparable. We generated simulated count data from negative binomial distributions and mixtures of negative binomial distributions. Since `MAST`, `scDD` and `SigEMD` require continuous input, we transformed the counts into continuous values while preserving as much as possible the count nature of the data (*i.e.* negative binomial assumption). To do so, we added a small Gaussian noise with mean equal to 0 and variance equal to 0.01 to the simulated counts. 500 simulated datasets were generated including 10,000 genes each, of which 1,000 are differentially expressed under a two conditions setting for 7 different sample sizes n (20, 40, 60, 80, 100, 160, 200). The

observations were equally divided into the two groups. Korthauer et al. [2016] classified four different patterns of unimodal or multi-modal distributions: differential expression in mean (DE), differential proportion (DP), differential modality (DM) and both differential modality and different component means within each condition (DB). We used these four differential distributions to create our own simulations. Specifically, we simulated 250 DD genes, 250 DM genes, 250 DP genes and 250 DB genes. Plus, 9,000 non-differentially expressed genes are simulated according to two non-differential scenarios (see Supplementary Materials for the simulation settings).

Figure 4-2 shows the Monte-Carlo estimation over the 500 simulations of the type-I error and the statistical power as well as the false discovery rate and true discovery rate, according to increasing samples sizes. The type-I error is computed as the proportion of significant genes among the true negative and the power as the proportion of significant genes among the true positive. After Benjamini-Hochberg correction for multiple testing, the FDR is computed as the proportion of false positives among the genes declared DD and the TDR as the proportion of true positives among the genes declared DD. The nominal p-value is fixed to 5%. The three state-of-the-art methods as well as `ccdf` exhibit good control of type-I error and no inflation of FDR. The four methods present a high overall True Discovery Rate. However, `MAST` shows a lack of power of about 23% for a sample size equal to 200. The True Positive rate (after Benjamini-Hochberg correction) for each scenario for all the methods is shown Figure 4-3. The three leading methods perform well in finding the traditional difference in mean (DE) as soon as a number of 60 observations is reached. `ccdf` is less powerful for a sample size of 20 and 40 cells but shows the same performances with a larger sample size. The difference in modality (DM) is well detected by all the methods with a slight advantage for `SigEMD` in low sample sizes (from 20 to 60 cells). The difference in proportion (DP) is not favorable for the asymptotic test of `ccdf` until 160 observations but the permutation test exhibits higher power. The asymptotic test requiring a sufficiently large number of observations to converge, the permutation test is more efficient for a lower number of cells. `SigEMD` and `scDD` are the most effective in detecting DP genes. Even though `MAST` shows good power for DE, DM and DP genes,

it fails to detect DB genes. In fact, `MAST` is designed to detect difference in the overall mean (traditional differential expression), which is absent in DB scenario. The difference in modality and in different component means is then overlooked as expected with its parametric model. `SigEMD`, `scDD` and both `ccdf`'s tests present competitive performances. Generally speaking, `ccdf` retains competitive performance with the state-of-the-art in this two condition benchmark, which makes it a method particularly adapted to traditional DEA. Even though `ccdf` is a non-parametric method, the asymptotic test is relatively reasonable in terms of computation times, especially compared to `SigEMD` (see Table 4.1). If the computation times seem too large, we recommend to use the adaptive thresholds strategy explained in the Supplementary Materials and in particular for the permutation test, we recommend to switch to the adaptive permutations also described in the Supplementary Materials.

Table 4.1: Computation times for the state-of-the-art methods and `ccdf` in the two condition case for one run and $n = 100$, using a MacBook Pro 2019 (2,3 GHz Intel Core i9 8 cores 16 threads)

Method	**Computation times in min**
`ccdf` asymptotic test	3
`ccdf` permutation test	1385
`ccdf` permutation test with adaptive permutations	750
`SigEMD`	1680
`MAST`	1.4
`scDD`	0.05

4.3.2 Multiple comparisons

This second scenario deals with a multiple comparisons design where cell observations were split into 4 different groups. Count data were generated from negative binomial distributions and mixtures of negative binomial distributions and transformed into continuous values as in section 4.3.1. 500 simulated datasets were generated including 10,000 genes each, of which 1,000 are differentially expressed under a four conditions setting for 7 different sample sizes n (40, 80, 120, 160, 200, 320, 400). The observations were then equally divided into four groups. Instead of generating two

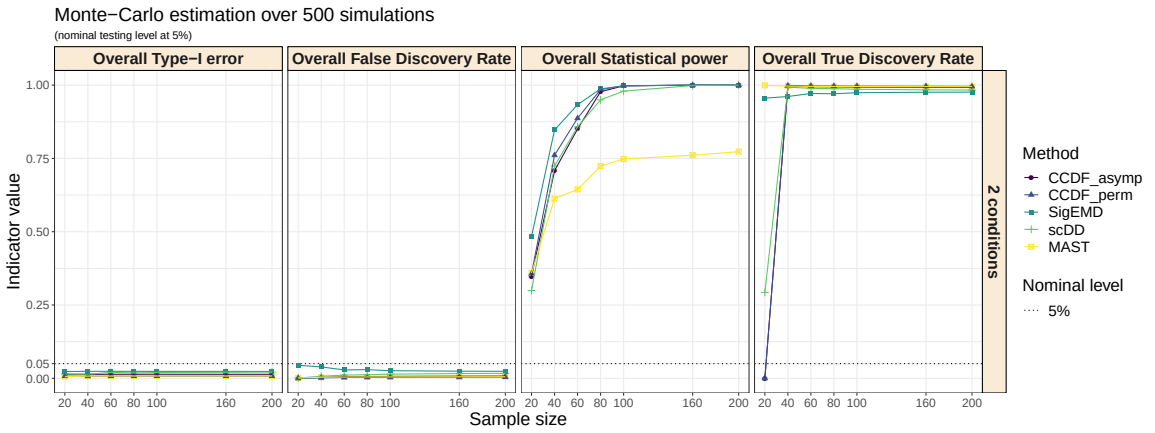


Figure 4-2: **Overall Type-I error, Power, FDR and TDR under the 2 conditions case with increasing sample size**. For `ccdf`, we perform the asymptotic test and the permutation test.

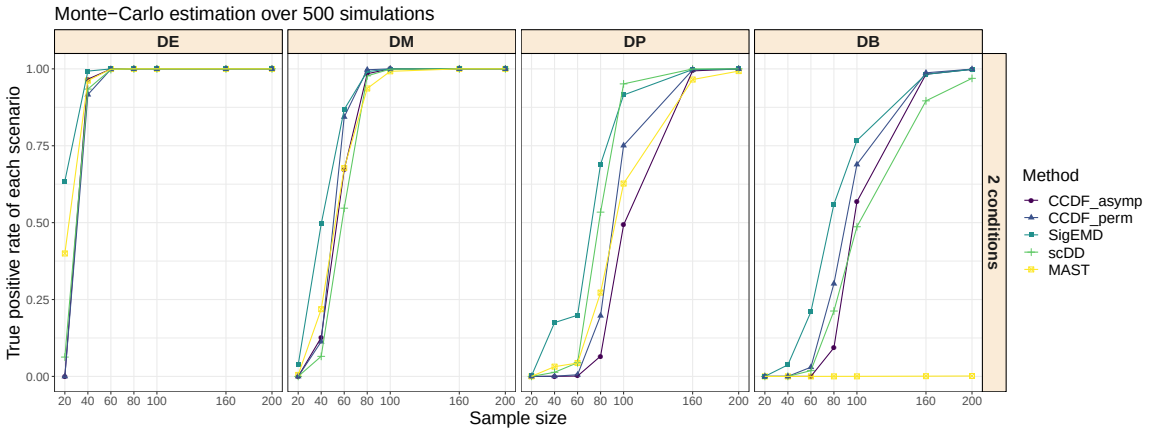


Figure 4-3: **True positive rate under the 2 conditions case for the four DD scenarios with increasing sample size.** DE: difference expression in mean. DM: difference in modality. DP: difference in proportion. DB: both differential modality and different component means within each condition. For `ccdf`, we perform the asymptotic test and the permutation test.

distributions for each gene as in section 4.3.1, we created four distributions and therefore new DD scenarios for this specific DEA simulation: multiple DE, multiple DP, multiple DM and multiple DB (more details in the Supplementary Materials). The non-differentially expressed genes are also simulated in a specific fashion described in the Supplementary Material. The idea of differential distribution patterns was converted into a multiple differential distribution setting. For example, the multiple DD scenario consists in four distributions with four different means. Since the other

approaches can not handle this type of design, only `ccdf` with the asymptotic test and the permutation test is run. Figure 4-5 shows that both versions of `ccdf` have great power to identify DE and DM genes as the sample size increases. DP and DB genes require larger number of cells to achieve sufficient power (from $n > 200$).

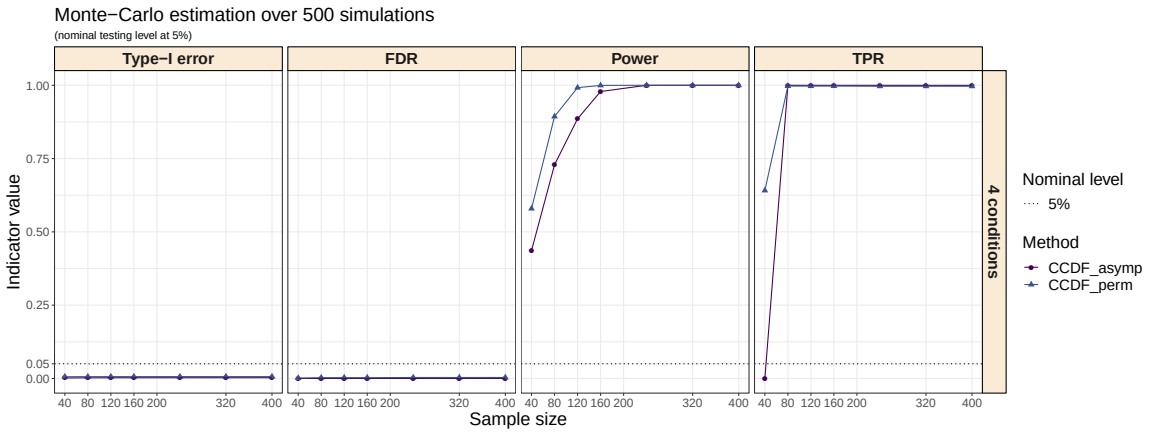


Figure 4-4: **Overall Type-I error, Power, FDR and TDR for `ccdf` under the 4 conditions case with increasing sample size**. `ccdf` is the only method capable of dealing with more than 2 conditions. We perform the asymptotic test and the permutation test.

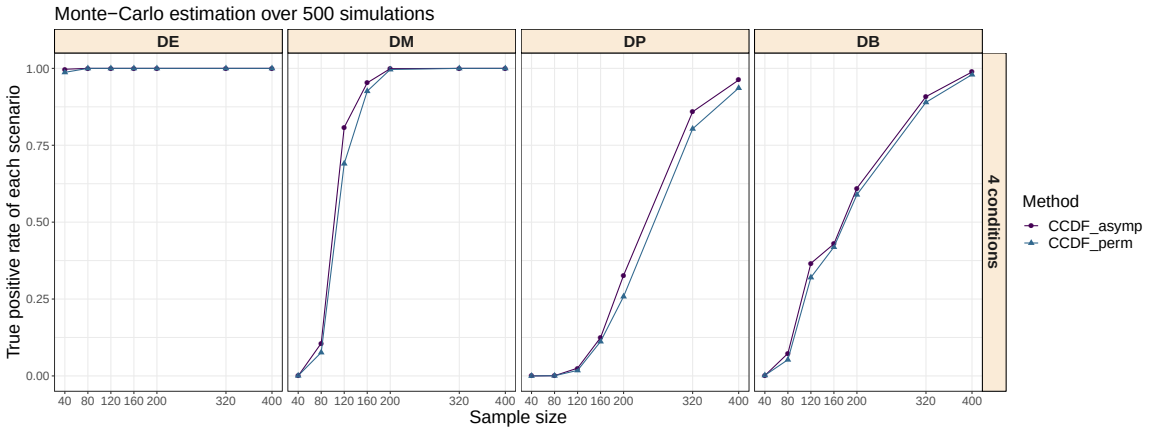


Figure 4-5: **True positive rate under the 4 conditions case for the four DD scenarios with increasing sample size.** DE: difference expression in mean. DM: difference in modality. DP: difference in proportion. DB: both differential modality and different component means within each condition.`ccdf` is the only method capable of dealing with more than 2 conditions. We perform the asymptotic test and the permutation test.

4.3.3 Two conditions comparison given a covariate Z

As represented in Figure 4.1, a potential confounding covariate Z can alter the test of the link between the outcome Y and the variable X. Then, it is needed to adjust to the confounding variable by carrying out a CIT. We aim to emphasize the importance of taking into account Z, thanks to our approach, by showing the erroneous results obtained not doing so. For this purpose, we simulated a confounding variable Z from a Normal distribution. The values of the variable to be tested X depends on the quartile of Z, creating a strong link between X and Z. Y is constructed from X for DE genes and from Z for non-DE genes, the last case is particularly misleading if one do not take into account Z (see simulation settings in Supplementary Material).

We simulated 10,000 genes of which 1,000 are differentially expressed for several sample sizes n (20, 40, 60, 80, 100, 160, 200). `ccdf` permutation test was excluded in this simulation scheme because of the large amount of time to compute. In fact, when we have to adjust for a covariate, the permutation test is not suited to such large sample sizes and number of genes because of the underlying permutation strategy.

`MAST` is able to adjust for covariates like `ccdf` so we expect good performances from both of them. Conversely, `scDD` and `SigEMD` can not control for confounding variables that might impact the detected DE genes. The results of the benchmark between `MAST`, `scDD`, `SigEMD` and `ccdf` are depicted in Figure 4-6. Under the alternative hypothesis, we created a large difference in the mean of the normal distributions between the two conditions in order to make DE genes easier to identify. We see therefore that `scDD`, `SigEMD` and `ccdf` exhibit high power at all sample sizes whereas `MAST` tends to be less powerful for a given size. As expected, the link between Y and Z is very confusing for `scDD`, `SigEMD` which interpret it as a link between Y and X, since X is constructed from Z. Consequently, we observe a consistent rise of Type-I error. `ccdf` and `MAST` perfectly control the Type-I error. In addition, `scDD` and `SigEMD` suffer greatly in terms of TDR as the number of cells is increasing. In fact, fewer real discoveries are found meaning that more genes are identified as significant while being actually false positive, which is in line with the drastically inflated FDR. Dealing

with this complex design, `ccdf` outperforms the leading methods by controlling the Type-I error as well as the FDR and by providing a powerful test. This simulation study highlights the importance of taking into account the confounding variables that may exist. Otherwise, DEA may lead to inaccurate results and a potential huge amount of false positives. It is worth mentioning that `ccdf` can adjust for more than one covariate using the asymptotic test while preserving relatively fast calculation times. The permutation test is for now limited to one adjustment variable and the computation times are obviously increased due to the permutation strategy.

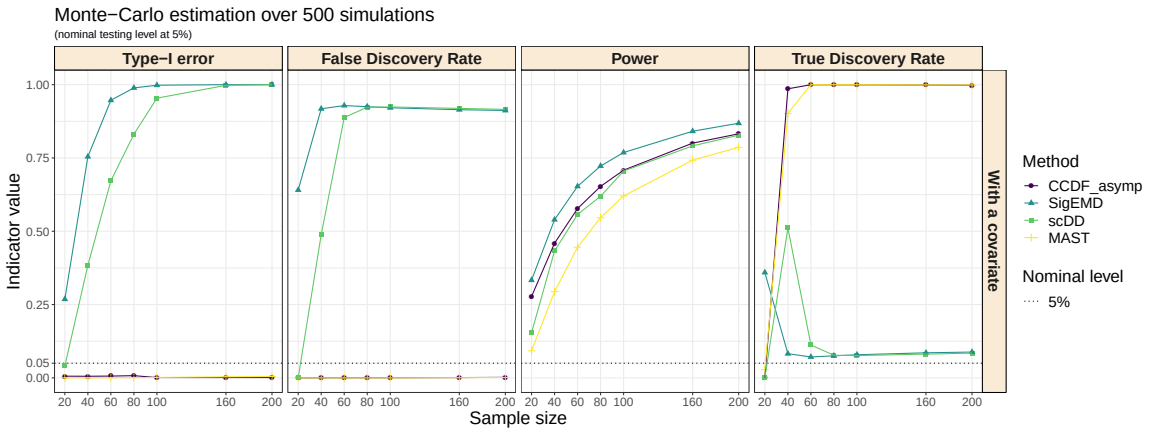


Figure 4-6: **Overall Type-I error, Power, FDR and TDR under the 2 conditions comparison given a confounding variable with increasing sample size**. For `ccdf`, we perform the asymptotic test and the permutation test.

4.4 Comparisons using real data benchmarks

To be in a more realistic context with a greater number of zeros, a positive control data set and a negative control data set described in Wang et al. [2019] were used in order to compare the performances of several methods. The genes with a variance equal to zero were removed from the datasets and counts were converted into log-transformed count per millions (CPM) values.

Table 4.2: Number of detected DE genes, and sensitivities of the state-of-the-art methods and `ccdf` tools using positive control real data for an adjusted p-value of 0.05

Method	Number of DE genes	(TP/1000 gold standard)
`ccdf`	9,353	0.734
`SigEMD`†	3,702	0.488
`scDD`†	2,638	0.351
`MAST`†	734	0.198

† results from Wang et al. [2019].

4.4.1 Positive control dataset

Islam et al. [2011] dataset includes 22,928 genes measured across 48 mouse embryonic stem cells and 44 mouse embryonic fibroblasts. We considered 1,000 genes validated through qRT-PCR experiments as a gold standard gene set in the same fashion as Wang et al. [2019] in order to compute true positive rate. A gene is defined as a true positive if it is found as DE by the method and belongs to the gold standard set [Moliner et al., 2008]. For `ccdf`, `SigEMD`, `scDD` and `MAST`, the number of DE genes for an adjusted p-value of 0.05 and the number of true positive over the 1,000 gold standard genes are given in Table 4.2. `ccdf` leads to the highest true positive rate (0.734) and enables to identify 24.6% more genes in common with the top 1,000 genes compared to `SigEMD` (0.488). `scDD` shows a true positive rate of 0.351 and `MAST` exhibits the lowest rate (0.198) of all the tools.

4.4.2 Negative control dataset

To get false positive rate, we used the dataset from Grün et al. [2014]. We selected 80 samples under the same condition. To create 10 datasets, we randomly divided these 80 cells into 2 groups of 40 cells. As there is no difference between the two groups, not a single gene is to be found in each dataset. Performance evaluations are compared across methods in Table 4.3. `ccdf` and `MAST` do not detect any genes which was expected. Although all the cells are under the same condition, `scDD` and `SigEMD` identified respectively 5 and 50 DE genes.

Table 4.3: Number of the detected DE genes and false positive rates (FPR) of the state-of-the-art methods and `ccdf` using negative control real data for an adjusted p-value of 0.05

Method	Number of detected DE genes	FPR
`ccdf`	0	0
`scDD`†	5	0.0007
`MAST`†	0	0
`SigEMD`†	50	0.007

† results from Wang et al. [2019].

4.5 Application to a scRNA-seq study in COVID-19 patients

Kusnadi et al. [2021] publicly released a large scRNA-seq data set on COVID-19, which is available from GEO (GSE153931). It includes UMI counts for 13,816 genes measured across 84,140 virus-reactive CD8+ T cells. Cellular gene expression was measured in 38 COVID-19 patients and 8 healthy controls. Among the COVID-19 patients, 9 were treated in Intensive Care Unit (ICU), 13 were treated in standard hospital wards (non-ICU) and 16 were not hospitalized at all. This data set features a particularly complex design for DEA (see table 4.5). Kusnadi et al. [2021] regrouped the 22 hospitalized patients as having a severe COVID-19 while the 16 non-hospitalized were considered as mild, and they compared the gene expression across CD8+ T cells between them. Instead, we went further in analyzing the differences between the three COVID-19 categories, namely ICU, non-ICU hospitalized and not hospitalized to provide better insights between the different severity of COVID-19.

As often done in scRNA-seq analyses, Kusnadi et al. [2021] clustered the 84,140 cells, resulting in 8 different clusters. Table 4.5 presents the numbers of cells assigned to each cluster (Kusnadi et al. [2021] excluded cluster 7 due to its small size). These clusters represent various CD8+ T cell sub-populations with different states of cell differentiation, with cluster 1 being composed by exhausted (dying) cells for instance. Some biological pathways, activated across some clusters, could also be associated to the disease severity (such as pro-survival features as reported in the original paper

[Kusnadi et al., 2021]). Hence, it is not clear whether the association of the expression of some genes with the severity of the disease is either reflecting a difference in abundance of some clusters of CD8+ T cell populations, or due to the activation of some pathways independently of the clusters. We propose to compare the gene expression according to the severity of the disease while adjusting or not for the clusters, which might be acting as a confounding factor. Both the outcome and the potential founding factor are categorical variables with more than 2 levels, making our analysis design relatively complex.

Table 4.4: Study design

	Healthy subjects	**COVID-19 patients**		
		Not admitted	**non-ICU (hospitalized)**	**ICU (hospitalized)**
Sex				
Male	7	7	10	7
Female	1	9	3	2
Total	8	16	13	9

Table 4.5: `Seurat` clustering

	Cluster							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number of cells	27086	17915	12270	9783	5688	5170	829	156

With our notations, Y is the gene expression, X the COVID-19 severity status and Z the cell cluster. We apply the asymptotic `ccdf` test for both analyses (with and without adjusting on Z) while testing each gene for differential expression across COVID-19 severity status. The number of thresholds to compute the marginal and conditional CDFs is set to 10. Following Kusnadi et al. [2021], only transcripts expressed in at least 0.1% of the cells were included in the differential analyses, yielding 10,525 genes in total to be tested and UMI data are then converted to log_2(CPM+1). 6,290 genes are found DE when adjusting on the clusters, while 8,181 genes are DE when those are ignored (see Figure 4-7).

To understand where the difference of 2,129 DE genes comes from, we focused specifically on the 221 genes included in the IFN gene set, one of the 7 pathways highlighted by Kusnadi et al. [2021] (see their supplementary material Table S5 for their definition). Among these genes, 188 were found significant regardless of the adjustment on the clusters while 7 become significant and 20 become not significant when adjusting. Upon a closer look at the adjusted p-values of each gene, we notice that adjusting for the cell populations changed the individual p-values. Figure 4-8 shows that some genes with low p-values in unadjusted DEA have much higher p-values after adjustment for clusters (e.g. IFIT3, IFI6, TDRD7, IFI44L, IFI44, IFITM2, OAS2, HERC6 and CD38) whereas others such as LCP2 do not exhibit a strong variation of the p-value. This could be expected when looking at the abundance of the genes according to the clusters (Table S4 in Kusnadi et al. [2021]) as for instance IFIT3, IFI6, IFI44 were much more abundant in cluster 1 than in other clusters, whereas LCP2 abundance was more balanced across clusters. As a visualization example, figure 4-9 display the various conditional CDFs estimated through both `ccdf` analyses. It illustrates the reduction of the difference between the conditional CDF on X and its marginal counterpart (when not conditioning on X) when additionally conditioning on Z the clusters. This is reflected by the p-value which increases from 0 to 6.2e-36 with the conditioning on Z: while IFI6 remains DE in both analyses, its significance is decreased when taking into account the different cell clusters. While the majority of IFN genes remain significant at a threshold of 5% in both DEA, there is a general increasing trend of the p-values when we adjust for the clusters. This supports some confounding effect on the significance of some genes within the Interferon pathway. Thus, taking into account the clusters does not fundamentally change the biological interpretation that Interferon pathway plays an important role in the severity status, but that part of its difference in expression is mediated through the identified clusters. See supplementary figures for additional visualizations of `ccdf` results.

An obvious limit of this real data analysis lies in the double use of the data. First Kusnadi et al. [2021] performed a clustering to define cell populations that were not

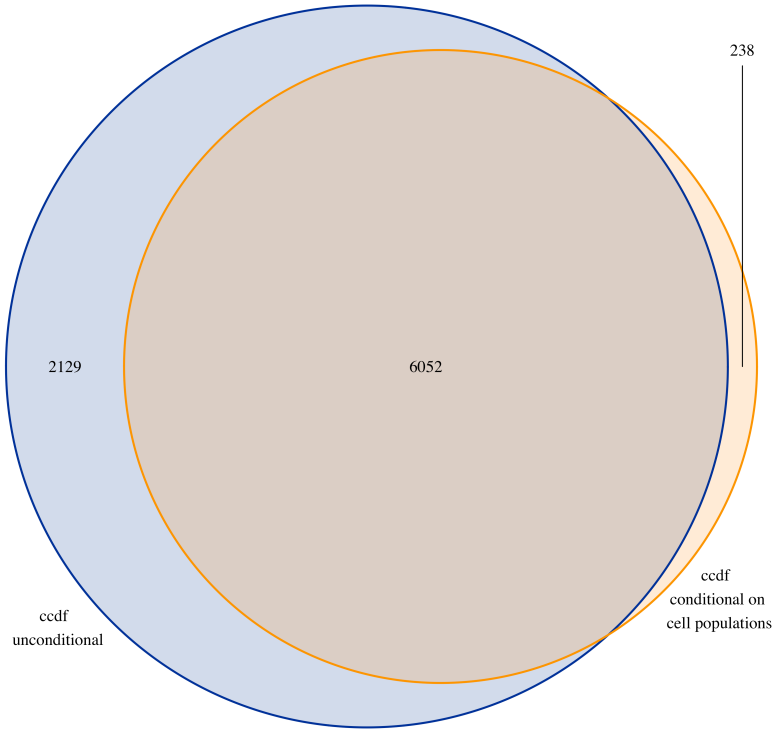


Figure 4-7: Venn diagram of the gene signature found by `ccdf` conditional on 7 cell populations and the gene signature found by `ccdf` without conditioning on the cell populations (`ccdf` unconditional)

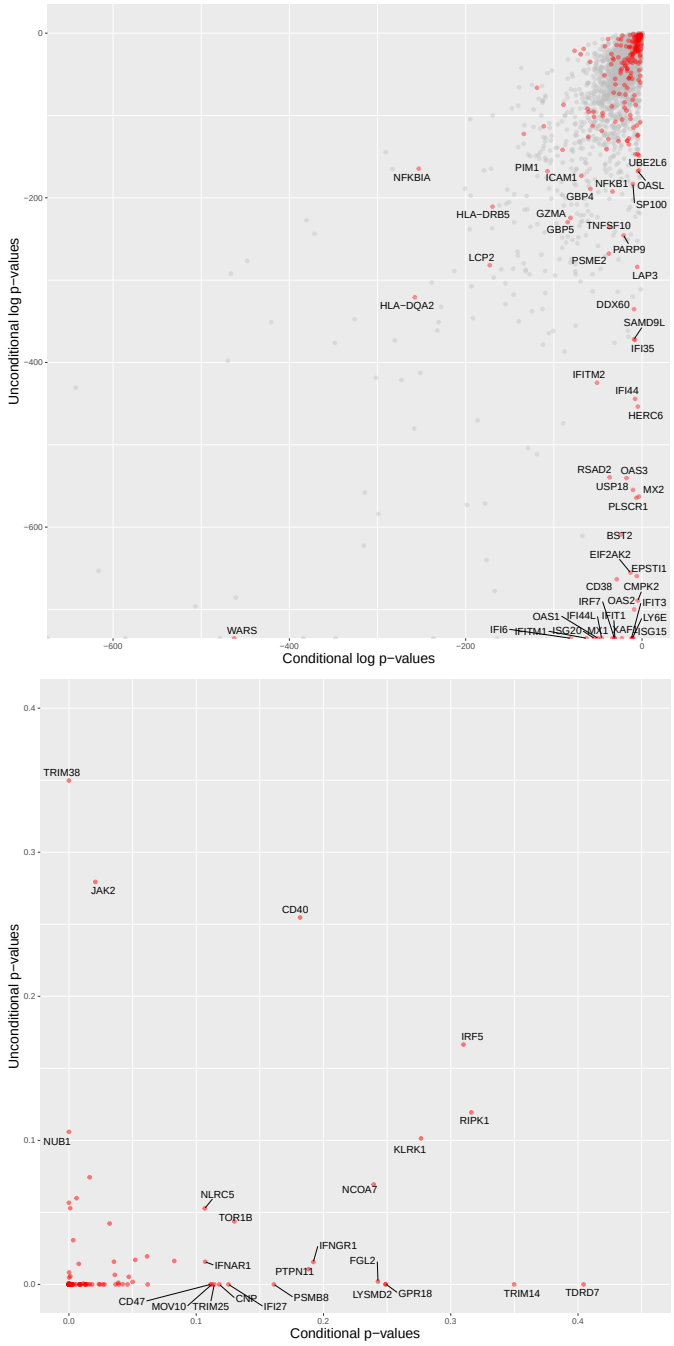


Figure 4-8: Top figure: scatter plot of the genes p-values in log-scale. The x-axis represents the log p-values obtained with `ccdf` conditional on cell populations. The y-axis represents the log p-values obtained with `ccdf` without conditioning on the cell populations. The red points correspond to the genes of the IFN list. The grey points correspond to the genes that are not part of the IFN list. Bottom figure: scatter plot of the IFN genes p-values between 0 and 0.5. The x-axis represents the p-values obtained with `ccdf` conditional on cell populations. The y-axis represents the p-values obtained with `ccdf` without conditioning on the cell populations.

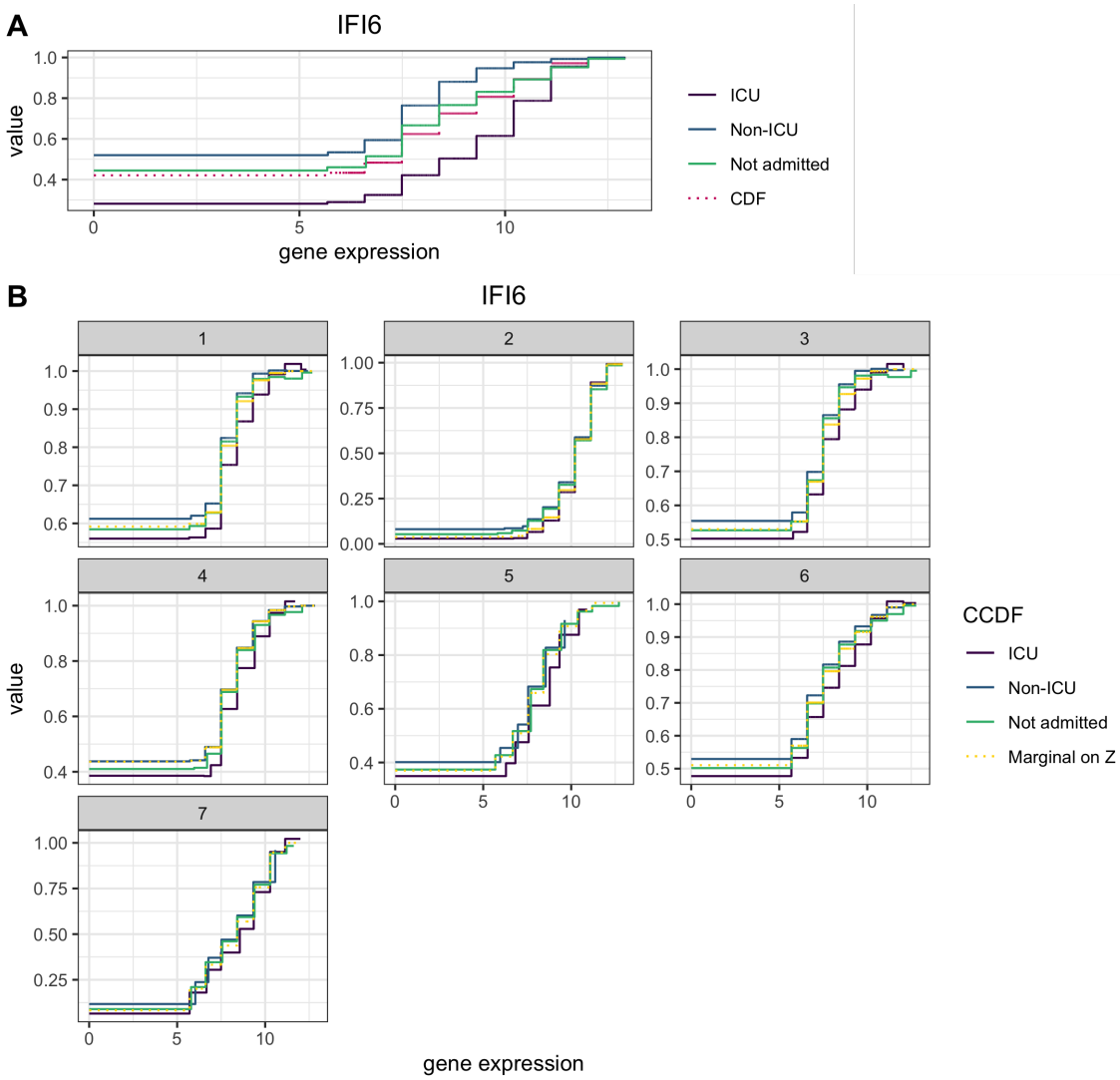


Figure 4-9: A: The solid lines represent the conditional CDF of IFI6 gene on ICU status (ICU, Non-ICU and Not admitted) and the dark pink dotted line represents the marginal CDF of IFI6 gene, *i.e.* without conditioning on ICU status. The underlying test performed by `ccdf` in the first DEA consists in comparing the marginal CDF with the conditional CDF. B: The solid lines represent the conditional CDF of IFI6 gene on both Z, the 7 clusters, and X, the severity status (ICU, Non-ICU and Not Admitted), while the dotted yellow line represents the marginal CDF of IFI6 expression without conditioning on X (but only conditioning on Z the clusters). The underlying test performed by `ccdf` now consists in comparing the marginal CDF on Z with the conditional CDF on both X and Z. The two CDFs (the dotted line and the solid lines) are much closer which means the variable X has less impact on the conditional CDF. The number of steps of the CDF matches the thresholds chosen in `ccdf` (10 in the analysis). The first value of each CDF is the proportion of zeroes.

annotated beforehand. Then, we use once again those clusters of cells as a covariate in our DEA. The clustering is based on the cell gene expression, therefore the association between the clusters and the gene expression – needed for the clusters to play the role of a confounding variable – is constructed during the first step. Although this example was used as an illustration of the `ccdf` method, it is quite realistic thanks to high-throughput technologies that allow nowadays to measure cell characteristics (surface and intracellular proteins) as well as single-cell gene expression.

4.6 Discussion

We propose a new framework for performing CIT, with an immediate application to differential expression analysis of scRNA-seq data. Our approach can accommodate complex designs, e.g. with more than two experimental conditions or with continuous responses while adjusting for several additional covariates. `ccdf` is capable of distinguishing differences in distribution by using a CIT based on the estimation of CCDFs through a linear regression. The resulting asymptotic test is attractive due to the low computation times, especially dealing with a high number of observations. Yet, for small samples sizes (e.g. due to experimental design or cost limitations in data acquisition), we cannot always rely on an asymptotic test. Consequently, a permutation test is proposed in such cases. Performing permutations is obviously time consuming, but as it is necessary only for small sample sizes, computation times remain reasonable in such settings. Nevertheless, easy parallelization of the permutation test can alleviate this problem. Furthermore, an adaptive procedure, for both the number of permutations and the number of thresholds, is implemented in order to accelerate `ccdf` while preserving a sufficient statistical power along with numerical precision for the lowest p-values. Per-gene asymptotic tests can also be computed in parallel to speed up computations. The proposed approach has been fully implemented in the user-friendly R package `ccdf` available on CRAN at `https://CRAN.R-project.org/package=ccdf`.

While `ccdf` can be applied to many types of data thanks to its flexibility, it has been specifically tailored for the need of scRNA-seq data DEA. In the simulation

study, `ccdf` exhibits great results and versatility in complex designs such as multiple conditions comparison, and also allows to analyze data when the experiment design includes confounding variables while most of the competing methods cannot. Finally `ccdf` maintains great power while ensuring an effective control of FDR. The application to a real single-cell RNA-seq data set enhances the importance of being able to handle complex experimental designs, especially with a potential confounding variable. Of note, `ccdf` being a distribution-free approach, it can support any normalization method (and this choice is ultimately left to the responsibility of the user).

Tiberi et al. [2020] recently proposed non-parametric permutation approach that also compares empirical CDF. While `distinct` shares some common ideas with `ccdf`, the two approaches widely differs in their test statistics and capabilities. On the one hand `distinct` requires multiple samples and only addresses the two groups comparison, while on the other hand `ccdf` provides an asymptotic test and can accommodate more complex experimental designs.

The number of evaluating thresholds considered in `ccdf` directly impacts both its computation time but also potentially its statistical power. To optimize this trade-off and to maintain the performances while reducing the computational cost, we propose to use a sequence of evenly spaced thresholds. Besides, instead of choosing a regular sequence of thresholds, one can prefer to manually chose the different thresholds at which CCDFs are computed and compared, e.g. using *prior* knowledge to emphasize some areas of the distribution.

Although the linear regression estimation of the CCDF is efficient, different parameterizations can be considered (e.g. logistic regression which would be more natural). Currently `ccdf` cannot analyse multi-sample data, such as biological replicates or repeated measurements. This would consist in modifying the structure of the working model in equation (4.1). `ccdf` would thus be able to perform multi-sample analysis while keeping the advantage of its asymptotic test, at the cost of an increased computational burden. Besides, other test statistics could be investigated based on the same CIT.

Software

The proposed method has been implemented in an open-source R package called `ccdf`, available on CRAN at https://CRAN.R-project.org/package=ccdf. All R scripts used for the simulations and the real data set analyses in this article were performed using R v3.6.3 and can be found on GitHub at https://github.com/Mgauth/ccdf.

Acknowledgments

MG is supported within the Digital Public Health Graduate's school, funded by the PIA 3 (Investments for the Future - Project reference: 17-EURE-0019). The project is supported through SWAGR Inria Associate-Team from the Inria@SiliconValley program. Computer time for this study was provided by the computing facilities MCIA (*Mésocentre de Calcul Intensif Aquitain*) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. BH & MG thank Franck Picard for helpful discussion about this work. *Conflict of Interest*: None declared.

4.7 Supplementary Materials

4.7.1 Parameter estimation

We have $\boldsymbol{Y} \in [\zeta_{\min}, \zeta_{\max}]$ for some known constants ζ_{min}, ζ_{max}. Let $\zeta_{\min} \leq \omega_1 < \omega_2 < ... < \omega_p < \zeta_{\max}$ is a sequence of p ordered and regular thresholds. Let the design matrix W be the matrix with i^{th} row $\boldsymbol{W}_i = (1, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $\tilde{\boldsymbol{Y}}_i = (\tilde{Y}_{11}, ..., \tilde{Y}_{1p})$ for the p thresholds. The full set of regression coefficients for the j^{th} regression is written $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{0j}, \widehat{\boldsymbol{\beta}}_{1j}, \widehat{\boldsymbol{\beta}}_{2j})^\intercal$. By making use of OLS, we can write

$$\widehat{\boldsymbol{\beta}}_j = (W^\intercal W)^{-1} W^\intercal \tilde{\boldsymbol{Y}}_j \tag{4.4}$$

$$= (n^{-1} W^\intercal W)^{-1} n^{-1} W^\intercal \tilde{\boldsymbol{Y}}_j + o_p(1) \tag{4.5}$$

where $\tilde{\boldsymbol{Y}}_j = (\tilde{Y}_{1j}, ..., \tilde{Y}_{nj})^\intercal$. Then,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j &= n^{-1} \sum_{i=1}^{n} \Psi^{-1} \boldsymbol{w}_i \tilde{Y}_{ij} + \left\{ (n^{-1} W^\intercal W)^{-1} - \Psi^{-1} \right\} n^{-1} \sum_{i=1}^{n} \boldsymbol{w}_i \tilde{Y}_{ij} \\ &= n^{-1} \sum_{i=1}^{n} \Psi^{-1} \boldsymbol{w}_i \tilde{Y}_{ij} + o_p(n^{-1/2}) \end{aligned} \tag{4.6}$$

where $\Psi = \mathbb{E}(\boldsymbol{w}_i \boldsymbol{w}_i^\intercal) = \lim_{n\to\infty} n^{-1} W^\intercal W$. Denote $\boldsymbol{h}_i = (h_i^1, ..., h_i^s)^\intercal$ such as $\boldsymbol{h}_i$ is the matrix product of the rows of Ψ^{-1} associated with $\boldsymbol{X}_i$ and $\boldsymbol{w}_i$. From (4.6), $\widehat{\boldsymbol{\beta}}_{1j}$ can be expressed by

$$\widehat{\boldsymbol{\beta}}_{1j} = n^{-1} \sum_{i=1}^{n} \boldsymbol{h}_i \tilde{Y}_{ij} \tag{4.7}$$

Recall that $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\boldsymbol{\beta}}_{11}, \widehat{\boldsymbol{\beta}}_{12}, ..., \widehat{\boldsymbol{\beta}}_{1p})$ is a $s \times p$ matrix then we deduce from (4.7) the expression of the matrix of the s coefficients corresponding to $\boldsymbol{X}$ for the thresholds ω_js, $\forall j = 1, ..., p$, as

$$\widehat{\boldsymbol{\beta}}_1 = n^{-1} \sum_{i=1}^{n} \boldsymbol{h}_i \tilde{\boldsymbol{Y}}_i \tag{4.8}$$

Last, in order to simplify, we transform the matrix $\widehat{\boldsymbol{\beta}}^1$ of size $s \times p$ into a vector $\widehat{\boldsymbol{\gamma}}^1$ of size ps by concatenating the s lines of $\widehat{\boldsymbol{\beta}}^1$ one after another.

Note that regression estimates are still consistent up to a multiplicative scalar Li and Duan [1989], so the use of linear regression instead of logistic regression remains valid.

4.7.2 Asymptotic test

After plugging in the estimate of β_{1j} for the estimated test statistic, we need to derive the resulting asymptotic distribution under the null hypothesis. Knowing $\widehat{\boldsymbol{\beta}}_1$, by the multivariate central limit theorem, we have

$$\sqrt{n}\,(\widehat{\boldsymbol{\gamma}_1} - \boldsymbol{\gamma}_1{}^*) \longrightarrow N(0, \Sigma) \tag{4.9}$$

where $\boldsymbol{\gamma}_1^*$ is the true expectation of $\boldsymbol{h}_i\tilde{\boldsymbol{Y}}_i$ (concatenated into a vector). Under the null hypothesis, $\boldsymbol{\gamma}_1{}^* = \mathbf{0}$. Σ is a symmetric positive semi-definite covariance matrix of size $ps \times ps$ that can be estimated by the method of moments: $\Sigma = n^{-1}\sum_{i=1}^{n} \mathrm{Cov}(\boldsymbol{h}_i\tilde{\boldsymbol{Y}}_i) = E\{\mathrm{Cov}(\boldsymbol{h}_i\tilde{\boldsymbol{Y}}_i)\}$ such as the $\big((k-1)p+j, (k'-1)p+j'\big)^{\text{th}}$ entry of $\mathrm{Cov}(\boldsymbol{h}_i\tilde{\boldsymbol{Y}}_i)$) is defined as:

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{h}_i\tilde{\boldsymbol{Y}}_i)_{(k-1)p+j,(k'-1)p+j'} &= \mathrm{Cov}\left(h_i^k \mathbb{1}_{\{Y_i \le \omega_j\}}, h_i^{k'} \mathbb{1}_{\{Y_i \le \omega_{j'}\}}\right) \\ &= \begin{cases} h_i^k h_i^{k'}\big(\pi_j - \pi_j\pi_{j'}\big), & \omega_j \le \omega_{j'} \\ h_i^k h_i^{k'}\big(\pi_{j'} - \pi_j\pi_{j'}\big), & \omega_{j'} < \omega_j \end{cases}\end{aligned}$$

where $\pi_j = P(Y_i \le \omega_j)$ for $j = 1, ..., p$ and h_i^k is the k^{th} element of $\boldsymbol{h}_i$ for $k = 1, ..., s$.

Let $\boldsymbol{\nu} = \sqrt{n}\widehat{\boldsymbol{\gamma}}_1$, then

$$n\widehat{\gamma}_1^{\intercal}\widehat{\gamma}_1 = \boldsymbol{\nu}^{\intercal}\boldsymbol{\nu} = \boldsymbol{\nu}^{\intercal}\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}\boldsymbol{\nu} = \boldsymbol{u}^{\intercal}\Sigma\boldsymbol{u}^{\intercal} \tag{4.10}$$

where $\boldsymbol{u}$ follows asymptotically a multivariate standard normal. Let $\Sigma = UAU^{\intercal}$, where U is an orthonormal set of eigenvectors and A is a diagonal matrix of eigenvalues of Σ. Because of the orthonormality of U, $\tilde{\boldsymbol{u}} = \boldsymbol{u}U$ is also asymptotically multivariate normal.

By (4.9) and (4.10), the observed test statistic $\widehat{D}_n$ is given by:

$$\widehat{D}_n = n\sum_{j=1}^{ps}\widehat{\gamma}_{1j}^2 \tag{4.11}$$

and

$$n\sum_{j=1}^{ps}\widehat{\gamma}_{1j}^2 = \tilde{\boldsymbol{u}}^{\intercal}A\tilde{\boldsymbol{u}} = \sum_{j=1}^{ps}\widehat{a}_j\tilde{\boldsymbol{u}}_j^2 \tag{4.12}$$

which is asymptotically a mixture of χ_1^2 random variables. The estimated mixing coefficient $\widehat{a}_j$ is the j^{th} diagonal element of A. Note that the form of the estimated test statistic (4.11) is equivalent to the one given in the main manuscript.

Finally, by (4.12), we have the following asymptotic distribution of $\widehat{D}_n$ under the

null hypothesis:

$$\widehat{D}_n \underset{n\to+\infty}{\longrightarrow} \sum_{j=1}^{ps} \widehat{a}_j \chi_1^2 \tag{4.13}$$

4.7.3 Conditional permutation algorithm

For now, the permutation test is only able to take into account an univariate variable X and an univariate covariate Z whereas the asymptotic test can handle a multivariate variable X and a multivariate covariate Z without increasing the computation times. Consequently, if one wants to adjust for many variables, the asymptotic test must be used. Even if the sample size is low, we show in the main manuscript that the asymptotic test remains powerful, so it is still reasonable to use this specific test in this case. When we need to adjust for a continuous covariate Z, the permutation test requires a specific shuffling. In order not to break the link between X the variable to be tested and Z the covariate, we permute the observations of X according to a probability distribution μ_i which takes into account the relationship between X and Z. μ_i is computed in the following way. First, we perform a linear regression of X on Z, then for all $i = 1, ..., n$, for all $j \neq i$, we compute

$$\mu_{ij} = \frac{|\widehat{X}_i - \widehat{X}_j|^{-1}}{\sum_{\ell=1}^{n} |\widehat{X}_i - \widehat{X}_\ell|^{-1}} \tag{4.14}$$

with $\widehat{X}_i$ the predicted value coming from the linear regression. We denote $\mu_i = (\mu_{i1}, \ldots, \mu_{in})$ the vector of probabilities computed according to (4.14). For a permutation step, the observation X_i is then replaced by $X_i^\star$ randomly drawn from $\{X_\ell, \ell \neq i\}$ given the probabilities μ_i. Therefore, for a given i, the observations close to X_i given Z_i have a higher probability to be chosen. By adding some randomness through μ_i, X_i is not replaced by the same observation at each permutation.

4.7.4 Practical considerations for computational speed up

Adaptive permutations

The disadvantage of using permutations could be the onerous computation times, especially when dealing with large sample size, which is more often encountered in single-cell DEA than in bulk DEA. The software computes 1,000 permutations by default for all the genes, but an adaptive procedure may provide similar accuracy at much lower computational cost. When calculation times appear to be too excessive, the user can switch to adaptive permutations. According to some pre-defined rules, the number of permutations is increased at each step to get sufficient numerical precision on the p-values only for certain genes. By default, the method computes 100 permutations for all genes, then we add 150 permutations for the genes with a p-value less than 0.1, bringing the total number of permutations for these genes to 250. Then, for the genes with an associated p-values less than 0.05, we perform 250 permutations more and finally the genes with a p-values less than 0.01, we add 500 permutations to reach 1,000 permutations for a reduced bunch of genes. If the computation times are still too long, the user can choose the number of p-values thresholds and the different limit values. The number of permutations executed at each step is also configurable.

Evaluation thresholds

Selecting the thresholds $\omega_1, \omega_2, \ldots, \omega_p$ where the CCDF is evaluated may be difficult in practice. If too few thresholds are selected, then important changes in $F_{Y|X,Z}(\omega_j \mid x, z)$ may not be detected. One could instead select the thresholds to match the unique observations of Y, even though this selection technically violates the assumption that the thresholds are fixed and independent of the data. Yet, this technical violation does not appear to adversely affect the performance of our approach in simulations (see Section 4.3), and `ccdf` selects the thresholds to match the unique observations of Y by default.

However, there may be an important computational cost to selecting so many thresholds: the number of linear regressions required to estimate all β_1s is then equal

to the sample size. So when analyzing data from a large number of cells, one gets an equally large number of regressions to estimate along with a large matrix Σ, significantly increasing the computation time. One solution to reduce computation times (both for the asymptotic test and the permutation test) is to decrease the number of evaluated thresholds and thus the number of estimated β_1s as well as the dimension of Σ.

Instead of going through all the unique values of Y, one can choose a regular sequence of thresholds. Since single-cell RNA-seq data are count data, we propose spacing these thresholds according to a logarithmic scale, i.e., to better focus on the values where the CCDFs will not be too close to 1. This way, `ccdf` statistical power is maximized as variations in distributions are more likely to appear for smaller values (this point is all the more important as the number of thresholds is small).

4.7.5 Simulations

The code to generate the dataset under the described setting is available from the GitHub repository [1]

Analyses settings

`ccdf` We perform the asymptotic test with as many thresholds as unique values and without using a logarithmic threshold. The permutation test is computed using the adaptive procedure by default (from 100 to 1,000 permutations).

`SigEMD` We compute the Earth's mover distance to compare the distributions without using the imputation of dropouts (because the number of simulated dropouts is very low). `SigEMD`'s test performs a permutation test relying on two parameters to tune: the number of permutations and the bin size (between 0 and 1) to fix the width of the histogram's intervals. Given the fact the method exhibits huge computation times, we select the parameters with a good trade-off between the time of execution

[1] `https://github.com/Mgauth/ccdf`

and the statistical power. Therefore, we arbitrarily fix the number of permutations to 500 and the bin size to 0.2.

`MAST` As described in the user guide [2], we fit a hurdle model and run a likelihood ratio test here, testing for differences in X.

`scDD` The method employs a Bayesian modeling framework requiring some hyper-parameters. We refer to the method's quick start [3] and use the following settings: alpha=0.01, mu0=0, s0=0.01, a0=0.01, b0=0.01. We used the default option involving the Kolmogorov-Smirnov test. This allows to have a faster method despite a slight decrease of the power as indicated by the authors. We plan to add to the benchmark the full `scDD` framework based on permutations.

2 conditions

Korthauer et al. Korthauer et al. [2016] classified four different patterns of unimodal or multi-modal distributions:

- **differential expression (DE)**: two unimodal distributions with a different mean in each condition.
- **differential proportion (DP)**: two bimodal distributions with equal component means across conditions; the proportion in the low mode is 0.3 for condition 1 and 0.7 for condition 2.
- **differential modality (DM)**: one unimodal distribution in condition 1 and one bimodal distribution in condition 2 with one overlapping component. Half of the cells in condition 2 belongs to each mode.
- **both differential modality and different component means within each condition (DB)**: one unimodal distribution in condition 1 and 1 bimodal distribution in condition 2. The distributions have no overlapping components.

[2] https://www.bioconductor.org/packages/release/bioc/vignettes/MAST/inst/doc/MAITAnalysis.html#4_Differential_Expression_using_a_Hurdle_model

[3] https://bioconductor.org/packages/release/bioc/vignettes/scDD/inst/doc/scDD.pdf

The mean of condition 1 is half-way between the overall means in condition 2. Half of the cells in condition 2 belongs to each mode.

The non-differentially expressed genes are divided into two categories:

- **(unimodal distribution with equivalent expression) (EE)**: two unimodal distributions with equal means.
- **bimodal distribution with equivalent proportions (EB)**: two bimodal distributions with equal component means across conditions and equal proportions in the low mode.

Since scRNA-seq data exhibits zero values (dropouts), we simulated beforehand a negative binomial distribution with mean equal to 0.5 in each condition to create zeroes or very low counts. Each distribution described above was simulated after uniformly having placed the zeroes and low counts between the 2 conditions, so that no difference in zero is created. With a sample size n, we chose a proportion of $n_0 = n/10$ for the zeroes and low counts equally divided between the conditions. The negative binomial distribution $NB(m, p)$ with size parameter equal to m and probability parameter equal to p has the following density: $\binom{x+m-1}{x} p^m (1-p)^x$ for $x \in \mathbb{N}^*$ and $p \in]0, 1]$. For simplicity, we kept constant $p = 0.5$ across all the scenarios. Then, for a sample size n, we denote $N = n - n_0$, gene expression is generated with the following parameters: $m_1 = \mathcal{U}(10, 20)$, $m_2 = 3m_1$, $m_3 = m_1 + m_2/2$, $\alpha_0 = 0.5$, $\alpha_1 = n/3$ and $\alpha_2 = 1 - \alpha_1$.

- **DE**: for $i = 1, \ldots, 250$,
$$y_{ij} = \begin{cases} NB(m_1, p) & \text{if } j = 1, \ldots, N/2 \\ NB(m_3, p) & \text{if } j = N/2 + 1, \ldots, N \end{cases}$$
- **DM**: for $i = 251, \ldots, 500$,

$$y_{ij} = \begin{cases} NB(m_2, p) & \text{if } j = 1, \ldots, N/2 \\ \alpha_1 NB(m_1, p) + \alpha_2 NB(m_2, p) & \text{if } j = N/2+1, \ldots, N \end{cases}$$

- **DP**: for $i = 501, \ldots, 750$,

$$y_{ij} = \begin{cases} \alpha_1 NB(m_1, p) + \alpha_2 NB(m_2, p) & \text{if } j = 1, \ldots, N/2 \\ \alpha_2 NB(m_1, p) + \alpha_1 NB(m_2, p) & \text{if } j = N/2+1, \ldots, N \end{cases}$$

- **DB**: for $i = 751, \ldots, 1000$,

$$y_{ij} = \begin{cases} \alpha_0 NB(m_1, p) + (1-\alpha_0) NB(m_2, p) & \text{if } j = 1, \ldots, N/2 \\ NB(m3, p) & \text{if } j = N/2+1, \ldots, N \end{cases}$$

- **EE**: for $i = 1001, \ldots, 5500$,

$$y_{ij} = NB(m_1, p) \quad \forall j = 1, \ldots, N$$

- **EB**: for $i = 5501, \ldots, 10000$,

$$y_{ij} = \alpha_0 NB(m_1, p) + (1-\alpha_0) NB(m_2, p) \quad \forall j = 1, \ldots, N$$

4 conditions

The four scenarios under 4 conditions comparison, inspired by Korthauer et al. Korthauer et al. [2016], are the following:

- **multiple DE**: 4 unimodal distributions with a different mean in each condition.
- **multiple DP**: 4 bimodal distributions with equal component means across conditions; the proportion in the low mode is 0.1 for condition 1 and 0.3 for condition 2, 0.9 for condition 3, 0.7 for condition 4.

- **multiple DM**: 2 unimodal distributions in condition 1 and 2 and 2 bimodal distributions in condition 3 and 4 with one overlapping component. Half of the cells in condition 3 and 4 belongs to each mode

- **multiple DB**: 1 unimodal distribution in condition 1; 1 bimodal distribution in condition 2, 1 distribution with three modes in condition 3 and 1 distribution with 4 modes in condition 4; distributions in condition 2 and 4 have no overlapping components with distributions in condition 1 and 3; distribution in condition 3 have the middle component overlapping the distribution in condition 1. The mean of condition 1 is half-way between the overall means in condition 2, 3 and 4. Half of the cells in condition 2 belongs to each mode, a third of the cells in condition 3 belongs to each mode and a quarter of the cells in condition 4 belongs to each mode.

The non-differentially expressed genes are divided into two categories:

- **multiple EE**: four unimodal distributions with equal means.

- **multiple EB**: four bimodal distributions with equal component means across conditions and equal proportions in the low mode.

Since scRNA-seq data exhibits zero values (dropouts), we simulated beforehand a negative binomial distribution with mean equal to 0.5 in each condition to create zeroes or very low counts. Each distribution described above was simulated after uniformly having placed the zeroes and low counts between the 4 conditions, so that no difference in zero is created. With a sample size n, we chose a proportion of $n_0 = n/10$ for the zeroes and low counts equally divided between the conditions. Then, for a sample size n, we denote $N = n - n_0$, gene expression is generated with the following parameters: $p = 0.5$, $m_1 = \mathcal{U}(10, 20)$, $m_2 = 2m_1$, $m_3 = m_1 + m_2/2$, $m_4 = 3m_1$, $\alpha_0 = 0.25$, $\alpha_1 = 0.1n/4$, $\alpha_2 = 1-\alpha_1$, $\alpha_3 = 0.3n/4$, $\alpha_4 = 1-\alpha_3$, $\delta = 0.9m_3$, $\gamma = 0.6m_3$ and $\epsilon = 0.3m_2$.

- **multiple DE**: for $i = 1, \ldots, 250$,

$$y_{ij} = \begin{cases} NB(m_1, p) & \text{if } j = 1, \ldots, N/4 \\ NB(m_2, p) & \text{if } j = N/4+1, \ldots, N/2 \\ NB(m_3, p) & \text{if } j = N/2+1, \ldots, 3N/4 \\ NB(m_4, p) & \text{if } j = 3N/4+1, \ldots, N \end{cases}$$

- **multiple DM**: for $i = 251, \ldots, 500$,

$$y_{ij} = \begin{cases} NB(m_2, p) & \text{if } j = 1, \ldots, N/4 \\ 1/2NB(m_1, p) + 1/2NB(m_2, p) & \text{if } j = N/4+1, \ldots, N/2 \\ 1/3NB(m_1, p) + 1/3NB(m_2, p) + 1/3NB(m_3, p) & \text{if } j = N/2+1, \ldots, 3N/4 \\ NB(m_1, p) & \text{if } j = 3N/4+1, \ldots, N \end{cases}$$

- **multiple DP**: for $i = 501, \ldots, 750$,

$$y_{ij} = \begin{cases} \alpha_1 NB(m_1, p) + \alpha_2 NB(m_2, p) & \text{if } j = 1, \ldots, N/4 \\ \alpha_2 NB(m_1, p) + \alpha_1 NB(m_2, p) & \text{if } j = N/4+1, \ldots, N/2 \\ \alpha_3 NB(m_1, p) + \alpha_4 NB(m_2, p) & \text{if } j = N/2+1, \ldots, 3N/4 \\ \alpha_4 NB(m_1, p) + \alpha_3 NB(m_2, p) & \text{if } j = 3N/4+1, \ldots, N \end{cases}$$

- **multiple DB**: for $i = 751, \ldots, 1000$,

$$y_{ij} = \begin{cases} NB(m_3, p) & \text{if } j = 1, \ldots, N/4 \\ 1/2NB(m_3 + \delta, p) + 1/2NB(m_3 + \delta, p) & \text{if } j = N/4+1, \ldots, N/2 \\ 1/3NB(m_3 - \gamma, p) + 1/3NB(m_3, p) + 1/3NB(m_3 + \gamma, p) & \text{if } j = N/2+1, \ldots, 3N/4 \\ 1/4NB(m_3 - 2\epsilon, p) + 1/4NB(m_3 - \epsilon, p) & \\ +1/4NB(m_3 + \epsilon, p) + 1/4NB(m_3 + 2\epsilon, p) & \text{if } j = 3N/4+1, \ldots, N \end{cases}$$

- **multiple EE**: for $i = 1001, \ldots, 5500$,

$$y_{ij} = NB(m_1, p) \quad \forall j = 1, \ldots, N$$

- **mutiple EB**: for $i = 5501, \ldots, 10000$,

$$y_{ij} = \alpha_0 NB(m_1, p) + (1 - \alpha_0) NB(m_2, p) \quad \forall j = 1, \ldots, N$$

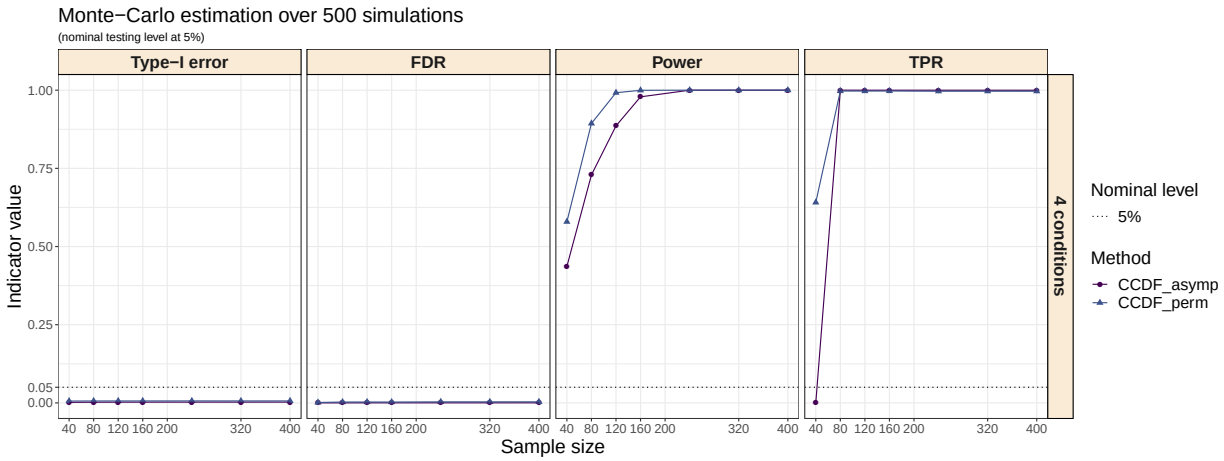


Figure 4-10: **Overall Type-I error, Power, FDR and TDR under the 4 conditions case with increasing sample size**.

Two conditions comparison given a covariate Z

We simulated a confounding variable Z from a Normal distribution $N(10, 2)$. The variable to be tested X was simulated in the following way:

$$X = \begin{cases} 1, & Z \leq Q_1 \quad \text{and} \quad Q_2 \leq Z \leq Q_3 \\ 2, & \text{otherwise} \end{cases}$$

where Q_p is the p^{th} quartile of Z.

We generated Y as a normally distributed variable:

$$Y = \begin{cases} AX + \epsilon_1, & \text{DE gene} \\ BZ + \epsilon_2, & \text{non-DE gene} \end{cases}$$

where $\mu \sim N(1, 0.5)$, $A \sim N(5 + \mu \mathbb{1}_{\{X=1\}}, 1)$, $B \sim U(0.3, 0.5)$, $\epsilon_1 \sim N(0, 1)$ and $\epsilon_2 \sim N(0, 0.5)$.

4.7.6 Comparisons using real data benchmarks

Positive control dataset

Single-cell RNA-seq data from Islam et al. [2011] are publicly available on GEO with the primary accession code GSE29087 for the positive control dataset. The top 1000 DE genes validated through qRT-PCR experiments and used as gold standard gene set Moliner et al. [2008] is available from the GitHub repository (https://github.com/Mgauth/ccdf). The matrix of raw counts contains 2,928 genes measured across 48 mouse embryonic stem cells and 44 mouse embryonic fibroblasts.

Negative control dataset

Single-cell RNA-seq data from Grün et al. [2014] are publicly available on GEO with the primary accession code GSE29087 for the negative control dataset. The matrix of raw counts includes 12,535 genes measured across 160 mouse embryonic stem cells.

`ccdf` settings

As showed in the main document, the asymptotic test is powerful enough from a sample size of 80 observations. Furthermore, regarding the great number of genes, it is more appropriate to use the asymptotic test instead of the permutations due to the difference of computation times. The number of thresholds is equal to half the number of unique values for each gene.

4.7.7 Processing all types of data

`ccdf` is designed in the first place to perform single-cell DEA but the flexibility and the absence of distributional assumption on the input data allow to apply `ccdf` to any kind of variables as long as they remain continuous.

Single-cell

The user must pay attention to two issues when dealing with single-cell measurements: the normalization and the dropouts.

Normalization

`ccdf` does not contain a prior normalization step. The user must normalize the raw counts beforehand to make the samples comparable. Since there is no consensus on which normalization is most appropriate, the choice is left to the user and we advise to get acquainted with Lytal et al. [2020] for a comparative review.

Dropouts

Single-cell data exhibit a huge amount of zeroes values. Even though `ccdf` is tailored to the presence or absence of zero inflation, if there is a large majority of zeroes (say more than 98%), one can question the accuracy of the results. We recommend to filter the genes by the number of dropouts that seems acceptable to the user and remove the selected ones from the analysis.

Other types of data

Any kind of quantitative variables can be considered as an input of `ccdf` since no distributional assumption is made on the data. We propose the following example to show the versatility of `ccdf`. The user may not necessarily need to perform multiple tests, as in genomics data.

The Boston Housing Dataset was originally published by Harrison and Rubinfeld Harrison Jr and Rubinfeld [1978]. Each observation corresponds to one of the 506 neighborhoods near Boston along with 13 variables (see details here [4]). We want to test, say, the independence between per capita crime rate by town CRIM and median value of owner-occupied homes in $1000's MEDV, given lower status of the population LSTAT. Given the large sample size, we can use the asymptotic test (thanks to the corresponding function) with as many thresholds as unique values, since we perform a single test, and without logarithmic scale. If the sample size had been very low, a sub-function for the permutation test is also available.

4.7.8 Application to a scRNA-seq study in COVID-19 patients

[4] `http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`

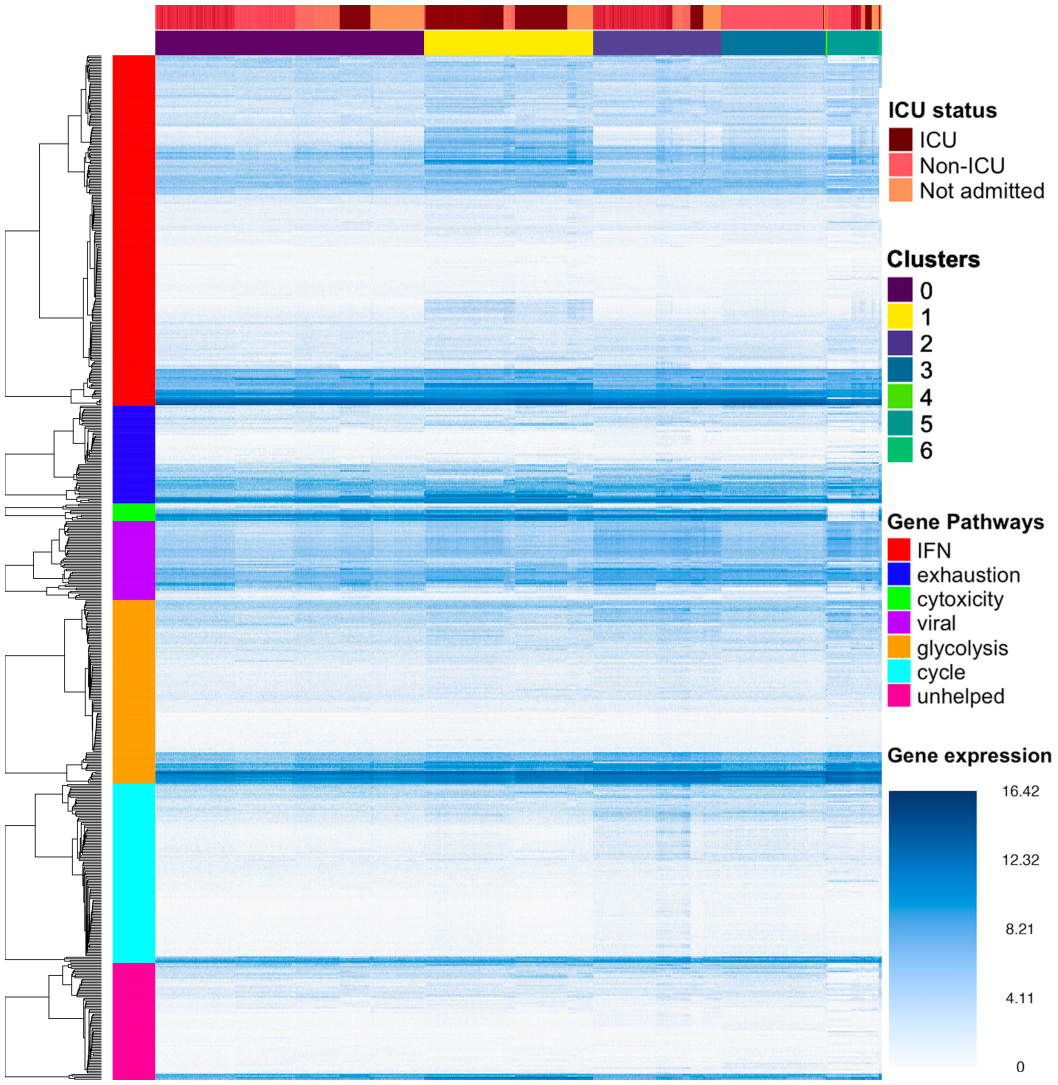


Figure 4-11: Heatmap of the log-CPM in the 7 gene pathways according to ICU status and clusters of cells.

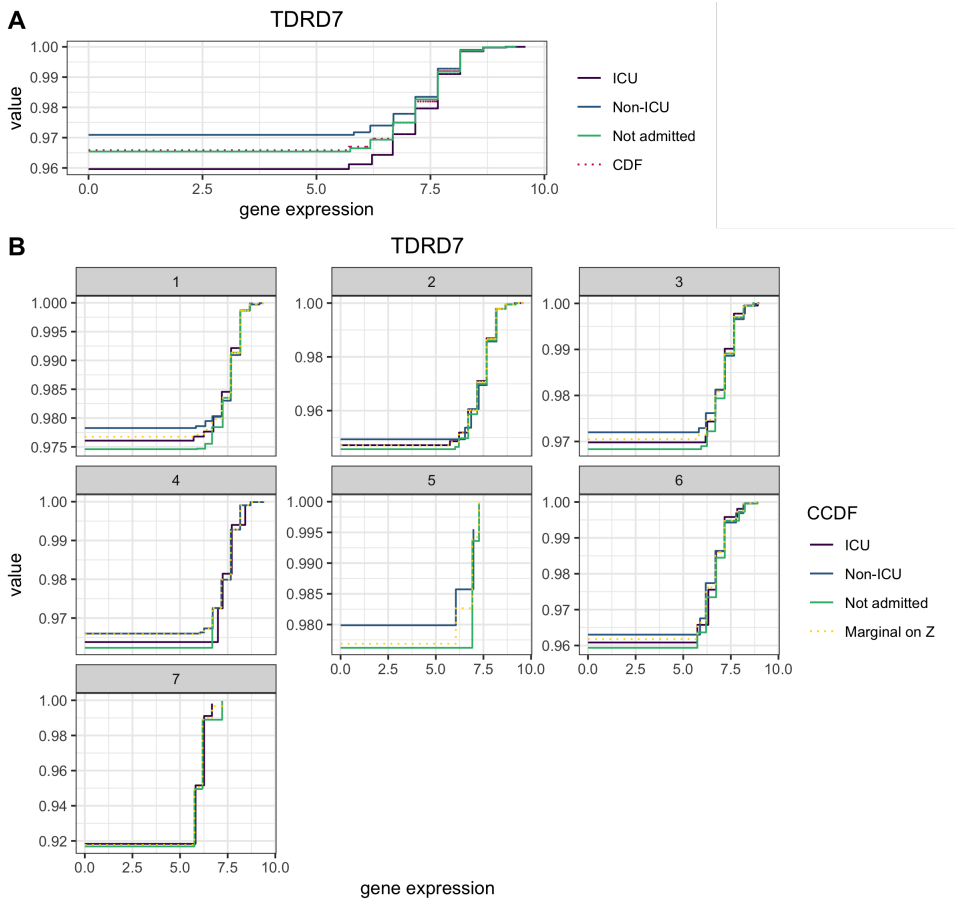


Figure 4-12: A: The solid lines represent the conditional CDF of TDRD7 gene on ICU status (ICU, Non-ICU and Not admitted) and the dark pink dotted line represents the marginal CDF of TDRD7 gene, *i.e.* without conditioning on ICU status. The underlying test performed by `ccdf` in the first DEA consists in comparing the marginal CDF with the conditional CDF. The p-value equals to 1.7e-06 not adjusting for the clusters. B: The solid lines represent the conditional CDF of TDRD7 gene on both Z, the 7 clusters, and X, the severity status (ICU, Non-ICU and Not Admitted), while the dotted yellow line represents the marginal CDF of TDRD7 expression without conditioning on X (but only conditioning on Z the clusters). The underlying test performed by `ccdf` now consists in comparing the marginal CDF on Z with the conditional CDF on both X and Z. The p-value equals to 0.4 when adjusting for the clusters. The number of steps of the CDF matches the thresholds chosen in `ccdf` (10 in the analysis). The first value of each CDF is the proportion of zeroes.

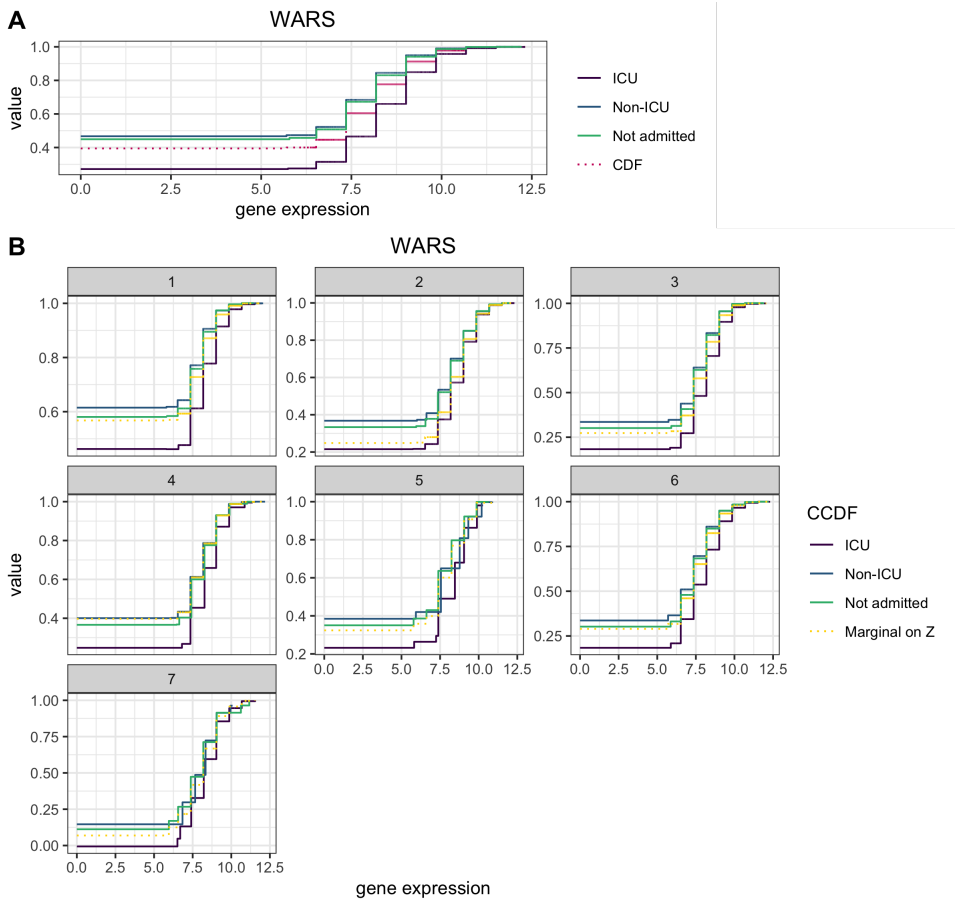


Figure 4-13: A: The solid lines represent the conditional CDF of WARS gene on ICU status (ICU, Non-ICU and Not admitted) and the dark pink dotted line represents the marginal CDF of WARS gene, *i.e.* without conditioning on ICU status. The underlying test performed by `ccdf` in the first DEA consists in comparing the marginal CDF with the conditional CDF. The p-value equals to 0 not adjusting for the clusters. B: The solid lines represent the conditional CDF of WARS gene on both Z, the 7 clusters, and X, the severity status (ICU, Non-ICU and Not Admitted), while the dotted yellow line represents the marginal CDF of WARS expression without conditioning on X (but only conditioning on Z the clusters). The underlying test performed by `ccdf` now consists in comparing the marginal CDF on Z with the conditional CDF on both X and Z. The p-value equals to 0.1e-201 when adjusting for the clusters. The number of steps of the CDF matches the thresholds chosen in `ccdf` (10 in the analysis). The first value of each CDF is the proportion of zeroes.

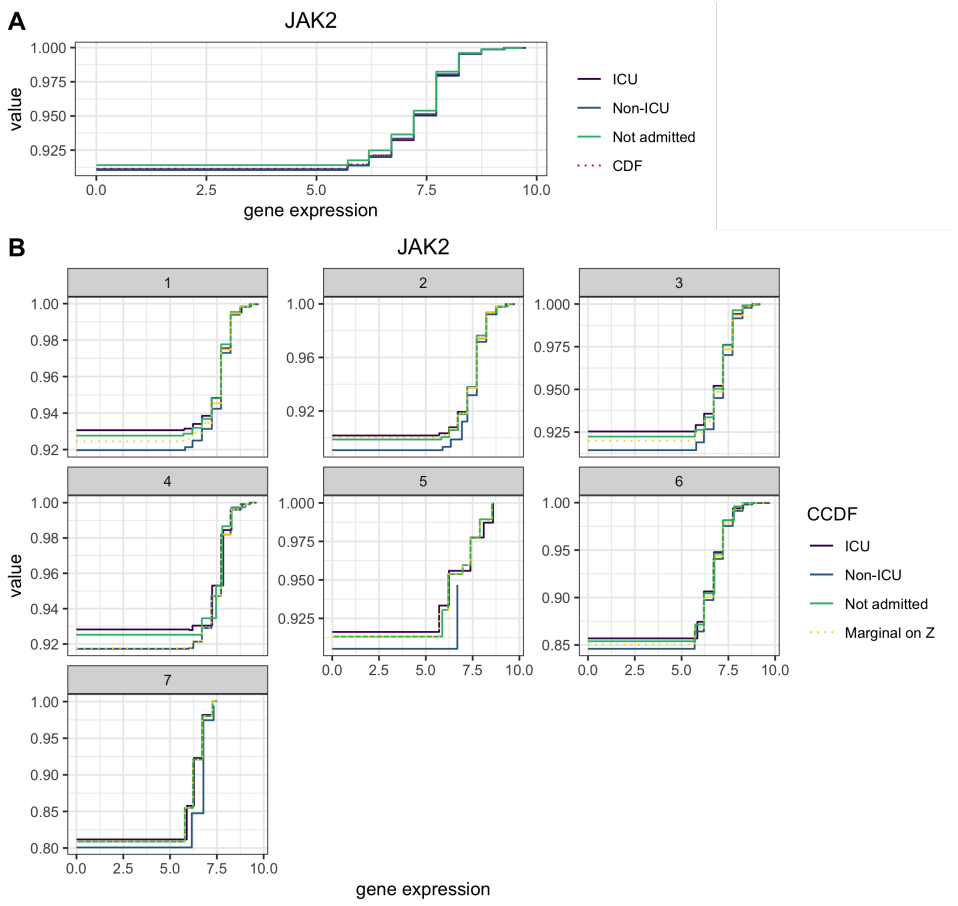


Figure 4-14: A: The solid lines represent the conditional CDF of JAK2 gene on ICU status (ICU, Non-ICU and Not admitted) and the dark pink dotted line represents the marginal CDF of JAK2 gene, *i.e.* without conditioning on ICU status. The underlying test performed by `ccdf` in the first DEA consists in comparing the marginal CDF with the conditional CDF. The p-value equals to 0.2 not adjusting for the clusters. B: The solid lines represent the conditional CDF of JAK2 gene on both Z, the 7 clusters, and X, the severity status (ICU, Non-ICU and Not Admitted), while the dotted yellow line represents the marginal CDF of JAK2 expression without conditioning on X (but only conditioning on Z the clusters). The underlying test performed by `ccdf` now consists in comparing the marginal CDF on Z with the conditional CDF on both X and Z. The p-value equals to 0.02 when adjusting for the clusters. The number of steps of the CDF matches the thresholds chosen in `ccdf` (10 in the analysis). The first value of each CDF is the proportion of zeroes.

Chapitre 5

Conclusion et perspectives

Dans cette thèse, nous avons proposé une première méthode d'analyse différentielle pour données RNA-seq en masse qui puise son orignalité dans l'utilisation d'un modèle linéaire à effets mixtes et un test en composante de variance. Que le schéma expérimental comprenne des données groupées ou non, longitudinales ou transversales, des variables d'intérêt et/ou covariables discrètes ou continues, notre approche reste adaptée tout en assurant un contrôle satisfaisant du FDR. De plus, l'utilisation d'un test asymptotique préférentiellement en présence d'échantillons de grande taille ou d'un test par permutations lorsque le nombre de mesures est faible rend notre approche particulièrement flexible. Nous gérons l'hétéroscédasticité intrinsèque aux données en s'inspirant des travaux de Law et al. [2014]. En se focalisant sur le FDR et sur la violation des hypothèses distributionnelles, nous avons mis en évidence les faiblesses des méthodes les plus utilisées. Par des simulations les plus réalistes possibles et l'application sur un jeu de données réelles portant sur la tuberculose, nous exposons les potentiels faux positifs que génèrent les méthodes les plus utilisées comme `edgeR`, `DESeq2` et `limma-voom` tandis que nous illustrons les performances significatives de notre méthode appelée `dearseq`, ne nécessitant pas d'hypothèse de distribution sur les données. `dearseq` a été entièrement implémentée dans un paquet R disponible sur Bioconductor.

Dans le prolongement des avancées technologiques sur le séquençage de l'ADN, nos recherches ont de ce fait épousé le mouvement du RNA-seq en masse vers le

RNA-seq en cellule unique. La seconde méthode a donc été conçue pour l'analyse de ces nouvelles données. Il nous paraissait primordial d'assurer la flexibilité du modèle ainsi que sa robustesse à toute hypothèse distributionnelle, tout en exploitant le nombre important de mesures grâce à l'échelle d'observation cellulaire. L'absence de consensus sur la meilleure distribution à adopter et sur le traitement des valeurs nulles renforcent la nécessité de se passer d'hypothèse forte sur ces aspects. C'est naturellement que nous utilisons les outils des statistiques non-paramétriques, spécifiquement les fonctions de répartition conditionnelles. Nous proposons une estimation originale de ces dernières via de multiple régressions linéaires ce qui permet l'inclusion de plusieurs variables à tester et de covariables, continues et/ou discrètes, et un temps de calcul réduit. Puis, nous généralisons le problème de l'analyse différentielle en se ramenant à un test d'indépendance conditionnelle. Nous axons notre étude de simulations sur la différence en distribution et nous l'élargissons à la comparaison à plus de deux conditions. Notre application sur un jeu de données réelles récent portant sur le COVID-19 donne un aperçu des possibilités offertes par notre méthode `ccdf`, et met en lumière l'importance de considérer de potentiels facteurs de confusion dans les analyses différentielles. Leur prise en compte peut également permettre l'exploration de nouvelles problématiques biologiques qui n'auraient pas été considérées sans un outil approprié.

Dans la méthode `dearseq`, la question de la précision numérique dans le calcul des p-valeurs se pose encore, et cela vaut aussi pour la méthode `ccdf`. En effet, la distribution nulle de ces deux tests asymptotiques est une somme pondérée de variables de χ^2 qui nécessite d'être estimée. Plus cette estimation sera précise, plus la p-valeur associée le sera aussi. La précision numérique des p-valeurs est d'autant plus importante lorsqu'elles prennent des faibles valeurs. L'application d'une correction pour la multiplicité des tests comme celles de Benjamini-Hocheberg [Benjamini and Hochberg, 1995] renforce le besoin d'une estimation la plus fine possible. Lors d'un test d'hypothèse, c'est alors essentiellement les petites p-valeurs, plutôt proches du seuil de significativité choisie, qui requièrent une attention particulière. Une précision grossière peut altérer la p-valeur et la rendre non-significative à quelques centièmes

près. De nombreuses méthodes sont proposées pour évaluer la probabilité de la queue de distribution à droite d'un mélange de χ^2 : les méthodes dites exactes [Davies, 1980; Farebrother, 1984], les méthodes des moments [Liu et al., 2009] et une approximation du point de selle [Kuonen, 1999]. Comme décrit dans Chen and Lumley [2019], certaines méthodes offrent une précision numérique plus ou moins satisfaisante. Par exemple, lorsque nous sommes souvent confrontés à un grand nombre de termes ($n > 1000$), particulièrement en RNA-seq en cellule unique, et de petites p-valeurs ($< 10^{-4}$), la méthode des moments est la plus imprécise. D'après les recommandations de Chen and Lumley [2019], nous avons opté pour l'approximation du point selle [Kuonen, 1999] (l'implémentation de `dearseq` a d'abord utilisé l'approche de Davies [1980] comme expliqué dans le Chapitre 3 puis nous avons remplacé cette procédure).

La précision numérique est également un enjeu contraignant pour les p-valeurs issues d'un test par permutations, puisque celle-ci est dépendante du nombre de permutations choisies. Bien que nous ayons proposé des permutations adaptatives pour augmenter seulement la précision des p-valeurs les plus petites et donc les plus significatives, l'accélération des temps de calcul de `dearseq` et `ccdf` permettrait d'évaluer les p-valeurs avec davantage de permutations, le résultat de l'analyse différentielle pouvant en être grandement impacté. Afin de répondre à cette double problématique des temps de calcul et de la précision, nous avons parallélisé notre code sur l'étape des permutations (et non des gènes). Cependant, cela n'est pour le moment pas suffisant pour une adoption de cette technique par la communauté scientifique lorsque la taille d'échantillon et le nombre de permutations nécessaires sont conséquents. A titre d'exemple, les résultats présentés dans ce travail ont nécessité l'équivalent de 2 578 136 heures de calcul sur le serveur du Mésocentre de Calcul Intensif Aquitain au cours de cette dernière année. Nous voyons alors deux manières d'accélérer les temps de calculs. La première est logicielle, il serait intéressant de considérer un langage de programmation de plus bas niveau et orienté vers le calcul haute performance, nous pensons ici à C et au plus récent Julia. La deuxième est matérielle, elle consiste à profiter de l'architecture des cartes graphiques (GPU). Un GPU est organisé en blocs, chaque bloc possède un certain nombre de coeurs. Dans cette configuration, il

est possible de paralléliser notre méthode à deux niveaux : chaque gène serait envoyé sur un bloc du GPU, puis dans chacun de ces blocs, nous pourrions paralléliser les permutations relatives au gène en question sur les différents coeurs du bloc. Pour donner un ordre d'idées, un CPU grand public possède aujourd'hui entre 16 et 32 coeurs, ce qui nous amène à effectuer 16 à 32 permutations en même temps pour un gène donné alors que les cartes graphiques grand public récentes possèdent plusieurs milliers de coeurs. Par exemple, la RTX 3090 de Nvidia possède plus de 10 000 coeurs répartis sur 80 blocs. Sur un tel GPU il serait donc possible de calculer la statistique de test pour 80 gènes simultanément tout en effectuant plus de 100 permutations en même temps pour chacun d'eux.

Les perspectives d'évolution de `ccdf` sont nombreuses. D'abord, il serait judicieux de proposer des arguments théoriques quant au choix du nombre optimal de seuils. Une idée serait de se placer dans le cas des estimateurs à histogramme régulier où les seuils seraient pris selon un pas régulier qui dépendrait seulement du nombre n d'observations. Bosq and Lecoutre [1987] étudient la consistance des estimateurs à histogrammes réguliers dans le cas de l'estimation de densités. Ils montrent sous certaines hypothèses de régularité de la densité que l'estimateur converge et donnent l'ordre de grandeur du pas optimal (pour un critère d'erreur quadratique) pour cet estimateur (exprimé seulement en fonction de n). Il serait alors question ici d'exprimer l'estimateur de la fonction de répartition conditionnelle de `ccdf` comme un estimateur par histogramme régulier et d'étudier le nombre optimal des seuils entièrement en fonction de n afin de minimiser un critère de convergence de type L^2. Cette perspective nous semble la plus appropriée puisqu'elle est nourrie d'une riche littérature, et permettrait de se passer d'une étape d'optimisation computationnellement très coûteuse nécessaire aux approches alternatives. De plus, nous pourrions réflechir à l'utilisation de seuils à pas non-réguliers afin de les répartir voire de les concentrer là où l'information est la plus pertinente. Cela permettrait de s'adapter à la distribution des gènes, par exemple, en basant les seuils sur les quantiles.

Motivés par les travaux de Tiberi et al. [2020] et le schéma expérimental du jeu de données réelles RNA-seq en cellule unique utilisé dans le Chapitre 4 portant sur

les différents stades de gravité du COVID-19 [Kusnadi et al., 2021], une extension de nos recherche sur les fonctions de répartition conditionnelles serait l'analyse multi-échantillons (*multi-sample*). En effet, les cellules sont annotées de façon à être associées aux patients. Cependant, nous sommes obligés d'omettre volontairement cette information puisque la méthode `ccdf` n'est, pour le moment, pas en mesure de gérer ce type de schéma expérimental. Une première idée inspirée directement de nos travaux du Chapitre 3 serait d'utiliser un modèle linéaire à effets mixtes à la place d'une régression linéaire. En ajoutant un effet aléatoire sur l'individu, on prendrait alors en compte la structure de corrélation existante pour un même sujet. Cependant, il est fort probable que nous devions modifier le test initial et que nous soyons confrontés à des problèmes de temps de calcul. Une autre difficulté réside dans le faible nombre d'individus inclus dans la majorité des études cliniques utilisant le RNA-seq en cellule unique. L'hypothèse de normalité portant sur l'effet aléatoire peut ne pas être vérifiée dans ces conditions. Une idée plus simple consisterait à conserver le modèle linéaire tout en réécrivant la distribution de β_{1j} avec une nouvelle structure de variance-covariance lorsque les observations sont groupées.

Les analyses gène par gène ont toutefois leurs limites. Typiquement, ce sont plus de 10 000 gènes qui composent la matrice des données. Il arrive qu'un grand nombre de gènes soit trouvés comme différentiellement exprimés rendant difficile l'interprétation biologique. Par ailleurs, on sait que certains gènes sont susceptibles d'être co-régulés et d'être liés fonctionnellement entre eux. On peut alors les aggréger. On parle alors de groupes de gènes ou d'ensembles de gènes (*gene sets*). Des groupes de gènes associés à des voies biologiques sont par exemple prédéfinis par KEGG [Kanehisa and Goto, 2000], Gene Ontology [Ashburner et al., 2000] ou les modules fonctionnels de Chaussabel et al. [2008] et permettent par la suite d'effectuer des analyses différentielles dans le but d'identifier non pas des gènes individuels mais des groupes de gènes différentiellement exprimés. L'analyse par groupe de gènes est censée être plus puissante puisqu'elle exploite le signal plus fort du groupe de gènes comparativement à celui d'un gène isolé. De plus, il est possible qu'un changement d'expression dans le groupe entier soit biologiquement davantage significatif que le changement d'expression d'un

seul et même gène. La seule et unique méthode d'analyse pour groupe de gènes pour données RNA-seq en cellule unique que nous ayons trouvée dans la littérature est la méthode `iDEA` [Ma et al., 2020] qui se présente comme une approche par enrichissement. Néanmoins, plusieurs limites apparaissent comme des hypothèses distributionnelles sur les nombreux paramètres du modèle et des estimations qui semblent être computationnellement intensives. Enfin, cette approche est en réalité constituée de deux étapes puisqu'elle nécessite l'obtention préalable d'un ensemble de paramètres estimés par une méthode d'analyse différentielle en gène par gène comme `MAST` [Finak et al., 2015] afin de calculer la statistique de test pour le groupe de gène. Elle aura de ce fait les défauts de l'approche choisie. Il semble donc y avoir un vide méthodologique s'agissant de l'analyse différentielle par groupe de gènes pour données RNA-seq en cellule unique. Dans notre cadre de travail des fonctions de répartition conditionnelles, nous pourrions imaginer vouloir calculer les fonctions de répartition des gènes constituant le groupe pour ensuite les agréger, par exemple par une fonction de répartition moyenne. Cela se traduirait dans notre test par l'agrégation des paramètres d'intérêt β_{1j}. Il s'agit enfin de s'assurer de la forme de la statistique de test ainsi que de la distribution asymptotique.

Travailler en collaboration avec des immunologistes, des biostatisticiens, des mathématiciens ou encore des informaticiens amène à une certaine souplesse des idées. C'est naturellement que les méthodes présentées dans cette thèse ont la particularité d'être flexibles et adaptées à des situations très différentes. Il semble aujourd'hui essentiel de mettre cette capacité d'adaptation au centre des développements scientifiques.

Bibliographie

Agniel, D. and Hejblum, B. P. [2017], 'Variance component score test for time-course gene set analysis of longitudinal RNA-seq data', *Biostatistics* **18**(4), 589–604.

Agniel, D., Xie, W., Essex, M. and Cai, T. [2018], 'Functional principal variance component testing for a genetic association study of HIV progression', *Annals of Applied Statistics* .

Anders, S. and Huber, W. [2010], 'Differential expression analysis for sequence count data', *Nature Precedings* pp. 1–1.

Anderson, T. W. [1962], 'On the distribution of the two-sample cramer-von mises criterion', *The Annals of Mathematical Statistics* pp. 1148–1159.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. [2000], 'Gene ontology : tool for the unification of biology', *Nature genetics* **25**(1), 25–29.

Assefa, A. T., De Paepe, K., Everaert, C., Mestdagh, P., Thas, O. and Vandesompele, J. [2018], 'Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data', *Genome Biology* **19**(1), 96.

Bacher, P., Rosati, E., Esser, D., Martini, G. R., Saggau, C., Schiminsky, E., Dargvainiene, J., Schröder, I., Wieters, I., Khodamoradi, Y. et al. [2020], 'Low-avidity cd4+ t cell responses to sars-cov-2 in unexposed individuals and humans with severe covid-19', *Immunity* **53**(6), 1258–1271.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M. and Kendziorski, C. [2017], 'Scnorm : robust normalization of single-cell rna-seq data', *Nature methods* **14**(6), 584.

Benjamini, Y. and Hochberg, Y. [1995], 'Controlling the false discovery rate : a practical and powerful approach to multiple testing', *Journal of the royal statistical society. Series B (Methodological)* pp. 289–300.

Berry, M. P., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A., Oni, T., Wilkinson, K. A., Banchereau, R., Skinner, J., Wilkinson, R. J. et al. [2010], 'An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis', *Nature* **466**(7309), 973.

Blum, J. R., Kiefer, J. and Rosenblatt, M. [1961], 'Distribution free tests of independence based on the sample distribution function', *The annals of mathematical statistics* **32**(2), 485–498.

Bolstad, B. M., Irizarry, R. A., Åstrand, M. and Speed, T. P. [2003], 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics* **19**(2), 185–193.

Bosq, D. and Lecoutre, J.-P. [1987], 'Théorie de l'estimation fonctionnelle. collection economie et statistiques avancées', *Economica, Paris* .

Bost, P., Giladi, A., Liu, Y., Bendjelal, Y., Xu, G., David, E., Blecher-Gonen, R., Cohen, M., Medaglia, C., Li, H. et al. [2020], 'Host-viral infection maps reveal signatures of severe covid-19 patients', *Cell* **181**(7), 1475–1488.

Bouezmarni, T., Rombouts, J. V. and Taamouti, A. [2012], 'Nonparametric copula-based test for conditional independence with applications to granger causality', *Journal of Business & Economic Statistics* **30**(2), 275–287.

Bradley, T., Ferrari, G., Haynes, B. F., Margolis, D. M. and Browne, E. P. [2018], 'Single-cell analysis of quiescent hiv infection reveals host transcriptional profiles that regulate proviral latency', *Cell reports* **25**(1), 107–117.

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. [2016], 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology* **34**(5), 525–527.

Breda, J., Zavolan, M. and van Nimwegen, E. [2021], 'Bayesian inference of gene expression states from single-cell rna-seq data', *Nature Biotechnology* pp. 1–9.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C. et al. [2013], 'Accounting for technical noise in single-cell rna-seq experiments', *Nature methods* **10**(11), 1093–1095.

Brier, G. W. [1950], 'Verification of forecasts expressed in terms of probability', *Monthey Weather Review* **78**(1), 1–3.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. and Stegle, O. [2015], 'Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells', *Nature biotechnology* **33**(2), 155–160.

Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. [2010], 'Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments', *BMC bioinformatics* **11**(1), 1–13.

Burden, C. J., Qureshi, S. E. and Wilson, S. R. [2014], 'Error estimates for the analysis of differential expression from rna-seq count data', *PeerJ* **2**, e576.

Bustin, S., Dhillon, H. S., Kirvell, S., Greenwood, C., Parker, M., Shipley, G. L. and Nolan, T. [2015], 'Variability of the reverse transcription step : practical implications', *Clinical chemistry* **61**(1), 202–212.

C. Elegans Sequencing Consortium [1998], 'Genome sequence of the nematode C. elegans : a platform for investigating biology', *Science* **282**(5396), 2012–2018.

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I. et al. [2008], 'A modular analysis framework for blood genomics studies : application to systemic lupus erythematosus', *Immunity* **29**(1), 150–164.

Cheerla, A. and Gevaert, O. [2019], 'Deep learning with multimodal representation for pancancer prognosis prediction', *Bioinformatics* **35**(14), i446–i454.

Chen, R. and Snyder, M. [2013], 'Promise of personalized omics to precision medicine', *Wiley Interdisciplinary Reviews : Systems Biology and Medicine* **5**(1), 73–82.

Chen, T. and Lumley, T. [2019], 'Numerical evaluation of methods approximating the distribution of a large quadratic form in normal variables', *Computational Statistics & Data Analysis* **139**, 75–81.

Ching, T., Zhu, X. and Garmire, L. X. [2018], 'Cox-nnet : an artificial neural network method for prognosis prediction of high-throughput omics data', *PLoS computational biology* **14**(4).

Choi, K., Chen, Y. and Skelly, D. [2020], 'Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics', *Genome Biology* **21**(183).

Cliff, J. M., Andrade, I. N., Mistry, R., Clayton, C. L., Lennon, M. G., Lewis, A. P., Duncan, K., Lukey, P. T. and Dockrell, H. M. [2004], 'Differential gene expression identifies novel markers of cd4+ and cd8+ t cell activation following stimulation by mycobacterium tuberculosis', *The Journal of Immunology* **173**(1), 485–493.

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S. and Yosef, N. [2019], 'Performance assessment and selection of normalization procedures for single-cell rna-seq', *Cell systems* **8**(4), 315–328.

Commenges, D. and Andersen, P. K. [1995], 'Score test of homogeneity for survival data', *Lifetime data analysis* **1**(2), 145–156.

Commenges, D. and Jacqmin-Gadda, H. [2015], *Modèles biostatistiques pour l'épidémiologie*, De Boeck Superieur.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. and Mortazavi, A. [2016], 'A survey of best practices for RNA-seq data analysis', *Genome Biol* **17**(1), 13–13.

Costa-Silva, J., Domingues, D. and Lopes, F. M. [2017], 'RNA-Seq differential expression analysis : An extended review and a software tool', *PLOS ONE* **12**(12).

Costa, V., Aprile, M., Esposito, R. and Ciccodicola, A. [2013], 'Rna-seq and human complex diseases : recent accomplishments and future perspectives', *European Journal of Human Genetics* **21**(2), 134–142.

Cotugno, N., Ruggiero, A., Santilli, V., Manno, E. C., Rocca, S., Zicari, S., Amodio, D., Colucci, M., Rossi, P., Levy, O. et al. [2019], 'Omic technologies and vaccine development : from the identification of vulnerable individuals to the formulation of invulnerable vaccines', *Journal of immunology research* **2019**.

Crick, F. [1970], 'Central dogma of molecular biology', *Nature* **227**(5258), 561–563.

Crick, F. H. [1958], On protein synthesis, *in* 'Symp Soc Exp Biol', Vol. 12, p. 8.

Davies, R. B. [1980], 'The distribution of a linear combination of $\chi 2$ random variables', *Journal of the Royal Statistical Society : Series C (Applied Statistics)* **29**(3), 323–333.

De Lecea, M. G. M. and Rossbach, M. [2012], 'Translational genomics in personalized medicine–scientific challenges en route to clinical practice', *The HUGO Journal* **6**(1), 1–9.

DeLaughter, D. M. [2018], 'The use of the Fluidigm C1 for RNA expression analyses of single cells', *Current protocols in molecular biology* **122**(1), e55.

DeLisi, C. [1988], 'The human genome project : the ambitious proposal to map and decipher the complete sequence of human dna', *American Scientist* **76**(5), 488–493.

Delmans, M. and Hemberg, M. [2016], 'Discrete distributional differential expression (d 3 e)-a tool for gene expression analysis of single-cell rna-seq data', *BMC bioinformatics* **17**(1), 110.

Demey, J. [2020], 'L'immunologiste Steve Pascolo : "L'ARN messager peut en théorie résoudre toutes les maladies', *Le Journal du Dimanche* . [En ligne ; accès le 11/10/2021].
URL: *https ://www.lejdd.fr/Societe/Sante/limmunologiste-steve-pascolo-larn-messager-peut-en-theorie-resoudre-toutes-les-maladies-4014813*

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J. et al. [2013], 'A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis', *Briefings in bioinformatics* **14**(6), 671–683.

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A. and Wang, W. [2015], 'Normalization and noise reduction for single cell RNA-seq experiments', *Bioinformatics* **31**(13), 2225–2227.

Doran, G., Muandet, K., Zhang, K. and Schölkopf, B. [2014], A permutation-based kernel conditional independence test, *in* 'Uncertainty In Artificial Intelligence : Proceedings of the Thirtieth Conference', UAI'14, AUAI Press, Arlington, Virginia, USA, p. 132–141.

Dorr, C., Wu, B., Guan, W., Muthusamy, A., Sanghavi, K., Schladt, D. P., Maltzman, J. S., Scherer, S. E., Brott, M. J., Matas, A. J. et al. [2015], 'Differentially expressed gene transcripts using rna sequencing from the blood of immunosuppressed kidney allograft recipients', *PloS one* **10**(5), e0125045.

Dudoit, S., Van Der Laan, M. J. and van der Laan, M. J. [2008], *Multiple testing procedures with applications to genomics*, Springer.

Eberwine, J., Sul, J.-Y., Bartfai, T. and Kim, J. [2014], 'The promise of single-cell sequencing', *Nature methods* **11**(1), 25–27.

Evans, C., Hardin, J. and Stoebel, D. M. [2018], 'Selecting between-sample rna-seq normalization methods from the perspective of their assumptions', *Briefings in bioinformatics* **19**(5), 776–792.

Farebrother, R. [1984], 'Algorithm as 204 : the distribution of a positive linear combination of χ 2 random variables', *Applied Statistics* pp. 332–339.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M. et al. [2015], 'Mast : a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data', *Genome biology* **16**(1), 1–13.

Frese, K. S., Katus, H. A. and Meder, B. [2013], 'Next-generation sequencing : from understanding biology to personalized medicine', *Biology* **2**(1), 378–398.

Furman, D., Hejblum, B. P., Simon, N., Jojic, V., Dekker, C. L., Thiébaut, R., Tibshirani, R. J. and Davis, M. M. [2014], 'Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination', *Proceedings of the National Academy of Sciences* **111**(2), 869–874.

Furman, D., Jojic, V., Kidd, B., Shen-Orr, S., Price, J., Jarrell, J., Tse, T., Huang, H., Lund, P., Maecker, H. T. et al. [2013], 'Apoptosis and other immune biomarkers predict influenza vaccine responsiveness', *Molecular systems biology* **9**(1), 659.

Geall, A. J., Mandl, C. W. and Ulmer, J. B. [2013], Rna : the new revolution in nucleic acid vaccines, *in* 'Seminars in immunology', Vol. 25, Elsevier, pp. 152–159.

Germain, P. L., Vitriolo, A., Adamo, A., Laise, P., Das, V. and Testa, G. [2016], 'RNAontheBENCH : Computational and empirical resources for benchmarking RNAseq quantification and differential expression methods', *Nucleic Acids Research* **44**(11), 5054–5067.

Goeman, J. J., van de Geer, S. A. and van Houwelingen, H. C. [2006], 'Testing against a high dimensional alternative', *Journal of the Royal Statistical Society Series B-Statistical Methodology* **68**, 477–493.

Golumbeanu, M., Cristinelli, S., Rato, S., Munoz, M., Cavassini, M., Beerenwinkel, N. and Ciuffi, A. [2018], 'Single-cell rna-seq reveals transcriptional heterogeneity in latent and reactivated hiv-infected cells', *Cell reports* **23**(4), 942–950.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. [2012], 'A kernel two-sample test', *The Journal of Machine Learning Research* **13**(1), 723–773.

Grün, D., Kester, L. and Van Oudenaarden, A. [2014], 'Validation of noise models for single-cell transcriptomics', *Nature methods* **11**(6), 637–640.

Harrison Jr, D. and Rubinfeld, D. L. [1978], 'Hedonic housing prices and the demand for clean air', *Journal of environmental economics and management* **5**(1), 81–102.

Harville, D. A. [1977], 'Maximum likelihood approaches to variance component estimation and to related problems', *Journal of the American statistical association* **72**(358), 320–338.

Hejblum, B. P., Skinner, J. and Thiébaut, R. [2015], 'Time-Course Gene Set Analysis for Longitudinal Gene Expression Data', *PLOS Computational Biology* **11**(6), e1004310.

Hicks, S. C., Townes, F. W., Teng, M. and Irizarry, R. A. [2018], 'Missing data and technical variability in single-cell RNA-sequencing experiments', *Biostatistics* **19**(4), 562–578.

Hill, K. [2013], 'Blueprints of nsa's ridiculously expensive data center in utah suggest it holds less info than thought', https://www.forbes.com/sites/kashmirhill/2013/07/24/blueprints-of-nsa-data-center-in-utah-suggest-its-storage-capacity-is-less-impressive-than-thought/?sh=55953b357457. [En ligne ; accès le 11/10/2021].

Huang, M., Sun, Y. and White, H. [2016], 'A flexible nonparametric test for conditional independence', *Econometric Theory* **32**(6), 1434–1482.

Huang, T.-M. et al. [2010], 'Testing conditional independence using maximal nonlinear conditional correlation', *The Annals of Statistics* **38**(4), 2047–2091.

Huang, Y.-T. and Lin, X. [n.d.], 'Gene set analysis using variance component tests', **14**(1), 210–210.

Huang, Z., Johnson, T. S., Han, Z., Helm, B., Cao, S., Zhang, C., Salama, P., Rizkalla, M., Christina, Y. Y., Cheng, J. et al. [2020], 'Deep learning-based cancer survival prognosis from rna-seq data : approaches and evaluations', *BMC medical genomics* **13**(5), 1–12.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P. and Linnarsson, S. [2011], 'Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq', *Genome research* **21**(7), 1160–1167.

Jiang, L., Chen, H., Pinello, L. and Yuan, G.-C. [2016], 'Giniclust : detecting rare cell types from single-cell gene expression data with gini index', *Genome biology* **17**(1), 1–13.

Jiang, R., Sun, T., Song, D. and Li, J. J. [2020], 'Zeros in scrna-seq data : good or bad ? how to embrace or tackle zeros in scrna-seq data analysis ?', *bioRxiv* .

Kadota, K., Nishiyama, T. and Shimizu, K. [2012], 'A normalization strategy for comparing tag count data', *Algorithms for Molecular Biology* **7**(1), 1–13.

Kanehisa, M. and Goto, S. [2000], 'Kegg : kyoto encyclopedia of genes and genomes', *Nucleic acids research* **28**(1), 27–30.

Katayama, S., Töhönen, V., Linnarsson, S. and Kere, J. [2013], 'Samstrt : statistical test for differential expression in single-cell transcriptome with spike-in normalization', *Bioinformatics* **29**(22), 2943–2945.

Kharchenko, P. V., Silberstein, L. and Scadden, D. T. [2014], 'Bayesian approach to single-cell differential expression analysis', *Nature methods* **11**(7), 740–742.

Kiefer, J. [1959], 'K-sample analogues of the kolmogorov-smirnov and cramér-v. mises tests', *The Annals of Mathematical Statistics* pp. 420–447.

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R. and Kendziorski, C. [2016], 'A statistical approach for identifying differential distributions in single-cell rna-seq experiments', *Genome biology* **17**(1), 222.

Kuksin, M., Morel, D., Aglave, M., Danlos, F.-X., Marabelle, A., Zinovyev, A., Gautheret, D. and Verlingue, L. [2021], 'Applications of single-cell and bulk RNA sequencing in onco-immunology', *European Journal of Cancer* **149**, 193–210.

Kuonen, D. [1999], 'Saddlepoint approximations for distributions of quadratic forms in normal variables', *Biometrika* **86**(4), 929–935.

Kusnadi, A., Ramírez-Suástegui, C., Fajardo, V., Chee, S. J., Meckiff, B. J., Simon, H., Pelosi, E., Seumois, G., Ay, F., Vijayanand, P. et al. [2021], 'Severely ill covid-19 patients display impaired exhaustion features in SARS-CoV-2-reactive CD8+ T cells', *Science immunology* **6**(55).

Łabaj, P. P. and Kreil, D. P. [2016], 'Sensitivity, specificity, and reproducibility of RNA-Seq differential expression calls', *Biology Direct* **11**(1), 66.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A. et al. [2020], 'Eleven grand challenges in single-cell data science', *Genome biology* **21**(1), 1–35.

Laird, N. M. and Ware, J. H. [1982], 'Random-effects models for longitudinal data', *Biometrics* pp. 963–974.

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M. and Maza, E. [2018], 'Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size', *Frontiers in Plant Science* **9**.

Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. [2014], 'Voom : Precision weights unlock linear model analysis tools for RNA-seq read counts.', *Genome biology* **15**(2), R29–R29.

León-Novelo, L., Fuentes, C. and Emerson, S. [2017], 'Marginal likelihood estimation of negative binomial parameters with applications to RNA-seq data', *Biostatistics* **18**(4), 637–650.

Lévy, Y., Wiedemann, A., Hejblum, B. P., Durand, M., Lefebvre, C., Surénaud, M., Lacabaratz, C., Perreau, M., Foucat, E., Déchenaud, M. et al. [2021], 'CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death', *Iscience* **24**(7), 102711.

Li, C. and Fan, X. [2020], 'On nonparametric conditional independence tests for continuous variables', *Wiley Interdisciplinary Reviews : Computational Statistics* **12**(3), e1489.

Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. [2012], 'Normalization, testing, and false discovery rate estimation for RNA-sequencing data', *Biostatistics* **13**(3), 523–538.

Li, K.-C. and Duan, N. [1989], 'Regression analysis under link violation', *The Annals of Statistics* pp. 1009–1052.

Li, P., Piao, Y., Shon, H. S. and Ryu, K. H. [2015], 'Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data', *BMC bioinformatics* **16**(1), 1–9.

Li, Q., Maasoumi, E. and Racine, J. S. [2009], 'A nonparametric test for equality of distributions with mixed categorical and continuous data', *Journal of Econometrics* **148**(2), 186–200.

Li, S., Łabaj, P. P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, P.-Y., Wang, M., Wang, C., Thierry-Mieg, D., Thierry-Mieg, J., Kreil, D. P. and Mason, C. E. [2014], 'Detecting and correcting systematic variation in large-scale RNA sequencing data', *Nature Biotechnology* **32**(9), 888–895.

Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D. S., Johnson, S. E., Milton, A., Rajam, G. et al. [2014], 'Molecular signatures of antibody responses derived from a systems biology study of five human vaccines', *Nature immunology* **15**(2), 195–204.

Lin, X. [1997], 'Variance component testing in generalised linear models with random effects', *Biometrika* **84**(2), 309–326.

Lin, Y., Golovnina, K., Chen, Z.-X., Lee, H. N., Negron, Y. L. S., Sultana, H., Oliver, B. and Harbison, S. T. [2016], 'Comparison of normalization and differential expression analyses using rna-seq data from 726 individual drosophila melanogaster', *BMC genomics* **17**(1), 1–20.

Linton, O., Gozalo, P. et al. [1996], 'Conditional independence restrictions : Testing and estimation', *Cowles Foundation Discussion Paper* .

Liu, H., Tang, Y. and Zhang, H. H. [2009], 'A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables', *Computational Statistics & Data Analysis* **53**(4), 853–856.

Love, M. I., Huber, W. and Anders, S. [2014], 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome biology* **15**(12), 1–21.

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., Levens, D. L., Lee, T. I. and Young, R. A. [2012], 'Revisiting global gene expression analysis', *Cell* **151**(3), 476–482.

Lumley, T. [2004], 'Analysis of complex survey samples', *Journal of Statistical Software* **9**(1), 1–19. R package verson 2.2.

Lun, A. T., Bach, K. and Marioni, J. C. [2016], 'Pooling across cells to normalize single-cell RNA sequencing data with many zero counts', *Genome biology* **17**(1), 75.

Lytal, N., Ran, D. and An, L. [2020], 'Normalization methods on single-cell RNA-seq data : an empirical survey', *Frontiers in genetics* **11**, 41.

Ma, Y., Sun, S., Shang, X., Keller, E. T., Chen, M. and Zhou, X. [2020], 'Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies', *Nature communications* **11**(1), 1–13.

Martínez-Camblor, P. and de Uña-Álvarez, J. [2009], 'Non-parametric k-sample tests : Density functions vs distribution functions', *Computational Statistics & Data Analysis* **53**(9), 3344–3357.

Massey Jr, F. J. [1951], 'The Kolmogorov-Smirnov test for goodness of fit', *Journal of the American statistical Association* **46**(253), 68–78.

Maza, E., Frasse, P., Senin, P., Bouzayen, M. and Zouine, M. [2013], 'Comparison of normalization methods for differential gene expression analysis in RNA-seq experiments : a matter of relative size of studied transcriptomes', *Communicative & integrative biology* **6**(6), e25849.

Mazzoni, G., Kogelman, L. J. A., Suravajhala, P. and Kadarmideen, H. N. [2015], 'Systems Genetics of Complex Diseases Using RNA-Sequencing Methods', *International Journal of Bioscience, Biochemistry and Bioinformatics* **5**(4), 264–279.

McCarthy, D. J. and Smyth, G. K. [2009], 'Testing significance relative to a fold-change threshold is a treat', *Bioinformatics* **25**(6), 765–771.

McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J. and Nuzhdin, S. V. [2011], 'RNA-seq : technical variability and sampling', *BMC genomics* **12**(1), 1–13.

Medaglini, D., Santoro, F. and Siegrist, C.-A. [2018], 'Correlates of vaccine-induced protective immunity against ebola virus disease', **39**, 65–72.

Menicucci, A. R., Sureshchandra, S., Marzi, A., Feldmann, H. and Messaoudi, I. [2017], 'Transcriptomic analysis reveals a previously unknown role for CD8+ T-cells in rVSV-EBOV mediated protection', *Scientific reports* **7**(1), 1–12.

Miao, Z., Deng, K., Wang, X. and Zhang, X. [2018], 'DEsingle for detecting three types of differential expression in single-cell rna-seq data', *Bioinformatics* **34**(18), 3223–3224.

Moliner, A., Ernfors, P., Ibanez, C. F. and Andäng, M. [2008], 'Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials', *Stem cells and development* **17**(2), 233–243.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. [2008], 'Mapping and quantifying mammalian transcriptomes by RNA-seq', *Nature methods* **5**(7), 621–628.

Muandet, K., Fukumizu, K., Sriperumbudur, B. and Schölkopf, B. [2016], 'Kernel mean embedding of distributions : A review and beyond', *arXiv preprint arXiv :1605.09522* .

Muandet, K., Fukumizu, K., Sriperumbudur, B. and Schölkopf, B. [2017], 'Kernel mean embedding of distributions : A review and beyond', *Foundations and Trends® in Machine Learning* **10**(1-2), 1–141.

Mullis, K. B. [1990], 'The unusual origin of the polymerase chain reaction', *Scientific American* **262**(4), 56–65.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. [1986], Specific enzymatic amplification of dna in vitro : the polymerase chain reaction, *in* 'Cold Spring Harbor symposia on quantitative biology', Vol. 51, Cold Spring Harbor Laboratory Press, pp. 263–273.

Nabavi, S., Schmolze, D., Maitituoheti, M., Malladi, S. and Beck, A. H. [2016], 'EMDomics : a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes', *Bioinformatics* **32**(4), 533–541.

Nakaya, H. I., Wrammert, J., Lee, E. K., Racioppi, L., Marie-Kunze, S., Haining, W. N., Means, A. R., Kasturi, S. P., Khan, N., Li, G.-M. et al. [2011], 'Systems biology of vaccination for seasonal influenza in humans', *Nature immunology* **12**(8), 786–795.

Nawy, T. [2014], 'Single-cell sequencing', *Nature methods* **11**(1), 18–18.

Neu, K. E., Tang, Q., Wilson, P. C. and Khan, A. A. [2017], 'Single-cell genomics : approaches and utility in immunology', *Trends in immunology* **38**(2), 140–149.

Oshlack, A. and Wakefield, M. J. [2009], 'Transcript length bias in rna-seq data confounds systems biology', *Biology direct* **4**(1), 1–10.

Papalexi, E. and Satija, R. [2018], 'Single-cell RNA sequencing to explore immune cell heterogeneity', *Nature Reviews Immunology* **18**(1), 35–45.

Parzen, E. [1962], 'On estimation of a probability density function and mode', *The annals of mathematical statistics* **33**(3), 1065–1076.

Paszek, P. [2007], 'Modeling stochasticity in gene regulation : characterization in the terms of the underlying distribution function', *Bulletin of Mathematical Biology* **69**(5), 1567–1601.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L. et al. [2014], 'Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma', *Science* **344**(6190), 1396–1401.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C. [2017], 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods* **14**(4), 417–419.

Phipson, B. and Smyth, G. K. [2010], 'Permutation p-values should never be zero : calculating exact p-values when permutations are randomly drawn', *Statistical applications in genetics and molecular biology* **9**(1).

Proserpio, V. and Mahata, B. [2016], 'Single-cell technologies to study the immune system', *Immunology* **147**(2), 133–140.

Pruitt, K. D. and Maglott, D. R. [2001], 'RefSeq and LocusLink : NCBI gene-centered resources', *Nucleic acids research* **29**(1), 137–140.

Pulendran, B. and Ahmed, R. [2011], 'Immunological mechanisms of vaccination', *Nature immunology* **12**(6), 509–517.

Querec, T. D., Akondy, R. S., Lee, E. K., Cao, W., Nakaya, H. I., Teuwen, D., Pirani, A., Gernert, K., Deng, J., Marzolf, B., Kennedy, K., Wu, H., Bennouna, S., Oluoch, H., Miller, J., Vencio, R. Z., Mulligan, M., Aderem, A., Ahmed, R. and Pulendran, B. [2009], 'Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans', *Nature Immunology* **10**(1), 116–125.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. and Betel, D. [2013], 'Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data', *Genome Biology* **14**(9), R95–R95.

Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U., Lohse, A., Krähling, V., Eickmann, M., Becker, S., Agnandji, S. T., Krishna, S., Kremsner, P. G., Brosnahan, J. S., Bejon, P., Njuguna, P., Addo, M. M., Becker, S., Krähling, V., Siegrist, C.-A., Huttner, A., Kieny, M.-P., Moorthy, V., Fast, P., Savarese, B., Lapujade, O., Thiébaut, R., Altfeld, M. and Addo, M. [2017], 'Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV', *Cell Reports* **20**(9), 2251–2261.

Rigaill, G., Balzergue, S., Brunaud, V., Blondet, E., Rau, A., Rogier, O., Caius, J., Maugis-Rabusseau, C., Soubigou-Taconnat, L., Aubourg, S. et al. [2018], 'Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis', *Briefings in bioinformatics* **19**(1), 65–76.

Risso, D., Ngai, J., Speed, T. P. and Dudoit, S. [2014], 'Normalization of rna-seq data using factor analysis of control genes or samples', *Nature biotechnology* **32**(9), 896–902.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.-P. [2018], 'A general and flexible method for signal extraction from single-cell RNA-seq data', *Nature Communications* **9**(1), 284.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. [2015], 'limma powers differential expression analyses for rna-sequencing and microarray studies', *Nucleic acids research* **43**(7).

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. [2010], 'edgeR : a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics* **26**(1), 139–140.

Robinson, M. D. and Oshlack, A. [2010], 'A scaling normalization method for differential expression analysis of rna-seq data', *Genome biology* **11**(3), 1–9.

Robinson, M. D. and Smyth, G. K. [2007], 'Moderated statistical tests for assessing differences in tag abundance', *Bioinformatics* **23**(21), 2881–2887.

Robinson, M. D. and Smyth, G. K. [2008], 'Small-sample estimation of negative binomial dispersion, with applications to SAGE data', *Biostatistics* **9**(2), 321–332.

Rocke, D. M., Ruan, L., Gossett, J. J., Durbin-Johnson, B. and Aviran, S. [2015], 'Controlling false positive rates in methods for differential gene expression analysis using rna-seq data', *Biorxiv* p. 018739.

Romano, J. P. and Wolf, M. [2017], 'Resurrecting weighted least squares', *Journal of Econometrics* **197**(1), 1–19.

Runge, J. [2018], Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 938–947.

Saliba, A.-E., Westermann, A. J., Gorski, S. A. and Vogel, J. [2014], 'Single-cell RNA-seq : advances and future challenges', *Nucleic acids research* **42**(14), 8845–8860.

Salomon, R., Kaczorowski, D., Valdes-Mora, F., Nordon, R. E., Neild, A., Farbehi, N., Bartonicek, N. and Gallego-Ortega, D. [2019], 'Droplet-based single cell RNAseq tools : a practical guide', *Lab on a Chip* **19**(10), 1706–1727.

Sanger, F., Nicklen, S. and Coulson, A. R. [1977], 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the national academy of sciences* **74**(12), 5463–5467.

Sarkar, A. and Stephens, M. [2021], 'Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis', *Nature Genetics* **53**(6), 770–777.

Scholz, F. W. and Stephens, M. A. [1987], 'K-sample anderson–darling tests', *Journal of the American Statistical Association* **82**(399), 918–924.

Schuster, S. C. [2008], 'Next-generation sequencing transforms today's biology', *Nature methods* **5**(1), 16–18.

Scoazec, J.-Y. [2006], La révolution de l'arn, *in* 'Annales de Pathologie', Vol. 26, Elsevier, pp. 275–280.

Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G. and Shakkottai, S. [2017], Model-powered conditional independence test, *in* 'Advances in neural information processing systems', pp. 2951–2961.

SEQC/MAQC-III Consortium [2014], 'A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium', *Nature Biotechnology* **32**(9), 903–14.

Seyednasrollah, F., Laiho, A. and Elo, L. L. [2015], 'Comparison of software packages for detecting differential expression in RNA-seq studies', *Briefings in Bioinformatics* **16**(1), 59–70.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N. et al. [2014], 'Single-cell RNA-seq reveals dynamic paracrine control of cellular variation', *Nature* **510**(7505), 363–369.

Singhania, A., Verma, R., Graham, C. M., Lee, J., Tran, T., Richardson, M., Lecine, P., Leissner, P., Berry, M. P., Wilkinson, R. J. et al. [2018], 'A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection', *Nature communications* **9**(1), 2308.

Smirnov, N. V. [1939], 'On the estimation of the discrepancy between empirical curves of distribution for two independent samples', *Bull. Math. Univ. Moscou* **2**(2), 3–14.

Smyth, G. K. [2004], 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments', *Statistical applications in genetics and molecular biology* **3**(1), 1–25.

Soneson, C. and Robinson, M. D. [2018], 'Bias, robustness and scalability in single-cell differential expression analysis', *Nature methods* **15**(4), 255–261.

Southern, E., Mir, K. and Shchepinov, M. [1999], 'Molecular interactions on microarrays', *Nature genetics* **21**(1), 5–9.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. and Robinson, G. E. [2015], 'Big data : astronomical or genomical ?', *PLoS biology* **13**(7), e1002195.

Storey, J. D. [2002], 'A direct approach to false discovery rates', *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* **64**(3), 479–498.

Stubbington, M. J., Rozenblatt-Rosen, O., Regev, A. and Teichmann, S. A. [2017], 'Single-cell transcriptomics to explore the immune system in health and disease', *Science* **358**(6359), 58–63.

Su, L. and White, H. [2007], 'A consistent characteristic function-based test for conditional independence', *Journal of Econometrics* **141**(2), 807–834.

Su, L. and White, H. [2012], 'Conditional independence specification testing for dependent processes with local polynomial quantile regression', *Advances in Econometrics* **29**, 355–434.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U. and Naef, F. [2011], 'Mammalian genes are transcribed with widely different bursting kinetics', *science* **332**(6028), 472–474.

Svensson, V. [2020], 'Droplet scRNA-seq is not zero-inflated', *Nature Biotechnology* **38**(2), 147–150.

Svensson, V., Vento-Tormo, R. and Teichmann, S. A. [2018], 'Exponential scaling of single-cell rna-seq in the past decade', *Nature protocols* **13**(4), 599–604.

Szabo, P. A., Levitin, H. M., Miron, M., Snyder, M. E., Senda, T., Yuan, J., Cheng, Y. L., Bush, E. C., Dogra, P., Thapa, P. et al. [2019], 'Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease', *Nature communications* **10**(1), 1–16.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A. et al. [2009], 'mRNA-seq whole-transcriptome analysis of a single cell', *Nature methods* **6**(5), 377–382.

Tang, M., Sun, J., Shimizu, K. and Kadota, K. [2015], 'Evaluation of methods for differential expression analysis on multi-group RNA-seq count data', *BMC Bioinformatics* **16**(1), 1–14.

Tay, S., Hughey, J. J., Lee, T. K., Lipniacki, T., Quake, S. R. and Covert, M. W. [2010], 'Single-cell nf-κb dynamics reveal digital activation and analogue information processing', *Nature* **466**(7303), 267–271.

Thiébaut, R., Hejblum, B. P., Hocini, H., Bonnabau, H., Skinner, J., Montes, M., Lacabaratz, C., Richert, L., Palucka, K., Banchereau, J. et al. [2019], 'Gene expression signatures associated with immune and virological responses to therapeutic vaccination with dendritic cells in hiv-infected individuals', *Frontiers in immunology* **10**, 874.

Tiberi, S., Crowell, H. L., Weber, L. M., Samartsidis, P. and Robinson, M. D. [2020], 'distinct : a novel approach to differential distribution analyses', *bioRxiv* .

Tibshirani, R. [1996], 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society : Series B (Methodological)* **58**(1), 267–288.

Townes, F. W., Hicks, S. C., Aryee, M. J. and Irizarry, R. A. [2019], 'Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model', *Genome biology* **20**(1), 1–16.

Vallejos, C. A., Marioni, J. C. and Richardson, S. [2015], 'Basics : Bayesian analysis of single-cell sequencing data', *PLoS computational biology* **11**(6), e1004333.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. and Marioni, J. C. [2017], 'Normalizing single-cell rna sequencing data : challenges and opportunities', *Nature methods* **14**(6), 565–571.

Verbeke, G. [2000], 'Linear mixed models for longitudinal data', *Springer Series in Statistics* pp. 30–50.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S. et al. [2017], 'Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors', *Science* **356**(6335).

Wagner, G. P., Kin, K. and Lynch, V. J. [2012], 'Measurement of mRNA abundance using rna-seq data : RPKM measure is inconsistent among samples', *Theory in biosciences* **131**(4), 281–285.

Wang, T., Li, B., Nelson, C. E. and Nabavi, S. [2019], 'Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data', *BMC bioinformatics* **20**(1), 1–16.

Wang, T. and Nabavi, S. [2018], 'SigEMD : A powerful method for differential gene expression analysis in single-cell rna sequencing data', *Methods* **145**, 25–32.

Wang, X.-D., Reeves, K., Luo, F. R., Xu, L.-A., Lee, F., Clark, E. and Huang, F. [2007], 'Identification of candidate predictive and surrogate molecular markers for dasatinib in prostate cancer : rationale for patient selection and efficacy monitoring', *Genome biology* **8**(11), 1–11.

Wang, Z., Gerstein, M. and Snyder, M. [2009], 'RNA-seq : a revolutionary tool for transcriptomics', *Nature reviews genetics* **10**(1), 57–63.

Wasserman, L. [2006], *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer-Verlag, New York.

Wen, W., Su, W., Tang, H., Le, W., Zhang, X., Zheng, Y., Liu, X., Xie, L., Li, J., Ye, J. et al. [2020], 'Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing', *Cell discovery* **6**(1), 1–18.

Wickramasinghe, S., Cánovas, A., Rincón, G. and Medrano, J. F. [2014], 'RNA-sequencing : a tool to explore new frontiers in animal genetics', *Livestock Science* **166**, 206–216.

Xu, G., Qi, F., Li, H., Yang, Q., Wang, H., Wang, X., Liu, X., Zhao, J., Liao, X., Liu, Y. et al. [2020], 'The differential immune responses to COVID-19 in peripheral and lung revealed by single-cell RNA sequencing', *Cell discovery* **6**(1), 1–14.

Yang, W., Rosenstiel, P. C. and Schulenburg, H. [2016], 'ABSSeq : A new RNA-Seq analysis method based on modelling absolute expression differences', *BMC Genomics* **17**(1), 541.

Yang, Y. H. and Thorne, N. P. [2003], 'Normalization for two-color cdna microarray data', *Lecture Notes-Monograph Series* pp. 403–418.

Yazdanpanah, Y., cohort investigators, F. C., study group, Roriz, M., Rispal, P., Redl, S., Lefebvre, L., Granier, P., Maulin, L., Joseph, C., Moyet, J. et al. [2020], 'Impact on disease mortality of clinical, biological, and virological characteristics at hospital admission and overtime in COVID-19 patients', *Journal of medical virology* **93**(4), 2149–2159.

Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C. and Wang, J. [2017], 'Linnorm : improved statistical analysis for single cell RNA-seq expression data', *Nucleic acids research* **45**(22), e179–e179.

Zhang, J., Chiodini, R., Badr, A. and Zhang, G. [2011], 'The impact of next-generation sequencing on genomics', *Journal of genetics and genomics* **38**(3), 95–109.

Zhang, J. and Wu, Y. [2007], 'k-sample tests based on the likelihood ratio', *Computational Statistics & Data Analysis* **51**(9), 4682–4691.

Zhang, J.-Y., Wang, X.-M., Xing, X., Xu, Z., Zhang, C., Song, J.-W., Fan, X., Xia, P., Fu, J.-L., Wang, S.-Y. et al. [2020], 'Single-cell landscape of immunological responses in patients with COVID-19', *Nature immunology* **21**(9), 1107–1118.

Zhang, K., Peters, J., Janzing, D. and Schölkopf, B. [2011], Kernel-based conditional independence test and application in causal discovery, *in* 'Uncertainty in Artificial Intelligence : Proceedings of the Twenty-seventh Conference', AUAI Press, pp. 804–813.

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., Robinson, G. J., Lundberg, A. E., Bartlett, P. F., Wray, N. R. and Zhao, Q.-Y. [2014], 'A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data', *Plos One* **9**(8), e103207.

Zhou, Y., Liu, J. and Zhu, L. [2020], 'Test for conditional independence with application to conditional screening', *Journal of Multivariate Analysis* **175**, 104557.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M. and Siatkowski, I. [2015], 'The impact of normalization methods on RNA-seq data analysis', *BioMed research international* **2015**.

Annexe A : Revue de la littérature sur les tests d'indépendance conditionnelle

Un test d'indépendance conditionnelle (CIT) généralise les tests d'indépendance en testant l'indépendance entre deux variables étant donnés une troisième ou un ensemble de variables. Deux variables aléatoires X et Y sont conditionnellement indépendantes étant donnée une variable Z si et seulement si $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$. En ce qui concerne notre problématique d'analyse différentielle pour données RNA-seq en cellule unique (cf. Chapitre 4), les tests d'indépendance conditionnelle doivent être rapides et flexibles dans les variables qu'ils peuvent traiter. En effet, les variables peuvent être discrètes ou continues. Pour les analyses d'expression génique où des milliers de gènes sont mesurés, cela implique des milliers de tests d'indépendance. Dans le cas de données catégorielles, le plus simple est de répéter le test non-conditionnel choisi pour une valeur fixe de $Z = z$, sauf s'il existe une structure supplémentaire parmi les catégories. Lorsque la variable de conditionnement Z est continue et qu'au moins X ou Y est catégorielle, nous avons une situation "hybride". La plupart des méthodes ne s'appliquent pas en raison de la nature catégorielle de X ou Y. Pour une variable de conditionnement Z continue, on ne peut pas vérifier explicitement l'indépendance conditionnelle en vérifiant l'indépendance pour chaque valeur de Z.

Un problème retrouvé en inférence statistique consiste à tester l'égalité de k distributions à partir d'échantillons aléatoires indépendants, sans aucune hypothèse paramétrique sur les populations sous-jacentes. Les tests les plus populaires sont basés

sur la fonction de répartition empirique. Plusieurs auteurs ont généralisé les tests traditionnels pour deux échantillons au test de k échantillons. Kiefer [1959] ont proposé une extension des tests de Kolmogorov-Smirnov et de Cramér-von Mises, tandis que Scholz and Stephens [1987] ont donné la généralisation du test d'Anderson-Darling où la fonction de répartition de la i-ième condition ($1 \leq i \leq k$) est comparée à la fonction de répartition des échantillons groupés. Zhang and Wu [2007] ont proposé trois nouveaux tests de k échantillons basés sur le rapport de vraisemblance. Sous l'hypothèse que les distributions sont absolument continues, on peut introduire des tests k échantillons basés sur la comparaison de densités par le biais des estimateurs à noyau (*Kernel Density Estimators*) [Parzen, 1962]. Martínez-Camblor and de Uña-Álvarez [2009] ont introduit un test de k échantillons permettant de compararer l'estimateur à noyau du i-ième échantillon ($1 \leq i \leq k$) avec l'estimateur à noyau des échantillons groupés à l'aide des distances $L1$, $L2$ et L^{∞}. Le problème principal et non négligeable lors du calcul des estimateurs à noyau est le choix de la fenêtre de lissage. Bien que Martínez-Camblor and de Uña-Álvarez [2009] aient proposé une sélection automatique de ce paramètre, ils reconnaissent que la puissance du tes est fortement influencée par la valeur de la fenêtre. De plus, son optimisation reste très gourmande en temps de calcul. On peut également cité Su and White [2007] qui ont développé des tests non-paramétriques basés sur les distances pondérées entre fonctions caractéristiques et entre densités. Dans un article très riche, Li et al. [2009] ont proposé un test d'égalité des densités basé sur l'intégrale carrée de la différence entre deux estimateurs à noyau, les variables pouvant être continues et/ou discrètes. Ils ont ensuite étendu la méthode au cas du test d'égalité de deux distributions conditionnelles, mais la variable de conditionnement ne peut quant à elle être seulement discrète. Les fenêtres de lissage sont choisis par validation croisée. Bien que ces développements autour des estimateurs à noyau soient essentiels, les temps de calcul associés à l'optimisation des nombreux paramètres multipliés par le nombre de tests que nous souhaitons effectuer en analyse différentielle, restent inapplicables.

En parcourant la littérature [Linton et al., 1996; Su and White, 2007, 2012; Bouezmarni et al., 2012], plusieurs limites se dessinent en ce qui concerne l'application des

tests d'indépendance conditionnelle à l'analyse différentielle pour données RNA-seq en cellule unique :

- l'optimisation de la fenêtre dans le cas des estimateurs à noyau
- trop de paramètres à estimer pour les autres types d'estimateurs
- la statistique de test et sa distribution asymptotique nécessite une estimation coûteuse des paramètres
- la statistique de test ne possède pas de distribution asymptotique impliquant donc l'utilisation d'un bootstrap intensif en temps de calcul
- un manque de flexibilité dans la nature des variables à tester

Zhou et al. [2020] ont montré que l'indépendance conditionnelle entre X et Y étant donné Z est équivalente à l'indépendance non-conditionnelle entre les distributions conditionnelles $F_1(X|Z)$ et $F_2(Y|Z)$. Ainsi, tester l'indépendance conditionnelle revient à tester l'indépendance non-conditionnelle. La corrélation de Blum-Kiefer-Rosenblatt [Blum et al., 1961] est utilisé pour construire la statistique de test comme suit. Soit $V = F_1(X \mid Z)$ et $V = F_1(Y \mid Z)$. Soit $F_V(v)$ et $F_W(w)$ les fonctions de distribution marginale de V et W et $F_{V,W}(v, w)$ la distribution jointe de (V, W). La corrélation de Blum-Kiefer-Rosenblatt est définie comme suit :

$$\rho^{CI} = \int_{\mathbb{R}} \int_{\mathbb{R}} (F_{V,W}(v, w) - F_V(v)F_W(w))^2 dF_V(v)dF_W(w)$$

$\rho = 0$ si seulement V et W sont indépendants. Les distributions conditionnelles sont estimées par l'estimateur de Nadaraya-Watson. La distribution nulle asymptotique résultante ne dépend pas des distributions de X, Y ou Z. Sous H_0,

$$n\widehat{\rho} \xrightarrow{d} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{\chi_{ij}^2(1)}{\pi^4 i^2 j^2} \text{ as } n \to \infty$$

Bien que cette méthode ait retenue notre attention un certain temps et aussi attrayante soit-elle, le test n'est valable que si les fonctions de distribution $F(Y \mid Z)$

et $F(X \mid Z)$ sont continues alors que X peut être discret en analyse d'expression différentielle. La comparaison entre deux conditions (X est une variable binaire) est fondamentale et ce test ne le permet pas, ce qui constitue son principal inconvénient.

Force est de constater que notre revue de la littérature ne propose pas de solution adéquate à notre problèmatique de recherche, nous apportons alors notre résolution dans le Chapitre 4 où nous explicitons un test d'indépendance conditionnelle que nous avons développé et qui permet de palier un grand nombre des limites décrites précédemment.

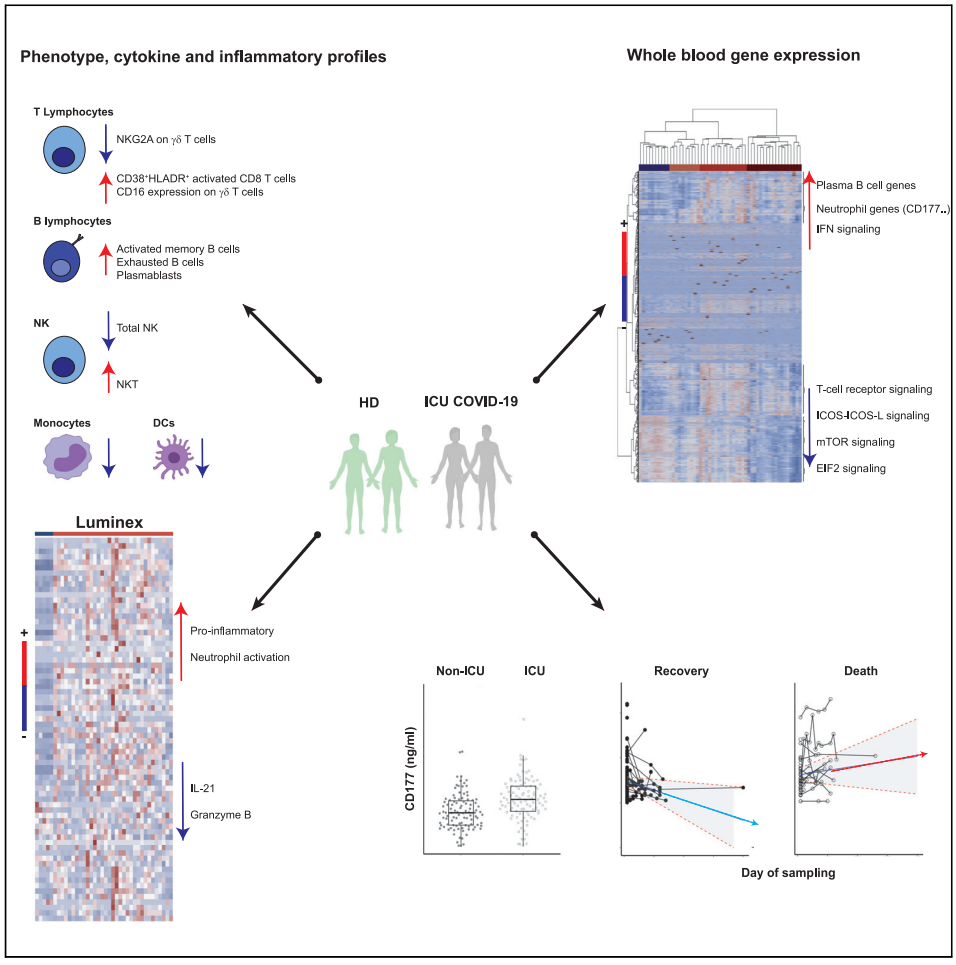
Annexe B : CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death

iScience

CellPress
OPEN ACCESS

Article

CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death



Yves Lévy, Aurélie Wiedemann, Boris P. Hejblum, ..., Giuseppe Pantaleo, Hakim Hocini, Rodolphe Thiébaut

yves.levy@aphp.fr (Y.L.)
rodolphe.thiebaut@u-bordeaux.fr (R.T.)

Highlights

Increase in B cells, activated CD8 T cells, NKT, and γδ T NKG2A + cells in severe COVID-19

Severe COVID-19 is characterized by an increase of neutrophil and inflammatory markers

Serum CD177 protein levels are increased in patients with COVID-19 in ICU

Sustained high levels of CD177 discriminated recovery and death of patients with COVID-19

Lévy et al., iScience 24, 102711 July 23, 2021
https://doi.org/10.1016/j.isci.2021.102711

iScience

CellPress
OPEN ACCESS

Article

CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death

Yves Lévy,[1,2,14,*] Aurélie Wiedemann,[1,11] Boris P. Hejblum,[1,3,11] Mélany Durand,[1,3] Cécile Lefebvre,[1] Mathieu Surénaud,[1] Christine Lacabaratz,[1] Matthieu Perreau,[4] Emile Foucat,[1] Marie Déchenaud,[1] Pascaline Tisserand,[1] Fabiola Blengio,[1] Benjamin Hivert,[3] Marine Gauthier,[3] Minerva Cervantes-Gonzalez,[5,6,7] Delphine Bachelet,[5,7] Cédric Laouénan,[5,7] Lila Bouadma,[8] Jean-François Timsit,[8] Yazdan Yazdanpanah,[6,7] Giuseppe Pantaleo,[1,4,9] Hakim Hocini,[1,12] Rodolphe Thiébaut,[1,3,10,12,*] and the French COVID cohort study group[13]

SUMMARY

The identification of patients with coronavirus disease 2019 and high risk of severe disease is a challenge in routine care. We performed cell phenotypic, serum, and RNA sequencing gene expression analyses in severe hospitalized patients (n = 61). Relative to healthy donors, results showed abnormalities of 27 cell populations and an elevation of 42 cytokines, neutrophil chemo-attractants, and inflammatory components in patients. Supervised and unsupervised analyses revealed a high abundance of *CD177*, a specific neutrophil activation marker, contributing to the clustering of severe patients. Gene abundance correlated with high serum levels of CD177 in severe patients. Higher levels were confirmed in a second cohort and in intensive care unit (ICU) than non-ICU patients ($P < 0.001$). Longitudinal measurements discriminated between patients with the worst prognosis, leading to death, and those who recovered ($P = 0.01$). These results highlight neutrophil activation as a hallmark of severe disease and CD177 assessment as a reliable prognostic marker for routine care.

INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic is caused by a newly described highly pathogenic beta coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Coronaviridae Study Group of the International Committee on Taxonomy of, 2020; Phelan et al., 2020). COVID-19 consists of a spectrum of clinical symptoms that range from mild upper respiratory tract disease in most cases to severe disease that affects approximately 15% of patients requiring hospitalization (Wu and McGoogan, 2020), some of whom require intensive care because of severe lower respiratory tract illness, acute respiratory distress syndrome, and extrapulmonary manifestations, leading to multiorgan failure and death. Several recent studies have provided important clues about the physiopathology of COVID-19 (Blanco-Melo et al., 2020; Chen et al., 2020a; Kuri-Cervantes et al., 2020; Mehta et al., 2020; Qin et al., 2020). Most compared the immune and inflammatory status of patients at different stages of the disease (Hadjadj et al., 2020; Mathew et al., 2020; Wilk et al., 2020). Thus, several important biomarkers associated with specific phases of the evolution of COVID-19 have thus far been identified (Ponti et al., 2020; Silvin et al., 2020). Inflammation, cytokine storms, and other dysregulated immune responses have been shown to be associated with severe disease pathogenesis (Lucas et al., 2020; Ong et al., 2020). Patients with severe COVID-19 are characterized by elevated numbers of monocytes and neutrophils and lymphopenia (Giamarellos-Bourboulis et al., 2020; Huang et al., 2020; Mathew et al., 2020; Zhou et al., 2020), and a high neutrophil-to-lymphocyte ratio predicts inhospital mortality of critically ill patients (Fu et al., 2020). High levels of proinflammatory cytokines, including interleukin (IL)-6,IL-1β, tumor necrosis factor (TNF), MCP-1, IP-10, and granulocyte colony-stimulating factor (G-CSF), in the plasma (Chen et al., 2020a, 2020b; Giamarellos-Bourboulis et al., 2020; Mathew et al., 2020; Zhou et al., 2020) and a possible defect in type I interferon (IFN) activity have been reported in patients with severe COVID-19 (Arunachalam et al., 2020; Hadjadj et al., 2020; Ong et al., 2020). However, these responses are dynamic, changing rapidly during the clinical course of the disease, which

[1]Vaccine Research Institute, Université Paris-Est Créteil, Faculté de Médecine, INSERM U955, Team 16, Hopital Henri Mondor, 51 Av Marechal de Lattre de Tassigny, 94010 Créteil, France

[2]Assistance Publique-Hôpitaux de Paris, Groupe Henri-Mondor Albert-Chenevier, Service Immunologie Clinique, Créteil, France

[3]Univ. Bordeaux, Department of Public Health, INSERM U1219 Bordeaux Population Health Research Centre, Inria SISTM, UMR 1219, 146 Rue Leo Saignat, 33076 Bordeaux, France

[4]Swiss Vaccine Research Institute, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland

[5]AP-HP, Hôpital Bichat, Département Épidémiologie Biostatistiques et Recherche Clinique, INSERM, Centre d'Investigation clinique-Epidémiologie Clinique 1425, F-75018 Paris, France

[6]AP-HP, Hôpital Bichat, Service de Maladies Infectieuses et Tropicales, F-75018 Paris, France

[7]Université de Paris, INSERM, IAME UMR 1137, F-75018 Paris, France

[8]APHP- Hôpital Bichat – Médecine Intensive et Réanimation des Maladies Infectieuses, Paris, France

[9]Immunology and Allergy Service, Department of Medicine, Lausanne

Continued

iScience *24*, 102711, July 23, 2021 © 2021 The Author(s).

may explain the high variability of the immunological spectrum described (Arunachalam et al., 2020; Bouadma et al., 2020; Lucas et al., 2020). This makes it difficult to deduce a unique profile of the pathophysiology of this infection, which is still undetermined. Furthermore, the high amplitude of the signals generated by the inflammation associated with the disease may hide other pathways that are involved.

From a clinical standpoint, clinicians face the daily challenge of predicting worsening patients owing to the peculiar clinical course of severe COVID-19, characterized by a sudden deterioration of the clinical condition 7 to 8 days after the onset of symptoms. Determination of the onset of the pathological process once infection has been established in a patient with a severe stage of infection is highly imprecise because of the possible paucisymptomatic or asymptomatic phase of the infection, as well as the low specificity of self-limited "flu" illness.

We used a systems immunology approach to identify host factors that were significantly associated with the time to illness onset, severity of the disease (intensive care unit [ICU] or transfer to ICU), and mortality of patients with COVID-19 enrolled in the multicentric French COVID cohort (Yazdanpanah, 2020). In addition to the depletion of T cells and mobilization of B cells, neutrophil activation, and severe inflammation, we show upregulation of *CD177* gene expression and protein levels in the blood of patients with COVID-19 in both the COVID-19 cohort and a "confirmatory" cohort, that is, Swiss cohort, relative to healthy subjects. CD177, a neutrophil activation marker, characterized critically ill patients and marked disease progression and death. Our finding highlights the major role of neutrophil activation through CD177 overexpression in the critical clinical transition point in the trajectory of patients with COVID-19.

RESULTS

Overview of the phenotype, cytokine, and inflammatory profiles of patients with COVID-19

Patient characteristics from the French COVID cohort enrolled in this analysis are shown in Table 1. All patients from this cohort were stratified as severe as per criteria of the French COVID cohort (clinicaltrials.gov NCT04262921) (Yazdanpanah, 2020), with 53 (87%) being hospitalized in an ICU (either initially or after clinical worsening or death) and eight were not. The median age was 60 years (interquartile range, 50–69) and 80% were men. Sampling for immunological analyses was performed within three days of entry and after a median of 11 days [7–14] after the onset of symptoms. We first assessed leukocyte profiles by flow cytometry using frozen peripheral blood mononuclear cells (PBMCs) from 50 patients with COVID-19 (with available PBMC samples) and 18 healthy donors (HDs) (14 or 15 HDs were used as controls per immune cell subset).

We analyzed 52 immune cell populations (gating strategies are shown in Figure S1); of which, 23 showed significant differences (Wilcoxon test adjusted for multiple comparisons) between patients with COVID-19 and HDs. We not only confirmed previously reported abnormalities but also revealed new immunological features of patients with COVID-19 (Figure S2). Patients with COVID-19 showed a significant reduction in the frequency of total $CD3^+$ T cells and $CD8^+$ T cells relative to HDs, as previously reported (De Biasi et al., 2020; Xu et al., 2020), that expressed an activated phenotype ($CD38^+HLA\text{-}DR^+$) (Figure 1A). Patients with COVID-19 also showed lower frequencies of resting memory B cells contrasting with higher frequencies of activated memory B cells and exhausted B cells (Figure 1B). As previously reported (Bouadma et al., 2020; Mathew et al., 2020), the proportion of plasmablasts was markedly higher in patients with COVID-19 (median [Q1-Q3]: 10.85% [3–23]) than HDs (0.76% [0.4–0.8]) ($P < 0.001$). Total natural killer (NK) cell frequencies, more precisely those of the $CD56^{bright}$ and $CD56^{dim}CD57^-$ NK cell subpopulations, were lower than in HDs ($P = 0.017$, $P < 0.001$, and $P = 0.004$, respectively) (Figure 1C), whereas a higher proportion of these NK cells, as well as NKT cells, were cycling, expressing Ki67 antigen ($CD56^{bright}$: 22% [13–30], $CD56^{dim}CD57^-$: 16.8% [11.5–27], and NKT: 10% [5.6–18.2]) ($P = 0.003$, $P = 0.004$, and $P = 0.001$ compared with HD) (Figure 1C). In addition, patients with COVID-19 showed significantly smaller classical ($CD14^+CD16^-$), intermediate ($CD14^+CD16^+$), and nonclassical ($CD14^-CD16^+$) monocyte subpopulations than HDs ($P = 0.013$, $P = 0.017$, $P < 0.001$, respectively) (Figure 1D). Interestingly, patients with COVID-19 tended to exhibit a higher frequency of γδ T cells than HDs (median 10.4% [7.5–16.1] vs 7.3% [6–10] in HDs; $P = 0.068$) (Figure 1E), with a significant proportion of γδ T cells showing higher expression of the activation marker CD16 ($P = 0.01$) and lower expression of the inhibitory receptor NKG2A ($P < 0.001$) than HDs (Figure 1E). Finally, we observed markedly smaller frequencies of dendritic cells (DCs) for all populations studied (pre-DC, plasmacytoid DC (pDC), and conventional DC (cDC1 and cDC2) in patients with COVID-19 than in HDs ($P < 0.001$, for all comparisons) (Figure 1F). Neutrophils count were available in 44 patients, and the concentration was more elevated in patients with COVID-19 belonging to group 2 and group 3 compared with group 1 (8.109/L vs 3.109/L, $p < 0.03$). It was also more elevated in patients hospitalized in an ICU (8.109/L vs 2.109/L, $p < 0.009$).

University Hospital, University of Lausanne, Lausanne, Switzerland

[10]CHU de Bordeaux, Pôle de Santé Publique, Service d'Information Médicale, Bordeaux, France

[11]These authors contributed equally

[12]Senior authors

[13]Members of the French COVID study group are in listed in Supplementary Information

[14]Lead contact

*Correspondence: yves.levy@aphp.fr (Y.L.), rodolphe.thiebaut@u-bordeaux.fr (R.T.)

https://doi.org/10.1016/j.isci.2021.102711

Table 1. Patient characteristics of the French COVID cohort (n=61)

	Number of patients	
Demographic characteristics		
Age – median (IQR) – years	61	60 (50–69)
Male sex – No./total No. (%)	61	49/61 (80)
ICU or transfer to ICU or death – No./total No. (%)	61	53/61 (87)
Outcome – No./total No. (%)	61	
Death		21/61 (34)
Discharge alive		40/61 (66)
Median interval from first symptoms on admission (IQR) – days	61	11 (7–14)
Comorbidities – No./total No. (%)		
Any	61	14/61 (23)
Chronic cardiac disease	61	9/61 (15)
Hypertension	61	22/61 (36)
Chronic pulmonary disease	61	5/61 (8)
Asthma	61	4/61 (7)
Chronic kidney disease	61	6/61 (10)
Chronic neurological disorder	61	2/61 (3)
Obesity	60	23/60 (38)
Diabetes	61	12/61 (20)
Smoking History – No./total No. (%)		
Smoking	61	5/61 (8)
Laboratory findings on admission - Median (IQR)		
Hemoglobin – g/dL	57	13 (11–14)
WBC count – $x10^9$/L	57	6 (5–9)
Platelet count – $x10^9$/L	57	189 (143–270)
C-reactive protein (CRP) – mg/L	57	120 (66–195)
Blood urea nitrogen (urea) – mmol/L	57	7 (5 – 12)
Symptoms on admission – No./total No. (%)		
Fever	59	51/59 (86)
Cough	57	40/57 (70)
Sore throat	56	4/56 (7)
Wheezing	54	6/54 (11)
Myalgia	56	21/56 (38)
Arthralgia	55	9/55 (16)
Fatigue	57	27/57 (47)
Dyspnea	57	46/57 (81)
Headache	57	11/57 (19)
Altered consciousness	56	3/56 (5)
Abdominal pain	53	8/53 (15)
Vomiting/nausea	56	10/56 (18)
Diarrhea	56	11/56 (20)
Clinical characteristics on admission – Median (IQR)		
SOFA score (ICU patients)	34	6 (4–8)

(*Continued on next page*)

Table 1. *Continued*

	Number of patients	
SAPS2 (ICU patients)	36	32 (27–53)
Heart rate – beats per minute	61	87 (76–104)
Respiratory rate – breaths per minute	55	24 (20–32)
Systolic blood pressure - mmHg	60	130 (109–145)
Diastolic blood pressure – mmHg	60	77 (70–87)
Oxygen saturation – percent	61	96 (91–98)
Oxygen saturation on – No./total No. (%)	56	
Room air		17/56 (30)
Oxygen therapy		39/56 (70)
Treatments – No./total No. (%)		
Antiviral	60	40/60 (66)
Antibiotic	60	46/60 (77)
Corticosteroids	60	33/60 (55)
Antifungal	60	9/60 (15)
Hydroxychloroquine	59	8/59 (14)

We then evaluated the levels of 71 serum cytokines, chemokines, and inflammatory factors in 33 patients with COVID-19 and 5 HDs. Forty-four analytes differed significantly (Wilcoxon test adjusted for multiple comparisons) between the patients with COVID-19 and HDs (shown in the heatmap in Figure 2 and detailed in Figure S3). The levels of 42 factors were higher, among them, proinflammatory factors (IL-1a, IL-6, IL-18, TNF alpha and β [TNF-α, TNF-β], IL-1ra, ST2/IL-1R4, the acute phase protein lipopolysaccharide-binding protein LBP, IFN-a2); Th1 pathway factors (IL-12 [p70], IFN-γ, IP-10, IL-2Ra); Th2/regulatory cytokines (IL-4, IL-10, IL-13); IL-17, which also promotes G-CSF-mediated granulopoiesis and recruits neutrophils to inflammatory sites; T cell proliferation and activation factors (IL-7, IL-15); growth factors (SCF, SCGF-b, HGF, b-FGF, b-NGF); and a significant number of cytokines and chemokines involved in macrophage and neutrophil activation and chemotaxis (RANTES (CCL5), MIP-1a and b (CCL3 and CCL4), MCP-1 (CCL2), MCP-3 (CCL7), M-CSF, MIF, Gro-a (CXCL1), monokine inducible by γ IFN MIG/CXCL9, IL-8, IL-9). Interestingly, we found higher levels of midkine, a marker usually not detectable in the serum, which enhances the recruitment and migration of inflammatory cells and contributes to tissue damage (Cai et al., 2020). In parallel, granzyme B and IL-21 levels were significantly lower in patients with COVID-19 patients than HDs ($P = 0.007$ and $P = 0.004$, respectively) (Figure S3).

Whole blood gene expression profiles show a specific signature for patients with COVID-19

The comparison of gene abundance in whole blood between patients with COVID-19 (n = 44) and HDs (n = 10) showed 4,079 differentially expressed genes (DEGs) with an absolute fold change ≥ 1.5, including 1,904 that were upregulated and 2,175 that were downregulated (Figure 3A). The main pathways associated with the DEG correspond to the immune response, including neutrophil and IFN signaling, T and B cell receptor responses, metabolism, protein synthesis, and regulators of the eIF2 and mammalian target of rapamycin (mTOR) signaling pathways (Figure 3A). Although several of these pathways involved multiple cell types, analysis of the neutrophil pathway showed higher abundance of genes mainly related to neutrophil activation, their interaction with endothelial cells, and migration (Figure 3B). Among the most highly expressed genes, this signature included *CD177*, a specific marker of neutrophil adhesion to the endothelium and transmigration (Bayat et al., 2010), *HP* (haptoglobin), a marker of granulocyte differentiation and released by neutrophils in response to activation (Theilgaard-Monch, 2006), *VNN1* (hematopoietic cell trafficking), *GPR84* (neutrophil chemotaxis), *MMP9* (neutrophil activation and migration), and *S100A8* and *S100A12* (neutrophil recruitment, chemotaxis, and migration). The *S100A12* protein is produced predominantly by neutrophils and is involved in inflammation and the upregulation of vascular endothelial cell adhesion molecules (Roth et al., 2003) (Figure 3B).

In parallel, we observed a higher abundance of several IFN-stimulating genes (ISGs) (*IFI27*, *IFITM3*, *IFITM1*, *IFITM2*, *IFI6*, *IRF7*, *IRF4*) (Figure 3C) and cytokines and cytokine receptors (*IL-1R*, *IL-18R1*, *IL-18RAP*, *IL-4R*, *IL-17R*, *IL-10*) (Figure 3D). Consistent with profound T cell lymphopenia, the expression of several families of

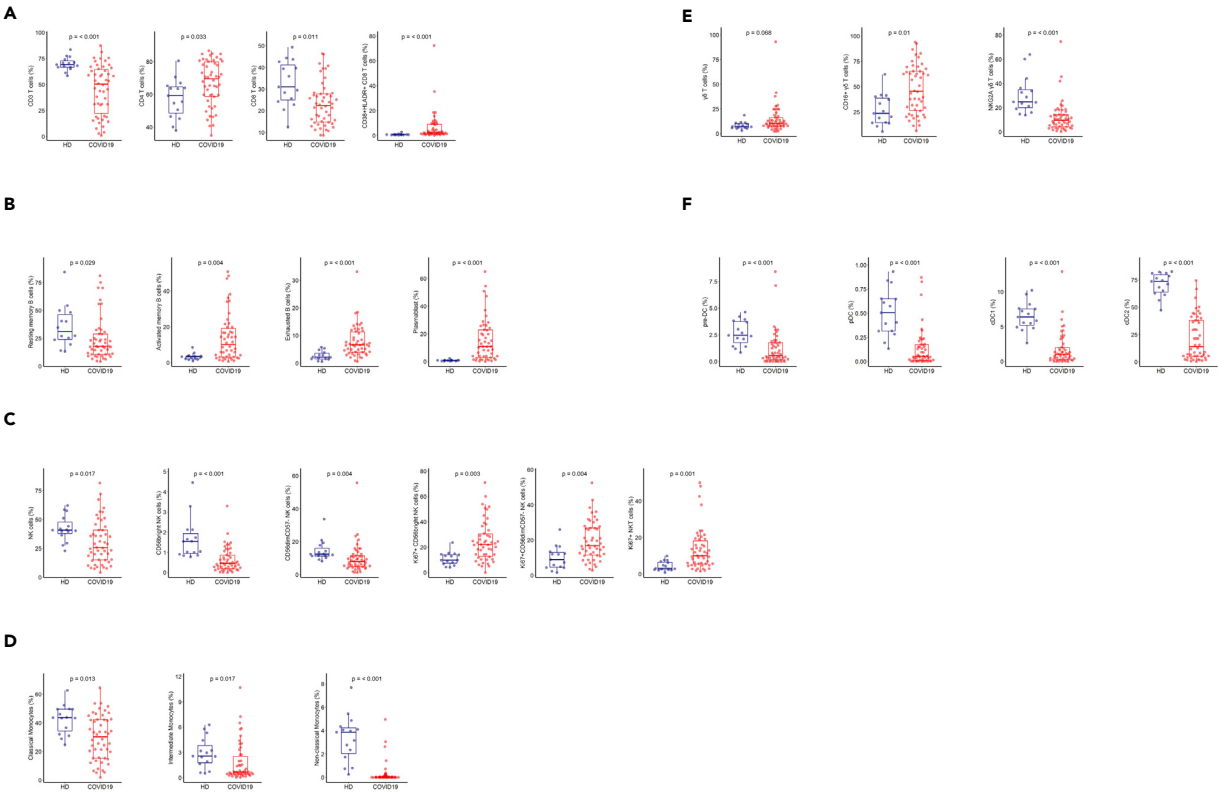


Figure 1. Frequency of immune-cell subsets between HDs (n = 18) and patients with COVID-19 (n = 50)

(A) Frequency of total CD3 T cells, CD4 and CD8 T cell subsets, and activated $CD38^{+}HLADR^{+}$ CD8 T cells.

(B) Frequency of B cell subsets ($CD21^{+}CD27^{+}$: resting memory, $CD21^{-}CD27^{+}$: activated memory, $CD21^{-}CD27^{-}$: exhausted) and plasmablasts ($CD38^{++}CD27^{+}$) gated on $CD19^{+}$ B cells.

(C) Frequency of NK cell subsets (gated on $CD3^{-}$ $CD14^{-}$) $CD56^{Bright}$: $CD56^{++}CD16^{+}$, $CD56^{dim}$: $CD56^{+}CD16^{++}CD57^{+/-}$, differentiated $Ki67^{+}$ NK cells (gated on $CD56^{Bright}$ or $CD56^{dim}CD57^{-}$ NK cells) and differentiated $Ki67^{+}$ NKT cells (gated on $CD3^{+}CD56^{+}$ cells).

(D) Monocyte subsets (gated on $CD3^{-}CD56^{-}$) (classical monocytes: $CD14^{+}CD16^{-}$, intermediate monocytes: $CD16^{+}CD14^{+}$, non-classical monocytes: $CD14^{-}CD16^{+}$).

(E) Frequency of γδ T cells (gated on $CD3^{+}$ T cells) and CD16 and NKG2A expression (gated on γδ CD3 T cells).

(F) Frequency of DC subsets (gated on $HLADR^{+}Lin^{-}$) (pDC: $CD45RA^{+}CD33^{-}CD123^{+}$, pre-DC: $CD123^{+}CD45RA^{+}$, cDC1: $CD33^{+}CD123^{-}CD141^{+}CD1c^{low}$, cDC2: $CD33^{+}CD123^{-}CD14^{+}CD1c^{+}$) detected by flow cytometry in PBMCs from n = 50 patients with COVID-19 and n = 18 HDs. The differences between the two groups were evaluated using Wilcoxon rank sum statistical tests. The lower and upper boundaries of the box represent the 25% and 75% percentiles, the whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range away from the box. Median values (horizontal line in the boxplot) are shown.

See also Figures S1 and S2.

T cell receptor genes was lower (Figure 3E). We observed severe dysregulation of T cell function that involved inhibition of serine/threonine kinase PKCθ signaling (*Z* score = −4.46), as well as the inducible T cell costimulator/ICOSL axis (*Z* score = −4.5) (Figure S4). In contrast to the results for T cells, the peripheral expansion of memory B cells and plasmablasts was associated with broad expansion of the B cell receptor (BCR) (Figure 3F and Table S2).

We also observed genes belonging to several crucial pathways and biological processes that had not been previously reported to characterize patients with COVID-19 to be underrepresented. These included eIF2 signaling, with many downregulated genes, such as ribosomal proteins and eukaryotic translation initiation factors (Figure S4A), common targets of the integrated stress response (ISR), including antiviral defense (Levin and London, 1978; Pakos-Zebrucka et al., 2016). In addition, we also found genes involved in signaling through mTOR (Figure S4B), a member of the phosphatidylinositol-3-kinase-related kinase family

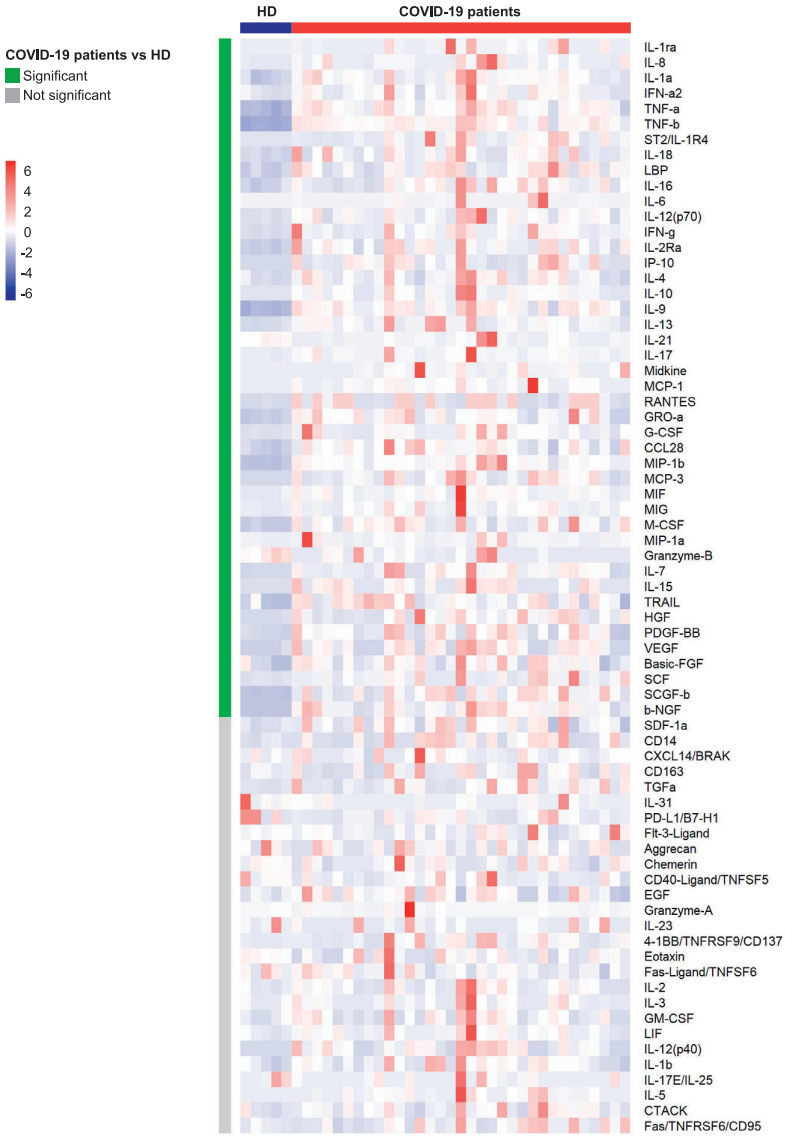


Figure 2. Heatmap of analyte abundance in serum

The colors represent standardized expression values centered around the mean, with variance equal to 1. HD: healthy donors (n = 5), COVID: patients with COVID-19 (n = 33). Each column represents a subject. Each line represents an analyte. See also Figure S3.

of protein kinases. Prediction analysis using Ingenuity pathways showed both lower eIF2 (*Z* score = - 6.8) and mTOR (*Z* score = −2.2) signaling in patients with COVID-19 than HDs.

Unsupervised whole blood gene expression profiles reveal distinct features of patients with COVID-19

Unsupervised classification of 44 patients with COVID-19 and 10 HDs identified three distinct groups of patients with COVID-19: 10 in group 1, 16 in group 2, and 18 in group 3 (Figure 4). Detailed patient characteristics as per the group are presented in Table S1. Among a large set of clinical and biological characteristics, the analysis showed the differential clustering to not be explained by the severity of the disease. Indeed, the median Sequential Organ Failure Assessment (SOFA) score and Simplified Acute Physiology

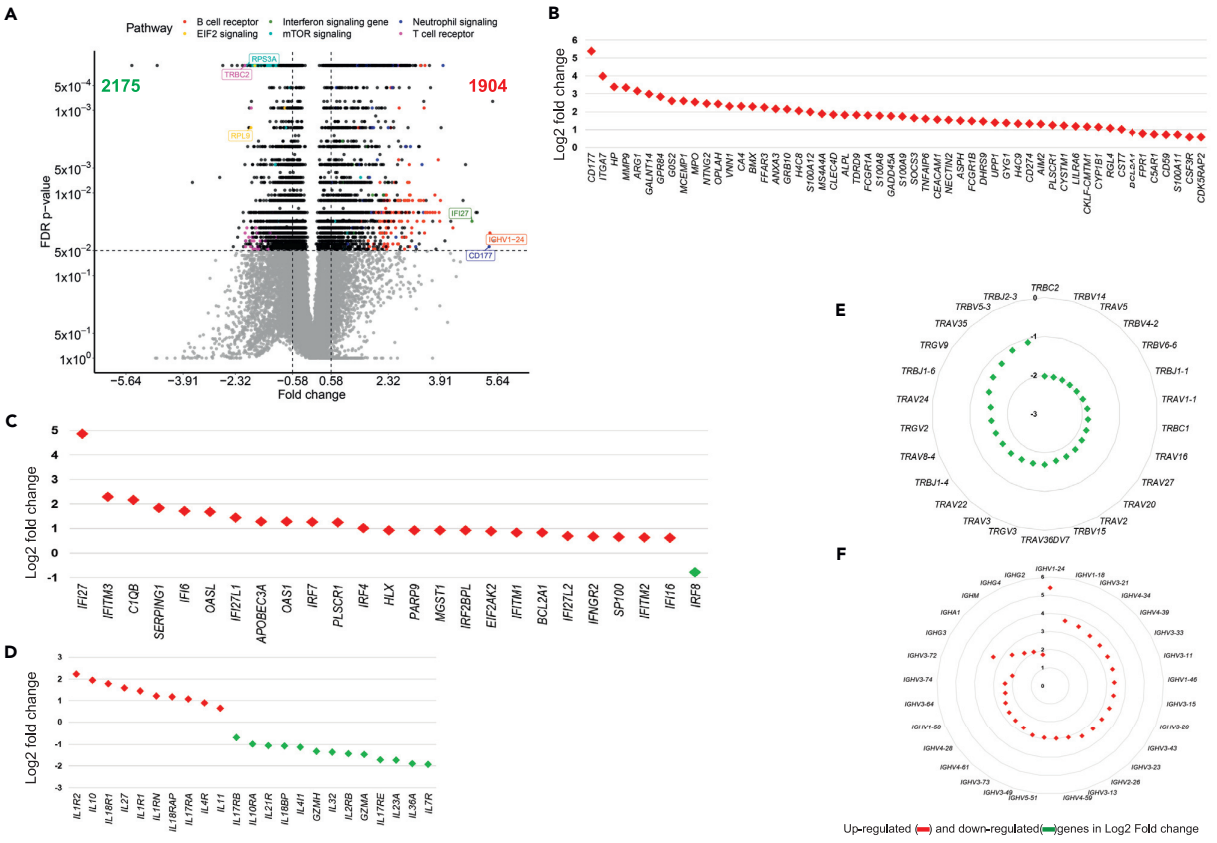


Figure 3. Whole blood gene expression in COVID-19 patients and HDs

(A) Volcano plot showing differentially expressed genes (DEG) as per the log2 fold change (log2 FC) and Benjamini-Hocberg False Discovery Rate (FDR) with thresholds at absolute log2 FC $\geq$ log2(1.5) and FDR $\leq$0.05.

(B) Main top DEG related to neutrophils.

(C and D) Main DEG related to *IFN* and interleukin responses, respectively.

(E) Main *TCRV* T cell repertoire DEG.

(F) Main B-cell *IGHV* repertoire DEG. Red symbols represent overabundant genes in COVID-19 relative to HD, green symbols represent underabundant genes.

See also Figure S4 and Table S2.

Score (SAPS2), which include a large number of physiological variables (Le Gall, 1993; Vincent et al., 1998) and evaluate the clinical severity of the disease (a high score is associated with a worse prognosis) of patients with COVID-19, were 6 [4–7] and 36 [28–53], respectively, with no significant differences between groups. Nevertheless, we observed a significant difference from symptoms onset to the admission, which ranged from 7 [6–11] days for patients in group 1 to 11 [10–14] and 13 [9–14] days for patients in groups 2 and 3, respectively (P = 0.04, Kruskal-Wallis test). Finally, group 1 which was the closest to HDs in terms of gene profile consisted of patients in the early days of the disease (Table S1).

Analysis of the genes contributing to the differences between groups confirmed and extended the findings described previously (Figures 3 and S4). Several pathways were highly represented in sectors of the heatmap defined as per gene abundance across patient groups. For example, 97% of the genes making up the BCR and 65% of those involved in neutrophil responses were represented among the genes showing a greater abundance in COVID-19 groups 2 and 3 than group 1 and HDs (Figure 4). Other pathways, such as those for IFN (64%), TCR (100%), iCOS-iCOS-L (88%), mTOR (81%), and eIF2 signaling (92%) were also highly represented. The IFN-signaling genes, such as *IFI44L*, *IFIT2*, and *IRF8*, a regulator of type I IFN (α, β), were significantly more abundant at earlier stages (in patients from group 2) and tended to be less abundant in group 3, at more advanced stages of the disease. Finally, the abundance of genes belonging to

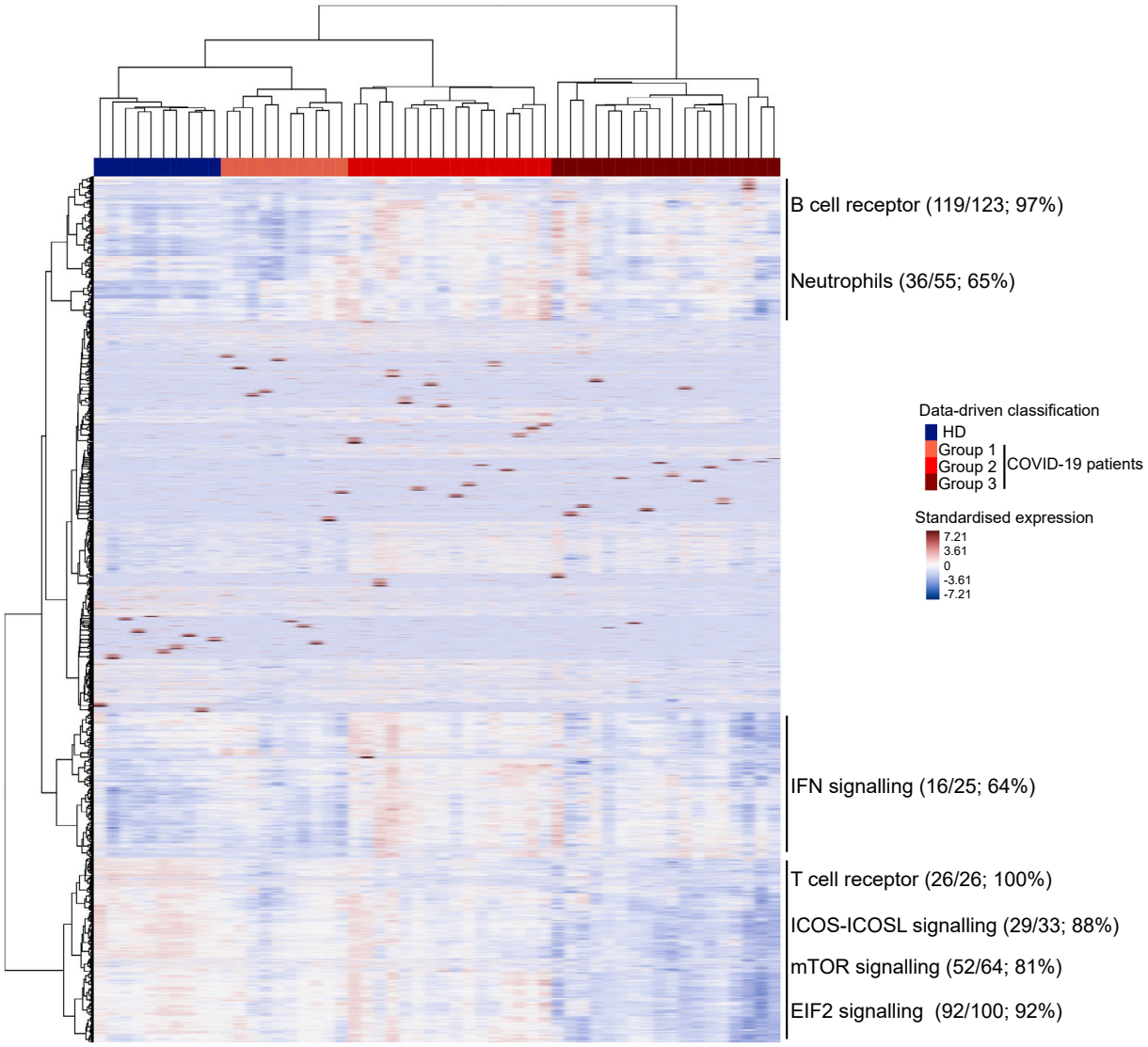


Figure 4. Heatmap of standardized gene expression

The colors represent standardized expression values centered around 0, with variance equal to 1. Each column represents a subject. This heatmap was built by unsupervised hierarchical clustering of log2-counts-per-million RNA-seq transcriptomic data from whole blood (29,302 genes) and subjects (n = 54) using the Euclidean distance and Ward's method. Seven blocks are highlighted as per the features of gene expression across the groups of individuals. Enrichment (number and % of genes of a given pathway selected in the block) of pathways of interest are shown for each block.
See also Table S1.

T cell pathways (TCR, iCOS-iCOSL signaling) or mTOR and eIF2 signaling was lower in group 3, that is to say, those who were analyzed after a longer time from symptom onset to the admission. The findings described previously highlight the heterogeneity of patients with COVID-19.

Integrative analysis of all biomarkers reveals the major contribution of CD177 in the clinical outcome of patients with COVID-19

We performed an integrative analysis using all available data to disentangle the relative contribution of the various markers at the scale of every patient. We thus pooled the data for 29,302 genes from whole blood RNA sequencing (RNA-seq), cell phenotypes (52 types), and cytokines (71 analytes) using the recently

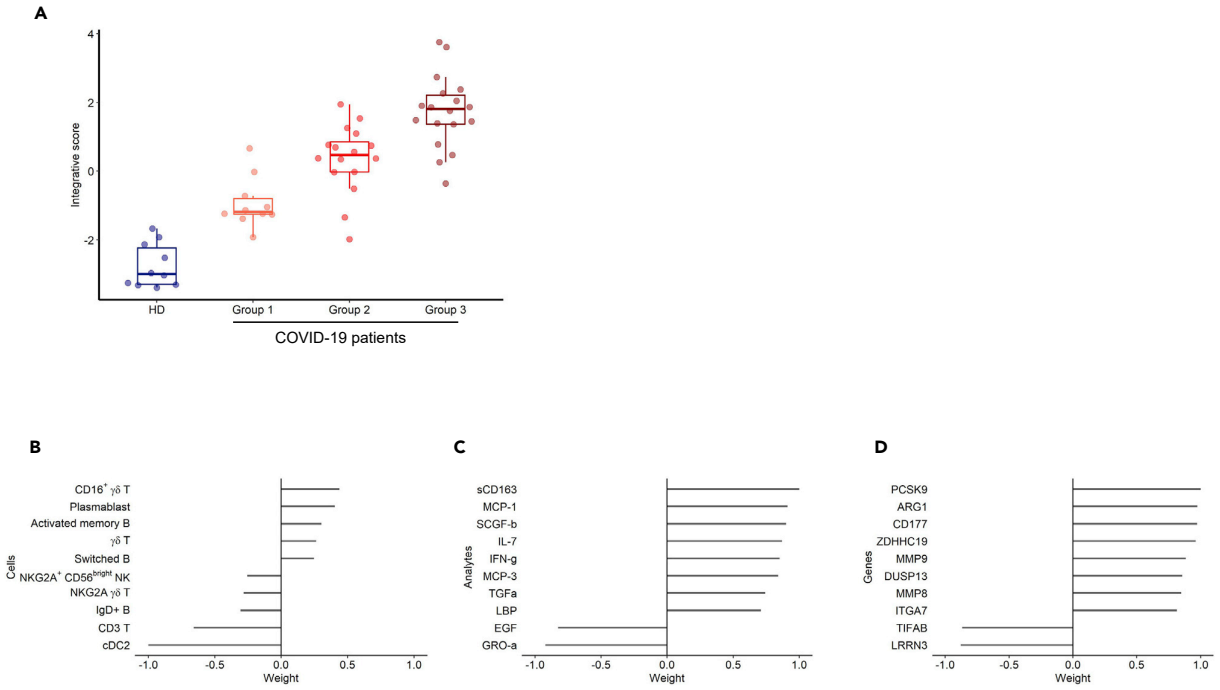


Figure 5. Integrative analysis. Integrative analysis of the data of RNA-seq (29,302 genes) from 44 patients with COVID-19, cell phenotype (52 types) from 45 patients with COVID-19, and serum analytes (71 analytes) from 33 patients with COVID-19 using a sparse principal component analysis approach, MOFA v2

(A–D)(A) Integrative score as per the patient groups defined by the hierarchical clustering of the RNA-seq data. Top 10 marker contributions (as per the weight from −1 to 1) of the cell phenotypes (B), serum analytes (C), and RNA-seq (D). The integrative score corresponds to the first factor of the analysis and allows the ordering of individuals along an axis centered at 0. Individuals with an opposite sign for the factor therefore have opposite characteristics. See also Figure S5.

described MOFA approach (Argelaguet et al., 2019), which is a statistical framework for dimension reduction adapted to the multiomics context. The data are reduced to components that are linear combinations of variables explaining interpatient variability across the three biological measurement modalities. The first component, that we called our integrative score, discriminated between the three groups of RNA sequencing (initially defined by hierarchical classification based on gene abundance only) and HDs (Figure 5A), although it only explained a portion of the variability within each of the three types of markers (14% of gene expression, 14% of cell phenotypes, and 5% of cytokines). Figures 5B–5D show the contribution of each type of markers to the integrative score. The main contributors for the cell phenotype were the significantly lower frequency of cDC2 and T cells and, marginally, the higher number of plasmablasts and CD16^{+} γδ T cells in patients with COVID-19 (Figure 5B). The contribution of soluble factors was marked by higher levels of soluble CD163 (sCD163), a marker of polarized M2 macrophages involved in tissue repair (Zhi et al., 2017), in more advanced COVID-19 groups (Figure 5C). Indeed, CD163 gene expression was also significantly higher in patients with COVID-19 than in HDs (log2 fold change = +1.55; FDR = 4.79 10^{-2}). sCD163 has also been reported to be a marker of disease severity in critically ill patients with various inflammatory or infectious conditions (Buechler et al., 2013). Interestingly, the genes that contribute the most to the synthesis of this factor were part of the neutrophil module (*CD177*, *ARG1*, *MMP9*) (Figure 5D). Integrated analysis also revealed higher expression of proprotein convertase subtilisin/kexin type 9 (*PCSK9*). High plasma PCSK9 protein levels highly correlate with the development and aggravation of subsequent multiple organ failure during sepsis (Boyd et al., 2016; Dwivedi et al., 2016). Of note, high PCSK9 levels have been recently associated with severe Dengue infection (Gan et al., 2020). Finally, the increasing abundance of CD177 gene expression as per the group was again clearly apparent (Figure S4). An additional cell-type-specific significance analysis has been performed to check the robustness of the CD177 differential expression according to the cell-type frequencies (Shen-Orr et al., 2010). We found that CD177 differential expression between patients with COVID-19 and HD was not fully explained by population variations. Indeed,

Table 2. Characteristics of patients involved in the CD177 analysis

	Number of patients	French cohort (n = 115)	Swiss cohort (n = 88)
Demographic characteristics			
Age – median (IQR) – years	200	62 (54–72)	63 (57–74)
Male sex – No./total No. (%)	201	82/113 (73)	56/88 (64)
ICU or transfer to ICU or death – No. /total No. (%)	200	61/112 (54)	40/88 (45)
Outcome - No. /total No. (%)	173		
Death		32/107 (30)	8/66 (12)
Discharge alive		75/107 (70)	58/66 (88)
Median interval from first symptoms on admission (IQT)	192	13 (9-18)	12 (9-17)
Comorbidities - No./total No. (%)			
Chronic cardiac disease	197	22/109 (20)	25/88 (28)
Chronic pulmonary disease	197	14/109 (13)	9/88 (10)
Diabetes	197	23/109 (21)	26/88 (30)
Laboratory findings on admission – Median (IQR)			
C-reactive protein (CRP) – mg/L	34	122 (62–196)	
Lactate dehydrogenase (LDH) UI/L	31	466 (337–533)	
Clinical characteristics on admission – Median (IQR)			
Score SOFA	41	4 (2–7)	
Score SAPS2	40	32 (27–49)	

it remained significant after deconvolution in several leukocytes subpopulations (notably FDR of 0.04 within T cells and 0.03 within monocytes).

Serum CD177 protein levels are associated with the clinical outcome of patients with COVID-19

Given the contribution of the neutrophil activation pathway in the clustering of patients with COVID-19, we sought neutrophil-activation features that could act as possible reliable markers of disease evolution. We focused on CD177 because i) it is a neutrophil-specific marker representative of neutrophil activation, ii) it was the most highly differentially expressed gene in patients, and iii) the protein can be measured in the serum, making its use as a marker clinically applicable. Thus, we used an enzyme-linked immunosorbent assay (ELISA) to quantify CD177 in the serum of 203 patients with COVID-19 (115 patients from the French cohort and 88 patients from the Swiss COVID-19 cohort that we used as "a confirmatory" cohort, patient characteristics are described in Table 2), 21% of whom the measurements were repeated (from 2 to 10 measurements per individual). First, we confirmed the significantly higher median serum protein level in the global cohort of patients with COVID-19 (4.5 [2.2–7.4]) relative to that of 16 HDs (2.2 [0.9–4.2]) ($P = 0.015$, Wilcoxon test) (Figure 6A). Second, we found a robust agreement between CD177 gene expression measured by RNA-seq and CD177 protein levels measured by ELISA (intraclass correlation coefficient 0.88) (Figure 6B).

Then, we examined the association of clinical characteristics and outcomes with serum CD177 concentration at the time of admission. The serum CD177 concentration was positively associated with the time from symptom onset to the admission ($r = 0.22$, $P = 0.0026$) (Figure 6C) and was higher for patients hospitalized in an ICU (6.0 ng/mL [3.5–9.4] vs 3.3 ng/mL [1.5–5.6], $P < 0.001$) (Figure 6D). The association between serum CD177 levels and hospitalization in an ICU was independent of the usual risk factors, such as age, sex, chronic cardiac or pulmonary diseases, or diabetes (multivariable logistic regression, adjusted odds ratio 1.14 per unit increase, $P < 0.001$). We observed a trend toward a positive association with the SOFA and SAPS2 risk scores that was not statistically significant ($P = 0.17$ and $P = 0.074$, respectively) (Figures S6A and S6B). CD177 levels were not associated with other conditions that contribute to a high risk of severe disease, such as diabetes ($P = 0.632$), chronic cardiac disease ($P = 0.833$), chronic pulmonary disease ($P = 0.478$), or age of the patient ($P = 0.83$).

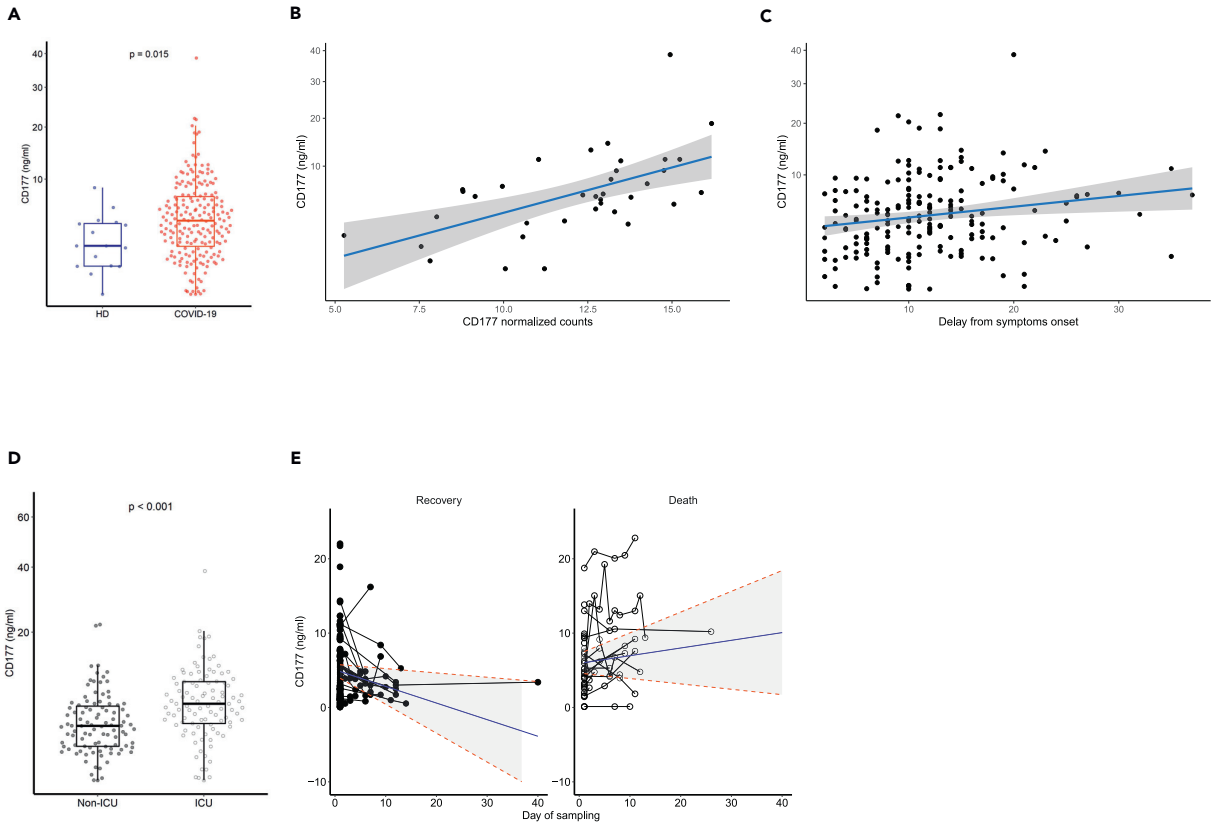


Figure 6. Distribution of the CD177 marker and association with clinical outcomes of patients with COVID-19

(A) Measurement of CD177 (ng/mL). HD: Healthy donors (n = 16), patients with COVID-19 (n = 203). The difference between the two groups was evaluated using Wilcoxon rank sum statistical tests. The median values (horizontal line in the boxplot) are shown. The lower and upper boundaries of the box represent the 25% and 75% percentiles.

(B) Correlation between normalized *CD177* values of gene expression measured by RNA-seq and CD177 protein by ELISA (ng/mL) from 36 patients with COVID-19. The blue line represents the linear regression line and the gray area the 95% prediction confidence interval.

(C) Association between CD177 serum concentration and time from symptom onset to the admission (n = 192). This association was tested using Spearman correlation tests. The blue line represents the linear regression line and the gray area the 95% confidence interval.

(D) Measurement of CD177 serum concentration in patients hospitalized in an intensive care unit (ICU) or not (n = 196). Wilcoxon rank tests were used. The median values (horizontal line in the boxplot) are shown. The lower and upper boundaries of the box represent the 25% and 75% percentiles.

(E) Change of CD177 concentration over time according to the occurrence of death for 172 patients with COVID-19 and a total of 248 measurements. Predictions were calculated using a mixed effect models for longitudinal data.

See also Figure S6.

We then examined the dynamics of the CD177 concentration in 172 patients with COVID-19, with longitudinal serum samples, using all available measurements (Figure 6E). At the first measurement, the average concentration of CD177 was not significantly different between the patients who died and those who recovered (5.93 vs 5.06, P = 0.26, Wald test). When looking at the change of CD177 concentration over time, it appears clearly that the concentration was decreasing in those who recovered (−0.22 ng/mL/day, 95% CI -0.307; −0.139), whereas it was stable in those who died later on (+0.10 ng/mL/day, 95% confidence interval -0.014; +0.192). These results show that the stability of CD177 protein levels in patients with severe COVID-19 during the course of the disease is a hallmark of a worse prognosis, leading to death.

DISCUSSION

Here, we investigated factors that influence the clinical outcomes of patients with severe COVID-19 involved in a multicentric French cohort combining standardized whole-blood RNA-seq analyses, in-depth phenotypic analysis of immune cells, and measurements of a large panel of serum analytes. An integrated

and global overview of host markers revealed several pathways associated with the course of COVID-19 disease, with a prominent role for neutrophil activation. This signature included CD177, a specific neutrophil marker of activation, adhesion to the endothelium, and transmigration. The correlation between *CD177* gene abundance and serum protein levels in the blood of patients with COVID-19 underscores the importance of this marker, making the measurement of CD177 protein levels a reliable approach that is largely accessible in routine care. We also demonstrated that the dynamics of serum CD177 levels is strongly associated with the severity of COVID-19 disease, ICU hospitalization, and survival in an additional cohort of patients. During follow-up, the CD177 protein levels decrease in patients who are recovering, while staying high in those who died.

CD177 is a glycosylphosphatidylinositol (GPI)-anchored protein expressed by a variable proportion of human neutrophils. It plays a key role in the regulation of neutrophils by modulating their migration and activation. For example, the CD177 molecule has been identified as the most dysregulated parameter in purified neutrophils from patients with septic shock (Demaret et al., 2016) and in severe influenza (Tang et al., 2019). Clinically, neutrophil chemotaxis, infiltration of endothelial cells, and extravasation into alveolar spaces have been described in lung autopsies from deceased patients with COVID-19 (Fu et al., 2020). Elevated *CD177* mRNA expression has also been described for patients with acute Kawasaki disease (Huang et al., 2019) and resistant to IV Ig therapy (Jing et al., 2020; Ko et al., 2019), a syndrome that has been described as a possible complication of SARS-CoV-2 infection in children (Toubiana et al., 2020; Viner and Whittaker, 2020). Our results are also consistent with those obtained using animal models, suggesting an important role for neutrophil activation in the severity of infection with respiratory viruses through their migration toward infected lungs, and in humans infected with influenza (Brandes et al., 2013; Narasaraju et al., 2011; Zaas et al., 2009).

The neutrophil activation signature is a specific feature of the homing of activated neutrophils toward infected lung tissue in acute lung injury (Juss et al., 2016), followed by the initiation of aggressive responses and the release of neutrophil extracellular traps (NETs), leading to an oxidative burst and the initiation of thrombus formation (Darbousset et al., 2012). Previous studies have reported elevated levels of circulating NETs in COVID-19 (Barnes et al., 2020a). Consistent with this finding and extending these data, we showed the differential expression of NET-related genes (Brandes et al., 2013; Narasaraju et al., 2011; Tang et al., 2019) (*S100A8*, *S100A9*, and *S100A12*), confirming the recently described elevated expression of calprotectin (heterodimer of *S100A8* and A9) in patients with severe COVID-19 (Silvin et al., 2020). The association of neutrophil activation signature with COVID-19 severity has also been described recently with *CD177* gene being one of the most differentially expressed gene in advanced disease (Aschenbrenner et al., 2021; Schulte-Schrepping et al., 2020). Likely, we believe that our data extended significantly these recent observations showing that CD177 is increased both at the level of coding RNA and at the protein level. Moreover, we show also that CD177 is not only a marker of severity but also of death as revealed by the longitudinal analysis which was confirmed in a second cohort.

Although, it is difficult to formally conclude whether CD177 is a causal factor of disease progression or a consequence of the severity of the disease, our data strongly show that CD177 is a valid hallmark of the physiopathology of COVID-19. This observation suggests that the activation of neutrophils, triggered by the infection, is a fundamental element of the innate response. However, persistent activation of this pathway may constitute, along with other factors (e.g., "cytokine storm"), fatal harm possibly associated with the critical turning point in the clinical trajectory of patients during the second week of the infection.

Neutrophil activation was associated with significant changes in the level of gene expression of several pathways, some concordantly associated with disturbances in immune-cell populations and cytokine and inflammatory profiles. We reveal a complex picture of the activation of innate immunity, assessed by significant changes in the expression of several genes involved in IFN signaling and the response to stress and the production of inflammatory/activation markers, with a balance between proinflammatory signals (increased expression of IL-1R1, IL-18R1, and its accessory chain IL-18 RAP) and anti-inflammatory cytokines or regulators (increased expression of IL-10, IL-4R, IL-27, IL-1RN) (Kim et al., 2005; Migliorini et al., 2020). The frequency of γδ T cells, a subpopulation of $CD3^+$ T cells that were first described in the lung (Augustin et al., 1989) and that play critical roles in anti-viral immune responses, tissue healing, and epithelial cell maintenance (Cheng and Hu, 2017), was elevated and they expressed an activation marker (CD16) and low levels of the inhibitory receptor NKG2A, suggesting possible killing capacity. In accordance with our observation that eIF2 signaling is significantly inhibited in patients with COVID-19, recent studies have shown that coronaviruses encode ISR antagonists, which act as competitive inhibitors of eIF2 signaling (Rabouw et al., 2016,

2020). Similarly, the inhibition of mTOR signaling that we found in patients with COVID-19 is consistent with the impaired mTOR signaling reported in blood myeloid dendritic cells of patients with COVID-19 (Arunachalam et al., 2020). Interestingly, we observed a markedly smaller proportion of all DC (pre-DC, pDC, cDC1, and cDC2) and monocyte cell populations, including classical, intermediate, and nonclassical subpopulations, in patients with COVID-19 than in HDs. Based on these observations, it can be hypothesized that the impairment in IFN-α production observed in patients with severe COVID-19 may be the result of both a decrease in the number of pDCs, which are natural IFN-producing cells (Ali et al., 2019), and inhibition of mTOR signaling, a regulator of IFN-α production, in these cells (Kaur et al., 2012).

We confirmed the previously reported expansion of B cell populations (Arunachalam et al., 2020; Bouadma et al., 2020) in patients with COVID-19, and our results showed that the anti-SARS-CoV-2 B cell repertoire is commonly mobilized. The marked upregulation of IGV gene families included the *IgHV1-24* family, described to be specific for COVID-19 (Brouwer et al., 2020). The expanded *VH4-39* family was also recently reported to be the most highly represented in S-specific SARS-CoV-2 sequences (Brouwer et al., 2020). We also found enrichment of *VH3-33*, previously described in a set of clonally related anti-SARS-CoV-2 receptor-binding domain antibodies (Barnes et al., 2020b).

Globally, these results show that the defense against SARS-CoV-2 after pathogen recognition triggers a fine-tuned program that not only includes the production of antiviral (IFN signaling) and proinflammatory cytokines but also signals the cessation of the response and a strong disturbance of adaptive immunity.

The same pathways (immune and stress responses through eIF2 signaling, neutrophil and IFN signaling, T and B cell receptor responses, and mTOR pathways) contributed to the ability to discriminate between three groups of patients with severe COVID-19 in an unsupervised analysis. One limitation of our study was that we did not repeat the RNA-seq analyses in these specific groups of patients. Nonetheless, it is noteworthy that these groups differed significantly by the time from disease onset. These findings provide clues in our understanding of the wide range of profiles previously described for COVID-19 by showing that these patterns may be mainly related to time-dependent changes in the blood during the course of the infection (Laing et al., 2020; Ong et al., 2020). For example, the observed lower abundance of IFN signaling genes in the last group of patients with COVID-19 may be owing to decreased abundance in more advanced disease and/or patients who constitutively present a lower abundance of ISG, as described in previous studies (Bastard et al., 2020).

Several months after the emergence of this new disease, treatment options for patients with severe disease requiring hospitalization are still limited to corticosteroids, which has emerged as the treatment of choice for critically ill patients (Prescott and Rice, 2020; Sterne et al., 2020). However, interventions that can be administered early during the course of infection to prevent disease progression and longer-term complications are urgently needed. A major obstacle for the design of "adapted" therapies to the various stages of disease evolution is a lack of markers associated with sudden worsening of the disease of patients with moderate to severe disease and markers to predict improvement. Our results show that the measurement of CD177 during the course of the disease may be helpful in following the response to treatment and revision of the prognosis. In addition, they suggest that therapies aiming to control neutrophil activation and chemotaxis should be considered for the treatment of hospitalized patients.

Limitations of the study

HDs were collected from the French Blood Donors Organization (Etablissement Français du sang) before the COVID-19 outbreak. Age to donate blood is limited between 18 and 65 years (with the necessary medical agreement beyond 60 years). Owing to the age of hospitalized patients with COVID-19 (median: 61 years), it was not possible to have age-matched HDs. So we could not exclude that differences observed between HD and patients with COVID-19 could be associated to the age difference. Samples were not available for the entire cohort of 61 patients for every experimentation but clinical characteristics were not different between included and excluded patients for each assay, that is cell phenotype (all $p > 0.97$), seric markers (all $p > 0.58$) and gene expression (all $p > 0.15$). Finally, although we have identified a set of gene characterizing activation of a neutrophil pathway in patients, as well as an association with increased serum CD177, a marker highly specific of neutrophils, we did not study the phenotype of blood neutrophils from patients included in our cohorts for practical reasons.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Subjects
- METHOD DETAILS
  - Quantification of serum analytes
  - Cell phenotyping
  - RNA sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESSOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102711.

ACKNOWLEDGMENTS

We thank the patients who donated their blood. We thank F. Mentre, S. Tubiana, the French COVID cohort, and REACTing (REsearch & ACtion emergING infectious diseases) for cohort management. We thank the scientific advisory board of the French COVID-19 cohort composed of Dominique Costagliola, Astrid Vabret, Hervé Raoul, and Laurence Weiss. We thank Romain Lévy for fruitful discussions. This work was supported by INSERM and the Investissements d'Avenir program, Vaccine Research Institute (VRI), managed by the ANR under reference ANR-10-LABX-77-01 and the CARE project funded from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 101005077. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and Bill & Melinda Gates Foundation, Global Health Drug Discovery Institute, University of Dundee. The French COVID Cohort is funded through the Ministry of Health and Social Affairs and Ministry of Higher Education and Research dedicated COVID19 fund and PHRC n°20–0424 and the REACTing consortium. Funding sources were not involved in the study design, data acquisition, data analysis, data interpretation, or writing of the manuscript.

AUTHOR CONTRIBUTIONS

YL and AW conceived and designed the study. MC, JFT, YY, LB, DB, CLa, GP, and MP participated in sample and clinical data collection. EF, MDe, MS, CLe, and PT performed the experiments. MD, MG, BH, and CLe analyzed the data. YL, RT, AW, HH, MS, BJH, and CL analyzed and interpreted the data. YL, RT, AW, and HH drafted the first version and wrote the final version of the manuscript. All authors approved the final version.

DECLARATION OF INTERESTS

None of the authors has any conflict of interest to declare.

Received: February 8, 2021
Revised: March 26, 2021
Accepted: June 8, 2021
Published: July 23, 2021

REFERENCES

Ali, S., Mann-Nüttel, R., Schulze, A., Richter, L., Alferink, J., and Scheu, S. (2019). Sources of type I interferons in infectious immunity: plasmacytoid dendritic cells not always in the driver's seat. Front. Immunol. *10*, 778.

Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y.L., Hanna, C.W., et al. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature *576*, 487.

Arunachalam, P.S., Wimmers, F., Mok, C.K.P., Perera, R., Scott, M., Hagan, T., Sigal, N., Feng, Y., Bristow, L., Tak-Yin Tsang, O., et al. (2020). Systems biological assessment of immunity to mild versus severe

COVID-19 infection in humans. Science *369*, 1210–1220.

Aschenbrenner, A.C., Mouktaroudi, M., Krämer, B., Oestreich, M., Antonakos, N., Nuesch-Germano, M., Gkizeli, K., Bonaguro, L., Reusch, N., Baßler, K., et al. (2021). Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. Genome Med. *13*, 7.

Augustin, A., Kubo, R.T., and Sim, G.K. (1989). Resident pulmonary lymphocytes expressing the gamma/delta T-cell receptor. Nature *340*, 239–241.

Barnes, B.J., Adrover, J.M., Baxter-Stoltzfus, A., Borczuk, A., Cools-Lartigue, J., Crawford, J.M., Daßler-Plenker, J., Guerci, P., Huynh, C., Knight, J.S., et al. (2020a). Targeting potential drivers of COVID-19: neutrophil extracellular traps. J. Exp. Med. *217*, e20200652.

Barnes, C.O., West, A.P., Jr., Huey-Tubman, K.E., Hoffmann, M.A.G., Sharaf, N.G., Hoffman, P.R., Koranda, N., Gristick, H.B., Gaebler, C., Muecksch, F., et al. (2020b). Structures of human antibodies bound to SARS-CoV-2 spike reveal common epitopes and recurrent features of antibodies. Cell *182*, 828–842 e816.

Bastard, P., Rosen, L.B., Zhang, Q., Michailidis, E., Hoffmann, H.-H., Zhang, Y., Dorgham, K., Philippot, Q., Rosain, J., Béziat, V., et al. (2020). Autoantibodies against type I IFNs in patients with life-threatening COVID-19. Science *370*, eabd4585.

Bayat, B., Werth, S., Sachs, U.J.H., Newman, D.K., Newman, P.J., and Santoso, S. (2010). Neutrophil transmigration mediated by the neutrophil-specific antigen CD177 is influenced by the endothelial S536N dimorphism of platelet endothelial cell adhesion molecule-1. J. Immunol. *184*, 3889–3896.

Blanco-Melo, D., Nilsson-Payant, B.E., Liu, W.C., Uhl, S., Hoagland, D., Moller, R., Jordan, T.X., Oishi, K., Panis, M., Sachs, D., et al. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell *181*, 1036–1045 e1039.

Bouadma, L., Wiedemann, A., Patrier, J., Surenaud, M., Wicky, P.H., Foucat, E., Diehl, J.L., Hejblum, B.P., Sinnah, F., de Montmollin, E., et al. (2020). Immune alterations in a patient with SARS-CoV-2-related acute respiratory distress syndrome. J. Clin. Immunol. *40*, 1082–1092.

Boyd, J.H., Fjell, C.D., Russell, J.A., Sirounis, D., Cirstea, M.S., and Walley, K.R. (2016). Increased plasma PCSK9 levels are associated with reduced endotoxin clearance and the development of acute organ failures during sepsis. J. Innate Immun. *8*, 211–220.

Brandes, M., Klauschen, F., Kuchen, S., and Germain, R.N. (2013). A systems analysis identifies a feedforward inflammatory circuit leading to lethal influenza infection. Cell *154*, 197–212.

Brouwer, P.J.M., Caniels, T.G., van der Straten, K., Snitselaar, J.L., Aldon, Y., Bangaru, S., Torres, J.L., Okba, N.M.A., Claireaux, M., Kerster, G., et al. (2020). Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. Science *369*, 643–650.

Buechler, C., Eisinger, K., and Krautbauer, S. (2013). Diagnostic and prognostic potential of the macrophage specific receptor CD163 in inflammatory diseases. Inflamm. Allergy Drug Targets *12*, 391–402.

Cai, Y.Q., Lv, Y., Mo, Z.C., Lei, J., Zhu, J.L., and Zhong, Q.Q. (2020). Multiple pathophysiological roles of midkine in human disease. Cytokine *135*, 155242.

Chen, G., Wu, D., Guo, W., Cao, Y., Huang, D., Wang, H., Wang, T., Zhang, X., Chen, H., Yu, H., et al. (2020a). Clinical and immunological features of severe and moderate coronavirus disease 2019. J. Clin. Invest. *130*, 2620–2629.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., et al. (2020b). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet *395*, 507–513.

Cheng, M., and Hu, S. (2017). Lung-resident gammadelta T cells and their roles in lung diseases. Immunology *151*, 375–384.

Coronaviridae Study Group of the International Committee on Taxonomy of, V. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544.

Darbousset, R., Thomas, G.M., Mezouar, S., Frère, C., Bonier, R., Mackman, N., Renné, T., Dignat-George, F., Dubois, C., and Panicot-Dubois, L. (2012). Tissue factor–positive neutrophils bind to injured endothelial wall and initiate thrombus formation. Blood *120*, 2133–2143.

De Biasi, S., Meschiari, M., Gibellini, L., Bellinazzi, C., Borella, R., Fidanza, L., Gozzi, L., Iannone, A., Lo Tartaro, D., Mattioli, M., et al. (2020). Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. Nat. Commun. *11*, 3434.

Demaret, J., Venet, F., Plassais, J., Cazalis, M.-A., Vallin, H., Friggeri, A., Lepape, A., Rimmelé, T., Textoris, J., and Monneret, G. (2016). Identification of CD177 as the most dysregulated parameter in a microarray study of purified neutrophils from septic shock patients. Immunol. Lett. *178*, 122–130.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Dwivedi, D.J., Grin, P.M., Khan, M., Prat, A., Zhou, J., Fox-Robichaud, A.E., Seidah, N.G., and Liaw, P.C. (2016). Differential expression of PCSK9 modulates infection, inflammation, and coagulation in a murine model of sepsis. Shock *46*, 672–680.

Fu, J., Kong, J., Wang, W., Wu, M., Yao, L., Wang, Z., Jin, J., Wu, D., and Yu, X. (2020). The clinical implication of dynamic neutrophil to lymphocyte ratio and D-dimer in COVID-19: a retrospective study in Suzhou China. Thromb. Res. *192*, 3–8.

Gan, E.S., Tan, H.C., Duyen, H.L.T., Trieu, H.T., Wills, B., Ooi, E.E., Seidah, N.G., and Yacoub, S. (2020). Dengue virus induces PCSK9 expression to alter antiviral responses and disease outcomes. J. Clin. Invest. *130*, 5223–5234.

Gauthier, M., Agniel, D., Thiébaut, R., and Hejblum, B.P. (2019). dearseq: a variance component score test for RNA-Seq differential analysis that effectively controls the false discovery rate. NAR Genom. Bioinform. *2*, lqaa093.

Giamarellos-Bourboulis, E.J., Netea, M.G., Rovina, N., Akinosoglou, K., Antoniadou, A., Antonakos, N., Damoraki, G., Gkavogianni, T., Adami, M.E., Katsaounou, P., et al. (2020). Complex immune dysregulation in COVID-19 patients with severe respiratory failure. Cell Host Microbe *27*, 992–1000 e1003.

Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., Pere, H., Charbit, B., Bondet, V., Chenevier-Gobeaux, C., et al. (2020). Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. Science *369*, 718–724.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet *395*, 497–506.

Huang, W.D., Lin, Y.T., Tsai, Z.Y., Chang, L.S., Liu, S.F., Lin, Y.J., and Kuo, H.C. (2019). Association between maternal age and outcomes in Kawasaki disease patients. Pediatr. Rheumatol. *17*, 46.

Jing, Y., Ding, M., Fu, J., Xiao, Y., Chen, X., and Zhang, Q. (2020). Neutrophil extracellular trap from Kawasaki disease alter the biologic responses of PBMC. Biosci. Rep. *40*, BSR20200928.

Juss, J.K., House, D., Amour, A., Begg, M., Herre, J., Storisteanu, D.M.L., Hoenderdos, K., Bradley, G., Lennon, M., Summers, C., et al. (2016). Acute respiratory distress syndrome neutrophils have a distinct phenotype and are resistant to phosphoinositide 3-kinase inhibition. Am. J. Resp Crit. Care *194*, 961–973.

Kaur, S., Sassano, A., Majchrzak-Kita, B., Baker, D.P., Su, B., Fish, E.N., and Platanias, L.C. (2012). Regulatory effects of mTORC2 complexes in type I IFN signaling and in the generation of IFN responses. Proc. Natl. Acad. Sci. *109*, 7723–7728.

Kim, S.H., Han, S.Y., Azam, T., Yoon, D.Y., and Dinarello, C.A. (2005). Interleukin-32: a cytokine and inducer of TNFalpha. Immunity *22*, 131–142.

Ko, T.M., Chang, J.S., Chen, S.P., Liu, Y.M., Chang, C.J., Tsai, F.J., Lee, Y.C., Chen, C.H., Chen, Y.T., and Wu, J.Y. (2019). Genome-wide transcriptome analysis to further understand neutrophil activation and lncRNA transcript profiles in Kawasaki disease. Sci. Rep. *9*, 328.

Kuri-Cervantes, L., Pampena, M.B., Meng, W., Rosenfeld, A.M., Ittner, C.A.G., Weisman, A.R., Agyekum, R., Mathew, D., Baxter, A.E., Vella, L., et al. (2020). Immunologic Perturbations in Severe COVID-19/sars-CoV-2 Infection (bioRxiv).

Laing, A.G., Lorenc, A., Del Molino Del Barrio, I., Das, A., Fish, M., Monin, L., Munoz-Ruiz, M., McKenzie, D.R., Hayday, T.S., Francos-Quijorna, I., et al. (2020). A dynamic COVID-19 immune signature includes associations with poor prognosis. Nat. Med. *26*, 1623–1635.

Le Gall, J.R. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/

North American multicenter study. JAMA *270*, 2957–2963.

Levin, D., and London, I.M. (1978). Regulation of protein synthesis: activation by double-stranded RNA of a protein kinase that phosphorylates eukaryotic initiation factor 2. Proc. Natl. Acad. Sci. U S A *75*, 1121–1125.

Lucas, C., Wong, P., Klein, J., Castro, T.B.R., Silva, J., Sundaram, M., Ellingson, M.K., Mao, T., Oh, J.E., Israelow, B., et al. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature *584*, 463–469.

Mathew, D., Giles, J.R., Baxter, A.E., Oldridge, D.A., Greenplate, A.R., Wu, J.E., Alanio, C., Kuri-Cervantes, L., Pampena, M.B., D'Andrea, K., et al. (2020). Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science *369*, eabc8511.

Mehta, P., McAuley, D.F., Brown, M., Sanchez, E., Tattersall, R.S., and Manson, J.J. (2020). COVID-19: consider cytokine storm syndromes and immunosuppression. Lancet *395*, 1033–1034.

Migliorini, P., Italiani, P., Pratesi, F., Puxeddu, I., and Boraschi, D. (2020). The IL-1 family cytokines and receptors in autoimmune diseases. Autoimmun. Rev. *19*, 102617.

Narasaraju, T., Yang, E., Samy, R.P., Ng, H.H., Poh, W.P., Liew, A.A., Phoon, M.C., van Rooijen, N., and Chow, V.T. (2011). Excessive neutrophils and neutrophil extracellular traps contribute to acute lung injury of influenza pneumonitis. Am. J. Pathol. *179*, 199–210.

Ong, E.Z., Chan, Y.F.Z., Leong, W.Y., Lee, N.M.Y., Kalimuddin, S., Haja Mohideen, S.M., Chan, K.S., Tan, A.T., Bertoletti, A., Ooi, E.E., et al. (2020). A dynamic immune response shapes COVID-19 progression. Cell Host Microbe *27*, 879–882 e872.

Pakos-Zebrucka, K., Koryga, I., Mnich, K., Ljujic, M., Samali, A., and Gorman, A.M. (2016). The integrated stress response. EMBO Rep. *17*, 1374–1395.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Phelan, A.L., Katz, R., and Gostin, L.O. (2020). The novel coronavirus originating in wuhan, China: challenges for global Health governance. JAMA *323*, 709–710.

Ponti, G., Maccaferri, M., Ruini, C., Tomasi, A., and Ozben, T. (2020). Biomarkers associated with COVID-19 disease progression. Crit. Rev. Clin. Lab. Sci. *57*, 389–399.

Prescott, H.C., and Rice, T.W. (2020). Corticosteroids in COVID-19 ARDS. JAMA *324*, 1292.

Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., Xie, C., Ma, K., Shang, K., Wang, W., et al. (2020). Dysregulation of immune response in patients with COVID-19 in Wuhan, China. Clin. Infect. Dis. *71*, 762–768.

Rabouw, H.H., Langereis, M.A., Knaap, R.C., Dalebout, T.J., Canton, J., Sola, I., Enjuanes, L., Bredenbeek, P.J., Kikkert, M., de Groot, R.J., et al. (2016). Middle East respiratory coronavirus accessory protein 4a inhibits PKR-mediated antiviral stress responses. PLoS Pathog. *12*, e1005982.

Rabouw, H.H., Visser, L.J., Passchier, T.C., Langereis, M.A., Liu, F., Giansanti, P., van Vliet, A.L.W., Dekker, J.G., van der Grein, S.G., Saucedo, J.G., et al. (2020). Inhibition of the integrated stress response by viral proteins that block p-eIF2-eIF2B association. Nat. Microbiol. *5*, 1361–1373.

Roth, J., Vogl, T., Sorg, C., and Sunderkötter, C. (2003). Phagocyte-specific S100 proteins: a novel group of proinflammatory molecules. Trends Immunol. *24*, 155–158.

Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., et al. (2020). Severe COVID-19 is marked by a dysregulated myeloid cell compartment. Cell *182*, 1419–1440.e1423.

See, P., Dutertre, C.A., Chen, J., Gunther, P., McGovern, N., Irac, S.E., Gunawan, M., Beyer, M., Handler, K., Duan, K., et al. (2017). Mapping the human DC lineage through the integration of high-dimensional techniques. Science *356*, eaag3009.

Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., and Butte, A.J. (2010). Cell type–specific gene expression differences in complex tissues. Nat. Methods *7*, 287–289.

Silvin, A., Chapuis, N., Dunsmore, G., Goubet, A.G., Dubuisson, A., Derosa, L., Almire, C., Henon, C., Kosmider, O., Droin, N., et al. (2020). Elevated Calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. Cell *182*, 1401–1418.e18.

Sterne, J.A.C., Murthy, S., Diaz, J.V., Slutsky, A.S., Villar, J., Angus, D.C., Annane, D., Azevedo, L.C.P., Berwanger, O., Cavalcanti, A.B., et al. (2020). Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19. JAMA *324*, 1330.

Tang, B.M., Shojaei, M., Teoh, S., Meyers, A., Ho, J., Ball, T.B., Keynan, Y., Pisipati, A., Kumar, A., Eisen, D.P., et al. (2019). Neutrophils-related host factors associated with severe disease and fatality in patients with influenza infection. Nat. Commun. *10*, 3422.

Theilgaard-Monch, K. (2006). Haptoglobin is synthesized during granulocyte differentiation, stored in specific granules, and released by neutrophils in response to activation. Blood *108*, 353–361.

Toubiana, J., Poirault, C., Corsia, A., Bajolle, F., Fourgeaud, J., Angoulvant, F., Debray, A., Basmaci, R., Salvador, E., Biscardi, S., et al. (2020). Kawasaki-like multisystem inflammatory syndrome in children during the covid-19 pandemic in Paris, France: prospective observational study. BMJ *369*, m2094.

Vincent, J.-L., de Mendonca, A., Cantraine, F., Moreno, R., Takala, J., Suter, P.M., Sprung, C.L., Colardyn, F., and Blecher, S. (1998). Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units. Crit. Care Med. *26*, 1793–1800.

Viner, R.M., and Whittaker, E. (2020). Kawasaki-like disease: emerging complication during the COVID-19 pandemic. Lancet *395*, 1741–1743.

Wilk, A.J., Rustagi, A., Zhao, N.Q., Roque, J., Martinez-Colon, G.J., McKechnie, J.L., Ivison, G.T., Ranganath, T., Vergara, R., Hollis, T., et al. (2020). A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nat. Med. *26*, 1070–1076.

Wu, Z.Y., and McGoogan, J.M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China summary of a report of 72 314 cases from the Chinese center for disease control and prevention. JAMA *323*, 1239–1242.

Xu, B., Fan, C.Y., Wang, A.L., Zou, Y.L., Yu, Y.H., He, C., Xia, W.G., Zhang, J.X., and Miao, Q. (2020). Suppressed T cell-mediated immunity in patients with COVID-19: a clinical retrospective study in Wuhan, China. J. Infect. *81*, e51–e60.

Yazdanpanah, Y. (2020). Impact on disease mortality of clinical, biological, and virological characteristics at hospital admission and overtime in COVID-19 patients. J. Med. Virol. *93*, 2149–2159.

Zaas, A.K., Chen, M., Varkey, J., Veldman, T., Hero, A.O., Lucas, J., Huang, Y., Turner, R., Gilbert, A., Lambkin-Williams, R., et al. (2009). Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell Host Microbe *6*, 207–217.

Zhi, Y., Gao, P., Xin, X., Li, W., Ji, L., Zhang, L., Zhang, X., and Zhang, J. (2017). Clinical significance of sCD163 and its possible role in asthma (Review). Mol. Med. Rep. *15*, 2931–2939.

Zhou, Z., Ren, L., Zhang, L., Zhong, J., Xiao, Y., Jia, Z., Guo, L., Yang, J., Wang, C., Jiang, S., et al. (2020). Heightened innate immune responses in the respiratory tract of COVID-19 patients. Cell Host Microbe *27*, 883–890 e882.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse Monoclonal anti-CD38 FITC	BD Biosciences	#340909
Mouse Monoclonal anti-HLADR PE	BD Biosciences	#347401
Mouse Monoclonal anti-CD4 BV421	BD Biosciences	#562424
Mouse Monoclonal anti-CD8 APCH7	BD Biosciences	#560179
Mouse Monoclonal anti-CD3 Alexa 700	BD Biosciences	#557943
Rat Monoclonal anti-CCR7 Alexa647	BD Biosciences	#557734
Mouse Monoclonal anti-CD21 PE	BD Biosciences	#555422
Mouse Monoclonal anti-CD27 APC	BD Biosciences	#337169
Mouse Monoclonal anti-CD45 Alexa 700	BD Biosciences	#560566
Mouse Monoclonal anti-CD56 PECF594	BD Biosciences	#564849
Mouse Monoclonal anti-HLADR BV605	BD Biosciences	#562845
Mouse Monoclonal anti-CD33 BV421	BD Biosciences	#562854
Mouse Monoclonal anti-CD141 BV711	BD Biosciences	#563155
Mouse Monoclonal anti-CD45RA PerCpCy5.5	BD Biosciences	#563429
Mouse Monoclonal anti-HLA ABC BV786	BD Biosciences	#740982
Mouse Monoclonal anti-CD86 PECF594	BD Biosciences	#562390
Mouse Monoclonal anti-PD1 BV605	BD Biosciences	#563245
Mouse Monoclonal anti-TCR gamma delta	BD Biosciences	#559878
Mouse Monoclonal anti-CD45RA PEefluor 610	ebiosciences	#61-0458-42
Mouse Monoclonal anti-Ki67 PercPe710	ebiosciences	#46-5698-82
Mouse Monoclonal anti-CD19 PC7	Beckman Coulter	#IM3628
Mouse Monoclonal anti-CD38 PercpCy5.5	Biolegend	#303522
Mouse Monoclonal anti-IgM Pacific Blue	Biolegend	#314514
Mouse Monoclonal anti-CD16 APC Cy7	Biolegend	#302018
Mouse Monoclonal anti-CD14 BV605	Biolegend	#301834
Mouse Monoclonal anti-CD1c PECy7	Biolegend	#331516
Mouse Monoclonal anti-CD40 PE	Biolegend	#334308
Mouse Monoclonal Lineage FITC	Biolegend	#348801
Mouse Monoclonal anti-CD57 PercPCy5.5	Biolegend	#359622
F(ab')2-Goat anti-Human IgD FITC	Invitrogen	#H15501
Mouse Monoclonal anti-CD123 APC	Miltenyi Biotec	#130-113-322
Mouse Monoclonal anti-NKG2A PEVio770	Miltenyi Biotec	#130-113-567
BD Cytofix/cytoperm fixation/permeabilization kit	BD Biosciences	# 554714
Biological Samples		
French COVID-19 patients	French COVID cohort	clinicaltrials.gov NCT04262921
Swiss COVID-19 patients	Swiss cohort	Swiss ethics protocol ID: 2020-00620
Critical Commercial Assays		
Human Magnetic Luminex Assay (CD163, ST2, and CD14, LBP)	R&D Systems	LXSAHM-2 kits
Human XL Cyt Disc Premixed Mag Luminex Perf Assay Kit	R&D Systems	LXSAHM-19 kit

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
48-Plex Bio-Plex Pro Human Cytokine	Bio-Rad	#12007283
CD177 ELISA Kit	ThermoFisher Scientific	EH80RBX5
Deposited Data		
Raw and analyzed data	This paper	GEO code: GSE171110
Software and Algorithms		
DIVA v6.2	BD Biosciences	https://www.bdbiosciences.com/en-us/instruments/research-instruments/research-software/flow-cytometry-acquisition/facsdiva-software
Bio-Plex Manager v6.1	Biorad	https://www.bio-rad.com/fr-fr/product/bio-plex-manager-software-standard-edition?ID=5846e84e-03a7-4599-a8ae-7ba5dd2c7684
hg19 human reference genome	This paper	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/
STAR - v. 2.5.3ar, and quantified relative to annotation model hg19 - GENCODE Genes - release 19	N/A	[https://www.gencodegenes.org/human/release_19.html]
Sequence Analysis Viewer (SAV) version 2.1.8.		
R (version 3.6)	The R Foundation for Statistical Computing, Vienna, Austria	https://www.r-project.org/
FlowJo v9	Treestar	https://www.flowjo.com/solutions/flowjo/downloads
SPICE v5.22	https://doi.org/10.1002/cyto.a.21015	(http://exon.niaid.nih.gov/spice)
Ingenuity Pathway software v.51963813.	Qiagen	Ingenuity Pathway Analysis (IPA) - QIAGEN Online Shop

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yves Lévy (yves.levy@aphp.fr).

Materials availability

No materials were newly generated for this paper.

Data and code availability

RNA sequencing data that support the findings of this study have been deposited in Gene Expression Omnibus repository with the accession codes GSE171110. Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact: Yves Lévy (yves.levy@aphp.fr)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects

We enrolled a subgroup of patients with COVID-19 of the prospective French COVID cohort in this immunological study which is part of the cohort main objectives. The median age of patients with COVID-19 was 60 years [50-69], and 80% were men. Ethics approval was given on February 5 by the French Ethics Committee CPP-Ile-de-France VI (ID RCB: 2020-A00256-33). Eligible patients were those who were hospitalized with virologically confirmed COVID-19. Briefly, nasopharyngeal swabs were performed on the day of inclusion for SARS-CoV-2 testing as per the World Health Organization (WHO) or French National Health Agency guidelines. Viral loads were quantified by real-time semiquantitative reverse transcriptase polymerase chain reactions using either the Charité WHO protocol (testing the E gene and RdRp) or the Pasteur

institute assay (testing the E gene and two other RdRp targets, IP2 and IP4). The study was conducted with the understanding and the consent of each participant or its surrogate covering the sampling, storage, and use of biological samples. The time from symptom onset to the admission has been retrospectively collected by the interview of patients enrolled in the national "French COVID-19 cohort." The Swiss cohort was approved by the ethical commission (CER-VD; Swiss ethics protocol ID: 2020-00620) and all subjects provided written informed consent. Blood from healthy donors (HDs) was collected from the French Blood Donors Organization (Etablissement Français du sang) before the COVID-19 outbreak. HD characteristics are shown in Table S3.

METHOD DETAILS

Quantification of serum analytes

In total, 71 analytes were quantified in heat-inactivated serum samples by multiplex magnetic bead assays or ELISA. Serum samples from five healthy donors were also assayed as controls. The following kits were used as per the manufacturers' recommendations: LXSAHM-2 kits for CD163,ST2,CD14 and LBP (R&D Systems); the LXSAHM-19 kit for IL-21, IL-23, IL-31, EGF, Flt-3 Ligand, Granzyme B, Granzyme A, IL-25, PD-L1/B7-H1, TGF-α, Aggrecan, 4-1BB/CD137, Fas, FasL,CCL-28, Chemerin, sCD40L, CXCL14, and Midkine (R&D Systems); and the 48-Plex Bio-Plex Pro Human Cytokine screening kit for IL-1β, IL-1rα, IL-2, IL-4, IL-5, IL-6, IL-7, IL-8 / CXCL8, IL-9, IL-10, IL-12 (p70), IL-13, IL-15, IL-17A / CTLA8, Basic FGF (FGF-2), Eotaxin / CCL11, G-CSF, GM-CSF, IFN-γ, IP-10/CXCL10, MCP-1 /CCL2, MIP-1α / CCL3, MIP-1β / CCL4, PDGF-BB (PDGF-AB/BB), RANTES/CCL5, TNF-α, VEFG (VEGF-A), IL-1a, IL-2Ra (IL-2R), IL-3, IL-12 (p40), IL-16, IL-18, CTACK / CCL27, GRO-a /CXCL1 (GRO), HGF, IFN-α2, LIF, MCP-3 / CCL7, M-CSF, MIF, MIG/CXCL9, b-NGF, SCF, SCGF-b ,SDF-1α ,TNF-b/LTA, and TRAIL (Bio-Rad). The data were acquired using a Bio-Plex 200 system. Extrapolated concentrations were used and the out-of-range values were entered at the highest or lowest extrapolated concentration. Values were standardized for each cytokine across all displayed samples (centered around the observed mean, with variance equal to 1). CD177 quantification was performed on non-inactivated serum samples (diluted 1:2 or 1:10) using a Human CD177 ELISA Kit (ThermoFisher Scientific), according to the manufacturer's instructions.

Cell phenotyping

Immune-cell phenotyping was performed using an LSR Fortessa 4-laser (488, 640, 561, and 405 nm) flow cytometer (BD Biosciences) and Diva software, version 6.2. FlowJo software, version 9.9.6 (Tree Star Inc.), was used for data analysis. CD4+ and CD8+ T cells were analyzed for CD45RA and CCR7 expression to identify the naive, memory, and effector cell subsets for coexpression of activation (HLA-DR and CD38) and exhaustion/senescence (CD57and PD1) markers. CD19+ B cell subsets were analyzed for the markers CD21 and CD27. ASC (plasmablasts) were identified as CD19+ cells expressing CD38 and CD27. We used CD16, CD56, and CD57 to identify NK cell subsets. γδ T cells were identified using an anti-TCR γδ antibody. HLA-DR, CD33, CD45RA, CD123, CD141, and CD1c were used to identify dendritic cell subsets, as previously described (See et al., 2017). Extracellular labeling was performed for all antibodies except for Ki 67 for which an intracellular labeling was performed with the BD cytofix/cytoperm kit (BD Biosciences).

RNA sequencing

Total RNA was purified from whole blood using the Tempus™ Spin RNA Isolation Kit (ThermoFisher Scientific). RNA was quantified using the Quant-iT RiboGreen RNA Assay Kit (Thermo Fisher Scientific) and quality control performed on a Bioanalyzer (Agilent). Globin mRNA was depleted using GLOBINclear Kit (Invitrogen) before mRNA library preparation with the TruSeq® Stranded mRNA Kit, as per the Illumina protocol. Libraries were sequenced on an Illumina HiSeq 2500 V4 system. Sequencing quality control was performed using Sequence Analysis Viewer. FastQ files were generated on the Illumina BaseSpace Sequence Hub. Transcript reads were aligned to the hg18 human reference genome using Salmon v0.8.2 (Patro et al., 2017) and quantified relative to annotation model "hsapiens_gene_ensembl" recovered from the R package biomaRt v2.42.1 (Durinck et al., 2009). Quality control of the alignment was performed via MultiQC v1.4 (Patro et al., 2017). Finally, counts were normalized as counts per million.

QUANTIFICATION AND STATISTICAL ANALYSIS

Subgroups of patients with COVID-19 were identified from unsupervised hierarchical clustering of log2-counts-per-million RNA-seq transcriptomics from whole blood using the Euclidean distance and Ward's method. Differential expression analysis was carried out using dearseq (Gauthier et al., 2019) to contribute

to the analysis of genes of which the abundance differed across the three subgroups of patient with COVID-19 and healthy subjects. Once the groups were defined by hierarchical clustering, the analysis of the genes contributing to each group was performed by selecting genes with an absolute fold change of ≥ 1.5 in the comparison of interest for which the difference in expression between HDs and patients with COVID-19 was significant ($P \leq 0.05$) (to avoid so called "double dipping" [https://arxiv.org/abs/2012.02936]). Pathway analyses of the genes involved in each comparison was performed using Ingenuity Pathway Analysis (IPA ®, Qiagen, Redwood City, California, Version 57662101, 2020). For canonical pathway analysis, a Z-score ≥ 2 was defined as the threshold for significant activation, whereas a Z-score ≤ -2 was defined as the threshold for significant inhibition.

The integrative analysis of the three types of biological data (RNA-seq, cell phenotypes, serum analytes) was performed using MOFA+ (Argelaguet et al., 2019), a sparse factor analysis method. It provides latent variables which are linear combination of the most influential factors for explaining interpatient variability across the three biological measurement modalities. The first component is presented and called integrative score here. The analyses of factors associated with CD177 protein concentration were performed using nonparametric Wilcoxon test or Spearman correlation coefficient when appropriate. To look at the independent association of CD177 with ICU, a logistic regression for the prediction of hospitalization in the ICU adjusted for age, sex, chronic cardiac disease, chronic pulmonary disease, and diabetes was fitted. The analysis of repeated measurements of CD177 over time was performed by using a linear mixed effect model adjusted for time from hospitalization and an interaction with survival outcome (death or recovery). The model included a random intercept and a random slope with an unstructured matrix for variance parameters. Predictions of marginal trajectories were performed. All analyses, if not stated otherwise, were performed using R software, version 3.6.3. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

ADDITIONAL RESSOURCES

A subgroup of patients with COVID-19 of the prospective French COVID cohort was enrolled in this study. French COVID cohort was registered at: https://clinicaltrials.gov/ct2/show/NCT04262921

Méthodes statistiques pour l'analyse différentielle de données RNA-seq en masse et en cellule unique appliquées en immunologie

Résumé : La technologie RNA-seq s'impose comme le nouveau standard pour la mesure de l'expression génique. Ses variations peuvent être mises en lien avec de nombreuses pathologies ou phénotypes et peuvent être détectées par des méthodes statistiques dites d'analyse différentielle. L'objectif de l'analyse différentielle est d'identifier les gènes dont le niveau d'expression est significativement associé à un ensemble de variables. La complexité grandissante des schémas expérimentaux exige des approches plus flexibles, par la nature des variables à tester et par la prise en compte de covariables, tout en maîtrisant le taux de fausses découvertes. Nous introduisons une nouvelle méthode d'analyse différentielle pour données RNA-seq en masse reposant sur un modèle linéaire à effets mixtes et un test du score en composante de variance. Par une étude de simulations et une analyse d'un jeu de données réelles sur la Tuberculose, il apparaît que notre méthode conserve une bonne puissance statistique et limite le nombre de potentiels faux positifs, comparativement aux méthodes les plus populaires. Tandis que les données RNA-seq en masse correspondent à l'expression moyenne d'une population cellulaire, l'émergence récente de la technologie RNA-seq en cellule unique a permis de mesurer le niveau d'expression des gènes à l'échelle de la cellule offrant ainsi une résolution biologique inédite. La particularité de ce nouveau type de données réside dans le nombre important de zéros et l'hétérogénéité des distributions, souvent multimodales, rendant la modélisation difficile. Afin d'allier flexibilité et absence d'hypothèse distributionnelle, nous proposons une approche basée sur un test d'indépendance conditionnelle qui s'appuie sur une estimation originale des fonctions de distribution conditionnelles par des régressions multiples. Nous l'appliquons à un jeu de données réelles de cellules T CD8+ réactives au SARS-CoV-2, afin d'identifier les gènes différentiellement exprimés dans trois groupes de gravité COVID-19 tout en tenant compte de sept sous-populations cellulaires différentes.
Mots-clés : Analyse d'expression différentielle, gène, RNA-seq en masse, RNA-seq en cellule unique, modèle mixte, test d'indépendance conditionnelle, immunologie

---

Statistical methods for differential analysis of bulk and single-cell RNA-seq data applied in immunology

Abstract: RNA-seq technology is the new standard for measuring gene expression. Its variations can be linked to many pathologies or phenotypes and can be detected by statistical methods called differential analysis. The purpose of differential analysis is to identify genes whose expression is significantly associated with a set of variables. The increasing complexity of experimental designs requires more flexible approaches, in terms of the nature of the variables to be tested and the covariates to take into account, while controlling the false discovery rate. We introduce a new differential analysis method for bulk RNA-seq data based on a linear mixed effects model and a variance component score test. Through a simulation study and the analysis of a real-world Tuberculosis data set, it is shown that our method retains good statistical power and limits the number of potential false positives, compared to the most popular methods. While bulk RNA-seq data represent the average expression of a cell population, the recent development of single-cell RNA-seq technology allows to measure gene expression at the cell level, providing a new biological resolution. The specificity of this type of data lies in the large number of zeros and the heterogeneity of the distributions, often multimodal, making modelling difficult. In order to combine flexibility and distribution-free tool, we propose an approach based on a conditional independence test which relies on an original estimation of conditional cumulative distribution functions using multiple regressions. We apply it to a real data set of SARS-CoV-2 reactive CD8+ T cells, in order to identify genes differentially expressed in three COVID-19 severity groups while considering seven different cell subpopulations.
Keywords: differential expression analysis, gene, bulk RNA-seq, single-cell RNA-seq, mixed model, conditional independence test, immunology

---

Discipline : Santé publique – option : Biostatistiques
Laboratoire : Unité INSERM U1219, Bordeaux Population Health center - INRIA - Université de Bordeaux 146 rue Léo Saignat 33076 Bordeaux, FRANCE