TESIS DE DOCTORADO

IDENTIFICATION OF GENOME WIDE HOST RNA BIOMARKERS FOR INFECTIOUS DISEASES

Ruth Barral Arca

ESCUELA DE DOCTORADO INTERNACIONAL

PROGRAMA DE DOCTORADO EN AVANCES Y NUEVAS ESTRATEGIAS EN CIENCIAS FORENSES

SANTIAGO DE COMPOSTELA

2020

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DEL AUTOR DE LA TESIS

Identification of genome wide host RNA biomarkers for infectious diseases

Dña. Ruth Barral Arca

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1) La tesis abarca los resultados de la elaboración de mi trabajo.
2) En su caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.
3) La tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.
4) Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.
5) Declaro no tener ningún conflicto de interés en relación con la tesis de doctorado

En Santiago de Compostela, XX de julio de 2020

Fdo Ruth Barral Arca

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DEL DIRECTOR / TUTOR DE LA TESIS

Identification of genome wide host RNA biomarkers for infectious diseases

D. Antonio Salas Ellacuriaga

D. Federico Martinón Torres

INFORMAN:

Que la presente tesis, corresponde con el trabajo realizado por Dña. Ruth Barral Arca, bajo mi dirección, y autorizo su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director de ésta no incurre en las causas de abstención establecidas en Ley 40/2015.

En Santiago de Compostela, XX de julio de 2020

Fdo. Antonio Salas Ellacuriaga

Fdo. Federico Martinón Torres

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Dña. Ruth Barral Arca declara no tener ningún conflicto de intereses en relación con la Tesis Doctoral titulada:

Identification of genome wide host RNA biomarkers for infectious diseases

En Santiago de Compostela a XX de julio de 2020

Fdo. Ruth Barral Arca

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Dedicatoria

A mis padres, por su apoyo incondicional

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Por años, disfrutar del error
y de su enmienda,
haber podido hablar, caminar libre,
no existir mutilada,
no entrar o sí en iglesias,
leer, oír la música querida,
ser en la noche un ser como en el día.

No ser casada en un negocio,
medida en cabras,
sufrir gobierno de parientes
o legal lapidación.
No desfilar ya nunca
y no admitir palabras
que pongan en la sangre
limaduras de hierro.

Descubrir por ti misma
otro ser no previsto
en el puente de la mirada.

Ser humano y mujer, ni más ni menos

Poema Fortuna de Ida Vitale

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimientos

Pensé que cuando me dispusiese a escribir los agradecimientos sería algo rápido y sencillo dado que sería de lo poco de esta tesis que se me permite escribir en castellano. Pero ahora siento que son como cerrar un capítulo de mi vida que guardo con muchísimo cariño. Y como si de un bistec lleno de tendones se tratase, estos agradecimientos se me hacen "bola".

A mi director Toño, gracias por darme la oportunidad de descubrir el mundo de la investigación y por tu apoyo, confianza y ánimos en los últimos cinco años. También me gustaría darle las gracias a mi segundo supervisor Federico, por compartir conmigo su experiencia en el campo de la medicina y por muchas veces, confiar en mí más de lo que yo confiaba en mí misma. Siento mucho no haber patentado algo que te quitase de trabajar como siempre pedías. Gracias a ambos por ayudarme a crecer como científica, y a superar uno de los mayores retos de mi vida.

Al equipo del Luis Concheiro, por recibirme con los brazos abiertos durante la primera etapa de esta tesis.

A Alberto, Sara, Miriam, Jacobo, Mariajo, Xabi, Mirian, Patricia, Irene, Belén y todo el equipo Genvip por ser compañeros en este viaje. Por compartir horas de esfuerzo y también momentos de esparcimiento. Os deseo el mayor de los éxitos.

A Iván por acompañarme durante este reto, y por su absoluto apoyo en todas las aventuras en las que me embarcado.

A Jenny, Helena, Ana y Rebeca. Siempre digo que me cuesta hacer amigas, pero las que tengo son unas verdaderas joyas. Gracias por estar siempre ahí, y ayudarme a desconectar de la tesis.

A mi tío José Manuel quién habría disfrutado de asistir a la lectura de esta tesis, gracias por esas tardes hablando de ciencia y literatura, gracias también por hablar siempre de mi lleno de orgullo como tu sobrina la científica. Te echamos de menos.

A massive thank go to Professor Levin's research group (Michael Levin, Jehtro Herberg, Victoria Wright, Stephanie Menikou, Dominique Habgood…) for welcoming during three months in their lab as one more of the team. I truly cherish my time there. I would like to particularly thanks Myrsini Kaforou for your kindness, for being my mentor, and all the time you spent with me while I was giving my first steps into transcriptomics and statistical modelling. You are truly an inspiration.

I would also like to thank Fabian Grandke, my colleague in my new role in industry, for always asking me about how my thesis was going and being a big big support during the last months of this thesis.

Pero el mayor agradecimiento va para mis padres por su amor y apoyo incondicional. Gracias papá por transmitirme tu amor a la ciencia, por desde que tengo memoria hablarme de átomos, galaxias lejanas, moléculas colisionando... Seguramente si gran parte de mi infancia y adolescencia no hubiese transcurrido en la cocina de casa oyéndote hablar de física y matemáticas no estaría hoy aquí. Gracias a mi madre por ser mi red de seguridad, por cuidarme de todas las formas imaginables, y por ser un ejemplo de mujer fuerte e independiente.

A todos aquellos que me estoy dejando en el tintero, pero que me han ayudado durante este reto. Mil gracias.

LIST OF PUBLICATIONS

Published and Accepted

1. Barral-Arca, R., Pardo-Seco, J., Martinón-Torres, F., & Salas, A. (2018). **A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms**. Scientific Reports, 8(1), 1-7.
2. Salas, A., Pardo-Seco, J., Barral-Arca, R., Cebey-López, M., Gómez-Carballa, A., Rivero-Calle, I.,... & Martinón-Torres, F. (2018). **Whole exome sequencing identifies new host genomic susceptibility factors in empyema caused by streptococcus pneumoniae in children: a pilot study**. Genes, 9(5), 240.
3. Barral-Arca, R., Pardo-Seco, J., Bello, X., Martinon-Torres, F., & Salas, A. (2019). **Ancestry patterns inferred from massive RNA-seq data**. RNA, 25(7), 857-868.
4. Gómez-Carballa, A., Cebey-López, M., Pardo-Seco, J., Barral-Arca, R., Rivero-Calle, I., Pischedda, S.,... & Salas, A. (2019). **A qPCR expression assay of IFI44L gene differentiates viral from bacterial infections in febrile children**. Scientific Reports, 9(1), 1-12.
5. Barral-Arca, R.; Gómez-Carballa, A.; Cebey-López, M.; Bello, X.; Martinón-Torres, F.; Salas, A. **A Meta-Analysis of Multiple Whole Blood Gene Expression Data Unveils a Diagnostic Host-Response Transcript Signature for Respiratory Syncytial Virus.** Int. J. Mol. Sci. 2020, 21, 1831.

6. Barral-Arca, R., Gómez-Carballa, A., Cebey-López, M., Currás-Tuala, M. J., Pischedda, S., Viz-Lasheras, S.,... & Salas, A. (2020). **RNA-Seq Data-Mining Allows the Discovery of Two Long Non-Coding RNA Biomarkers of Viral Infection in Humans**. International Journal of Molecular Sciences, 21(8), 2748.

Acceptation pending

1. Host transcriptomic response following administration of rotavirus vaccine in infants' mimics wildtype. Barral-Arca,R., Gómez-Carballa, A., Cebey-López, M., Currás-Tuala,MJ., Pischedda ,S., Habgood Coote,D., Herberg,J., Kaforou,M., Martinón-Torres,F., Salas,A.

CONTENT

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Content of Figures, Tables and Equations

Figures

Tables

Equations

Resumen

Según la Organización Mundial de la Salud (OMS), las enfermedades infecciosas siguen siendo una de las principales causas de mortalidad y hospitalización infantil en todo el mundo. Hasta hace poco, se creía que generalmente la mayoría de las muertes eran causadas por enfermedades bacterianas, pero durante la última década, hay evidencias crecientes que apuntan al hecho de que las infecciones virales también son responsables de una morbilidad y mortalidad significativas en los niños.

Hoy en día, distinguir entre infecciones virales y bacterianas sigue siendo un desafío, ya que los resultados de los cultivos bacterianos pueden tardar días en estar disponibles y, a menudo, son negativos cuando la infección se encuentra en lugares inaccesibles o los niños han recibido tratamiento con antibióticos. Por tanto, la mayoría de los médicos por temor a pasar por alto una infección bacteriana potencialmente mortal deciden admitir niños febriles en el hospital y administrar antibióticos mientras esperan los resultados de los cultivos. En consecuencia, numerosas infecciones virales se tratan erróneamente con antibióticos, lo que contribuye al desarrollo de bacterias resistentes a los antibióticos. El uso excesivo de antibióticos, junto con la ausencia de medicamentos antimicrobianos de nueva generación, está creando una "era post-antibiótica", en la que las infecciones comunes y las lesiones menores podrían ser mortales, un escenario probable en el futuro cercano según lo declarado por la OMS.

El desarrollo de ensayos basados en la PCR (reacción en cadena de la polimerasa), junto con los microarrays, ha aumentado

notablemente nuestra capacidad de diagnosticar de forma precisa infecciones virales conocidas y emergentes además de posibilitar la detección de múltiples patógenos en una sola prueba. Cabe destacar sin embargo que los ensayos moleculares son menos eficaces en la detección de infecciones bacterianas, especialmente las causadas por bacterias invasivas. Además, la PCR podría no decirnos quién es el agente causal primario, ya que solo indica la presencia de ácidos nucleicos del patógeno. Pero el patógeno detectado podría no ser viable y su presencia podría atender a una enfermedad reciente pero no relacionada o una simple colonización asintomática. Además, tan solo aquellos patógeno que hayan sido considerados en el diseño del panel de PCR podrán ser detectado, lo que pueden darse falsos negativos para el agente causal. Por lo tanto, mejorar nuestras herramientas de diagnóstico para que sean más eficaces y rápidas es uno de los mayores desafíos en la atención médica actual, como ha quedado demostrado en la actual pandemia del coronavirus SARS-CoV-2.

En este trabajo de tesis, nos centramos en el estudio de las firmas de expresión génica, ya que el transcriptoma, al ser la conexión entre la información contenida en nuestros genes y el fenotipo puede actuar como el "canario" del genoma indicándonos los genes del huésped que tienen potencial como biomarcadores de enfermedades infecciosas. Mientras que el genoma de todas las células dentro del cuerpo humano es el mismo, el transcriptoma es una capa dinámica de información, que cambia entre los tipos de células y las condiciones del organismo. Por lo tanto, cuantificar la cantidad de ARN mensajero ayuda a comprender la actividad celular a nivel molecular, facilitando el análisis de la respuesta del paciente al virus, bacterias y/o parásitos. Por lo tanto, los enfoques basados en el estudio de la transcriptómica del huésped tienen un gran potencial para arrojar luz sobre la patogénesis de las enfermedades infecciosas y pueden permitir nuevos

enfoques de diagnóstico en comparación a aquellos basados en el estudio del genoma.

La mayoría de las firmas transcriptómicas han sido descubiertas utilizando datos obtenidos a partir de microarrays de expresión. Sin embargo, en los últimos años la técnica de RNA-seq está aumentando su popularidad y sustituyendo rápidamente a los microarrays. Debido a que permite el descubrimiento de nuevos transcritos, interpretar eventos de *splicing* alternativo, estudiar la expresión de cada hebra de ADN por separado etc. Pero más interesante es el hecho de que permite realizar estudios meta-transcriptómicos, esto es secuenciar todas las moléculas de ARN de las distintas especies que viven en el mismo ecosistema al mismo tiempo, posibilitando ejemplo permite estudiar el transcriptoma del microbioma intestinal junto con el huésped

Como el número de moléculas medidas por microarrays y RNA-seq puede ser enorme, es necesario utilizar algoritmos supervisados de aprendizaje automático para seleccionar los transcritos con potencial como biomarcadores. Un breve resumen del proceso sería: Primero obtener el número de moléculas (*counts*) por cada transcrito en el caso de RNA-seq o la intensidad con la que emite fluorescencia cada sonda en el caso de los microarrays. A continuación, es necesario normalizar los datos para estimar de forma precisa el nivel de expresión y así poder hacer comparaciones precisas de la expresión génica entre muestras; este paso es necesario ya que, aunque en general cuantas más moléculas/intensidad por transcrito mayor expresión, existen otros factores como la profundidad de secuenciación, tamaño del gen o incluso el contenido guanina/citosina que influyen en la cantidad de transcritos/intensidad detectada. Después realizar un análisis de expresión diferencial utilizando la distribución binomial negativa en el caso de datos RNA-seq o una distribución normal si se trabaja con microarrays. Finalmente, se obtiene una tabla de transcritos

diferencialmente expresados. Con esta lista, se puede realizar análisis adicionales en función del objetivo del experimento, como el análisis de enriquecimiento de rutas metabólicas o usar algoritmos para buscar firmas

Afortunadamente, hoy por hoy disponemos de muchos algoritmos que pueden aplicarse a la búsqueda de firmas transcriptómicas; por citar algunos: *lasso*, *elastic net*, máquinas de soporte de vectores etc. Cuando hay muchos transcritos en un modelo de predicción, estos algoritmos permiten elegir automáticamente la mejor combinación de transcritos para construir un modelo de predicción óptimo. Eliminar las transcritos menos relevantes ayuda a encontrar un modelo más parsimonioso, simple y fácil de entender. Cuando el rendimiento es el mismo, los modelos más simples siempre se deben priorizar frente a aquellos más complejos.

El objetivo principal es encontrar un número reducido de transcritos con una especificidad y sensibilidad lo suficientemente altas como para ser utilizadas en entornos clínicos. La forma más común de evaluar el rendimiento del diagnóstico de la firma transcriptómica es utilizar las curvas ROC (acrónimo de *Receiver Operating Characteristic Curve*) junto con valores predictivos positivos y negativos en una cohorte test, que debe ser diferente a la utilizada para entrenar el algoritmo, cohorte de entrenamiento.

Para que las firmas transcriptómicas sean aplicables en entornos clínicos, se deben cumplir dos condiciones: reducir el coste de la prueba al mínimo y reducir el procesamiento de muestras y datos a tiempos que se adapten a las necesidades clínicas de tratamiento y monitorización del paciente. La forma más fácil de conseguirlo sería usar qPCR en lugar de microarrays o RNA-seq junto con un software sencillo que estimase el resultado de la prueba basándose en la concentración de ARNm de la firma transcriptómica.

La principal ventaja de la secuenciación de nueva generación (NGS) es que permitiría analizar simultáneamente los niveles de expresión del huésped y la presencia de ácidos nucleicos de múltiples patógenos en una misma prueba. Pero, a pesar de que el coste y el tiempo de procesamiento requeridos para NGS han disminuido en la última década, solo unas pocas tecnologías de NGS están totalmente adaptadas para el diagnóstico en entornos clínicos o a pie de cama.

Es importante señalar que las firmas transcriptómicas no tienen como objetivo reemplazar las pruebas diagnósticas basados en la detección de patógenos ni los estudios microbiológicos, sino que se convertirán en una nueva fuente de información para el tratamiento de los pacientes infectados. El desarrollo de firmas basadas en el huésped sería particularmente útil para desentrañar la etiología de pacientes febriles, o para distinguir enfermedades infecciosas de enfermedades autoinmunes como el Kawasaki.

A pesar de las evidencias y resultados prometedores que se pueden encontrar en la literatura, ninguna firma transcriptómica de enfermedades infecciosas se ha transformado en una prueba de diagnóstico en el punto de atención. En esta tesis doctoral se demuestra la viabilidad del uso de algoritmos de aprendizaje automático para el descubrimiento de marcadores de huésped bacterianos y virales, y la viabilidad para traducirlos a pruebas de qPCR que pueden ser implementadas en muchos centros fácilmente, ya que la mayoría hospitales ya están haciendo pruebas basadas en qPCR para muchas condiciones diferentes. Por lo tanto, esta tesis puede considerarse un paso hacia delante en el desarrollo de nuevas pruebas diagnósticas basadas en firmas transcriptómicas y algoritmos de aprendizaje automático.

Las hipótesis de partida de esta tesis que fueron evaluadas en los diferentes estudios fueron:

1. Existe una predisposición genética en humanos a la susceptibilidad y severidad de las infecciones. O dicho de otra forma los patógenos no infectan a quien quieren sino a quien pueden.
2. Durante las infecciones agudas, el transcriptoma del huésped sufre cambios que son específicos del patógeno ("huella/firma transcriptómica").
3. El ARN extraído de sangre periférica de pacientes con infecciones agudas se puede utilizar para descubrir firmas de expresión en huésped específicas para distintos patógenos o condiciones.
4. La ancestralidad tiene un impacto en la expresión génica, puede ser un factor de confusión, y debe tenerse en cuenta al buscar firmas de expresión del huésped en todo el genoma
5. La ancestralidad se puede inferir de los datos de secuencia de ARN

En concreto, los principales objetivos fijados para esta tesis fueron:

1. Desarrollar un pipeline de análisis bioinformático aplicable a los datos de microarrays de expresión génica y secuencias de ARN (RNA-seq) para descubrir genes expresados diferencialmente
2. Emplear algoritmos de aprendizaje automático para identificar pequeños grupos de transcritos con potencial como biomarcadores, que permitan discriminar entre infecciones fenotípicamente similares.
3. Evaluar el rendimiento de los conjuntos de biomarcadores y validar su poder clasificatorio en cohortes independientes.

4. Comparar el rendimiento de nuestros biomarcadores descubiertos con otros encontrados por otros investigadores, así como las pruebas de diagnóstico actuales empleadas en la rutina clínica.
5. Desarrollar un *pipeline* de análisis bioinformático para inferir polimorfismos de nucleótido único (SNP) a partir de lecturas de RNA-seq. Y a partir de estos SNPs estimar la ancestralidad del paciente.
6. Evaluar cómo las variables de confusión, particularmente la ancestralidad del paciente, pueden afectar la expresión génica.

Durante esta tesis, nos concentramos en la identificación de biomarcadores para enfermedades infecciosas transcriptómicos en el huésped, tanto en un sentido amplio (infecciones virales frente a bacterianas) pero también investigamos patógenos concretos el como rotavirus, virus respiratorio sincitial (VRS), *S. pneumoniae*, etc.

A pesar de que tanto RNA-seq como los microarrays son dos herramientas muy potentes para descubrir firmas de ARN, ambos tienen problemas inherentes, como una tasa de error más alta que la secuencia de Sanger tradicional, problemas de estandarización y reproducibilidad, etc. Por lo tanto, antes de que cualquier biomarcador sea traducido a una prueba clínica, debe validarse utilizando tecnologías precisas y menos propensas a falsos positivos que la secuenciación de nueva generación y los microarrays como la qPCR o Nanostring®.

El desarrollo de una prueba diagnóstica a pie de cama basada en biomarcadores transcriptómicos es un objetivo difícil de conseguir actualmente debido a las limitaciones técnicas existentes. Sin embargo, este objetivo podría lograrse en el futuro cercano gracias a las nuevas tecnologías emergentes. Prueba de ello es que en el campo

de la oncología ya se comercializan firmas de expresión génica para evaluar las posibilidades de recurrencia del cáncer de seno. Desafortunadamente, esta innovadora prueba de diagnóstico se basa en el análisis de microarrays de ADNc de muestras tumorales, por lo que solo puede ser realizada por personal de laboratorio experto en centros con infraestructura moderna

Un buen ejemplo de estas nuevas tecnologías que permiten la detección sencilla. sensible y cualitativa de la expresión génica es el Oxford Nanopore MinION® podría ser una herramienta interesante para traducir la firma de ARN del huésped del genoma completo en una prueba de diagnóstico de cabecera de rutina. Dado que MinION® es un secuenciador portátil, permite la adquisición de datos en tiempo real (no es necesario esperar a que finalice la ejecución para comenzar a analizar los datos) y el protocolo puede optimizarse para obtener el resultado aproximadamente seis horas. Además, ya existen dispositivos para la preparación de librerías que permitirían realizar las pruebas sin la necesidad de un laboratorio.

Independientemente de sus limitaciones, la presente tesis representa un paso adelante hacia el uso de firmas transcriptómicas en la práctica clínica. La aplicación de los biomarcadores ómicos encontrados en esta tesis en una prueba clínica para diagnóstico o pronóstico necesita validaciones adicionales. Incluyendo el diseño de estudios clínicos completos para evaluar escenarios, como diferentes severidades, puntos temporales en el curso de la enfermedad infecciosa, infecciones parasitarias, otras enfermedades inflamatorias, etc.

Sería deseable que estos futuros estudios utilicen un enfoque holístico y combinen análisis moleculares, inmunológicos, cultivos celulares microbiológicos tradicionales, enfoques metagenómicos y firmas de genes del huésped, para así poder integrar la información del

patógeno, la respuesta inmune / transcriptómica del huésped y los síntomas clínicos de la enfermedad.

Esto nos permitirá, por un lado, ampliar el conocimiento de la patogenia de las infecciones, y, por otro lado, ayudará a estimar si un patógeno detectado en una prueba clínica es responsable de la patogénesis observada o si es solo una colonización inofensiva. Cabe destacar que las firmas de genes del huésped no están destinadas a reemplazar el diagnóstico basado en microbiología, sino que emergen como una herramienta complementaria para obtener más información. El desarrollo de pruebas de diagnóstico / pronóstico basadas tanto en el patógeno como en la respuesta del huésped podría revolucionar el tratamiento de pacientes con sospecha de sepsis, fiebre de etiología incierta y también ayudar a distinguir a los pacientes con un mayor riesgo de desarrollar una enfermedad infecciosa grave o invasiva que permita tratamiento farmacológico temprano y / o aumento de la vigilancia del paciente.

El objetivo final debería ser desentrañar el papel de estos biomarcadores y sus rutas metabólicas asociadas ya que podrían ayudar a desentrañar los mecanismos críticos en la defensa del huésped contra patógenos específicos. Lo que también ayudará a desarrollar nuevos enfoques terapéuticos.

Aunque todavía hay numerosas adversidades que superar antes de que las firmas de expresión génica del huésped puedan introducirse en la rutina del diagnóstico molecular. Las firmas basadas en biomarcadores de expresión génica en sangre del huésped tienen un gran potencial para el diagnóstico de enfermedades infecciosas, y probablemente pronto veremos sus primeras aplicaciones clínicas. Como la mayoría de los estudios de firmas transcriptómicas se han basado en analizar ARN a partir de sangre periférica, es probable que las primeras pruebas que se comercialicen utilicen este tipo de muestra, pero sería interesante también evaluar el uso de otras

muestras menos invasivas como saliva o hisopos nasofaríngeos especialmente en contextos pediátricos.

En los próximos años, la comunidad científica probablemente construirá una biblioteca de firmas genéticas para todas las condiciones y patógenos comunes. Lo que, en paralelo con el desarrollo de nuevas tecnologías que puedan determinar de forma rápida y precisa la expresión génica de un pequeño número de genes, conduciría a un diagnóstico más rápido y a reducir el mal uso de antibióticos. Las firmas transcriptómicas tienen el potencial de permitir obtener un diagnóstico antes de la aparición de los primeros síntomas de la enfermedad, el diagnóstico rápido de enfermedades infecciosas no solo mejora los resultados de los pacientes, sino que ayuda a retrasar la transmisión paciente-paciente o que será de crucial para la especie humana en un mundo cada vez más globalizado como demostró la pandemia de SARS-CoV2 de 2019.

Finalmente, enumero las principales conclusiones extraídas de la investigación realizada en el presente doctorado:

1. Nuestro estudio de asociación de genoma completo (GWAS) encontró dos SNPS rs201967957 (gen *MEIS1*) y rs576099063 (gen *TSPAN15*) en el huésped asociadas con la neumonía neumocócica. Utilizando una prueba de carga de patogenicidad encontramos otros cuatro genes, a saber, *OR9G9*, *MUC6*, *MUC3A* y *APOB*, que acumulan variantes patogénicas en una proporción mayor que los controles. Al analizar varios repositorios de datos transcriptómicos, confirmamos que los genes *MEIS1*, *TSPAN15* y *APOBR* (que codifica el receptor de la proteína APOB) tienen un rol en el desarrollo de neumonía en modelos de ratón y humanos.

2. Nuestros resultados sugieren que los transcritos de los genes *IFI44L* y *FAM89A* son suficientes para diferenciar las infecciones bacterianas de las virales. La señal de estos genes no se ve afectada por ancestralidad de los pacientes, y es útil para discriminar una amplia gama de patógenos y diferentes niveles de gravedad. Además, de acuerdo con nuestro ensayo q-PCR, medir la expresión del gen *IFI44L* sería suficiente para discriminar entre enfermedades virales y bacterianas con una gran sensibilidad y especificidad.
3. Demostramos que los datos de expresión del huésped (RNA-seq o microarrays) pueden traducirse con éxito en un ensayo de qPCR in vitro rápido, altamente preciso y relativamente económico que podría implementarse en la rutina clínica sin un gran esfuerzo.
4. Demostramos que es posible inferir la ancestralidad de los pacientes a partir de lecturas RNA-seq, pero cabe destacar la precisión de los resultados depende de la calidad de las lecturas durante secuenciación.
5. Este trabajo de tesis supone un toque de atención para los estudios de expresión génica, acerca de la importancia de controlar y modelizar la ancestralidad de los pacientes para evitar efectos de confusión.
6. Identificamos una firma de 17 transcritos en el huésped específica para la infección por VRS comparando la huella transcriptómica de este virus contra otros virus respiratorios.
7. Encontramos dos RNAs largos no-codificantes vinculados a infecciones virales. Estas moléculas son biomarcadores prometedores (ENSG00000254680 y ENSG00000273149) por lo que patentamos su uso como biomarcadores virales.

8. Demostramos por primera vez el potencial de los RNAs largos no-codificantes como biomarcadores para el diagnóstico de infecciones virales humanas.
9. La vacuna contra el rotavirus induce cambios en el transcriptoma del huésped similares, pero no completamente equivalentes, a los que causa la infección natural en los niños. Alterando los patrones de expresión de genes asociadas con el ciclo celular, diarrea, náuseas, vómitos, invaginación intestinal y morfología anormal del intestino medio.
10. Encontramos una firma de 9 transcritos, que identifica con precisión a los niños que han sido vacunados de rotavirus, lo que puede ser de gran utilidad para identificar a los niños no vacunados o en los que la vacunación no haya proporcionado una adecuada nivel de protección.
11. Identificamos el micro ARN hsa-mir-149 que parece desempeñar un papel en la defensa del huésped contra los patógenos virales y puede ser una potencial diana terapéutica.
12. En conjunto, los resultados de esta tesis sugieren que cada enfermedad infecciosa está asociada con un patrón único de genes que se activan o desactivan generando una "firma o huella transcriptómica", que se puede utilizar para identificar patógenos y enfermedades.

1. Introduction

According to the World Health Organization (WHO), infectious diseases are still among the major causes child mortality and are responsible for a big number of medical visits and hospitalizations all around the globe[1, 2]. Until recently it was commonly accepted that the most of the casualties were caused by bacterial infectious diseases, but during the last decade, increasing evidence point towards the fact that viral infections are also responsible for significant morbidity and mortality in children [3].

Nowadays distinguishing between viral and bacterial infections is still a challenge, as the bacterial culture results can take some days and are often negative when the infection is located in unreachable sites [4] or children have received antibiotic treatment[5]. Therefore, most clinicians out of the fear of missing a life-threatening bacterial infection decide to admit febrile children to the hospital and administer antibiotics while awaiting the culture results [6]. Consequently, numerous viral infections are erroneously treated with antibiotics, contributing to the development of antibiotic-resistant bacteria [7]. Antibiotics have allowed humanity to live longer and healthier, but the antibiotic over-use, in conjunction with the absence of new generation anti-microbial drugs, is making a “post-antibiotic era”, in which common infections and minor injuries would once again kill, a likely scenario in the near future as stated by the WHO.

The development of PCR (Polymerase chain reaction) assays, together with microarrays, have noticeably increased our capability of accurate diagnosis old and emerging viral infections [8] while allowing the interrogation of multiple viruses in a single test [9].

Unfortunately, molecular assays have been less efficient in detecting bacterial infections especially those caused by invasive bacteria [10]. Furthermore, PCR might not tell us who is the primary causative agent, as it only indicates the presence of nucleic acids. But the detected pathogen might no longer be viable and its presence may attend to a recent but unrelated illness [11] or asymptomatic colonization. Therefore, improving our diagnostic tools is one of the biggest challenges in current healthcare.

Currently, the different omics approaches such as proteomics, metabolomics, genomics, epigenomics and transcriptomics are a powerful tool our knowledge of the relationship between the pathogens and the host also leading to the discovery of omic signatures [12].

Microarrays and RNA-seq allow studying the gene expression of many samples in a short period, producing highly dimensional datasets that require sophisticated bioinformatics pipelines and mathematical algorithms [12, 13]. RNA-seq is rising in popularity and is quickly substituting microarrays because it allows the discovery of new transcripts, interpret complex alternative splicing events, study allele-specific expression etc [12]. Furthermore, it also allows performing meta-transcriptomic studies and sequencing RNA molecules from distinct species that live in the same ecosystem at the same time [14].

In this thesis we focused on the study of gene expression signatures as the transcriptome can act as the "canary" of the genome, as it is a bridge between the information content within our genes and the phenotype. Whereas the genome of all the cells within the human body is the same the transcriptome is a dynamic layer of information that changes between cell types and the conditions of the organism. Therefore, quantifying the amount of messenger RNA (mRNA) helps to comprehend the cellular activity at a molecular level, then

facilitating the analysis of the patient response to viruses, bacteria and parasites. Thus, host transcriptomics approaches hold the potential to shed further light into the pathogenesis of infectious diseases and may enable new diagnosis approaches [12], Figure 1.

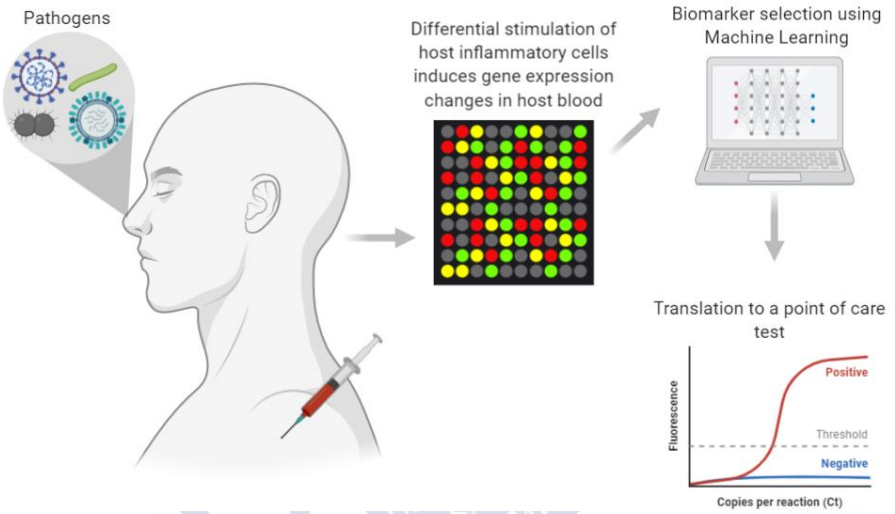


Figure 1. Different pathogens trigger unique host transcriptomic responses that can be evaluated from whole blood.

The following steps are the core of any bioinformatic pipeline for gene expression biomarker discovery: (1) data quality control and pre-processing, (2) biomarker selection (3) application of a prediction model (4) evaluation of the model performance[12]. Even Though the data pre-processing step for the RNA-seq data is far more complex and computationally demanding than for microarray data [13], once the expression data is normalized and pre-processed employing machine-learning algorithms is possible to identify relevant transcripts for prognosis and diagnosis.

As the number of molecules measured by microarrays and RNA-seq can be huge [13], using algorithms for data mining and feature

selection is a necessary step. Fortunately, there are many supervised machine learning algorithms that can be used for biomarker selection (lasso, elastic net, support vector machine, etc) and the computational capability to address this problem [15, 16]. The main objective is to find a reduced number of transcripts [5, 12] with high specificity and sensibility to be used in clinical settings. The most common way to evaluate the performance of diagnosing transcriptomic signature is to use Receiver operating characteristic (ROC) curves, together with positive and negative predictive values in a different cohort than the one used to train the algorithm.

The first ones to use this revolutionary approach were Zaas and colleagues [17] in 2009. They proposed a ground-breaking approach, instead of focusing on the infectious agent they focused on studying the host transcriptomic response and were first to evaluate the potential of host gene signatures for diagnosis of infections. After studying the blood transcriptomic profile of patients affected by three different respiratory viruses rhinovirus, respiratory syncytial virus (RSV) and influenza they found three transcriptomic signatures: 1) *OAS2*, *CXCL10*, and *SOCS1* for *HRV*, 2) *FCRGR1A*, *GBP1*, and *LAP3* for RSV, and *TNFAIP1*, *IFI27*, and *SEPT4* for influenza) able to distinguish between healthy individuals and infected patients [17].

From 2009 to present-day many research groups focused on the study in mRNA host response signatures for acute infections diagnosis [5, 18-25]; Figure 2.

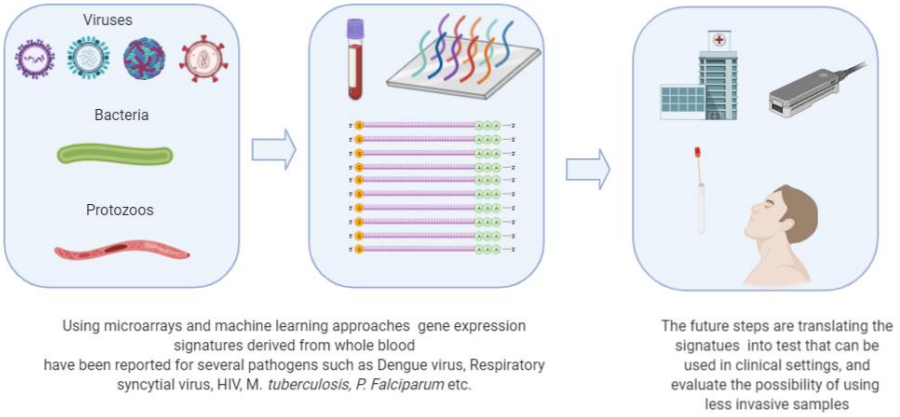


Figure 2. State of art of the art of transcriptomic signatures and future steps.

Taking together these results, on the one hand, confirm the enormous potential of host transcriptomic signatures for the diagnosis of infectious and inflammatory diseases, and on the other hand, support the idea that the different species of viruses and bacteria trigger a characteristic host immune response leading to particular gene expression patterns [10].

It deserves attention the fact that blood signatures can reach high accuracy even for viral infections that are believed to be restricted to the gut or the respiratory tract [7, 26], which is good news as obtaining a blood sample is far less invasive than obtaining infected tissue, especially in a pediatric context.

As we discussed before, transcriptomic signatures have already proven their potential to diagnose infectious diseases, but further research into the host response to the different pathogens would allow broadening their scope and clinical applications. For example, correlating symptoms of a bad prognosis with changes in host

transcriptome would allow the design of severity signatures that might help clinicians to determine the priority of patient's treatments based on the severity of their condition or likelihood of recovery [27]. Which would be of particular interest when facing a pandemic such as the one caused by the SARS-CoV-2 in 2019 [28]. For the discovery of these severity signatures, it might be also interesting including measurements of the immune response, and microbiology tests such as qPCR or cultures as it will help us to have a complete picture of the host response and its correlation with the symptoms of the disease. Furthermore, this combined approach would be of great interest when studying coinfections [10].

Here and now, the challenge is identifying and validating new the diagnostic and prognostic signatures for the most common human pathogens for which experimental design and clinical recruitment are key pieces. The most common approach to ensure reproducibility and validity of the biomarkers is using a training group for the algorithm and evaluate its performance in an independent validation group [12].

Nevertheless, careful patient recruitment is essential in both the training and discovery sets to avoid bias when the algorithm chooses the markers. It has been described that just a few false assignments in the discovery set might result in distorted biomarker identification [29].

Even though rigorous patient phenotyping increases the possibilities of training the model appropriately and find clinical relevant markers, this is not always easy as some diseases are complex and lack of gold standard tests as in the case of tuberculosis [20]. Furthermore, patient recruitment needs to be representative of the population including cofounder variables that might impact the results of study such as endemic infections (such as HIV), physical conditions (obesity), different groups of age, gender, ancestry [30] etc.

As these conditions might have an impact in the transcriptome and bias the results, so they should be considered in the algorithms.

Currently, using new parsimonious approaches [31] together with multi-cohort meta-analyses [19], scientists have been able to identify expression signatures using minimal numbers of transcripts, a key stepping stone [12] towards the translation of transcriptomic signatures into clinical diagnostic tests for disease progression and/or treatment effectiveness [32].

Gene expression signatures hold the potential to allow to obtain a diagnosis before the appearance of disease symptoms. Quick diagnosis of infectious diseases improves patient outcomes together with aiding to slow patient-patient transmission [12], which would be of crucial importance in scenarios such as the Covid-19 pandemic.

To turn gene expression signatures into a clinical test is necessary to use a technology that allows obtaining a reliable measurement of gene expression [12]. The good news is that the stage is already settled, as in the research area of oncology, gene expression signatures are being used to assess the chances of breast cancer recurrence [33]. Unfortunately, this innovative diagnostic test is based on microarray analysis of tumor sample cDNA, thus it can only be conducted by experts in laboratories with modern infrastructure [12].

In order to make transcriptomic signatures applicable in clinical settings two conditions need to be fulfilled: a) reducing its cost, and b) reduce the sample and data processing to the minimum in order to be relevant in clinical settings.

The easiest way would be using qPCR instead of microarrays or RNA-seq together with simple software that will just yield the predicted result according to the mRNA concentration the transcript signature. Unfortunately, the qPCR technique is required to be

performed in a fully equipped laboratory by expert laboratory technicians making it impossible to use the field.

The main advantage of NGS is that it would allow to simultaneously target the host and multiple pathogens in the same run. But, despite the fact that the cost and processing time required for NGS have dropped in the last decade, only a few of NGS technologies are suited for diagnosis in point of care settings, such as Ion Torrent Genexus System ® which allows obtaining results in a single day with a hands-off automated workflow. Another promising tool is Oxford Nanopore's MinION sequencer®, which allows for portable sequencing opening the door to the development of bedside NGS tests.

It must be remarked that transcriptome signatures do not aim to replace microbe focused diagnosis approaches, they will become a new source of information for managing the infected patients. The development of host-based signatures would be particularly useful for unravelling the etiology of febrile patients, or for distinguishing infectious diseases from autoimmune conditions [10, 34].

Up till now, no transcriptional signature has been transformed into a point-of-care test (Turner, Gupta et al. 2020). The findings summarized in the present thesis documents prove the viability of using machine learning algorithms for the discovery of bacterial and viral host markers and the feasibility to translate them to a qPCR test. This would facilitate its implementation, as most hospitals are already doing qPCR-based tests for many different conditions. Therefore, this thesis can be considered a stepping-stone towards the development of more transcriptomic specific signatures, by generating more data together mining the vast public repositories of transcriptomic data already available following machine learning approaches.

2. Hypothesis

This thesis has three main hypotheses that were evaluated in the research articles:

There is a genetic predisposition in humans to susceptibility and severity of infections. Not all individuals in close contact with pathogens get infected and develop the disease, in general, most patients show mild to moderate symptoms, and just a minority develop severe disease. We suspect that the genome influences a patient's likelihood to suffer from different infectious diseases including whether they become infected and how severe their outcome will be.

During Acute infections the transcriptome undergoes gene expression changes that are pathogen-specific. Therefore, the RNA from whole blood of patients with acute infections can be used to discover pathogen-specific genome-wide host expression signatures for diagnosis and prognosis.

Ancestry has an impact on gene expression and should be taken into account when looking for genome-wide host expression signatures as it could act as cofounding effect like in genome-wide association studies. We expect that variant calling from RNA—seq data should yield enough SNPs to infer a sample ancestry without the need of complementary experiments

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3. Objectives

The objectives put forward in the different studies were

1. To develop a bioinformatic analysis pipeline applicable to gene expression microarray data and RNA-seq to discover differentially expressed genes
2. To employ machine learning algorithms to identify the smallest sets of biomarkers that would allow discriminating between phenotypically similar infections.
1. To assess the performance of the biomarker sets and validate their classificatory power in independent cohorts.
2. To compare the performance of our discovered biomarkers with other found by other researchers, as well as current diagnostic tests employed in the clinical routine.
3. To develop a bioinformatic analysis pipeline applicable to gene expression RNA-seq to infer SNPs and from them estimate the patient's ancestry.
4. Evaluate how cofounder effects, particularly the patient's ancestry, may affect gene expression.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4. Methods

In this section, I will broadly summarize the main steps followed in the current thesis. For a more detailed description of the methods, protocols, pipelines and tools please refer to the methods section of each article in the results section (Figure 3).

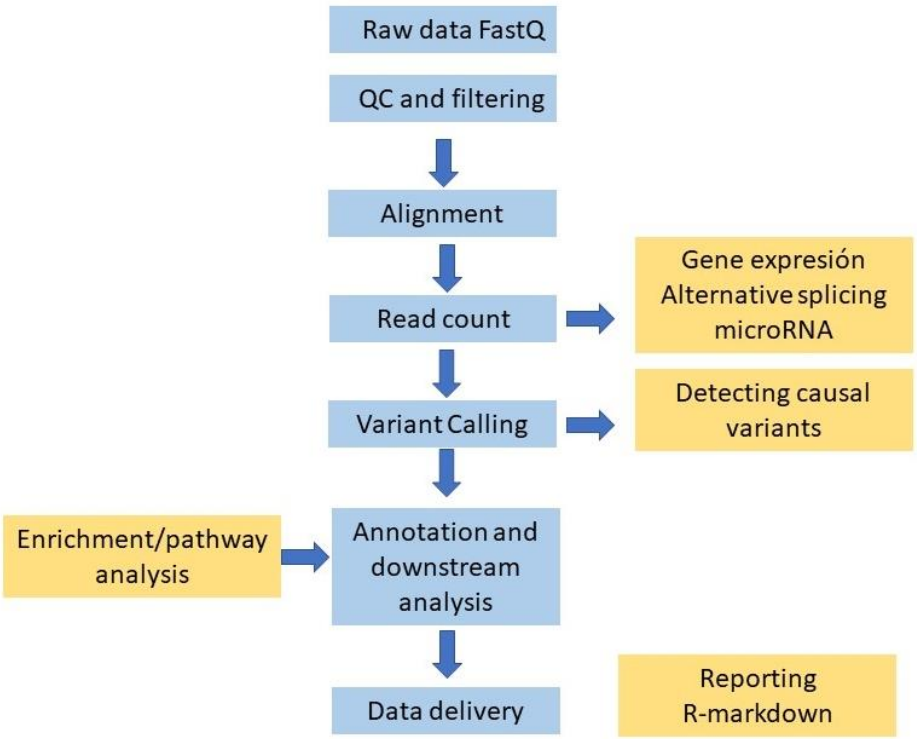


Figure 3. NGS data analysis overview

4.1 What is next-generation sequencing?

Next-generation sequencing [35], is a term that comprises a broad range of DNA sequencing technologies developed after the Human Genome Project (end on 2000) as an alternative method of the classical Sanger sequencing method. NGS has fully transformed the field of genetics. All these methods use massively parallel approaches to generate large amounts of reads from several samples at high throughput and coverage to increase the sequencing accuracy. The first NGS technologies were[36]:

- Illumina sequencing: which works by detecting the nucleotides fluorescent signal when they are added to the nucleotide chain.
- Roche 454 sequencing: which as the previous method is based on fluorescence, but it detects the pyrophosphate release when nucleotides are incorporated by polymerase to the nucleotide chain.
- Ion Torrent sequencing: which differs from the previous two methods as it does not measure light, rather it measures pH changes. Ion Torrent works detecting the release of protons produced when a nucleotide is incorporated to the DNA strand.

Currently, many companies are offering different and new NGS approaches such as Thermo Fisher, Illumina, Oxford Nanopore, PacBiO, BGI etc. Thanks to these relatively new technologies researchers can sequence DNA and RNA much faster and cheaper compared to Sanger sequencing. For instance, nowadays is it possible to sequence a complete human genome in just a single day using NGS, whereas over a decade was required to deliver the final draft of the human genome using Sanger sequencing [37]

Therefore, in the last decades, NGS has become extremely popular being used by researchers all over the world, generating numerous revolutionary discoveries completely revolutionizing the genomic field. Nevertheless, it has several intrinsic limitations such as:

1. Cost: Even though the cost has dropped dramatically, it is still higher than other diagnostic techniques such as cell cultures or qPCR. Nevertheless, NGS use for molecular diagnosis in clinical routine is becoming more widespread, especially in hospitals and fertility clinics. Up to the point that some companies have started to offer direct-to-consumer genetic testing without the involvement of any health care provider.
2. NGS is complex, it requires skilled and trained personal to both perform the laboratory protocols and analyze the data which complicates its application in clinical settings.
3. Depending on the platform and the protocol the error rate of NGS can be high, therefore especially in clinical scenarios, Sanger sequencing needs to be used for confirmatory purposes [38].

4.2 NGS platforms

Several companies are producing NGS machines with different technologies properties, advantages and disadvantages (Table 1), but regardless of the chosen platform. NGS consists of sequencing the DNA (or cDNA) of as given and sample and transform the obtained signal into a text file containing “reads”. These “reads” are text lines that contain chunks of the targeted sequences and therefore composed of the characters AGTC (or N when sequencer cannot detect which nitrogenated base is present).

Platform	Illumina	Solid	ion Torrent	454 GS FLX	PacBio RS II	Oxford nanopore minION
Technology	Sequencing By Synthesis	Primer hybridization	Change in pH	Pyrosequencing	Single-molecule Real-Time	Single-molecule Real-Time
Read lenght (bases)	50- 250 bp	75 - 100 bp	200 bp	700 bp	3kb	10kb
Error Rate	~ 0,1%	~ 0,1	~ 1%	~ 0.1%	~ 13%	5-20%
Run Time	1 - 6 days	21 days	2 - 7 hours	10 - 24 hours	1 - 2 hours	1min-48h
Gb per run	15 - 1000	220	2 - 10	0.04-0.7	3	50
Pros	Most used High throughput platform	The two-base encoding provides inherent error correction	Less expensive equipment. Fast run times	Long reads fast run times	Longest read length. Fast run times	Long reads. Portable. Cheap equipment
Cons	Short reads. Expensive equipment	Long run times. Expensive equipment	Homopolymeric errors. Lower throughput	High reagent cost. Homopolymeric errors. Low throughput	very expensive equipment, high error rate	high error rate

Table 1. NGS platforms comparison adapted from [39]

4.3 NGS data preprocessing

Regardless of the sequencing technology used to sequence the patient's samples, the raw data is presented as unmapped reads (FastQ format) or mapped reads (BAM format). The first step before starting the data analysis, it to perform quality control of the samples and discard the quality outliers if present.

The following data preprocessing steps depend on the kind of study; for instance, for RNA-seq it would be necessary to perform the sequence alignment of the FastQ reads to the human reference to quantify the gene expression and data normalization. Whereas for genome-wide association studies the following steps would be the variant calling to retrieve the SNPs present in the samples and dep filter the obtained markers according to minimum allele frequency, Hardy-Weinberg equilibrium, call rate…

4.3.1 Quality control of FASTQ files

Although there are quite a few formats for NGS raw sequences (such as Oxford Nanopore ® FAST5), the most used and common one is the FastQ file. Which consist of a text file that contains the nucleotide sequence with the quality information of each nucleotide base in addition to other technical data.

Reads are represented by four lines on a FastQ file:

- Line1: Starts with an '@' followed by the sequence identifier and an optional description. The description contains the platform information and coordinates within the flow cell.
- Line2: This line contains the raw sequence letters.
- Line3: It begins with a + character. It is not mandatory. It may contain extra information.

- Line4: Corresponds to the quality values for each nucleotide of Line 2, it is encoded with ASCII characters.

Quality score boxplot (Figure 4) displays the distribution of base quality values as a box plot for each position in the read. The background color (green, orange or red) specifies ranges of high, medium and low qualities. The plot is offered only for FastQ files (BAM files usually contain base quality information, but reads contained in the BAM, which are aligned reads, are not necessarily a representative subset of all the sequenced reads).

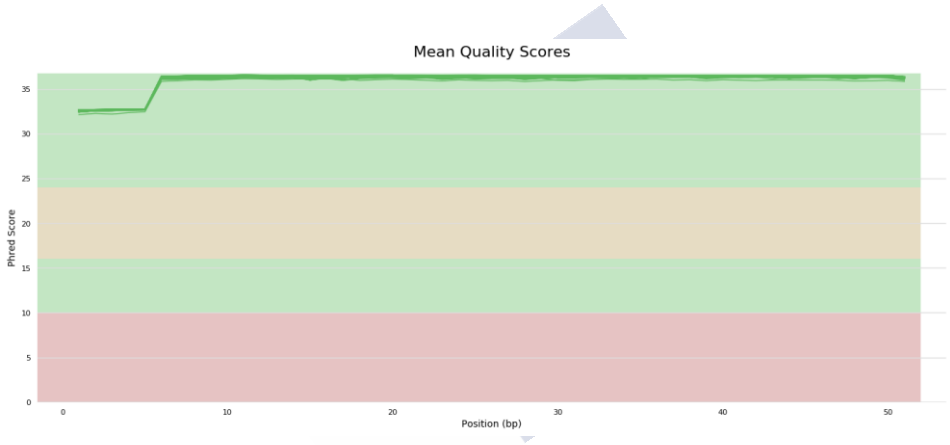


Figure 4. Quality score boxplot. In this example [40], all read positions have excellent quality (green area). It is common to observe a reduction of quality at the 3' end of the reads especially in long reads (>150bp), if the quality decreases till reaching the red region it is a common practice to truncate the 5' and 3' extremes of the reads.

Example of a FastQC Script

```
#!/bin/bash
#SBATCH -t 24:00:00
#SBATCH -n 24
#SBATCH -p thinnodes
module load fastqc/0.11.5
cd /home/usc/ap/rba/lustre/fastaQ ## the directory of our FastaQ files
fastqc./*.gz
```

Example of a MultiQC script

```
module load gcc/5.3.0
module load python/2.7.11
multiqc --force /home/usc/ap/rba/lustre/fastaQC
```

The software FastQC [41] performs an automated quality assessment, executing the most common quality control analysis. Another useful tool is multiQC [42] which aggregates the results from FastQC across many samples into a single report to easily visualize problematic samples.

Usually, the quality of sequencing drops towards the 3' end of the reads. Therefore, it is a common practice to remove trim low-quality 3' ends (below 20) from reads in addition to adapter removal. There are different tools to perform these tasks such as Cutadapt [43]or the R package QuasR [44]

4.4 Mapping NGS reads

After the FastQ files have been pre-processed, the reads have to be aligned (or mapped) against the human reference genome. There are different aligners freely available such as *TopHat* [45], *Bowtie2* [46] and *STAR* [47]. The aligner that has been used in the present thesis was STAR as it has been specially designed for RNA-seq data and it has the faster running time.

After the alignment is finished, STAR generates ".BAM" files containing the mapping information. The ".SAM" and ".BAM" files are the standard file formats for storing information on reads including read quality and how and where do they align to the reference genome. The ".BAM" files are the binary versions of ".SAM" files. Consequently, even though ".BAM" is not human-readable it is processed much faster and it is much less computationally demanding. The ".SAM" file has the same information but in a plain text format.

If the option --quantMode GeneCounts is chosen, STAR counts the number of reads per gene while mapping (note that a read is only counted if it overlaps just a single gene). The result of this counting process is stored in a text file per sample with the termination ".ReadsPerGene.out.tab" file containing four columns corresponding to:

- column 1: gene ID
- column 2: counts for unstranded RNA-seq
- column 3: counts for the 1st read strand aligned with RNA
- column 4: counts for the 2nd read strand aligned with RNA

Depending on the experiment that is being carried out, reads can be mapped to:

- Genes if the study focuses on gene expression.
- Exons if the study focuses on alternative splicing.
- Other features such as miRNAs, tRNAs, piRNAs, etc.

STAR example

Generating the reference genome Index

```
cd /home/ruth/STAR-2.7.0f

 STAR --runMode genomeGenerate --runThreadN 16

--genomeDir /home/ruth/star_rna --genomeFastaFiles
/home/ruth/star_rna/GRCh38.all.fa

 --sjdbGTFfile /home/ruth/star_rna/Homo_sapiens.GRCh38.96.gtf --sjdbOverhang 99
```

Example of a STAR scripts

Single End Reads

```
cd /home/ruth/STAR-2.7.0f/bin/Linux_x86_64
dirReads=/home/ruth/rota/hilde
for ff in `cd $dirReads; ls *.fastq.gz`
do
 echo $ff
 mkdir $ff
 cd $ff
 /home/ruth/STAR-2.7.0f/bin/Linux_x86_64/STAR --runThreadN 16 --
genomeDir /home/ruth/star_rna --readFilesIn $dirReads/$ff --
readFilesCommand zcat --runMode alignReads --quantMode GeneCounts --
outFilterType BySJout --alignSJoverhangMin 10 --alignSJDBoverhangMin 5 --
outSAMtype BAM SortedByCoordinate --outSAMattributes NH HI NM MD --
outReadsUnmapped Fastx --outBAMcompression 8 --twopassMode Basic --
outFileNamePrefix $ff
 cd ..
done
```

Paired End Reads

```
cd /home/usc/ap/rba/lustre/newfiles/bams/ASIA/INDIA
dirReads=/home/ruth/files
for ff in `cd $dirReads; ls *1.fastq.gz`
do
 base=${ff:0:10}
     var1="_1.fastq.gz"
     var2="_2.fastq.gz"
     file1=$base$var1
     file2=$base$var2
 echo $file1
     echo $file2
 mkdir $base
 cd $base
 /home/ruth/STAR-2.7.0f/bin/Linux_x86_64/STAR --runThreadN 24 --
genomeDir /home/ruth/star_rna --readFilesIn $dirReads/$file1 $dirReads/$file2
--runMode alignReads --quantMode GeneCounts --outFilterType BySJout --
alignSJoverhangMin 10 --alignSJDBoverhangMin 5 --outSAMtype BAM
SortedByCoordinate --outSAMattributes NH HI NM MD --outReadsUnmapped
Fastx --outBAMcompression 8 --twopassMode Basic --outFileNamePrefix $ff --
readFilesCommand zcat
 cd ..
done
```

4.5 Quality Control of BAM files

Occasionally, all the previous steps may be executed from the sequencing services and they directly provide the ".BAM" files. In this scenario or when downloading data from public repositories, it is crucial to ensure which genome and which version was used for the alignment, as well as the parameters employed.

Regardless of the way the ".BAM" files are obtained, there are different procedures to carry out the quality control of the aligned data, and to ensure that the alignment process was successful and that they are not biases in a way that may impact downstream analyses. The previously described tool MultiQC together with STAR provide nice and informative reports that help in the decision-making process especially when outliers are present.

Duplication level plots (Figure 5) computed for each sample display the proportion of reads with different duplication levels. In a diverse library, the majority of sequences will appear only once. Therefore a low level of duplication implies a high level of coverage of the sequence of interest, but a very high level of duplication would be probably caused by technical bias e.g. PCR over-amplification [41]. Nevertheless, due to the nature of RNA-seq data a certain level of duplication is expected [48]. Moreover, the most common sequences are shown. This could be useful to identify contaminations from another organism, over sequencing issues or fragments of adapters still attached to the reads.

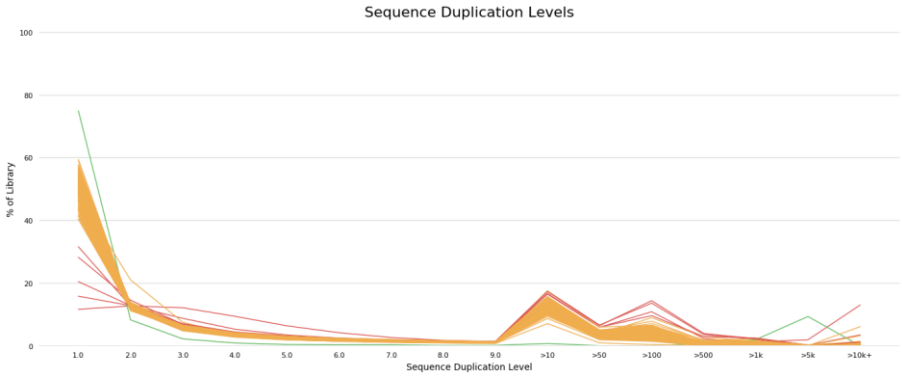


Figure 5. Plot of the relative level of duplication found for every sequence. The X-axis represents the proportion of duplicated reads whereas the Y-axis refers to the %of library.

Mapping statistics (Figure 6) displays the proportion of reads that were [42] mappable to the reference genome. It is useful to evaluate: a) library complexity as it shows the proportion of reads that map only to one locus, or multiple loci, b) the alignment efficiency if a sample has a higher percentage of unmapped reads it might need to be removed in downstream analysis. These statistics values depend on the library size (total number of reads in a library); therefore it is wiser to only make the comparison between libraries of similar sizes.

Unfortunately, it is becoming uncommon to find quality plots in the supplementary material on most biomedical omic studies. As inferred from our recent study [48], journals should encourage authors to include their quality plots (in the supplementary files).

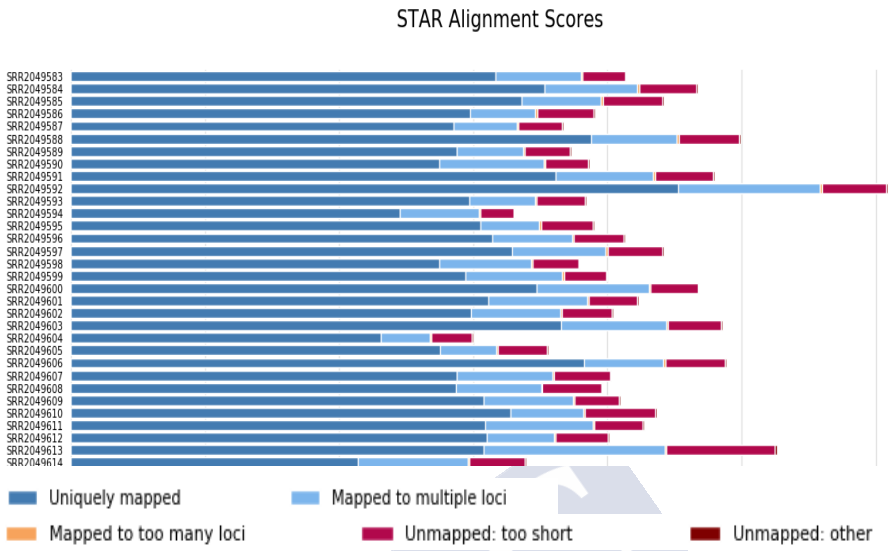


Figure 6. Mapping statistics summary

4.6 RNA-seq analysis obtaining differentially expressed genes

4.6.1 Summary

RNA-seq is a new method to perform expression analysis *via* high-throughput sequencing. It is the high-throughput sequencing counterpart of the gene expression microarray technology [49]. RNA-seq main advantage is that it gives a better estimation of gene expression than microarrays because the amount of data obtained is much higher as all of the RNA within a sample is sequenced (some kinds of RNA species such as miRNA may need special laboratory protocols to construct the library) rather than just those transcripts that are complementary to the microarray proves.

RNA-seq can be summarized in the following steps: 1) sequencing every RNA molecule in a sample, 2) counting the number of times each transcript has been sequenced. With current technologies it is not possible to sequence the whole transcriptome at once, therefore mRNA molecules are fragmented, converted to cDNA and amplified by PCR. After this step, the cDNA fragments are sequenced and the FastQ files are generated [13].

RNA-seq works by counting the number of reads (counts) that map a feature of interest (gene, tránscrito, exon, etc.). Consequently, RNA-seq data are discrete and do not follow a normal distribution, whereas microarray data are continuous and follows the normal distribution [50].

There are two main approaches when dealing with RNA-seq data:

1- Normalize the counts transforming the data into continuous data (using the R package *limma* and the function *voom*), and analyze it using microarray pipelines [50]. Even though this approach is popular especially for combining RNA-seq and microarrays data it has some inherent drawbacks:

- Microarray normalization implies taking logarithms to transform the data, but as some transcripts will likely have cero counts it will raise an error. Therefore, it is mandatory to add “pseudo counts” to the count data.
- There are other strategies for data transformation, but it is still necessary to treat RNA-seq data with the microarray algorithms designed to eliminate background noise and normalize. The main issue is that it is not possible to be sure if the assumptions made by microarray data make sense for RNA-seq data and also is complex to figure out if data transformation is successfully representing the original RNA-seq.

2- Taking this scenario into account, unless microarray and RNA-seq data are being compared in the same study, the vast majority of scientists prefer to analyze RNA-seq data applying statistical distribution for count data such as the negative binomial (Figure 7).

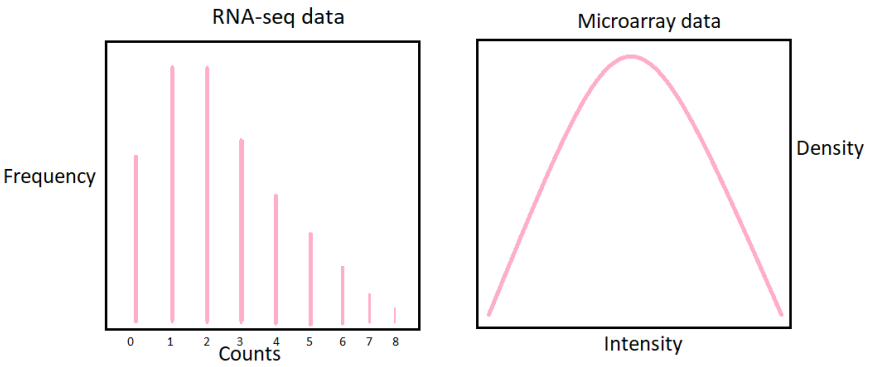


Figure 7. Expression level: microarray data vs RNA-seq data distribution

In summary, the first step is to obtain the number of counts; then it is necessary to normalize the data, between and within samples. The last step would be the differential expression analysis using the negative binomial distribution. Finally, a table of differentially expressed transcripts is obtained. With this list, there is further analysis that can be performed depending on the experimental objective such as pathway enrichment analysis, or test for over-representation [51] of microRNA-targets [52].

4.6.2 Replicates

As in all laboratory experiments, it will be common to have biological and/or technical replicates. The number of replicates that must be included in an RNA-seq experiment depends on two factors:

1) the biological variability of the measurements study which depends on the technical and the biological variability, and 2) the desired statistical power. Currently, the rule of the thumb is that at least three replicates should be employed for any inferential analysis [13].

Thus, for finding genes differentially expressed during a disease biological replicates in disease and controls are needed. Nevertheless, in some scenarios (e.g. when studying very rare diseases), this is often impossible. Therefore a few numbers of cases and controls are available, to raise the statistical power, the best strategy is to employ technical replicates. Considering for each biological replicate 2-3 technical replicates. But if the number of samples is adequate including technical replicates is not encouraged as in RNA-seq reproducibility between technical replicates is usually high (Spearman $r^2 > 0.9$) [13].

Careful design of RNA sequencing experiments is crucial to avoid technical biases and is as crucial as a good experimental design. Particularly when the study includes a large number of samples that need to be pre-processed in several batches and sequenced in different runs. No statistician neither bioinformatician regardless of their experience will be able to save a poorly designed experiment [13].

4.6.3 Normalization

The goal of normalization is to eliminate technical biases that may occur in the data to avoid that they have a negative impact on the final results. Normalization between samples is a necessary step because the number of read counts is associated with several factors, for instance:

- Sequencing depth: Is necessary to take care of the fact that the different samples might have different sequencing depth. As if this is not taken into account samples with higher sequence

reads would overestimate the gene expression level (more reads per gene will be expected in these samples) while samples with low sequence depth will underestimate the gene expression level (fewer reads per gene will be expected in these samples) [13].

- Gene length and mRNA expression level: the longer the gene, the more number of reads it will have in the library [53].
- The GC-content: Some sequencing technologies and library preparation protocols may cause an enrichment in GC rich sequences. The higher the GC content the higher the proportion of genes having further coverage. Therefore the GC-content has an important impact on gene expression measurements that is sample-specific and may increase the risk of having false positives in differentially expression analysis [54].

At present, there are many methods and tools for normalizing RNA-seq data, here we will discuss some of the more popular approaches:

Copies Per Million (CPM)

It is the easiest normalization approach; it consists of dividing the counts per gene by library size (the number of reads in the sample). The result is the proportion of reads mapping at each gene. It is common to multiply the proportion per one million to have copies per million.

Reads Per Million Mapped Reads (RPKM)

It resembles the CPM approach, but it also takes into account the gene length. Consequently, it needs a gene annotation file (gff3/gtf) to assess the length of each gene. It is important to take into account that only a small fraction of the genome is expressed in a given tissue or

during a certain condition. Therefore, it is expected that when comparing samples, the vast majority of transcripts would not be differentially expressed. Under this assumption, to evaluate if the normalization process has been successful, we could plot the distribution of log-ratios between the expression levels of the two samples log(RPKMs-sample1 / RPKMs-sample2), and the resulting distribution should be centred at 0. We should also expect that some ratios will not be centred at 0, because they will correspond to differentially expressed transcripts [55].

Trimmed Mean of M values (TMM)

The percentage of counts mapped to a certain gene in a library is not only dependent on the gene expression level even though it is affected by the expression features of the entire sample. The trimmed mean of M-values (TMM) is a method that has been created to solve this problem [53]. It is implemented in the *EDAseq* R package [56], the *tweeDEseq* package [57] and it tries to correct for library size, gene length and GC-content.

The method can be summarized as follows:

- Ygk: observed reads for gene g in library k summarized from the raw reads
- Gk: true and unknown number transcripts,
- Lg: gene g length
- Nk: total number of reads for library k
- Sk: total RNA output of a sample [58].

The expected value of Ygk is:

$$E[Ygk] = \frac{\mu gkLg}{Sk} Nk \; where \; Sk = \sum_{g=1}^{G} \mu gkLg$$

Equation 1. TMM formula

Conditional quantile normalization (CQN)

CQN normalization approach is based on an algorithm that uses a robust generalized regression that takes into account the library size, the gene length, and the GC-content. It is implemented in the *cqn* R package[54].

According to this approach, the number of counts (Y) given the mean value of a given gene of a certain sample could be modelled using the following function: Let Yi;j be the observed counts, let Xg = (Xg;1…..Xg;p) be the covariates (GC-content, gene length, etc), let hi be the technical variability whereas fi;j accounts for sample-dependent systematic bias which is modelled using cubic splines mi is the sequencing depth for each sample

$$Y_{g,i} | \mu_{g,i} = \sim Poisson(\mu g, i)$$

$$\mu_{g,i} = exp\left\{ h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}) + \log(m_i) \right\}$$

Equation 2. CQN formula

Relative Log Expression normalization (RLE)

It is implemented in the DESeq2 package. The method can be summarized as following: define Ygkr be the observed reads for gene g ∈ (1, …, G)f, in condition k ∈)1, …, K) for the number of replicates r ∈ (1, …, R)For a more detailed explanation of the algorithm please refer to [58, 59]

$$Y_g^{RLE} = \sqrt[KR]{(\pi_{k=1}^{K} \pi_{r=1}^{R} X_{gkr})}$$

$$\tau_{kf}^{RLE} = median_g(\frac{X_{gkr}}{Y_g^{RLE}})$$

Equation 3. RLE(DESeq2) formula

The first step is to calculate the geometric mean for each gene between all individuals. Then the counts per gene in each individual are divided by this mean. The size factor for an individual is calculated as the median of these ratios. This approach corrects for library size and RNA composition bias, which may appear for instance when only a reduced proportion of genes are very highly expressed in one condition but not in the others.

As discussed before is necessary to have biological replicates condition of study to accurately calculate the dispersion. If the number of biological replicates is small is very difficult to assess group variance reliably, therefore *DESeq2* uses shrinkage estimation for dispersions and fold changes. A dispersion value is calculated for each gene by fitting a model. If there are no biological replicates for one of the studied conditions, *DESeq* will measure the dispersion using the samples from the other conditions as replicates [59, 60]. According to

several studies, TMM and RLE yield virtually the same results with real and simulated data sets [58].

It is important to bear in mind that RNA-seq is still an evolving field, and every year new normalization methods and tools are being published. Here the most common ones were summarized, but new methods are being developed that might become popular in the future.

Differential Expression

In RNA-seq the differential expression analysis aims to find genes differentially expressed between phenotypical conditions from a statistical point of view [59]. As explained in previous sections it is crucial to have at least 3 biological replicates for those conditions (or technical replicates if biological are impossible to obtain)[13]. The differential expression analysis consists of comparing the mean value of genes across all the studied conditions.

As explained before, the main issue when analyzing RNA-seq is that the data do not follow a normal distribution, because it is count data (one, two, three times, etc). At the very beginning the number of counts mapped to gene, or feature of interest, was modelled using a Poisson distribution [61]. Poisson distribution is useful to describe variables are counted, but it has an important drawback: it assumes that mean and variance are the same. Therefore, the Poisson distribution could only be used if when analyzing count data when the mean and the variance are the same, which usually is a too strong/limiting assumption. As it has been described in real RNA-seq count data, several transcripts have a variance that is much higher than the mean. Consequently, the biological variability of RNA-seq data cannot be correctly described using the Poisson distribution [62]. This is the reason why currently most scientists use the Negative Binomial distribution (NB) which assumes the count data is over-dispersed (variance > mean) to model

RNA-seq data. The most popular algorithms designed to find differentially expressed genes use NB distribution to model the RNA-seq count data but they differ in the approaches used for measuring/ assessing the data overdispersion.[53, 59, 63].

Whereas Poisson distribution has only one parameter, the mean value of the number of counts that map to a gene, NB is a two-parameter distribution that models overdispersion. Overdispersion can be defined as the ratio between the mean and the variance when the variance and the mean are equal the overdispersion equals one, consequently in this particular scenario, the negative binomial and the Poisson would be the same [64, 65].

Enrichment/Pathway Analysis

Enrichment pathway analysis is usually the final step of any kind of NGS data analysis. The input for this analysis is a list of features of interest such as genes differentially expressed, genes associated with a phenotype, most expressed genes etc. There are several analysis and tools available (both free-to-use and private) to perform this kind of analysis once a obtained a list of genes of interest.

Functional Annotations databases

For this purpose the Bioconductor project [66] offers annotations packages ordered into the subsequent groups:

- Organism-level annotations: org.Hs.eg.db
- Microarray annotations: illuminaHumanv4.db, hgu133a.db etc.
- Gene-set annotations: GO.db, KEGG.db etc.

These annotation packages have the information stored as SQLite databases [67]. This information can be accessed via SQL commands

or using the R language as an interface that writes these statements instead of the user. Annotation packages are updated each time the Bioconductor software is updated every six months.

For illustrating how functional annotation using geneset databases is performed, I will focus on the famous Gene Ontology Project (http://www.geneontology.org) [68]. However, all the following step could be done using other databases such as The Kioto Encyclopedia of genes (KEGG)[69], Reactome[70], or Ingenuity Pathway Analysis ® (IPA)[71] etc. The analysis design would be exactly the same the only thing would change would be the database used.

GO is a major bioinformatics project whose aims is to unify the representation of gene and gene product attributes across all species[72]. This vocabulary is based on the called "GO terms", which are duos composed of a term identifier (GO ID) limma and a description:

- GO:0050727: Regulation of inflammatory response
- GO:0000016: Lactase activity
- GO:0050755: Chemokine metabolic process

Each GO term belongs to one of the next three ontologies: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Note that a certain gene product could be associated with more than one cellular component and participate in several biological processes. Therefore, each gene will be associated with several GO terms.

Overview of a functional analysis

After annotating the genes of interest to the chosen database it will be desirable to know if the genes of interest have any function in common or belong to the same pathways, which is the same as saying if there is an enrichment in any pathway or function.

The main statistical analysis to evaluate the existence of over/under-represented functions or pathways are 1)Enrichment, or overrepresentation which is by far the most used [73]. 2) The aggregate score also is known as GSEA-like [74]

Functional Enrichment or Over/Under-representation analysis.

In summary, given a group of genes of interest, the easiest way to determine if they are associated with any pathway is to apply a hypergeometric test. Which will estimate the probability that the proportion of genes belonging to a given function/pathway within the genes of interest list is statistically significantly higher than the proportion expected by chance.

In this scenario, the data will follow a discrete distribution as its obtained from taking data from a basket without replacement which corresponds to a binomial distribution. Though, as there is a limited number of genes for each pathway, the data instead of following a binomial distribution will follow a hypergeometric distribution[75]. The hypergeometric distribution is a discrete probability distribution that describes the number of successes n draws from a finite population without replacement, whereas the binomial distribution describes the number of successes for n draws with replacement.

4.7 Variant Calling

The variant calling analysis is used for finding genomic variants such as single nucleotide polymorphisms (SNPs) or Copy Number Variations (CNVs) between the samples of our study and the reference genome. Consequently, the first step would be to get the DNA-seq

reads aligned to a reference genome, as described previously, using a DNA-seq aligner such as Bowtie[46, 76].

Then use a variant caller algorithm such as Gatk [77] or Platypus (https://github.com/andyrimmer/Platypus). The inputs of variant caller algorithms usually are .BAM files, whereas the output will be a variant call format (vcf) file containing the detected variants. Finally using a reference gene database and different tools such as Annovar [78], Oncotator [79], or Nirvana[80] we can obtain information about the discovered variants such as which genes are associated with them.

Example of a Platypus script

```
cd /home/ruth/rotavirus/filtered

python /home/ruth/rota/filtered/Platypus_0.8.1/Platypus.py
  callVariants
--bamFiles=listOfFiles.txt --
  refFile=/home/ruth/star_rna/GRCh38.all.fa
--filterDuplicates=0 --minMapQual=0 --minFlank=0 --
  maxReadLength=500
 --minGoodQualBases=10 --minBaseQual=20 --nCPU=16 --
  output=rota.GRCh38.vcf
```

4.8 Microarrays

Microarrays consist of a predefined library of synthetic nucleic acid molecules (known as probes or oligos) that are immobilized and spatially arranged in rows and columns on solid support[81].

Depending on the nature of the biomolecules there are different types of microarrays being the most common the DNA and RNA microarrays.

	DNA microarray	RNA microarray
Sample:	**DNA**	**RNA**
Uses	• One sample • Paired samples: Case/control • Familial studies • Genome-Wide Association Studies (GWAS)	• Gene expression profiling studies • Discovery of biomarkers involved in biological processes • Detection of genes and pathways involved in diseases and pharmacological treatments • Pharmacogenomics and toxicogenomics studies • Dose-effect studies • Classification of samples based on gene signatures • Predictive models associated with genes
What information yield?	• Gained/lost genomic region • Molecular Karyotype • Loss of heterozygosity • Uniparental disomies detection • Mosaicism • SNP Genotyping	• Differentially Expressed Genes between conditions • Functional ENRICHMENT for each comparison (GO, IPA, KEGG, etc) • Alternative splicing

Table 2. Comparison of DNA vs RNA microarrays

Microarrays are a versatile technique as there are many kinds of studies that can be performed using this technology such as mRNA expression, miRNA, DNA copy number variation, single nucleotide polymorphisms, Methylation, DNA-protein interaction etc. (Table 2).

A general pipeline of an RNA microarray can be summarized in the following steps [81]:

1. Isolate and purify mRNA from samples

2. Reverse transcription: This step consists of performing a reverse transcription of the mRNAs to obtain complementary DNA strand (cDNA) while incorporating a fluorescent dye linked to a DNA nucleotide, obtaining a fluorescent labelled cDNA strand. If Cases and controls samples are labelled with different colour dyes the can be hybridized in the same microarray.
3. Hybridization: Labelled cDNAs are placed on a DNA microarray where they will hybridize to the complementary DNA. Non-hybridized cDNAs are washed away.
4. Scanning: The hybridized fluorescent labelled cDNAs are excited with a laser beam to produce a signal. The intensity of the fluorescent signal correlates with the number of cDNAs hybridized to the probes. In other words, the fluorescence intensity correlates to the expression level in the sample. The microarray vendor software detects, quantify and transform the signal into proportions that can be bioinformatically analyzed.

A detailed explanation of how microarray data analysis was conducted in the current thesis can be found in the articles method section of the results. But in a few words: the first step is to analyze the image of the array to see if there has been an error in the scanner. After that, a quality control analysis is performed to check all the parameters such as internal control probes etc. If no problems are found, the next step is the normalization of the microarray. After this step the analysis pipeline is different for DNA and RNA arrays:

In genomic array after the normalization step, different pipelines can be followed depending on the feature of interest: copy number, mosaicism, genotyping... Whereas in expression arrays after a previous step of filtering there are other pipelines to be followed such as differentially expression analysis, alternative splicing, enrichment

analysis etc. Whenever possible is particularly interesting to integrate both kinds of arrays to enhance the possibility of discovering results with biological meaning [82].

For microarray preprocessing and analysis there are many commercial tools and R packages, for instance, some popular R packages are Affy[83] for preprocessing Affymetrix® arrays and *Lumi* for preprocessing Illumina® microarrays [84].In the present thesis, we used *Limma* [50] for the detection of differentially expressed genes measured with microarrays between groups. *Limma* is an R package that combines a moderated t-test statistics, an empirical Bayes approach together with linear models [85] for analysing data from gene expression experiments including microarrays and RNA-seq for a detailed explanation of *limma* algorithm refers to Richie et. al article [50]

The first microarray was invented in 1995 [86] but is not until the appearance of commercial arrays together with the sequence of the human genome that their use becomes widespread. Due to recent developments and an exponential reduction in sequencing costs, the use of NGS has skyrocketed. Having in mind this scenario, it may look like microarrays are an obsolete technique but nothing could be further from reality, as proven by the popularity of recently developed array based technologies such as Nanostring® [87, 88]. Currently, microarrays are cheaper than NGS, especially if the cost of analysis is taken into account because data management and analysis are far more complex for NGS than for microarray, up to the point that it typically requires the use of a computational cluster. On the contrary, microarrays data analysis can usually be performed in user-oriented computers. Besides, the rapid and continuous evolution NGS technologies and pipelines make many scientists feel like NGS a less mature approach than microarrays. Even though they are a powerful tool, microarrays have drawbacks being the main one that their design

requires *a priori* knowledge of the reference genome sequence or have a prior hypothesis about which genes to include. Furthermore like NGS microarrays may have reproducibility issues and the results should be validated using Sanger sequencing or qPCR[89] [35].

4.9 Variable subset selection

When there are many transcripts in a prediction model, model selection algorithms allow to choose automatically the best combination of transcripts for constructing an optimal prediction model. Eliminating non-relevant transcripts helps to find a simpler and easy to understand the model. When the performance is the same, simpler models are always preferred over complex ones.

Furthermore, the use of variable subset selection algorithms is crucial when dealing with omic data, as the main challenge comes from the fact that (p) the number of variables (genes, transcripts, proteins etc) is much larger than the number of samples (n). This problem is called the “dimensionality course”.

It has been described that when p>>n it is relatively straightforward to find genes that perform great on the data set used to train the model but fail miserably when external validation is conducted. Leading to poor prediction models. Moreover, there can be a lot of variability in the least-squares fit, generating overfitting and subsequently bad predictions on future observations not included in the training data set.

The conceptually easiest strategy consists of evaluating all the possible combination of genes and then picking the best model. This method called best subsets regression is computationally demanding which makes it unfeasible when dealing with many predictor variables.

Consequently, a better approach for omic data is the Step Wise regression, which consists of adding and removing genes to find the model with the best performance using a reduced set of genes. Other suitable methods for high dimensional omic data are penalized regression (*ridge*, *lasso*, *elastic net*, etc) and the principal component-based regression methods principal component regression (PCR) and partial least squares (PLS) [90].

4.9.1 Standard Linear Model / Ordinary Least Squares Method

Assuming a linear model of n samples and p predictors (genes), the equation of a simple model would like this:

$$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$$

Let Y be the dependent variable (phenotype)

Let β the vector of coefficient estimates for different independent variables X.

The fitting procedure needs a loss function which is called residual sum of squares (RSS). The coefficients of the model are calculated in order to minimize this loss function.

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \ x_{ij})^2$$

Equation 4. Residual Sum of Squares Formula

The coefficients will be adjusted using the training dataset. If there is noise in the training data, then the estimated coefficients won't

generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

The standard linear model (equation 4) performs poorly when dealing with high dimensional data p>>n [90].

4.9.2 Penalized Regression/Shrinkage Methods /Regulation methods

A more suitable alternative for dealing with omic data is the penalized regression as it allows to obtain a model that is penalized for having too many genes in the model (the lesser number of genes the easiest to translate a transcriptomic signature into a clinical test) by adding a penalty in the algorithm.

This penalty reduces the values of the coefficients of the less informative genes towards zero or equal zero. This shrinkage needs the selection of a regulation parameter called lambda that fixes the quantity of penalty. In the following subsections, we will summarize the penalty regression methods employed in this thesis [90] [91].

4.9.2.1 Ridge Regression

Ridge regression reduces the regression coefficients of the less informative genes towards zero. The reduction of the coefficients is obtained by penalizing the regression model with the sum of the square coefficients. The amount of penalty can be established using a constant called lambda λ. When lambda equals zero the penalty parameter has no effect and ridge regression will behave as the least square method. Therefore, the higher the lambda the higher the impact of the shrinkage on the coefficients of the model.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j\ x_{ij})^{\,2} + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

Equation 5. Ridge Regression Formula

The above formula corresponds to ridge regression, where the RSS has been changed by adding the quantity of shrinkage. The coefficients are calculated by minimizing this function. λ corresponds to the tuning parameter that chooses how much the flexibility of the model is penalized. The increase in flexibility corresponds to the increase of its coefficients.

The main advantage of this regression approach compared to the previous one is that it performs reasonably well with high dimensional omic data (p>>n). Nevertheless, it has an important disadvantage that made this approach unfeasible for reaching our objectives, as it yields extremely complex models as a result because it keeps all the genes in the final model. Ridge regression reduces the coefficients towards zero, but any of them will be zero. The lasso approach is an alternative that avoids this problem [90].

4.9.2.2 Least Absolute Shrinkage and Selection Operator regression (Lasso regression)

Lasso is another regression model which is based on minimizing the function:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j\ x_{ij})^{\,2} + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\left|\beta_j^2\right|.$$

Equation 6. Lasso Regression Formula

It reduces the regression coefficients towards zero by penalizing the regression model with the sum of the absolute coefficients. The advantage of the lasso compared to the ridge regression is that the penalty causes that genes with a minimum contribution to the model have a coefficient equal to zero. Which means that lasso allows on the one hand to shrink the model coefficients and on the other hand to perform variable selection. As a consequence lasso yield as result models simpler and easier to interpret models, that incorporate only a fraction of the genes, than the ridge regression [90, 91].

4.9.2.3 Elastic Net:

Elastic net produces a regression model that is penalized with both 1) sum of the square coefficients and 2) the sum of the absolute coefficients.

$$\beta\hat{}^{enet} = argmin\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j\ x_{ij}))^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right\}$$

Equation 7. Elastic Net Formula

When compared to the other methods one of the advantages of the elastic net, is that it has the capability to yield more than n non-zero coefficients. This is particularly interesting when dealing with omic data as it helps to avoid problems due to n>>p [90, 91].

4.9.2.3 Parallel Regularised Regression Model Search (PREMS)

PReMS is particularly interesting methods for discovering small biomarker signatures as it has been designed to estimate small prediction models.

$$\theta_i = \int p(y_i \mid \beta)\, p(\beta \mid \tau, D) d\beta = \frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \beta_s^*)$$

Equation 8. Parallel Regularised Regression Model Search (PREMS) Formula

When given a group of biomarker candidates PReMS looks through numerous logistic regression models built from optimal subgroups of the candidate biomarkers, iteratively increasing the model size. Similarly, to elastic net the method estimates both the best shrinkage coefficient parameter and well as the optimal model size. The main advantage of PREMS over lasso and elastic net is that it tends to select smaller models, normally without compromising much the model performance in terms of AUC [92].

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. Results

This thesis has been written following the article compendium modality. Consequently, results are displayed as research articles. Six articles are presented as the main results and constitute the essential core of the thesis. Before each article, the quality indicators of the publications summarized. All the articles have undergone a peer-reviewing process and were published during the doctoral period.

The results have been divided into three broad blocks that correspond to three crucial aspects of genome-wide RNA biomarkers discovery:

1. Discovery of biomarkers
2. Validation of biomarkers
3. Study of confounding effects

Additionally, as complementary results other unpublished article is included. This article has been sent to journals, but as publication experienced a backlog due to the COVID19 outbreak, it is still undergoing a peer-reviewing process.

5.1 Article 1: Discovery of new biomarkers

Original title:

Whole exome sequencing identifies new host genomic susceptibility factors in empyema caused by streptococcus pneumoniae in children: a pilot study.

Authors:

Antonio Salas, Jacobo Pardo-Seco, **Ruth Barral-Arca**, Miriam Cebey-López, Alberto Gómez-Carballa, Irene Rivero-Calle, Sara Pischedda, María-José Currás-Tuala, Jorge Amigo, José Gómez-Rial, Federico Martinón-Torres

Identification of the article:

https://doi.org/10.3390/genes9050240

Identification of the journal:

Genes

ISSN: 2073-4425

Current Impact factor: 3.331

Quartil / Research area: Q2 / 'Genetics & Heredity'

Doctoral student contributions

In relation to the present article, the doctoral student took part in the data analysis and the writing process.

5.1.1 Evidence of Quality

Genes (ISSN 2073-4425; CODEN: GENEG9) is a peer-reviewed open-access journal of genetics and genomics published monthly

online by MDPI. Impact Factor: 3.331 (2018); 5-Year Impact Factor: 3.484 (2018)

Citations up to September 2019:

1. Salas, A. (2019). The natural selection that shapes our genomes. Forensic Science International: Genetics, 39, 57-60.
2. Szyroka, J. (2019). Regulation of the 'molecular scissor'ADAM10 by tetraspanin Tspan15 (Doctoral dissertation, University of Birmingham).
3. Martinón-Torres, F., Bosch, X., Rappuoli, R., Ladhani, S., Redondo, E., Vesikari, T.,... & Martín, C. (2019). TIPICO IX: report of the 9th interactive infectious disease workshop on infectious diseases and vaccines. Human vaccines & immunotherapeutics, 1-11.

5.1.2 Article abstract including the main results

Pneumonia is the leading cause of death amongst infectious diseases. *Streptococcus pneumoniae* is responsible for about 25% of pneumonia cases worldwide, and it is a major cause of childhood mortality.

We carried out a whole-exome sequencing (WES) study in eight patients with complicated cases of pneumococcal pneumonia (empyema). An initial assessment of the statistical association of WES variation with pneumonia was carried out using data from the 1000 Genomes Project (1000G) for the Iberian Peninsula [91] as reference controls. Pseudo-replication statistical analyses were carried out using different European control groups. Association tests pointed to single nucleotide polymorphism (SNP) rs201967957 (gene *MEIS1*; chromosome 2; p-value$_{IBS}$ = 3.71 × 10−13) and rs576099063 (gene

TSPAN15; chromosome 10; p-value$_{IBS}$ = 2.36 × 10−8) as the best candidate variants associated to pneumococcal pneumonia.

A burden gene test of pathogenicity signalled four genes, namely, *OR9G9, MUC6, MUC3A* and *APOB*, which carry significantly increased pathogenic variation when compared to controls.

By analysing various transcriptomic data repositories, we found strong supportive evidence for the role of *MEIS1*, *TSPAN15* and *APOBR* (encoding the receptor of the APOB protein) in pneumonia in mouse and human models. Furthermore, the association of the olfactory receptor gene *OR9G9* has recently been related to some viral infectious diseases, while the role of mucin genes (*MUC6* and *MUC3A*), encoding mucin glycoproteins, are well-known factors related to chronic obstructive airway disease.

WES emerges as a promising technique to disentangle the genetic basis of host genome susceptibility to infectious respiratory diseases.

Salas, A., Pardo-Seco, J., Barral-Arca, R., Cebey-López, M., Gómez-Carballa, A., Rivero-Calle, I.,... & Martinón-Torres, F. (2018).

Whole exome sequencing identifies new host genomic susceptibility factors in empyema caused by streptococcus pneumoniae in children: a pilot study.

Genes, 9(5), 240.

https://www.mdpi.com/2073-4425/9/5/240

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5.2 Article 2: Discovery of new biomarkers

Original title:

RNA-Seq Data-Mining Allows the Discovery of Two Long Non-Coding RNA Biomarkers of Viral Infection in Humans

Authors:

Barral-Arca, R., Gómez-Carballa, A., Cebey-López, M., Currás-Tuala, M. J., Pischedda, S., Viz-Lasheras, S.,... & Salas, A.

Identification of the article:

doi: 10.3390/ijms21082748.

Identification of the journal:

Int J Mol Sci

EISSN: 1422-0067,

Current Impact factor: 4.183

Doctoral student contributions

In relation to the present article, the doctoral student took part in the data analysis and the writing process.

5.2.1 Evidence of Quality

International Journal of Molecular Sciences (ISSN 1422-0067; CODEN: IJMCFK; ISSN 1661-6596 for printed edition) is an international peer-reviewed open access journal providing an advanced forum for biochemistry, molecular and cell biology, molecular biophysics, molecular medicine, and all aspects of molecular research

in chemistry, and is published semi-monthly online by MDPI. Impact Factor: 4.183 (2018)

5.2.2 Article abstract including the main results

There is a growing interest in unravelling gene expression mechanisms leading to viral host invasion and infection progression. Current findings reveal that long non-coding RNAs (lncRNAs) are implicated in the regulation of the immune system by influencing gene expression through a wide range of mechanisms. By mining whole-transcriptome shotgun sequencing (RNA-seq) data using machine learning approaches, we detected two lncRNAs (ENSG00000254680 and ENSG00000273149) that are downregulated in a wide range of viral infections and different cell types, including blood mononuclear cells, umbilical vein endothelial cells, and dermal fibroblasts. The efficiency of these two lncRNAs was positively validated in different viral phenotypic scenarios. These two lncRNAs showed a strong downregulation in virus-infected patients when compared to healthy control transcriptomes, indicating that these biomarkers are promising targets for infection diagnosis. To the best of our knowledge, this is the very first study using host lncRNAs biomarkers for the diagnosis of human viral infections.

5.2.3 European patent derived from this study

European Patent Office
80298 MUNICH
GERMANY

Questions about this communication ?
Contact Customer Services at www.epo.org/contact

BARRAL ARCA, Ruth
Universidade de Santiago de Compostela
EDIFICIO EMPRENDIA - CAMPUS VIDA
15782 SANTIAGO DE COMPOSTELA
ESPAGNE

Date
29.03.19

Reference	Application No./Patent No.
	19382084.2 - 1118
Applicant/Proprietor	
Universidade de Santiago de Compostela, et al	

Designation as inventor - communication under Rule 19(3) EPC

You have been designated as inventor in the above-mentioned European patent application. Below you will find the data contained in the designation of inventor and further data mentioned in Rule 143(1) EPC:

DATE OF FILING	: 05.02.19
PRIORITY	: / /
TITLE	: IN VITRO METHOD FOR THE DIAGNOSIS OF VIRAL INFECTIONS
DESIGNATED STATES	: AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

INVENTOR (PUBLISHED = 1, NOT PUBLISHED = 0):

1/BARRAL ARCA, Ruth/Universidade de Santiago de Compostela EDIFICIO EMPRENDIA - CAMPUS VIDA/15782 SANTIAGO DE COMPOSTELA/ES
1/MARTINÓN TORRES, Federico/SERVIZO GALEGO DE SAÚDE TRAVESA DA CHOUPANA, S/N/15706 SANTIAGO DE COMPOSTELA/ES
1/SALAS ELLACURIAGA, Antonio/Universidade de Santiago de Compostela EDIFICIO EMPRENDIA - CAMPUS VIDA/15782 SANTIAGO DE COMPOSTELA/ES

DECLARATION UNDER ARTICLE 81 EPC:
The applicant(s) has (have) acquired the right to the European patent as employer(s).

Receiving Section

EPO Form 1048 10.09

page 1 of 1

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Barral-Arca, R., Gómez-Carballa, A., Cebey-López, M., Currás-Tuala, M. J., Pischedda, S., Viz-Lasheras, S.,... & Salas, A. (2020).

RNA-Seq Data-Mining Allows the Discovery of Two Long Non-Coding RNA Biomarkers of Viral Infection in Humans.

International Journal of Molecular Sciences,

21(8), 2748

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7215422/

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5.3 Article 3: Discovery of new biomarkers

Original title:

A Meta-Analysis of Multiple Whole Blood Gene Expression Data Unveils a Diagnostic Host-Response Transcript Signature for Respiratory Syncytial Virus. Int. J. Mol. Sci. 2020, 21, 1831.

Authors:

Identification of the article:

doi: 10.3390/ijms21051831

Identification of the journal:

Int J Mol Sci

EISSN: 1422-0067,

Current Impact factor: 4.183

Doctoral student contributions

In relation to the present article, the doctoral student took part in the study design, data analysis and the writing process.

5.3.1 Evidence of quality.

International Journal of Molecular Sciences (ISSN 1422-0067; CODEN: IJMCFK; ISSN 1661-6596 for printed edition) is an international peer-reviewed open access journal providing an advanced forum for biochemistry, molecular and cell biology, molecular biophysics, molecular medicine, and all aspects of

molecular research in chemistry, and is published semi-monthly online by MDPI. Impact Factor: 4.183 (2018)

5.3.2 Article abstract including the main results

Respiratory syncytial virus (RSV) is one of the major causes of acute lower respiratory tract infection worldwide. The absence of a commercial vaccine and the limited success of current therapeutic strategies against RSV make further research necessary. We used a multi-cohort analysis approach to investigate host transcriptomic biomarkers and shed further light on the molecular mechanism underlying RSV-host interactions. We meta-analyzed seven transcriptome microarray studies from the public Gene Expression Omnibus (GEO) repository containing a total of 922 samples, including RSV, healthy controls, coronaviruses, enteroviruses, influenzas, rhinoviruses, and coinfections, from both adult and paediatric patients.

We identified > 1500 genes differentially expressed when comparing the transcriptomes of RSV-infected patients against healthy controls. Functional enrichment analysis showed several pathways significantly altered, including immunologic response mediated by RSV infection, pattern recognition receptors, cell cycle, and olfactory signaling. Besides, we identified a minimal 17-transcript host signature specific for RSV infection by comparing transcriptomic profiles against other respiratory viruses. These multi-genic signatures might help to investigate future drug targets against RSV infection.

Barral-Arca, R.; Gómez-Carballa, A.; Cebey-López, M.; Bello, X.; Martinón-Torres, F.; Salas, A.

A Meta-Analysis of Multiple Whole Blood Gene Expression Data Unveils a Diagnostic Host-Response Transcript Signature for Respiratory Syncytial Virus.

Int. J. Mol. Sci. 2020, 21, 1831.

https://www.mdpi.com/1422-0067/21/5/1831

5.4 Article 4: Validation of biomarkers

Original title:

A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms

Authors:

R. Barral-Arca, J. Pardo-Seco, F. Martinón-Torres & A. Salas

Identification of the article:

https://doi.org/10.1038/s41598-018-26239-1

Identification of the journal:

Scientific Reports

ISSN: 2045-2322

Current Impact factor: 4.011

Doctoral student contributions

In relation to the present article, the doctoral student took part in the study design, data analysis and the writing process.

5.4.1 Evidence of quality.

Scientific Reports is an open-access journal publishing original research from across all areas of the natural and clinical sciences. Scientific Reports is the 11th most-cited journal in the world, with more than 300,000 citations in 2018, and receives widespread attention in policy documents and the media. Scientific Reports is led by the same ethical and editorial policy guidelines as other Nature Research journals to ensure that all the research we publish is

scientifically robust, original, and of the highest quality. Journal metrics 2018: 2-year impact factor: 4.011,5-year impact factor: 4.525

5.4.2 Article abstract including the main results

Recently, a biomarker signature consisting of 2-transcript host RNAs was proposed for discriminating bacterial from viral infections in febrile children. We evaluated the performance of this signature in a different disease scenario, namely a cohort of Mexican children (n = 174) suffering from acute diarrhoea of different infectious etiologies.

We first examined the admixed background of the patients, indicating that most of them have a predominantly Native American genetic ancestry with a variable amount of European background (ranging from 0% to 57%).

The results confirm that the RNA test can discriminate between viral and bacterial causes of infection (t-test; P-value = 6.94×10−11; AUC = 80%; sensitivity: 68% [95% CI: 55%–79%]; specificity: 84% [95% CI: 78%–90%]), but the strength of the signal differs substantially depending on the causal pathogen, with the stronger signal being that of Shigella (P-value = 3.14 × 10−12; AUC = 89; sensitivity: 70% [95% CI: 57%–83%]; specificity: 100% [95% CI: 100%–100%]). The accuracy of this test improves significantly when excluding mild cases (P-value = 2.13 × 10−6; AUC = 85%; sensitivity: 79% [95% CI: 58%–95%]; specificity: 78% [95% CI: 65%–88%]).

The results broaden the scope of previous studies by incorporating different pathogens, variable levels of disease severity, and different ancestral background of patients, and add confirmatory support to the clinical utility of these 2-transcript biomarkers.

Barral-Arca, R., Pardo-Seco, J., Martinón-Torres, F., & Salas, A. (2018).

A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms.

Scientific Reports, 8(1), 1-7.

https://www.nature.com/articles/s41598-018-26239-1

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5.5 Article 5: Validation of biomarkers

Original title:

A qPCR expression assay of IFI44L gene differentiates viral from bacterial infections in febrile children

Authors:

Alberto Gómez-Carballa, Miriam Cebey-López, Jacobo Pardo-Seco, **Ruth Barral-Arca**, Irene Rivero-Calle, Sara Pischedda, María José Currás-Tuala, José Gómez-Rial, Francisco Barros, Federico Martinón-Torres & Antonio Salas

Identification of the article:

doi.org/10.1038/s41598-019-48162-9

Identification of the journal:

Scientific Reports

ISSN: 2045-2322

Current Impact factor: 4.011

Doctoral student contributions

In relation to the present article, the doctoral student took part in the data analysis

5.5.1 Evidence of quality.

Scientific Reports is an open-access journal publishing original research from across all areas of the natural and clinical sciences. Scientific Reports is the 11th most-cited journal in the world, with more than 300,000 citations in 2018, and receives widespread

attention in policy documents and the media. Scientific Reports is led by the same ethical and editorial policy guidelines as other Nature Research journals to ensure that all the research we publish is scientifically robust, original, and of the highest quality. Journal metrics 2018: 2-year impact factor: 4.011,5-year impact factor: 4.525

5.5.2 Article abstract including the main results

The diagnosis of bacterial infections in hospital settings is currently performed using bacterial culture from a sterile site, but they are lengthy and limited. Transcriptomic biomarkers are becoming promising tools for diagnosis with potential applicability in clinical settings. We evaluated an RT-qPCR assay for a 2-transcript host expression signature (*FAM89A* and *IFI44L* genes) inferred from microarray data that allow differentiating between viral and bacterial infection in febrile children. This assay was able to discriminate viral from bacterial infections (P-value = $1.04 \times 10{-}4$; AUC = 92.2%; sensitivity = 90.9%; specificity = 85.7%) and showed very high reproducibility regardless of the reference gene(s) used to normalize the data. Unexpectedly, the monogenic *IFI44L* expression signature yielded better results than those obtained from the 2-transcript test (P-value = $3.59 \times 10{-}5$; AUC = 94.1%; sensitivity = 90.9%; specificity = 92.8%).

We validated this *IFI44L* signature in previously published microarray and whole-transcriptome data from patients affected by different types of viral and bacterial infections, confirming that this gene alone differentiates between both groups, thus saving time, effort, and costs. Herein, we demonstrate that host expression microarray data can be successfully translated into a fast, highly accurate and relatively inexpensive in vitro assay that could be implemented in the clinical routine.

Gómez-Carballa, A., Cebey-López, M., Pardo-Seco, J., Barral-Arca, R., Rivero-Calle, I., Pischedda, S.,... & Salas, A. (2019).

A qPCR expression assay of IFI44L gene differentiates viral from bacterial infections in febrile children.

Scientific Reports, 9(1), 1-12.

https://www.nature.com/articles/s41598-019-48162-9

5.6 Article 6: Studying Confounding effects

Original title:

Ancestry patterns inferred from massive RNA-seq data

Authors:

Barral-Arca, R., Pardo-Seco, J., Bello, X., Martinon-Torres, F., & Salas, A

Identification of the article:

doi: 10.1261/rna.070052.118

Identification of the journal:

RNA

ISSN: 1469-9001

Current Impact factor: 3.63

Doctoral student contributions

In relation to the present article, the doctoral student took part in the study design, data analysis and the writing process.

5.6.1 Evidence of quality

The journal RNA established in 1995 serves as an international forum for publishing original reports on RNA research in the broadest sense. RNA is a monthly journal which provides rapid publication of significant original research in all areas of RNA structure and function in eukaryotic, prokaryotic, and viral systems. It covers a broad range of subjects in RNA research

5.6.2 Article abstract including the main results

There is a growing body of evidence suggesting that patterns of gene expression vary within and between human populations. However, the impact of this variation in human diseases has been poorly explored, in part owing to the lack of a standardized protocol to estimate biogeographical ancestry from gene expression studies.

Here we examine several studies that provide new solid evidence indicating that the ancestral background of individuals impacts gene expression patterns. Next, we test a procedure to infer genetic ancestry from RNA-seq data in 25 data sets where information on ethnicity was reported. Genome data of reference continental populations retrieved from The 1000 Genomes Project were used for comparisons. Remarkably, only eight out of 25 data sets passed FastQC default filters. We demonstrate that, for these eight population sets, the ancestral background of donors could be inferred very efficiently, even in data sets including samples with complex patterns of admixture (e.g., American-admixed populations). For most of the gene expression data sets of suboptimal quality, ancestral inference yielded odd patterns.

The present study thus brings a cautionary note for gene expression studies highlighting the importance to control for the potential confounding effect of ancestral genetic background.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6. Complementary results (non-published articles): Article 7 Discovery of new biomarkers

Original title:

Host transcriptomic response following administration of rotavirus 2 vaccine in infants' mimics wild type infection

Authors:

Barral-Arca R., Gómez-Carballa A., Cebey-López M., Currás-Tuala MJ., Pischedda S., Gómez-Rial J., Habgood-Coote D., Herberg J., Kaforou M., Martinón-Torres F., Salas A.

Doctoral student contributions

In relation to the present article, the doctoral student took part in the study design, data analysis and the writing process.

Abstract

Background. Rotavirus (RV) is an enteric pathogen that has a devastating impact on childhood morbidity and mortality worldwide. The immunologic mechanism underlying the protection achieved after RV vaccination is not yet fully understood. We compared the transcriptome of children affected by community-acquired RV infection, and children immunized with a live attenuated RV vaccine (RotaTeq®). RV vaccination mimics the wild type infection-causing similar changes in children's transcriptome, including transcripts associated with cell cycle, diarrhoea, nausea, vomiting,

intussusception, and abnormal morphology of midgut. A machine learning approach allowed to detect a combination of 9-transcripts that differentiates vaccinated from convalescent-naturally infected children (AUC: 0.9; 95%CI: 0.7–1) and distinguishes between acute-infected and healthy control children (in both cases, AUC: 1; 95%CI: 1–1).

We identified a miRNA hsa-mir-149 that seems to play a role in the host defense against viral pathogens and may have an antiviral role. Our findings might shed further light in the understanding of RV infection, its functional link to intussusception causes, as well as guide development of antiviral treatments and safer and more effective vaccines. The 9-transcript signature may constitute a marker of vaccine protection and helps to differentiate vaccinated from naturally infected or susceptible children.

Journal of Infectious Diseases – Major Article

Host transcriptomic response following administration of rotavirus vaccine in infants' mimics wild type infection

Ruth Barral-Arca[1,2,3], Alberto Gómez-Carballa[1,2,3], Miriam Cebey-López[1,2,3], Maria José Currás-Tuala[1,2,3], Sara Pischedda[1,2,3], José Gómez-Rial[1,3], Dominic Habgood-Coote[4], Jethro A. Herberg[4], Myrsini Kaforou[4], Federico Martinón-Torres[1,2*], Antonio Salas[1,3,*,#]

[1]Genetics, Vaccines and Pediatric Infectious Diseases Research Group (GENVIP), Instituto de Investigación Sanitaria de Santiago (IDIS) and Universidad de Santiago de Compostela (USC), Galicia, Spain

[2]Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, Galicia, Spain

[3]Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigaciones Sanitarias (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain

[4]Section of Pediatric Infectious Diseases, Imperial College London, London, United Kingdom

Running title: Transcriptomic response in rotavirus vaccinated infants

Abstract word count: 215; Text word count: 3495

*Equally contributed; #Corresponding author: Antonio Salas

Instituto de Investigación Sanitaria - IDIS

Planta -2, laboratorio de investigación nº19

Hospital Clínico Universitario de Santiago de Compostela (CHUS)

c/ A Choupana s.n. / 15706 Santiago de Compostela (SPAIN)

Email: antonio.salas@usc.es

Funding

This study received support from the Instituto de Salud Carlos III: project GePEM (Instituto de Salud Carlos III(ISCIII)/PI16/01478/Cofinanciado FEDER), DIAVIR (Instituto de Salud Carlos III(ISCIII)/DTS19/00049/Cofinanciado FEDER; Proyecto de Desarrollo Tecnológico en Salud) and Resvi-Omics (Instituto de Salud Carlos III(ISCIII)/PI19/01039/Cofinanciado FEDER) given to A.S.; and project ReSVinext (Instituto de Salud Carlos III(ISCIII)/PI16/01569/Cofinanciado FEDER), and Enterogen (Instituto de Salud Carlos III(ISCIII)/ PI19/01090/Cofinanciado FEDER) given to F.M.-T. D.H-C, J.H, and M.K. receive support from the NIHR Imperial College BRC. M.K. is funded by the Wellcome Trust (Sir Henry Wellcome Fellowship grant no. 206508/Z/17/Z)

Conflicts of interest

FM-T has received honoraria from GSK, Pfizer, Sanofi Pasteur, MSD, Seqirus, and Janssen for taking part in advisory boards and expert meetings, and for acting as speaker in congresses outside the scope of the submitted work. JGR has received honoraria from GSK, Pfizer and MSD for taking part in advisory boards and expert meetings, and for acting as speaker in congresses outside the scope of the submitted work. FM-T has also acted as principal investigator in RCTs of the above-mentioned companies as well as Ablynx, Regeneron, Roche, Abbott, Novavax, and MedImmune, with honoraria paid to his institution. The rest of the authors declare there are not commercial conflicts of interest in the present study.

Abstract

Background. Rotavirus (RV) is a enteric pathogen that has devastating impact on childhood morbidity and mortality worldwide. The immunologic mechanism underlaying the protection achieved after RV vaccination is not yet fully understood.

Methods. We compared the transcriptome of children affected by community-acquired RV infection, and children immunized with a live attenuated RV vaccine (RotaTeq®).

Results. RV vaccination mimics the wild type infection causing similar changes in children's transcriptome, including transcripts associated with cell cycle, diarrhea, nausea, vomiting, intussusception, and abnormal morphology of midgut. A machine learning approach allowed to detect a combination of 9-transcripts that differentiates vaccinated from convalescent-naturally infected children (AUC: 0.9; 95%CI: 0.7–1), and distinguishes between acute-infected and healthy control children (in both cases, AUC: 1; 95%CI: 1–1). We identified a miRNA hsa-mir-149 that seems to play a role in the host defense against viral pathogens and may have an antiviral role.

Discussion. Our findings might shed further light in the understanding of RV infection, its functional link to intussusception causes, as well as guide development of antiviral treatments and safer and more effective vaccines. The 9-transcript signature may constitute a marker of vaccine protection and helps to differentiate vaccinated from naturally infected or susceptible children.

Keywords: biomarkers, RNA-seq, transcriptomics, vaccination, miRNA, rotavirus, machine learning, intussusception, midgut morphology

Background

Infectious acute gastroenteritis is one of the major causes of hospitalization in children, with rotavirus (RV) being the most frequent etiologic agent in severe disease [1]. RV is also one of the leading causes of infant death in developing countries; it was estimated that RV was responsible for the death of more than 600.000 children per year worldwide before the introduction of vaccines, and 128.000 after the introduction of vaccines in children younger than five years [2; 3; 4]. As there are no antiviral therapies available, the treatment of RV infection is based on avoiding dehydration and replacing the electrolyte losses of affected children. The development and introduction of RV vaccines has resulted in significant fewer cases of severe gastroenteritis in those countries where RV vaccination is included in the routine schedule [5; 6].

Two different vaccines are licensed in Europe for the immunization against RV: (a) the live attenuated pentavalent human-bovine reassorted vaccine RotaTeq® (RV5, Merck and Co, Inc, Pennsylvania, USA), and (b) the live attenuated human vaccine Rotarix™ (RV1, GSK Biologicals, Rixensart, Belgium) [7]. RV5 is composed of a combination of five human/bovine reassorted RV that replicate poorly in the gut [3]. RV1 is made from a single human live attenuated strain, that replicates easily in the intestine [3; 7]. Both vaccines confer protection and have shown real-life effectiveness and impact, however the exact immunologic mechanism conferring protection against RV gastroenteritis is not fully understood [8]. The development of future RV vaccines or the improvement of current formulations is limited by our incomplete knowledge of the mechanisms responsible for RV pathogenesis and the host susceptibility [9]. Possible heterologous effects of RV vaccination are also the focus of attention (see [10; 11; 12]

and references therein). It has been recently reported that RV infection is able to provoke global changes in the transcriptome of infected cells to evade the innate host response; likewise, the host develops mechanisms to avoid viral invasion, including a strong inhibition of glycophorin genes [13].

Despite the importance of these interactions and the burden that RV means to human health, only a few human blood gene expression studies have been published to date [13; 14]; none of them have investigated how vaccines influence the blood transcriptome. There is therefore a lack of knowledge on how RV interacts with the host [13] and the mechanism that underlies the acquired immunity after RV vaccination.

To the best of our knowledge, this is the first transcriptomic investigation of RV vaccine response in whole blood, and we present a comparison of vaccinated infants *versus* wild type RV infected children, and age-matched healthy controls.

Methods

Samples and ethical approval

The Spanish cohort of 32 western-European children, prospectively collected between 2013-2014 at the Hospital Clínico Universitario of Santiago de Compostela (Galicia; Western Spain) (**Figure 1A**) comprised: (*i*) 6 healthy age-matched controls (with all the vaccines of the Spanish immunization schedule up to date but no rotavirus vaccine), (*ii*) 14 RV5 vaccinated infants, i.e., all the regular vaccines up to date plus three RV5 doses (RV5V group), and (*iii*) 6 RV infected children required medical attention due to moderate or severe symptomatology (RVinf group) at two different time-points, namely, acute (during medical attendance) and convalescent phases (40±10 days after clinical

recovery) (**Table S1**). A blood sample was obtained from these children using a PAXgene RNA tube (PreAnalytiX GmbH). Ages ranged from nearly 2 to 34 months (male/female ratio = 0.77). The mean time elapsed from hospital admission to blood collection in infected children was three days; whereas, in RV vaccinated children the blood sample was taken prior to vaccination and one month after the last RV5 dose. There were no remarkable clinical features in the individuals recruited. A subset of these controls and infected children were previously analyzed in [13].

All researchers were specifically trained in the study protocol for patient recruitment, sampling processing, and storage. The study was conducted following the principles of Good Clinical Practice and of the Declaration of Helsinki. Written informed consent was obtained from a parent or legal guardian for each subject before study inclusion. The project was approved by the Ethical Committee of Clinical Investigation of Galicia (CEIC ref. 2012/301).

Quality control of total RNA, libraries preparation and RNA-seq

We followed the same quality standards described before [13]. Briefly, Bioanalyzer 2100 and Qubit 2.0 were employed to evaluate the quality and the quantity of the collected RNA. We used GLOBINclearTM-Human Blood Globin Reduction Kit (Life Technologies; CA, USA) to eliminate globin mRNA and obtain a clearer signal from mRNAs from leukocytes. Poly(A)+ mRNA fraction was isolated from total RNA, and cDNA libraries were obtained following Illumina's recommendations. Equimolar pooling of the libraries was performed before clusters generation using *cbot* from Illumina. An Illumina HiSeq 2000 sequencer was used to sequence the pool of cDNA libraries using paired-end sequencing (100×2).

RNA-seq bioinformatic analysis

We first performed quality control of the raw data using FastQC [15]. Next, we used MultiQC [16] to ensure that no biases exist in the data which may affect the downstream analysis. Afterward, the whole transcriptome paired-end reads were mapped against the human genome provided by Ensembl v. GRCh37_r87/release 87 using the aligner STAR (https://github.com/alexdobin/STAR). We used *STAR* to count the number of reads that map to each gene. Using R v3.4.3 (http:/www.r-project.org), we computed Reads Per Million Mapped reads or RPKM [17] and TMM [18] implemented in the *edgeR* package [19], and Conditional Quantile Normalization or CQN [20] and *Deseq2* [21] using the library *tweeDEseq* package [22]. We finally chose the Deseq2 package as this package was also used to perform the downstream analysis.

The samples were analyzed for their ancestral background in Barral-Arca et al. [23] indicating their main European ancestry, then matching self-reported ethnicity.

We used the Negative Binomial distribution [18] implemented in the *DESeq* package together with the Surrogate Variable Analysis (SVA) method implemented in the *sva* R package to estimate differential expressed genes (DEG) and reduce batch effects. A generalized linear model was fitted in each cohort, and a t-statistic was calculated for each gene. *P*-values were corrected for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) approach.

Table S2 also contains the information regarding the genes differentially expressed between vaccinated and infected children.

Statistical analysis

We used Principal Component Analysis (PCA) to visualize the global transcriptome patterns of RNA-seq data and to identify outliers. PCA was carried out using the library *DESeq* R package.

We investigated the differentially expressed genes (DEG) for over-representation of common functions and/or pathways, using a hypergeometric test that calculates the probability that the proportion of genes within a given function/pathway might be found by chance within our selection of genes. We used two different public databases: (*i*) the Gene Ontology Project (GO; [24]), and (*ii*) the Kyoto Encyclopedia of Genes and Genomes or KEGG [25]. Ingenuity Pathway Analysis (IPA; https://www.qiagenbioinformatics.com/) tool was used to estimate the most significantly altered pathways and generate networks of biomarkers.

Among the DEG between RV5V and controls, we focused on those reported to be associated with intussusception according to the Disgenet database ([26]), namely: *STK11*, *PTEN* and *ARID1B*. We also investigated the *APC* gene as it was also reported to be associated to intussusception in the literature [27].

The R package *CORNA* [28] was used to test for significant associations between the genes differentially expressed in vaccinated children and microRNAs.

The Heatmaps of the genes associated with nausea, vomiting, and diarrhea according to Ingenuity® and the genes associated with hsa-mir-149 according to CORNA, were generated using hierarchical clustering and the R package *ggplot2*.

We used the Fisher enrichment test to detect enrichment in immunity-related genes within those DEG when comparing vaccinated-controls *vs.* wild type controls. The list of

immunity-related genes considered were those included in the nCounter® Immunology Panel (Human V2; https://www.nanostring.com).

We used a linear discriminant analysis to identify a transcript signature that distinguishes unvaccinated children from vaccinated children using Parallel Regularized Regression Model Search or PReMS [29]. The ability of the predicted model to discriminate vaccinated children was assessed by computing the Area Under the Curve (AUC), and the sensitivity, and the specificity at the optimal cutpoint according to the Youden index was calculated with the R package *Optimal Cutpoints* [30]. PReMS was initially built splitting the whole dataset into a training set (80% of the samples) and a test set (20% of the samples taken at random).

The performance of the proposed signatures was evaluated using Receiver Operating Characteristic (ROC) curves that represent the true positive rate (TPR) against the FPR at different threshold cutpoints. ROC curves were built in R using the package *pROC* [31].

The proportions of different cell types in peripheral blood may differ naturally, and in consequence, mRNA measurements can vary as well [32]. We used the Cell-type COmputational Differential Estimation (CellCODE) [33] method implemented in the R package of the same name, which assigns expression alterations to their cell type of origin with high accuracy, to analyze if there were any difference between the cell-type proportions in the blood of our three groups under study.

Results

RNA-seq results

To study the changes experienced in the transcriptome of RV5V and RVinf we performed large-scale expression screening using RNA-seq. A PCA of the whole transcriptome identified one outlier among the acutely infected children, which was eliminated from the followed-up analysis. After eliminating this outlier, the first principal component of the PCA (PC1; accounting for most of the variation, 73%; **Figure 1B**), shows two main clusters separating healthy controls from vaccinated children plus infected children, suggesting that both RV wild type and the vaccine attenuated virus modify the global transcriptome in a similar manner.

We obtained 9,503 DEG in the vaccinated *versus* controls comparison, and 8,958 in the infected children (RVinf acute phase and RVinf convalescent phase) *versus* controls (**Table S2**, **Figure 1C**). It is interesting to note that more than half (~52%; **Figure 1C**) of the DEG of vaccinated children against healthy controls overlap with those differentially expressed in infected children against healthy controls (**Figure 1C**).

Three out of four genes related to intussusception (*ARID1B, APC, PTEN, and STK11*) according to Disgenet were significantly differentially expressed between RV5V and controls (**Figure 2**). The three genes were up-regulated in the RV5V group: *ARID1B* (logFC: 0.76; *P*-value 2.1 x 10^{-11}), *PTEN* (logFC: 0.64; *P*-value = 3.7 x 10^{-5}), and *APC* (logFC: 1.32; *P*-value = 7.7 x 10^{-14}).

Pathway analysis

Analysis of differential regulation using IPA showed that many of the DEG in vaccinated *versus* healthy children were associated with gastrointestinal disease, inflammatory disease, organ injury and abnormalities (*P*-value = 3.3 x 10^{-4}; **Table S3; Figure 3**), including fecal incontinence (*P*-value = 3.1 x 10^{-3}; **Figure 3B**), diarrhea (*P*-value = 1.7 x

225 10^{-2}; **Figure 3C**), and nausea and vomiting (*P*-value = 2.7 x 10^{-2}; **Figure 3D**).
226 Furthermore, IPA also identified a statistically significant overexpression of pathways
227 and genes associated to the humoral immunity component of the adaptive immune
228 system which is responsible for secreting antibodies (**Figure 4**). This result is consistent
229 with the Fisher analysis showing that there is an enrichment in genes associated with
230 the immune system in both comparisons RV5V *vs*. controls (*P*-value [Fisher exact test] =
231 5.5×10^{-14}; OR = 2.12) and RVinf *vs*. controls (*P*-value [Fisher exact test] = 8.1×10^{-15};
232 OR = 2.20).

233 In addition, IPA also identified (**Figure S1**) the pathway "abnormal morphology of
234 midgut" (nine genes involved) as significantly enriched in RV5V *versus* controls (*P*-value
235 = 2.0×10^{-3}); the heatmap of **Figure S1** shows the differential expression of these
236 genes.

237 Enrichment of humoral immunity component of the adaptive immune system is also
238 present when comparing RVinf against controls (**Figure 4**).

239 Overall, the results suggest that the activation of the immune system produced by
240 the vaccine is comparable to the one caused by the wild type infection.

241 GO analysis indicates the overexpression of biomarkers associated with
242 gastrointestinal injury and abnormalities (**Table S4**), including bacterial invasion of the
243 epithelium (hsa05100, hsa05120) and a noticeable down-expression of genes
244 associated to cell-to-cell adhesion: GO:0007155, GO:0022610, GO:0016337, hsa04540,
245 hsa04530, hsa04520.

Furthermore, the pathway analysis results yielded by KEGG (**Table S5**) and GO ontology (**Table S4**) showed an enrichment of genes related to the regulation of cell cycle (hsa04110, GO:0051726, GO:0007049).

Cell Deconvolution

Cell deconvolution analysis indicates a statistically significant increase of B and T lymphocytes in vaccinated children compared to controls (CD4T: *P*-value = 8.4×10^{-5}; CD8T: *P*-value = 2.0×10^{-3}; B cells: *P*-value = 5.0×10^{-3}; **Figure S2**), in agreement with the IPA results. (**Table S3**;).

The results also indicate that the relative proportion of innate and adaptive immune cells of the infected against the vaccinated children is statistically significant in several cell types (**Figure S2**); in particular when comparing convalescent-infected children against vaccinated.

MiRNA enrichment analysis

The association test for over-representation of microRNA-target between vaccinated children and controls yielded one remarkable result: of a total of 9,503 DEG, there were a total of 216 genes (**Figure 5**) differentially expressed between these two groups that are targets of the microRNA hsa-mir-149 (*P*-value = 3.7×10^{-2}; expectation = 173; observations = 216).

A 9-transcript RNA signature to differentiate vaccinated *versus* unvaccinated

We used the PReMS algorithm [29] to create a signature able to distinguish between vaccinated and unvaccinated children (including healthy controls, RV acute and convalescent children). The algorithm found a 9-transcripts signature (**Table 1**) that allows to accurately separate these three classes (**Figure 6**). ROC curves indicate that our model classified correctly all the observations as the AUC was 100% for the training set and >90% for the test set (**Figure 6**). Convalescent children *vs.* vaccinated are the two classes that are more difficult to differentiate using this signature.

At the optimal cutpoint, and according to the Youden statistic (-1.9967), the sensitivity was 1; whereas the specificity was 0.97 with an AUC of 0.98 (0.94-1 IC) when comparing vaccinated against unvaccinated children in the whole dataset.

Discussion

RV vaccination causes global long-lasting changes in the transcriptome of peripheral blood cells, affecting the expression of more than 9,000 genes. Although the vast majority of children do not experience any adverse effects after vaccination [34], we found altered expression of biomarkers associated with vomit, diarrhea, fecal incontinence and nausea. This suggests that the vaccine actually mimics a mild version of the disease.

Due to the reported association of intussusception and earlier RV vaccines in the past (risk of 1.5 [95%CI: 0.2-3.2] with the first dose according to Yih et al. [35]), large safety studies were conducted on the current vaccines RV5 and RV1 before they were

approved. Nevertheless, the link between RV and intussusception remains unclear, up to the point that several studies have not found an increase in intussusception cases after administration of RotaTeq [36]. There is now a general agreement in the medical community indicating that the benefits of RV vaccination substantially surpass the low risk of intussusception that might be associated with vaccination [37]. We found that several DEG between RV5V, and control children have been reported to be associated with intussusception (**Figure 2)** and abnormal morphology of midgut (**Figure S1**); e.g. gene *APC*, that is up-regulated in RV5V and RVinf, has been described to play a role in the development of a jejunal adenoma causing intussusception [27]. This gene expression pattern may contribute to explain the reported increase of intussusception risk in vaccinated children. These genes could be targeted for the development of future safer vaccines and specifically analyzed in those children experiencing intussusception after vaccination.

Children vaccinated against RV over-expressed cell cycle related genes, this mechanism is used by many other viruses to facilitate their replication [38]. Transcription of these genes may be a consequence of the increase of B2 lymphocytes observed in RV5V children (**Figure S2**; **Table S3**). As RV5 is a live-attenuated vaccine whose viral particles replicate in the gut, these results are in good agreement with our previous findings indicating that the host cell cycle is affected by RV infection [13; 39].

Previous studies suggested that antibody-based responses are necessary for acute control of RV infection, and for immunological memory [40]. We found that vaccinated children have a significant increase in B cell proportion in peripheral blood (IPA analysis [**Table S3**]: *P*-value = 2.7×10^{-2}; cell deconvolution analysis [**Figure S2**]: *P*-value = 5.0×10^{-3}). This signal persists for a month after the last dose of the vaccine (the time the

sample was taken in vaccinated children), in concordance with the role of B cells in long term protection against RV reinfections. Several studies claim that both B cells and CD8+-T cells play an important role in long term protection against RV reinfection [40; 41; 42]. Consistently, our results also indicate that vaccinated children have higher levels of T cells (**Figure** S2**B** and S2**C**) compared to the healthy controls. Furthermore, vaccinated children express biomarkers associated with the differentiation of pre-T lymphocytes *CEBPA*, *MYH11*, *RAG2* and T cell receptor signaling (hsa04660) (**Table S3** and **S5**). Also interesting is the fact that in general the innate and adaptive response of convalescent infected children seem to be more remarkable that the response provoked by the vaccine (see B-cells and natural killer in **Figure S2**); this can be due to (*i*) the stronger impact on the immune system of the wild infection compared to the vaccine, and/or (*ii*) the fact that the sampling time point for convalescent is about 3.7 months while for vaccinated children is roughly 5.2 months. In this time period, we cannot discard the possibility of new infections among convalescents. It is expected however that such reinfections would modify the transcriptome in the same direction as the transcriptome of acute infected children; actually, this might be the case of one of the convalescent child in the PCA plot (see yellow dot within the cluster of infected children; **Figure 1B**).

Response to RV vaccination is also characterized by an over-expression of genes associated with gastrointestinal disease and inflammation (**Table S3**). RV5, like the RV, has a lytic cycle that burst epithelial cells to liberate the viral particles. Therefore, the presence of those biomarkers in vaccinated children possibly reflects that the intestinal barrier is being compromised due to the attenuated RV virus replication. This hypothesis is also supported by the fact that several pathways associated to cell-cell adhesion (e.g.

334 GO:0007155, GO:0016337, hsa04530, hsa04520, hsa04540) are significantly down-
335 regulated in the vaccinated cohort (**Table S4** and **S5**).

336 Bioinformatic miRNA target enrichment analysis showed that the expression levels
337 of >200 genes differentially expressed between RV5V and healthy controls (**Figure 5**)
338 can be explained by the regulatory effects of the miRNA hsa-mir-149. Hsa-mir-149 is
339 known to target the HIV gene *Vpr* [43], and also RV genes [44]. Most recently, it has
340 been described that hsa-mir-149 is able to significantly reduce polio replication within
341 host cells [45]. Further investigation of the relationship between RV and host mirna hsa-
342 mir-149 may elucidate mechanisms of RV pathogenesis.

343 While RV5 is an oral vaccine containing reassorted RV strains that replicates poorly
344 in the gut, we were able to see its effects in the blood transcriptome. This fact
345 strengthens the hypothesis that RV causes a systematic infection, rather than one
346 limited to the intestine [46; 47].

347 The PReMS method yielded a 9-transcript signature that distinguished vaccinated
348 and unvaccinated children with an accuracy ~90%. Although the signature shows a
349 good performance in the training and test sets, it would be necessary to validate this
350 signature in an external cohort of vaccinated children. A signature that identifies children
351 who have mounted a successful vaccine response might be of particular interest to
352 detect vaccine failures, to prevent severe RV reinfections, to perform epidemiological
353 control, and to evaluate immune response in the development of new RV vaccines.
354 While the number of transcripts might be too large for a ready to use qPCR assay [48],
355 other technologies would allow to easily test a 9-transcript panel that could be used for
356 epidemiological surveillance or vaccine research purposes.

A limitation of the present study is the limited number of subjects analyzed, even though the sample size lies within the standard range of transcriptome functional studies [49]. We analyzed the transcriptome of peripheral blood samples, away from the principal target of infection on the intestinal epithelium, and it would be of particular interest to compare the impact of RV vaccination on these different tissues.

To conclude, the response to RV vaccination is characterized by the over-expression of genes associated with gastrointestinal disease, inflammation, activation of the immune system and gene over-expression of the cell cycle. Although the alterations of the transcriptome caused by RV vaccination strongly resemble the ones caused by community-acquired disease, there are DEG that allow accurate discrimination of vaccinated and acute/convalescent infected children. Further research on these differences may help to unravel the molecular mechanisms of immune protection against RV, heterologous effects of the vaccine [50], and key features that allow the development of safer and more effective vaccines and novel antiviral drugs. Finally, we describe a 9-transcript signature/panel able to distinguish vaccinated children from unvaccinated, which may aid in the detection of vaccination failures.

Acknowledgments

We would like to acknowledge CESGA (Supercomputing Centre of Galicia, Santiago de Compostela, Spain) for its supercomputing availability, web hosting, and support.

Conflict of Interests

The authors declare that they have no competing of interest.

References

[1] T. Vesikari, Rotavirus vaccination: a concise review. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 18 Suppl 5 (2012) 57-63.

[2] J.T. Patton, Rotavirus diversity and evolution in the post-vaccine world. Discovery medicine 13 (2012) 85-97.

[3] D.I. Bernstein, Rotavirus overview. The Pediatric infectious disease journal 28 (2009) S50-3.

[4] C. Troeger, I.A. Khalil, P.C. Rao, S. Cao, B.F. Blacker, T. Ahmed, G. Armah, J.E. Bines, T.G. Brewer, D.V. Colombara, G. Kang, B.D. Kirkpatrick, C.D. Kirkwood, J.M. Mwenda, U.D. Parashar, W.A. Petri, Jr., M.S. Riddle, A.D. Steele, R.L. Thompson, J.L. Walson, J.W. Sanders, A.H. Mokdad, C.J.L. Murray, S.I. Hay, and R.C. Reiner, Jr., Rotavirus Vaccination and the Global Burden of Rotavirus Diarrhea Among Children Younger Than 5 Years. JAMA pediatrics 172 (2018) 958-965.

[5] B.K. Sederdahl, J. Yi, R.C. Jerris, S.E. Gillespie, L.F. Westblade, C.S. Kraft, A.L. Shane, and E.J. Anderson, Trends in rotavirus from 2001 to 2015 in two paediatric hospitals in Atlanta, Georgia. Epidemiology and infection 146 (2018) 465-467.

[6] E. Burnett, C.L. Jonesteller, J.E. Tate, C. Yen, and U.D. Parashar, Global Impact of Rotavirus Vaccination on Childhood Hospitalizations and Mortality From Diarrhea. The Journal of infectious diseases 215 (2017) 1666-1672.

400 [7] P.H. Dennehy, Rotavirus vaccines: an overview. Clinical microbiology reviews 21
401 (2008) 198-208.
402 [8] H.B. Greenberg, and M.K. Estes, Rotaviruses: from pathogenesis to vaccination.
403 Gastroenterology 136 (2009) 1939-51.
404 [9] J. Angel, M.A. Franco, and H.B. Greenberg, Rotavirus vaccines: recent
405 developments and future considerations. Nat Rev Microbiol 5 (2007) 529-39.
406 [10] J. Gómez-Rial, S. Sánchez-Batán, I. Rivero-Calle, J. Pardo-Seco, J.M. Martinón-
407 Martínez, A. Salas, and M.-T. F, Further considerations on rotavirus vaccination
408 and seizure-related hospitalization rates. Infect Drug Resist 12 (2019) 989—991.
409 [11] J. Gomez-Rial, S. Sanchez-Batan, I. Rivero-Calle, J. Pardo-Seco, J.M. Martinon-
410 Martinez, A. Salas, and F. Martinon-Torres, Rotavirus infection beyond the gut.
411 Infect Drug Resist 12 (2019) 55-64.
412 [12] A. Salas, J. Pardo-Seco, M. Cebey-López, J.M. Martinón-Martinez, J. Gómez-Rial,
413 M.J. Currás-Tuala, S. Pischedda, R. Barral-Arca, A. Justicia-Grande, I. Rivero-
414 Calle, J. Vilar, and F. Martinón-Torres, Impact of rotavirus vaccination on
415 childhood hospitalizations for seizures: Heterologous or unforeseen direct
416 vaccine effects? Vaccine 37 (2019) 3362-3368.
417 [13] A. Salas, G. Marco-Puche, J.C. Trivino, A. Gomez-Carballa, M. Cebey-Lopez, I.
418 Rivero-Calle, L. Vilanova-Trillo, C. Rodriguez-Tenreiro, J. Gomez-Rial, and F.
419 Martinon-Torres, Strong down-regulation of glycophorin genes: A host defense
420 mechanism against rotavirus infection. Infect. Genet. Evol. 44 (2016) 403-11.
421 [14] H.A. DeBerg, M.B. Zaidi, M.C. Altman, P. Khaenam, V.H. Gersuk, F.D. Campos, I.
422 Perez-Martinez, M. Meza-Segura, D. Chaussabel, J. Banchereau, T. Estrada-
423 Garcia, and P.S. Linsley, Shared and organism-specific host responses to

childhood diarrheal diseases revealed by whole blood transcript profiling. PLoS One 13 (2018) e0192082.

[15] J. Brown, M. Pirrung, and L.A. McCue, FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics (2017).

[16] P. Ewels, M. Magnusson, S. Lundin, and M. Kaller, MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32 (2016) 3047-8.

[17] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5 (2008) 621-8.

[18] M.D. Robinson, and A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology 11 (2010) R25.

[19] M.D. Robinson, D.J. McCarthy, and G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (2010) 139-40.

[20] K.D. Hansen, R.A. Irizarry, and Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13 (2012) 204-16.

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

[22] M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, and J.R. Gonzalez, A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. BMC Bioinformatics 14 (2013) 254.

[23] R. Barral-Arca, J. Pardo-Seco, X. Bello, F. Martinón-Torres, and A. Salas, Ancestry patterns inferred from massive RNAseq data. RNA 25 (2019) 857-868.

[24] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25 (2000) 25-9.

[25] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40 (2012) D109-14.

[26] J. Piñero, A. Bravo, N. Queralt-Rosinach, A. Gutierrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L.I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 45 (2017) D833-D839.

[27] H. Ishida, T. Iwama, S. Inokuma, I. Takeuchi, D. Hashimoto, and M. Miyaki, APC gene mutations in a jejunal adenoma causing intussusception in a patient with familial adenomatous polyposis. J Gastroenterol 37 (2002) 1057-61.

[28] X. Wu, and M. Watson, CORNA: testing gene lists for regulation by microRNAs. Bioinformatics 25 (2009) 832-3.

[29] C.J. Hoggart, PReMS: Parallel Regularised Regression Model Search for sparse bio-signature discovery. BioRxiv (2018) 355479.

[30] M. López-Ratón, M.X. Rodríguez-Álvarez, C. Cadarso-Suárez, and F. Gude-Sampedro, OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests. J. Stat. Softw. 61 (2014) 1-36.

471 [31] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, and M. Muller,
472 pROC: an open-source package for R and S+ to analyze and compare ROC
473 curves. BMC Bioinform. 12 (2011) 77.
474 [32] S.S. Shen-Orr, R. Tibshirani, P. Khatri, D.L. Bodian, F. Staedtler, N.M. Perry, T.
475 Hastie, M.M. Sarwal, M.M. Davis, and A.J. Butte, Cell type-specific gene
476 expression differences in complex tissues. Nature methods 7 (2010) 287-9.
477 [33] M. Chikina, E. Zaslavsky, and S.C. Sealfon, CellCODE: a robust latent variable
478 approach to differential expression analysis for heterogeneous cell populations.
479 Bioinformatics 31 (2015) 1584-91.
480 [34] E. Burnett, U. Parashar, and J. Tate, Rotavirus Vaccines: Effectiveness, Safety, and
481 Future Directions. Paediatr Drugs 20 (2018) 223-233.
482 [35] W.K. Yih, T.A. Lieu, M. Kulldorff, D. Martin, C.N. McMahill-Walraven, R. Platt, N.
483 Selvam, M. Selvan, G.M. Lee, and M. Nguyen, Intussusception risk after rotavirus
484 vaccination in U.S. infants. N Engl J Med 370 (2014) 503-12.
485 [36] K. Soares-Weiser, H. Bergman, N. Henschke, F. Pitan, and N. Cunliffe, Vaccines for
486 preventing rotavirus diarrhoea: vaccines in use. The Cochrane database of
487 systematic reviews 2019 (2019).
488 [37] C. Centers for Disease, and Prevention, Postmarketing monitoring of
489 intussusception after RotaTeq vaccination--United States, February 1, 2006-
490 February 15, 2007. MMWR Morb Mortal Wkly Rep 56 (2007) 218-22.
491 [38] K.A. Schafer, The cell cycle: a review. Veterinary pathology 35 (1998) 461-78.
492 [39] R. Bhowmick, G. Banik, S. Chanda, S. Chattopadhyay, and M. Chawla-Sarkar,
493 Rotavirus infection induces G1 to S phase transition in MA104 cells via
494 Ca(+)(2)/Calmodulin pathway. Virology 454-455 (2014) 270-9.

[40] K. Wen, T. Bui, M. Weiss, G. Li, J. Kocher, X. Yang, P.M. Jobst, T. Vaught, J. Ramsoondar, S. Ball, S. Clark-Deener, D. Ayares, and L. Yuan, B-Cell-Deficient and CD8 T-Cell-Depleted Gnotobiotic Pigs for the Study of Human Rotavirus Vaccine-Induced Protective Immune Responses. Viral immunology 29 (2016) 112-27.

[41] N.A. Kuklin, L. Rott, J. Darling, J.J. Campbell, M. Franco, N. Feng, W. Muller, N. Wagner, J. Altman, E.C. Butcher, and H.B. Greenberg, alpha(4)beta(7) independent pathway for CD8(+) T cell-mediated intestinal immunity to rotavirus. The Journal of clinical investigation 106 (2000) 1541-52.

[42] U. Desselberger, and H.I. Huppertz, Immune responses to rotavirus infection and vaccination and associated correlates of protection. The Journal of infectious diseases 203 (2011) 188-95.

[43] M. Hariharan, V. Scaria, B. Pillai, and S.K. Brahmachari, Targets for human encoded microRNAs in HIV genes. Biochem Biophys Res Commun 337 (2005) 1214-8.

[44] P.W. Hsu, L.Z. Lin, S.D. Hsu, J.B. Hsu, and H.D. Huang, ViTa: prediction of host microRNAs targets on viruses. Nucleic Acids Res 35 (2007) D381-5.

[45] N.L. Orr-Burks, B.S. Shim, W. Wu, A.A. Bakre, J. Karpilow, and R.A. Tripp, MicroRNA screening identifies miR-134 as a regulator of poliovirus and enterovirus 71 infection. Sci Data 4 (2017) 170023.

[46] R.F. Ramig, Pathogenesis of intestinal and systemic rotavirus infection. Journal of virology 78 (2004) 10213-20.

[47] I. Rivero-Calle, J. Gomez-Rial, and F. Martinon-Torres, Systemic features of rotavirus infection. J Infect 72 Suppl (2016) S98-S105.

[48] A. Gómez-Carballa, M. Cebey-López, J. Pardo-Seco, R. Barral-Arca, I. Rivero-Calle, S. Pischedda, M.J. Curras-Tuala, J.M. Gómez-Rial, F. Barros, F. Martinón-Torres, and A. Salas, A qPCR expression assay of *IFI44L* gene differentiates viral from bacterial infections in febrile children. Submitted (2019).

[49] Y. Liu, J. Zhou, and K.P. White, RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30 (2014) 301-4.

[50] H.S. Goodridge, S.S. Ahmed, N. Curtis, T.R. Kollmann, O. Levy, M.G. Netea, A.J. Pollard, R. van Crevel, and C.B. Wilson, Harnessing the beneficial heterologous effects of vaccination. Nat Rev Immunol 16 (2016) 392-400.

Legend to the figures

Figure 1. (A) Scheme of sampling and project design; (B) PCA constructed with the top 500 most highly expressed genes; (C) Venn plot of the genes differentially expressed when comparing healthy controls *vs.* vaccinated children and healthy controls *vs.* acute and convalescent infected (community-acquired) children, corrected by age and gender.

Figure 2. Heatmap of genes associated to intussusception.

Figure 3. (A) Network of biomarkers associated to fecal incontinence, diarrhea, nausea, and vomiting within the genes differentially expressed between vaccinated and healthy controls according to IPA; (B) Heatmap of the genes associated to fecal incontinence within the genes differentially expressed between vaccinated and healthy controls according to IPA; (C) Heatmap of the genes associated to diarrhea within the genes differentially expressed between vaccinated and healthy controls according to IPA; and (D) Heatmap of the genes associated to nausea and vomit within the genes differentially expressed between vaccinated and healthy controls according to IPA.

Figure 4. Heatmap of genes associated to humoral immunity.

Figure 5. Heatmap of the genes regulated by the miRNA hsa-mir-149 within the genes differentially expressed between vaccinated and healthy controls.

Figure 6. Classification performance to distinguish RV5V from control children based on a 9-transcript model.

548 **Table 1.** Genes included in the 9-transcript signature. Genes with positive logistic
549 regression coefficient values are upregulated in vaccinated children relative to
550 unvaccinated, whereas genes with negative values are downregulated. LR: logistic
551 regression.
552
553

Ensembl ID	Gene name	Gene	LR coefficient
ENSG00000118113	*MMP8*	Matrix metalloproteinase-8	-7.31×10^{-03}
ENSG00000128512	*DOCK4*	Dedicator of cytokinesis 4	5.80×10^{-03}
ENSG00000131142	*CCL25*	C-C motif chemokine ligand 25	-5.54×10^{-02}
ENSG00000172738	*TMEM217*	Transmembrane protein 217	-9.37×10^{-02}
ENSG00000175894	*TSPEAR*	Thrombospondin type laminin G domain and EAR repeat	2.41×10^{-02}
ENSG00000196565	*HBG2*	Hemoglobin subunit gamma 2	-6.18×10^{-06}
ENSG00000197768	*STPG3*	Sperm-tail PG-rich repeat containing 3	-2.41×10^{-01}
ENSG00000198435	*NRARP*	Notch-regulated ankryrin repeat protein	-1.63×10^{-01}
ENSG00000255423	*EBLN2*	Endogenous Bornavirus like nucleoprotein 2	2.24×10^{-02}

554 ## Supplementary Data

555 **Figure S1.** Heatmap of genes associated to abnormal morphology of the midgut.

556 **Figure S2**. Box and whiskers plots of the proportion of blood cells according to cell
557 deconvolution analysis. (A) Dendritic cells, (B) CD4T lymphocytes, (C) CD8T
558 lymphocytes, (D) plasma cells, (E) monocytes, (F) natural killer cells, (G) B lymphocytes,
559 and (H) neutrophils. For clarity, statistically significant values are only given for
560 comparisons between all conditions (acute and convalescent infected and healthy
561 controls) against vaccinated children. **Table S1**. Detailed sample information.

562 **Table S2**. Differentially expressed genes between controls and vaccinated/RV infected
563 according to Deseq2 corrected by age and gender.

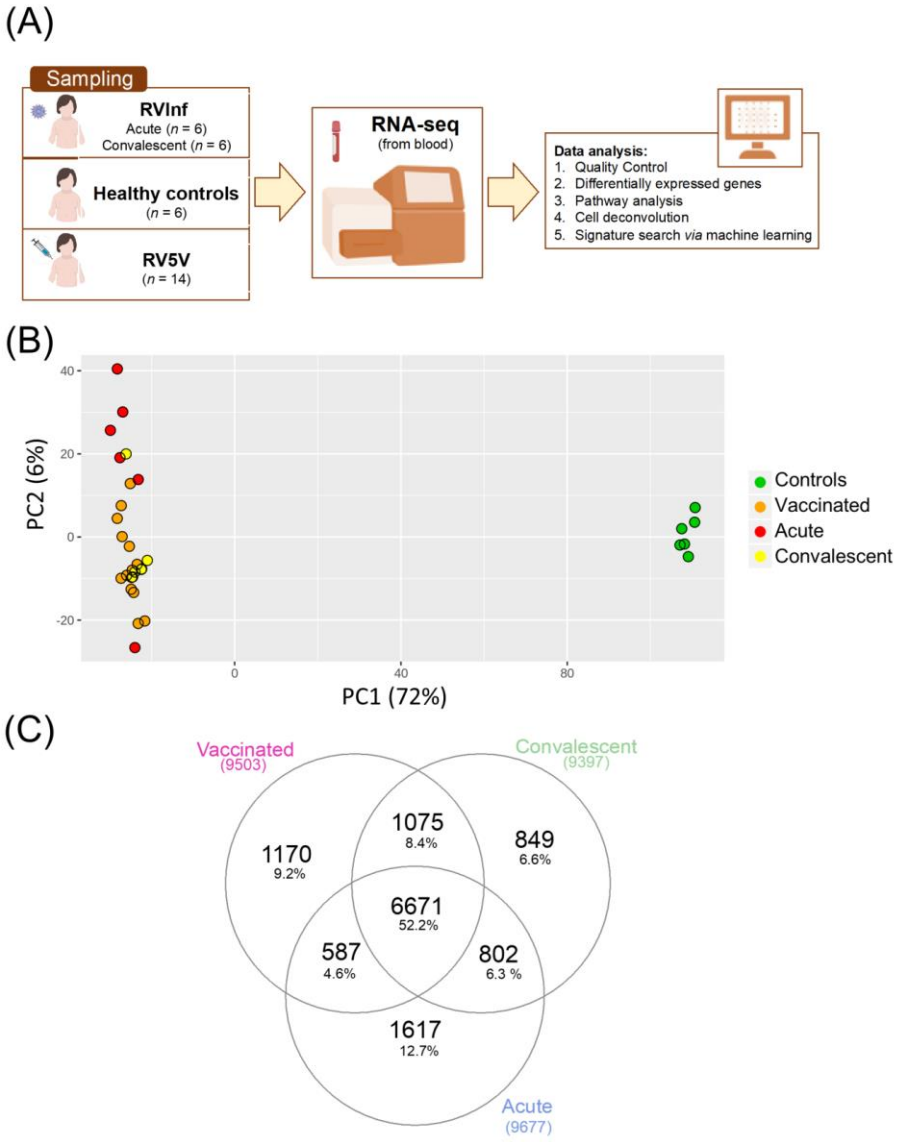
564 **Table S3**. Ingenuity canonical pathway analysis of the differentially expressed genes
565 between controls and vaccinated children. The top diseases and functions are indicated
566 as well as a detailed list of pathways specifically related to gastrointestinal and
567 immunological diseases.

568 **Table S4**. KEGG pathway enrichment analysis of the differentially expressed genes
569 between controls and vaccinated children.

570 **Table S5**. GO pathway enrichment analysis of the differentially expressed genes
571 between controls and vaccinated children.

572

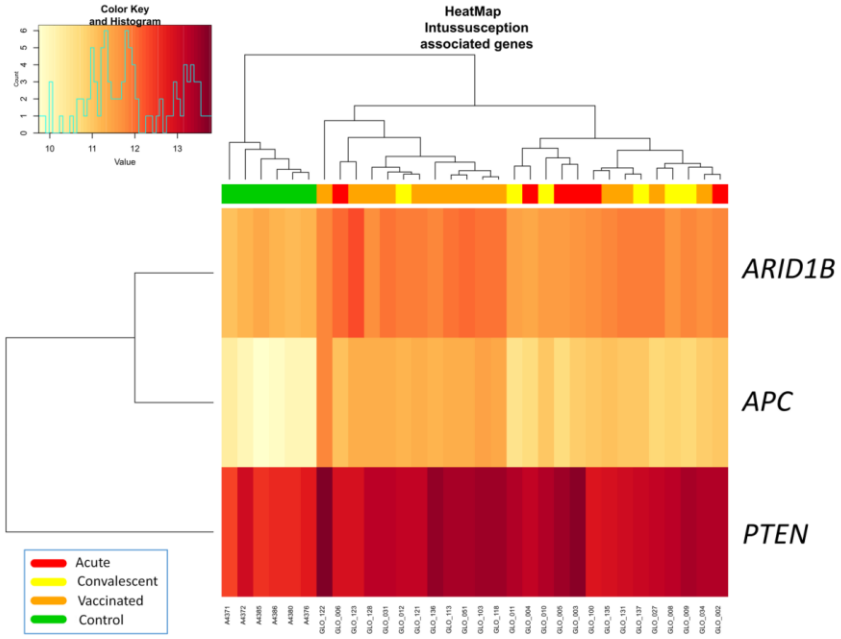
573 **Figure 1.**



574

575 **Figure 2.**

576



577

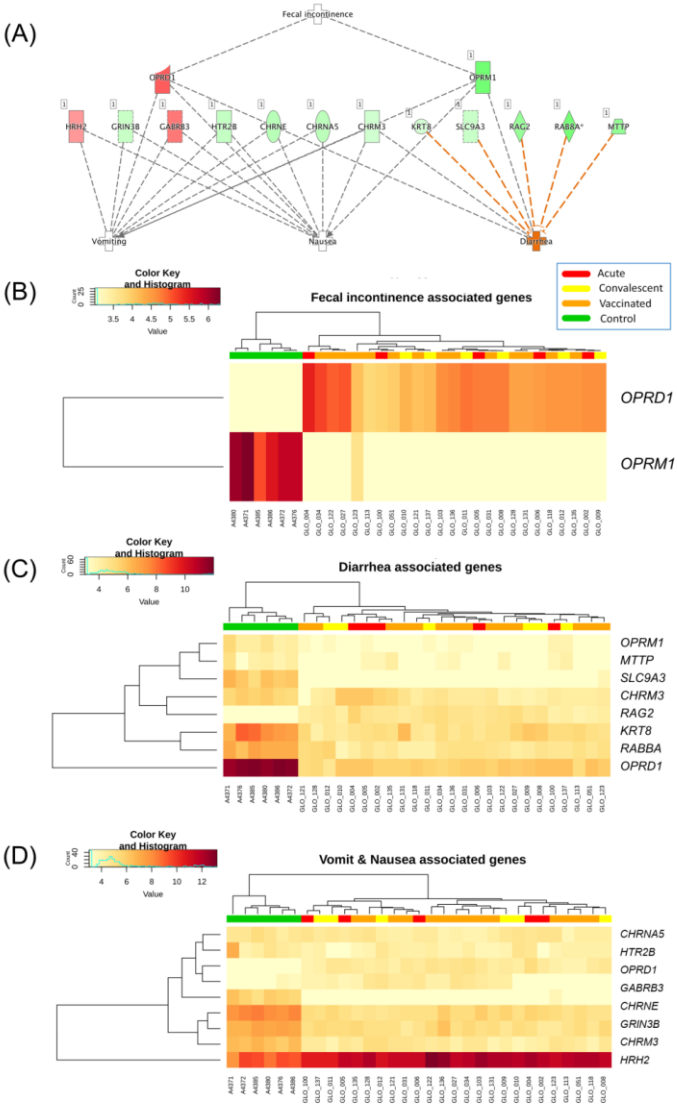
578

579

580

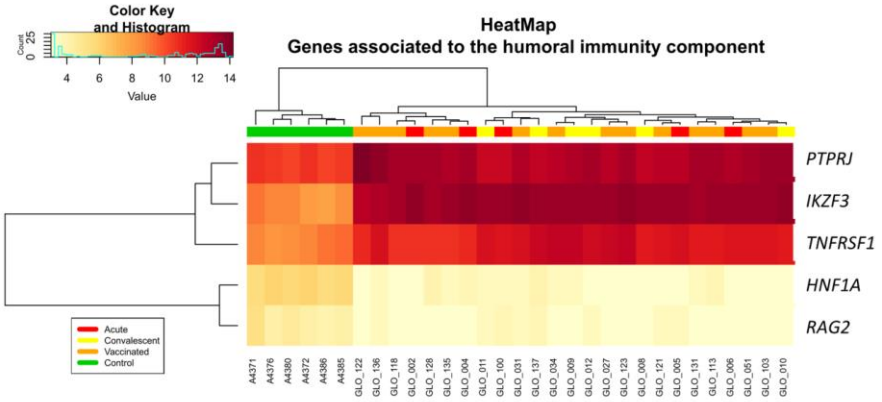
581

582 **Figure 3.**



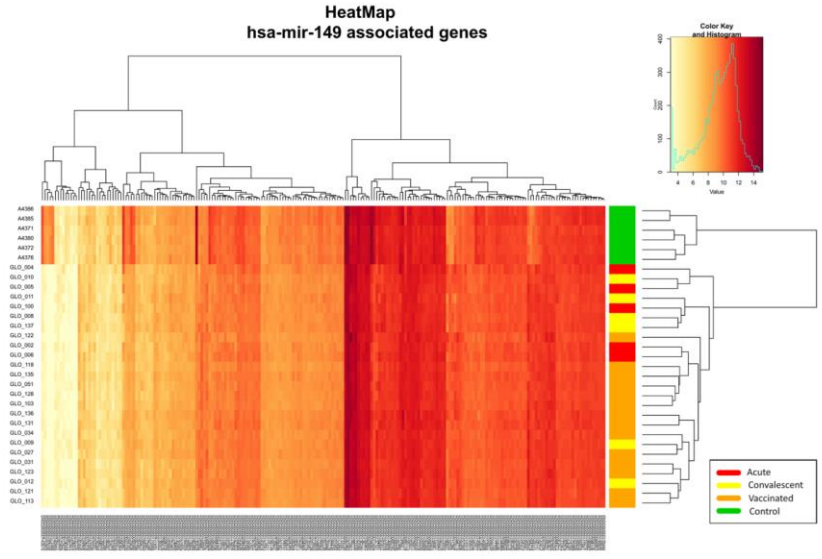
583

584 **Figure 4**



585

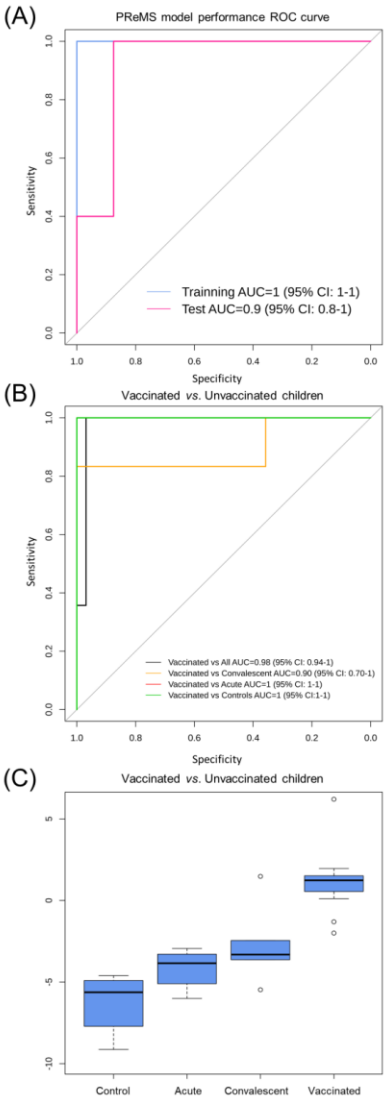
586 **Figure 5.**



587

588

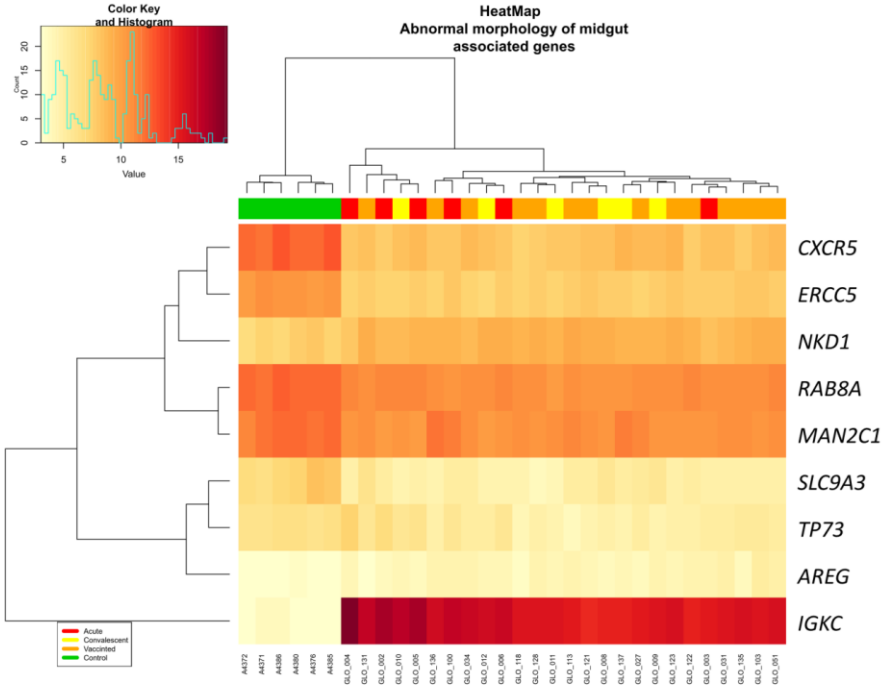
589 **Figure 6.**



590

591 **Figure S1.**

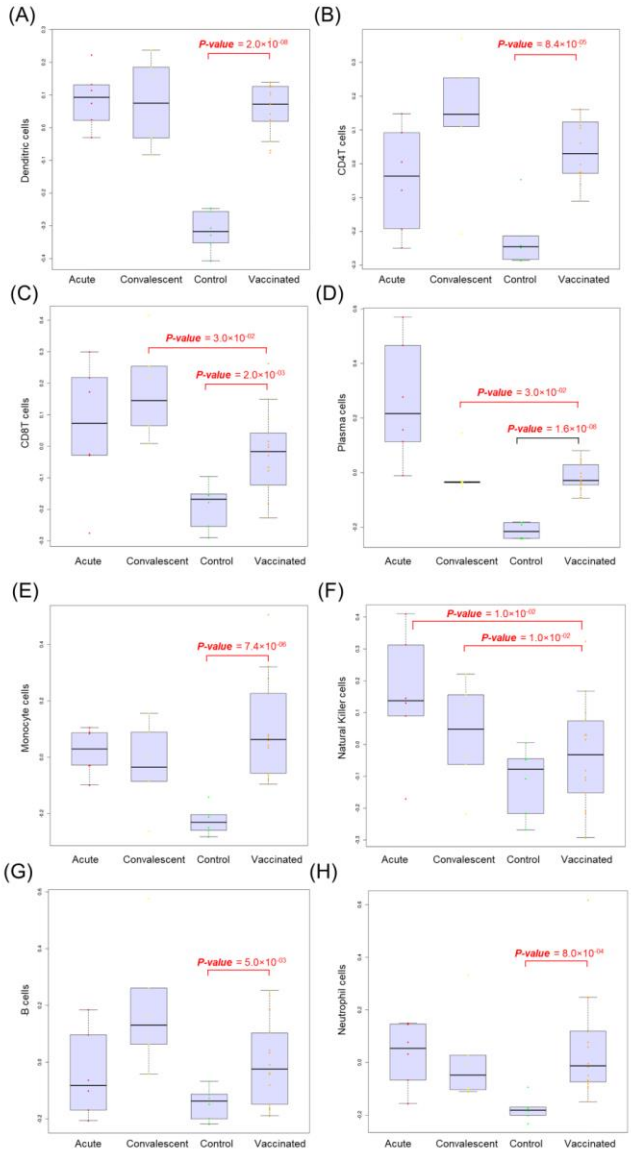
592



593

594

595 **Figure S2.**



596

7. Conclusions

During this thesis, I focused on the identification of transcriptomic host biomarkers for infectious diseases in a broad sense viral vs bacterial infections but also particular diseases such as rotavirus, respiratory syncytial virus, *S. pneumoniae* etc.. Here I list the main conclusions extracted from the research conducted in the present PhD:

1. Our whole Exome association study pointed to the SNPs rs201967957 (gene *MEIS1*) and rs576099063 (gene *TSPAN15*) as host variants associated with pneumococcal pneumonia. A burden gene test of pathogenicity signaled four other genes, namely, *OR9G9*, *MUC6*, *MUC3A* and *APOB*, which carry significantly increased pathogenic variation when compared to controls. By analyzing various transcriptomic data repositories, we found strong supportive evidence for the role of *MEIS1*, *TSPAN15* and *APOBR* (encoding the receptor of the *APOB* protein) in pneumonia in mouse and human models.

2. Our results suggest that the two transcripts *IFI44L* and *FAM89A*, provide a strong signal to differentiate bacterial from viral infections, a non-population dependent signal, and useful for discriminating a wide range of pathogens and different levels of severity. Furthermore, according to our q-PCR assay measuring the expression of the gene *IFI44L* would be enough to discriminate between viral and bacterial diseases.

3. We demonstrate that host expression data (RNA-seq or microarray) can be successfully translated into a fast, highly accurate and relatively inexpensive in vitro qPCR assay that could be implemented in the clinical routine.
4. We proved that is possible to infer genetic ancestry information from RNA-seq data, but the accuracy of the results depends on the quality of the sequencing reads.
5. This thesis brings a cautionary note for gene expression studies highlighting the importance to control for the potential confounding effect of ancestral genetic background.
6. We identified a minimal 17-transcript host signature specific for RSV infection by comparing transcriptomic profiles against other respiratory viruses.
7. We found 2-lncRNAs that arise as promising targets for infection diagnosis and therapy (ENSG00000254680 and ENSG00000273149) and patented their use as viral biomarkers.
8. To the best of our knowledge, this is the very first thesis to use host lncRNAs biomarkers for the diagnosis of human viral infections.
9. Rotavirus vaccination mimics the wild type infection-causing similar, but not completely equivalent changes in children's transcriptome, including transcripts associated with cell cycle, diarrhea, nausea, vomiting, intussusception, and abnormal morphology of midgut
10. We found a 9-transcript signature, that accurately identifies vaccinated against convalescent-infected children and against acute-infected and healthy control children, which may be helpful to identify unvaccinated or RV susceptible children in

which the vaccination may have not provided an adequate level of protection.

11. We identified a miRNA hsa-mir-149 that seems to play a role in the host defense against viral pathogens and may have an antiviral role

Taken together these results suggest that each infectious disease is associated with a unique pattern of genes that turn on or off by forming a "molecular signature," which can be used to quickly identify each disease.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8. Discussion

Regardless of its limitations, the present thesis represents a step forward towards the use of host gene signature in clinical practice. The translation of the omic biomarkers found in this thesis into a clinical test for diagnosis, prognosis or risk assessment needs additional validation. And this validation would require the designing of thoughtful clinical studies in order to evaluate scenarios. For instance, different severities, time points in the course of the infectious disease, parasitic infections, other inflammatory diseases etc.

It would be desirable that these future studies combine molecular, immunological analysis, traditional microbiologic cell cultures, metagenomic approaches, and omic host gene signatures. As using a holistic approach to integrate the pathogen information, the host immune/transcriptional response, and the clinical symptoms of the disease will allow us on the one hand to expand the infectious pathogenesis knowledge. And on the other hand, will help to estimate if a pathogen detected in a clinical test is responsible for the observed pathogenesis or if it is just harmless colonization.

It has to be remarked that host gene signatures are not intended to replace microbiology based diagnosis, instead, they emerge as a complementary tool to obtain further information. The development of diagnostic/prognosis tests based on both the pathogen and the host response could potentially revolutionize the management of patients with suspected sepsis, fever of uncertain aetiology and also help to distinguish patients with a higher risk of developing a severe or invasive infectious disease enabling earlier drug treatment and/or increasing patient surveillance.

The final goal should be unravelling the role of these biomarkers and their associated pathways under the hypothesis that they might help to discover the critical mechanisms in host defence against specific pathogens. Which will also help develop new therapeutic approaches.

Even though there are still numerous adversities to overcome before host gene expression signatures can be introduced into molecular diagnosis routine. Signatures based on host gene expression biomarkers have a great potential for diagnosis of infectious diseases, and probably in the following years, we will see skyrocket their use in clinical diagnostic tests.

In the next years, the scientific community will likely build a genetic signature library that covers all common conditions. Which in parallel with developing new technologies that will be able to quickly and accurately determine gene expression of a small number of genes would lead to quicker diagnosis and avoid unnecessary antibiotic treatments.

Even though RNA-seq and microarrays are the most powerful screening approaches for the discovery of host RNA signatures of infectious diseases, both have inherent problems such as a higher error rate than traditional Sanger sequencing, standardization and reproducibility issues etc. Therefore, before any biomarker is translated into a clinical test, it needs to be validated using technologies more precise and less prone to false positives than NGS and microarrays like qPCR. Consequently, a reasonable next step in the future would be to validate via qPCR or Nanostring ® the host biomarkers discovered in this thesis.

qPCR is currently the “gold standard” in gene expression studies. Many studies have proven that qPCR is a valid method to validate microarrays and RNA-seq findings. Furthermore, qPCR based assays are already widely used in hospitals because it is a technique with high accuracy, relatively cheap and fast. It has been described that there is a

strong correlation between microarray and qPCR results. Nevertheless, it must be taken into account that this association can be deeply affected by methodological and analytical factors. Therefore establishing a detailed laboratory qPCR protocol, that includes a careful selection of housekeeping genes for each specific condition, and ensuring good laboratory practices is crucial to successfully convert a host transcriptional signature into a qPCR assay to be used in diagnostics laboratory routine[7]

And last but not least, even though the development os bedside test based on transcriptomic biomarkers would be hoped-for at present is a goal extremely difficult to achieve technical limitations. Nonetheless, this objective might be achieved in the near future thanks to new emergent technologies that allow sensitive and qualitative detection of gene expression. For instance, the Oxford Nanopore MinION® could be an interesting tool to translate genome-wide host RNA signature into a routine bedside diagnosis test. Since MinION® is a portable sequencer, allows real-time data acquisition (there is no need to wait for the run to end to start analysing the data) and the protocol can be optimized to yield the result roughly six hours[93]. Furthermore, Oxford Nanopore offers devices for library preparation that would allow running the test without the need for a laboratory. Another interesting approach is the NanoString nCounter® system to measure multiple mRNA transcripts at the same time. Even if extremely reproducible and powerful, this approach requires expensive equipment and sophisticated bioinformatic data analysis, which makes unsuitable for its use in point of care routine diagnosis[12]

It is likely that in the next years see the application of the first host gene expression diagnosis tests for in infectious diseases in clinical settings and, more importantly, an improvement in the diagnosis of infectious diseases.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

9. References

1. Schlapbach, L.J., *Paediatric sepsis.* Current opinion in infectious diseases, 2019. **32**(5): p. 497-504.
2. Martinón-Torres, F., et al., *Life-threatening infections in children in Europe (the EUCLIDS Project): a prospective cohort study.* The lancet child & adolescent health, 2018. **2**(6): p. 404-414.
3. Hall, C.B., et al., *The burden of respiratory syncytial virus infection in young children.* New England journal of medicine, 2009. **360**(6): p. 588-598.
4. Le Doare, K., et al., *Very low rates of culture-confirmed invasive bacterial infections in a prospective 3-year population-based surveillance in Southwest London.* Archives of disease in childhood, 2014. **99**(6): p. 526-531.
5. Herberg, J.A., et al., *Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children.* Jama, 2016. **316**(8): p. 835-845.
6. Barral-Arca, R., et al., *A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms.* Scientific reports, 2018. **8**(1): p. 1-7.
7. Gómez-Carballa, A., et al., *A qPCR expression assay of IFI44L gene differentiates viral from bacterial infections in febrile children.* Scientific reports, 2019. **9**(1): p. 1-12.
8. Chu, D.K., et al., *Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia.* Clinical chemistry, 2020.

9. Mahony, J.B., *Detection of respiratory viruses by molecular methods.* Clinical microbiology reviews, 2008. **21**(4): p. 716-747.

10. Ramilo, O. and A. Mejías, *Shifting the paradigm: host gene signatures for diagnosis of infectious diseases.* Cell host & microbe, 2009. **6**(3): p. 199-200.

11. Chang, A.B., et al., *Improving the diagnosis, management, and outcomes of children with pneumonia: where are the gaps?* Frontiers in pediatrics, 2013. **1**: p. 29.

12. Gliddon, H.D., et al., *Genome-wide host RNA signatures of infectious diseases: discovery and clinical translation.* Immunology, 2018. **153**(2): p. 171-178.

13. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.

14. Westermann, A.J., S.A. Gorski, and J. Vogel, *Dual RNA-seq of pathogen and host.* Nature Reviews Microbiology, 2012. **10**(9): p. 618-630.

15. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**(Mar): p. 1157-1182.

16. Frohlich, H., O. Chapelle, and B. Scholkopf. *Feature selection for support vector machines by means of genetic algorithm*. in *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*. 2003. IEEE.

17. Zaas, A.K., et al., *Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans.* Cell host & microbe, 2009. **6**(3): p. 207-217.

18. Scicluna, B.P., et al., *A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission.* American journal of respiratory and critical care medicine, 2015. **192**(7): p. 826-835.

19. Sweeney, T.E., et al., *Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis.* The Lancet Respiratory Medicine, 2016. **4**(3): p. 213-224.

20. Anderson, S.T., et al., *Diagnosis of childhood tuberculosis and host RNA expression in Africa.* New England Journal of Medicine, 2014. **370**(18): p. 1712-1723.

21. Almansa, R., et al., *Critical COPD respiratory illness is linked to increased transcriptomic activity of neutrophil proteases genes.* BMC research notes, 2012. **5**(1): p. 401.

22. Burnham, K.L., et al., *Shared and distinct aspects of the sepsis transcriptomic response to fecal peritonitis and pneumonia.* American journal of respiratory and critical care medicine, 2017. **196**(3): p. 328-339.

23. Mayhew, M.B., et al., *A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections.* Nature Communications, 2020. **11**(1): p. 1-10.

24. Pankla, R., et al., *Genomic transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of septicemic melioidosis.* Genome biology, 2009. **10**(11): p. R127.

25. Parnell, G.P., et al., *A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia.* Critical Care, 2012. **16**(4): p. R157.

26. Barral-Arca, R., et al., *A Meta-Analysis of Multiple Whole Blood Gene Expression Data Unveils a Diagnostic Host-Response Transcript Signature for Respiratory Syncytial Virus.* International Journal of Molecular Sciences, 2020. **21**(5): p. 1831.

27. Chaussabel, D., et al., *A modular framework for biomarker and knowledge discovery from blood transcriptional profiling studies: application to systemic lupus erythemathosus.* Immunity, 2008. **29**(1): p. 150-164.

28. Chen, N., et al., *Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study.* The Lancet, 2020. **395**(10223): p. 507-513.

29. Cox, J.A., et al., *Autopsy causes of death in HIV-positive individuals in sub-Saharan Africa and correlation with clinical diagnoses.* AIDS Rev, 2010. **12**(4): p. 183-94.

30. Serrano-Gomez, S.J., et al., *Ancestry as a potential modifier of gene expression in breast tumors from Colombian women.* PloS one, 2017. **12**(8).

31. Mahajan, P., et al., *Association of RNA biosignatures with bacterial infections in febrile infants aged 60 days or younger.* Jama, 2016. **316**(8): p. 846-857.

32. Mejias, A., et al., *Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection.* PLoS medicine, 2013. **10**(11).

33. Buyse, M., et al., *Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer.* Journal of the National Cancer Institute, 2006. **98**(17): p. 1183-1192.

34. Allantaz, F., et al., *Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade.* The Journal of experimental medicine, 2007. **204**(9): p. 2131-2144.

35. Hurd, P.J. and C.J. Nelson, *Advantages of next-generation sequencing versus the microarray in epigenetic research.* Briefings in Functional Genomics and Proteomics, 2009. **8**(3): p. 174-183.

36. Heather, J.M. and B.J.G. Chain, *The sequence of sequencers: The history of sequencing DNA.* 2016. **107**(1): p. 1-8.

37. Behjati, S. and P.S. Tarpey, *What is next generation sequencing?* Archives of Disease in Childhood-Education and Practice, 2013. **98**(6): p. 236-238.

38. Wall, J.D., et al., *Estimating genotype error rates from high-coverage next-generation sequence data.* . Genome research, 2014. **24**(11): p. 1734-1739.

39. Farhan, S.M. and R.A. Hegele, *Exome sequencing: new insights into lipoprotein disorders.* Current cardiology reports, 2014. **16**(7): p. 507.

40. Barral-Arca, R., et al., *A 2-transcript host cell signature distinguishes viral from bacterial diarrhea and it is influenced by the severity of symptoms.* Sci Rep, 2018. **8**(1): p. 8043.

41. Andrews, S., *FastQC: a quality control tool for high throughput sequence data.*, in https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.

42. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics.* Bioinformatics, 2016. **32**(19): p. 3047-3048.

43. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. 10-12.

44. Gaidatzis, D., A. Lerch, Hahne, F., and M.B. Stadler, *QuasR: quantification and annotation of short reads in R.* Bioinformatics, 2014. **31**(7): p. 1130-1132.

45. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

46. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

47. Dobin, A., et al., *.STAR: ultrafast universal RNA-seq aligner. Bioinformatics.* 2013. **29**(1): p. 15-21.

48. Barral-Arca, R., et al., *Ancestry patterns inferred from massive RNA-seq data.* RNA, 2019. **25**(7): p. 857-868.

49. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

50. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

51. Falcon, S. and R. Gentleman, *Using GOstats to test gene lists for GO term association.* Bioinformatics, 2006. **23**(2): p. 257-258.

52. Wu, X. and M. Watson, *CORNA: testing gene lists for regulation by microRNAs.* Bioinformatics, 2009. **25**(6): p. 832-833.

53. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome biology, 2010. **11**(3): p. R25.

54. Hansen, K.D., R.A. Irizarry, and Z. Wu, *Removing technical variability in RNA-seq data using conditional quantile normalization.* Biostatistics, 2012. **13**(2): p. 204-216.

55. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621.

56. Risso, D., et al., *GC-content normalization for RNA-Seq data.* BMC bioinformatics, 2011. **12**(1): p. 480.

57. Gonzalez, J.R., et al., *Package 'tweeDEseq'.* 2013.

58. Maza, E., *In Papyro comparison of TMM (edgeR), RLE (DESeq2), and MRN normalization methods for a simple two-conditions-without-replicates RNA-Seq experimental design.* Frontiers in genetics, 2016. **7**: p. 164.

59. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.

60. Kallio, M.A., et al., *Chipster: user-friendly analysis software for microarray and other high-throughput data.* BMC genomics, 2011. **12**(1): p. 507.

61. Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome research, 2008. **18**(9): p. 1509-1517.

62. Di, Y., et al., *The NBP negative binomial model for assessing differential gene expression from RNA-Seq.* Statistical Applications in Genetics and Molecular Biology, 2011. **10**(1).

63. Rapaport, F., et al., *Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.* Genome biology, 2013. **14**(9): p. 3158.

64. Fang, Z., J. Martin, and Z. Wang, *Statistical methods for identifying differentially expressed genes in RNA-Seq experiments.* Cell & bioscience, 2012. **2**(1): p. 26.

65. Robinson, M.D. and G.K. Smyth, *Moderated statistical tests for assessing differences in tag abundance.* Bioinformatics, 2007. **23**(21): p. 2881-2887.

66. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome biology, 2004. **5**(10): p. R80.

67. Owens, M., *The definitive guide to SQLite*. 2006: Apress.

68. Consortium, G.O., *The Gene Ontology (GO) database and informatics resource.* Nucleic acids research, 2004. **32**(suppl_1): p. D258-D261.

69. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

70. Croft, D., et al., *The Reactome pathway knowledgebase.* Nucleic acids research, 2013. **42**(D1): p. D472-D477.

71. Krämer, A., et al., *Causal analysis approaches in ingenuity pathway analysis.* Bioinformatics, 2013. **30**(4): p. 523-530.

72. Consortium, G.O., *The gene ontology project in 2008.* Nucleic acids research, 2007. **36**(suppl_1): p. D440-D444.

73. Alexa, A., J. Rahnenführer, and T. Lengauer, *Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.* Bioinformatics, 2006. **22**(13): p. 1600-1607.

74. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.

75. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2008. **37**(1): p. 1-13.

76. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome biology, 2009. **10**(3): p. R25.

77. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome research, 2010. **20**(9): p. 1297-1303.

78. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic acids research, 2010. **38**(16): p. e164-e164.

79. Ramos, A.H., et al., *Oncotator: cancer variant annotation tool.* Human mutation, 2015. **36**(4): p. E2423-E2429.

80. Stromberg, M., et al. *Nirvana: clinical grade variant annotator*. in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2017. ACM.

81. Bumgarner, R., *Overview of DNA microarrays: types, applications, and their future.* Curr Protoc Mol Biol, 2013. **Chapter 22**: p. Unit 22 1.

82. Rohart, F., et al., *mixOmics: An R package for 'omics feature selection and multiple data integration.* PLoS computational biology, 2017. **13**(11): p. e1005752.

83. Gautier, L., et al., *affy—analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-315.

84. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-1548.

85. Smyth, G., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments. 2004; 3: Article3.* Stat Appl Genet Mol Biol. **3**.

86. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-470.

87. Cesano, A., *nCounter® PanCancer immune profiling panel (NanoString technologies, Inc., Seattle, WA).* Journal for immunotherapy of cancer, 2015. **3**(1): p. 42.

88. Kulkarni, M.M., *Digital multiplexed gene expression analysis using the NanoString nCounter system.* Current protocols in molecular biology, 2011. **94**(1): p. 25B. 10.1-25B. 10.17.

89. Gomez-Carballa, A., et al., *A qPCR expression assay of IFI44L gene differentiates viral from bacterial infections in febrile children.* Sci Rep, 2019. **9**(1): p. 11780.

90. Kassambara, A., *Machine Learning Essentials: Practical Guide in R*. 2018: sthda.

91. Friedman, J., T. Hastie, and R. Tibshirani, *glmnet: Lasso and elastic-net regularized generalized linear models.* R package version, 2009. **1**(4).

92. Hoggart, C.J., *PRReMS: Parallel Regularised Regression Model Search for bio-signature discovery.* bioRxiv, 2018: p. 355479.

93. Charalampous, T., et al., *Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection.* 2019. **37**(7): p. 783-792.