Universidad
de La Laguna

Escuela de Doctorado
y Estudios de Posgrado

TÍTULO DE LA TESIS DOCTORAL

Genomic approaches to understand the pathogenesis of the acute repiratory distress syndrome

AUTOR/A

BEATRIZ	GUILLEN	GUIO

DIRECTOR/A

Carlos Alberto	Flores	Infante

CODIRECTOR/A

DEPARTAMENTO O INSTITUTO UNIVERSITARIO

FECHA DE LECTURA

06/02/20

Universidad
de La Laguna

Doctoral program in Health Science

DOCTORAL THESIS

Genomic approaches to understand the pathogenesis of the acute respiratory distress syndrome

Beatriz Guillén Guío

Research unit, University Hospital Nuestra Señora de Candelaria

Santa Cruz de Tenerife, November 2019

El Doctor D. Carlos A. Flores Infante, director de la tesis doctoral presentada por Dña. Beatriz Guillén Guío

Certifica que:

La memoria presentada por la licenciada Beatriz Guillén Guío titulada “Genomic approaches to understand the pathogenesis of the acute respiratory distress syndrome” ha sido realizada bajo su dirección en la Unidad de Investigación del Hospital Universitario Nuestra Señora de Candelaria, y considerando que reúne las condiciones de calidad y rigor científico, se autoriza para que pueda ser presentada y defendida ante la comisión nombrada al efecto para optar al grado de Doctor por la Universidad de La Laguna.

Fdo.: Dr. D. Carlos A. Flores Infante

Santa Cruz de Tenerife, noviembre de 2019

Esta tesis se ha realizado fundamentalmente gracias a la financiación de los proyectos de investigación PI14/00844 y PI17/00610 del Instituto de Salud Carlos III cofinanciados por los fondos Europeos de Desarrollo Regional "Una manera de hacer Europa". Mi contrato de investigación predoctoral (TESIS2015010057) ha sido financiado por la Agencia Canaria de Investigación, Innovación y Sociedad de la Información (ACIISI) de la Consejería de Economía, Industria, Comercio y Conocimiento y por el Fondo Social Europeo (FSE) Programa Operativo Integrado de Canarias 2014-2020, Eje 3 Tema Prioritario 74 (85%).

Agradecimientos

Ahora que echo la vista atrás, son tantas las personas a las que me gustaría agradecer por su ayuda y apoyo durante estos años que me resulta complicado resumirlo en pocas palabras.

En primer lugar, me gustaría agradecer al Dr. Carlos Flores, mi director de tesis, por el tiempo y esfuerzo que ha puesto en mi formación y en que esta Tesis saliera adelante. Gracias por guiarme durante todos estos años y por tu incalculable ayuda en absolutamente todos los retos que se me han presentado. Gracias por ser un gran mentor, por tus consejos, por transmitirme la pasión por la ciencia y el espíritu crítico y por confiar en mí. Ha sido un lujo realizar la tesis siguiendo tus pasos.

Quiero agradecer también a mis compañeros, tanto a los que estuvieron en la etapa inicial de mi tesis, como a los que me han acompañado en la fase final. Es muy fácil trabajar cuando te rodeas de personas tan excepcionales. Gracias a Natalia por estar a mi lado desde el principio, por comprenderme mejor que nadie durante estos años y por ser amiga y confidente. Gracias a Fabi por sus consejos que me empujaron a emprender esta aventura y por hacerme sentir parte del grupo desde el primer día. Gracias a Mar y Marialbert por ser las mejores hermanas mayores y un gran ejemplo de trabajo duro y dedicación. Gracias a Itahisa por su ayuda y apoyo durante la última etapa de mi Tesis y por transmitirme tranquilidad cuando la necesitaba. Gracias a Almudena por su gran labor en el procesamiento de las muestras y en otras innumerables tareas. Gracias a Tamara por su ayuda y colaboración en los trabajos presentados en mi Tesis, así como a Héctor, José Miguel, Amalia, Rafa, Ana, Laura y al resto de compañeros que han contribuido, de forma directa o indirecta, a que yo haya llegado hasta aquí. Gracias también a la Dra. Lina Pérez, al Dr. Juan Navarro y al resto de personas de la Unidad de investigación que me han apoyado en este proceso, especialmente a Isabel González y a Lari. Gracias a todos por hacer que este camino, a veces complicado, haya sido una experiencia tan positiva. Gracias también al Dr. Manuel Espinosa por redirigirme al grupo adecuado en el momento oportuno.

Quiero agradecer al Dr. Jesús Villar su gran ayuda en la clasificación de los pacientes, crucial para la elaboración de una base de datos clínica sólida. Agradezco también a todas las personas implicadas en el reclutamiento de pacientes y en la recogida de muestras: a todos los médicos que forman parte del estudio GEN-SEP, a los departamentos de anestesiología y microbiología del Hospital Universitario Nuestra Señora de Candelaria y al CDC de Canarias. Especialmente, gracias a los donantes de las muestras incluidas en los estudios y a sus familiares. Additionally, I would like to thank our

collaborators from the University of Virginia, the University of Pennsylvania and the Jena University Hospital for their collaboration in the GWAS, particularly Drs. Shwu-Fan Ma, André Scherag and Nuala J. Meyer. I would also like to thank the Genetic Epidemiology Group at the University of Leicester for welcoming me during my predoctoral stay and for their help during those months, specially Professor Louise Wain and Dr. Richard Allen.

Finalmente, si esta Tesis ha sido posible es también gracias al apoyo que he recibido de los grandes pilares de mi vida. Gracias a Dami por su paciencia y comprensión durante estos años, por transmitirme confianza en los momentos más complicados, por apoyarme en todo momento y en toda decisión y por hacerme ver siempre el vaso medio lleno. Gracias también a mis padres y a mi hermana por su apoyo durante esta etapa, por ser pacientes, por confiar en mí y por hacer siempre que todo fuese más fácil. Mamá, papá, la persona y la profesional que soy ahora os lo debo a vosotros, a la educación que me habéis transmitido desde pequeña, a vuestro esfuerzo y dedicación, a vuestros consejos y a vuestro amor incondicional. La palabra “gracias” se me queda corta.

A mis padres

Abstract

The acute respiratory distress syndrome (ARDS) is an acute lung inflammatory process that commonly develops as a consequence of severe infections, being sepsis its main cause of development. Despite the fatality of the syndrome, there is a lack of specific therapeutic options and effective prognostic methods for patients. Since many studies support the influence of genetic factors and microbiome shifts in the origin and evolution of ARDS, here we have aimed to address its pathophysiology using different genomic approaches. We have performed a genome-wide association study in European patients with sepsis, revealing a novel gene associated with ARDS susceptibility. Additionally, we have sequenced the bacterial DNA extracted from lung aspirates from a subset of the individuals with sepsis, reporting the association of the reduction of bacterial diversity with intensive care unit mortality during the first 8 h of sepsis diagnosis. Finally, the exploration of the genomic variation of a recently admixed population has pointed out genomic regions related to the ethnicity and harboring novel genes associated with response to infections and with the severe acute respiratory syndrome, among many other traits. All these findings have allowed us to further understand the pathogenesis of the syndrome and of main risk factors, as well as i) to propose VEGFR-1 as a potential therapeutic target, ii) to suggest the bacterial diversity as an early prognostic biomarker in critical patients, and iii) to lay the foundations for designing fine and admixture mapping studies in Canary Islanders to identify novel risk genes for complex traits such as sepsis and ARDS.

Abbreviations

1KGP: 1000 Genomes Project

ACE: angiotensin-converting enzyme

ALI: acute lung injury

APACHE II: Acute Physiology and Chronic Health Evaluation II

ARDS: acute respiratory distress syndrome

ATI: alveolar type I

ATII: alveolar type II

eQTL: expression quantitative trait loci

EUR: European

FKBP5: FKBP prolyl isomerase 5

FiO_2: fraction of inspired oxygen

FLT1: Fms related tyrosine kinase 1

GWAS: genome-wide association study

HLA: human leukocyte antigen

HMP: Human Microbiome Project

HRC: Haplotype Reference Consortium

ICU: intensive care unit

IL: interleukin

LD: linkage disequilibrium

MYLK: myosin light chain kinase

MV: mechanical ventilation

NAF: North African

NGS: next-generation DNA sequencing

PaO_2: partial pressure of oxygen in arterial blood

PCDH: protocadherin

PCR: polymerase chain reaction

PEEP: positive end-expiratory pressure

SARS: severe acute respiratory syndrome

SELPLG: selectin p ligand

SNPs: single nucleotide polymorphisms

SSA: sub-Saharan Africans

TLR: Toll-like receptor

OTU: operational taxonomic unit

SIRS: systemic inflammatory response syndrome

VEGFA: vascular endothelial growth factor A

VEGFR-1: vascular endothelial growth factor receptor 1

WES: whole-exome sequencing

WGS: whole-genome sequencing

Table of Contents

1. Introduction

1. Introduction

1.1 Acute respiratory distress syndrome (ARDS)

The acute respiratory distress syndrome (ARDS) is an acute pulmonary inflammatory process that commonly manifests as a response to severe infections, trauma, and several other factors. This heterogeneous syndrome is caused by a direct or indirect insult to the alveolar-capillary barrier that leads to an increased vascular permeability and the consequent formation of pulmonary edema (Bernard et al. 1994). Clinically, patients with ARDS present severe hypoxemia, which is assessed by means of the ratio of the partial pressure of oxygen in arterial blood (PaO_2) and the fraction of inspired oxygen (FiO_2), as well as bilateral pulmonary infiltrates and decreased lung compliance (Villar 2011).

Since 1967, when ARDS was first described (Ashbaugh et al. 1967), numerous studies have addressed the characteristics of the syndrome to identify a global diagnostic criterion for patients. According to the most recent definition of ARDS, the Berlin definition (ARDS Definition Task Force et al. 2012), ARDS is defined as an acute onset characterized by a PaO_2/FiO_2 ratio of less than or equal to 300 mmHg, bilateral pulmonary opacities on chest radiograph, non-cardiogenic edema, and by the use of a minimum positive end-expiratory pressure (PEEP) or a continuous positive airway pressure (CPAP) of 5 cm H_2O during mechanical ventilation (MV) (ARDS Definition Task Force et al. 2012). Additionally, based on the degree of hypoxemia, the syndrome can be classified as mild (200 mm Hg < $PaO_2/FiO_2 \leq$ 300 mm Hg), moderate (100 mm Hg < $PaO_2/FiO_2 \leq$ 200 mm Hg), or severe ($PaO_2/FiO_2 \leq$ 100 mm Hg) (ARDS Definition Task Force et al. 2012). The mild form of ARDS was previously known as acute lung injury (ALI) (ARDS Definition Task Force et al. 2012). However, this definition is being questioning and a clear diagnostic consensus has not been reached yet (Barbas et al. 2014; Villar et al. 2016).

The most common risk factor of ARDS development is sepsis, a severe systemic inflammatory response to infections of both pulmonary (pneumonia) and non-pulmonary origin (Cohen 2002; Rubenfeld et al. 2005). Additional risk factors for ARDS include severe trauma, aspiration, acute pancreatitis, and transfusion, among others (Stapleton et al. 2005; Gajic et al. 2011). Furthermore, age, gender, alcoholism, obesity, and diabetes also modify the predisposition to the syndrome (Gajic et al. 2011).

1.2 Epidemiology of ARDS

The incidence of ARDS is a matter of debate. It broadly varies among countries and even between studies, detecting the highest values in the USA (up to 82 cases per 100,000 person-years) (Li et al. 2011). The annual incidence of ARDS estimated in Europe is approximately ten times lower, mostly ranging between five and eight cases per 100,000 persons (Villar et al. 2014), although it can reach higher values in other population-based studies from Europe (Hughes et al. 2003). The fluctuations in the estimate of incidence between USA and Europe may be due to the fact that European intensive care units (ICUs) admit much less patients that those from USA (Cavallazzi et al. 2010). In Spain, the annual incidence of ARDS was estimated in seven cases per 100,000 persons (Villar et al. 2011), which is similar to the rate found in other European studies (Villar et al. 2014).

ARDS is also an important cause of morbidity and mortality in the ICUs worldwide, although mortality rates also differ between studies. The estimate of overall hospital mortality by ARDS is 40% on average (Pham and Rubenfeld 2017), reaching the highest values in patients with the most severe forms of the syndrome (ARDS Definition Task Force et al. 2012). This rate is also influenced by the ethnicity, with higher values among African-Americans, and by the gender of patients (Moss and Mannino 2002; Kangelaris et al. 2012). Furthermore, those patients who survive suffer from different ailments as a result of the syndrome, including muscle weakness and cognitive impairment (Mikkelsen et al. 2009; Fan et al. 2014).

1.3 Pathophysiology of ARDS

The progression to ARDS is a complex process marked by a severe inflammation that affects to the alveolar-capillary barrier. This barrier plays a key role in the proper gas exchange between the lung alveoli and the blood in the capillaries, and it is constituted by alveolar epithelial cells, capillary endothelial cells, and the extracellular matrix (Han and Mallampalli 2015; Herrero et al. 2018) (**Figure 1**). The pulmonary epithelium is a mechanical barrier where alveolar type I (ATI) and alveolar type II (ATII) pneumocytes protect from lung insults and contribute to the maintenance of the alveolar integrity (Guillot et al. 2013). Meanwhile, the pulmonary endothelium is a dynamic layer with metabolic properties that regulates the vascular homeostasis and plays an important role in inflammatory events (Block 1992). Furthermore, the alveoli contain resident macrophages that participate in the regulation of the immune response and inflammation in the lungs (Divangahi et al. 2015).

The pathogenesis of ARDS has been simplified in two phases: exudative and fibroproliferative. The exudative phase is characterized by alveolar inflammation mediated by different inflammatory markers, including tumor necrosis factor-alpha (TNF-α), interleukin (IL)-1, and IL-6 (Ware and Matthay 2000; Blondonnet et al. 2016) (**Figure 1**). In this acute stage, epithelial and endothelial damage occurs in the lungs, followed by the accumulation of protein-rich fluid in the interstitial and in the alveolar spaces that results in gas exchange impairment (Herrero et al. 2018). These changes are also accompanied by necrosis of ATI epithelial cells, the formation of hyaline membranes, and by an increase in the number of neutrophils in the lungs mediating the inflammatory response (Bellingan 2002). Together with neutrophils, monocytes and macrophages have also been found to have a key role in inflammation during ARDS (Aggarwal et al. 2014).

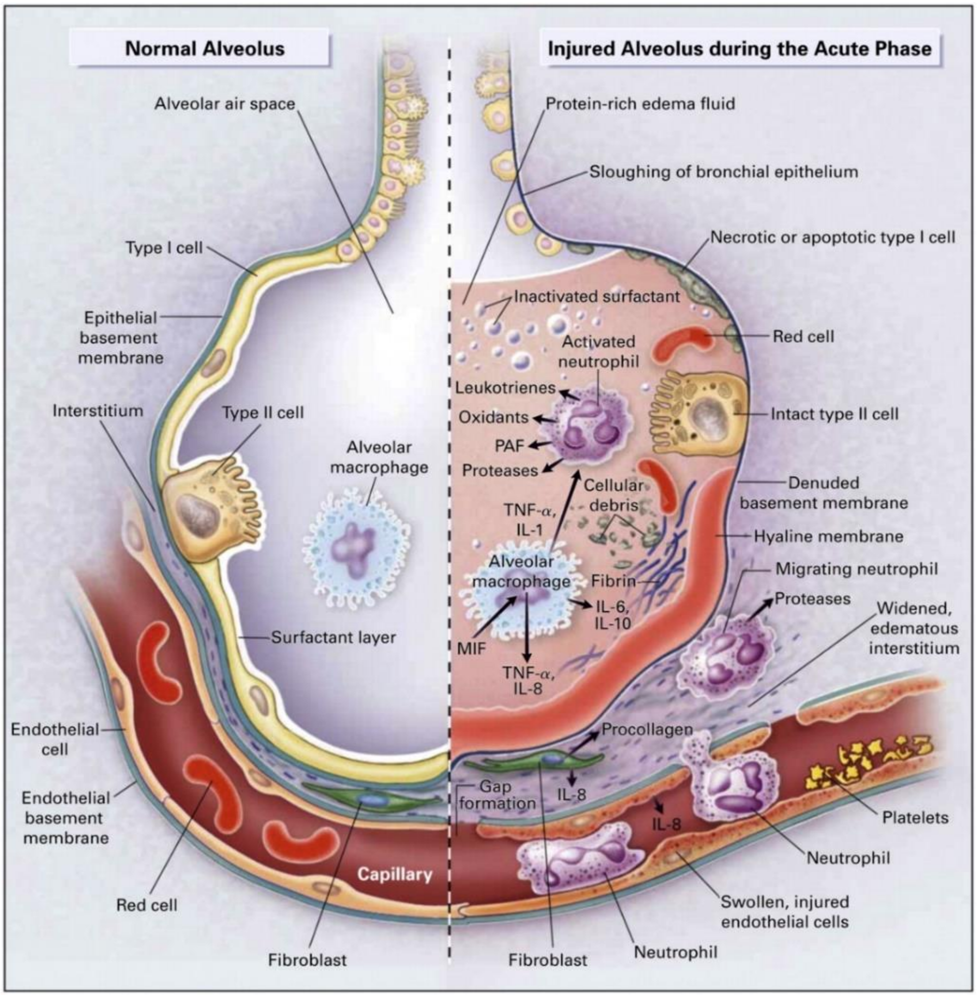


Figure 1. *Schematic representation of a healthy alveolus (left) and the alveolus after the lung injury (right). Reproduced with permission from Ware and Matthay, N Engl J Med 2000* (Ware and Matthay 2000)*, Copyright Massachusetts Medical Society.*

The fibroproliferative phase is characterized by a fibrotic process and the reduction of alveolar edema (Bellingan 2002). During this phase, ATII cells proliferate and cover the damaged epithelial surface, where they differentiate into ATI pneumocytes to reconstitute the alveolar epithelium (Bellingan 2002). ATII cells also secrete pulmonary surfactant lipids and proteins to contribute to the maintenance of lung homeostasis (Fujino et al. 2012). Additionally, fibroblasts proliferate and differentiate into myofibroblasts, driving the deposition of collagen and other extracellular matrix components in the lung (Quesnel et al. 2010) (**Figure 2**). The fibroproliferative phase can be followed by a resolution stage where the normal alveolar structure is restored (Hendrickson et al. 2015). This is facilitated by the decrease of the alveolar edema and a clearance of apoptotic cells, including neutrophils (Bellingan 2002). Nevertheless, a dysregulation of the fibrotic process can also occur, resulting in an excessive accumulation of collagen that is commonly associated with disease aggravation (Hendrickson et al. 2015) (**Figure 2**).

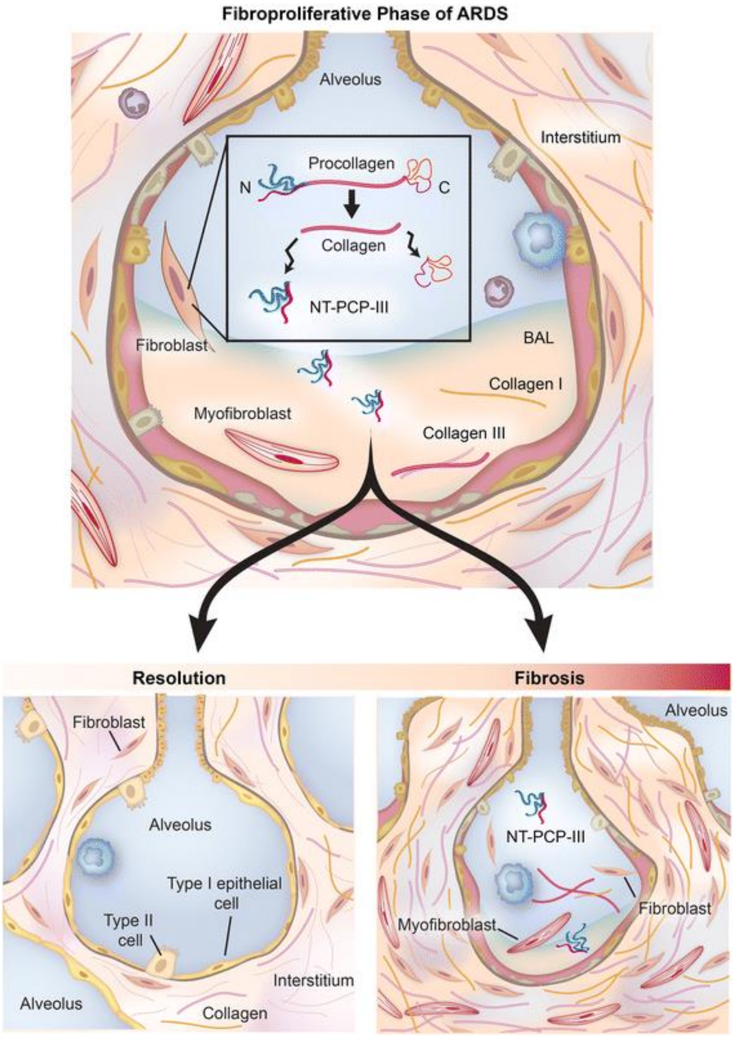


__Figure 2__. Fibroproliferative phase of ARDS and development. Reproduced with permission from Hendrickson and colleagues, Intensive Care Medicine 2015 (Hendrickson et al. 2015), Springer Nature License.

Despite the advances leading to an overall improvement of the incidence rates and mortality by ARDS, the mechanisms underlying this heterogeneous syndrome remain elusive, complicating the development of specific therapies (Bellingan 2002; Shaver and Bastarache 2014). The only therapeutic option currently available for ARDS patients is the use of MV under a regime of low tidal volume and a high level of PEEP (Acute Respiratory Distress Syndrome Network et al. 2000; Guo et al. 2018). Additionally, prone positioning of patients with ARDS can also improve their oxygenation and survival (Guérin et al. 2013). For this reason, numerous studies focus their efforts on the identification of novel therapeutic targets and biomarkers, necessary to develop more effective therapeutic strategies for ARDS patients.

1.4 Genetics and genomics of ARDS

As for other complex diseases, ARDS susceptibility and survival are expected to be conditioned by genetic factors, in addition to the environment. This is supported by twin studies and by genetic association studies (Sørensen et al. 1988; Acosta-Herrera et al. 2014). Accordingly, it has been reported that clinical factors alone do not predict the ARDS development or severity (Reilly et al. 2017). Thus, the identification of genes associated with the syndrome is crucial to further understanding the physiopathology of the disease, and to develop novel therapeutic, predictive, and prognostic options, with the aim of implementing precision medicine strategies. The genetic association studies are the most common genetic studies in ARDS. These are based in the comparison of allele frequencies for specific loci, typically single nucleotide polymorphisms (SNPs), between cases with ARDS and controls that do not develop the syndrome (Flores et al. 2008; Acosta-Herrera et al. 2014).

1.4.1 Candidate gene association studies of ARDS

Historically, candidate gene association studies are the most common approaches applied to unravel the genetics of ARDS, while an assessment at genomic levels remains practically unexplored (Hernández-Beeftink et al. 2019). These studies focus on particular genes selected based on a previous biological hypothesis of their implication in the disease, constituting a very narrow strategy with low replicability rates and difficulty of interpretation (Marigorta et al. 2018). Candidate gene studies in ARDS have involved genes linked to immune response, chemotaxis, response to the oxidative stress, cell proliferation, and cell signal transduction (Flores et al. 2008; Acosta-Herrera et al. 2014). Despite the limitations of these studies, a few associations with ARDS have been validated in different independent studies (Meyer et al. 2012). Some of these genes are *IL6*, *IL10*, vascular endothelial growth factor A (*VEGFA*), and angiotensin-converting enzyme (*ACE*) (Acosta-Herrera et al. 2014). IL-6, a cytokine encoded by *IL6*, is an important pro-inflammatory mediator of the exudative phase of ARDS (Blondonnet et al. 2016) and has been associated with ARDS development and with bad prognosis in patients with sepsis or ARDS (Meduri et al. 1995; Remick et al. 2005; Aisiku et al. 2016). Likewise, the serum levels of IL-10, encoded by *IL10*, have also been linked to ARDS development (Aisiku et al. 2016; Chen et al. 2018).

On the other hand, *ACE* and *VEGFA* are centrally involved in vascular permeability. The protein encoded by *ACE* catalyzes the conversion of angiotensin I into angiotensin II, which is a key mediator of arterial blood pressure (Patel et al. 2016). The renin-angiotensin system has been implicated in ARDS pathogenesis (Vrigkou et al. 2017) and reduced serum ACE levels have been correlated with severity of lung injury during ARDS (Fourrier et al. 1985). More interestingly, the role of *VEGFA* has been broadly related to ARDS development and progression (Barratt et al. 2014). The protein encoded by this gene (known as VEGF or VEGF-A) is highly expressed in healthy lungs (**Figure 3**) (Medford and Millar 2006; Voelkel et al. 2006).

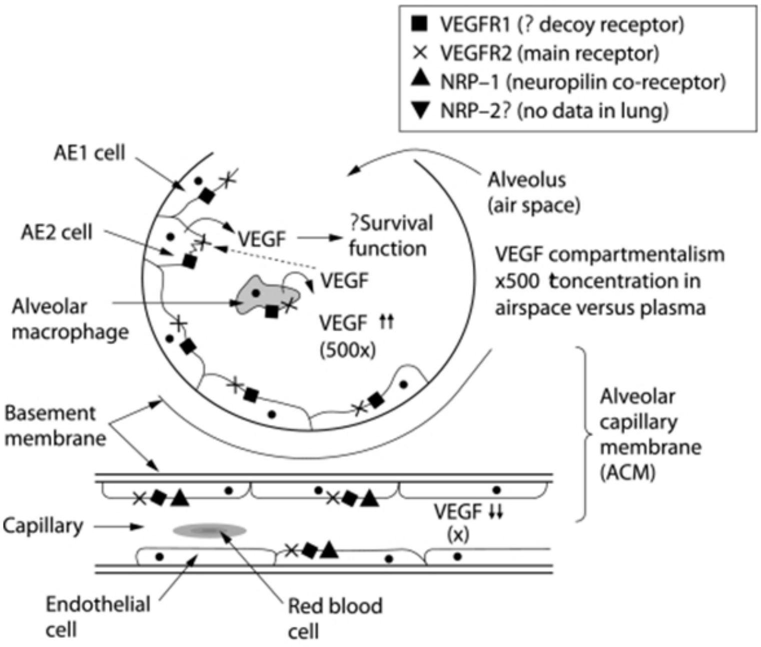


***Figure 3**. Expression of VEGF in the healthy lung. Reproduced and edited with permission from Medford, Thorax 2006* (Medford and Millar 2006)*, BMJ Publishing Group Ltd.*

In the context of ARDS, VEGF-A has been related to increased vascular permeability in lungs, as well as to the fibrosis process during the fibroproliferative phase of the syndrome (Barratt et al. 2014; Murray et al. 2017). However, despite the VEGF family seems to be a key element in ARDS physiopathology (Barratt et al. 2014), several studies report contradictory findings about the underlying mechanisms (Medford and Millar 2006), which makes difficult the development of therapeutic strategies targeting this pathway in critical patients.

1.4.2 Genomic studies of ARDS

Although candidate gene association studies have allowed to outline a catalogue of genes associated with ARDS, there is still a long way to unravel the genetics of ARDS. In this sense, those studies at genome-wide level have a greater potential to identify novel genes associated with the syndrome (Reilly et al. 2017). Genome-wide association studies (GWAS) represent a good alternative because a previous biological hypothesis is not required, analysis protocols are more standardized, and the chances of replicability are higher (Marigorta et al. 2018). Under the "common disease-common variant" hypothesis, which proposes that frequent genetic variants underlie common diseases in a population, GWAS allow to test the genetic association of a high proportion of frequent variants with a complex disease (Dehghan 2018). These studies are based in the use of commercial SNP arrays that allow genotyping hundreds of thousands of selected SNPs located along the genome (Dehghan 2018). Additionally, to improve the statistical power of the analyses, many other million variants that are known to correlate with the genotyped variants (i.e. that are in high linkage disequilibrium (LD)) can be imputed using reference datasets from different studies, including the Haplotype Reference Consortium (HRC) (McCarthy et al. 2016) and the 1000 Genomes Project (1KGP) (1000 Genomes Project Consortium et al. 2015).

Despite their evident potential, there are only two GWAS of ARDS published to date. Christie and colleagues performed a GWAS of trauma-associated ARDS on 2,866 individuals of European ancestry (600 cases and 2,266 controls) and replicated the association of SNPs with $p<0.01$ in an independent dataset (Christie et al. 2012). Results were followed up by a functional evaluation phase based on expression quantitative trait loci (eQTL) analyses, which allowed them to identify a variant associated with the mRNA expression of the PTPRF interacting protein alpha 1 (*PPFIA1*) gene. On the other hand, Bime and colleagues performed a GWAS on 232 African-American ARDS patients and 162 at-risk controls followed by a biological pathway analysis to prioritize variants (Bime et al. 2018). They identified coding variants in the selectin p ligand (*SELPLG*) gene that were associated with ARDS risk. However, there was no evidence of replication for any of the assessed SNPs in the independent sample. The authors also conducted solid functional analyses using animal models that reinforced the role of *SELPLG* in ARDS susceptibility, likely involving an increase in *SELPLG* gene expression.

In addition to the GWAS, next-generation DNA sequencing (NGS)-based approaches are increasingly used to reveal novel genes related to complex diseases (Petersen et al. 2017). These studies allow to determine the sequence of nucleotides in DNA extracted from different biological samples. As a result, the whole spectrum of risk allele frequencies can be assessed, including rare genetic variants that cannot be determined by GWAS because they are not catalogued in the reference datasets, which provides a better understanding of the genetics of a disease (Petersen et al. 2017). Nowadays, and because of the cost-efficiency and the insight into the interpretable genome, whole-exome sequencing (WES) is the NGS approach that is most frequently used for the study of genetic risk factors associated with human diseases. WES allows to obtain the nucleotide sequence of the exons from most of the protein-coding genes, which are estimated to harbor about 85% of mutations related to human diseases (Majewski et al. 2011).

Two small WES studies have been performed so far in ARDS patients. On the one hand, Lee and colleagues sequenced the exome of 88 individuals with sepsis-associated ARDS and reported that the myosin light chain kinase (*MYLK*) gene was the top-ranked when correlated with ARDS severity, measured by ventilator-free days (VFD) (Lee et al. 2012). This gene had already been associated with ARDS in previous candidate gene association studies (Gao et al. 2006; Christie et al. 2008), being related to the inflammatory response during ARDS. On the other hand, Shortt and colleagues performed WES in DNA samples from 96 patients with ARDS and compared it with data from 1KGP (Shortt et al. 2014). As a result, they identified three novel genes related to ARDS susceptibility, severity and outcomes, including the arylsulfatase D gene (*ARSD*), the XK, Kell blood group complex subunit-related family, member 3 gene (*XKR3*), and the zinc-finger protein 335 (*ZNF335*). Whole-genome sequencing (WGS) studies of ARDS are still lacking from the literature (Hernández-Beeftink et al. 2019), likely due to the high associated costs.

1.5 The microbiome and disease

1.5.1 The human microbiome

It has been estimated that the human body is colonized by up to 100 trillion symbiotic microbial cells (Qin et al. 2010), including bacteria, viruses, archaea, fungi, and other eukaryotes (Lloyd-Price et al. 2016). This collection of microorganisms is referred to as the human microbiota, which is organized in complex communities that can adapt to environmental changes and is involved in human processes such as metabolic functions, epithelial development, and the immune response (Wang et al. 2017). To conduct these functions, the human microbiota carries more than 100-fold more genes that the human genome (Qin et al. 2010). The catalog of genes these microbial cells harbor is known as human microbiome (Ursell et al. 2012).

In 2009, an initiative called the Human Microbiome Project (HMP) emerged to characterize the human microbiome by studying samples from different human tissues of healthy individuals using high-throughput technologies (Peterson et al. 2009). Additionally, the HMP seeks to determine the correlation between changes in microbiomes and health or disease, as well as to provide a standard database of microbial genomes and of new technologies and tools to enable the data analyses (Peterson et al. 2009). Most of research studies performed so far have been focused on the gastrointestinal tract, where most of the human microbiota resides (Lloyd-Price et al. 2016). Besides, other body sites such as the skin, oral cavity, placenta, urogenital system, and lung also have their specific microbiomes, which are also being studied because of their implication in disease (Lloyd-Price et al. 2016).

1.5.2 Metagenomics

Historically, the study of the human microbiota has focused on traditional culture-based methods of single microorganisms from complex biological samples. These cultures require prior knowledge of the metabolic necessities of the bacteria to be grown, which implies that a huge proportion of the bacteria is yet uncultivable. Specifically, only 1% of the total microbial population can be cultured (Pham and Kim 2012). Furthermore, microbial cultures require restricted media and cultivation conditions to grow each microorganism, limiting the study of their complex natural environments (Pham and Kim 2012). As an alternative, a culture-independent approach based on the study of microbial DNA has emerged to overcome the deficiencies of conventional microbiological methods based on isolated microorganisms. This approach is known as metagenomics and consists in the genomic study of the collective microorganisms present in environmental samples, such as biological tissues and fluids (Handelsman 2004; Tringe and Rubin 2005). Metagenomics is now based in the use of NGS technologies that allow to obtain the DNA sequence of microorganisms, including uncultured microbials, with the aim of revealing the microbial composition of complex systems and studying microbial changes between groups of samples (Pflughoeft and Versalovic 2012). Metagenomic studies have frequently utilized random DNA sequencing (shotgun) or targeted gene sequencing (Ursell et al. 2012), both mainly focused on the bacterial DNA assessment. Shotgun consists in the untargeted sequencing of microbial DNA extracted from an environmental sample and subsequently sheared into small fragments (Quince et al. 2017). This results in overlapping sequence segments (i.e. reads) that are preprocessed and classified to obtain the microbiome profile. Finally, statistical analyses are performed to evidence differences between the different samples (Quince et al. 2017). This approach is independent of DNA amplification and allows to examine thousands of organisms in parallel, as well as a taxonomic classification at (sub)species level (Nayfach and Pollard 2016; Ranjan et al. 2016).

Historically, metagenomic studies have been performed using targeted gene sequencing technologies, although, strictly speaking, these do not involve the analysis of the whole genome (Ursell et al. 2012). The 16S ribosomal RNA bacterial gene (16S rRNA) (~1,500 bp) is the most commonly used marker in these studies because of its utility to differentiate among bacterial taxa (Janda and Abbott 2007). The RNA encoded by this gene is a component of the 30S small subunit of the prokaryotic ribosome, involved in protein synthesis (Mizrahi-Man et al. 2013). The 16S rRNA contains many conserved regions and nine hypervariable regions (V1-V9) that allow to distinguish among different bacteria (Chakravorty et al. 2007). The 16S rRNA sequencing is based on the amplification by polymerase chain reaction (PCR) of one or few of these hypervariable regions of 16S rRNA using flanking primers, with V3 and V4 being the most frequently evaluated regions, as these have proven to be the most informative (Mizrahi-Man et al. 2013). Once the amplified products (amplicons) are sequenced, a bioinformatic analysis must be performed. Among others, sequence reads must be pre-processed, filtered, and grouped into operational taxonomic units (OTUs) (Mizrahi-Man et al. 2013). Finally, the taxa assignment is conducted based on reference data and diversity analyses are performed using specific software to compare samples (Caporaso et al. 2010). A schematic representation of a typical 16S rRNA sequencing procedure is shown in **Figure 4**. For the study of fungi, a specific region of the 18S rRNA of these organisms is sequenced instead, although it fails to adequately fully cover fungal diversity (Soeta et al. 2009; Wang et al. 2014; Budden et al. 2019).

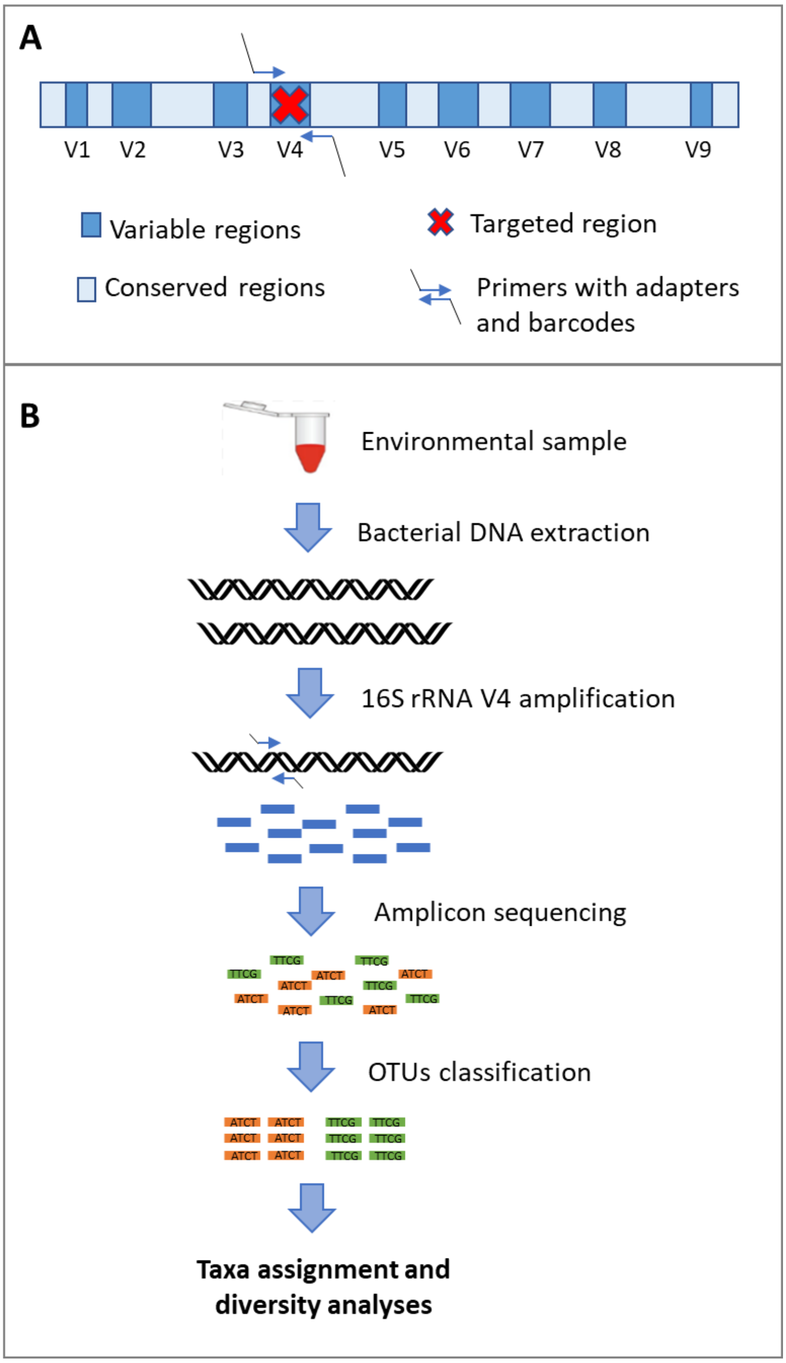


***Figure 4**. Schematic strategy for 16S rRNA V4 sequencing. A) Primers design for amplification of 16S rRNA V4. B) Procedure to obtain the microbiome profile from environmental samples through 16S rRNA V4 sequencing. OTUs, operational taxonomic units.*

1.5.3 The microbiome and critical illness

In the last decades, the study of the implication of the human microbiome in health and disease has been rapidly increased thanks to the use of high-throughput sequencing technologies and the possibilities of bioinformatics analyses (Cox et al. 2013). The importance of these studies is based in the need to identify novel prognostic methods and more efficient therapies. In this sense, perturbations in human microbial populations (microbial dysbiosis) have been broadly related to the development of complex diseases and to immune dysregulation (Pflughoeft and Versalovic 2012). Diseases such as Crohn's disease, diabetes, obesity, inflammatory bowel disease, atopic dermatitis, and metabolic syndrome have been associated with changes of the normal microbiome (Pflughoeft and Versalovic 2012; Althani et al. 2016). Furthermore, numerous studies have linked this microbial dysbiosis to infectious diseases and critical illness (Caverly et al. 2015; McDonald et al. 2016; Jacobs et al. 2017), and a reduced microbial diversity has been related to patient severity in the ICU (Lamarche et al. 2018).

Among critical diseases, a large part of the studies conducted to date have focused on the implications of the microbiome in patients with sepsis, by mainly assessing the gut microbiome (Haak and Wiersinga 2017). Accordingly, the gut flora has been linked to sepsis complications and mortality by systemic inflammatory response syndrome (SIRS) (Shimizu et al. 2011). Additionally, recent studies support that microbial dysbiosis processes in blood, nasal cavity, and the lungs are related to sepsis susceptibility and/or severity (Dickson 2016; Gosiewski et al. 2017; Tan et al. 2019). Furthermore, Dickson and colleagues reported that the lung microbiome of critical patients is significantly altered and enriched in bacteria commonly found in the gastrointestinal tract, probably because the translocation of microorganisms from the gut to the patient's lungs (Dickson et al. 2016). Given the relevance of these results, further studies must be performed to explore their potential translation into clinical practice.

1.6 Implication of the genetic ancestry in disease

1.6.1 Genetic ancestry and critical illness

The prevalence of numerous diseases has been shown to be different across ethnic groups (National Research Council 2004). Particularly, several studies highlight the relationship between the ancestry of an individual and the risk to develop critical illnesses, including sepsis and ARDS (Moss and Mannino 2002; Soto et al. 2013; Sandoval and Chang 2016). In this context, it was reported that African Americans were more likely to be admitted to the ICU than individuals of European ancestry (Dombrovskiy et al. 2005). African Americans also had the highest risk for sepsis development (Martin et al. 2003; Barnato et al. 2008; Mayr et al. 2010). Furthermore, higher rates of hospital mortality by sepsis have been evidenced for African-American and Latin-Americans compared to European patients (Martin et al. 2003; Jones et al. 2017). Accordingly, the mortality rates of African-American patients with ARDS are higher than those of individuals of European ancestry (Moss and Mannino 2002; Erickson et al. 2009).

These disparities cannot be explained solely by socioeconomic factors (Moss and Mannino 2002; Esper et al. 2006; Soto et al. 2013). Recent studies have shown that genetic ancestry can influence the development and outcomes of complex diseases (National Research Council 2004), including respiratory diseases (Kumar et al. 2010; Flores et al. 2012; Rumpel et al. 2012; Vergara et al. 2013; Hernandez-Pacheco et al. 2016) and critical illnesses (Soto et al. 2013). In this sense, genetic variants linked to the ancestry could be affecting the response to infection and inflammatory processes in critical care patients (Soto et al. 2013). For example, genetic variants known to increase the expression of proinflammatory cytokines (such as IL-1A, IL-1B, IL-6, and IL-18) and to reduce the expression of the anti-inflammatory cytokine IL-10 were more frequent in African-American women compared to women of European ancestry (Ness et al. 2004). Additionally, polymorphisms in *MYLK* that were related with a higher risk of sepsis and ARDS have shown distinct allele frequencies between populations of different ethnicities (Gao et al. 2006; Christie et al. 2008), and a genetic variant in the Duffy antigen/receptor for chemokines (*DARC*) gene was associated with worse clinical outcomes in African American patients with mild ARDS (Kangelaris et al. 2012).

Furthermore, polymorphisms in the human Toll-like receptor (TLR) 2 gene (*TLR2*), which is involved in pathogen recognition and inflammatory responses, have been revealed to confer differences between racial groups (Yim et al. 2004). Accordingly, a study assessing *TLR4* polymorphism haplotypes revealed higher allele frequencies in sub-Saharan African populations, suggesting that it was related to protection against mortality from malaria as a consequence of an evolutionary pressure in this population (Ferwerda et al. 2007). Interestingly, this gene had been correlated with susceptibility to infectious diseases (Agnese et al. 2002) and increased mortality to septic shock (Lorenz et al. 2002). A similar scenario, where polymorphism frequencies have shifted because of natural selection processes, has been found in other studies (Stephens et al. 1998; Taylor et al. 2012; Meyer et al. 2018). Remarkably, it is well known that signatures of natural selection are found in the human leukocyte antigen (HLA) system, involving both positive and balancing selection (Meyer et al. 2018). As is the case of *TLR4*, alleles within *HLA-B* have been related to malaria protection in African populations (Sanchez-Mazas et al. 2017).

Based on the evidence, it has become apparent that the genetic makeup of a population has important consequences in the predisposition to complex diseases and for drug response (Wilson et al. 2001; Botigué et al. 2013). In fact, the ethnicity of an individual is starting to be considered in clinical practice for precision medicine (Li et al. 2009; Dean 2012). For example, polymorphisms in vitamin k epoxide reductase complex subunit 1 (*VKORC1*) and in the cytochrome P450 family 1 subfamily c member 9 (*CYP2C9*), which influence the metabolism of warfarin, a widely used anticoagulant, have been found at distinct frequencies across populations (Li et al. 2009). As a result, the warfarin dose can be adjusted in patients based on their ethnicity and genotypes (Li et al. 2009). A similar situation is found for the clopidogrel therapy, an antiplatelet agent, and the *CYP2C19* genotype (Dean 2012).

1.6.2 Estimation of the genetic ancestry in a recently admixed population

The stratification of the genetic variation among populations across the world is mainly explained by genetic drift and migration, but also by the existence of selective pressures that are related to past or ongoing local adaptations (Seldin et al. 2011). Therefore, genetic studies in recently admixed populations have a huge potential to evaluate the influence of the genetic ancestry in diseases, with the final objective of identifying novel candidates to be evaluated as genetic risk factors of diseases, including critical illnesses (Seldin et al. 2011). In this sense, a profound characterization of the genetic structure of the assessed population is needed (**Figure 5**) (Thornton and Bermejo 2014).

Genetic ancestry estimators involve the use of genetic data from the putative ancestral populations that were mixed in the past recent history. Thus, prior knowledge of the historical admixture of the population is required (Alexander et al. 2009). There are two main types of ancestry estimators of the genome: global and local. The global ancestry is the overall genetic ancestry of an individual (**Figure 5A**) (Thornton and Bermejo 2014) and can be estimated by means of programs such as ADMIXTURE, one of the most commonly used algorithms, which uses a maximum likelihood model to obtain individual ancestry proportions based on multiple unlinked SNPs (Alexander et al. 2009). The local ancestry refers to the genetic ancestry of an individual at a given chromosomal locus (Thornton and Bermejo 2014). As a result of the admixture of different populations, the genome of recently admixed individuals becomes a mosaic composed by different chromosomal fragments or ancestry blocks, each derived from an ancestral population (**Figure 5B**) (Tang et al. 2006). The longer the time since the admixture, the shorter the size of the ancestry blocks. Therefore, the local ancestry can be estimated analyzing this mosaic comparing it with genomes from reference populations. Three of the algorithms that are commonly used to estimate the local ancestry are LAMP-LD (Baran et al. 2012), RFMIX (Maples et al. 2013), and ELAI (Guan 2014). LAMP-LD and RFMIX are based on haplotype transitions from the parental population; hence they depend on a previous step for haplotype reconstruction. These software tools have been broadly used to estimate the local ancestry in Latino populations (Padhukasahasram 2014; Eyheramendy et al. 2015; Sofer et al. 2017; Spear et al. 2019). Contrarily, ELAI is a more recent method that directly uses genotype data, without needing a phasing step. This software can detect small ancestry tracts and has been used to characterize the complex admixture of South African populations (Pierron et al. 2018; Williams et al. 2018).

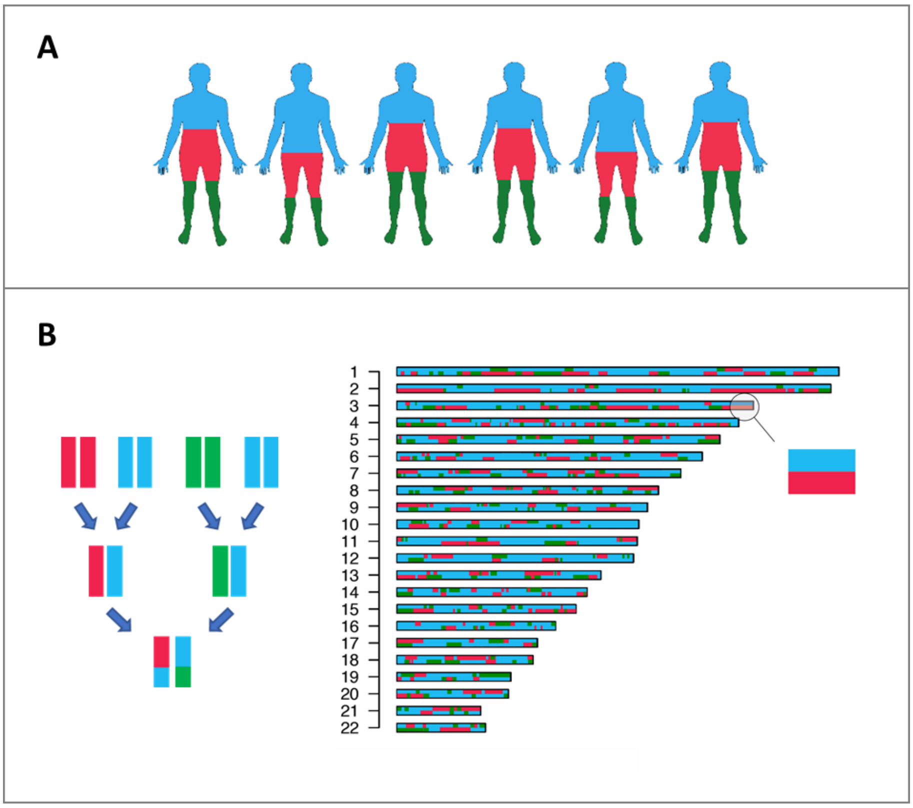


***Figure 5**. Schematic representation of the genetic ancestry of a recently admixed population with three ancestral populations (plotted in pink, blue, and green colors). A) Global ancestry proportions for six admixed individuals. B) Recombination events between haplotypes of different ancestries (left) and local ancestry estimations for an individual genome (right).*

In case a disease risk variant has distinct allele frequencies across ancestral populations and confers a different disease susceptibility, a deviation in local ancestry can be found at that locus in a recently admixed population (Mani 2017; Shriner 2017). Consequently, local ancestry estimates can be used to identify genomic regions where ancestry tends to be coinherited with a specific disease. This analysis is known as admixture mapping, which allows to reveal novel disease genes that show differential risk by ancestry (Patterson et al. 2004; Shriner 2017). This kind of studies can be performed only in recently admixed populations. However, it has the advantage over general association studies in that, as local ancestry blocks are larger than haplotype blocks, the correction by multiple tests is less restrictive and the statistical power to detect disease signals is increased, allowing to use more reduced sample sizes to attain a given power (Shriner et al. 2011). Conversely, large regions are identified by means of this method, and additional studies are required to identify the underlying risk variants (Shriner 2017).

1.6.3 The Canary Islands population in genetic ancestry studies

The current inhabitants of the Canary Islands have a unique genetic admixture that makes it suitable to be considered in genetic ancestry studies. The previous evidence supports that the aboriginal population from the Canarian archipelago (collectively known as Guanches) had a North African (NAF) origin (Hooton 1970; Onrubia Pintado 1987). The Spanish conquest took place during the XVth century (de Abreu Galindo and Cioranescu 1977), which resulted in an admixture of the aborigines with the European population (EUR) (de Abreu Galindo and Cioranescu 1977), as well as with sub-Saharan Africans (SSA), due to the flourishing slave trade occurring at that historical moment (Lobo-Cabrera 1993). This admixture has been shown by classical molecular studies focused on blood groups, red blood cell enzymes, or with a few of polymorphic *Alu* insertions (Flores et al. 2001; Maca-Meyer et al. 2004). These estimated that ancestry proportions of the Canary Islanders were 62-78% EUR, 20-38% NAF, and 3-10% SSA. Accordingly, other studies analysing a reduced number of SNPs or a limited sample size (Pino-Yanes et al. 2011; Botigué et al. 2013) revealed comparable proportions of ancestry, finding 75-83% EUR, 17-23% NAF, and less than 2% SSA. However, none of these studies have evaluated the disease implications of such admixture scenario in the Canary Islanders.

The Canary Islands population have an increased prevalence of different chronic diseases, including cardiovascular diseases, such as diabetes, obesity, and hypertension (Cabrera de León et al. 2006; Bueno et al. 2008; Marcelino-Rodríguez et al. 2016), and respiratory diseases, such as asthma (Sánchez-Lerma et al. 2009; Juliá-Serdá et al. 2011), when compared to other mainland Spanish populations. This disease burden could be affected, in addition to environmental factors, by the distinctive genetic admixture of this population. Furthermore, due to the historical isolation of the Canary Islands, one would expect that the genomes of the current inhabitants have an enrichment in low-frequency functional variants (Xue et al. 2017), resulting in an increased number of recessive variants that could confer risk to specific complex diseases (Campbell et al. 2007; Moltke et al. 2014; Ghani et al. 2015). Additionally, mutations underlying monogenic disorders would be expected in the Canary Islanders. For example, a founder mutation in the alanine-glyoxylate and serine-pyruvate aminotransferase (*AGXT*) gene has been associated with a high prevalence of type 1 primary hyperoxaluria in La Gomera (Santana et al. 2003; Lorenzo et al. 2006; Lorenzo et al. 2014). Additionally, a high percentage of patients from La Palma carry the most frequent mutation in the Fanconi anemia complementation group A (*FANCA*) gene, which perhaps explains the high incidence of sickle-cell anemia in this population (Castella et al. 2011). Finally, an increased prevalence of Wilson disease has been related to a variant of the ATPase copper transporting beta (*ATP7B*) gene in individuals from Gran Canaria, and other variants have been associated with cardiovascular traits in the same population, pinpointing again the singular genetic characteristics of Canary Islanders (García-Villarreal et al. 2000; Rodríguez-Esparragón et al. 2017). Therefore, the characterization of the genome of the Canary Islanders would be important for revealing gene regions that are distinctive of this population and that may be linked to particular disease risks, including critical illnesses, as well as for designing subsequent admixture mapping studies in this population.

2. Hypothesis and objectives

2. Hypothesis and objectives

Despite numerous studies have tried to disentangle the biological complexity of ARDS, there is still a lack of specific treatments and effective prognostic methods for critical care patients. Given that this syndrome is influenced by both environmental and genetic factors, **the hypothesis of this work** is that the use of different genomic approaches will provide complementary information to better understand the pathophysiology of ARDS and of main risk factors, revealing novel therapeutic and prognostic options.

The specific objectives of this work are:

1. To perform a systematic review of all published studies reporting associations of genetic variants with ARDS susceptibility and outcomes.

2. To identify novel common genetic variants associated with ARDS susceptibility by means of the first GWAS in patients with sepsis-associated ARDS.

3. To identify if there are lung microbiome shifts in patients with sepsis associated with aggravation and to evaluate its prognostic utility.

4. To characterize the recent evolutionary history of current inhabitants of the Canary Islands based on genome-wide data to identify links between genetic ancestry and risks of sepsis and ARDS, and to lay the foundation for designing admixture mapping studies in this particular population.

3. Chapters

The methods of Chapter 1 are detailed on the introductory page to the chapter. Chapters 2, 3, and 4 include their specific introduction, methods, results, discussion and conclusions. All chapters also include a section of references.

Chapter 1.

Systematic review of the genetics of ARDS

A bibliographic review of all the genetic association studies in ARDS published from September 2012 to December 2015 is reported in this chapter. The quality assessment of these studies was based on diverse criteria considered in previous works performed by this research group, where the quality of genetic association studies from 1996 to 2012 had been previously evaluated in two different articles (Flores et al. 2008; Acosta-Herrera et al. 2014). The final objective of this work was to complete a systematic search of all genes reported to be significantly associated with ARDS susceptibility or outcomes.

We conducted a search in PubMed using the following combinations of terms: “acute respiratory distress syndrome” AND “polymorphism”, “acute respiratory distress syndrome” AND “genetic variant”, “ARDS” AND “polymorphism”, “ARDS” AND “genetic variant”, “acute lung injury” AND “polymorphism”, and “acute lung injury” AND “genetic variant”. All references were manually revised, and those genes harboring genetic variants nominally associated with ALI/ARDS susceptibility or outcomes (significance of $p \leq 0.05$) were considered. As a result, we found a total of 81 candidate genes that had been associated with ARDS until December 2015, most of them involved in immune response and vascular permeability. The association of only seven of these genes was validated in at least four independent samples. This assessment supports the low replicability of this type of studies and the difficulties in the interpretation of results, as well as the need of implementing genomic approaches to identify novel ARDS risk genes, accompanying the results with functional studies. In this sense, as reported in this chapter, only one GWAS and two WES studies in ARDS patients had been published until 2015.

This chapter was published in 2016 in *eLS* with the title "Genetics of Acute Respiratory Distress Syndrome" (doi: 10.1002/9780470015902.a0026533). It is reproduced under the terms of John Wiley and Sons License.

Genetics of Acute Respiratory Distress Syndrome

Advanced article

Online posting date: 15th April 2016

Beatriz Guillén-Guío, *Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain*

Marialbert Acosta-Herrera, *CIBER de Enfermedades Respiratorias (Instituto de Salud Carlos III), Research Unit, Hospital Universitario de Gran Canaria Dr. Negrín, Las Palmas de Gran Canaria, Spain*

Jesús Villar, *CIBER de Enfermedades Respiratorias (Instituto de Salud Carlos III), Research Unit, Hospital Universitario de Gran Canaria Dr. Negrín, Las Palmas de Gran Canaria, Spain*

Carlos Flores, *CIBER de Enfermedades Respiratorias (Instituto de Salud Carlos III), Research Unit, Hospital Universitario Nuestra Señora de Candelaria, Santa Cruz de Tenerife, Spain*

The acute respiratory distress syndrome (ARDS), a diffuse lung inflammation leading to an acute hypoxemia, is a complex syndrome induced by a systemic inflammatory response commonly caused by severe infections or trauma. Despite the heterogeneous mechanisms underlying disease causality, genetic risk factors involved in ARDS susceptibility and outcomes are being identified. However, a full characterisation of the genetic architecture of this syndrome remains to be completed. Novel 'omics' tools have the promise for recognising ARDS subtypes, which will allow the identification of novel risk factors that will translate into individualised patient management and treatments, potentially leading to better individual prognosis. An important direction for future research is to encourage the use of large and well-characterised samples, ensure that patients of diverse ancestry are included in genetic studies and establish the clinical utility of risk variants identified for prevention, therapy or risk stratification.

eLS subject area: Genetics & Disease

How to cite:
Guillén-Guío, Beatriz; Acosta-Herrera, Marialbert; Villar, Jesús; and Flores, Carlos (April 2016) Genetics of Acute Respiratory Distress Syndrome. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0026533

Introduction

Definition

The acute respiratory distress syndrome (ARDS), a severe form of acute lung injury (ALI), is a clinical disease process characterised by an acute onset of diffuse lung inflammation caused by an insult to the alveolar–capillary membrane that results in increased permeability of pulmonary capillaries and alveoli, and subsequent formation of interstitial and alveolar oedema (Bernard *et al*., 1994). ARDS usually develops in adult patients with predisposing conditions that induce a systemic inflammatory response. Sepsis, both of pulmonary and nonpulmonary origin, is the most common cause of ARDS. ARDS is induced in 46% of the cases by pulmonary entities, but there are other clinical conditions, including severe trauma or acute pancreatitis, and some secondary factors such as alcoholism and obesity, which also increase its risk (Moss *et al*., 1996; Gajic *et al*., 2011). The mechanisms by which a wide variety of insults can lead to this syndrome are not clear. Independent of the clinical disorders associated with ARDS, its pathogenesis can be a result of both direct and indirect insults (i.e. by an acute systemic inflammatory response) to the lungs.

There is no typical ARDS patient. Diagnosis of ARDS is based on a combination of clinical, oxygenation, hemodynamic and radiographic criteria. ARDS is now defined, based on the Berlin definition, as an acute hypoxemia and the presence of bilateral pulmonary infiltrates on chest radiographs not fully explained by cardiac failure or fluid overload. Acute hypoxemia is defined by a PaO_2/FiO_2 ratio ≤300 mmHg, where PaO_2 is the partial pressure of oxygen in arterial blood, and FiO_2 is the fractional concentration of inspired oxygen. Thereby, ARDS can be classified as mild ($200 < PaO_2/FiO_2 \leq 300$), moderate ($100 < PaO_2/FiO_2 \leq 200$) or severe ($PaO_2/FiO_2 \leq 100$). Until the Berlin definition was established, the group of patients with mild

ARDS was termed ALI. However, there is no consensus in the scientific literature and both terms continue to be used in clinical (Ranieri *et al*., 2012) and animal model studies.

Epidemiology

Despite recent advances in the study of ARDS, its incidence and mortality remain without consensus. This may be due to the presence of several definitions of the syndrome, differences in demographics and variations in study designs. Estimates of ARDS incidence in diverse recent clinical observational studies vary, by country and year, from 4.9 to 82.4 cases per 100,000 person-years, roughly being seven times higher in the United States than in remaining assessed countries (Villar *et al*., 2011a; Buregeya *et al*., 2014) (**Table 1**).

ARDS is a common cause of death in adult intensive care units (ICU) causing an overall hospital mortality of 44% on average. Mortality varies depending on factors such as age, aetiology of the lung injury and by the presence of nonpulmonary organ dysfunction (Suchyta *et al*., 1997; Erickson *et al*., 2009; Villar *et al*., 2011a). Besides, the ethnicity might influence the mortality risk, which may be attributable to underlying genetic differences among populations and/or to distinct environmental factors. It has been reported that African-American ARDS patients have a higher mortality risk than European descent patients (Moss and Mannino, 2002; Kangelaris *et al*., 2012).

While ARDS has been classically considered an acute complication, recent studies have shown that patients surviving the acute process develop long-term complications, frequently experiencing a reduction in their quality of life. They usually suffer long-term cognitive impairment, as well as physical impairment or psychiatric disorders. However, there are currently no strategies to improve quality-of-life outcomes after ARDS (Spragg *et al*., 2010).

Pathophysiology

Although the molecular mechanisms leading to ARDS are complex and remain unclear, the syndrome develops owing to an aggressive inflammatory process, resulting in increased vascular permeability, oedema formation, surfactant depletion and pulmonary fibrosis (Ware, 2006). Similar to any form of inflammation, ARDS represents a complex process in which multiple pathways can propagate or inhibit lung injury.

The pathophysiologic process of ARDS can be divided into three phases: exudative, fibroproliferative and chronic/resolution. In the exudative phase, there is evidence of a rapid interstitial and alveolar oedema, which reflects the injury of the lung capillary endothelium and the alveolar epithelium. It leads to increased permeability of the alveolar-capillary barrier and alveolar flooding by a protein-rich fluid as well as decreased surfactant production, so that normal gas exchange is impaired. In this phase, there is an acute inflammatory response accompanied by a marked accumulation of active neutrophils in the lungs. Some patients recover in this acute phase with gradual resolution of the oedema and the acute parenchymal inflammation without fibrosis (Ware, 2006; Carlucci *et al*., 2014).

In the fibroproliferative phase, the alveolar oedema decreases, the alveolar space becomes filled with neutrophils and macrophages that intensify the release of inflammatory mediators, and the alveolar epithelium is repopulated by the proliferation and differentiation of alveolar epithelial type II cells. Finally, it takes place chronic inflammation, neovascularisation and fibrosis, as recognised by the deposition of collagen and other material from the extracellular matrix. Epithelium is repaired by type II alveolar epithelial cells that proliferate to cover the injured basement membrane and differentiate into type I cells. Furthermore, there is a resolution phase of the neutrophil-mediated inflammation through the clearance of neutrophils from the injured lung (Ware, 2006; Carlucci *et al*., 2014). However, clinical studies have established that alveolar fluid clearance is impaired in most patients with ARDS (Ware and Matthay, 2001), which may have long-term consequences in patients surviving ARDS. It is unknown why some patients can rapidly resolve the acute inflammation while others progress to the chronic phase or why fibroproliferative changes are rapidly developed in some cases and not in others.

Molecular mechanisms involved in ARDS

There is currently no specific treatment for ARDS. Once it develops, lung protective mechanical ventilation (MV) is the only

Table 1 Estimates of incidence and hospital mortality in acute respiratory distress syndrome in clinical observational studies[a]

Country	Study sample	Incidence (per 100.000/year)	Hospital mortality (%)	Reference
Finland	59	4.9	42.0	Valta *et al*. (1999)
Sweden, Denmark and Iceland	1.515	13.5	41.2	Luhr *et al*. (1999)
Australia	1.977	28.0	34.0	Bersten *et al*. (2002)
United Kingdom	38.116	16.0	60.9	Hughes *et al*. (2003)
United States	4.251	58.7	41.1	Rubenfeld *et al*. (2005)
Spain	3.462	7.2	47.8	Villar *et al*. (2011a)
United States	8.034	82.4 (2001)–38.9 (2008)[b]	34.8	Li *et al*. (2011)
Brazil	7.133	6.3	55.5	Caser *et al*. (2014)

[a] Data referred to studies utilising the definition based on The American-European Consensus Conference on ARDS (Bernard *et al*., 1994).
[b] Data corresponding to estimates obtained in 2001 and 2008.

lung-directed intervention known to affect patient survival at the moment.

In the past years, animal models have provided a way to expose the optimal mode to ventilate critically ill patients. A commonly studied model is the ventilator-induced lung injury (VILI), which reproduces most physiological and pathological hallmarks of ARDS, recognising that the ventilator can cause also ALI that is undistinguishable from ARDS. These studies suggested very early that MV with high tidal volume (HVT) was able to induce lung injury and initiate or augment the inflammatory response (Villar *et al*., 2011b). Currently, with the use of low tidal volume (LVT) MV in the range of 4–8 mL/kg predicted body weight and moderate to high levels of positive end-expiratory pressure (PEEP) in the range of 8–14 cmH_2O, injured lungs can be partially protected, resulting in the reduction of ventilator-associated complications and better hospital survival rate of patients with ARDS (The Acute Respiratory Distress Syndrome Network, 2000).

Functional genomics had an essential role in our current understanding of the molecular changes occurring in the setting of ARDS and the identification of key genes potentially contributing to its predisposition. One of the most commonly used tools for functional genomics studies of ARDS has been the assessment of gene expression in tissue samples (from clinical, animal or *in vitro* studies) in DNA microarrays. This technique aims to analyse simultaneously the mRNA expression levels of thousands of genes to identify molecular mechanisms and pathways involved in this complex disease. **See also**: **DNA Chip Revolution**

To date, numerous microarray studies have been performed to unravel the molecular processes involved in ARDS development (Wurfel, 2007). Copland *et al*. (2003) studied the overall lung tissue gene expression profiling in a rat model of VILI comparing HVT (20 mL/kg) with LVT ventilation (6 mL/kg). The authors found a series of genes that were upregulated by HVT, including the nuclear receptor subfamily 4, group A, member 1 (*Nur77*), early growth response 1 (*Egr1*), BTG family, member 2 (*Btg2*) and jun proto-oncogene (*c-Jun*). The upregulation of *Egr1* was related to the activation of protein kinase C (PKC)-mediated pathways, while *Nur77*, *Btg2* and *c-Jun* were implicated in the innate immunity and inflammation, apoptosis and the activation of several growth factor signalling pathways. The authors also observed upregulation of the gene encoding the heat shock protein family A member 4 (*Hsp70*), a cytoprotector, and the interleukin 1 beta (*Il1b*), an important mediator of the inflammatory response during lung injury and fibrosis when proverexpressed in the lung (Copland *et al*., 2003).

Different studies have induced VILI by a double-hit model, applying both an HVT MV and the injection of bacterial lipopolysaccharide (LPS), revealing a synergistic effect of both harmful elements in increasing lung injury (Martin *et al*., 1997), and suggesting that the presence of microbial products that induce inflammatory responses may aggravate the pathophysiology of VILI and the molecular response. Altemeier *et al*. (2004, 2005) compared the lung tissue gene expression profiling of four groups of rodents: control, suboptimal MV (tidal volume = 10 mL/Kg), intratracheal LPS and suboptimal MV + LPS. They found many differentially regulated genes in the MV + LPS group in comparison with the LPS-only group and 10 times more of differentially regulated genes compared with the MV-only group. Several genes identified by these investigators were involved in immunity (e.g. *Ccl3*, *Cxcl2*, *Il6*, *Il1β*), in the response to stress (e.g. *Gadd45g*), or were transcription factors (e.g. *Irf7*, *Atf3*). These results suggested that inappropriate MV contributes to the inflammatory response leading to the development of ARDS. Further studies identified distinct deregulated transcription factors in the MV-only and LPS-only groups, while the resulting list of deregulated transcription factors in the MV + LPS group was a combination of the results from either condition alone (Gharib *et al*., 2006).

Moreover, Grigoryev *et al*. (2004) conducted a computer analysis to simultaneously evaluate the orthologous lung tissue gene expression profiles of multispecies animal models of VILI and human lung vascular endothelial cells exposed to high-level cyclic stretch, with the assumption that overlapping responses to mechanical stretch among the models would enable to identify the genes that were deregulated identically in all species, therefore providing robust links with ARDS development. They identified a list of 69 genes that were differentially regulated and shared among models, of which only 12 were already related to lung injury in other previous studies (e.g. *IL1B*, *IL6*, *SERPINE1* and *AQP1*). In addition, they indicated that the most recognisable upregulated biological processes during VILI were those involved in the immune response, blood coagulation and cell cycle arrest.

Dolinay *et al*. (2006) compared the gene expression patterns of isolated lungs in mice subjected to overventilation (25 cmH_2O), LPS-treatment and low-pressure ventilation (10 cmH_2O). The isolated lungs, within a negative pressure chamber, were perfused *ex vivo* with cell-free culture media, eliminating peripheral leukocytes from the pulmonary system and allowing the determination of lung-specific changes (from lung structural cells) in gene expression. The authors found several genes that were regulated in an additive manner by overventilation and LPS (e.g. *Il1β*, *Il6*, *Csf2*, *Mif* and *Ccl2*), being mainly involved in immunity and stress responses. In addition, they identified novel genes that were upregulated by overventilation including *Areg*, *Akap12*, *Cyr61* and *Il11*, as well as others that were previously evidenced in independent studies such as *Nur77*.

Ye *et al*. (2005a,b) also contributed with a few other significant studies. They studied microarray-based lung gene expression profiling in canine and murine models of VILI and identified an increased expression of pre-B-cell colony-enhancing factor (*Pbef*), also known as nicotinamide phosphoribosyltransferase (*Nampt*) (Ye *et al*., 2005a). Subsequent studies performed by the same investigators suggested that *Pbef* is involved in the regulation of inflammatory cytokines, playing a role in endothelial permeability (Ye *et al*., 2005b).

The above-mentioned studies have focused on the physiologic and molecular mechanisms that occur in the lungs after the application of an injurious MV strategy. However, few studies have explored the molecular mechanisms involved in the lung protection associated with the use of LVT and PEEP. Acosta-Herrera *et al*. (2015) analysed the molecular mechanisms and pathways in lung tissues using microarray and microRNA sequencing in an animal model of sepsis-induced ARDS. They compared three groups of septic animals under different ventilator strategies: nonventilated spontaneous breathing animals, LVT (6 mL/kg) plus 10 cmH_2O PEEP and HVT (20 mL/kg) plus 2 cmH_2O PEEP. They found that the most significantly upregulated biological process in nonventilated and HVT animals was the 'response to microorganisms', whereas the 'neuron projection morphogenesis' process was one of the most significantly deregulated in LVT. Further studies led the authors to suggest that a coderegulation of 'VEGF signalling' together with 'neuron projection morphogenesis' participate in the protective mechanism. Besides, given that small noncoding RNA sequences that bind to target mRNAs (i.e. microRNAs) play a central role in the regulation of gene expression, including the regulation of susceptibility genes involved in ARDS as well as in other complex diseases (Zhou *et al*., 2011), the authors also did complementary experiments to show key microRNA species underlying the biological responses. **See also**: **MicroRNAs and Human Disease**. A microRNA analysis further supported by small RNA-seq studies, allowed detecting four deregulated species (mir-27a, Mir-103, mir-17-5p and mir-130a) that targeted a total of 159 genes overlapping the biological processes evidenced by microarray studies.

While more studies are required to elucidate gene interaction networks involved in the pathogenesis of the syndrome, owing to their significant role in gene regulation and because no invasive methods are required for routine clinical testing (i.e. as from serum), microRNA studies constitute a truly promising option to evidence ARDS biomarkers with diagnostic or prognostic utility.

Genetics of ARDS

In Hippocrates' words, 'it is far more important to know what person the disease has than what disease the person has'. Given that clinical factors alone do not allow predicting which patients will develop ARDS or which patients will die as a result of the syndrome, and the fact that genetic factors are known to play an important role in infection disease outcomes (Petersen *et al*., 2010), including ARDS development (Leikauf *et al*., 2002; Villar *et al*., 2003), there is a huge interest in the study of genetic factors as predictors of risk and prognosis. Identifying the ARDS predisposing genes will potentially: (1) offer new perspectives of disease pathogenesis and increase our capacity to identify other risk factors; (2) improve risk stratification models and patient care and (3) define individual patterns of disease, leading to the development of novel therapies and better individual treatment.

ARDS is a complex syndrome, which includes multiple genetic risk factors, and constitutes a largely heterogeneous clinical entity. In this scenario, it is the aggregation of several risk factors (genetic and environmental) that could probably explain the susceptibility to the syndrome (Risch and Merikangas, 1996). **See also**: **Disease-related Genes: Identification**. Therefore, conventional genomic approaches as the linkage analysis and the positional cloning studies have limited potential to detect the loci of interest (Risch and Merikangas, 1996). This is because such analyses require the recruitment of large numbers of families with affected members for effective detection of risk loci, and ARDS develops most commonly at older age, limiting the access to DNA and clinical data from relatives. As for many other complex diseases, genetic risk effects on ARDS susceptibility and outcomes have been historically identified through case–control association studies, as these do not necessitate the collection of family samples. Briefly, these studies compare the allele frequencies at known polymorphic loci, usually single nucleotide polymorphisms (SNPs), between unrelated ARDS cases and controls without the syndrome (Leikauf *et al*., 2002; Villar *et al*., 2003). The hypothesis underlying this design is that the closer the variant analysed is to a causally related unknown risk variant, the more likely is that they are coinherited among subjects in the population, and the larger the difference in frequency between cases and controls.

Candidate gene association studies

While current technologies allow to study simultaneously a large number variants from virtually all genes of our genome for their association with a disease or trait, most studies completed to date in ARDS susceptibility and outcomes have been performed focusing on the analysis of variation at specific genes, with a biological plausibility, or at putative functional loci. Such candidate genes have been usually identified as key elements in the molecular mechanisms or pathways related to the pathogenesis unravelled in animal models or clinical observations. However, candidate gene association studies in other complex traits have demonstrated a large number of questionable allele–trait associations because of nonreproducibility of findings in independent studies that are likely due to a combination of methodological and analytical problems (Clark and Baudouin, 2006). As a result, given that such studies will continue to be important in the field, among the best practices suggested, emphasis is made on the replication of findings in independent study samples (Chanock *et al*., 2007).

So far, a total of 81 distinct genes in 68 independent studies have been completed in ARDS (**Figure 1**), most of them being performed in samples from populations of European descent according to a recent study (Acosta-Herrera *et al*., 2014). These genes are mainly involved in response to external stimulus and cell signal transduction, although there are also genes implicated in cell proliferation, immune response and chemotaxis (Flores *et al*., 2008, 2010). The genes that have the largest number of independent study findings supporting a significant association (in at least four study samples) encode the interleukin 6 (*IL6*), interleukin 10 (*IL10*), vascular endothelial growth factor A (*VEGFA*; also known as *VEGF*), angiotensin-converting enzyme (ACE), mannose-binding lectin (protein C) 2, soluble (*MBL2*), interleukin 1 receptor antagonist (*IL1RN*) and *NAMPT*. These studies can be found in recent assessments (Flores *et al*., 2008; Acosta-Herrera *et al*., 2014) and in subsequent reports (Meyer *et al*., 2013).

The gene encoding IL-6, a cytokine that takes part in inflammation and in the maturation of B cells, is an excellent candidate as cross-species gene expression pattern comparisons in experimental models of ARDS have shown that *IL6* is highly upregulated (Grigoryev *et al*., 2004). Furthermore, high circulating concentration of IL-6 in ARDS patients has been found in independent

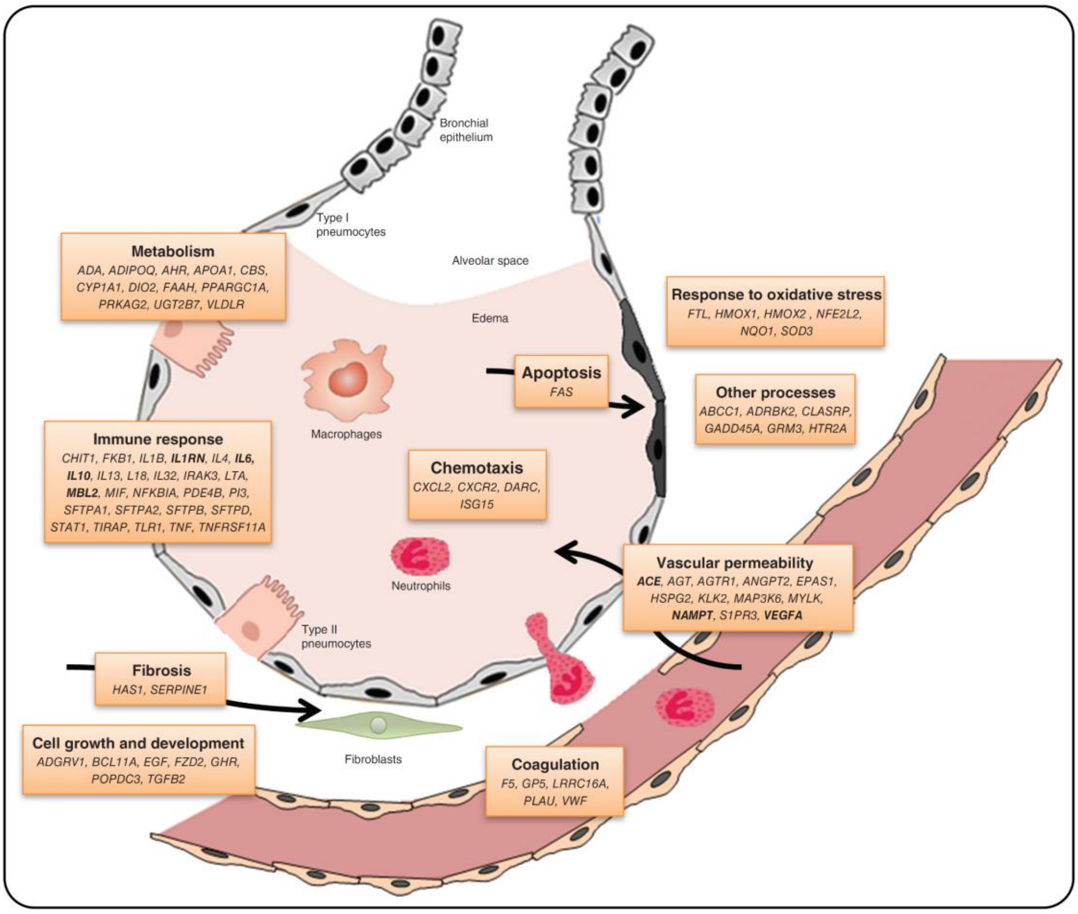

Bronchial
epithelium
Type I
pneumocytes
Alveolar space
Metabolism
ADA, ADIPOQ, AHR, APOA1, CBS, CYP1A1, DIO2, FAAH, PPARGC1A, PRKAG2, UGT2B7, VLDLR
Response to oxidative stress
FTL, HMOX1, HMOX2 , NFE2L2, NQO1, SOD3
Edema
Apoptosis
FAS
Other processes
ABCC1, ADRBK2, CLASRP, GADD45A, GRM3, HTR2A
Macrophages
Immune response
CHIT1, FKB1, IL1B, **IL1RN**, IL4, **IL6, IL10**, IL13, L18, IL32, IRAK3, LTA, **MBL2**, MIF, NFKBIA, PDE4B, PI3, SFTPA1, SFTPA2, SFTPB, SFTPD, STAT1, TIRAP, TLR1, TNF, TNFRSF11A
Chemotaxis
CXCL2, CXCR2, DARC, ISG15
Neutrophils
Vascular permeability
ACE, AGT, AGTR1, ANGPT2, EPAS1, HSPG2, KLK2, MAP3K6, MYLK, **NAMPT**, S1PR3, **VEGFA**
Type II
pneumocytes
Fibrosis
HAS1, SERPINE1
Cell growth and development
ADGRV1, BCL11A, EGF, FZD2, GHR, POPDC3, TGFB2
Fibroblasts
Coagulation
F5, GP5, LRRC16A, PLAU, VWF

Figure 1 Schematic representation of an alveolus and a pulmonary capillary depicting the potential links between the ARDS susceptibility genes and key biological processes involved in the pathogenesis. In bold, genes with evidence of association in at least four study samples. Gene symbols correspond to the acronyms provided by the NCBI website.

clinical studies (Meduri *et al*., 1995). Other variants in *IL10*, encoding another cytokine involved in immunoregulation and inflammation, have been related to development and outcome of ARDS. Similarly, variants in *IL1RN*, which encodes a protein that inhibits the activities of IL-1, have been associated with a reduced susceptibility to ARDS and with higher IL1RN levels in plasma among patients with severe trauma or septic shock.

Another broadly studied gene that has been associated with ARDS is *VEGFA*. This gene encodes a protein with different effects, including a role in increasing vascular permeability, inducing angiogenesis and regulating vasculogenesis. Frequent variants downstream from *VEGFA* gene constitute major determinants of inter-individual variation in the circulating levels of VEGFA (Debette *et al*., 2011). To date, diverse studies of lung injury in humans show that *VEGFA* polymorphisms are associated with an increased mortality in patients with ARDS and a reduction in plasma levels of VEGFA at early stages (Zhai *et al.*, 2007; Yang *et al*., 2011). However, VEGFA is highly expressed in the lung tissue of recovering ARDS patients, suggesting that it may have a protective role in the resolution of the syndrome (Medford *et al*., 2009).

The gene encoding ACE, which catalyses the conversion of angiotensin I into a physiologically active peptide angiotensin I, is known to harbour frequent variants that influence the circulating ACE activity (Chung *et al*., 2011). Variants of this gene have also been associated with ARDS mortality in several studies. Nevertheless, the enzyme ACE2 has an opposite function to ACE protecting against lung damage caused by ARDS (Nicholls and Peiris, 2005).

Finally, variants in *NAMPT*, previously mentioned in the studies of Ye *et al*. (2005a), have been associated with susceptibility to ARDS. This gene encodes a protein involved in the biosynthesis of nicotinamide adenine dinucleotide that has been found at higher concentrations in serum from patients with ARDS. In addition, *MBL2* polymorphisms were related with severity in ARDS. MBL2 plays an important role in the innate immune system as it is responsible for recognising mannose and *N*-acetylglucosamine of many microorganisms.

Genomic studies

Genome-wide association studies (GWAS) constitute a powerful option to find different susceptibility genes with no dependence on pre-existing knowledge of disease pathogenesis. GWAS have used commercially designed microarrays with probes that are specific for an SNP content genotyping in the order of 0.5–2.5 million, allowing to infer (by means of imputation methods) many other millions of variants that were not genotyped but that are known to correlate with them based on data from reference studies such as The 1000 Genomes Project. **See also**: **Genome-Wide Association Studies**. Because of the reference data that sustained the selection of the SNPs contained is well recognised that GWAS are well powered to identify risk variants that are relatively common in the population (i.e. usually for those with a minor allele frequency above 1–5%).

To date, only one GWAS on ARDS has been published (Christie *et al*., 2012). It consists on a cost-effective two-stage design with a third stage that aimed to assist in the identification of susceptibility genes based on the correlation of SNPs with gene expression values (i.e. eQTL analysis) obtained from B-lymphoblastoid cell lines. In the first stage, Christie and colleagues obtained data from roughly 2.5 million SNPs across the genome in 600 trauma-induced ARDS patients and 2266 population-based controls of European descent. For the second stage, the authors followed up the top 1% most significant SNPs in another case–control sample consisting of 212 trauma-induced ARDS patients and 283 at-risk controls. Frequent variants from the protein tyrosine phosphatase, receptor type, f polypeptide, interacting protein (liprin) and alpha 1 (*PPFIA1*) gene were finally identified with the assistance of the eQTL analysis, therefore providing support for an unanticipated ARDS susceptibility gene. *PPFIA1* encodes the liprin alpha, a protein involved in cell adhesion and cell–matrix interactions as well as other functions. Furthermore, variants in other genes including *IL10*, myosin light chain kinase (*MYLK*), angiopoietin 2 (*ANGPT2*) and Fas cell surface death receptor (*FAS*) were also replicated in the study (Christie *et al*., 2012).

It is anticipated that more GWAS on ARDS will be completed in the near future. However, both the actual definition and the complex genetic architecture of ARDS will continue to complicate the task of identifying genetic risk factors. Therefore, it will be unsurprising to find GWAS of intermediate traits, such as the clinical responses taking place during ARDS development (e.g. circulating biomarker levels). Given that these traits are closer expressions of the effects of genetic variants, which are less subject to clinical interpretations, these studies will be needed to assist in the identification of novel susceptibility genes.

Although GWAS have been key to find new susceptibility genes for many distinct complex diseases, these studies have the limitation of analysing only a fraction of the actual variants in our genome (i.e. those that are more frequent in the population), not allowing to study the whole spectrum of risk allele frequency. Therefore, many other variants with frequencies below 1%, which are more likely to be deleterious, remain unanalysed in these studies. In order to fill this gap, new technologies such as next-generation DNA sequencing (NGS) are required to improve the power to detect disease-associated loci and to improve our knowledge of the genetic determinants of ARDS susceptibility and severity. While the road ahead allows envisioning that, in a couple of years, whole-genome sequencing (WGS) will be the approach of choice for such studies, the demanding costs that are currently associated with the complete characterisation of genetic variation in a large study sample makes this a challenging endeavour for now. Whole-exome sequencing (WES) studies constitute for now a cost-effective alternative. In fact, WES has already been established as a powerful and robust tool to elucidate genetic variants underlying human diseases (Goh and Choi, 2012), and two studies have been completed to date in the context of ARDS.

Lee *et al.* (2012) performed a WES study using a design of extreme phenotypes in 88 individuals with sepsis-induced ARDS. The selection of patients was based on ‘ventilator-free days’ (VFD), which is correlated with the severity of the syndrome. The authors selected roughly half and half of patients from the extremes of the distribution of VFD (i.e. compared subjects with very high VFD against those with very low VFD) and identified 130,000 SNP variants. Because of sample limitations, they analysed those infrequent variants present in 6488 genes, allowing the identification of the *MYLK* gene as one of the most significantly associated with VFD in ARDS patients. This result provided a strong support for the association of *MYLK* with the syndrome, as it was previously linked with ARDS susceptibility based on distinct independent genetic studies (Flores *et al*., 2008; Acosta-Herrera *et al*., 2014). This gene encodes a key element of the inflammatory response, critically involved in the lung vascular permeability and the leukocyte diapedesis, promoting oedema formation when is activated by pro-inflammatory stimuli.

Besides, Shortt *et al.* (2014) also performed WES in 96 sepsis-induced ARDS samples that were compared against the data deposited in The 1000 Genomes Project. While this study did not found either a strong support for a novel susceptibility gene for ARDS or provide evidence of replication, their results were remarkable as they supported the association of variants in genes encoding class I (*HLA-B*) and II molecules of the major histocompatibility complex (*HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB5*), critically involved in the immune system. In addition, they highlighted three other genes, including the arylsulfatase D gene (*ARSD*), the XK, Kell blood group complex subunit-related family, member 3 gene (*XKR3*) and the zinc-finger protein 335 (*ZNF335*), whose variants may be associated with the susceptibility, severity and outcome of ARDS and that had not been previously associated with the syndrome.

Conclusions

Several studies highlight the implication of genetic component in ARDS susceptibility and outcome. However, although they have provided relevant insights into ARDS pathogenesis, new technological advances will offer the opportunity to reinterpret genomic information on ARDS. In this sense, new high-throughput technologies including GWAS, WES and WGS are potent tools to assay DNA on a genomic scale and to improve our understanding of ARDS pathophysiology and treatment options. Besides, given that ARDS is a complex and heterogeneous syndrome and that the disease precipitates or aggravates because of the interaction among many genetic and nongenetic factors, it will also be important to improve diagnosis, prognosis and the knowledge of risk factors by using alternative 'omics' approaches such as proteomics, metabolomics and metagenomics. In summary, the knowledge of the genetic factors involved in ARDS susceptibility is in its infancy. Further studies in larger patient populations of different ethnicities are necessary to identify genetic factors associated with ARDS to develop a personalised medicine approach.

Acknowledgements

This work was supported by grants CB06/06/1088 and PI14/00844 from the ISCIII, Spain and cofinanced by the European Regional Development Funds, 'A way of making Europe' from the European Union. MAH was supported with a fellowship from the ISCIII (FI11/00074).

References

Acosta-Herrera M, Pino-Yanes M, Perez-Mendez L, *et al.* (2014) Assessing the quality of studies supporting genetic susceptibility and outcomes of ARDS. *Frontiers in Genetics* **5**: 20.

Acosta-Herrera M, Lorenzo-Diaz F, Pino-Yanes M, *et al.* (2015) Lung transcriptomics during protective ventilatory support in sepsis-induced acute lung injury. *PLoS One* **10** (7): e0132296.

Altemeier WA, Matute-Bello G, Frevert CW, *et al.* (2004) Mechanical ventilation with moderate tidal volumes synergistically increases lung cytokine response to systemic endotoxin. *American Journal of Physiology. Lung Cellular and Molecular Physiology* **287** (3): L533–L542.

Altemeier WA, Matute-Bello G, Gharib SA, *et al.* (2005) Modulation of lipopolysaccharide-induced gene transcription and promotion of lung injury by mechanical ventilation. *Journal of Immunology* **175** (5): 3369–3376.

Bernard GR, Artigas A, Brigham KL, *et al.* (1994) The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *American Journal of Respiratory and Critical Care Medicine* **149** (3 Pt 1): 818–824.

Bersten AD, Edibam C, Hunt T, *et al.* (2002) Incidence and mortality of acute lung injury and the acute respiratory distress syndrome in three Australian States. *American Journal of Respiratory and Critical Care Medicine* **165** (4): 443–448.

Buregeya E, Fowler RA, Talmor DS, *et al.* (2014) Acute respiratory distress syndrome in the global context. *Global Heart* **9** (3): 289–295.

Carlucci M, Graf N, Simmons JQ, *et al.* (2014) Effective management of ARDS. *Nurse Practitioner* **39** (12): 35–40.

Caser EB, Zandonade E, Pereira E, *et al.* (2014) Impact of distinct definitions of acute lung injury on its incidence and outcomes in Brazilian ICUs: prospective evaluation of 7,133 patients. *Critical Care Medicine* **42** (3): 574–582.

Clark MF and Baudouin SV (2006) A systematic review of the quality of genetic association studies in human sepsis. *Intensive Care Medicine* **32** (11): 1706–1712.

Copland IB, Kavanagh BP, Engelberts D, *et al.* (2003) Early changes in lung gene expression due to high tidal volume. *American Journal of Respiratory and Critical Care Medicine* **168** (9): 1051–1059.

Chanock SJ, Manolio T, Boehnke M, *et al.* (2007) Replicating genotype-phenotype associations. *Nature* **447** (7145): 655–660.

Christie JD, Wurfel MM, Feng R, *et al.* (2012) Genome wide association identifies PPFIA1 as a candidate gene for acute lung injury risk following major trauma. *PLoS One* **7** (1): e28268.

Chung CM, Lin TH, Chen JW, *et al.* (2011) A genome-wide association study reveals a quantitative trait locus of adiponectin on CDH13 that predicts cardiometabolic outcomes. *Diabetes* **60** (9): 2417–2423.

Debette S, Visvikis-Siest S, Chen MH, *et al.* (2011) Identification of cis- and trans-acting genetic variants explaining up to half the variation in circulating vascular endothelial growth factor levels. *Circulation Research* **109** (5): 554–563.

Dolinay T, Kaminski N, Felgendreher M, *et al.* (2006) Gene expression profiling of target genes in ventilator-induced lung injury. *Physiological Genomics* **26** (1): 68–75.

Erickson SE, Martin GS, Davis JL, *et al.* (2009) Recent trends in acute lung injury mortality: 1996–2005. *Critical Care Medicine* **37** (5): 1574–1579.

Flores C, Pino-Yanes Mdel M and Villar J (2008) A quality assessment of genetic association studies supporting susceptibility and outcome in acute lung injury. *Critical Care* **12** (5): R130.

Flores C, Pino-Yanes MM, Casula M, *et al.* (2010) Genetics of acute lung injury: past, present and future. *Minerva Anestesiologica* **76** (10): 860–864.

Gajic O, Dabbagh O, Park PK, *et al.* (2011) Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort study. *American Journal of Respiratory and Critical Care Medicine* **183** (4): 462–470.

Gharib SA, Liles WC, Matute-Bello G, *et al.* (2006) Computational identification of key biological modules and transcription factors in acute lung injury. *American Journal of Respiratory and Critical Care Medicine* **173** (6): 653–658.

Goh G and Choi M (2012) Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics & Informatics* **10** (4): 214–219.

Grigoryev DN, Ma SF, Irizarry RA, *et al.* (2004) Orthologous gene-expression profiling in multi-species models: search for candidate genes. *Genome Biology* **5** (5): R34.

Hughes M, MacKirdy FN, Ross J, *et al.* (2003) Acute respiratory distress syndrome: an audit of incidence and outcome in Scottish intensive care units. *Anaesthesia* **58** (9): 838–845.

Kangelaris KN, Sapru A, Calfee CS, *et al.* (2012) The association between a DARC gene polymorphism and clinical outcomes in African American patients with acute lung injury. *Chest* **141** (5): 1160–1169.

Lee S, Emond MJ, Bamshad MJ, *et al.* (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics* **91** (2): 224–237.

Leikauf GD, McDowell SA, Wesselkamper SC, *et al.* (2002) Acute lung injury: functional genomics and genetic susceptibility. *Chest* **121** (3 Suppl): 70S–75S.

Li G, Malinchoc M, Cartin-Ceba R, *et al.* (2011) Eight-year trend of acute respiratory distress syndrome: a population-based study in Olmsted County, Minnesota. *American Journal of Respiratory and Critical Care Medicine* **183** (1): 59–66.

Luhr OR, Antonsen K, Karlsson M, *et al.* (1999) Incidence and mortality after acute respiratory failure and acute respiratory distress syndrome in Sweden, Denmark, and Iceland. *American Journal of Respiratory and Critical Care Medicine* **159** (6): 1849–1861.

Martin TR, Rubenfeld GD, Ruzinski JT, *et al.* (1997) Relationship between soluble CD14, lipopolysaccharide binding protein, and the alveolar inflammatory response in patients with acute respiratory distress syndrome. *American Journal of Respiratory and Critical Care Medicine* **155** (3): 937–944.

Medford AR, Godinho SI, Keen LJ, *et al.* (2009) Relationship between vascular endothelial growth factor + 936 genotype and plasma/epithelial lining fluid vascular endothelial growth factor protein levels in patients with and at risk for ARDS. *Chest* **136** (2): 457–464.

Meduri GU, Headley S, Kohler G, *et al.* (1995) Persistent elevation of inflammatory cytokines predicts a poor outcome in ARDS. Plasma IL-1 beta and IL-6 levels are consistent and efficient predictors of outcome over time. *Chest* **107** (4): 1062–1073.

Meyer NJ, Feng R, Li M, *et al.* (2013) IL1RN coding variant is associated with lower risk of acute respiratory distress syndrome and increased plasma IL-1 receptor antagonist. *American Journal of Respiratory and Critical Care Medicine* **187** (9): 950–959.

Moss M, Bucher B, Moore FA, *et al.* (1996) The role of chronic alcohol abuse in the development of acute respiratory distress syndrome in adults. *JAMA* **275** (1): 50–54.

Moss M and Mannino DM (2002) Race and gender differences in acute respiratory distress syndrome deaths in the United States: an analysis of multiple-cause mortality data (1979–1996). *Critical Care Medicine* **30** (8): 1679–1685.

Nicholls J and Peiris M (2005) Good ACE, bad ACE do battle in lung injury, SARS. *Nature Medicine* **11** (8): 821–822.

Petersen L, Andersen PK and Sorensen TI (2010) Genetic influences on incidence and case-fatality of infectious disease. *PLoS One* **5** (5): e10603.

Ranieri VM, Rubenfeld GD, Thompson BT, *et al.* (2012) Acute respiratory distress syndrome: the Berlin Definition. *JAMA* **307** (23): 2526–2533.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* **273** (5281): 1516–1517.

Shortt K, Chaudhary S, Grigoryev D, *et al.* (2014) Identification of novel single nucleotide polymorphisms associated with acute respiratory distress syndrome by exome-seq. *PLoS One* **9** (11): e111953.

Spragg RG, Bernard GR, Checkley W, *et al.* (2010) Beyond mortality: future clinical research in acute lung injury. *American Journal of Respiratory and Critical Care Medicine* **181** (10): 1121–1127.

Suchyta MR, Clemmer TP, Elliott CG, *et al.* (1997) Increased mortality of older patients with acute respiratory distress syndrome. *Chest* **111** (5): 1334–1339.

The Acute Respiratory Distress Syndrome Network (2000) Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New England Journal of Medicine* **342** (18): 1301–1308.

Valta P, Uusaro A, Nunes S, *et al.* (1999) Acute respiratory distress syndrome: frequency, clinical course, and costs of care. *Critical Care Medicine* **27** (11): 2367–2374.

Villar J, Flores C and Mendez-Alvarez S (2003) Genetic susceptibility to acute lung injury. *Critical Care Medicine* **31** (4 Suppl): S272–S275.

Villar J, Blanco J, Anon JM, *et al.* (2011a) The ALIEN study: incidence and outcome of acute respiratory distress syndrome in the era of lung protective ventilation. *Intensive Care Medicine* **37** (12): 1932–1941.

Villar J, Blanco J, Zhang H, *et al.* (2011b) Ventilator-induced lung injury and sepsis: two sides of the same coin? *Minerva Anestesiologica* **77** (6): 647–653.

Ware LB and Matthay MA (2001) Alveolar fluid clearance is impaired in the majority of patients with acute lung injury and the acute respiratory distress syndrome. *American Journal of Respiratory and Critical Care Medicine* **163** (6): 1376–1383.

Ware LB (2006) Pathophysiology of acute lung injury and the acute respiratory distress syndrome. *Seminars in Respiratory and Critical Care Medicine* **27** (4): 337–349.

Wurfel MM (2007) Microarray-based analysis of ventilator-induced lung injury. *Proceedings of the American Thoracic Society* **4** (1): 77–84.

Yang S, Cao S, Li J, *et al.* (2011) Association between vascular endothelial growth factor + 936 genotype and acute respiratory distress syndrome in a Chinese population. *Genetic Testing and Molecular Biomarkers* **15** (10): 737–740.

Ye SQ, Simon BA, Maloney JP, *et al.* (2005a) Pre-B-cell colony-enhancing factor as a potential novel biomarker in acute lung injury. *American Journal of Respiratory and Critical Care Medicine* **171** (4): 361–370.

Ye SQ, Zhang LQ, Adyshev D, *et al.* (2005b) Pre-B-cell-colony-enhancing factor is critically involved in thrombin-induced lung endothelial cell barrier dysregulation. *Microvascular Research* **70** (3): 142–151.

Zhai R, Gong MN, Zhou W, *et al.* (2007) Genotypes and haplotypes of the VEGF gene are associated with higher mortality and lower VEGF plasma levels in patients with ARDS. *Thorax* **62** (8): 718–722.

Zhou T, Garcia JG and Zhang W (2011) Integrating microRNAs into a system biology approach to acute lung injury. *Translational Research* **157** (4): 180–190.

Further Reading

Flores C, Ma SF, Maresso K, *et al.* (2006) Genomics of acute lung injury. *Seminars in Respiratory and Critical Care Medicine* **27** (4): 389–395.

Meyer NJ (2013) Future clinical applications of genomics for acute respiratory distress syndrome. *Lancet Respiratory Medicine* **1** (10): 793–803.

Meyer NJ (2014) Beyond single-nucleotide polymorphisms: genetics, genomics, and other ‘omic’ approaches to acute respiratory distress syndrome. *Clinics in Chest Medicine* **35** (4): 673–684.

Rubenfeld GD, Caldwell E, Peabody E, *et al.* (2005) Incidence and outcomes of acute lung injury. *New England Journal of Medicine* **353** (16): 1685–1693.

Web links

The 1000 Genomes Project: This website provides detailed information of the project, sampled populations and methods, as well as tools to browse the genetic variation across ethnicities. http://www.1000genomes.org.

NCBI: This website provides the official full names of all described genes as well as additional information related to each of them. http://www.ncbi.nlm.nih.gov/gene.

Chapter 2.

Genome-wide association study of sepsis-associated ARDS in individuals of European ancestry

This second chapter reports the results of the first GWAS of susceptibility to sepsis-associated ARDS, performed in 1,935 individuals with sepsis of European ancestry. Given that the genetic catalog of ARDS remains largely unknown, the aim of this case-control study was to identify novel genes associated with ARDS and to propose new therapeutic options. The GWAS design consisted in a discovery stage to identify suggestive signals of association, a replication stage to validate the results, and a meta-analysis of both stages together. The discovery stage included 672 patients admitted into a network of Spanish ICUs, while the replication stage included 1,345 individuals from two independent datasets of American and German cohorts. Finally, functional analyses of significant signals were performed, involving RNA-sequencing from lung biopsies, *in silico* analyses, and luciferase reporter assays.

Our analyses revealed common genetic variants associated with susceptibility to sepsis-associated ARDS. These variants were located within the Fms Related Tyrosine Kinase 1 (*FLT1*) gene, which encodes the VEGF receptor 1 (VEGFR-1), broadly involved in vascular permeability and immunity, among other processes. The functional assessment revealed a higher expression of *FLT1* in peripheral blood from ARDS patients compared to other critical care patient groups and supported the role of these variants in the regulation of the *FLT1* promoter. Particularly, protective alleles reduced the promoter activity in a monocyte cell line. These results corroborated the implication of the VEGF signaling pathway in ARDS pathophysiology and suggested VEGFR-1 as a potential therapeutic target. Based on the evidence, we suggest that the repurpose of marketed drugs targeting VEGFR-1 should be considered as novel potential treatments for ARDS patients.

This study has been published in *The Lancet Respiratory Medicine* with the title "Sepsis-associated acute respiratory distress syndrome in individuals of European ancestry: a genome-wide association study" (doi: 10.1016/S2213-2600(19)30368-6).

Genome-wide association study of sepsis-associated acute respiratory distress syndrome in individuals of European ancestry

Beatriz Guillen-Guio, MSc[1]; Jose M. Lorenzo-Salazar, MSc[2]; Shwu-Fan Ma, PhD[3]; Pei-Chi Hou, PhD[3]; Tamara Hernandez-Beeftink, MSc[1,4]; Almudena Corrales, LT[1,5]; M. Isabel García-Laorden, PhD[4,5]; Jonathan Jou, MD[6]; Elena Espinosa, MD[7]; Arturo Muriel, MD[8]; David Domínguez, MD[7]; Leonardo Lorente, MD[9]; María M. Martín, MD[10]; Carlos Rodríguez-Gallego, MD[11]; Jordi Solé-Violán, MD[5,12]; Alfonso Ambrós, MD[13]; Demetrio Carriedo, MD[14]; Jesús Blanco, MD[5,8]; José M. Añón, MD[5,15]; John P. Reilly, MD[16]; Tiffanie K. Jones, MD[16]; Caroline A.G. Ittner, PhD[16]; Rui Feng, PhD[17]; Franziska Schöneweck, MSc[18]; Michael Kiehntopf, MD[19]; Imre Noth, MD[3]; Markus Scholz, PhD[20]; Frank M. Brunkhorst, MD[21]; André Scherag, PhD[22]; Nuala J. Meyer, MD[16]; Jesús Villar, MD[4,5,23] and Carlos Flores, PhD[1,2,5,24*]

[1]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain

[2]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain

[3]Division of Pulmonary & Critical Care Medicine, Department of Medicine, University of Virginia, Charlottesville, VA, USA

[4]Research Unit, Hospital Universitario de Gran Canaria Dr. Negrín, Las Palmas de Gran Canaria, Spain

[5]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain

[6]University of Illinois College of Medicine at Peoria, Peoria, IL, USA

[7]Department of Anesthesiology, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain

[8]Intensive Care Unit, Hospital Universitario Rio Hortega, Valladolid, Spain

[9]Intensive Care Unit, Hospital Universitario de Canarias, La Laguna, Tenerife, Spain

[10]Intensive Care Unit, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain

[11]Department of Immunology, Hospital Universitario de Gran Canaria Dr Negrín, Las Palmas de Gran Canaria, Spain

[12]Intensive Care Unit, Hospital Universitario de Gran Canaria Dr Negrín, Las Palmas de Gran Canaria, Spain

[13]Intensive Care Unit, Hospital General de Ciudad Real, Ciudad Real, Spain

[14]Intensive Care Unit, Complejo Hospitalario Universitario de León, León, Spain

[15]Intensive Care Unit, Hospital Universitario La Paz, I*di*PAZ, Madrid, Spain

[16]University of Pennsylvania Perelman School of Medicine; Division of Pulmonary, Allergy, and Critical Care Medicine, Philadelphia PA, USA

[17]University of Pennsylvania, Department of Biostatistics, Epidemiology, and Informatics, Philadelphia PA, USA

[18]Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany

[19]Institute of Clinical Chemistry and Laboratory Diagnostics and Integrated Biobank Jena (IBBJ), Jena University Hospital, Jena, Germany

[20]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany

[21]Center for Clinical Studies and Head of Paul-Martini-Clinical Sepsis Research Unit, Department of Anesthesiology and Intensive Care Medicine, Jena University Hospital, Jena, Germany

[22]Institute of Medical Statistics, Computer and Data Sciences and Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany

[23] Keenan Research Center for Biomedical Sciences at the Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Ontario, Canada

[24]Instituto de Tecnologías Biomédicas (ITB), Universidad de La Laguna, Santa Cruz de Tenerife, Spain.

***Corresponding autor**

Dr. Carlos Flores, E-mail: cflores@ull.edu.es, Tel: +34 601239957

Research Unit, Hospital Universitario N.S. de Candelaria

Carretera del Rosario s/n, 38010 Santa Cruz de Tenerife, Spain

Abstract

Background. The acute respiratory distress syndrome (ARDS) is a lung inflammatory process mainly caused by sepsis. Most previous studies that identified genetic risks for ARDS were focused on biological candidates. We aimed to identify novel genetic variants associated with ARDS susceptibility and to provide complementary functional evidence.

Methods. We conducted a case-control genome-wide association study (GWAS) in 1,935 European subjects, using sepsis-associated ARDS patients as cases and sepsis patients without ARDS as controls. The discovery stage included 672 patients admitted into a network of Spanish intensive care units. The replication stage comprised 1,345 individuals from two independent datasets involving the MESSI cohort study (U.S.A.) and the VISEP/MAXSEP trials of the SepNet study (Germany). We used RNAseq-based gene expression data from lung biopsies, *in silico* analyses, and luciferase reporter assays to assess functionality.

Findings. We identified a novel genome-wide significant association with sepsis-associated ARDS susceptibility (rs9508032, odds ratio [OR]=0·61 [95% CI=0·41-0·91], *p*-value=5·18x10^{-8}) located within the Fms Related Tyrosine Kinase 1 (*FLT1*) gene encoding the vascular endothelial growth factor (VEGF) receptor 1 (VEGFR-1). The region containing the sentinel variant and its best proxies acted as a silencer for *FLT1* promoter, and alleles with protective effects in ARDS further reduced promoter activity (*p*=4·66x10^{-3}). A literature mining of all previously described ARDS genes validated the association of *VEGFA* (*p*=4·69x10^{-5}; OR=0·55 [95%CI = 0·41-0·73]).

Interpretation. A common variant within *FLT1* gene is associated with sepsis-associated ARDS. Our findings support the central role of VEGF signaling pathway in ARDS pathogenesis and provides a potential therapeutic target.

Funding. Instituto de Salud Carlos III, European Regional Development Funds, Instituto Tecnológico y de Energías Renovables, Agencia Canaria de Investigación, Innovación y Sociedad de la Información, European Social Fund, Thuringian Ministry of Education, Science and Culture, the public funded Thuringian Foundation for Technology, Innovation and Research, the German Sepsis Society, German Ministry of Education and Research, NIH, and the American Thoracic Society Foundation.

Word Count: Abstract (250), Main text (3,873)

2 Tables, 4 Figures, 40 References, 1 Supplementary document

Research in context

Evidence before this study: We conducted a literature search on PubMed for all studies reporting genes which were significantly associated with ARDS up to November 2018. Most previous genetic studies in ARDS have focused on biological candidate genes mainly involved in the immune response, vascular permeability, and metabolism. Two small whole-exome sequencing studies and two genome-wide association studies (GWAS) of ARDS have been published to date, although none of them focused exclusively on sepsis-associated ARDS.

Added value of this study: To our knowledge, we report the results of the first GWAS of sepsis-associated ARDS susceptibility conducted on 1,935 European patients with sepsis. We reveal a novel protective genome-wide significant association with sepsis-associated ARDS within the Fms Related Tyrosine Kinase 1 (*FLT1*) gene, encoding the vascular endothelial growth factor receptor 1 (VEGFR-1). We also report that SNP alleles with protective effects in ARDS reduce *FLT1* promoter activity. These findings reinforce the need to target VEGF signaling in ARDS pathogenesis, a pathway linked to vascular permeability and immune and inflammatory responses.

Implications of all the available evidence: Our results support the central role of VEGF signaling in ARDS pathogenesis and suggest VEGFR-1 as a potential therapeutic target. There are effective drugs targeting this protein that are being used in other diseases and they could be potentially repurposed for ARDS.

Introduction

Acute respiratory distress syndrome (ARDS) is a serious complication of sepsis of pulmonary or non-pulmonary origin.[1] This syndrome is defined as an acute inflammatory process of the lung caused by injury to the alveolar-capillary barrier, resulting in increased alveolar-capillary permeability and protein-rich pulmonary edema. This leads to severe hypoxemia (assessed by PaO_2/FiO_2 ratio), bilateral pulmonary infiltrates, and decreased lung compliance.

The annual incidence of ARDS ranges from five to 80 cases per 100,000 individuals,[2,3] with an overall hospital mortality of approximately 40%.[4] In fact, ARDS is a cause of morbidity and mortality in adult intensive care units (ICUs) worldwide. Survivors often develop physical and cognitive impairments, including neuropsychiatric disorders, that diminish their quality of life.[5,6] At present, there are no available methods to treat or rapidly rehabilitate lungs of affected patients. Effective therapeutic options remain elusive, likely due to the heterogeneity of the syndrome. Currently, the only available interventions that impact patient survival involve specific strategies for mechanical ventilation (MV) and patient position to minimize ventilator-induced lung injury.[7,8]

Given the limited therapeutic options, there is a strong interest in identifying genetic factors that modify ARDS risks and which may serve as potential therapeutic targets. Several studies have reviewed the implication of genetic factors in ARDS susceptibility and outcomes.[9] Overall, most genetic studies have focused on biologically motivated candidate genes mainly involved in the immune response, vascular permeability and metabolism.[9] In addition, two small whole-exome sequencing studies in ARDS patients revealed that the *MYLK* gene was associated with ARDS severity as measured by ventilator-free days,[10] and that three other genes (*ARSD*, *XKR3*, and *ZNF335*) were associated with ARDS susceptibility, severity and mortality outcomes.[11]

Two genome-wide association studies (GWAS) of ARDS have been published to date, one used trauma-associated ARDS cases of European ancestry[12] and the other used all-cause ARDS African-American cases.[13] These studies revealed two potential ARDS genes, *PPFIA1* and *SELPLG*, although the reported variants both failed to reach genome-wide significance. Despite the marginal associations, prior GWAS results were paired with functional analyses either based on expression quantitative trait loci (eQTL) or on animal models to reinforce the role of prioritized genes in ARDS susceptibility.

Nonetheless, the genetics of ARDS susceptibility remains largely elusive. Thus, further studies on a genomic scale and larger sample sizes are needed. To our knowledge, here we performed the first

GWAS of susceptibility to ARDS in 1,935 individuals of European ancestry, using sepsis-associated ARDS patients as cases and sepsis patients without ARDS as controls. With the hypothesis that frequent genetic variants in the population associate with disease risk, we aimed to identify genetic variants associated with ARDS susceptibility and provide complementary functional evidence using *in silico* analyses, gene expression data, and luciferase reporter assays.

Methods

Study design and sample description

We performed a case-control GWAS of ARDS in sepsis patients of European ancestry. A discovery stage was designed to prioritize variants based on their suggestive association. At the conclusion of this stage (during November of 2017), investigators from two independent cohorts were contacted and their data were used to validate the associations in the replication stage. Finally, a meta-analysis combining the discovery and replication association results was performed during September of 2018 to identify variants significantly associated with ARDS.

The GEN-SEP cohort was used in the discovery stage (Figure 1 and Appendix). It consisted of 672 unrelated adult patients with sepsis[14] who were followed for the development of ARDS according to the Berlin definition criteria.[15] Controls constituted patients with all-cause sepsis who did not develop ARDS during their ICU stay. DNA was extracted from peripheral blood of all patients (Appendix).

The replication stage was conducted on two independent datasets from European-ancestry ICU patients, where sepsis-associated ARDS were used as cases and sepsis without ARDS were considered as at-risk controls (Figure 1). The first dataset was derived from 605 patients (268 cases and 337 controls) out of 1,263 patients of multiple ancestries from the “Molecular Epidemiology of Sepsis in the ICU” (MESSI) cohort study (U.S.A.). The second dataset was obtained from 740 patients (91 cases and 649 controls) out of 880 patients from the VISEP and MAXSEP trials of the SepNet study group (Germany). Patients in both studies meet the Berlin definition criteria for ARDS.[15] These datasets, thereafter referred to as MESSI and SepNet datasets, have been previously described.[16,17]

Genotyping and statistical analyses

For the discovery stage, a total of 587,352 SNPs were genotyped using the Axiom Genome-Wide Human CEU 1 Array (Affymetrix). Additionally, a principal component (PC) analysis (PCA) was conducted to reduce the effects of population stratification in the analysis (Appendix and Supplementary Figure 1). SNPs were genotyped using the Affymetrix Axiom TxArray v.1 (Affymetrix) in the MESSI study, while HumanOmniExpressExome arrays (Illumina, Inc.) were used in the SepNet study. Genotyping procedures are detailed in the Appendix.

After variant imputation in GEN-SEP data, logistic regression models assumed an additive inheritance. Sex, age, and the Acute Physiology and Chronic Health Evaluation II (APACHE II) score were included as covariates to address potential bias. Variants with low allele frequency (MAF<1%) or with a low imputation quality (Rsq<0·3) were excluded from the analysis. Details of imputation and association procedures are described in Appendix. Independent variants showing a p<5·0x10^{-5} were followed up in the replication stage.

In the MESSI study, logistic regression models were performed assuming additive inheritance considering the first two PCs, age, and sex as covariates. For the SepNet study, logistic regressions were performed including the first three PCs, sex, age, and APACHE II as covariates. Meta-analysis was performed on the results of these two studies. For variants that showed a nominal association (p<0·05) with ARDS susceptibility in the replication stage, a meta-analysis including discovery and replication stages was also performed. Genome-wide significance was declared with a meta-analysed significance of p<9·26x10^{-8} according to the most recent empirical estimations in European populations.[18]

FLT1 and VEGFA gene expression and functional annotation of genetic variants

In silico and *in vitro* approaches were used to investigate potential biological consequences of variants associated with ARDS. First, *FLT1* and *VEGFA* expressions were assessed in nine lung biopsies from healthy individuals by means of RNA-sequencing (Appendix). In parallel, we accessed public gene expression data (GSE32707) from 88 critically-ill adult patients that were evaluated for sepsis and ARDS (Appendix). Next, to highlight the functional role of the associated variant and of SNPs that were LD proxies in Europeans (r^2=1·0), we applied several *in silico* tools for variant prioritization [DeepSea, DSNetwork, Open Targets Genetics] and to predict potential regulatory genomic regions including epigenetic modifications [DeepSea, HaploReg, RegulomeDB], long-distance physical interactions [Capture Hi-C Plotter (CHiCP)], and tissue specific local expression quantitative trait loci (cis eQTLs) [GTEx, TIVAN]. Additional tools [VEP, SNPdelScore] were used to predict the likelihood of deleteriousness of each SNP. See Appendix for further details.

Dual-luciferase reporter assays

The potential regulatory effect of the ARDS-associated variant on promoter activity was investigated using a Dual-Luciferase Reporter Assay System® (Promega, Madison, WI). Experiments were performed using human lung epithelial (A549) and peripheral blood monocyte (THP-1) cell lines, both known to have an active *FLT1* promoter activity and expressing VEGFR-1.[19] Two types of constructs were generated: 1) a reporter construct including a fragment of the *FLT1* promoter inserted into a promoterless pGL4.10 [luc2] luciferase reporter vector, and 2) two regulatory constructs including a region of intron 10 containing either the reference or alternative alleles of the most significantly associated variant within *FLT1* and its perfect LD proxies, which were inserted into the reporter construct. Promoter activities were expressed as a relative response ratio of *Firefly* luciferase/Renilla luciferase signals. See Appendix for further details.

Literature mining of previously reported ARDS-associated genes

A literature search for all studies reporting genes which were significantly associated with ARDS was conducted. Association results in the discovery stage were extracted and an effective number of independent signals per gene was measured in order to adjust for multiple testing. See Appendix for further details.

Role of the funding source

The funders had no role in the study design, data collection, analysis, interpretation of data, decision to publish, or preparation of the manuscript. CF was involved in all stages of study development and delivery, had full access to all data in the study, and had final responsibility for the decision to submit for publication.

Results

GWAS of sepsis-associated ARDS

After filtering steps and the quality control, a total of 515,657 SNPs from 590 patients (274 sepsis-associated ARDS cases and 316 controls with sepsis) were used for the discovery stage (Figure 1). Demographic and clinical features of these patients are shown in Table 1. Genotype imputation on the HRC r.1.1 allowed us to perform the association testing of this stage on 7·98 million variants with MAF≥1%. The genomic inflation factor (λ=0·98) did not show signs of inflation of the results

(Supplementary Figure 2). Suggestive associations (*p*<5·0x10^{-5}) were detected for 229 variants residing in 53 independent loci (lowest *p*=2·6x10^{-7}) (Figure 2, Supplementary Table 1).

The replication stage in a total of 359 patients with sepsis-associated ARDS and 986 controls with sepsis focused on the sentinel variants (variants with the smallest *p*-values) of 52 autosomal loci (Figure 1, Supplementary Table 2). Because of the difficulties in accessing data, we did not follow up the X chromosome variants in the replication stage. Association testing in the replication stage revealed four SNPs that were nominally significant (uncorrected *p*<0·05; Table 2), although none of them were significantly associated with ARDS susceptibility after a Bonferroni correction (threshold *p*=9·62x10^{-4}). The first signal is an intronic variant (rs9508032) of the *FLT1* gene (Figure 3), encoding the transmembrane receptor known as the VEGFR-1. The other three SNPs were located intergenic (rs11195238) to the genes encoding the structural maintenance of chromosomes 3 (*SMC3*) and the RNA binding motif protein 20 (*RBM20*); intergenic (rs8001184) to the genes encoding slit and neurotrophic tyrosine kinase (NTRK) like family member 5 (*SLITRK5*) and glypican 5 (*GPC5*); and in intron one (rs2734600) of the gene encoding serine protease 3 (*PRSS3*). Meta-analysis of results from the discovery and replication stages for these four SNPs revealed that the sentinel variant rs9508032, located intronic to the *FLT1* gene, was the only SNP that reached genome-wide significance (Table 2). The *FLT1* variant showed consistent direction of effects, with an odds ratio (OR) for the T allele of 0·61 (95% confidence interval (CI) = 0·41-0·91), and a *p*-value of 5·18x10^{-8}. A sensitivity analysis of the association of rs9508032 at *FLT1* supported that the association was robust to adjustment for comorbidities, isolated pathogen, and the disease severity (Supplementary Table 3), though clinical data was missing for a significant proportion of subjects for some variables (up to 55%). The rs9508032-ARDS association demonstrated similar effect sizes and directions even when the sample size was significantly reduced due to missing clinical data. Furthermore, there were 16 additional variants residing in *FLT1* among the 226 SNPs with suggestive associations in the discovery stage (Supplementary Table 4). All but one was nominally significant in the replication stage, and five achieved genome-wide significance after meta-analysis of discovery and replication stages. In *ad hoc* analyses, we evaluated if the association of the sentinel variant persisted when unselected population controls were used instead of the at-risk controls. Based on the genotypes from 927 unrelated Spanish individuals that were genotyped with the same array in previous studies,[20,21] results also supported a significant association of rs9508032 with ARDS (OR= 0·73, 95% CI= 0·58-0·90, *p*-value=3·86x10^{-3}). We also evaluated if the sentinel variant predicted ICU mortality. However, our results indicated that it did not predict ICU survival among sepsis or ARDS patients from the GEN-SEP cohort (Supplementary Table 5). This evidence further supports that the *FLT1* association with sepsis-associated ARDS was genuine. Finally, at this stage, we assessed if the sentinel variant (and perfect LD proxies) of the *FLT1* also associated with ARDS after trauma, but none of them was present in the GWAS of Christie and colleagues (Appendix).[12]

Gene expression and functional impact predictions at variant sites

Transcriptomic data from lung biopsies obtained from non-ARDS control subjects revealed a high expression of *FLT1* (9,977 counts per million on average ± 5,228) and *VEGFA* (19,221 counts per million on average ± 16,165) in lung tissues, which is in agreement with GTEx information supporting a prominent expression of these genes in the lung. Among the eight *FLT1* isoforms that were evaluated on the RNAseq dataset, the canonical isoform encoding a membrane-spanning protein (FLT1-201, ENST00000282397) and the next one in size (FLT1-207, ENST00000615840), which encodes a secreted VEGF-binding protein of 687 amino acids,[22] accumulated more than 10 times more reads in average than the rest of the gene isoforms (Supplementary Table 6). Among the 29 *VEGFA* isoforms we assessed, those that had higher expression in the lungs were VEGFA-205 (ENST00000372067), VEGFA-229 (ENST00000621747), VEGFA-227 (ENST00000615393), VEGFA-222 (ENST00000520948), VEGFA-206 (ENST00000372077), VEGFA-212 (ENST00000480614), and VEGFA-215 (ENST00000497139) (Supplementary Table 6). We also accessed array expression data from peripheral blood obtained from a cohort of critically-ill patients that included donors with sepsis, with and without ARDS, as well as non-sepsis patients. These data strongly supported that the mean *FLT1* expression level in peripheral blood varied significantly among patient groups (ANOVA, *p*=0·002), with a higher average *FLT1* gene expression among ARDS patients than in ICU controls without sepsis or SIRS (t-test, *p*=0·001) (Supplementary Figure 3). On the contrary, the expression levels for the three available probes of *VEGFA* did not vary significantly among ICU patient groups (ANOVA; ILMN_2375879, *p*=0·638; ILMN_1693060, *p*=0·435; and ILMN_1803882, *p*=0·214) (Supplementary Figure 3). Next, we performed an *in silico* bioinformatic approach to explore the functional features of rs9508032 and the other five variants of *FLT1* that reached genome-wide significance after meta-analysis (Supplementary Table 4). Relevant functional information was found for rs9508032 and two of its proxies (rs722503 and rs8002446), all of them from intron 10, as these three SNPs were located in enhancer and promoter histone marks, in DNase I hypersensitive sites (DHS) of many cell types, and were related to the alteration of regulatory motifs (Supplementary Table 7). Additionally, rs722503 and rs8002446 have predicted effects on transcription factor binding. The algorithmic framework of DeepSEA predicted a significant functional effect for rs722503 (*p*=0·045). DSNetwork predicted similar results where rs722503 was prioritized as the best candidate variant for further functional analysis in this region. In contrast, Open Targets Genetics prioritized rs8002446 as potentially functional based on information of DHS and enhancer- transcription start sites data. Using GTEx, no significant eQTLs were identified for rs9508032 or its proxies, although we did observe that both rs9508032 and rs722503 had high CellulAr dePendent dEactivating (CAPE) scores for eQTL and DNase I QTLs in human umbilical vein endothelial cells, fibroblasts, epithelial and immune (monocyte) cells (Supplementary Table 7). Using the CHiCP to visualize capture Hi-C experiments conducted by Mifsud and colleagues,[23] we observed the existence of physical interactions between the region containing the three variants and the *FLT1* promoter region in a lymphoblastoid cell line (GM12878).

In vitro luciferase assays

Based on the above evidence, we then performed luciferase promoter assays to assess the effect of the intron 10 region containing the genome-wide significance SNPs on *FLT1* promoter activity (Figure 4A). Our results showed that the *FLT1* intron 10 region containing these variants repressed gene promoter activity with a consistent effect on both peripheral blood monocytes (65·1 ± 10·7% reduction) and human lung epithelial cells (48·7 ± 4·1% reduction) (Wilcoxon test, p=4·10x10^{-4} and p=0·02, respectively) (Figure 4B). When we compared the constructs with reference vs. alternative alleles for all positions within intron 10 of *FLT1*, we found that the presence of alternative alleles (protective for ARDS) were associated with a further decrease (48·6 ± 7·2% reduction) of the *FLT1* promoter activity in peripheral blood monocytes (Wilcoxon test, p=4·66x10^{-3}) (Figure 4C). No significant reduction of the *FLT1* promoter activity was found for pneumocytes (Wilcoxon test, p=0·89).

Association of previously reported ARDS genes

Finally, we performed a thorough literature mining on genes previously associated with ARDS in our discovery stage. Results of our search merged with previous reviews identified 96 genes with prior reported association with ARDS susceptibility or outcome (Supplementary Table 8). Although none of the 96 genes surpassed a study-wise Bonferroni-corrected threshold in the discovery (p=2·18x10^{-6}), the *VEGFA* gene reached a gene-wise significance after Bonferroni correction in the discovery study (top signal: OR= 0·55, 95% CI = 0·41-0·73; p=4·69x10^{-5}) (Supplementary Table 8). Not surprisingly, VEGF-A is one of the main ligands of VEGFR-1.[19]

Discussion

To our knowledge, here we reported the results of the first GWAS of sepsis-associated ARDS completed to date, where we identified a locus located in *FLT1* associated with ARDS that reached genome-wide significance in a combined meta-analysis of all cohorts. Of note, the sentinel SNP of *FLT1* (rs9508032) and the perfect LD proxies were all located in close proximity within intron 10, a region which we observed acting as a silencer of the *FLT1* promoter activity in monocyte and human lung epithelial cell lines. In conjunction with human transcriptomics data, we also determined that different *FLT1* isoforms are expressed in lung tissues, and that its expression in peripheral blood is positively correlated with the severity of illness, with the highest levels detected in ARDS patients. Evidence from our studies suggested a possible functional role of the sentinel SNP (rs9508032) and two of its perfect LD proxies in Europeans. Findings also revealed allelic effects of intron 10 on the *FLT1* promoter activity, which was particularly significant on monocytes. Interestingly, *FLT1* and other nearby genes (*FLT3* and *PAN3*) were strongly associated with monocyte counts in the UK Biobank.[24] All these findings reinforce the concept that monocytes are also crucial in the VEGF-mediated lung response.[19] Variants from *FLT1* had never been associated with ARDS susceptibility or outcomes in previous independent studies, although Kim and colleagues[25] have reported the association of *FLT1* with all-cause pulmonary complications. Additionally, there is evidence of association of *FLT1* with other complex diseases, such as coronary arterial disease[26] and preeclampsia,[27,28] where the endothelium plays an important pathophysiological role.

FLT1 encodes VEGFR-1, a tyrosine-protein kinase that acts as a transmembrane receptor of VEGF-A, other VEGF family members, and the placental growth factor (PLGF). VEGF was originally identified as a vascular permeability factor,[19] although it has diverse and pleiotropic activities beyond the regulation of the alveolar-capillary barrier.[29] VEGF has been involved in the fibroproliferative phase of ARDS[30], as well as in resolution of ventilator-associated pneumonia.[31] However, Ware and colleagues found that levels of VEGF were similar in undiluted edema fluid from hydrostatic and ARDS patients.[32] Although its role remains unclear, abundant evidence supports a negative regulatory role of an alternatively spliced soluble form of VEGFR-1 (sFLT-1) sequestering part of VEGF bioactivity.[22] High levels of sFLT-1 in the alveolar space are associated in humans with the occurrence of late ARDS in trauma,[33] as well as with sepsis severity, organ dysfunction, and ICU survival.[34,35,36] In parallel, we have found that *FLT1* expression varied between ARDS and other ICU patients in peripheral blood, while *VEGFA* expression did not show differences. Taken together, this suggests that disease-related VEGF bioavailability could be dependent on the receptor isoforms. Interestingly, the array-based transcriptomics experiment specifically targeted exon 30 of *FLT1* (Supplementary Figure 3), which critically involves the canonical receptor (FLT1-201), one of the few highly expressed isoforms. These observations offer a potential mechanistic link between the GWAS results and ARDS pathophysiology, suggesting that the *FLT1* SNPs could be linked to the expression of the VEGFR-1 transmembrane isoform. The decrease of *FLT1* promoter activity *in vitro* in the presence of intron 10 alleles associated with ARDS protection may translate in a reduction of the canonical VEGFR-1 expression and, thus, in a decrease of VEGF signalling. This hypothesis reconciles with the attenuation of many of the VEGF-mediated pathophysiologic effects in ARDS, including the formation of pulmonary edema. However, given the limitations to distinguish expression levels from gene isoforms in array-based transcriptomics experiments and that the cell type(s) that mechanistically link *FLT1* SNPs with the ARDS pathophysiology remains unknown, this scenario is purely speculative.

Despite the central role of VEGF in ARDS and the availability of VEGF-targeting drugs, clinical trials using drugs directed towards VEGF pathways for ARDS patients are scarce. There is one entry of clinical trial of the efficacy of bevacizumab (anti-VEGF antibody) to prevent sepsis-associated ARDS (ClinicalTrials.gov identifier: NCT01314066). However, it was withdrawn without a single patient enrolled due to a lack of funding. A search in DrugBank[37] and additional *in silico* explorations in Gene2drug[38] allowed us to systematically identify available drugs targeting this pathway. Although most of them are currently in use for cancer treatment (none of them under evaluation in ARDS patients), nintedanib constitutes one of the few effective antifibrotic therapies as it targets VEGFR-1 and slows the rate of forced vital capacity decline of idiopathic pulmonary fibrosis.[39] In addition, the antifungal drug itraconazole is known to inhibit the glycosylation of VEGFR-1 and VEGFR-2, affecting their migration pattern and signaling activity.[40] Based on this and our findings, nintedanib and itraconazole potentially might be repurposed as ARDS drugs and warrant further investigation.

We acknowledge there are strengths and limitations of our study. The main strength is that, to our knowledge, our study is the first GWAS of sepsis-associated ARDS, a complex acute syndrome with a high morbidity and mortality in ICUs worldwide. We contrast our ARDS cases with similarly well-characterized critically ill sepsis patients that did not develop ARDS to address the heterogeneity of the syndrome. This approach allowed us to identify reproducible associations at one locus. We provide strong evidence (transcriptomics data, functional annotations and *in vitro* experiments) to sustain a functional implication of *FLT1* variants in ARDS physiopathology. However, this study also has some limitations. The main weakness is the small sample size overall, limiting the power for detecting variants of smaller effects or of lower frequency. The limited sample size can be attributed to the low incidence and high heterogeneity of the syndrome, which makes sample collection difficult and slow. In this respect, it is plausible that rare variants in or near the identified regions remain undetected because of technological limitations. Whole-exome and genome sequencing analyses would offer better resolution to achieve that aim. Therefore, more ARDS loci are to be expected as the genomic studies of ARDS increase in size and marker resolution. Additionally, this study focused only on European ancestry patients. Further studies are needed to identify whether *FLT1* variation also impacts ARDS risk in non-European populations. We used the A549 cell line as a model for human alveolar epithelial cell, which inherently entails experimental limitations because of its cancerous nature. Further experiments should evaluate primary human alveolar type 2 cells to assess the impact of this choice in our observations. Finally, because the X chromosome is usually filtered out from most GWAS because it adds a level of difficulty to the analyses, we were unable to follow-up a variant in *OPHN1* gene (encoding a Rho-GTPase-activating protein) in the replication stage.

In summary, we describe the results of a GWAS of sepsis-associated ARDS. We report one novel locus located in *FLT1* involved in ARDS susceptibility. Based on these results and the accumulated evidence, this study provides an orthogonal demonstration of the genuine central role of VEGF signalling pathway in ARDS susceptibility and strongly favours that VEGFR-1 is a therapeutic target for preventing ARDS. Independent studies should aim to validate our findings, including independent association studies in non-sepsis ARDS patients.

Acknowledgments

We would like to thank all patients for their willingness to participate in the study. We thank all the investigators of GENetics of SEPsis-induced ARDS (GEN-SEP) Network and the nurses that assisted in the collection of donor samples. Similarly, we thank all members of the SepNet and MESSI study groups for their willingness to support data sharing. This study was funded by the Instituto de Salud Carlos III (CB06/06/1088; PI10/0393; PI14/00844; PI16/00049; PI16/00759; PI17/00610; FI17/00177) and co-financed by the European Regional Development Funds, “A way of making Europe” from the European Union; and by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology. BGG was supported by a fellowship from Agencia Canaria de Investigación, Innovación y Sociedad de la Información (TESIS2015010057) co-funded by European Social Fund. The genotyping service of GEN-SEP was carried out at CEGEN-PRB3-ISCIII; it is supported by grant PT17/0019, of the PE I+D+i 2013-2016, funded by ISCIII and ERDF. The SepNet GWAS project was supported by the Paul-Martini-Sepsis Research Group, funded by the Thuringian Ministry of Education, Science and Culture (ProExcellence; grant PE 108-2); the public funded Thuringian Foundation for Technology, Innovation and Research (STIFT) and the German Sepsis Society (GSS); the Jena Center of Sepsis Control and Care (CSCC), funded by the German Ministry of Education and Research (BMBF; 01 EO 1002, 01 EO 1502). The VISEP and MAXSEP trials from the SepNet Study Group had been supported by a BMBF grant (01 KI 0106) and by unrestricted grants from B. Braun, HemoCue, Novo Nordisk, Astra Zeneca GmbH, Wedel, Germany and Bayer HealthCare, Leverkusen, Germany. Finally, AS was supported by the BMBF (01 ZZ 1803C) within the SMITH (“Smart Medical Information Technology for Healthcare”) consortium. The MESSI cohort was funded by NIH

HL137006 and HL137915 (Meyer) and by an American Thoracic Society Foundation award. Finally, we would like to thank the six anonymous reviewers for their critical advice to improve this manuscript.

Authors' contributions

CF designed and supervised the study. BGG, JMLS, SFM, PCH, THB, JJ, RF, FS, AS, NJM, JV, and CF performed the analyses, experiments, and data interpretation. AC, MIGL, EE, AM, DD, LL, MMM, CRG, JSV, AA, DC, JB, JMA, JPR, TKJ, CAGI, MK, IN, MS, FMB, AS, NJM, and JV participated in data collection. BGG and CF wrote the draft of the manuscript. All authors participated in the critical revision and final approval of the manuscript.

Conflict of interests

JPR reports grants from NIH, during the conduct of the study. MK reports that his institution has granted patents (EP2592421, CN104204808A, EP2780719) and pending patent applications (US20140248631, JP2014533368, EP3239712, WO2017186842). MS receives funding from Pfizer Inc. for a project not related to this research. NJM reports grants from GlaxoSmithKline, Inc, advisory board membership from SOBI, Inc, and consulting fees from Competitive Drug Development International on behalf of Bayer, Inc, outside the submitted work.

Ethics committee approval

All participating studies were performed according to The Code of Ethics of the World Medical Association (Declaration of Helsinki), and informed consent was obtained from all subjects or their representatives. The Research Ethics Committees at participating centers approved this study.

References

1 Bernard GR, Artigas A, Brigham KL, *et al.* The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am J Respir Crit Care Med* 1994; **149**: 818–24.

2 Villar J, Blanco J, Añón JM, *et al.* The ALIEN study: incidence and outcome of acute respiratory distress syndrome in the era of lung protective ventilation. *Intensive Care Med* 2011; **37**: 1932–41.

3 Buregeya E, Fowler RA, Talmor DS, Twagirumugabe T, Kiviri W, Riviello ED. Acute respiratory distress syndrome in the global context. *Glob Heart* 2014; **9**: 289–95.

4 Bellani G, Laffey JG, Pham T, *et al.* Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. *JAMA* 2016; **315**: 788–800.

5 Spragg RG, Bernard GR, Checkley W, *et al.* Beyond mortality: future clinical research in acute lung injury. *Am J Respir Crit Care Med* 2010; **181**: 1121–7.

6 Huang M, Parker AM, Bienvenu OJ, *et al.* Psychiatric Symptoms in Acute Respiratory Distress Syndrome Survivors. *Crit Care Med* 2016; **44**: 954–65.

7 Acute Respiratory Distress Syndrome Network, Brower RG, Matthay MA, *et al.* Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med* 2000; **342**: 1301–8.

8 Guérin C, Reignier J, Richard J-C, *et al.* Prone positioning in severe acute respiratory distress syndrome. *N Engl J Med* 2013; **368**: 2159–68.

9 Guillén-Guío B, Acosta-Herrera M, Villar J, Flores C. Genetics of Acute Respiratory Distress Syndrome. *eLS. John Wiley Sons* 2016. Doi: 10.4046/trd.2001.51.1.5

10 Lee S, Emond MJ, Bamshad MJ, *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; **91**: 224–37.

11 Shortt K, Chaudhary S, Grigoryev D, *et al.* Identification of novel single nucleotide polymorphisms associated with acute respiratory distress syndrome by exome-seq. *PLoS One* 2014; **9**: e111953.

12 Christie JD, Wurfel MM, Feng R, *et al.* Genome wide association identifies PPFIA1 as a candidate gene for acute lung injury risk following major trauma. *PLoS One* 2012; **7**: e28268.

13 Bime C, Pouladi N, Sammani S, *et al.* Genome-Wide Association Study in African Americans with Acute Respiratory Distress Syndrome Identifies the Selectin P Ligand Gene as a Risk Factor. *Am J Respir Crit Care Med* 2018; **197**: 1421–32.

14 Singer M, Deutschman CS, Seymour CW, *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016; **315**: 801–10.

15 ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, *et al.* Acute respiratory distress syndrome: the Berlin Definition. *JAMA* 2012; **307**: 2526–33.

16 Scherag A, Schöneweck F, Kesselmeier M, *et al.* Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28Day Mortality. *EBioMedicine* 2016; **12**: 239–46.

17 Reilly JP, Wang F, Jones TK, *et al.* Plasma angiopoietin-2 as a potential causal marker in sepsis-associated ARDS development: evidence from Mendelian randomization and mediation analysis. *Intensive Care Med* 2018; **44**: 1849–58.

18 Kanai M, Tanaka T, Okada Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J Hum Genet* 2016; **61**: 861–6.

19 Barratt S, Medford AR, Millar AB. Vascular endothelial growth factor in acute lung injury and acute respiratory distress syndrome. *Respiration* 2014; **87**: 329–42.

20 Barreto-Luis A, Pino-Yanes M, Corrales A, *et al*. Genome-wide association study in Spanish identifies ADAM metallopeptidase with thrombospondin type 1 motif, 9 (ADAMTS9), as a novel asthma susceptibility gene. *J Allergy Clin Immunol* 2016; **137**:964–6.

21 Guillen-Guio B, Lorenzo-Salazar JM, González-Montelongo R, *et al*. Genomic analyses of human European diversity at the southwestern edge: isolation, African influence and disease associations in the Canary Islands. *Mol Biol Evol* 2018; **35**:3010–3026.

22 Kendall RL, Thomas KA. Inhibition of vascular endothelial cell growth factor activity by an endogenously encoded soluble receptor. *Proc Natl Acad Sci U S A* 1993; **90**: 10705–9.

23 Mifsud B, Tavares-Cadete F, Young AN, *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* 2015; **47**: 598–606.

24 Astle WJ, Elding H, Jiang T, *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 2016; **167**: 1415–1429.e19.

25 Kim JY, Hildebrandt MAT, Pu X, *et al.* Variations in the vascular endothelial growth factor pathway predict pulmonary complications. *Ann Thorac Surg* 2012; **94**: 1079-84; discussion 1084–5.

26 The CARDIoGRAMplusC4D Consortium. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013; **45**: 25–33.

27 McGinnis R, Steinthorsdottir V, Williams NO, *et al.* Variants in the fetal genome near FLT1 are associated with risk of preeclampsia. *Nat Genet* 2017; **49**:1255–1260.

28 Gray KJ, Saxena R, Karumanchi SA. Genetic predisposition to preeclampsia is conferred by fetal DNA variants near FLT1, a gene involved in the regulation of angiogenesis. *Am J Obstet Gynecol* 2018; **218**:211–218.

29 Becker PM, Verin AD, Booth MA, Liu F, Birukova A, Garcia JG. Differential regulation of diverse physiological responses to VEGF in pulmonary endothelial cells. *Am J Physiol Lung Cell Mol Physiol* 2001; **281**: L1500–11.

30 Nagy JA, Dvorak AM, Dvorak HF. VEGF-A(164/165) and PlGF: roles in angiogenesis and arteriogenesis. *Trends Cardiovasc Med* 2003; **13**: 169–75.

31 Strouvalis I, Routsi C, Adamopoulou M, *et al.* Early increase of VEGF-A is associated with resolution of ventilator-associated pneumonia: Clinical and experimental evidence. *Respirology* 2018; **23**:942–949.

32 Ware LB, Kaner RJ, Crystal RG, *et al*. VEGF levels in the alveolar compartment do not distinguish between ARDS and hydrostatic pulmonary oedema. *Eur Respir J* 2005; **26**:101–5.

33 Guo J, Yan W, Yang Y, Wang Z, Tian F. Monitoring of vascular endothelial growth factor and its soluble receptor levels in early trauma. *J Trauma Acute Care Surg* 2017; **82**: 766–70.

34 Shapiro NI, Schuetz P, Yano K, *et al.* The association of endothelial cell signaling, severity of illness, and organ dysfunction in sepsis. *Crit Care* 2010; **14**: R182.

35 Skibsted S, Jones AE, Puskarich MA, *et al.* Biomarkers of endothelial cell activation in early sepsis. *Shock* 2013; **39**: 427–32.

36 Hou PC, Filbin MR, Wang H, *et al.* Endothelial Permeability and Hemostasis in Septic Shock: Results From the ProCESS Trial. *Chest* 2017; **152**: 22–31.

37 Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018; **46**: D1074–82.

38 Napolitano F, Carrella D, Mandriani B, *et al.* gene2drug: a computational tool for pathway-based rational drug repositioning. *Bioinformatics* 2018; **34**: 1498–505.

39 Richeldi L, du Bois RM, Raghu G, *et al.* Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* 2014; **370**: 2071–82.

40 Nacev BA, Grassi P, Dell A, Haslam SM, Liu JO. The Antifungal Drug Itraconazole Inhibits Vascular Endothelial Growth Factor Receptor 2 (VEGFR2) Glycosylation, Trafficking, and Signaling in Endothelial Cells. *J Biol Chem* 2011; **286**: 44045–56.

Figure 1. **Flow chart of quality control steps for the samples and genotyped SNPs in the discovery and replication stages.** CR, Call rate; DQC, Affymetrix dish quality control; FLD, Fisher´s Linear Discriminant; HetSO, Heterozygous Cluster Strength Offset; HWE, Hardy-Weinberg Equilibrium; MAF, minor allele frequency; mtDNA, mitochondrial DNA; Y-chr, Y chromosome.

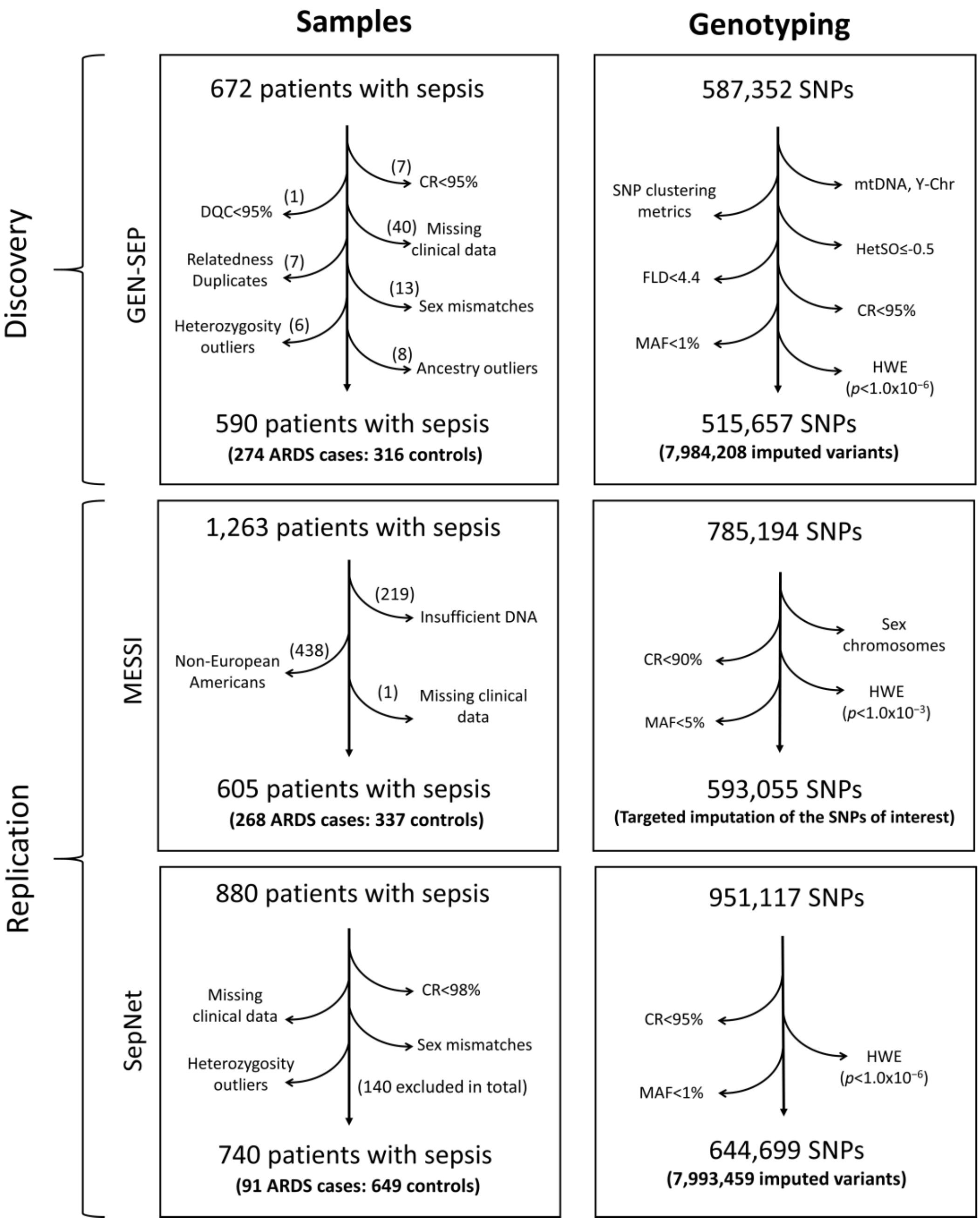


Figure 2. Manhattan plot of GWAS results for the discovery stage. Axes display the -$\log_{10}$ transformed p-values by position in each chromosome. The horizontal line indicates the threshold considered for prioritizing variants for the replication stage (*p*=5·0x10^{-5}).

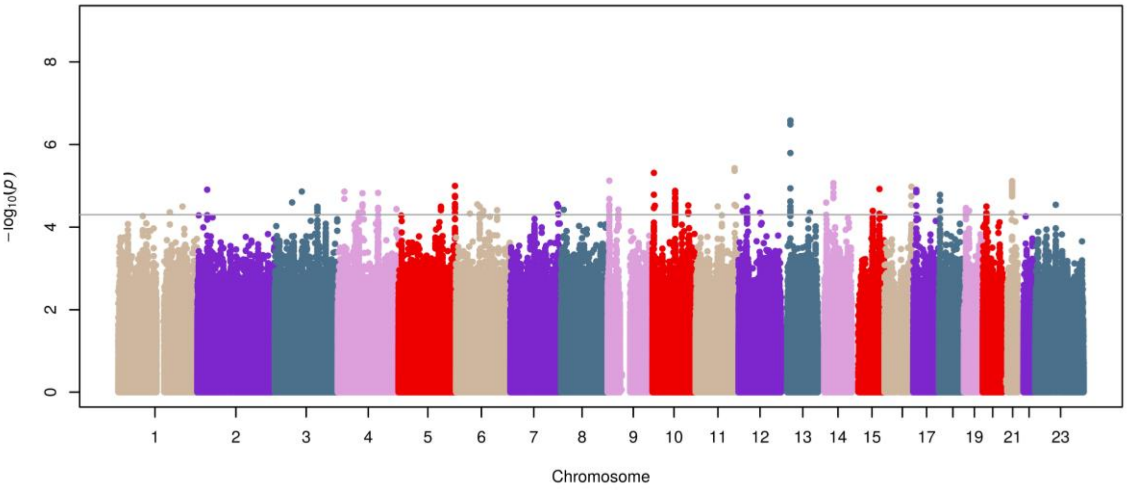


Figure 3. Regional plot of association results for the genome-wide significant locus. The -$\log_{10}$ transformed p-values for association tests are plotted by position. The SNP rs number indicated on the plot denotes the sentinel SNP. The remaining SNPs are color coded to reflect their degree of linkage disequilibrium with the indicated SNP based on pairwise r^2 values from the European population from The 1000 Genomes Project. Estimated recombination rates (light blue line) are plotted on the right y-axis.

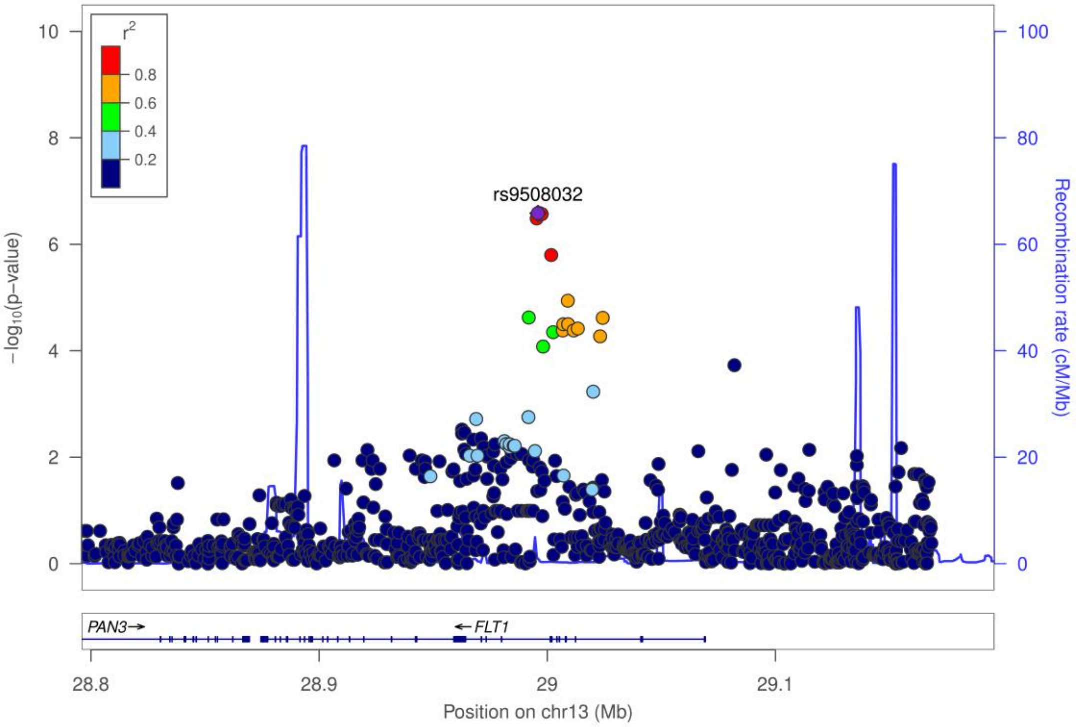


Figure 4. Luciferase reporter assay to assess the role of intron 10 and of rs9508032 and its perfect LD proxies on *FLT1* promoter activity. A) Scheme of vector constructs. B) Experimental data showing that the intron 10 fragment harboring the reference alleles suppresses the *FLT1* promoter activity in A549 and THP-1 cells. C) Experimental data showing that the intron 10 fragment harboring the alternative alleles further decreased the *FLT1* promoter activity, showing a significant difference in THP-1 cells. Significance was assessed by Wilcoxon signed-rank tests (**p*<0·05, #*p*<0·005, §*p*<0·0005). Ref and Alt indicate risk and protective alleles, respectively.

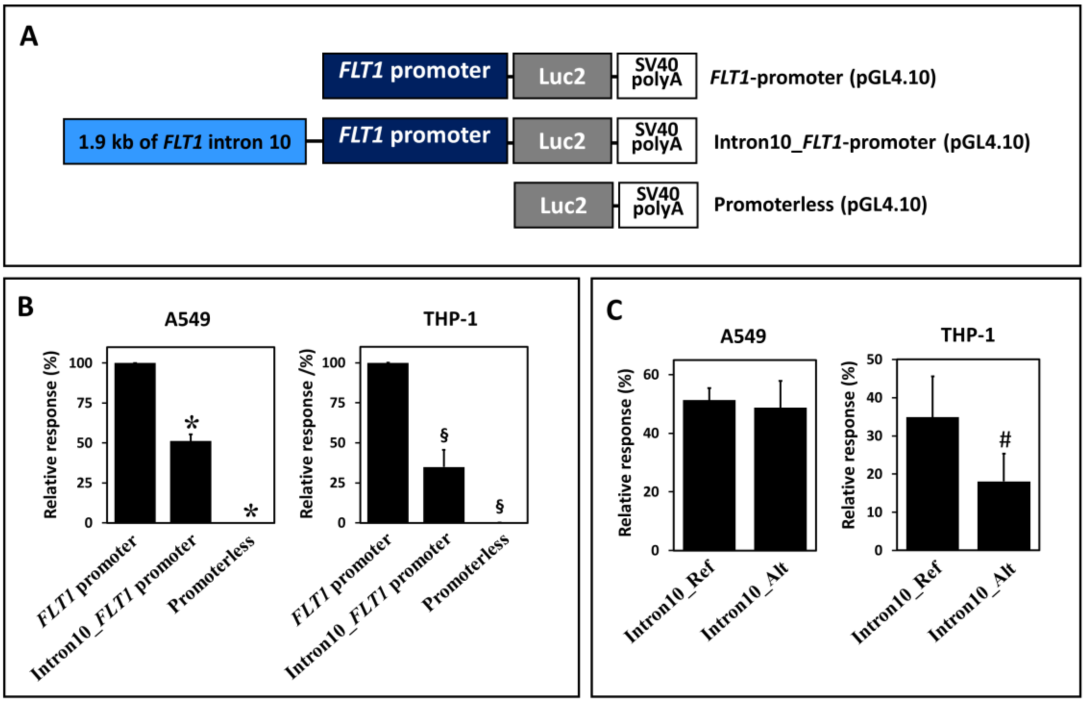


Table 1. Demographic and clinical features of the GEN-SEP study.

	Controls (n=316)	**Cases (n=274)**	***p*-value†**
Sex (n males/N)	197/316 (62·3%)	194/274 (70·8%)	0·04
Mean age (years)*	63·0 ± 15·0	62·5 ± 14·1	0·47
Hypertension (n/N)	60/144 (41·7%)	73/160 (45·6%)	0·56
Smokers (n/N)	61/188 (32·4%)	59/175 (33·7%)	0·88
Previous surgery (%)	32/127 (25·2%)	35/137 (25·5%)	1·00
Ischemic cardiac disease (n/N)	31/285 (10·9%)	19/210 (9·0%)	0·61
Pulmonary sepsis (n/N)	83/267 (31·1%)	128/252 (50·8%)	$7{\cdot}5\text{x}10^{-6}$
APACHE II (median) (P_{25}–P_{75})*	20 (15-24)	22 (18-27)	$2{\cdot}2\text{x}10^{-5}$
ICU mortality (n/N)	79/310 (25·5%)	115/268 (42·9%)	$1{\cdot}5\text{x}10^{-5}$
Pathogen (n/N)			
Gram-positive	48/178 (27·0%)	58/162 (35·8%)	0·09
Gram-negative	74/178 (41·6%)	59/162 (36·4%)	0·41
Gram-positive and Gram-negative	26/178 (14·6%)	17/162 (10·5%)	0·34
Fungi	5/178 (2·8%)	3/162 (1·9%)	0·83
Virus	2/178 (1·1%)	9/162 (5·6%)	0·04
Polymicrobial	16/178 (9·0%)	11/162 (6·8%)	0·45
Organ dysfunction (n/N)			
Circulatory	232/270 (85·9%)	238/255 (93·3%)	0·01
Coagulation	62/270 (23·0%)	68/255 (26·7%)	0·38
Hepatic	48/269 (17·8%)	41/255 (16·1%)	0·67
Neurologic	54/270 (20·0%)	59/254 (23·2%)	0·43
Renal	124/316 (39·2%)	108/274 (39·4%)	1·00

n=number of individuals with data available, N=group size. *All individuals have age and APACHE II data. Percentages refer only to the individuals with available data for each clinical feature. †Mean age and APACHE II comparisons were conducted by the Wilcoxon signed-rank test; the other variables were compared by a chi-square test. APACHE II, Acute Physiology and Chronic Health Evaluation II; ARDS, acute respiratory distress syndrome; ICU, intensive care unit; P25, percentile 25; P75, percentile 75.

Table 2. Results for the SNPs that were nominally significant in the replication stage.

					Discovery (274:316)*			Replication (359:986)*			Meta-analysis (633:1,302)*	
Variant	Chr	Position	Gene	A1/A2	MAF	OR [95% CI]	*p*-value	MAF†	OR [95% CI]	*p*-value	OR [95% CI]	*p*-value
rs9508032	**13**	**28995940**	***FLT1***	**T/C**	**0·29**	**0·49 [0·38, 0·65]**	**2·62x10^{-7}**	**0·25**	**0·74 [0·60, 0·92]**	**5·98x10^{-3}**	**0·61 [0·41, 0·91]**	**5·18x10^{-8}**
rs11195238	10	112388857	*SMC3-RBM20*	T/C	0·15	0·47 [0·33, 0·67]	2·97x10^{-5}	0·13	0·69 [0·52, 0·91]	9·55x10^{-3}	0·60 [0·48, 0·74]	3·73x10^{-6}
rs2734600	9	33753355	*PRSS3*	T/C	0·15	0·48 [0·33, 0·68]	3·76x10^{-5}	0·12	0·85 [0·36, 1·97]	0·02	0·52 [0·37, 0·72]	7·76x10^{-5}
rs8001184	13	90603540	*SLITRK5-GPC5*	A/C	0·48	1·65 [1·30, 2·10]	4·48x10^{-5}	0·49	1·26 [1·04, 1·52]	0·02	1·40 [1·20, 1·62]	1·19x10^{-5}

*Cases:Controls; † For rs9508032, MAF was 0·27 in MESSI and 0·22 in SepNet; For rs11195238, MAF was 0·14 in MESSI and 0·12 in SepNet. Fort the other two SNPs, MAFs were identical in the two replication cohorts. A1, Effect allele; A2, Non-effect allele; CI, Confidence Interval; MAF, Minor Allele Frequency; OR, Odd Ratio. Top hit is indicated in bold. Alleles referred to the positive strand of hg19. The imputation quality (Rsq/INFO) for the four variants ranged from 0·93 to 1·00 in all studies.

Genome-wide association study of sepsis-associated acute respiratory distress syndrome in individuals of European ancestry

Beatriz Guillen-Guio, Jose M. Lorenzo-Salazar, Shwu-Fan Ma, Pei-Chi Hou, Tamara Hernandez-Beeftink, Almudena Corrales, M. Isabel García-Laorden, Jonathan Jou, Elena Espinosa, Arturo Muriel, David Domínguez, Leonardo Lorente, María M. Martín, Carlos Rodríguez-Gallego, Jordi Solé-Violán, Alfonso Ambrós, Demetrio Carriedo, Jesús Blanco, José M. Añón, John P. Reilly, Tiffanie K. Jones, Caroline A.G. Ittner, Rui Feng, Franziska Schöneweck, Michael Kiehntopf, Imre Noth, Markus Scholz, Frank M. Brunkhorst, André Scherag, Nuala J. Meyer, Jesús Villar, and Carlos Flores.

Appendix

The GEN-SEP cohort

GEN-SEP is a national, multicenter, observational study conducted in a Spanish network of 11 intensive care units (ICUs) between 2,002 and 2,017. The list of Spanish hospitals involved in this study are: Hospital Universitario de Canarias, Tenerife; Hospital Universitario NS de Candelaria, Tenerife; Hospital Universitario Río Hortega, Valladolid; Hospital Universitario Dr. Negrin, Gran Canaria; Hospital General de Ciudad Real, Ciudad Real; Complejo Hospitalario Universitario de León, León; Hospital Virgen de la Luz, Cuenca; Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela; Fundació Althaia-Manresa, Barcelona; Hospital Clinic, Barcelona; and Hospital Clínico de Valladolid¸ Valladolid.

A total of 672 patients with sepsis[1] were included in this stage and diagnosed with ARDS based on Berlin definition criteria:[2] 1) acute onset with PaO2/FiO2 <300 mmHg, 2) bilateral pulmonary infiltrates on frontal chest radiograph, 3) use of invasive mechanical ventilation, and 4) no evidence of cardiac failure. Controls were those sepsis patients that did not develop ARDS during their ICU stay.

Four ml of peripheral blood were withdrawn at the time of inclusion into the study and stored at -20°C until use. DNA was extracted using the Illustra™ blood genomicPrep Mini Spin Kit (GE Healthcare), quantified with a Qubit 3.0 fluorometer (Thermo Fisher Scientific), and stored at -20°C until use. Samples with a low concentration of DNA (<10 ng/µl) were cleaned and concentrated using the RealClean & concentrator microspin kit (Real Laboratory).

Genotyping and quality control in discovery and replication stages

For the discovery stage, a total of 587,352 SNPs were genotyped in the National Genotyping Centre (CeGen) using the Axiom Genome-Wide Human CEU 1 Array (Affymetrix). Variant calling was performed in a single batch for all samples using AffyPipe[3] following the authors' recommendations to fine tune the filtering of low quality SNPs and samples. PLINK v1.90[4] and R 3.3.2[5] tools were used to conduct quality controls. Samples with missing clinical information, genotype call rates < 95%, sex mismatches between genotypes and the clinical data, samples with high degree of kinship (PIHAT>0·2), and heterozygosity outliers were removed. Variants with low minor allele frequency (MAF<0·01), genotype call rates (CR) < 95%, or deviated from Hardy Weinberg equilibrium expectations (HWE, $p<1{\cdot}0\text{x}10^{-6}$) were excluded. Additionally, a Principal Component (PC) analysis (PCA) was conducted to reduce the effects of population stratification in the analysis. For this purpose, we removed SNPs located at known regions that are in long-distance linkage disequilibrium (LD). We then pruned SNPs in high LD using the function "*indep-pairwise*" of PLINK, setting a r^2 of 0·15 to keep approximately 100,000 independent variants. After excluding eight ancestry outliers, the two first PCs for the discovery sample were plotted overlaid with the HapMap3 populations.[6] The PCA evidenced the similarity between the GEN-SEP samples included in the discovery stage and the European population from HapMap (Supplementary Figure 1).

In the MESSI study, SNPs were genotyped using the Affymetrix Axiom TxArray v.1 (Affymetrix). As described elsewhere,[7] variants were filtered if they were located on sex chromosomes, had a MAF<5%, had a missing genotype rate of >10%, or if deviated from HWE ($p<1{\cdot}0\text{x}10^{-3}$). In the SepNet study, HumanOmniExpressExome arrays (Illumina, Inc.) were used for variant genotyping. As described elsewhere,[8] individuals with sex mismatches, missing sex records, CR<98%, implausible heterozygosity (<20% and >26%), and ancestry outliers based on the PCA were removed. Variants with CR≤95%, MAF<1%, or deviated from HWE ($p<1{\cdot}0\text{x}10^{-6}$) were also excluded.

Statistical analyses for discovery stage

For variant imputation, phasing was conducted with SHAPEIT v2.r790[9] and the Haplotype Reference Consortium (HRC) version r.1.1 data were used as the reference panel[10] on the Michigan Imputation Server[11]. Logistic regression models assuming an additive inheritance were carried out using EPACTS v3.2.6[12] based on the Wald test. We included sex, age, and the Acute Physiology and Chronic Health Evaluation II (APACHE II) score as covariates. For the variants in the X chromosome, variant imputation and association tests were conducted separately in males and females, and results were subsequently meta-analysed with METASOFT v2.0.1[13]. Fixed-effects or Han and Eskin's Random Effects models were used depending on the significance of the Cochran's Q statistic. Variants with low allele frequency (MAF<1%) or with a low imputation quality (Rsq<0·3) were excluded from the analysis. The genomic inflation factor (λ) of the results was calculated with the GenABEL package v1.8-0.[14]

GCTA-COJO 1.26.0[15] was used for conditional regression analyses to identify independent loci taking into account the underlying LD structure in the study sample. Independent variants showing a p<5·0x10^{-5} were followed up in the replication stage. Regional association plots were generated using LocusZoom[16] based on LD information of European populations from the 1000 Genomes Project (1KGP)[17] and gene information from the UCSC browser data.

Statistical analyses for replication stage

Pre-phasing and variant imputation in MESSI were conducted with MACH v1.0,[18] using the European population from 1KGP Phase 1 v2 as the reference panel. Logistic regression models were performed assuming additive inheritance using R 3.3.2 stats package (glm function, binomial distribution),[5] considering the first two PCs, age, and sex as covariates. All the investigated variants in MESSI had a MAF >1% and Rsq >0·3. As described elsewhere,[8] for the SepNet study SHAPEIT v2.r790[9] was utilized for pre-phasing, and IMPUTE2 v2.3.1[19] was used for variant imputation considering the 1KGP Phase 1 v3 data as the reference panel. Logistic regressions were performed with SNPTEST v2.5,[19] which included the first three PCs, sex, age, and APACHE II as covariates. All the assessed variants in SepNet had MAF >1%, no evidence for deviations from HWE (p>1·0x10^{-10}) and an INFO Score >0·8. To combine the results from MESSI and SepNet, a meta-analysis was assessed using METASOFT v2.0.1.[13]

Meta-analysis of discovery and replication stages

A meta-analysis across discovery and replication stages was performed with METASOFT v2.0.1[13] to estimate the overall effect size of the SNPs reaching nominal significance in replication stage. Fixed-effects or Han and Eskin's Random-effects models were used based on the Cochran's Q test significance. Genome-wide significance was declared with a meta-analysed significance of p<9·26x10^{-8} according to the most recent empirical estimations in European populations.[20] The same approximation was used for the sensitivity analysis of the association of the sentinel variant.

Statistical power

Statistical power was estimated using the Genetic Association Study (GAS) Power Calculator.[21] We assumed a multiplicative model, a GWAS with a sample size of 630 cases and 1,302 controls, a relative risk of 1·5 and a prevalence of 0·1, the study had 80·4% statistical power for detecting associated variants with MAF of 0·30 or greater at significance level of p<9·26x10^{-8}.

FLT1 and *VEGFA* gene expression and functional annotation of genetic variants

Total RNAs from nine lung biopsies of healthy individuals obtained from the Gift of Hope Network Regional Organ Bank of Il (GOH/ROBI) were isolated and subjected to RNA-sequencing. Expression levels of *FLT1* and *VEGFA* were expressed in counts per million (Shwu-Fan Ma and Imre Noth, personal communication). Additionally, the ExAtlas tool[22] was used to explore public gene expression data available (GSE32707) from a peripheral blood transcriptomics analysis in 88 critically-ill adult patients that were evaluated for sepsis and ARDS.[23] For this analysis we used ANOVA followed by pairwise Student's *t*-tests to assess the differences in average intensities of the array probe targeting *FLT1* (ILMN_1752307, which targets exon 30, that is found in the canonical isoform FLT1-201) and *VEGFA* (ILMN_2375879, ILMN_1693060, and ILMN_1803882) between ICU controls (n=34), systemic inflammatory response syndrome (SIRS, n=21), sepsis (n=30), and sepsis-associated ARDS patients (n=18) at study inclusion. We report uncorrected two-sided *p*-values.

Next, we used a combination of tools and datasets to evaluate the regulatory potential of the associated variants in gene expression (through epigenetic mechanisms, long-distance physical interactions, and tissue-specific cis- expression quantitative trait loci (eQTLs)), and the likelihood of deleteriousness. These included Capture Hi-C Plotter (CHiCP),[24] DeepSea,[25] DSNetwork,[26] GTEx Analysis Release V7,[27] HaploReg v4.1,[28] Open Targets Genetics,[29] RegulomeDB,[30] SNPdelScore,[31] TIVAN,[32] and Variant Effect Predictor (VEP).[33]

CHiCP allows for the determination of empirically-observed physical interactions between distal DNA regulatory elements and gene promoters in multiple tissues. DeepSea predicts the epigenetics state of a sequence and prioritize regulatory variants by calculating functional significance scores, while DSNetwork allows for the selection of the most probable functional SNP from a list of variants according to nearly sixty prediction approaches. The GTEx Portal allows for the study of Single-Tissue eQTL and tissue-specific gene expression and regulation. HaploReg, Open Targets Genetics, and RegulomeDB explore annotations of coding and non-coding variants integrating data from chromatin states, regulatory motifs, eQTLs, pQTLs, DNase I hypersensitive sites, enhancer-transcription start sites, and promoter capture Hi-C experiments from different cell lines. Open Targets Genetics puts functional information in the context of the UK Biobank association evidence, allowing one to link each variant to its proximal and distal target gene(s), using a single evidence score. SNPdelScore combines different methods to address deleterious effects of noncoding variants, including the CellulAr dePendent dEactivating (CAPE) mutations predictor. Finally, TIVAN allows for the prediction of tissue-specific cis-eQTL single nucleotide variants, and VEP determines the effect of the variants analysed on genes, transcripts, protein sequence, and regulatory regions.

Constructs, transient transfections, and dual-luciferase reporter assays

A Dual-Luciferase Reporter Assay System® (Promega, Madison, WI) was used to evaluate the potential regulatory effect of the ARDS-associated variant on promoter activity. The reporter construct was generated by synthesizing (GenScript Inc, Piscataway, NJ) a fragment of 1,032 bp of the *FLT1* promoter (Ensembl ID: ENSR00000060438) plus 284 bp of the 5' UTR of exon one and 784 bp of the upstream sequence (chr13: 29,068,982-29,070,013; GRCh37/hg19 coordinates), and inserting it into a promoterless pGL4.10 [luc2] luciferase reporter vector (Promega). This *FLT1* promoter region was chosen for having the highest activity *in vitro* in a previous characterization of the gene promoter.[34] In addition, two regulatory constructs were generated by synthesizing (GenScript Inc, Piscataway, NJ) a 1·9 kb intron 10 fragment containing either the reference or alternative alleles of the most significantly associated variant within *FLT1* and its perfect LD proxies (i.e. rs9508032, rs9508034, rs722503, rs8002446, rs9513111, r^2=1·0) in Europeans (chr13: 28,995,800-28,997,700; GRCh37/hg19 coordinates) and inserting them into the reporter construct.

The constructed plasmids (50 ng DNA each) and the control plasmid pGL4.74 [hRluc/TK] (10 ng DNA) were transiently co-transfected into human lung epithelial (A549) or peripheral blood monocyte (THP-1) cell lines using the TransIT®-LT1 Transfection Reagent (Mirus Bio LLC, Madison, WI) following manufacturer's protocol. A549 and THP-1 cells were separately grown in 10% DMEM or 10% RPM 1640 media, respectively, and were plated into white 96-well plates until confluency. Twenty-four hours after transfection, cells were collected and luminescence was measured by Dual-Luciferase Reporter Assay System according to manufacturer's instructions using a Cytation5 plate reader (BioTek, Winooski, VT). Luminescence experiments were performed four to eight times, with each transfection in triplicate. Following manufacturer's instructions,[35] to reduce variability, simplify comparisons and improve significance, promoter activities were expressed as the relative response ratio of *Firefly* luciferase/Renilla luciferase luminescence according to the formula:

$$Relative\ response\ ratio = \frac{(experimental\ sample\ ratio) - (negative\ control\ ratio)}{(positive\ control\ ratio) - (negative\ control\ ratio)}$$

We considered the construct including only the $FLT1$ promoter as the positive control and the promoterless construct as the negative control (see figure 4A). Mean differences among the independent experimental groups were assessed by non-parametric Wilcoxon signed-rank test. Again, we report uncorrected two-sided p-values.

Literature mining of previously reported ARDS-associated genes

In order to evaluate genes that were previously associated with ARDS, we conducted a bibliographic search on PubMed for all studies reporting genes which were significantly associated with ARDS from December 2015 to November 2018. This updated result was merged with a list of all published studies we collected up to December 2015 available elsewhere.[36–38] For that search, we used combinations of the terms “acute respiratory distress syndrome”, “ARDS” OR “acute lung injury” with “polymorphism” OR “genetic variant” and retrieved seven publications reporting five additional candidate ARDS genes in adults. Association results in the discovery stage were extracted and an effective number of independent signals per gene was measured using the Genetic Type I Error Calculator[39] in order to adjust for multiple testing. Significant association was declared if any of the individual variants surpassed one of two Bonferroni-corrected significant levels. We considered a study-wise adjustment accounting for all the independent tests across all genes, and a gene-wise adjustment just accounting - i.e. adjusting for the independent variants mapping at individual genes.

Evaluation of *FLT1* variants in a trauma-associated ARDS cohort

We evaluated if the *FLT1* sentinel variant and perfect proxies also associated with non-sepsis ARDS. For that, we accessed the table S2 of the only GWAS of trauma-associated ARDS published to date,[40] containing publicly available (but incomplete) summary data. We found that none of the *FLT1* variants that achieved genome-wide significance in sepsis-associated ARDS were present in that study because of the reference panel used for variant imputation. Despite this, it was reassuring to find that out of 13 *FLT1* SNPs listed (all within a region of 31 kb and showing nominally significant associations with ARDS after trauma), six were also located in intron 10 (p-values in the range of $9{\cdot}15\text{x}10^{-4}$ to $2{\cdot}44\text{x}10^{-3}$). However, their LD with the sentinel variant of our study was weak in Europeans (r^2=0·13).

References

1 Singer M, Deutschman CS, Seymour CW, *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016; **315**: 801–10.

2 ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, *et al.* Acute respiratory distress syndrome: the Berlin Definition. *JAMA* 2012; **307**: 2526–33.

3 Nicolazzi EL, Iamartino D, Williams JL. AffyPipe: an open-source pipeline for Affymetrix Axiom genotyping workflow. *Bioinformatics* 2014; **30**: 3118–9.

4 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 7.

5 R Core Team. R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria* 2013. http://www.r-project.org/. Accessed September 2017.

6 International HapMap 3 Consortium, Altshuler DM, Gibbs RA, *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–8.

7 Reilly JP, Wang F, Jones TK, *et al.* Plasma angiopoietin-2 as a potential causal marker in sepsis-associated ARDS development: evidence from Mendelian randomization and mediation analysis. *Intensive Care Med* 2018; **44**: 1849–58.

8 Scherag A, Schöneweck F, Kesselmeier M, *et al.* Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28Day Mortality. *EBioMedicine* 2016; **12**: 239–46.

9 Delaneau O, Howie B, Cox AJ, Zagury J-F, Marchini J. Haplotype estimation using sequencing reads. *Am J Hum Genet* 2013; **93**: 687–96.

10 McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; **48**: 1279–83.

11 Das S, Forer L, Schönherr S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016; **48**: 1284–7.

12 Kang H. Efficient and parallelisable association container toolbox (EPACTS). http://genome.sph.umich.edu/wiki/EPACTS 2014. Accessed October 2017.

13 Han B, Eskin E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *Am J Hum Genet* 2011; **88**: 586–98.

14 Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association

analysis. *Bioinformatics* 2007; **23**: 1294–6.

15 Zhu Z, Zheng Z, Zhang F, *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun* 2018; **9**: 224.

16 Pruim RJ, Welch RP, Sanna S, *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; **26**: 2336–7.

17 1000 Genomes Project Consortium RA, Auton A, Brooks LD, *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

18 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–34.

19 Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39**: 906–13.

20 Kanai M, Tanaka T, Okada Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J Hum Genet* 2016; **61**: 861–6.

21 Johnson JL, Abecasis GR. GAS Power Calculator: web-based power calculator for genetic association studies. *bioRxiv* 2017. Doi: https://doi.org/10.1101/164343.

22 Sharov AA, Schlessinger D, Ko MSH. ExAtlas: An interactive online tool for meta-analysis of gene expression data. *J Bioinform Comput Biol* 2015; **13**: 1550019.

23 Dolinay T, Kim YS, Howrylak J, *et al.* Inflammasome-regulated cytokines are critical mediators of acute lung injury. *Am J Respir Crit Care Med* 2012; **185**: 1225–34.

24 Schofield EC, Carver T, Achuthan P, *et al.* CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics* 2016; **32**: 2511–3.

25 Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015; **12**: 931–4.

26 Lemaçon A, Scott-Boyer M-P, Soucy P, Ongaro-Carcy R, Simard J, Droit A. DSNetwork: An integrative approach to visualize predictions of variants' deleteriousness. *bioRxiv 526335* 2019. Doi:https://doi.org/10.1101/526335.

27 GTEx Consortium J, Thomas J, Salvatore M, *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013; **45**: 580–5.

28 Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and

regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012; **40**: D930-4.

29 Carvalho-Silva D, Pierleoni A, Pignatelli M, *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res* 2019; **47**: D1056–65.

30 Boyle AP, Hong EL, Hariharan M, *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012; **22**: 1790–7.

31 Vera Alvarez R, Li S, Landsman D, Ovcharenko I. SNPDelScore: combining multiple methods to score deleterious effects of noncoding mutations in the human genome. *Bioinformatics* 2017; **34**: 289–91.

32 Chen L, Wang Y, Yao B, Mitra A, Wang X, Qin X. TIVAN: Tissue-specific cis-eQTL single nucleotide variant annotation and prediction. *Bioinformatics* 2018; published online Oct 10. Doi:10.1093/bioinformatics/bty872.

33 McLaren W, Gil L, Hunt SE, *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 2016; **17**: 122.

34 Morishita K, Johnson DE, Williams LT. A novel promoter for vascular endothelial growth factor receptor (flt-1) that confers endothelial-specific gene expression. *J Biol Chem* 1995; **270**: 27948–53.

35 Eggers C, Hook B, Lewis S, Strayer C, Landreman A. Designing a Bioluminescent Reporter Assay: Normalization. *Promega Corp. Web site.* http://www.promega.es/resources/pubhub/designing-a-bioluminescent-reporter-assay-normalization/ 2016. Accessed April 23, 2019.

36 Flores C, Pino-Yanes M del M, Villar J. A quality assessment of genetic association studies supporting susceptibility and outcome in acute lung injury. *Crit Care* 2008; **12**: R130.

37 Acosta-Herrera M, Pino-Yanes M, Perez-Mendez L, Villar J, Flores C. Assessing the quality of studies supporting genetic susceptibility and outcomes of ARDS. *Front Genet* 2014; **5**: 20.

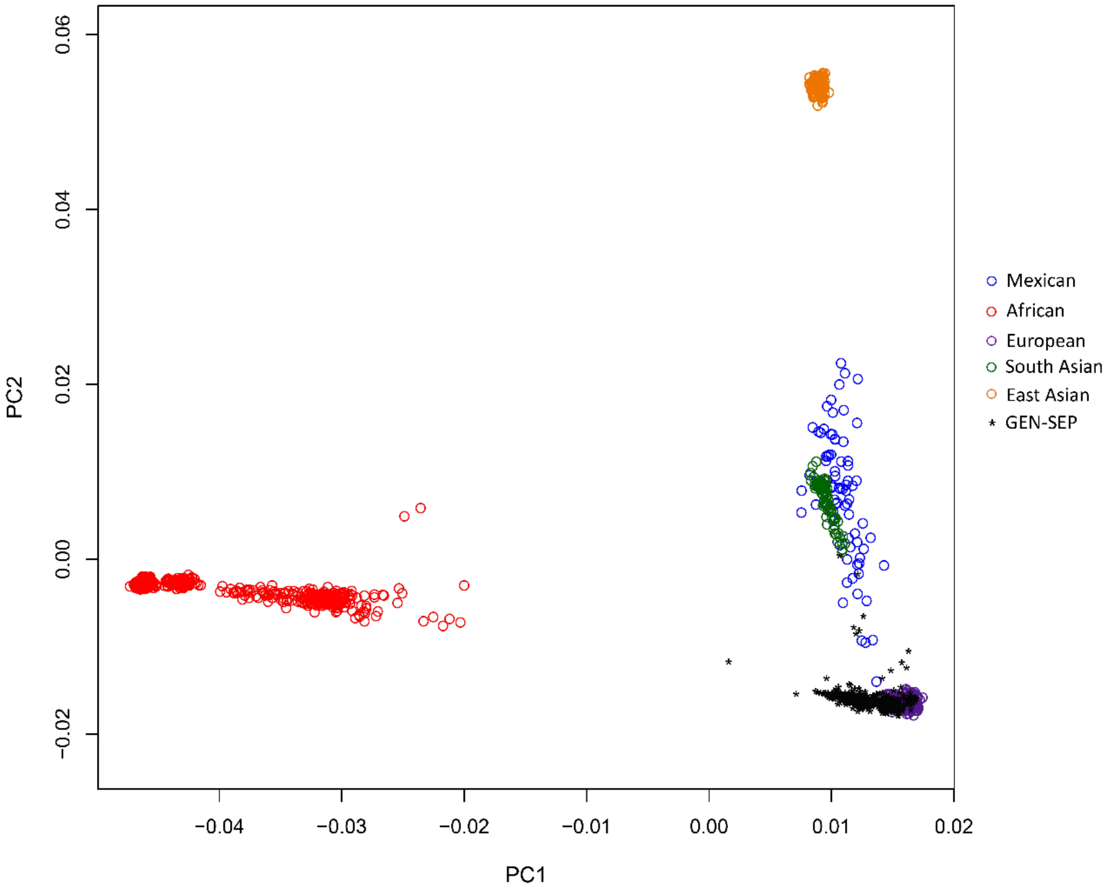
38 Guillén-Guío B, Acosta-Herrera M, Villar J, Flores C. Genetics of Acute Respiratory Distress Syndrome. *eLS. John Wiley Sons* 2016. DOI: 10.4046/trd.2001.51.1.5

39 Li M-X, Yeung JMY, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet* 2012; **131**: 747–56.

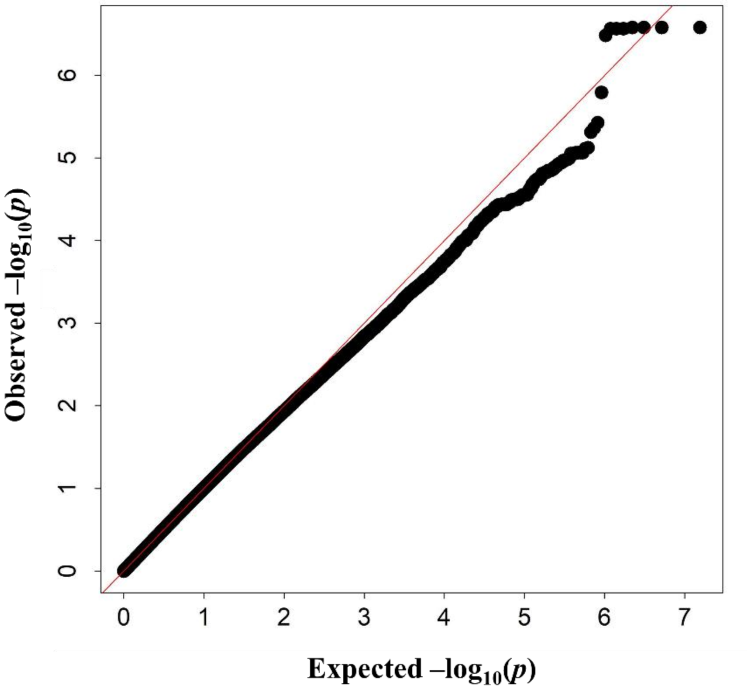
40 Christie JD, Wurfel MM, Feng R, et al. Genome wide association identifies PPFIA1 as a candidate gene for acute lung injury risk following major trauma. PLoS One 2012; **7**: e28268.

Supplementary Figures

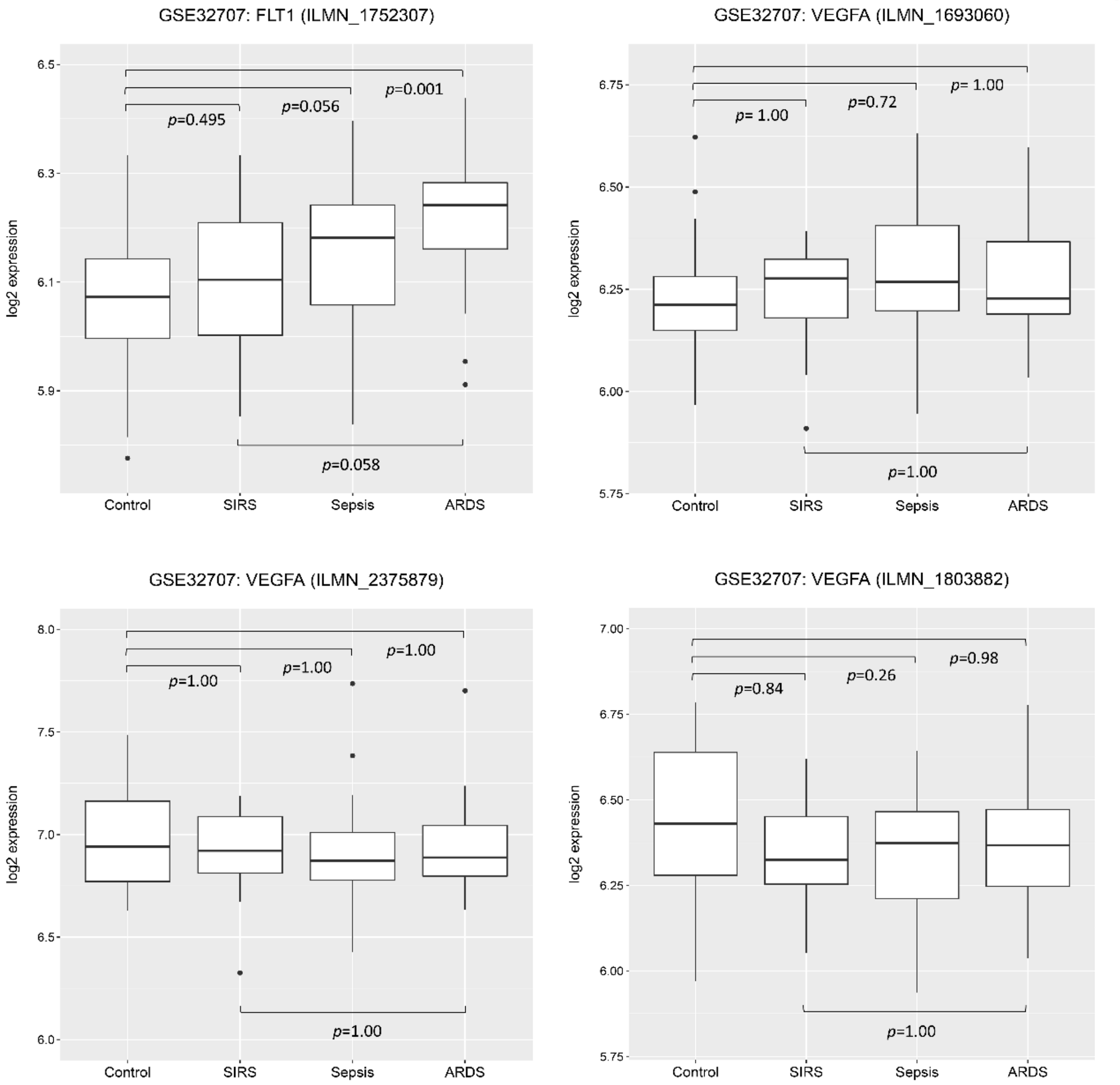
Supplementary Figure 1. Principal component analysis. Plot of the first two principal components (PC) of individuals analyzed in the discovery stage were projected on the HapMap3 reference dataset**.**



Supplementary Figure 2. Quantile-Quantile (Q-Q) plot. Observed versus expected -log$_{10}$ *p*-values for the GWAS results of the discovery stage.



Supplementary Figure 3. *FLT1* and *VEGFA* gene expression in critical care patients. Probe intensities of expression arrays obtained from peripheral blood samples from ICU controls (n=34), systemic inflammatory response syndrome (SIRS, n=21), sepsis (n=30) and sepsis-associated ARDS patients (n=18) at study inclusion. Note that the probe ILMN_1752307 targets *FLT1* exon 30, which is found in the canonical isoform (FLT1-201), one of the most highly expressed isoforms of the gene. Differences in average intensities were assessed using ANOVA followed by t-tests. GEO accession: GSE32707.[23]



Supplementary Tables

Supplementary Table 1. Top 53 independent SNPs associated with ARDS in the discovery stage (p<5·0x10^{-5}).

Variant	Chr	Position (hg19)	Gene	A1/A2	MAF	OR [95% CI]	*p*-value
rs598782	1	202572596	*SYT2*	T/C	0·173	0·49 [0·35, 0·69]	3·16x10^{-5}
rs10917581	1	162624504	*DDR2*	G/A	0·262	0·56 [0·42, 0·74]	4·37x10^{-5}
rs56865040	2	30907832	*LCLAT1-CAPN13*	G/A	0·066	3·35 [1·95, 5·77]	1·24x10^{-5}
rs58982889	3	85080936	*CADM2*	C/G	0·483	0·57 [0·44, 0·73]	1·37x10^{-5}
rs12494792	3	54631523	*CACNA2D3*	A/G	0·251	1·82 [1·38, 2·40]	2·52x10^{-5}
rs71331755	3	134040335	*RYK-AMOTL2*	C/G	0·237	1·83 [1·37, 2·43]	3·20x10^{-5}
rs76763432	4	20933002	*KCNIP4*	T/C	0·114	2·36 [1·60, 3·48]	1·38x10^{-5}
rs12513121	4	126763999	*FAT4-INTU*	A/C	0·302	0·55 [0·43, 0·72]	1·48x10^{-5}
rs11097547	4	77763070	*SHROOM3-SOWAHB*	G/A	0·200	1·95 [1·44, 2·64]	1·51x10^{-5}
rs78119818	4	78068598	*CCNI-CCNG2*	A/T	0·048	3·79 [2·03, 7·06]	2·81x10^{-5}
rs10518480	4	126898260	*FAT4-INTU*	G/A	0·202	1·93 [1·41, 2·64]	3·46x10^{-5}
rs66691935	4	184540486	*ING2-RWDD4*	T/C	0·175	1·97 [1·43, 2·71]	3·67x10^{-5}
rs62300402	4	66422737	*EPHA5*	G/A	0·276	0·56 [0·42, 0·74]	4·46x10^{-5}
rs66486976	5	177602232	*NHP2-GMCL2*	T/C	0·313	0·55 [0·42, 0·71]	1·00x10^{-5}
rs58681704	5	133268913	*FSTL4-C5orf15*	A/G	0·202	0·51 [0·37, 0·70]	3·18x10^{-5}
rs62390494	5	177493565	*FAM153C-N4BP3*	T/C	0·199	1·90 [1·40, 2·60]	4·53x10^{-5}
rs9453845	6	67330152	*EYS-ADGRB3*	T/G	0·107	0·41 [0·27, 0·62]	2·81x10^{-5}
rs3003179	6	74677167	*CD109-COL12A1*	A/G	0·279	1·76 [1·35, 2·30]	3·20x10^{-5}
rs58277258	6	129194762	*LAMA2*	C/T	0·084	2·69 [1·68, 4·30]	3·88x10^{-5}
rs12197618	6	85969855	*TBX18-NT5E*	A/G	0·053	3·58 [1·95, 6·57]	4·01x10^{-5}
rs9367172	6	43709993	*MRPS18A-VEGFA*	A/G	0·237	0·55 [0·41, 0·73]	4·69x10^{-5}
rs72611587	7	146905995	*CNTNAP2*	T/C	0·140	2·16 [1·51, 3·09]	2·78x10^{-5}
rs7777943	7	150483237	*GIMAP5-TMEM176B*	G/A	0·277	0·57 [0·44, 0·74]	3·18x10^{-5}
rs12678166	8	8520530	*PRAG1-CLDN23*	T/C	0·152	2·05 [1·46, 2·89]	3·80x10^{-5}
rs796455145	9	5487547	*PLGRKT*	C/T	0·411	1·74 [1·37, 2·22]	7·49x10^{-6}
rs4740791	9	4611901	*SPATA6L*	T/C	0·087	2·62 [1·66, 4·12]	3·14x10^{-5}
rs2734600	9	33753355	*PRSS3*	T/C	0·152	0·48 [0·33, 0·68]	3·76x10^{-5}
rs1751276	10	4477665	*KLF6-AKR1E2*	A/G	0·123	2·53 [1·70, 3·77]	4·85x10^{-6}
rs1867966	10	71187839	*TACR2-TSPAN15*	G/A	0·430	0·58 [0·45, 0·74]	1·32x10^{-5}
rs11195238	10	112388857	*SMC3-RBM20*	T/C	0·151	0·47 [0·33, 0·67]	2·97x10^{-5}
rs10795549	10	7582855	*SFMBT2-ITIH5*	C/A	0·478	0·60 [0·47, 0·76]	3·02x10^{-5}
rs10736526	11	122589092	*UBASH3B*	C/T	0·207	0·50 [0·37, 0·67]	3·73x10^{-6}
rs61710829	11	126566557	*KIRREL3*	G/C	0·378	1·71 [1·33, 2·20]	3·07x10^{-5}
rs602124	11	69388853	*MYEOV-CCND1*	C/G	0·297	1·77 [1·35, 2·31]	3·15x10^{-5}
rs76921243	12	26606699	*ITPR2*	A/G	0·056	0·25 [0·13, 0·47]	1·81x10^{-5}
rs1861180	12	12958559	*DDX47*	T/C	0·088	0·40 [0·25, 0·62]	4·07x10^{-5}
rs1904566	12	68125847	*DYRK2-IFNG*	C/A	0·420	1·69 [1·31, 2·17]	4·49x10^{-5}
rs9508032	13	28995940	*FLT1*	T/C	0·288	0·49 [0·38, 0·65]	2·62x10^{-7}
rs8001184	13	90603540	*SLITRK5-GPC5*	A/C	0·483	1·65 [1·30, 2·10]	4·48x10^{-5}
rs946626	14	49140883	*MDGA2-RPS29*	A/G	0·428	1·73 [1·36, 2·20]	8·59x10^{-6}
rs7161717	14	26389695	*STXBP6-NOVA1*	C/T	0·132	0·44 [0·30, 0·64]	2·54x10^{-5}
rs4887263	15	86626153	*AGBL1*	A/C	0·096	2·73 [1·74, 4·28]	1·19x10^{-5}
rs12902176	15	65518664	*CILP-PARP16*	G/A	0·268	1·78 [1·35, 2·35]	4·04x10^{-5}
rs11647343	16	84454267	*ATP2C2*	C/A	0·384	1·75 [1·36, 2·24]	1·05x10^{-5}
rs244783	16	84360055	*WFDC1*	T/G	0·212	1·97 [1·45, 2·69]	1·77x10^{-5}
rs4791367	17	9724374	*GLP2R*	G/A	0·092	0·35 [0·22, 0·56]	1·25x10^{-5}
rs9675656	18	2947220	*LPIN2*	C/G	0·109	2·44 [1·63, 3·66]	1·65x10^{-5}
rs397195	19	6619504	*CD70-TNFSF14*	G/C	0·354	1·71 [1·32, 2·20]	3·44x10^{-5}
rs285251	19	16415993	*AP1M1-KLF2*	C/T	0·286	0·56 [0·42, 0·74]	4·00x10^{-5}

rs6040856	20	11702045	*JAG1-BTBD3*	G/C	0·354	1·68 [1·32, 2·15]	3·16x10^{-5}
rs2831537	21	29516376	*ADAMTS5-N6AMT1*	T/C	0·462	0·57 [0·45, 0·73]	7·65x10^{-6}
rs4817154	21	28477085	*ADAMTS5-N6AMT1*	A/G	0·082	2·59 [1·64, 4·10]	4·57x10^{-5}
rs1155955	X	67297091	*OPHN1*	G/A	0·120	3·12 [1·83, 5·31]	2·87x10^{-5}

A1, Effect allele; A2, Non-effect allele; CI, Confidence Interval; MAF, Minor Allele Frequency; OR, Odd Ratio.

Supplementary Table 2. Main demographic and clinical features of individuals included in the study

	GEN-SEP			MESSI			SepNet			Comparison between studies	
	Controls (n=316)	**Cases (n=274)**	*p*-value*	**Controls (n=337)**	**Cases (n=268)**	*p*-value*	**Controls (n=649)**	**Cases (n=91)**	*p*-value*	*Controls* *p*-value*	*Cases* *p*-value*
Sex (n males/N)	197/316 (62·3%)	194/274 (70·8%)	0·04	200/337 (59·3%)	163/268 (60·8%)	0·74	379/649 (58·4%)	59/91 (64·8%)	0·50	0·50	0·05
Mean age (years)	63·0 ± 15·0	62·5 ± 14·1	0·47	61·5 ± 13·8	58·8 ± 14·1	0·02	65·4 ± 14.1	62·1 ± 13·6	0·02	0·04	0·88
Pneumonia (n/N)	83/267 (31·1%)	128/252 (50·8%)	7·50x10^{-6}	134/337 (39.8%)	166/268 (61·9%)	1·2x10^{-7}	249/646 (38·5%)	49/91 (53·8%)	5·0x10^{-3}	0·05	0·03
APACHE (median) (P_{25}–P_{75})†	20 (15-24)	22 (18-27)	2·22x10^{-5}	71 (57-88)	85 (67-104)	1·4x10^{-7}	20 (16-24)	19 (16-23)	0·47	0·42	1·6x10^{-3}
Mortality (n/N)‡	79/310 (25.5%)	115/268 (42.9%)	1·5x10^{-5}	127/337 (37·7%)	182/268 (67·9%)	2·2x10^{-13}	137/649 (21·1%)	12/91 (13·2%)	0·08	-	-

n=number of individuals with data available, N=group size. All individuals have age and APACHE data, except for MESSI cohort, where only 229 cases and 269 controls have APACHE data. *Categorical data compared by chi square test, continuous data compared by Wilcoxon test (two-sample comparison) or Kruskal-Wallis test (three-sample comparison); †APACHE III was measured for the MESSI Cohort and APACHE II for GENSEP and SepNet Cohorts. Therefore, only APACHE II scores were considered in the comparison between studies; ‡Patient mortality was not compared between studies since ICU mortality was considered for the GENSEP cohort, 30-day mortality was considered for the MESSI Cohort, and 28-day mortality for the SepNet cohort. APACHE, Acute Physiology and Chronic Health Evaluation; P25, percentile 25; P75, percentile 75.

Supplementary Table 3. Sensitivity analysis for the rs9508032 in the three cohorts together.

	OR [95% CI]	*p*-value
Unadjusted*	0·62 [0·43, 0·90]	1·07x10^{-7}
Sex	0·62 [0·43, 0·90]	1·11x10^{-7}
Age	0·62 [0·43, 0·90]	1·30x10^{-7}
APACHE†	0·61 [0·40, 0·93]	1·81x10^{-8}
Smokers‡	0·58 [0·42, 0·80]	9·72x10^{-4}††
Previous surgery§	0·64 [0·50, 0·83]	7·35x10^{-4}††
Ischemic cardiac disease§	0·56 [0·45, 0·70]	3·05x10^{-7}‡‡
Pulmonary sepsis	0·61 [0·41, 0·91]	9·12x10^{-8}‡‡
Mortality¶	0·62 [0·41, 0·94]	2·51x10^{-7}
Pathogen‡	0·48 [0·34, 0·68]	2·34x10^{-5}††
Multi organ dysfunction‖	0·61 [0·41, 0·91]	1·03x10^{-7}
Comorbidities**	0·62 [0·40, 0·94]	1·56x10^{-7}‡‡

*Unadjusted data for GEN-SEP, adjustment for 2 PC for MESSI, and adjustment for 3 PC for SepNet; †APACHE III was measured for the MESSI Cohort and APACHE II for GEN-SEP and SepNet Cohorts; ‡There was only information available for GEN-SEP study; §There was not information available for MESSI study; ¶ICU mortality was considered for the GENSEP cohort, 30-day mortality was considered for the MESSI Cohort, and 28-day mortality for the SepNet cohort; ‖Two or more affected organs; **For the GEN-SEP and SepNet studies, comorbidities considered are autoimmune diseases, cancer, chronic diseases, diabetes, hepatopathies, immunosuppression, kidney diseases, morbid obesity, pregnancy, severe infections, severe brain damage, and valvulopathies. For MESSI, comorbidities are defined as immunocompromise (cancer receiving treatment, hematologic malignancy, AIDS, metastatic cancer, or receiving immunosuppressive medication), cirrhosis, congestive heart failure, or chronic renal insufficiency including dialysis; ††Missing data for more than 35% of individuals from the GEN-SEP study; ‡‡Missing data for the 15-20% of individuals from the GEN-SEP study. APACHE, Acute Physiology and Chronic Health Evaluation; ICU, intensive care unit; OR, Odd Ratio; CI, Confidence Interval.

Supplementary Table 4. Association results for all SNPs within *FLT1* that have data available in all studies.

							Discovery (274:316)[*]		Replication (359:986)[*]		Meta-analysis (633:1302)[*]	
rs ID	Chr	Position	Location[†]	A1/A2	MAF	r^2	OR [95% CI]	*p*-value	OR [95% CI]	*p*-value	OR [95% CI]	*p*-value
rs9508032	**13**	**28995940**	**Intron 10**	**T/C**	**0·290**	**1·00**	**0·49 [0·38, 0·65]**	**2·62x10^{-7}**	**0·74 [0·60, 0·92]**	**5·98x10^{-3}**	**0·61 [0·41, 0·91]**	**5·18x10^{-8}**
rs8002446	13	28997400	Intron 10	A/G	0·288	1·00	0·50 [0·38, 0·65]	2·71x10^{-7}	0·75 [0·61, 0·92]	6·82x10^{-3}	0·61 [0·41, 0·92]	6·03x10^{-8}
rs9513111	13	28997563	Intron 10	T/C	0·288	1·00	0·50 [0·38, 0·65]	2·71x10^{-7}	0·75 [0·61, 0·92]	6·82x10^{-3}	0·61 [0·41, 0·92]	6·03x10^{-8}
rs722503	13	28997052	Intron 10	T/C	0·288	1·00	0·50 [0·38, 0·65]	2·71x10^{-7}	0·75 [0·61, 0·92]	6·94x10^{-3}	0·61 [0·41, 0·92]	6·11x10^{-8}
rs9508034	13	28996604	Intron 10	C/A	0·290	1·00	0·49 [0·38, 0·65]	2·62x10^{-7}	0·75 [0·61, 0·93]	7·58x10^{-3}	0·61 [0·41, 0·92]	6·47x10^{-8}
rs9508033	13	28996568	Intron 10	C/T	0·290	0·99	0·49 [0·38, 0·65]	2·62x10^{-7}	0·75 [0·61, 0·92]	7·19x10^{-3}	0·61 [0·41, 0·92]	6·18x10^{-8}
rs6491284	13	28995390	Intron 10	T/C	0·288	0·98	0·50 [0·38, 0·65]	3·26x10^{-7}	0·76 [0·61, 0·94]	9·91x10^{-3}	0·62 [0·41, 0·93]	1·05x10^{-7}
rs2281827	13	29001721	Intron 9	T/C	0·265	0·86	1·97 [1·49, 2·60]	1·60x10^{-6}	1·37 [1·10, 1·70]	4·20x10^{-3}	1·63 [1·14, 2·32]	1·69x10^{-7}
rs7324510	13	29007035	Intron 6	A/C	0·242	0·72	0·54 [0·40, 0·72]	3·17x10^{-5}	0·69 [0·55, 0·86]	1·01x10^{-3}	0·63 [0·53, 0·75]	2·81x10^{-7}
rs9513115	13	29011570	Intron 4	A/C	0·246	0·72	0·55 [0·41, 0·73]	4·17x10^{-5}	0·69 [0·55, 0·86]	1·12x10^{-3}	0·63 [0·53, 0·75]	3·69x10^{-7}
rs9513114	13	29009059	Intron 4	T/C	0·245	0·71	0·52 [0·39, 0·70]	1·15x10^{-5}	0·70 [0·56, 0·88]	1·76x10^{-3}	0·63 [0·53, 0·75]	2·62x10^{-7}
rs8001784	13	29009213	Intron 4	G/A	0·242	0·71	0·54 [0·40, 0·72]	3·17x10^{-5}	0·69 [0·56, 0·87]	1·28x10^{-3}	0·63 [0·53, 0·75]	3·68x10^{-7}
rs4771249	13	29013414	Intron 3	C/G	0·241	0·71	0·54 [0·41, 0·73]	3·84x10^{-5}	0·70 [0·56, 0·87]	1·36x10^{-3}	0·63 [0·53, 0·76]	4·53x10^{-7}
rs9508035	13	29009099	Intron 4	A/C	0·242	0·71	0·54 [0·40, 0·72]	3·17x10^{-5}	0·70 [0·56, 0·88]	1·76x10^{-3}	0·64 [0·53, 0·76]	5·37x10^{-7}
rs3794405	13	29006847	Intron 6	T/C	0·243	0·71	0·54 [0·41, 0·73]	4·16x10^{-5}	0·70 [0·56, 0·87]	1·29x10^{-3}	0·64 [0·53, 0·76]	4·56x10^{-7}
rs9505994	13	29024346	Intron 3	G/A	0·239	0·71	0·53 [0·40, 0·71]	2·41x10^{-5}	0·69 [0·55, 0·86]	1·10x10^{-3}	0·63 [0·53, 0·75]	2·53x10^{-7}
rs34961350	13	28991902	Intron 10	G/C	0·220	0·57	1·89 [1·41, 2·54]	2·38x10^{-5}	1·27 [1·00, 1·60]	5·02x10^{-2}	1·53 [1·03, 2·27]	2·55x10^{-5}

[*] Cases:Controls; [†] According to the principal *FLT1* transcript (ENST00000282397.8) ; A1, Effect allele; A2, Non-effect allele; CI, Confidence Interval; MAF, Minor Allele Frequency (GEN-SEP cohort) ; OR, Odd Ratio; r^2, squared coefficient of correlation (with respect to rs9508032). Alleles referred to the positive strand of hg19. The top hit is indicated in bold.

Supplementary Table 5. Prediction of ICU survival for the sentinel variant at *FLT1* in GEN-SEP

	OR [95% CI]	***p*-value**
Sepsis	1·00 [0·81, 1·24]	0·974
ARDS	1·16 [0·87, 1·54]	0·307

Data were obtained using Cox regression models adjusted for age, sex, and APACHE II scores. APACHE II, Acute Physiology and Chronic Health Evaluation II; ICU, intensive care unit; OR, Odd Ratio; CI, Confidence Interval.

Supplementary Table 6. Expression of *FLT1* and *VEGFA* isoforms in lung tissue.

Name	Average	SD
FLT1-201	5,928	2,803
FLT1-202	5	5
FLT1-203	6	4
FLT1-204	20	12
FLT1-205	6	7
FLT1-206	12	12
FLT1-207	3,998	2,492
FLT1-208	2	2
VEGFA-203	0	0
VEGFA-226	11	28
VEGFA-209	0	0
VEGFA-207	12	14
VEGFA-205	3,804	3,543
VEGFA-213	0	0
VEGFA-228	5	12
VEGFA-229	2,748	2,824
VEGFA-208	0	0
VEGFA-204	0	0
VEGFA-227	1,804	3,203
VEGFA-220	5	7
VEGFA-202	0	0
VEGFA-224	0	0
VEGFA-222	1,536	1,704
VEGFA-218	94	81
VEGFA-225	9	27
VEGFA-219	0	1
VEGFA-223	1	4
VEGFA-201	69	75
VEGFA-206	1,232	1,325
VEGFA-210	1	4
VEGFA-221	1	1
VEGFA-212	3,393	2,676
VEGFA-215	4,193	3,631
VEGFA-211	52	41
VEGFA-214	83	105
VEGFA-217	8	15
VEGFA-216	161	158

No data was available for the isoforms FLT1-209, VEGFA-230, VEGFA-231, or VEGFA-232. SD, standard deviation. Values are given in counts per million.

Supplementary Table 7. Functional annotation of the *FLT1* top hit (rs9508032) and the most promising two proxies.

	rs9508032	**rs722503**	**rs8002446**
Functional significance score predicted with DeepSEA	0·13	<0·05	0·10
regulomedB Score	(5) TF binding or Dnase peak	(3a) TF binding + any motif + Dnase peak	(4) TF binding + Dnase peak
Enhancer histone marks	H3K4me1[*], H3K27ac[†]	H3K4me1[‡], H3K27ac[§]	H3K4me1[¶], H3K27ac[ǀǀ]
Promoter histone marks	H3K4me3[**], H3K9ac[††]	H3K4me3[‡‡], H3K9ac[§§]	H3K4me3[¶¶], H3K9ac[§§]
DNAse	HSC & B-cell, Monocytes-CD14+ RO01746 Primary Cells	HSC & B-cell, ES-deriv	IMR90, iPSC, Blood & T-cell, HSC & B-cell, Epithelial, Thymus, Muscle, Fetal Kidney, Fetal Lung, Ovary, Placenta, GM12878 Lymphoblastoid Cells, Monocytes-CD14+ RO01746 Primary Cells
Altered regulatory motifs	Cdc5, Gfi-1, HNF1, Mef2	CCNT2, MAZR, NF-kappaB, Spz1	None
Proteins bound	None	POL2, NFKB	BCL11A, EBF1, EBF1, ELF1, PAX5C20, PAX5N19, PU1, SP1, PU1, POL2, CMYC, MAX
CHICP	<u>CD34</u>: *POMP* (11·36); <u>GM12878</u>: *POMP* (9·71), *FLT1* (10·97), *SLC46A3/RNU6-53P* (9·91), *PAN3* (9·13), *SLC46A3/CYP51A1P2* (8·63)		
Open Targets Genetics	None	None	*FLT3* (top ranked), *POMP*, *PAN3*
eQTLs	None	None	None
Score CAPE dsQTL >0·5	HUVEC, A549 EtOH 0.02pct Lung Carcinoma Cell Line	None	None
Score CAPE eQTL >0·5	None	HUVEC, NHLF Lung Fibroflast Primary Cells, NHDF-Adult Dermal Fibroflast Primary Cells, Monocytes-CD14+ RO01746 Primary Cells, A549 EtOH 0·02pct Lung Carcinoma Cell Line, Foreskin Fibroblast Primary Cells skin01, IMR90 fetal lung fibroblasts Cell Line	A549 EtOH 0·02pct Lung Carcinoma Cell Line

CAPE, Cellular dependent deactivating mutations; CD34, human hematopoietic progenitor cell line; GM12878, lymphoblastoid cell line; HUVEC, Human umbilical vein endothelial cell; IMR90, Human foetal lung cells; NHDF, Normal Human Dermal Fibroblasts; NHLF, Normal human lung fibroblasts. [*]IMR90, ESC, iPSC, ES-deriv, Blood & T-cell, HSC & B-cell, Epithelial, Brain, Adipose, Muscle, Heart, Fetal Lung, Fetal Adrenal Gland, Liver, Spleen, GM12878 Lymphoblastoid Cells, HUVEC Umbilical Vein Endothelial Primary Cells, Monocytes-CD14+ RO01746 Primary Cells. [†]iPSC, ES-deriv, HSC & B-cell, Epithelial, Adipose, Spleen, GM12878 Lymphoblastoid Cells, HUVEC Umbilical Vein Endothelial Primary Cells. [‡]IMR90, ESC, iPSC, ES-deriv, Blood & T-cell, HSC & B-cell, Epithelial, Thymus, Brain, Adipose, Muscle, Heart, Digestive, Fetal Lung, Fetal Adrenal Gland, Placenta, Liver, Lung, Spleen, Dnd41 TCell Leukemia Cell Line, GM12878 Lymphoblastoid Cells, HUVEC Umbilical Vein Endothelial Primary Cells, K562 Leukemia Cells, Monocytes-CD14+ RO01746 Primary Cells. [§]iPSC, HSC & B-cell, Brain, Adipose, Heart, Digestive, Liver, Dnd41 TCell Leukemia Cell Line, GM12878 Lymphoblastoid Cells, HUVEC Umbilical Vein Endothelial Primary Cells, Monocytes-CD14+ RO01746 Primary Cells. [¶]ESC, iPSC, ES-deriv, Blood & T-cell, HSC & B-cell, Epithelial, Thymus, Brain, Adipose, Muscle, Heart, Digestive, Fetal Lung, Fetal Adrenal Gland, Placenta, Liver, Lung, Spleen, Dnd41 TCell Leukemia Cell Line, GM12878 Lymphoblastoid Cells, HUVEC Umbilical Vein Endothelial Primary Cells, Monocytes-CD14+ RO01746 Primary Cells. [ǀǀ]iPSC, HSC & B-cell, Epithelial, Brain, Adipose, Heart, Digestive, Ovary, Liver, Dnd41 TCell Leukemia Cell Line, GM12878 Lymphoblastoid Cells, HUVEC Umbilical Vein Endothelial Primary Cells, Monocytes-CD14+ RO01746 Primary Cells. [**]HSC & B-cell, Monocytes-CD14+ RO01746 Primary Cells. [††]iPSC, Adipose, Digestive. [‡‡]Blood & T-cell, HSC & B-cell, Brain, Adipose, Heart, Digestive, Liver, Dnd41 TCell Leukemia Cell Line, GM12878 Lymphoblastoid Cells, Monocytes-CD14+ RO01746 Primary Cells. [§§]Blood & T-cell, Adipose, Sm. Muscle, Dnd41 TCell Leukemia Cell Line, GM12878 Lymphoblastoid Cells, Monocytes-CD14+ RO01746 Primary Cells. [¶¶]Blood & T-cell, HSC & B-cell, Brain, Dnd41 TCell Leukemia Cell Line, GM12878 Lymphoblastoid Cells, Monocytes-CD14+ RO01746 Primary Cells

Supplementary Table 8. Signals of replication at gene level in the GEN-SEP dataset within 100 kb of previously reported candidate genes.

Gene	Independent signals	Gene-wise Bonferroni *p*-value threshold	SNP min *p*-value	A1/A2	OR [95% CI]	*p*-value
ABCC1	275·18	1·82x10^{-4}	rs246233	G/T	1·98 [1·21, 3·25]	6·79x10^{-3}
ACE	72·94	6·85x10^{-4}	rs9857615	T/C	0·70 [0·49, 0·98]	3·89x10^{-2}
ADA	137·74	3·63x10^{-4}	rs17687734	G/A	2·60 [1·31, 5·16]	6·31x10^{-3}
ADGRV1	640·97	7·80x10^{-5}	rs6094023	A/G	0·48 [0·29, 0·79]	3·65x10^{-3}
ADIPOQ	288·87	1·73x10^{-4}	rs114210898	A/G	1·34 [1·05, 1·72]	2·06x10^{-2}
ADRBK2	318·26	1·57x10^{-4}	rs1467387	C/T	0·67 [0·50, 0·89]	5·96x10^{-3}
*AGER**	363·32	1·38x10^{-4}	rs61746206	T/C	0·32 [0·13, 0·83]	1·83x10^{-2}
AGT	303·17	1·65x10^{-4}	rs1078499	G/A	0·69 [0·53, 0·91]	8·68x10^{-3}
AGTR1	281·4	1·78x10^{-4}	rs275643	A/G	1·86 [1·24, 2·80]	2·70x10^{-3}
AHR	201·39	2·48x10^{-4}	rs140084506	T/C	4·69 [1·26, 17·4]	2·10x10^{-2}
ANGPT2	456·67	1·09x10^{-4}	rs2442570	A/G	0·43 [0·24, 0·76]	3·76x10^{-3}
APOA1	236·49	2·11x10^{-4}	rs2513094	C/T	1·36 [1·03, 1·79]	2·99x10^{-2}
ARSD	159·08	3·14x10^{-4}	rs1698814	C/T	1·45 [1·08, 1·96]	1·46x10^{-2}
BCL11A	262·4	1·91x10^{-4}	rs76064527	A/C	1·90 [1·08, 3·37]	2·69x10^{-2}
CBS	160·35	3·12x10^{-4}	rs2401154	T/C	1·49 [1·06, 2·08]	2·03x10^{-2}
CELF2	669·82	7·46x10^{-5}	rs76209150	T/G	2·65 [1·33, 5·26]	5·50x10^{-3}
CHIT1	269·23	1·86x10^{-4}	rs1845466	T/G	0·64 [0·49, 0·83]	8·63x10^{-4}
CLASRP	113·15	4·42x10^{-4}	rs10405859	C/T	0·71 [0·56, 0·91]	6·50x10^{-3}
CXCL2	96·08	5·20x10^{-4}	rs28574621	C/G	0·30 [0·09, 0·94]	3·85x10^{-2}
CXCR2	98·44	5·08x10^{-4}	rs12989315	A/G	2·26 [1·26, 4·07]	6·36x10^{-3}
CYP1A1	82·88	6·03x10^{-4}	rs17861120	G/A	0·51 [0·32, 0·83]	6·20x10^{-3}
DARC	202·96	2·46x10^{-4}	rs55833893	T/C	0·31 [0·14, 0·68]	3·66x10^{-3}
DIO2	158·36	3·16x10^{-4}	rs17176215	G/A	0·24 [0·07, 0·87]	2·96x10^{-2}
EGF	240·96	2·08x10^{-4}	rs146141236	C/T	0·22 [0·09, 0·55]	1·25x10^{-3}
*EGLN1**	144·18	3·47x10^{-4}	rs141921538	T/C	4·56 [1·44, 14·39]	9·75x10^{-3}
EPAS1	417·3	1·20x10^{-4}	rs11888926	G/C	1·65 [1·19, 2·28]	2·63x10^{-3}
F5	269·74	1·85x10^{-4}	rs144628673	A/G	5·37 [1·51, 19·05]	9·37x10^{-3}
FAAH	187·13	2·67x10^{-4}	rs78918625	G/C	0·17 [0·04, 0·78]	2·30x10^{-2}
FAS	239·16	2·09x10^{-4}	rs61852572	G/A	0·60 [0·41, 0·89]	1·11x10^{-2}
*FER**	644·44	7·76x10^{-5}	rs10515395	C/T	1·82 [1·22, 2·71]	3·28x10^{-3}
FTL	143·23	3·49x10^{-4}	rs140747916	T/A	1·70 [1·08, 2·66]	2·12x10^{-2}
FZD2	125·26	3·99x10^{-4}	rs9900767	T/C	0·50 [0·28, 0·90]	2·09x10^{-2}
GADD45A	184·53	2·71x10^{-4}	rs344923	G/A	1·33 [1·05, 1·70]	1·92x10^{-2}
GHR	322·68	1·55x10^{-4}	rs41271073	A/G	0·34 [0·15, 0·74]	6·48x10^{-3}
GP5	210·56	2·37x10^{-4}	rs7611390	T/C	0·57 [0·41, 0·81]	1·63x10^{-3}
GRM3	285·81	1·75x10^{-4}	rs6974073	A/C	1·59 [1·02, 2·47]	4·01x10^{-2}
HAS1	225·37	2·22x10^{-4}	rs113174648	G/C	1·51 [1·12, 2·03]	7·03x10^{-3}
HECTD2	174·41	2·87x10^{-4}	rs11186608	T/G	0·71 [0·56, 0·89]	3·23x10^{-3}
HMOX1	129·66	3·86x10^{-4}	rs4645773	T/C	0·35 [0·16, 0·75]	6·49x10^{-3}
HMOX2	119·94	4·17x10^{-4}	rs190300249	T/C	0·34 [0·12, 0·93]	3·51x10^{-2}
HSPG2	351·59	1·42x10^{-4}	rs72662414	C/A	3·05 [1·03, 9·07]	4·50x10^{-2}
HTR2A	286·1	1·75x10^{-4}	rs1923886	T/C	1·54 [1·21, 1·96]	5·36x10^{-4}
IL10	185·14	2·70x10^{-4}	rs79474100	A/T	0·27 [0·10, 0·73]	9·99x10^{-3}
IL13	86·87	5·76x10^{-4}	rs60153262	T/C	2·94 [1·50, 5·77]	1·74x10^{-3}
IL18	143·25	3·49x10^{-4}	rs360723	T/A	0·69 [0·49, 0·96]	2·87x10^{-2}
IL1RN	281·1	1·78x10^{-4}	rs6746416	G/A	1·37 [1·07, 1·76]	1·28x10^{-2}
IL32	70·73	7·07x10^{-4}	rs12598558	G/T	0·58 [0·38, 0·89]	1·18x10^{-2}
IL4	91·11	5·49x10^{-4}	rs60153262	T/C	2·94 [1·50, 5·77]	1·74x10^{-3}

IL6	236·93	2·11x10^{-4}	rs75897827	A/G	2·70 [1·18, 6·21]	1·91x10^{-2}
IL8	154·71	3·23x10^{-4}	rs7686667	G/A	2·43 [1·06, 5·61]	3·69x10^{-2}
IRAK3	186·89	2·68x10^{-4}	rs569436368	A/G	0·26 [0·09, 0·81]	2·05x10^{-2}
ISG15	8·12	6·16x10^{-3}	rs12093451	A/C	0·50 [0·29, 0·88]	1·55x10^{-2}
KLK2	201·25	2·48x10^{-4}	rs1701934	T/C	5·03 [1·65, 15·4]	4·61x10^{-3}
LRRC16A	781·8	6·40x10^{-5}	rs2690123	G/A	1·41 [1·11, 1·8]	5·64x10^{-3}
LTA	459·6	1·09x10^{-4}	rs45552734	T/C	0·58 [0·41, 0·82]	2·40x10^{-3}
*MAP3K1**	262·39	1·91x10^{-4}	rs1910019	T/C	1·77 [1·25, 2·50]	1·30x10^{-3}
MAP3K6	86·03	5·81x10^{-4}	rs12742921	C/T	0·72 [0·52, 0·99]	4·53x10^{-2}
MBL2	250·05	2·00x10^{-4}	rs34546527	C/A	0·44 [0·28, 0·68]	2·50x10^{-4}
MIF	141·4	3·54x10^{-4}	rs75761219	T/C	0·22 [0·06, 0·76]	1·72x10^{-2}
*MUC5B**	71·73	6·97x10^{-4}	rs2071175	T/C	2·10 [1·18, 3·74]	1·16x10^{-2}
MYLK	307·19	1·63x10^{-4}	rs16834826	G/A	1·48 [1·15, 1·91]	2·45x10^{-3}
NAMPT	238·31	2·10x10^{-4}	rs56844330	G/A	0·64 [0·47, 0·87]	5·02x10^{-3}
NFE2L2	196·18	2·55x10^{-4}	rs2588866	T/C	0·51 [0·30, 0·87]	1·25x10^{-2}
NFKB1	210·58	2·37x10^{-4}	rs76615823	G/A	3·21 [1·13, 9·14]	2·88x10^{-2}
NFKBIA	254·35	1·97x10^{-4}	rs75208350	T/C	2·44 [1·15, 5·19]	2·01x10^{-2}
NOS3	195·41	2·56x10^{-4}	rs41307316	A/G	0·22 [0·10, 0·51]	4·03x10^{-4}
NQO1	143·23	3·49x10^{-4}	rs116423606	A/G	0·30 [0·09, 0·95]	3·98x10^{-2}
PDE4B	530·28	9·43x10^{-5}	rs6664875	C/A	0·64 [0·48, 0·85]	1·96x10^{-3}
PI3	142·59	3·51x10^{-4}	rs877608	A/T	1·98 [1·08, 3·63]	2·71x10^{-2}
PLAU	116·57	4·29x10^{-4}	rs72816344	A/G	3·50 [1·08, 11·36]	3·73x10^{-2}
POPDC3	181·91	2·75x10^{-4}	rs1051484	T/C	0·62 [0·46, 0·82]	1·12x10^{-3}
PPARGC1A	978·39	5·11x10^{-5}	rs6847465	T/C	3·92 [1·51, 10·15]	4·96x10^{-3}
PPFIA1-SHANK2	881·08	5·67x10^{-5}	rs11602848	C/T	1·71 [1·20, 2·43]	3·01x10^{-3}
PRKAG2	648·11	7·71x10^{-5}	rs10231047	C/T	0·71 [0·56, 0·89]	2·91x10^{-3}
S1PR3	243·08	2·06x10^{-4}	rs150901384	G/T	4·57 [1·44, 14·5]	9·79x10^{-3}
SELPLG	176·88	2·83x10^{-4}	rs8179106	A/G	1·77 [1·17, 2·66]	6·44x10^{-3}
SERPINE1	207·41	2·41x10^{-4}	rs73168394	A/G	0·33 [0·17, 0·67]	2·22x10^{-3}
SFTPA1	143·17	3·49x10^{-4}	rs17886197	G/T	1·60 [1·14, 2·26]	7·27x10^{-3}
SFTPA2	166·37	3·01x10^{-4}	rs17886197	G/T	1·60 [1·14, 2·26]	7·27x10^{-3}
SFTPB	165·98	3·01x10^{-4}	rs75830997	T/G	3·64 [1·51, 8·73]	3·86x10^{-3}
SFTPD	222·05	2·25x10^{-4}	rs7082484	C/A	2·27 [1·22, 4·24]	9·55x10^{-3}
SOD3	175·95	2·84x10^{-4}	rs2361079	C/T	0·59 [0·40, 0·85]	4·55x10^{-3}
STAT1	161·06	3·10x10^{-4}	rs4853453	A/G	1·58 [1·12, 2·24]	9·30x10^{-3}
TGFB2	244·72	2·04x10^{-4}	rs75854892	C/T	5·78 [1·60, 20·9]	7·43x10^{-3}
TIA1	104·96	4·76x10^{-4}	rs11694045	G/T	1·42 [1·11, 1·82]	5·48x10^{-3}
TIRAP	187·4	2·67x10^{-4}	rs12283024	A/G	0·34 [0·15, 0·78]	1·08x10^{-2}
TLR1	227·88	2·19x10^{-4}	rs193202734	C/T	5·63 [1·86, 17·01]	2·20x10^{-3}
TNF	448·83	1·11x10^{-4}	rs45552734	T/C	0·58 [0·41, 0·82]	2·40x10^{-3}
TNFRSF11A	246·87	2·03x10^{-4}	rs7235828	A/G	0·65 [0·47, 0·89]	7·73x10^{-3}
TRAF6	185·4	2·70x10^{-4}	rs2458928	A/G	0·68 [0·52, 0·89]	5·39x10^{-3}
UGT2B7	186·83	2·68x10^{-4}	rs139914109	C/T	7·69 [1·65, 35·87]	9·42x10^{-3}
VEGFA	**262·98**	**1·90x10^{-4}**	**rs9367172**	**A/G**	**0·55 [0·41, 0·73]**	**4·69x10^{-5}**
VLDLR	375·76	1·33x10^{-4}	rs10491716	C/A	1·55 [1·19, 2·02]	1·20x10^{-3}
VWF	403·43	1·24x10^{-4}	rs2239160	G/A	0·42 [0·26, 0·68]	3·28x10^{-4}
XKR3	97·09	5·15x10^{-4}	rs5994042	A/T	3·23 [1·20, 8·71]	2·03x10^{-2}
ZNF335	141·92	3·52x10^{-4}	rs1736493	G/A	0·53 [0·31, 0·89]	1·72x10^{-2}

A1, Effect allele; A2, Non-effect allele; CI, Confidence Interval; OR, Odds ratio for the effect alleles. In bold, genes harboring variants reaching the bonferroni threshold. *Genes identified for this study (December 2015 to November 2018) based on the search of terms “acute respiratory distress syndrome”, “ARDS” OR “acute lung injury” with “polymorphism” OR “genetic variant”.

Chapter 3.

Early bacterial dysbiosis in the lungs predicts ICU mortality in non-pulmonary sepsis patients

Sepsis is the main factor leading to ARDS and an important cause of mortality in ICU. However, there is a lack of efficient prognostic methods for sepsis patients. In this chapter, we assessed the lung microbiome of patients with non-pulmonary sepsis by means of the analysis of lung aspirate samples collected at three different time points (at 8 h of sepsis diagnosis, and after 48 h and 72 h). We used NGS technologies to sequence the hypervariable region V4 of the 16S rRNA gene, and bioinformatics and statistical tools to determine the bacterial abundance and to perform diversity analyses.

In this single-center study, we observed that bacterial diversity in lung aspirates was significantly linked to patient mortality within 8 h of sepsis diagnosis, being much lower in deceased patients and presenting a predictive value (area under the curve) of 86.5% in our data, higher than the value obtained considering the Acute Physiology and Chronic Health Evaluation II (APACHE II) score. Additionally, we observed that lung aspirates from deceased patients presented commensal gut bacteria genera and were depleted in healthy lung bacteria genera. Therefore, these results suggest the potential of using the microbial diversity as an early prognostic biomarker in patients with sepsis, as well as the utility of NGS techniques in clinical practice as a complement to culture-dependent methods.

This study has been published in *Intensive Care Medicine* with the title "Could lung bacterial dysbiosis predict ICU mortality in patients with extra-pulmonary sepsis? A proof-of-concept study" (doi: 10.1007/s00134-020-06190-4). It is reproduced under the Springer Nature terms of use.

Early bacterial dysbiosis in the lungs predicts ICU mortality in non-pulmonary sepsis patients

Beatriz Guillen-Guio,[1] Tamara Hernandez-Beeftink,[1,2] David Domínguez,[3] Adrian Baez-Ortega,[4] Almudena Corrales,[1,5] Raúl Hernández-Bisshopp,[3] Jorge Arias,[3] Luis Soto,[3] David Viera Camacho,[3] Gabriela Noemí González,[3] Javier Belda,[6] Elena Espinosa,[3] Julia Alcoba-Florez,[7] Rafaela González-Montelongo,[8] Jesús Villar,[2,5] Carlos Flores,[1,5,8,9*]

[1]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [2]Research Unit, Hospital Universitario Dr. Negrin, Las Palmas de Gran Canaria, Spain; [3]Department of Anesthesiology, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain; [4]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK; [5]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain; [6]Department of Anesthesiology, Hospital Clínico Universitario de Valencia, Valencia, Spain; [7]Department of Microbiology, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain; [8]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain; [9]Instituto de Tecnologías Biomédicas (ITB), Universidad de La Laguna, Santa Cruz de Tenerife, Spain.

*Corresponding author: E-mail: cflores@ull.edu.es

Abstract

Introduction: Sepsis is an important cause of mortality in adult intensive care units (ICUs). The lack of efficient prognostic methods for patients with sepsis makes clear the necessity of identifying novel early biomarkers. The lung microbiome has a key role in the maintenance of lung immune homeostasis, although its link with sepsis outcomes remains unknown. Here we hypothesized that lung bacterial dysbiosis is associated with ICU mortality in patients with sepsis.

Materials and methods: A total of 36 patients with non-pulmonary sepsis admitted into a single medical-surgical ICU were included in the study. We analyzed 69 tracheal aspirates collected from these patients at sepsis diagnosis (within 8 h), and after 24 h and 72 h. Bacterial abundance was obtained by DNA sequencing of the V4 hypervariable region of the 16S rRNA gene. Sequence data preprocessing, taxonomic assignment, Shannon diversity estimates, and statistical comparisons were performed with QIIME and R programming.

Results: Bacterial diversity did not vary significantly between the three ICU collection times. However, diversities were extremely different very early between deceased and surviving patients (*p*=0.001). Among bacteria detected in deceased patients, we found gut commensals and a depletion of bacterial genera commonly found in healthy lungs. The predictive value of ICU mortality by the diversity index was 86.5% in our data, greater than the obtained by the APACHE II score at inclusion.

Conclusions: The reduction of bacterial lung diversity in patients with non-pulmonary sepsis was found to be associated with ICU mortality within 8 h of diagnosis, supporting its role as a potential novel early prognostic biomarker.

Introduction

Sepsis is a complex disorder currently defined as a life-threatening organ dysfunction that results from a systemic inflammatory response due to infections (Singer et al. 2016). Sepsis is one of the most common causes of death in adult patients admitted into intensive care units (ICUs) (Angus et al. 2001), and it is the major risk factor of acute respiratory distress syndrome (ARDS) development, a fatal condition with poor prognosis (Bellani et al. 2016). Several biomarkers alone or combined with clinical symptoms have been related to the mortality of sepsis patients (Gibot et al. 2012; Larsen and Petersen 2017). However, despite all the efforts to establish clinically-relevant prognostic methods for sepsis, these have showed a limited power to predict patient severity (Gibot et al. 2012). Therefore, there is an urgent necessity to identify earlier and more accurate biomarkers of sepsis prognosis.

Several studies have reported the important role of the microbiome in complex diseases and immunity (Thaiss et al., 2016). Particularly, shifts in the microbial diversity (commonly known as dysbiosis) have been recently linked to the natural disease development and to interventions in the critically ill patient (Dickson 2016; McDonald et al. 2016; Jacobs et al. 2017; Lamarche et al. 2018). Given the infectious nature of sepsis, these observations may be clinically relevant in the pathogenesis or the aggravation of this critical condition. In fact, injurious ventilation regimes have been associated with an increase in the spreading of infections in animal models of non-pulmonary sepsis (Rodríguez-González et al. 2014). This can be reconciled with the observations of MacFie and colleagues (MacFie et al. 1999), who reported the association of bacterial translocation and septic morbidity.

Microbial dysbiosis in the lung, blood and the upper airways has been correlated with sepsis development and severity (Dickson 2016; Gosiewski et al. 2017; Tan et al. 2019). Interestingly, an enrichment of the gut-associated bacteria in the lung microbiome has been found in patients with sepsis and ARDS, possibly explained by the systemic translocation of the intestinal microbes in these patients (Dickson et al. 2016; Mukherjee and Hanidziar 2018). Changes in the gut microbiome have been linked to sepsis morbidity and mortality (Haak and Wiersinga 2017), as well as to outcomes in patients with systemic inflammatory response syndrome (SIRS) (Shimizu et al. 2011). Lamarche and colleagues also reported a reduced microbial diversity at three anatomical sites, including the trachea, associated with severity among a heterogeneous ICU patient population (Lamarche et al. 2018). Therefore, although it is known that the lung microbiome is severely altered in critically ill patients, a specific association of lung dysbiosis with sepsis mortality remains to be determined.

To test this possibility while avoiding the potential confounder effects in lung dysbiosis due to pneumonia infection, which generally leads to bacterial overgrowth of single bacterial species in the respiratory tract (de Steenhuijsen Piters et al. 2016), we focused on intubated non-pulmonary sepsis patients from a single medical-surgical ICU. To gain insight into the bacterial shifts and increase the power to detect the association, we collected lung aspirates at three times over 72 h while the patient was intubated. Our sequencing analyses based on the 16S ribosomal RNA (16 rRNA) gene strongly supported the existence of an early reduction in bacterial lung diversity associated with ICU mortality.

Methods

Sample description

The study was conducted on 36 mechanically-ventilated adult patients of European ancestry diagnosed with non-pulmonary sepsis (Singer et al. 2016) that were admitted to a medical-surgical ICU at University Hospital Nuestra Señora de Candelaria (Santa Cruz de Tenerife, Spain) between January 2015 and January 2019. Tracheal aspirate samples were collected from these patients at three different times whenever possible while the patients remained intubated: within the first 8 h of sepsis diagnosis, at 48 h after sepsis onset, and after 72 h. A total of 69 aspirates were finally obtained from the patients and stored at -20 ºC. Bacterial DNA was extracted from the aspirates using the QIAamp® UCP Pathogen Mini Kit (Qiagen), quantified with a Qubit 3.0 fluorometer using a High Sensitivity DNA Analysis Kit (Thermo Fisher Scientific), and stored at -20 ºC until use.

The study was approved by the Research Ethics Committee of the hospital and performed according to The Code of Ethics of the World Medical Association (Declaration of Helsinki). An informed consent was obtained from all patients or from their representatives.

Amplification and sequencing of V4 16S rRNA

The V4 hypervariable region of the 16 rRNA gene was amplified by polymerase chain reaction (PCR) in 20-µl reactions. We used a HotStarTaq DNA Polymerase (Qiagen) along with fusion primers including 12 bp Golay barcodes and the Illumina adaptor sequences (Caporaso et al. 2012). Purification and size-selection of PCR products was performed with the AxyPrep™ Mag FragmentSelect-I purification kit (Axygen), using a 1.4 ratio of magnetic beads/PCR product. Purified PCR products were normalized with the SequalPrep™ Normalization Plate (96) Kit (Thermo Fisher Scientific) and pooled to 96-plex at 25 ng per sample. In addition to the lung aspirates, libraries from a mock community (ZymoBIOMICS™ Microbial Community DNA Standard, Zymo Research) and from PCR-grade water were also included in each of the pools to serve as positive and negative controls, respectively. The pooled libraries were quantified by a 7500 Fast Real-Time PCR System (Life Technologies) using the KAPA Library

Quantification Kit Illumina® platforms (KapaBiosystems) and by the Qubit High Sensitivity DNA Analysis Kit. The size distribution of amplicons was evaluated with the Agilent 4200 TapeStation system using the High Sensitivity D1000 ScreenTape Assay kit (Agilent Technologies Inc.). Libraries were loaded at 12 pM and sequenced using the MiSeq Reagent V2 kit (300 cycles paired-end) in a MiSeq sequencer (Illumina Inc.), including a 10% of PhiX library as a sequencing control. Sequencing experiments were performed at Instituto Tecnológico y de Energías Renovables (Santa Cruz de Tenerife, Spain).

Bioinformatics and statistical analysis

Analyses of the 16S rRNA sequencing data were conducted with the Quantitative Insights Into Microbial Ecology (QIIME) v1.9 package (Caporaso et al. 2010), by means of a custom BASH pipeline. After joining forward and reverse read pairs, considering a minimum overlap of 47 bp, raw data were demultiplexed by barcode and low-quality reads (Phred quality score<30) were filtered. Operational taxonomic unit (OTU) clustering was performed with an open-reference approach using UCLUST (Edgar 2010) and chimeric sequences were detected and removed using ChimeraSlayer (Haas et al. 2011). The taxonomic assignment with a 97% sequence identity was based on the Greengenes database (DeSantis et al. 2006). Alignment was conducted with PyNAST (Caporaso et al. 2010), and a phylogenetic tree was then built using FastTree (Price et al. 2010). Human mitochondrial OTUs and the OTUs with at least 10 reads in the negative controls, were removed from downstream analyses. Diversity analyses were also conducted with QIIME and focused on the core lung microbiome, defined by the OTUs that were present in >50% of the patients. Although the analyses were focused on the Shannon diversity index, a high correlation between it and other diversity indices was found in our results (**Supplementary Table 1**). In order to normalize across patients and time points, analyses were conducted using a random sampling of 1,000 reads per library. The OTU abundance was determined and taxa abundance plots were generated. Alpha diversity metrics were computed and subsequently compared between survivors and deceased groups based on a nonparametric two-sample t-test using Monte Carlo permutations. To compare diversities of more than three groups, Kruskal-Wallis tests were performed with R version 3.3.2 (R Core Team 2013). Additional logistic regressions were performed in R to evaluate the effect of different clinical and demographic covariates in the model. The online software Calypso (Zakrzewski et al. 2017) was used to compare the alpha diversity scores between deceased and survivors and to identify significant differences in the relative abundances of individual taxa by means of the Linear Discriminant Analysis Effect Size (LEfSe) algorithm. Beta diversity was also assessed by means of weighted and unweighted UniFrac distance matrices with QIIME, in order to generate Principal Coordinates Analysis (PCoA) plots. An additional principal component (PC) analysis (PCA) was performed from relative abundances using the R *FactoMineR* package (Lê et al. 2008). A PCA biplot showing both PC scores of samples and of the loadings from bacterial taxa was generated with the *factoextra* R package. The area under the ROC curve (AUC) was calculated using the R *pROC* package (Robin et al. 2011) to evaluate the predictive value of the bacterial lung diversity index at 8 h of sepsis diagnosis, compared to that of the Acute Physiology and Chronic Health Evaluation II (APACHE II) score at inclusion, to discriminate between the sepsis patients surviving or dying in the ICU. Finally, based on the model of Cumsill and colleagues (Cumsille et al. 2000), and assuming a probability threshold of 0.5 in a binary logistic regression, we estimated a cutoff point for the bacterial diversity index in the lung aspirates in order to predict ICU mortality.

Statistical power

Based on internal observations of the dissimilarity in the microbiome profiles from lung aspirates from patients with respiratory infections vs. controls (associated with modified Cramer's φ >0.20), and assuming a minimum of 10,000 reads per sample and a significance level of 0.01, a thousand simulations performed following the procedures described elsewhere (La Rosa et al. 2012) indicated that a minimum of 5 patients per comparison group provides a statistical power >80% to detect differences in the microbiome profiles.

Results

Study sample and sequencing performance

A total of 69 lung aspirates from 36 individuals with non-pulmonary sepsis were finally included in the study. Main demographic and clinical features of these patients are provided in **Table 1**, suggesting that survivors and deceased patients were significantly discordant only for the age and the APACHE II score at inclusion. Sequencing and filtering the V4 16S rRNA data left us with a total of 7,646,140 high-quality paired-end reads for further analyses (110,813 on average per sample). The presence of bacterial DNA was evidenced in all samples. The core lung microbiome of sepsis patients was composed by 54 OTUs with a frequency higher than 0.1%, 38 OTUs when only the samples collected within 8 h of diagnosis were considered.

Longitudinal analysis in the ICU

We first focused in the changes in bacterial abundance over collection times. Because of the dropout of patients from the study because of ICU discharge, extubation or death, we analyzed 36 lung aspirates collected within the first 8 h of sepsis diagnosis, 17 aspirates collected at 48 h after sepsis onset, and 16 aspirates collected at 72 h after sepsis onset. A PCoA based on weighted UniFrac distances showed that there was a lack of a clear clustering of the lung aspirates by the day of collection

(**Supplementary Figure 1**). Supporting this observation, the bacterial diversity did not vary significantly among the three sample collection times (*p*=0.99) (**Table 2**). To maximize the power of the study and based on these results and on the clinical relevance of identifying prognostic biomarkers of sepsis in the earliest possible disease stage, subsequent analyses focused exclusively on the lung aspirates within 8 h of sepsis diagnosis.

Bacterial dysbiosis and ICU mortality

Diversity and abundance estimates in the bacterial lung communities within 8 h of sepsis diagnosis were compared between the patients that were discharged alive from the ICU and those who died during the ICU stay. We found that bacterial diversity decreased significantly in deceased compared to surviving patients (Shannon diversity index, *p*=0.001) (**Figure 1**). Results were similar if mortality was considered at 28, 60, or 90-day. Furthermore, a sensitivity analysis was performed to evaluate the effect of possible confounding variables in the association of bacterial diversity with ICU mortality. We observed that the association was robust to adjustments for age, APACHE II score, ARDS incidence, comorbidities, isolated pathogen and infection source, among others (**Table 3**).

A total of 13 and 36 OTUs were found in deceased patients and survivors, respectively. The most abundant taxa comprised 95.1% of the core lung microbiome among deceased patients and included the genera *Achromobacter, Enterococcus*, *Proteus, Pseudomonas*, and *Staphylococcus*, as well as the family *Enterobacteriaceae* (classified from reads than did not enable the classification at genus level) **(Figure 2)**. Accordingly, the PCA with the loadings showing how strongly each taxon contributed to each PC also showed a strong relationship between these bacteria and ICU mortality (**Figure 3**). Remarkably, these taxa represented only 13.7% of the core bacterial lung microbiome among survivors. The most abundant taxa in the lungs of the survivors were *Acinetobacter*, *Haemophilus*, and *Streptococcus* (58.4% of the total), which were barely detected among the lung aspirates from deceased patients (0.64%) (**Figure 2**). Comparatively, the LEfSe analysis prioritized some genera as the most likely observations to explain differences between deceased and survivors (**Figure 4**). We observed that the genus *Proteus* was significantly enriched among deceased patients, while *Streptococcus*, *Prevotella*, *Veillonella*, and *Leptotrichia* were significantly enriched among survivors (**Figure 4**).

Predicting ICU mortality by the bacterial diversity in lung aspirates

The ROC assessment revealed that the predictive value of the lung bacterial diversity at 8 h of diagnosis on the ICU mortality was 86.5%, even higher than that provided by the APACHE II score at inclusion (**Figure 5**). In fact, APACHE II was not significantly associated with ICU mortality in the study (*p*=0.078). Finally, we estimated a suggestive cutoff point of 0.42 for the bacterial diversity index to optimally predict ICU mortality. A visual inspection of the bacterial diversity distributions suggested that all sepsis patients with a diversity index above 2.50 survived, while those with an index ranging between 0.42 and 2.50 were difficult to classify based on the model (**Supplementary Figure 2**).

Discussion

In clinical practice, the prognostic stratification system is based on severity scores as the APACHE II at inclusion (Giamarellos-Bourboulis et al. 2012) and the Sequential Organ Failure Assessment (SOFA) (Singer et al. 2016), in combination with plasma biomarkers like the procalcitonin, the C-reactive protein or the lactate (Larsen and Petersen 2017). However, it is well known that none of these scores or the plasma biomarkers have a good predictive level of the severity of critical patients (Vincent et al. 2010). Therefore, there is a huge necessity of identifying effective biomarkers of disease prognosis. To the best of our knowledge, this is the first time that lung bacterial dysbiosis is associated with ICU mortality in patients with non-pulmonary sepsis. Based on NGS methods, we evidenced a strong reduction in the lung bacterial diversity in the patients that died during the ICU stay in comparison to those who survived. Most importantly, such lung dysbiosis was identified as early as within the 8 h of the sepsis diagnosis, supporting it as a novel early biomarker for fatal outcomes in sepsis. In fact, the APACHE II score at inclusion did not significantly predict ICU mortality in our study. On the contrary, the lung diversity index calculated within 8 h of sepsis diagnosis was able to precisely predict ICU mortality in our study (86.5%). These observations in sepsis are analogous to others supporting the association of gut dysbiosis with SIRS patient mortality and complications (Shimizu et al. 2011) or of lung dysbiosis with idiopathic pulmonary fibrosis (IPF) mortality (Molyneaux et al. 2014; O'Dwyer et al. 2019), further supporting the importance of assessing the microbiome in complex conditions for prognostic purposes.

The lung microbiome profile of deceased patients with sepsis was mainly composed of pathogenic bacteria, including the genera *Pseudomonas*, *Proteus*, *Achromobacter*, *Enterococcus*, and *Staphylococcus*, and the family *Enterobacteriaceae*. Consistently, *Staphylococcus aureus*, *Escherichia coli*, and *Pseudomonas aeruginosa* are among the most frequently isolated pathogenic bacteria species from patients with sepsis (Opal et al. 2003; Minasyan 2019). Among the five genera that were abundant among the deceased patients, only the abundance of the genus *Proteus* differed significantly by ICU mortality. *Proteus* is a member of the *Enterobacteriaceae* family consisting of Gram-negative bacilli commonly found in the normal intestinal flora (Hamilton et al. 2018). Some *Proteus* spp (mainly *P. mirabilis*) are broadly involved in urinary tract infections, although they can also invade the bloodstream and lead to bacteremia (Chen et al. 2012). *Proteus* spp have been linked also to respiratory tract infections including ventilator-associated pneumonia (Finegold and Johnson 1985; Xia et al. 2015). The fact that bacterial taxa that are generally located in the gut (such as *Proteus* and *Enterococcus*) have been found in the lungs of patients with sepsis is not surprising given that the translocation of gut commensals into the lungs has been previously described in critical care patients (Dickson et al. 2016; Mukherjee and Hanidziar 2018). Moreover, previous studies revealed that the development of ARDS, a deadly syndrome commonly caused by sepsis, is highly correlated with an enrichment of *Enterobacteriaceae* spp in the patient’s lungs (Panzer et al. 2018).

Significant differences in abundance between survivors and deceased patients were also found for the Gram-positive genus *Streptococcus* and the Gram-negative genera *Prevotella*, *Veillonella* and *Leptotrichia*, which were all present at a higher proportion in surviving patients. Interestingly, the two most abundant bacteria genera described in the low respiratory tract of healthy subjects are *Prevotella* spp and *Veillonella* spp (Dickson et al. 2017), which are also the most abundant taxa in the oral cavity, together with *Streptococcus* spp (Dickson et al. 2017). The latter is also found among the most commonly isolated bacteria in patients with sepsis (Minasyan 2019). In this sense, the reduction of *Prevotella* spp and *Veillonella* spp in the lungs accompanied by the presence of potential pathogenic bacteria has been related to asthma risk (Hilty et al. 2010; Moffatt and Cookson 2017), and a lower abundance of *Prevotella*, *Veillonella*, and *Leptotrichia* spp in the upper respiratory tract has been associated with pneumonia in the elderly (de Steenhuijsen Piters et al. 2016). In agreement with this, Park and found a reduction of *Prevotella* spp in oropharyngeal swab samples from patients with chronic obstructive pulmonary disease or with asthma (Park et al. 2014; Yadava et al. 2016). On the other hand, the abundance of *Streptococcus*, *Veillonella* and *Prevotella* spp in the bronchoalveolar lavage fluid has been associated with less airway inflammation (Zemanick et al. 2017). This is particularly relevant for *Prevotella* spp, for which a key role in the immunologic homeostasis of the airways has been suggested (Huffnagle et al. 2017). Furthermore, the presence of *Streptococcus* and *Veillonella* spp in the lungs has been associated with IPF risk (Molyneaux et al. 2014), suggesting that these bacteria could have a role in lung fibrosis. Therefore, through their links with the immunological homeostasis of the lung, the depletion of these bacterial genera in sepsis patient’s lungs could have a central role in their ICU survival.

Our study has a number of strengths and limitations. The main strength is that our findings are based on patients from a single medical-surgical ICU, allowing to limit the heterogeneous environmental exposures of the patients enrolled. Additionally, the inclusion of patients with non-pulmonary sepsis guaranteed that the observed lung dysbiosis was a consequence of the sepsis *per se* and not due to the overgrowth of single bacteria that takes place during pneumonia. Our results were robust to confounding factors, addressed through a sensitivity analysis. Furthermore, the findings were similar when 28, 60 or 90-day ICU mortality were considered instead of ICU mortality, indicating that results are robust and independent of the patient follow up records of mortality. On the other hand, the analyses were based on next-generation sequencing technologies, which allowed to detect bacteria in all samples, reducing the bias derived from conventional microbiological methods such as microbiological cultures. Among the main limitations, we acknowledge that the analysis was based on a limited sample size and the lack of an analysis of patients from an independent ICU to validate our results. Additionally, because of the utilized approach, taxa were only classified at genus level, and the rich information that may come from the detection of bacterial species and strains remains unexplored. In this sense, alternative approaches, such as the shotgun sequencing, would offer much better resolution of the taxa. Furthermore, we focused on bacterial DNA, not implying that those bacteria were alive at the moment of sampling or that they were pathogens. Finally, although bacteria are the main microorganisms involved in sepsis (Faria et al. 2018), it can be also caused by viruses and fungi, which were not analyzed in this study.

Conclusions

The results of this study revealed that a decreased bacterial lung diversity within 8 h of sepsis was associated with ICU mortality among non-pulmonary sepsis patients. Additionally, both the presence of gut commensals in the lungs and the reduction of healthy lung bacteria were related to ICU mortality. These results support a central role of the host-microbial interactions in maintaining lung homeostasis and provide a novel early prognostic biomarker for sepsis.

Funding

Funded by Instituto de Salud Carlos III (PI14/00844; PI17/00610; PI17/00610; FI17/00177) and co-financed by the European Regional Development Funds, “A way of making Europe” from the European Union, Cabildo Insular de Tenerife (CGIEU0000219140), and by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology. BGG was supported by a fellowship from the ACIISI (TESIS2015010057) co-funded by European Social Fund.

References

Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. 2001. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit. Care Med. 29:1303–1310.

Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, Gattinoni L, van Haren F, Larsson A, McAuley DF, et al. 2016. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. JAMA 315:788–800.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7:335–336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 6:1621–1624.

Chen C-Y, Chen Y-H, Lu P-L, Lin W-R, Chen T-C, Lin C-Y. 2012. Proteus mirabilis urinary tract infection and bacteremia: risk factors, clinical presentation, and outcomes. J. Microbiol. Immunol. Infect. 45:228–236.

Cumsille F, Bangdiwala SI, sen PK, Kupper LL. 2000. Effect of dichotomizinlg a continuous variable on the model structure in multiple linear regression models. Commun. Stat. Methods 29:643–654.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069–5072.

Dickson RP. 2016. The microbiome and critical illness. Lancet. Respir. Med. 4:59–72.

Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB, Curtis JL. 2017. Bacterial Topography of the Healthy Human Lower Respiratory Tract. MBio 8.

Dickson RP, Singer BH, Newstead MW, Falkowski NR, Erb-Downward JR, Standiford TJ, Huffnagle GB. 2016. Enrichment of the lung microbiome with gut bacteria in sepsis and the acute respiratory distress syndrome. Nat. Microbiol. 1:16113.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Faria MMP, Winston BW, Surette MG, Conly JM. 2018. Bacterial DNA patterns identified using paired-end Illumina sequencing of 16S rRNA genes from whole blood samples of septic patients in the emergency room and intensive care unit. BMC Microbiol. 18:79.

Finegold SM, Johnson CC. 1985. Lower respiratory tract infection. Am. J. Med. 79:73–77.

Giamarellos-Bourboulis EJ, Norrby-Teglund A, Mylona V, Savva A, Tsangaris I, Dimopoulou I, Mouktaroudi M, Raftogiannis M, Georgitsi M, Linnér A, et al. 2012. Risk assessment in sepsis: a new prognostication rule by APACHE II score and serum soluble urokinase plasminogen activator receptor. Crit. Care 16:R149.

Gibot S, Béné MC, Noel R, Massin F, Guy J, Cravoisy A, Barraud D, De Carvalho Bittencourt M, Quenot J-P, Bollaert P-E, et al. 2012. Combination biomarkers to diagnose sepsis in the critically ill patient. Am. J. Respir. Crit. Care Med. 186:65–71.

Gosiewski T, Ludwig-Galezowska AH, Huminska K, Sroka-Oleksiak A, Radkowski P, Salamon D, Wojciechowicz J, Kus-Slowinska M, Bulanda M, Wolkow PP. 2017. Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy volunteers using next-generation sequencing method - the observation of DNAemia. Eur. J. Clin. Microbiol. Infect. Dis. 36:329–336.

Haak BW, Wiersinga WJ. 2017. The role of the gut microbiota in sepsis. lancet. Gastroenterol. Hepatol. 2:135–143.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward D V, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, et al. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 21:494–504.

Hamilton AL, Kamm MA, Ng SC, Morrison M. 2018. Proteus spp. as Putative Gastrointestinal Pathogens. Clin. Microbiol. Rev. 31:3.

Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A, Poulter L, Pachter L, et al. 2010. Disordered microbial communities in asthmatic airways. PLoS One 5:1.

Huffnagle GB, Dickson RP, Lukacs NW. 2017. The respiratory tract microbiome and lung inflammation: a two-way street. Mucosal Immunol. 10:299–306.

Jacobs MC, Haak BW, Hugenholtz F, Wiersinga WJ. 2017. Gut microbiota and host defense in critical illness. Curr. Opin. Crit. Care 23:257–263.

Lamarche D, Johnstone J, Zytaruk N, Clarke F, Hand L, Loukov D, Szamosi JC, Rossi L, Schenck LP, Verschoor CP, et al. 2018. Microbial dysbiosis and mortality during mechanical ventilation: a prospective observational study. Respir. Res. 19:245.

Larsen FF, Petersen JA. 2017. Novel biomarkers for sepsis: A narrative review. Eur. J. Intern. Med. 45:46–50.

Lê S, Josse J, Husson F. 2008. FactoMineR: An R Package for Multivariate Analysis. J. Stat. Software, Artic. 25:1–18.

MacFie J, O'Boyle C, Mitchell CJ, Buckley PM, Johnstone D, Sudworth P. 1999. Gut origin of sepsis: a prospective study investigating associations between bacterial translocation, gastric microflora, and septic morbidity. Gut 45:223–228.

McDonald D, Ackermann G, Khailova L, Baird C, Heyland D, Kozar R, Lemieux M, Derenski K, King J, Vis-Kampen C, et al. 2016. Extreme Dysbiosis of the Microbiome in Critical Illness. mSphere 1.

Minasyan H. 2019. Sepsis: mechanisms of bacterial injury to the patient. Scand. J. Trauma. Resusc. Emerg. Med. 27:19.

Moffatt MF, Cookson WO. 2017. The lung microbiome in health and disease. Clin. Med. 17:525–529.

Molyneaux PL, Cox MJ, Willis-Owen SAG, Mallia P, Russell KE, Russell A-M, Murphy E, Johnston SL, Schwartz DA, Wells AU, et al. 2014. The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis. Am. J. Respir. Crit. Care Med. 190:906–913.

Mukherjee S, Hanidziar D. 2018. More of the Gut in the Lung: How Two Microbiomes Meet in ARDS. Yale J. Biol. Med. 91:143–149.

O'Dwyer DN, Ashley SL, Gurczynski SJ, Xia M, Wilke C, Falkowski NR, Norman KC, Arnold KB, Huffnagle GB, Salisbury ML, et al. 2019. Lung Microbiota Contribute to Pulmonary Inflammation and Disease Progression in Pulmonary Fibrosis. Am. J. Respir. Crit. Care Med. 199:1127–1138.

Opal SM, Garber GE, LaRosa SP, Maki DG, Freebairn RC, Kinasewitz GT, Dhainaut J-F, Yan SB, Williams MD, Graham DE, et al. 2003. Systemic Host Responses in Severe Sepsis Analyzed by Causative Microorganism and Treatment Effects of Drotrecogin Alfa (Activated). Clin. Infect. Dis. 37:50–58.

Panzer AR, Lynch SV., Langelier C, Christie JD, McCauley K, Nelson M, Cheung CK, Benowitz NL, Cohen MJ, Calfee CS. 2018. Lung Microbiota Is Related to Smoking Status and to Development of Acute Respiratory Distress Syndrome in Critically Ill Trauma Patients. Am. J. Respir. Crit. Care Med. 197:621–631.

Park H, Shin JW, Park S-G, Kim W. 2014. Microbial communities in the upper respiratory tract of patients with asthma and chronic obstructive pulmonary disease. PLoS One 9:e109710.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

R Core Team. 2013. R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria. Available online from: http://www.r-project.org/

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M, Siegert S. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77.

Rodríguez-González R, Martín-Barrasa JL, Ramos-Nuez Á, Cañas-Pedrosa AM, Martínez-Saavedra MT, García-Bello MÁ, López-Aguilar J, Baluja A, Álvarez J, Slutsky AS, et al. 2014. Multiple system organ response induced by hyperoxia in a clinically relevant animal model of sepsis. Shock 42:148–153.

La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD. 2012. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS One 7:e52078.

Shimizu K, Ogura H, Hamasaki T, Goto M, Tasaki O, Asahara T, Nomoto K, Morotomi M, Matsushima A, Kuwagata Y, et al. 2011. Altered gut flora are associated with septic complications and death in critically ill patients with systemic inflammatory response syndrome. Dig. Dis. Sci. 56:1171–1177.

Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM, et al. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 315:801–810.

de Steenhuijsen Piters WAA, Huijskens EGW, Wyllie AL, Biesbroek G, van den Bergh MR, Veenhoven RH, Wang X, Trzciński K, Bonten MJ, Rossen JWA, et al. 2016. Dysbiosis of upper respiratory tract microbiota in elderly pneumonia patients. ISME J. 10:97–108.

Tan X, Liu H, Long J, Jiang Z, Luo Y, Zhao X, Cai S, Zhong X, Cen Z, Su J, et al. 2019. Septic patients in the intensive care unit present different nasal microbiotas. Future Microbiol. 14:383–395.

Thaiss CA, Zmora N, Levy M, Elinav E. 2016. The microbiome and innate immunity. Nature 535:65–74.

Vincent JL, Opal SM, Marshall JC. 2010. Ten reasons why we should NOT use severity scores as entry criteria for clinical trials or in our treatment decisions. Crit. Care Med. 38:283–287.

Xia LP, Bian LY, Xu M, Liu Y, Tang AL, Ye WQ. 2015. 16S rRNA gene sequencing is a non-culture method of defining the specific bacterial etiology of ventilator-associated pneumonia. Int. J. Clin. Exp. Med. 8:18560–18570.

Yadava K, Pattaroni C, Sichelstiel AK, Trompette A, Gollwitzer ES, Salami O, von Garnier C, Nicod LP, Marsland BJ. 2016. Microbiota Promotes Chronic Pulmonary Inflammation by Enhancing IL-17A and Autoantibodies. Am. J. Respir. Crit. Care Med. 193:975–987.

Zakrzewski M, Proietti C, Ellis JJ, Hasan S, Brion M-J, Berger B, Krause L. 2017. Calypso: a user-friendly web-server for mining and visualizing microbiome-environment interactions. Bioinformatics 33:782–783.

Zemanick ET, Wagner BD, Robertson CE, Ahrens RC, Chmiel JF, Clancy JP, Gibson RL, Harris WT, Kurland G, Laguna TA, et al. 2017. Airway microbiota across age and disease spectrum in cystic fibrosis. Eur. Respir. J. 50:1700832.

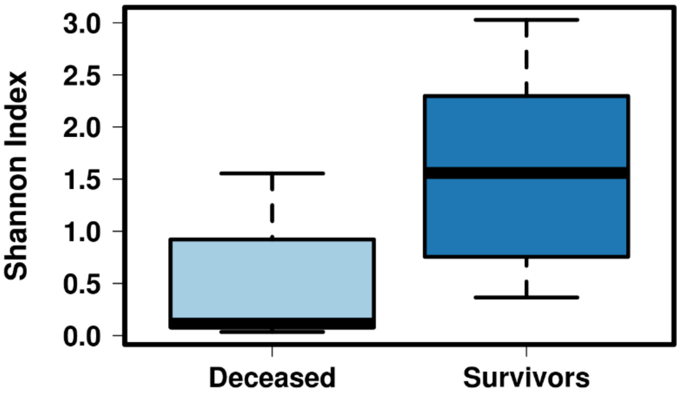


Figure 1. Bacterial lung diversity in ICU patients with sepsis at 8 h of diagnosis.

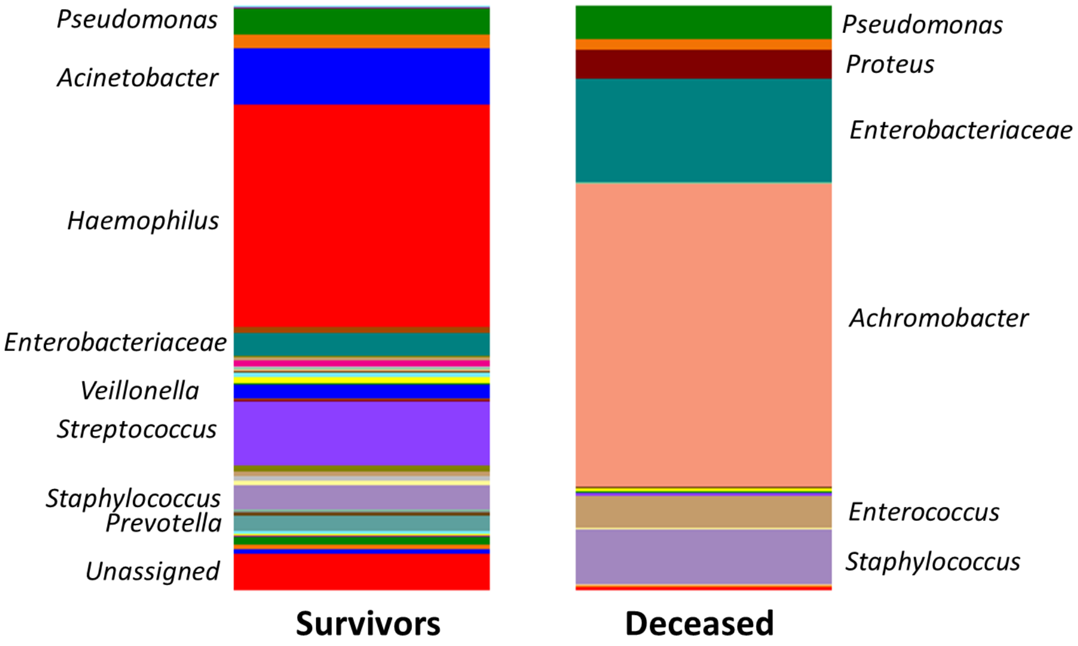


Figure 2. Main bacterial taxa identified in lung aspirates from ICU patients with sepsis at 8 h of diagnosis.

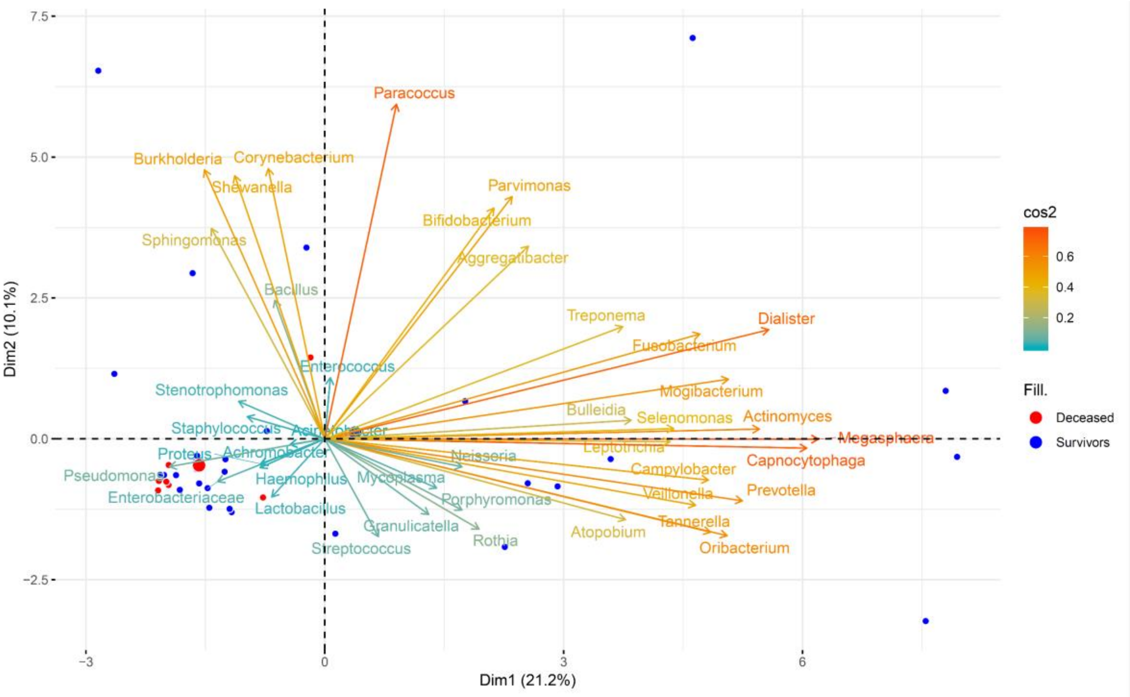


Figure 3. PCA biplot summarizing the two main dimensions of patient differentiation (colored dots) and their correlation with the bacterial taxa (arrows) identified in lung aspirates collected at 8 h of diagnosis.

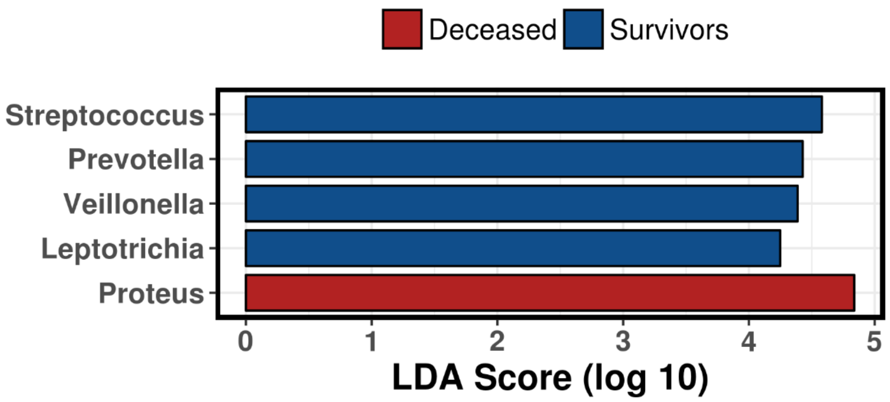


Figure 4. Prioritized lung bacterial genera explaining the differences between deceased and survivors at 8 h of sepsis diagnosis based on the LEfSe analysis.

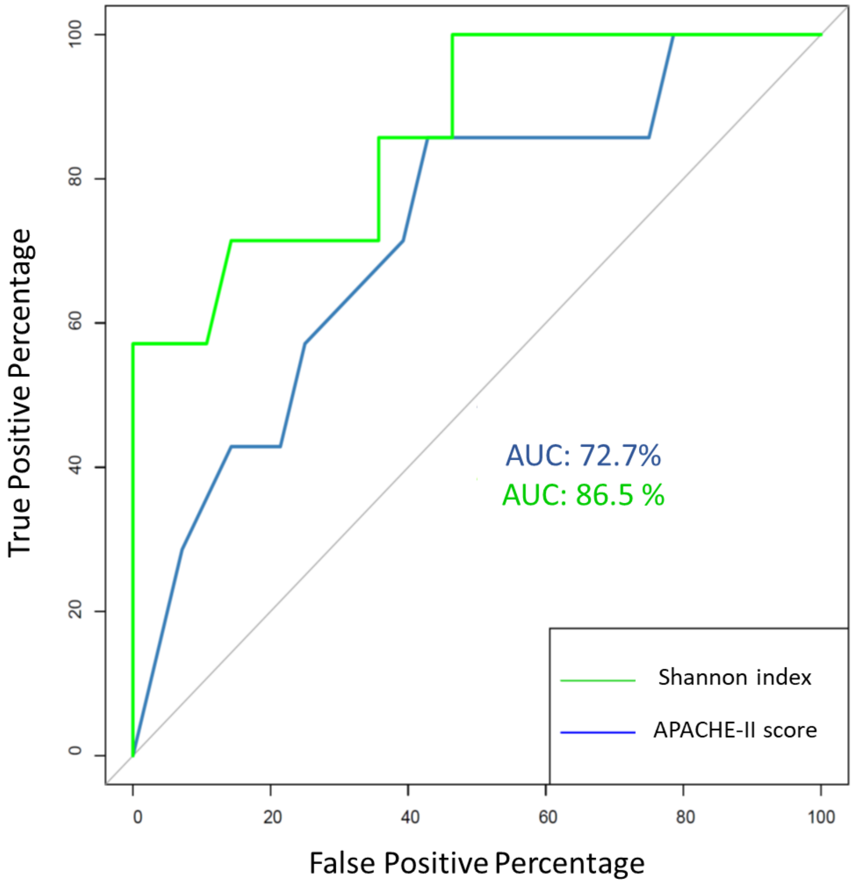


Figure 5. Comparison of ROC curves and AUC estimates of the bacterial lung diversity at 8 h of diagnosis (Shannon index) and the APACHE II score at inclusion as predictors of ICU mortality.

Table 1. Demographic and clinical features of the sample.

	Survivors (n=29)	**Deceased (n=7)**	***p*-value***
Sex (% male)	75.9	57.1	0.60
Mean age (years)	66.3 ± 11.8	71.6 ± 5.9	4.44E-14
Hypertension (%)	55.2	57.1	1.00
Smokers (%)	20.7	28.6	1.00
ARDS (%)	13.8	0.0	0.71
Previous severe infections	13.8	0.0	0.71
Antibiotic treatment (%)†	82.8	85.7	1.00
Comorbidities	32.0	57.1	0.44
APACHE II (median) (P_{25}–P_{75})	24 (19-27)	27 (26-31)	1.73E-13
Infection source (%)			
Abdominal and gastrointestinal tract	86.2	100.0	0.71
Genitourinary system	10.3	0.0	0.90
Bones and soft tissues	3.4	0.0	1.00
Brain and central nervous system	0.0	0.0	-
Pathogen (%)			
Gram-positive bacteria	16.7	28.6	0.71
Gram-negative bacteria	37.5	28.6	1.00
Gram-positive and Gram-negative	37.5	14.3	1.00
Fungi	4.2	0.00	1.00
Virus	0.00	0.00	-
Polymicrobial	29.2	28.6	1.00
Organ dysfunction (%)			
Circulatory	89.7	100.0	0.90
Coagulation	34.5	42.9	0.88
Hepatic	20.7	14.3	1.00
Neurologic	65.5	71.4	1.00
Renal	34.5	14.3	0.56

*Mean age and APACHE II comparisons were conducted by the Wilcoxon signed-rank test; the other variables were compared by a chi-square test.

†Percentage of patients with an active antibiotic treatment at 8 h of sepsis diagnosis.

APACHE II, Acute Physiology and Chronic Health Evaluation II at inclusion; ARDS, acute respiratory distress syndrome; P25, percentile 25; P75, percentile 75. Percentages refer only to the individuals with available data for each clinical feature.

Table 2. Differences in diversity between the three sample collection times

	8h (7:29)[a]	**48h (3:14)**[a]	**72h (2:14)**[a]	***p*-value**[b]
Shannon diversity index (mean ± SD)	2.02 ± 1.34	2.01 ± 1.34	1.97 ± 1.31	0.99

[a]Deceased:Survivors with available data; [b]Data compared by Kruskal-Wallis test.

Table 3. Sensitivity analyses for the Shannon diversity index

	Adjusted model	
	OR [95% CI]	***p*-value**
Sex	0.23 [0.07, 0.84]	0.026
Age	0.21 [0.05, 0.81]	0.023
APACHE II	0.24 [0.07, 0.85]	0.027
ARDS	0.22 [0.06, 0.78]	0.019
Smokers	0.23 [0.07, 0.80]	0.020
Previous severe infections	0.20 [0.06, 0.74]	0.016
Infection source	0.25 [0.08, 0.82]	0.023
Isolated pathogen	0.23 [0.07, 0.81]	0.022
Antibiotic treatment*	0.17 [0.04, 0.74]	0.018
Multi organ dysfunction†	0.20 [0.05, 0.80]	0.022
Comorbidities‡	0.23 [0.06, 0.85]	0.028
Arterial hypertension	0.22 [0.06, 0.80]	0.022

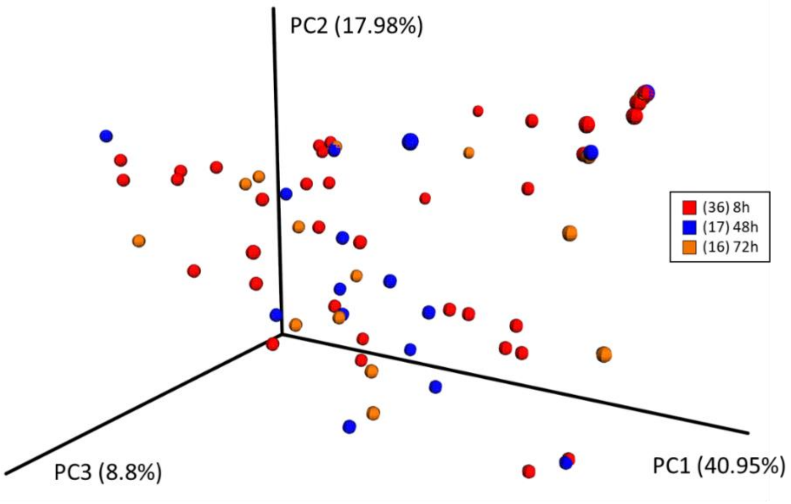
*Active antibiotic treatment at 8 h of sepsis diagnosis.

†Two or more affected organs.

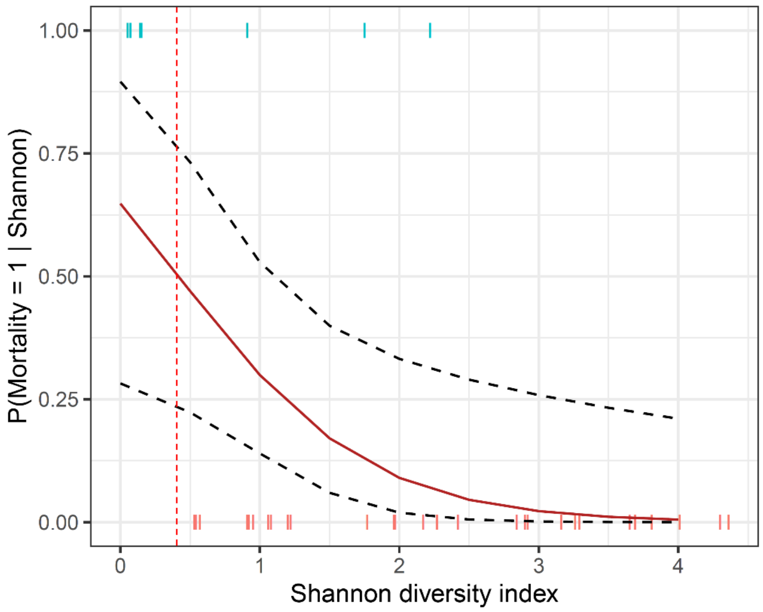
‡Presence of comorbidities (autoimmune diseases, cancer, chronic diseases, diabetes, hepatopathies, immunosuppression, kidney diseases, morbid obesity, pregnancy, severe infections, severe brain damage, valvulopathies).

APACHE II, Acute Physiology and Chronic Health Evaluation II at inclusion; ARDS, Acute Respiratory Distress Syndrome; OR, Odds Ratio; CI, Confidence Interval.

Supplementary Material



Supplementary Figure 1. Plot of the three main principal components (PC) of patient differentiation at the three distinct sample collection times (8 h, 48 h and 72 h) based on weighted UniFrac distances from the abundance of lung bacterial taxa. In parenthesis, fraction of variance explained by each PC. The legend shows the number of samples with available data at each collection time.



Supplementary Figure 2. ICU mortality probability (with 95% confidence interval) estimated based on bacterial lung diversity at 8 h of sepsis diagnosis. The model estimated a cutoff point of Shannon diversity index at 0.42 (vertical broken red line), assuming a probability threshold of 0.5 in a binary logistic regression. Individual bacterial diversity estimates in deceased (blue) and survivors (red) are indicated.

Supplementary Table 1. Spearman correlation between the different diversity indices estimated in the study patients.

	Shannon	Simpson	Chao1	Observed species
Shannon	1	-	-	-
Simpson	0.99	1	-	-
Chao1	0.85	0.82	1	
Observed species	0.94	0.91	0.92	1

Chapter 4.

Genomic Analyses of Human European Diversity at the Southwestern Edge: Isolation, African Influence and Disease Associations in the Canary Islands

Since the genetic ancestry is associated with the development and outcomes of complex diseases, likely including critical illnesses, the study of the genetic makeup of a recently admixed population is crucial to identify genomic regions where ancestry tend to be coinherited with specific diseases. This fourth chapter contains the results of the genomic characterization of the recent evolutionary history of Canary Islanders by using SNP array data and WGS. We estimated the global and local genetic ancestries of this population and assessed the links between particular regions and disease risks.

The results of this study revealed that up to 34% of the genome of current Canary Islanders is of recent African descent. Additionally, we identified eight genomic regions with large local ancestry deviations in African or European ancestry that harbored genes linked to prevalent diseases, such as asthma and diabetes, to infectious diseases and to severe acute respiratory syndrome (SARS), among other traits. Interestingly, some of these genomic regions were located near well-known targets of natural selection, including the lactase (*LCT*) gene and the HLA region. We also estimated that the last African admixture in this population occurred ~14 generations ago, and that the average number of ancestry blocks per haploid genome equals 276. These findings lay the foundations for designing admixture mapping studies in the Canary Islands population to identify novel disease risk genes for complex traits such as sepsis and ARDS.

This chapter was published in 2018 in *Molecular Biology and Evolution* with the title "Genomic Analyses of Human European Diversity at the Southwestern Edge: Isolation, African Influence and Disease Associations in the Canary Islands" (doi: 10.1093/molbev/msy190). This is an Open Access article that is reproduced under the terms of John Wiley and Sons License.

Genomic Analyses of Human European Diversity at the Southwestern Edge: Isolation, African Influence and Disease Associations in the Canary Islands

Beatriz Guillen-Guio,[1] Jose M. Lorenzo-Salazar,[2] Rafaela González-Montelongo,[2] Ana Díaz-de Usera,[2] Itahisa Marcelino-Rodríguez,[1] Almudena Corrales,[1,3] Antonio Cabrera de León,[1] Santos Alonso,[4] and Carlos Flores*,[1,2,3]

[1]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain

[2]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain

[3]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain

[4]Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country UPV/EHU, Leioa, Bizkaia, Spain

***Corresponding author:** E-mail: cflores@ull.edu.es.

Associate editor: Connie Mulligan

Abstract

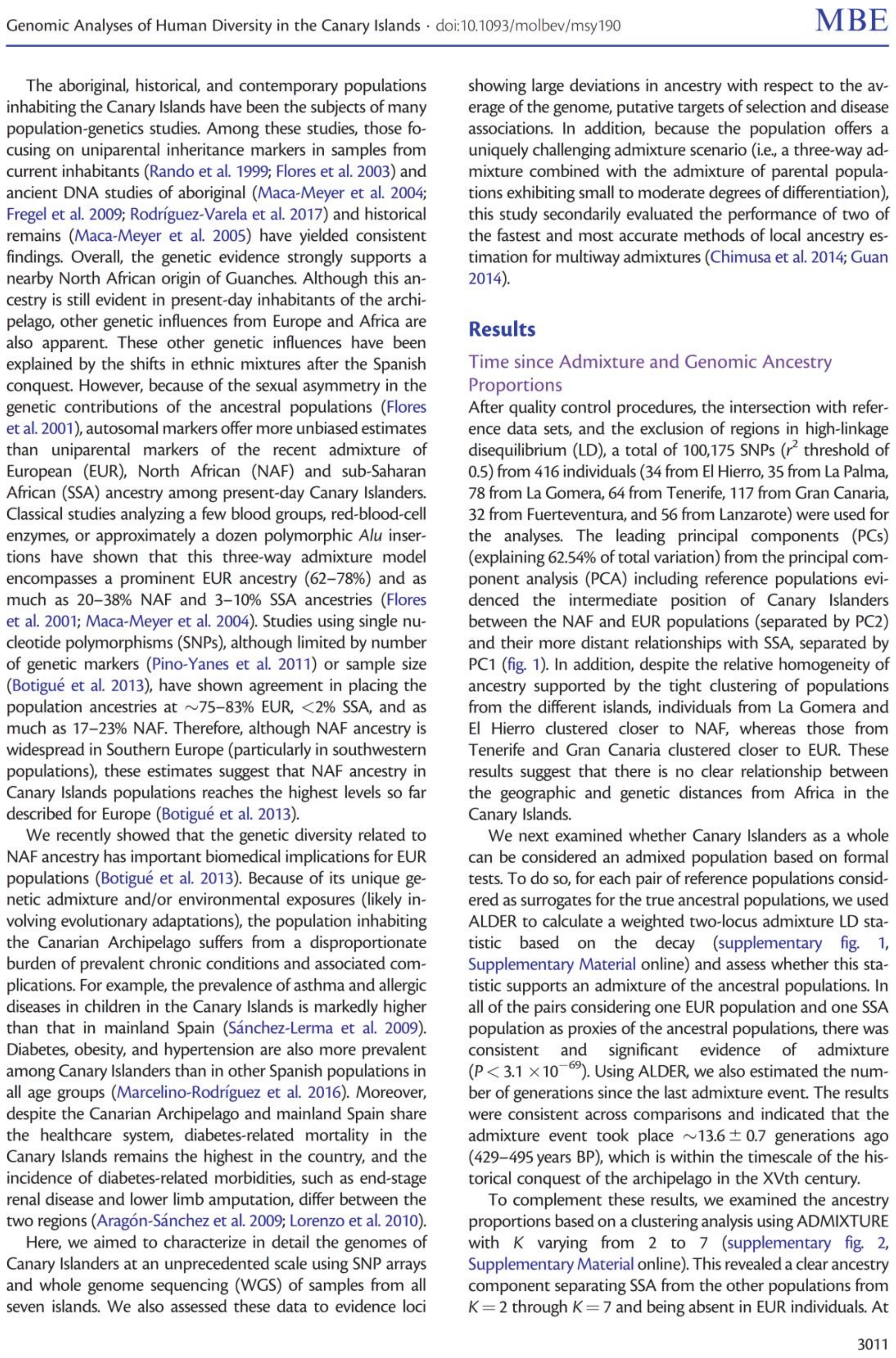
Despite the genetic resemblance of Canary Islanders to other southern European populations, their geographical isolation and the historical admixture of aborigines (from North Africa) with sub-Saharan Africans and Europeans have shaped a distinctive genetic makeup that likely affects disease susceptibility and health disparities. Based on single nucleotide polymorphism array data and whole genome sequencing (30×), we inferred that the last African admixture took place ∼14 generations ago and estimated that up to 34% of the Canary Islander genome is of recent African descent. The length of regions in homozygosis and the ancestry-related mosaic organization of the Canary Islander genome support the view that isolation has been strongest on the two smallest islands. Furthermore, several genomic regions showed significant and large deviations in African or European ancestry and were significantly enriched in genes involved in prevalent diseases in this community, such as diabetes, asthma, and allergy. The most prominent of these regions were located near *LCT* and the HLA, two well-known targets of selection, at which 40–50% of the Canarian genome is of recent African descent according to our estimates. Putative selective signals were also identified in these regions near the *SLC6A11-SLC6A1*, *KCNMB2*, and *PCDH20-PCDH9* genes. Taken together, our findings provide solid evidence of a significant recent African admixture, population isolation, and adaptation in this part of Europe, with the favoring of African alleles in some chromosome regions. These findings may have medical implications for populations of recent African ancestry.

Key words: **consanguinity, ROH, local ancestry, MHC, natural selection.**

Introduction

The Canarian Archipelago consists of seven main islands located in the Atlantic Ocean ∼100 km from the northwest African coast. By no later than 2,500 years B.P. (Onrubia Pintado 1987) and until the XVth century, when the Spanish conquest began, the Canary Islands were inhabited by the indigenous Guanche population (Crosby 1999). Many anthropological, archaeological, and cultural traits indicate that the most likely origin of Guanche aborigines was the Berber population from North Africa (Hooton 1970), and supporting evidence indicates more than one aboriginal settlement from the Maghreb and the Sahara likely occurred (Arco and Navarro 1987). The Spanish conquest can be divided into two stages: 1) occupation of the less populated islands, which was concluded rapidly and peacefully, and 2) a subsequent slower and violent invasion of the more populated islands, which ended in 1496 (de Abreu Galindo and Cioranescu 1977). Despite the devastating effect of the conquest, many aborigines remained in the territory, either freed or enslaved. The strategic location of the Canary Islands (located between the Americas and Africa) stimulated a continuous immigration between the XVIth and XIXth centuries from Europe and sub-Saharan Africa, with immigration from the latter occurring as a result of the slave trade (Lobo-Cabrera 1993). By the XVIth century, the chronicles estimated the population size as ∼35,000 inhabitants, of which nearly 11,000 were potentially of aboriginal or sub-Saharan African origin (Wölfel 1930). Physical anthropology studies of the inhabitants provided evidence of the continuity of indigenous traits during the XXth century (Falkenburger 1942; Ara 1959; Berthelot 1978).

3010 *Mol. Biol. Evol.* 35(12):3010–3026 doi:10.1093/molbev/msy190 Advance Access publication October 5, 2018



The aboriginal, historical, and contemporary populations inhabiting the Canary Islands have been the subjects of many population-genetics studies. Among these studies, those focusing on uniparental inheritance markers in samples from current inhabitants (Rando et al. 1999; Flores et al. 2003) and ancient DNA studies of aboriginal (Maca-Meyer et al. 2004; Fregel et al. 2009; Rodríguez-Varela et al. 2017) and historical remains (Maca-Meyer et al. 2005) have yielded consistent findings. Overall, the genetic evidence strongly supports a nearby North African origin of Guanches. Although this ancestry is still evident in present-day inhabitants of the archipelago, other genetic influences from Europe and Africa are also apparent. These other genetic influences have been explained by the shifts in ethnic mixtures after the Spanish conquest. However, because of the sexual asymmetry in the genetic contributions of the ancestral populations (Flores et al. 2001), autosomal markers offer more unbiased estimates than uniparental markers of the recent admixture of European (EUR), North African (NAF) and sub-Saharan African (SSA) ancestry among present-day Canary Islanders. Classical studies analyzing a few blood groups, red-blood-cell enzymes, or approximately a dozen polymorphic *Alu* insertions have shown that this three-way admixture model encompasses a prominent EUR ancestry (62–78%) and as much as 20–38% NAF and 3–10% SSA ancestries (Flores et al. 2001; Maca-Meyer et al. 2004). Studies using single nucleotide polymorphisms (SNPs), although limited by number of genetic markers (Pino-Yanes et al. 2011) or sample size (Botigué et al. 2013), have shown agreement in placing the population ancestries at ~75–83% EUR, <2% SSA, and as much as 17–23% NAF. Therefore, although NAF ancestry is widespread in Southern Europe (particularly in southwestern populations), these estimates suggest that NAF ancestry in Canary Islands populations reaches the highest levels so far described for Europe (Botigué et al. 2013).

We recently showed that the genetic diversity related to NAF ancestry has important biomedical implications for EUR populations (Botigué et al. 2013). Because of its unique genetic admixture and/or environmental exposures (likely involving evolutionary adaptations), the population inhabiting the Canarian Archipelago suffers from a disproportionate burden of prevalent chronic conditions and associated complications. For example, the prevalence of asthma and allergic diseases in children in the Canary Islands is markedly higher than that in mainland Spain (Sánchez-Lerma et al. 2009). Diabetes, obesity, and hypertension are also more prevalent among Canary Islanders than in other Spanish populations in all age groups (Marcelino-Rodríguez et al. 2016). Moreover, despite the Canarian Archipelago and mainland Spain share the healthcare system, diabetes-related mortality in the Canary Islands remains the highest in the country, and the incidence of diabetes-related morbidities, such as end-stage renal disease and lower limb amputation, differ between the two regions (Aragón-Sánchez et al. 2009; Lorenzo et al. 2010).

Here, we aimed to characterize in detail the genomes of Canary Islanders at an unprecedented scale using SNP arrays and whole genome sequencing (WGS) of samples from all seven islands. We also assessed these data to evidence loci showing large deviations in ancestry with respect to the average of the genome, putative targets of selection and disease associations. In addition, because the population offers a uniquely challenging admixture scenario (i.e., a three-way admixture combined with the admixture of parental populations exhibiting small to moderate degrees of differentiation), this study secondarily evaluated the performance of two of the fastest and most accurate methods of local ancestry estimation for multiway admixtures (Chimusa et al. 2014; Guan 2014).

Results

Time since Admixture and Genomic Ancestry Proportions

After quality control procedures, the intersection with reference data sets, and the exclusion of regions in high-linkage disequilibrium (LD), a total of 100,175 SNPs (r^2 threshold of 0.5) from 416 individuals (34 from El Hierro, 35 from La Palma, 78 from La Gomera, 64 from Tenerife, 117 from Gran Canaria, 32 from Fuerteventura, and 56 from Lanzarote) were used for the analyses. The leading principal components (PCs) (explaining 62.54% of total variation) from the principal component analysis (PCA) including reference populations evidenced the intermediate position of Canary Islanders between the NAF and EUR populations (separated by PC2) and their more distant relationships with SSA, separated by PC1 (fig. 1). In addition, despite the relative homogeneity of ancestry supported by the tight clustering of populations from the different islands, individuals from La Gomera and El Hierro clustered closer to NAF, whereas those from Tenerife and Gran Canaria clustered closer to EUR. These results suggest that there is no clear relationship between the geographic and genetic distances from Africa in the Canary Islands.

We next examined whether Canary Islanders as a whole can be considered an admixed population based on formal tests. To do so, for each pair of reference populations considered as surrogates for the true ancestral populations, we used ALDER to calculate a weighted two-locus admixture LD statistic based on the decay (supplementary fig. 1, Supplementary Material online) and assess whether this statistic supports an admixture of the ancestral populations. In all of the pairs considering one EUR population and one SSA population as proxies of the ancestral populations, there was consistent and significant evidence of admixture ($P < 3.1 \times 10^{-69}$). Using ALDER, we also estimated the number of generations since the last admixture event. The results were consistent across comparisons and indicated that the admixture event took place ~13.6 ± 0.7 generations ago (429–495 years BP), which is within the timescale of the historical conquest of the archipelago in the XVth century.

To complement these results, we examined the ancestry proportions based on a clustering analysis using ADMIXTURE with K varying from 2 to 7 (supplementary fig. 2, Supplementary Material online). This revealed a clear ancestry component separating SSA from the other populations from $K = 2$ through $K = 7$ and being absent in EUR individuals. At

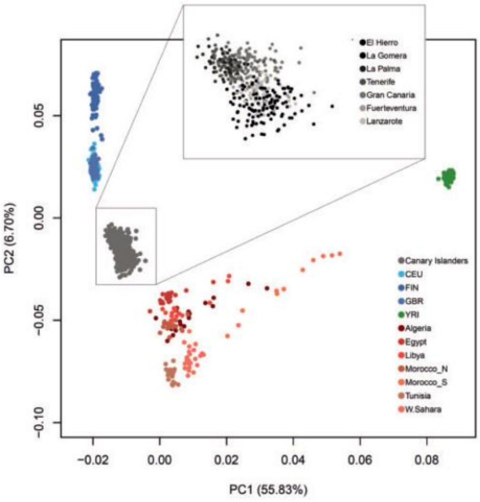


Fig. 1. Plot of the first two principal components (explaining 62.5% of variability) from PCA of Canary Islanders and samples from reference populations from Europe, North Africa, and sub-Saharan Africa. The inset depicts a detailed view per island. Results are based on a subsample of 100,175 SNPs excluding those in high LD (pairwise r^2 threshold = 0.5).

$K \geq 3$, a new ancestry cluster became apparent, which reached its maximum frequency in EUR and Canary Islanders. At $K = 4$ and above, ∼40–60% of the ancestry clusters detected in NAF individuals were assigned to a new cluster with its maximum in these populations (>95% on an average in Tunisians). This NAF-related ancestry was also evident in Canary Islanders. Other ancestry clusters emerged for $K > 4$, which generally reflected both further ancestry sharing or unresolved clustering in NAF and Canary Islanders and additional ancestry subdivisions among NAF populations compatible with the mixed pattern of ancestry components evidenced elsewhere (Henn et al. 2012; Arauna et al. 2017). Cross-validation error was lowest for $K = 4$ (fig. 2). At this value, the clusters roughly corresponded to the ancestries, reaching their maximum frequencies in SSA and NAF plus two other ancestry clusters defined for EUR. The results for EUR are largely compatible with previous observations (Seldin et al. 2006). To provide further support to the ancestry clusters evidenced by the ADMIXTURE analysis, we evaluated the fitting of the admixture model with the optimal measures of haplotype sharing between groups using badMIXTURE. For $K = 4$, we observed small residuals with no systematic patterns on the Canary Islanders, whereas larger residuals were concentrated in one of the ancestral components (NAF) (supplementary fig. 3, Supplementary Material online). Therefore, regardless of the genetic heterogeneity resulting from the merging of different North African populations into a single group, the structure of the residuals indicate that the admixture proportions provided by the ADMIXTURE analysis constitute robust inferences of the ancestries in Canary Islanders.

Based on this evidence, we interpreted the EUR-, NAF-, and SSA-related fractions indicated by the ADMIXTURE analysis as the admixture proportions in Canary Islanders. For $K = 4$, the overall ancestry proportions among Canary Islanders were, on an average, 22% NAF and 3% SSA (table 1). These estimates are highly concordant with our previous results obtained under standard settings (Botigué et al. 2013) despite the latter being based on a data set containing SNPs with moderate levels of LD (r^2 threshold set at 0.5). However, in this study, which comprised a larger and more diverse sample of Canary Islanders, we found a wider NAF-related ancestry interval of 14.9–29.9%, and a slightly higher SSA-related ancestry of as much as 9.2% (table 1 and fig. 2). The use of alternative African data sets from The 1000 Genomes Project (1KGP) as SSA representatives in the ADMIXTURE analysis did not qualitatively change the results as the ancestry estimates were highly correlated (for $K = 4$, $R^2 \geq 0.995$; $P < 2.2 \times 10^{-16}$ in all comparisons). Similarly, balancing the number of individuals from the reference populations ($n = 75$ for each) resulted in similar ancestry estimates (averages of 26% NAF and 1% SSA). Overall, these results demonstrate that the Canary Islanders are closely related to EUR but show substantial influences from NAF and distant relatedness with SSA.

There was no difference in the NAF-related ancestry between the eastern (Fuerteventura, Lanzarote, and Gran Canaria) and western islands (Tenerife, La Gomera, La Palma, and El Hierro) (22.4% vs. 21.8%, respectively; Wilcoxon test $P = 0.160$), although the SSA-related ancestry differed significantly between the eastern and western islands (3.1% vs. 2.8%, respectively; Wilcoxon test $P = 6.4 \times 10^{-5}$). However, the largest differences were observed between the islands that were conquered first and more peacefully by the Spanish (Fuerteventura, Lanzarote, La Gomera, and El Hierro) and those in which the conquest took longer (Tenerife, Gran Canaria, and La Palma) (24.0% vs. 20.3%, respectively, for the NAF-related fraction, and 3.5% vs. 2.4%, respectively, for the SSA-related fraction; Wilcoxon test $P < 3.0 \times 10^{-5}$ for both comparisons). This observation agrees with findings based on mitochondrial DNA (mtDNA) lineages that suggest increased African affinities on the former islands (Rando et al. 1999). However, they differ from those based on the nonrecombining portion of the Y chromosome (NRY) that indicate limited NAF affinities of paternal genetic markers on all of the Canary Islands but particularly the populations from the westernmost islands (Flores et al. 2003). Such different distributions of NAF and SSA ancestries, previously evidenced by only uniparental markers, are expected given the sexual asymmetry of parental contributions detected in the current and historical populations of the archipelago (Flores et al. 2001; Fregel et al. 2009).

Population Isolation

Recent studies have indicated that runs of homozygosity (ROHs) are common to all world populations, are longer than expected, and have profiles that can indicate distinctive demographic histories of the population and of inbreeding (Kirin et al. 2010; Pemberton et al. 2012). Here, we have

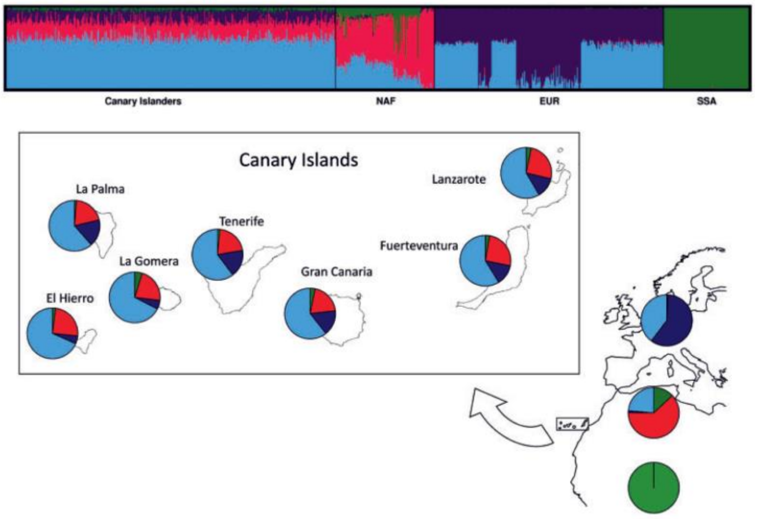


Fig. 2. ADMIXTURE estimates for $K = 4$ for Canary Islanders and samples from reference populations from Europe, North Africa, and sub-Saharan Africa.

Table 1. Genomic Ancestry Proportions (from ADMIXTURE, $K = 4$) in Canary Islanders.

	North African			Sub-Saharan African		
	Min.	Mean	Max.	Min.	Mean	Max.
Fuerteventura	0.218	0.255	0.296	0.011	0.027	0.046
Lanzarote	0.214	0.254	0.296	0.014	0.032	0.057
Gran Canaria	0.155	0.200	0.264	0.005	0.032	0.082
Tenerife	0.149	0.208	0.255	0.002	0.015	0.057
La Gomera	0.160	0.221	0.289	0.013	0.048	0.092
La Palma	0.170	0.200	0.245	0.000	0.013	0.032
El Hierro	0.192	0.246	0.299	0.005	0.020	0.032

analyzed, for the first time, genome-wide ROH patterns in Canary Islanders to reveal the level of population isolation. Focusing first on the average total sum of ROHs over all subjects sampled, we found that <10 Mb of the Canary Islanders genome on an average was in ROHs of >2 Mb in length. For smaller ROHs (≤1.6 Mb), which have been associated with geographic distances from East Africa (Pemberton et al. 2012), the profiles were very similar across the island populations (fig. 3). However, among all ROHs >1.6 Mb, the samples from La Gomera and El Hierro showed consistently higher average total ROHs than did the samples from the remaining islands, suggestive of increased recent inbreeding in these two islands.

An exploration of average total ROHs by island again indicated El Hierro and La Gomera as population outliers. The genomes from these two islands showed the largest number of fragments in ROHs and the longest average total ROH length (fig. 4), the latter demonstrating significant differences from all the other islands (Wilcoxon test $P < 1.0\times10^{-3}$ for all pairwise comparisons) except La Palma. Interestingly, La Gomera and El Hierro showed average total ROH lengths of 114 and 134 Mb. The conclusions remained unchanged when considering only ROHs >1.6 Mb for the average total ROHs estimates (Wilcoxon test $P < 1.0\times10^{-3}$ for all pairwise comparisons), whereas the differences disappeared when using the average total of ROHs ≤1.6 Mb (lowest $P = 4.5\times10^{-3}$). Given that variant ascertainment of the array is less of a problem for ROHs >1Mb (Ceballos et al. 2018), we interpret the ROH patterns observed for La Gomera and El Hierro as signatures of genetic isolation, reduced population size and consequently, endogamy within the archipelago. In this respect, although the ROH patterns suggested a tendency toward larger average total ROH length in the western and smaller islands of the archipelago, the mean number of genomic regions in ROHs significantly increased with geographical longitude (Spearman's rank correlation rho = 0.93, $P = 6.7\times10^{-3}$).

Local Ancestry: Assessing Estimators and the Average Size of Blocks

In cases of recent admixture, the genomes of admixed individuals become a mosaic of chromosomal stretches originating from the ancestral populations, which can be detected by examining locus-specific ancestry (usually termed local ancestry). Many local ancestry methods perform well in two-way admixture scenarios with large genetic differences between the ancestral populations (e.g., in African–Americans). However, the difficulty increases with the number of ancestral populations (e.g., three-way admixtures), particularly if a small to moderate degree of differentiation exists between some of the ancestral populations (e.g., between NAF and EUR). Given that this study constitutes the first time that the patterns of ancestry blocks in Canary Islander genomes have been assessed, we first evaluated two of the fastest and most

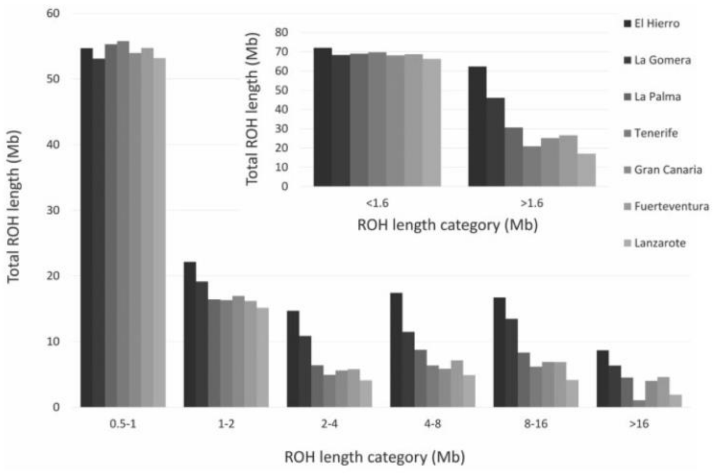


FIG. 3. Average length (Mb) of ROHs using two classifications into categories in the populations from the Canary Islands.

accurate methods of local ancestry estimation for multiway admixtures, ELAI and LAMP-LD, both of which have been fruitfully utilized with Hispanic populations (Baran et al. 2012; Zhou et al. 2016). We compared their results of these methods with those obtained previously by ADMIXTURE, pooling the two EUR-related ancestry fractions for the comparisons. The Pearson correlations between the estimates of ADMIXTURE and either ELAI or LAMP-LD were strongly significant ($P < 5.0\times10^{-10}$ in all cases) (table 2). However, ELAI offered better fitting estimates than LAMP-LD based on the overall estimates of NAF- and SSA-related ancestries obtained. Whereas the ELAI estimates were similar to those of ADMIXTURE (estimates for NAF and SSA of 23.6% and 2.3%, respectively), LAMP-LD provided an inflated estimate of NAF-related ancestry (estimates for NAF and SSA of 32.7 and 1.4%, respectively) (fig. 5). A least squares estimator indicated that individual fractions were between 2 (SSA) and 4 times (EUR and NAF) as different between ADMIXTURE and LAMP-LD than between ADMIXTURE and ELAI (not shown). In addition, whereas the lengths of the ancestry blocks provided by ELAI and LAMP-LD were correlated for the three ancestries (fig. 6), ELAI provided greater sensitivity for the detection of smaller stretches of ancestry. Based on these results, all further local ancestry analyses were conducted with ELAI estimates.

Block sizes followed lognormal distributions (supplementary fig. 4, Supplementary Material online), with the largest blocks on an average corresponding to the EUR-related component (13.05 Mb) followed by the NAF (8.46 Mb) and SSA (7.48 Mb) components, and with all pairwise comparisons yielding significant differences (Wilcoxon test $P < 3.0\times10^{-8}$). Based on these block length estimates and the ADMIXTURE proportions calculated for all Canary Islanders, we would expect averages of ∼181 EUR-, 82 NAF-, and 13 SSA-related blocks per haploid genome (i.e., 276 ancestry blocks in total), which are lower than but within the range of the estimates previously suggested for African–Americans (Shriner et al. 2011).

Deviations in Local Ancestry and Selection

Based on the previous results, we then used the ELAI estimates to assess the existence of regional deviations in any of the three ancestries in Canary Islanders. We detected eight peaks with large ancestry deviations: two enriched in EUR ancestry, and six associated with higher proportions of African ancestries (fig. 7, table 3, and supplementary fig. 5, Supplementary Material online). Notably, the EUR- and NAF-related peak positions largely overlapped. The EUR-related peaks were located on chr2 (with the lead SNP showing an average local ancestry of 55% and flanked by the *CXCR4* and *THSD7B* genes) and chr6 (with all lead SNPs located within *HLA-B*, with an average local ancestry of 49%). The two peaks of NAF-related ancestry enrichment were located on chr2 (with the lead SNP showing an average local ancestry of 43.2% and flanked by the same genes flanking the EUR-related peaks) and chr6 (with the lead SNP showing an average local ancestry of 46.3% and flanked by the *NFKBIL1* and *LTA* genes). Peaks associated with SSA-related ancestry were situated on chr3 (two hits: one with the lead SNP located within *SLC6A11*, the other with the lead SNP located within *KCNMB2*, both presenting an average local ancestry of ∼4.0%), chr6 (with the lead SNP within *ZNRD1*, showing an average local ancestry of 5.9%), and chr13 (with the lead SNP near *PCDH9* and flanked by two long intergenic noncoding RNA genes: *LINC00355* and *LINC01052*, showing an average local ancestry of 4.4%). These regions span several Mbs (range of 1.2–12.3 Mb), meaning that, while the observed ancestry deviations may indicate adaptive processes, any adaptation may be unrelated to the gene harboring the lead SNP. However, it is likely that the *LCT* and HLA loci caused the peaks on chr2 and chr6. One regulatory variant of *LCT* is the major determinant for lactase persistence in EUR populations

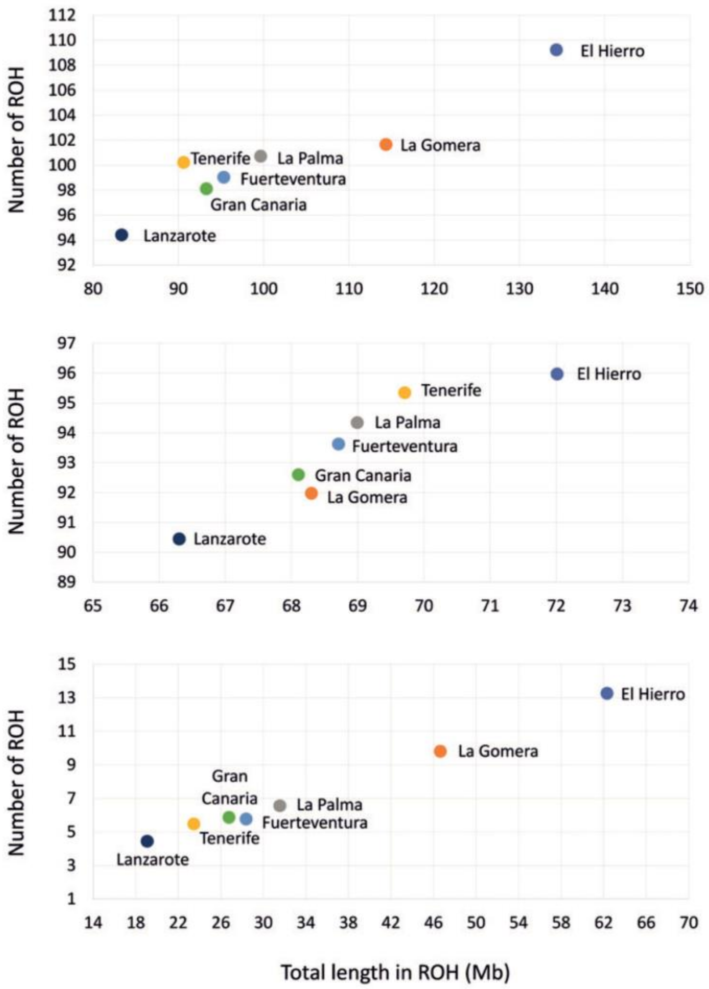


Fig. 4. Average number of genome regions in ROHs with respect to the average total (top), ≤1.6 Mb (middle), and >1.6 Mb (bottom) lengths of ROHs per island.

Table 2. Pearson Correlation between ADMIXTURE ($K = 4$) and ELAI or LAMP-LD Estimates of Ancestry in Canary Islanders.

	ELAI		LAMP-LD	
Ancestry	Coef.	95% CI	Coef.	95% CI
European	0.95	0.94–0.96	0.83	0.80–0.86
North African	0.88	0.85–0.90	0.74	0.68–0.78
Sub-Saharan African	0.93	0.92–0.94	0.89	0.87–0.91

and underlies the strongest signal of selection in the human genome identified to date (Mathieson et al. 2015). Similarly, this region has been shown to be under positive selection in Bantu-speaking populations (Patin et al. 2017). As for HLA, the gene-rich human leukocyte antigen region is also a well-known target of selection in EUR and African populations, likely involving balancing selection (Gurdasani et al. 2015; Mathieson et al. 2015; Patin et al. 2017) as well as directional selective processes, such as that described for *HLA-B* alleles and malaria protection in the Sahel (Sanchez-Mazas et al. 2017). Based on this evidence, we reasoned that all or some of the remaining regions with ancestry deviations detected (two on chr3 and one on chr13) could also have resulted from selective processes. One of the hallmarks of selection is increased homozygosity and reduced diversity in the surrounding regions (Pickrell et al. 2009). Therefore, we then assessed the mean heterozygosity and the number of samples with ROHs containing SNPs from the ancestry-enriched regions. We performed this assessment for all peaks associated with African ancestry (including those in chr2 and chr6) to obtain references for comparisons (table 4) and found that one of the regions on chr3 and the region on chr13 showed signals of reduced diversity among Canary Islanders. To formally test for the existence of selective signals in chr3 and chr13 loci, we assessed the Tajima's *D* and the Population Branch Statistic (PBS) in a subset of individuals for which WGS data was available. Deviations from neutrality at three locations were revealed by at least one test for the three loci, with frequent support for more than one hit obtained for some regions after accounting for the number of comparisons ($P < 0.017$ and $P < 0.01$ for Tajima's *D* and PBS, respectively, after Bonferroni correction). These three locations were as follows: in the intergenic region between the *SLC6A11* and *SLC6A1* genes on chr3 (Tajima's $D = -1.733$, $P < 0.061$; $\text{PBS} = 0.108$, $P < 2.0 \times 10^{-4}$), near *KCNMB2* on the same chromosome (Tajima's $D = -1.511$, $P < 0.130$; $\text{PBS} = 0.064$, $P < 2.0 \times 10^{-4}$), and in the intergenic region between *PCDH20* and *PCDH9* on chr13 (Tajima's $D = -2.505$, $\text{PBS} = 0.136$, $P < 2.0 \times 10^{-4}$ in both comparisons) (table 5). Interestingly, common genetic variants previously found to be associated with height (He et al. 2015), bone traits (Kiel et al. 2007), and asthma (Ferreira et al. 2011) reside in two of these regions (table 5). This observation is in agreement with the observation that variants for inflammatory diseases located via genome-wide association studies are significantly enriched in signatures of positive selection in European populations (Raj et al. 2013).

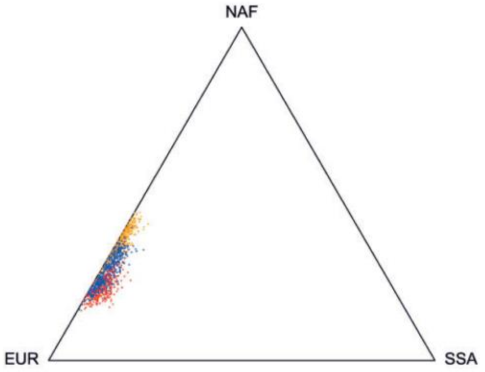


Fig. 5. Triangle plot of individual genomic admixture proportions in Canary Islanders as estimated by ADMIXTURE with $K = 4$ (red), ELAI (blue), and LAMP-LD (orange).

One striking observation relates to the chr2 peak, which indicates a higher proportion of NAF-related alleles in Canary Islanders in this region. Given that at least three other *LCT* variants have been linked with lactase persistence in other populations, we then explored whether any of those variants existed in this population and could offer a potential explanation for this peak. By accessing the WGS data available for a subset of 14 subjects, we found that the only detected variant

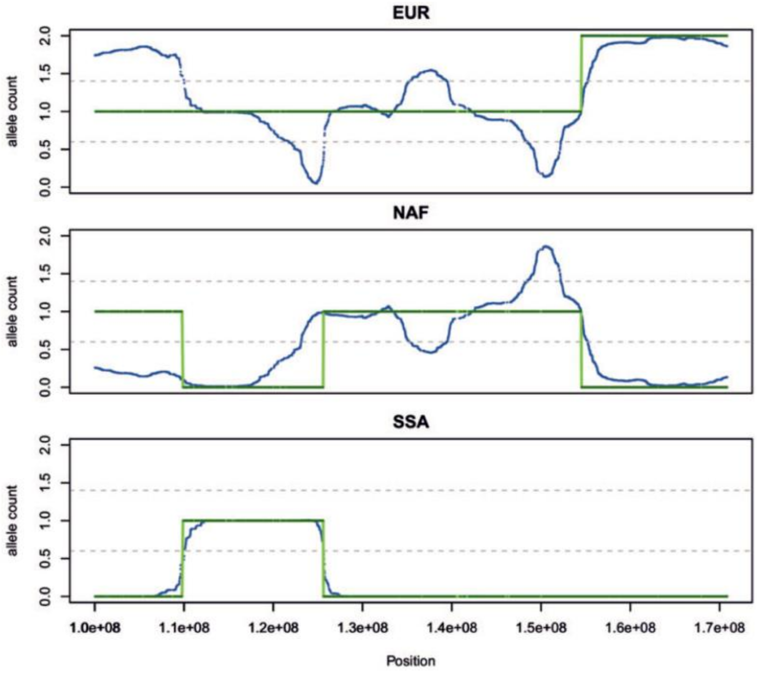


Fig. 6. Inference of local ancestry by ELAI and LAMP-LD. The plot shows an example of inference in a chromosome region (one panel of each parental population) comparing ELAI (blue) with LAMP-LD (green) allele dosages.

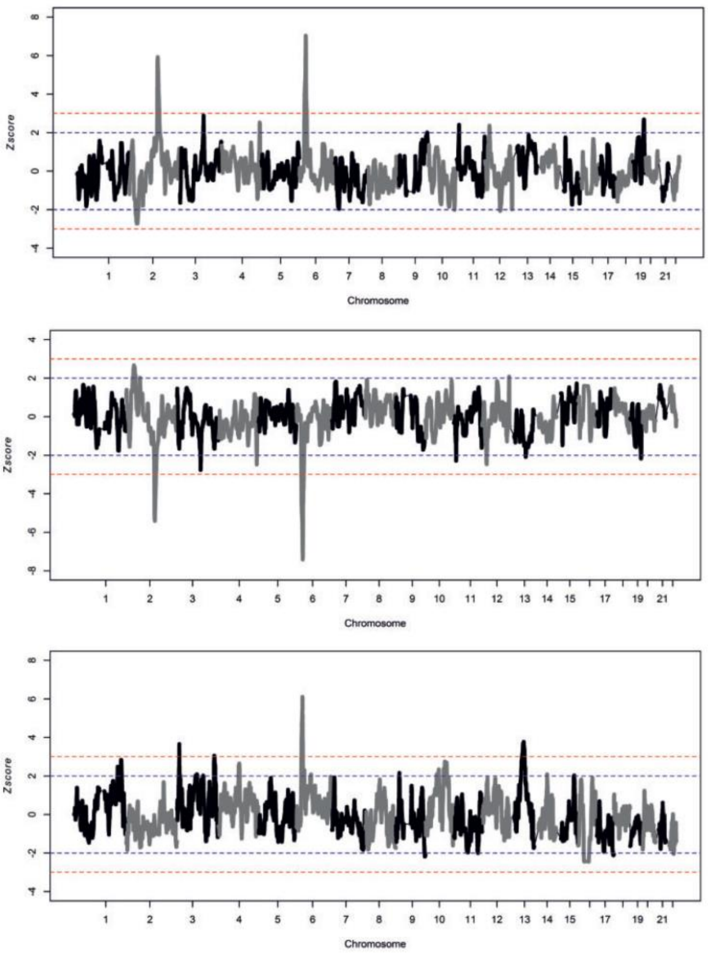


Fig. 7. Genome-wide Z-score scan of ELAI local admixtures in Canary Islanders for NAF (top), EUR (middle), and SSA (bottom). Horizontal broken lines (blue > |2|; red > |3|) indicate score thresholds.

Table 3. Genomic Regions with Supported Deviations in Ancestry among Canary Islanders.

Ancestry	Region	Lead SNP	Z Score	Mean Ancestry
North African	chr2: 133,952,040-144,266,489	rs10177911	5.92	43.2
	chr6: 24,703,442-36,288,651	rs2844484	7.03	46.3
European	chr2: 134,088,150-142,882,593	rs4954402	−5.41	55.0
	chr6: 24,120,456-36,653,597	(*)	−7.41	49.0
Sub–Saharan African	chr3: 10,539,482-11,710,471	rs17033567	3.65	4.3
	chr3: 177,443,968-178,679,751	rs13061192	3.02	3.9
	chr6: 24,790,462-32,192,083	rs16896944	6.10	5.9
	chr13: 57,962,413-70,091,195	rs9540226	3.77	4.4

(*) rs2524095, rs16899203, rs16899205, rs16899207, rs2524089, rs2394967, rs2524066, rs9366778.

position associated with lactase persistence was rs4988235 (a.k.a. -13,910), which is the major determinant for lactase persistence in Europe. The frequency of this lactase persistence allele in Canary Islanders (-13,910*T, 40%) is intermediate between the frequency reported in other central and northern EUR populations (60–80%) and that in NAF populations (24%) (Bersaglieri et al. 2004; Ben Halima et al. 2017) (http://www.ucl.ac.uk/mace-lab/resources/glad).

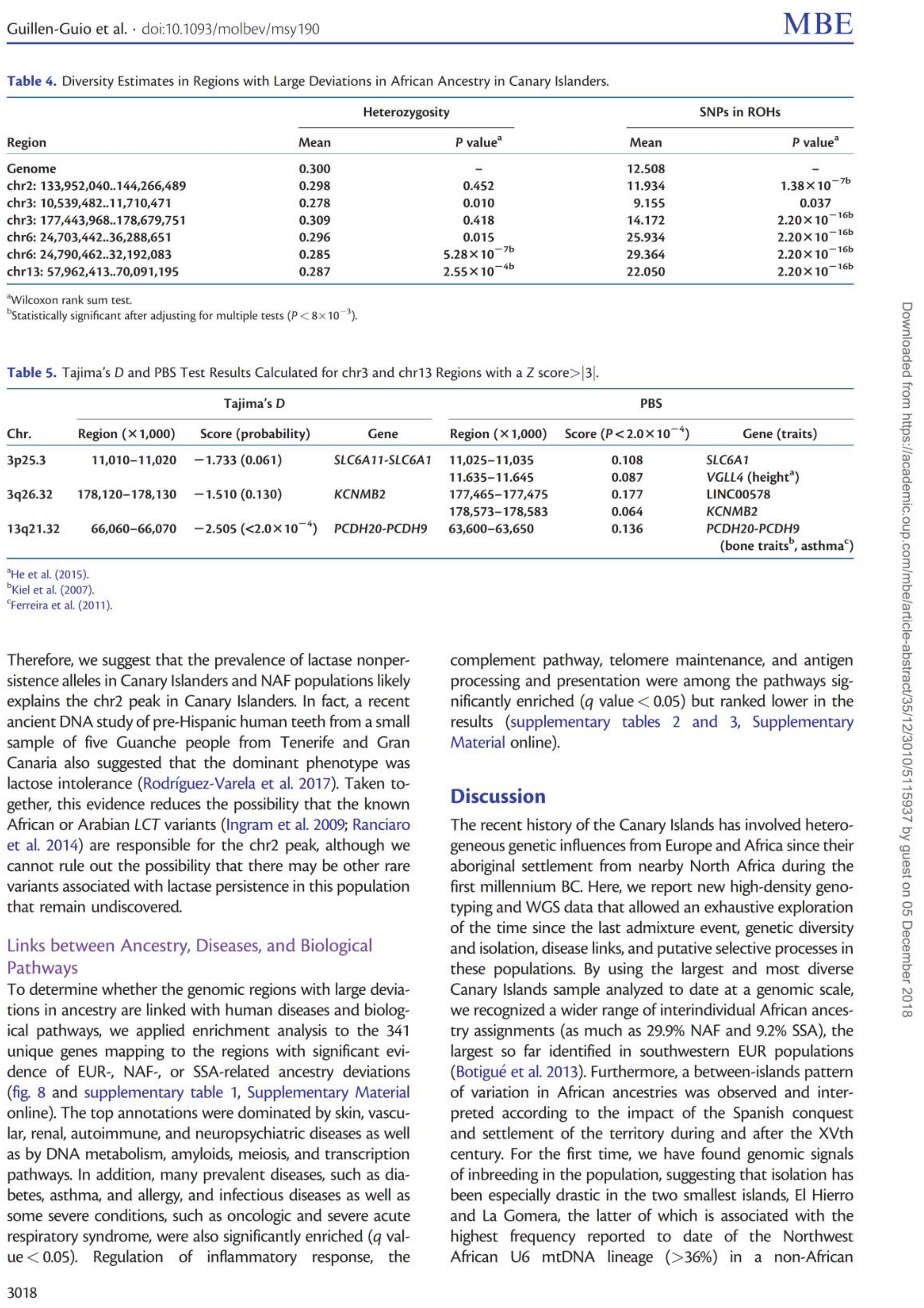


Table 4. Diversity Estimates in Regions with Large Deviations in African Ancestry in Canary Islanders.

	Heterozygosity		SNPs in ROHs	
Region	Mean	*P* value[a]	Mean	*P* value[a]
Genome	0.300	–	12.508	–
chr2: 133,952,040..144,266,489	0.298	0.452	11.934	1.38×10^{-7b}
chr3: 10,539,482..11,710,471	0.278	0.010	9.155	0.037
chr3: 177,443,968..178,679,751	0.309	0.418	14.172	2.20×10^{-16b}
chr6: 24,703,442..36,288,651	0.296	0.015	25.934	2.20×10^{-16b}
chr6: 24,790,462..32,192,083	0.285	5.28×10^{-7b}	29.364	2.20×10^{-16b}
chr13: 57,962,413..70,091,195	0.287	2.55×10^{-4b}	22.050	2.20×10^{-16b}

[a] Wilcoxon rank sum test.
[b] Statistically significant after adjusting for multiple tests ($P < 8\times10^{-3}$).

Table 5. Tajima's *D* and PBS Test Results Calculated for chr3 and chr13 Regions with a *Z* score>|3|.

	Tajima's *D*			PBS		
Chr.	Region (×1,000)	Score (probability)	Gene	Region (×1,000)	Score ($P < 2.0\times10^{-4}$)	Gene (traits)
3p25.3	11,010–11,020	−1.733 (0.061)	*SLC6A11-SLC6A1*	11,025–11,035	0.108	*SLC6A1*
				11.635–11.645	0.087	*VGLL4* (height[a])
3q26.32	178,120–178,130	−1.510 (0.130)	*KCNMB2*	177,465–177,475	0.177	LINC00578
				178,573–178,583	0.064	*KCNMB2*
13q21.32	66,060–66,070	−2.505 (<2.0×10^{-4})	*PCDH20-PCDH9*	63,600–63,650	0.136	*PCDH20-PCDH9* (bone traits[b], asthma[c])

[a] He et al. (2015).
[b] Kiel et al. (2007).
[c] Ferreira et al. (2011).

Therefore, we suggest that the prevalence of lactase nonpersistence alleles in Canary Islanders and NAF populations likely explains the chr2 peak in Canary Islanders. In fact, a recent ancient DNA study of pre-Hispanic human teeth from a small sample of five Guanche people from Tenerife and Gran Canaria also suggested that the dominant phenotype was lactose intolerance (Rodríguez-Varela et al. 2017). Taken together, this evidence reduces the possibility that the known African or Arabian *LCT* variants (Ingram et al. 2009; Ranciaro et al. 2014) are responsible for the chr2 peak, although we cannot rule out the possibility that there may be other rare variants associated with lactase persistence in this population that remain undiscovered.

Links between Ancestry, Diseases, and Biological Pathways

To determine whether the genomic regions with large deviations in ancestry are linked with human diseases and biological pathways, we applied enrichment analysis to the 341 unique genes mapping to the regions with significant evidence of EUR-, NAF-, or SSA-related ancestry deviations (fig. 8 and supplementary table 1, Supplementary Material online). The top annotations were dominated by skin, vascular, renal, autoimmune, and neuropsychiatric diseases as well as by DNA metabolism, amyloids, meiosis, and transcription pathways. In addition, many prevalent diseases, such as diabetes, asthma, and allergy, and infectious diseases as well as some severe conditions, such as oncologic and severe acute respiratory syndrome, were also significantly enriched (*q* value < 0.05). Regulation of inflammatory response, the complement pathway, telomere maintenance, and antigen processing and presentation were among the pathways significantly enriched (*q* value < 0.05) but ranked lower in the results (supplementary tables 2 and 3, Supplementary Material online).

Discussion

The recent history of the Canary Islands has involved heterogeneous genetic influences from Europe and Africa since their aboriginal settlement from nearby North Africa during the first millennium BC. Here, we report new high-density genotyping and WGS data that allowed an exhaustive exploration of the time since the last admixture event, genetic diversity and isolation, disease links, and putative selective processes in these populations. By using the largest and most diverse Canary Islands sample analyzed to date at a genomic scale, we recognized a wider range of interindividual African ancestry assignments (as much as 29.9% NAF and 9.2% SSA), the largest so far identified in southwestern EUR populations (Botigué et al. 2013). Furthermore, a between-islands pattern of variation in African ancestries was observed and interpreted according to the impact of the Spanish conquest and settlement of the territory during and after the XVth century. For the first time, we have found genomic signals of inbreeding in the population, suggesting that isolation has been especially drastic in the two smallest islands, El Hierro and La Gomera, the latter of which is associated with the highest frequency reported to date of the Northwest African U6 mtDNA lineage (>36%) in a non-African

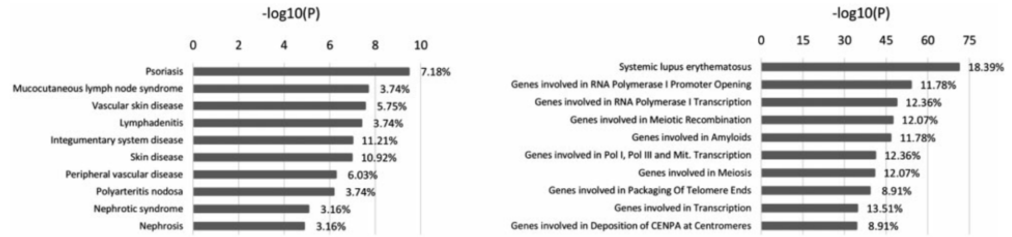


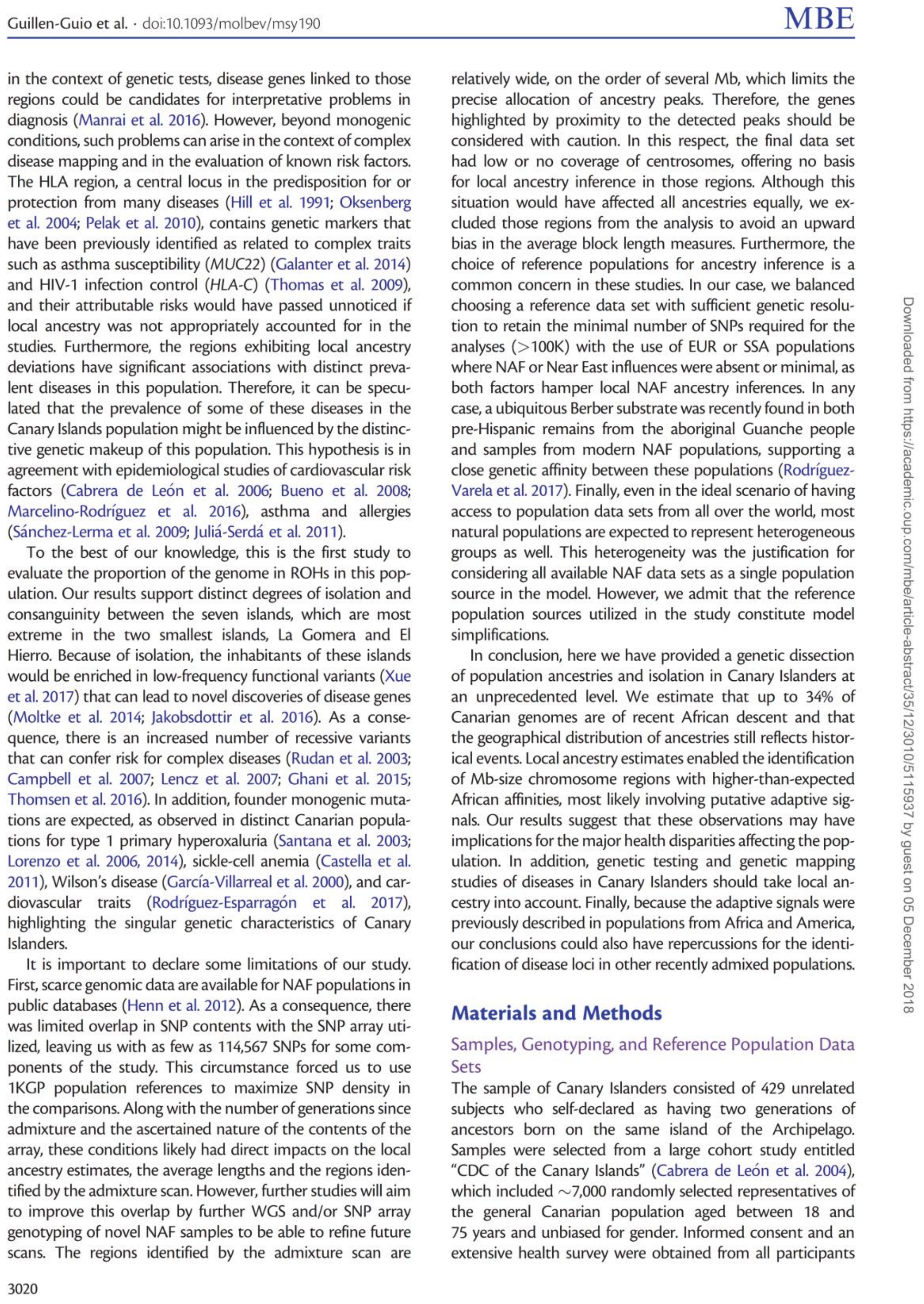
Fig. 8. Enrichment analysis on regions with large deviations for any ancestry. Top ten significantly enriched human diseases (left) and MSigDB pathways (right).

population (Rando et al. 1999). In addition, we uncovered the mosaic nature of Canary Islander genomes, finding that a typical haploid genome would consist of fewer than 300 ancestry blocks and the number of EUR-related blocks would be approximately twice and 10 times that of the NAF- and SSA-related blocks, respectively. Finally, we used this information to focus on particular chromosome regions that showed an overrepresentation of one of the recognized ancestries. These regions included the *LCT* gene, the HLA, and other genes that appear to have undergone selective processes. Most importantly, chromosome regions with large deviations in the ancestries were enriched in genes underlying important diseases that disproportionately affect the archipelago, including respiratory diseases and diabetes, among others.

Despite the profound impact of the European immigration starting in the XVth century, our results indicate that significant genetic footprints of African ancestry persist in the current inhabitants of the Canary Islands. Our data also imply that gene flow between the island populations must have been high to maintain the relative homogeneity observed for the African ancestries, despite the significant discontinuity favoring African affinities in the populations of El Hierro, La Gomera, Fuerteventura, and Lanzarote. Strikingly, such a pattern was also revealed in mtDNA studies of independent samples (Rando et al. 1999), suggesting that African ancestry has been better preserved in the populations that were conquered at the beginning of the historic process, where the literature attests to a more peaceful occupation of those territories (Suárez et al. 1988). In contrast to our previous study, which was limited by the number of autosomal markers utilized at the time (Pino-Yanes et al. 2011), in the present study, we were able to measure the existence of a minor but substantial SSA-related influence in the populations of the seven islands. SSA migrations into NAF populations are well known (Rando et al. 1998; Wright 2007), suggesting that continuous gene flow started before the aboriginal settlement of the islands. In addition, our results evidenced significant differences in the mean ancestry block lengths interpreted as of recent SSA and NAF origin in Canary Islanders. While this result may be indicative of admixture events involving different African components, as has been supported by many previous studies (Arco and Navarro 1987; de Castro 1987; Navarro Mederos 1997; Flores et al. 2003), independent analyses should aim to establish the time periods of such events. For example, in a recent genomic study of ancient DNA, a very low proportion of SSA ancestry was observed in pre-Hispanic human remains despite their genetic resemblance to other modern NAF populations (Rodríguez-Varela et al. 2017). This finding suggests that the proportion of SSA ancestry we observed in Canary Islanders likely originated in the postconquest importation of enslaved African people.

Previous genetic studies of pre-Hispanic remains, burial remains from the XVIIth century, and samples from current inhabitants indicate declining frequencies of the NRY lineages ascribed to aboriginal people (Fregel et al. 2009). However, a continuity of aboriginal mtDNA lineages, possibly evolving locally into Canarian-specific U6b subhaplogroups (Rando et al. 1999; Maca-Meyer et al. 2004), has been evidenced in the aboriginal remains and current inhabitants, suggesting an important sexual asymmetry involved in the demographic process of the population (Flores et al. 2001). In this study, we found that five other regions in the autosomal genome, including the HLA and *LCT*, paralleling this scenario, where genetic affinities with African populations remain regionally preserved today. Independent studies and our own data support that putative selective processes are the cause of such autosomal affinities. However, given that the bulk of human adaptation occurs via selection on standing variation (Schrider and Kern 2017), it is possible that these signals are not due to private alleles but to preexisting variations in any of the parental populations involved in the admixture. Indeed, extensive evidence indicates pervasive selection processes in the HLA and *LCT* regions in different populations; these processes have been explained by dietary and cultural changes occurring with the Neolithic dispersion, including population density increases and the establishment of permanent settlements, resulting in novel pathogen exposures (Marciniak and Perry 2017).

Additionally, the *SLC6A11* and *KCNMB2* genes have been suggested to be under selective pressures in African and American populations (Hu 2012; Watkins et al. 2012), and literature evidence supports *DIAPH3*, not *PCDH9*, as the closest candidate target for selection in the chr13 region (Pickrell et al. 2009). One important consequence of this observation is that those regions contain genes in which mutations are linked to rare disorders that are usually screened in families with genetic risks, such as *PCDH9*, *SLC6A11*, *SLC6A1*, and *COL11A2*. Therefore, as local ancestry will act as a confounder



in the context of genetic tests, disease genes linked to those regions could be candidates for interpretative problems in diagnosis (Manrai et al. 2016). However, beyond monogenic conditions, such problems can arise in the context of complex disease mapping and in the evaluation of known risk factors. The HLA region, a central locus in the predisposition for or protection from many diseases (Hill et al. 1991; Oksenberg et al. 2004; Pelak et al. 2010), contains genetic markers that have been previously identified as related to complex traits such as asthma susceptibility (*MUC22*) (Galanter et al. 2014) and HIV-1 infection control (*HLA-C*) (Thomas et al. 2009), and their attributable risks would have passed unnoticed if local ancestry was not appropriately accounted for in the studies. Furthermore, the regions exhibiting local ancestry deviations have significant associations with distinct prevalent diseases in this population. Therefore, it can be speculated that the prevalence of some of these diseases in the Canary Islands population might be influenced by the distinctive genetic makeup of this population. This hypothesis is in agreement with epidemiological studies of cardiovascular risk factors (Cabrera de León et al. 2006; Bueno et al. 2008; Marcelino-Rodríguez et al. 2016), asthma and allergies (Sánchez-Lerma et al. 2009; Juliá-Serdá et al. 2011).

To the best of our knowledge, this is the first study to evaluate the proportion of the genome in ROHs in this population. Our results support distinct degrees of isolation and consanguinity between the seven islands, which are most extreme in the two smallest islands, La Gomera and El Hierro. Because of isolation, the inhabitants of these islands would be enriched in low-frequency functional variants (Xue et al. 2017) that can lead to novel discoveries of disease genes (Moltke et al. 2014; Jakobsdottir et al. 2016). As a consequence, there is an increased number of recessive variants that can confer risk for complex diseases (Rudan et al. 2003; Campbell et al. 2007; Lencz et al. 2007; Ghani et al. 2015; Thomsen et al. 2016). In addition, founder monogenic mutations are expected, as observed in distinct Canarian populations for type 1 primary hyperoxaluria (Santana et al. 2003; Lorenzo et al. 2006, 2014), sickle-cell anemia (Castella et al. 2011), Wilson's disease (García-Villarreal et al. 2000), and cardiovascular traits (Rodríguez-Esparragón et al. 2017), highlighting the singular genetic characteristics of Canary Islanders.

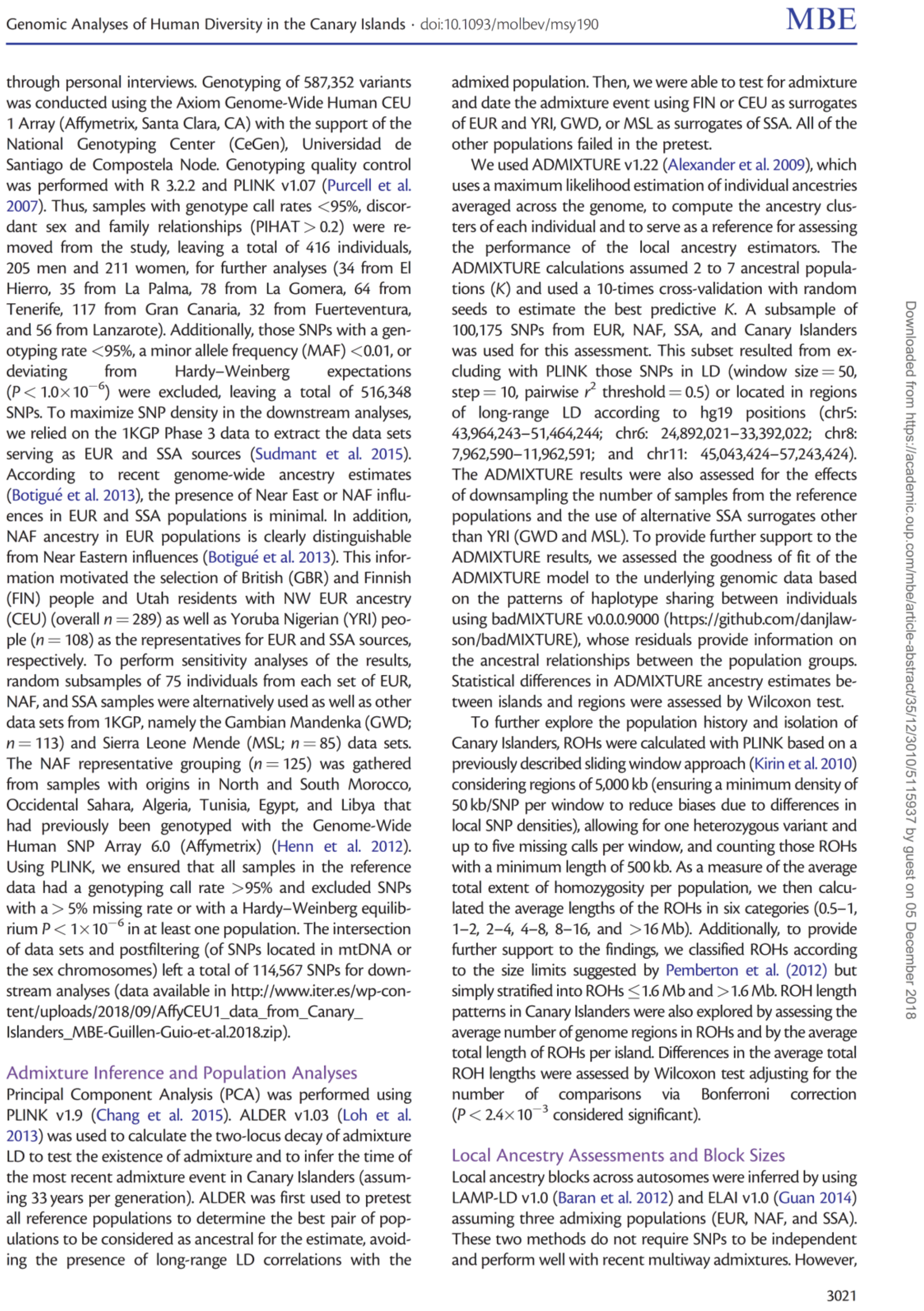
It is important to declare some limitations of our study. First, scarce genomic data are available for NAF populations in public databases (Henn et al. 2012). As a consequence, there was limited overlap in SNP contents with the SNP array utilized, leaving us with as few as 114,567 SNPs for some components of the study. This circumstance forced us to use 1KGP population references to maximize SNP density in the comparisons. Along with the number of generations since admixture and the ascertained nature of the contents of the array, these conditions likely had direct impacts on the local ancestry estimates, the average lengths and the regions identified by the admixture scan. However, further studies will aim to improve this overlap by further WGS and/or SNP array genotyping of novel NAF samples to be able to refine future scans. The regions identified by the admixture scan are relatively wide, on the order of several Mb, which limits the precise allocation of ancestry peaks. Therefore, the genes highlighted by proximity to the detected peaks should be considered with caution. In this respect, the final data set had low or no coverage of centrosomes, offering no basis for local ancestry inference in those regions. Although this situation would have affected all ancestries equally, we excluded those regions from the analysis to avoid an upward bias in the average block length measures. Furthermore, the choice of reference populations for ancestry inference is a common concern in these studies. In our case, we balanced choosing a reference data set with sufficient genetic resolution to retain the minimal number of SNPs required for the analyses (>100K) with the use of EUR or SSA populations where NAF or Near East influences were absent or minimal, as both factors hamper local NAF ancestry inferences. In any case, a ubiquitous Berber substrate was recently found in both pre-Hispanic remains from the aboriginal Guanche people and samples from modern NAF populations, supporting a close genetic affinity between these populations (Rodríguez-Varela et al. 2017). Finally, even in the ideal scenario of having access to population data sets from all over the world, most natural populations are expected to represent heterogeneous groups as well. This heterogeneity was the justification for considering all available NAF data sets as a single population source in the model. However, we admit that the reference population sources utilized in the study constitute model simplifications.

In conclusion, here we have provided a genetic dissection of population ancestries and isolation in Canary Islanders at an unprecedented level. We estimate that up to 34% of Canarian genomes are of recent African descent and that the geographical distribution of ancestries still reflects historical events. Local ancestry estimates enabled the identification of Mb-size chromosome regions with higher-than-expected African affinities, most likely involving putative adaptive signals. Our results suggest that these observations may have implications for the major health disparities affecting the population. In addition, genetic testing and genetic mapping studies of diseases in Canary Islanders should take local ancestry into account. Finally, because the adaptive signals were previously described in populations from Africa and America, our conclusions could also have repercussions for the identification of disease loci in other recently admixed populations.

Materials and Methods

Samples, Genotyping, and Reference Population Data Sets

The sample of Canary Islanders consisted of 429 unrelated subjects who self-declared as having two generations of ancestors born on the same island of the Archipelago. Samples were selected from a large cohort study entitled "CDC of the Canary Islands" (Cabrera de León et al. 2004), which included ~7,000 randomly selected representatives of the general Canarian population aged between 18 and 75 years and unbiased for gender. Informed consent and an extensive health survey were obtained from all participants



through personal interviews. Genotyping of 587,352 variants was conducted using the Axiom Genome-Wide Human CEU 1 Array (Affymetrix, Santa Clara, CA) with the support of the National Genotyping Center (CeGen), Universidad de Santiago de Compostela Node. Genotyping quality control was performed with R 3.2.2 and PLINK v1.07 (Purcell et al. 2007). Thus, samples with genotype call rates <95%, discordant sex and family relationships (PIHAT > 0.2) were removed from the study, leaving a total of 416 individuals, 205 men and 211 women, for further analyses (34 from El Hierro, 35 from La Palma, 78 from La Gomera, 64 from Tenerife, 117 from Gran Canaria, 32 from Fuerteventura, and 56 from Lanzarote). Additionally, those SNPs with a genotyping rate <95%, a minor allele frequency (MAF) <0.01, or deviating from Hardy–Weinberg expectations ($P < 1.0 \times 10^{-6}$) were excluded, leaving a total of 516,348 SNPs. To maximize SNP density in the downstream analyses, we relied on the 1KGP Phase 3 data to extract the data sets serving as EUR and SSA sources (Sudmant et al. 2015). According to recent genome-wide ancestry estimates (Botigué et al. 2013), the presence of Near East or NAF influences in EUR and SSA populations is minimal. In addition, NAF ancestry in EUR populations is clearly distinguishable from Near Eastern influences (Botigué et al. 2013). This information motivated the selection of British (GBR) and Finnish (FIN) people and Utah residents with NW EUR ancestry (CEU) (overall $n = 289$) as well as Yoruba Nigerian (YRI) people ($n = 108$) as the representatives for EUR and SSA sources, respectively. To perform sensitivity analyses of the results, random subsamples of 75 individuals from each set of EUR, NAF, and SSA samples were alternatively used as well as other data sets from 1KGP, namely the Gambian Mandenka (GWD; $n = 113$) and Sierra Leone Mende (MSL; $n = 85$) data sets. The NAF representative grouping ($n = 125$) was gathered from samples with origins in North and South Morocco, Occidental Sahara, Algeria, Tunisia, Egypt, and Libya that had previously been genotyped with the Genome-Wide Human SNP Array 6.0 (Affymetrix) (Henn et al. 2012). Using PLINK, we ensured that all samples in the reference data had a genotyping call rate >95% and excluded SNPs with a > 5% missing rate or with a Hardy–Weinberg equilibrium $P < 1 \times 10^{-6}$ in at least one population. The intersection of data sets and postfiltering (of SNPs located in mtDNA or the sex chromosomes) left a total of 114,567 SNPs for downstream analyses (data available in http://www.iter.es/wp-content/uploads/2018/09/AffyCEU1_data_from_Canary_Islanders_MBE-Guillen-Guio-et-al.2018.zip).

Admixture Inference and Population Analyses

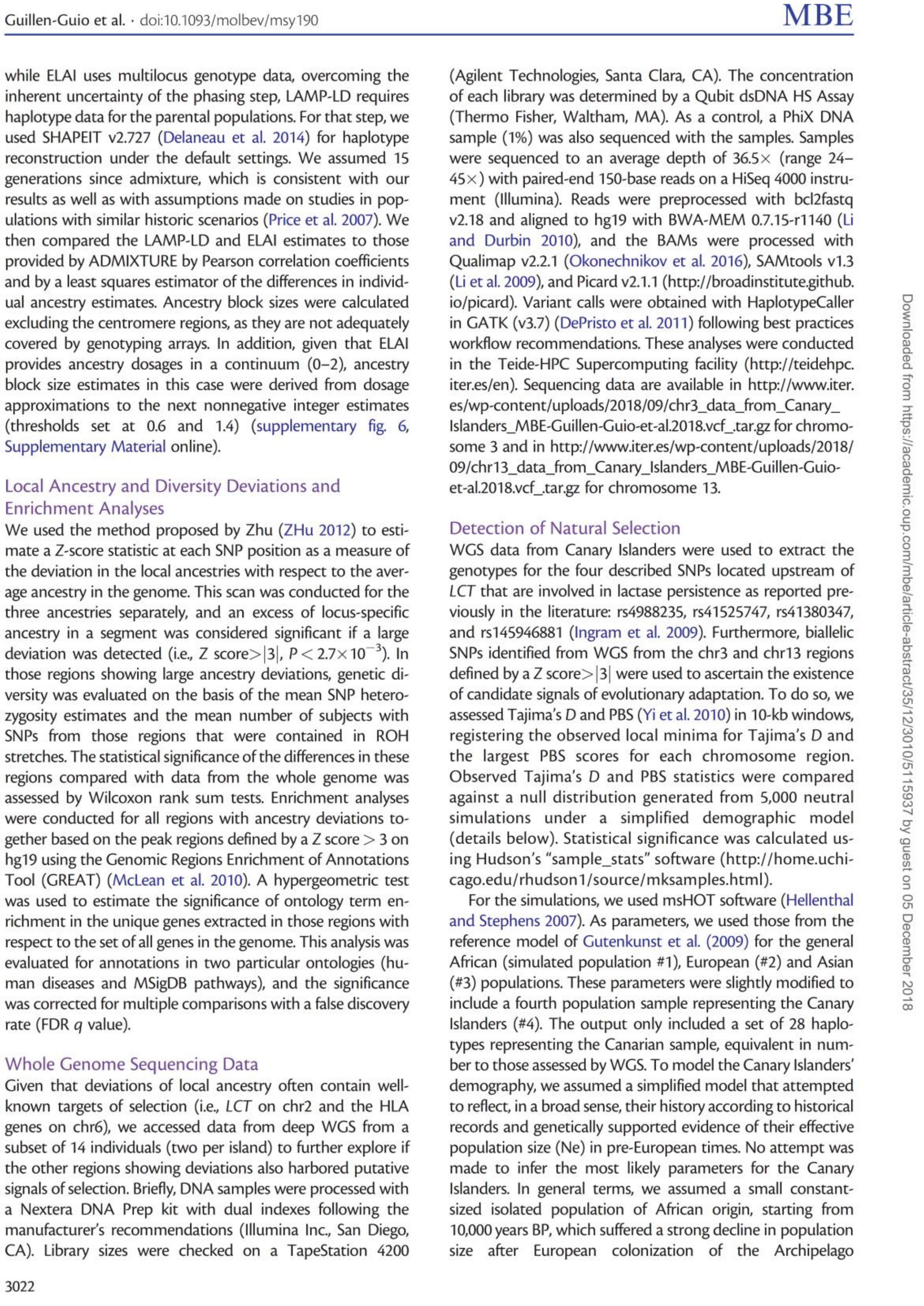
Principal Component Analysis (PCA) was performed using PLINK v1.9 (Chang et al. 2015). ALDER v1.03 (Loh et al. 2013) was used to calculate the two-locus decay of admixture LD to test the existence of admixture and to infer the time of the most recent admixture event in Canary Islanders (assuming 33 years per generation). ALDER was first used to pretest all reference populations to determine the best pair of populations to be considered as ancestral for the estimate, avoiding the presence of long-range LD correlations with the admixed population. Then, we were able to test for admixture and date the admixture event using FIN or CEU as surrogates of EUR and YRI, GWD, or MSL as surrogates of SSA. All of the other populations failed in the pretest.

We used ADMIXTURE v1.22 (Alexander et al. 2009), which uses a maximum likelihood estimation of individual ancestries averaged across the genome, to compute the ancestry clusters of each individual and to serve as a reference for assessing the performance of the local ancestry estimators. The ADMIXTURE calculations assumed 2 to 7 ancestral populations (*K*) and used a 10-times cross-validation with random seeds to estimate the best predictive *K*. A subsample of 100,175 SNPs from EUR, NAF, SSA, and Canary Islanders was used for this assessment. This subset resulted from excluding with PLINK those SNPs in LD (window size = 50, step = 10, pairwise r^2 threshold = 0.5) or located in regions of long-range LD according to hg19 positions (chr5: 43,964,243–51,464,244; chr6: 24,892,021–33,392,022; chr8: 7,962,590–11,962,591; and chr11: 45,043,424–57,243,424). The ADMIXTURE results were also assessed for the effects of downsampling the number of samples from the reference populations and the use of alternative SSA surrogates other than YRI (GWD and MSL). To provide further support to the ADMIXTURE results, we assessed the goodness of fit of the ADMIXTURE model to the underlying genomic data based on the patterns of haplotype sharing between individuals using badMIXTURE v0.0.0.9000 (https://github.com/danjlawson/badMIXTURE), whose residuals provide information on the ancestral relationships between the population groups. Statistical differences in ADMIXTURE ancestry estimates between islands and regions were assessed by Wilcoxon test.

To further explore the population history and isolation of Canary Islanders, ROHs were calculated with PLINK based on a previously described sliding window approach (Kirin et al. 2010) considering regions of 5,000 kb (ensuring a minimum density of 50 kb/SNP per window to reduce biases due to differences in local SNP densities), allowing for one heterozygous variant and up to five missing calls per window, and counting those ROHs with a minimum length of 500 kb. As a measure of the average total extent of homozygosity per population, we then calculated the average lengths of the ROHs in six categories (0.5–1, 1–2, 2–4, 4–8, 8–16, and >16 Mb). Additionally, to provide further support to the findings, we classified ROHs according to the size limits suggested by Pemberton et al. (2012) but simply stratified into ROHs ≤1.6 Mb and >1.6 Mb. ROH length patterns in Canary Islanders were also explored by assessing the average number of genome regions in ROHs and by the average total length of ROHs per island. Differences in the average total ROH lengths were assessed by Wilcoxon test adjusting for the number of comparisons via Bonferroni correction ($P < 2.4 \times 10^{-3}$ considered significant).

Local Ancestry Assessments and Block Sizes

Local ancestry blocks across autosomes were inferred by using LAMP-LD v1.0 (Baran et al. 2012) and ELAI v1.0 (Guan 2014) assuming three admixing populations (EUR, NAF, and SSA). These two methods do not require SNPs to be independent and perform well with recent multiway admixtures. However,



while ELAI uses multilocus genotype data, overcoming the inherent uncertainty of the phasing step, LAMP-LD requires haplotype data for the parental populations. For that step, we used SHAPEIT v2.727 (Delaneau et al. 2014) for haplotype reconstruction under the default settings. We assumed 15 generations since admixture, which is consistent with our results as well as with assumptions made on studies in populations with similar historic scenarios (Price et al. 2007). We then compared the LAMP-LD and ELAI estimates to those provided by ADMIXTURE by Pearson correlation coefficients and by a least squares estimator of the differences in individual ancestry estimates. Ancestry block sizes were calculated excluding the centromere regions, as they are not adequately covered by genotyping arrays. In addition, given that ELAI provides ancestry dosages in a continuum (0–2), ancestry block size estimates in this case were derived from dosage approximations to the next nonnegative integer estimates (thresholds set at 0.6 and 1.4) (supplementary fig. 6, Supplementary Material online).

Local Ancestry and Diversity Deviations and Enrichment Analyses

We used the method proposed by Zhu (ZHu 2012) to estimate a *Z*-score statistic at each SNP position as a measure of the deviation in the local ancestries with respect to the average ancestry in the genome. This scan was conducted for the three ancestries separately, and an excess of locus-specific ancestry in a segment was considered significant if a large deviation was detected (i.e., Z score$>|3|$, $P<2.7\times10^{-3}$). In those regions showing large ancestry deviations, genetic diversity was evaluated on the basis of the mean SNP heterozygosity estimates and the mean number of subjects with SNPs from those regions that were contained in ROH stretches. The statistical significance of the differences in these regions compared with data from the whole genome was assessed by Wilcoxon rank sum tests. Enrichment analyses were conducted for all regions with ancestry deviations together based on the peak regions defined by a Z score >3 on hg19 using the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al. 2010). A hypergeometric test was used to estimate the significance of ontology term enrichment in the unique genes extracted in those regions with respect to the set of all genes in the genome. This analysis was evaluated for annotations in two particular ontologies (human diseases and MSigDB pathways), and the significance was corrected for multiple comparisons with a false discovery rate (FDR *q* value).

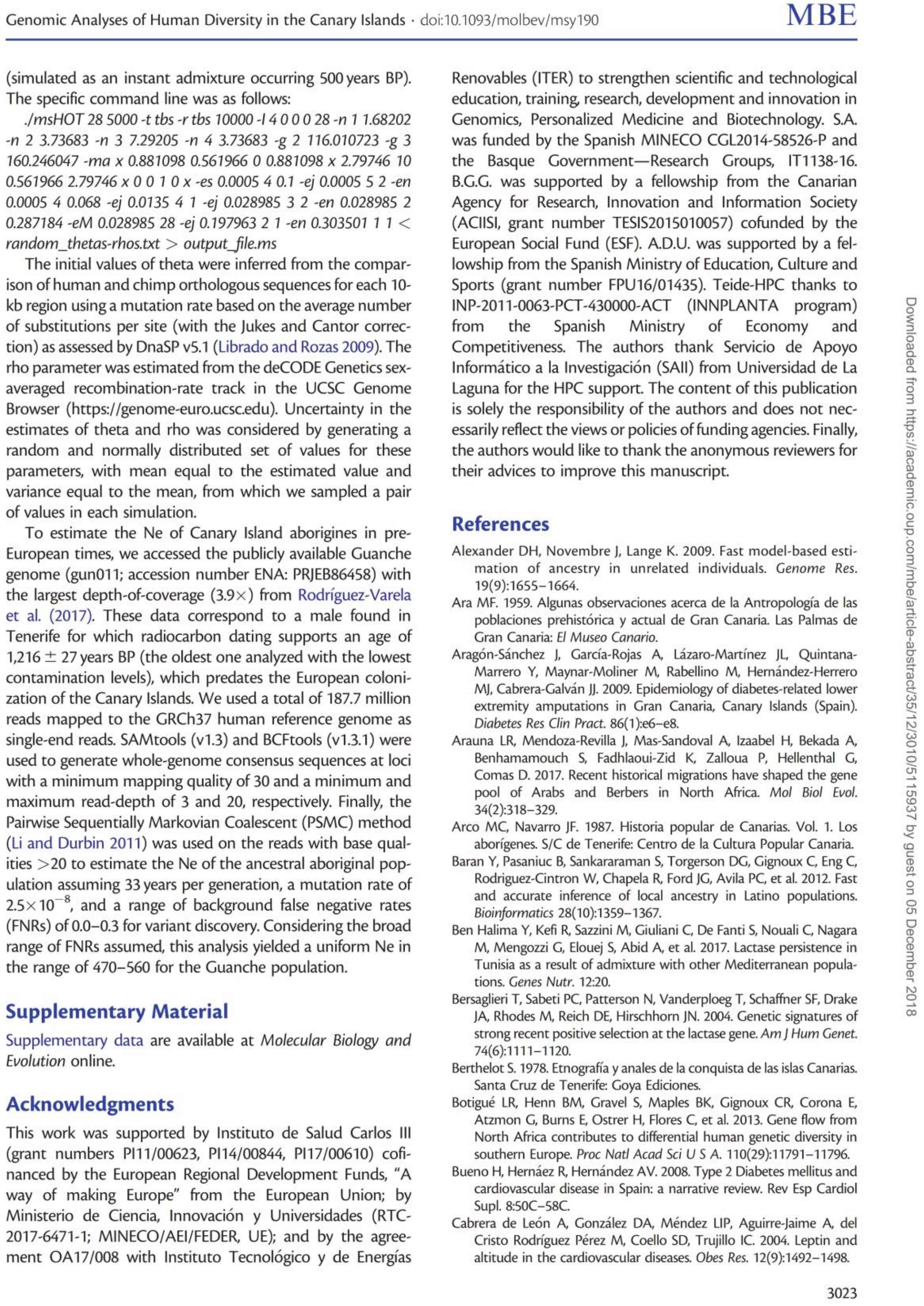
Whole Genome Sequencing Data

Given that deviations of local ancestry often contain well-known targets of selection (i.e., *LCT* on chr2 and the HLA genes on chr6), we accessed data from deep WGS from a subset of 14 individuals (two per island) to further explore if the other regions showing deviations also harbored putative signals of selection. Briefly, DNA samples were processed with a Nextera DNA Prep kit with dual indexes following the manufacturer's recommendations (Illumina Inc., San Diego, CA). Library sizes were checked on a TapeStation 4200 (Agilent Technologies, Santa Clara, CA). The concentration of each library was determined by a Qubit dsDNA HS Assay (Thermo Fisher, Waltham, MA). As a control, a PhiX DNA sample (1%) was also sequenced with the samples. Samples were sequenced to an average depth of 36.5× (range 24–45×) with paired-end 150-base reads on a HiSeq 4000 instrument (Illumina). Reads were preprocessed with bcl2fastq v2.18 and aligned to hg19 with BWA-MEM 0.7.15-r1140 (Li and Durbin 2010), and the BAMs were processed with Qualimap v2.2.1 (Okonechnikov et al. 2016), SAMtools v1.3 (Li et al. 2009), and Picard v2.1.1 (http://broadinstitute.github.io/picard). Variant calls were obtained with HaplotypeCaller in GATK (v3.7) (DePristo et al. 2011) following best practices workflow recommendations. These analyses were conducted in the Teide-HPC Supercomputing facility (http://teidehpc.iter.es/en). Sequencing data are available in http://www.iter.es/wp-content/uploads/2018/09/chr3_data_from_Canary_Islanders_MBE-Guillen-Guio-et-al.2018.vcf_.tar.gz for chromosome 3 and in http://www.iter.es/wp-content/uploads/2018/09/chr13_data_from_Canary_Islanders_MBE-Guillen-Guio-et-al.2018.vcf_.tar.gz for chromosome 13.

Detection of Natural Selection

WGS data from Canary Islanders were used to extract the genotypes for the four described SNPs located upstream of *LCT* that are involved in lactase persistence as reported previously in the literature: rs4988235, rs41525747, rs41380347, and rs145946881 (Ingram et al. 2009). Furthermore, biallelic SNPs identified from WGS from the chr3 and chr13 regions defined by a Z score$>|3|$ were used to ascertain the existence of candidate signals of evolutionary adaptation. To do so, we assessed Tajima's *D* and PBS (Yi et al. 2010) in 10-kb windows, registering the observed local minima for Tajima's *D* and the largest PBS scores for each chromosome region. Observed Tajima's *D* and PBS statistics were compared against a null distribution generated from 5,000 neutral simulations under a simplified demographic model (details below). Statistical significance was calculated using Hudson's "sample_stats" software (http://home.uchicago.edu/rhudson1/source/mksamples.html).

For the simulations, we used msHOT software (Hellenthal and Stephens 2007). As parameters, we used those from the reference model of Gutenkunst et al. (2009) for the general African (simulated population #1), European (#2) and Asian (#3) populations. These parameters were slightly modified to include a fourth population sample representing the Canary Islanders (#4). The output only included a set of 28 haplotypes representing the Canarian sample, equivalent in number to those assessed by WGS. To model the Canary Islanders' demography, we assumed a simplified model that attempted to reflect, in a broad sense, their history according to historical records and genetically supported evidence of their effective population size (Ne) in pre-European times. No attempt was made to infer the most likely parameters for the Canary Islanders. In general terms, we assumed a small constant-sized isolated population of African origin, starting from 10,000 years BP, which suffered a strong decline in population size after European colonization of the Archipelago



(simulated as an instant admixture occurring 500 years BP). The specific command line was as follows:

```
./msHOT 28 5000 -t tbs -r tbs 10000 -I 4 0 0 0 28 -n 1 1.68202
-n 2 3.73683 -n 3 7.29205 -n 4 3.73683 -g 2 116.010723 -g 3
160.246047 -ma x 0.881098 0.561966 0 0.881098 x 2.79746 10
0.561966 2.79746 x 0 0 1 0 x -es 0.0005 4 0.1 -ej 0.0005 5 2 -en
0.0005 4 0.068 -ej 0.0135 4 1 -ej 0.028985 3 2 -en 0.028985 2
0.287184 -eM 0.028985 28 -ej 0.197963 2 1 -en 0.303501 1 1 <
random_thetas-rhos.txt > output_file.ms
```

The initial values of theta were inferred from the comparison of human and chimp orthologous sequences for each 10-kb region using a mutation rate based on the average number of substitutions per site (with the Jukes and Cantor correction) as assessed by DnaSP v5.1 (Librado and Rozas 2009). The rho parameter was estimated from the deCODE Genetics sex-averaged recombination-rate track in the UCSC Genome Browser (https://genome-euro.ucsc.edu). Uncertainty in the estimates of theta and rho was considered by generating a random and normally distributed set of values for these parameters, with mean equal to the estimated value and variance equal to the mean, from which we sampled a pair of values in each simulation.

To estimate the Ne of Canary Island aborigines in pre-European times, we accessed the publicly available Guanche genome (gun011; accession number ENA: PRJEB86458) with the largest depth-of-coverage (3.9×) from Rodríguez-Varela et al. (2017). These data correspond to a male found in Tenerife for which radiocarbon dating supports an age of 1,216 ± 27 years BP (the oldest one analyzed with the lowest contamination levels), which predates the European colonization of the Canary Islands. We used a total of 187.7 million reads mapped to the GRCh37 human reference genome as single-end reads. SAMtools (v1.3) and BCFtools (v1.3.1) were used to generate whole-genome consensus sequences at loci with a minimum mapping quality of 30 and a minimum and maximum read-depth of 3 and 20, respectively. Finally, the Pairwise Sequentially Markovian Coalescent (PSMC) method (Li and Durbin 2011) was used on the reads with base qualities >20 to estimate the Ne of the ancestral aboriginal population assuming 33 years per generation, a mutation rate of 2.5×10^{-8}, and a range of background false negative rates (FNRs) of 0.0–0.3 for variant discovery. Considering the broad range of FNRs assumed, this analysis yielded a uniform Ne in the range of 470–560 for the Guanche population.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by Instituto de Salud Carlos III (grant numbers PI11/00623, PI14/00844, PI17/00610) cofinanced by the European Regional Development Funds, "A way of making Europe" from the European Union; by Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE); and by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology. S.A. was funded by the Spanish MINECO CGL2014-58526-P and the Basque Government—Research Groups, IT1138-16. B.G.G. was supported by a fellowship from the Canarian Agency for Research, Innovation and Information Society (ACIISI, grant number TESIS2015010057) cofunded by the European Social Fund (ESF). A.D.U. was supported by a fellowship from the Spanish Ministry of Education, Culture and Sports (grant number FPU16/01435). Teide-HPC thanks to INP-2011-0063-PCT-430000-ACT (INNPLANTA program) from the Spanish Ministry of Economy and Competitiveness. The authors thank Servicio de Apoyo Informático a la Investigación (SAII) from Universidad de La Laguna for the HPC support. The content of this publication is solely the responsibility of the authors and does not necessarily reflect the views or policies of funding agencies. Finally, the authors would like to thank the anonymous reviewers for their advices to improve this manuscript.

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19(9):1655–1664.

Ara MF. 1959. Algunas observaciones acerca de la Antropología de las poblaciones prehistórica y actual de Gran Canaria. Las Palmas de Gran Canaria: *El Museo Canario*.

Aragón-Sánchez J, García-Rojas A, Lázaro-Martínez JL, Quintana-Marrero Y, Maynar-Moliner M, Rabellino M, Hernández-Herrero MJ, Cabrera-Galván JJ. 2009. Epidemiology of diabetes-related lower extremity amputations in Gran Canaria, Canary Islands (Spain). *Diabetes Res Clin Pract*. 86(1):e6–e8.

Arauna LR, Mendoza-Revilla J, Mas-Sandoval A, Izaabel H, Bekada A, Benhamamouch S, Fadhlaoui-Zid K, Zalloua P, Hellenthal G, Comas D. 2017. Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. *Mol Biol Evol*. 34(2):318–329.

Arco MC, Navarro JF. 1987. Historia popular de Canarias. Vol. 1. Los aborígenes. S/C de Tenerife: Centro de la Cultura Popular Canaria.

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC, et al. 2012. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28(10):1359–1367.

Ben Halima Y, Kefi R, Sazzini M, Giuliani C, De Fanti S, Nouali C, Nagara M, Mengozzi G, Elouej S, Abid A, et al. 2017. Lactase persistence in Tunisia as a result of admixture with other Mediterranean populations. *Genes Nutr*. 12:20.

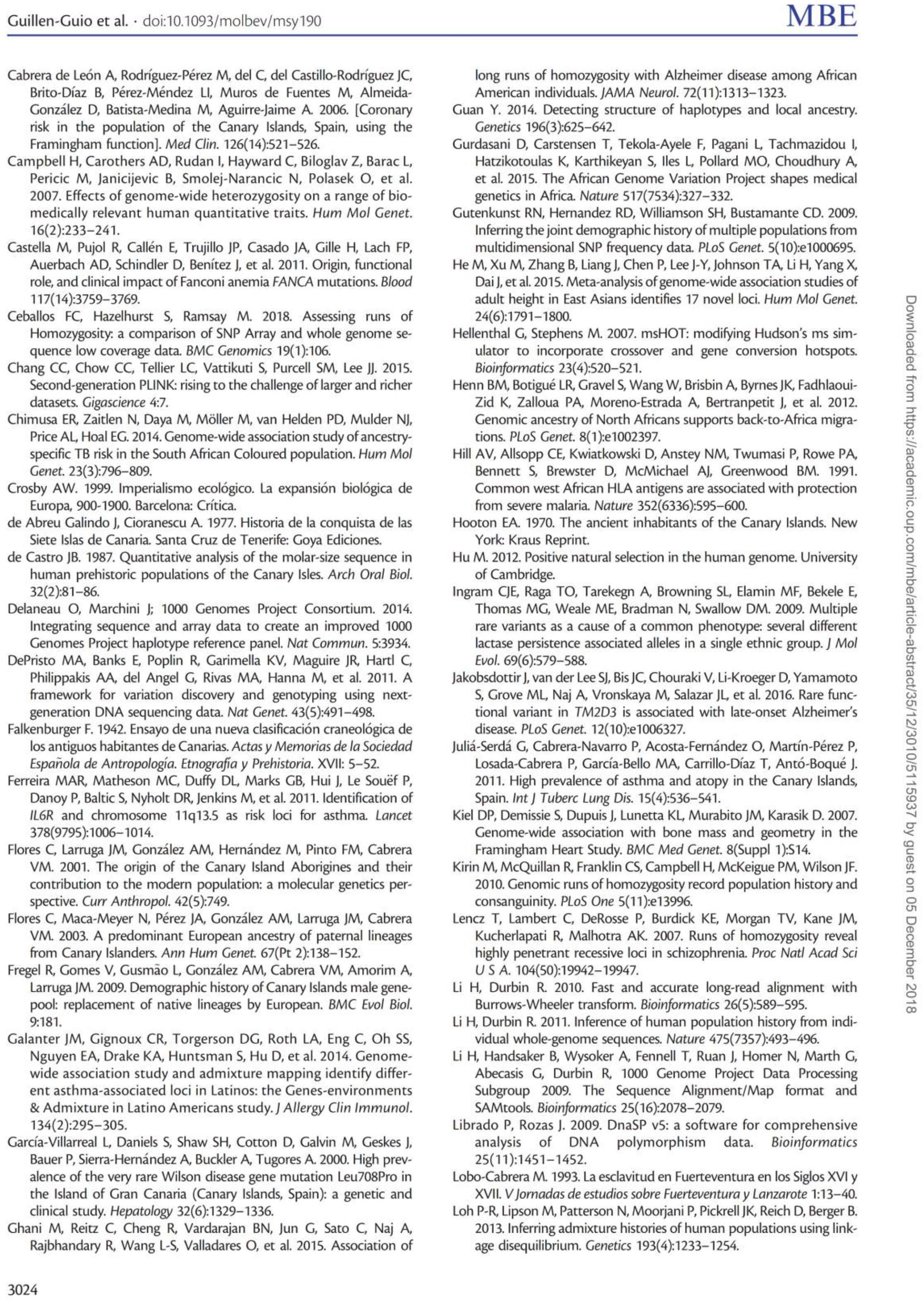
Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. 74(6):1111–1120.

Berthelot S. 1978. Etnografía y anales de la conquista de las islas Canarias. Santa Cruz de Tenerife: Goya Ediciones.

Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrer H, Flores C, et al. 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc Natl Acad Sci U S A*. 110(29):11791–11796.

Bueno H, Hernáez R, Hernández AV. 2008. Type 2 Diabetes mellitus and cardiovascular disease in Spain: a narrative review. Rev Esp Cardiol Supl. 8:50C–58C.

Cabrera de León A, González DA, Méndez LIP, Aguirre-Jaime A, del Cristo Rodríguez Pérez M, Coello SD, Trujillo IC. 2004. Leptin and altitude in the cardiovascular diseases. *Obes Res*. 12(9):1492–1498.



Cabrera de León A, Rodríguez-Pérez M, del C, del Castillo-Rodríguez JC, Brito-Díaz B, Pérez-Méndez LI, Muros de Fuentes M, Almeida-González D, Batista-Medina M, Aguirre-Jaime A. 2006. [Coronary risk in the population of the Canary Islands, Spain, using the Framingham function]. *Med Clin.* 126(14):521–526.

Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, Pericic M, Janicijevic B, Smolej-Narancic N, Polasek O, et al. 2007. Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Hum Mol Genet.* 16(2):233–241.

Castella M, Pujol R, Callén E, Trujillo JP, Casado JA, Gille H, Lach FP, Auerbach AD, Schindler D, Benítez J, et al. 2011. Origin, functional role, and clinical impact of Fanconi anemia *FANCA* mutations. *Blood* 117(14):3759–3769.

Ceballos FC, Hazelhurst S, Ramsay M. 2018. Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. *BMC Genomics* 19(1):106.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

Chimusa ER, Zaitlen N, Daya M, Möller M, van Helden PD, Mulder NJ, Price AL, Hoal EG. 2014. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Hum Mol Genet.* 23(3):796–809.

Crosby AW. 1999. Imperialismo ecológico. La expansión biológica de Europa, 900-1900. Barcelona: Crítica.

de Abreu Galindo J, Cioranescu A. 1977. Historia de la conquista de las Siete Islas de Canaria. Santa Cruz de Tenerife: Goya Ediciones.

de Castro JB. 1987. Quantitative analysis of the molar-size sequence in human prehistoric populations of the Canary Isles. *Arch Oral Biol.* 32(2):81–86.

Delaneau O, Marchini J; 1000 Genomes Project Consortium. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun.* 5:3934.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491–498.

Falkenburger F. 1942. Ensayo de una nueva clasificación craneológica de los antiguos habitantes de Canarias. *Actas y Memorias de la Sociedad Española de Antropología. Etnografía y Prehistoria.* XVII: 5–52.

Ferreira MAR, Matheson MC, Duffy DL, Marks GB, Hui J, Le Souëf P, Danoy P, Baltic S, Nyholt DR, Jenkins M, et al. 2011. Identification of *IL6R* and chromosome 11q13.5 as risk loci for asthma. *Lancet* 378(9795):1006–1014.

Flores C, Larruga JM, González AM, Hernández M, Pinto FM, Cabrera VM. 2001. The origin of the Canary Island Aborigines and their contribution to the modern population: a molecular genetics perspective. *Curr Anthropol.* 42(5):749.

Flores C, Maca-Meyer N, Pérez JA, González AM, Larruga JM, Cabrera VM. 2003. A predominant European ancestry of paternal lineages from Canary Islanders. *Ann Hum Genet.* 67(Pt 2):138–152.

Fregel R, Gomes V, Gusmão L, González AM, Cabrera VM, Amorim A, Larruga JM. 2009. Demographic history of Canary Islands male gene-pool: replacement of native lineages by European. *BMC Evol Biol.* 9:181.

Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, Nguyen EA, Drake KA, Huntsman S, Hu D, et al. 2014. Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. *J Allergy Clin Immunol.* 134(2):295–305.

García-Villarreal L, Daniels S, Shaw SH, Cotton D, Galvin M, Geskes J, Bauer P, Sierra-Hernández A, Buckler A, Tugores A. 2000. High prevalence of the very rare Wilson disease gene mutation Leu708Pro in the Island of Gran Canaria (Canary Islands, Spain): a genetic and clinical study. *Hepatology* 32(6):1329–1336.

Ghani M, Reitz C, Cheng R, Vardarajan BN, Jun G, Sato C, Naj A, Rajbhandary R, Wang L-S, Valladares O, et al. 2015. Association of long runs of homozygosity with Alzheimer disease among African American individuals. *JAMA Neurol.* 72(11):1313–1323.

Guan Y. 2014. Detecting structure of haplotypes and local ancestry. *Genetics* 196(3):625–642.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517(7534):327–332.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5(10):e1000695.

He M, Xu M, Zhang B, Liang J, Chen P, Lee J-Y, Johnson TA, Li H, Yang X, Dai J, et al. 2015. Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum Mol Genet.* 24(6):1791–1800.

Hellenthal G, Stephens M. 2007. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* 23(4):520–521.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, et al. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* 8(1):e1002397.

Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ, Greenwood BM. 1991. Common west African HLA antigens are associated with protection from severe malaria. *Nature* 352(6336):595–600.

Hooton EA. 1970. The ancient inhabitants of the Canary Islands. New York: Kraus Reprint.

Hu M. 2012. Positive natural selection in the human genome. University of Cambridge.

Ingram CJE, Raga TO, Tarekegn A, Browning SL, Elamin MF, Bekele E, Thomas MG, Weale ME, Bradman N, Swallow DM. 2009. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *J Mol Evol.* 69(6):579–588.

Jakobsdottir J, van der Lee SJ, Bis JC, Chouraki V, Li-Kroeger D, Yamamoto S, Grove ML, Naj A, Vronskaya M, Salazar JL, et al. 2016. Rare functional variant in *TM2D3* is associated with late-onset Alzheimer's disease. *PLoS Genet.* 12(10):e1006327.

Juliá-Serdá G, Cabrera-Navarro P, Acosta-Fernández O, Martín-Pérez P, Losada-Cabrera P, García-Bello MA, Carrillo-Díaz T, Antó-Boqué J. 2011. High prevalence of asthma and atopy in the Canary Islands, Spain. *Int J Tuberc Lung Dis.* 15(4):536–541.

Kiel DP, Demissie S, Dupuis J, Lunetta KL, Murabito JM, Karasik D. 2007. Genome-wide association with bone mass and geometry in the Framingham Heart Study. *BMC Med Genet.* 8(Suppl 1):S14.

Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010. Genomic runs of homozygosity record population history and consanguinity. *PLoS One* 5(11):e13996.

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK. 2007. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci U S A.* 104(50):19942–19947.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589–595.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11):1451–1452.

Lobo-Cabrera M. 1993. La esclavitud en Fuerteventura en los Siglos XVI y XVII. *V Jornadas de estudios sobre Fuerteventura y Lanzarote* 1:13–40.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193(4):1233–1254.

Lorenzo V, Alvarez A, Torres A, Torregrosa V, Hernández D, Salido E. 2006. Presentation and role of transplantation in adult patients with type 1 primary hyperoxaluria and the I244T *AGXT* mutation: single-center experience. *Kidney Int*. 70(6):1115–1119.

Lorenzo V, Boronat M, Saavedra P, Rufino M, Maceira B, Novoa FJ, Torres A. 2010. Disproportionately high incidence of diabetes-related end-stage renal disease in the Canary Islands. An analysis based on estimated population at risk. *Nephrol Dial Transplant*. 25(7):2283–2288.

Lorenzo V, Torres A, Salido E. 2014. Primary hyperoxaluria. *Nefrologia* 34(3):398–412.

Maca-Meyer N, Arnay M, Rando JC, Flores C, González AM, Cabrera VM, Larruga JM. 2004. Ancient mtDNA analysis and the origin of the Guanches. *Eur J Hum Genet*. 12(2):155–162.

Maca-Meyer N, Cabrera VM, Arnay M, Flores C, Fregel R, González AM, Larruga JM. 2005. Mitochondrial DNA diversity in 17th–18th century remains from Tenerife (Canary Islands). *Am J Phys Anthropol*. 127(4):418–426.

Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, Margulies DM, Loscalzo J, Kohane IS. 2016. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med*. 375(7):655–665.

Marcelino-Rodríguez I, Elosua R, del Cristo Rodríguez Pérez M, Fernández-Bergés D, Guembe MJ, Alonso TV, Félix FJ, González DA, Ortiz-Marrón H, Rigo F, et al. 2016. On the problem of type 2 diabetes-related mortality in the Canary Islands, Spain. The DARIOS Study. *Diabetes Res Clin Pract*. 111:74–82.

Marciniak S, Perry GH. 2017. Harnessing ancient genomes to study the history of human adaptation. *Nat Rev Genet*. 18(11):659–674.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528(7583):499–503.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*. 28(5):495–501.

Moltke I, Grarup N, Jørgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, Korneliussen TS, Andersen MA, Nielsen TS, Krarup NT, et al. 2014. A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* 512(7513):190–193.

Navarro Mederos JF. 1997. *Arqueología de las Islas Canarias. Espacio, tiempo y forma, Serie I, Prehistoria y arqueología* 10:447–478.

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32(2):292–294.

Oksenberg JR, Barcellos LF, Cree BAC, Baranzini SE, Bugawan TL, Khan O, Lincoln RR, Swerdlin A, Mignot E, Lin L, et al. 2004. Mapping multiple sclerosis susceptibility to the *HLA-DR* locus in African Americans. *Am J Hum Genet*. 74(1):160–167.

Onrubia Pintado J. 1987. Les cultures préhistoriques des Îles Canaries, état de la question. *Anthropologie* 91:653–678.

Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356(6337):543–546.

Pelak K, Goldstein DB, Walley NM, Fellay J, Ge D, Shianna KV, Gumbs C, Gao X, Maia JM, Cronin KD, et al. 2010. Host determinants of HIV-1 control in African Americans. *J Infect Dis*. 201(8):1141–1149.

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet*. 91(2):275–292.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*. 19(5):826–837.

Pino-Yanes M, Corrales A, Basaldúa S, Hernández A, Guerra L, Villar J, Flores C. 2011. North African influences and potential bias in case-control association studies in the Spanish population. *PLoS One* 6(3):e18389.

Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G, et al. 2007. A genomewide admixture map for Latino populations. *Am J Hum Genet*. 80(6):1024–1036.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81(3):559–575.

Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. 2013. Common risk alleles for inflammatory diseases are targets of recent positive selection. *Am J Hum Genet*. 92(4):517–529.

Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, et al. 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am J Hum Genet*. 94(4):496–510.

Rando JC, Cabrera VM, Larruga JM, Hernández M, González AM, Pinto F, Bandelt HJ. 1999. Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann Hum Genet*. 63(Pt 5):413–428.

Rando JC, Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM, Bandelt HJ. 1998. Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. *Ann Hum Genet*. 62(Pt 6):531–550.

Rodríguez-Esparragón F, López-Fernández JC, Buset-Ríos N, García-Bello MA, Hernández-Velazquez E, Cappiello L, Rodríguez-Pérez JC. 2017. Paraoxonase 1 and 2 gene variants and the ischemic stroke risk in Gran Canaria population: an association study and meta-analysis. *Int J Neurosci*. 127(3):191–198.

Rodríguez-Varela R, Günther T, Krzewińska M, Storå J, Gillingwater TH, MacCallum M, Arsuaga JL, Dobney K, Valdiosera C, Jakobsson M, et al. 2017. Genomic analyses of pre-european conquest human remains from the canary islands reveal close affinity to modern North Africans. *Curr Biol*. 27(21):3396–3402.e5.

Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, Rudan P. 2003. Inbreeding and the genetic complexity of human hypertension. *Genetics* 163(3):1011–1021.

Sánchez-Lerma B, Morales-Chirivella FJ, Peñuelas I, Blanco Guerra C, Mesa Lugo F, Aguinaga-Ontoso I, Guillén-Grima F. 2009. High prevalence of asthma and allergic diseases in children aged 6 to [corrected] 7 years from the Canary Islands. [corrected]. *J Investig Allergol Clin Immunol*. 19:383–390.

Sanchez-Mazas A, Černý V, Di D, Buhler S, Podgorná E, Chevallier E, Brunet L, Weber S, Kervaire B, Testi M, et al. 2017. The *HLA-B* landscape of Africa: signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. *Mol Ecol*. 26(22):6238–6252.

Santana A, Salido E, Torres A, Shapiro LJ. 2003. Primary hyperoxaluria type 1 in the Canary Islands: a conformational disease due to I244T mutation in the P11L-containing alanine: glyoxylate aminotransferase. *Proc Natl Acad Sci U S A*. 100(12):7277–7282.

Schrider DR, Kern AD. 2017. Soft sweeps are the dominant mode of adaptation in the human genome. *Mol Biol Evol*. 34(8):1863–1877.

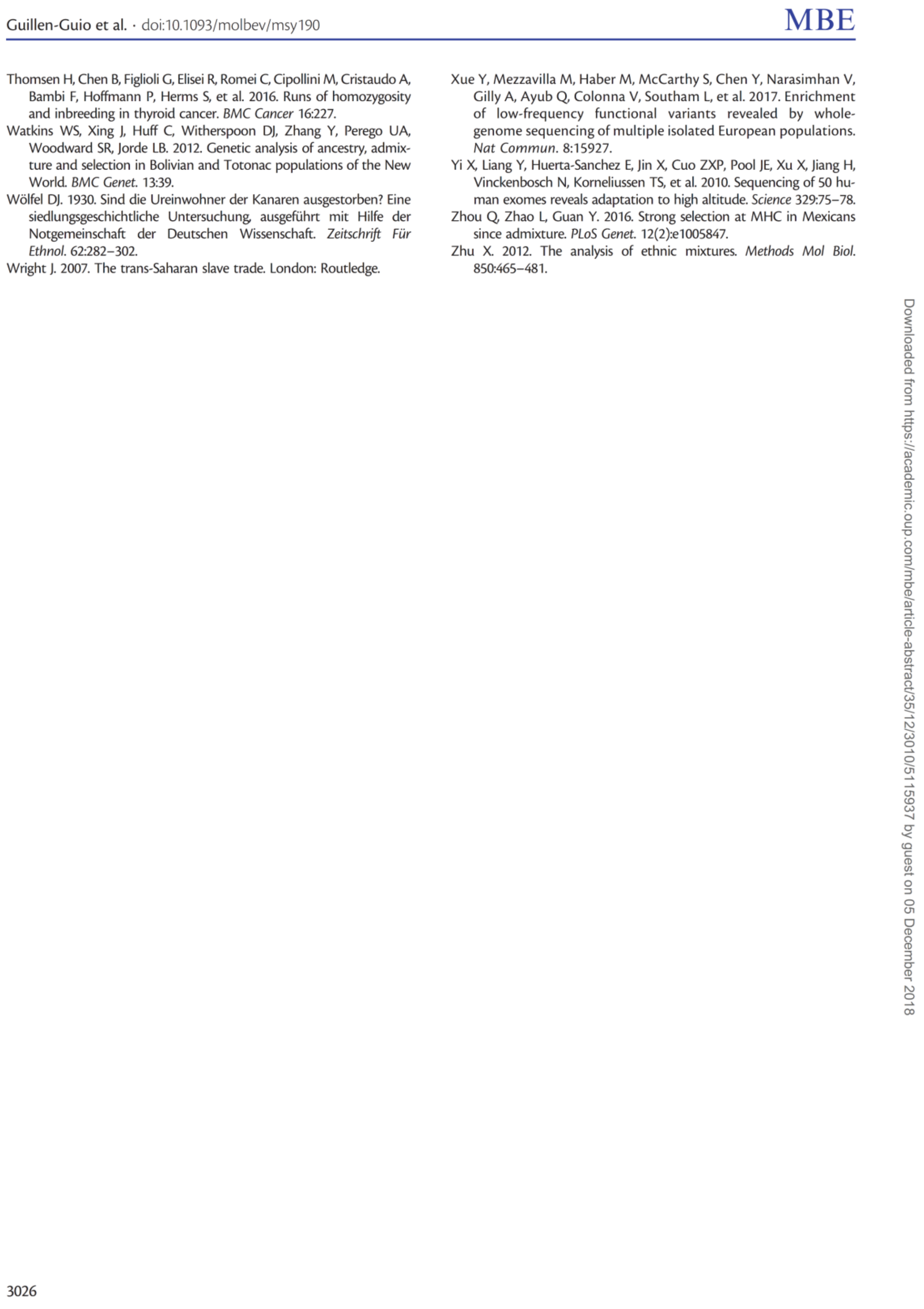
Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK. 2006. European population substructure: clustering of northern and southern populations. *PLoS Genet*. 2(9):e143.

Shriner D, Adeyemo A, Rotimi CN. 2011. Joint ancestry and association testing in admixed individuals. *PLoS Comput Biol*. 7(12):e1002325.

Suárez JJ, Rodríguez F, Quintero CL. 1988. Historia popular de Canarias. Vol. 2. Conquista y colonización. Santa Cruz de Tenerife: Gobierno de Canarias. Centro de Cultura Popular Canaria.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526(7571):75–81.

Thomas R, Apps R, Qi Y, Gao X, Male V, O'hUigin C, O'Connor G, Ge D, Fellay J, Martin JN, et al. 2009. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of *HLA-C*. *Nat Genet*. 41(12):1290–1294.



Thomsen H, Chen B, Figlioli G, Elisei R, Romei C, Cipollini M, Cristaudo A, Bambi F, Hoffmann P, Herms S, et al. 2016. Runs of homozygosity and inbreeding in thyroid cancer. *BMC Cancer* 16:227.

Watkins WS, Xing J, Huff C, Witherspoon DJ, Zhang Y, Perego UA, Woodward SR, Jorde LB. 2012. Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. *BMC Genet.* 13:39.

Wölfel DJ. 1930. Sind die Ureinwohner der Kanaren ausgestorben? Eine siedlungsgeschichtliche Untersuchung, ausgeführt mit Hilfe der Notgemeinschaft der Deutschen Wissenschaft. *Zeitschrift Für Ethnol.* 62:282–302.

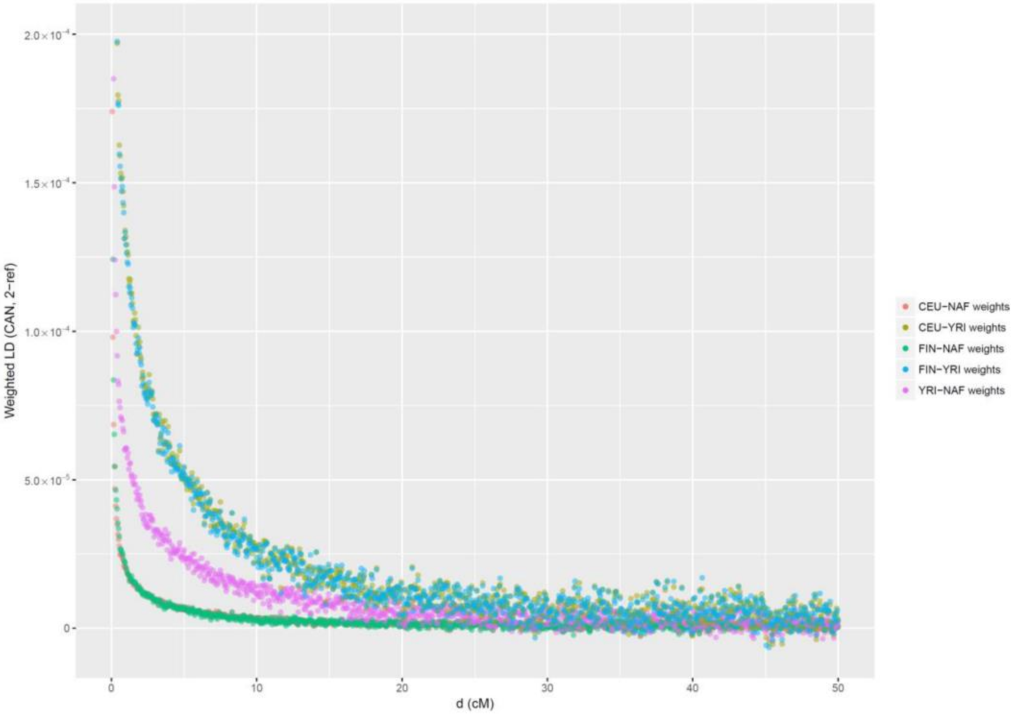
Wright J. 2007. The trans-Saharan slave trade. London: Routledge.

Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, Gilly A, Ayub Q, Colonna V, Southam L, et al. 2017. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun.* 8:15927.

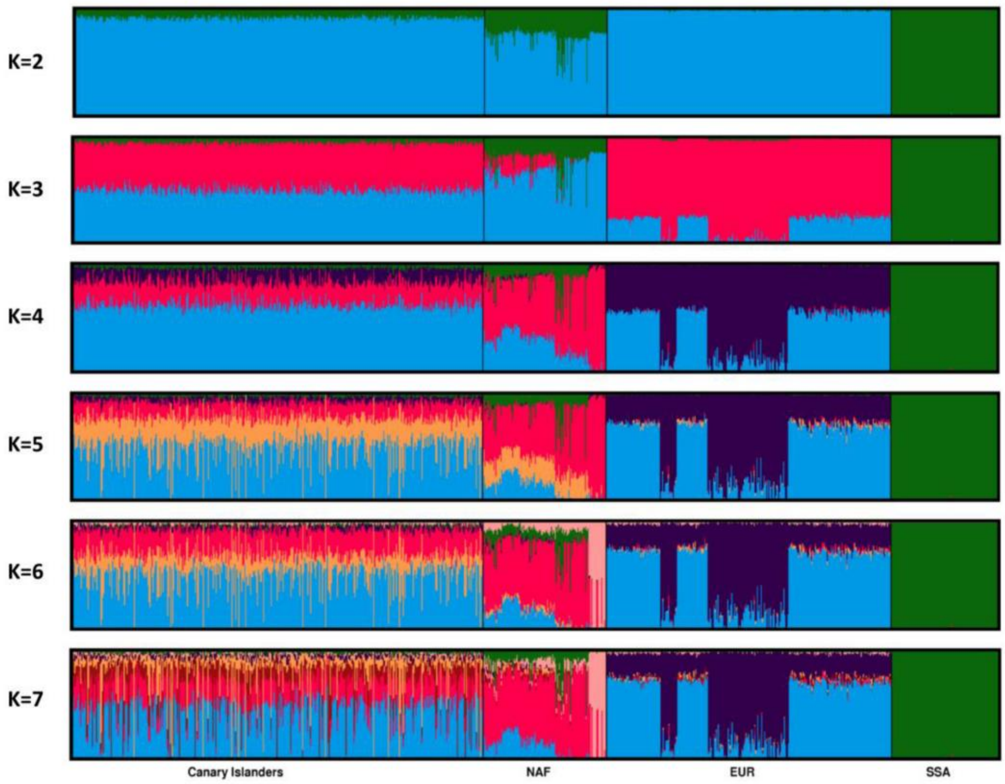
Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.

Zhou Q, Zhao L, Guan Y. 2016. Strong selection at MHC in Mexicans since admixture. *PLoS Genet.* 12(2):e1005847.

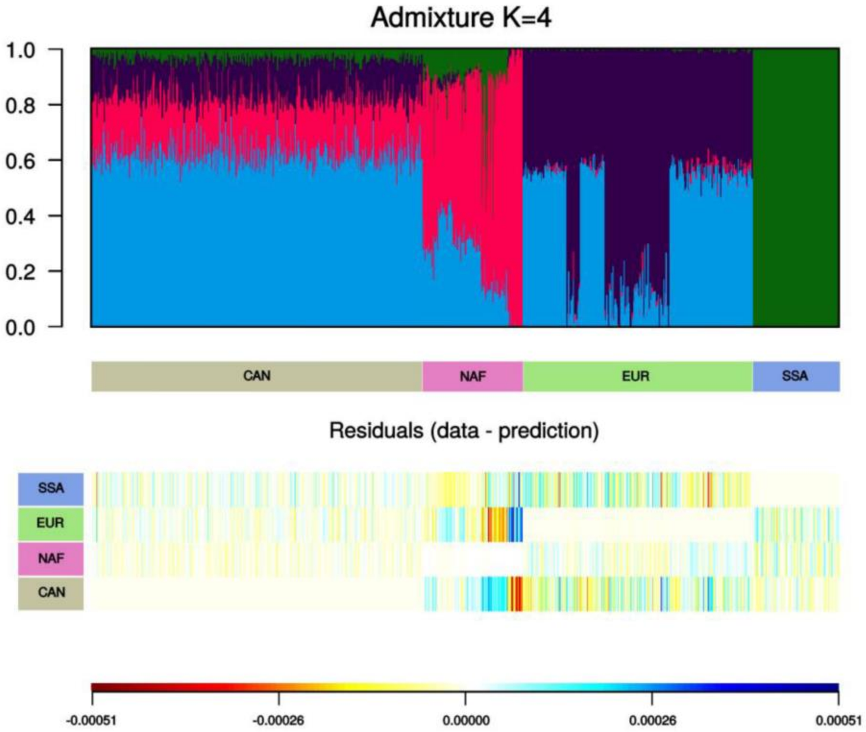
Zhu X. 2012. The analysis of ethnic mixtures. *Methods Mol Biol.* 850:465–481.



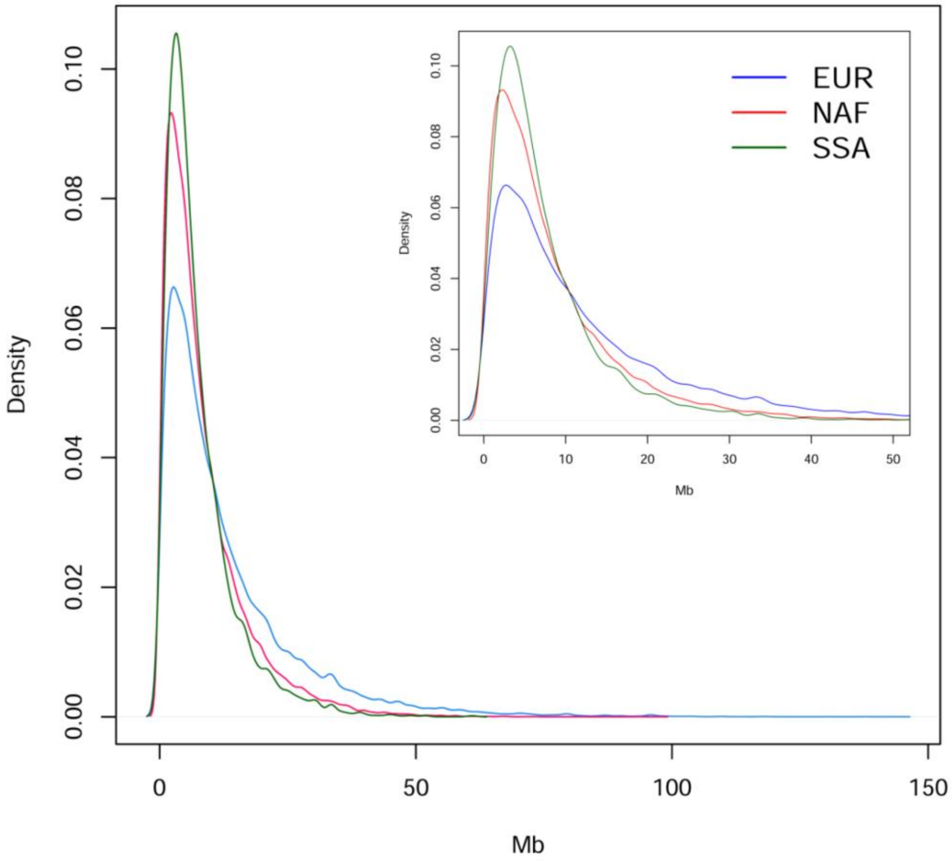
Supplementary Figure 1. ALDER results for 2-reference weighted LD computations in exemplar population pairs.



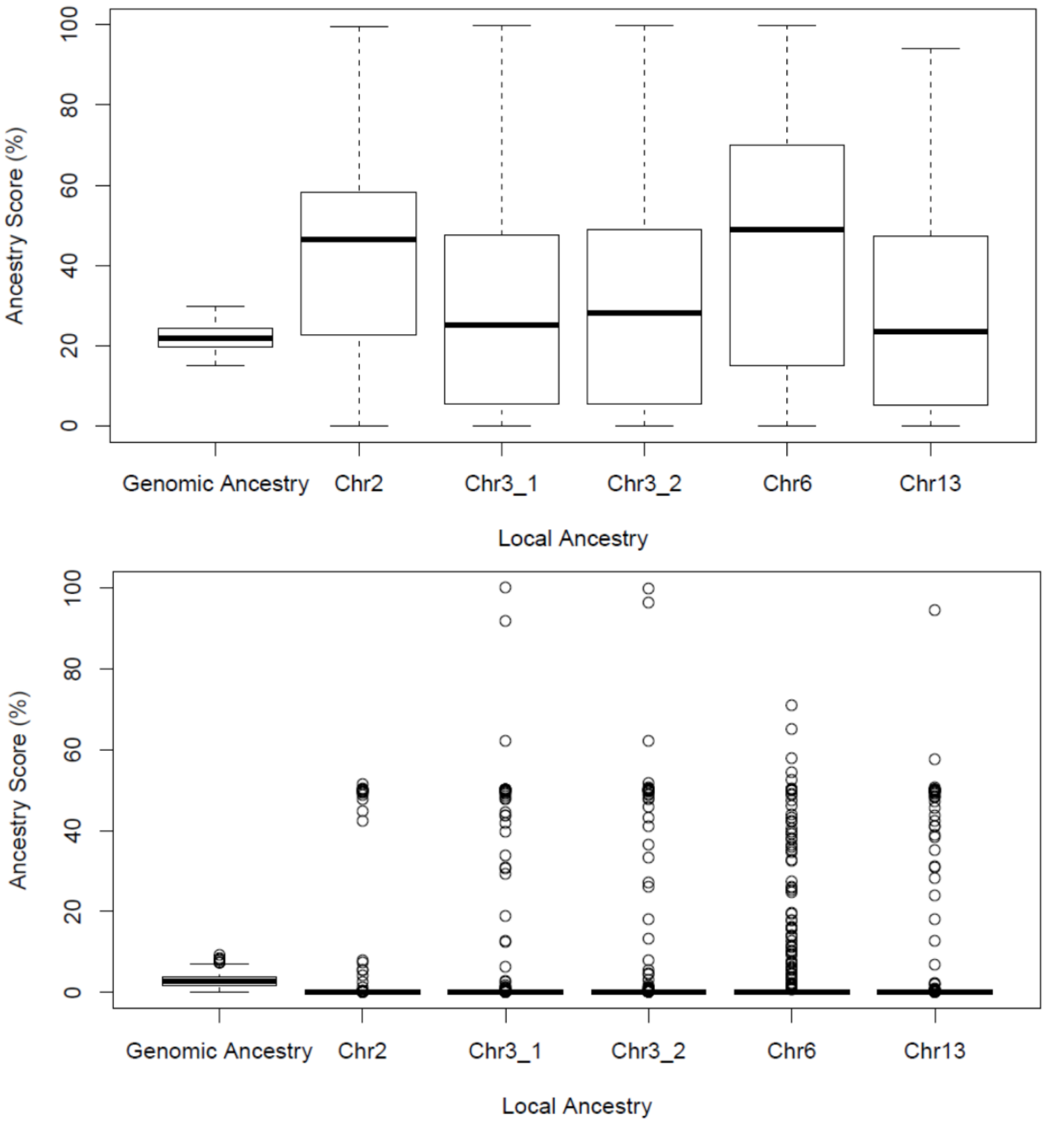
Supplementary Figure 2. ADMIXTURE results from K= 2 through 7. Individuals are represented as vertical lines, and each K ancestral genetic cluster is represented by a color. The lowest CV error was obtained at K=4, differentiating two ancestry clusters in Europeans. For K>5 a new genetic cluster arises mainly assigned to the Tunisian population, revealing novel ancestry clusters in populations other than SSA.



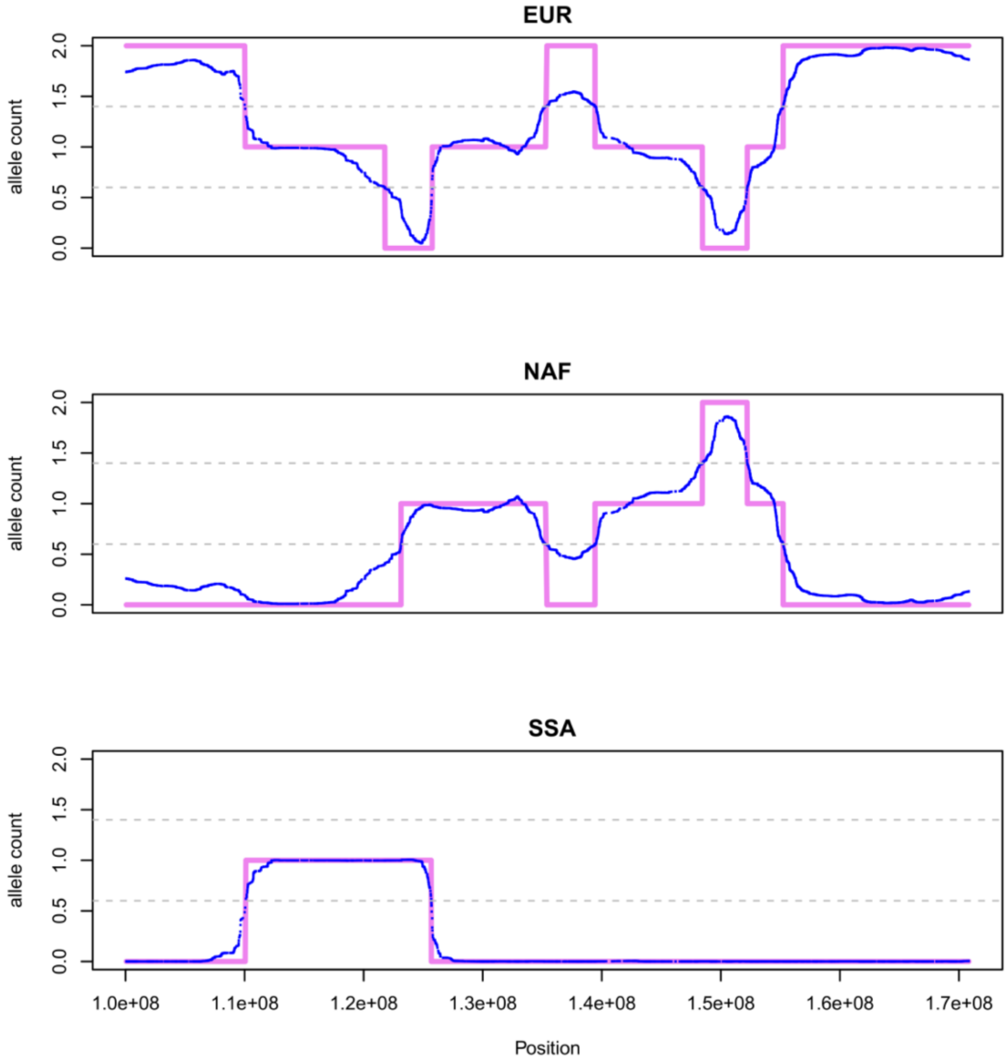
Supplementary Figure 3. Admixture model fitting as provided by badMIXTURE. Upper panel: Ancestry clusters as estimated by ADMIXTURE (K=4). Lower panel: Residuals from the goodness of fit of the model with CHROMOPAINTER measures of haplotype sharing with individuals from NAF, EUR and SSA groups.



Supplementary Figure 4. Density plot of approximate ELAI block size estimates.



Supplementary Figure 5. Regional plots of ancestry scores for the regions with large deviations in NAF (top) and SSA (bottom) ancestries. Chr3_1 and Chr3_2 correspond to chr3:10,539,482-11,710,471 and chr3:177,443,968-178,679,751, respectively.



Supplementary Figure 6. Inference of local ancestry by ELAI. The plot shows an example of inference in a chromosome region (one panel of each parental population) comparing ELAI allele dosages (blue) with the approximations to the next non-negative integer (pink) to assist in the estimation of ELAI block lengths.

Supplementary Table 1. Genes mapping to the regions with large deviations in ancestry.

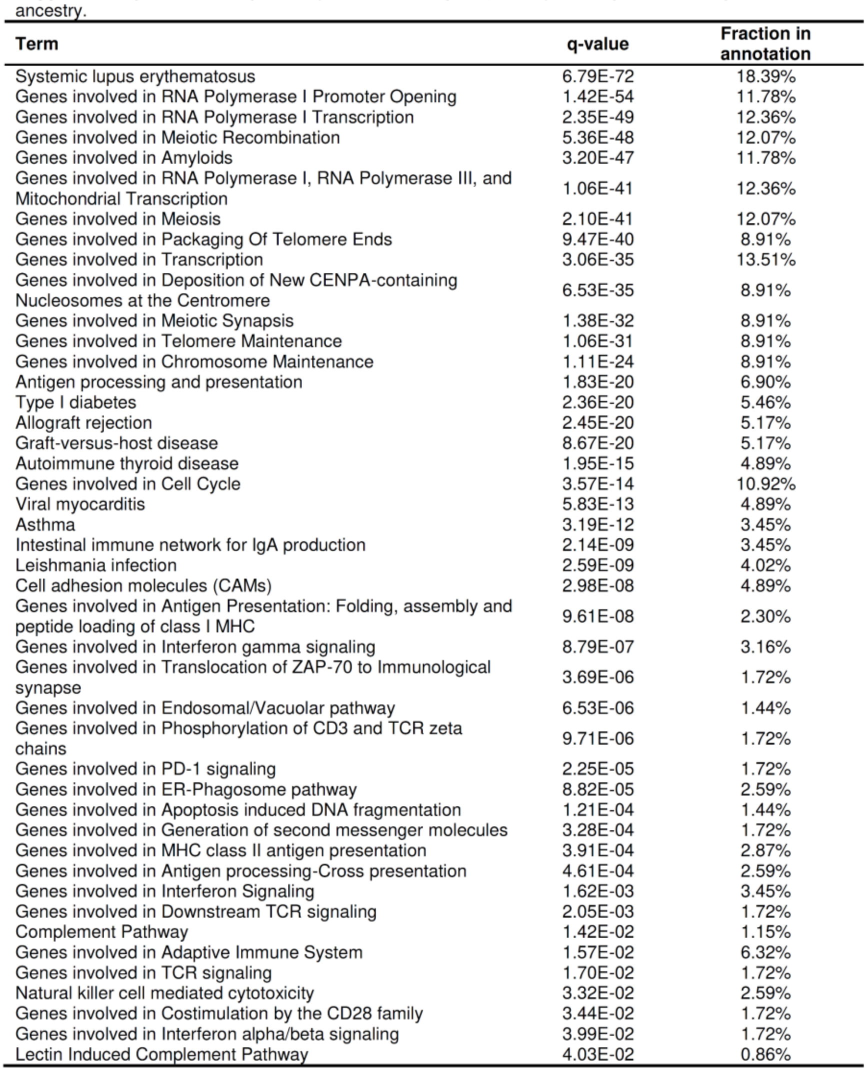
ABCF1, ABHD16A, ABT1, ACMSD, ACOT13, AGER, AGPAT1, AIF1, ALDH5A1, ANKS1A, APOM, ARHGAP15, ARMC12, ATAT1, ATF6B, ATG7, ATP2B2, ATP6V1G2, B3GALT4, BAG6, BAK1, BRD2, BRPF3, BTN1A1, BTN2A1, BTN2A2, BTN3A1, BTN3A2, BTN3A3, BTNL2, C2, C4A, C4B_2, C6orf1, C6orf10, C6orf106, C6orf136, C6orf15, C6orf222, C6orf25, C6orf47, C6orf48, C6orf62, CCHCR1, CCNT2, CDKN1A, CDSN, CFB, CLIC1, CLPS, CLPSL1, CLPSL2, COL11A2, CSNK2B, CUTA, CXCR4, CYP21A2, DARS, DAXX, DCDC2, DDAH2, DDR1, DDX39B, DEF6, DHX16, DIAPH3, DPCR1, EGFL8, EHMT2, ETV7, FAM65B, FANCE, FKBP5, FKBPL, FLOT1, GABBR1, GMNN, GNL1, GPANK1, GPLD1, GPSM3, GPX5, GPX6, GRM4, GTF2H4, HFE, HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1D, HIST1H1E, HIST1H1T, HIST1H2AA, HIST1H2AB, HIST1H2AC, HIST1H2AD, HIST1H2AE, HIST1H2AG, HIST1H2AH, HIST1H2AJ, HIST1H2AK, HIST1H2AL, HIST1H2AM, HIST1H2BA, HIST1H2BB, HIST1H2BC, HIST1H2BD, HIST1H2BE, HIST1H2BF, HIST1H2BG, HIST1H2BH, HIST1H2BI, HIST1H2BJ, HIST1H2BK, HIST1H2BL, HIST1H2BM, HIST1H2BN, HIST1H2BO, HIST1H3A, HIST1H3B, HIST1H3C, HIST1H3D, HIST1H3E, HIST1H3F, HIST1H3G, HIST1H3H, HIST1H3I, HIST1H3J, HIST1H4A, HIST1H4B, HIST1H4C, HIST1H4D, HIST1H4E, HIST1H4F, HIST1H4G, HIST1H4H, HIST1H4I, HIST1H4J, HIST1H4K, HIST1H4L, HLA-A, HLA-B, HLA-C, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB5, HLA-E, HLA-F, HLA-G, HMGA1, HMGN4, HNMT, HRH1, HSD17B8, HSPA1A, HSPA1B, HSPA1L, IER3, IP6K3, ITPR3, KAAG1, KCNMB2, KCTD20, KIAA0319, KIFC1, KYNU, LCT, LEMD2, LHFPL5, LRP1B, LRRC16A, LSM2, LST1, LTA, LTB, LY6G5B, LY6G5C, LY6G6C, LY6G6D, LY6G6F, MAP3K19, MAPK13, MAPK14, MAS1L, MCCD1, MCM6, MDC1, MGAT5, MICA, MICB, MLN, MOG, MRPS18B, MRS2, MUC21, MUC22, NCKAP5, NCR3, NELFE, NEU1, NFKBIL1, NKAPL, NOTCH4, NRM, NRSN1, NUDT3, NXPH2, OR10C1, OR11A1, OR12D2, OR12D3, OR14J1, OR2B2, OR2B3, OR2B6, OR2H1, OR2H2, OR2J2, OR2J3, OR2W1, OR5V1, PACSIN1, PBX2, PCDH17, PCDH20, PCDH9, PFDN6, PGBD1, PHF1, PNPLA1, POM121L2, POU5F1, PPARD, PPP1R10, PPP1R11, PPP1R18, PPT2, PRR3, PRRC2A, PRRT1, PRSS16, PSMB8, PSMB9, PSORS1C1, PSORS1C2, PXT1, R3HDM1, RAB3GAP1, RGL2, RING1, RNF39, RNF5, RPL10A, RPP21, RPS10, RPS18, RXRB, SCGN, SCUBE3, SFTA2, SKIV2L, SLC17A1, SLC17A2, SLC17A3, SLC17A4, SLC26A8, SLC39A7, SLC44A4, SLC6A1, SLC6A11, SNRPC, SPDEF, SPOPL, SRPK1, SRSF3, STK19, STK38, SYNGAP1, TAF11, TAP1, TAP2, TAPBP, TCF19, TCP11, TDP2, TDRD3, TEAD3, THSD7B, TMEM163, TNF, TNXB, TRIM10, TRIM15, TRIM26, TRIM27, TRIM31, TRIM38, TRIM39, TRIM39-RPP21, TRIM40, TUBB, TULP1, UBD, UBXN4, UHRF1BP1, VARS, VARS2, VGLL4, VPS52, VWA7, WDR46, ZBTB12, ZBTB22, ZBTB9, ZFP57, ZKSCAN3, ZKSCAN4, ZKSCAN8, ZNF165, ZNF184, ZNF311, ZNF322, ZNF391, ZNF76, ZNRD1, ZRANB3, ZSCAN16, ZSCAN23, ZSCAN26, ZSCAN31, ZSCAN9

Supplementary Table 2. Significantly enriched human diseases in regions with large deviations in ancestry.

Term	q-value	Fraction in annotation
Psoriasis	3.44E-10	7.18%
Mucocutaneous lymph node syndrome	2.13E-08	3.74%
Vascular skin disease	2.88E-08	5.75%
Lymphadenitis	4.16E-08	3.74%
Integumentary system disease	1.03E-07	11.21%
Skin disease	1.10E-07	10.92%
Peripheral vascular disease	5.50E-07	6.03%
Polyarteritis nodosa	6.91E-07	3.74%
Nephrotic syndrome	8.71E-06	3.16%
Nephrosis	1.37E-05	3.16%
Graves' disease	3.56E-05	3.16%
Sarcoidosis	4.96E-05	3.16%
Autoimmune disease	7.32E-05	10.92%
Hyperthyroidism	2.33E-04	3.16%
Thyrotoxicosis	2.44E-04	3.16%
Hepatitis	3.08E-04	6.90%
Goiter	4.04E-04	3.16%
Autoimmune disease of endocrine system	4.38E-04	3.16%
Cytomegalovirus infectious disease	4.53E-04	3.45%
(+)ssRNA virus infectious disease	4.56E-04	5.46%
Psoriatic arthritis	4.65E-04	2.01%
Pemphigus	4.65E-04	2.01%
Allergy	4.66E-04	4.31%
Duodenal ulcer	4.81E-04	1.72%
Cystic echinococcosis	5.30E-04	0.86%
Leprosy	5.44E-04	2.01%
Flaviviridae infectious disease	5.65E-04	4.89%
Echinococcosis	6.07E-04	1.15%
Demyelinating disease	6.73E-04	5.46%
Behcet's disease	6.78E-04	2.59%
Hepatitis C	7.20E-04	4.60%
DNA virus infectious disease	9.47E-04	7.76%
Intrahepatic cholestasis	1.04E-03	1.44%
Multiple sclerosis	1.04E-03	5.17%
Demyelinating disease of central nervous system	1.22E-03	5.17%
Hepatitis B	1.24E-03	4.02%
Pulmonary tuberculosis	1.30E-03	2.01%
Cavernous hemangioma	1.36E-03	0.86%
Liver disease	1.84E-03	8.05%
Congenital adrenal hyperplasia	1.84E-03	1.15%
Primary bacterial infectious disease	1.86E-03	4.60%
Arthritis	1.89E-03	8.05%
Bone inflammation disease	1.91E-03	8.33%
Dengue disease	2.20E-03	1.44%
Rheumatic fever	2.20E-03	1.44%
Urticaria	2.41E-03	2.01%
Food allergy	2.60E-03	2.59%
Glucose metabolism disease	3.26E-03	10.06%
Diabetes mellitus	3.26E-03	9.77%
Peptic ulcer	3.53E-03	2.01%
Nasopharynx carcinoma	3.77E-03	3.74%
Viral infectious disease	4.04E-03	10.06%
Brucellosis	4.31E-03	1.72%
Chagas cardiomyopathy	4.48E-03	0.86%
Neuromyelitis optica	5.01E-03	1.15%
Parasitic infectious disease	5.06E-03	3.16%
Systemic lupus erythematosus	5.14E-03	4.60%
Severe acute respiratory syndrome	5.67E-03	1.72%
Bacterial infectious disease	5.82E-03	4.89%
RNA virus infectious disease	5.89E-03	6.32%
Rheumatoid arthritis	6.17E-03	6.32%
Nidovirales infectious disease	6.18E-03	1.72%
Vitiligo	6.32E-03	2.01%

Prostatitis	6.41E-03	0.86%
Celiac disease	6.83E-03	2.30%
Hepatobiliary disease	7.70E-03	8.33%
Aplastic anemia	7.83E-03	2.01%
Bullous skin disease	7.83E-03	2.01%
Herpesviridae infectious disease	8.06E-03	4.02%
Vasculitis	8.27E-03	2.87%
Purpura	8.88E-03	2.01%
Pure red-cell aplasia	8.92E-03	0.57%
Alcoholic pancreatitis	9.01E-03	0.86%
Autoimmune disease of gastrointestinal tract	9.07E-03	3.45%
Mycobacterium infectious disease	1.03E-02	3.16%
Hypotrichosis	1.07E-02	1.72%
Adrenal hyperplasia	1.15E-02	1.15%
Parasitic helminthiasis infectious disease	1.16E-02	1.44%
Gastrointestinal system disease	1.18E-02	14.66%
Dengue shock syndrome	1.19E-02	0.86%
Primary Actinomycetales infectious disease	1.28E-02	3.16%
Lupus erythematosus	1.30E-02	4.89%
Autoimmune disease of the nervous system	1.33E-02	1.72%
Dermatitis	1.40E-02	4.31%
Hair disease	1.43E-02	1.72%
Upper respiratory tract disease	1.67E-02	3.45%
Open-angle glaucoma	1.83E-02	1.72%
Myasthenia gravis	1.89E-02	1.44%
Neuromuscular junction disease	2.05E-02	1.44%
Nose disease	2.06E-02	2.30%
Facial neoplasm	2.10E-02	0.57%
Bacterial prostatitis	2.10E-02	0.57%
Spondylitis	2.64E-02	1.72%
Ankylosing spondylitis	2.64E-02	1.72%
Flavivirus infectious disease	2.74E-02	1.44%
Arthropathy	2.77E-02	2.01%
Primitive neuroectodermal tumor	2.77E-02	6.03%
Alopecia	2.94E-02	1.44%
Sickle cell anemia	2.96E-02	1.15%
Pancreatitis	3.03E-02	2.30%
Diabetes mellitus type 1	3.03E-02	0.86%
Uveal disease	3.17E-02	1.44%
Uveitis	3.27E-02	1.15%
Vascular hemostatic disease	3.35E-02	3.74%
dsDNA virus infectious disease	3.37E-02	5.17%
Embryonal cancer	3.44E-02	6.03%
Collagen disease	3.47E-02	2.87%
Sjogren's syndrome	3.47E-02	1.72%
Rubella	3.48E-02	0.86%
Dengue hemorrhagic fever	3.48E-02	0.86%
Alloimmunization	3.49E-02	0.57%
Alcoholic fatty liver	3.49E-02	0.57%
Complex regional pain syndrome	3.49E-02	0.57%
Crohn's disease	3.50E-02	2.87%
Tuberculosis	3.72E-02	2.59%
Nephritis	3.73E-02	3.16%
Spondyloarthropathy	4.28E-02	1.72%
Allergic rhinitis	4.33E-02	2.01%
Retinal degeneration	4.57E-02	3.45%
Rhinitis	4.78E-02	2.01%

Supplementary Table 3. Significantly enriched MSigDB pathways in regions with large deviations in ancestry.



Term	q-value	Fraction in annotation
Systemic lupus erythematosus	6.79E-72	18.39%
Genes involved in RNA Polymerase I Promoter Opening	1.42E-54	11.78%
Genes involved in RNA Polymerase I Transcription	2.35E-49	12.36%
Genes involved in Meiotic Recombination	5.36E-48	12.07%
Genes involved in Amyloids	3.20E-47	11.78%
Genes involved in RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription	1.06E-41	12.36%
Genes involved in Meiosis	2.10E-41	12.07%
Genes involved in Packaging Of Telomere Ends	9.47E-40	8.91%
Genes involved in Transcription	3.06E-35	13.51%
Genes involved in Deposition of New CENPA-containing Nucleosomes at the Centromere	6.53E-35	8.91%
Genes involved in Meiotic Synapsis	1.38E-32	8.91%
Genes involved in Telomere Maintenance	1.06E-31	8.91%
Genes involved in Chromosome Maintenance	1.11E-24	8.91%
Antigen processing and presentation	1.83E-20	6.90%
Type I diabetes	2.36E-20	5.46%
Allograft rejection	2.45E-20	5.17%
Graft-versus-host disease	8.67E-20	5.17%
Autoimmune thyroid disease	1.95E-15	4.89%
Genes involved in Cell Cycle	3.57E-14	10.92%
Viral myocarditis	5.83E-13	4.89%
Asthma	3.19E-12	3.45%
Intestinal immune network for IgA production	2.14E-09	3.45%
Leishmania infection	2.59E-09	4.02%
Cell adhesion molecules (CAMs)	2.98E-08	4.89%
Genes involved in Antigen Presentation: Folding, assembly and peptide loading of class I MHC	9.61E-08	2.30%
Genes involved in Interferon gamma signaling	8.79E-07	3.16%
Genes involved in Translocation of ZAP-70 to Immunological synapse	3.69E-06	1.72%
Genes involved in Endosomal/Vacuolar pathway	6.53E-06	1.44%
Genes involved in Phosphorylation of CD3 and TCR zeta chains	9.71E-06	1.72%
Genes involved in PD-1 signaling	2.25E-05	1.72%
Genes involved in ER-Phagosome pathway	8.82E-05	2.59%
Genes involved in Apoptosis induced DNA fragmentation	1.21E-04	1.44%
Genes involved in Generation of second messenger molecules	3.28E-04	1.72%
Genes involved in MHC class II antigen presentation	3.91E-04	2.87%
Genes involved in Antigen processing-Cross presentation	4.61E-04	2.59%
Genes involved in Interferon Signaling	1.62E-03	3.45%
Genes involved in Downstream TCR signaling	2.05E-03	1.72%
Complement Pathway	1.42E-02	1.15%
Genes involved in Adaptive Immune System	1.57E-02	6.32%
Genes involved in TCR signaling	1.70E-02	1.72%
Natural killer cell mediated cytotoxicity	3.32E-02	2.59%
Genes involved in Costimulation by the CD28 family	3.44E-02	1.72%
Genes involved in Interferon alpha/beta signaling	3.99E-02	1.72%
Lectin Induced Complement Pathway	4.03E-02	0.86%

4. Discussion

4. Discussion

This thesis work addressed the complex pathophysiology of ARDS from distinct angles by using different genomic approaches. As part of this study, we have: i) performed a systemic review of the genetics of ARDS; ii) assessed the genetic variation in cases with ARDS and controls with sepsis, the major cause of ARDS development, to identify novel disease genes; iii) studied the lung microbiome shifts of patients with sepsis as a biomarker of ICU mortality; and iv) characterized the genomic diversity of a recently admixed European population with the aim of revealing relationships between genetic ancestry, adaptations and disease risks. As a result of these studies, we have proposed a new therapeutic target for ARDS and an early prognostic biomarker for non-pulmonary sepsis patients. We have also evidenced that genomic regions with large deviations in local ancestry in the Canary Islands population harbor genes related to critical illnesses and have characterized the number and size of their ancestry blocks, an information that is necessary to assist the design and analyses of subsequent admixture mapping studies in this population.

4.1. Genetic association studies in ARDS: difficulties and challenges

In Chapter 1, we revised all published genetic association studies in ARDS until December 2015 to detect the genes that were most likely involved in susceptibility or outcomes based on the number of independent studies that reported a significant association. We observed that, while current approaches allow to scan genetic variations across the genome in relation to a disease (Reilly et al. 2017), most of genetic studies in ARDS have focused on candidate genes based on their biological plausibility, and only a few studies have evaluated the genetic variation at genomic level (Guillén-Guío et al. 2016). Those genes that had the largest number of independent study findings were mainly involved in the immune response, such as interleukin 1 receptor antagonist (*IL1RN*), *IL6*, *IL10*, and mannose-binding lectin (protein C) 2, soluble (*MBL2*); and in vascular permeability, including *ACE*, *VEGFA*, and the nicotinamide phosphoribosyltransferase (*NAMPT*) (Guillén-Guío et al. 2016). Remarkably, the *MYLK* gene was also revealed as a robust ARDS gene based on candidate gene, GWAS and WES studies (Gao et al. 2006; Christie et al. 2008; Christie et al. 2012; Lee et al. 2012), suggesting again the important role of vascular permeability in ARDS pathophysiology. Unfortunately, although candidate gene association studies have revealed important insights into ARDS pathogenesis, there are no effective treatments designed based on the reported genes, and identifying novel genetic factors involved in ARDS remains a necessity.

The absence of potential therapeutic targets in ARDS is affected by the low replicability of candidate gene association studies, which makes the results less unreliable (Chanock et al. 2007). In fact, a large proportion of the associations between genetic variants and other complex diseases revealed by these kind of studies have been questioned because of problems inherent to the approach and to the difficulties in interpreting the results (Clark and Baudouin 2006). Furthermore, the lack of reproducibility could also be caused by other factors. On the one hand, ARDS is a heterogeneous trait with different sources and a complex pathology that hinders a precise patient classification. Since its initial definition (Ashbaugh et al. 1967), the efforts to better classify ARDS have been constant. Nowadays, the most recent ARDS classification follows the Berlin definition (ARDS Definition Task Force et al. 2012), although it is far from being settled (Barbas et al. 2014; Villar et al. 2016). This divergence of ARDS definitions through time can be also affecting the genetic studies, hindering analyses and influencing the non-replicability of the findings. Furthermore, sample sizes used in these studies are limited and, hence, their statistical power to reveal factors with weak effects is reduced (Columb and Atkinson 2015).

In this sense, the use of larger sample sizes and high-throughput technologies that assess genetic variation at a genomic scale, including GWAS, WES, and WGS, would be more powerful alternatives to disentangle the genetic variation associated with ARDS. Their effectiveness has become evident by the results obtained in the GWAS of sepsis-associated ARDS described in Chapter 2, where a novel ARDS gene has emerged and a potential therapeutic target has been proposed. Moreover, a replication phase in an independent sample should be practically mandatory to validate the association results, regardless of the approach chosen. Ideally, the results should also be accompanied by sensitivity analyses to control the effects of confounder factors, as well as by functional studies that can shed light on how the risk alleles affect the disease. In addition, given that most association studies in ARDS have been performed in European populations (Acosta-Herrera et al. 2014), further studies including patients of other ethnicities are relevant to unmask novel variants that are prevalent in other ancestries. In this sense, genetic studies in recently admixed populations, such as that of the Canary Islands, constitute a promising complement to the most common studies conducted in admixed American populations. Finally, given the complex nature of ARDS, which is associated with genetic and environmental factors, the use of alternative “omics” in the context of ARDS (such as metagenomics, metabolomics, epigenetics, and proteomics) is crucial to improve the knowledge of this syndrome, as well as to design new therapeutic and prognostic options.

4.2. The role of *FLT1* in sepsis-associated ARDS

Our results from the first GWAS of sepsis-associated ARDS, described in Chapter 2, revealed a novel genome-wide significant association (rs9508032) with ARDS susceptibility in individuals of European ancestry. This variant was located within the *FLT1* gene, which encodes VEGFR-1, one of the main receptors of VEGF-A (also known as VEGF). *FLT1* had never been specifically associated with ARDS before in an independent study, although it has been related to pulmonary complications altogether (Kim et al. 2012). It was also associated with other complex traits where, as in ARDS, the endothelium has a key role, including preeclampsia (McGinnis et al. 2017; Gray et al. 2018) and coronary arterial disease (CARDIoGRAMplusC4D Consortium et al. 2013; Konta et al. 2016). Conversely, most of the candidate gene studies focusing the VEGF pathway had paid attention to the *VEGFA* gene, revealing numerous independent associations with ARDS. Accordingly, when we performed the look-up of genes that had been previously associated with ARDS in our GWAS findings, only genetic variants within *VEGFA* were found to be significantly associated at the gene level after Bonferroni correction. Taken together, this led to support the hypothesis that other genes from the VEGF pathway should be promising candidates for further detailed genetic and functional studies in order to reveal novel ARDS risks.

Additional functional analyses supported the role of the top-most significant variant in our GWAS within *FLT1* and its perfect LD proxies (all located within the intron 10) in the regulation of the *FLT1* promoter. Specifically, luciferase reporter assays showed that the protective alleles of these variants were associated with a reduced promoter activity in monocyte cells. Monocytes, which express VEGFR-1 at high levels (Shibuya 2001), have been linked to the pathophysiology of ARDS (Herold et al. 2013; Aggarwal et al. 2014; Abdulnour et al. 2018). During the syndrome, peripheral blood monocytes are recruited into the alveolar compartment and differentiate into macrophages, mediating the inflammatory response (Huang et al. 2018). In fact, the regulation of the function of macrophages and monocytes could be a potential therapeutic option in ARDS patients (Huang et al. 2018). Interestingly, *FLT1* and other genes in the same locus (*FLT3* and the poly(A) specific ribonuclease subunit PAN3 (*PAN3*)) have been associated with monocyte counts (Astle et al. 2016). Furthermore, previous studies suggest that the VEGF-A signaling mediated by VEGFR-1 is involved in the migration of human monocytes (Barleon et al. 1996; Clauss et al. 1996; Barratt et al. 2014).

The activity of the VEGF signaling pathway has been extensively linked to the ARDS pathophysiology (Medford and Millar 2006; Barratt et al. 2014). VEGF-A is a key regulator of vascular permeability and it is involved in angiogenesis, chemotaxis, and proliferation and migration of vascular endothelial cells

(Olsson et al. 2006; Barratt et al. 2014). Studies in animal models have also supported the role of VEGF in inflammation, permeability, and fibrosis (Hamada et al. 2005). In the context of ARDS, a study in an animal model of sepsis-induced ALI suggested that the VEGF signaling is centrally involved in the response mechanism during sepsis-associated ARDS (Acosta-Herrera et al. 2015). In this sense, an increase in the VEGF gene expression was found in lungs of lipopolysaccharides-induced ALI models (Karmpaliotis et al. 2002). Additionally, the patient's VEGF-A levels have been related to increased vascular permeability in the lungs and with the fibrotic process that occurs in the fibroproliferative stage of ARDS (Medford and Millar 2006; Barratt et al. 2014; Murray et al. 2017). Interestingly, it has been reported that the soluble form encoded by *FTL1*, known as sFLT-1, sequesters VEGF-A and inhibits its biological function, acting as a competitive inhibitor of VEGF-A (Kendall and Thomas 1993). In this sense, high plasma levels of sFLT1 have been associated with organ dysfunction and sepsis severity in ICU patients (Shapiro et al. 2010; Hou et al. 2017). Accordingly, increased sFLT1 levels have been detected in the bronchoalveolar lavage from patients with ARDS (Perkins et al. 2005).

Our RNA sequencing analysis in control subjects revealed a high expression of *FLT1* isoforms in the lungs, including the forms encoding the transmembrane receptor and the soluble form. Moreover, results of the gene expression analysis in peripheral blood from ICU patients revealed a higher expression of *FLT1* (most likely of the transmembrane receptor) in peripheral blood from ARDS patients compared to other critically ill patient groups. However, the expression of *VEGFA* did not vary significantly among groups. Accordingly, a previous study supported that VEGF-A levels in pulmonary edema were reduced both in ARDS and in hydrostatic pulmonary edema, without finding significant differences between them (Ware et al. 2005). Based on the evidence, we speculate that protective genetic variants within the intron 10 from *FLT1* could be silencing the promoter activity of this gene. This would lead to the decreased expression of the transmembrane receptor and, consequently, to the reduction of the VEGF signaling activity. As a result, those pathological events triggered by VEGF during ARDS would be dimmed. However, further studies would be necessary to disentangle the biological relation between the reported genetic variants of *FLT1* and the pathophysiology of the syndrome.

As previously indicated, there is a lack of specific treatments for ARDS patients. Despite the central role of the VEGF signaling pathway in ARDS (Medford and Millar 2006; Barratt et al. 2014), the mechanisms by which these activities influence the pathophysiology of the syndrome remain unclear, which makes the development of new therapies particularly complicated. Only one clinical trial targeting VEGF has been reported, although it was retired because of a lack of funding (ClinicalTrials.gov identifier: NCT01314066). Interestingly, many of the drugs commercialized to treat cancer target the VEGF signaling pathway, although they are extremely invasive and would not be suitable for patients with ARDS. Conversely, there are two safe and non-invasive drugs targeting VEGFR-1 that could be repurposed for ARDS: nintenadib and itraconazole. The former is a tyrosine kinase inhibitor currently used to treat idiopathic pulmonary fibrosis that blocks the autophosphorylation of the VEGF receptors and the downstream signaling cascades (Richeldi et al. 2014; Wollin et al. 2015; Hajari Case and Johnson 2017). Given the important fibrotic process occurring during the fibroproliferative phase of ARDS, this drug could have beneficial effects on the recovering after the acute stage of ARDS. Accordingly, Li and colleagues reported that nintenadib reduced the epithelial-mesenchymal transition caused by MV under a regime of high tidal volume and the pulmonary fibrosis after mild ARDS induced by bleomycin (Li et al. 2017). With respect to itraconazole, it is a drug used to treat fungal infections, including severe blastomycosis that can lead to ARDS (Smith and Kauffman 2010; Schwartz et al. 2016). Although other antifungals have been tested in ARDS with no satisfactory results (Thompson 2000), itraconazole could be a promising alternative, since this safe drug inhibits the glycosylation of VEGFR-1 and VEGFR-2 and, hence, their trafficking and signaling (Nacev et al. 2011).

4.3. Bacterial lung dysbiosis as a prognostic marker in non-pulmonary sepsis

In line with the discussion about the identification of a novel genetic factor of ARDS susceptibility, the identification of effective prognostic markers in critically ill patients is also promising for clinical practice. Nowadays, the prognosis of critical care patients is based on the use of severity scores, such as the simplified acute physiology score (SAPS), the APACHE II score, which is usually calculated just at patient admission, and the sequential organ failure assessment (SOFA) score (Vincent et al. 2010; Giamarellos-Bourboulis et al. 2012; Singer et al. 2016). Nevertheless, these score systems are limited and should be updated, being necessary the identification of novel prognostic markers (Vincent et al. 2010). For that purpose, in Chapter 3, we have assessed the lung microbiome in a subset of the patients with sepsis described in the discovery phase of the GWAS, with the aim of studying the implication of the bacterial diversity in ICU mortality by using NGS targeting the 16S rRNA V4 region. Specifically, we collected lung aspirates from 36 patients with non-pulmonary sepsis at three different collection times and found a significant reduction of the relative bacterial abundance of the lung in deceased patients with respect to survivors, even during the first 8 h after sepsis diagnosis. These findings agree with those of other studies where microbiome shifts had already been linked to mortality in patients with other infectious and respiratory diseases (Shimizu et al. 2011; Molyneaux et al. 2014; Lamarche et al. 2018; O'Dwyer et al. 2019). In fact, the predictive value of the diversity index for ICU mortality in our data was surprisingly elevated (of 86.5%), higher than the one obtained with the commonly used APACHE II score (72.7%). Based on the evidence, the reduction of the bacterial diversity in lung could be an early prognostic marker for patients with non-pulmonary sepsis. However, one must take into account that a validation of the prediction is lacking from our study. Therefore, further studies in independent samples should be performed to obtain a more precise estimate of the prediction.

The comparison of the microbiome profiles revealed five bacterial genera most likely to explain the differences between deceased and survival patients. Among them, the genus *Proteus* was significantly enriched in lung aspirates from deceased individuals, while the genera *Streptococcus*, *Prevotella*, *Veillonella* and *Leptotrichia* were more abundant in survivors. *Proteus* species belong to the *Enterobacteriaceae* family and some of them are normal commensals of the intestinal flora, although their abundance is low (Yatsunenko et al. 2012; Hamilton et al. 2018). When the healthy host-microorganism balance is altered, *Proteus* spp have been related to diseases including urinary tract infections (Schaffer and Pearson 2015), intestinal diseases (Hamilton et al. 2018), bacteremia (Chen et al. 2012), and ventilator-associated pneumonia (Xia et al. 2015). Our findings agree with previous studies reporting an extravasation of gut bacteria into the lungs of critical patients (Dickson et al. 2016; Mukherjee and Hanidziar 2018). This evidence supports the existence of a crosstalk between the lung and other organs that affects the severity of sepsis, justifying the focus on patients with non-pulmonary sepsis to assess the extrapulmonary impact on pulmonary homeostasis. Additionally, Panzer and colleagues revealed an enrichment of *Enterobacteriaceae* species linked to ARDS development and to the severity of the injury (Panzer et al. 2018). To our knowledge, for the first time, our results link these observations to the prognosis of critical care patients.

Conversely, *Prevotella* spp, *Veillonella* spp, and *Streptococcus* spp, which were found to be significantly reduced in lung aspirates of deceased patients, have been reported as the most abundant bacterial genera of the healthy low respiratory tract and the oral cavity (Dickson et al. 2017). Accordingly, previous studies have also detected a reduction in the lung abundance of these bacterial genera (mainly of *Prevotella* spp) in other respiratory conditions such as asthma, pneumonia in the elderly, and chronic obstructive pulmonary disease (Park et al. 2014; de Steenhuijsen Piters et al. 2016; Yadava et al. 2016; Moffatt and Cookson 2017). These bacterial genera could have an important role in the immune response during critical illness, since *Streptococcus*, *Veillonella* and *Prevotella* spp have been related to less airway inflammation, and *Prevotella* spp could be involved in homeostatic processes that regulate pulmonary immune responses (Huffnagle et al. 2017; Zemanick et al. 2017).

The implication of microbial shifts in the dysregulation of the immune response in humans has been well described, supporting the central role of the host-microbiome interactions (Rooks and Garrett 2016; Belkaid and Harrison 2017). In this sense, previous studies have linked host genetic variants within immunity-related genes with the microbiome composition (Benson et al. 2010; Blekhman et al. 2015). Accordingly, numerous genetic variants have been related to an increased risk of bacterial infections (Boyd et al. 2014). For example, Li and colleagues revealed genetic polymorphisms associated with the interindividual variation of cytokine responses to specific pathogens (Li et al. 2016). Additionally, Lee and colleagues reported common alleles associated with effects on inter-individual variation in pathogen sensing and suggested that the pathogen-sensing pathway could have an important role in inflammatory diseases (Lee et al. 2014). Furthermore, many of the genes that have been previously associated with sepsis and ARDS are involved in immune response (Giamarellos-Bourboulis and Opal 2016; Guillén-Guío et al. 2016). Based on all this evidence, further studies linking the host genetics and the microbiome could help to improve the knowledge of the physiopathology of these critical care conditions.

4.4. The study of the genetic ancestry in Canary Islanders as an approach to evidence novel risk factors in critical illness

Most of genetic association studies in ARDS have been performed in European populations. Therefore, many genetic risks for ARDS that are more prevalent in other ethnicities might remain undiscovered. This is especially important for populations with recent African ancestry, since this ancestry has been linked to large disparities in diverse complex diseases such as respiratory and critical illnesses (National Research Council 2004; Ness et al. 2004; Kumar et al. 2010; Flores et al. 2012; Rumpel et al. 2012; Soto et al. 2013; Vergara et al. 2013; Hernandez-Pacheco et al. 2016). In Chapter 4, we reported the results of the largest and more detailed genomic characterization of the current inhabitants of the Canary Islands, a southern European population with a recent African admixture. Based on SNP array data and WGS (30X), we estimated a high percentage of African descent of their genome (up to 34%), evidenced signals of genetic isolation and of adaptation, and assessed the implications of the admixture in disease.

As a result of our analyses, we calculated that the last African admixture in this population occurred ~14 generations ago. Additionally, we identified genomic signals of inbreeding, reflecting the historical isolation of the inhabitants from El Hierro and La Gomera, the two smallest island populations that were analyzed. This is especially relevant in the context of disease, since inbreeding can lead to an increase in the allelic frequency of deleterious recessive variants due to the increased homozygosity rate, as has been described for hypertension (Rudan et al. 2003), schizophrenia (Lencz et al. 2007), Alzheimer disease (Ghani et al. 2015), thyroid cancer (Thomsen et al. 2016), and quantitative traits such as systolic and diastolic blood pressure, LDL cholesterol, and forced expiratory flow (Campbell et al. 2007). Linked to the existence of founder mutations leading to monogenic diseases (García-Villarreal et al. 2000; Lorenzo et al. 2006; Castella et al. 2011; Rodríguez-Esparragón et al. 2017), this makes the Canary Islands population attractive for subsequent genetic studies of disease. Besides, these results could be used to develop a biogeographical map of homozygosity (as a proxy for genetic risks) for the inhabitants of the islands, especially the smallest ones, which would be helpful for the Healthcare system, for example, to prioritize carrier screenings of monogenic diseases.

Furthermore, local ancestry analyses revealed eight regions with large ancestry deviations that contained genes related to prevalent diseases, especially in the Canary Islands population, such as asthma or diabetes (Sánchez-Lerma et al. 2009; Marcelino-Rodríguez et al. 2016), and genes linked to renal and neuropsychiatric diseases, as well as to infection response and to SARS, a condition that implies the occurrence of respiratory failure and from which about 25% of patients progress to ARDS (Lew et al. 2003). Thus, further genetic studies focused on these regions are projected in order to reveal novel genetic variants associated with diseases. Based on this, we have assessed the results of the discovery phase of the sepsis-associated ARDS GWAS for these regions of interest focusing only on the Canary Islander patients. Although no signals reached the Bonferroni threshold considering all independent SNPs within the African deviated regions (p<1.19x10^{-6}), unpublished results revealed three independent SNPs with a suggestive association with sepsis-associated ARDS (p<5.0x10^{-4}), where effect alleles conferred protection from the syndrome. The best ranking SNP (rs4954479) was an intronic variant of the thrombospondin type 1 domain containing 7B (*THSD7B*) gene, one of the genes flanking the lead SNP of the EUR-related peak in chr2, as revealed in Chapter 4, that has been previously associated with pulmonary function in the UK Biobank (Kichaev et al. 2019)(http://www.nealelab.is/uk-biobank/). The second ranked SNP (rs9592430) was located in the intergenic region between the protocadherin (PCDH) 20 *(PCDH20)* and *PCDH9* genes. Interestingly, in the same chapter, we described putative selective signals in this intergenic region, which also contained variants previously associated with asthma (Ferreira et al. 2011). This agrees with the fact that genetic variants linked to inflammatory diseases in European populations are significantly enriched in signatures of positive selection (Raj et al. 2013). Additionally, another intronic variant (rs2766532) within the FKBP Prolyl Isomerase 5 (*FKBP5*) gene also ranked high in the GWAS results. *FKBP5* encodes a member of the immunophilin protein family that plays a major role in immunoregulation and has been previously associated with asthma, with eosinophil, leukocyte, lymphocyte, and monocyte counts, and with lung function in the UK Biobank (Astle et al. 2016; Ferreira et al. 2019; Kichaev et al. 2019)(http://www.nealelab.is/uk-biobank/), constituting a truly interesting candidate for further genetic studies in ARDS.

In addition to *PCDH20-PCDH9*, the genomic regions identified in Chapter 4 included other putative signals of natural selection. One can anticipate that the initial settlement of the Canary Islands by aborigines was accompanied with a process of adaptation to particular climatic conditions and pathogens, which usually entails frequency shifts in genetic variants (Sabeti et al. 2006; Novembre and Di Rienzo 2009; Vasseur and Quintana-Murci 2013). In this sense, the peak in chromosome 6 includes a well-recognized target of selection: the HLA region. HLA contains genes that are robustly associated with many traits, including asthma and the response to infections (Thomas et al. 2009; Galanter et al. 2014; Sanchez-Mazas et al. 2017). Finally, based on the ancestry block length estimates, we obtained the average number of blocks for a Canarian population haploid genome (i.e. 276 ancestry blocks), laying the foundation for performing future admixture mapping studies in this population that will allow to unravel novel disease risk factors. Future studies will need to assess this estimate considering a varying number of chromosomes, likely through simulation studies, and a larger representation of the population diversity.

4.5. Strengths and limitations

This thesis has several strengths that have allowed us to provide novel and robust insights into the pathophysiology of ARDS through different approaches, including human genomics, metagenomics, and a genetic ancestry study. Firstly, the selection of the donors in all studies has been systematic, and detailed information has been collected from them. In all GWAS stages, patients with clinically-characterized sepsis were included in the study and followed up collecting signs of aggravation, including the development of ARDS according to the Berlin definition. Data from gender, age, APACHE II scores, and sources of infection, among other demographic and clinical parameters, were collected from all patients. The metagenomic study used a subset of these donors, from a single center and a single hospital service, in order to control potential environmental differences between sites. With this, we tried to ensure a homogeneous sample of mechanically-ventilated patients with non-pulmonary sepsis that were all under the same environmental conditions. Additionally, to limit the effects of recent migrations, we ensured the donors used in Chapter 4 were selected for having two generations of ancestors born on the same island. In all studies, individuals with a high degree of kinship were excluded based on genetic estimates.

Secondly, we report the results of the first GWAS of sepsis-associated ARDS published to date, where we utilized a SNP array designed for European population and assessed almost eight million of imputed genetic variants from well-characterized European patients with sepsis. As part of this study, we performed genetic association analyses, after robust quality controls, followed by a replication stage to validate the results. We also conducted complementary gene expression and functional analyses that strongly supported the important role of *FLT1* and its genetic variants in ARDS pathophysiology. Thirdly, as a result of the metagenomic study, the reduction of the lung microbial diversity was linked to the mortality by sepsis in ICUs, providing a potential early prognostic marker for patients with non-pulmonary sepsis. We used an NGS approach that allowed to infer bacteria in all samples analyzed, overcoming the bias derived from microbiological cultures and the limitations of the microbial characterization of infections in sepsis patients (40-60% of microbiological cultures are negative) (de Prost et al. 2013). Remarkably, sensitivity analyses were included in both the GWAS and the 16S rRNA metagenomic study to assess the effects of confounding factors. Finally, using SNP array data and WGS analyses, we identified genomic regions enriched in African or European alleles that harbored signals of natural selection and links to disease. As a main strength, we used simulations to assess the significance of the signals of selection detected, for which we had to estimate the effective number of aborigines in pre-European times based on WGS data available from the literature and corresponding to an individual from Tenerife (Rodríguez-Varela et al. 2017).

We acknowledge this thesis has also a number of limitations. Firstly, the sample size utilized in these studies is limited, mainly in the 16S rRNA metagenomic study, restricting the statistical power of the analyses. In the GWAS, this translates into limitations to detect low frequency variants and SNPs with subtle effects. In the metagenomic study, the reduced sample size together with the absence of a validation stage implies that further studies on independent samples are necessary, mainly to optimally assess the predictive value of the lung bacterial dysbiosis. In the genetic ancestry study, the main limitation with respect to the sample was the absence of a proper NAF dataset with higher marker resolution that allowed the analysis of a greater number of genetic markers after overlapping the study samples with the reference datasets. Therefore, future studies should include genetic data of NAF individuals obtained from WGS or SNP arrays with a larger number of markers, with the aim of optimizing the overlap with the other population datasets and, hence, of improving the local ancestry estimation.

Secondly, only European individuals were included in our GWAS and in the 16S rRNA metagenomic study, and populations of diverse ethnicities should be also assessed. In this sense, the evaluation of recently admixed populations is also a useful option in the context of disease studies. Besides, other causes of ARDS development should also be considered to validate our association results. In addition, the technology used in the metagenomic study is limited in terms of detection of bacterial species and strains (only allows confident detection at the genus level) and of evaluation of their functionality. NGS technology is also limited in terms of DNA extraction. We opted for a DNA extraction method that allowed us to recuperate the greatest quantity of bacterial DNA based on a previous laboratory comparison of different bacterial DNA extraction kits. Besides, DNA amplification is also a critical step, because the primers utilized have a better base pair complementarity with the sequence of some bacteria than with others, in addition to the bias introduced by the DNA polymerase. Additionally, despite the use of antibiotics did not change the observations in the lung aspirates, the heterogeneity in the specific antibiotics used per patient was not modelled in the statistical analyses due to the small sample size. Another limitation is that the number of 16S rRNA copies, which is known to vary among microbes, was not controlled in the analyses. Thus, bacterial abundances should be evaluated with caution. Furthermore, we did not analyze the DNA of viruses and fungi that may be present in the lung aspirates of our study. Finally, given the extension of the genomic regions identified in the genetic ancestry study, those genes that we highlight in the chapter should also be considered with caution. Besides, we did not analyze the local ancestry in the centrosomes, where we may be missing important information. Moreover, the number of ancestry blocks that were estimated based on average block length measures will necessitate simulation studies to reach a more accurate estimation.

4.6. Future directions

Despite the advances reported in this thesis, further studies that support and/or complement our results will be necessary. These should include larger sample sizes and individuals of other ethnicities. In this sense, the laboratory is currently recruiting more cases of sepsis that will be genotyped and exome-sequenced and will also undergo future metagenomic studies. Additionally, other triggering factors linked to ARDS development, rather than sepsis, should be considered. As we have described in Chapter 2, we accessed the only publicly available GWAS of ARDS data entailing patients from an insult other than sepsis, consisting on trauma-associated ARDS patients (Christie et al. 2012). However, none of the *FLT1* variants that we reported as significantly associated with sepsis-associated ARDS were present in the reference panel used by that study for the imputation. Additionally, the use of NGS technologies, such as WGS or WES, will help to assess the effects of rare variants that, due to the limited statistical power of the GWAS, remained obscure in this thesis. Furthermore, because of the cancerous nature of the A549 cell line, additional luciferase reporter assays in primary human ATII cells should be performed to evaluate the implications of *FLT1* variants in the gene promoter activity.

On the other hand, alternative metagenomic approaches would be necessary to validate our results linking the lung microbiome decreased with the ICU mortality. Strategies that are being currently assessed in the laboratory in the context of critical illness include shotgun sequencing and third-generation sequencing. The use of a shotgun approach would also allow the taxonomic classification at (sub)species level, without the need for prior amplification of DNA or without relying on the 16S rRNA gene. Additionally, the use of third-generation sequencers, as those marketed by the company Oxford Nanopore Technologies, will allow to sequence larger DNA fragments in real time, permitting the identification of bacteria at species level and reducing the time of analyses. At this moment, the laboratory is trying to reproduce the analyses described in Chapter 3 by using the MinION sequencer and independent taxa classifiers, targeting the full 16S rRNA gene. In addition to the speed of analysis, this is a portable sequencer that does not require large facilities or enormous amounts of DNA material, presenting a great potential for clinical practice. All these strategies would also facilitate the study of viruses, fungi, and bacterial pathogenic elements if adapted to alternative amplicon-based or shotgun applications. Moreover, additional studies linking the lung microbiome with the genetics of patients with sepsis would be interesting to further help understand the pathophysiology of sepsis and ARDS.

Finally, the genomic characterization of Canary Islanders offers a range of possibilities for subsequent studies. Firstly, the identification of genomic regions enriched in African alleles in this population opens the door to future fine mapping studies in those specific sites to disentangle the genetic variation associated with specific diseases, including sepsis and ARDS. Additionally, the estimation of the number of ancestry blocks provides the basis for performing admixture mapping studies of disease. Accordingly, we are currently conducting an admixture mapping study of sepsis in the Canary Islander ICU patients, with the aim of identifying genomic regions where the genetic ancestry and the syndrome are linked.

5. Conclusions

5. Conclusions

1) Most genes associated with ARDS susceptibility or outcomes have been revealed based on a candidate gene approach. These genes are mainly involved in the immune response and vascular permeability.

2) The candidate genes associated with ARDS that have the largest number of independent study findings are: *ACE, IL1RN, IL6, IL10, MBL2, NAMPT,* and *VEGFA*.

3) Common genetic variants within the *FLT1* gene are associated with susceptibility to sepsis-associated ARDS.

4) The expression of *FLT1* gene in peripheral blood differed significantly among critical care patient groups. ARDS patients showed the highest average expression of the *FLT1* gene.

5) *In silico* and *in vitro* analyses of the function of the *FLT1* variants associated with ARDS supported their transcriptional role by affecting the regulation of the *FLT1* promoter. The alleles with protective effects in ARDS reduced the *FLT1* promoter activity in a monocyte cell line.

6) The bacterial diversity in lung aspirates was reduced within 8 h of diagnosis in patients with non-pulmonary sepsis who died in the ICU compared to those who survived.

7) The bacterial dysbiosis of lung aspirates from patients with non-pulmonary sepsis had a higher predictive value of ICU mortality than the APACHE II score in our study.

8) Lung aspirates from the patients with non-pulmonary sepsis deceased in the ICU presented commensal gut bacterial genera and were depleted in healthy lung bacterial genera.

9) The genome of present-day Canary Islanders harbors eight regions with large local ancestry deviations that contain putative signals of selection. These regions are enriched in genes related to prevalent diseases, to the response to infections and to SARS, among many other traits.

10) The last African admixture in the Canary Islanders was estimated to take place ~14 generations ago. Based on the admixture estimates in this population, we calculated a total of 276 ancestry blocks on average per haploid genome. This provides the basis for designing admixture mapping studies of complex traits in the Canary Islands, including sepsis and ARDS.

6. References

6. References

1000 Genomes Project Consortium RA, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. Nature 526:68–74.

Abdulnour REE, Gunderson T, Barkas I, Timmons JY, Barnig C, Gong M, Kor DJ, Gajic O, Talmor D, Carter RE, et al. 2018. Early intravascular events are associated with development of acute respiratory distress syndrome a substudy of the lips-A clinical trial. Am. J. Respir. Crit. Care Med. 197:1575–1585.

de Abreu Galindo J, Cioranescu A. 1977. Historia de la conquista de las Siete Islas de Canaria. Goya Edici. Santa Cruz de Tenerife.

Acosta-Herrera M, Lorenzo-Diaz F, Pino-Yanes M, Corrales A, Valladares F, Klassert TE, Valladares B, Slevogt H, Ma S-F, Villar J, et al. 2015. Lung Transcriptomics during Protective Ventilatory Support in Sepsis-Induced Acute Lung Injury. PLoS One 10:e0132296.

Acosta-Herrera M, Pino-Yanes M, Perez-Mendez L, Villar J, Flores C. 2014. Assessing the quality of studies supporting genetic susceptibility and outcomes of ARDS. Front. Genet. 5:20.

Acute Respiratory Distress Syndrome Network, Brower RG, Matthay MA, Morris A, Schoenfeld D, Thompson BT, Wheeler A. 2000. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N. Engl. J. Med. 342:1301–1308.

Aggarwal NR, King LS, D'Alessio FR. 2014. Diverse macrophage populations mediate acute lung inflammation and resolution. Am. J. Physiol. Lung Cell. Mol. Physiol. 306.

Agnese DM, Calvano JE, Hahm SJ, Coyle SM, Corbett SA, Calvano SE, Lowry SF. 2002. Human toll-like receptor 4 mutations but not CD14 polymorphisms are associated with an increased risk of gram-negative infections. J. Infect. Dis. 186:1522–1525.

Aisiku IP, Yamal JM, Doshi P, Benoit JS, Gopinath S, Goodman JC, Robertson CS. 2016. Plasma cytokines IL-6, IL-8, and IL-10 are associated with the development of acute respiratory distress syndrome in patients with severe traumatic brain injury. Crit. Care 20.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Althani AA, Marei HE, Hamdi WS, Nasrallah GK, El Zowalaty ME, Al Khodor S, Al-Asmakh M, Abdel-Aziz H, Cenciarelli C. 2016. Human Microbiome and its Association With Health and Diseases. J. Cell. Physiol. 231:1688–1694.

ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L, Slutsky AS. 2012. Acute respiratory distress syndrome: the Berlin Definition. JAMA 307:2526–2533.

Ashbaugh D, Bigelow D, Petty T, Levine B. 1967. Acute respiratory distress in adults. Lancet 2:319–23.

Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, et al. 2016. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167:1415–1429.e19.

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC, et al. 2012. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28:1359–1367.

Barbas CSV, Ísola AM, Caser EB. 2014. What is the future of acute respiratory distress syndrome after the Berlin definition? Curr. Opin. Crit. Care 20:10–16.

Barleon B, Sozzani S, Zhou D, Weich HA, Mantovani A, Marmé D. 1996. Migration of human monocytes in response to vascular endothelial growth factor (VEGF) is mediated via the VEGF receptor flt-1. Blood 87:3336–3343.

Barnato AE, Alexander SL, Linde-Zwirble WT, Angus DC. 2008. Racial variation in the incidence, care, and outcomes of severe sepsis: Analysis of population, patient, and hospital characteristics. Am. J. Respir. Crit. Care Med. 177:279–284.

Barratt S, Medford AR, Millar AB. 2014. Vascular endothelial growth factor in acute lung injury and acute respiratory distress syndrome. Respiration 87:329–342.

Belkaid Y, Harrison OJ. 2017. Homeostatic Immunity and the Microbiota. Immunity 46:562–576.

Bellingan GJ. 2002. The pulmonary physician in critical care * 6: The pathogenesis of ALI/ARDS. Thorax 57:540–546.

Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, Zhang M, Oh PL, Nehrenberg D, Hua K, et al. 2010. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc. Natl. Acad. Sci. U. S. A. 107:18933–18938.

Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, Lamy M, Legall JR, Morris A, Spragg R.

1994. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. Am. J. Respir. Crit. Care Med. 149:818–824.

Bime C, Pouladi N, Sammani S, Batai K, Casanova N, Zhou T, Kempf CL, Sun X, Camp SM, Wang T, et al. 2018. Genome-Wide Association Study in African Americans with Acute Respiratory Distress Syndrome Identifies the Selectin P Ligand Gene as a Risk Factor. Am. J. Respir. Crit. Care Med. 197:1421–1432.

Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, et al. 2015. Host genetic variation impacts microbiome composition across human body sites. Genome Biol. 16.

Block ER. 1992. Pulmonary endothelial cell pathobiology: implications for acute lung injury. Am. J. Med. Sci. 304:136–144.

Blondonnet R, Constantin JM, Sapin V, Jabaudon M. 2016. A Pathophysiologic Approach to Biomarkers in Acute Respiratory Distress Syndrome. Dis. Markers 2016:3501373.

Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, Atzmon G, Burns E, Ostrer H, Flores C, et al. 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern europe. Proc. Natl. Acad. Sci. U. S. A. 110:11791–11796.

Boyd JH, Russell JA, Fjell CD. 2014. The meta-genome of sepsis: host genetics, pathogens and the acute immune response. J. Innate Immun. 6:272–283.

Budden KF, Shukla SD, Rehman SF, Bowerman KL, Keely S, Hugenholtz P, Armstrong-James DPH, Adcock IM, Chotirmall SH, Chung KF, et al. 2019. Functional effects of the microbiota in chronic respiratory disease. Lancet Respir. Med. 7:907–920.

Bueno HA, Hernaez R, Hernandez A V. 2008. Type 2 Diabetes Mellitus and Cardiovascular Disease in Spain: A Narrative Review.Rev. Esp. Cardiol. Supl. 8:50C-58C.

Cabrera de León A, Rodríguez-Pérez M del C, del Castillo-Rodríguez JC, Brito-Díaz B, Pérez-Méndez LI, Muros de Fuentes M, Almeida-González D, Batista-Medina M, Aguirre-Jaime A. 2006. Coronary risk in the population of the Canary Islands, Spain, using the Framingham function. Med. Clin. (Barc). 126:521–526.

Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, Pericic M, Janicijevic B, Smolej-Narancic N, Polasek O, et al. 2007. Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum. Mol. Genet. 16:233–241.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7:335–336.

CARDIoGRAMplusC4D Consortium P, Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, Ingelsson E, Saleheen D, Erdmann J, et al. 2013. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. 45:25–33.

Castella M, Pujol R, Callén E, Trujillo JP, Casado JA, Gille H, Lach FP, Auerbach AD, Schindler D, Benítez J, et al. 2011. Origin, functional role, and clinical impact of fanconi anemia fanca mutations. Blood 117:3759–3769.

Cavallazzi R, Marik PE, Hirani A, Pachinburavan M, Vasu TS, Leiby BE. 2010. Association between time of admission to the ICU and mortality: A systematic review and metaanalysis. Chest 138:68–75.

Caverly LJ, Zhao J, LiPuma JJ. 2015. Cystic fibrosis lung microbiome: Opportunities to reconsider management of airway infection. Pediatr. Pulmonol. 50:S31–S38.

Chakravorty S, Helb D, Burday M, Connell N, Alland D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods 69:330–339.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, et al. 2007. Replicating genotype-phenotype associations. Nature 447:655–660.

Chen C-Y, Chen Y-H, Lu P-L, Lin W-R, Chen T-C, Lin C-Y. 2012. Proteus mirabilis urinary tract infection and bacteremia: risk factors, clinical presentation, and outcomes. J. Microbiol. Immunol. Infect. 45:228–236.

Chen Z, Hu Y, Xiong T, Chen C, Su XX, Huang Y, Zhang L. 2018. IL-10 promotes development of acute respiratory distress syndrome via inhibiting differentiation of bone marrow stem cells to alveolar type 2 epithelial cells. Eur. Rev. Med. Pharmacol. Sci. 22:6085–6092.

Christie JD, Ma S-F, Aplenc R, Li M, Lanken PN, Shah C V, Fuchs B, Albelda SM, Flores C, Garcia JGN. 2008. Variation in the myosin light chain kinase gene is associated with development of acute lung injury after major trauma. Crit. Care Med. 36:2794–2800.

Christie JD, Wurfel MM, Feng R, O'Keefe GE, Bradfield J, Ware LB, Christiani DC, Calfee CS, Cohen MJ, Matthay M, et al. 2012. Genome wide association identifies PPFIA1 as a candidate gene for acute lung injury risk following major trauma. PLoS One 7:e28268.

Clark MF, Baudouin SV. 2006. A systematic review of the quality of genetic association studies in human sepsis. Intensive Care Med. 32:1706–1712.

Clauss M, Weich H, Breier G, Knies U, Röckl W, Waltenberger J, Risau W. 1996. The vascular endothelial growth factor receptor Flt-1 mediates biological activities. Implications for a functional role of placenta growth factor in monocyte activation and chemotaxis. J. Biol. Chem. 271:17629–17634.

Cohen J. 2002. The immunopathogenesis of sepsis. Nature 420:885–891.

Columb MO, Atkinson MS. 2015. Statistical analysis: sample size and power estimations. BJA Educ. 16:159–161.

Cox MJ, Cookson WOCM, Moffatt MF. 2013. Sequencing the human microbiome in health and disease. Hum. Mol. Genet. 22.

Dean L. 2012. Clopidogrel Therapy and CYP2C19 Genotype. Bethesda: Medical Genetics Summaries.

Dehghan A. 2018. Genome-Wide Association Studies. Methods Mol. Biol. 1793:37–49.

Dickson RP. 2016. The microbiome and critical illness. Lancet. Respir. Med. 4:59–72.

Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB, Curtis JL. 2017. Bacterial Topography of the Healthy Human Lower Respiratory Tract. MBio 8.

Dickson RP, Singer BH, Newstead MW, Falkowski NR, Erb-Downward JR, Standiford TJ, Huffnagle GB. 2016. Enrichment of the lung microbiome with gut bacteria in sepsis and the acute respiratory distress syndrome. Nat. Microbiol. 1:16113.

Divangahi M, King IL, Pernet E. 2015. Alveolar macrophages and type I IFN in airway homeostasis and immunity. Trends Immunol. 36:307–314.

Dombrovskiy VY, Martin AA, Sunderram J, Paz HL. 2005. Facing the challenge: decreasing case fatality rates in severe sepsis despite increasing hospitalizations. Crit. Care Med. 33:2555–2562.

Erickson SE, Martin GS, Davis JL, Matthay MA, Eisner MD, NIH NHLBI ARDS Network. 2009. Recent trends in acute lung injury mortality: 1996-2005. Crit. Care Med. 37:1574–1579.

Esper AM, Moss M, Lewis CA, Nisbet R, Mannino DM, Martin GS. 2006. The role of infection and comorbidity: Factors that influence disparities in sepsis. Crit. Care Med. 34:2576–2582.

Eyheramendy S, Martinez FI, Manevy F, Vial C, Repetto GM. 2015. Genetic structure characterization of Chileans reflects historical immigration patterns. Nat. Commun. 6.

Fan E, Dowdy DW, Colantuoni E, Mendez-Tellez PA, Sevransky JE, Shanholtz C, Himmelfarb CRD, Desai S V., Ciesla N, Herridge MS, et al. 2014. Physical complications in acute lung injury survivors: A two-year longitudinal prospective study. Crit. Care Med. 42:849–859.

Ferreira MAR, Matheson MC, Duffy DL, Marks GB, Hui J, Le Souëf P, Danoy P, Baltic S, Nyholt DR, Jenkins M, et al. 2011. Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. Lancet 378:1006–1014.

Ferreira MAR, Mathur R, Vonk JM, Szwajda A, Brumpton B, Granell R, Brew BK, Ullemar V, Lu Y, Jiang Y, et al. 2019. Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. Am. J. Hum. Genet. 104:665–684.

Ferwerda B, McCall MBB, Alonso S, Giamarellos-Bourboulis EJ, Mouktaroudi M, Izagirre N, Syafruddin D, Kibiki G, Cristea T, Hijmans A, et al. 2007. TLR4 polymorphisms, infectious diseases, and evolutionary pressure during migration of modern humans.1. Proc. Natl. Acad. Sci. U. S. A. 104:16645–16650.

Flores C, Larruga JM, González AM, Hernández M, Pinto FM, Cabrera VM. 2001. The Origin of the Canary Island Aborigines and Their Contribution to the Modern Population: A Molecular Genetics Perspective. Curr. Anthropol. 42:749–755.

Flores C, Ma SF, Pino-Yanes M, Wade MS, Pérez-Méndez L, Kittles RA, Wang D, Papaiahgari S, Ford JG, Kumar R, et al. 2012. African ancestry is associated with asthma risk in African Americans. PLoS One 7.

Flores C, Pino-Yanes M del M, Villar J. 2008. A quality assessment of genetic association studies supporting susceptibility and outcome in acute lung injury. Crit. Care 12:R130.

Fourrier F, Chopin C, Wallaert B, Mazurier C, Mangalaboyi J, Durocher A. 1985. Compared evolution of plasma fibronectin and angiotensin-converting enzyme levels in septic ARDS. Chest 87:191–195.

Fujino N, Ota C, Takahashi T, Suzuki T, Suzuki S, Yamada M, Nagatomi R, Kondo T, Yamaya M, Kubo H. 2012. Gene expression profiles of alveolar type II cells of chronic obstructive pulmonary disease: A case-control study. BMJ Open 2.

Gajic O, Dabbagh O, Park PK, Adesanya A, Chang SY, Hou P, Anderson H, Hoth JJ, Mikkelsen ME, Gentile NT, et al. 2011. Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort study. Am. J. Respir. Crit. Care Med. 183:462–470.

Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, Nguyen EA, Drake KA, Huntsman S, Hu D, et al. 2014. Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: The Genes-environments & Admixture in Latino Americans study. J. Allergy Clin. Immunol. 134:295–305.

Gao L, Grant A, Halder I, Brower R, Sevransky J, Maloney JP, Moss M, Shanholtz C, Yates CR, Meduri GU, et al. 2006. Novel polymorphisms in the myosin light chain kinase gene confer risk for acute lung injury. Am. J. Respir. Cell Mol. Biol. 34:487–495.

García-Villarreal L, Daniels S, Shaw SH, Cotton D, Galvin M, Geskes J, Bauer P, Sierra-Hernández A, Buckler A, Tugores A. 2000. High prevalence of the very rare Wilson disease gene mutation Leu708Pro in the island of Gran Canaria (Canary Islands, Spain): A genetic and clinical study. Hepatology 32:1329–1336.

Ghani M, Reitz C, Cheng R, Vardarajan BN, Jun G, Sato C, Naj A, Rajbhandary R, Wang LS, Valladares O, et al. 2015. Association of long runs of homozygosity with Alzheimer disease among African American individuals. JAMA Neurol. 72:1313–1323.

Giamarellos-Bourboulis EJ, Norrby-Teglund A, Mylona V, Savva A, Tsangaris I, Dimopoulou I, Mouktaroudi M, Raftogiannis M, Georgitsi M, Linnér A, et al. 2012. Risk assessment in sepsis: a new prognostication rule by APACHE II score and serum soluble urokinase plasminogen activator receptor. Crit. Care 16:R149.

Giamarellos-Bourboulis EJ, Opal SM. 2016. The role of genetics and antibodies in sepsis. Ann. Transl. Med. 4:328.

Gosiewski T, Ludwig-Galezowska AH, Huminska K, Sroka-Oleksiak A, Radkowski P, Salamon D, Wojciechowicz J, Kus-Slowinska M, Bulanda M, Wolkow PP. 2017. Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy volunteers using next-generation sequencing method - the observation of DNAemia. Eur. J. Clin. Microbiol. Infect. Dis. 36:329–336.

Gray KJ, Saxena R, Karumanchi SA. 2018. Genetic predisposition to preeclampsia is conferred by fetal DNA variants near FLT1, a gene involved in the regulation of angiogenesis. Am. J. Obstet. Gynecol. 218:211–218.

Guan Y. 2014. Detecting structure of haplotypes and local ancestry. Genetics 196:625–642.

Guérin C, Reignier J, Richard J-C, Beuret P, Gacouin A, Boulain T, Mercier E, Badet M, Mercat A, Baudin O, et al. 2013. Prone positioning in severe acute respiratory distress syndrome. N. Engl. J. Med. 368:2159–2168.

Guillén-Guío B, Acosta-Herrera M, Villar J, Flores C. 2016. Genetics of Acute Respiratory Distress Syndrome. eLS. John Wiley Sons.

Guillot L, Nathan N, Tabary O, Thouvenin G, Le Rouzic P, Corvol H, Amselem S, Clement A. 2013. Alveolar epithelial cells: master regulators of lung homeostasis. Int. J. Biochem. Cell Biol. 45:2568–2573.

Guo L, Xie J, Huang Y, Pan C, Yang Y, Qiu H, Liu L. 2018. Higher PEEP improves outcomes in ARDS patients with clinically objective positive oxygenation response to PEEP: A systematic review and meta-analysis. BMC Anesthesiol. 18.

Haak BW, Wiersinga WJ. 2017. The role of the gut microbiota in sepsis. lancet. Gastroenterol. Hepatol. 2:135–143.

Hajari Case A, Johnson P. 2017. Clinical use of nintedanib in patients with idiopathic pulmonary fibrosis. BMJ Open Respir. Res. 4.

Hamada N, Kuwano K, Yamada M, Hagimoto N, Hiasa K, Egashira K, Nakashima N, Maeyama T, Yoshimi M, Nakanishi Y. 2005. Anti-vascular endothelial growth factor gene therapy attenuates lung injury and fibrosis in mice. J. Immunol. 175:1224–1231.

Hamilton AL, Kamm MA, Ng SC, Morrison M. 2018. Proteus spp. as Putative Gastrointestinal Pathogens. Clin. Microbiol. Rev. 31.

Han S, Mallampalli RK. 2015. The acute respiratory distress syndrome: from mechanism to translation. J. Immunol. 194:855–860.

Handelsman J. 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev. 68:669–685.

Hendrickson CM, Crestani B, Matthay MA. 2015. Biology and pathology of fibroproliferation following the acute respiratory distress syndrome. Intensive Care Med. 41:147–150.

Hernández-Beeftink, Guillen-Guio, Villar, Flores. 2019. Genomics and the Acute Respiratory Distress Syndrome: Current and Future Directions. Int. J. Mol. Sci. 20:4004.

Hernandez-Pacheco N, Flores C, Oh SS, Burchard EG, Pino-Yanes M. 2016. What Ancestry Can Tell Us About the Genetic Origins of Inter-Ethnic Differences in Asthma Expression. Curr. Allergy Asthma Rep. 16.

Herold S, Gabrielli NM, Vadász I. 2013. Novel concepts of acute lung injury and alveolar-capillary barrier dysfunction. Am. J. Physiol. Lung Cell. Mol. Physiol. 305.

Herrero R, Sanchez G, Lorente JA. 2018. New insights into the mechanisms of pulmonary edema in acute lung injury. Ann. Transl. Med. 6:32–32.

Hooton E. 1970. The ancient inhabitants of the Canary Islands. Kraus Repr. New York

Hou PC, Filbin MR, Wang H, Ngo L, Huang DT, Aird WC, Yealy DM, Angus DC, Kellum JA, Shapiro NI, et al. 2017. Endothelial Permeability and Hemostasis in Septic Shock: Results From the ProCESS Trial. Chest 152:22–31.

Huang X, Xiu H, Zhang S, Zhang G. 2018. The role of macrophages in the pathogenesis of ali/ards. Mediators Inflamm.

Huffnagle GB, Dickson RP, Lukacs NW. 2017. The respiratory tract microbiome and lung inflammation: a two-way street. Mucosal Immunol. 10:299–306.

Hughes M, MacKirdy FN, Ross J, Norrie J, Grant IS, Scottish Intensive Care Society. 2003. Acute respiratory distress syndrome: an audit of incidence and outcome in Scottish intensive care units. Anaesthesia 58:838–845.

Jacobs MC, Haak BW, Hugenholtz F, Wiersinga WJ. 2017. Gut microbiota and host defense in critical illness. Curr. Opin. Crit. Care 23:257–263.

Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. J. Clin. Microbiol. 45:2761–2764.

Jones JM, Fingar KR, Miller MA, Coffey R, Barrett M, Flottemesch T, Heslin KC, Gray DT, Moy E. 2017. Racial Disparities in Sepsis-Related In-Hospital Mortality: Using a Broad Case Capture Method and Multivariate Controls for Clinical and Hospital Variables, 2004-2013. Crit. Care Med. 45:e1209–e1217.

Juliá-Serdá G, Cabrera-Navarro P, Acosta-Fernández O, Martín-Pérez P, Losada-Cabrera P, García-Bello MA, Carrillo-Díaz T, Antó-Boqué J. 2011. High prevalence of asthma and atopy in the Canary Islands, Spain. Int. J. Tuberc. Lung Dis. 15:536–541.

Kangelaris KN, Sapru A, Calfee CS, Liu KD, Pawlikowska L, Witte JS, Vittinghoff E, Zhuo H, Auerbach AD, Ziv E, et al. 2012. The association between a Darc gene polymorphism and clinical outcomes in African American patients with acute lung injury. Chest 141:1160–1169.

Karmpaliotis D, Kosmidou I, Ingenito EP, Hong K, Malhotra A, Sunday ME, Haley KJ. 2002. Angiogenic growth factors in the pathophysiology of a murine model of acute lung injury. Am. J. Physiol. Lung Cell. Mol. Physiol. 283:L585-95.

Kendall RL, Thomas KA. 1993. Inhibition of vascular endothelial cell growth factor activity by an endogenously encoded soluble receptor. Proc. Natl. Acad. Sci. U. S. A. 90:10705–10709.

Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL. 2019. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. 104:65–75.

Kim JY, Hildebrandt MAT, Pu X, Ye Y, Correa AM, Vaporciyan AA, Wu X, Roth JA. 2012. Variations in the vascular endothelial growth factor pathway predict pulmonary complications. Ann. Thorac. Surg. 94:1079-84; discussion 1084-5.

Konta A, Ozaki K, Sakata Y, Takahashi A, Morizono T, Suna S, Onouchi Y, Tsunoda T, Kubo M, Komuro I, et al. 2016. A functional SNP in FLT1 increases risk of coronary artery disease in a Japanese population. J. Hum. Genet. 61:435–441.

Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, Colangelo L, Galanter J, Gignoux C, Hu D, Sen S, et al. 2010. Genetic ancestry in lung-function predictions. N. Engl. J. Med. 363:321–330.

Lamarche D, Johnstone J, Zytaruk N, Clarke F, Hand L, Loukov D, Szamosi JC, Rossi L, Schenck LP, Verschoor CP, et al. 2018. Microbial dysbiosis and mortality during mechanical ventilation: a prospective observational study. Respir. Res. 19:245.

Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, Imboywa SH, Chipendo PI, Ran FA, Slowikowski K, et al. 2014. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science 343:1246980.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team DC, Christiani DC, Wurfel MM, Lin X. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. 91:224–237.

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK. 2007. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc. Natl. Acad. Sci. U. S. A. 104:19942–19947.

Lew TWK, Kwek T-K, Tai D, Earnest A, Loo S, Singh K, Kwan KM, Chan Y, Yim CF, Bek SL, et al. 2003. Acute respiratory distress syndrome in critically ill patients with severe acute respiratory syndrome. JAMA 290:374–380.

Li G, Malinchoc M, Cartin-Ceba R, Venkata C V., Kor DJ, Peters SG, Hubmayr RD, Gajic O. 2011. Eight-year trend of acute respiratory distress syndrome: A population-based study in Olmsted County, Minnesota. Am. J. Respir. Crit. Care Med. 183:59–66.

Li J, Wang S, Barone J, Malone B. 2009. Warfarin pharmacogenomics. P T 34:422–427.

Li LF, Kao KC, Liu YY, Lin CW, Chen NH, Lee CS, Wang CW, Yang CT. 2017. Nintedanib reduces ventilation-augmented bleomycin-induced epithelial–mesenchymal transition and lung fibrosis through suppression of the Src pathway. J. Cell. Mol. Med. 21:2937–2949.

Li Y, Oosting M, Deelen P, Ricaño-Ponce I, Smeekens S, Jaeger M, Matzaraki V, Swertz MA, Xavier RJ, Franke L, et al. 2016. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. Nat. Med. 22:952–960.

Lloyd-Price J, Abu-Ali G, Huttenhower C. 2016. The healthy human microbiome. Genome Med. 8.

Lobo-Cabrera M. 1993. No TitleLa esclavitud en Fuerteventura en los Siglos XVI y XVII. V Jornadas Estud. sobre Fuerteventura y Lanzarote 1:13–40.

Lorenz E, Mira JP, Frees KL, Schwartz DA. 2002. Relevance of mutations in the TLR4 receptor in patients with gram-negative septic shock. Arch. Intern. Med. 162:1028–1032.

Lorenzo V, Alvarez A, Torres A, Torregrosa V, Hernández D, Salido E. 2006. Presentation and role of transplantation in adult patients with type 1 primary hyperoxaluria and the I244T AGXT mutation: Single-center experience. Kidney Int. 70:1115–1119.

Lorenzo V, Torres A, Salido E. 2014. Primary hyperoxaluria. Nefrologia 34:398–412.

Maca-Meyer N, Arnay M, Rando JC, Flores C, González AM, Cabrera VM, Larruga JM. 2004. Ancient mtDNA analysis and the origin of the Guanches. Eur. J. Hum. Genet. 12:155–162.

Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. 2011. What can exome sequencing do for you? J. Med. Genet. 48:580–589.

Mani A. 2017. Local Ancestry Association, Admixture Mapping, and Ongoing Challenges. Circ. Cardiovasc. Genet. 10.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93:278–288.

Marcelino-Rodríguez I, Elosua R, Pérez M del CR, Fernández-Bergés D, Guembe MJ, Alonso TV, Félix FJ, González DA, Ortiz-Marrón H, Rigo F, et al. 2016. On the problem of type 2 diabetes-related mortality in the Canary Islands, Spain. The DARIOS Study. Diabetes Res. Clin. Pract. 111:74–82.

Marigorta UM, Rodríguez JA, Gibson G, Navarro A. 2018. Replicability and Prediction: Lessons and

Challenges from GWAS. Trends Genet. 34:504–517.

Martin GS, Mannino DM, Eaton S, Moss M. 2003. The epidemiology of sepsis in the United States from 1979 through 2000. N. Engl. J. Med. 348:1546–1554.

Mayr FB, Yende S, Linde-Zwirble WT, Peck-Palmer OM, Barnato AE, Weissfeld LA, Angus DC. 2010. Infection rate and acute organ dysfunction risk as explanations for racial differences in severe sepsis. JAMA - J. Am. Med. Assoc. 303:2495–2503.

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48:1279–1283.

McDonald D, Ackermann G, Khailova L, Baird C, Heyland D, Kozar R, Lemieux M, Derenski K, King J, Vis-Kampen C, et al. 2016. Extreme Dysbiosis of the Microbiome in Critical Illness. mSphere 1.

McGinnis R, Steinthorsdottir V, Williams NO, Thorleifsson G, Shooter S, Hjartardottir S, Bumpstead S, Stefansdottir L, Hildyard L, Sigurdsson JK, et al. 2017. Variants in the fetal genome near FLT1 are associated with risk of preeclampsia. Nat. Genet. 49:1255–1260.

Medford ARL, Millar AB. 2006. Vascular endothelial growth factor (VEGF) in acute lung injury (ALI) and acute respiratory distress syndrome (ARDS): paradox or paradigm? Thorax 61:621–626.

Meduri GU, Headley S, Kohler G, Stentz F, Tolley E, Umberger R, Leeper K. 1995. Persistent elevation of inflammatory cytokines predicts a poor outcome in ARDS. Plasma IL-1 beta and IL-6 levels are consistent and efficient predictors of outcome over time. Chest 107:1062–1073.

Meyer D, Vitor VR, Bitarello BD, Débora DY, Nunes K. 2018. A genomic perspective on HLA evolution. Immunogenetics 70:5–27.

Meyer NJ, Daye ZJ, Rushefski M, Aplenc R, Lanken PN, Shashaty MGS, Christie JD, Feng R. 2012. SNP-set analysis replicates acute lung injury genetic risk factors. BMC Med. Genet. 13.

Mikkelsen ME, Shull WH, Biester RC, Taichman DB, Lynch S, Demissie E, Hansen-Flaschen J, Christie JD. 2009. Cognitive, mood and quality of life impairments in a select population of ARDS survivors. Respirology 14:76–82.

Mizrahi-Man O, Davenport ER, Gilad Y. 2013. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. PLoS One 8.

Moffatt MF, Cookson WO. 2017. The lung microbiome in health and disease. Clin. Med. 17:525–529.

Moltke I, Grarup N, Jørgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, Korneliussen TS, Andersen MA, Nielsen TS, Krarup NT, et al. 2014. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature 512:190–193.

Molyneaux PL, Cox MJ, Willis-Owen SAG, Mallia P, Russell KE, Russell A-M, Murphy E, Johnston SL, Schwartz DA, Wells AU, et al. 2014. The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis. Am. J. Respir. Crit. Care Med. 190:906–913.

Moss M, Mannino DM. 2002. Race and gender differences in acute respiratory distress syndrome deaths in the United States: an analysis of multiple-cause mortality data (1979- 1996). Crit. Care Med. 30:1679–1685.

Mukherjee S, Hanidziar D. 2018. More of the Gut in the Lung: How Two Microbiomes Meet in ARDS. Yale J. Biol. Med. 91:143–149.

Murray LA, Habiel DM, Hohmann M, Camelo A, Shang H, Zhou Y, Coelho AL, Peng X, Gulati M, Crestani B, et al. 2017. Antifibrotic role of vascular endothelial growth factor in pulmonary fibrosis. JCI insight 2.

Nacev BA, Grassi P, Dell A, Haslam SM, Liu JO. 2011. The Antifungal Drug Itraconazole Inhibits Vascular Endothelial Growth Factor Receptor 2 (VEGFR2) Glycosylation, Trafficking, and Signaling in Endothelial Cells. J. Biol. Chem. 286:44045–44056.

National Research Council. 2004. Critical Perspectives on Racial and Ethnic Differences in Health in Late Life. Washington, DC: The National Academies Press

Nayfach S, Pollard KS. 2016. Toward Accurate and Quantitative Comparative Metagenomics. Cell 166:1103–1116.

Ness RB, Haggerty CL, Harger G, Ferrell R. 2004. Differential distribution of allelic variants in cytokine genes among African Americans and White Americans. Am. J. Epidemiol. 160:1033–1038.

Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. Nat. Rev. Genet. 10:745–755.

O'Dwyer DN, Ashley SL, Gurczynski SJ, Xia M, Wilke C, Falkowski NR, Norman KC, Arnold KB, Huffnagle GB, Salisbury ML, et al. 2019. Lung Microbiota Contribute to Pulmonary Inflammation and Disease Progression in Pulmonary Fibrosis. Am. J. Respir. Crit. Care Med. 199:1127–1138.

Olsson A-K, Dimberg A, Kreuger J, Claesson-Welsh L. 2006. VEGF receptor signalling - in control of vascular function. Nat. Rev. Mol. Cell Biol. 7:359–371.

Onrubia Pintado J. 1987. Les cultures préhistoriques des Îles Canaries, état de la question. Anthropologie 91:653–678.

Padhukasahasram B. 2014. Inferring ancestry from population genomic data and its applications. Front. Genet. 5.

Panzer AR, Lynch SV., Langelier C, Christie JD, McCauley K, Nelson M, Cheung CK, Benowitz NL, Cohen MJ, Calfee CS. 2018. Lung Microbiota Is Related to Smoking Status and to Development of Acute Respiratory Distress Syndrome in Critically Ill Trauma Patients. Am. J. Respir. Crit. Care Med. 197:621–631.

Park H, Shin JW, Park S-G, Kim W. 2014. Microbial communities in the upper respiratory tract of patients with asthma and chronic obstructive pulmonary disease. PLoS One 9:e109710.

Patel VB, Zhong JC, Grant MB, Oudit GY. 2016. Role of the ACE2/angiotensin 1-7 axis of the renin-angiotensin system in heart failure. Circ. Res. 118:1313–1326.

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, et al. 2004. Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. 74:979–1000.

Perkins GD, Roberts J, McAuley DF, Armstrong L, Millar A, Gao F, Thickett DR. 2005. Regulation of vascular endothelial growth factor bioactivity in patients with acute lung injury. Thorax 60:153–158.

Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. 2017. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 18.

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, et al. 2009. The NIH Human Microbiome Project. Genome Res. 19:2317–2323.

Pflughoeft KJ, Versalovic J. 2012. Human Microbiome in Health and Disease. Annu. Rev. Pathol. Mech. Dis. 7:99–122.

Pham T, Rubenfeld GD. 2017. Fifty years of research in ards the epidemiology of acute respiratory distress syndrome a 50th birthday review. Am. J. Respir. Crit. Care Med. 195:860–870.

Pham VHT, Kim J. 2012. Cultivation of unculturable soil bacteria. Trends Biotechnol. 30:475–484.

Pierron D, Heiske M, Razafindrazaka H, Pereda-Loth V, Sanchez J, Alva O, Arachiche A, Boland A, Olaso R, Deleuze JF, et al. 2018. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. Nat. Commun. 9.

Pino-Yanes M, Corrales A, Basaldúa S, Hernández A, Guerra L, Villar J, Flores C. 2011. North African influences and potential bias in case-control association studies in the Spanish population. PLoS One 6.

de Prost N, Razazi K, Brun-Buisson C. 2013. Unrevealing culture-negative severe sepsis. Crit. Care 17:1001.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65.

Quesnel C, Nardelli L, Piednoir P, Leçon V, Marchal-Somme J, Lasocki S, Bouadma L, Philip I, Soler P, Crestani B, et al. 2010. Alveolar fibroblasts in acute lung injury: Biological behaviour and clinical relevance. Eur. Respir. J. 35:1312–1321.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35:833–844.

Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. 2013. Common risk alleles for inflammatory diseases are targets of recent positive selection. Am. J. Hum. Genet. 92:517–529.

Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. 469:967–977.

Reilly JP, Christie JD, Meyer NJ. 2017. Fifty years of research in ARDS genomic contributions and opportunities. Am. J. Respir. Crit. Care Med. 196:1113–1121.

Remick DG, Bolgos G, Copeland S, Siddiqui J. 2005. Role of interleukin-6 in mortality from and physiologic response to sepsis. Infect. Immun. 73:2751–2757.

Richeldi L, du Bois RM, Raghu G, Azuma A, Brown KK, Costabel U, Cottin V, Flaherty KR, Hansell DM, Inoue Y, et al. 2014. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. N. Engl. J. Med. 370:2071–2082.

Rodríguez-Esparragón F, López-Fernández JC, Buset-Ríos N, García-Bello MA, Hernández-Velazquez E, Cappiello L, Rodríguez-Pérez JC. 2017. Paraoxonase 1 and 2 gene variants and the ischemic stroke risk in Gran Canaria population: an association study and meta-analysis. Int. J. Neurosci. 127:191–198.

Rodríguez-Varela R, Günther T, Krzewińska M, Storå J, Gillingwater TH, MacCallum M, Arsuaga JL, Dobney K, Valdiosera C, Jakobsson M, et al. 2017. Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. Curr. Biol. 27:3396–3402.e5.

Rooks MG, Garrett WS. 2016. Gut microbiota, metabolites and host immunity. Nat. Rev. Immunol. 16:341–352.

Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, Stern EJ, Hudson LD. 2005. Incidence and outcomes of acute lung injury. N. Engl. J. Med. 353:1685–1693.

Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, Janicijevic B, Jin L, Chakraborty R, Deka R, et al. 2003. Inbreeding and risk of late onset complex disease. J. Med. Genet. 40:925–932.

Rumpel JA, Ahmedani BK, Peterson EL, Wells KE, Yang M, Levin AM, Yang JJ, Kumar R, Burchard EG, Williams LK. 2012. Genetic ancestry and its association with asthma exacerbations among African American subjects with asthma. J. Allergy Clin. Immunol. 130:1302–1306.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614–1620.

Sánchez-Lerma B, Morales-Chirivella FJ, Peñuelas I, Blanco Guerra C, Mesa Lugo F, Aguinaga-Ontoso I, Guillén-Grima F. 2009. High prevalence of asthma and allergic diseases in children aged 6 to [corrected] 7 years from the Canary Islands. [corrected]. J. Investig. Allergol. Clin. Immunol. 19:383–390.

Sanchez-Mazas A, Černý V, Di D, Buhler S, Podgorná E, Chevallier E, Brunet L, Weber S, Kervaire B, Testi M, et al. 2017. The HLA-B landscape of Africa: Signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. Mol. Ecol. 26:6238–6252.

Sandoval E, Chang DW. 2016. Association Between Race and Case Fatality Rate in Hospitalizations for Sepsis. J. racial Ethn. Heal. disparities 3:625–634.

Santana A, Salido E, Torres A, Shapiro LJ. 2003. Primary hyperoxaluria type 1 in the Canary Islands: a conformational disease due to I244T mutation in the P11L-containing alanine:glyoxylate aminotransferase. Proc. Natl. Acad. Sci. U. S. A. 100:7277–7282.

Schaffer JN, Pearson MM. 2015. Proteus mirabilis and Urinary Tract Infections. Microbiol. Spectr. 3.

Schwartz IS, Embil JM, Sharma A, Goulet S, Light RB. 2016. Management and outcomes of acute respiratory distress syndrome caused by blastomycosis a retrospective case series. Med. (Baltimore) 95:e3538.

Seldin MF, Pasaniuc B, Price AL. 2011. New approaches to disease mapping in admixed populations. Nat. Rev. Genet. 12:523–528.

Shapiro NI, Schuetz P, Yano K, Sorasaki M, Parikh SM, Jones AE, Trzeciak S, Ngo L, Aird WC. 2010. The association of endothelial cell signaling, severity of illness, and organ dysfunction in sepsis. Crit. Care 14:R182.

Shaver CM, Bastarache JA. 2014. Clinical and biological heterogeneity in acute respiratory distress syndrome: Direct versus indirect lung injury. Clin. Chest Med. 35:639–653.

Shibuya M. 2001. Structure and dual function of vascular endothelial growth factor receptor-1 (Flt-1). Int. J. Biochem. Cell Biol. 33:409–420.

Shimizu K, Ogura H, Hamasaki T, Goto M, Tasaki O, Asahara T, Nomoto K, Morotomi M, Matsushima A, Kuwagata Y, et al. 2011. Altered gut flora are associated with septic complications and death in critically ill patients with systemic inflammatory response syndrome. Dig. Dis. Sci. 56:1171–1177.

Shortt K, Chaudhary S, Grigoryev D, Heruth DP, Venkitachalam L, Zhang LQ, Ye SQ. 2014. Identification of novel single nucleotide polymorphisms associated with acute respiratory distress syndrome by exome-seq. PLoS One 9:e111953.

Shriner D. 2017. Overview of admixture mapping. Curr. Protoc. Hum. Genet. 2017:1.23.1-1.23.8.

Shriner D, Adeyemo A, Ramos E, Chen G, Rotimi CN. 2011. Mapping of disease-associated variants in admixed populations. Genome Biol. 12.

Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM, et al. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 315:801–810.

Smith JA, Kauffman CA. 2010. Blastomycosis. Proc. Am. Thorac. Soc. 7:173–180.

Soeta N, Terashima M, Gotoh M, Mori S, Nishiyama K, Ishioka K, Kaneko H, Suzutani T. 2009. An improved rapid quantitative detection and identification method for a wide range of fungi. J. Med. Microbiol. 58:1037–1044.

Sofer T, Baier LJ, Browning SR, Thornton TA, Talavera GA, Wassertheil-Smoller S, Daviglus ML, Hanson R, Kobes S, Cooper RS, et al. 2017. Admixture mapping in the Hispanic Community Health

Study/Study of Latinos reveals regions of genetic associations with blood pressure traits. PLoS One 12.

Sørensen TI, Nielsen GG, Andersen PK, Teasdale TW. 1988. Genetic and environmental influences on premature death in adult adoptees. N. Engl. J. Med. 318:727–732.

Soto GJ, Martin GS, Gong MN. 2013. Healthcare disparities in critical illness. Crit. Care Med. 41:2784–2793.

Spear ML, Hu D, Pino-Yanes M, Huntsman S, Eng C, Levin AM, Ortega VE, White MJ, McGarry ME, Thakur N, et al. 2019. A genome-wide association and admixture mapping study of bronchodilator drug response in African Americans with asthma. Pharmacogenomics J. 19:249–259.

Stapleton RD, Wang BM, Hudson LD, Rubenfeld GD, Caldwell ES, Steinberg KP. 2005. Causes and timing of death in patients with ARDS. Chest 128:525–532.

de Steenhuijsen Piters WAA, Huijskens EGW, Wyllie AL, Biesbroek G, van den Bergh MR, Veenhoven RH, Wang X, Trzciński K, Bonten MJ, Rossen JWA, et al. 2016. Dysbiosis of upper respiratory tract microbiota in elderly pneumonia patients. ISME J. 10:97–108.

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L, et al. 1998. Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am. J. Hum. Genet. 62:1507–1515.

Tan X, Liu H, Long J, Jiang Z, Luo Y, Zhao X, Cai S, Zhong X, Cen Z, Su J, et al. 2019. Septic patients in the intensive care unit present different nasal microbiotas. Future Microbiol. 14:383–395.

Tang H, Coram M, Wang P, Zhu X, Risch N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. 79:1–12.

Taylor SM, Parobek CM, Fairhurst RM. 2012. Haemoglobinopathies and the clinical epidemiology of malaria: A systematic review and meta-analysis. Lancet Infect. Dis. 12:457–468.

Thomas R, Apps R, Qi Y, Gao X, Male V, O'Huigin C, O'Connor G, Ge D, Fellay J, Martin JN, et al. 2009. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. Nat. Genet. 41:1290–1294.

Thompson BT. 2000. Ketoconazole for early treatment of acute lung injury and acute respiratory distress syndrome: A randomized controlled trial. J. Am. Med. Assoc. 283:1995–2002.

Thomsen H, Chen B, Figlioli G, Elisei R, Romei C, Cipollini M, Cristaudo A, Bambi F, Hoffmann P, Herms

S, et al. 2016. Runs of homozygosity and inbreeding in thyroid cancer. BMC Cancer 16.

Thornton TA, Bermejo JL. 2014. Local and global ancestry inference and applications to genetic association analysis for admixed Populations. Genet. Epidemiol. 38.

Tringe SG, Rubin EM. 2005. METAGENOMICS : DNA SEQUENCING. Genetics 6:805–814.

Ursell LK, Metcalf JL, Parfrey LW, Knight R. 2012. Defining the human microbiome. Nutr. Rev. 70.

Vasseur E, Quintana-Murci L. 2013. The impact of natural selection on health and disease: Uses of the population genetics approach in humans. Evol. Appl. 6:596–607.

Vergara C, Murray T, Rafaels N, Lewis R, Campbell M, Foster C, Gao L, Faruque M, Oliveira RR, Carvalho E, et al. 2013. African Ancestry is a Risk Factor for Asthma and High Total IgE Levels in African Admixed Populations. Genet. Epidemiol. 37:393–401.

Villar J. 2011. What Is the Acute Respiratory Distress Syndrome? Respir Care. 56:1539–1545.

Villar J, Blanco J, Añón JM, Santos-Bouza A, Blanch L, Ambrós A, Gandía F, Carriedo D, Mosteiro F, Basaldúa S, et al. 2011. The ALIEN study: incidence and outcome of acute respiratory distress syndrome in the era of lung protective ventilation. Intensive Care Med. 37:1932–1941.

Villar J, Blanco J, Kacmarek RM. 2016. Current incidence and outcome of the acute respiratory distress syndrome. Curr. Opin. Crit. Care 22:1–6.

Villar J, Sulemanji D, Kacmarek RM. 2014. The acute respiratory distress syndrome: incidence and mortality, has it changed? Curr. Opin. Crit. Care 20:3–9.

Vincent JL, Opal SM, Marshall JC. 2010. Ten reasons why we should NOT use severity scores as entry criteria for clinical trials or in our treatment decisions. Crit. Care Med. 38:283–287.

Voelkel NF, Vandivier RW, Tuder RM. 2006. Vascular endothelial growth factor in the lung. Am. J. Physiol. Lung Cell. Mol. Physiol. 290.

Vrigkou E, Tsangaris I, Bonovas S, Tsantes A, Kopterides P. 2017. The evolving role of the renin-angiotensin system in ARDS. Crit. Care 21.

Wang B, Yao M, Lv L, Ling Z, Li L. 2017. The Human Microbiota in Health and Disease. Engineering 3:71–82.

Wang ZK, Yang YS, Stefka AT, Sun G, Peng LH. 2014. Review article: Fungal microbiota and digestive diseases. Aliment. Pharmacol. Ther. 39:751–766.

Ware L, Matthay M. 2000. The acute respiratory distress syndrome. N. Engl. J. Med. 342:1334–1349.

Ware LB, Kaner RJ, Crystal RG, Schane R, Trivedi NN, McAuley D, Matthay MA. 2005. VEGF levels in the alveolar compartment do not distinguish between ARDS and hydrostatic pulmonary oedema. Eur. Respir. J. 26:101–105.

Williams LM, Qi Z, Batai K, Hooker S, Hall NJ, Machado RF, Chen A, Campbell-Lee S, Guan Y, Kittles R, et al. 2018. A locus on chromosome 5 shows African ancestry-limited association with alloimmunization in sickle cell disease. Blood Adv. 2:3637–3647.

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB. 2001. Population genetic structure of variable drug response. Nat. Genet. 29:265–269.

Wollin L, Wex E, Pautsch A, Schnapp G, Hostettler KE, Stowasser S, Kolb M. 2015. Mode of action of nintedanib in the treatment of idiopathic pulmonary fibrosis. Eur. Respir. J. 45:1434–1445.

Xia LP, Bian LY, Xu M, Liu Y, Tang AL, Ye WQ. 2015. 16S rRNA gene sequencing is a non-culture method of defining the specific bacterial etiology of ventilator-associated pneumonia. Int. J. Clin. Exp. Med. 8:18560–18570.

Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, Gilly A, Ayub Q, Colonna V, Southam L, et al. 2017. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. Nat. Commun. 8.

Yadava K, Pattaroni C, Sichelstiel AK, Trompette A, Gollwitzer ES, Salami O, von Garnier C, Nicod LP, Marsland BJ. 2016. Microbiota Promotes Chronic Pulmonary Inflammation by Enhancing IL-17A and Autoantibodies. Am. J. Respir. Crit. Care Med. 193:975–987.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, et al. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

Yim JJ, Ding L, Schäffer AA, Park GY, Shim YS, Holland SM. 2004. A microsatellite polymorphism in intron 2 of human Toll-like receptor 2 gene: functional implications and racial differences. FEMS Immunol. Med. Microbiol. 40:163–169.

Zemanick ET, Wagner BD, Robertson CE, Ahrens RC, Chmiel JF, Clancy JP, Gibson RL, Harris WT, Kurland G, Laguna TA, et al. 2017. Airway microbiota across age and disease spectrum in cystic fibrosis. Eur. Respir. J. 50:1700832.