Clinica Chimica Acta 451 (2015) 39–45

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/clinchim

Exome sequencing and pathway analysis for identification of genetic variability relevant for bronchopulmonary dysplasia (BPD) in preterm newborns: A pilot study

Paola Carrera [a,b,*], Chiara Di Resta [a], Chiara Volonteri [c], Emanuela Castiglioni [a], Silvia Bonfiglio [d], Dejan Lazarevic [d], Davide Cittaro [d], Elia Stupka [d], Maurizio Ferrari [a,b,c], Marco Somaschini [a], for the, BPD and Genetics Study Group

Rosario Magaldi [1], Matteo Rinaldi [1], Gianfranco Maffei [1], Mauro Stronati [2], Chryssoula Tzialla [2], Alessandro Borghesi [2], Paolo Tagliabue [3], Tiziana Fedeli [3], Marco Citterio [3], Fabio Mosca [4], Mariarosa Colnaghi [4], Anna Lavizzari [4], Massimo Agosti [5], Gaia Francescato [5], Giulia Pomero [6], Cristina Dalmazzo [6], Antonio Boldrini [7], Rosa Scaramuzzo [7], Enrico Bertino [8], Silvia Borgione [8], Claudio Martano [8], Virgilio Carnielli [9], Stefano Nobile [9], Antonietta Auriemma [10], Cristina Bellan [10], Giuseppe Carrera [11], Chiara Zambetti [11], Riccardo Pucello [12], Sara Palatta [12]

[1] *Foggia, Italy*
[2] *Pavia, Italy*
[3] *Monza, Italy*
[4] *Milano, Italy*
[5] *Varese, Italy*
[6] *Cuneo, Italy*
[7] *Pisa, Italy*
[8] *Torino, Italy*
[9] *Ancona, Italy*
[10] *Seriate, Italy*
[11] *Lodi, Italy*
[12] *Roma, Italy*

[a] *Unit of Genomics for Diagnosis of Human Pathologies, Division of Genetics and Cell Biology, IRCCS Ospedale San Raffaele, Milano, Italy*
[b] *Laboratory of Clinical Molecular Biology, IRCCS Ospedale San Raffaele, Milano, Italy*
[c] *Vita-Salute San Raffaele University, Milano, Italy*
[d] *Centre for Translational Genomics and Bioinformatics, IRCCS Ospedale San Raffaele, Milano, Italy*

ARTICLE INFO

Article history:
Received 15 October 2014
Received in revised form 30 December 2014
Accepted 2 January 2015
Available online 8 January 2015

Keywords:
Bronchopulmonary dysplasia
Exome sequencing
Genetics

ABSTRACT

Background: Bronchopulmonary dysplasia (BPD) is the most common chronic lung disease in infancy, affecting preterm children with low birth weight. The disease has a multifactorial aetiology with a significant genetic component; until now published association studies have identified several candidate genes but only few of these data has been replicated. In this pilot study, we approached exome sequencing aimed at identifying non-common variants, which are expected to have a stronger phenotypic effect.
Materials and methods: We performed this study on 26 Italian severely affected BPD preterm unrelated newborns, homogeneously selected from a large prospective cohort. We used an Illumina HiSeq 2000 for sequencing. Data analysis was focussed on genes previously associated to BPD susceptibility and to new candidates in related pathways, highlighted by a prioritization analysis performed using ToppGene Suite.
Results: By exome sequencing, we identified 3369 novel variants, with a median of 400 variations per sample. The top candidate genes highlighted were *NOS2, MMP1, CRP, LBP* and the toll-like receptor (*TLR*) family. All of them have been confirmed with Sanger sequencing.

* Corresponding author. Tel.: +39 02 26434759.
E-mail address: carrera.paola@hsr.it (P. Carrera).

http://dx.doi.org/10.1016/j.cca.2015.01.001

Conclusions: Potential candidate genes have been discovered in this preliminary study; the pathogenic role of identified variants will need to be confirmed with functional and segregation studies and possibly with further methods, able to evaluate the collective influence of rare variants.
Moreover, additional candidates will be tested and genetic analysis will be extended to all affected children.

1. Introduction

In recent years, next generation sequencing (NGS) technology has been exploited to gain insight into human genetic field, identifying new variants with functional and pathological relevance. With NGS, whole exome can be sequenced in a short time for reasonably low cost. Exome sequencing explores both common and rare coding variants, which may impact directly on the protein's structure and function. This application can represent a great challenge to identify new candidate genes not only in Mendelian but also in complex diseases, whose associated susceptibility loci are still unknown. In complex diseases numerous genome-wide association studies have been performed to explore the role of common genetic variants but they were able to explain only a relatively small proportion of heritability. Indeed association of a single rare variant with a phenotype requires enormous sample sizes, and methods to evaluate the collective influences of rare variants across a gene or across multiple genes in the same pathway are being developed [1]. In this study, we exploited the great potential of NGS to identify new causative variants associated to bronchopulmonary dysplasia (BPD) susceptibility.

BPD is the most common chronic lung disease in preterm newborns [2]. It affects about 30% of surviving infants and it is related to gestational age (GA; 22–28 weeks) and birth weight (501–1500 g) [3,4]. BPD was initially described as a chronic lung disease caused by injury of mechanical ventilation and oxygen exposure. Such a disease occurred in relatively large premature infants and was histologically characterized by intense airway inflammation and lung fibrosis. Since 1990 this respiratory disease defined as “old BPD” has been replaced by the “new BPD” that occurs in very low birth weight infants, whose survival rate has significantly improved by the advancement of perinatal care including antenatal steroids, routine surfactant replacement, and introduction of less invasive ventilation modalities. Unlike the original form of the disease, this new form often develops in preterm newborns, who may have needed little or no ventilator support and had low inspired oxygen concentrations during the postnatal period. At autopsy, lung histology of these infants has regions of more uniform and milder injury, but impaired alveolar and vascular growth remains prominent [5]. From a clinical point of view there are three different forms of severity of the clinical phenotype (mild, moderate and severe), classified according to Jobe and Bancalari consensus criteria [6]. BPD surely has a multifactorial aetiology and today it is clear that genetic predisposition plays a critical role in BPD pathogenesis, particularly in association with lower gestational age mostly in moderate and severe BPD forms [7,8]. In the past years, several association studies have been performed to discover genetic variants associated with BPD, either on a chosen set of SNPs/genes or on whole genome. Association studies on candidates were focused on genes encoding surfactant proteins, genes involved in vascular development, inflammation-related genes, matrix remodelling proteins, adhesion molecules, and antioxidant enzymes [7,9]. All of them focused on moderate and severe BPDs, considering that extreme phenotype has a stronger genetic susceptibility [7,9]. Results were encouraging but few of these studies have been replicated and in most of them there was a limited statistical power because of the small sample size. To improve the current knowledge on the genetic basis of BPD, we exploited the next generation sequencing (NGS) technology. At first we performed a pilot exome analysis on 26 severely-BPD affected patients; these patients have an extreme phenotype and our hypothesis is to have a higher chance to detect rare variants with moderate-to-high impact in this subgroup [10,11].

2. Materials and methods

2.1. Selected cohort of patients

For this pilot study, we selected 26 unrelated newborns with a clinical diagnosis of severe BPD, chosen among the collected cohort of 366 premature children admitted to the neonatal intensive care units of 12 Italian medical centres participating in the study in the past 5 years. Institutional review boards for each participating centre approved this study, led by San Raffaele Hospital. Written informed consent was obtained from the legal tutors of all enrolled infants. All patients were recruited prospectively. Clinical, personal and familial epidemiological data have been previously reported and recalled below [12]. The eligibility criteria for the enrolment of newborns are i) the gestational age up to 32 weeks, ii) the European origin to avoid biases on ethnic group in genetic analysis, iii) the survival at 36 weeks postmenstrual age (PMA), and iv) the absence of major congenital malformations of the lung. Children who developed the pathology were included in the case group ($n = 141$) while the unaffected infants were included in the control group ($n = 225$) [12]. Clinicians classified BPD-affected newborns into three different phenotype groups (mild, severe and moderate), according to Jobe and Bancalari consensus criteria [6]. For more complex cases, the referring centre confirmed the clinical diagnosis. We recorded clinical features, personal and family data and other information about perinatal events of all enrolled patients and controls into a database shared among members of the network. Since we planned to perform molecular analysis on DNA extracted from blood cells, if an infant had a blood transfusion, we did not accept blood sampling within 20 days from last transfusion to be sure that genetic analyses were performed on the case's DNA and not on the donor's. In particular for this study, we analysed 26 out of 36 collected newborns with severe BPD, requiring $O_2 > 30\%$ at 36 weeks PMA and/or VM/N-CPAP support [6]. In the subgroup of the 26, the mean birth-weight (778 g) was comparable to that (740 g) observed in the whole group of newborns with the severe BPD ($n = 36$). No difference in the occurrence of IUGR was observed between the patients and controls. The cohort of patients sequenced in this work has Caucasian origin, specifically 77% of patients are of Italian origin and 23% originate from Romania, Albania, Serbia and Slovenia.

2.2. Exome sequencing

Genomic DNA (gDNA) was extracted from 800 μl of peripheral blood using the automated extractor Maxwell® 16 Research System (Promega, Madison, WI, USA); the concentration and high quality of gDNA ($A_{260/280}$ 1.8–2.0) was determined using a Nanodrop™ Spectrophotometer 1000 (Thermo Fisher Scientific, Wilmington, DE, USA). A Covaris™ E220™ (Covaris, Inc., Woburn, MA, USA) was employed to shear 1–3 μg of each DNA sample. The exome sequencing protocol requires the mean of DNA fragments close to 250 bp and it was verified through an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldron, Germany). Exome sequencing was carried out on an Illumina HiSeq 2000 platform (Illumina, Inc., San Diego, CA. USA) using Illumina TruSeq for DNA sample preparation and exome enrichment protocols.

2.3. Sequencing data analysis

Image analysis and base calling was performed by converting light signal intensities into sequences of nucleotides as FastQ files. The quality

of obtained sequence data was reported in FastQC files, which evaluates base quality statistics, contamination sources (i.e. adaptors, concatamers) and sequence duplication levels.

FastQ data were mapped on the NCBI human reference genome hg19 build with the Wheeler alignment tool (BWA) [13] with default parameters, providing high-speed alignment and good performance in terms of precision/recall. Alignments can be seen with IGV2.1 software, released from the Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA [14]. All the reads with more than 5 mismatches or those with mapping quality (MAPQ) less than 15 were filtered out. Duplicate reads due to clonal amplification during library preparation were removed in order to avoid allele frequency errors, using Picard's MarkDuplicates (http://broadinstitute.github.io/picard). Single nucleotide variation (SNV) and insertion/deletion (INDEL) were called with a pipeline involving GATK [15]. Resulting variant frequencies were compared to dbSNP (v 136) and an internal sequence variant database. Possible impact of variations was evaluated using SnpEff [16] that is useful to classify variants "high-", "moderate-" or "low-impact" on the basis of their predicted effect on the protein. All variants have been grouped in three tiers, each one being a subgroup of the previous one: tier 1 includes all variants with a VQRS LOD score >2.62, variants that do not fulfil this condition were not taken into further consideration; tier 2 selects variants that SnpEff predicts as high and moderate; tier 3 comprises variants from tier 2 whose MAF is <0.02 or unknown, and those that are novel.

2.4. Selection of potentially causative variants

In order to select potentially interesting variants that can increase the susceptibility of BPD, we adopted two different strategies: i) Genes previously associated to BPD. We filtered all variants of tier 2 with an high or moderate effect according to SnpEff, identified in genes previously related to the disease [7,8]. Indeed a number of association studies have already been performed, with a focus on surfactant proteins, genes involved in vascular and lung development, inflammation-related genes, matrix remodelling proteins, adhesion molecules, and antioxidant enzymes [6,17–21].

ii) Prioritization analysis. We focused on rare and novel variants that SnpEff predicted as having high or moderate effects and included in tier 3. We selected variants that both PolyPhen-2 (PP-2) and SIFT predicted as potentially detrimental (possibly damaging, P, and probably damaging, D, for PP-2 and SIFT score <0.05). Variants with neither PP-2 nor SIFT predictions were also taken into account. After that, looking for potentially interesting genes, we focused on those with more than one variant, be it the same variant in more than one sample, homozygosis in a single sample or different variants in the same gene. We then used ToppGene Prioritization software [22] for pathway analysis on this selection at two different levels: 1) using a complete training list, with genes participating into different pathways that published literature implicates in BPD pathogenesis; 2) separating the chosen training genes according to the main pathway they belong to and then comparing genes' rank in different pathways. In order to identify genes with a broader and stronger effect, we chose as interesting candidates those that ranked ≤5 with the first kind of analysis and those ranked ≤100 in all pathways. Top variants have been validated using Sanger sequencing (ABI 3730 DNA analyzer, Applied Biosystems, Monza, Italy). Primer sequence and PCR conditions are available on request. Possible molecular pathways and interactions among genes identified by the two strategies were then studied by String 9.122 [23].

3. Results

In this study we performed the exome sequencing on 26 patients with a diagnosis of severe BPD, in agreement with consensus criteria [6], requiring oxygen supplementation (oxygen ≥30%) or a respiratory support at 36 weeks PMA. For our study, we considered the extreme phenotypes because a stronger genetic component is hypothesized. The exome sequencing was performed and only the 21 samples with a mean coverage ≥15× were considered for variants' analyses. Haplotype Caller identified a total of 1,229,601 single nucleotide variants or small insertions/deletions (tier 1). Each patient had around 200,000 variants. Upon filtering for predicted effect, the count dropped to 27,673 (about 8000 per individual). This number includes variants with moderate impact (non-synonymous coding, codon insertions or deletions, codon changes) and variants with high effect (gain or loss of stop codons, frameshift variants, loss of start codons, variants affecting splice donor or splice acceptor sites). It means that about 2% of all variants found probably have a functional impact on protein structure or expression. Tier 3, with rare and novel variants, included a total of 9427 variants, of which 3369 have never been described before (novelty rate of about 0.3%) while 1524 were predicted as probably damaging (0.1% of all identified variants). In each sample, data analyses identified about 400 genetic unknown variations and almost 100 variants with a strong impact on protein structure and function.

We chose different strategies for the selection of putative candidate genes. i) Genes previously associated to BPD. Considering variants included in tier 2, we focused our analysis on variants identified in genes previously associated to BPD. In particular, we considered surfactant metabolism genes (*SFTPA1, SFTPA2, SFTPB, STPTC, SFTPD, ABCA3*), matrix remodelling genes (*MMP2, MMP16, MMP14, SPOCK2*), vascular development genes (*VEGFC, ACE, GSTP1*), oxidative stress-related genes (*MTHFR, EPHX1, EPHX2*), and inflammation related genes (genes belonging to the *TLR*-family, *MBL1, MBL2*). We identified a total of 61 variants in 19 genes and confirmed them with Sanger; 31 are common polymorphism, 25 are rare and classified as dbSNP rs with a MAF <0.05 and 6 are novel (Suppl. Table 1). All these variants are non-synonymous coding with exception of the codon insertion rs71553864, the stop gained rs5744168 and a novel frame-shift. Among variants, 4 of them, in *TLR4, TLR5, MBL2*, were predicted to be pathogenic in dbSNP database while 6 were predicted with a damaging effect by both PP-2 and SIFT. Interestingly, we identified 2 novel variants (p.R1035C and p.Q403P) in the *ABCA3* gene, known for its association to pulmonary surfactant metabolism dysfunction type 3 (MIM610921). The variation p.R1035C was identified in two unrelated patients. Considering all the variants, the most mutated genes are those belonging to the *TLR*-family (*TLR10, TLR1, TLR4*), to oxidative stress-related genes (*EPHX2, MTHFR, EPHX1*) and to surfactant metabolism genes (*SFTPD, ABCA3*). ii) Prioritization analysis. Considering variants included in tier 3, our approach was focused on pathway analysis with tools from ToppGene Suite [22].

We chose our training set of 65 genes from our laboratory's previous works, published literature data, including association studies and expression studies in pulmonary tissues and in animal models of BPD or oxidative injury in lungs [12,17–21,24–43]. All of these genes are involved in different pathways that may underlie BPD. We defined the pathway they belonged to in agreement with data from the said literature, from annotation in the National Centre for Biotechnology Information's gene database and in the ToppGene Suite's database. The list of training set of genes for pathway analyses and the pathways' partition is reported in Suppl. Table 2. With the purpose of finding variants with a strong/moderate effect, we focused on tier 3. In particular, we selected deleterious variants for both used prediction tools and those with no SIFT or PolyPhen-2 prediction. This last category comprised 262 novel single nucleotide variants in a total of 2536 variants in 2139 genes. In order to filter out private variants whose relationship with BPD could hardly be proved, we focused on those that were present in more than one sample (607 variants distributed among 501 genes) and on variants that belonged to the same gene (701 variants in 304 genes). Since 70 genes presented both multiple different variants and at least one variant in multiple samples, our selection was reduced to 735 genes and 1132 variants. Ten additional genes found were added because they had a single variant in a single patient but it was in homozygous state. This list of 745 genes was used as the testing set for

Table 1
ToppGene Prioritization.

Gene	SF	OxStress	Angiogenesis	Tissue remodelling	Immunity/inflammation	Lung development	ALL
TLR1	564	137	16	155	1	7	1
MMP1	487	92	2	1	6	20	2
NOS2[*]	100	20	8	24	7	4	3
CRP	41	94	113	126	10	76	4
LBP	63	660	108	15	17	140	5

List of genes that ranked in the first five positions, and ranking in each pathway.
SF: surfactant system; OxStress: oxidative stress; ALL: complete training list (65 genes).
[*] NOS2 ranked in the first 100 positions in each pathway-specific training list.

pathway analyses (Suppl. Table 3). ToppGene Prioritization tool ranks the testing genes according to their agreement with the training set (Suppl. Table 2). This agreement is based on many different gene ontology annotations[21]. We tested the selection of 745 genes i) against the complete training list (Table 1 shows the top 5 genes in this rank); and ii) against the different pathways; six different training sets were used, each one related to a different pathway (data not shown). In order to zoom in on genes that could have the broadest effect on BPD pathogenesis, we decided to focus first on the list of the top 5 genes, which was obtained with the ToppGene Prioritization tool using the complete training list (Table 1) and considering them as potentially interesting candidates from 735 genes of testing set. In this phase of data analysis we validated rare and novel variants identified in the top candidate genes: *TLR1, MMP1, NOS2, CRP* and *LBP*, using Sanger sequencing (Table 2). These variants were shared among at least 2 BPD patients, with the exception of the novel p.K730E missense in the *NOS2* gene, confirmed in only one patient. The variant was absent in a sample of 180 control chromosomes from our collection of healthy preterm newborn subjects, corroborating the hypothesis of a possible role in BPD.

To evaluate the possible interaction between candidate genes, the ones previously associated to BPD and showing variants in our sample (*ABCA3, SFTPD, SPOCK2, ACE, MTHFR, EPHX1, EPHX2, TLR5, TLR10, TLR1, TLR6, TLR4, GSTP1, MBL2, TLR10, TLR2*) and the top 5 genes (*NOS2, TLR1, MMP1, CRP, LBP*) highlighted by the ToppGene analysis, we used String 9.122 (url: http://string-db.org/). The results are reported in Fig. 1 allowing the possibility of a networking with a main focus on genes involved in inflammation.

4. Discussion

BPD is still the most common complication of prematurity showing an increased risk of death, respiratory sequelae in childhood and adulthood, and long term neurodevelopmental problems [44,45]. Treatments have helped to modestly improve BPD outcomes and most of the current therapies are supportive. This lack of significant clinical benefit from therapies has raised interest in searching biological markers useful in targeting therapeutic interventions. Despite its high incidence on preterm newborns, its pathogenesis is not yet clearly understood. BPD is a multifactorial disease characterized by impaired alveolar and vascular development. Environmental factors such as high concentrations of oxygen exposure, mechanical ventilation, perinatal infections and inflammation play major roles in pathogenesis; however, a genetic component, accounting for the more than a half of the variance in liability to BPD has been proved [8]. Association studies did not lead to any major breakthroughs. They rarely identify the true causal variants; in fact, they only examine common variants, according to the "common variant-common disease" hypothesis. Therefore, NGS studies are appealing because of their ability to test also non-common variants which are expected to have a stronger phenotypic effect [10]. If complex diseases do rely on rare functional variants, sequencing is necessary for causative variants identification. Recently, the paradigm for complex disease genetics has shifted towards this hypothesis, posing that rare variants with strong functional effects underlie the majority of these disorders [10].

Here we performed a study on a cohort of BPD patients exploiting for the first time the exome sequencing applied to this complex disorder. The studied preterm newborns with the severe BPD belong to a cohort prospectively enrolled and diagnosed uniformly in agreement with the defined consensus criteria [6]. These subjects were selected as extreme phenotypes, where a stronger genetic component is hypothesized and our aim was to detect rare variants with moderate-to-high impact. To date there is a lack of guidelines for management and analysis of NGS data especially for complex diseases. In the present work, a pilot study was performed, with interpretation of results based on two approaches. First, we analysed all genes known to be associated to BPD susceptibility, with a particular interest for all novel and rare variants; second, as a strategy for selection of additional candidates, we performed a prioritization study based on pathway analysis of recurring genes with rare or novel variants that were predicted to have a strong impact on protein function and/or structure. Pathway analyses are particularly interesting for complex diseases; identification of variants in gene networks is more likely than discovery of variants in the same gene when different rare genetic variants underlie these disorders [46]. They are extremely useful to make sense of a huge amount of data on diverse genes. ToppGene Prioritization is a free ready-to-use software that allows to look for correlations between a questioning set of genes and a training set that might or might not belong to the same pathway. Its most interesting feature is that it ranks testing genes by comparing them to a training list of choice. Since BPD is a complex disease, whose pathogenesis likely involves many different pathways, how to exploit such a tool at its best still is in a burgeoning phase. Therefore, we chose two different strategies and subsequently compared their results. When we used the complete training list, the top 5 genes included some but not all of the first-ranked genes in separate-pathway analysis.

Table 2
List of selected candidate variants.

Gene	CHR	POS	ID	REF	ALT	AF	GMAF	EFF.CODOON,AA,EFFECT
NOS2	chr17	26093594	None	T	C	0.011		Aag/Gag,K730E,NON_SYNONYMOUS_CODING;
TLR1	chr4	38798255	rs5743621	G	A	0.011	0.092	cCc/cTc,P733L,NON_SYNONYMOUS_CODING;
TLR1	chr4	38799955	None	CTT	CA	0.011		–,–166,FRAME_SHIFT;
MMP1	chr11	102667446	rs17879749	CA	C	0.053	0.005	–,–125,FRAME_SHIFT;–,–125,FRAME_SHIFT;
CRP	chr1	159683814	rs77832441	G	A	0.032	0.0005	aCg/aTg,T59M,NON_SYNONYMOUS_CODING;
LBP	chr20	36993333	rs2232607	A	G	0.032	0.0073	gAt/gGt,D283G,NON_SYNONYMOUS_CODING;

Legend: CHR: chromosome number; POS: genomic position; ID: variant's reference name; REF: reference allele; ALT: alternative allele; AF: allele frequency in our population; GMAF: frequency in 1000 Genome project; EFF.CODOON,AA,EFFECT:effect on variant on codon and on aminoacid.

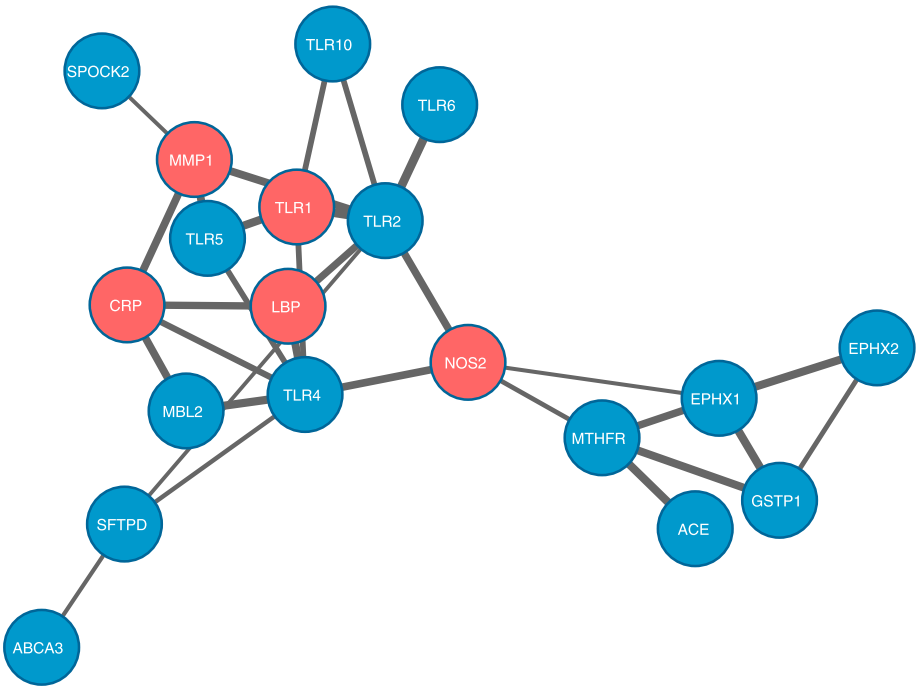

Fig. 1. STRING pathway analysis plot (confidence score = 0.5). Network included 15 genes previously associated to BDP susceptibility showing variations in our cohort and the 5 candidate genes highlighted by ToppGene analysis.

This result might suggest that those pathways (angiogenesis, inflammation and tissue remodelling) are more relevant to BPD pathogenesis. Another possibility is that the training set we used was more specific for these pathways and imprecise for the others. Shared genes among pathways might also account for this result. Whichever the reason, the resulting top ranked genes should anyway be the most relevant to BPD pathogenesis thanks to the fact that all likely implicated genes were used for training of the algorithm. On the contrary, when the training set was split in multiple pathways, different genes ranked as first. In order to focus on those ones with the broadest effect, we decided that our candidate genes would have a high rank in all pathways. We arbitrarily chose 100 as a threshold that would represent about 15% of the testing set, and it only gave back 1 confirmed gene (0.1%), (Table 1). Therefore, this strategy is remarkably efficient in focusing on a small number of genes. However, it is possible that the disruption of a single pathway is enough to cause BPD. On account of this, ranks from every pathway will also be kept into consideration for future developments of this study. As a general reflection, we would keep using both strategies for complex diseases, where no unique causative pathway is described.

Our training list was based on internal previous results and on published literature of association studies, expression studies and animal models of BPD. It is to note that adding or removing one gene from the training set considerably changed ranking of the testing set. Since all of the studies we extracted this list from have weak points (sample size, absence of replication, different condition of lung injury…), this list might be improved overtime. Anyway, for the time being, we believe it is the most representative one of pathways involved in BPD pathogenesis.

Our selection of testing genes was extremely focused, for we singled out genes with novel or rare variants that two softwares, predicting the functional impact, classified as deleterious. Moreover, in consideration of the need of confirming variants' functional relevance to BPD pathogenesis, we excerpted genes with more than one variant allele (more different variants in the same gene or more than one sample with the same variant or one sample's homozygosis). In this manner we might have left out important genes, but as a preliminary study we aimed to sort out variants that might be causative per se. This strategy for analysis led us to 1132 variants from the 9427 in tier 3, that reduced our list to 745 potentially interesting genes.

Candidate genes: i) interestingly, we identified two novel missense variants in ATP-binding cassette 3 (*ABCA3*) gene in 3 patients. *ABCA3* participates in the transport of phospholipids to lamellar bodies, the organelles where surfactant is stored before secretion in the alveolus. Recessive mutations in *ABCA3* have been identified in full-term infants with unexplained distress that clinically and radiographically resembles Respiratory Distress Syndrome (RDS) in preterm infants. The two missense variants identified are novel and not reported either in genomic dbSNP, 1000 genomes or in disease specific databases (HGMD, ABCMdb, http://abcmutations.hegelab.org). Thus they seem good candidates because they could contribute to the phenotype in association with other genetic and environmental factors. Our findings also suggest that BPD and RDS may share some genetic aspect, and might be considered as not completely distinct entities, at least in patients subgroups; and ii) several genes belonging to the toll-like receptor (*TLR*) family resulted altered in our cohort (Table 1 and Supp. Table 1). Toll-like receptors are innate immunity receptors that play a fundamental role in pathogen recognition and clearance, regulating inflammatory response and tissue repair. It is known in literature that an alteration of inflammatory response could be a risk factor predisposing to development of pulmonary disorders [47]. Other interesting candidate genes highlighted are *CRP*, C-reactive protein, involved in several host defence related functions [48], *LBP*, encoding a lipopolysaccharide binding protein, LPS, with a key role into the immune response [49], and *MMP1*, a collagenase associated to lung infection [50]. These genes, besides the toll-like receptor family described above, seem to corroborate the theory of a role of inflammatory events in enhancing BPD susceptibility; and iii) *NOS2* was the gene ranked in the first five positions and with a rank above 100 in all pathways. *NOS2* codes for a protein named iNOS, an isoform of nitric oxide synthase (NOS) whose expression is inducible by lipopolysaccharide (LPS) in combination with various cytokines. Once induced, iNOS produces nitric oxide (NO) at a high rate. NO's role in lung development and inflammation is controversial. Indeed, it mediates transition from foetal to neonatal life. For this reason, NO is

sometimes used as a therapy for BPD since its administration in animal models promoted lung development and reduced inflammation [51]. NO is a reactive free radical involved in numerous molecular signalling processes [52]. With regard to its role in immunity responses in the lung, iNOS activity has been proved to be of fundamental importance for alveolar macrophages' activation [53] and cytokine signalling [54]. Moreover, *NOS2* expression by pulmonary microvascular endothelial cells inhibits apoptosis in infiltrating neutrophils [55]. Evidently this property has a double implication: on the one hand, iNOS might contribute to persistence of inflammatory state in the lung, but, on the other hand, it might favour resolution of infections. In this sense, many studies focused on animal models of lung injury by LPS administration. These results support the theory that early proinflammatory events brought on by *NOS2* expression are important for induction of repair mechanisms [53,56]. Moreover a number of studies have also tried to define iNOS activity in hyperoxic lung injury. While it has been proved that high oxygen concentrations induce upregulation of *NOS2* in lungs, their role in lung injury isn't clear. Studies in *NOS2* knock-out mice indicate that they are mainly involved in responding to an inflammatory state induced by reactive oxygen species [57].

In consideration of the results obtained in this pilot study, we can conclude that our approach may be interesting to initiate the dissection of genetic pathogenesis of BPD. Our preliminary results encourage us to pursue along this project, in order to explore also the other candidates picked up with pathway analysis and to extend the study to all the affected patients in our cohort. Potential candidate genes discovered in this preliminary study and in further developments will need to be confirmed with additional evidences derived from appropriate functional studies, from further genetic analyses (i.e. family-based studies) as well as from methods able to evaluate the collective influence of variants. For instance, expression studies on infants' broncho-alveolar washes could be performed as a first validation of the variants' effect. Moreover it will be important also to screen a comparable healthy control population for the list of these putative genes. Not least, with our study we would like to rise the actual knowledge on BPD mechanism and translate it into the clinic, allowing improvement of the condition and shifting to less empirical treatment in affected patients. To this respect, fostering integration of genomic studies either with other “omic” approaches, such as epigenomics, proteomics and metabolomics, or with studies on intrauterine development [58] would be very important to reach a better understanding of pathogenic mechanisms and biochemical pathways involved in BPD.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cca.2015.01.001.

Acknowledgements

We would like to thank all the families participating to the study as well as the Italian association “Un Respiro Nel Futuro Onlus”.

References

[1] Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011; 12:745–55. http://dx.doi.org/10.1038/nrg3031.
[2] Merritt TA, Deming DD, Boynton BR. The “new” bronchopulmonary dysplasia: challenges and commentary. Semin Fetal Neonatal Med 2009;14:345–57. http://dx.doi.org/10.1016/j.siny.2009.08.009.
[3] Horbar JD, Carpenter JH, Badger GJ, Kenny MJ, Soll RF, Morrow KA, et al. Mortality and neonatal morbidity among infants 501 to 1500 grams from 2000 to 2009. Pediatrics 2012;129:1019–26. http://dx.doi.org/10.1542/peds. 2011-3028.
[4] Stoll BJ, Hansen NI, Bell EF, Shankaran S, Laptook AR, Walsh MC, et al. Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network. Pediatrics 2010;126:443–56. http://dx.doi.org/10.1542/peds. 2009-2959.
[5] Gien J, Kinsella JP. Pathogenesis and treatment of bronchopulmonary dysplasia. Curr Opin Pediatr 2011;23:305–13. http://dx.doi.org/10.1097/MOP.0b013e328346577f.
[6] Jobe AH, Bancalari E. Bronchopulmonary dysplasia. Am J Respir Crit Care Med 2001; 163:1723–9. http://dx.doi.org/10.1164/ajrccm.163.7.2011060.
[7] Bhandari V, Gruen JR. The genetics of bronchopulmonary dysplasia. Semin Perinatol 2006;30:185–91. http://dx.doi.org/10.1053/j.semperi.2006.05.005.
[8] Lavoie PM, Pham C, Jang KL. Heritability of bronchopulmonary dysplasia, defined according to the consensus statement of the national institutes of health. Pediatrics 2008;122:479–85. http://dx.doi.org/10.1542/peds. 2007-2313.
[9] Shaw GM, O'Brodovich HM. Progress in understanding the genetics of bronchopulmonary dysplasia. Semin Perinatol 2013;37:85–93. http://dx.doi.org/10.1053/j.semperi.2013.01.004.
[10] Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010;11:415–25. http://dx.doi.org/10.1038/nrg2779.
[11] Marian AJ. Molecular genetic studies of complex phenotypes. Transl Res 2012;159: 64–79. http://dx.doi.org/10.1016/j.trsl.2011.08.001.
[12] Somaschini M, Castiglioni E, Volonteri C, Cursi M, Ferrari M, Carrera P. Genetic predisposing factors to bronchopulmonary dysplasia: preliminary data from a multicentre study. J Matern Fetal Neonatal Med 2012;25(Suppl. 4):127–30. http://dx.doi.org/10.3109/14767058.2012.714995.
[13] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25:1754–60. http://dx.doi.org/10.1093/bioinformatics/btp324.
[14] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24–6. http://dx.doi.org/10.1038/nbt.1754.
[15] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303. http://dx.doi.org/10.1101/gr.107524.110.
[16] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) n.d.;6:80–92. http://dx.doi.org/10.4161/fly.19695.
[17] Rova M, Haataja R, Marttila R, Ollikainen V, Tammela O, Hallman M. Data mining and multiparameter analysis of lung surfactant protein genes in bronchopulmonary dysplasia. Hum Mol Genet 2004;13:1095–104. http://dx.doi.org/10.1093/hmg/ddh132.
[18] Ryckman KK, Dagle JM, Kelsey K, Momany AM, Murray JC. Genetic associations of surfactant protein D and angiotensin-converting enzyme with lung disease in preterm neonates. J Perinatol 2012;32:349–55. http://dx.doi.org/10.1038/jp.2011.104.
[19] Hadchouel A, Durrmeyer X, Bouzigon E, Incitti R, Huusko J, Jarreau P-H, et al. Identification of SPOCK2 as a susceptibility gene for bronchopulmonary dysplasia. Am J Respir Crit Care Med 2011;184:1164–70. http://dx.doi.org/10.1164/rccm.201103-0548OC.
[20] Kazzi SNJ, Jacques SM, Qureshi F, Quasney MW, Kim UO, Buhimschi IA. Tumor necrosis factor-alpha allele lymphotoxin-alpha + 250 is associated with the presence and severity of placental inflammation among preterm births. Pediatr Res 2004;56:94–8. http://dx.doi.org/10.1203/01.PDR.0000130474.12948.A4.
[21] Kwinta P, Bik-Multanowski M, Mitkowska Z, Tomasik T, Legutko M, Pietrzyk JJ. Genetic risk factors of bronchopulmonary dysplasia. Pediatr Res 2008;64:682–8. http://dx.doi.org/10.1203/PDR.0b013e318184edeb.
[22] Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 2009;37:W305–11. http://dx.doi.org/10.1093/nar/gkp427.
[23] Snel B. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 2000;28:3442–4. http://dx.doi.org/10.1093/nar/28.18.3442.
[24] Alejandre-Alcázar MA, Kwapiszewska G, Reiss I, Amarie OV, Marsh LM, Sevilla-Pérez J, et al. Hyperoxia modulates TGF-beta/BMP signaling in a mouse model of bronchopulmonary dysplasia. Am J Physiol Lung Cell Mol Physiol 2007;292: L537–49. http://dx.doi.org/10.1152/ajplung.00050.2006.
[25] Benjamin JT, Carver BJ, Plosa EJ, Yamamoto Y, Miller JD, Liu J-H, et al. NF-kappaB activation limits airway branching through inhibition of Sp1-mediated fibroblast growth factor-10 expression. J Immunol 2010;185:4896–903. http://dx.doi.org/10.4049/jimmunol.1001857.
[26] Bhattacharya S, Go D, Krenitsky DL, Huyck HL, Solleti SK, Lunger VA, et al. Genome-wide transcriptional profiling reveals connective tissue mast cell accumulation in bronchopulmonary dysplasia. Am J Respir Crit Care Med 2012;186:349–58. http://dx.doi.org/10.1164/rccm.201203-0406OC.
[27] Bourbon J, Boucherat O, Chailley-Heu B, Delacourt C. Control mechanisms of lung alveolar development and their disorders in bronchopulmonary dysplasia. Pediatr Res 2005;57:38R–46R. http://dx.doi.org/10.1203/01.PDR.0000159630.35883.BE.
[28] Brasch F, Ochs M, Kahne T, Guttentag S, Schauer-Vukasinovic V, Derrick M, et al. Involvement of napsin A in the C- and N-terminal processing of surfactant protein B in type-II pneumocytes of the human lung. J Biol Chem 2003;278:49006–14. http://dx.doi.org/10.1074/jbc.M306844200.
[29] Bühling F, Kouadio M, Chwieralski CE, Kern U, Hohlfeld JM, Klemm N, et al. Gene targeting of the cysteine peptidase cathepsin H impairs lung surfactant in mice. PLoS One 2011;6:e26247. http://dx.doi.org/10.1371/journal.pone.0026247.
[30] Chauhan M, Bombell S, McGuire W. Tumour necrosis factor (−308A) polymorphism in very preterm infants with bronchopulmonary dysplasia: a meta-analysis. Arch Dis Child Fetal Neonatal Ed 2009;94:F257–9. http://dx.doi.org/10.1136/adc.2008.153122.
[31] Hikino S, Ohga S, Kinjo T, Kusuda T, Ochiai M, Inoue H, et al. Tracheal aspirate gene expression in preterm newborns and development of bronchopulmonary dysplasia. Pediatr Int 2012;54:208–14. http://dx.doi.org/10.1111/j.1442-200X.2011.03510.x.
[32] Iosef C, Alastalo T-P, Hou Y, Chen C, Adams ES, Lyu S-C, et al. Inhibiting NF-κB in the developing lung disrupts angiogenesis and alveolarization. Am J Physiol Lung Cell Mol Physiol 2012;302:L1023–36. http://dx.doi.org/10.1152/ajplung.00230.2011.
[33] Kho AT, Bhattacharya S, Tantisira KG, Carey VJ, Gaedigk R, Leeder JS, et al. Transcriptomic analysis of human lung development. Am J Respir Crit Care Med 2010;181:54–63. http://dx.doi.org/10.1164/rccm.200907-1063OC.

[34] Londhe VA, Maisonet TM, Lopez B, Jeng J-M, Xiao J, Li C, et al. Conditional deletion of epithelial IKKβ impairs alveolar formation through apoptosis and decreased VEGF expression during early mouse lung morphogenesis. Respir Res 2011;12:134. http://dx.doi.org/10.1186/1465-9921-12-134.
[35] Mallakin A, Kutcher LW, McDowell SA, Kong S, Schuster R, Lentsch AB, et al. Gene expression profiles of Mst1r-deficient mice during nickel-induced acute lung injury. Am J Respir Cell Mol Biol 2006;34:15–27. http://dx.doi.org/10.1165/rcmb.2005-0093OC.
[36] Manar MH, Brown MR, Gauthier TW, Brown LAS. Association of glutathione-S-transferase-P1 (GST-P1) polymorphisms with bronchopulmonary dysplasia. J Perinatol 2004;24:30–5. http://dx.doi.org/10.1038/sj.jp.7211020.
[37] Ray M, Yu S, Sharda DR, Wilson CB, Liu Q, Kaushal N, et al. Inhibition of TLR4-induced IκB kinase activity by the RON receptor tyrosine kinase and its ligand, macrophage-stimulating protein. J Immunol 2010;185:7309–16. http://dx.doi.org/10.4049/jimmunol.1000095.
[38] Roos AB, Berg T, Nord M. A relationship between epithelial maturation, bronchopulmonary dysplasia, and chronic obstructive pulmonary disease. Pulm Med 2012;2012:196194. http://dx.doi.org/10.1155/2012/196194.
[39] Sadakata T, Washida M, Morita N, Furuichi T. Tissue distribution of Ca2 +-dependent activator protein for secretion family members CAPS1 and CAPS2 in mice. J Histochem Cytochem 2007;55:301–11. http://dx.doi.org/10.1369/jhc.6A7033.2006.
[40] Sampath V, Garland JS, Le M, Patel AL, Konduri GG, Cohen JD, et al. A TLR5 (g.1174C > T) variant that encodes a stop codon (R392X) is associated with bronchopulmonary dysplasia. Pediatr Pulmonol 2012;47:460–8. http://dx.doi.org/10.1002/ppul.21568.
[41] Ueno T, Linder S, Na C-L, Rice WR, Johansson J, Weaver TE. Processing of pulmonary surfactant protein B by napsin and cathepsin H. J Biol Chem 2004;279:16178–84. http://dx.doi.org/10.1074/jbc.M312029200.
[42] Wu S, Platteau A, Chen S, McNamara G, Whitsett J, Bancalari E. Conditional overexpression of connective tissue growth factor disrupts postnatal lung development. Am J Respir Cell Mol Biol 2010;42:552–63. http://dx.doi.org/10.1165/rcmb.2009-0068OC.
[43] Zhou Y-Q, Chen Y-Q, Fisher JH, Wang M-H. Activation of the RON receptor tyrosine kinase by macrophage-stimulating protein inhibits inducible cyclooxygenase-2 expression in murine macrophages. J Biol Chem 2002;277:38104–10. http://dx.doi.org/10.1074/jbc.M206167200.
[44] Carraro S, Filippone M, Da Dalt L, Ferraro V, Maretti M, Bressan S, et al. Bronchopulmonary dysplasia: the earliest and perhaps the longest lasting obstructive lung disease in humans. Early Hum Dev 2013;89(Suppl. 3):S3–5. http://dx.doi.org/10.1016/j.earlhumdev.2013.07.015.
[45] Herriges M, Morrisey EE. Lung development: orchestrating the generation and regeneration of a complex organ. Development 2014;141:502–13. http://dx.doi.org/10.1242/dev.098186.
[46] Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet 2012;28:323–32. http://dx.doi.org/10.1016/j.tig.2012.03.004.
[47] Brasch F, Schimanski S, Mühlfeld C, Barlage S, Langmann T, Aslanidis C, et al. Alteration of the pulmonary surfactant system in full-term infants with hereditary ABCA3 deficiency. Am J Respir Crit Care Med 2006;174:571–80. http://dx.doi.org/10.1164/rccm.200509-1535OC.
[48] Fruchter O, Rosengarten D, Goldberg E, Ben-Zvi H, Tor R, Kramer MR. Airway bacterial colonization and serum C-reactive protein are associated with chronic obstructive pulmonary disease exacerbation following bronchoscopic lung volume reduction. Clin Respir J 2014. http://dx.doi.org/10.1111/crj.12211.
[49] Eckert JK, Kim YJ, Kim JI, Gürtler K, Oh D-Y, Sur S, et al. The crystal structure of lipopolysaccharide binding protein reveals the location of a frequent mutation that impairs innate immunity. Immunity 2013;39:647–60. http://dx.doi.org/10.1016/j.immuni.2013.09.005.
[50] Joos L, He J-Q, Shepherdson MB, Connett JE, Anthonisen NR, Paré PD, et al. The role of matrix metalloproteinase polymorphisms in the rate of decline in lung function. Hum Mol Genet 2002;11:569–76.
[51] Schulzke SM, Pillow JJ. The management of evolving bronchopulmonary dysplasia. Paediatr Respir Rev 2010;11:143–8. http://dx.doi.org/10.1016/j.prrv.2009.12.005.
[52] Hampl V, Herget J. Role of nitric oxide in the pathogenesis of chronic pulmonary hypertension. Physiol Rev 2000;80:1337–72.
[53] D'Alessio FR, Tsushima K, Aggarwal NR, Mock JR, Eto Y, Garibaldi BT, et al. Resolution of experimental lung injury by monocyte-derived inducible nitric oxide synthase. J Immunol 2012;189:2234–45. http://dx.doi.org/10.4049/jimmunol.1102606.
[54] Okamoto T, Gohil K, Finkelstein EI, Bove P, Akaike T, van der Vliet A. Multiple contributing roles for NOS2 in LPS-induced acute airway inflammation in mice. Am J Physiol Lung Cell Mol Physiol 2004;286:L198–209. http://dx.doi.org/10.1152/ajplung.00136.2003.
[55] Wang L, Mehta S, Gillis C, Law C, Taneja R. Modulation of neutrophil apoptosis by murine pulmonary microvascular endothelial cell inducible nitric oxide synthase. Biochem Biophys Res Commun 2010;401:207–12. http://dx.doi.org/10.1016/j.bbrc.2010.09.029.
[56] Zeidler PC, Millecchia LM, Castranova V. Role of inducible nitric oxide synthase-derived nitric oxide in lipopolysaccharide plus interferon-gamma-induced pulmonary inflammation. Toxicol Appl Pharmacol 2004;195:45–54. http://dx.doi.org/10.1016/j.taap.2003.10.005.
[57] Potter CF, Kuo NT, Farver CF, McMahon JT, Chang CH, Agani FH, et al. Effects of hyperoxia on nitric oxide synthase expression, nitric oxide activity, and lung injury in rat pups. Pediatr Res 1999;45:8–13. http://dx.doi.org/10.1203/00006450-199901000-00003.
[58] Fanos V, Pintus MC, Lussu M, Atzori L, Noto A, Stronati M, et al. Urinary metabolomics of bronchopulmonary dysplasia (BPD): preliminary data at birth suggest it is a congenital disease. J Matern Fetal Neonatal Med 2014;27(Suppl. 2):39–45. http://dx.doi.org/10.3109/14767058.2014.955966.