

# Gene expression studies in cells from primary ciliary dyskinesia patients identify 208 potential ciliary genes

Maciej Geremek · Marcel Bruinenberg ·  
Ewa Ziętkiewicz · Andrzej Pogorzelski ·  
Michał Witt · Cisca Wijmenga

Received: 5 September 2010 / Accepted: 15 November 2010 / Published online: 7 December 2010  
© Springer-Verlag 2010

**Abstract** Cilia are small cellular projections that either act as sensors (primary cilia) or propel fluid over the epithelia of various organs (motile cilia). The organellum has gained much attention lately because of its involvement in a group of human diseases called ciliopathies. Primary ciliary dyskinesia (PCD) is an autosomal recessive ciliopathy caused by mutations in cilia motility genes. The disease is characterized by recurrent respiratory tract infections due to the lack of an efficient mucociliary clearance. We performed whole-genome gene expression profiling in bronchial biopsies from PCD patients. We used the quality threshold clustering algorithm to identify

groups of genes that revealed highly correlated RNA expression patterns in the biopsies. The largest cluster contained 372 genes and was significantly enriched for genes related to cilia. The database and literature search showed that 164 genes in this cluster were known cilia genes, strongly indicating that the remaining 208 genes were likely to be new cilia genes. The tissue expression pattern of the 208 new cilia genes and the 164 known genes was consistent with the presence of motile cilia in a given tissue. The analysis of the upstream promotor sequences revealed evidence for RFX transcription factors binding site motif in both subgroups. Based on the correlated expression patterns in PCD-affected tissues, we identified 208 genes that we predict to be involved in cilia biology. Our predictions are based directly on the human material and not on model organisms. This list of genes provides candidate genes for PCD and other ciliopathies.

M. Witt and C. Wijmenga equally contributed to the manuscript.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00439-010-0922-4) contains supplementary material, which is available to authorized users.

M. Geremek (✉) · M. Witt  
International Institute of Molecular and Cell Biology,  
Warsaw, Poland  
e-mail: macgerem@gmail.com

M. Geremek · M. Bruinenberg · C. Wijmenga  
Department of Genetics, University Medical Center Groningen,  
9700 RB Groningen, The Netherlands

M. Geremek · C. Wijmenga  
Complex Genetics Group, Department of Biomedical Genetics,  
University Medical Center Utrecht, 3508 AB Utrecht,  
The Netherlands

E. Ziętkiewicz · M. Witt  
Department of Molecular and Clinical Genetics,  
Institute of Human Genetics, Poznan, Poland

A. Pogorzelski  
Institute of Tuberculosis and Lung Diseases, Rabka, Poland

## Introduction

Cilia are built on a scaffold of nine peripheral microtubule doublets. In motile cilia, the nine peripheral microtubule doublets are accompanied by two central microtubules (the 9+2 structure). Radial spokes connect the peripheral and central microtubules, while peripheral doublets are connected with each other by nexin links. Outer and inner dynein arms are anchored to the peripheral microtubule doublet and produce the force needed for the movement of cilia. Synchronized beating of the cilia generates the flow of mucus and cerebrospinal fluid in the respiratory tract and in the brain, respectively, whereas in the fallopian tubes cilia help in moving the ovum toward the uterus. Flagella, built like motile cilia with a similar scheme of 9+2 microtubules, provide motility to unicellular organisms or

cells, such as spermatozoa. Primary cilia—which lack the central pair of microtubules (the 9+0 structure), dynein arms and radial spokes—act as sensory organelles, displaying receptor molecules and sensing chemical and mechanical stimuli (Satir and Christensen 2008). However, it is important to note that motile cilia perform a variety of mechanosensory or chemosensory functions and primary cilia located, for example, on the embryonic node are motile (Bloodgood 2010).

Nevertheless, our understanding of the molecular composition of cilia is far from complete. Several approaches have been taken to characterize the cilia genome and proteome. Database search for tissue-specific expression pattern limited to ciliated tissues was used to predict 99 cilia related murine genes (McClintock et al. 2008). Liquid chromatography-mass spectrometry of human epithelial cilia identified 164 axonemal proteins (Ostrowski et al. 2002). A mass spectrometry study of flagella of the unicellular alga *Chlamydomonas reinhardtii* identified 360 proteins very likely to be involved in cilia formation, and 292 that are probably involved (Pazour et al. 2005). In addition, when comparing the genomes of non-ciliated organisms to ciliated organisms predicted 688 and 183 cilia-related genes in *Chlamydomonas reinhardtii* (Li et al. 2004) and *Drosophila melanogaster* (Avidor-Reiss et al. 2004), respectively. The presence of X-box promoter elements, which are targets of cilia-related transcription factors, has led to the identification of additional ciliary genes in *Caenorhabditis elegans* (Blacque et al. 2005; Efimenko et al. 2005). The results of these different studies have been assembled in online databases: the Ciliome Database (Inglis et al. 2006) and the Cilia Proteome database (Gherman et al. 2006), which also contains basal bodies proteins. As far as human motile cilia are concerned, the database is probably far from complete since only a single high-throughput proteomics study was performed on mammalian cilia (Ostrowski et al. 2002). Most of what we know about cilia comes from work on the unicellular biflagellate alga *Chlamydomonas reinhardtii*. Although the evolutionary conservation of ciliary proteins is remarkably high, it can be assumed that mammalian cilia are more complex. Moreover, the genes related to the presence of multiple cilia in one epithelial cell, the genes associated with the coordinated beating of multiple cilia, as well as the genes coding for receptor molecules responsible for communication with the environment of the human body, are likely to be missing from the database.

Defects in both primary and motile cilia have been associated with a group of pleiotropic, clinically overlapping, human diseases called ciliopathies, such as primary ciliary dyskinesia (PCD), polycystic liver and kidney disease, retinitis pigmentosa, nephronophthisis, Bardet–Biedl syndrome, Alstrom syndrome, and Meckel–Gruber syndrome

(Badano et al. 2006; Valente et al. 2006). PCD is a rare, genetically heterogeneous disorder caused by mutations in genes encoding proteins important for cilia motility (Geremek and Witt 2004; Pennarun et al. 1999; Olbrich et al. 2002; Bartoloni et al. 2002; Loges et al. 2002; Omran et al. 2008; Castleman et al. 2009; Duquesnoy et al. 2009; Loges et al. 2009). The disease is characterized by recurrent respiratory tract infections, bronchiectasis and infertility. Pulmonary symptoms occur due to the lack of an efficient mucociliary clearance caused by kinetic dysfunction of respiratory cilia. Male infertility is caused by immotility of flagella in spermatozooids. *Situs inversus*, a mirror reversal of thoracic organs positioning, is present in approximately half of the PCD patients because of immotility of primary cilia of the embryonic node. Nodal cilia are motile and generate leftward flow of extraembryonic fluid in the nodal pit. The flow has been identified as the initial left–right symmetry breaking event during embryogenesis. According to one of the hypotheses, the flow generates movement of nodal vesicular parcels to the left edge of the node. The parcels contain signaling molecules that activate signaling pathways. In case of immotility of nodal cilia, as in PCD, left–right determination is randomized causing *situs inversus* in approximately half of the patients (Nobutaka et al. 2009).

We observed a down-regulation of the expression of dynein genes and other ciliary genes in PCD patients as compared to the controls (unpublished observations). More importantly, we noticed that more genes followed the pattern of dyneins expression in PCD patients, suggesting a functional relationship between the co-regulated genes. We postulated that gene expression patterns in PCD patients could be used to identify mammalian cilia genes. For this study, we performed gene expression analysis on bronchial biopsies from PCD patients to identify groups of highly correlated cilia-related genes. In contrast to most of previous studies our predictions are based directly on the human material and not on model organisms. We report a cluster of 164 genes previously linked to cilia and 208 new genes that we predict will be involved in cilia-related processes.

## Materials and methods

### Bronchial biopsies

We collected bronchial biopsies of six clinically diagnosed PCD patients and nine control subjects that were referred to the hospital for unrelated reasons. All patients had a saccharine test and light microscopy imaging of cilia motility characteristic for PCD. An electron microscopy evaluation of ciliary defects was performed, however, in two specimens an insufficient number of epithelial cells was recovered for

inspection (Supplementary table 1). The concentration of nitric oxide (NO) in the nasal cavity was measured with a chemiluminescence analyzer, with a threshold value of 200 ppb for diagnosing PCD (Karadag et al. 1999). Patient #1 had a concentration of nitric oxide in the nasal cavity little below the threshold and no electron microscopy imaging of cilia. Therefore, a second measurement of nitric oxide was performed confirming low concentration. He had a typical course of the disease with a resection of middle lobe due to necrosis and typical for PCD picture in bronchoscopic examination and nasal cavity examination. The specimens from non-PCD controls were obtained through the same protocol; these individuals were referred to the Institute of Tuberculosis and Lung Diseases in Rabka, Poland for regular check-ups, with no symptoms of acute disease, with no bronchoscopic signs of the disturbance of mucociliary transport and with normal ciliary beating in the light microscopy. The study was ethically approved by the institutional review board. Informed consent was obtained from all the subjects.

#### Gene expression

Anti-sense RNA was synthesized, amplified and purified using the Ambion Illumina TotalPrep Amplification Kit (Ambion, USA) following the manufacturer's protocol. Complementary RNA was hybridized to Illumina HumanRef-12 Whole Genome BeadChips and scanned on the Illumina BeadArray Reader. The gene expression data has been submitted to GEO under accession number GSE11501.

#### Data preprocessing

The initial steps of data preprocessing and quantile normalization were performed in the Illumina BeadStudio Gene Expression module v3.2. Expression values below 5 were thresholded to 5 and scaled by base 2 logarithm. We limited our data to Ensembl database v52 coding genes. The probes with an expression value variance in the lower 25% of the data were removed. To further limit the computations and to filter out genes not stably expressed in bronchial tissue, we removed probes not detected as present in 4/9 control individuals who had a diagnostic bronchoscopy, but did not display PCD symptoms. The remaining 13,811 probes were subject to clustering.

#### Clustering

The quality threshold (QT) clustering algorithm was implemented in C according to the algorithm described previously. The QT clustering algorithm performs a computationally extensive search for groups of correlated

genes. It looks for largest cluster by iteratively putting every gene in the center of a potential cluster and adding to it genes with the highest Jackknife correlation coefficients in a way that minimizes the cluster diameter  $d$ , until no further genes can be added without exceeding a predefined  $d$  value (Heyer et al. 1999; Coppe et al. 2009). The quality of clusters was ensured by keeping the cluster diameter  $<0.3$ .

#### Database search

For gene annotation enrichment analysis, we used the DAVID (Database for Annotation, Visualization, and Integrated Discovery: <http://www.david.abcc.ncifcrf.gov/home.jsp>) (Huang da et al. 2007; Kouwenhoven et al. 2010). The tool calculates over-representation of specific gene ontology terms with respect to the total number of genes assayed and annotated. DAVID applies a modified Fisher exact test, to establish if the proportion of genes falling into an annotation category significantly differs for a particular group of genes and the background group of genes. Ensembl Gene IDs of the cluster A genes were used as queries and the whole set of genes on the Illumina 12HT chip was used as the background group. The Ciliome Database ([http://www.sfu.ca/~leroux/ciliome\\_home.htm](http://www.sfu.ca/~leroux/ciliome_home.htm)) was queried with Ensembl Gene IDs of cluster A genes and Mouse Ensembl Gene IDs of their orthologues to find known ciliary genes (Inglis et al. 2006). The Ciliary Proteome Database (Gherman et al. 2006) (<http://www.ciliaproteom.com>) was queried with UniProt/SwissProt accessions of clusterA genes. Biomart module of the Ensembl database was used for ID/accession conversions. In addition, Pubmed was searched for publications linking individual genes to cilia or flagella.

#### Tissue-specific expression

The number of EST transcripts for analyzed gene in 30 different tissue types or organ pools was fetched from the UniGene database. Tissue expression enrichment scores were calculated as described in (Yu et al. 2006; Colecchia et al. 2009). In brief, enrichment score  $ES_i(g) = o_i(g)/e_i(g)$  is the ratio between observed to expected number of ESTs for gene  $g$  in tissue  $i$ . The total number of ESTs in UniGene for gene  $g$  is  $T(g) = \sum_i o_i(g)$ . Given the total size of EST libraries in tissue  $i$ ,  $s_i$ , the expected number of ESTs in tissue  $i$  for each gene is proportional to  $p_i = s_i/\sum_i s_i$ . For gene  $g$ , if it is expressed equally across all tissues, the expected number of ESTs in tissue  $i$  is equal to  $e_i = T(g)p_i$ . The mean enrichment scores of the analyzed groups of genes in each of the tissues were calculated and plotted for visualization of the tissue-specific expression patterns.

## Transcription factor-binding sites

Over-representation of transcription factor binding sites was evaluated in PAINT v3.9 (Vadigepalli et al. 2003; Riehle et al. 2008) interfaced with the TRANSFAC Professional v 2009.2 database (Wingender et al. 1996). The 500 base pairs upstream from transcription starting sites were extracted from Ensembl database v52. A search for transcription factor binding sites was performed by the Match program with matrices deposited in the TRANSFAC database and filter option set to minimize false positives.  $p$  values were calculated using all human upstream sequences in the PAINT database as control group and false-discovery rate as the multiple testing correction.

## Results

### Clustering of genes based on the expression profiling of PCD tissue

We performed whole-genome gene expression profiling in bronchial tissue of six PCD patients using Illumina HT-12 bead arrays. From the 48,803 probes on the arrays, 13,811 probes passed the filtering steps and were used for further analysis. We first used the QT (quality threshold) clustering algorithm to identify groups of correlated genes in PCD patients. We identified 12 clusters with more than 100 genes; the clusters ranged in size from 100 (panel L, Fig. 1) to 372 genes (panel A, Fig. 1). We used DAVID (Database for Annotation, Visualization, and Integrated Discovery) (Huang da et al. 2007; Kouwenhoven et al. 2010) to perform functional annotation enrichment analysis on all cluster members. DAVID applies a modified Fisher exact test to establish if the proportion of genes falling into an annotation category significantly differs for a particular group of genes and the background group of genes. The whole set of genes on the Illumina 12HT chip was used as the background group. The terms significantly enriched in cluster A were related to cilia, flagella and microtubules ( $p < 0.05$  after Bonferroni's correction) (Table 1). The remaining 11 clusters were not significantly enriched for terms that would have indicated a relation to a specific functional process. We also used DAVID to investigate the tissue expression patterns of the 372 genes in cluster A and found that they are mainly expressed in tissues known to have ciliated epithelium, such as testis, lung and trachea (Table 2).

### Characterization of cluster A genes

We queried the publicly available Ciliome Database (Inglis et al. 2006) to check how many members of our cluster A

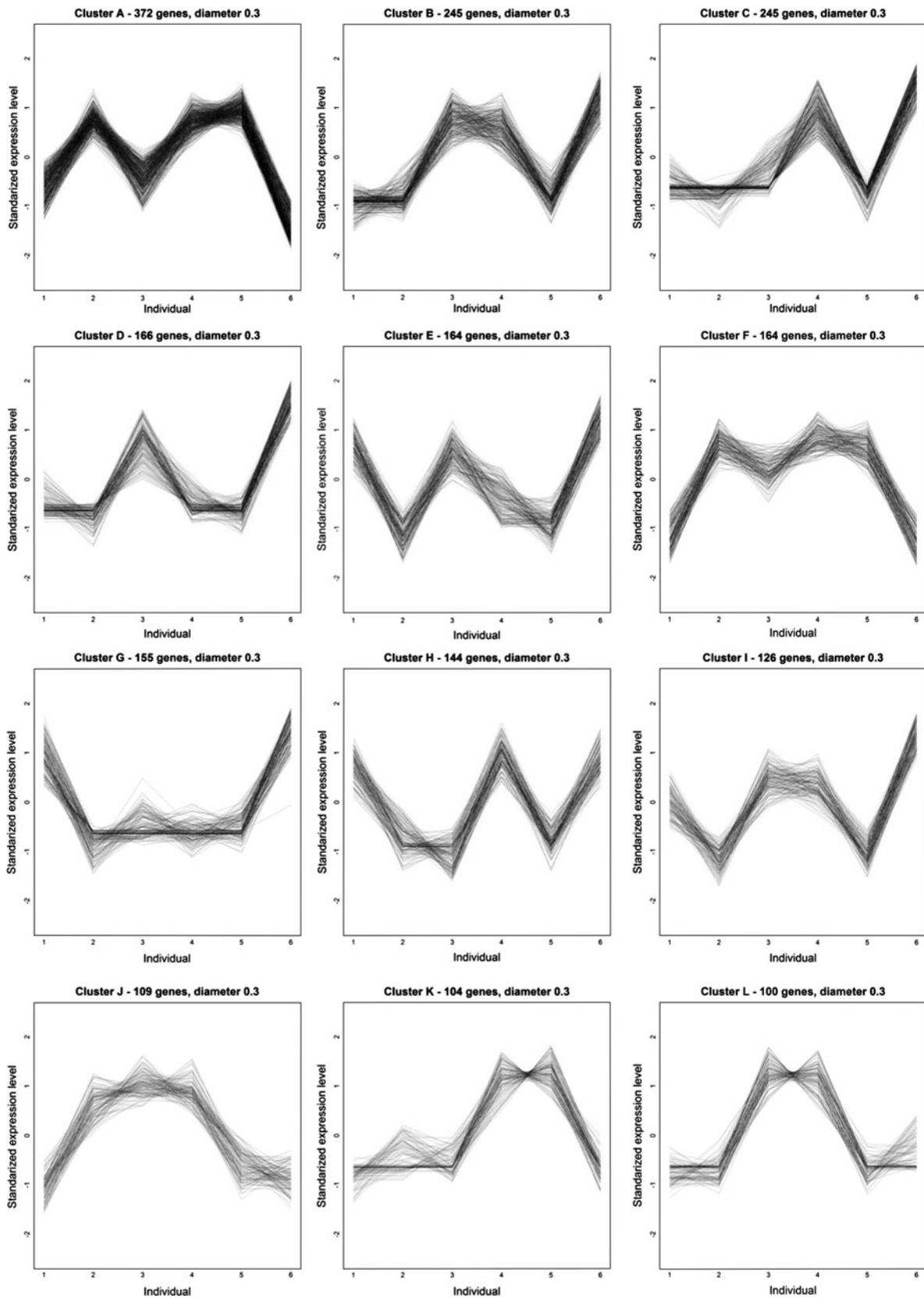
had been previously linked to cilia and found 121 genes to be present in this database.

A similar search in the Ciliary Proteome Database (Gherman et al. 2006) resulted in classification of 115 genes as ciliary, 14 of which were not present in the Ciliome Database. A search in PubMed for publications linking cilia or flagella to any of the individual genes from cluster A led to classification of an additional 29 ciliary genes, resulting in a total of 164 known ciliary genes (i.e. 44% of cluster A; Supplementary table 2). These results strongly suggest that cluster A is a ciliary gene cluster (binomial  $p < 1.10^{-8}$  if we assume 10% of human genes to be related to cilia). This would also suggest that the remaining 208 genes are related to cilia and these may represent new cilia genes not previously reported (Supplementary table 3). We observed the highest percentage of shared genes with the part of the Ciliome Database that was built on the results of experimental studies of motile cilia (Table 3). Cluster F (Fig. 1) has an expression curve shape very similar to cluster A and 30% of the genes in cluster F also proved to be linked to cilia. However, we limited our analysis to cluster A genes only.

To further investigate tissue expression patterns of cluster A genes, we calculated expression enrichment scores for the 30 tissues from the Unigenes dbEST database (Yu et al. 2006; Colecchia et al. 2009). We performed the calculations both on the full set of cluster A genes and on two subsets of cluster A members: genes previously linked to cilia and the new genes identified in this study as potential ciliary genes. Each of the genes was given an enrichment score based on the ratio between the number of ESTs representing it in a given tissue and the expected number of ESTs for that gene in the tissue. The expected number of ESTs for a given gene was calculated assuming equal expression across all tissues. The mean enrichment scores for all cluster A genes and the subsets were calculated for each tissue to present the tissue-specific expression pattern of the genes (Fig. 2).

The mean enrichment score for cluster A genes and the two subsets was the highest for testis, trachea and lung, which are tissues known to have motile cilia. Brain tissue, where motile cilia are present in the ependyma, kidney and connective tissue having primary cilia, and the eye where rods and cones have a small connecting cilium, were ranked high; 8 out of the 10 highest ranking tissues were the same in the two subgroups of known ciliary genes and genes not previously linked to cilia.

We screened the 500 base pairs (bp) upstream of each gene from cluster A for over-representation of transcription factor binding sites using PAINT v 3.9 (Vadigepalli et al. 2003; Riehle et al. 2008) and the TRANSFAC Professional v 2009.2 database (Wingender et al. 1996) (Table 3). The binding site for regulatory factor X (RFX) family of



**Fig. 1** The 12 largest clusters of correlated genes obtained from bronchial tissue of 6 PCD patients. The expression was mean centered and divided by the standard deviation

**Table 1** Distribution of gene ontology (GO) terms in cluster A analyzed by DAVID

GO term	Number of genes	%	<i>p</i>	Bonferroni's corrected <i>p</i> value
Microtubule-based process	15	3.9	1.10E−06	6.00E−03
Microtubule cytoskeleton	28	7.3	5.50E−12	4.80E−09
Axoneme	9	2.3	2.30E−10	2.00E−07
Microtubule	20	5.2	3.10E−10	2.70E−07
Cilium	10	2.6	6.50E−09	5.60E−06
Dynein complex	9	2.3	2.40E−08	2.10E−05
Cell projection	10	2.6	4.10E−08	3.60E−05
Microtubule associated	13	3.4	8.90E−08	7.70E−05
Cell projection	19	4.9	9.80E−07	8.50E−04
Axonemal dynein	6	1.6	1.10E−06	9.30E−04
Axoneme part	6	1.6	1.50E−06	1.30E−03
Cytoskeletal part	27	7	3.70E−06	3.20E−03
Microtubule motor	11	2.9	4.20E−07	1.20E−03

**Table 2** Tissue-specific expression enrichment analysis

UNIGENE EST term	Count	%	<i>p</i>	Bonferroni's corrected <i>p</i> value
Testis third quartile	230	59.9	8.90E−29	6.80E−27
Lung third quartile	202	52.6	9.30E−27	7.10E−25
Uterine tumor third quartile	153	39.8	5.50E−16	4.20E−14
Trachea third quartile	121	31.5	7.90E−13	6.00E−11
Uterus third quartile	135	35.2	1.10E−08	8.60E−07
Brain third quartile	171	44.5	2.20E−08	1.70E−06
Kidney third quartile	146	38	5.00E−08	3.80E−06

Presence of a given gene in a tissue third quartile means that 75% of genes have a lower number of transcripts (ESTs) in the given tissue. The count and percentage of genes from cluster A are shown as well as statistical significance of over-representation in cluster A as compared to the genomic control.

transcription factors was significantly over-represented (false-discovery (FDR) rate adjusted  $p < 10^{-6}$ ) in the upstream sequences of cluster A genes, and in both the known and new ciliary genes in cluster A as well as in the “motile part” of the Ciliome Database (Table 4; Supplementary table 4).

The expression of the cluster A genes in the control samples was stable, but not highly correlated (Fig. 3a). The QT clustering in 6 controls returned 13 clusters of more than 100 genes. None of the clusters was significantly enriched in annotation terms related to cilia. The biggest control cluster contained 383 genes (Fig. 3b) and was not significantly enriched in any annotation term. Thirteen percent of genes in this cluster have been previously related to cilia, which was not statistically significant. The tissue expression pattern was different from cluster.

A particularly the control cluster genes had much lower expression in testis. A similar pattern could be observed also for other control clusters (Fig. 3d). The clustering performed in 6 patients together with 6 controls returned a

cluster of 205 genes with similar to cluster A pattern of expression in the patients and 171(83%) of genes present also in cluster A (Fig. 3c). Forty-four percent of genes in this cluster have been previously related to cilia.

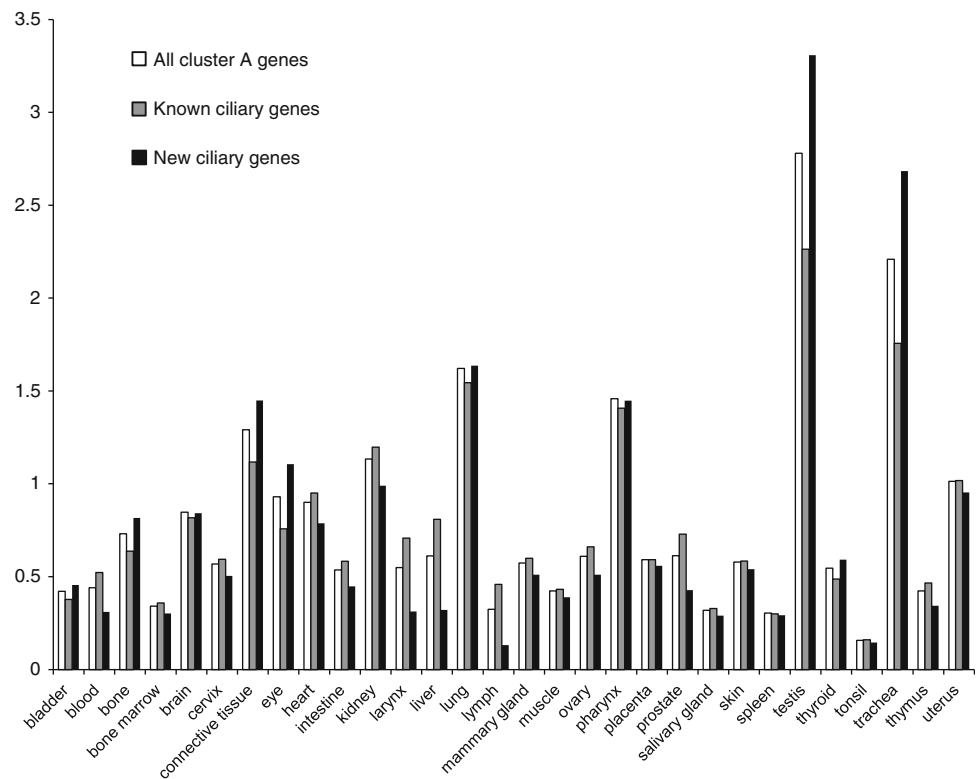
## Discussion

We have identified a group of 208 new genes that are likely to be involved in cilia-related processes. We used gene expression profiling of bronchial tissue from PCD patients to obtain a cluster of 372 genes with highly correlated expression. Forty-four percent of the cluster members have been previously related to cilia based on high-throughput studies or individual experimental studies. The remainder of the cluster members is very likely to be cilia-related genes as they showed tissue-specific expression patterns in accordance with the presence of cilia in the given tissues and they also showed over-presentation of RFX-binding sites in the upstream sequences. The tissue-specific

**Table 3** The intersection of cluster A members and the Ciliome Database

	Type of cilia			
	Primary (non-motile)		Motile cilia	
	Target organism			
	<i>Drosophila melanogaster</i>	<i>Caenorhabditis elegans</i>	<i>Chlamydomonas reinhardtii</i>	<i>Tetrahymena thermophila</i>
Study	Avidor-Reiss et al. (2004)	Blacque et al. (2005)	Efimenko et al. (2005)	Li et al. (2004)
Method	CG	SAGE	XBOX	XBOX
Number of genes mapped to human genome	160	681	768	184
Number of genes in cluster A	31	16	24	11
Percentage of genes from each study in cluster A	19.3	2.3	3.1	5.9
			MS	MS
			91	128
			38	44
			41.7	34.3
			20.7	36.6
			60	60
			323	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6
			67	22
			18.1	36.6
			309	60
			38	22
			41.7	36.6
			20.7	36.6

**Fig. 2** Mean expression enrichment scores of known ciliary genes in cluster A, new cilia related genes in cluster A and all cluster A genes in 30 different tissues. The enrichment score for a gene represents its relative expression in the tissue comparing with other tissues. The highest relative expression in ciliated tissues can be observed for the three groups of genes



**Table 4** Over-representation of transcription factor-binding sites in the 500-bp upstream sequences of cluster A genes and motile cilia genes from the Ciliome Database

Transcription factor <sup>a</sup>	Cluster A ( <i>n</i> = 372)		New ciliary prediction in cluster A ( <i>n</i> = 208)		Known ciliary genes in cluster A ( <i>n</i> = 164)		Motile ciliome ( <i>n</i> = 648)		Number of sites in the reference group ( <i>n</i> = 44,236) <sup>b</sup>
	Number of sites	FDR adjusted <i>p</i> value	Number of sites	FDR adjusted <i>p</i> value	Number of sites	FDR adjusted <i>p</i> value	Number of sites	FDR adjusted <i>p</i> value	
Kid3	336	0.000000	189	0.015281	147	0.025793	572	0.000000	36,299
MAZ	81	0.000172	43	0.030229	38	0.011130	113	0.003006	5,854
RFX	128	0.000000	70	0.000000	58	0.000000	172	0.000000	8,768
RFX1	49	0.000000	23	0.002173	26	0.000000	61	0.000000	1,918
Sp1	91	0.000000	45	0.002853	46	0.000000	163	0.000000	5,329
v-Myb	107	0.000000	58	0.015938	49	0.005906	204	0.000000	8,273
ZF5	118	0.000000	62	0.000000	56	0.000000	216	0.000000	7,490
ETF	88	0.000000	45	0.015281	43	0.000318	165	0.000000	5,911

<sup>a</sup> Only the transcription factors that remained significant after multiple testing correction in all groups of genes are shown

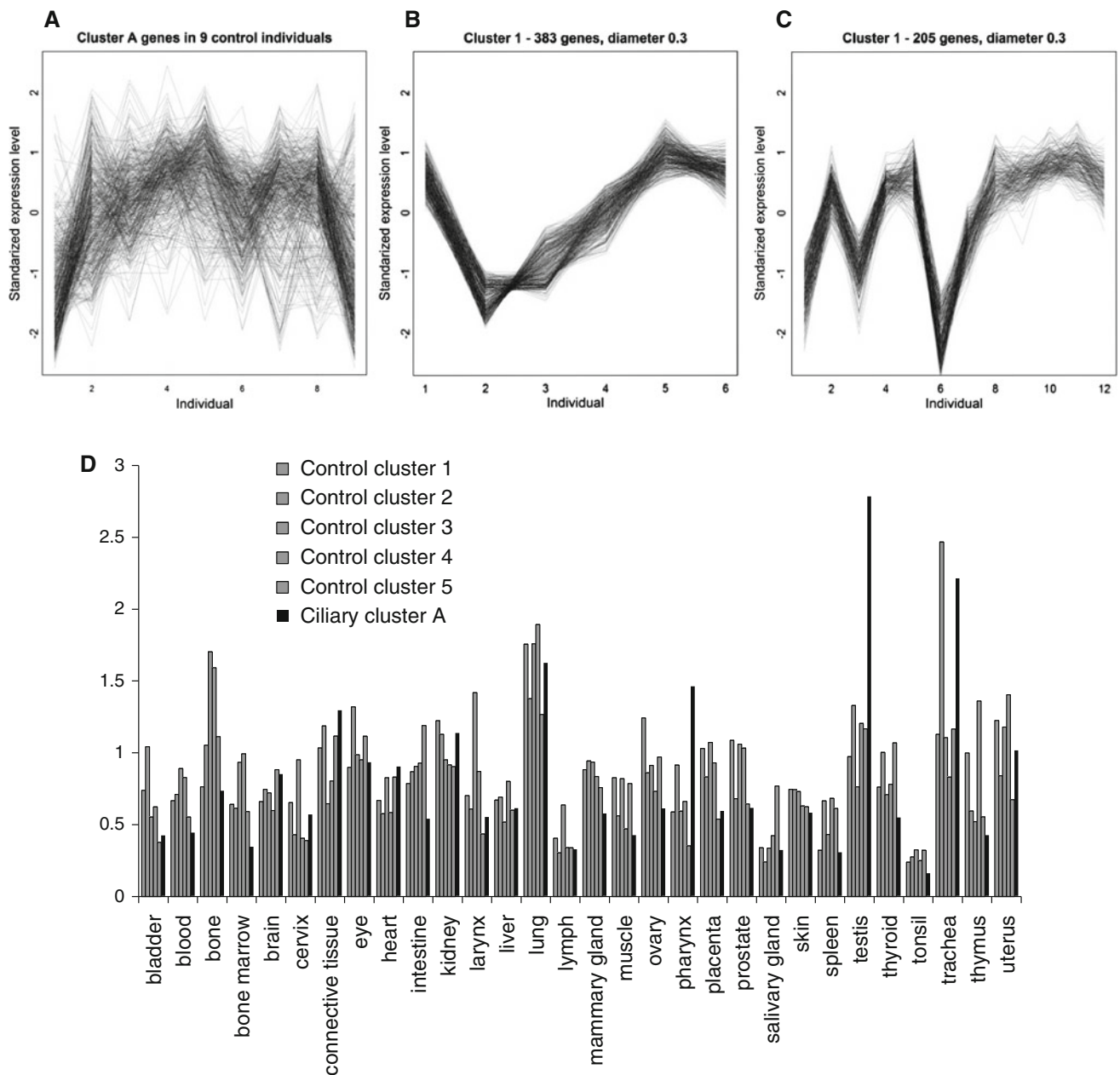
<sup>b</sup> All human promoters sequences in PAIN databases were used as reference set. FDR false-discovery rate

*BBS5*, *BBS11*), Meckel–Gruber syndrome (*MKS1*), nephronoptosis (*NPHP1*, *NPHP2*, *NPHP3*) and PCD (*DNAI1*, *DNAI2*, *RSPH4A*, *RSPH9*, *LRRC50*). Cluster A also contains two limb–girdle muscular dystrophy genes (*FUKUTIN*, *TRIM32*), one of which is also mutated in Bardet–Biedl syndrome (*TRIM32*), and three genes associated with retinitis pigmentosa (*PDE6B*, *Pronin* and *RPGR*). *RPGR* is known to cause a complex phenotype with both

retinitis pigmentosa and PCD in some patients (Moore et al. 2006).

The ciliary proteome has been studied extensively using different high-throughput methods, including mass spectrometry proteomic studies and comparative genomics. All studies, but one (Ostrowski et al. 2002) were performed in lower organisms and required a Blast search to identify possible human orthologs. Our ciliary set is based on the





**Fig. 3** The expression of 372 cluster A genes in 9 non-PCD control individuals (a). The biggest cluster of correlated gene expression in 6 control individuals (b). The biggest cluster in six control individuals and six patients showing a similar curve as cluster A (c). Tissue-

specific expression pattern of five biggest control clusters (grey) and the cluster A (black). Cluster A is the only one with high relative expression in testis

human transcriptome data, and this is one of our study's strengths. Our dataset, however, might contain false-positive genes that matched the ciliary pattern by chance or are related uniquely to a pathology present in PCD. We based our prediction exclusively on PCD patients' cells, what also limits our dataset to that part of a ciliary puzzle altered by this particular disease process.

PCD is a highly heterogeneous disorder, for which nine genes are known to be mutated. Yet, in the majority of

patients, the causative genes remain undiscovered. Our gene set could be considered a list of PCD candidate genes. Because one of the PCD loci is on chromosome 19q (Meeks et al. 2000; Blouin et al. 2000) and we found three genes (*CACNG6*, *C19ORF51* and *AURKC*) from our cluster that map to this region, they might well be considered disease candidate genes. A 1-bp deletion in the aurora kinase c gene (*AURKC*) has been associated with male infertility characterized by large-headed,

multi-flagellar, polyploid spermatozoa, what makes this gene an interesting candidate for further investigation in PCD patients (Dieterich et al. 2007).

The expression of cluster A genes was not highly correlated in the control samples. We observed a disease-specific pattern of expression for 372 functionally related genes. This phenomenon occurs in response to the mutation of a single gene, but most probably not the same gene in each patient studied, since PCD is a genetically heterogeneous disease. It indicates that information from the cytoplasm of the PCD epithelium, where cilia are misassembled (Fliegauf et al. 2005), is transferred to the nucleus and transformed on the genomic level into a regulatory signal altering expression of ciliary genes. Our results show how a monogenic disease affecting multi-protein complexes can be used to study more complex genomic networks.

In conclusion, we identified a group of 208 new cilia-related genes. This list of genes provides candidate genes for PCD and other ciliopathies.

**Acknowledgments** The cooperation of all the Polish families who participated in this study was invaluable. We thank Jackie Senior for editing the manuscript, Ewa Rutkiewicz for laboratory assistance, and the staff of the Genomics Facility, UMCG, for scientific and technical work. MG was supported by the International PhD Program of Utrecht University, The Netherlands coordinated by the International Institute of Molecular and Cell Biology, Warsaw, Poland.

## References

- Avidor-Reiss T, Maer AM, Koundakjian E, Polyanovsky A, Keil T et al (2004) Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* 117(4):527–539
- Badano JL, Mitsuma N, Beales PL, Katsanis N (2006) The ciliopathies: an emerging class of human genetic disorders. *Annu Rev Genomics Hum Genet* 7:125–148
- Bartoloni L, Blouin J-L, Pan Y, Gehrig C, Maiti AK, Scamuffa N, Rossier C, Jorissen M, Armengot M, Meeks M, Mitchison HM, Chung EMK, Delozier-Blanchet CD, Craigen WJ, Antonarakis SE (2002) Mutations in the DNAH11 (axonemal heavy chain dynein type 11) gene cause one form of situs inversus totalis and most likely primary ciliary dyskinesia. *Proc Natl Acad Sci* 99:10282–10286
- Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C et al (2005) Functional genomics of the cilium, a sensory organelle. *Curr Biol* 15(10):935–941
- Bloodgood RA (2010) Sensory reception is an attribute of both primary cilia and motile cilia. *J Cell Sci* 123(Pt 4):505–509
- Blouin JL, Meeks M, Radhakrishna U et al (2000) Primary ciliary dyskinesia: a genome-wide linkage analysis reveals extensive locus heterogeneity. *Eur J Hum Genet* 8(2):109–118
- Castleman VH, Romio L, Chodhari R, Hirst RA, de Castro SCP, Parker KA, Ybot-Gonzalez P, Emes RD, Wilson SW, Wallis C, Johnson CA, Herrera RJ et al (2009) Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. *Am J Hum Genet* 84:197–209
- Colecchia F, Kottwitz D, Wagner M, Pfenninger CV, Thiel G, Tamm I, Peterson C, Nuber UA (2009) Tissue-specific regulatory network extractor (TS-REX): a database and software resource for the tissue and cell type-specific investigation of transcription factor-gene networks. *Nucleic Acids Res* 37(11):e82
- Coppe A, Ferrari F, Bisognin A, Danieli GA, Ferrari S, Biccato S, Bortoluzzi S (2009) Motif discovery in promoters of genes colocalized and co-expressed during myeloid cells differentiation. *Nucleic Acids Res* 37(2):533–549
- Dieterich K, Soto Rifo R, Faure AK et al (2007) Homozygous mutation of AURKC yields large-headed polyploid spermatozoa and causes male infertility. *Nat Genet* 39(5):661–665
- Duquesnoy P, Escudier E, Vincensini L, Freshour J, Bridoux AM, Coste A, Deschildre A, de Blic J, Legendre M, Montantin G, Tenreiro H, Vojtek A-M, Loussert C, Clement A, Escalier D, Bastin P, Mitchell DR, Amselem S (2009) Loss-of-function mutations in the human ortholog of *Chlamydomonas reinhardtii* ODA7 disrupt dynein arm assembly and cause primary ciliary dyskinesia. *Am J Hum Genet* 85:890–896
- Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR et al (2005) Analysis of *xbx* genes in *C. elegans*. *Development* 132(8):1923–1934
- El Zein L, Ait-Lounis A, Morlé L, Thomas J, Chhin B, Spassky N, Reith W, Durand B (2009) RFX3 governs growth and beating efficiency of motile cilia in mouse and controls the expression of genes involved in human ciliopathies. *J Cell Sci* 122(Pt 17):3180–3189
- Fliegauf M, Olbrich H, Horvath J, Wildhaber JH, Zariwala MA et al (2005) Mislocalization of DNAH5 and DNAH9 in respiratory cells from patients with primary ciliary dyskinesia. *Am J Respir Crit Care Med* 171(12):1343–1349
- Geremek M, Witt M (2004) Primary ciliary dyskinesia: genes, candidate genes and chromosomal regions. *J Appl Genet* 45(3):347–361
- Gherman A, Davis EE, Katsanis N (2006) The ciliary proteome database: an integrated community resource for the genetic and functional dissection of cilia. *Nat Genet* 38(9):961–962
- Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 9(11):1106–1115
- Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG et al (2007) The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 8(9):R183
- Inglis PN, Boroevich KA, Leroux MR (2006) Piecing together a ciliome. *Trends Genet* 22(9):491–500
- Iwama A, Pan J, Zhang P, Reith W, Mach B, Tenen DG, Sun Z (1999) Dimeric RFX proteins contribute to the activity and lineage specificity of the interleukin-5 receptor alpha promoter through activation and repression domains. *Mol Cell Biol* 19(6):3940–3950
- Karadag B, James AJ, Gultekin E (1999) Nasal and lower airway level of nitric oxide in children with primary ciliary dyskinesia. *Eur Respir J* 13:1402–1405
- Kouwenhoven EN, van Heeringen SJ, Tena JJ, Oti M, Dutilh BE, Alonso ME, de la Calle-Mustienes E, Smeenk L, Rinne T, Parsaulian L, Bolat E, Jurgelenaite R, Huynen MA, Hoischen A, Veltman JA, Brunner HG, Roscioli T, Oates E, Wilson M, Manzanares M, Gómez-Skarmeta JL, Stunnenberg HG, Lohrum M, van Bokhoven H, Zhou H (2010) Genome-wide profiling of p63 DNA-binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus. *PLoS Genet* 19(6(8):e1001065
- Kubo A, Yuba-Kubo A, Tsukita S, Tsukita S, Amagai M (2008) Sentan: a novel specific component of the apical structure of vertebrate motile cilia. *Mol Biol Cell* 19(12):5338–5346

- Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM et al (2004) Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. *Cell* 117(4):541–552
- Loges NT, Olbrich H, Fenske L, Mussaffi H, Horvath J, Fliegauf M, Kuhl H, Baktai G, Peterffy E, Chodhari R, Chung EMK, Rutman A et al (2002) DNAI2 mutations cause primary ciliary dyskinesia with defects in the outer dynein arm. *Am J Hum Genet* 83:547–558
- Loges NT, Olbrich H, Becker-Heck A, Haffner K, Heer A, Reinhard C, Schmidts M, Kispert A, Zariwala MA, Leigh MW, Knowles MR, Zentgraf H, Seithe H, Nurnberg G, Nurnberg P, Reinhardt R, Omran H (2009) Deletions and point mutations of LRRC50 cause primary ciliary dyskinesia due to dynein arm defects. *Am J Hum Genet* 85:883–889
- McClintock TS, Glasser CE, Bose SC, Bergman DA (2008) Tissue expression patterns identify mouse cilia genes. *Physiol Genomics* 32(2):198–206
- Meeks M, Walne A, Spiden S et al (2000) A locus for primary ciliary dyskinesia maps to chromosome 19q. *J Med Genet* 37(4):241–244
- Moore A, Escudier E, Roger G, Tamalet A, Pelosse B et al (2006) RPGR is mutated in patients with a complex X linked phenotype combining primary ciliary dyskinesia and retinitis pigmentosa. *J Med Genet* 43(4):326–333
- Nobutaka H, Yosuke T, Yasushi O (2009) Left–right determination: involvement of molecular motor KIF3, cilia, and nodal flow. *Cold Spring Harb Perspect Biol* 1(1):a000802
- Olbrich H, Haffner K, Kispert A, Volkel A, Volz A, Sasmaz G, Reinhardt R, Hennig S, Lehrach H, Konietzko N, Zariwala M, Noone PG, Knowles M, Mitchison HM, Meeks M, Chung EMK, Hildebrandt F, Sudbrak R, Omran H (2002) Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left–right asymmetry. *Nat Genet* 30:143–144
- Omran H, Kobayashi D, Olbrich H, Tsukahara T, Loges NT, Hagiwara H, Zhang Q, Leblond G, O’Toole E, Hara C, Mizuno H, Kawano H et al (2008) Ktu/PF13 is required for cytoplasmic pre-assembly of axonemal dyneins. *Nature* 456:611–616
- Ostrowski LE, Blackburn K, Radde KM, Moyer MB, Schlatzer DM et al (2002) A proteomic analysis of human cilia: identification of novel components. *Mol Cell Proteomics* 1(6):451–465
- Pazour GJ, Agrin N, Leszyk J, Witman GB (2005a) Proteomic analysis of a eukaryotic cilium. *J Cell Biol* 170(1):103–113
- Pazour GJ, Agrin N, Walker BL, Witman GB (2005b) Identification of predicted human outer dynein arm genes: candidates for primary ciliary dyskinesia genes. *J Med Genet* 43(1):62–73
- Pennarun G, Escudier E, Chapelin C, Bridoux A-M, Cacheux V, Roger G, Clement A, Goossens M, Amselem S, Duriez B (1999) Loss-of-function mutations in a human gene related to *Chlamydomonas reinhardtii* dynein IC78 result in primary ciliary dyskinesia. *Am J Hum Genet* 65:1508–1519
- Regamey N, Hilliard TN, Saglani S, Zhu J, Scallan M, Balfour-Lynn IM, Rosenthal M, Jeffery PK, Alton EW, Bush A, Davies JC (2007) Quality, size, and composition of pediatric endobronchial biopsies in cystic fibrosis. *Chest* 131(6):1710–1717
- Riehle KJ, Campbell JS, McMahan RS, Johnson MM, Beyer RP, Bammler TK, Fausto N (2008) Regulation of liver regeneration and hepatocarcinogenesis by suppressor of cytokine signaling 3. *J Exp Med* 205(1):91–103
- Satir P, Christensen ST (2008) Structure and function of mammalian cilia. *Histochem Cell Biol* 129(6):687–693
- Smith JC, Northey JG, Garg J, Pearlman RE, Siu KW (2005) Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of *Tetrahymena thermophila*. *J Proteome Res* 4(3):909–919
- Stolc V, Samanta MP, Tongprasit W, Marshall WF (2005) Genome-wide transcriptional analysis of flagellar regeneration in *Chlamydomonas reinhardtii* identifies orthologs of ciliary disease genes. *Proc Natl Acad Sci USA* 102(10):3703–3707
- Vadigepalli R, Chakravarthula P, Zak DE, Schwaber JS, Gonye GE (2003) PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. *OMICS* 7(3):235–252
- Valente EM, Silhavy JL, Brancati F, Barrano G, Krishnaswami SR, Castori M, Lancaster MA, Boltshauser E, Boccone L, Al-Gazali L, Fazzi E, Signorini S, Louie CM, Bellacchio E, Bertini E, Dallapiccola B, Gleeson JG (2006) Mutations in CEP290, which encodes a centrosomal protein, cause pleiotropic forms of Joubert syndrome. *Nat Genet* 38(6):623–625
- Wingender E, Dietze P, Karas H, Knüppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 24(1):238–241
- Yu X, Lin J, Zack DJ, Qian J (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res* 34(17):4925–4936