



OPEN

The transcriptional signature associated with human motile cilia

Anirudh Patir, Amy M. Fraser, Mark W. Barnett, Lynn McTeir, Joe Rainger, Megan G. Davey & Tom C. Freeman

Cilia are complex microtubule-based organelles essential to a range of processes associated with embryogenesis and tissue homeostasis. Mutations in components of these organelles or those involved in their assembly may result in a diverse set of diseases collectively known as ciliopathies. Accordingly, many cilia-associated proteins have been described, while those distinguishing cilia subtypes are poorly defined. Here we set out to define genes associated with motile cilia in humans based on their transcriptional signature. To define the signature, we performed network deconvolution of transcriptomics data derived from tissues possessing motile ciliated cell populations. For each tissue, genes coexpressed with the motile cilia-associated transcriptional factor, *FOXJ1*, were identified. The consensus across tissues provided a transcriptional signature of 248 genes. To validate these, we examined the literature, databases (CilDB, CentrosomeDB, CiliaCarta and SysCilia), single cell RNA-Seq data, and the localisation of mRNA and proteins in motile ciliated cells. In the case of six poorly characterised signature genes, we performed new localisation experiments on *ARMC3*, *EFCAB6*, *FAM183A*, *MYCBPAP*, *RIBC2* and *VWA3A*. In summary, we report a set of motile cilia-associated genes that helps shape our understanding of these complex cellular organelles.

Cilia and flagella are related organelles that facilitate an array of cellular functions. In eukaryotes, the core structural components of cilia includes: the *axoneme*, a microtubular protrusion from the cell surface composed of an array of microtubules; a *centrosomal core*, comprised of a mother (basal body) and daughter centriole^{1,2} anchored to the base of the axoneme, and the centriole-associated distal and sub-distal appendages³. Generally, cilia can be subdivided into non-motile primary cilia, in which nine microtubules constitute the axoneme (9 + 0) and motile cilia, characterised by an additional central pair of microtubules (9 + 2)⁴⁻⁶. Primary cilia are found on most cell types, where their principal role is as a sensor of the cell's microenvironment⁷. In contrast, motile cilia are restricted to specific cell populations. Flagellum function as a single large 'propeller' and in eukaryotes are found exclusively on spermatozoa where they drive cell motility. Other motile cilia are found in large numbers on the apical surface of certain types of epithelial cells, where their co-ordinated beating displaces the luminal contents over the epithelial surface, e.g. the clearance of mucus in the respiratory tract. Whilst there are a set of core proteins common to all cilia, there are also structural and regulatory elements unique to motile cilia which underpin their distinct functional activity^{8,9}.

Motile cilia play a vital role in human development and homeostasis, and there is a growing list of ciliopathies (cilia-related diseases) associated with mutations of ciliary assembly proteins and protein components of these organelles. These include defects in left-right patterning during embryogenesis¹⁰, infertility¹¹, asthma¹² and hydrocephalus¹³. Perhaps the most notable and well-characterised ciliopathy is primary ciliary dyskinesia, an autosomal recessive disorder which has an estimated prevalence of 1 in 10,000^{5,14}. Causative mutations leading to primary ciliary dyskinesia include those in genes encoding the motile ciliary components of radial spokes (*RSPH1*, *RSPH9* and *RSPH4A*)^{15,16}, dynein arms, specifically the outer dynein arm (*DNAI1*, *DNAI2* and *DNAH11*)¹⁷⁻¹⁹, proteins involved in their assembly (*CCDC103*, *LRRC6* and *ZMYND10*)²⁰⁻²² and the key transcriptional regulator of motile ciliogenesis *FOXJ1*²³. Patients carrying mutations in these genes are often treated for respiratory symptoms, including chronic respiratory infections, due to their inefficient clearance of mucus from the lungs^{24,25}.

There have already been considerable efforts made to characterise the molecular components of cilia. *FOXJ1* and the RFX family of genes have been identified as the key transcription factors which regulate motile ciliogenesis, which in turn, depending on species, have shown to be regulated by the Wnt, Hedgehog and Notch signalling pathways⁹. Moreover, in conjunction with other transcriptional regulators, such as *HNF1B* and *SOX5*, further

The Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, Scotland, EH25 9RG, UK. e-mail: tfreeman@roslin.ed.ac.uk

ciliary diversity is introduced for mechanosensory renal cilia and bronchiolar cilia, respectively^{26,27}. Proteomic profiling studies have sought to define the components of motile cilia by dysregulating such transcriptional regulators and analysing the proteome of isolated cilia preparations using mass spectrometry^{28–32}. Each of these studies has produced a list of cilia-associated proteins and accordingly a number of databases have been established. The most relevant to the study of human cilia include: *CentrosomeDB*, a set of human (and *Drosophila*) genes encoding proteins that are localized in the centrosome, either as centrosome constituents or as centrosome visitors³³; *CilDB* a database dedicated to proteins involved in centrioles, centrosomes, basal bodies, cilia and flagella in eukaryotes³⁴; *SysCilia* a curated list of cilia genes many of which are associated with disease³⁵; and *CiliaCarta* which employs a naive Bayesian classifier to predict cilia candidate genes across a diverse set of datasets³⁶. These resources list between 303 and 3,376 genes and have greatly broadened our understanding of the complexity of cilia while attempting to define the role of these genes in the context of development, ciliogenesis and ciliopathies.

Here we have sought to provide a consensus human motile cilia gene signature conserved across known motile cilia containing tissues and compare it with the relevant databases. We have used a network deconvolution approach to define gene coexpression clusters containing the transcriptional regulator *FOXJ1*³⁷ using transcriptomics data from the Genotype-Tissue Expression (GTEx)³⁸ project for human tissues known to possess motile ciliated cells. In support of these analyses, we have also examined various lines of evidence in order to validate the set of genes identified. These include a comparison with cilia and centrosomal databases mentioned above, studies of their expression profile across motile and primary cilia containing cells and tissues, and a number of new expression studies examining several poorly characterised genes identified by this work, namely *ARMC3*, *EFCAB6*, *FAM183A*, *MYCBPAP*, *RIBC2* and *VWA3A*. Overall the study proposes a set of motile cilia-associated genes that are tightly coexpressed across tissues, including certain but not all cilia-associated centrosomal genes previously identified. The signature genes have been summarised graphically based on their function and/or known localization.

Results

Derivation of the human motile cilia signature. The GTEx RNA-Seq (v7) dataset is the largest transcriptomics data resource for non-pathological human tissues currently available and was used here to derive a human motile cilia gene signature. Data derived from tissues known to contain cell populations possessing motile cilia, i.e. ependymal cells in brain regions likely to adjoin the cerebrospinal fluid-filled ventricular space (n = 863), bronchial epithelia of the lung (n = 427), spermatocytes in testis (n = 259) and tubal epithelial cells in fallopian tube/ endocervix (n = 12) were downloaded; in total this represented tissue RNA-Seq data from 1,561 samples derived from 566 donors (Fig. 1). To identify genes associated with motile cilia, we examined coexpressing genes, i.e. genes which have a similar expression profile across samples from the same tissue. This similarity was measured using the Pearson correlation coefficient and highly correlated genes were used to construct a gene correlation network (GCN). The network comprises of nodes representing genes, and those correlated beyond the selected threshold are connected within the network. Subsequently, the network was subjected to network cluster analysis to define groups of coexpressed genes, and in the case of each tissue GCN, those that clustered with *FOXJ1* (Table S1). *FOXJ1* associated clusters ranged in size from 597 to 6,126 genes. Such variation in the size of the motile cilia cluster across tissues is likely explained by a varying number of samples and the different tissue biology, e.g. the expression landscape of the testis is dominated by transcriptional signal associated with spermatogenesis, making the flagellum-specific gene module difficult to separate from other sperm-associated gene clusters³⁹. 1,517 genes from the different tissue derived signatures overlapped with one another (Fig. 2B), and for the final signature we considered only the 248 genes present in all tissue-derived gene lists. However, we acknowledge that the extended list, i.e. 479 genes found in three of the four motile cilia tissue clusters contains many additional validated cilia genes and therefore is likely to contain many other novel motile cilia-associated genes (Table S1).

Comparison with database genes and their expression profiles across cells/tissue. Enrichment analysis of the 248 gene signature was conducted for GO terms, pathways, gene families, transcription factor binding sites, and human phenotypes (Table S2). Enriched gene families included components of the 'dynein regulatory complex' (q value = 2.2×10^{-15}), 'axonemal, dyneins', (q value = 3.2×10^{-13}) and 'tektins' (q value = 1.4×10^{-7}), with the corresponding enrichment of biological processes such as 'cilium movement' (q value = 3.4×10^{-52}), 'cilium organization' (q value = 5.1×10^{-40}) and 'cilium-dependent cell motility' (q value = 2.7×10^{-26}). Additionally, binding sites for RFX1 and MIF were also found to be enriched for these genes (q value < 10^{-7}). Human disease phenotypes associated with disorders of motile cilia included 'abnormal respiratory motile cilium morphology' (q value = 1.2×10^{-24}), ' situs inversus totalis' (q value = 3.4×10^{-21}), 'bronchiectasis' (q value = 1.5×10^{-13}), and 'male infertility' (q value = 3.7×10^{-13}). Signature genes associated with human ciliopathies were identified from OMIM, a knowledgebase of human genetic disorders⁴⁰. The database listed 33 of the signature genes linked to human ciliopathies, majority being primary ciliary dyskinesia (Table S2). To further investigate the ciliary/centrosomal association of the signature genes we first conducted a literature search on all signature genes. Of the genes identified, the literature supported 133 (54%) as having direct experimental evidence supporting their spatial localization or functional association with cilia (Table S1). A further 87 (35%) genes were found associated with cilia through coexpression analysis but without any direct evidence of their localization within ciliary structures. For 28 genes (11%) no prior association with cilia could be identified.

We then sought to examine the signature's overlap with public databases of cilia/centrosome proteins. Including the signature reported here, a total of 4,333 genes have been implicated previously with cilia and/or centrosomes. These include the *CentrosomeDB*³³, *CilDB*³⁴, *CiliaCarta*³⁶ and *SysCilia* (gold standard)³⁵ (Fig. 2A and Table S3). There were only four genes which were common to all the databases and the derived signature (*DNAAF1*, *FOXJ1*, *KIF24*, and *MAK*). In support of our literature search, the majority of the signature genes (196 genes) overlapped with genes listed in *CilDB*, including well known motile cilia genes such as members of the

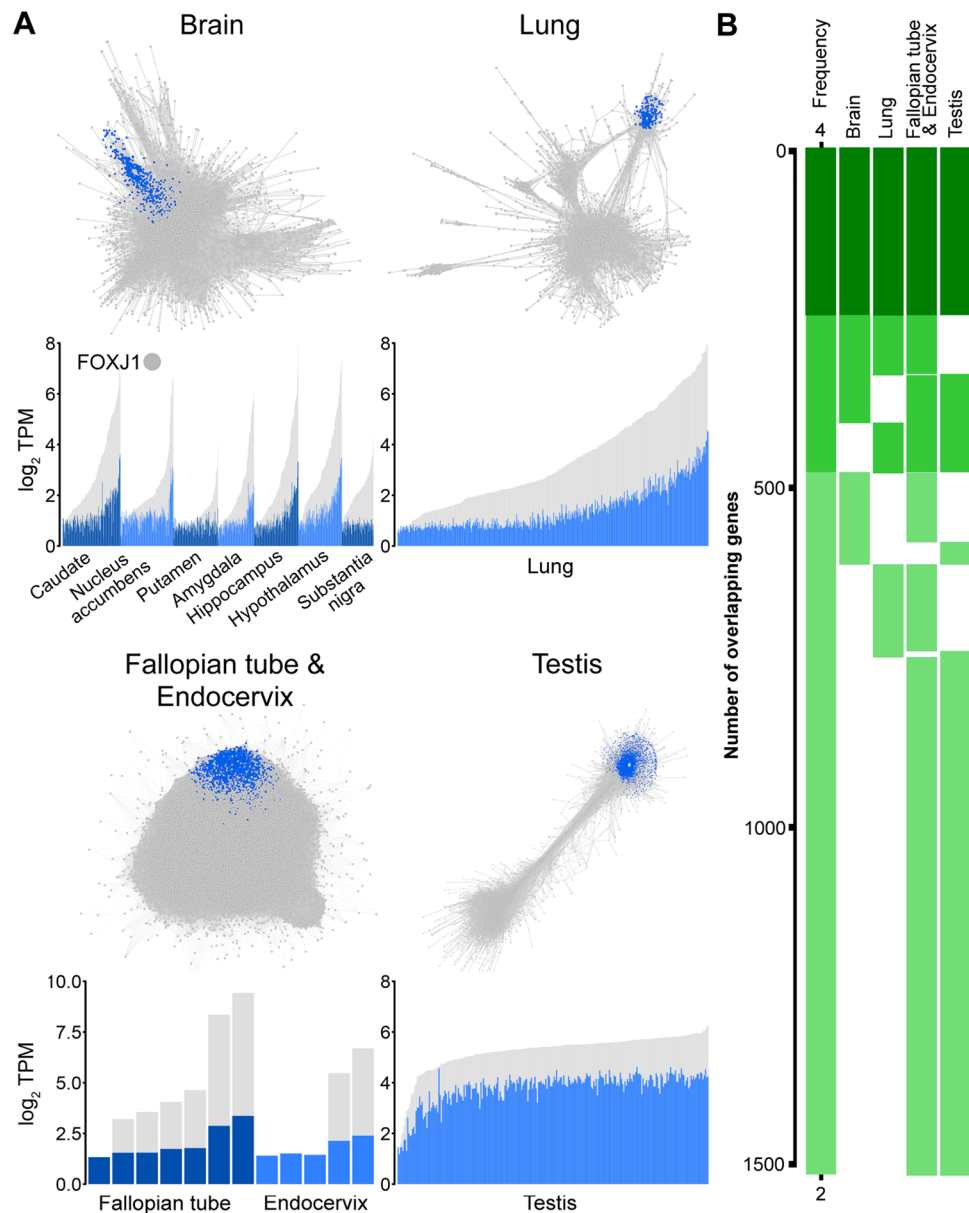


Figure 1. Derivation of the human motile cilia transcriptomic signature. **(A)** Gene coexpression networks for each of the four selected tissue types generated from RNA-Seq data from the GTEx project. Each node represents a gene, with genes sharing similar expression profiles being highly connected to one another; genes that co-clustered with *FOXJ1* have been highlighted in blue in the network. The corresponding expression profile is shown for *FOXJ1*-containing gene clusters (blue) with the average expression of its genes and individually shown for *FOXJ1* (grey). Here, each column represents individual samples of the respective tissue. **(B)** 1,517 genes were found to be present in more than one of the four *FOXJ1*-containing gene clusters. Genes present in all four tissue clusters are depicted in dark green and represent the final 248 tissue-derived motile cilia gene signature.

dynein regulatory complex (*DRC1*, *TCTE1* and *IQCD*), axonemal dynein (*DNAH2*, *DNALI1* and *DNAI1*) and tektin gene family (*TEKT1*, *TEKT2*, and *TEKT4*), whilst also included genes with poor evidence supporting an association with cilia, e.g. *MYCBPAP*, *ARMC3* and *EFCAB6*. Relative to the databases, 52 genes were found to be unique to the current study and included genes not associated with human motile cilia previously, e.g. *FAM183A* and *VWA3A*. By contrast, 84 genes recorded by all database resources were absent from the derived signature. Upon inspection, these largely represented genes associated with the cell cycle⁴¹ and ciliary assembly and maintenance, e.g. members of the centrin family (*CETN1*, *CETN2* and *CETN3*), BBSome complex members (*BBS1*, *BBS4*, *BBS5* and *BBS7*) and IFT genes (*IFT20*, *IFT74* and *IFT81*)⁴².

As a further analysis, we examined the global expression patterns of all signature genes and those recorded in databases for their expression across all 51 tissue types in the GTEx resource. GCN analysis was again used to visualise and explore the expression profile of signature and database genes across human tissues (Fig. 2B–D).

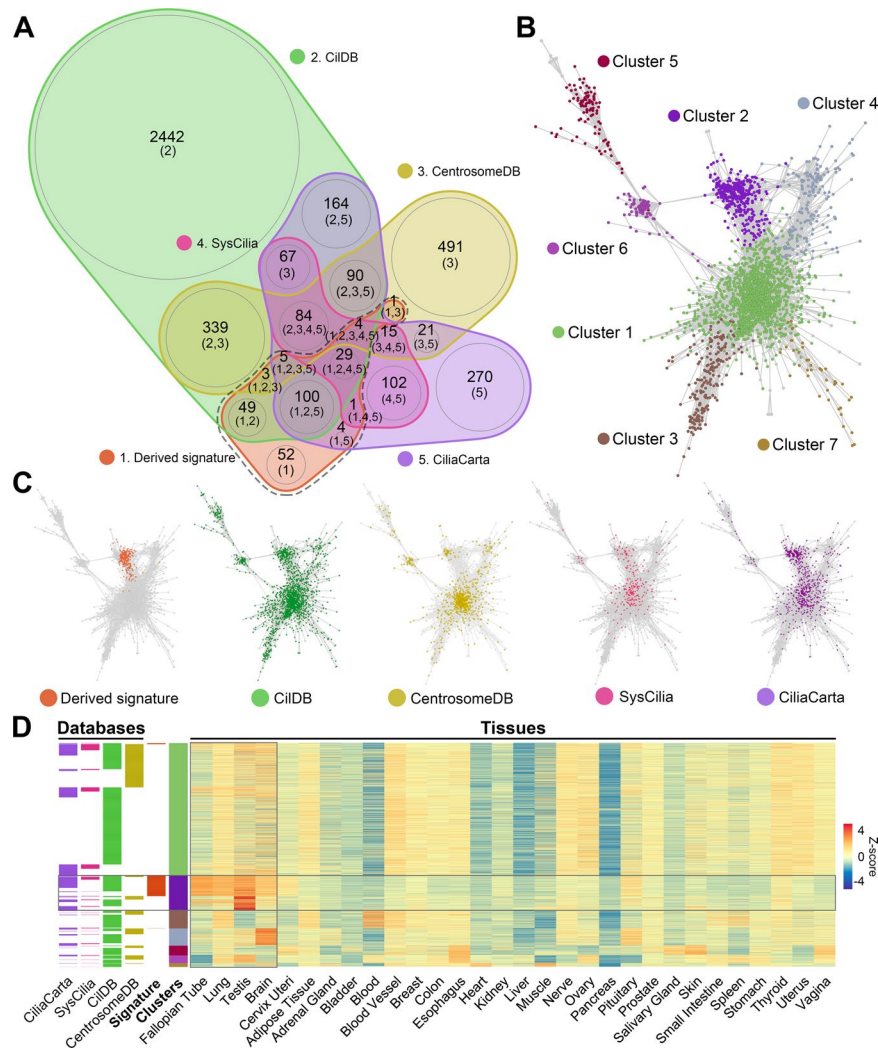


Figure 2. Comparison of signature and database genes and their expression across tissues. **(A)** Venn diagram of overlapping genes between the derived signature and databases using nVenn v0.2.1 (<https://cran.r-project.org/web/packages/nVennR/index.html>)⁷⁷. Here the overlap between gene lists can be observed through the overlap of their respective segments and the numbers shown in brackets, are indicative of the gene list number. The derived signature is encircled by a dashed line. **(B)** Gene coexpression network constructed in Graphia v1 (<https://kajeka.com/graphia/>) for all these genes as defined by their expression in the context of GTEx tissues. Nodes are coloured by cluster analysis of this network. **(C)** Genes from the signature and included in each of the databases have been highlighted individually on the network. **(D)** Heatmap of the signature and database genes across GTEx tissues ordered based on the gene clustering from **(B)**.

Cluster analysis was used to broadly group genes together based on their underlying expression pattern (Fig. 2B). Highlighting the genes from each database showed them in each case to be distributed across the network. In contrast to the distribution of signature genes which were far more localized. This is indicative of their tight coexpression across all tissues (Fig. 2C), attributed by their relatively high expression in tissues known to have motile ciliated cells (Fig. 2D). Conversely, genes for each of the databases were scattered throughout the graph suggesting that they had very different expression profiles, ranging from a broad expression across all tissues as represented by cluster 1, to being highly expressed in certain tissues such as blood (cluster 3) or brain (cluster 4). Additionally, cluster 3 included many immune genes, e.g. associated with MHC class 1 and 2, TLR receptors and TNF family of genes (Table S3). As a more direct comparator, the expression of signature genes was examined in single cell RNA-Seq data derived from the mouse brain and lung (Fig. 3). Here motile ciliated ependymal and bronchial cells, respectively, showed a significantly higher average expression of signature genes (q value < 0.001) when compared to other cell types, again supporting their specific association with motile cilia possessing cells.

The localisation of candidate proteins in motile ciliated tissue. In order to provide an additional level of validation for the 248 signature genes, we examined the immunohistochemistry (IHC) data in the Human Protein Atlas (HPA) resource⁴³ in tissues containing motile ciliated cells (Fig. 4, Table S1). Based on our own criteria (independent of the confidence scores set by HPA), genes were placed into three groups: high confidence

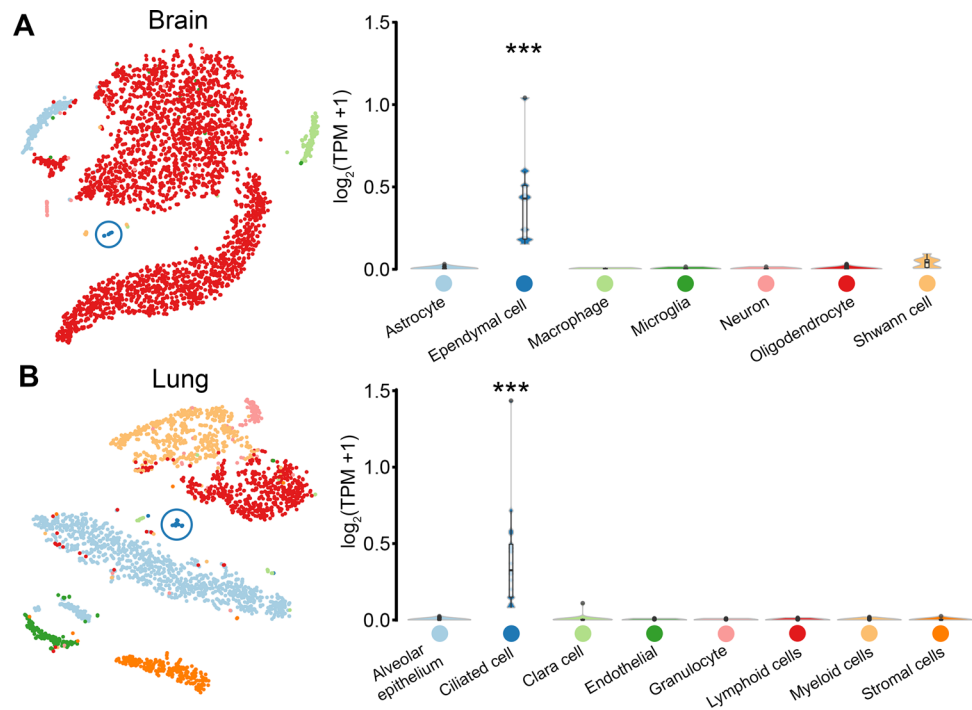


Figure 3. Expression of signature genes in cells from the brain and lung. Average expression of signature genes across cells of the (A) brain and (B) lung taken from the mouse cell atlas dataset⁷¹.

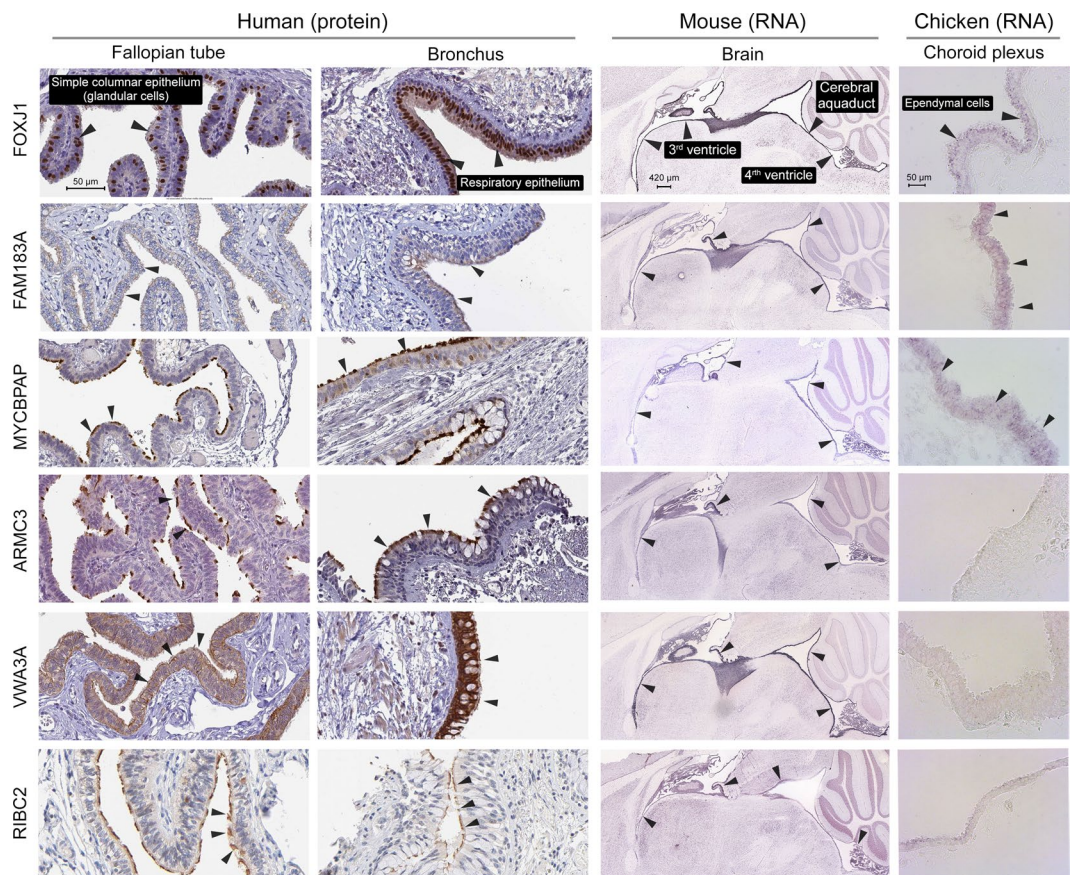


Figure 4. IHC and ISH staining from signature genes across species and tissues. IHC and ISH staining of tissue sections from the HPA and Allen brain atlas of the mouse brain for encoded proteins and transcribed RNA of signature genes, respectively. The final column consists of new ISH staining performed in choroid plexus sections from chicken embryos (stage 35).

components are upregulated⁴¹, and primary cilia, which contain many of the same proteins, are present in most cell types.

Here we have attempted to harness the power of GCN's and employ the principle of 'guilt-by-association' to identify genes specifically associated with motile cilia. This is based on the fact that genes specifically associated with a given cell type or biological process frequently vary in expression with their relative abundance or activity within a sample and consequently across a large sample set the expression of these genes is tightly correlated. This approach has been used previously to identify genes associated with specific cell populations and processes, from tissue and cell-level transcriptomics data^{47–50}. Here we analysed the brain, lung and the female reproductive tract, all of which contain populations of multi-ciliated cells which function to move luminal contents (cerebrospinal fluid, mucus) over the epithelial surface. In addition, we examined the testis, where cilia proteins are associated with the flagellum of sperm, a fundamentally different type of motile cilia but comprised of many of the same molecular components. In the current study, we first identified genes from each of the selected tissues which co-clustered with *FOXJ1*, a key transcriptional regulator of the motile ciliogenic program^{37,51,52}. In the case of the brain and lung, a clear transcriptional module associated with multiciliated epithelial cells was defined due to the marked variation in the abundance of these cell populations across the samples. For the female reproductive tract and testis, however, such modules were harder to define accurately, as there were either only a few samples available and with flagellum-related genes being strongly associated with genes involved in spermatogenesis³⁹, respectively. To circumvent these limitations and filter out any cell type-specific genes, we compared the gene clusters from each tissue to arrive at a consensus signature of 248 genes. It should be noted, however, that the list of genes associated with three of the tissue clusters (an additional 231 genes) also contained many other known cilia proteins and therefore by inference genes encoding other uncharacterised cilia components (Table S1).

Validation of the gene signature included enrichment analyses, annotation based on a literature review and cilia-associated databases, and exploration of other resources describing the cellular expression of genes and proteins, confirmed the majority to be known components of motile cilia or associated regulatory systems (summarised in Fig. 5). The binding site for the transcriptional factor RFX1, a member of the RFX gene family⁵³, was enriched. This gene has shown to be involved in development, based on a mouse knockout model and regulates the basal body-associated protein ALMS1, defects in which cause Alström syndrome ciliopathy^{54,55}. Furthermore, another member, RFX2 has been proposed to work in conjunction with FOXJ1 to regulate cilia gene expression, including RIBC2 also identified and examined in this study^{56,57}. However, of the RFX family only RFX3 was found in three of the tissue-derived motile cilia signatures, making it a likely candidate in regulating motile-ciliogenesis, an observation supported by a recent study⁵⁸. In addition, the transcriptional binding site for MIF was also enriched in signature gene promoters. MIF is known to affect cell motility through the regulation of microtubule formation^{59,60}.

The signature genes were also cross-referenced with the four cilia/centrosome gene databases; CentrosomeDB³³, CilDB³⁴, CiliaCarta³⁶ and SysCilia (gold standard)³⁵ databases. The majority (79%) of genes in the signature were corroborated by one or more of the databases. Notably, many well-known primary cilia genes involved in ciliary assembly and signalling and listed by the databases were absent from the signature. This included members of the BBSome complex, IFT chain^{61,62} and many associated with the centrosome. As known components of primary cilia, these genes are regulated through the stages of cell cycle and are ubiquitously expressed across cell types, and would therefore be expected to have a different expression profile in the tissues examined relative to the genes coexpressing with *FOXJ1*. To explore the expression of signature and database genes across tissues, GCN analysis was used for all the 51 tissues from the GTEx project. Interestingly, although the signature was derived from separate analyses of individual tissues, in general their coexpression was highly conserved across the 51 tissue types, being highly expressed in motile ciliated tissues relative to others. Interestingly, these also included a number of centrosomal genes indicative of a specialised centrosomal system for motile cilia assembly and function. Genes listed by the various databases coexpressing with those of the signature included known motile cilia components like dyneins (*DNAH3*, *DNAH7*, and *DNAH8*) and members of the α -tubulin gene family (*TUBA3D*, *TUBA3E*, and *TUBA3C*)^{63,64}. In contrast, analysis of databases showed genes within a given database to be distributed across the GCN, and exhibit little evidence of co-expression, suggestive of representing different biology across tissues. Closer inspection showed some to be immune-related genes, e.g. TLR and MHC genes, and their presence in the databases is likely an artefact of the approaches used to define them⁶⁵. As a direct validation of their specificity of expression, single-cell transcriptomics data derived from the mouse brain and lung showed the signature genes to be highly and specifically expressed in ependymal and ciliated epithelial cells, respectively, of these tissues. Hence, the signature as a whole could help in identifying motile cilia biology in bulk and single cell transcriptomics data due to its conserved nature across several cells and tissues. Furthermore, in representing a homeostatic motile cilia, the signature would be a reference to motile cilia in disease models.

The HPA resource was used to further verify the validity of signature genes based on IHC analysis. Genes for which there was data were ranked as being of either high or medium confidence based on their expression pattern matching that expected for proteins associated with motile cilia. Our analysis of the HPA data showed it to validate the majority of signature genes; 48% were scored as high confidence genes and 20% as medium confidence, based on the criteria outlined in the methods. Nothing could be concluded for the 32% genes for which no data was available or the data was of very poor quality. Remarkably, none of the protein localization data directly contradicted an association of any gene with motile cilia. We then set out to further investigate six genes, *EFCAB6* having no HPA data and five high confidence genes based on their HPA IHC results but with poor evidence of their association with cilia in humans based on literature: *ARMC3*, *FAM183A*, *MYCBPAP*, *RIBC2* and *VWA3A*. ISH experiments performed on the chicken choroid plexus, showed *FAM183A*, *EFCAB6* and *MYCBPAP* to be expressed by motile ciliated cells lining this tissue, although this was not apparent for *RIBC2*, *VWA3A* and *ARMC3*. Apart from being a possible false negative, this discrepancy could be indicative of the diversity of

ciliary components in eukaryotes⁶⁶. A recent study using evolutionary proteomics has predicted *MYCBPAP* to be associated with cilia, and ISH of the *FAM183A* orthologue has shown to positively stain motile ciliated tissue in *Xenopus laevis*^{67,68}. In support of our observations, ISH data from the Allen brain atlas for the selected genes showed positive staining of ciliated cells lining the ventricles of the mouse brain. Finally, we have summarized our findings graphically based on database and literature mining of the known associations of signature genes with cilia structures and function, as well as human ciliopathies. Clearly, many of the known components of the motile cilia machinery have been identified by this study, and many others have evidence supporting their association but not with specific components of the organelle. The curation of the list clearly highlights the many potentially novel cilia genes/proteins identified by this work and those which can be further examined.

In summary, we have used coexpression analyses to identify a set of 248 genes highly associated with the presence of motile ciliated cells within human tissue. Significant efforts were then made to validate the genes identified based on further coexpression analyses, extensive searches of the literature, online resources of information on the cellular and tissue expression data for genes and proteins, as well as public databases of cilia related genes across different species. Along with a graphical description of signature genes within cilia, the signature highlights similar genes from cilia and centrosome databases, helping in the categorization of known cilia genes. In the case of a number of poorly described genes we identified i.e. *ARMC3*, *EFCAB6*, *FAM183A*, *MYCBPAP*, *RIBC2* and *VWA3A*, we have been able to provide new evidence supporting their association with motile cilia. Such analyses serve to extend and refine the list of genes/proteins specifically associated with motile cilia, allowing more targeted analyses of their localisation and functional role within these complex and important organelles.

Material and methods

Data pre-processing, signature derivation. Pre-normalized RNA-Seq data from the GTEx project³⁸ was downloaded (version 7) and log-transformed. Data for tissues known to possess motile ciliated cells were sub-sampled. These included samples taken from seven regions of the brain (n = 863), lung (n = 427), testis (n = 259), fallopian tube (n = 7) and endocervix (n = 5). Due to the small number of samples of fallopian tube and endocervix, data from these tissues were combined. As such, the relative content of motile cilia containing cells varied considerably across samples, with the expression of genes specifically associated with these structures varying accordingly. Motile cilia-associated genes were identified for each individual tissue by GCN analysis. In order to generate a GCN, a gene-to-gene Pearson correlation matrix was calculated between all genes using the network analysis software, Graphia v1 (Kajeka Ltd., Edinburgh, UK, <https://kajeka.com/graphia/>). A threshold of $r \geq 0.8$ was then applied such that only genes correlated to others above this threshold were connected by an edge. In each case, a structured GCN was generated with modules of coexpressed genes forming highly connected cliques within the network. These were defined as clusters using the Markov clustering algorithm (MCL)⁶⁹, using an inflation value $MCLi = 2.2$ (which defines the granularity of clustering). Putative motile cilia-associated genes were defined as those present in the same cluster as *FOXJ1*. Accordingly, four gene clusters were obtained, one for each tissue type. This approach has been adopted previously to identify co-regulated genes with a related function or association with a given cell type^{48–50}. From the four tissue-derived signatures, those genes common to all four signatures were considered for the final human motile cilia signature. Evidence for an association with cilia was explored through literature mining and enrichment analysis, conducted using ToppGene⁷⁰. Disease association for genes was found through the OMIM database⁴⁰.

Functional annotation of motile cilia signature genes and comparison with databases. Evidence for an association of the 248 motile cilia signature genes with cilia was explored through literature mining and enrichment analysis, conducted using ToppGene⁷⁰. Signature genes were then compared to genes listed in the databases of cilia and centrosomal components, i.e. CentrosomeDB³³, CilDB³⁴, CiliaCarta³⁶ and SysCilia (gold standard)³⁵ were collated based on their Ensembl gene IDs and compared to the signature list derived here (Table S3). The expression profile of this combined list was examined across 51 tissues (excluding samples derived from pooled cells) of the GTEx dataset (n = 11,215, donors = 713). A GCN was then generated using these genes only, again using a correlation threshold of $r \geq 0.8$ and the resultant graph was clustered using a low inflation value ($MCLi = 1.2$) so as to provide a coarse grain segmentation of the graph comprising of 17 clusters.

To explore the expression of the signature genes at a cellular level, single cell transcriptomics data from the mouse brain and lung were taken from the Mouse Cell Atlas⁷¹ and analysed. These tissues were selected as they include populations of motile ciliated cells. The batch corrected expression matrices based on unique molecular identifiers were downloaded from the mouse cell atlas database (https://figshare.com/articles/MCA_DGE_Data/5435866). This included batch 1 of the brain (n = 3285 cells) and lung (n = 2501 cells) cell data. Additionally, for the latter, 11 cells annotated as “dividing cells” were excluded, as it was unclear which cell types these referred to. Corresponding mouse orthologues for signature genes were identified based on their Ensembl gene ID using BioMart⁷². The average expression of signature genes was then tested for significance in the ciliated cell populations (ependymal cells of the brain and ciliated epithelial cells of the lung), versus all other cell types as defined in the mouse cell atlas. The non-parametric Wilcoxon signed-rank test was adopted for these comparisons.

Immunohistochemistry and RNA *in situ* hybridisation. The tissue distribution of mRNA and proteins for all signature genes were investigated using publicly available resources. IHC staining of human tissue sections from the bronchus and fallopian tube were examined in the HPA⁴³. In both tissues, positive staining of the ciliated epithelial cells lining the tissue was considered as validity evidence.

The expression of a number of novel genes were further examined in the choroid plexus of chicken embryos (stage 35, day 9) by ISH. Chicken embryos were obtained from wild type Isa Brown fertile chicken eggs and were incubated for 9 days and sacrificed as per approval by The Roslin Institute, University of Edinburgh under

supervision and approval of a Named Animal Care Welfare Officer in accordance with the regulations prescribed for animals under the Animals (Scientific Procedures) Act 1986, UK. Although not subject to this act, fertilised eggs were monitored throughout incubation to ensure high welfare standards. Researchers involved in this project have the necessary UK Home Office Project and Personal licenses for undertaking experimental protocols involving the use of animals in this project. For selected genes, clones which covered the majority of exons near the centre of the gene were preferentially selected (*ARMC3*: ChEST208k22, *EFCAB6*: ChEST912jB, *FAM183A*: ChEST261m5, *FOXJ1*, *MYCBPAP*: ChEST864g6, and *RIBC2*: ChEST770c15) using the UCSC Genome Browser⁷³ and where available obtained (Source BioSciences, UK)⁷⁴. Fertilised chicken eggs were incubated for nine days at which point the embryos were sacrificed, the choroid plexus dissected and tissues fixed overnight in 4% paraformaldehyde (PFA) at 4 °C. Samples were then rinsed in PBS and equilibrated overnight in 15% sucrose/PBS before embedding in sucrose-gelatin (15%:7.5%) and snap frozen in isopentane at −70 °C. Cryostat sections (10 µm) were cut and stored overnight at −20 °C. Sections were then rinsed in PBS and fixed overnight in 4% PFA. After successive rinses with PBS, the tissue was permeabilized by incubation in proteinase-K (20 ng/ml K-03115836001 Roche) for 10 min at room temperature. Sections were treated consecutively with 4% PFA, acetic anhydride solution (0.25% acetic anhydride and 1.3% triethanolamine) with intermittent washing. Finally, 5 nM probe in hybridisation buffer (50% formamide, 5xSSC pH 4.5, 0.05 µg/ml yeast RNA, 0.05 µg/ml heparin, and 1% SDS) was applied to the slides. Following an overnight hybridization with probes at 65 °C, sections went through a series of post-hybridization washes and then maleic acid buffer-tween (0.15 M NaCl, 0.1 M maleic acid, 0.18 M NaOH and 0.02% tween). After blocking (20% heat-inactivated FBS/KTBT) for 1 h, sections were incubated overnight with 1:1000 anti-digoxigenin-alkaline phosphatase (11093274910 Roche, 1:1000) at 4 °C. Following a final series of washing with maleic acid buffer-tween, sections were incubated with staining solution (3.5 µl/ml nitro blue tetrazolium, N-6876 Sigma; 3.5 µl/ml BCIP, B-8503). After staining for 1–2 h (depending on the probe), the reaction was stopped by several rinses in PBS. ISH staining of brain sections for these genes was also examined in data from the mouse Allen brain atlas⁷⁵, where the choroid plexus and ventricular system were present (which is lined with motile ciliated ependymal cells).

Data availability

The tissue and single cell RNA-Seq data that support the findings of this study are available in the Genotype-Tissue Expression portal³⁸ (doi: 10.1038/ng.2653., <https://gtexportal.org/home/>) and from the Mouse Cell Atlas⁷¹ figshare repository (doi: 10.1016/j.cell.2018.02.001, https://figshare.com/articles/MCA_DGE_Data/5435866) respectively.

The cilia associated gene lists used in this study were publically available and taken from the CilDB³⁴ website (<http://cildb.cgm.cnrs-gif.fr/>), CentrosomeDB³³ website (doi: 10.1093/nar/gkt1126, <http://centrosome.cnb.csic.es/>), CiliaCarta article³⁶ (Table S3, doi: 10.1371/journal.pone.0216705), and Syscilia³⁵ website (doi: 10.1186/2046-2530-2-7, <http://www.syscilia.org/goldstandard.shtml>).

Immunohistochemistry of human tissues that support the findings of this study were publically available in the Human Protein Atlas⁷⁶ website (doi: 10.1126/science.1260419., <https://www.proteinatlas.org/>). *In situ* hybridization images of the mouse brain were publically available in the Allen Brain Atlas website for mouse⁷⁵ (doi: 10.1038/nature05453, <https://mouse.brain-map.org/>). *In situ* hybridization experiments conducted on the chicken choroid plexus have been made available within the article and are also available from the corresponding author upon reasonable request.

Received: 17 December 2019; Accepted: 15 April 2020;

Published online: 02 July 2020

References

- Satir, P. & Christensen, S. T. Overview of Structure and Function of Mammalian Cilia. *Annu Rev Physiol* **69**, 377–400, <https://doi.org/10.1146/annurev.physiol.69.040705.141236> (2007).
- Mizuno, N., Taschner, M., Engel, B. D. & Lorentzen, E. Structural Studies of Ciliary Components. *J Mol Biol* **422**, 163–180, <https://doi.org/10.1016/j.jmb.2012.05.040> (2012).
- Uzbekov, R. & Alieva, I. Who are you, subdistal appendages of centriole? *Open biology* **8**, <https://doi.org/10.1098/rsob.180062> (2018).
- Reiter, J. F. & Leroux, M. R. Genes and molecular pathways underpinning ciliopathies. *Nature Reviews Molecular Cell Biology* **18**, 533, <https://doi.org/10.1038/nrm.2017.60> (2017).
- Mitchison, H. M. & Valente, E. M. Motile and non-motile cilia in human pathology: from function to phenotypes. *J Pathol* **241**, 294–309, <https://doi.org/10.1002/path.4843> (2017).
- Hoyer-Fender, S. in *Cilia and Nervous System Development and Function* (eds. Kerry L., Tucker & Tamara, Caspary) 1–53 (Springer Netherlands, 2013).
- Anvarian, Z., Mykytyn, K., Mukhopadhyay, S., Pedersen, L. B. & Christensen, S. T. Cellular signalling by primary cilia in development, organ function and disease. *Nature Reviews. Nephrology* **15**, 199–219, <https://doi.org/10.1038/s41581-019-0116-9> (2019).
- Heydeck, W., Fievet, L., Davis, E. E. & Katsanis, N. The complexity of the cilium: spatiotemporal diversity of an ancient organelle. *Current opinion in cell biology* **55**, 139–149, <https://doi.org/10.1016/j.ceb.2018.08.001> (2018).
- Choksi, S. P., Lauter, G., Swoboda, P. & Roy, S. Switching on cilia: transcriptional networks regulating ciliogenesis. *Development* **141**, 1427–1441, <https://doi.org/10.1242/dev.074666> (2014).
- Bisgrove, B. W. & Yost, H. J. The roles of cilia in developmental disorders and disease. *Development* **133**, 4131, <https://doi.org/10.1242/dev.02595> (2006).
- Lyons, R. A., Saridogan, E. & Djahanbakhch, O. The reproductive significance of human Fallopian tube cilia. *Human reproduction update* **12**, 363–372, <https://doi.org/10.1093/humupd/dml012> (2006).
- Tilley, A. E., Walters, M. S., Shaykhiyev, R. & Crystal, R. G. Cilia dysfunction in lung disease. *Annu Rev Physiol* **77**, 379–406, <https://doi.org/10.1146/annurev-physiol-021014-071931> (2015).
- Narita, K. & Takeda, S. Cilia in the choroid plexus: their roles in hydrocephalus and beyond. *Front Cell Neurosci* **9**, 39, <https://doi.org/10.3389/fncel.2015.00039> (2015).

14. Praveen, K., Davis, E. E. & Katsanis, N. Unique among ciliopathies: primary ciliary dyskinesia, a motile cilia disorder. *F1000Prime Rep* **7**, 36–36, <https://doi.org/10.12703/P7-36> (2015).
15. Castleman, V. H. *et al.* 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, <https://doi.org/10.1016/j.ajhg.2009.01.011> (2009).
16. Kott, E. *et al.* Loss-of-Function Mutations in RSPH1 Cause Primary Ciliary Dyskinesia with Central-Complex and Radial-Spoke Defects. *Am J Hum Genet* **93**, 561–570, <https://doi.org/10.1016/j.ajhg.2013.07.013> (2013).
17. Guichard, C. *et al.* Axonemal Dynein Intermediate-Chain Gene (DNAI1) Mutations Result in Situs Inversus and Primary Ciliary Dyskinesia (Kartagener Syndrome). *Am J Hum Genet* **68**, 1030–1035, <https://doi.org/10.1086/319511> (2001).
18. Loges, N. T. *et al.* DNAI2 Mutations Cause Primary Ciliary Dyskinesia with Defects in the Outer Dynein Arm. *Am J Hum Genet* **83**, 547–558, <https://doi.org/10.1016/j.ajhg.2008.10.001> (2008).
19. Bartoloni, L. *et al.* Mutations in the DNAH11 (axonemal heavy chain dynein type 11) gene cause one form of situs inversus totalis and most likely primary ciliary dyskinesia. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10282–10286, <https://doi.org/10.1073/pnas.152337699> (2002).
20. Kott, E. *et al.* Loss-of-Function Mutations in LRRC6, a Gene Essential for Proper Axonemal Assembly of Inner and Outer Dynein Arms, Cause Primary Ciliary Dyskinesia. *Am J Hum Genet* **91**, 958–964, <https://doi.org/10.1016/j.ajhg.2012.10.003> (2012).
21. Zariwala, M. A. *et al.* ZMYND10 Is Mutated in Primary Ciliary Dyskinesia and Interacts with LRRC6. *Am J Hum Genet* **93**, 336–345, <https://doi.org/10.1016/j.ajhg.2013.06.007> (2013).
22. Panizzi, J. R. *et al.* CCDC103 mutations cause primary ciliary dyskinesia by disrupting assembly of ciliary dynein arms. *Nat Genet* **44**, 714, <https://doi.org/10.1038/ng.2277> (2012).
23. Wallmeier, J. *et al.* De Novo Mutations in FOXJ1 Result in a Motile Ciliopathy with Hydrocephalus and Randomization of Left/Right Body Asymmetry. *The American Journal of Human Genetics* **105**, 1030–1039, <https://doi.org/10.1016/j.ajhg.2019.09.022> (2019).
24. Mirra, V., Werner, C. & Santamaria, F. Primary Ciliary Dyskinesia: An Update on Clinical Aspects, Genetics, Diagnosis, and Future Treatment Strategies. *Frontiers in pediatrics* **5**, 135, <https://doi.org/10.3389/fped.2017.00135> (2017).
25. Leigh, M. W. *et al.* Clinical and genetic aspects of primary ciliary dyskinesia/Kartagener syndrome. *Genet Med* **11**, 473, <https://doi.org/10.1097/GIM.0b013e3181a53562> (2009).
26. Gresh, L. *et al.* A transcriptional network in polycystic kidney disease. *EMBO J* **23**, 1657, <https://doi.org/10.1038/sj.emboj.7600160> (2004).
27. Kiselak, E. A. *et al.* Transcriptional Regulation of an Axonemal Central Apparatus Gene, Sperm-associated Antigen 6, by a SRY-related High Mobility Group Transcription Factor, S-SOX5. *J Biol Chem* **285**, 30496–30505, <https://doi.org/10.1074/jbc.M110.121590> (2010).
28. Choksi, S. P., Babu, D., Lau, D., Yu, X. & Roy, S. Systematic discovery of novel ciliary genes through functional genomics in the zebrafish. *Development* **141**, 3410–3419, <https://doi.org/10.1242/dev.108209> (2014).
29. Blackburn, K., Bustamante-Marin, X., Yin, W., Goshe, M. B. & Ostrowski, L. E. Quantitative Proteomic Analysis of Human Airway Cilia Identifies Previously Uncharacterized Proteins of High Abundance. *Journal of proteome research* **16**, 1579–1592, <https://doi.org/10.1021/acs.jproteome.6b00972> (2017).
30. Ostrowski, L. E. *et al.* A proteomic analysis of human cilia: identification of novel components. *Molecular & cellular proteomics: MCP* **1**, 451–465 (2002).
31. El Zein, L. *et al.* 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**, 3180–3189, <https://doi.org/10.1242/jcs.048348> (2009).
32. Campbell, E. P., Quigley, I. K. & Kintner, C. Foxn4 promotes gene expression required for the formation of multiple motile cilia. *Development* **143**, 4654–4664, <https://doi.org/10.1242/dev.143859> (2016).
33. Nogales-Cadenas, R., Abascal, F., Díez-Pérez, J., Carazo, J. M. & Pascual-Montano, A. CentrosomeDB: a human centrosomal proteins database. *Nucleic Acids Res* **37**, D175–D180, <https://doi.org/10.1093/nar/gkn815> (2009).
34. Arnaiz, O. *et al.* Cildb: a knowledgebase for centrosomes and cilia. *Database* **2009**, bap022–bap022, <https://doi.org/10.1093/database/bap022> (2009).
35. van Dam, T. J., Whewey, G., Slaats, G. G., Huynen, M. A. & Giles, R. H. The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. *Cilia* **2**, 7, <https://doi.org/10.1186/2046-2530-2-7> (2013).
36. van Dam, T. J. P. *et al.* CiliaCarta: An integrated and validated compendium of ciliary genes. *PLoS One* **14**, e0216705, <https://doi.org/10.1371/journal.pone.0216705> (2019).
37. Yu, X., Ng, C. P., Habacher, H. & Roy, S. Foxj1 transcription factors are master regulators of the motile ciliogenic program. *Nat Genet* **40**, 1445, <https://doi.org/10.1038/ng.263> (2008).
38. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).
39. Zheng, W. *et al.* Identification and functional analysis of spermatogenesis-associated gene modules in azoospermia by weighted gene coexpression network analysis. *Journal of cellular biochemistry* **120**, 3934–3944, <https://doi.org/10.1002/jcb.27677> (2019).
40. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33**, D514–D517, <https://doi.org/10.1093/nar/gki033> (2005).
41. Giotti, B. *et al.* Assembly of a parts list of the human mitotic cell cycle machinery. *J Mol Cell Biol*, <https://doi.org/10.1093/jmcb/mjy063> (2018).
42. Lechtreck, K. F. IFT–Cargo Interactions and Protein Transport in Cilia. *Trends Biochem Sci* **40**, 765–778, <https://doi.org/10.1016/j.tibs.2015.09.003> (2015).
43. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **28**, 1248, <https://doi.org/10.1038/nbt1210-1248> (2010).
44. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* **41**, D996–d1008, <https://doi.org/10.1093/nar/gks1042> (2013).
45. Stephen, L. A. *et al.* Failure of centrosome migration causes a loss of motile cilia in talpid(3) mutants. *Developmental dynamics: an official publication of the American Association of Anatomists* **242**, 923–931, <https://doi.org/10.1002/dvdy.23980> (2013).
46. Sanchez, T., Welch, D., Nicastro, D. & Dogic, Z. Cilia-Like Beating of Active Microtubule Bundles. *Science* **333**, 456–459, <https://doi.org/10.1126/science.1203963> (2011).
47. Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C. & Hume, D. A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics* **14**, 632 (2013).
48. Patir, A., Shih, B., McColl, B. W. & Freeman, T. C. A core transcriptional signature of human microglia: Derivation and utility in describing region-dependent alterations associated with Alzheimer’s disease. *Glia* **67**, 1240–1253, <https://doi.org/10.1002/glia.23572> (2019).
49. Nirmal, A. J. *et al.* Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors. *Cancer Immunol Res* **6**, 1388–1400, <https://doi.org/10.1158/2326-6066.cir-18-0342> (2018).
50. Shih, B. B. *et al.* Derivation of marker gene signatures from human skin and their use in the interpretation of the transcriptional changes associated with dermatological disorders. *J Pathol* **241**, 600–613 (2017).
51. Vij, S. *et al.* Evolutionarily Ancient Association of the FoxJ1 Transcription Factor with the Motile Ciliogenic Program. *PLoS Genet* **8**, e1003019, <https://doi.org/10.1371/journal.pgen.1003019> (2012).

52. Newton, F. G. *et al.* Forkhead Transcription Factor Fd3F Cooperates with Rfx to Regulate a Gene Expression Program for Mechanosensory Cilia Specialization. *Developmental Cell* **22**, 1221–1233, <https://doi.org/10.1016/j.devcel.2012.05.010> (2012).
53. Piasecki, B. P., Burghoorn, J. & Swoboda, P. Regulatory Factor X (RFX)-mediated transcriptional rewiring of ciliary genes in animals. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12969–12974, <https://doi.org/10.1073/pnas.0914241107> (2010).
54. Feng, C., Xu, W. & Zuo, Z. Knockout of the regulatory factor X1 gene leads to early embryonic lethality. *Biochem Biophys Res Commun* **386**, 715–717, <https://doi.org/10.1016/j.bbrc.2009.06.111> (2009).
55. Purvis, T. L. *et al.* Transcriptional regulation of the Alstrom syndrome gene ALMS1 by members of the RFX family and Sp1. *Gene* **460**, 20–29, <https://doi.org/10.1016/j.gene.2010.03.015> (2010).
56. Quigley, I. K. & Kintner, C. Rfx2 Stabilizes Foxj1 Binding at Chromatin Loops to Enable Multiciliated Cell Gene Expression. *PLOS Genetics* **13**, e1006538, <https://doi.org/10.1371/journal.pgen.1006538> (2017).
57. Chung, M. I. *et al.* Coordinated genomic control of ciliogenesis and cell movement by RFX2. *eLife* **3**, e01439, <https://doi.org/10.7554/eLife.01439> (2014).
58. Chen, B. *et al.* Auto-fatty acylation of transcription factor RFX3 regulates ciliogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E8403–e8412, <https://doi.org/10.1073/pnas.1800949115> (2018).
59. Pick, E., Honig, S. & Griffel, B. The Mechanism of Action of Soluble Lymphocyte Mediators. *Int Arch Allergy Immunol* **58**, 149–159, <https://doi.org/10.1159/000232187> (1979).
60. Winner, M. *et al.* A Novel, Macrophage Migration Inhibitory Factor Suicide Substrate Inhibits Motility and Growth of Lung Cancer Cells. *Cancer Res* **68**, 7253, <https://doi.org/10.1158/0008-5472.CAN-07-6227> (2008).
61. Wei, Q. *et al.* The BBSome controls IFT assembly and turnaround in cilia. *Nat Cell Biol* **14**, 950, <https://doi.org/10.1038/ncb2560> (2012).
62. Ye, F., Nager, A. R. & Nachury, M. V. BBSome trains remove activated GPCRs from cilia by enabling passage through the transition zone. *J Cell Biol* **217**, 1847–1868, <https://doi.org/10.1083/jcb.201709041> (2018).
63. Konno, A., Shiba, K., Cai, C. & Inaba, K. Branchial Cilia and Sperm Flagella Recruit Distinct Axonemal Components. *PLoS One* **10**, e0126005, <https://doi.org/10.1371/journal.pone.0126005> (2015).
64. Fischer, A. J. *et al.* Differential Gene Expression in Human Conducting Airway Surface Epithelia and Submucosal Glands. *American journal of respiratory cell and molecular biology* **40**, 189–199, <https://doi.org/10.1165/rcmb.2008-0240OC> (2009).
65. Ross, A. J., Dailey, L. A., Brighton, L. E. & Devlin, R. B. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *American journal of respiratory cell and molecular biology* **37**, 169–185, <https://doi.org/10.1165/rcmb.2006-0466OC> (2007).
66. Nevers, Y. *et al.* Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. *Molecular biology and evolution* **34**, 2016–2034, <https://doi.org/10.1093/molbev/msx146> (2017).
67. Beckers, A. *et al.* The evolutionary conserved FOXJ1 target gene Fam183b is essential for motile cilia in *Xenopus* but dispensable for ciliary function in mice. *Sci Rep* **8**, 14678, <https://doi.org/10.1038/s41598-018-33045-2> (2018).
68. Sigg, M. A. *et al.* Evolutionary Proteomics Uncovers Ancient Associations of Cilia with Signaling Pathways. *Dev Cell* **43**, 744–762, <https://doi.org/10.1016/j.devcel.2017.11.014> (2017).
69. van Dongen, S. & Abreu-Goodger, C. in *Bacterial Molecular Networks: Methods and Protocols* (eds Jacques van Helden, Ariane Toussaint, & Denis Thieffry) 281–295 (Springer New York, 2012).
70. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305–W311 (2009).
71. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091–1107, <https://doi.org/10.1016/j.cell.2018.02.001> (2018).
72. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res* **30**, 38–41 (2002).
73. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51–54, <https://doi.org/10.1093/nar/gkg129> (2003).
74. Boardman, P. E. *et al.* A Comprehensive Collection of Chicken cDNAs. *Curr Biol* **12**, 1965–1969, [https://doi.org/10.1016/S0960-9822\(02\)01296-4](https://doi.org/10.1016/S0960-9822(02)01296-4) (2002).
75. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168, <https://doi.org/10.1038/nature05453> (2006).
76. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419, <https://doi.org/10.1126/science.1260419> (2015).
77. Pérez-Silva, J. G., Araujo-Voces, M. & Quesada, V. nVenn: generalized, quasi-proportional Venn and Euler diagrams. *Bioinformatics* **34**, 2322–2324, <https://doi.org/10.1093/bioinformatics/bty109> (2018).

Acknowledgements

T.C.F., M.G.D., M.W.B., J.R. and L.M. receive funding from an Institute Strategic Programme Grant awarded by the Biotechnology and Biological Sciences Research Council, Institute Strategic Program Grant [BB/P013732/1: ‘BluePrints for Healthy Animals’]. A.M.F. is supported by a BBSRC EastBio Studentship.

Author contributions

A.P. carried out the formal analysis, wrote the manuscript and generated the figures and tables. A.M.F., M.W.B., L.M. and J.R. were involved in the methodology and validation. M.G.D. and T.C.F. conceptualized and supervised the project. All authors were involved in reviewing and editing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-66453-4>.

Correspondence and requests for materials should be addressed to T.C.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020