

1 CiliaCarta: an integrated and validated compendium 2 of ciliary genes

3 Authors

4 Teunis J. P. van Dam^{1*#}, Julie Kennedy², Robin van der Lee¹, Erik de Vrieze^{3,4}, Kirsten A. Wunderlich⁵,
5 Suzanne Rix⁶, Gerard W. Dougherty⁷, Nils J. Lambacher², Chunmei Li⁸, Victor L. Jensen⁸, Michel R.
6 Leroux⁸, Rim Hjeij⁷, Nicola Horn⁹, Yves Texier^{9§}, Yasmin Wissinger⁹, Jeroen van Reeuwijk¹⁰,
7 Gabrielle Wheway^{11†}, Barbara Knapp⁵, Jan F. Scheel^{5¶}, Brunella Franco^{12,13}, Dorus A. Mans¹⁰, Erwin
8 van Wijk^{3,4}, François Képès¹⁴, Gisela G. Slaats¹⁵, Grischa Toedt¹⁶, Hannie Kremer^{3,4,17}, Heymut
9 Omran⁷, Katarzyna Szymanska¹¹, Konstantinos Koutroumpas¹⁴, Marius Ueffing⁹, Thanh-Minh T.
10 Nguyen¹⁰, Stef J.F. Letteboer¹⁰, Machteld M. Oud¹⁰, Sylvia E. C. van Beersum¹⁰, Miriam Schmidts^{10,18},
11 Philip L. Beales⁶, Qianhao Lu^{19,20}, Rachel H. Giles¹⁵, Radek Szklarczyk^{1‡}, Robert B. Russell^{19,20}, Toby J.
12 Gibson¹⁶, Colin A. Johnson¹¹, Oliver E. Blacque², Uwe Wolfrum⁵, Karsten Boldt⁹, Ronald Roepman¹⁰,
13 Victor Hernandez-Hernandez⁶, Martijn A. Huynen^{1*}

14 Affiliations

15 ¹Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, PO Box
16 9101, 6500 HB Nijmegen, the Netherlands

17 ²School of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4,
18 Ireland

19 ³Department of Otorhinolaryngology, Radboud University Medical Center, PO Box 9101, 6500 HB
20 Nijmegen, the Netherlands

21 ⁴Donders Centre for Cognitive Neurosciences, 6525 AJ Nijmegen, the Netherlands

22 ⁵Department of Cell and Matrix Biology, Institute of Zoology, Johannes Gutenberg University of
23 Mainz, Mainz, Germany

24 ⁶Institute of Child Health, University College London, London, UK

25 ⁷Department of General Pediatrics, University Hospital Muenster, Muenster, Germany 48149

26 ⁸Department of Molecular Biology and Biochemistry and Centre for Cell Biology, Development and
27 Disease, Simon Fraser University, Burnaby, BC, Canada

28 ⁹Medical Proteome Center, Institute for Ophthalmic Research, University of Tuebingen, 72074
29 Tuebingen, Germany

30 ¹⁰Department of Human Genetics and Radboud Institute for Molecular Life Sciences, Radboud
31 University Medical Center, PO Box 9101, 6500 HB Nijmegen, the Netherlands

32 ¹¹Section of Ophthalmology & Neurosciences, Leeds Institute of Molecular Medicine, University of
33 Leeds, Leeds, LS9 7TF, UK.

34 ¹²Telethon Institute of Genetics and Medicine (TIGEM), Naples, Italy

1 ¹³Medical Genetics, Department of Translational Medicine, Federico II University of Naples, Naples,
2 Italy

3 ¹⁴Institute of Systems and Synthetic Biology (iSSB), Genopole, CNRS, Univ. Evry, France

4 ¹⁵Dept. Nephrology and Hypertension, Regenerative Medicine Center, University Medical Center
5 Utrecht, Uppsalalaan 6, 3584CT Utrecht, the Netherlands

6 ¹⁶Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstr.
7 1, 69117 Heidelberg, Germany

8 ¹⁷Department of Human Genetics, Radboud University Medical Center, PO Box 9101, 6500 HB
9 Nijmegen, the Netherlands

10 ¹⁸Pediatric Genetics Division, Center for Pediatrics and Adolescent Medicine, University Hospital
11 Freiburg, Mathildenstrasse 1, 79106 Freiburg, Germany

12 ¹⁹Biochemie Zentrum Heidelberg (BZH), Ruprecht-Karls Universitaet Heidelberg, Im Neuenheimer
13 Feld 328, Heidelberg, Germany

14 ²⁰Bioquant/Cell Networks, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

15 # Current address: Theoretical Biology and Bioinformatics, Science faculty, Utrecht University,
16 Padualaan 8, 3584 CH Utrecht, the Netherlands

17 § Current address: Agilent Technologies Sales & Services GmbH & Co. KG, Hewlett-Packard-Str. 8,
18 76337 Waldbronn

19 †Current address: Centre for Research in Biosciences, University of the West of England,
20 Coldharbour Lane, Bristol, BS16 1QY, UK.

21 ¶Current address: Max Planck Institute for Chemistry, Department of Multiphase Chemistry, Mainz,
22 Germany

23 ‡ Current address: Department of Clinical Genetics, Unit Clinical Genomics, Maastricht University
24 Medical Centre, P.O. Box 616, 6200 MD Maastricht, the Netherlands

25 *Corresponding authors

26 Teunis J. P. van Dam, PhD

27 Email: t.j.p.vandam@uu.nl

28 Prof. Martijn A. Huynen, PhD

29 Email: martijn.huijnen@radboudumc.nl

1 **Short title**

2 CiliaCarta

3 **Abstract**

4 The cilium is an essential organelle at the surface of most mammalian cells whose dysfunction
5 causes a wide range of genetic diseases collectively called ciliopathies. The current rate at which
6 new ciliopathy genes are identified suggests that many ciliary components remain undiscovered.
7 We generated and rigorously analyzed genomic, proteomic, transcriptomic and evolutionary data
8 and systematically integrated these using Bayesian statistics into a predictive score for ciliary
9 function. This resulted in 285 candidate ciliary genes. We found experimental evidence of ciliary
10 associations for 24 out of 36 analyzed candidate proteins. In addition, we show that OSCP1, which
11 has previously been implicated in two distinct non-ciliary functions, causes a cilium dysfunction
12 phenotype when depleted in zebrafish. The candidate list forms the basis of CiliaCarta, a
13 comprehensive ciliary compendium covering 836 genes. The resource can be used to objectively
14 prioritize candidate genes in whole exome or genome sequencing of ciliopathy patients and can be
15 accessed at <http://bioinformatics.bio.uu.nl/john/syscilia/ciliacarta/>.

16 **Keywords**

17 Ciliary genes, the cilium, Bayesian classifier, CiliaCarta, OSCP1, ciliary proteome, data integration

1 Cilia are microtubule-based organelles extending from the surface of most eukaryotic cells, serving
2 critical functions in cell and fluid motility, as well as the transduction of a plethora of sensory and
3 biochemical signals associated with developmental processes¹. Cilium disruption leads to a wide
4 range of human disorders, known as ciliopathies, characterized by defects in many different tissues
5 and organs leading to symptoms such as cystic kidneys, blindness, bone malformation, nervous
6 system defects and obesity. The cilium is a complex and highly organized structure, typically
7 comprised of a ring of nine microtubule doublets extending from a centriole-derived basal body,
8 and enveloped by an extension of the plasma membrane. Importantly, cilia are compartmentalized
9 structures, with a membrane and internal composition that differs significantly from that of the
10 plasma membrane and cytoplasm^{2,3}. Several hundred proteins are thought to be involved in the
11 formation and function of ciliary structures and associated signaling and transport pathways.
12 Between Gene Ontology (GO)⁴ and the SYSCILIA Gold Standard (SCGS)⁵ about 600 genes have
13 already been associated with ciliary function. However, it is likely that many more ciliary proteins
14 remain to be functionally characterized, as new ciliary genes are uncovered on a regular basis, often
15 in relation to cilia-related genetic disorders.

16 Although omics data sets provide a rich source of information to determine the proteins that
17 constitute an organelle, they are inherently imperfect. They tend to be biased towards proteins with
18 specific properties, or lack coverage and sensitivity. It is well established that large scale
19 approaches often miss key players, for example proteomics fares better at finding intracellular
20 versus extracellular or membrane proteins⁶. Combining data sets into a single resource that
21 exploits their complementary nature is a logical step. The power of such an approach has been
22 demonstrated for various cellular systems such as small RNA pathways⁷, the innate antiviral
23 response⁸, and, most notably, by MitoCarta⁹. The latter is a compendium of mitochondrial proteins
24 based on the integration of various types of genomics data that has been extensively used by the
25 biomedical community¹⁰.

26 Like the mitochondrion, the cilium has been subject to approaches that exploit signals in genomics
27 data to predict new ciliary genes, e.g. the specific occurrence of genes in species with a cilium^{11,12},
28 the sharing of specific transcription factor binding sites like the X-box motif¹³ and the
29 spatiotemporal co-expression of ciliary genes¹⁴. Furthermore, an existing database, CilDB, contains
30 data from individual genomics, transcriptomics, and proteomics experiments related to the
31 cilium¹⁵. Although such databases are invaluable to the researcher, it is not obvious how to handle
32 the sometimes-conflicting information presented by independent data sources: i.e. how to weigh
33 the data relative to each other. A powerful solution lies in statistical and probabilistic integration of
34 data sets. Multiple methods exist to combine genomics data, ranging from simply taking the genes
35 that are predicted by most data sources, to machine learning methods that take into account non-
36 linear combinations of the data (reviewed in¹⁶). In this spectrum, naive Bayesian classifiers take a
37 middle ground. They exploit the relative strengths of the various data sets while maintaining

1 transparency of the integration. For each gene, the contribution of each data set to the prediction
2 can be determined, and new, independent data sources can then simply be added to those already
3 used to improve predictive value and coverage.

4 Here, we present CiliaCarta, an experimentally benchmarked compendium of ciliary genes based on
5 literature, annotation, and genome-wide Bayesian integration of a wide range of experimental data,
6 including a recent large-scale protein-protein interaction data set specifically focused on ciliary
7 proteins¹⁷. We extend the currently known set of cilia-related genes with 228 putative cilia-related
8 genes to a total of 836 genes. Based on the outcome of the probabilistic integration and the results
9 from the experimental validations, we estimate the total size of the human ciliome to be
10 approximately 1200 genes. Furthermore, we show that objective data integration using Bayesian
11 probabilities is capable of overcoming biases based on other sources or literature. As an exemplar
12 of our approach, we show that OSCP1, which several publications have described as a solute carrier
13 or tumor suppressor, is also a ciliary component, shedding new light on previous observations.

14 Results

15 Data collection and curation

16 We collected and constructed a total of six new data sets from proteomics, genomics, expression
17 and evolutionary data and complemented these with two public data sets (Table 1). The data sets
18 include three new protein-protein interaction (PPI) data sets based on three methods: tandem-
19 affinity purification coupled to mass spectrometry (TAP-MS)¹⁷, stable isotope labeling of amino
20 acids in cell culture (SILAC), and yeast two-hybrid (Y2H) screens. The latter two data sets are
21 published as part of this work and include 1666 proteins (1301 and 365, resp.) describing 4659
22 interactions (4160 and 499, resp.) and an estimated positive predictive value of X. Because the TAP-
23 MS and SILAC data sets are based on similar methodology and have a large bait overlap (14 out of
24 16 SILAC baits were used in the TAP-MS data set), we merged them into a single data set (Mass-
25 spec based PPI) to avoid bias (see methods). In addition, we created three bioinformatic data sets:
26 (i) a data set describing the presence or absence of conserved cilia-specific RFX and FOXJ1
27 transcription factor binding sites (TFBS) in the promoter regions of human genes based on the 29
28 mammals project¹⁸, (ii) a gene expression screen for genes that co-express with a set of known
29 ciliary genes, and (iii) a comprehensive co-evolution data set from the presence-absence
30 correlation patterns of genes with the presence of the cilium over a representative data set of
31 eukaryotic species. Supplementing these six data sets we included two published large-scale data
32 sets, covering primary cilia and motile cilia: (i) the Liu *et al.* data set is a proteomics data set
33 describing the protein content of sensory cilia derived from isolated murine photoreceptor cells¹⁹;
34 (ii) the Ross *et al.* data set is a dynamic gene expression data set describing the up-regulation of
35 genes during ciliogenesis in a time series after shearing off cilia in human lung epithelial cells²⁰.

1 Although both data sets are not among the strongest predictors for ciliary function, they are of good
2 quality and, importantly, have a much higher coverage of the human proteome than other
3 published datasets (Supplementary fig. 1). In addition, the eight data sets were selected to ensure
4 independence of the types of evidence and comprehensive coverage of molecular signatures for
5 ciliary genes (for more details see methods). Each data set contains a highly significant signal for
6 the discovery of ciliary genes (Table 1, p-values ranging from 2.1E-14 to 2.4E-69). A full description
7 on each data set is given in the methods section.

8 **Bayesian integration of omics data sets provides gene-specific probabilities for** 9 **cilia involvement**

10 By combining the complementary sources of evidence for cilium function, we can in principle,
11 obtain a data-driven, objective, high-confidence compendium of ciliary genes. To integrate the data
12 sets in a probabilistic manner we assessed their efficacy at predicting ciliary genes using the SCGS, a
13 manually curated set of known ciliary genes (the ‘positive’ set)⁵, and a set of genes whose proteins
14 show non-ciliary subcellular localization and are most likely not involved in ciliary function (the
15 ‘negative’ set, see methods). We divided each data set into appropriate sub-categories that reflect
16 increasing propensities to report ciliary genes (Fig. 1a). Then, for each data category we calculated
17 the likelihood ratios of predicting ciliary genes versus predicting non-ciliary genes using Bayes’
18 theorem. The final log-summed likelihood ratio, which includes the prior expected probability to
19 observe a ciliary gene in the genome, we here call the CiliaCarta Score. This score therefore
20 represents the likelihood for a gene to be ciliary, based on all the data sets considered (see
21 methods; CiliaCarta and individual data set scores can be found in Supplementary table 1). The
22 integrated score readily distinguishes between the positive and negative set (Fig. 1b, p-value: 2.8e-
23 85 Mann-Whitney U test). A ten fold cross-validation demonstrates that the results are highly
24 robust, with an area under the curve of 0.86 (Supplementary Fig. 2). The top ranking genes are
25 highly enriched for known ciliary genes (Fig. 1c). The Bayesian classifier outperforms any
26 individual data set, while achieving genome-wide coverage (Fig. 1d).

27 **Experiments validate ciliary function for 67% of selected candidate genes**

28 In order to validate the quality of our approach we performed a series of experimental tests for
29 ciliary function or localization of newly predicted candidate genes and their proteins. We set our
30 inclusion threshold at a False Discovery Rate of 25% (cFDR, corrected for the prior expectation to
31 observe a cilium gene, Methods). 404 genes fall within this threshold, 285 of which were not in the
32 training sets and thus constitute novel candidate ciliary genes. Eight genes from the negative set
33 (HSPH1, CALM3, COL21A1, CALM1, PTGES3, HSPD1, COL28A1, and PPOX) occur in the top 404
34 genes. Literature searches revealed no previous connection to the cilium. Some high scoring non-
35 ciliary genes are expected due to stochasticity in the underlying experimental data, but it is possible
36 that some of these genes will turn out to have a ciliary function in the future. The FDR for the 285

1 candidate genes is 33% when excluding genes from the training sets (Methods). From the
2 candidates we selected a total of 36 genes, spread evenly across the top predictions, for
3 experimental validation that were not known to be ciliary at the time (Table 2). The gene selection
4 was performed as unbiased as possible and was restricted by orthology (zebrafish, *C. elegans*) and
5 the resources available to the participating labs (see methods). We performed validations in
6 human, mouse, zebrafish and *C. elegans* by applying six distinct approaches to determine ciliary
7 localization and function.

8 *Validation by phenotype*

9 In *C. elegans* we investigated candidate gene associations with cilium formation and function *in vivo*.
10 In the *C. elegans* hermaphrodite (959 cells), cilia are found on 60 sensory neurons, most of which
11 are located in the head. These cilia are immotile, eight of which extend from the dendritic endings,
12 forming sensory pores in the nematode cuticle²¹. 68 of the 285 candidate genes have one-to-one
13 orthologs in *C. elegans*. The apparent low number of orthologs in *C. elegans* could have a number of
14 reasons, one of which is the absence of motile cilia, which resulted in the loss of ciliary genes
15 related to motility. For 21 of these genes, viable alleles were available from the *Caenorhabditis*
16 Genetics Center (University of Minnesota, USA) (Supplementary table 2). The available alleles were
17 all nonsense or deletion mutations and predicted to severely disrupt gene function. For each
18 available mutant, we employed dye-filling assays - to indirectly assess the integrity of a subset of
19 cilia (six amphid pairs and one phasmid pair) - foraging (roaming) and osmotic avoidance assays to
20 investigate sensory behaviors^{22,23}. Notably, none of the mutants showed an abnormal osmotic
21 avoidance response, or a highly penetrant dye uptake phenotype (Supplementary table 2), implying
22 that the associated genes are not involved in global regulation of cilium formation or function. In
23 addition all worm mutants had normal physical appearance. Roaming defects were observed for
24 nine mutants, three of which also displayed a mild defect in dye uptake (Fig. 2a & b, Supplementary
25 Fig. 3, and Supplementary table 2).

26 Three candidate genes with clear orthologs in zebrafish, *srgap3*, *ttc18* and *rab36* were investigated
27 in zebrafish for cilia-related phenotypes after knockdown by morpholinos. The development of cilia
28 in the Kupffer's vesicle was examined (Fig. 3a). *rab36* and *ttc18* morphants show significantly
29 reduced cilia length and number in the Kupffer's vesicle ($p < 0.001$ and $p < 0.05$ resp.), while the
30 *srgap3* morphant shows an increased cilia length ($p < 0.001$), but no effect on cilia number.
31 Pronephric ducts are several orders of magnitude larger in all three morphants during the
32 pharyngula period (24 hpf) compared to wild type (Fig. 3b, p -value < 0.001). Furthermore, all three
33 morphants exhibit body-axis defects associated with cilium dysfunction^{24,25} (Fig. 3c). Although,
34 zebrafish morpholino phenotypes, and *C. elegans* phenotypes, do not provide definitive proof that
35 these genes are all ciliary, they provide quantitative support for a substantial enrichment of ciliary
36 genes in the CiliaCarta.

1 *Validation by subcellular localization*

2 Subcellular localization studies were performed for a total of nine proteins in ciliated hTERT-RPE1
3 cells using eCFP-tagged overexpression constructs (Table 2, Supplementary Fig.4). Eight proteins
4 localized to the basal body and/or the axoneme. To account for possible localization artifacts, two
5 representative photos are taken per eCFP fusion protein, containing at least one ciliated cell from
6 the same slide. Cells transfected for c15orf22::eCFP and c16orf80 ciliated only when expression of
7 the eCFP fusion protein was low. Fig. 4a shows representative examples, with C20orf26 and
8 CCDC147 localizing to the basal body, whilst IQCA1 was not enriched at the cilium.

9 Four proteins (RIBC2, ARMC3, CCDC113 and C6orf165) were investigated for ciliary localization of
10 the endogenous protein by immunofluorescence microscopy. RIBC2, ARMC3 and CCDC113 were
11 observed to co-localize with acetylated alpha tubulin, a marker for the ciliary axoneme in human
12 respiratory cells, while C6orf165 specifically localized to the base of the cilium (Fig. 4b). We also
13 tested these four proteins in other model systems to investigate if their ciliary localization is model-
14 specific. Location was validated in murine retinal sections by immunofluorescence and at the
15 ultrastructural level via immunogold electron microscopy (Fig. 4c-f). All four proteins associated
16 with the sensory cilia of photoreceptor cells. However, in hTERT-RPE1 cells only RIBC2 showed
17 localization at the basal body suggesting that ciliary localization can be cell type specific (Fig. 4g).
18 Concurrent to our efforts other labs recently identified ciliary involvement for RIBC2²⁶ and
19 CCDC113²⁷.

20 *Overall validation performance*

21 Combining the results of all the validation experiments, we observed a ciliary localization or
22 putative cilium-associated phenotype for 24 of the 36 candidates tested (Table 2). There appears to
23 be a notable performance difference between the localization- and phenotype-based validation
24 assays (PPV of 0.92 and 0.56 respectively, Fig. 4h). This difference is likely attributable to the fact
25 that knockdown of a ciliary gene does not necessarily lead to an observable phenotype. Although
26 we should keep in mind that the nematode and zebrafish ciliary phenotypes require confirmation
27 via rescue experimentation, the overall experimentally determined positive predictive value (PPV)
28 of the Bayesian classifier will be within 40 to 70%, approaching 67% depending on conservative or
29 optimistic interpretations of the presence of false positives or false negatives in our validations (Fig.
30 4h).

31 The observed experimental PPV corresponds with the theoretical PPV of 67%, which corresponds
32 to the FDR of 33% calculated for the candidate genes from the integrated CiliaCarta score (expected
33 validation rate: 24.1 out of 36, $p(x=24 \text{ given } 33\% \text{ FDR})=0.15$, hypergeometric test, Fig. 4h). Even if
34 we assume the lowest PPV of 40%, the experimental validation rate is significantly higher than
35 expected by chance, based on the estimated prior distribution of ciliary versus non-ciliary genes in

1 the human genome (expected 1.8 out of 36, $p=3.34e-23$ for PPV=67% and $p=7.64e-10$ for
2 PPV=40%; hypergeometric test). The experimental verification of our candidate genes therefore
3 validates our Bayesian classifier.

4 **OSCP1, a novel ciliary protein**

5 Unbiased integration of large-scale genomics data can give rise to apparent inconsistencies with
6 previous literature reports. For instance, organic solute carrier partner 1 (OSCP1, or oxidored-nitro
7 domain-containing protein 1, NOR1), scores high on our CiliaCarta list (ranked 402), despite
8 reports of varied functions not obviously consistent with a ciliary role, such as regulation of
9 inflammation, apoptosis, proliferation and tumor suppression²⁸⁻³⁰. OSCP1 was first implicated as a
10 tumor suppressor in nasopharyngeal cancer²⁸ and was later found to modulate transport rates of
11 organic solutes over the plasma membrane in rodents^{31,32}. These two facets of OSCP1 function have
12 never been connected to each other in literature. Subsequent independent research indicated a role
13 for OSCP1 in regulation of inflammation and apoptosis²⁹, and that it is specifically expressed in
14 mouse testes³³. The availability of a substantial set of literature on OSCP1 without any indication of
15 ciliary involvement as predicted by our method, would have suggested that OSCP1 could have been
16 a false positive in our Bayesian method. However, our *C. elegans* phenotype screen suggests that *C.*
17 *elegans oscp-1* (R10F2.5) mutant alleles may possess modest sensory cilia defects (Fig. 2a & b).
18 Therefore, we targeted OSCP1 for more detailed investigation of OSCP1 in *C. elegans*, zebrafish and
19 mammalian cells.

20 *OSCP1 locates at the base and axoneme of the cilium in human, mouse and C. elegans cells*

21 In *C. elegans*, expression of GFP-tagged OSCP-1, driven by its endogenous promoter, is restricted to
22 ciliated sensory neurons (the only ciliated cells in the nematode), indicating a high likelihood of a
23 cilium-associated function (Fig. 5a; Supplementary Fig. 5). Analysis of the subcellular localization
24 pattern confirmed this by showing that OSCP-1::GFP localizes specifically to ciliary axonemes,
25 including the ciliary base (Fig. 5a; Supplementary Fig. 5). In human hTERT-RPE1 cells, OSCP1::eCFP
26 localizes at the basal body and daughter centriole in ciliated cells and at the centrioles of non-
27 ciliated cells, indicating a basal body/centriole role for OSCP1 in this epithelial cell type (Fig. 5b). In
28 some cells a clear punctate localization in the cytoplasm can also be observed.

29 To assess the localization of the endogenous OSCP1 protein in various mammalian cells, we used
30 commercially available antibodies from three suppliers (Proteintech, Atlas and Biorbyt). In human
31 multi-ciliated respiratory cells, OSCP1 shows an increased concentration at the base of the cilia (Fig.
32 5c). In murine photoreceptors OSCP1 is localized at the inner segments (Fig. 5d), where it is
33 particularly abundant at the base of the connecting cilium (the equivalent of the transition zone in
34 primary cilia), and at low concentration at the adjacent centriole. Immunoelectron microscopy
35 analysis also shows OSCP1 at the basal body, occasionally at the connecting cilium and sporadically

1 within the inner segment (Fig. 5d). Serum-starved IMCD3 cells (murine collecting duct cells)
2 showed ubiquitous punctate cytoplasmic staining with increased signal at the basal body and along
3 the ciliary axoneme (Fig. 5e; Supplementary Fig. 6). In ATDC5 cells (murine pre-chondrocyte cells)
4 OSCP1 is mainly localized to the ciliary axonemes, although the ciliary base signal is less apparent
5 (Supplementary Fig. 6).

6 Together, these exhaustive subcellular localization analyses place OSCP1 at ciliary structures,
7 including the ciliary base and axoneme, although there are some subtle differences between cell
8 types. We also frequently observed extensive non-ciliary signals for OSCP1, consistent with
9 published reports for OSCP1 localizations at other organelles (ER, Golgi, Mitochondria) and within
10 the cytosol^{34,35}.

11 *OSCP1 is required for cilium formation in multiple zebrafish tissues, but dispensable in C. elegans*
12 *sensory neurons*

13 To investigate possible ciliogenesis roles for OSCP1, we first examined cilia in zebrafish morphants
14 depleted for *oscp1* in fertilized eggs. In embryos at 2 days post fertilization (dpf) developmental
15 phenotypes were observed that are often associated with ciliary defects: a curved body axis, small
16 eyes and melanocyte migration defects (Fig. 5f). Likewise, the 4 dpf morphants presented with
17 pronephric cysts, small eyes, heart edema, small heads and short bodies (Fig. 5f & g). The cilia in the
18 medial portion of the pronephric ducts of *oscp1* morphants were shortened and disorganized (Fig.
19 5h&i), and in the Kupffer's vesicle cilium length and number was decreased compared to control
20 injected larvae (Fig. 5j). Co-injecting human *OSCP1* mRNA together with the morpholino partially
21 rescued the observed phenotype, indicating that the observed phenotype is specific for loss of *oscp1*
22 function (Fig. 5k and Supplementary Fig. 7). Therefore, in zebrafish, *oscp1* is required for cilium
23 formation and associated functions in many tissue types and organs.

24 In contrast, analysis of the *oscp-1(gk699)* null allele in *C. elegans* revealed that OSCP-1 is not
25 required for cilium formation. Using fluorescence reporters and transmission electron microscopy,
26 the amphid and phasmid channel cilia appeared to be intact, and full length (Supplementary Fig. 5
27 and 9). In addition, *oscp-1* does not appear to be functionally associated with the transition zone at
28 the ciliary base; disruption of this prominent ciliary domain ('gate') in the *mks-5* mutant³⁶ does not
29 affect OSCP-1::GFP localization, and loss of *oscp-1* itself does not influence the localization of several
30 transition zone proteins (Supplementary Fig. 5). Thus, OSCP1 is differentially required for cilium
31 formation in worms and zebrafish, reflecting species distinctions in the ciliary requirement for this
32 protein. These distinctions could reflect redundancy of OSCP1 function with another ciliary protein
33 in the nematode, but not in zebrafish, or differences in cell type requirements (nematode sensory
34 neurons versus zebrafish epithelial cells). Clearly, based on its restricted expression in ciliated cells
35 and localization with the ciliary axoneme, *C. elegans* OSCP-1 is serving a ciliary function, although
36 the specifics of this role remain to be elucidated.

1 Discussion

2 With the current interest in cilia biology it is certain that many new genes will be implicated in
3 ciliary function for several years to come. With the advent of systems biology and the need to
4 understand the cilium as a whole the research community requires an inventory of genes and
5 proteins involved in ciliary structure and function. The obvious sources for such an inventory, GO⁴
6 and the SCGS⁵, currently only cover 608 human genes (510 in GO as of December 3rd 2015, 302 for
7 the SCGS). By applying a naive Bayesian integration of heterogeneous large-scale ciliary data sets
8 we have expanded this set by 38% to 836 human genes, adding 228 putative genes to the known
9 cilium gene repertoire (Supplementary Fig. 8). GO term enrichment analysis indicates that these
10 putative ciliary genes are enriched for genes that are “unclassified” (86 genes, 1.69 fold enrichment,
11 p-value < e-100) suggesting possible new ciliary biology to be discovered. We put forward these
12 putative ciliary genes, together with the SCGS and the GO annotated genes, as the “CiliaCarta”, a
13 compendium of ciliary components with an estimated FDR of 10% (Supplementary table 3). This
14 community resource can be used to facilitate the discovery of new cilium biology and to identify the
15 genetic causes of cilia related genetic disease. CiliaCarta therewith has a very different purpose than
16 the SCGS, it serves as a tool for discovery of new genes, rather than as a reference of known cilium
17 genes.

18 CiliaCarta is still likely to be incomplete. Estimates of the ciliary proteome range from one to two
19 thousand proteins depending on the techniques used and on the types of cilia and the species
20 studied^{19,37}. In addition to obtaining the CiliaCarta list of proteins, our Bayesian analysis allows us
21 to obtain an objective estimate of the total number of ciliary proteins. Using the posterior
22 probabilities and the outcome from the validation experiments we estimate the size of the ciliome
23 to be approximately 1200 genes (Methods).

24 Predicting the genes responsible for an organelle structure or function poses the question of where
25 we draw the boundary between that organelle and the rest of the cell. For example, does the basal
26 body in its entirety belong to the cilium or are some components to be considered exclusive to the
27 centrioles? Furthermore, one can argue that in the case of the cilium, proteins that are not part of
28 the cilium, but play a role in the transport of proteins to the ciliary base or regulation of cilium gene
29 expression, can be regarded as components of the ‘ciliary system’. In practice, what we regard here
30 as a ciliary component depends on the definition within the SCGS, which is used to weigh the data
31 sets, and on our experimental validations. In both we have taken a rather inclusive approach by
32 regarding genes whose disruption cause ciliary phenotypes as ciliary genes. This approach makes
33 our predictions relevant to human disease. Indeed KIAA0753, which falls within our 25% cFDR list
34 of cilium genes, was recently shown to interact with OFD1 and FOR20 at pericentriolar satellites
35 and centrosomes, and gives rise to oral-facial-digital syndrome³⁸, a phenotype associated with
36 disruption in the ciliary transition zone and the basal body.

1 Data integration through objective quantification of the predictive value of individual large-scale
2 data sets allows one to find new functions and associations, without bias from previous studies. As
3 a case in point we report here that OSCP1, not previously implicated in ciliary functions based on
4 the existing published literature, is validated as a ciliary protein in four species as determined by
5 multiple independent experimental methods. Re-evaluation of previous experimental evidence on
6 OSCP1 function does not exclude ciliary involvement. Nevertheless, the cilium contains several
7 specific ion-channels in its membrane³⁹ and the organelle has been implicated to play a role in the
8 development of cancer⁴⁰⁻⁴². Therefore connecting OSCP1 to the cilium might provide the missing
9 link to connect previously observed effects of OSCP1 on organic solute in-/efflux and its role in
10 nasopharyngeal cancer. Our results therefore provide a cellular target and biomolecular framework
11 to further unravel OSCP1 function.

12 Systematic integration of heterogeneous large-scale cilia data sets by employing Bayesian statistics
13 combined with medium throughput experimental validation is a powerful approach to identify
14 many new ciliary genes and provides a molecular definition of the cilium. The experimental
15 observations of a potential ciliary role for selected high-confidence candidates, together with the
16 results from the cross validation, indicates that the top tier of the entire ranked human genome is
17 highly enriched for ciliary genes. The genome-wide CiliaCarta Score and ranking as provided here,
18 should therefore make it possible to efficiently and objectively prioritize candidate genes in order
19 to discover new ciliary genes and ciliary functions.

20 Materials and methods

21 Data set collection and mapping

22 All data sets were mapped to the ENSEMBL human gene set version 71, release April 2013⁴³.
23 Resources using other identifiers (i.e. Entrez, Uniprot) were mapped to ENSEMBL gene IDs using
24 ENSEMBL BioMART (version 71). Orthologs from non-human data sets were mapped using the
25 ENSEMBL Compara ortholog catalogue from ENSEMBL 71, with exception of Sanger sequences
26 from Y2H screens based on the Bovine cDNA library that were mapped to Bovine genome
27 sequences by the BLAT tool⁴⁴ in the UCSC genome browser⁴⁵ and subsequently mapped to human
28 orthologs using the 'non-cow RefSeq genes track' from this browser.

29 DNA constructs

30 Bait protein selection was based on the association of proteins with ciliopathies (including mutant
31 vertebrates showing ciliopathy features), involvement in IFT or part of our candidate list of ciliary
32 proteins. Gateway-adapted cDNA constructs were obtained from the Ultimate™ ORF clone
33 collection (Thermo Fisher Scientific) or generated by PCR from IMAGE clones (Source BioScience)
34 or human marathon-ready cDNA (Clontech) as template and cloning using the Gateway cloning

1 system (Thermo Fisher Scientific) according to the manufacturer's procedures followed by
2 sequence verification.

3 **Yeast two-hybrid system**

4 A GAL4-based yeast two-hybrid system was used to screen for binary protein-protein interactions
5 with proteins expressed from several different cDNA libraries (see below) as described
6 previously⁴⁶. Yeast two-hybrid constructs were generated according to the manufacturer's
7 instructions using the Gateway cloning technology (Thermo Fisher Scientific) by LR recombination
8 of GAL4-BD Gateway destination vectors with sequence verified Gateway entry vectors containing
9 the cDNA's of selected bait proteins.

10 Constructs encoding full-length or fragments of bait proteins fused to a DNA-binding domain
11 (GAL4-BD) were used as baits to screen human oligo-dT primed retinal, brain (Human Foetal Brain
12 Poly A+ RNA, Clontech), kidney (Human Adult Kidney Poly A+ RNA, Clontech) or testis cDNA
13 libraries, or a bovine random primed retinal cDNA library, fused to a GAL4 activation domain
14 (GAL4-AD). The retina and testis two-hybrid libraries were constructed using HybriZAP-2.1
15 (Stratagene), the brain and kidney two-hybrid libraries were constructed using the "Make Your
16 Own Mate & Plate™ Library System" (Clontech).

17 The yeast strains PJ96-4A and PJ96-4a (opposing mating types), which carry the *HIS3* (histidine),
18 *ADE2* (adenine), *MEL1* (α -galactosidase), and *LacZ* (β -galactosidase) reporter genes, were used as
19 hosts. Interactions were identified by reporter gene activation based on growth on selective media
20 (*HIS3* and *ADE2* reporter genes), α -galactosidase colorimetric plate assays (*MEL1* reporter gene),
21 and β -galactosidase colorimetric filter lift assays (*LacZ* reporter gene).

22 **Affinity purification of protein complexes using SILAC**

23 *DNA constructs and cell culture*

24 Experiments were essentially performed as described before⁴⁷. In short, N-terminally SF-TAP-
25 tagged bait proteins that were obtained by LR recombination of TAP-destination vector with
26 sequence verified Gateway entry vectors containing the cDNA's of selected bait proteins using
27 Gateway cloning technology (Thermo Fisher Scientific). HEK293T cells were seeded, grown
28 overnight, and then transfected with SF-TAP-tagged bait protein constructs using Effectene
29 (Qiagen) according to the manufacturer's instructions. HEK293T cells were grown in SILAC cell
30 culture medium as described⁴⁷.

31 *Affinity purification of protein complexes*

32 For one-step Strep purifications, SF-TAP-tagged proteins and associated protein complexes were
33 purified essentially as described previously⁴⁷. In short, SILAC labeled HEK293T cells, transiently

1 expressing the SF-TAP tagged constructs were lysed in lysis buffer containing 0.5% Nonidet-P40,
2 protease inhibitor cocktail (Roche), and phosphatase inhibitor cocktails II and III (Sigma-Aldrich) in
3 TBS (30 mM Tris-HCl, pH 7.4, and 150 mM NaCl) for 20 minutes at 4°C. After sedimentation of
4 nuclei at 10,000 g for 10 minutes, the protein concentration was determined using a standard
5 Bradford assay. Equal protein amounts were used as input for the experiments to be compared. The
6 lysates were then transferred to Strep-Tactin-Superflow beads (IBA) and incubated for 1 hour
7 before the resin was washed 3 times with wash buffer (TBS containing 0.1% NP- 40 and
8 phosphatase inhibitor cocktails II and III). The protein complexes were eluted by incubation for 10
9 minutes in Strep-elution buffer (IBA). The eluted samples were combined and concentrated using
10 10-kDa cutoff VivaSpin 500 centrifugal devices (Sartorius Stedim Biotech) and prefractionated
11 using SDS-PAGE and in-gel tryptic cleavage as described elsewhere⁴⁸.

12 *Quantitative mass spectrometry*

13 After precipitation of the proteins by methanol-chloroform, a tryptic in-solution digestion was
14 performed as described previously⁴⁹. LC-MS/MS analysis was performed on a NanoRSLC3000 HPLC
15 system (Dionex) coupled to a LTQ OrbitrapXL, respectively coupled to a LTQ Orbitrap Velos mass
16 spectrometer (Thermo Fisher Scientific) by a nano spray ion source. Tryptic peptide mixtures were
17 automatically injected and loaded at a flow rate of 6 µl/min in 98% buffer C (0.1% trifluoroacetic
18 acid in HPLC-grade water) and 2% buffer B (80% acetonitrile and 0.08% formic acid in HPLC-grade
19 water) onto a nanotrap column (75 µm i.d. × 2 cm, packed with Acclaim PepMap100 C18, 3 µm, 100
20 Å; Dionex). After 5 minutes, peptides were eluted and separated on the analytical column (75 µm
21 i.d. × 25 cm, Acclaim PepMap RSLC C18, 2µm, 100 Å; Dionex) by a linear gradient from 2% to 35%
22 of buffer B in buffer A (2% acetonitrile and 0.1% formic acid in HPLC-grade water) at a flow rate of
23 300 nl/min over 33 minutes for EPASIS samples, and over 80 minutes for SF-TAP samples.
24 Remaining peptides were eluted by a short gradient from 35% to 95% buffer B in 5 minutes. The
25 eluted peptides were analyzed by using a LTQ Orbitrap XL, or a LTQ OrbitrapVelos mass
26 spectrometer. From the high-resolution mass spectrometry pre-scan with a mass range of 300–
27 1,500, the 10 most intense peptide ions were selected for fragment analysis in the linear ion trap if
28 they exceeded an intensity of at least 200 counts and if they were at least doubly charged. The
29 normalized collision energy for collision-induced dissociation was set to a value of 35, and the
30 resulting fragments were detected with normal resolution in the linear ion trap. The lock mass
31 option was activated and set to a background signal with a mass of 445.12002⁵⁰. Every ion selected
32 for fragmentation was excluded for 20 seconds by dynamic exclusion.

33 For quantitative analysis, MS raw data were processed using the MaxQuant software⁵¹ (version
34 1.5.0.3). Trypsin/P was set as cleaving enzyme. Cysteine carbamidomethylation was selected as
35 fixed modification and both methionine oxidation and protein acetylation were allowed as variable
36 modifications. Two missed cleavages per peptide were allowed. The peptide and protein false

1 discovery rates were set to 1%. The initial mass tolerance for precursor ions was set to 6 ppm and
2 the first search option was enabled with 10 ppm precursor mass tolerance. The fragment ion mass
3 tolerance was set to 0.5 Da. The human subset of the human proteome reference set provided by
4 SwissProt (Release 2012_01 534,242 entries) was used for peptide and protein identification.
5 Contaminants like keratins were automatically detected by enabling the MaxQuant contaminant
6 database search. A minimum number of 2 unique peptides with a minimum length of 7 amino acids
7 needed to be detected to perform protein quantification. Only unique peptides were selected for
8 quantification.

9 **Protein-protein interaction data processing**

10 Based on three ciliary protein-protein interaction (PPI) data sets, we inferred proteins to "interact
11 with ciliary components" as a proxy for being part of the cilium. First, we obtained protein complex
12 purification data from a large-scale study on the identification of ciliary protein complexes by
13 tandem-affinity purification coupled to mass spectrometry (TAP-MS) for 181 proteins known or
14 predicted to be involved in ciliary functions¹⁷. Since integration of this data set into the CiliaCarta
15 many pull-downs were repeated and reverse experiments included. As a result, our data set
16 includes 539 found interactors that are not part of the now published final data set (4702 proteins,
17 Table S2 from Boldt *et al.*¹⁷), and does not include 679 new interactors that have been identified
18 since June 2013. The complete and current data is available at <http://landscape.syscilia.org/> and
19 IntAct [IM-25054]. Second, we obtained affinity purification data for 16 bait proteins from a more
20 sensitive and quantitative approach using affinity purification combined with stable isotope
21 labeling of amino acids in cell culture (SILAC). In total 1301 interactors were identified by SILAC in
22 57 experiments (Supplementary table 4). Third, we also obtained direct protein-protein interaction
23 data from several independent yeast two-hybrid (Y2H) screens against cDNA libraries derived from
24 hTERT-RPE1 (retinal pigment epithelial) cells, as well as brain, kidney, retina and testis tissue. In
25 total 69 Y2H screens were performed using 27 baits, identifying a total of 343 interacting proteins
26 (Supplementary table 5).

27 The SILAC and Y2H studies were focused on finding new interactors for selected ciliary proteins of
28 interest and were not part of a systematic analysis. Parts of the resulting PPIs were published in
29 previous studies (four out of 16 baits for SILAC^{47,52,53} and nine out of 27 baits for Y2H⁵³⁻⁶¹), however
30 here we consider the entire PPI data sets. The complete data sets are publicly available in
31 Supplementary tables 4 & 5 and at the IntAct database⁶² under references {DB reference 1} (SILAC)
32 and {DB reference 2} (Y2H) (note to reviewers: datasets will be submitted to IntAct before
33 publication). Because the TAP-MS and SILAC data sets are based on similar methodology and have a
34 large bait overlap (14 out of 16 SILAC baits were used in the TAP-MS data set), we merged them
35 into a single data set (Mass-spec based PPI) with 4799 unique proteins identified to interact with
36 184 bait proteins.

1 The Y2H, SILAC and TAP-MS data were transformed to genome wide data sets by defining genes as
2 1 (“found”) when the gene product was found to interact with the baits, and as 0 (“not found”)
3 when the gene product was not found to interact. Due to the large overlap in baits and the largely
4 similar methods used in the TAP-MS and SILAC data sets we decided to combine the data sets in
5 order to avoid counting the interacting proteins multiple times and thereby artificially
6 overestimating their CiliaCarta Scores. The mass-spectrometry data sets and Y2H data sets were
7 found to be sufficiently different to include them as separate data sets (positive set correlation is
8 0.13, Supplementary Fig. 10).

9 **Expression screen data set**

10 In expression screening⁶³ separate gene-expression data sets are weighted for their potential to
11 predict new genes for a system by measuring, per data set, the level of co-expression of the known
12 genes. We have already successfully applied this method to predict TMEM107 as part of the ciliary
13 transition zone ⁶⁴ and now extend the approach to the complete cilium. An integrated cilium co-
14 expression data set was constructed by applying the weighted co-expression method WeGet⁶⁵ to
15 ciliary genes in 465 human expression data sets available in the NCBI Gene Expression Omnibus⁶⁶.
16 For individual genes, correlations of their expression profiles were determined with expression
17 profiles of the set of ciliary components. The contribution of each data set to a final co-expression
18 score per gene was weighed by how consistently the set of cilia components were expressed
19 together, i.e. how well the data set in question is able to detect ciliary components⁶⁵. To avoid
20 circularity with the training of the Bayesian classifier the expression screen was performed using a
21 gene set of ciliary components from GO (GO:Cilium) and removed any overlap from the positive set
22 used to evaluate this data set for the Bayesian classifier. The data set from Ross *et al.*²⁰, which has
23 been included in the Bayesian classifier, has been excluded from the microarray data sets used in
24 the expression screen.

25 **Ciliary co-evolution data set**

26 Given the large number of independent losses of the cilium in eukaryotic evolution (we counted
27 eight independent loss events throughout the eukaryotic kingdom)⁶⁷⁻⁶⁹, presence/absence profiles
28 have a high value for predicting new cilium genes^{12,64}. We constructed a comprehensive co-
29 evolution data set from a comparative genomics analysis of presence-absence correlation patterns
30 over a representative data set of eukaryotic species. We correlated the occurrence of orthologs of
31 22,000 human genes in 52 eukaryotic genomes to that of cilia or flagella using differential Dollo
32 parsimony (DDP)⁷⁰. A perfectly matched profile pair would obtain a DDP of 0 (all events match, no
33 differences), while mismatching profile pairs would receive a DDP equal to the number of
34 evolutionary events that did not occur at the same time in evolution (e.g. the gene was lost in a
35 lineage still maintaining a cilium, or the gene was maintained in a lineage in which the cilium has
36 been lost). Thus we obtained an objective measure for each human gene that describes how well its

1 evolutionary trajectory (i.e. point of origin and independent loss events) matches that of the ciliary
2 system. A number of mismatches are expected for some ciliary genes; *Plasmodium falciparum* for
3 instance has maintained a cilium, but has lost all genes of the IFT machinery. Due to the topology of
4 the species tree we observed a complicated distribution of genes from the positive and negative
5 sets (Supplementary Fig. 11a): for low DDP scores (0-6) we did not observe a single negative gene,
6 which would result in unrealistic log odd scores (i.e. infinity). To avoid these unrealistic log odds,
7 we decided to combine the DDP scores into two categories, namely genes with a $DDP \leq 9$ and genes
8 with a $DDP \geq 10$. Genes with a score between 0 and 9 were generally overrepresented among ciliary
9 genes (Supplementary Fig. 11b).

10 Transcription factor binding sites data set

11 The RFX and FOXJ1 transcription factors play an important role in the regulation of ciliogenesis^{71,72}.
12 X-box (RFX) or a FOXJ1 transcription factor binding site (TFBS) have been used to predict novel
13 ciliary genes in *Caenorhabditis elegans* (nematode) and *Drosophila melanogaster* (fruit fly)⁷³⁻⁷⁵. We
14 processed the publicly available data sets from the 29 mammals project¹⁸ to obtain human genes
15 with a conserved X-box or FOXJ1 TFBS in their promoters, which were defined as 4 kilobase (kb)
16 windows centered (i.e. 2kb upstream and 2kb downstream) at all annotated transcription start
17 sites of the gene. The restriction that the TFBS motifs are conserved among mammalian species
18 infers a higher level of confidence that these motifs are indeed relevant and not spurious hits. The
19 final data set was constructed by defining two categories, namely: “Gene has a X-box and/or FOXJ1
20 TFBS”, represented as 1, and “Gene does not have a X-box and/or FOXJ1 TFBS”, represented as 0.
21 We found relatively limited overlap with the previous invertebrate X-box TFBS data sets: 13% of
22 the genes with an X-box in *C. elegans* (225 out of 1695 genes)⁷³⁻⁷⁵ and 15% in *D. melanogaster* (71
23 out of 470)⁷⁵ have a conserved X-box in human. This low overlap may result from differences in the
24 sensitivity detecting functional X-box sequences. It might also indicate that the X-box motifs are
25 transient in the genome; i.e. often gained and lost, as has also been observed in vertebrate evolution
26 of other TF binding sites⁷⁶.

27 Published data sets

28 There are a number of high-throughput cilia data sets available from the CilDB¹⁵ that can
29 complement the data sets mentioned above. We only included data sets from mammalian species to
30 avoid significant issues with orthology (i.e. avoid mapping to paralogous genes) and which had a
31 broad coverage of the entire genome/proteome (Supplementary fig. 1). We excluded data sets that
32 focused specifically on the centriole/basal body, since this structure is also affected by the cell
33 cycle^{77,78} and therefore could potentially skew the Bayesian classifier towards this process. We
34 avoided redundancy in the data sets by selecting only one data set per experiment type (e.g.
35 proteomics, expression). We only considered proteomics datasets specifically generated for the
36 cilium as opposed to whole cell proteomics data sets to minimize false positives. The data from

1 Ross *et al.*²⁰ (expression) and Liu *et al.*¹⁹ (proteomics) were selected based on these criteria. These
2 data sets were extracted from CilDB¹⁵ and implemented using the predefined confidence categories
3 from CilDB (Low confidence, medium confidence, high confidence). Genes not covered by these data
4 sets were assigned to the “not found” category.

5 **Training sets**

6 We used the SYSCILIA Gold Standard (SCGS)⁵ as our positive training set. We did not use GO
7 annotations as we regard the SCGS, that has been annotated by experts in the cilium field, of higher
8 quality Furthermore, having an independent cilium genes dataset allowed us to prevent circularity
9 in the expression screening analysis (see below). We are currently in the process of improving cilia-
10 related GO terminology and transferring our SCGS annotations⁷⁹. The SCGS, or positive set, contains
11 a total of 302 manually curated human ciliary genes. We constructed a negative set by selecting
12 genes annotated in GO to function in processes and cellular compartments we deemed least likely
13 to be (in)directly involved in ciliary processes. We selected genes annotated with at least one of the
14 following GO Cellular Component terms: extracellular, lysosome, endosome, peroxisome, ribosome,
15 and nucleolus. We ensured that the positive and negative training sets do not overlap by removing
16 genes found in both from the negative set. Since the majority of human genes are expected to be
17 non-ciliary the similarity between the score distribution of the negative set and the remaining
18 “other genes” indicates that the negative set overall gives an excellent representation of what we
19 reasonably can expect to be non-ciliary genes. The final negative set contains 1275 genes.

20 We adapted the positive training set for the PPI data sets as well as the expression screen data set
21 to avoid overtraining. Since many of the baits used in the PPI data are known ciliary components
22 and therefore part of the positive training set, this could lead to a potential overestimation of the
23 predictive value of the data sets. Therefore we excluded the bait proteins from the positive set for
24 evaluating the Y2H, TAP-MS and SILAC data. The training of the expression screen was performed
25 using a gene set of ciliary components from GO (GO:Cilium) and to avoid overtraining we
26 subtracted the overlap from the positive set used for the Bayesian classifier. In this way we avoided
27 inflation of the predictive values for these data sets.

28 **Bayesian classifier**

29 *Performance and false discovery rate calculations*

30 The performance of each data set for predicting ciliary genes as well as the integrated Bayesian
31 classifier was evaluated using the positive and negative training sets. We determined the fraction,
32 or recall, of the positive training set, which is also known as the sensitivity or true positive rate
33 (TPR).

1
$$\text{Sensitivity (SN)} = \text{True Positive Rate (TPR)} = \frac{TP}{TP + FN}$$

2 Where True Positives (TP) are true ciliary genes that are correctly discovered by the data set, and
3 False Negatives (FN) are true ciliary genes that were not discovered by the data set. We also
4 determined the fraction of the negative set retrieved by the data set, also known as the false
5 positive rate (FPR).

6
$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}$$

7 Where True Negatives (TN) are non-ciliary genes that are correctly excluded from the data set, and
8 False Positives (FP) are non-ciliary genes incorrectly discovered in the data set. The FPR together
9 with the TPR are used to calculate the predictive ability for each data set (see below).

10 The FPR is related to the specificity (SP), another well-known metric for the quality of a data set, as
11 follows:

12
$$\text{Specificity (SP)} = \frac{TN}{FP + TN} = 1 - TPR$$

13 The false discovery rate (FDR) denotes the chance of encountering a false positive among a set of
14 predictions and thus reflects the trustworthiness of a prediction. The FDR is given by:

15
16
$$\text{False Discovery Rate (FDR)} = \frac{FP}{TP + FP} = \frac{1 - SP}{1 - SP + SN}$$

17 We use the FDR to determine the overall performance of the classifier given a specified threshold of
18 the CiliaCarta Score. However, the FDR depends on both training sets and is sensitive to deviations
19 in set size compared to the actual populations of positives and negatives for the whole genome, and
20 thus we need to correct the canonical FDR equation for the differences in the population and set
21 size:

22
$$FDR_{corrected} = \frac{1 - SP}{1 - SP + SN \cdot O_{prior}}$$

23 Throughout the text we refer to this adjusted FDR as the corrected FDR (cFDR). The positive
24 predictive value (PPV) is directly related to the FDR and reflects the probability that a gene within
25 the set threshold will indeed be ciliary:

26
$$\text{Positive Predictive Value (PPV)} = 1 - FDR$$

1 The PPV and FDR therefore reflect the chance of success or failure within a set of genes rather than
2 for individual genes; i.e. they do not reflect the per gene probability for it being ciliary or not.
3 Instead, the per-gene probabilities are reflected by the posterior odds, the CiliaCarta Score, see
4 below.

5
6 At a cFDR cut-off of 25% our rank-ordered list of predictions covers 404 genes. However, this set
7 contains genes from both the negative and the positive training sets, and after excluding those we
8 obtain 285 predictions. This remaining set will however have a different FDR since we removed the
9 training set genes. We obtain the FDR for the remaining 285 genes by: (i) calculating the number of
10 TP and FP based on the cFDR threshold, (ii) subtracting the number of genes from the training sets
11 (TP - genes in positive set, FP - genes in negative set) and (iii) recalculating the FDR using these
12 adjusted TP and FP values. This results in a FDR of 33% for the 285 candidate predictions.

13
14 In our experimental validations 24 out of 36 genes are positive for ciliary phenotype and/or
15 localization. We can derive an observed FDR directly from these numbers, i.e. 24 TP and 12 FP. The
16 observed FDR is thus $12/36 = 33\%$, the same as the estimated FDR for our candidate genes.

17 *Calculation of the CiliaCarta Score using naive Bayesian integration*

18 Naive Bayesian integration allows a direct comparison and weighing of many and diverse data sets
19 describing the properties of ciliary genes and integrates these into a single probabilistic score for
20 each gene accommodating for missing data^{7,8,80,81}. This approach has been successfully used to
21 predict for instance mitochondrial⁸¹ and innate immunity⁸ genes.

22 For a given gene in the human genome we calculate the conditional probability that the gene is
23 involved in ciliary processes given the observed evidence in the data sets. For our purposes, since
24 we have only two possible outcomes (i.e. ciliary vs. non-ciliary) it is more convenient and
25 appropriate to use odds instead. We can write the probability that a gene is ciliary given the
26 outcome of the experiment in data set i (D_i) for all data sets j :

$$27 \frac{P(\text{ciliary gene}|D_{1..j})}{P(\text{non-ciliary gene}|D_{1..j})} = \prod_{i=1}^j \frac{P(\text{ciliary gene}|D_i)}{P(\text{non-ciliary gene}|D_i)}$$

28 These odds cannot be calculated directly, but we can obtain them using Bayes' theorem by
29 approximating the reverse likelihood ratio L that a gene is observed in the data sets given it is
30 either ciliary, or non-ciliary:

$$31 L(D_{1..j}) = \prod_{i=1}^j \frac{P(D_i|\text{ciliary gene})}{P(D_i|\text{non-ciliary gene})}$$

1 We can calculate this likelihood directly from the distribution of the training sets. This equates to
2 the ratio between the data sets' true positive rate for ciliary genes (i.e. the proportion of known
3 ciliary genes retrieved), and the false positive rate for non-ciliary genes (i.e. the proportion of
4 known non-ciliary genes retrieved). Using Bayes' theorem we can now obtain the final (posterior)
5 odds from the likelihood ratio L as follows:

$$6 \frac{P(\text{ciliary gene} | D_{1...j})}{P(\text{non-ciliary gene} | D_{1...j})} = O_{\text{prior}} \cdot L(D_{1...j}) = O_{\text{posterior}}$$

7 Where O_{prior} is the prior odd, i.e. the odds for a gene to be ciliary if one would randomly sample
8 from the genome:

$$9 O_{\text{prior}} = \frac{P(\text{ciliary gene})}{P(\text{non-ciliary gene})}$$

10 And thus we obtain the posterior probability for data sets $D_{i,j}$:

$$11 O_{\text{posterior}} = \frac{P(\text{ciliary gene})}{P(\text{non-ciliary gene})} \cdot \prod_{i=1}^j \frac{P(D_i | \text{ciliary gene})}{P(D_i | \text{non-ciliary gene})}$$

12 Finally we obtain the CiliaCarta Score by \log_2 transformation of the individual terms to get an
13 additive score:

$$14 \text{CiliaCarta Score} = \log_2 \left(\frac{P(\text{ciliary gene})}{P(\text{non-ciliary gene})} \right) + \sum_{i=1}^j \log_2 \left(\frac{P(D_i | \text{ciliary gene})}{P(D_i | \text{non-ciliary gene})} \right)$$

15 The additive nature of the log transformed posterior odds equation makes the contribution of each
16 data set to the final score more insightful. It also makes the score robust against rounding errors
17 otherwise encountered during multiplication of the untransformed odds. The log-odds for each
18 individual data set per sub-category are listed in Supplementary table 7.

19 *Determining the prior*

20 One of the elements in Bayesian calculations is estimating the prior: the *a priori* expectation of how
21 many ciliary genes there are in the human genome. Although the ranking of genes does not depend
22 on the prior, it is required to obtain a corrected false discovery rate (see above). Furthermore it
23 gives a meaning to the posterior log odd scores (the CiliaCarta score), i.e. a positive log odd means
24 that the gene is more likely to be ciliary than non-ciliary and a negative log odd means that it is
25 more likely to be non-ciliary.

26 To our knowledge there is no substantiated estimate for the number of genes involved in the cilium.
27 Currently 608 human proteins have been annotated as being part of the cilium in a combination of
28 GO (GO:0005929 Cell Component Cilium & GO:0042384 Biological Process Cilium Assembly,

1 Ensembl biomart as of December 3rd 2015) and the SCGS, but we can reasonably assume the total
2 number of ciliary genes to be much higher. For instance, the ciliary proteome as identified by Liu *et*
3 *al.* in the mouse photoreceptor sensory cilium entails 1185 to 1968 proteins, depending on the
4 stringency of the filters applied¹⁹. We have chosen a prior that we deem to be both reasonable and
5 conservative: 5% of the human genome (i.e. 1135 ciliary genes). The O_{prior} then becomes:

$$6 \quad O_{prior} = \frac{P(\text{ciliary gene})}{1 - P(\text{ciliary gene})} = \frac{0.05}{1 - 0.05} \approx 0.0526$$

7 *Conditional independence*

8 An assumption of our naïve Bayesian approach is that the outcome of one data set is independent of
9 the outcome of another. This assumption of independence is not always attainable, since for
10 instance gene expression is to some extent biologically correlated to the presence of proteins in the
11 proteomics data sets. Violations of the independence assumption can bias the predictions and can
12 lead to an overestimation of the likelihood scores. However previous work has shown that,
13 regardless of biological correlations between data sets, naïve Bayesian integration of genomics data
14 is highly effective to predict novel genes involved in a molecular system^{7,81}. Analysis of the
15 correlations suggests that the data sets used to predict ciliary genes are largely complementary
16 (Supplementary Fig. 10). Several data sets have high correlations, such as ciliary co-evolution and
17 co-expression. However these data sets are methodologically and experimentally completely
18 unrelated and thus the high correlation is purely based on the ability of the methods to predict
19 ciliary genes.

20 *Ciliome size estimation*

21 The Bayesian framework can be used to obtain a systematic estimate for the total number of ciliary
22 genes in two ways. The first approach involves fitting a new O_{prior} based on the observed validation
23 rate of our experiments; that is, we make use of the discrepancy between the expected and
24 experimentally determined number of hits. We can estimate this new O_{prior} by equating the cFDR
25 of the Bayesian integration, which depends on our original O_{prior} to the observed FDR of the
26 validation experiments:

$$27 \quad FDR_{observed} = FDR_{corrected}$$

28 Where $FDR_{observed}$ is calculated from the TP and FP determined from the validation experiments
29 (24 and 12 resp.). The SP and SN for the cFDR can be obtained from the training sets. We then try
30 to find an $O_{estimated}$ for which the following equation holds:

$$31 \quad \frac{FP_{obs}}{TP_{obs} + FP_{obs}} = \frac{1 - SP_{training}}{1 - SP_{training} + SN_{training} \cdot O_{estimated}}$$

1 Atypically, in our study the observed FDR and the cFDR are equal (i.e. 33%), which indicates that
2 $O_{estimated}$ essentially equals our initially chosen O_{prior} (1135 ciliary genes).

3 The second approach to estimate the ciliome size is based on the Bayesian posterior probabilities
4 obtained for each gene (the CiliaCarta Scores). To determine the expected value (i.e. the number of
5 true positives) among a set of ciliary candidates, we considered each gene to be a random variable
6 with a binary outcome (success, a true ciliary gene, or failure, not a ciliary gene) whose probability
7 of success is defined by $P(ciliary\ gene|D_1D_i)$. Effectively this corresponds to a Bernoulli process
8 with different probabilities for success or failure for each successive trial. Assuming independence,
9 the expected value E for a set of n binary random variables (i.e. ciliary candidates) is equal to the
10 sum of the expected values of the individual variables. The expected value for an individual binary
11 random variable, in turn, equals its probability of success. From the posterior odds:

$$12 \quad O_{posterior} = \frac{P(ciliary\ gene|D_{1...j})}{P(non\ ciliary\ gene|D_{1...j})} = \frac{P(ciliary\ gene|D_{1...j})}{1 - P(ciliary\ gene|D_{1...j})}$$

13 We can obtain the probability that a gene is ciliary by rewriting the above as:

$$14 \quad P(ciliary\ gene|D_{1...j}) = \frac{O_{posterior}}{1 + O_{posterior}}$$

15 Since the CiliaCarta Score (CCS) is the \log_2 of the posterior odds we finally get:

$$16 \quad P(ciliary\ gene|D_1...D_j) = \frac{2^{CCS}}{2^{CCS} + 1}$$

17 And thus to obtain the expected value E becomes:

$$18 \quad E_{ciliary\ genes} = \sum_{i=1}^n \frac{2^{CCS_i}}{2^{CCS_i} + 1}$$

19 The expected value for the number of ciliary genes based on the Bayesian posterior probabilities
20 was found to be 1273. It should be noted that the posterior CiliaCarta probabilities, and hence the
21 expected value for a set of genes, depend on the prior estimation of the number of ciliary genes. We
22 have above already established that our chosen prior was accurate based on our validation
23 outcome. Indeed the posterior expected number of ciliary genes is close to the prior expected
24 number of ciliary genes (1135, difference of 138). Averaging these two estimates we arrive at a
25 total number of expected ciliary genes of approximately 1200 genes.

26 **Validation and OSCP1 experiments**

27 *Candidate selection*

28 The number of genes tested per method, the model organisms and cell-lines were determined by
29 time and resources available to the participating labs. Orthologs in *C. elegans* were identified for the

1 candidate list. For 21 orthologs null alleles were available from the *Caenorhabditis* Genetics Center
2 (University of Minnesota, USA). All 21 null mutants were obtained and tested. Candidates for the
3 eCFP localization studies in hTERT-RPE1 cells were randomly selected. The candidates for the
4 immunofluorescence in human lung epithelial and murine retina were selected based on the
5 availability of a suitable antibody in the collaborating labs. Candidates tested in zebrafish were
6 randomly selected based on the presence of unambiguous 1-1 orthologs. An equal spread of the
7 selected candidates throughout the ranked list was taken into account as much as possible as is
8 shown in Fig. 4h.

9 *Localization studies in hTERT-RPE1 cells*

10 Expression constructs were created with Gateway Technology (Life Technologies) according to the
11 manufacturer's instructions. These constructs encoded eCFP fusion proteins of OSCP1 (transcript
12 variant 1; NM_145047.4), CFAP61 (C20orf26, NM_015585), CFAP20 (C16orf80, NM_013242.2),
13 CYB5D1 (NM_144607), CCDC147 (M_001008723.1), CFAP161 (C15orf26, NM_173528), IQCA1
14 (NM_024726.4), IPO5 (NM_002271), HSPA1L (NM_005527), and TEKT1 (NM_053285). The
15 sequences for all entry clones were verified by Sanger sequencing. Human TERT-immortalized
16 retinal pigment epithelium 1 (hTERT- RPE1) cells were cultured as previously described⁸². Cells
17 were seeded on coverslips, grown to 80% confluency, and subsequently serum starved for 24 hr in
18 medium containing only 0.2% foetal calf serum for inducing cilium growth. The cells were then
19 transfected with eCFP expression construct using Lipofectamine 2000 (Life Technologies)
20 according to the manufacturer's instructions. Cells were fixed in 4% paraformaldehyde for 20 min,
21 treated with 1% Triton X-100 in PBS for 5 min, and blocked in 2% BSA in PBS for 20 min. Cells were
22 incubated with the primary antibody GT335 (cilium and basal body marker, 1:500) diluted in 2%
23 BSA in PBS, for 1 hr. After washing in PBS, the cells were incubated with the secondary antibody for
24 45 min. Secondary antibody, goat anti-mouse Alexa 568 (1:500; Life Technologies) was diluted in
25 2% BSA in PBS. Cells were washed with PBS and briefly with milliQ before being mounted in
26 Vectashield containing DAPI (Vector Laboratories). The cellular localization of eCFP-fused proteins
27 was analyzed with a Zeiss Axio Imager Z1 fluorescence microscope equipped with a 63x objective
28 lens. Optical sections were generated through structured illumination by the insertion of an
29 ApoTome slider into the illumination path and subsequent processing with AxioVision (Zeiss) and
30 Photoshop CS6 (Adobe Systems) software.

31 *Immunofluorescence microscopy of cells*

32 IMCD3 and ATDC5 cells were growth to 80% confluence in DMEM-Glutamax medium with 10%
33 Foetal Bovine Serum. Then cells were Serum-starved for 24 hours and fixed in cold methanol for 5
34 minutes, PBS washed and blocked with 1% Bovine Serum Albumin for 1 hour before incubating
35 with primary antibodies overnight at room temperature. Antibodies and concentrations were anti-
36 acetylated -tubulin (Sigma 6-11B-1, T7451) 1/200, anti-gamma-tubulin (Sigma GTU-88, T6557)

1 1/200, anti-OSCP1 Proteintech 12598-1-AP 1/100, anti-OSCP1 ATLAS HPA028436 1/100 and anti-
2 OSCP1 Biorbyt 185681 1/100. Human respiratory cells were analyzed by immunofluorescence
3 microscopy as previously described⁸³. The following rabbit polyclonal antibodies were purchased
4 from Atlas antibodies: anti-CCDC113 (HPA040869), anti-RIBC2 (HPA003210), anti-ARMC3
5 (HPA037824), anti-C6orf165 (HPA044891) and anti-OSCP1 (HPA028436).

6 *Mouse handling and experiments*

7 C57Bl/6J wild-type mice were kept on a 12 h light-dark cycle with unlimited access to food and
8 water. All procedures were in accordance with the guidelines set by the ARVO statement for the use
9 of animals in Ophthalmic and Vision Research and the local laws on animal protection. The
10 following antibodies were used for immunofluorescence/immune-EM analysis of murine retina
11 sections: rabbit anti-CCDC113 (1:500/1:500), anti-RIBC2 (1:250/1:250), anti-EFHC1 (1:500/-),
12 anti-WDR69 (1:250/-), C6orf165 (1:50/1:200), anti-ARMC3 (1:250/1:500), anti-OSCP1 (for IF
13 1:100, biorbyt, Cambridge, UK; 1:100, proteintech, Manchester, UK; for EM: 1:100, Atlas, Stockholm,
14 Sweden), mouse anti-centrin-3 (1:100;⁸⁴), rabbit anti-rootletin (1:100, ⁸⁵). Sections were
15 counterstained with DAPI (1 mg/ml) Sigma-Aldrich, Munich, Germany), and, where applicable, with
16 FITC-labeled peanut agglutinin (PNA, 1:400, Sigma-Aldrich, Munich, Germany). Secondary
17 antibodies conjugated to Alexa 488®, Alexa 555®, and Alexa 568® (1:400) were purchased from
18 Invitrogen (Karlsruhe, Germany) and CF™-640 (1:400) from Biotrend Chemikalien GmbH (Cologne,
19 Germany). For pre-embedding electron microscopy, we used biotinylated secondary antibodies
20 (1:150; Vector Laboratories, Burlingame, CA, USA). For immunofluorescence microscopy, the eyes
21 of adult C57Bl/6J mice were cryofixed sectioned, and immunostained as described previously ⁸⁶.
22 We double-stained the cryosections for CCDC113, RIBC2, EFHC1, WDR69, C6orf165, ARMC3,
23 OSCP1, and centrin-3 as a molecular marker for the connecting cilium, the basal body, and the
24 adjacent centriole of photoreceptor cells {Trojan et al. 2008} at 4°C overnight. Sections stained for
25 C6orf165 were counterstained with FITC-labeled cone photoreceptor marker peanut agglutinin
26 (PNA⁸⁷⁻⁸⁹). After one-hour incubation at room temperature with the according secondary
27 antibodies and the nuclear marker DAPI, sections were mounted in Mowiol 4.88 (Hoechst,
28 Frankfurt, Germany). Images were obtained and deconvoluted with a Leica LEITZ DM6000B
29 microscope (Leica, Wetzlar, Germany) and processed with Adobe Photoshop CS with respect to
30 contrast and color correction as well as bicubic pixel interpolation. We applied a pre-embedding
31 labeling protocol as previously introduced for immunoelectron microscopy of mouse photoreceptor
32 cells ⁹⁰⁻⁹². Ultrathin sections were analyzed with a transmission electron microscope (TEM) (Tecnai
33 12 BioTwin; FEI, Eindhoven, The Netherlands). Images were obtained with a charge-coupled device
34 camera (SIS MegaView3, Olympus, Shinjuka, Japan) and processed with Adobe Photoshop CS
35 (brightness and contrast).

1 *Zebrafish handling and experiments*

2 Wild-type (AB × Tup LF) zebrafish were maintained and staged as described previously in⁹³.
3 Antisense MO oligonucleotides (Gene Tools) were designed against the start codons and against
4 splice sites, as described in Supplementary table 6. MOs were injected (4–6 ng) into embryos at the
5 1- to 2-cell stage and reared at 28.5°C until the desired stage. For cilia immunostaining, 6 somite-
6 stage or 24 hpf (hours post fertilization) embryos were dechorionated and fixed in 4% PFA
7 overnight (O/N) at 4°C, dehydrated through 25%, 50% and 75%, methanol/PBT (1% Triton X-100
8 in PBS) washes and stored in 100% methanol –20°C. The embryos were rehydrated again through
9 75%, 50% and 25% methanol/PBT washes. Embryos of 24 hpf were permeabilized with Proteinase
10 K (10ug/ml in PBT) for 10 minutes at 37°C, and subsequently refixed in 4% PFA. Prior to
11 immunostaining, embryos were incubated in block buffer (5% goat serum in PBT) blocked with 5%
12 goat serum (in PBT) for 1 h and subsequently incubated O/N at 4°C with mouse monoclonal anti- γ -
13 tubulin (1:200, GTU-88, Sigma) and anti-acetylated tubulin (1:800, 6-11B-1, Sigma) diluted in
14 blocking buffer. Secondary antibodies used were Alexa Fluor goat anti-mouse IgG1 488, Alexa Fluor
15 donkey anti-mouse IgG2b 568, and Alexa Fluor goat anti-mouse IgG2b 594 (Molecular Probes).
16 Nuclei were stained with Hoechst and embryos were mounted in Citofluor. Z-stack images were
17 captured using a Zeiss 710 Confocal Microscope. For rescue experiments, the aforementioned
18 human OSCP1 cDNA was cloned into pCS2+ using gateway technology. OSCP1 plasmids were
19 linearized using NotI and mRNA was synthesized using Ambion mMessage mMachine kit for the
20 sense strand. 100 pg of mRNA was injected into the cell of one cell-stage embryos. These embryos
21 were subsequently injected with 4 ng *oscp1* MO at the two-cell stage, and embryos were allowed to
22 develop at 28.5°C.

23 *C. elegans handling and experiments*

24 *C. elegans* were maintained and cultured at 20°C using standard techniques. Mutant strains were
25 obtained from the *Caenorhabditis* Genetics Center (University of Minnesota, USA); the alleles used
26 are shown in Supplementary table 2. Assays for dye uptake (DiI), roaming and osmotic avoidance
27 were performed as previously described²³. Briefly, for the dye-filling assay, worms were placed into
28 a DiI solution (diluted 1:200 with M9 buffer) for 1 hour, allowed to recover on NGM plates, and then
29 imaged (40x objective, Texas Red filter set) on a compound epi-fluorescence microscope (Leica
30 DM5000b), fitted with an Andor EMCCD camera. For the osmotic avoidance assay, young adult
31 worms were placed within a ring-shaped barrier of 8M glycerol and scored during 10 minutes for
32 worms that crossed the barrier. For the roaming assay, single young adult worms were placed for
33 16 hours onto seeded plates and track coverage assessed using a grid reference. For transmission
34 electron microscopy, *oscp-1(gk699)* worms were first backcrossed 2 times with wild type worms
35 (to remove unlinked mutations), using primers that flank the *gk699* deletion. Day 1 adults were
36 fixed, sectioned and imaged as described previously²³. Translational reporters were introduced into

1 *oscp-1(gk699)* by standard mating methods. The translational construct for *oscp-1* (R10F2.5 and
2 R10F2.4) was generated by fusing the genomic region, including 853 bp of the native promoter, to
3 GFP with the *unc-54* 3' UTR. Standard mating procedures were used to introduce OSCP-1::GFP into
4 the *mks-5(tm3100)* mutant background.

5 References

- 6 1. Yuan, S. & Sun, Z. Expanding horizons: ciliary proteins reach beyond cilia. *Annu. Rev. Genet.*
7 **47**, 353–76 (2013).
- 8 2. Tyler, K. M. *et al.* Flagellar membrane localization via association with lipid rafts. *J. Cell Sci.*
9 **122**, 859–66 (2009).
- 10 3. Takao, D. & Verhey, K. J. Gated entry into the ciliary compartment. *Cell. Mol. Life Sci.* (2015).
11 doi:10.1007/s00018-015-2058-0
- 12 4. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology
13 Consortium. *Nat. Genet.* **25**, 25–9 (2000).
- 14 5. van Dam, T. J., Wheway, G., Slaats, G. G., Huynen, M. A. & Giles, R. H. The SYSCILIA gold
15 standard (SCGSv1) of known ciliary components and its applications within a systems
16 biology consortium. *Cilia* **2**, 7 (2013).
- 17 6. Josic, D. & Clifton, J. G. Mammalian plasma membrane proteomics. *Proteomics* **7**, 3010–29
18 (2007).
- 19 7. Tabach, Y. *et al.* Identification of small RNA pathway genes using patterns of phylogenetic
20 conservation and divergence. *Nature* **493**, 694–8 (2013).
- 21 8. van der Lee, R. *et al.* Integrative Genomics-Based Discovery of Novel Regulators of the Innate
22 Antiviral Response. *PLOS Comput. Biol.* **11**, e1004553 (2015).
- 23 9. Pagliarini, D. J. *et al.* A mitochondrial protein compendium elucidates complex I disease
24 biology. *Cell* **134**, 112–23 (2008).
- 25 10. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian
26 mitochondrial proteins. *Nucleic Acids Res.* gkv1003 (2015). doi:10.1093/nar/gkv1003
- 27 11. Avidor-Reiss, T. *et al.* Decoding Cilia Function Defining Specialized Genes Required for
28 Compartmentalized Cilia Biogenesis. *Cell* **117**, 527–539 (2004).
- 29 12. Li, J. B. *et al.* Comparative Genomics Identifies a Flagellar and Basal Body Proteome that
30 Includes the BBS5 Human Disease Gene. *Cell* **117**, 541–552 (2004).
- 31 13. Piasecki, B. P., Burghoorn, J. & Swoboda, P. Regulatory Factor X (RFX)-mediated
32 transcriptional rewiring of ciliary genes in animals. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12969–
33 74 (2010).
-

- 1 14. Ivliev, A. E., 't Hoen, P. A. C., van Roon-Mom, W. M. C., Peters, D. J. M. & Sergeeva, M. G.
2 Exploring the Transcriptome of Ciliated Cells Using In Silico Dissection of Human Tissues.
3 *PLoS One* **7**, e35618 (2012).
- 4 15. Arnaiz, O. *et al.* Cildb: a knowledgebase for centrosomes and cilia. *Database* **2009**, bap022
5 (2009).
- 6 16. Touw, W. G. *et al.* Data mining in the Life Sciences with Random Forest: a walk in the park or
7 lost in the jungle? *Brief. Bioinform.* **14**, 315–26 (2013).
- 8 17. Boldt, K. *et al.* An organelle-specific protein landscape identifies novel diseases and
9 molecular mechanisms. *Nat. Commun.* **7**, 11491 (2016).
- 10 18. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29
11 mammals. *Nature* **478**, 476–82 (2011).
- 12 19. Liu, Q. *et al.* The proteome of the mouse photoreceptor sensory cilium complex. *Mol. Cell.*
13 *Proteomics* **6**, 1299–317 (2007).
- 14 20. Ross, A. J., Dailey, L. A., Brighton, L. E. & Devlin, R. B. Transcriptional profiling of mucociliary
15 differentiation in human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **37**, 169–85
16 (2007).
- 17 21. Inglis, P. N., Ou, G., Leroux, M. R. & Scholey, J. M. The sensory cilia of *Caenorhabditis elegans*.
18 *WormBook* 1–22 (2007). doi:10.1895/wormbook.1.126.2
- 19 22. Starich, T. A. *et al.* Mutations affecting the chemosensory neurons of *Caenorhabditis elegans*.
20 *Genetics* **139**, 171–188 (1995).
- 21 23. Sanders, A. A. W. M., Kennedy, J. & Blacque, O. E. Image analysis of *Caenorhabditis elegans*
22 ciliary transition zone structure, ultrastructure, molecular composition, and function.
23 *Methods Cell Biol.* **127**, 323–47 (2015).
- 24 24. Coon, B. G. *et al.* The Lowe syndrome protein OCRL1 is involved in primary cilia assembly.
25 *Hum. Mol. Genet.* **21**, 1835–47 (2012).
- 26 25. Hernandez-Hernandez, V. *et al.* Bardet-Biedl syndrome proteins control the cilia length
27 through regulation of actin polymerization. *Hum. Mol. Genet.* **22**, 3858–68 (2013).
- 28 26. Chung, M.-I. *et al.* Coordinated genomic control of ciliogenesis and cell movement by RFX2.
29 *Elife* **3**, e01439 (2014).
- 30 27. Firat-Karalar, E. N., Sante, J., Elliott, S. & Stearns, T. Proteomic analysis of mammalian sperm
31 cells identifies new components of the centrosome. *J. Cell Sci.* **127**, 4128–33 (2014).
- 32 28. Nie, X. *et al.* Cloning, expression, and mutation analysis of NOR1, a novel human gene down-
33 regulated in HNE1 nasopharyngeal carcinoma cell line. *J. Cancer Res. Clin. Oncol.* **129**, 410–4
34 (2003).
-

- 1 29. Huu, N. T., Yoshida, H. & Yamaguchi, M. Tumor suppressor gene OSCP1/NOR1 regulates
2 apoptosis, proliferation, differentiation, and ROS generation during eye development of
3 *Drosophila melanogaster*. *FEBS J.* **282**, 4727–46 (2015).
- 4 30. Shan, Z. *et al.* Overexpression of oxidored-nitro domain containing protein 1 induces growth
5 inhibition and apoptosis in human prostate cancer PC3 cells. *Oncol. Rep.* **32**, 1939–46 (2014).
- 6 31. Izuno, H. *et al.* Rat organic solute carrier protein 1 (rOscp1) mediated the transport of
7 organic solutes in *Xenopus laevis* oocytes: isolation and pharmacological characterization of
8 rOscp1. *Life Sci.* **81**, 1183–92 (2007).
- 9 32. Kobayashi, Y. *et al.* Isolation and characterization of polyspecific mouse organic solute
10 carrier protein 1 (mOscp1). *Drug Metab. Dispos.* **35**, 1239–45 (2007).
- 11 33. Hiratsuka, K. *et al.* Intratesticular localization of the organic solute carrier protein, OSCP1, in
12 spermatogenic cells in mice. *Mol. Reprod. Dev.* **75**, 1495–504 (2008).
- 13 34. Huu, N. T., Yoshida, H., Umegawachi, T., Miyata, S. & Yamaguchi, M. Structural
14 characterization and subcellular localization of *Drosophila* organic solute carrier partner 1.
15 *BMC Biochem.* **15**, 11 (2014).
- 16 35. Hiratsuka, K. *et al.* Neuronal expression, cytosolic localization, and developmental regulation
17 of the organic solute carrier partner 1 in the mouse brain. *Histochem. Cell Biol.* **135**, 229–38
18 (2011).
- 19 36. Jensen, V. L. *et al.* Formation of the transition zone by Mks5/Rpgrip1L establishes a ciliary
20 zone of exclusion (CIZE) that compartmentalises ciliary signalling proteins and controls PIP2
21 ciliary abundance. *EMBO J.* **34**, 2537–56 (2015).
- 22 37. Gherman, A., Davis, E. E. & Katsanis, N. The ciliary proteome database: an integrated
23 community resource for the genetic and functional dissection of cilia. *Nat. Genet.* **38**, 961–2
24 (2006).
- 25 38. Chevrier, V. *et al.* OFIP/KIAA0753 forms a complex with OFD1 and FOR20 at pericentriolar
26 satellites and centrosomes and is mutated in one individual with Oral-Facial-Digital
27 Syndrome. *Hum. Mol. Genet.* (2015). doi:10.1093/hmg/ddv488
- 28 39. Slaats, G. G. *et al.* Screen-based identification and validation of four new ion channels as
29 regulators of renal ciliogenesis. *J. Cell Sci.* **128**, 4550–9 (2015).
- 30 40. Seeley, E. S. & Nachury, M. V. Constructing and deconstructing roles for the primary cilium in
31 tissue architecture and cancer. *Methods Cell Biol.* **94**, 299–313 (2009).
- 32 41. Michaud, E. J. & Yoder, B. K. The primary cilium in cell signaling and cancer. *Cancer Res.* **66**,
33 6463–7 (2006).
- 34 42. Basten, S. G. & Giles, R. H. Functional aspects of primary cilia in signaling, cell cycle and
35 tumorigenesis. *Cilia* **2**, 6 (2013).
-

- 1 43. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44**, D710-6 (2015).
- 2 44. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-64 (2002).
- 3 45. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).
- 4 46. Letteboer, S. J. F. & Roepman, R. Versatile screening for binary protein-protein interactions
5 by yeast two-hybrid mating. *Methods Mol. Biol.* **484**, 145-59 (2008).
- 6 47. Boldt, K. *et al.* Disruption of intraflagellar protein transport in photoreceptor cilia causes
7 Leber congenital amaurosis in humans and mice. *J. Clin. Invest.* **121**, 2169-80 (2011).
- 8 48. Gloeckner, C. J., Boldt, K. & Ueffing, M. Strep/FLAG tandem affinity purification (SF-TAP) to
9 study protein interactions. *Curr. Protoc. Protein Sci.* **Chapter 19**, Unit19.20 (2009).
- 10 49. Boldt, K., van Reeuwijk, J., Gloeckner, C. J., Ueffing, M. & Roepman, R. Tandem affinity
11 purification of ciliopathy-associated protein complexes. *Methods Cell Biol.* **91**, 143-60
12 (2009).
- 13 50. Olsen, J. V *et al.* Parts per million mass accuracy on an Orbitrap mass spectrometer via lock
14 mass injection into a C-trap. *Mol. Cell. Proteomics* **4**, 2010-21 (2005).
- 15 51. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-
16 range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-
17 72 (2008).
- 18 52. Cevik, S. *et al.* Active transport and diffusion barriers restrict Joubert Syndrome-associated
19 ARL13B/ARL-13 to an Inv-like ciliary membrane subdomain. *PLoS Genet.* **9**, e1003977
20 (2013).
- 21 53. Dona, M. *et al.* NINL and DZANK1 Co-function in Vesicle Transport and Are Essential for
22 Photoreceptor Development in Zebrafish. *PLoS Genet.* **11**, e1005574 (2015).
- 23 54. Chaki, M. *et al.* Exome capture reveals ZNF423 and CEP164 mutations, linking renal
24 ciliopathies to DNA damage response signaling. *Cell* **150**, 533-48 (2012).
- 25 55. Coene, K. L. M. *et al.* OFD1 Is Mutated in X-Linked Joubert Syndrome and Interacts with
26 LCA5-Encoded Lebercilin. *Am. J. Hum. Genet.* **85**, 465-481 (2009).
- 27 56. Roepman, R. *et al.* The retinitis pigmentosa GTPase regulator (RPGR) interacts with novel
28 transport-like proteins in the outer segments of rod photoreceptors. *Hum. Mol. Genet.* **9**,
29 2095-105 (2000).
- 30 57. Roepman, R. *et al.* Interaction of nephrocystin-4 and RPGRIP1 is disrupted by
31 nephronophthisis or Leber congenital amaurosis-associated mutations. *Proc. Natl. Acad. Sci.*
32 **102**, 18520-18525 (2005).
- 33 58. Otto, E. A. *et al.* Candidate exome capture identifies mutation of SDCCAG8 as the cause of a
-

- 1 retinal-renal ciliopathy. *Nat. Genet.* **42**, 840–850 (2010).
- 2 59. Kersten, F. F. *et al.* The mitotic spindle protein SPAG5/Astrin connects to the Usher protein
3 network postmitotically. *Cilia* **1**, 2 (2012).
- 4 60. Eblimit, A. *et al.* Spata7 is a retinal ciliopathy gene critical for correct RPGRIP1 localization
5 and protein trafficking in the retina. *Hum. Mol. Genet.* **24**, 1584–601 (2015).
- 6 61. Huang, L. *et al.* TMEM237 is mutated in individuals with a Joubert syndrome related disorder
7 and expands the role of the TMEM family at the ciliary transition zone. *Am. J. Hum. Genet.* **89**,
8 713–30 (2011).
- 9 62. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**,
10 D841–6 (2012).
- 11 63. Baughman, J. M. *et al.* A computational screen for regulators of oxidative phosphorylation
12 implicates SLIRP in mitochondrial RNA homeostasis. *PLoS Genet.* **5**, e1000590 (2009).
- 13 64. Lambacher, N. J. *et al.* TMEM107 recruits ciliopathy proteins to subdomains of the ciliary
14 transition zone and causes Joubert syndrome. *Nat. Cell Biol.* **18**, 122–131 (2015).
- 15 65. Szklarczyk, R. *et al.* WeGET: predicting new genes for molecular systems by weighted co-
16 expression. *Nucleic Acids Res.* gkv1228 (2015). doi:10.1093/nar/gkv1228
- 17 66. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and
18 hybridization array data repository. *Nucleic Acids Res.* **30**, 207–10 (2002).
- 19 67. van Dam, T. J. P. *et al.* Evolution of modular intraflagellar transport from a coatomer-like
20 progenitor. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 6943–8 (2013).
- 21 68. Barker, A. R., Renzaglia, K. S., Fry, K. & Dawe, H. R. Bioinformatic analysis of ciliary transition
22 zone proteins reveals insights into the evolution of ciliopathy networks. *BMC Genomics* **15**,
23 531 (2014).
- 24 69. Briggs, L. J., Davidge, J. A., Wickstead, B., Ginger, M. L. & Gull, K. More than one way to build a
25 flagellum: comparative genomics of parasitic protozoa. *Curr. Biol.* **14**, R611–2 (2004).
- 26 70. Kensche, P. R., van Noort, V., Dutilh, B. E. & Huynen, M. A. Practical and theoretical advances
27 in predicting the function of a protein by its phylogenetic distribution. *J. R. Soc. Interface* **5**,
28 151–70 (2008).
- 29 71. Swoboda, P., Adler, H. T. & Thomas, J. H. The RFX-type transcription factor DAF-19 regulates
30 sensory neuron cilium formation in *C. elegans*. *Mol. Cell* **5**, 411–21 (2000).
- 31 72. You, Y. *et al.* Role of f-box factor foxj1 in differentiation of ciliated airway epithelial cells. *Am.*
32 *J. Physiol. Lung Cell. Mol. Physiol.* **286**, L650–7 (2004).
- 33 73. Efimenko, E. *et al.* Analysis of xbx genes in *C. elegans*. *Development* **132**, 1923–34 (2005).
-

- 1 74. Chen, N. *et al.* Identification of ciliary and ciliopathy genes in *Caenorhabditis elegans* through
2 comparative genomics. *Genome Biol.* **7**, R126 (2006).
- 3 75. Laurençon, A. *et al.* Identification of novel regulatory factor X (RFX) target genes by
4 comparative genomics in *Drosophila* species. *Genome Biol.* **8**, R195 (2007).
- 5 76. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of
6 transcription factor binding. *Science* **328**, 1036–40 (2010).
- 7 77. Sorokin & S. Centrioles and the formation of rudimentary cilia by fibroblasts and smooth
8 muscle cells. *J. Cell Biol.* **15**, 363–77 (1962).
- 9 78. Hoyer-Fender, S. Centriole maturation and transformation to basal body. *Semin. Cell Dev.*
10 *Biol.* **21**, 142–7 (2010).
- 11 79. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049-56 (2014).
- 12 80. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions
13 from genomic data. *Science* **302**, 449–53 (2003).
- 14 81. Calvo, S. *et al.* Systematic identification of human mitochondrial disease genes through
15 integrative genomics. *Nat. Genet.* **38**, 576–82 (2006).
- 16 82. Graser, S. *et al.* Cep164, a novel centriole appendage protein required for primary cilium
17 formation. *J. Cell Biol.* **179**, 321–30 (2007).
- 18 83. Hjeij, R. *et al.* CCDC151 mutations cause primary ciliary dyskinesia by disruption of the outer
19 dynein arm docking complex formation. *Am. J. Hum. Genet.* **95**, 257–74 (2014).
- 20 84. Trojan, P. *et al.* Centrins in retinal photoreceptor cells: regulators in the connecting cilium.
21 *Prog. Retin. Eye Res.* **27**, 237–59 (2008).
- 22 85. Yang, J. *et al.* Rootletin, a novel coiled-coil protein, is a structural component of the ciliary
23 rootlet. *J. Cell Biol.* **159**, 431–40 (2002).
- 24 86. Overlack, N. *et al.* Direct interaction of the Usher syndrome 1G protein SANS and myomegalin
25 in the retina. *Biochim. Biophys. Acta* **1813**, 1883–92 (2011).
- 26 87. Blanks, J. C. & Johnson, L. V. Specific binding of peanut lectin to a class of retinal
27 photoreceptor cells. A species comparison. *Invest. Ophthalmol. Vis. Sci.* **25**, 546–57 (1984).
- 28 88. Reiners, J., Märker, T., Jürgens, K., Reidel, B. & Wolfrum, U. Photoreceptor expression of the
29 Usher syndrome type 1 protein protocadherin 15 (USH1F) and its interaction with the
30 scaffold protein harmonin (USH1C). *Mol. Vis.* **11**, 347–55 (2005).
- 31 89. Wunderlich, K. A. *et al.* Retinal functional alterations in mice lacking intermediate filament
32 proteins glial fibrillary acidic protein and vimentin. *FASEB J.* **29**, 4815–28 (2015).
-

- 1 90. Sedmak, T. & Wolfrum, U. Intraflagellar transport molecules in ciliary and nonciliary cells of
2 the retina. *J. Cell Biol.* **189**, 171–186 (2010).
- 3 91. Maerker, T. *et al.* A novel Usher protein network at the periciliary reloading point between
4 molecular transport machineries in vertebrate photoreceptor cells. *Hum. Mol. Genet.* **17**, 71–
5 86 (2008).
- 6 92. Sedmak, T., Sehn, E. & Wolfrum, U. Immunoelectron microscopy of vesicle transport to the
7 primary cilium of photoreceptor cells. *Methods Cell Biol.* **94**, 259–72 (2009).
- 8 93. Westerfield, M. *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio Rerio*)*.
9 (Institute of Neuro Science, 1994).
- 10
- 11

1 Acknowledgements

2 We kindly acknowledge Prof. Dr. N. Katsanis and J. R. Willer from Duke University USA for
3 providing several of the clone constructs used as baits in the SILAC and Y2H data sets. We also
4 thank the *Caenorhabditis elegans* Genetics Center (Minnesota, USA), the National Bioresource
5 project (Tokyo, Japan), the International *C. elegans* gene knockout consortium, and the *C. elegans*
6 Million Mutation Project for nematode reagents. We thank Emine Bolat, Ideke Lamers, Moniek
7 Riemersma, and Karlien Coene for performing yeast two-hybrid and SILAC experiments. This work
8 was financially supported via the European Community's Seventh Framework Programme
9 FP7/2009 (SYSCILIA grant agreement no: 241955). TJPvD, RvdL and MAH acknowledge support by
10 the Virgo consortium, funded by the Dutch government (FES0908), and by the Netherlands
11 Genomics Initiative (050-060-452). MRL acknowledges research funding from the Canadian
12 Institutes of Health Research (CIHR; grants MOP-142243 and MOP-82870) and a senior scholar
13 award from Michael Smith Foundation for Health Research (MSFHR). VLJ acknowledges
14 postdoctoral fellowships from MSFHR and KRESCENT. RHG, PLB and RR acknowledge the Dutch
15 Kidney Foundation "KOUNCIL" consortium CP11.18. M.U. and K.B. acknowledge FP7 grant
16 agreement no. 278568, PRIMES. RBR and QL acknowledge support by the Excellence grant
17 CellNetworks funded by the Deutsche Forschungsgemeinschaft (DFG). RS acknowledges support by
18 Metakids Foundation and European Union 7th Framework Programme [PROPANE Study, 602273].
19 PLB and VH-H acknowledge support by the NIHR Great Ormond Street Hospital Biomedical
20 Research Center. PLB is an NIHR Senior Investigator. MS acknowledges support by a Radboudumc
21 Hypatia Tenure Track Fellowship, a Radboud University Excellence Fellowship and the German
22 Research Foundation (DFG) (collaborative research center grant SFB-1411 KIDGEM). RR
23 acknowledges support by the Netherlands Organization for Scientific Research (NWO Vici-
24 865.12.005).

25 Authors' contributions

26 TJPvD and MAH conceived and led the study. TJPvD and MAH, wrote the manuscript with
27 significant contributions from RBR, OEB, RHG, and RvdL. RvdL and TJPvD developed the algorithm
28 and performed the bioinformatic analyses. JK, EdV, KAW, SR, GWD, NJL, CL, VLJ, RH, and VH-H
29 performed the validation experiments and experiments on OSCP1. EdV, NH, YT, YW, JR, GW, BK, JFS,
30 DAM, EvW, GGS, KS, TMTN, SJFL, SECvB, and KB performed experiments leading to the Y2H and
31 SILAC data sets. TJPvD, RvdL and RS created the bioinformatics data sets. TJPvD, JvR, KK, GT, QL
32 collected, quality assessed, formatted and mapped the data sets. MRL, FK, HK, HO, MU, PLB, BF, MS,
33 RHG, RBR, TJG, CAJ, OEB, UW, KB, RR, VH-H, GWD, and MAH suggested strategies and supervised
34 work. All authors read and approved the final manuscript.

1 **Competing financial interests**

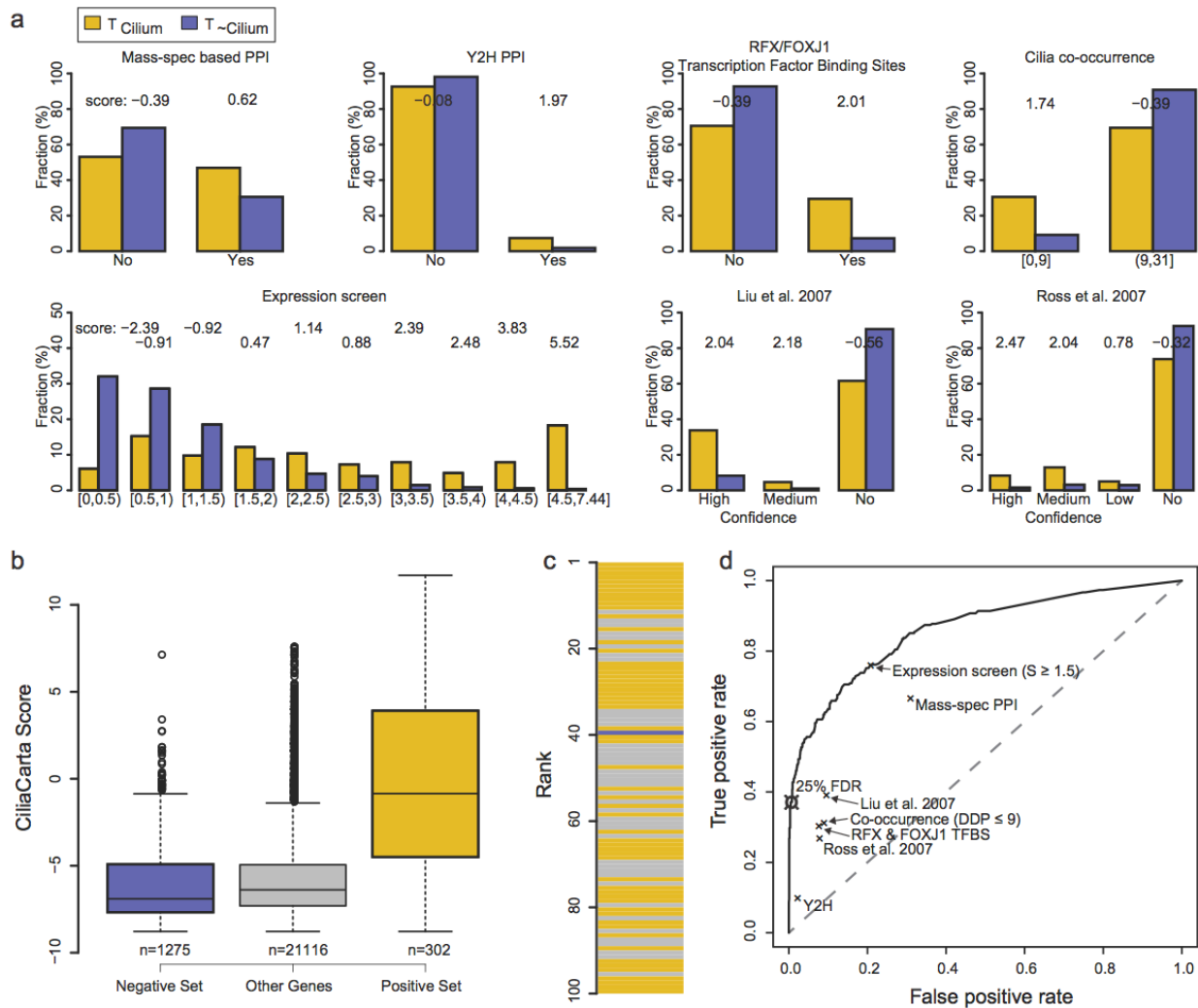
2 The authors declare that they have no competing financial interests.

3 **Materials & Correspondence**

4 Material requests and correspondence should be addressed to T.J.P. van Dam or Martijn A. Huynen.

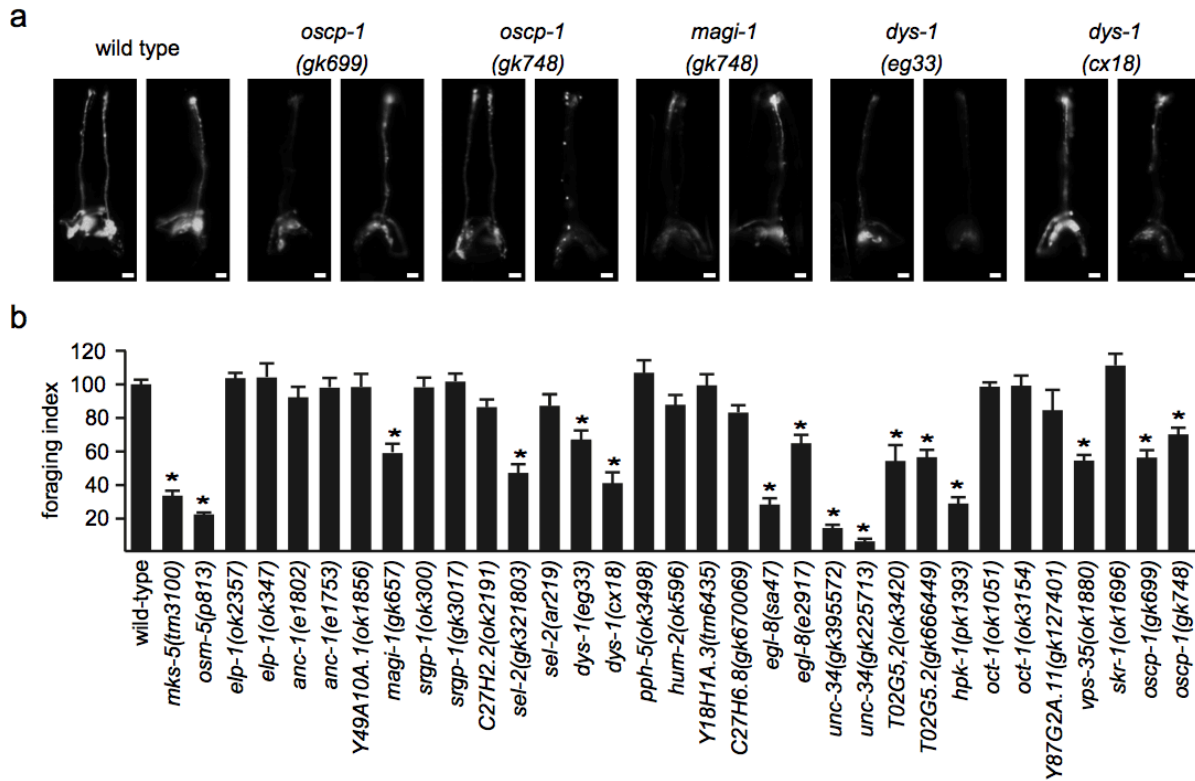
5

1 Figures



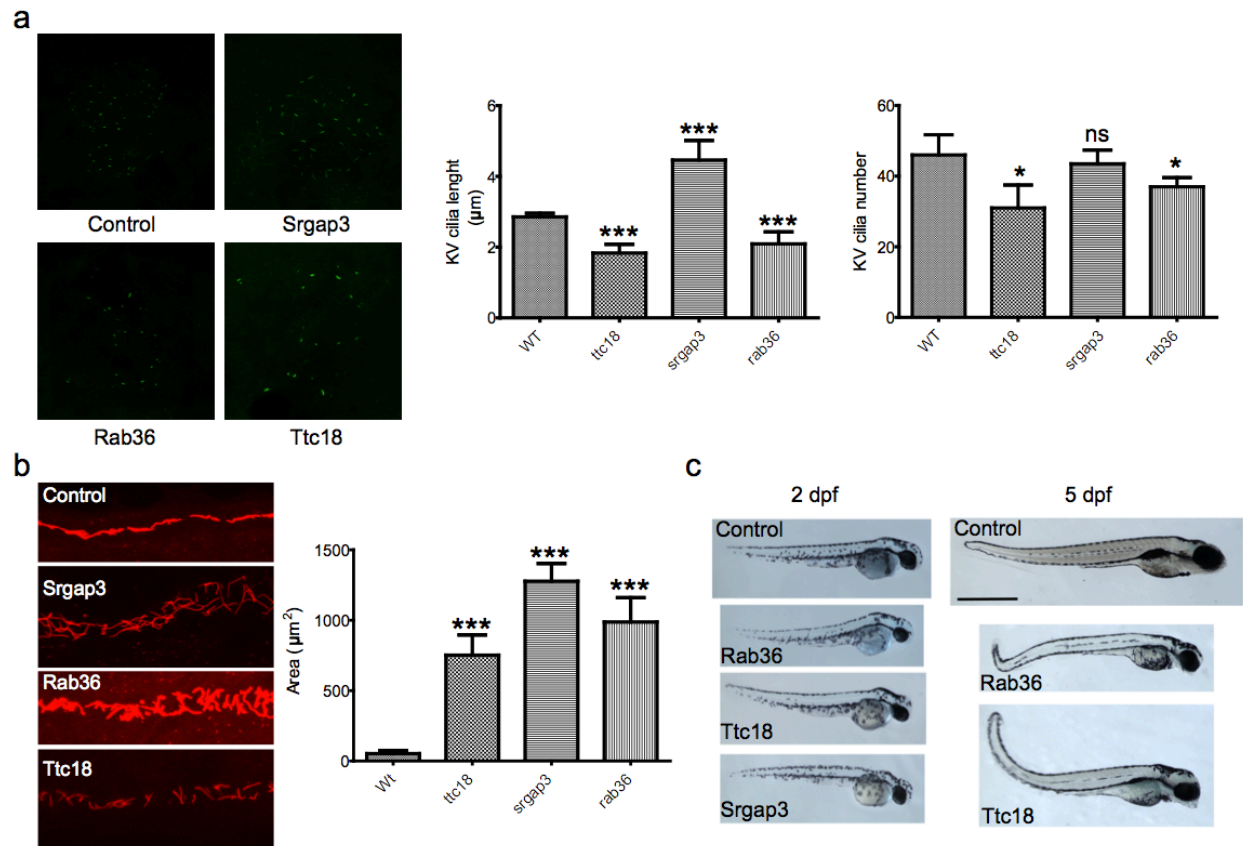
2
3
4
5
6
7
8
9
10

Figure 1: Data sets and performance of the Bayesian classifier for predicting ciliary genes. a) For each data set the fraction of positive (T_{Cilium}) and negative gene sets (T_{~Cilium}) and the log-likelihood scores are displayed per sub-category. b) Distributions of the integrated CiliaCarta scores for the negative set, the positive set, and the remaining unassigned genes. The positive set has significantly higher scores than the negative set (p-value: 2.8e-85 Mann-Whitney U test). c) Top 100 scoring genes. Known ciliary genes from the positive set are in yellow, genes from the negative set are in blue. High scoring genes in grey are prime candidate novel ciliary genes. d) Receiver-Operator Characteristics curve showing the performance of the Bayesian classifier as a function of the CiliaCarta Score and the performance of the individual data sets.



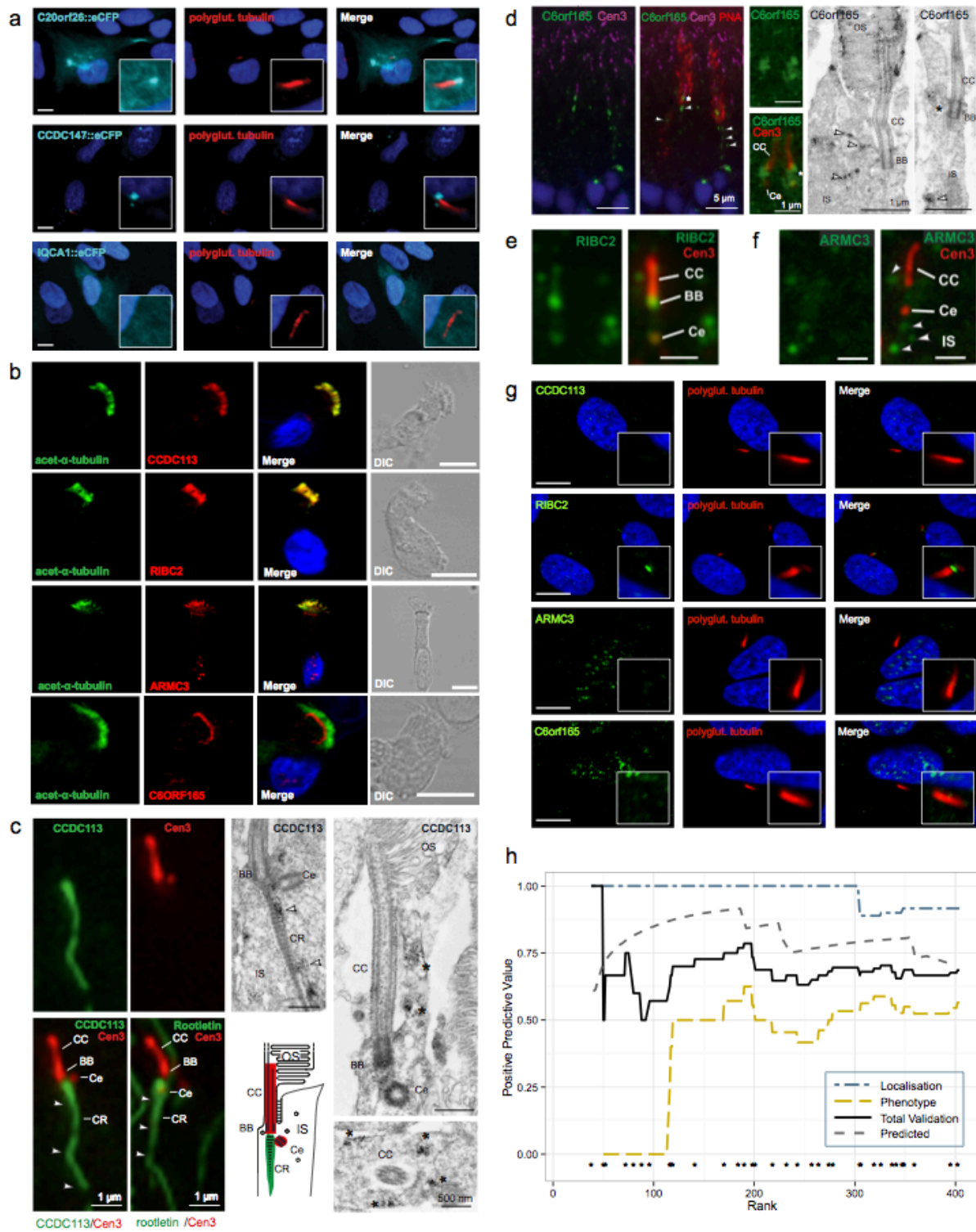
1
2 **Figure 2: Validation by worm phenotype.** a) *C. elegans* dye uptake assay. In wild type worms, DiI dye is
3 taken up by 6 pairs of amphid (head) and one pair of phasmid (tail; not shown) neurons via their
4 environmentally exposed sensory cilia. In the mutants shown, the amount of incorporated dye is modestly
5 reduced, although most or all neurons still uptake the dye. Scale bars: 20 μ m. b) Single worm roaming assays.
6 Bars represent mean \pm S.E.M (n \geq 20) independent experiments), normalized to wild type control. * p<0.05
7 (unpaired t-test; vs. WT).

8



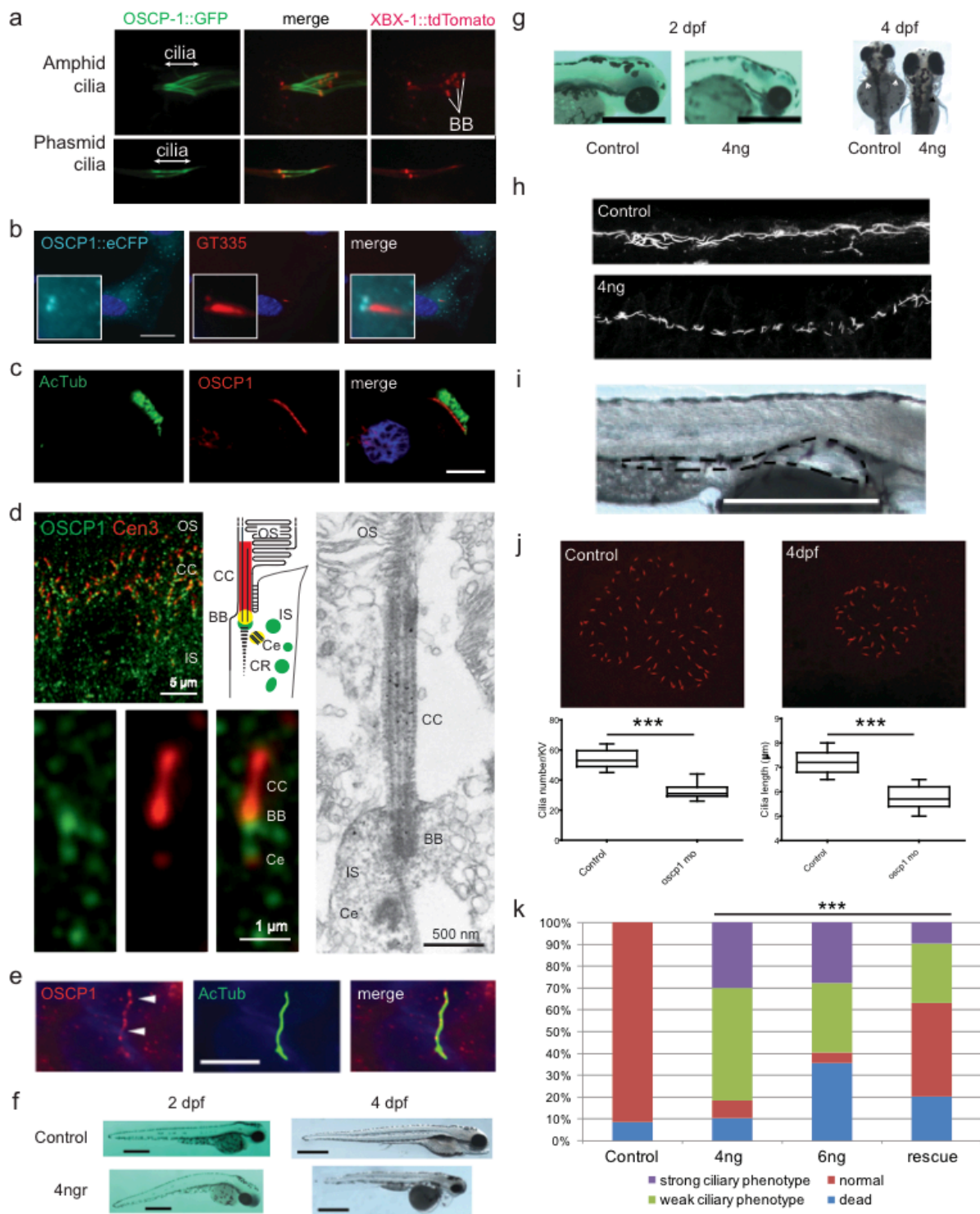
1
2 **Figure 3: Validation by zebrafish phenotype.** a) Cilia length and number in zebrafish Kupffer's vesicles. The
3 length and number of cilia in *ttc18* and *rab36* morphants are significantly reduced ($p < 0.001$ and $p < 0.05$
4 resp.). Cilia in *srgap3* morphants are elongated ($p < 0.001$) but the number of cilia is normal. Bars represent
5 mean \pm S.E.M. b) Pronephric ducts in 24 hpf morphants. Cilia are stained with antibodies against acetylated
6 alpha tubulin. The pronephric ducts are significantly enlarged for all three morphants compared to wild type
7 ($p < 0.001$). Bars represent mean \pm S.E.M. c) Whole embryo phenotype 2 days post fertilization (dpf) and 5 dpf
8 zebrafish control and morphant embryos. All morphants exhibit the body curvature that is characteristic for
9 cilia dysfunction. Note that in our screening we did not manage to obtain surviving *srgap3* morphants past 3
10 dpf.

11



1

1 **Figure 4: Validation by ciliary localization.** a) Fluorescence microscopy of eCFP fused to C20orf26,
2 CCDC147 and IQCA1 in hTERT-RPE1 cells. Acetylated alpha tubulin (red) is used to mark the axoneme. DAPI
3 (blue) staining is used to mark the cell nucleus. IQCA1 does not appear to co-localize with acetylated alpha
4 tubulin. b) Localization of CCDC113, RIBC2, ARMC3 and C6orf165 (red) compared with acetylated alpha
5 tubulin (green) in human lung epithelial cells. c) Localization of CCDC113 in the primary sensory cilium of
6 mature mouse photoreceptor cells. On the left: indirect 2-color immunofluorescence of CCDC113 (green) and
7 centrin-3 (Cen3, red), a marker protein for the connecting cilium (CC), the basal body (BB) and the adjacent
8 centriole (Ce), and of the ciliary rootlet (CR) marker rootletin (green) and Cen3 (red) indicates the
9 localization of CCDC113 at the ciliary base and the CR projecting into the inner segment (IS). On the right:
10 immunoelectron microscopy of CCDC113 confirms the localization of CCDC113 at the CR (arrowheads) and
11 demonstrates accumulation of CCDC113 in the periciliary region of photoreceptor cells (asterisks). d) On the
12 left: localization of C6orf165 in the primary sensory cilium of mature mouse cone photoreceptor cells.
13 Indirect double immunofluorescence of C6orf165 (green) and Cen3 (magenta) in combination with the
14 counterstaining with fluorescent PNA (red) revealed the localization of C6orf165 at the BB and the Ce of cone
15 photoreceptors and a punctate staining in the IS (arrowheads). At the center: higher magnification of the
16 double immunofluorescence of C6orf165 (green) and Cen3 (red). On the right: immunoelectron microscopy
17 of the ciliary region of photoreceptors confirmed the ciliary and periciliary localization (asterisks) of
18 C6orf165, but also demonstrated its presence in outer segments (OS) of cones. e) Double
19 immunofluorescence of RIBC2 (green) and Cen3 (red) of a photoreceptor cilium showed localization of RIBC2
20 throughout the connecting cilium (CC) and the adjacent centriole (Ce) as seen by co-localization with the
21 ciliary marker Cen3. f) Double immunofluorescence of ARMC3 (green) and Cen3 (red) of a photoreceptor
22 cilium revealed the absence of ARMC3 from the CC (counterstained for Cen3) but a punctate staining in the
23 periciliary region of the photoreceptor IS (arrowheads). g) Localization of CCDC113, RIBC2, ARM3 and
24 C6orf165 compared with polyglutamylated tubulin in hTERT-RPE1 cells. h) Positive predictive value (PPV) of
25 the Bayesian classifier based on the experimental validation outcomes plotted against CiliaCarta gene rank.
26 The PPV of the combined validation converges to 0.67, which equals the predicted PPV (0.67, given 0.33 FDR).
27 The asterisks (*) above the x-axis denote the ranks of the candidate genes and proteins tested for ciliary
28 function or localization.



1 **Figure 5. OSCP1 localizes to the cilium and regulates ciliary function in vivo.** a) GFP-tagged OSCP-1
2 driven by its endogenous promoter is specifically expressed in ciliated sensory neurons in *C. elegans*, and the
3 GFP-fusion protein is concentrated along the length of the cilium. Shown are fluorescent images of OSCP-
4 1::GFP and XBX-1::tdTomato (ciliary marker) localisation in amphid and phasmid cilia. Basal bodies (bb) and
5 cilia are indicated. b) OSCP1::eCFP localization in hTERT-RPE1 cells. OSCP1 localizes to the basal body and
6 daughter centriole as well as in the cytosol in a punctate manner. c) OSCP1 localization in human respiratory
7 cells (red) co-stained with acetylated tubulin (green). OSCP1 localizes to the cytosol, but specifically to the
8 base of the ciliated crown of these multi-ciliated cells. d) Indirect high magnification immunofluorescence of
9 OSCP1 (green) and centrin-3 (Cen3, red), a marker protein for the connecting cilium (CC), the basal body (BB)
10 and the adjacent centriole (Ce), in the photoreceptor cilium region of an adult mouse. Immunoelectron
11 microscopy of CCDC113 confirms the localization of OSCP1 at the base of the cilium. The schematic
12 represents a zoom of the ciliary region of a photoreceptor stained for OSCP1 and Cen3 in the according
13 colours. IS, inner segment; OS, outer segment. e) Immunostaining of serum starved murine IMCD3 cells with
14 OSCP1 antibodies. OSCP1 is expressed in a punctate pattern along the axoneme. Acetylated tubulin and γ -
15 tubulin (green), OSCP1 (Proteintech, 12598-1-AP, red) and nuclei (blue). Scale bar; 5 μ m. f) Zebrafish
16 embryos injected with 4 ng of *oscp1* splice morpholino 2 and 4 days post-fertilization (dpf). The
17 characteristic ciliary phenotype with a curved body, small eyes and melanocyte migration defects at 2dpf. 4
18 dpf morphants display obvious pronephric cysts, small eyes, heart edemas, small heads and short bodies.
19 Scale Bars 500 μ m. g) Left panel: details of the head of 2 dpf zebrafish embryos showing small eyes and small
20 head in the *oscp1* morphants. Scale bar 500 μ m. Right panel: dorsal view of 4 dpf zebrafish embryos. Left
21 embryo is a 4dpf *oscp1* morphant. Right embryo is a control. Scale bar 200 μ m. Note the small eyes,
22 melanocyte migration defects and small fin buds (white arrows) in the *oscp1* morphants compared with the
23 control fin buds (black arrows). h) Immunofluorescence staining of pronephric cilia at 24 hpf (acetylated α -
24 tubulin). 4 ng *oscp1* morphants display shortened and disorganized cilia in the medial portion of the
25 pronephric ducts. i) Detail of a pronephric cyst (outlined) in a 4 dpf zebrafish morphant. Scale bar 500 μ m. j)
26 *Oscp1* morphants Kupffer's vesicle cilia staining. *Oscp1* morphants show smaller Kupffer's vesicles with
27 reduced cilia number per Kupffer's vesicle (56 controls vs. 33 *oscp1* morphants) and shorter cilia (controls;
28 7.1 μ m vs *oscp1* morphants; 5.7 μ m. Significance was determined by t-test p-value<0.01. k) Dose dependent
29 phenotype of *oscp1* morphants. After injecting 4 ng of *oscp1* morpholino the percentage of embryos with a
30 weak phenotype is 51% and embryos with a strong phenotype is 28%. Those percentages change when 6 ng
31 of morpholino are used with 31% of embryos showing with a weak phenotype and 27% with a strong
32 phenotype, (strong and weak phenotypes described in Supplementary Fig. 7). The number of dead embryos
33 increases when 6ng of *oscp1* morpholino are used (38%) compared with 4 ng of *oscp1* morpholino (10%).
34 We injected zebrafish embryos at one cell stage with 4 ng of *oscp1* morpholino and 100 pg of human OSCP1
35 mRNA. The rescue increased the normal phenotype percentage from 8% to 43% and decreased the weak
36 phenotype from 51% to 27% and the strong phenotype from 28% to 9.5%. Significance was determined by
37 χ^2 test, p<0.0001.

38

1 Tables

Dataset	# of genes in dataset	Coverage of genome	Coverage of Gold Standard	p-value
TAP-MS	4410	19.4%	64.9%	1.5E-67
SILAC	1397	6.2%	21.5%	3.8E-19
Mass-spec based PPI (TAP-MS + SILAC)	4799	21.1%	65.9%	2.0E-64
Y2H PPI	343	1.5%	9.2%	2.1E-14
RFX/FOXJ1 TFBS	2201	9.7%	29.4%	2.2E-22
Cilia co-occurrence (DDP \leq 9)	1485	6.5%	30.5%	2.8E-37
Expression screen ($S \geq 1.5$)	5448	24.0%	75.5%	2.4E-69
Liu et al. 2007	2085	9.2%	38.4%	1.4E-43
Ross et al. 2007	1204	5.3%	26.2%	2.4E-33

2
3 **Table 1: Coverage and predictive power of the cilium data sets.** Coverage columns denote the fraction of
4 the genome or SCGS that are actually identified by each approach. P-values indicate significant
5 overrepresentation of the SCGS compared to random by Fisher's exact test. Mass-spec based PPI represents
6 the union of the TAP and SILAC data sets, resulting in the integration of five new and two published data sets.

Gene Rank	Gene Symbol	Phenotype based			eCFP fusion	Immunofluorescence/EM				Ciliary phen./loc.?	Published after June 2013 (PubMed IDs)
		Roaming (C. elegans)	Dyefilling (C. elegans)	Zebrafish morpholinos	Localisation based						
					hTERT-RPE1	Lung epithelial cells (human)	hTERT-RPE1 cells	Retina cross-sections (mouse)			
38	C20orf26				Ax. & BB					yes	
50	EML1	++	++							no	
52	RIBC2					Ax.	BB	CC, BB, AC, PCM		yes	24424412
72	ARMC3					Ax.	NCL	CC, BB, PCR, PS		yes	
80	SYNE1	++	++							no	
88	EFHC2	++	++							no	
96	C16orf80				Ax. & BB					yes	
116	MAGI2	-	+							yes	24608321
117	SRGAP3	++	+ / ++	increased cilia length in kv, dilated pronephric duct						yes	26104135
119	FAM65B	++	++							no	
141	CCDC113					Ax.	NCL	Rtlt, BB, AC, PCR		yes	25074808
170	NBEA	-	++							yes	
184	CYB5D1				Ax. & BB					yes	
186	C6orf165					BB	NCL	Cone specific. PIS, PCR, BB.		yes	
190	DMD	++	+							yes	
198	PPP5C	++	++							no	
202	MYO5B	++	++							no	
218	RALGAPA1	++	++							no	
232	CCDC147				BB					yes	
243	C12orf10	++	++							no	
257	C15orf27				BB					yes	
264	PLCB4	-	++							yes	
274	ENAH	-	++							yes	
278	EFCAB7	-	++							yes	24582806
305	IQCA1				NCL					no	
306	HIPK1	-	++							yes	
319	TTC18			decreased cilia length in kv, decreased cilia number in kv, dilated pronephric duct						yes	Predicted in 17971504 (2008)
337	SLC22A4	++	++							no	
341	TSSC1	++	++							no	
347	IPO5				BB					yes	23914977
348	HSPAL1				BB					yes	
349	VPS35	-	++							yes	
359	SKP1	++	++							no	
379	TEKT1				BB					yes	24521320
395	RAB36			decreased cilia length in kv, decreased cilia number in kv, dilated pronephric duct						yes	
402	OSCP1	-	+							yes	

1

1 **Table 2: Genes selected for validation and the validation outcomes.** Empty cells means not tested. Seven
2 proteins (marked by *) have been published as ciliary proteins since the start of the validation experiments.
3 Ax.: axonemal, BB: basal body, NCL: non-ciliary localization, CC: connecting cilium, AC: adjacent centriole,
4 PCM: periciliary membrane, PCR: periciliary region, PS: photoreceptor synapse, Rtl1: Rootlet, PIS:
5 photoreceptor inner segment. Roaming: “++” normal, “-“ defective (p<0.0001). Dye uptake: “++” normal,
6 “+ / ++” slightly reduced uptake, “+” mild reduced dye uptake, “-“ defective dye uptake.