Cilia

REVIEW Open Access

The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium

Teunis JP van Dam[1*], Gabrielle Wheway[2], Gisela G Slaats[3], SYSCILIA Study Group, Martijn A Huynen[1] and Rachel H Giles[3]

Abstract

The multinational SYSCILIA consortium aims to gain a mechanistic understanding of the cilium. We utilize multiple parallel high-throughput (HTP) initiatives to develop predictive models of relationships between complex genotypes and variable phenotypes of ciliopathies. The models generated are only as good as the wet laboratory data fed into them. It is therefore essential to orchestrate a well-annotated and high-confidence dataset to be able to assess the quality of any HTP dataset. Here, we present the inaugural SYSCILIA gold standard of known ciliary components as a public resource.

Review

High-throughput (HTP) experiments and their computational analyses are becoming increasingly important as basic fundamental research tools. However, concerns have been raised with respect to the quality of the earliest comparative analyses of genomics data [1]. For example, the quality of HTP experiments and their bioinformatic analyses is typically undocumented and indeed often unknown. Quality, sensitivity and accuracy are important parameters to consider when deciding how to carry out HTP methods, determine cut-off thresholds and objectively evaluate the results. Within the SYSCILIA consortium, we aim to systematically evaluate the quality of our HTP experiments, such as genome-wide siRNA screening, as well as develop powerful bioinformatic tools and analytical tools to exploit the large datasets produced by HTP procedures across multiple centers. Here, we present one such tool we have generated, the SYSCILIA gold standard (SCGS) of known ciliary genes.

The SCGS is a standardized list of verified ciliary genes, which can be used as a reference dataset of cilia genes for quality metric analyses of experiments, and analyses investigating the cilium and its components. This list is not meant to be comprehensive but rather to be highly reliable; we err on the side of caution to ensure that the genes in this publically available list all encode well-characterized ciliary components. Such a gold standard is a very powerful tool for the comparison of datasets produced by HTP methods, allowing the quantification of the quality of our experiments in terms of sensitivity, specificity and related metrics (for example true positive rate and false discovery rate (FDR)).

Within the field of cilia and ciliopathy research, existing sets of databases, such as Cildb [2] and Cilia Proteome [3], are already widely consulted and represent an immense asset to ciliary research. This is reflected by the frequency of use of these resources by many cilia research groups (cited 14 and 140 times, respectively, in Thomson Reuters Web of Knowledge, 22 May 2013). However, all studies contributing data to these databases are considered equally informative despite some studies likely suffering from a higher number of false positives than others. Objective estimation of the quality or predictive power of each dataset would be a valuable addition. Calculating the sensitivity and specificity of each dataset will provide an objective indicator of whether to include or exclude datasets for a particular purpose, or how to weigh their contribution in Bayesian data integration. Additionally,

* Correspondence: teunis.j.p.vandam@gmail.com
[1]Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen 6500 HB, The Netherlands
Full list of author information is available at the end of the article

comparison of datasets to the SCGS can also facilitate determination of objective cut-off thresholds via receiver operator characteristic (ROC) curves. With the SCGS, we deliver a valuable resource to scientists in the wider field of cilia biology and anticipate a pivotal role for the SCGS in our multi-centre systems biology approach.

The SYSCILIA gold standard of ciliary genes

As a statistical tool, the SCGS needs to be a high-confidence list of sufficient size, but does not need to be comprehensive; the SCGS does not need to contain all possible ciliary genes to be effective. In order to obtain the most reliable results, the SCGS preferably needs to be free of experimental or other biases and contain no incorrectly assigned genes. For this reason, inclusion of genes based solely on recovery by single HTP experiments or sources with similar potentially high FDRs should be avoided; while genes extensively characterized as ciliary genes in individual 'gene-specific' publications, or multiple publications, are highly desirable. Nevertheless, the advantage of HTP results is that they offer a comprehensive starting point to start assembly, without the need to, for example, scan through the whole human genome for cilium genes. An efficient way of combining detailed expert cilia biology knowledge with the comprehensive nature of HTP experiments is to generate an automatically compiled gene list from potentially high quality datasets, curate it manually and combine it with expert knowledge for genes that were missed in the HTP experiments (Figure 1).

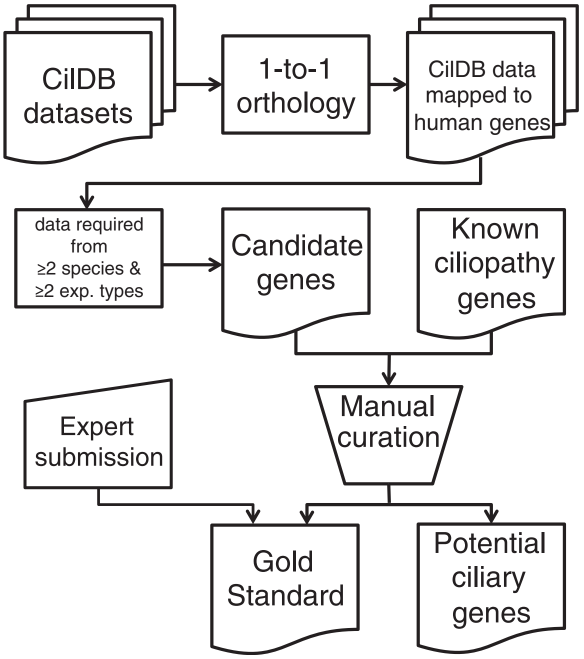

Figure 1 Flow diagram describing the processes to create the SCGS.

To compile the SCGS we collected 27 ciliary studies [2,4-29] from Cildb [2], which holds the largest collection of ciliary datasets (for an overview of the ciliary datasets see Additional file 1). Only datasets based on experimental methods were considered; datasets based on comparative genomics predictions were excluded. The remaining studies covered nine eukaryotic species. All datasets were mapped to human genes by combining two orthology methods, namely OrthoMCL [30] and InParanoid [31]. We only considered one-to-one orthologues between the species of a given dataset and human to avoid cases where after gene duplication, one of the daughter proteins no longer plays a role in the cilium. We defined one-to-one orthologues as defined by InParanoid when both genes are also contained within the same OrthoMCL orthologous group. If InParanoid did not report any human orthologues for a given gene, then the gene reported by OrthoMCL was taken. OrthoMCL performs better in retrieving distant homologues than InParanoid [32], which, with datasets from the distantly related species *Trypanosoma brucei* and *Chlamydomonas reinhardtii*, is particularly invaluable. All other genes in the datasets were excluded, leaving 3,575 genes. The remaining list was then filtered in two ways: data mapped by orthology to a human gene was required to originate from at least two different species and to be shown to be ciliary-related in at least two types of experiments (for example in expression data and proteomics data). A total of 503 genes remained. Finally, a set of 97 medically relevant ciliopathy genes was added from Reeuwijk *et al.* [33]. After removal of overlapping genes, this resulted in a total of 567 potential ciliary genes.

The resulting list of genes was then curated manually. Experts within the SYSCILIA consortium annotated genes as either 'known-ciliary', 'unknown', or 'non-ciliary' based on literature searches. Additionally, members submitted 123 known ciliary genes to this list. Genes were considered ciliary if evidence was published for ciliary localization (including basal body), function in ciliogenesis (including cilium-specific transcription) and involvement in ciliopathies. The final SCGS contains 303 curated ciliary genes.

We are confident that, by combining experimental datasets, a good proportion of the SCGS can be retrieved by commonly used experimental methods. By requiring at least two types of experimental evidence we limit inclusion of experimental biases particular for one type of experiment, like mass spectrometry, which often fails to retrieve membrane proteins [34]. We put effort into annotating the localization of each gene in the SCGS and the SCGS covers all the cilium components (Figure 2). These annotations can be used to quickly compile subsets based on localization.

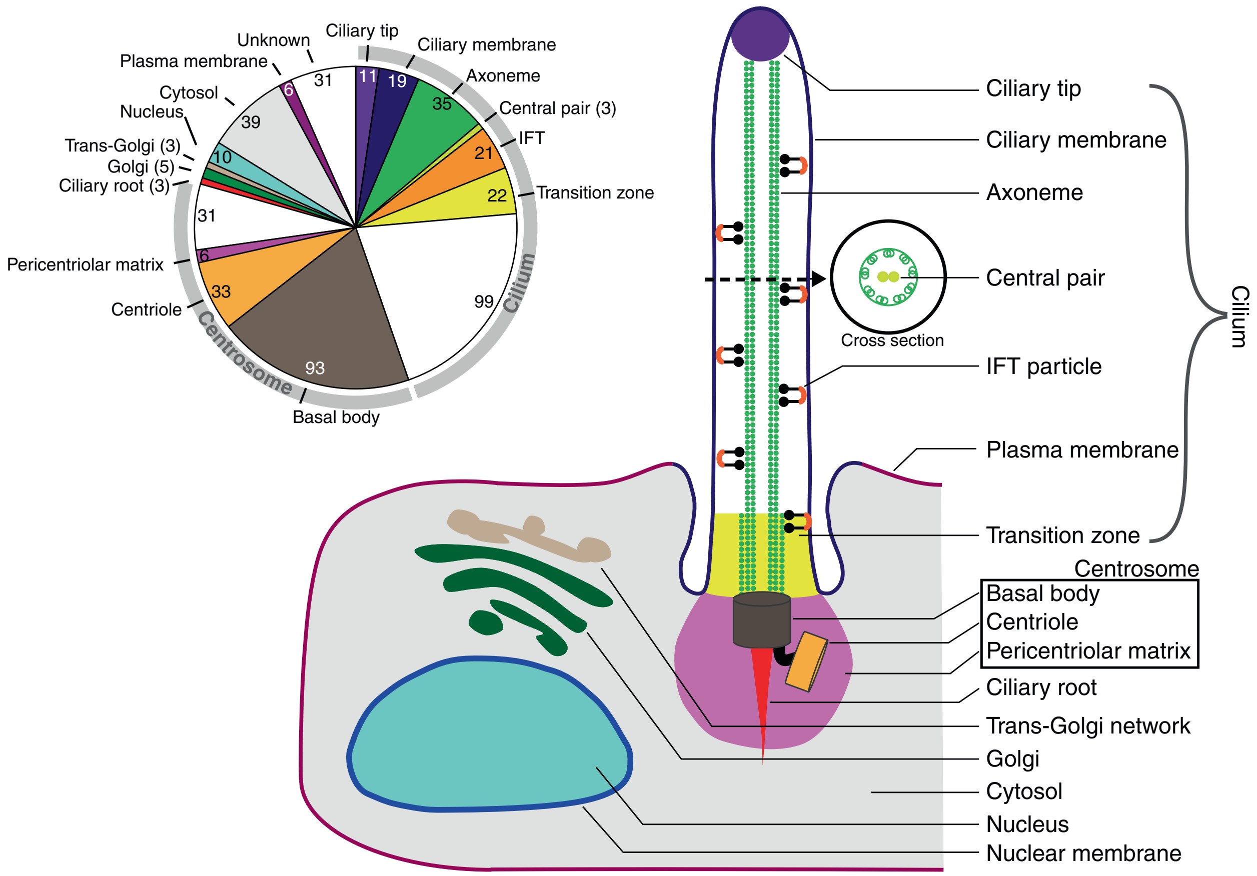

Figure 2 Schematic overview of ciliary components annotated in the gold standard. The schematic depiction of the eukaryotic cilium and its components as annotated in the SCGS, based on Basten *et al.* [36]. The pie chart represents the occurrence of ciliary component localization in the SCGS. The numbers in the individual colors represent the number of individual entries for each location. Note that many genes have been ascribed multiple localizations. SCGS, SYSCILIA gold standard.

Conclusion

Currently, the SCGS is actively used within our consortium for purposes ranging from optimization of experimental methods, to training and evaluating of bioinformatics tools, and as a reference resource. Because of its broad use and importance to the cilia community, we have made the SCGS publicly available (see Additional file 2 and http://www.syscilia.org/goldstandard.shtml). Our list of known ciliary genes is not exhaustive and we expect that the number of newly identified ciliary genes will increase greatly over the next two years. The high stringency applied to the filtering of datasets has led to a small but high-confidence dataset, which we will continue to expand and improve on the basis of novel published cilium genes. Regular updates of the SCGS can be accessed at our consortium website. For many of the metrics discussed above a negative control dataset is also required, that is a list of validated non-ciliary genes. We will also endeavor to make a negative control dataset available in the future. However, it is hard to definitively prove that a gene is never cilia-associated and some genes assigned as negative controls will likely change with new insights. A negative set is therefore volatile; nevertheless SYSCILIA has also recently published such a resource for negative protein-protein interactions [35].

We invite everyone to contribute or curate new and known ciliary genes, to combine and further our collective knowledge on ciliary biology, and use the SCGS to enhance research.

Availability of supporting data

The SYSCILIA gold standard is provided as an excel file in the supplementary material and online at http://www.syscilia.org/goldstandard.shtml.

Additional files

Additional file 1: Table of ciliary datasets used to compile the gene list, and curate the SCGS and references.

Additional file 2: Excel spread sheet of the SYSCILIA gold standard version 1 (SCGSv1), listing 303 curated genes involved in ciliary biology and listing potential ciliary genes.

Abbreviations

FDR: False discovery rate; HTP: High-throughput; ROC: Receiver operator characteristic; SCGS: SYSCILIA gold standard; siRNA: Small interfering RNA.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TJPD and the SYSCILIA Study Group designed the research; TJPD and GW performed the research; TJPD, GW and the SYSCILIA Study Group curated (analysed) the data; and TJPD, GW, GGS, MAH and RHG generated the paper and figures. All authors read and approved the final manuscript.

Acknowledgements

We thank Matthew Betts for assistance with the webpage and online submission form. The SYSCILIA Study Group consists of Gordana Apic (Cambridge Cell Networks Ltd, Cambridge UK), Philip Beales (University College London, London, UK), Oliver E Blacque (University College Dublin, Dublin, Ireland), Brunella Franco (Telethon Institute of Genetics and Medicine, Naples, Italy), Toby J Gibson (European Molecular Biology Laboratory, Heidelberg, Germany), Rachel H Giles (Universitair Medisch Centrum Utrecht, Utrecht, The Netherlands), Martijn A Huynen (Radboud University Medical Centre, Nijmegen, The Netherlands), Colin A Johnson (Leeds Institute of Molecular Medicine, Leeds, UK), Nicholas Katsanis (Duke University Medical Center, Durham, NC, USA), François Képès (Centre National de la Recherche Scientifique, Paris, France), Hannie Kremer (Radboud University Medical Centre, Nijmegen, The Netherlands), Heymut Omran (Westfälische Wilhelms-Universität Münster, Münster, Germany), Marco Pontoglio (Institut National de la Sante et de la Recherche Medicale, Paris, France), Ronald Roepman (Radboud University Medical Centre, Nijmegen, The Netherlands), Rob B Russell (Ruprecht-Karls Universität Heidelberg, Heidelberg, Germany), Marius Ueffing (Eberhard Karls Universität Tübingen, Tübingen, Germany), Gerd Walz (Universitätsklinikum Freiburg, Freiburg, Germany) and Uwe Wolfrum (Johannes Gutenberg Universität Mainz, Mainz, Germany). The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2009, grant agreement number 241955, SYSCILIA.

Author details

[1]Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen 6500 HB, The Netherlands. [2]Section of Ophthalmology and Neuroscience, Leeds Institute of Molecular Medicine, Wellcome Trust Brenner Building, St James's University Hospital, Leeds LS9 7TF, UK. [3]Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht 3508 GA, The Netherlands.

Received: 2 April 2013 Accepted: 28 May 2013
Published: 31 May 2013

References

1. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399–403
2. Arnaiz O, Malinowska A, Klotz C, Sperling L, Dadlez M, Koll F, Cohen J (2009) Cildb: a knowledgebase for centrosomes and cilia. Database (Oxford): bap022
3. Gherman A, Davis EE, Katsanis N (2006) Theciliary proteome database: an integrated community resource for the genetic and functional dissection of cilia. Nat Genet 38:961–962
4. Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C, Warner A, Khattra J, Holt RA, Ou G, Mah AK, McKay SJ, Huang P, Swoboda P, Jones SJM, Marra MA, Baillie DL, Moerman DG, Shaham S, Leroux MR (2005) Functional genomics of the cilium, a sensory organelle. Curr Biol 15:935–941
5. Chen N, Mah A, Blacque OE, Chu J, Phgora K, Bakhoum MW, Newbury CRH, Khattra J, Chan S, Go A, Efimenko E, Johnsen R, Phirke P, Swoboda P, Marra M, Moerman DG, Leroux MR, Baillie DL, Stein LD (2006) Identification of ciliary and ciliopathy genes in Caenorhabditiselegans through comparative genomics. Genome Biol 7:R126
6. Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR, Ruvkun G, Thomas JH, Swoboda P (2005) Analysis of xbx genes in C. elegans. Development 132:1923–1934
7. Boesger J, Wagner V, Weisheit W, Mittag M (2009) Analysis of flagellar phosphoproteins from Chlamydomonas reinhardtii. Eukaryot Cell 8:922–932
8. Keller LC, Romijn EP, Zamora I, Yates JR, Marshall WF (2005) Proteomic analysis of isolated chlamydomonas centrioles reveals orthologs of ciliary-disease genes. Curr Biol 15:1090–1098
9. Pazour GJ, Agrin N, Leszyk J, Witman GB (2005) Proteomic analysis of a eukaryotic cilium. J Cell Biol 170:103–113
10. Stolc V, Samanta MP, Tongprasit W, Marshall WF (2005) Genome-wide transcriptional analysis of flagellar regeneration in Chlamydomonas reinhardtii identifies orthologs of ciliary disease genes. Proc Nat Acad Sci U S A 102:3703–3707
11. Reinders Y, Schulz I, Gräf R, Sickmann A (2006) Identification of novel centrosomal proteins in Dictyostelium discoideum by comparative proteomic approaches. J Proteome Res 5:589–598
12. Laurençon A, Dubruille R, Efimenko E, Grenier G, Bissett R, Cortier E, Rolland V, Swoboda P, Durand B (2007) Identification of novel regulatory factor X (RFX) target genes by comparative genomics in Drosophila species. Genome Biol 8:R195
13. Müller H, Schmidt D, Steinbrink S, Mirgorodskaya E, Lehmann V, Habermann K, Dreher F, Gustavsson N, Kessler T, Lehrach H, Herwig R, Gobom J, Ploubidou A, Boutros M, Lange BMH (2010) Proteomic and functional analysis of the mitotic Drosophila centrosome. EMBO J 29:3344–3357
14. Kim J, Lee JE, Heynen-Genel S, Suyama E, Ono K, Lee K, Ideker T, Aza-Blanc P, Gleeson JG (2010) Functional genomic screen for modulators of ciliogenesis and cilium length. Nature 464:1048–1051
15. Kubo A, Yuba-Kubo A, Tsukita S, Tsukita S, Amagai M (2008) Sentan: a novel specific component of the apical structure of vertebrate motile cilia. Mol Biol Cell 19:5338–5346
16. Nogales-Cadenas R, Abascal F, Díez-Pérez J, Carazo JM, Pascual-Montano A (2009) CentrosomeDB: a human centrosomal proteins database. Nucleic Acid Res 37:D175–D180
17. Ostrowski LE, Blackburn K, Radde KM, Moyer MB, Schlatzer DM, Moseley A, Boucher RC (2002) A proteomic analysis of human cilia: identification of novel components. Mol Cell Proteomics 1:451–465
18. Ross AJ, Dailey LA, Brighton LE, Devlin RB (2007) Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. Am J Respir Cell Mol Biol 37:169–185
19. Cao W, Gerton GL, Moss SB (2006) Proteomic profiling of accessory structures from the mouse sperm flagellum. Mol Cell Proteomics 5:801–810
20. Liu Q, Tan G, Levenkova N, Li T, Pugh EN, Rux JJ, Speicher DW, Pierce EA (2007) The proteome of the mouse photoreceptor sensory cilium complex. Mol Cell Proteomics 6:1299–1317
21. McClintock TS, Glasser CE, Bose SC, Bergman DA (2008) Tissue expression patterns identify mouse cilia genes. Physiol Genomics 32:198–206
22. Arnaiz O, Goût J-F, Bétermier M, Bouhouche K, Cohen J, Duret L, Kapusta A, Meyer E, Sperling L (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. BMC Genomics 11:547
23. Mayer U, Ungerer N, Klimmeck D, Warnken U, Schnölzer M, Frings S, Möhrlen F (2008) Proteomic analysis of a membrane preparation from rat olfactory sensory cilia. Chem Senses 33:145–162
24. Mayer U, Küller A, Daiber PC, Neudorf I, Warnken U, Schnölzer M, Frings S, Möhrlen F (2009) The proteome of rat olfactory sensory cilia. Proteomics 9:322–334
25. Wigge PA, Jensen ON, Holmes S, Souès S, Mann M, Kilmartin JV (1998) Analysis of the Saccharomyces spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. J Cell Biol 141:967–977
26. Kilburn CL, Pearson CG, Romijn EP, Meehl JB, Giddings TH, Culver BP, Yates JR, Winey M (2007) New Tetrahymena basal body protein components identify basal body domain structure. J Cell Biol 178:905–912
27. Smith JC, Northey JGB, Garg J, Pearlman RE, Siu KWM (2005) Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of Tetrahymena thermophila. J Proteome Res 4:909–919
28. Broadhead R, Dawe HR, Farr H, Griffiths S, Hart SR, Portman N, Shaw MK, Ginger ML, Gaskell SJ, McKean PG, Gull K (2006) Flagellar motility is required for the viability of the bloodstream trypanosome. Nature 440:224–247
29. Stubbs JL, Oishi I, Izpisúa Belmonte JC, Kintner C (2008) The forkhead protein Foxj1 specifies node-like cilia in Xenopus and zebrafish embryos. Nat Genet 40:1454–1460
30. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189

31. O'Brien KP, Remm M, Sonnhammer ELL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acid Res 33:D476–D480
32. Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Computat Biol 5:e1000262
33. van Reeuwijk J, Arts HH, Roepman R (2011) Scrutinizing ciliopathies by unraveling ciliary interaction networks. Hum Mol Genet 20:R149–R157
34. Josic D, Clifton JG (2007) Mammalian plasma membrane proteomics. Proteomics 7:3010–3029
35. Trabuco LG, Betts MJ, Russell RB (2012) Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. Methods 58:343–348
36. Basten S, Giles R (2013) Functional aspects of primary cilia in signaling, cell cycle and tumorigenesis. Cilia 2:6

doi:10.1186/2046-2530-2-7
Cite this article as: van Dam *et al.*: The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. *Cilia* 2013 **2**:7.