Topological data analysis reveals genotype-phenotype relationships in primary ciliary dyskinesia

Shoemark, Amelia; Rubbo, Bruna; Legendre, Marie; Fassad, Mahmood R.; Haarman, Eric G.; Best, Sunayna

Published in:
European Respiratory Journal

DOI:
10.1183/13993003.02359-2020

Publication date:
2021

Document Version
Peer reviewed version

Topological data analysis reveals genotype-phenotype relationships in primary ciliary dyskinesia

Shoemark, Amelia.[1,2,*], Rubbo, Bruna.[3,4,*], Legendre, Marie.[5,6], Fassad, Mahmood.R.[7,8], Haarman, Eric.G.[9], Best, Sunayna.[7,10], Bon, Irma.C.M.[9], Brandsma, Joost.[4], Burgel, Pierre-Regis.[11,12], Carlsson, Gunnar.[13], Carr, Siobhan.B.[1], Carroll, Mary.[3,4], Edwards, Matt.[14], Escudier, Estelle.[5,6], Honoré, Isabelle.[11], Hunt, David.[15], Jouvion, Gregory.[5,6], Loebinger, Michel.R.[16], Maitre, Bernard.[17,18], Morris- Rosendahl, Deborah.[14], Papon, Jean-Francois.[19,20,21,22], Parsons, Camille .M.[23], Patel, Mitali.P.[7], Thomas, N.Simon[24,25], Thouvenin, Guillaume.[26,274], Walker, Woolf .T.[3,4], Wilson, Robert.[16], Hogg, Claire.[1], Mitchison, Hannah.M.[6,28,‡], Lucas, Jane.S.[3,4,‡]

*Equal first author contribution

[1] PCD Diagnostic Centre and Department of Paediatric Respiratory Medicine, Royal Brompton and Harefield NHS Trust, London SW3 6NP, UK.
[2] Division of Molecular and Clinical Medicine, University of Dundee, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK.
[3] Primary Ciliary Dyskinesia Centre, University Hospital Southampton NHS Foundation Trust, Southampton SO17 1BJ, UK.
[4] School of Clinical and Experimental Sciences, University of Southampton Faculty of Medicine, Southampton SO17 1BJ, UK

[5] Département de Génétique Médicale, Hôpital Trousseau, Assistance Publique-Hôpitaux de Paris (AP-HP), F-75012 Paris, France.
[6] Sorbonne Université, Institut National de la Santé et de la Recherche Médicale INSERM, U933, Hôpital Trousseau, F-75012 Paris, France.
[7] Genetics and Genomic Medicine Department, University College London, UCL Great Ormond Street Institute of Child Health, London WC1N 1EH, UK.
[8] Department of Human Genetics, Medical Research Institute, Alexandria University, Egypt. 165 El-Horreya Avenue, Alexandria 21561, Egypt.
[9] Department of Pediatric Pulmonology, Emma Children's Hospital, Amsterdam UMC, Vrije Universiteit Amsterdam, the Netherlands.
[10] Leeds Institute of Medical Research, Faculty of Medicine and Health, University of Leeds, Leeds, LS9 7TF, UK.

[11] Service de Pneumologie, Hôpital Cochin, Assistance Publique-Hôpitaux de Paris (AP-HP), F-75014 Paris, France.
[12] Université de Paris, Institut National de la Santé et de la Recherche Médicale INSERM, U1016, Institut Cochin, F-75014 Paris, France.
[13] Department of Mathematics, Stanford University, Stanford, California 94305, USA
[14] Clinical Genetics and Genomics, Royal Brompton and Harefield NHS Foundation Trust, London SW3 6NP, UK.
[15] Wessex Clinical Genetics Service, University Hospitals Southampton, Princess Anne Hospital, Coxford Road, Southampton SO16 5YA, UK.
[16] Host Defence Unit, Department of Respiratory Medicine, Royal Brompton and Harefield NHS Foundation Trust, London, UK. NHLI, Imperial College, London SW3 6NP, UK.
[17]Service de Pneumologie, DHU A-TVB, Centre Hospitalier Intercommunal de Créteil, Université Paris Est, F-94000 Créteil, France.
[18] Institut Mondor de Recherche Biomédicale (IMRB), Unité Inserm U955, F-94000 Créteil, France.
[19] Service d'ORL et Chirurgie Cervico-Faciale, Hôpital Kremlin-Bicêtre, Assistance Publique-Hôpitaux de Paris (AP-HP), F-94270 Le Kremlin-Bicêtre, France.
[20] Faculté de Médecine, Université Paris-Saclay, F-94070 Le Kremlin-Bicêtre, France.
[21] CNRS, ERL 7240, F-94010 Créteil, France.
[22] Institut National de la Santé et de la Recherche Médicale INSERM, U955, F-94010 Créteil, France.
[23] MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton SO17 1BJ, UK.
[24] Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury SP2 8BJ, UK.
[25] Human Genetics and Genomic Medicine, University of Southampton Faculty of Medicine, Southampton SO17 1BJ, UK
[26] Service de Pneumologie Pédiatrique, Hôpital Trousseau, Assistance Publique-Hôpitaux de Paris (AP-HP), F-75012 Paris, France
[27] Sorbonne Université, INSERM, Centre de Recherche Saint-Antoine, F-75012 Paris, France.
[28] NIHR Great Ormond Street Hospital Biomedical Research Centre, London WC1N 3JH, UK.

[‡]**Joint corresponding authors:**

Professor Jane Lucas, Southampton University Hospital, Mailpoint 803 F level, Tremona Road, Southampton, SO16 6YD, UK. E-mail: jlucas1@soton.ac.uk Phone: +44 238120 6160

Professor Hannah M. Mitchison, Genetics and Genomic Medicine, University College London, UCL Great Ormond Street Institute of Child Health, London WC1N 1EH, UK.
E-mail: h.michison@ucl.ac.uk. Phone: +44 207 905 2866

Running head: Genotype-phenotype in primary ciliary dyskinesia

Key words: primary ciliary dyskinesia, genotype, phenotype, cilia, diagnosis, genetic testing

Tweetable ERS abstract:

Topological data analysis of 396 primary ciliary dyskinesia patients shows genetic mutations of worse (*CCDC39*), variable (*DNAH5*) and milder (*DNAH11*) effects on lung function, offering the potential for more accurately targeted disease management.

Author contributions

Concept and design of the study: JSL, CH, HMM, AS, BR

Genotyping: HMM, MRF, MMP, SNT, DH, ME, DM-R, ML

Clinical characterisation: JSL, CH, WTW, MC, SBC, MRL, RW, EGH, J-FP, BM, GT, P-RB, IH

TDA models: BR, JB, GC

Data collection: BR, AS, SB, WTW, CH, J-FP, BM, GT, P-RB, IH, EGH, ICMB, EE, GJ, ML

Planned and performed the statistical analyses: BR, AS, CMP, JSL

Laboratory analyses and data collection: EE, GJ, AS, ML, SB, HMM, MRF

Interpretation of data analyses: BR, AS, JSL, HMM, CH

Drafted the manuscript. AS, BR, JSL, HMM, CH

Revised the manuscript. JSL, CH, HMM, AS, BR, NST, ML, MF, WTW, SBC, IH, EE

All authors have read and approved the final manuscript. HMM and JSL had full access to all data and take final responsibility for the decision to submit for publication.

Funding: The PCD Centres in Southampton and London, the Wessex Regional Genetics Laboratory and Wessex Clinical Genetics Service are funded by the National Health Service for England (NHSE). Clinical research in Southampton was supported by NIHR Southampton Respiratory BRC and NIHR Southampton Wellcome Trust Clinical Research Facility. H.M.M. acknowledges support from Action Medical Research, Great Ormond Street Children's Charity and the NIHR Great Ormond Street Hospital Biomedical Research Centre. M.R.F was also supported by NIHR GOSH BRC and a PhD studentship from the British Council Newton-Mosharafa Fund and Ministry of Higher Education in Egypt. In France this work was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), the RaDiCo funded by the French National Research Agency under the specific programme "Investments for the Future" (Cohort grant agreement ANR-10-COHO-0003) and the Legs Poix grant from the Chancellerie des Universités of Sorbonne Universités. The funders had no role in the writing of the manuscript or the decision to submit it for publication. We have received no payment to write this article. JSL and HMM had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Abstract

Background Primary ciliary dyskinesia (PCD) is a heterogeneous inherited disorder caused by mutations in approximately 50 cilia-related genes. PCD genotype-phenotype relationships have mostly arisen from small case series because existing statistical approaches to investigate relationships have been unsuitable for rare diseases.

Methods We applied a topological data analysis (TDA) approach to investigate genotype-phenotype relationships in PCD. Data from separate training and validation cohorts included 396 genetically defined individuals carrying pathogenic variants in PCD genes. To develop the TDA models, twelve clinical and diagnostic variables were included. TDA-driven hypotheses were subsequently tested using traditional statistics.

Results Disease severity at diagnosis measured by FEV_1 z-score was (i) significantly worse in individuals with *CCDC39* mutations compared to other gene mutations and (ii) better in those with *DNAH11* mutations; the latter also reported less neonatal respiratory distress. Patients without neonatal respiratory distress had better preserved FEV_1 at diagnosis. Individuals with *DNAH5* mutations were phenotypically diverse. Cilia ultrastructure and beat pattern defects correlated closely to specific causative gene groups, confirming these tests can be used to support a genetic diagnosis.

Conclusions This large scale multi-national study presents PCD as a syndrome with overlapping symptoms and variation in phenotype, according to genotype. TDA modelling confirmed genotype-phenotype relationships reported by smaller studies (e.g. FEV_1 worse with *CCDC39* mutations), and identified new relationships, including FEV_1 preservation with *DNAH11* mutations and diversity of severity with *DNAH5* mutations.

Introduction

Primary ciliary dyskinesia (PCD) is clinically and genetically heterogeneous. Symptoms relate to dysfunction of multiple motile cilia and can include neonatal respiratory distress syndrome (NRDS), wet cough, recurring upper and lower respiratory tract infections, otitis media, bronchiectasis, infertility, situs inversus and congenital heart disease (CHD) [1]. Mutations in 50 ciliary genes have been described so far [2, 3].

Understanding of genotype-phenotype relationships informs diagnostic decisions and treatment, but due to the rarity (≈1:10 000) and diversity of PCD, and the constraints of traditional statistical methods, a large patient cohort has never been studied for genotype-phenotype relationships. Evidence for clinically relevant genotype-phenotype associations is mostly limited to small case series for a specific gene or clinical characteristic. For example, individuals with variants in *HYDIN*, a radial spoke head gene, or in multiciliogenesis gene variants like *MCIDAS* and *CCNO* are unlikely to have situs inversus, as nodal cilia are not affected [4-7]. Using traditional statistical approaches, cohort studies have been underpowered to investigate by single genes, and instead have combined functionally similar genes for analysis. A North American study of 137 children reported worse lung disease in those with central apparatus or microtubular disorganisation with inner dynein arm ultrastructural defects, most of whom have *CCDC39* and *CCDC40* variants, than in patients with outer dynein arm defects caused by *DNAH5* variants [8, 9].

Topological data analysis (TDA) allows for the visual exploration of data without establishing *a priori* hypotheses [10]. It can be used to explore the underlying patterns in complex datasets by generating clusters of individuals with similar features in multiple dimensions in an unsupervised manner, as extensively validated in several clinical studies [11-13]. TDA can be used to highlight

small groups of interest in large or complex datasets, that could be overlooked when applying traditional clustering methods that are typically more constrained by a requirement for pre-selection of parameters (e.g. definition of the number of clusters) to drive data analyses [10, 14]. In doing so, TDA can uncover patient subgroups more likely to benefit from a particular therapeutic intervention [12, 15-17]. It thereby provides a promising approach to investigate genotype-phenotype associations in heterogeneous patients with rare diseases.

We aimed to investigate relationships between clinical, diagnostic and genetic data, hypothesising that different subgroups of PCD patients with particular clinical and diagnostic phenotypes could be identified according to their underlying genotypes.

Methods

Ethics

Local and national research and ethical approvals were obtained and adhered to (NRES Committee South Central Hampshire Ethics 06/Q1702/109, London Bloomsbury Research Ethics Committee 08/H0713/82 and Ile-de-France Ethics Committee CPP07729).

Study Design

Clinical and diagnostic data were retrospectively collected from patients with a confirmed genetic diagnosis of PCD i.e. carrying autosomal bi-allelic variants or an X-linked variant classified as pathogenic according to international guidelines [18, 19]. **Supplementary table E1** shows the data coding for the clinical characteristics included in the study.

The study design was based on previous TDA studies and is outlined in **figure 1** [15]. TDA was performed in order to generate hypotheses, which could be tested using more traditional statistical testing. TDA was applied to a discovery cohort of 199 patients (cohort details and genetics can be found in **supplementary tables E2, E3, E4**) and validated using a second cohort of 197 patients (cohort details and genetics can be found in **supplementary figure E1** and **tables E5, E6**). An overview of the PCD genes affected by mutations in the full study population is shown in **supplementary figure E2**.

Topological data analysis

Topological models were developed using a licensed version of TDA software through the Symphony AyasdiAI cloud-based platform (www.ayasdi.com, v 2.0, Ayasdi Inc., Menlo Park, CA). More details of TDA are in the supplementary file.

The phenotypic data used for clustering were body mass index (BMI), forced expiratory volume in 1 second (FEV_1) z-score, forced vital capacity (FVC) z-score, neonatal respiratory distress (NRDS), wet cough, rhinitis, glue ear, cardiac situs, congenital heart disease (CHD), nasal nitric oxide (nNO), ciliary beat pattern (CBP) and transmission electron microscopy (TEM). Genetic data were not used to generate the topological models, as these were the study's main variable of interest; genes of interest were later mapped onto the models to develop hypotheses regarding genotype-phenotype associations.

Models were generated using an automated analysis option. Locally linear embedding (LLE) is a non-linear dimensionality reduction technique, on which highly complex data are summarised and compressed into smaller representations of their variability. The topological model with the best-defined clusters upon visual inspection used two LLE lenses and the correlation distance as metric (i.e. distance function). These identical parameters were applied to develop the discovery and validation models.

The Mapper algorithm was used to identify coherent groups of samples [20]. Each node of the topology model constitutes patients who have combinations of features that are similar between each other, with connecting lines (edges) representing data points that are shared between nodes. The size of the node represents the number of subjects with that specific combination of features.

Genotypes were mapped onto the model to visualise hypothesised associations between genotype and phenotypic clusters. Validation of hypotheses suggested by TDA were then performed using standard statistical analysis. Generating hypotheses using TDA prevented the requirement for multiple comparisons and loss of statistical power.

TDA is an effective method to apply in clinical studies as it can allow for missing data[21]. More detailed explanation of TDA can be found in the **supplementary material**.

Statistical analysis

Selection of variables for hypothesis testing was guided by the topological models to limit the number of comparisons. Further methodological details are provided in the **supplementary material**.

The derived hypotheses were tested through statistical analyses of the whole dataset and of the validation dataset alone. Where the same outcome was tested twice, p-values were adjusted using the Bonferroni correction ($p \leq 0.049$ was found to be significant). Continuous data were compared using student t-tests, ANOVA and Kruskal-Wallis, and categorical data were compared using chi-square or Fisher's exact tests. Tukey's test was used for pairwise comparisons following ANOVA and Dunn's test with Holm-Sidak adjustment following Kruskal-Wallis.Multiple regression models were used to model FEV_1 z-scores, adjusting for age at diagnosis, history of NRDS and presence of CHD. Normality of residuals was investigated using kernel density estimations, and visual inspection of histograms and residuals versus fits graph plots. Number of observations (*n*), regression coefficients (r) with 95% confidence intervals (CI) and model's goodness-of-fitness (adjusted R^2) were reported for each model. Data were analysed in STATA (version 14.0, StataCorp, College Station, TX).

Results

Data-driven genotype-phenotype associations using topological data analysis in a discovery group of 199 PCD patients

Genotype and diagnostic test phenotype associations

TEM defect and CBP mapped visually very closely to corresponding gene group (**figure 2**).

Genotype and FEV_1 associations

Systematic exploration of each of the features collected for this study showed that patients with defects in the 'radial spoke/central complex' and 'nexin-dynein regulatory complex (N-DRC)/molecular ruler' gene functional groups had worse FEV_1 z-scores at diagnosis (as indicated in **figure 3.B** by dark blue coloured nodes) than those with dynein structural gene mutations (higher FEV_1 z-scores, indicated in white coloured nodes in **figure 3.B**). Interestingly, in the cluster with predominantly poor FEV_1 (**figure 3.B** in dark blue), which corresponds to N-DRC or molecular ruler genes (*CCDC39, CCDC40, CCDC65, DRC1;* **figure 3A**), there was a defined group showing absence of history of rhinitis (**supplementary figure E3.B**).

The group with predominantly preserved lung function at diagnosis (**figure 3.B** in white) corresponds to a cluster of individuals with absence of NRDS (**figure 3.C** in white) and an area associated with gene defects of dynein structure (**figure 3.A.** in blue). Further exploration of the topological model showed that within this dynein structural defects group, it was predominantly *DNAH11* patients that had preserved lung function at diagnosis and absence of NRDS (**figure 3.E** in green).

In contrast, individuals with variants in *DNAH5* (the commonest genetic cause of PCD and most predominant patient group in the cohort) were a phenotypically diverse group regarding lung function, with no clear cluster observed (**figure 3.F**).

Genotype and other clinical phenotype associations

The model shows a group of patients with central complex and N-DRC/molecular ruler gene mutations without situs inversus but increased likelihood of glue ear (**supplementary figures E3.A** in yellow and orange, **E3.C** in red) [7, 22]; and a lack of laterality defects associated to *MCIDAS* and *CCNO* in the 'other function' gene group (**supplementary figure E3.D**; red) [6, 23]. Conversely, TDA revealed a cluster of patients with absence of glue ear; this was a genetically diverse group of individuals with dynein structural and assembly defects (**supplementary figures E3.A** in blue and green and **E3.C** in white).

Validation using topological data analysis in a replication group of 197 PCD patients

A validation topological model was generated by analysis of a replication cohort of 197 additional patients: 61 from the UK, 28 from the Netherlands and 108 from France (**supplementary tables E5, E6**). This confirmed the discovery group findings, with *CCDC39* mutation patients clustering in an area of the structure with lower FEV_1 z-scores at diagnosis (**figure 4.B** in dark blue and **figure 4.D** in green) and a higher proportion of reported NRDS (**figure 4.C** in red), while *DNAH11* mutation patients clustered in an area with higher FEV_1 z-scores (**figure 4.E** in green and **figure 4.B** in light blue and white) and less reported NRDS (**figure 4.C** in red and white). The model also confirmed the absence of a clear cluster of patients

with *DNAH5* mutations (**figure 4.F** in green). Additional features of the validation cohort are shown in **supplementary figure E4.**

When analysing gene groups, those with mutations in the 'dynein regulatory/molecular ruler' genes category had worse FEV_1 z-scores (**figure 4.A** in orange and **figure 4.B** in dark blue) and less rhinitis (data not shown) at diagnosis, as seen in the discovery model. The cluster with preserved lung function was mostly formed by patients with dynein structure gene variants (**figure 4.B** in light blue and white and **figure 4.A** in blue), particularly *DNAH11* (**figure 4.E** in green).

However, we could not confirm the inverse association between upper airway (rhinitis and glue ear) and lower airway disease (FEV_1 and NRDS) observed in the discovery model (Figure E4).

The distribution of gene variants in the total 396 patients from both cohorts, in 31 PCD genes, is shown in **figure 5** and the clinical and diagnostic characteristics in **supplementary tables E7 & E8.**

Validation of hypothesis suggested by TDA using standard statistical analysis

Two genes, *CCDC39* and *DNAH11*, fulfilled the criteria for further hypothesis-driven statistical analysis. This required the identification of clearly defined clusters of patients with mutations in each gene showing distinct features, in both the hypothesis-driving discovery (**figure 3**) and the validation (**figure 4**) topological models, along with sufficient patients in each phenotype to allow standard statistical approaches (n = 35 and 48, respectively, **figure 5**). These two genes clustered in areas with extreme values of FEV_1 z-scores in both topological models, leading to

the hypothesis that *CCDC39* and *DNAH11* patients had a distinct respiratory phenotype compared to the rest of the study population.

Testing these hypotheses using traditional statistical analyses, *CCDC39* mutation patients had significantly lower FEV_1 z-scores at diagnosis compared to all other patient genotypes grouped together (r = -1.2; 95% CI, -1.88 to -0.55, adjusted R^2 = 8.0%, p<0.001 *n* = 205), adjusted for age at diagnosis, NRDS and CHD. Conversely, those with *DNAH11* had significantly higher FEV_1 z values at diagnosis (r = 0.09; 95% CI, 0.27 to 1.53; adjusted R^2 = 5.8%, p = 0.003, *n* = 205) and reported less NRDS compared to patients with mutations in any of the other genes (41.03% vs 63.91%, p=0.008).

In contrast, there were no statistically significant differences in NRDS for patients with *CCDC39* mutations (67.86% vs 60.29% for any of the other genes), or in upper airway symptoms (i.e. rhinitis and glue ear) for patients with *CCDC39* (96.77%) or *DNAH11* mutations (97.67%) compared to any of the other genes (93.44% and 93.18%, respectively).

Discussion

This is the first large-scale study to systematically investigate associations between genotype and phenotype in the genetically heterogeneous disorder PCD. It demonstrates the use of a new methodology for the visualisation of data and generation of hypotheses complementing more traditional statistical approaches, where used alone these would not be sufficiently powered, even in multinational cohorts. TDA cluster modelling in nearly 400 individuals from three European countries identified several previously unknown genotype-phenotype relationships, in addition to confirming previously reported genetic associations [7, 22, 24]. PCD, a disease with many well-defined features and 50 causal genes, leant itself to TDA and machine learning for the identification of distinct phenotypic clusters that might share an underlying genetic mutation. TDA was able to identify clinical patterns amongst relatively small numbers of patients (<40) with mutations in a particular gene. We suggest the approach might be beneficial for similar rare diseases, where traditional statistical methods are not suitable.

The TDA model confirmed well-established associations between diagnostic tests (TEM, CBP) and genetics, as seen by the similar colour patterns in the topological models (**figure 2**) where TEM defect and CBP mapped visually very closely to corresponding gene group. This confirms a strong association that is in agreement with the published PCD literature [2, 21]. Distinct genetic findings were also associated with disease severity. We found *CCDC39* patients had significantly worse lung function at diagnosis (FEV_1 z-score) when compared to all other groups, as has previously been observed in individuals with microtubular defects [8, 9, 25, 26]. Furthermore, modelling identified other findings not reported before, including that individuals with *DNAH11* mutations were significantly less likely to have NRDS and, in turn, that the absence of NRDS is associated with better lung function at diagnosis. These findings were

consistent between discovery and validation groups, and when using traditional statistical approaches.

The underlying pattern of the discovery topological model data suggests that patients with compromised lower airways at diagnosis (i.e. decreased lung function and history of NRDS) reported less upper airway symptoms (i.e. history of glue ear and rhinitis). However, these findings could not be verified in the validation model; as they may result from over-fitting of the model, this requires independent validation in an adequately powered independent dataset.

Comparison to previous literature

Our findings confirm and add to evidence from other PCD genotype-phenotype studies. The largest of these have been two cross-sectional and longitudinal studies from the USA and Canada (Genetic Disorders of Mucociliary Clearance Consortium) which also showed that patients with microtubular defects have worse lung function, based on ultrastructural phenotype and limited genotype information [8, 9]. We also confirmed associations previously described in smaller studies, such as the absence of situs inversus in individuals with radial spoke, central complex, N-DRC/molecular ruler gene mutations [4, 5, 22, 27, 28].

A previous study using lung clearance index as a more sensitive measure of lung function showed preserved lung function in a small group of patients from our cohort with normal ultrastructure, of which the majority have *DHAH11* defects [26]. We have further confirmed that this genotype is associated with milder lung disease by showing that these patients clustered in an area with higher values of FEV_1 z-scores. Traditional statistics also showed better preserved lung function in patients with *DNAH11* variants compared to those with mutations in any of the other genes.

Notably, patients carrying mutations in *DNAH5* were phenotypically diverse. The reasons for this are unclear, but may likely be connected to the variety of different mutations within this large gene. *DNAH5* was the gene found to have the widest spectrum of gene variants in our overall cohort. This diversity and high number of different mutations is in line with *DNAH5* being the commonest overall genetic cause of PCD and most frequently mutated gene in affected individuals, with at least 100 different pathogenic mutations recorded worldwide [29]. It is likely in PCD that there will be patient phenotypic differences associated not just with the specific gene, but also the nature and location of the mutations within that gene. These genotype related differences are already emerging on a smaller scale. For example in *DNAH5*, diagnostic results are known to vary somewhat depending on the mutation type, e.g. premature stop codon (nonsense) vs missense [30]. Differences are also associated with missense *versus* truncation mutations in *CCDC103*, where a milder diagnostic and clinical phenotype was described in individuals with p.His154Pro missense mutations [18].

Strengths and weaknesses

This is the largest study investigating genotype-phenotype associations in PCD to date. Using a new methodology of hypothesis-free TDA to examine underlying patterns in the dataset, genotype–phenotype patterns were identified from relatively few patients, something that would be difficult with usual clustering methods. The use of temporally and geographically distinct training and validation groups is highly recommended for such topological clustering approaches [31]. Initial UK discovery findings were validated in the mixed internal and external dataset, including by replication of several important previously published associations, suggesting these results are generalisable to other PCD populations.

The major weakness of our study remains the statistical power required to tease out relationships in a heterogeneous rare condition. To avoid problems with multiple comparisons and loss of statistical power, TDA-led hypothesis testing was performed for only two genes (*CCDC39* and *DNAH11*) and this required combining the discovery and validation datasets. A multinational dataset larger than any existing cohort will be required to ascertain further differences, especially to analyse whether variant types (stop-gain, frameshift, splicing, missense, copy number variants) explain some of the differences seen in the phenotypic data.

Another limitation of our study was potential recall bias for neonatal and early life events, with reliance on parental memory to report symptoms at the time of diagnosis. Not all medical records were complete and therefore missing data were recorded for some of these variables; however, TDA is particularly robust to missing data (see supplementary for additional information) [14]. Finally, we acknowledge that TDA is not completely hypothesis free, as we chose variables to enter into the models and there may be confounding variables affecting our models that have not been identified.

Potential impact for clinical management and research

A better understanding of genotype–phenotype associations from studies such as these should inform education and counselling for PCD patients and their families and will alter disease management in the future. Identifying patients that may require more aggressive or personalised treatment due to underlying genetics will allow for better and targeted care. High risk groups, such as patients with *CCDC39* mutations, might benefit from more intense and targeted therapies.

The identification of mutations in known PCD-causative genes confirms a diagnosis of PCD. The topological models highlighted previously described links between the affected gene, TEM defect and CBP from high-speed video analysis (HSVA), indicating that TEM and HSVA diagnostic tests can play an important supportive role in the classification (likely causal nature) of novel gene variants and variants of uncertain clinical significance [2,19]. These tests can also direct genetic testing to target a specific sub-set of genes.

Our approach for exploring genotype-phenotype associations might be useful for future longitudinal trials in PCD, by including longitudinal parameters such as lung function in the model. It is a model-generating approach that could also be usefully applied to other rare diseases and to more common conditions. More accurate mapping of clinical characteristics, including severity, will allow a more targeted approach to treatments, with associated improvements in patient outcomes.

Overall, these clinically important findings can be useful in counselling parents and when considering prognosis and ongoing therapeutic interventions.

Acknowledgements

We thank the patients and their families for participating in the study and acknowledge the PCD Family Support Group. Dr Borislav Dimitrov, University of Southampton, led initial discussions regarding statistical and TDA approaches to explore genotype-phenotype relationships. He sadly died before the analyses began. We thank the following for their clinical and laboratory contributions to data used in this manuscript: Lucy Jenkins, Thomas Cullup, Alexandros Onoufriadis, Patricia Goggin, Claire L Jackson, Janice Coles, James Thompson, Amanda Harris, Amanda Friend, Mellisa Dixon, Sarah Ollosson, Andrew V Rogers, Emily Frost, Charlotte Richardson, Farheen Daudvohra, Paul Griffin, Thomas Burgoyne. The researchers are supported by the BEAT-PCD: Better Evidence to Advance Therapeutic options for PCD network (COST Action 1407 and European Respiratory Society Clinical Research Collaboration). Several authors of this publication are members of the European Reference Network for Rare Respiratory Diseases (ERN-LUNG) - Project ID No 739546.

Figure legends

Figure 1. Study Design. TDA models were used to identify clusters of clinical and diagnostic characteristics. Gene groups and individual genes were mapped onto these clusters to develop hypotheses, which could subsequently be tested using traditional statistical approaches such as ANOVA. Without the use of TDA then comparison of FEV_1 across >20 genes would require multiple comparisons and statistical power would be lost, whereas using this method we were able to directly test a single directed-hypothesis.

Figure 2. Topological discovery model. Topology analysis display of the results of unbiased clustering of several levels of data, here showing the connections amongst the patients according to their underlying gene defect and the resulting cilia structure and motility defect. Each node represents combinations of features. The size of the nodes represents the number of subjects. The connections represent that there are patients shared between the two nodes. Models A-C are coloured by the following features: A. Gene group; B. Transmission electron microscopy (TEM) results; C. ciliary beat pattern (CBP) by high-speed video analysis (HSVA). Within each of the three models, patients are grouped according to five different classes of gene, TEM and CBP in each of the models respectively. CC= central complex defect, ODA = outer dynein arm, IDA = inner dynein arm, MTD = microtubular disorganisation. Asterisk indicates abbreviation for the nexin-dynein regulatory complex/molecular ruler group.

Figure 3. Topological discovery model. Each node represents combinations of features. The size of the nodes represents the number of subjects. The connections represent that there are patients shared between the two nodes. Models a-f are coloured by the following features: A. Gene group; B. FEV_1 z-scores; C. Neonatal respiratory distress syndrome (NRDS); D. *CCDC39* mutations; E. *DNAH11* mutations; F. *DNAH5* mutations. Asterisk indicates abbreviation for the nexin-dynein regulatory complex/molecular ruler group.

Figure 4. Topological validation model. Each node represents combinations of features. The size of the nodes represents the number of subjects. The connections represent that there are patients shared between the two nodes. Models a-f are coloured by the following features: A. Gene group; B. FEV_1 z-scores; C. Neonatal respiratory distress syndrome (NRDS); D. *CCDC39* mutations; E. *DNAH11* mutations; F. *DNAH5* mutations. Asterisk indicates abbreviation for the nexin-dynein regulatory complex/molecular ruler group.

Figure 5. Total patient population according to genotype (n = 396). Mutations in 31 PCD genes were included for analysis. Bars are coloured according to gene group: blue represents genes involved in dynein structure, green in dynein assembly, yellow in radial spoke and central complex, orange in nexin-dynein regulatory complex/molecular ruler, and red in other functions such as ciliogenesis.

References

1. Goutaki, M., et al., *Clinical manifestations in primary ciliary dyskinesia: systematic review and meta-analysis.* Eur Respir J, 2016. **48**(4): p. 1081-1095.
2. Lucas, J.S., et al., *Primary ciliary dyskinesia in the genomics age.* Lancet Respir Med, 2020. **8**(2): p. 202-216.
3. Wallmeier, J., et al., *Motile ciliopathies.* Nat Rev Dis Primers, 2020. **6**(1): p. 77.
4. Olbrich, H., et al., *Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry.* Am J Hum Genet, 2012. **91**(4): p. 672-84.
5. Castleman, V.H., et al., *Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities.* Am J Hum Genet, 2009. **84**(2): p. 197-209.
6. Boon, M., et al., *MCIDAS mutations result in a mucociliary clearance disorder with reduced generation of multiple motile cilia.* Nat Commun, 2014. **5**: p. 4418.
7. Best, S., et al., *Risk factors for situs defects and congenital heart disease in primary ciliary dyskinesia.* Thorax, 2019. **74**(2): p. 203-205.
8. Davis, S.D., et al., *Clinical features of childhood primary ciliary dyskinesia by genotype and ultrastructural phenotype.* Am J Respir Crit Care Med, 2015. **191**(3): p. 316-24.
9. Davis, S.D., et al., *Primary Ciliary Dyskinesia: Longitudinal Study of Lung Disease by Ultrastructure Defect and Genotype.* Am J Respir Crit Care Med, 2019. **199**(2): p. 190-198.
10. Lum, P.Y., et al., *Extracting insights from the shape of complex data using topology.* Scientific Reports, 2013. **3**.
11. Nielson, J.L., et al., *Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury.* Nat Commun, 2015. **6**: p. 8581.
12. Frattini, V., et al., *A metabolic function of FGFR3-TACC3 gene fusions in cancer.* Nature, 2018. **553**(7687): p. 222-227.
13. Bruno, J.L., et al., *Longitudinal identification of clinically distinct neurophenotypes in young children with fragile X syndrome.* Proc Natl Acad Sci U S A, 2017. **114**(40): p. 10767-10772.
14. Offroy, M. and L. Duponchel, *Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry.* Anal Chim Acta, 2016. **910**: p. 1-11.
15. Nicolau, M., A.J. Levine, and G. Carlsson, *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(17): p. 7265-7270.
16. Li, L., et al., *Identification of type 2 diabetes subgroups through topological analysis of patient similarity.* Sci Transl Med, 2015. **7**(311): p. 311ra174.
17. Siddiqui, S., et al., *Airway pathological heterogeneity in asthma: Visualization of disease microclusters using topological data analysis.* J Allergy Clin Immunol, 2018. **142**(5): p. 1457-1468.
18. Lucas, J.S., et al., *European Respiratory Society guidelines for the diagnosis of primary ciliary dyskinesia.* European Respiratory Journal, 2017. **49**(1): p. 1601090.
19. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* Genet Med, 2015. **17**(5): p. 405-24.
20. Singh, G., F. Mémoli, and F. Carlsson, *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, in *Eurographics Symposium on Point Based Graphics*, M. Botsch, et al., Editors. 2007, The Eurographics Association. p. 91-100.
21. Glushakov, S. and I. Kotenko. *Handling Missing Data in Clinical Trials Using Topological Data Analysis*. 2018.

22. Pruliere-Escabasse, V., et al., *Otologic features in children with primary ciliary dyskinesia.* Arch Otolaryngol Head Neck Surg, 2010. **136**(11): p. 1121-6.
23. Wallmeier, J., et al., *Mutations in CCNO result in congenital mucociliary clearance disorder with reduced generation of multiple motile cilia.* Nat Genet, 2014. **46**(6): p. 646-51.
24. Davis, S.D., et al., *Clinical Features of Childhood Primary Ciliary Dyskinesia by Genotype and Ultrastructural Phenotype.* American Journal of Respiratory and Critical Care Medicine, 2015. **191**(3): p. 316-324.
25. Shah, A., et al., *A longitudinal study characterising a large adult primary ciliary dyskinesia population.* European Respiratory Journal, 2016. **48**(2): p. 441-450.
26. Irving, S., et al., *Primary Ciliary Dyskinesia Due to Microtubular Defects is Associated with Worse Lung Clearance Index.* Lung, 2018. **196**(2): p. 231-238.
27. Knowles, M.R., et al., *Mutations in RSPH1 cause primary ciliary dyskinesia with a unique clinical and ciliary phenotype.* Am J Respir Crit Care Med, 2014. **189**(6): p. 707-17.
28. Vallet, C., et al., *Primary ciliary dyskinesia presentation in 60 children according to ciliary ultrastructure.* Eur J Pediatr, 2013. **172**(8): p. 1053-60.
29. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Res, 2018. **46**(D1): p. D1062-D1067.
30. Kispert, A., et al., *Genotype-phenotype correlations in PCD patients carrying DNAH5 mutations.* Thorax, 2003. **58**(6): p. 552-554.
31. Moons, K.G., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration.* Ann Intern Med, 2015. **162**(1): p. W1-73.