Protein interaction perturbation profiling at amino-acid resolution

Jonathan Woodsmith[1,2] , Luise Apelt[1,3], Victoria Casado-Medrano[1,3], Ziya Özkan[1], Bernd Timmermann[1] & Ulrich Stelzl[1,2]

The identification of genomic variants in healthy and diseased individuals continues to rapidly outpace our ability to functionally annotate these variants. Techniques that both systematically assay the functional consequences of nucleotide-resolution variation and can scale to hundreds of genes are urgently required. We designed a sensitive yeast two-hybrid-based 'off switch' for positive selection of interaction-disruptive variants from complex genetic libraries. Combined with massively parallel programmed mutagenesis and a sequencing readout, this method enables systematic profiling of protein-interaction determinants at amino-acid resolution. We defined >1,000 interaction-disrupting amino acid mutations across eight subunits of the BBSome, the major human cilia protein complex associated with the pleiotropic genetic disorder Bardet–Biedl syndrome. These high-resolution interaction-perturbation profiles provide a framework for interpreting patient-derived mutations across the entire protein complex and thus highlight how the impact of disease variation on interactome networks can be systematically assessed.

The recent explosion in the amount of next-generation sequencing data is beginning to reveal the scope of genomic variation in healthy and diseased individuals. As such, an important and growing challenge in human genetics is identifying and characterizing variants of unknown significance that may lead to phenotypic change. As sequence variation data sets continue to grow, so does the statistical power to identify possible disease- or phenotype-associated mutations[1,2]. However, statistical methods inherently lack experimental evidence for any mechanistic outcome of identified mutations and techniques to systematically provide a functional context for nucleotide-resolution variation lag further and further behind. This can be explained in large part by the technical challenge of coupling systematic gene perturbation at nucleotide resolution with screening tens or hundreds of genes for a functional output that could drive phenotypic change.

At the molecular level, phenotypic change is principally driven through genetic variation that alters expression, localization, chemical activity or interaction profiles of associated RNA or proteins[3,4]. As such, systematically profiling any of these four generic functions will provide a mechanistic basis for characterizing known genetic variation. Toward this end, two recent studies screened large collections of disease-associated genetic point mutants for alterations in protein expression or localization[5], protein–DNA interactions[4] and protein–protein interactions[4,5]. These studies individually examined hundreds of single-protein variants, yet intrinsically such studies cannot keep up with rapidly increasing disease mutation annotations. As our knowledge of genetic variation in health and disease expands, the time and resources needed to create required clone sets become increasingly burdensome.

These issues can be circumvented by employing complex genetic mutant libraries, but variants that perturb protein function must be actively enriched before identification. Library generation is facilitated by recent advances in massively parallel programmed mutagenesis[6,7], and these advances allow construction of the entire mutational repertoire of a protein sequence in a targeted manner. However, to the best of our knowledge, no systematic method currently exists that can efficiently enrich nonsynonymous variants that perturb any of the four generic functional properties of proteins. This consequently prohibits large-scale screening approaches.

Here we address this by developing Int-Seq, a scalable method that can dissect binary protein–protein interactions at individual amino-acid resolution. We designed a reporter-gene promoter that turns off in response to a positive yeast two-hybrid (Y2H) signal. We combined this with massively parallel programmed mutagenesis and a sequencing readout to allow highly efficient selection and systematic identification of interaction-disrupting mutations. We use Int-Seq to probe the interaction landscape of components of Bardet–Biedl syndrome (BBS), a pleiotropic genetic disorder whose causative genes in large part make up an evolutionary conserved 0.5-megadalton (MDa) protein complex, the BBSome[8]. The BBSome complex plays a major role in transport of cellular cargo to and within the primary cilium, a crucial organelle involved in extracellular environment sensing across processes ranging from development to homeostasis[9,10]. Here we

provide an amino-acid-resolution interaction profile map of the BBSome, which defines residues crucial in mediating intracomplex interactions and aids prioritization of known BBS mutations on a mechanistic basis.

RESULTS

Int-Seq strains

In contrast to classical Y2H approaches, reverse two hybrid (R2H) technologies utilize protein–protein interactions to repress yeast growth (**Fig. 1a**). As such, the R2H system generates the opposite readout of the Y2H system to promote growth of noninteracting pairs on selective media. Two variations of the R2H method are to turn off auxotrophic reporter genes in response to Tet repressor (TetR) production[11] or to induce lethality by a toxic metabolite of the URA3 enzyme and the substrate 5-FOA[12,13]. These technologies have never been shown to be scalable to screening formats, despite the two decades since inception and recent methodological variations[14].

Therefore, to facilitate screening we combined modern synthetic biology principles with decades-old knowledge of yeast biosynthetic gene promoter architecture[15]. We integrated multiple tetO repeats into a novel synthetic promoter based on a minimal feature set required for *ADE2* gene expression to generate a repressive, switch-like response to interaction-driven *TetR* reporter gene activation (see Online Methods). The final diploid genome contains two copies of the LexA[8]::HIS3 Y2H reporter, four copies of the LexA[8]::TetR reporter and two copies of the *ADE2* gene under the control of the synthetic promoter containing five tetO sequences (**Fig. 1a**). As this genetic engineering approach was undertaken in our standard Y2H strains[16,17], four protein pairs could produce reciprocal readouts on agar lacking either histidine or adenine (**Fig. 1a** and **Supplementary Fig. 1e**). Int-Seq strain pairs are therefore functional for both network generation and dissection, which reduces screening workload. They allow efficient enrichment of interaction-disruptive variants over a large background of growth-repressed variants and can be coupled to a sequencing-based identification readout[17] (**Fig. 1b,c**).

BBSome interaction network

We first mapped binary interactions between nonchaperone human BBSome proteins using full-length and domain-based clones. This resulted in a highly specific, reproducible Y2H network involving eight proteins in nine interactions (**Fig. 2** and **Supplementary Fig. 2**; see Online Methods).
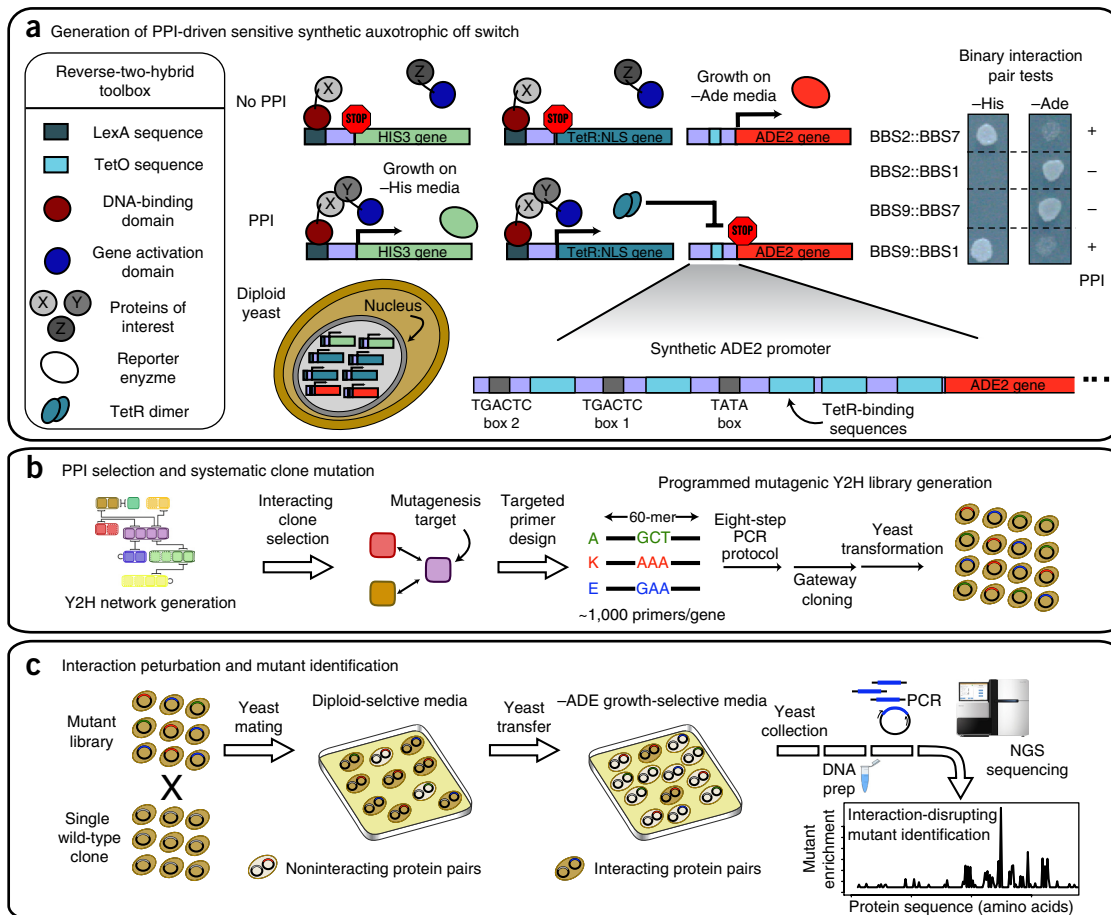

Figure 1 | Int-Seq pipeline to identify regions mediating protein interactions at amino-acid resolution. (**a**) Schematic of the interaction-driven reporter genes used here with an enlarged view of the ADE2 synthetic promoter that enables this technology. Inset top right, examples of Int-Seq yeast strain phenotypes driven by two interacting (+) and two noninteracting (–) protein pairs. PPI, protein–protein interaction. (**b**) Interacting construct selection and programmed mutant yeast library preparation. Mutant primers for on-chip synthesis contained the target codon mutated to an A, K, or E within a 60-bp DNA sequence. Mutant library was generated via eight-step PCR protocol then Gateway cloned into Y2H vectors before yeast transformation. (**c**) Mating approach to select for noninteracting mutant proteins followed by identification via second generation sequencing analysis.
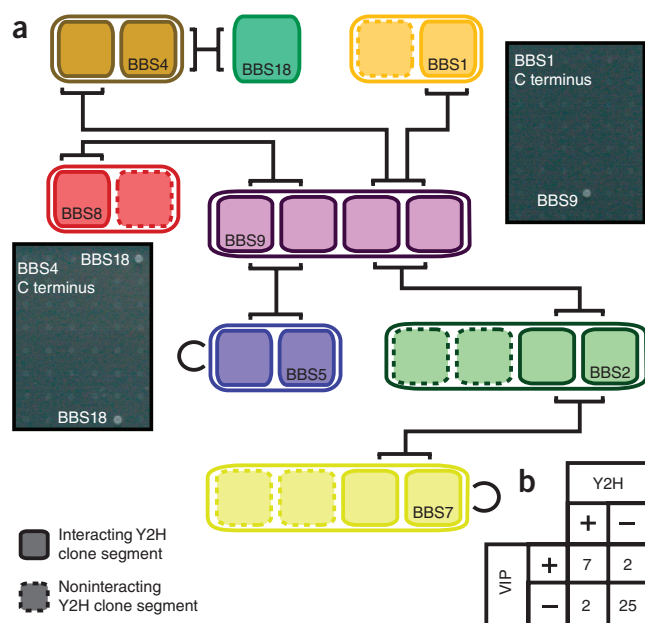
Figure 2 | Domain-based binary interaction network of the BBSome. (**a**) Proteins are depicted N terminal to C terminal left to right. Examples of Y2H results inset: BBS4 bait interacting with two independent BBS18 clones, BBS1 bait interacting with BBS9. (**b**) Comparison between the Y2H interaction data and a fluorescence-based immunoprecipitation approach (VIP, Katoh *et al.* 2015[20]).

The binary data are in agreement with the overall structural information of the main BBSome complex in the literature[10,18], with a core complex of BBS9, BBS2 and BBS7 surrounded by more peripheral members that mainly connect in through BBS9. Our domain-based approach allowed us to narrow down interaction sites, with the C-terminal-domain sections of BBS2,7 and 9 being necessary and sufficient for binary interactions within the core complex. While BBS8 and BBS5 only interact with full-length BBS9, the C-terminal domain of BBS9 is responsible for linking to the peripheral BBS1 and BBS4 subunits (**Fig. 2a**). The N-terminal domains of BBS8 and 4 are required for BBS9 interaction. While BBS3 is known to functionally interact with the N-terminal domain of BBS1[19], this domain yielded no interactions in our assay. We also reconfirmed a previously published BBS4 C-terminal interaction with BBS18 at its C terminus[18,20].

A previous study investigated binary BBSome interactions in cultured HEK cells[20], allowing a systematic comparison to assay reliability of our Y2H network. 78% of the interactions observed in ref. 20 in the visual immune precipitation (VIP) assay were recapitulated here, while 93% of the reported noninteracting pairs were also confirmed as negative (**Fig. 2b**). This high concordance rate between the two independent screens highlights the quality of the data set presented here.

Int-Seq constructs and targeted mutagenesis

The majority of the Y2H-positive interactions gave robust growth repression on media lacking adenine, which highlighted the scalability of the approach (**Fig. 3a**). We then applied massively parallel mutagenesis to systematically mutate all BBS components within our network[6]. In all programmed mutagenesis approaches, a tradeoff exists between the number of residues (and, as such,

genes) that can be mutated and the number of alterations programmed per residue. Here we aimed to identify individual amino acids required for maintaining protein interactions across many full-length proteins or large domains. We reasoned that mutating each residue individually to either a glutamic acid (E, negative switch) or a lysine (K, positive switch) could disrupt multiple types of local interaction, an approach termed hydrophile scanning[21]. In addition, an alanine (A) scanning approach could act as a reference point to identify amino acids that are generally susceptible to side-chain alterations. Using targeted AKE mutagenesis, we can systematically scan the majority of the BBS Y2H network with a single on-chip primer synthesis reaction. We generated six comprehensively mutated BBS constructs and added two random mutant libraries (see Online Methods), and we achieved an estimated AKE mutation coverage of 85–100% for comprehensive Int-Seq interrogation (**Supplementary Fig. 3**; see Online Methods).

Interaction perturbation landscape of the BBSome

We mated each mutant library against its wild-type interacting counterpart, and we enriched mutants that perturb the physical interaction through growth on –adenine agar (**Fig. 1c**). Yeast were then collected, and disruptive mutants were identified through next-generation-sequence analysis (see Online Methods). To ensure we could accurately identify mutants, we also sequenced pDONR and Y2H vector libraries under no interaction-selective pressure. In total we collected 159,937,962 paired-end reads distributed across 12 controls and 35 interaction samples (**Supplementary Data 1**).

The sequence data contained both mutations coded by the mutagenesis protocol (Ala, Lys and Glu) and, to a much lesser extent, noncoded 'random' mutants (**Fig. 3b** and **Supplementary Fig. 4**; see Online Methods). The recall frequency of sequences that mapped to any given mutant were dependent on both the amino-acid codon (**Fig. 3b**) and the position of the mutant sequenced (**Fig. 3c**), with mutations coded during the mutagenesis most frequently sequenced (**Fig. 3b**). This relationship extends to all amino-acid mutants and allows us to statistically model the variation present between samples (noncoded mutants, **Fig. 3d**; coded mutants, **Fig. 3e**). When comparing a Y2H-input library to sequences obtained from a mutant library tested against a wild-type interaction partner, we observe a clear deviation from the linear model across multiple positions in the AKE mutations but very little deviation from all other sequenced mutations (**Fig. 3f,g**). We modeled mutation enrichment for all interaction pairs (see Online Methods) and generated high-confidence interaction-perturbation profiles for 11 of the 14 directed interactions tested. This highlights the adaptability of the Int-Seq approach to most interactions detected by Y2H screening approaches. With the exception of the BBS18–BBS4(C terminal) pair that was perturbed through random mutagenesis, interactions were affected 8–67-fold more by coded AKE mutations than by all other mutations (**Fig. 3h**).

As a simple example, we resolved a 34-amino-acid stretch at the N terminus of BBS8 that, when mutated, could no longer mediate an interaction with BBS9 (**Fig. 4**). While the entire Y2H construct was covered by the mutant library, disruptive variants were almost exclusively confined to this region. This suggests mutations in BBS8 distal to this region do not destabilize this interaction-mediating region, and that the AKE mutagenesis
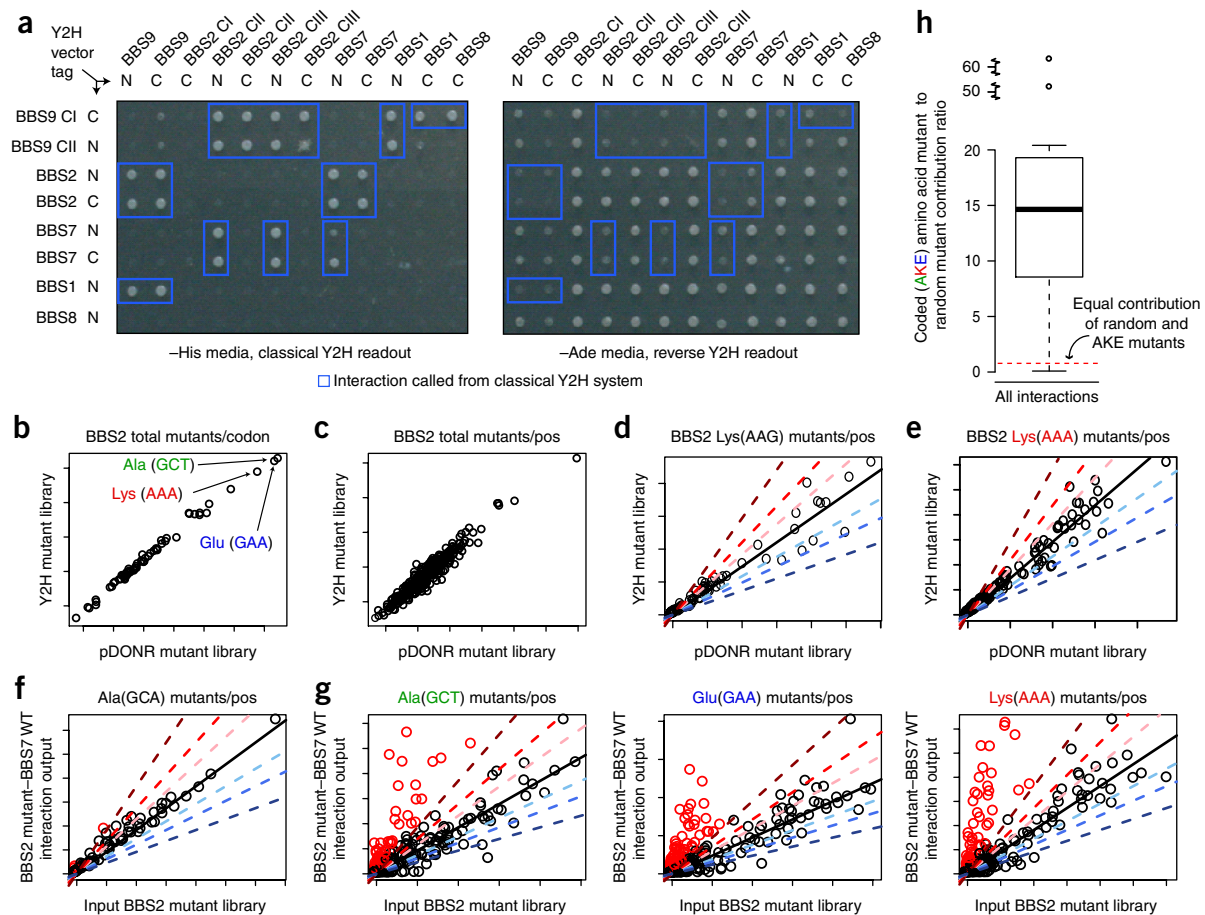
Figure 3 | The Int-Seq pipeline is scalable across many interactions and can identify disruptive single amino-acid mutations. (**a**) Comparison of the forward- and reverse-two-hybrid signals generated from the Int-Seq yeast strains. CI, CII, and CIII refer to differing constructs used in the approach for a given gene. For graphs **b**–**g**, *x* and *y* axes display the number of times a mutation was sequenced for a given codon or at a given position for BBS2 examples (here termed recall statistics). Recall statistics for any given mutation sequenced in control vectors are dependent on both the amino acid (**b**) and codon position (**c**). Recall statistics for individual nonprogrammed (**d**) and programmed mutant codons (**e**) across all positions within the target gene follow a linear model under no selective pressure. Recall statistics for individual nonprogrammed (**f**) and all programmed mutant codons (**g**) across all positions within the target gene in the interaction output when compared with vector control. Dashed lines in all graphs represent 1.25-, 1.5- and 2-fold higher (red) or lower (blue) than the linear model. (**h**) Coded mutations show a much higher contribution to the total interaction score per residue than do noncoded mutations. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers.

strategy is not hugely disruptive to the overall protein tertiary structure. This region is predicted to have a helical secondary structure that is separated from the TPR-domain-containing section of the BBS8 protein by disordered linker region (**Fig. 4a**). Int-Seq highlights the key role of a glutamic acid at position 6 that, when mutated to a lysine, disrupts the BBS9 interaction. The interaction-mediating region then centers around four positively charged amino acids at positions 16–19 that show large signals upon charge switching to glutamic acid (**Fig. 4b**). To validate the Int-Seq result, we individually re-cloned selected variants and performed a classical Y2H experiment. All ten tested interaction-disrupting mutations validated, as did two more C-terminal mutations that showed no signal in the sequencing results (**Fig. 4b**, inset).

We next examined in detail the BBS1–BBS9 interaction from the perspective of the BBS1, the most commonly identified BBS mutant gene. We found that BBS9 interacted with BBS1 through its C terminus, a region that has no annotated functional domain yet is predicted to have low disorder and mixed secondary

structure (**Fig. 5a**). Using Int-Seq, we identified a complex mutagenic profile spanning a 71-amino-acid interaction-mediating region that envelopes a cluster of uncharacterized disease-associated mutations (**Fig. 5b**). The interaction-disrupting mutations were mainly, but not exclusively, to large hydrophobic residues with low relative solvent accessibility; and roughly equal weight was carried by both charged mutant residues.

To validate the Int-Seq results, we created 20 mutations that were identified as interaction disruptive and three mutations that are annotated in areas of no signal (**Fig. 5b** and **Supplementary Fig. 5a**). Of these individual mutations, 21 reconfirmed the phenotype reported by Int-Seq in individual Y2H retests. Interestingly, a mutation derived from a patient with BBS lying directly between two areas of Int-Seq signal did not disrupt the interaction with BBS9 (R512H, Mutant 3, **Fig. 5b**). Individually validated Int-Seq mutations flank R512 just nine and seven amino acids N- and C-terminally, respectively; this highlights the requirement for high-resolution data to support mechanistic interpretations of genetic variation. This mutation was identified in the patient with BBS as
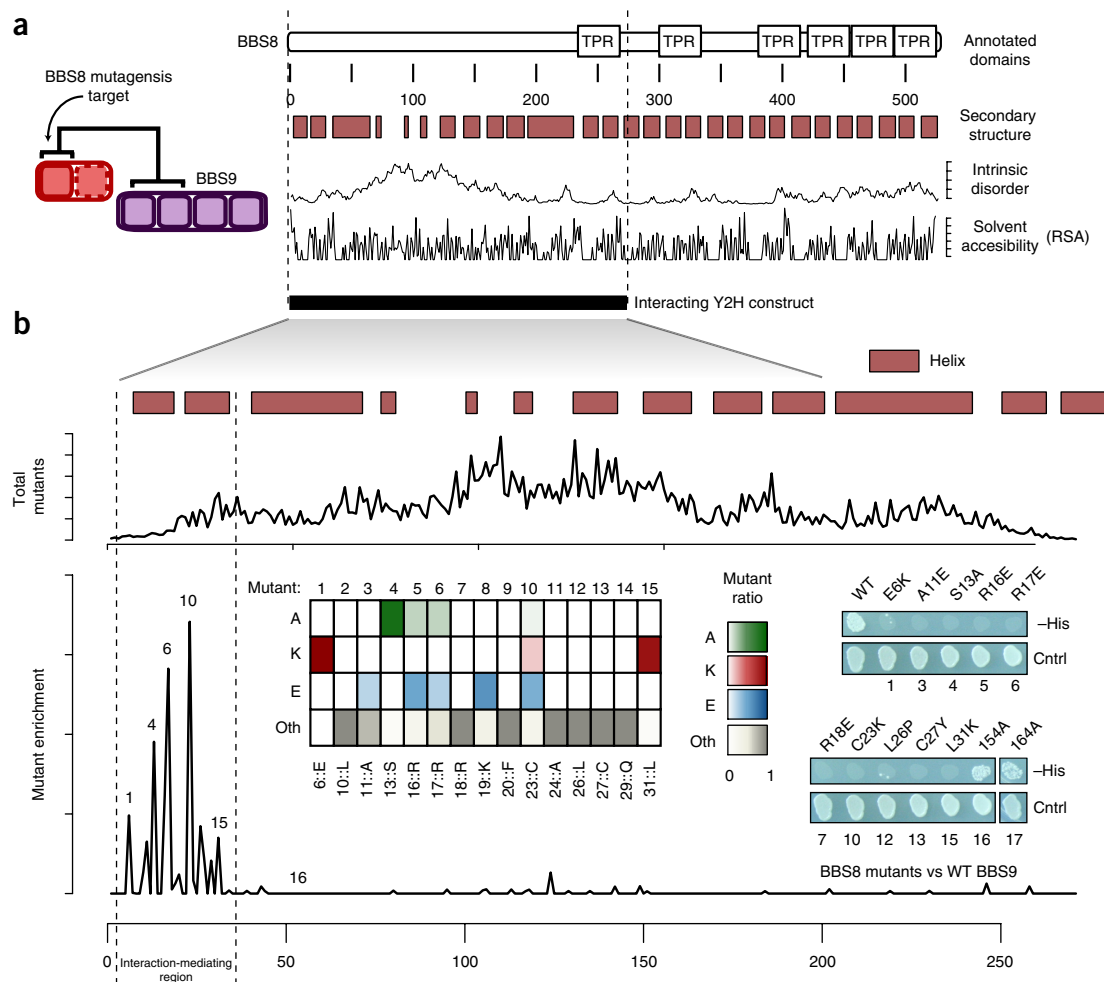
Figure 4 | Int-Seq identifies the N-terminal alpha helical region of BBS8 as required for BBS9 interaction. (**a**) The N-terminal half of BBS8 interacted with the full-length BBS9 construct and was subject to Int-Seq analysis. (**b**) Int-Seq mutant enrichment profile. Heatmap shows the proportion of each type of mutation that contributed to single peaks; example peaks are numbered according to their corresponding mutant. Inset, Y2H validation of selected mutations from the Int-Seq analysis with ten disruptive and two neutral mutations (mutants 16, 17). Mutants 2, 8, 9, 11, and 14 on heatmap were not individually retested.

heterozygous and alongside two other mutations in the chaperone BBS12[22]. Our method provides direct experimental evidence that can be used in conjunction with genetic data to aid prioritization of patient-sequenced mutations or direct hypothesis-driven studies. In contrast to the R512 interaction null variant, both the Int-Seq-identified and BBS- patient-sequenced mutations at leucine 503 prevented the interaction with BBS9. The L503H variant was identified as heterozygous in a patient assigned no other BBS mutations from a targeted genomic sequencing approach[23].

To further scrutinize the Int-Seq results, we employed a medium-throughput immunoprecipitation LUMIER-type assay to validate the impact of the mutations on protein interactions in human cell culture[24]. The immunoprecipitation results exactly mirrored the individual Y2H retests, with 20 out of 22 tested mutations reconfirming (**Fig. 5b**, inset (ii) and **Supplementary Fig. 5a**). In the iMCD3 cell line, the primary cilium can be observed as a single rod-like structure extruding from the cell membrane using anti monoacetylated tubulin antibody staining. While BBSome components have been shown to localize to the body and base of the primary cilium[25], in our system wild-type

constructs localized solely to the base (**Fig. 6a,b**). This localization was lost for the BBS1-L503E mutation. As such, Int-Seq provides direct mechanistic evidence to support the hypothesis that patient-derived mutation data at L503 may cause loss of BBS1 from the BBSome complex which, at least in part, may be causative of the displayed clinical phenotype.

Within the entire BBS amino-acid-resolution network, we identified 1,036 mutations across 680 residues that disrupted interactions between BBSome components. These residues can be generally characterized as preferentially locating to predicted buried regions of protein sequence (**Fig. 6c**), with residues showing stronger signals generally predicted as less solvent accessible than those with weaker signals (**Fig. 6d**). However, solvent accessibility alone is not predictive of interaction-disrupting potential, as many more buried residues show no signal in our assay (**Fig. 6c**, inset). To globally validate the approach, we generated a total of 56 mutations across all components of the BBS network covering 11 interaction-mediating protein regions. Mutants were generated both to validate the Int-Seq profiles and to select annotated disease mutations. Through systematic Y2H retesting we observed
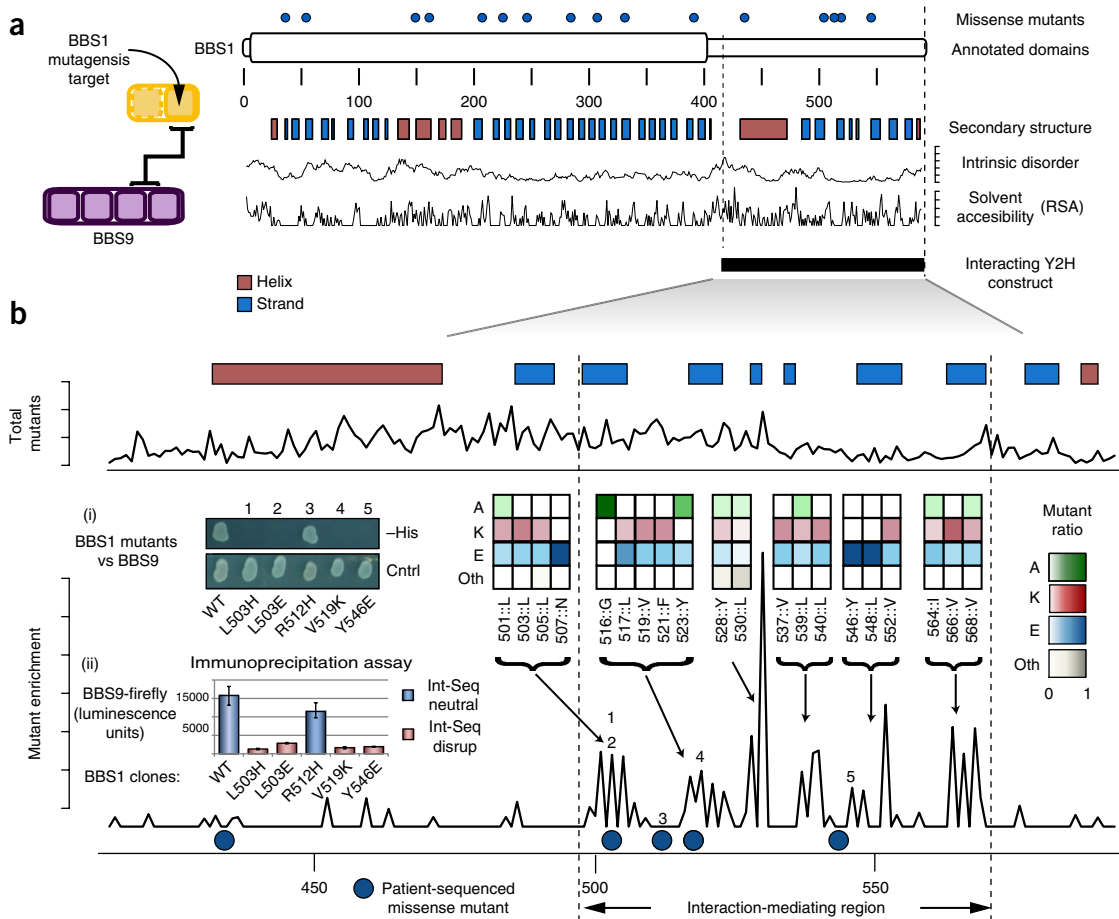
Figure 5 | Int-Seq functionally characterizes patient-derived BBS1 mutations. (**a**) The C-terminal section of BBS1 interacted with the C-terminal BBS9 construct and was subjected to Int-Seq analysis. (**b**) Int-Seq mutant enrichment profile. Heatmap shows the proportion of each type of mutation that contributed to single peaks. Inset (i), Y2H validation of three selected disruptive mutations from the Int-Seq analysis (mutants 2, 4, 5), one patient mutation predicted to be disruptive (mutant 1) and one patient mutation predicted to be neutral (mutant 3). Inset (ii), LUMIER-style immunoprecipitation assay of multiple BBS1 clones performed in HEK293T cells. Error bars represent highest and lowest luminescence reading from experimental triplicate. Full Y2H validation and LUMIER panels presented in **Supplementary Figure 5**.

a validation rate of 91% for positive and 86% for negative Int-Seq interactions (**Fig. 6e** and **Supplementary Data 1**). We further assayed 42 of these mutations in immunoprecipitation assays, and 39 of the 45 tested interactions reconfirmed the original Int-Seq phenotype (**Supplementary Fig. 5b** and **Supplementary Data 1**). This high reconfirmation rate in a completely orthologous interaction system not only highlights the quality of the Int-Seq data set, but it also validates our approach to identifying functionally relevant mutations in human cells. Furthermore, the majority of the mutations were not destabilizing the protein, as 37 of the 42 (88%) mutant proteins tested in cell culture did not have substantially reduced expression levels in comparison to the wild-type protein (**Supplementary Fig. 5c**).

In this global network we were able to identify and mechanistically characterize a second cluster of disease mutations as mediating the interaction between BBS4 and the small subunit BBS18 (**Fig. 6f** and **Supplementary Fig. 6**). Seven mutations in three BBS4 TPR domains were enveloped by the Int-Seq signal, with four mutants showing disruption at the specific disease mutant residue—two of which (L327P and D368G (heterozygous and homozygous in patients, respectively)) exactly matched disease

mutations (**Supplementary Fig. 6**). As BBS18, like BBS1, is one of the more peripheral subunits of the BBSome, we tested the localization of two BBS18 mutants disrupting the BBS4 interaction in the iMCD3 cells. Instead of localization to the base of the cilium, diffuse cytoplasmic signals were observed for the K27Y and L34A EYFP-tagged variants (**Fig. 6b**).

Furthermore, four mutations distributed across BBS5 (L50R, R56G, G72S and N184S) and one mutation in BBS2 (A504V) can be also mechanistically characterized as disrupting the BBS9 interaction (**Supplementary Figs. 7** and **8**). Overall, we have characterized residues required to maintain interactions for both peripheral BBSome components and the complex interplay within the BBS2-7-9 core. We believe this to be the first amino-acid-resolution perturbation map of an entire multisubunit human protein complex.

DISCUSSION
Here we have dissected the binary protein interaction map of the BBSome at an amino-acid resolution using Int-Seq. Generation of a synthetic promoter 'off switch' using the *ADE2* reporter gene in standard Y2H strains allows efficient, scalable forward and reverse
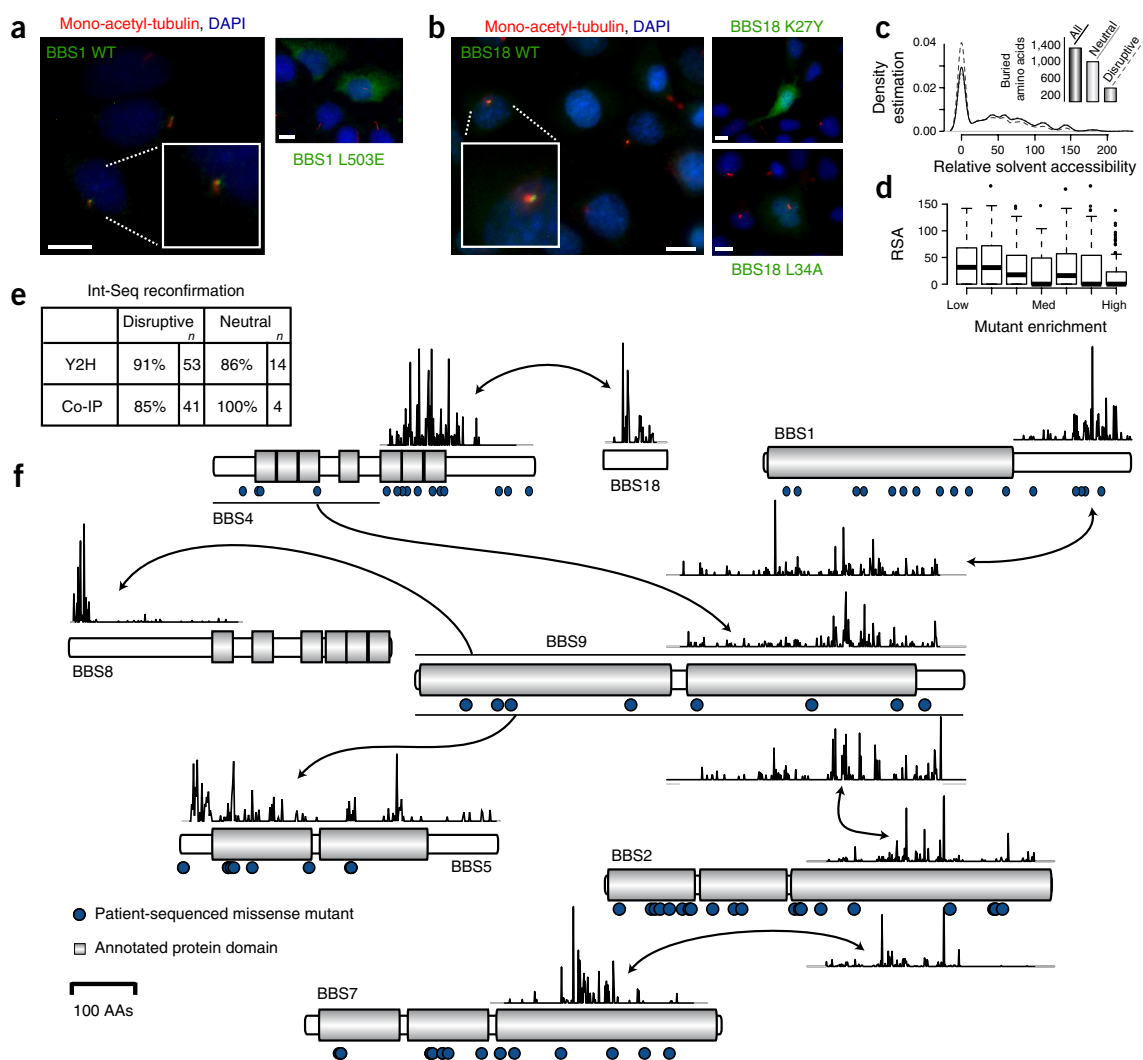
Figure 6 | The amino-acid-resolution interaction landscape of the BBSome. Subcellular localization of EYFP BBS1 (**a**) and BBS18 (**b**) mutants identified by the Int-Seq approach in serum-starved iMCD3 cells. The primary cilium was stained with anti-monoacetylated tubulin antibodies (red). BBS components in green; DAPI stain in blue. White scale bars, 10 μm. (**c**) Density distribution of residues with a mutant enrichment score >0 (Int-Seq disruptive) in comparison to neutral residues (mutant enrichment score = 0) over solvent accessibility. Inset, histogram of buried amino acid residues defined by Int-Seq as neutral or disruptive. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers. (**d**) Distribution of solvent accessibility values (RSA) across the range of Int-Seq-defined mutant enrichment scores. (**e**) Reconfirmation percentages for the global Int-Seq network for both the pairwise Y2H and coimmunoprecipitation (co-IP) assay of individually generated point mutations. *n* represents the number of interactions tested. (**f**) The entire amino-acid-resolution landscape of the BBSome as defined by Int-Seq. Mutant enrichment across each construct depicted by peak height. Gray lines at the end of each construct are residues that did not conform to the linear model. AA, amino acid.

interaction screens in a single experimental setup. To the best of our knowledge, R2H systems[11,13] have never been employed systematically on account of extensive background yeast growth. Therefore, the majority of assayed interactions would display weak or no growth phenotypes and lead to limited enrichment of disruptive residues from complex libraries, analogous to the passive dropout of disruptive mutants during a classic forward Y2H interaction screen[26,27]. For example, taking this forward approach to the high-affinity heterodimeric BARD1–BRCA1 interaction only led to small effect sizes outside of four metal ion chelating residues known to be crucial for BRCA1 structural integrity[26]. Using Int-Seq, our genetic perturbation profiles spanned dozens of amino acids across multiple interacting protein pairs, a result that also contrasts the

recent approach characterizing a panel of individual mutations spread across many proteins[4] (**Supplementary Data 1**).

While these profiles, with over 1,000 mutations across eight proteins showing a disruptive Y2H signal, can be used for a posteriori variant prioritization, they were generated as a proof of principle using three programmed mutations per residue. As such, they likely represent minimal perturbation profiles that can be refined with more extensive directed mutagenesis experiments. More amino-acid variants at each position or combinations of mutations could be adopted in future developments, which would likely produce more comprehensive profiles. Indeed, several deep mutagenesis scanning approaches have been designed to assay all possible amino-acid variants within small protein sequences.

Protein–protein interactions have been leveraged as functional filters of these complex variant libraries to investigate fundamental biological processes such as protein domain thermodynamic stability[28], the evolution of protein interaction specificity[29] and ligand-specificity class switching[30]. Comprehensive mutant libraries have also been used in genetic rescue experiments to identify key functional residues in the yeast RNA-binding protein Pab1[31] and the human nuclear receptor PPARγ[32]. These studies have pioneered deep mutagenesis scanning approaches, yet they were restricted to testing specific hypotheses and are not easily scalable. In contrast, as Int-Seq uses Y2H technology, it is rapidly scalable to proteome-wide analyses, and it operates with fewer restrictions on target systems within the field of protein–protein interactions[17,33].

The functional Int-Seq profiles provide a resource independent of structural information to define exact residues required to maintain physiologically relevant molecular interactions. This contrasts with previous approaches that used 3D structure information to model protein interaction networks. These networks were integrated with genetic data to suggest mechanistic roles for disease-related missense mutations annotated at interaction interfaces[34,35]. Despite these advances, limitations remain with characterizing genetic variation using structural approaches. Protein structures, and by extension their molecular interactions, are maintained through extensive interlinked amino-acid contact networks. Mutations distal from binding sites were shown to alter peptide-binding specificity in protein kinases[36] or PDZ domains[30] and also to impact protein homo-oligomerisation[37]. Defining amino-acid residues that are required for maintaining molecular interactions away from direct interface contacts using structural data alone remains a major challenge. Furthermore, up to 40% of the human proteome is defined as intrinsically unstructured[38] and may not be amenable to structural approaches at all. Int-Seq could generate selective yet comprehensive perturbation profiles across dozens of amino acids without any structure information. However, one drawback of this approach is that for single protein pairs it is difficult to distinguish mutations that ablate all protein interactions through protein structure perturbation from interaction-specific disrupting mutations. As more proteins are scanned using this approach, comparative analysis of multiple perturbation profiles for single proteins should facilitate finer dissection of mutational impact both within the classical structure–function paradigm and on the complex molecular interactions mediated by disordered protein sequences[38].

Current sequencing projects can identify variants associated with rare and genetically complex traits[39]. These projects will likely extend the panel of known modifier loci and inform models of complex trait inheritance in BBS[40–42]. To aid a better understanding of trait variation and disease phenotypes, we will need to characterize the molecular impact of large panels of genetic variants of unknown significance[43], including nucleotide variants unique to an individual. Through systematically interrogating protein–protein interactions at amino-acid resolution, we have presented a method that can address a large part of the genetic variation that impacts this universal protein function. Int-Seq can function across many distinct interactions and can provide *de novo* perturbation profiles as framework to aid prioritization of past and novel genetic variants in healthy and diseased individuals.

METHODS
Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS
U.S. supervised the project. J.W., L.A., V.C.-M. and Z.Ö. created the yeast strains. J.W. and L.A. performed the Y2H BBSome network construction. J.W. and U.S. developed the Int-Seq experimental protocol. J.W. performed the Int-Seq experimental protocol, Y2H mutant validation, LUMIER experiments and fluorescence microscopy. B.T. contributed tools and reagents. J.W. designed and undertook all bioinformatics analysis and figure generation. J.W. and U.S. wrote the paper. All authors contributed to paper feedback.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. **Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.**

1. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
2. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015).
3. Woodsmith, J. & Stelzl, U. Studying post-translational modifications with protein interaction networks. *Curr. Opin. Struct. Biol.* **24**, 34–44 (2014).
4. Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
5. Wei, X. *et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* **10**, e1004819 (2014).
6. Kitzman, J.O., Starita, L.M., Lo, R.S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).
7. Wrenbeck, E.E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* **13**, 928–930 (2016).
8. Nachury, M.V. *et al.* A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. *Cell* **129**, 1201–1213 (2007).
9. Malicki, J.J. & Johnson, C.A. The cilium: cellular antenna and central processing unit. *Trends Cell Biol.* **27**, 126–140 (2017).
10. Nachury, M.V., Seeley, E.S. & Jin, H. Trafficking to the ciliary membrane: how to get across the periciliary diffusion barrier? *Annu. Rev. Cell Dev. Biol.* **26**, 59–87 (2010).
11. Shih, H.M. *et al.* A positive genetic selection for disrupting protein–protein interactions: identification of CREB mutations that prevent association with the coactivator CBP. *Proc. Natl. Acad. Sci. USA* **93**, 13896–13901 (1996).
12. Ear, P.H. & Michnick, S.W. A general life-death selection strategy for dissecting protein functions. *Nat. Methods* **6**, 813–816 (2009).
13. Vidal, M., Braun, P., Chen, E., Boeke, J.D. & Harlow, E. Genetic characterization of a mammalian protein–protein interaction domain by using a yeast reverse two-hybrid system. *Proc. Natl. Acad. Sci. USA* **93**, 10321–10326 (1996).
14. Gronemeyer, T. *et al.* A split-ubiquitin based strategy selecting for protein complex-interfering mutations. *G3 (Bethesda)* **6**, 2809–2815 (2016).
15. Gedvilaite, A. & Sasnauskas, K. Control of the expression of the *ADE2* gene of the yeast *Saccharomyces cerevisiae*. *Curr. Genet.* **25**, 475–479 (1994).
16. Worseck, J.M., Grossmann, A., Weimann, M., Hegele, A. & Stelzl, U. A stringent yeast two-hybrid matrix screening approach for protein-protein interaction discovery. *Methods Mol. Biol.* **812**, 63–87 (2012).
17. Weimann, M. *et al.* A Y2H-seq approach defines the human protein methyltransferase interactome. *Nat. Methods* **10**, 339–342 (2013).

18. Zhang, Q., Yu, D., Seo, S., Stone, E.M. & Sheffield, V.C. Intrinsic protein-protein interaction-mediated and chaperonin-assisted sequential assembly of stable bardet-biedl syndrome protein complex, the BBSome. *J. Biol. Chem.* **287**, 20625–20635 (2012).

19. Mourão, A., Nager, A.R., Nachury, M.V. & Lorentzen, E. Structural basis for membrane targeting of the BBSome by ARL6. *Nat. Struct. Mol. Biol.* **21**, 1035–1041 (2014).

20. Katoh, Y., Nozaki, S., Hartanto, D., Miyano, R. & Nakayama, K. Architectures of multisubunit complexes revealed by a visible immunoprecipitation assay using fluorescent fusion proteins. *J. Cell Sci.* **128**, 2351–2362 (2015).

21. Boersma, M.D., Sadowsky, J.D., Tomita, Y.A. & Gellman, S.H. Hydrophile scanning as a complement to alanine scanning for exploring and manipulating protein-protein recognition. Application to the Bim BH3 domain. *Protein Sci.* **17**, 1232–1240 (2008).

22. Chen, J. *et al.* Molecular analysis of Bardet–Biedl syndrome families: report of 21 novel mutations in 10 genes. *Invest. Ophthalmol. Vis. Sci.* **52**, 5317–5324 (2011).

23. Beales, P.L. *et al.* Genetic interaction of *BBS1* mutations with alleles at other *BBS* loci can result in non-Mendelian Bardet-Biedl syndrome. *Am. J. Hum. Genet.* **72**, 1187–1199 (2003).

24. Hegele, A. *et al.* Dynamic protein-protein interaction wiring of the human spliceosome. *Mol. Cell* **45**, 567–580 (2012).

25. Liew, G.M. *et al.* The intraflagellar transport protein IFT27 promotes BBSome exit from cilia through the GTPase ARL6/BBS3. *Dev. Cell* **31**, 265–278 (2014).

26. Starita, L.M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).

27. Melamed, D., Young, D.L., Miller, C.R. & Fields, S. Combining natural sequence variation with high throughput mutational data to reveal protein interaction sites. *PLoS Genet.* **11**, e1004918 (2015).

28. Araya, C.L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. USA* **109**, 16858–16863 (2012).

29. Aakre, C.D. *et al.* Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).

30. Raman, A.S., White, K.I. & Ranganathan, R. Origins of allostery and evolvability in proteins: a case study. *Cell* **166**, 468–480 (2016).

31. Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).

32. Majithia, A.R. *et al.* Prospective functional classification of all possible missense variants in *PPARG*. *Nat. Genet.* **48**, 1570–1575 (2016).

33. Trigg, S.A. *et al.* CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. *Nat. Methods* **14**, 819–825 (2017).

34. Wang, X. *et al.* Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–164 (2012).

35. Mosca, R. *et al.* dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods* **12**, 167–168 (2015).

36. Creixell, P. *et al.* Unmasking determinants of specificity in the human kinome. *Cell* **163**, 187–201 (2015).

37. Perica, T. *et al.* Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science* **346**, 1254346 (2014).

38. Babu, M.M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44**, 1185–1200 (2016).

39. Boycott, K.M., Vanstone, M.R., Bulman, D.E. & MacKenzie, A.E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14**, 681–691 (2013).

40. Abu-Safieh, L. *et al.* In search of triallelism in Bardet––Biedl syndrome. *Eur. J. Hum. Genet.* **20**, 420–427 (2012).

41. Davis, E.E. *et al. TTC21B* contributes both causal and modifying alleles across the ciliopathy spectrum. *Nat. Genet.* **43**, 189–196 (2011).

42. Zhang, Y. *et al.* BBS mutations modify phenotypic expression of *CEP290*-related ciliopathies. *Hum. Mol. Genet.* **23**, 40–51 (2014).

43. Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* **14**, 168–178 (2013).

ONLINE METHODS

Int-Seq yeast strain generation. From the two major variations of reverse-two-hybrid systems, TetR-based auxotrophic 'off' switches have only been shown to function with single interactions such as CREB binding to its coactivator CBP[11], while each individual interacting pair requires careful toxic metabolite titration when inhibiting the URA3 enzyme to induce lethality[12,13]. We reasoned that incorporating modern synthetic biology principles into the design of a TetR-driven repressive switch based on our classic Y2H system would be more amenable to screening approaches.

We first conjugated LexA[4] and LexA[8] binding cassettes to a minimal HIS3 promoter that was PCR amplified from the genome of L40 Y2H strains previously used in the lab[24,44]. These cassettes were then both PCR stitched to the Tet repressor gene containing a C-terminal nuclear localization signal and attached to a transcription termination sequence to generate the final single LexA[x]::TetR cassettes. These were then cloned into the pAG25 vector to stitch the TetR to the nourseothricin-resistance cassette. The entire cassette was then PCR amplified and inserted in the *MET2* locus via homologous recombination, and nourseothricin-resistant colonies were selected and tested for TetR production in strains transformed with an autoactive bait via western blot (**Supplementary Fig. 1a**). The LexA[8] DNA-binding domain drove higher expression levels of the Tet repressor, and doubling the number of LexA[8]::TetR reporter genes further sensitized the interaction-mediated repression (**Supplementary Fig. 1a,b**). Two LexA[8]::TetR_NLS constructs were therefore cloned in parallel into the final pAG25 vector to generate the Int-Seq strains (**Supplementary Fig. 9a**). The TetR binds to its cognate tetO DNA sequence, with both position and number of tetO repeats having been shown to be important in target gene repression[45]. We therefore sought to create a sensitive TetR-mediated repressive switch in a minimal ADE2 promoter. In agreement with previous literature[15], ADE2 promoter truncations were tolerated up to two crucial TGACTC boxes (**Supplementary Fig. 1c**). We then optimized interaction-mediated repression of the *ADE2* gene through strategic placement of increasing tetO repeats while maintaining wild-type growth on media lacking adenine (**Supplementary Fig. 1d**). In total, five copies of the tetO sequence were inserted into the minimal promoter of the *ADE2* gene and replaced endogenous DNA sequence (*S. cerevisiae* coordinates chrXV: 565986.564476) using a combination of stitch PCR overlap extension and conventional cloning (cloning strategy presented in **Supplementary Fig. 9b**). The modified ADE2 cassette was then cloned into the HO-hisG-URA3 vector using conventional cloning with the restriction enzymes BamHI and EcoRI. The *ADE2*-containing vector was then linearized using a single NotI digest and inserted into the *HO* locus before selection for URA positive colonies[46]. Colonies were then tested for yeast growth on media lacking adenine to verify normal function of the synthetic promoter. Final MATa and MATα strains have the following genotypes: *MATa: his3Δ200,trp1-901, leu2 3-112, ade2, lys2-801am, LYS2::(lexOp)[4]-HIS3TATA-HIS3, met2::((lexOp)[8]-HIS3TATA-TetR_NLS)[2]::KanMX, HO-hisG-URA3::TetO[5]-ADE2. MATα: his3-Δ200, trp1-901, ade2, leu2 3-112, LYS2::(lexOp)[4]-HIS3TATA-HIS3, met2::((lexOp)[8]-HIS3TATA-TetR_NLS)[2]::KanMX, HO-hisG-URA3::TetO[5]-ADE2.*

Bardet–Biedl syndrome constructs generation. For large genes we created protein fragments based on tertiary and secondary structure prediction. We used HHpred for homology-based tertiary structure prediction (https://toolkit.tuebingen.mpg.de/#/tools/hhpred). We then used a secondary structure prediction tool (http://www.predictprotein.org/) to design cut sites in predicted loops. In addition, several truncations were based on constructs that expressed well in *E. coli* (data not shown), or when no domain boundaries were found divided roughly in equal size constructs based on secondary structure predictions.

Yeast two-hybrid network generation. Genomic variation in protein coding sequences has been shown to impact both entire protein function and specific protein–protein interactions alone[4,47]. It is therefore crucial to generate a comprehensive interaction network in order to cover as much of the potential interaction space as possible within a set of proteins. As such, we screened all clones in both N- and C-terminal tag bait and prey vectors. First, we constructed a novel C-terminal prey plasmid (pCBDU) based on our ultralow-expression N-terminal tag pACT4 prey vector used extensively elsewhere[17]. We then performed a 96-well-format matrix screen as previously described[16,44]. Each pairwise mating was repeated at least twice, and only reproducible interactions were taken forward into the final network.

On-chip primer design and mutagenic library preparation. Here we adapted the method of Kitzman *et al.*[6] to make it more amenable to high-throughput screening using Gateway Cloning; however, other recent advances in massively parallel programmed mutagenesis could streamline this process further[7]. Each DNA oligonucleotide sequence was designed as a 60-mer flanked by two distinct 22-mer primer tags specific to each open reading frame (ORF) targeted. The mutant codon triplicates were selected to be optimal for yeast expression (A: GCT, K: AAA, E: GAA) and were flanked by 27 WT nucleotides 5′ and 30 WT nucleotides 3′. The first and last five amino acids from each ORF were not targeted by mutagenesis on account of constrictions imposed by primer design. Mutagenesis was essentially carried out as in Kitzman *et al.*[6] with the following alterations. To create WT template strands, each ORF was amplified from a gateway destination vector using generic att site primers, which allowed direct transfer of the final PCR products into a gateway donor vector via a BP reaction. A standard midi-prep was carried out on bacterial colonies collected post each cloning step, with the number of colonies collected approximating 100× the number of mutations programmed per ORF. Vectors used did not contain a unique barcode tag to identify individual amplicons as previous reported[6]; rather, experimentally identified enriched mutants were later computationally reconstructed from the sequencing data (see "Sequence analysis pipeline"). The full-length 887 amino acid BBS9 mutant library could not be generated because of PCR failures in the course of the protocol; as such, only the C-terminal 482 amino acids were mutated here, which resulted in the BBS8 and BBS5 interactions only being tested in one direction.

Random mutagenesis clone generation. Random mutagenesis was undertaken using a standard taq polymerase (Bioline Biotaq) with a low MnCl$_2$ concentration (0.1 mM) spiked in to a 9 cycle (BBS18) or 12 cycle (BBS4) PCR reaction. To vastly reduce the

number of nonfunctional truncated Y2H constructs caused by premature stop codons, we cloned the resulting PCR products into pDONR Express, which allows enrichment of full-length clones through C-terminal aminoglycoside phosphotransferase tagging of ORFs followed by selection on kanamycin LB agar. pDONR Express cloning was carried out as described in the product manual (Invitrogen). Kanamycin concentration was first optimized for each construct through IPTG induction of wild-type protein expression before colony collection.

To estimate the mutant coverage of the libraries generated using both programmed and random mutagenesis strategies, we analyzed all unique 150-mer sequences at differing quality cutoffs. The reads were first grouped depending on the number of amino acid alterations identified (1,2,3,4 or ≥5 mutations). The number of times each codon mutation was identified was then collated for each of the subgroups. To generate high-confidence coverage estimates, only mutations that were observed ≥3 times were taken forward to the final calculations. Using 150-mers that contain only one mutation, the coverage of residues ranged from 85–100% for AKE mutations across the programmed mutant libraries (**Supplementary Fig. 3a**). As expected, the AKE mutation coverage in the random mutant libraries was lower, at 50% and 64% (**Supplementary Fig. 3a**, red dots). Using 150-mers that contain only one mutation, the percentage of all possible AKE mutations sequenced ranged from 48–99% for the programmed mutant libraries (**Supplementary Fig. 3b**). The random mutant libraries showed a much lower percentage of possible mutants (**Supplementary Fig. 3a**, red dots) and showed a slight increase in the percentage of all other mutations in comparison to the programmed mutant libraries (**Supplementary Fig. 3c**, with the exception of one programed library).

Int-Seq yeast mating protocol. For a step-by-step guide, please see the **Supplementary Protocol**[48]. Mutant libraries were transformed into yeast using a standard lithium acetate protocol, with the number of colonies collected approximating 100× the number of mutations programmed. Yeast colonies were collected in nitrogen base (NB) minus amino acid (–AA) liquid media to give a highly dense mutant library solution. Interaction counterpart wild-type construct yeast were freshly grown through overnight incubation in growth-selective media at 30 °C before centrifugation and resuspension in NB – AA media to concentrate the yeast cells to an ~OD_{600} 40–80. To saturate the mutant library, wild-type and mutant libraries were then mixed in a 2:1 ratio by cell number and mated on YPDA media before incubation at 30 °C for 24 h. For higher throughput experiments such as reported here, yeast mating can be undertaken in a six-well plate using 40 μl and 20 μl of equal concentration wild-type and mutant libraries, respectively. Yeast cells were then collected in 350 μl NB – AA media before spreading 30 μl on each of two large agar diploid-selective plates (NB-tryptophan,-leucine) in a total volume of 3 ml NB – AA per plate. Yeast were then incubated at 30 °C for 48 h to allow independent, noncompetitive colony growth of yeast to maintain library variation. Yeast were then collected in 10 ml NB – AA per plate, diluted to ~OD_{600} 0.2 in 6 ml NB – AA and plated onto 2× NB-adenine,-leucine,-tryptophan large agar plates media before incubation at 30 °C for 24–48 h followed by yeast colony collection. Yeast samples were then lysed using a standard alkaline lysis protocol, and the plasmid DNA was isolated through ethanol precipitation followed by a phenol extraction. The mutant ORF libraries were then amplified through a proof-reading KOD polymerase PCR (Novagen) using either vector-specific or ORF-specific primers.

Next-generation sequencing. PCR products were fragmented, barcoded and sequenced in a 150-base-pair paired-end read run on an Illumina NextSeq 500. 159, 937 and 962 paired end reads were distributed across 12 controls and 35 sample interactions.

Sequence analysis pipeline. Initially, unique 150-mer single and paired-end sequence fasta files were generated from the raw fastq files. Unique single reads represented ~41% of the total number of 150-mer reads present in the unique paired-end files. Therefore they were used to create the initial mutant profiles, as the alignments were over twice as efficient (**Supplementary Fig. 10a**). Unique paired-end reads were collected to allow more comprehensive reconstruction of the enriched mutants later in the pipeline. 150-mer sequences were then aligned against a specific construct used in each interaction assay using the rmapper tool in the sensitive SCHRiMP alignment software package (**Supplementary Fig. 10b**[49]). Aligned rmapper files were then condensed to obtain recall statistics of mutant codons in each position into a matrix format for importing into the R statistical analysis software (https://www.r-project.org/ and **Supplementary Fig. 10c**). Previous single-amino-acid-resolution data analysis discarded all reads that contained two or more mutations[6] and would here also result in discarding between 40% and 80% of the data per sample. As such, we sought to develop a sequencing analysis pipeline that could model variation in the data independent of the number of mutations per 150-mer. Furthermore, such an approach would be insensitive to the read length of the sequencing run undertaken (longer reads necessarily containing more mutations). Noncoded mutations identified here could be generated by random errors in the on-chip nucleotide synthesis, inserted during the PCR steps required to generate full-length mutants or be artifacts generated during the sequencing pipeline. While these errors are minimized, it is not currently possible to entirely eradicate them. We therefore modeled variation across all mutations, and we took forward all those that showed deviation from the linear model.

We observed that for any given position there was an expected number of sequences for any given codon mutation (**Fig. 3b–d**). This allowed us to empirically model the variation for each possible mutation over all positions and robustly identify enriched mutations. The R-based linear model function "lm" was used to model 99% of the data (to prevent extreme outliers from strongly influencing the model), and the enrichment was calculated as:

Amino acid codon enrichment = observed total sequences for codon x / expected total sequences for codon x

Furthermore, a high-confidence cutoff was used to remove sequence with very low recall statistics in both the input and output. For the noncoded random mutations, a cutoff of 10× the median codon sequences across all positions was used; and for the coded AKE mutations, a cutoff of 50 sequences was used. For some samples, the start and end positions of the PCR products poorly fitted the linear model and were as such discarded from the final analysis. After processing, each codon had an enrichment score across all positions in a given clone (**Supplementary Fig. 10d**).

The vast majority of variation between nonselective samples was within two-fold variation of the linear model. Therefore, to produce a high-confidence data set, only exact codons that showed repeated enrichment above a two-fold linear model enrichment cutoff at any given position were taken forward. Between two and four samples were averaged per interaction to give an initial mutagenic perturbation profile, taking mutations into the final profiles that were present in at least two experimental repetitions. For full recall statistics see **Supplementary Data 1**.

As all reads were used in this initial analysis regardless of the number of mutations, the obtained profiles were filtered for enriched mutations that stem from reads containing double mutations or premature stop codons caused by indels. To this end, the paired-end reads for each enriched mutant were extracted and reanalyzed for cosegregating mutations. Here, paired-end reads were used to cover a higher percentage of the construct of interest per unique sequence pair. The proportion of paired-end reads containing only the enriched mutant of interest was plotted against the proportion of paired-end reads that contained the most frequent secondary mutation (**Supplementary Fig. 10e**). Most of the enriched mutants contain a very small number of cosegregating secondary mutations, irrespective of the proportion of paired-end reads that contained only the mutant of interest (**Supplementary Fig. 10e**). Enriched mutation paired-end reads that contained high-frequency secondary mutations were then discarded before generation of the final, high-confident mutant profile (**Supplementary Fig. 10f**–**h**). An identical analysis was carried out to remove any peaks co-occurring with insertions or deletions, with an added filter that removes peaks that are routinely identified as insertion targets regardless of their co-occurrence in the data set.

LUMIER coimmunoprecipitation experiments. BBS ORFs were transferred to either firefly luciferase-V5 fusion vectors (pcDNA3.1V5-Fire) or protein-A fusion vectors (pcDNA3.1PA-D57) using standard Gateway cloning procedures. For co-IP assays, 3×10^4 HEK293T cells were transiently cotransfected with firefly (75 ng) and protein A (PA; 75 ng) plasmid DNA using Lipofectamine 2000 (Invitrogen) in each well of a 96-well plate. Cells were lysed 36 h after transfection in 100 µl HEPES buffer (50 mM HEPES pH 7.4, 150 mM NaCl, 1 mM EDTA, 10% glycerol, 1% Triton X-100 and protease inhibitor (Roche) for 30 min at 4 °C. Protein complexes were precipitated from 80 µl cleared cell extract in IgG-coated microtiter plates for 3 h at 4 °C and rapidly washed three times with 100 µl PBS. The binding of the firefly-V5-tagged fusion protein (co-IP) to the PA-tagged fusion protein was assessed by measuring the firefly luciferase activity in a luminescence plate reader (Beckmann DTX880, Bright-Glo Luciferase Assay (Promega)). Assays were performed twice and in triplicate transfections. PA protein binding/expression (IP) was assessed via western blotting using anti PA-antibody (#61-1620, Thermofisher).

Mutant validation. Enriched mutants that perturb the wild-type interaction were generated in pDONR constructs via site-directed mutagenesis before Gateway cloning into the Y2H vectors required for retesting of the initial interaction.

Cell culture and microscopy. mIMCD-3 (ATCC CRL-2123) cells were maintained in a humidified incubator at 37 °C with 5% CO2 and cultured in DMEM/F12 media supplemented with 10% fetal bovine serum (FBS). Cells were seeded on glass coverslips and transfected with Lipofectamine 2000 at a 3:1 ratio (Lipofectamine(µl): DNA(µg)) in serum-free media. 4 h post-transfection, the transfection media were replaced with DMEM/F12 media supplemented with 10% FCS. 24 h post-transfection, media were replaced with serum-free DMEM/F12 media to induced ciliation. Cells were then fixed with 4% paraformaldehyde before permeabilization with 0.3% Triton X-100 in PBS for 10 min. EYFP signal was detected using a rabbit polyclonal anti-GFP antibody (ab6556, Abcam) followed by a donkey anti-rabbit Alexa-Flour-488 secondary antibody (711-485-152, Jackson ImmunoResearch Laboratories). Primary cilia were visualized using a mouse monoclonal anti-mono-acetylated tubulin antibody (6-11B-1, Sigma) followed by a donkey anti-mouse Alexa Fluor 647 secondary antibody (ab150111, Abcam). Nuclei were stained with DAPI (0.5 µg/ml, Sigma). Cells strongly expressing any wild type of mutant constructs showed a general cytoplasmic distribution and were therefore excluded from any further analysis.

Information on experimental design and reagents can be found in the accompanying **Life Sciences Reporting Summary**.

Materials availability. Yeast strains are available on request.

Code availability. All computational code used in the analysis is available on request.

Data availability statement. A summary of the mutant-enrichment scores calculated for each interaction is present in **Supplementary Data 1**. Raw sequencing data files have been deposited at BioProject under the identifier ID PRJNA407860.

44. Apelt, L. *et al.* Systematic protein-protein interaction analysis reveals intersubcomplex contacts in the nuclear pore complex. *Mol. Cell Proteomics* **15**, 2594–2606 (2016).
45. Murphy, K.F., Balázsi, G. & Collins, J.J. Combinatorial promoter design for engineering noisy gene expression. *Proc. Natl. Acad. Sci. USA* **104**, 12726–12731 (2007).
46. Voth, W.P., Richards, J.D., Shaw, J.M. & Stillman, D.J. Yeast vectors for integration at the *HO* locus. *Nucleic Acids Res.* **29**, E59 (2001).
47. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
48. Woodsmith, J., Apelt, L., Casado-Medrano, V., Özkan, Z., Timmermann, B. & Stelzl, U. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.110.
49. Rumble, S.M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**, e1000386 (2009).

nature research

Corresponding author(s): Ulrich Stelzl and Jonathan Woodsmith

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. **Sample size**

   Describe how sample size was determined.

   > To ensure reproducibility of the final profiles, only mutations enriched across at least two samples were taken forward into the final analysis.

2. **Data exclusions**

   Describe any data exclusions.

   > No data was excluded from the analysis

3. **Replication**

   Describe whether the experimental findings were reliably reproduced.

   > Linear modelling showed reproducibility of identifying amino acid mutations using a sequencing approach.

4. **Randomization**

   Describe how samples/organisms/participants were allocated into experimental groups.

   > NA

5. **Blinding**

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   > NA

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. **Statistical parameters**

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	☐	The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

   See the web collection on statistics for biologists for further resources and guidance.

▶ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The linear regression function in the statistical software package R was used. As was the rmapper alignment software in the SCHRiMP tool.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▶ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No restrictions

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Rabbit anti-GFP antibody (ab6556, Abcam). mouse monoclonal anti-mono-acetylated tubulin antibody (6-11B-1,Sigma). Anti protein A antibody: Rabbit anti-goat HRP conjugate (61-1620, Invitrogren)

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

ATCC: mIMCD-3 (ATCC® CRL-2123™) cells. Invitrogen: HEK293T cells

b. Describe the method of cell line authentication used.

ATCC standard methods of morphology, karyotyping, and PCR based approaches to confirm the identity of cell lines and to rule out both intra- and interspecies contamination.

c. Report whether the cell lines were tested for mycoplasma contamination.

Visualisation of DAPI staining using a 63x objective lens showed no staining other than nuclear, indicating cells are free from mycoplasma

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

None of the cell lines used are listed as commonly misidentified by ICLAC.

▶ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

NA

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

NA