Deep Learning-based Identification of Intraocular Pressure-Associated Genes Influencing Trabecular Meshwork Cell and Organelle Morphology

Connor J Greatbatch,[1] Qinyi Lu,[1] Sandy Hung,[2] Son N Tran,[3] Kristof Wing,[1] Helena Liang,[2] Xikun Han,[4] Tiger Zhou,[5] Owen M Siggs,[10,11] David A Mackey,[1,6] Guei-Sheung Liu,[1] Anthony L Cook,[7] Joseph E Powell,[8,9*] Jamie E Craig,[5*] Stuart MacGregor,[4*] Alex W Hewitt.[1,2*]

Affiliations:

1. Menzies Institute for Medical Research, University of Tasmania, Australia.
2. Centre for Eye Research Australia, University of Melbourne, Australia.
3. Department of Information and Communication Technology, University of Tasmania, Australia.
4. QIMR Berghofer Medical Research Institute, Brisbane, Australia.
5. Department of Ophthalmology, Flinders University, Flinders Medical Centre, Bedford Park, Australia.
6. Lions Eye Institute, Centre for Ophthalmology and Visual Science, University of Western Australia, Australia.
7. Wicking Dementia Research and Education Centre, University of Tasmania, Hobart, TAS, Australia.
8. Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, NSW, Australia.
9. UNSW Cellular Genomics Futures Institute, UNSW, Sydney, NSW, Australia
10. Garvan Institute of Medical Research, Sydney, New South Wales, Australia
11. School of Clinical Medicine, Faculty of Medicine and Health, UNSW Sydney, Australia

* These authors jointly supervised this work

Word count: 3492

Acknowledgements

D.A.M., J.E.C., J.E.P, S.M., and A.W.H. are supported by the Australian National Health and Medical Research Council (NHMRC) Fellowships. X.H. was supported by the University of Queensland Research Training Scholarship and QIMR Berghofer Medical Research Institute PhD Top Up Scholarship. We are grateful for funding from Australian Vision Research; a NHMRC Program grant (1150144), Partnership grant (1132454) and the Clifford Craig Foundation.

Address for Correspondence:
Professor Alex Hewitt
Menzies Institute for Medical Research
University of Tasmania, 17 Liverpool St, Hobart, TAS, 7000
Phone: 0407359824
Email: hewitt.alex@gmail.com

ABSTRACT

PURPOSE: The exact pathogenesis of primary open-angle glaucoma (POAG) is poorly understood. Genome-wide association studies (GWAS) have recently uncovered many loci associated with variation in intraocular pressure (IOP); a crucial risk factor for POAG. Artificial intelligence (AI) can be used to interrogate the effect of specific genetic knockouts on the morphology of trabecular meshwork cells (TMCs), the regulatory cells of IOP.

METHODS: Sixty-two genes at fifty-five loci associated with IOP variation were knocked out in primary TMC lines. All cells underwent high-throughput microscopy imaging after being stained with a five-channel fluorescent cell staining protocol. A convolutional neural network (CNN) was trained to distinguish between gene knockout and normal control cell images. The area under the receiver operator curve (AUC) metric was used to quantify morphological variation in gene knockouts to identify potential pathological perturbations.

RESULTS: Cells where *RALGPS1* had been perturbed demonstrated the greatest morphological variation from normal TMCs (AUC 0.851, SD 0.030), followed by *LTBP2* (AUC 0.846, SD 0.029) and *BCAS3* (AUC 0.845, SD 0.020). Of seven multi-gene loci, five had statistically significant differences in AUC ($p<0.05$) between genes, allowing for pathological gene prioritisation. The mitochondrial channel most frequently showed the greatest degree of morphological variation (33.9% of cell lines).

CONCLUSIONS: We demonstrate a robust method for functionally interrogating genome-wide association signals using high-throughput microscopy and AI. Genetic variations inducing marked morphological variation can be readily identified, allowing for the gene-based dissection of loci associated with complex traits.

INTRODUCTION

Primary open-angle glaucoma (POAG) is a blinding disease characterised by progressive degeneration of the optic nerve and retinal nerve fibre layer.[1,2] POAG is one of the leading causes of blindness globally.[3] Whilst the precise pathophysiology of glaucoma is unknown, the most important modifiable risk factor is raised intraocular pressure (IOP).[1,4] Raised IOP in POAG is primarily caused by dysfunctional aqueous humour drainage through the trabecular meshwork.[1] Family heritage studies and genome-wide association studies (GWAS) have demonstrated a genetic contribution to trabecular meshwork dysfunction in POAG; however, the exact cellular and genetic processes involved remain unknown.[1] Current treatments for POAG focus on reducing IOP by decreasing the production of aqueous humour or increasing outflow, with medications, or through the use of pressure-lowering surgery. However, there is currently no definitive cure for all patients with POAG.[5] For novel pressure-lowering treatments to be developed, the pathophysiology of raised IOP in POAG must be understood, and molecular pathways for this vision-threatening disease uncovered.

Previous research has implicated a number of genes that contribute to POAG development and variation in IOP.[1,6] Linkage analysis identified variants in the *MYOC* gene as being strongly associated with POAG.[7–9] Disease-causing mutations in this gene have been shown to cause accumulation of a misfolded protein (myocilin), resulting in endoplasmic reticulum stress in trabecular meshwork cells (TMCs) and a subsequent rise in IOP.[6] GWAS have identified numerous genetic variants associated with raised IOP, many of which have also been associated with POAG.[10,11] However, further investigation into these genetic variants is required to identify which individual genes may be affected by these variants and, thus, what cellular mechanisms may be involved. The ongoing development of artificial intelligence (AI) and deep-

learning tools such as convolutional neural networks (CNNs) provides a unique opportunity to investigate the genes of interest highlighted in GWAS and their effect on single cell morphology.

Deep learning is a rapidly advancing field of machine learning that relies on neural networks to learn abstract representations of data. A CNN is a specialised deep-learning model designed to learn features of image data. In supervised learning, the original images are labelled, allowing CNNs to learn the correct representation for a given label. Given the effectiveness of CNNs at image classification[12], they have been extensively used in the analysis of cellular morphology, which is relevant in many domains of biology and medicine such as phenotype analysis,[13,14] drug screening,[15,16] and cell sorting.[17,18]

This study aimed to train a CNN to distinguish between primary TMCs that had specific genes from selected IOP-associated loci,[10,11] knocked out using CRISPR/Cas and control TMCs transfected with non-targeting guide RNAs. The accuracy, as measured by the area under the receiver operator curve (AUC) metric, was used to quantify variation in morphological profiles between target gene knockouts and control cells. This high throughput approach uncovered genes at IOP loci, which, when perturbed, lead to marked variation in TMC morphology.

METHODS

Cell culture and passaging

Primary TMCs were collected from donors through the Lions Eye Donation service (Human Research Ethics Committee of the Royal Victorian Eye and Ear Hospital - reference number 13-1151H). Cells were cultured in Dulbecco's Minimal Essential Medium (Gibco, 11965118) with 10% Foetal Bovine Serum (Gibco, 16000044) and 0.5% antibiotic-antimycotic (Gibco, 15240-062) (herein referred to as 'culture medium') at 37℃ with 5% CO_2. Cells were passaged by

121 removing the culture medium and washing twice with Phosphate Buffered Saline (Gibco,
122 14190144). Trypsin 0.25% diluted in PBS (Gibco, 25200056) was then added, and the cells
123 were incubated for 3 minutes at 37℃ with 5% CO2. The trypsin was deactivated with cell culture
124 medium, and cells were then aspirated into tubes and centrifuged at 1000 rpm for 5 minutes.
125 The supernatant was aspirated, and the cell pellet was resuspended in culture medium before
126 being plated at the desired ratio for ongoing culture. All TMCs were cultured in tissue culture
127 treated polystyrene plates (Corning, 3516, 3524). Cell lines were tested for mycoplasma on a
128 second weekly basis using the PCR Mycoplasma Test Kit (PromoKine, PK-CA91-1096)
129
130 ***Cell transfection and CRISPR gene knockout***
131 A total of 67 TMC lines were generated using a library of 124 targeting single guide RNAs
132 (sgRNAs) (two for each target gene), together with 10 non-targeting sgRNAs as negative
133 controls. SgRNAs were designed using GUIDES[19] and are displayed in **Supplementary Table**
134 **1**. Following synthesis, sgRNAs were cloned into a novel construct that had previously been
135 developed for the pooled single-cell RNA sequence profiling of primary cells (CROPseq-Guide-
136 pEFS-SpCas9-p2a-puro; Addgene: #99248).[20] The lentivirus was then packaged by transfecting
137 HEK 293FT cells with pCMV delta 8.91, pMDG, and the recombinant plasmid via lipofectamine
138 2000. Lentivirus was chosen as the optimal viral vector due to its large size of ~8.5kB allowing
139 sgRNA, Cas9 and puromycin resistance genes to be packaged into one viral vector.[21]
140
141 Passage one primary TMCs were transfected with 50 µL of lentiviral plasmid and each
142 CRISPR/Cas9/sgRNA/puromycin plasmid in an arrayed format. Individually cloned
143 CRISPR/Cas9/sgRNA/puromycin plasmids were separately added to 450 µL of TMC cells in
144 culture mixed with 1:100 lentiblast (OZ Bioscience, LB01500) in 24 well plates. Cell cultures
145 were incubated for three days before 1 µg/mL puromycin selection occurred over four days.
146 Transfected TMCs underwent standard cell passaging and were then resuspended in 100µl-

500µl DMEM depending on initial cell density. Initial cell density was qualitatively checked with brightfield microscopy before seeding. The predicted on-target editing efficiency for each sgRNA was generated for each sgRNA (**supplementary table 1**).The mRNA expression of each gene knockout can be quantified from RNA sequencing data, however, whilst CRISPR introduces indels into the targeted sequence, the transcription of mRNA for each target gene still occurs. Thus, directly editing efficiency is not able to be quantified using RNA sequencing data.

Cell painting and imaging protocols

Cells were seeded at random in triplicates across 96-well plates at a density of 4.0 X 10^3 cells per well using a Beckman Coulter MoFlo Astrios EQ fluorescence-activated cell sorter (FACS) to ensure an equal distribution of cells. The Cell Painting protocol as described by Bray and colleagues was then followed.[22] Six fluorophores were used to highlight eight cellular components, which were imaged with high content microscopy taken at 20x magnification across five fluorescent channels on a Zeiss CellDiscoverer7 as outlined in **Table 1**. Images were auto-focussed using the definite focus strategy (a set focus point for each image) at 25 sites per well as shown in **Figure 1**.

Image preprocessing and quality control

All images were separated into multiple single-cell images using the “Save Cropped Objects” function in CellProfiler (version 3.1.9, Broad Institute, Massachusetts Institute of Technology).[23,24] This was undertaken to ensure that single-cell morphology was the only feature of the image, and classification was not influenced by overall cell confluency. An image quality filter was then applied using CellProfiler, which flagged any low-quality images that may contain artefacts or were inadequate for analysis, and these were subsequently removed. CellProfiler analysis data was used to calculate Spearman's rank correlation of individual cells for all cell lines. Non-correlated cells from each line were then removed by setting a Spearman

correlation cutoff value of 0.15 to reduce well-to-well and batch-to-batch variation.

CNN architecture, training and evaluation

The CNN architecture is outlined in **Supplementary Table 2** and accessible via GitHub. The dataset was first split into training (80%), validation (10%), and testing (10%) sets. A separate CNN was trained for each fluorescent channel of each gene across five replicates (each with a different random seed to create individual datasets). Training was conducted for 100 epochs, with the model being saved at each epoch. An Adam optimiser was used with a learning rate of 0.0001. For evaluation, the best-performing model of the 100 epochs as per the loss function was selected and evaluated on the test set. The AUC metric was used to quantify CNN performance and thus the degree of morphological variation induced by genetic variations. The highest-performing models were all selected prior to reaching 100 epochs where model overfitting began to reduce model accuracy.

RESULTS

Image Filtering and Data Split

Filtering using CellProfiler and by Spearman correlation reduced the total dataset size from 225,095 images per channel to 114,830 images per channel, yielding a total of 574,150 images for analysis. The proportion of images removed via Spearman filtering varied across groups from 22.1% (*ANTXR1*) to 70.0% (non-targeting group one). The five non-targeting control lines had the greatest proportion of images removed via Spearman filtering as shown in **Figure 2**. The total number of cell images after filtering ranged from 221 (*ADAMTS6*) to 4323 (*ANTXR1*). This inter-group variability was balanced during training with image rotation data augmentation (0, 90, 180, 270, with or without horizontal mirroring) to reach approximately 3,000 images per

group. A random selection of non-targeting control images was then selected to produce a balanced dataset of gene knockout and non-targeting control images. The same non-targeting images were chosen for each knockout comparison. The dataset was split into training (80%), validation (10%), and testing (10%) sets.

Overall morphological variation induced by genetic knockouts

The AUC metric was used to assess the ability of the CNN to distinguish genetic knockout lines from non-targeting control lines thus providing a quantifiable value of morphological variation induced by gene knockouts. The mean AUC of five replicates across five channels was calculated to produce an overall AUC for each target gene. Knockout of *RALGPS1* produced the most morphologically distinct TMCs (AUC 0.851, SD 0.030), followed by *LTBP2* (AUC 0.846, SD 0.029) and *BCAS3* (AUC 0.845, SD 0.020). The overall AUCs ranged from 0.564 (*LMO7*) to the most distinguishable at 0.851 (*RALGPS1*) as displayed in **Figure 3**.

Morphological variation induced in individual organelles

Twenty one (33.9%) gene knockout groups had greater morphological distinction in the mitochondrial channel (mean AUC 0.760 of all cell lines, SD 0.070) compared to other organelles, illustrating that mitochondrial variation occurs in a large proportion of the gene knockouts. The relative AUC of each gene across all organelles is shown in **Figure 4**. Endoplasmic reticulum showed the next greatest morphological variation evident in 16 (25.8%) of the gene knockout lines (mean AUC 0.756, SD 0.079). The F-actin/cell membrane/Golgi body channel showed the highest morphological variation in 13 (20.9%) gene knockout lines (mean AUC 0.751, SD 0.073) followed by 11 (17.7%) knockout lines in the cytoplasmic RNA/nucleolus channels (mean AUC 0.753, SD 0.078). Finally, only the *ANAPC1* knockout showed morphological variation most in the nucleus (mean AUC 0.677, SD 0.079).

Gene prioritisation

Finally, we used the trained CNN AUC metrics to investigate TMC morphological variation for genes at multi-gene loci.[10,11] **Table 2** displays the AUC (knock-out of target gene compared to non-targeting control) for 15 genes across seven loci. For five of these loci, we identified gene knockouts (*ALDH9A1, CAV2, ME3, RALGPS1* (present in two loci)) which resulted in greater morphological variation than knockout of their neighbouring gene counterparts. Knockout of genes at two multi-gene loci (*EMID1-KREMEN1* and *GNB1L-TXNRD2*) generated TMCs that were morphologically similar and thus, could not be prioritised.

DISCUSSION

There has been a shift in recent years towards using high-throughput profiling to undertake large-scale studies investigating the cellular basis of disease. This shift has been accelerated by advancements in computational technology and AI as a method of rapidly analysing large, complex datasets. In this study, we utilised a convolutional neural network to perform a high-throughput morphological analysis of genetic variations associated with IOP variation in primary human TMCs. By training the CNN to distinguish gene knockout cells from healthy control cells, we could use the AUC as a metric to quantify differences in cellular morphology induced by various genetic variations. Therefore, the AUC can be used to identify which variations invoke a greater degree of morphological change and, thus, which are more likely to be involved in IOP dysregulation and the pathogenesis of POAG.

Of the genes known to cause primary congenital glaucoma or anterior segment dysgenesis, *LTBP2* and *TEK* showed marked differentiation from normal control morphology. The *LTBP2* knockout cell line was readily distinguished from normal control TMCs (AUC 0.846) with the

247 greatest degree of difference occurring in mitochondrial morphology indicating that *LTBP2* may
248 play a role in mitochondrial function. *LTBP2* encodes for latent transforming growth factor beta
249 binding protein 2 which is an extracellular matrix protein associated with fibrillin-1 containing
250 microfibrils and is hypothesised to modulate extracellular matrix production.[25] Variations in
251 *LTBP2* have been previously associated with primary congenital glaucoma, microspherophakia,
252 megalocornea and Weill-Marchesani syndrome.[25–28] A previous study has identified that *LTBP2*
253 knockout may contribute to the development of POAG via dysregulation of the extracellular
254 matrix; a crucial component of the trabecular meshwork.[29] Studies looking at dilated
255 cardiomyopathy and right ventricular failure have also implicated *LTBP2* function in fibrosis
256 regulation which may indicate a role in the pathogenesis of trabecular meshwork
257 dysfunction.[30,31]
258
259 The *TEK* knockout cell line also showed significant differentiation (AUC 0.768) most prominent
260 in the cytoplasmic RNA and nucleolus channel. This gene encodes for a tyrosine-kinase
261 receptor and is highly involved in the regulation of angiogenesis and vascular stability.[32] It also
262 acts as a receptor for *ANGPT1* which has been shown to be crucial for development of
263 Schlemm's canal.[33–35] Variations in *TEK* have been associated with raised IOP and congenital
264 glaucoma primarily due to disruption of Schlemm's canal, indicating a potential interaction with
265 *ANGPT1* in the development of glaucoma.[35–38] Curiously, *MYOC, CYP1B1, GMDS,* and *FOXC1*
266 knockouts resulted in only mild differentiation from control TMC morphology (AUC 0.615, 0.612,
267 0.704, 0.665, respectively) despite an association with glaucoma and anterior segment
268 dysgenesis.[7,39–42] These gene knocktous may not invoke significant morphological variation as
269 they are primarily involved trabecular meshwork development rather than the maintenance.[43]
270 Furthermore, some gene mutations associated with congenital glaucoma are gain-of-function
271 mutations and therefore will not show significant change when knocked out. Another reason for
272 not seeing change in cellular morphology is that these genes may primarily act extracellularly

such as *MYOC* which has been shown to demonstrate accumulation of extracellular products in specific mutations.[44]

The knockout of *RALGPS1* resulted in the greatest degree of differentiation (AUC 0.851) compared to other cell lines and was most prominent in the mitochondrial channel. This gene encodes for ras-specific guanine nucleotide-releasing factor RalGPS1, which is involved in Ras protein activation.[45] Not only has *RALGPS1* been associated with raised IOP[10,11], but previous studies have also highlighted a link to high myopia as well as a role in optic nerve regeneration.[46,47] The *BCAS3* gene knockout also produced a high degree of differentiation (AUC 0.845), which was also greatest in the mitochondrial channel. This gene encodes for breast carcinoma-amplified sequence 3 and has been shown to play a role in angiogenesis.[48,49] *BCAS3* variants have been previously associated with glaucoma and optic nerve head parameters.[50–52]

Overall, the mitochondrial channel most frequently displayed the greatest degree of differentiation (33.9% of all cell lines). Previous studies have highlighted an association between glaucoma and mitochondrial dysfunction, likely related to the high energy requirements of retinal ganglion cells.[53–55] Studies have also shown direct evidence of mitochondrial dysfunction in POAG affected eyes indicated by increased mitochondrial respiratory activity and elevated retinal mitochondrial flavoprotein; both of which are associated with mitochondrial dysfunction.[56–58] The endoplasmic reticulum channel also showed the most morphological variation in a large proportion of cell lines (25.8%), which is in keeping with many studies that have highlighted a link between glaucoma and endoplasmic reticulum stress.[59–61] Interestingly, the *ANAPC1* knockout was the only cell line to display the greatest differentiation in the nucleus channel compared to other organelles. Similarly, this gene is involved in progression through cellular mitosis.[62]

This work introduced a novel method for prioritising genes at overlapping loci identified in GWAS using CNN analysis.[10,11] The results show that *ALDH9A1, RALGPS1, CAV2* and *ME3* show statistically greater differentiation from control cells than the respectively associated gene at the same locus. Studies have previously associated POAG with genetic variants at the inter-genomic region of *TMCO1* and *ALDH9A1*.[63–65] The results of this study point toward *ALDH9A1* being the implicated gene in POAG due to inducing a greater degree of morphological change compared to *TMCO1* (p-value 7.78e−05). The mitochondrial channel in *ALDH9A1* displayed the greatest degree of differentiation, highlighting the potential role of mitochondrial dysfunction in *ALDH9A1* interruption in POAG. This is supported by the role of *ALDH9A1* in carnitine synthesis, which takes place in the mitochondrial matrix.[66] There have also been numerous studies illustrating an association between POAG and variations at the inter-genomic region of *CAV1* and *CAV2*.[67–70] This analysis prioritised *CAV2* as a potential causative gene, with a higher degree of morphological change from control cells than *CAV1* (p-value 4.00e−03). The *CAV2* knockout cell line displayed the most prominent changes in the F-actin, Golgi complex, cell membrane fluorescent channel. Supporting this, previous studies have highlighted the interaction between *CAV2* and the Golgi complex.[71–73] The genomic region containing *ME3* and *PRSS23* has previously been associated with open-angle glaucoma.[74] Our study highlighted a statistically greater degree of morphological change in the *ME3* cell line providing evidence for prioritisation over *PRSS23* in the pathogenesis of POAG. The remaining genes at overlapping loci (*EMID1 vs KREMEN1* and *GNB1L vs TXNRD2*) showed no statistically significant differences in morphology. They will require further investigation to prioritise which of these may be the causative gene.

A further application of AI-based analysis of single cell morphology is to predict gene expression as demonstrated in prior studies. For example, Chlis and colleagues developed a machine

learning model to predict gene expression of human mononuclear blood cells and mouse myeloid progenitor cells based on cellular morphology.[75] Our study further highlights the complex interaction between cell morphology and gene expression and the opportunity that AI poses as a means of analysing the large amounts of data produced. Further investigation into this field could uncover the genetic drivers behind pathological changes in morphology that drive disease processes and allow for identification of novel therapeutic targets.[75,76]

One of the main limitations of this study lies in the intrinsic difficulty in interpreting the decision-making process of CNNs. This means it can be difficult to establish if morphological features learned by the CNN are truly pathological or simply due to systematic bias. For example, if wells had lower cell density, the cells may grow to a larger size, thus cell size may inadvertently influence the decision-making of the CNN.

In summary, this study used a powerful approach to quantify morphological change induced by genetic variations associated with POAG. *RALGPS1* produced the greatest morphological variation. As well as this, we could prioritise genes at overlapping loci identified to have an association with IOP. However, there are some limitations due to the difficulty in removing systematic bias from the methodology. This bias may result in the CNN learning features that are not directly associated with IOP physiology. This study highlights a new avenue for utilising CNNs trained on single-cell morphology to further interpret the results of GWAS.

REFERENCES

1. Weinreb RN, Leung CKS, Crowston JG, et al. Primary open-angle glaucoma. Nat Rev Dis Primers 2016;2:16067.

2. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. JAMA 2014;311:1901–1911.

3. GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021;9:e144–e160.

4. Lusthaus J, Goldberg I. Current management of glaucoma. Med J Aust 2019;210:180–187.

5. Beidoe G, Mousa SA. Current primary open-angle glaucoma treatments and future directions. Clin Ophthalmol 2012;6:1699–1707.

6. Ojha P, Wiggs JL, Pasquale LR. The genetics of intraocular pressure. Semin Ophthalmol 2013;28:301–305.

7. Stone EM, Fingert JH, Alward WL, et al. Identification of a gene that causes primary open angle glaucoma. Science 1997;275:668–670.

8. Hewitt AW, Mackey DA, Craig JE. Myocilin allele-specific glaucoma phenotype database. Hum Mutat 2008;29:207–211.

9. Burdon KP, Graham P, Hadler J, et al. Specifications of the ACMG/AMP variant curation guidelines for myocilin: Recommendations from the clingen glaucoma expert panel. Hum Mutat 2022. Available at: http://dx.doi.org/10.1002/humu.24482.

10. MacGregor S, Ong J-S, An J, et al. Genome-wide association study of intraocular pressure uncovers new pathways to glaucoma. Nat Genet 2018;50:1067–1071.

11. Khawaja AP, Cooke Bailey JN, Wareham NJ, et al. Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. Nat Genet 2018;50:778–782.

12. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc.; 2012:1097–1105.

13. Oei RW, Hou G, Liu F, et al. Convolutional neural network for cell classification using microscope images of intracellular actin networks. PLoS One 2019;14:e0213626.

14. Kensert A, Harrison PJ, Spjuth O. Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes. SLAS Discov 2019;24:466–475.

15. Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery. Drug Discov Today 2018;23:1241–1250.

16. Gibson CC, Zhu W, Davis CT, et al. Strategy for Identifying Repurposed Drugs for the Treatment of Cerebral Cavernous Malformation. Circulation 2015;131:289–299–299.

17. Nitta N, Sugimura T, Isozaki A, et al. Intelligent Image-Activated Cell Sorting. Cell 2018;175:266–276.e13.

18. Ota S, Horisaki R, Kawamura Y, et al. Ghost cytometry. Science 2018;360:1246–1251.

19. Meier JA, Zhang F, Sanjana NE. GUIDES: sgRNA design for loss-of-function screens. Nat Methods 2017;14:831–832.

20. Datlinger P, Rendeiro AF, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods 2017;14:297–301.

21. Anon. Lentiviral Guide. Addgene. Available at: https://www.addgene.org/guides/lentivirus/ [Accessed November 29, 2022].

22. Bray M-A, Singh S, Han H, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 2016;11:1757–1774.

23. Carpenter AE, Jones TR, Lamprecht MR, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 2006;7:R100.

24. McQuin C, Goodman A, Chernyshev V, et al. CellProfiler 3.0: Next-generation image processing for biology. PLoS Biol 2018;16:e2005970.

25. Haji-Seyed-Javadi R, Jelodari-Mamaghani S, Paylakhi SH, et al. LTBP2 mutations cause Weill-Marchesani and Weill-Marchesani-like syndrome and affect disruptions in the extracellular matrix. Hum Mutat 2012;33:1182–1187.

26. Désir J, Sznajer Y, Depasse F, et al. LTBP2 null mutations in an autosomal recessive ocular syndrome with megalocornea, spherophakia, and secondary glaucoma. Eur J Hum Genet 2010;18:761–767.

27. Ali M, McKibbin M, Booth A, et al. Null mutations in LTBP2 cause primary congenital glaucoma. Am J Hum Genet 2009;84:664–671.

28. Kumar A, Duvvari MR, Prabhakaran VC, et al. A homozygous mutation in LTBP2 causes isolated microspherophakia. Hum Genet 2010;128:365–371.

29. Suri F, Yazdani S, Elahi E. LTBP2 knockdown and oxidative stress affect glaucoma features including TGFβ pathways, ECM genes expression and apoptosis in trabecular meshwork cells. Gene 2018;673:70–81.

30. Pang X-F, Lin X, Du J-J, Zeng D-Y. LTBP2 knockdown by siRNA reverses myocardial oxidative stress injury, fibrosis and remodelling during dilated cardiomyopathy. Acta Physiol 2020;228:e13377.

31. Potus F, Hindmarch CCT, Dunham-Snary KJ, et al. Transcriptomic Signature of Right Ventricular Failure in Experimental Pulmonary Arterial Hypertension: Deep Sequencing Demonstrates Mitochondrial, Fibrotic, Inflammatory and Angiogenic Abnormalities. Int J Mol Sci 2018;19. Available at: http://dx.doi.org/10.3390/ijms19092730.

420 32. Zhang Y, Kontos CD, Annex BH, Popel AS. Angiopoietin-Tie Signaling Pathway in
421 Endothelial Cells: A Computational Model. iScience 2019;20:497–511.

422 33. Thomson BR, Souma T, Tompson SW, et al. Angiopoietin-1 is required for Schlemm's canal
423 development in mice and humans. J Clin Invest 2017;127:4421–4436.

424 34. Thackaberry EA, Zhou Y, Zuch de Zafra CL, et al. Rapid Development of Glaucoma Via ITV
425 Nonselective ANGPT 1/2 Antibody: A Potential Role for ANGPT/TIE2 Signaling in Primate
426 Aqueous Humor Outflow. Invest Ophthalmol Vis Sci 2019;60:4097–4108.

427 35. Kim J, Park D-Y, Bae H, et al. Impaired angiopoietin/Tie2 signaling compromises Schlemm's
428 canal integrity and induces glaucoma. J Clin Invest 2017;127:3877–3896.

429 36. Thomson BR, Carota I, Souma T, et al. Defects in Angiopoietin-Tie2 signaling lead to dose-
430 dependent glaucoma in mice. Invest Ophthalmol Vis Sci 2016;57. Available at:
431 https://iovs.arvojournals.org/article.aspx?articleid=2564010 [Accessed May 10, 2021].

432 37. Kabra M, Zhang W, Rathi S, et al. Angiopoietin receptor TEK interacts with CYP1B1 in
433 primary congenital glaucoma. Hum Genet 2017;136:941–949.

434 38. Souma T, Tompson SW, Thomson BR, et al. Angiopoietin receptor TEK mutations underlie
435 primary congenital glaucoma with variable expressivity. J Clin Invest 2016;126:2575–2587.

436 39. Kaur K, Reddy ABM, Mukhopadhyay A, et al. Myocilin gene implicated in primary congenital
437 glaucoma. Clin Genet 2005;67:335–340.

438 40. Pasutto F, Chavarria-Soley G, Mardin CY, et al. Heterozygous loss-of-function variants in
439 CYP1B1 predispose to primary open-angle glaucoma. Invest Ophthalmol Vis Sci 2010;51:249–
440 254.

441 41. Gharahkhani P, Burdon KP, Fogarty R, et al. Common variants near ABCA1, AFAP1 and
442 GMDS confer risk of primary open-angle glaucoma. Nat Genet 2014;46:1120–1125.

443 42. Siggs OM, Souzeau E, Pasutto F, et al. Prevalence of FOXC1 Variants in Individuals With a
444 Suspected Diagnosis of Primary Congenital Glaucoma. JAMA Ophthalmol 2019;137:348–355.

445 43. Gauthier AC, Wiggs JL. Childhood glaucoma genes and phenotypes: Focus on FOXC1
446 mutations causing anterior segment dysgenesis and hearing loss. Exp Eye Res
447 2020;190:107893.

448 44. Ueda J, Wentz-Hunter K, Yue BYJT. Distribution of myocilin and extracellular matrix
449 components in the juxtacanalicular tissue of human eyes. Invest Ophthalmol Vis Sci
450 2002;43:1068–1076.

451 45. Rebhun JF, Chen H, Quilliam LA. Identification and characterization of a new family of
452 guanine nucleotide exchange factors for the ras-related GTPase Ral. J Biol Chem
453 2000;275:13406–13410.

454 46. Kloss BA, Tompson SW, Whisenhunt KN, et al. Exome Sequence Analysis of 14 Families
455 With High Myopia. Invest Ophthalmol Vis Sci 2017;58:1982–1990.

456 47. Lindborg JA, Tran NM, Chenette DM, et al. Optic nerve regeneration screen identifies
457 multiple genes restricting adult neural repair. Cell Rep 2021;34:108777.

458 48. Joshi D, Inamdar MS. Rudhira/BCAS3 couples microtubules and intermediate filaments to
459 promote cell migration for angiogenic remodeling. Mol Biol Cell 2019;30:1437–1450.

460 49. Siva K, Venu P, Mahadevan A, et al. Human BCAS3 expression in embryonic stem cells
461 and vascular precursors suggests a role in human embryogenesis and tumor angiogenesis.
462 PLoS One 2007;2:e1202.

463 50. Hauser MA, Liu Y, Gibson J, et al. Variants In CDKN2B, SIX1, and BCAS3 Are Associated
464 With Risk Of POAG. Invest Ophthalmol Vis Sci 2011;52:5328–5328. Available at: [Accessed
465 April 17, 2021].

466 51. Springelkamp H, Iglesias AI, Mishra A, et al. New insights into the genetics of primary open-
467 angle glaucoma based on meta-analyses of intraocular pressure and optic disc characteristics.
468 Hum Mol Genet 2017;26:438–453.

469 52. Craig JE, Han X, Qassim A, et al. Multitrait analysis of glaucoma identifies new risk loci and
470 enables polygenic prediction of disease susceptibility and progression. Nat Genet 2020;52:160–
471 166.

472 53. Ju W-K, Liu Q, Kim K-Y, et al. Elevated hydrostatic pressure triggers mitochondrial fission
473 and decreases cellular ATP in differentiated RGC-5 cells. Invest Ophthalmol Vis Sci
474 2007;48:2145–2151.

475 54. Mittag TW, Danias J, Pohorenec G, et al. Retinal damage after 3 to 4 months of elevated
476 intraocular pressure in a rat glaucoma model. Invest Ophthalmol Vis Sci 2000;41:3451–3459.

477 55. Ju W-K, Kim K-Y, Lindsey JD, et al. Intraocular pressure elevation induces mitochondrial
478 fission and triggers OPA1 release in glaucomatous optic nerve. Invest Ophthalmol Vis Sci
479 2008;49:4903–4911.

480 56. Abu-Amero KK, Morales J, Bosley TM. Mitochondrial abnormalities in patients with primary
481 open-angle glaucoma. Invest Ophthalmol Vis Sci 2006;47:2533–2541.

482 57. Zhou DB, Castanos MV, Geyman L, et al. Mitochondrial Dysfunction in Primary Open-Angle
483 Glaucoma Characterized by Flavoprotein Fluorescence at the Optic Nerve Head. Ophthalmol
484 Glaucoma 2022;5:413–420.

485 58. Chen AX, Conti TF, Hom GL, et al. Functional imaging of mitochondria in retinal diseases
486 using flavoprotein fluorescence. Eye 2021;35:74–92.

487 59. Peters JC, Bhattacharya S, Clark AF, Zode GS. Increased Endoplasmic Reticulum Stress in
488 Human Glaucomatous Trabecular Meshwork Cells and Tissues. Invest Ophthalmol Vis Sci
489 2015;56:3860–3868.

490 60. Anholt RRH, Carbone MA. A molecular mechanism for glaucoma: endoplasmic reticulum
491 stress and the unfolded protein response. Trends Mol Med 2013;19:586–593.

492 61. Yang Z, Ge Y, Pease M, et al. Role of Endoplasmic Reticulum Stress in Retinal Ganglion
493 Cell Death in Glaucoma and Optic Nerve Injury. Invest Ophthalmol Vis Sci 2010;51:5812–5812.
494 Available at: [Accessed November 5, 2022].

495 62. Jin L, Williamson A, Banerjee S, et al. Mechanism of ubiquitin-chain formation by the human
496 anaphase-promoting complex. Cell 2008;133:653–665.

497 63. Verkuil L, Danford I, Pistilli M, et al. SNP located in an AluJb repeat downstream of TMCO1,
498 rs4657473, is protective for POAG in African Americans. Br J Ophthalmol 2019;103:1530–1536.

499 64. Burdon KP, Macgregor S, Hewitt AW, et al. Genome-wide association study identifies
500 susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. Nat Genet
501 2011;43:574–578.

502 65. Hysi PG, Cheng C-Y, Springelkamp H, et al. Genome-wide analysis of multi-ancestry
503 cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma. Nat
504 Genet 2014;46:1126–1130.

505 66. Končitíková R, Vigouroux A, Kopečná M, et al. Kinetic and structural analysis of human
506 ALDH9A1. Biosci Rep 2019;39. Available at: http://dx.doi.org/10.1042/BSR20190558.

507 67. Wiggs JL, Kang JH, Yaspan BL, et al. Common variants near CAV1 and CAV2 are
508 associated with primary open-angle glaucoma in Caucasians from the USA. Hum Mol Genet
509 2011;20:4707–4713.

510 68. Chen F, Klein AP, Klein BEK, et al. Exome array analysis identifies CAV1/CAV2 as a
511 susceptibility locus for intraocular pressure. Invest Ophthalmol Vis Sci 2014;56:544–551.

512 69. Chen LJ, Tam POS, Pang CP. Revisit the association of CAV1/CAV2 with primary open-
513 angle glaucoma. Invest Ophthalmol Vis Sci 2013;54:6236–6236. Available at: [Accessed May 7,
514 2021].

515 70. Loomis SJ, Kang JH, Weinreb RN, et al. Association of CAV1/CAV2 genomic variants with
516 primary open-angle glaucoma overall and by gender and pattern of visual field loss.
517 Ophthalmology 2014;121:508–516.

518 71. Mora R, Bonilha VL, Marmorstein A, et al. Caveolin-2 localizes to the golgi complex but
519 redistributes to plasma membrane, caveolae, and rafts when co-expressed with caveolin-1. J
520 Biol Chem 1999;274:25708–25717.

521 72. Aga M, Bradley JM, Wanchu R, et al. Differential effects of caveolin-1 and -2 knockdown on
522 aqueous outflow and altered extracellular matrix turnover in caveolin-silenced trabecular
523 meshwork cells. Invest Ophthalmol Vis Sci 2014;55:5497–5509.

524 73. Parolini I, Sargiacomo M, Galbiati F, et al. Expression of caveolin-1 is required for the
525 transport of caveolin-2 to the plasma membrane. Retention of caveolin-2 at the level of the golgi
526 complex. J Biol Chem 1999;274:25718–25725.

527 74. Shin H-T, Yoon BW, Seo JH. Analysis of risk allele frequencies of single nucleotide
528 polymorphisms related to open-angle glaucoma in different ethnic groups. BMC Med Genomics
529 2021;14:80.

530 75. Chlis N-K, Rausch L, Brocker T, et al. Predicting single-cell gene expression profiles of
531 imaging flow cytometry data with machine learning. Nucleic Acids Res 2020;48:11335–11346.

532 76. Cutiongco MFA, Jensen BS, Reynolds PM, Gadegaard N. Predicting gene expression using
533 morphological cell responses to nanotopography. Nat Commun 2020;11:1384.

534

535

Cell Painting reagent	Fluorescent channel	Excitation filter (nm)	Emission filter (nm)	Organelles
Hoechst 33342	DAPI	387/11	417-477	Nucleus
Concanavalin A/Alexa Fluor 488 conjugate	EGFP	472/30	503-538	Endoplasmic reticulum
SYTO 14 Green Fluorescent Nucleic Acid stain	AF514	531/40	573-613	Cytoplasmic RNA, nucleolus
Phalloidin/AlexaFluor 568 **Wheat-Germ Agglutinin/AlexaFluor 555 conjugate**	AF594	581/609 (phalloidin) 590/617 (WGA)	622-662	F-actin, golgi complex, cell membrane
MitoTracker Deep Red	AF647	628/40	672-712	Mitochondria

536

537 **Table 1 - Cell Painting reagents, fluorescent channels and associated cellular organelles**
538 The Cell Painting protocol was designed to allow a maximum number of cellular organelles to be
539 visualised with minimal overlap of fluorescent channels.
540
541
542
543

Top GWAS SNP	Overlapping Genes (mean AUC)	P-value
rs7518099	*ALDH9A1* (AUC 0.709, SD 1.93e-02) *TMCO1* (AUC 0.634, SD 4.76e-02)	**7.78e−05**
rs11795066	*RALGPS1* (AUC 0.851, SD 3.05e-02) *ANGPTL2* (AUC 0.811, SD 2.50e-02)	**4.12e−04**
rs6478746	*LMX1B* (AUC 0.803, SD 2.03e-02) *RALGPS1* (0.851)	**5.5e−06**
rs10281637 rs55892100	*CAV1* (AUC 0.726, SD 5.53e-02) *CAV2* (AUC 0.817, SD 2.71e-02) *TES* (AUC 0.704, SD 5.79e-02)	4.49e−01 (*CAV1*, *TES*) **3.00e−03** (*CAV2*, *TES*) **4.00e−03** (*CAV1*, *CAV2*)
rs9608740	*EMID1* (AUC 0.834, SD 6.50e-02) *KREMEN1* (AUC 0.824, SD 5.70e-02)	5.73e−01
rs8141433	*GNB1L* (AUC 0.729, SD 5.97e-02) *TXNRD2* (AUC 0.695, SD 4.47e-02)	3.75e−01
rs746491	*ME3* (AUC 0.803, SD 2.45e-02) *PRSS23* (AUC 0.725, SD 4.25e-02)	**3.47e−04**

Table 2: Comparison of CNNs to morphologically distinguish TMCs with knockout of genes at overlapping IOP-associated loci.[10]
The mean AUC across all fluorescent channels of target knockouts versus non-targeting control cells was compared for genes at the same locus. A higher AUC indicates a larger degree of morphological variation compared to normal control cells. This allows for prioritisation of overlapping genes at given loci.

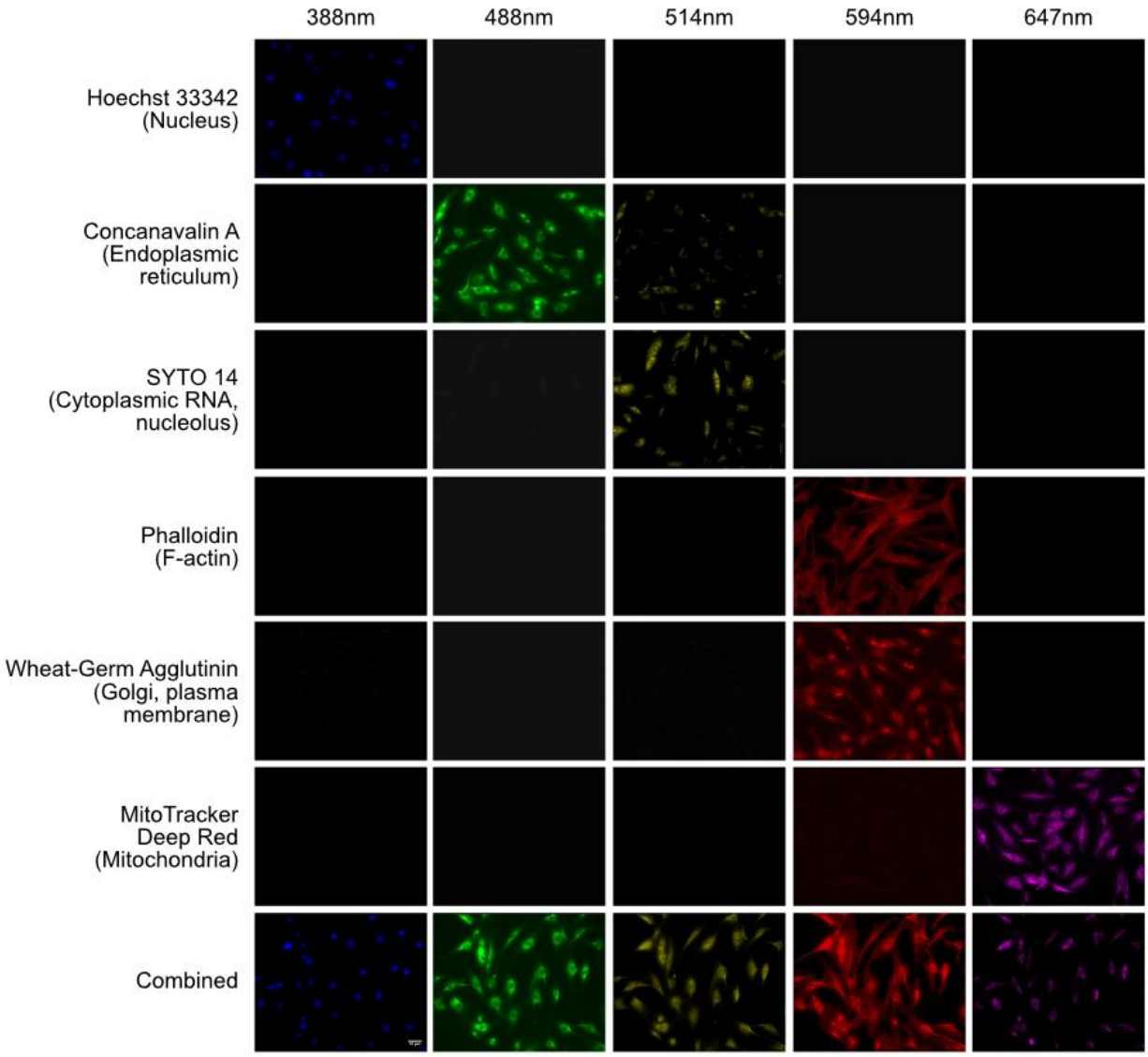

Figure 1: Cell Painting Assay

Example image of TMCs stained with the Cell Painting protocol in which six fluorophores are imaged over five channels to identify eight distinct intracellular organelles for morphological profiling. Each row shows cells stained with the indicated dye, or with all dyes combined (bottom row); columns indicate excitation wavelengths. Single channel testing shows minimal overlap across channels except for the Phalloidin and Wheat-Germ Agglutinin stains which are analysed together. This ensures that only a single stain will fluoresce when exposed to a particular wavelength of light. This figure shows whether a single stain would contaminate other emission channels and whether the signal of the light emission channel was dominated by the dyes we selected.

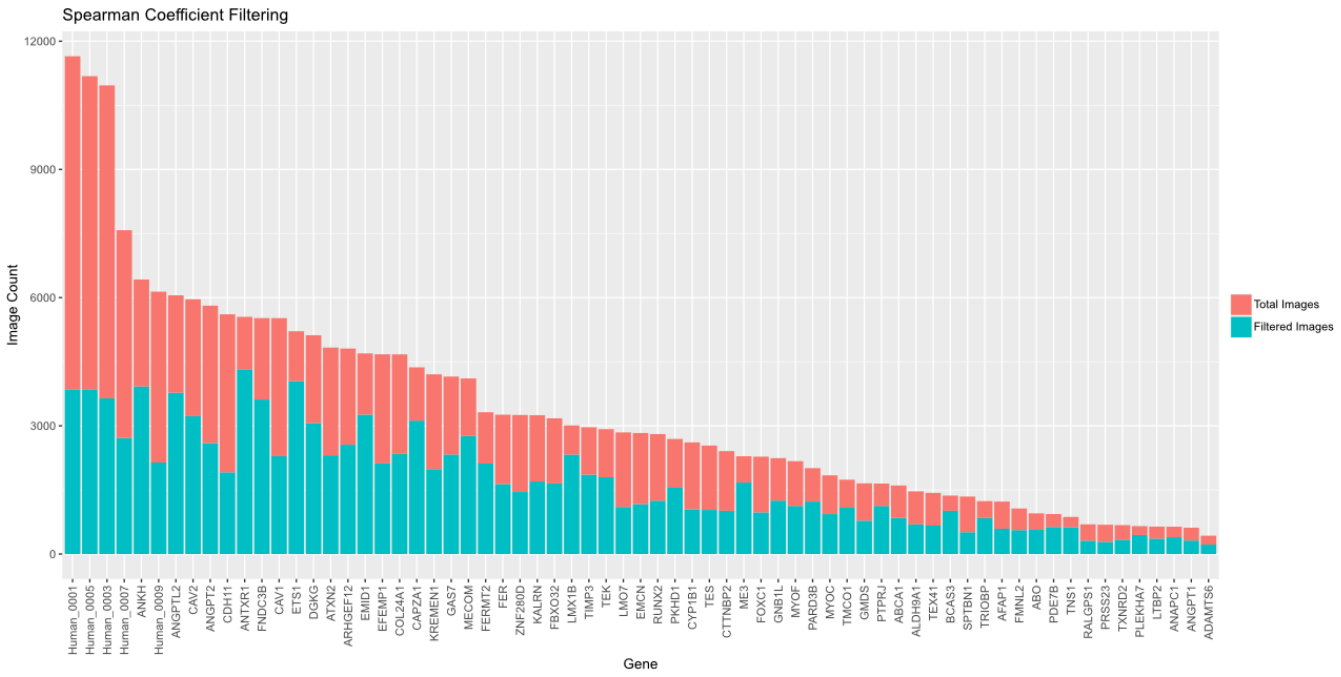

Figure 2: Total number of images for each arrayed cell line following Spearman correlation filtering.

Images were removed from the dataset if the Spearman correlation was >0.15 in order to improve the quality of the dataset and reduce the effect of well-to-well and batch-to-batch variation. Ultimately, the percentage of cells removed ranged from 67% (control line 1) to 22% (*ANTXR1*).

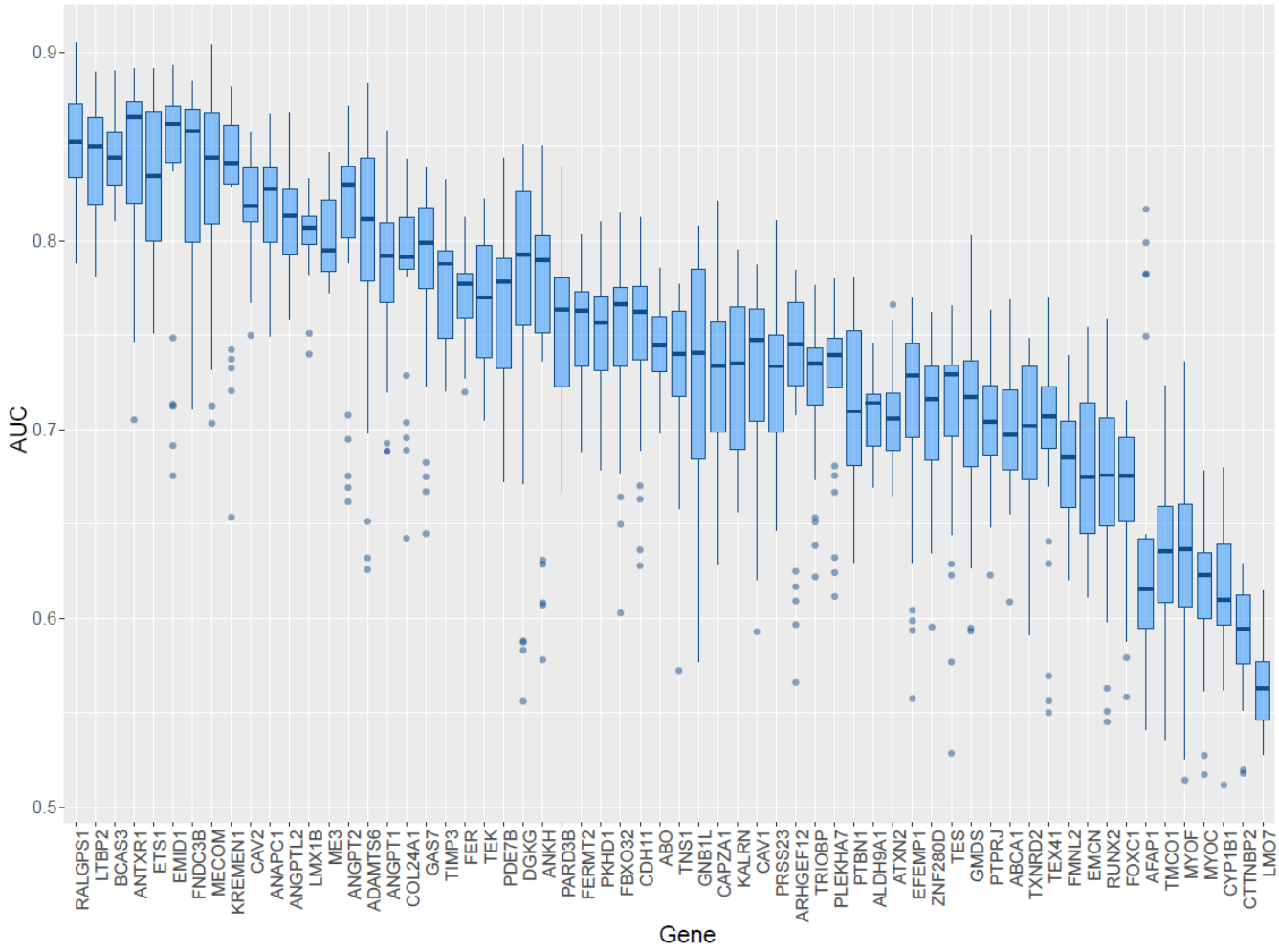

Figure 3: Mean CNN AUC scores for each gene-knockout cell-line

The mean AUC score when training a CNN to distinguish between gene-knockout cell-lines and non-targeting control cell-lines. A higher AUC indicates a more distinct morphological variation induced by a particular gene-knockout. The gene knockouts are ordered in decreasing order of mean AUC across all organelles. The bars represent the median AUC with upper and lower quartile boxes. Outliers are displayed as single dots.

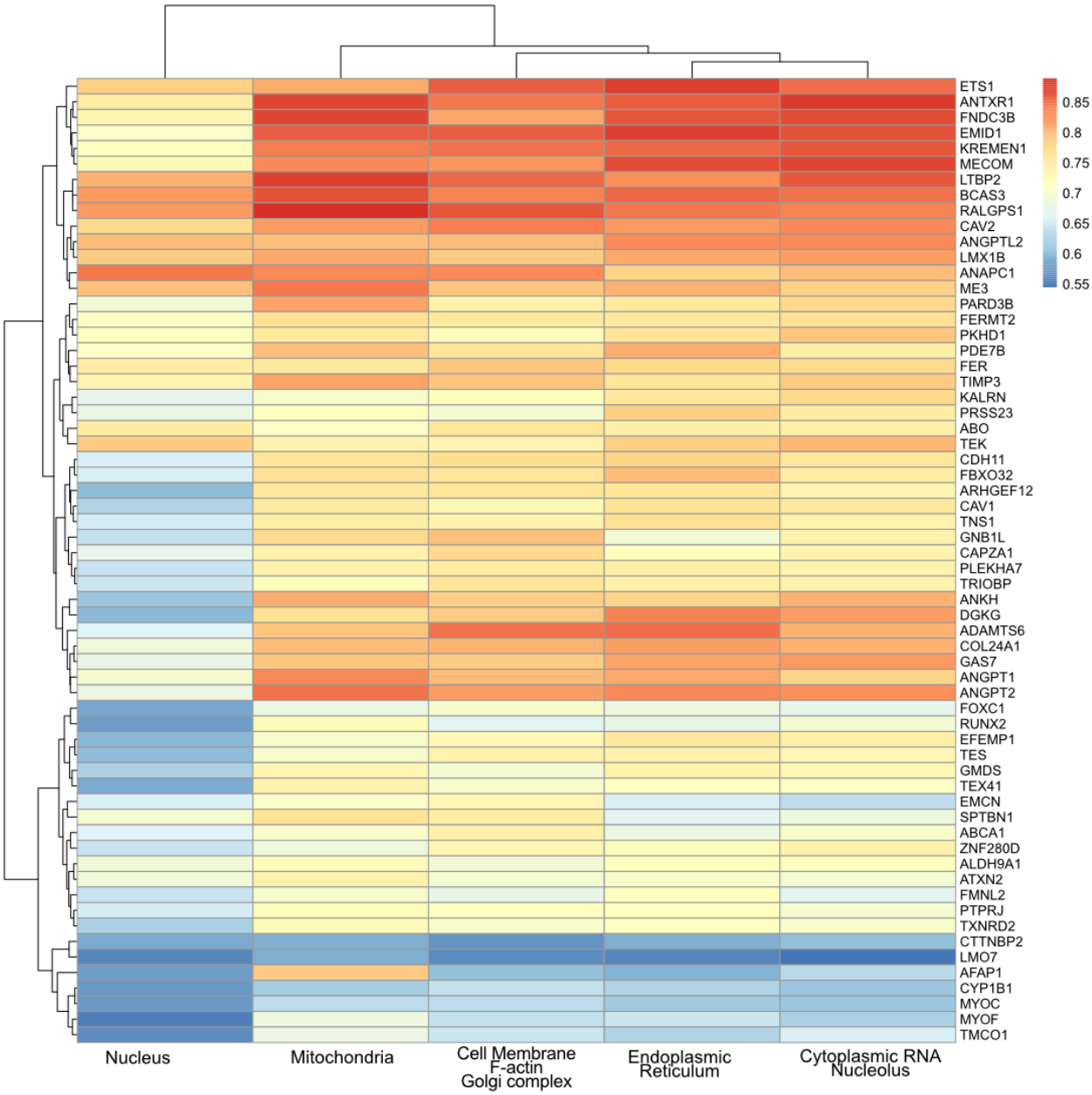

Figure 4: Gene knockout cell line AUC for each organelle.
Heatmap of the morphological variation (AUC) across individual fluorescent channels for each gene knockout. Red shading indicates a higher degree of morphological variation as indicated by a higher AUC.

SUPPLEMENTARY

Code availability

The Python functions utilised for data preparation, CNN training and evaluation are available on GitHub: https://github.com/ConnorG1/TMC_CNN

Data availability

Data is available at the European Bioimage Institute Bioimage Archive: Accession S-BSST841

Supplementary table 1

sgRNA name	Sequence	Gene	Exon	Protein domain targeted	Predicted On-target efficiency
GUIDES_sg001	GTTGACTGGGAGAGAACACG	*ABCA1*	46	ABC_tran	0.707186352
GUIDES_sg002	GTGTTCTAAAAGAGAAACAC	*ABCA1*	50	-	0.684107302
GUIDES_sg003	GTGCAGTGTCTCTCCTACAC	*ADAMTS6*	28	-	0.720054847
GUIDES_sg004	ACCAGTCATGTCCACCACAG	*ADAMTS6*	26	-	0.71972146
GUIDES_sg005	CAAGGGTAAAAAGCCCCCCG	*AFAP1*	16	-	0.759587174
GUIDES_sg006	GGAAAGAAAAGACCTTCGAG	*AFAP1*	17	-	0.753862205
GUIDES_sg007	CAACTATAACGTCAGCCCAG	*ALDH9A1*	10	Aldedh	0.764185557
GUIDES_sg008	TATGAACAATGCTGTAAAGG	*ALDH9A1*	6	Aldedh	0.704113168
GUIDES_sg009	GTGCTGTGAGCTGGGAAGTG	*ANAPC1*	39	-	0.69840945
GUIDES_sg010	ATGGCTCTTCCTGTAGGACG	*ANAPC1*	27	-	0.641611628
GUIDES_sg011	CCCCAGCCAATATTCACCGG	*ANGPT1*	9	Fibrinogen_C	0.666179972
GUIDES_sg012	AATATGGATGTCAATGGGGG	*ANGPT1*	8	Fibrinogen_C	0.630390739
GUIDES_sg013	TAACGTGTAGATGCCATTCG	*ANGPT2*	5	Fibrinogen_C	0.692502215
GUIDES_sg014	TGTGACATGGAAGCTGGAGG	*ANGPT2*	6	Fibrinogen_C	0.647042655
GUIDES_sg015	ACTCGCTCTCAGGTTCCAGG	*ANGPTL2*	5	Fibrinogen_C	0.761119961
GUIDES_sg016	CACCAGCATGTCACGCACAG	*ANGPTL2*	2	RasGEF	0.753686491
GUIDES_sg017	CTTTGTGGGAGAATCCACCA	*ANKH*	14	ANKH	0.705623542
GUIDES_sg018	TGAGGGCGCATCTCACCGGG	*ANKH*	13	ANKH	0.671178514
GUIDES_sg019	CTTCCGACATGCCCGCAACG	*ANTXR1*	10	Anth_Ig	0.66965417
GUIDES_sg020	CAGAACTGGAGATAAAAGAG	*ANTXR1*	12	Anth_Ig	0.661995855
GUIDES_sg021	CTGCTGGACCAGAAATTCGG	*ARHGEF12*	39	-	0.715629912
GUIDES_sg022	TCTCTGGGGTCATAATCATG	*ARHGEF12*	38	-	0.682338139
GUIDES_sg023	TACCAAATATGCCCCAACAG	*ATXN2*	21	-	0.746394842
GUIDES_sg024	ATTACAGGACTATAGACATG	*ATXN2*	22	-	0.705381212
GUIDES_sg025	ATGGGCCCAGGACTTCCAGG	*BCAS3*	35	-	0.763805308
GUIDES_sg026	TGAACTGGATGAGATAACTG	*BCAS3*	36	-	0.749363581
GUIDES_sg027	AGTTTTTAGGCTGAAACTGG	*CAPZA1*	6	F-actin_cap_A	0.671121892
GUIDES_sg028	GGAATAATGGTCTTTCACAT	*CAPZA1*	5	F-actin_cap_A	0.61837895
GUIDES_sg029	TAAACACCTCAACGATGACG	*CAV1*	3	Caveolin	0.697766064
GUIDES_sg030	TGGGGGCAAATACGTAGACT	*CAV1*	1	Caveolin	0.638532867
GUIDES_sg031	GATGTGCAGACAGCTGAGGG	*CAV2*	23	Caveolin	0.699222046
GUIDES_sg032	CGGCGTACTCGAGGCCGCTG	*CAV2*	22	Caveolin	0.674769146
GUIDES_sg033	TCAGGAAGCCAAAGTCCCAG	*CDH11*	19	Cadherin	0.765145054
GUIDES_sg034	GGATTGTGAATGATTTCAGG	*CDH11*	20	Cadherin	0.686839641
GUIDES_sg035	GGGTTATCTCGTGTGCCAAG	*COL24A1*	60	COLFI	0.662833491
GUIDES_sg036	GAAATTGCAGAAAACCTCAA	*COL24A1*	61	COLFI	0.614109727
GUIDES_sg037	AAGCGGCCAGACTTCCTGCG	*CTTNBP2*	25	-	0.737788504
GUIDES_sg038	GCCAGGTTGTCTTTTCACAG	*CTTNBP2*	24	-	0.682279464
GUIDES_sg039	AACATTCCCAGCATGTACGG	*DGKG*	22	DAGK_acc	0.746132205
GUIDES_sg040	GTACTTTGAATTTGGCACCT	*DGKG*	21	DAGK_acc	0.610555363
GUIDES_sg041	GACCACAAATGAATGCCGGG	*EFEMP1*	9	EGF_CA	0.7383367
GUIDES_sg042	TCACCACTTGGTATCCCTGG	*EFEMP1*	8	EGF_CA	0.711733541

GUIDES_sg043	ACGCTCTCTTTATCAGACTG	*EMCN*	19	Endomucin	0.716120299
GUIDES_sg044	GTTTTAGAAGGTGATGCATC	*EMCN*	15	Endomucin	0.500046723
GUIDES_sg045	AGGACTCCCAGGGACACCTG	*EMID1*	11	Collagen	0.633805274
GUIDES_sg046	GCTGCCCAGCAGAGCCTTGG	*EMID1*	13	Collagen	0.50154336
GUIDES_sg047	GCAGTGGACCAATCCAGCTA	*ETS1*	13	Ets	0.590477455
GUIDES_sg048	CACTAAAGAACAGCAACGAC	*ETS1*	8	SAM_PNT	0.507782957
GUIDES_sg049	ATGTTCAGTTGTAGGCACAA	*FBXO32*	7	-	0.680491123
GUIDES_sg050	AACTTGTCCGATGTTACCCA	*FBXO32*	8	-	0.665922839
GUIDES_sg051	ACAGTTCAGAGAGTGACGTG	*FER*	26	Pkinase	0.744231001
GUIDES_sg052	ATGTCTCGTCAAGAGGATGG	*FER*	25	Pkinase	0.666176654
GUIDES_sg053	GCCTGCAGATTAGCCTCCAA	*FERMT2*	17	PH	0.635135622
GUIDES_sg054	CTGAGGTTCATCTGATGAGC	*FERMT2*	15	PH	0.496718402
GUIDES_sg055	GGAGTTCATCCTCAACAATG	*FMNL2*	23	FH2	0.656016612
GUIDES_sg056	TTCACAAACCGGACAAAGAC	*FMNL2*	24	FH2	0.52306371
GUIDES_sg057	TGTGTACACACTACAGCTGG	*FNDC3B*	31	fn3	0.726749897
GUIDES_sg058	GCTCTTCCCAGTTCAGTACA	*FNDC3B*	30	fn3	0.69274705
GUIDES_sg059	GCTGAGCAACAAGACAGAGG	*GAS7*	19	-	0.715444147
GUIDES_sg060	GCTTGCGAAGGTCGGCAATG	*GAS7*	18	-	0.692490099
GUIDES_sg061	GCCCACACTCTCCAAGCACA	*GNB1L*	3	-	0.656080017
GUIDES_sg062	GCAGGCTCCAGATGTGTACC	*GNB1L*	2	WD40	0.572759585
GUIDES_sg063	TGAACAGAGAGACTTCTGAG	*KALRN*	59	PH	0.713929927
GUIDES_sg064	CTTCCTGAGATACAGTGAGA	*KALRN*	56	RhoGEF	0.590046749
GUIDES_sg065	TGATTACTGGAAGTACGGGG	*KREMEN1*	5	WSC	0.732200895
GUIDES_sg066	TTACTGGTGCCAGTTAGAGG	*KREMEN1*	4	WSC	0.660612633
GUIDES_sg067	CAGGGACTCGATGATCATGG	*LMO7*	34	LIM	0.767439497
GUIDES_sg068	GATCCTGACTTCAGCTCCTG	*LMO7*	35	LIM	0.666458078
GUIDES_sg069	CTTCGACGAGACCTCGAAGG	*LMX1B*	4	Homeobox	0.713403988
GUIDES_sg070	GTGCAAGGGTGACTACGAGA	*LMX1B*	3	LIM	0.675803927
GUIDES_sg071	ACGTCTCGGATGGTGCTGAG	*ME3*	18	Malic_M	0.711900963
GUIDES_sg072	AGAGAAAGAAGGTGTACCGA	*ME3*	14	Malic_M	0.683948993
GUIDES_sg073	TAGTACTTCCCATGTGCCAG	*MECOM*	24	-	0.69054325
GUIDES_sg074	ACTGTGGCAAGATTTTTCCA	*MECOM*	20	zf-C2H2	0.638498746
GUIDES_sg075	GGACTTCTGCTCAAAGAGGG	*MYOF*	56	-	0.689398449
GUIDES_sg076	TGCATGGGTTGGTGAACCAG	*MYOF*	58	-	0.68114604
GUIDES_sg077	AAAGGTACTCTGAAACATGG	*PARD3B*	24	-	0.748915706
GUIDES_sg078	TGGTCTCTTTCTGGAGACAG	*PARD3B*	25	-	0.687131379
GUIDES_sg079	TCTGGGAGATGAGCAAGCAG	*PDE7B*	11	PDEase_I	0.63710316
GUIDES_sg080	TCTTTCTGTTGATTACAAAG	*PDE7B*	12	PDEase_I	0.626787508
GUIDES_sg081	CGTGAGACTCCAGTCACAGG	*PKHD1*	20	-	0.721146822
GUIDES_sg082	ATGGGATAGCCCCAAGCAGG	*PKHD1*	16	-	0.648027027
GUIDES_sg083	TACTCAGGGGATCACCAGCG	*PLEKHA7*	28	-	0.733493753
GUIDES_sg084	CCCCGAACTCTACAGCCCAG	*PLEKHA7*	25	-	0.727012157
GUIDES_sg085	GCTGCCCACTGCATACACGA	*PRSS23*	3	Trypsin	0.707785666
GUIDES_sg086	AACATCAGTGAAGTTATCCA	*PRSS23*	3	Trypsin	0.589826381
GUIDES_sg087	CAGTGGTGTCGGGAACACCG	*PTPRJ*	26	Y_phosphatase	0.752294442
GUIDES_sg088	GTTCGGTAAAGGTCCTTGTG	*PTPRJ*	24	Y_phosphatase	0.742428198

GUIDES_sg089	TGGCAAAAAGGTTTCCATCG	*RALGPS1*	25	PH	0.63983389
GUIDES_sg090	CGAAAGAAGATAATTACAAG	*RALGPS1*	11	RasGEF	0.632295378
GUIDES_sg091	AGAGGTACCAGATGGGACTG	*RUNX2*	5	Runt	0.707376125
GUIDES_sg092	CATGGCGGAAGCATTCTGGA	*RUNX2*	11	RunxI	0.681198759
GUIDES_sg093	TGGAATTCCCTACCACAGCG	*SPTBN1*	36	PH	0.712246448
GUIDES_sg094	TCAGTCTTAACCATTCCCAT	*SPTBN1*	31	Spectrin	0.683179693
GUIDES_sg095	GGGCTGGCTATGATAAACTG	*TES*	6	LIM	0.774116567
GUIDES_sg096	CCATGAGTTGTCTCCCAGAG	*TES*	5	PET	0.735126362
GUIDES_sg097	GAAGCTTCCGAGAGTCTCTG	*TIMP3*	3	TIMP	0.703685264
GUIDES_sg098	CTATGATGGCAAGATGTACA	*TIMP3*	4	TIMP	0.591387438
GUIDES_sg099	AGTCCTTGGATGTAAGAAAG	*TMCO1*	9	DUF841	0.652850785
GUIDES_sg100	GAAACAATAACAGAGTCAGC	*TMCO1*	5	DUF841	0.622258097
GUIDES_sg101	AGAGACTTTGAAGTGAACGA	*TNS1*	42	PTB	0.715095311
GUIDES_sg102	CAGAAGGTGACAGTGTTGAG	*TNS1*	43	PTB	0.675343539
GUIDES_sg103	GCCGACTGGTGACCTCATGG	*TRIOBP*	3	-	0.713915639
GUIDES_sg104	GGGAGCAGGAGGCAGGAACG	*TRIOBP*	4	-	0.656603289
GUIDES_sg105	TAAACCACTGGAGTTCACGG	*TXNRD2*	20	Pyr_redox_dim	0.785275668
GUIDES_sg106	TCATCATTGCTACTGGAGGG	*TXNRD2*	8	Pyr_redox_2	0.706088692
GUIDES_sg107	GGTGAAGCTCCTGATTGCAG	*ZNF280D*	27	-	0.700021186
GUIDES_sg108	GAAGAAAGTAAAAGAAGTTG	*ZNF280D*	15	-	0.599965492
GUIDES_sg109	ATGGAGTTCCGCGACCACGT	*ABO*	7	CDS	0.6563
GUIDES_sg110	CCGGTCCCCAGCGTCACGCG	*ABO*	7	CDS	0.6687
GUIDES_sg111	CCACCTGGTACATCGCCTCA	*TEX41*	2	TRANSCRIPT	0.6631
GUIDES_sg112	AACTCAAGACATTGGAACCA	*TEX41*	5	TRANSCRIPT	0.6251
GUIDES_sg113	AATGTGGTAGCCCAAGACAG	*CYP1B1*	5	p450	0.775320729
GUIDES_sg114	GTGGCCACTGATCGGAAACG	*CYP1B1*	3	p450	0.726256031
GUIDES_sg115	GCAAGCCATGAGCCTGTACG	*FOXC1*	1	-	0.747984594
GUIDES_sg116	TCGTCGTCCCTGAGTCACGG	*FOXC1*	1	-	0.730741126
GUIDES_sg117	GATTGTGGTGAACTTCCGTG	*GMDS*	8	Epimerase	0.727535334
GUIDES_sg118	GTTGCAGAATGATGAGCCGG	*GMDS*	10	Epimerase	0.65599947
GUIDES_sg119	CCTCCCGCACGCGCACACAG	*LTBP2*	35	EGF	0.755863507
GUIDES_sg120	CAGGCAGACATAACCAGGCA	*LTBP2*	31	EGF_CA	0.708717
GUIDES_sg121	GGTCATACTCAAAAACCTGG	*MYOC*	3	OLF	0.763937898
GUIDES_sg122	ATGCCAGTATACCTTCAGTG	*MYOC*	1	-	0.722928246
GUIDES_sg123	TCTTGCGAAGGAAGTCCAGA	*TEK*	17	Pkinase	0.627785812
GUIDES_sg124	ATCTAATGAGACAATGCTGG	*TEK*	22	Pkinase	0.626682424
NonTargeting Human_0001	ACGGAGGCTAAGCGTCGCAA	-	-	-	-
NonTargeting Human_0002	CGCTTCCGCGGCCCGTTCAA	-	-	-	-
NonTargeting Human_0003	ATCGTTTCCGCTTAACGGCG	-	-	-	-
NonTargeting Human_0004	GTAGGCGCGCCGCTCTCTAC	-	-	-	-
NonTargeting Human_0005	CCATATCGGGGCGAGACATG	-	-	-	-
NonTargeting Human_0006	TACTAACGCCGCTCCTACAG	-	-	-	-

NonTargeting Human_0007	TGAGGATCATGTCGAGCGCC	-	-	-	-
NonTargeting Human_0008	GGGCCCGCATAGGATATCGC	-	-	-	-
NonTargeting Human_0009	TAGACAACCGCGGAGAATGC	-	-	-	-
NonTargeting Human_0010	ACGGGCGGCTATCGCTGACT				

593

Supplementary table 2 - Tensorflow CNN architecture

Layer	Output shape	Trainable parameters
conv2d (Conv2D)	(None, 54, 54, 96)	34,944
activation (relu)	(None, 54, 54, 96)	0
max_pooling2d (MaxPooling2D)	(None, 27, 27, 96)	0
conv2d_1 (Conv2D)	(None, 17, 17, 256)	2,973,952
activation_1 (relu)	(None, 17, 17, 256)	0
max_pooling2d_1 (MaxPooling2D)	(None, 8, 8, 256)	0
conv2d_2 (Conv2D)	(None, 6, 6, 384)	885,120
activation_2 (relu)	(None, 6, 6, 384)	0
conv2d_3 (Conv2D)	(None, 4, 4, 384)	1,327,488
activation_3 (relu)	(None, 4, 4, 384)	0
conv2d_4 (Conv2D)	(None, 2, 2, 256)	884,992
activation_4 (relu)	(None, 2, 2, 256)	0
max_pooling2d_2 (MaxPooling2D)	(None, 1, 1, 256)	0
flatten (Flatten)	(None, 256)	0
dense (Dense)	(None, 4096)	1,052,672
activation_5 (relu)	(None, 4096)	0
dropout (Dropout)	(None, 4096)	0
dense_1 (Dense)	(None, 4096)	16,781,312
activation_6 (relu)	(None, 4096)	0
dropout_1 (Dropout)	(None, 4096)	0
dense_2 (Dense)	(None, 1000)	4,097,000
activation_7 (relu)	(None, 1000)	0
dropout_2 (Dropout)	(None, 1000)	0
dense_3 (Dense)	(None, 2)	2,002
activation_8 (softmax)	(None, 2)	0
Total trainable parameters		28,039,482