Targeted single-cell RNA sequencing of transcription factors facilitates biological insights from human cell experimental models

Alexandra Pokhilko1,^, Adam E. Handel1,^, Fabiola Curion2, Viola Volpato3, Emma S. Whiteley4, Sunniva Bøstrand4, Sarah E. Newey4, Colin J. Akerman4, Caleb Webber3, Michael B. Clark5,6,*, Rory Bowden2,7,8,*, M. Zameel Cader1,*

^first authors

*corresponding authors

1 Translational Molecular Neuroscience Group, Weatherall Institute of Molecular Medicine, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DS, United Kingdom.

2 Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom

3 UK Dementia Research Institute, Cardiff University, Cardiff, UK.

4 Department of Pharmacology, Mansfield Road, University of Oxford, Oxford, OX1 3QT, United Kingdom.

5 Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, OX3 7JX United Kingdom.

6 Centre for Stem Cell Systems, Department of Anatomy and Neuroscience, The University of Melbourne, Parkville, VIC, 3010, Australia

7 The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052 Australia

8 University of Melbourne, Department of Medical Biology, 1G Royal Parade, Parkville, VIC 3052 Australia

Corresponding author details: Zameel Cader, Translational Molecular Neuroscience Group, Weatherall Institute of Molecular Medicine, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DS, United Kingdom. Telephone: 01865 231858; E-mail zameel.cader@ndcn.ox.ac.uk

Running title: Single cell capture sequencing

Key words: single cell RNA sequencing, iPSC neurons, targeted sequencing, transcription factors, neuronal differentiation.
Abstract

Single cell RNA sequencing (scRNA-seq) is a widely used method for identifying cell types and trajectories in biologically heterogeneous samples, but is limited in its detection and quantification of lowly expressed genes. This results in missing important biological signals, such as the expression of key transcription factors (TFs) driving cellular differentiation. We show that targeted sequencing of ~1000 TFs (scCapture-seq) in iPSC-derived neuronal cultures greatly improves the biological information garnered from scRNA-seq. Increased TF resolution enhanced cell-type identification, developmental trajectories and gene regulatory networks. This allowed us to resolve differences amongst neuronal populations, which were generated in two different labs using the same differentiation protocol. ScCapture-seq improved TF gene-regulatory network inference and thus identified divergent patterns of neurogenesis into either excitatory cortical neurons or inhibitory interneurons. Furthermore, scCapture-seq revealed a role for of retinoic acid signalling in the developmental divergence between these different neuronal populations. Our results demonstrate that TF targeting improves the characterization of human cellular models and allows identification of the essential differences between cellular populations, which would otherwise be missed in traditional scRNA-seq. ScCapture-seq TF targeting represents a cost-effective enhancement of scRNA-seq, which could be broadly applied to improve scRNA-seq resolution.

Introduction

Single cell RNA-seq (scRNA-seq) is widely used to elucidate the biology of complex heterogeneous samples. However, scRNA-seq libraries commonly suffer from high dropouts (false zero expression estimates) and variability due to both technical variation and the biological stochasticity of gene expression in individual cells (Kolodziejczyk et al. 2015). This means many scRNA-seq reads come from highly expressed genes that may not be informative for cell fate or function, while the biological information conveyed by other genes is lost. Low abundance essential genes, a class which includes many transcription factors (TFs), are particularly affected by the limitations of scRNA-seq (Lambert et al. 2018). TFs are prime regulators of gene expression that underpin much of cell biology from cell fate specification to disease processes. Therefore, a method to
improve the detection and quantification of TFs in single cells would significantly increase the ability of scRNA-seq to address key biological and disease questions.

Human induced pluripotent stem cell (iPSC)-derived neuronal cultures are used in medical and research applications (Fernandez Tde et al. 2013). However, high biological heterogeneity in developing neuronal cultures might hinder their wide-spread application (Volpato et al. 2018). To ensure reproducible generation of specific cell types from iPSC progenitors under the same protocol, it is crucially important to be able to compare differentiation outcomes within and between laboratories. We previously utilised scRNA-seq to examine the performance of an iPSC model of forebrain corticogenesis across multiple laboratories, identifying clear laboratory-dependant variability (Volpato et al. 2018). Therefore, we hypothesised in-depth single cell profiling of TFs could allow us to decipher the differences in cell types and differentiation trajectories responsible for this variability.

We previously described RNA Capture-Seq, a method able to improve transcript detection and quantification in bulk samples with targeted RNA-seq (Mercer et al. 2014). Capture-Seq utilises oligonucleotide probes targeted to gene or genomic regions of interest to enrich for expression from these regions, and is especially useful when targeting lowly expressed transcripts (Clark et al. 2015; Bartonicek et al. 2017), even when samples are limiting (Curion et al. 2020). scRNA-seq enrichment methods to-date have focused almost exclusively on the small family of highly-expressed immune receptor genes, while Capture-Seq potentially provides the scalability and sensitivity required to be widely applicable (Riemondy et al. 2019; Saikia et al. 2019; Singh et al. 2019). Moreover, another potential advantage of Capture-Seq is a reduction in sequencing costs, as fewer reads are required per cell when sequencing is targeted to a small portion of the transcriptome. We now describe the application of a targeted approach to single-cell sequencing libraries, which we term scCapture-seq, using probes against 972 TFs (Curion et al. 2020). We applied our method to existing libraries generated from cultures of iPSC-derived cortical neurons, differentiated in two different labs using the same standard protocol (Volpato et al. 2018). We aimed, firstly, to demonstrate scCapture-seq could overcome some of the limitations of scRNA by improving its sensitivity and, secondly, that targeting TFs with...
scCapture-seq would uncover biology not visible with standard scRNA-seq by enhancing the identification of cell types and trajectories.

Results

Improved identification of transcription factors in single cell RNA sequencing

Bulk RNA-seq and scRNA-seq was previously performed on iPSC-derived cortical neurons generated by two independent labs using identical cell lines and following the same established methods (referred to as Lab D and Lab E). This study revealed large variability in molecular phenotypes, highlighting factors underlying inter-lab variation. Cells for scRNA-seq were collected after 85 days differentiation and maturation in vitro (Volpato et al. 2018). We implemented TF capture on the single cell libraries and analysed 279 cells, out of 376 captured cells, which passed quality control (Methods, Fig. 1a). scCapture-seq resulted in an increase in the median number of mapped reads from 63.1% to 89.5% post-capture (Fig. 1b), whilst reads mapped to targeted TFs increased from 2.2% to 78.3% post-capture, resulting in a 36-fold enrichment for the transcripts of interest (Fig. 1c). There was consistent capture performance across all targeted genes (>10-fold enrichment; Supplemental Fig. S1a). Target gene expression levels were on average increased ~150-fold post-capture (Supplemental Fig. S1b) and highly correlated (R=0.97) with pre-capture expression. In the pre-capture data, 585 of the targeted TFs were expressed with non-zero mean counts, whilst post-capture there was a 25% increase in the number of TFs detected, for a total of 731 expressed TFs. Furthermore, the number of TFs detected in each cell increased more than 4-fold (Fig. 1d). These results demonstrate excellent performance of scCapture-seq on single cell libraries, allowing sensitive detection of the key transcriptional regulators.

TF capture improves cell clustering and identification

Clustering across 13299 expressed genes with pre-capture data identified 4 populations of cells (Fig. 2a, Supplemental Fig. S2). Using markers of human brain cells (van de Leemput et al. 2014; Darmanis et al.)
2015; Song et al. 2017), cluster 1 (lab D) and clusters 3 and 4 (lab E) were identified as neurons and cluster 2 (mainly lab E) as glial cells (Supplemental Fig. S2). As master transcriptional regulators, clustering by TFs (which in scRNA-seq datasets are under-represented due to generally low expression), might improve assignments of cell identity. Clustering based on both pre-capture and post-capture TF expression (585 vs 731 expressed TFs) divided cells into three groups, which we termed N1 (neuronal-like 1), N2 (neuronal-like 2) and G (glial-like) (Fig. 2b, Supplemental Fig. S3a,b). The capture improves the clustering by creating a clearer distinction between the clusters (Supplemental Fig. S3a,b) and increasing the number of differentially expressed TFs from 129 to 155 post-capture (Supplemental Fig. S3d,e). The neurons from lab E, (pre-capture clusters 3 and 4), coalesced into a single post-capture cluster N1 (Fig. 2c). This merge is consistent with the differential expression of only 4 TFs (BHLHE22, NFIX, ZBTB20 and NHLH1) between pre-capture clusters 3 and 4 and the observed similarity between these clusters pre-capture (Fig. 2a, Supplemental Fig. S2a,c).

TF capture improved the precision of hierarchical clustering (Supplemental Fig. S3), increasing the number of differentially expressed genes (DEGs) in the N2 cluster (lab D) from 9 pre-capture to 53 post-capture. In addition, partial re-assignment of G cells from lab E to the N1 cluster (Supplemental Fig. S3a,c) increases the number of post-capture N1 DEGs, while decreasing the number of DEGs in cluster G (Supplemental Fig. S3d,e). In further support of the biological value of the capture we next analysed the expression of 107 key TFs involved at early stages of neuronal differentiation (Inoue et al. 2019). We found that twice as many cells expressed key neural induction TFs post-capture and, moreover, there was more homogenous expression of critical TFs in each cluster (Supplemental Fig. S4).

Closer inspection of TF DEGs revealed that cells of the glial-like cell cluster G express TFs that included HES1, HESS, PAX6, NR2E1, TCF7L2, OTX1, as well as GFAP, consistent with a radial glial or proliferative neural progenitor identity (Supplemental Fig. S5)(Darmanis et al. 2015; Telley et al. 2016). This was unexpected, since a reliance upon the standard cell identity markers suggested an astrocytic glial identity (as previously reported (Volpato et al. 2018)). Neuronal-like N1 cells expressed neuronal TFs such as NEUROD2, TBR1, NEUROG2, TBR2 (EOMES), MYT1L and BHLHE22, consistent with the development of deep layer excitatory cortical neurons. In contrast, neuronal-like N2 cells expressed a distinct set of TFs, which included MEIS1, SP9, DLX2.
and DLX6, signifying an interneuron identity (Lim et al. 2018). Hence TF based cell-identity classification using capture sequencing data uncovered unappreciated details about the cell types present in the iPSC neuronal cultures. Specifically, an apparent glial cell cluster was in fact a neural progenitor population and neuronal cell clusters from two labs were distinguished as excitatory cortical neurons and inhibitory interneurons. It is therefore likely that different differentiation trajectories in Lab D and E were the principal cause of variation in phenotypic outcomes for the experiment described by Volpata et al. (Volpato et al. 2018) rather than differences between neuron and glial ratios, as had been concluded.

The generation of GABAergic interneurons from a cortical differentiation protocol that has been validated to produce glutamatergic excitatory neurons and astrocytes, was not expected (Shi et al. 2012a; Shi et al. 2012b). However, although was potentially consistent with recent findings (Strano et al. 2020). Therefore, to corroborate our transcriptomic observations, we performed immunohistochemistry and patch clamp electrophysiology on neurons differentiated in Lab D. In accordance with our scCapture-seq findings, these cultures contained GAD65 immunopositive neurons (Fig. 3 a-c) and exhibited functional GABAergic synaptic connections (Fig. 3 d-f). Our data further support the validity of using TF targeting for improved classification of neuronal populations.

TF capture improves the resolution of Gene Regulatory Networks

To investigate whether our TF scCapture-seq might enhance TF-gene networks underpinning the cell clusters, we built a co-regulation network of co-expressed genes using the whole transcriptome. Including imputation of post-capture TF expression, re-scaled accordingly to the differences in library sizes, resulted in improved resolution of the gene regulatory network (GRN), which contained 1.5 times more nodes and 2 times more edges (Methods). We also constructed a network focusing solely upon DEGs between clusters. The pre-capture GRN (Fig. 4a) had poor representation of N2 cluster genes, while N1 cluster genes were disconnected from the main network. Utilising the post-capture information, the imputed GRN had 517 nodes (121 TFs) and 1939 edges (Fig. 4b), with much improved N2 cluster representation (47 nodes, 591 edges); whilst the N1 cluster genes were now integrated inside the network. Both pre-capture and imputed GRNs
were enriched in the targeted TFs (P < 2 \times 10^{-7}, calculated using network enrichment analysis, NEA(Alexeyenko et al. 2012)). However, the imputed network identified substantially more connections between genes, greatly increasing network degrees (numbers of genes connected to each node), in particular for the hub genes (Fig. 4c). Furthermore, the number of TF targets present in the GRNs was significantly increased, from 50% pre-capture to 70% post-capture (Fig. 4d; Fisher’s exact test P = 4 \times 10^{-9}; Supplemental Table S1), suggesting that newly identified TFs post-capture are functionally related to DEGs between clusters.

Moreover, our analysis of known protein-protein interactions (PPIs) between TFs and other proteins (Kuleshov et al. 2016) revealed that TF PPIs links, which are present in the GRNs were significantly enriched in the imputed post-capture GRN clusters (Fig. 4e, P value of exact Wilcoxon-signed-rank test for the effect size is 1.9 \times 10^{-6}; Supplemental Table S2). Overall, incorporation of the capture data improved the resolution of the GRNs and enabled mechanistic insights into TF-mediated biological networks.

**TF capture facilitates the identification of TFs associated with developmental trajectories**

To explore the differentiation trajectories and TFs involved in cell fate determination in each cluster, we ordered cells in pseudotime (Angerer et al. 2016). The expression profiles of TFs with known maturation kinetics (Telley et al. 2016) were used to establish a pseudotime direction (Fig. 5a-d, Supplemental Fig. S5), in which cells of the G cluster occupied an earlier position, consistent with their neural progenitor identity (Darmanis et al. 2015; Telley et al. 2016). Using either pre- or post-capture genes resulted in a qualitatively similar ordering of cells (with G cells being the earliest on pseudotime, followed by two later branches of N1 and N2 cells), but the precision of expression change estimates for individual genes was greatly improved post-capture (Fig. 5e,f; Supplemental Fig. S5). This was manifested by the increased numbers of cells with detected TFs. For example, the number of cells expressing HES1 increased from 56 to 224 post-capture (Fig. 5e,f).

Pseudotime ordering revealed distinct developmental trajectories for neurons from Lab D (N2) and Lab E (N1). Also, TF capture, but not pre-capture data, distinguished the G cells from the two laboratories, suggesting that a common progenitor state was not represented in the cultures at the point of harvest (Fig. 5a-d).
To identify the set of TFs that might be involved in the developmental switch between the two neuronal subpopulations, we performed a differential expression between the respective clusters of N1 and N2 neurons (Fig. 6a,b, Supplemental Fig. S6a), or between G cells from different labs (Supplemental Fig. S6b). TFs were differentially expressed between N1 and N2 post-capture (Fig. 6b), compared to only 65 TFs pre-capture (Supplemental Fig. S6a). Importantly, amongst these post-capture exclusive DEGs, there were multiple TFs whose kinetics over pseudotime differed between the labs (SALL3, GSX2, BHLHE40, SOX3, TFAP2C, GLI53, POU3F2, SP4, CREBRF; Fig. 6c,d). The post-capture expression of lab-specific TFs revealed different trends in pseudo-temporal kinetics in two labs. Lab E cells showed a clear temporal ordering of TF expression, recapitulating the expected progression during development. Initially many neural progenitor genes were strongly expressed in Lab E cells, before their expression was extinguished and cells instead successively expressed cortical excitatory neuron TFs such as TBR2 (EOMES), TBR1 and FOXG1. In contrast, fewer Lab D cells expressed neural progenitor markers and the N2 neurons had less pronounced pseudotemporal patterns, suggesting a more uniform, mature developmental state. The G cells from Lab D strongly expressed HES5 and SALL3, whilst Lab D neurons expressed DLX2, DLX6 and SP9, which define a particular interneuron program (Lim et al. 2018).

TF capture identifies retinoic acid signalling as a candidate mechanism driving differences in developmental trajectories

We hypothesised that scCapture-seq may reveal previously unappreciated details of how a differentiation process could generate variability when the same protocol is applied to the same cell line. Indeed, the TF PPIs of the N2 cells showed significant evidence for retinoic acid (RA) signalling (RARA, RXRA and CEBPB), that was not apparent with pre-capture data alone (FDR = 1 × 10^-5 post-capture vs 0.02 pre-capture for RARA; Fig. 4e). We therefore used the post-capture imputed GRN from Fig. 4b to explore the potential RA-related effects. The RA subnetwork highlighted a highly interconnected network, with multiple direct connections between RA TFs (e.g., RORA, NR1H2, ZNF423) and N2-specific GABAergic (interneuron) genes such as GAD1, GAD2 and SLC6A1, as well as interneuron TFs such as DLX2 and DLX6 (Supplemental Fig.
S7a). The RA TFs were negatively correlated to the interneuron TFs, consistent with the downregulation of RA TFs in N2 neurons (Supplemental Fig. S7b) and available evidence on the induction of interneuronal genes MEIS and DLX by the abrogation of RA signalling (Wahl et al. 2018). Together, these observations identify RA signalling as a candidate mechanism, potentially mediating the inter-lab differences in cell fate choices.

In summary, these data establish that scCapture-seq targeting of TFs can improve the sensitivity for resolving cell-type differences amongst neuronal populations and identify the changes in gene regulatory networks that underlie these differences.

Validation of scCapture-seq approach with an independent droplet-based scRNA-seq sample

To further validate the scCapture-seq approach, we applied it to an independent pool of scRNA-seq libraries made using the 10x Genomics droplet-based protocol. This sample of intestinal stromal cells from ulcerative colitis (UC) patients has been previously described and is characterized by the expression of inflammatory genes (Kinchen et al. 2018). We performed scCapture-seq with our TF panel on the 126 out of 191 single-cell libraries that passed quality control (Methods). TF expression levels in post-capture libraries were on average increased 60-fold compared to pre-capture libraries, with a high correlation (R=0.91) between pre- and post-capture expression (Supplemental Fig. S8a). All the targeted TFs were enriched post-capture (> 10-fold, Supplemental Fig. S8b), leading to ~ 3-fold increase of the number of captured TFs per cell (Supplemental Fig. S8c). We next clustered the pre-capture cells using all genes with non-zero mean counts, which revealed two distinct clusters amongst the 125 cells (Supplemental Fig. S8d). Importantly, two very similar clusters were observed post-capture (Supplemental Fig. S8d), further confirming that using post-capture libraries, expressing only the targeted TFs, is sufficient to characterize cellular populations. Moreover, we found that several TFs, which are expected to be involved in intestinal inflammatory processes, such as TCF7L2, PBX1, TCF21 and FOXO3 (Chawla et al. 2013; Kim et al. 2017; Crow and Gillis 2018) were differentially expressed between the two clusters post-capture, but not pre-capture (Supplemental Table S4, Supplemental Fig. S8 e,f). Overall, using an independent scRNA-seq dataset we confirmed that scCapture-seq greatly improved the identification of the targeted TFs, potentially facilitating the discovery of important regulatory
TFs underlying the differences between cellular populations. These results also establish that scCapture-seq works robustly on both whole-transcript plate-based and 3’-end droplet-based sample libraries, although further optimisation for non-Smart-seq libraries types would be possible (see Discussion).

Validation of scCapture-seq with a separate target gene panel

We next validated the scCapture-seq approach with the NeuroGWAS (NG) panel (Curion et al. 2020), whose targets genes include markers for neuronal development and identity and candidate risk genes for schizophrenia and Parkinson’s disease. We analysed 359 out of 376 captured iPSC-derived cells from cortical neuron cultures (Volpato et al. 2018), which passed quality control. 87 out of 101 NG genes were expressed post-capture. Similarly to TF capture, there was a linear correlation between pre- and post-capture expression, >10-fold enrichment of NG genes and a 3.8-fold increase in the number of captured NG genes per cell (Supplemental Fig. S9a-c). Moreover, clustering identified 5 similar cellular communities in pre- and post-capture data (Supplemental Fig. S9d), confirming the validity of using post-capture for cell type identification. Using the cell type markers we confirmed that, similarly to TF capture cultures, NG capture cultures were represented by neuronal cells from lab D (clusters 1,2) and lab E (clusters 3,4), as well as glial-type cells (cluster 5; Supplemental Fig. S10). Differential expression of NG genes revealed that most of the candidate disease genes, e.g., BAG3, RAB29, STX4 and FAM126A potentially associated with Parkinson’s disease and CACNB2, EMX1 and AKT3 (potentially associated with schizophrenia) had higher expression in lab E neurons and/or glial cells, with only HIP1R having higher expression in lab D neurons (Supplemental Fig. S9e,f). One partial reason for these observations could be the less mature, more developing state of lab E neurons. Crucially, these results demonstrate that scCapture-seq can also be successfully applied to other sets of capture pools, including NG genes, to facilitate the identification of the disease genes in different cell types.

Discussion
We applied targeted sequencing to a large set of biologically critical genes that are typically poorly represented in single cell data, to investigate if this method can enhance the biological information discoverable using scRNA-seq. As a proof of principle, we targeted most known human TFs (~1000) and applied our approach to iPSC-derived neurons. We found that capture was highly effective at recovering TF single cell gene expression. Compared to pre-capture data we observed a 36-fold enrichment for TF reads, increasing the total number of TFs detected in the sample, while each cell was shown to express a broader range of TFs. The high correlation between post-capture and pre-capture expression suggested there was little bias in relative TF expression levels due to capture. Many of the key TFs were poorly represented pre-capture, precluding the downstream analysis of the neuronal fate specification based on pre-capture data alone. Our results demonstrate that performing TF scCapture-seq greatly improved our understanding of the underlying biology in the system and when combined with standard, full transcriptome scRNA-seq, allowed, at low additional cost, construction of a comprehensive gene regulatory network and additional insights into the processes driving neuronal differentiation. We also validated our approach on a different scRNA-seq dataset of intestinal stromal cells prepared using the 10x droplet-based approach, and on a different panel targeting neurological disease genes. In both cases we confirmed the enrichment of the targeted genes and the potential for scCapture-seq to uncover new biological insights.

Previous application of RNA Capture to bulk and low input samples has demonstrated that this method improves sequencing sensitivity, allowing the detection and quantification of transcripts which are poorly represented in standard libraries (Mercer et al. 2014; Clark et al. 2015; Curion et al. 2020). Here we demonstrate capture has similar advantages when applied to single cells, which are well known to have low sensitivity and a high noise threshold. Therefore, the benefit sequence capture provides to biological analyses of single cells may be greater than for bulk samples. Previous targeted single cell methods have demonstrated the enrichment of individual cells of interest by PCR for specific cell barcodes (Ranu et al. 2019), or immune receptor genes using LNA or DNA capture probes (Riemondy et al. 2019; Singh et al. 2019). Whilst these studies demonstrated that an improved biological understanding could be obtained through targeted single cell sequencing, the number of targeted cells or genes was very limited. PCR and LNA-probe based methods have practical and cost limitations for scaling to large number of targets, while targeting highly expressed
immune receptor genes left the applicability of single cell capture to lowly expressed target unresolved. Our results now demonstrate the utility of applying single cell capture to large numbers of genes, including those with low expression levels.

ScCapture-seq was primarily designed for Smart-seq2 libraries. However, we have also demonstrated that this method is compatible with 10x Genomics 3’ profiling droplet-based libraries from intestinal stromal cells. To ensure efficient target enrichment across a wide range of scRNA-seq library types we recommend utilising adaptor blocking oligonucleotides specifically designed for compatibility with the barcode and index features of each methodology. Additionally, for those single cell techniques that produce transcript counting (e.g., 10x Genomics, MARS-seq, etc.) rather than whole transcript libraries, capture probe designs could be optimised to cover only those 3’ or 5’ regions likely to be sequenced (Supplemental Fig. S11). One important difference between scCapture-seq and 10x Genomics Targeted Gene Expression is that our method can be applied to a range of single cell methodologies and library types. It can provide similar advantages to the 10x protocol on 10x libraries but is also compatible with methods like Smart-seq that preserve information across the full-length of transcripts and so potentially can be used for the identification of splice isoforms in a way that is not possible using 3’ or 5’ 10x Genomics Targeted Gene Expression measurements. Thus, TF capture resulted in a median 3-fold increase in detected splice junctions in our dataset, similarly to previously reported (Curion et al. 2020). scCapture-seq can be used as a cost-effective substitution for standard scRNA-seq, especially for larger projects, or multiple projects where the same capture design can be utilised. While cost savings depend upon many variables, including local costs and the number of additional samples able to be combined in a sequencing lane, we find capture-seq enables 5-10x less sequencing to be performed and saves money once 5 lanes of sequencing have been avoided. One important limitation to all single-cell enrichment method is that they can only enrich for genes present in the sequencing library and cannot overcome the inefficiencies of single cell library generation. As single cell library method efficiencies continue to improve (Hagemann-Jensen et al. 2020) the sensitivity of single cell capture will progressively increase.

The differentiation of iPSCs into defined neuronal populations provides an excellent opportunity to understand developmental processes, since these cultures typically contain a diversity of cell subtypes and
different stages of maturation. Protocols developed to produce forebrain cortical excitatory neurons are now well established and widely used. Typically, they involve dual SMAD inhibition to induce a neural ectoderm fate followed by maturation into a default forebrain cortical neuronal specification (Picelli et al. 2013). scRNA-seq has proved to be an excellent tool to characterise the heterogeneity of cultures, but without access to critical sets of genes, namely TFs, the potential to reveal mechanistic processes is significantly hampered. The data obtained from our TF capture highlights the importance of TF data in understanding cell biology and revealed unexpected differences in developmental programs arising from an identical differentiation protocol conducted on the same donor iPSC lines in two different labs. We found that Lab E neurons expressed TFs involved in the development of cortical excitatory neurons, while Lab D neurons expressed inhibitory interneuron TFs. This was unexpected, since the protocol employed does not include any factors to direct cells to an interneuron fate. We found that Lab D culture differentiate into a particular type of immature interneurons, expressing $DLX$ ($DLX2$, $DLX6$, $DLX1$) and $SP$ ($SP8$, $SP9$, $SP4$) TFs (Lim et al. 2018). Our results are in agreement with recent findings that cortical neuron specification from iPSCs is strongly modified by the fluctuations of factors affecting regional patterning of the brain, including RA levels and the activity of Wnt signalling pathway (Strano et al. 2020). For example, it was shown that insufficiency of Wnt signalling is capable of switching the culture from the formation of dorsal excitatory neurons to ventral inhibitory interneurons, particularly to the highly ventralised NKX2-1 type of interneurons (Strano et al. 2020). We found that Lab D culture differentiations included $DLX/SP$ expressing interneurons, which are distinct from NKX2-1-dependent interneurons (Lim et al. 2018), and may be less influenced by Wnt signalling (Strano et al. 2020). Other RA-mediated effects may therefore be more relevant for cell fate choices in our experiment.

To investigate the potential mechanism underlying the different development trajectories of neurons from different labs, we used the post-capture-inferred co-expression networks. We found that the interneuron subnetwork was significantly enriched in genes of retinoic acid (RA) signalling (e.g., $RORA$, $NR1H2$ and $ZNF423$). The RA genes were mainly downregulated in these neurons, but highly connected and anti-correlated to the interneuronal genes (e.g., $DLX2$ and $GAD1$), suggesting that interneuronal specialization is related to the suppression of the RA pathway in Lab D cultures. Reduction in RA signalling has previously been reported to upregulate key interneuronal TFs of the $DLX$ family (Wahl et al. 2018). RA is also known to be an
important factor in neuronal differentiation, having complex and concentration-dependant effects (Crandall et al. 2011). A potential explanation for the reduced RA signalling in lab D compared to lab E is differences in culture growth. Our previous work showed that passage number before differentiation, media volume changes and other factors contributed to the variability between labs, potentially affecting culture growth and differentiation rates (Volpato et al. 2018). In addition, our pseudotime analysis suggests that lab E cells are actively developing, while lab D cells are more stable. Therefore, it is possible that faster initial growth and development of lab D neurons caused the depletion of RA and altered cell fates. Overall, our analysis suggests that careful control of RA concentration during iPSC differentiation into cortical neurons might improve the reproducibility of the differentiation protocol.

To conclude, we showed that targeted TF sequencing, scCapture-seq, greatly improves the resolution of biological information derived from scRNA-seq. scCapture-seq alone can be used as cost-effective substitution for standard scRNA-seq, improving cell type characterization and the analysis of developmental trajectories. We applied scCapture-seq to previously published data on iPSC-derived neurons, and demonstrated that, due to greatly improved TF detection, scCapture-seq enabled the identification of key developmental TFs driving cellular heterogeneity. We subsequently showed that combining TF scCapture-seq with standard, whole transcriptome scRNA-seq recovered more comprehensive gene regulatory networks than scRNA-seq alone, in this case implicating retinoic acid signalling as a key factor in cellular heterogeneity.

Our approach has widespread application, since correct detection of TF expression in single cells will enable improved identification of cell types, trajectories and GRNs present during development, physiological or pathological states, or in response to drug perturbations. ScCapture-seq could also be adapted to capture other potentially interesting genes using bespoke oligonucleotide probes. Our approach could be immediately applied to any existing Smart-seq2 single-cell RNA-seq library to cost-effectively enhance the biological insights available with scRNA-seq.
We re-sequenced our previous scRNA-seq libraries (Volpato et al. 2018), using targeted sequencing (scCapture-seq) for 972 known human TFs, as described in (Curion et al. 2020). Briefly, the cell cultures were differentiated from human iPSCs derived from dermal fibroblasts using the dual SMAD inhibition protocol and collected on day 55 after plating. A potential source of vitamin A for endogenous RA synthesis in the cells of these cultures was a media B27 supplement (Volpato et al. 2018). Libraries were prepared from 376 day 55 neurons (and 8 “mini-bulk” samples) using the Smart-seq2 protocol (Picelli et al. 2013) and then utilized for both pre- and post-capture sequencing. Each sample was spiked with the equivalent of 1 μL of a 1:10,000,000 dilution of the ERCC RNA Spike-In Mix 1 (Thermo Fisher Scientific). Libraries were pooled at 384-plex and each pool sequenced on 1 lane of a HiSeq 4000 at 75 bp paired end. The average size of post-capture libraries were 4 times larger than the size of pre-capture libraries. The cells were from two different labs (named D and E in (Volpato et al. 2018)) and our analysis confirmed the previous observation of the differences between the cell populations originating from different labs (batches). Since the cells from different labs showed substantial differences, we did not correct for the batch effect in our analysis.

TF-Capture

The oligonucleotide capture probes were designed to target 972 human DNA-binding TF genes (Curion et al. 2020). These high-confidence TFs were previously chosen based on Tfcheckpoint.org database (Chawla et al. 2013), utilising a manually curated list of sequence-specific DNA-binding transcription factors (DbTF). While the TF capture design could be extended to include additional recently identified TFs, it is important to be mindful of the proportion of the transcriptome captured, as this affects the enrichment. Our previous modelling of the effect of the capture pool size upon enrichment suggested that the chosen set of ~1000 TFs enables successful enrichment, whilst targeting a large number of TFs that represents the biological diversity in human cells (Curion et al. 2020). Probes were also present for 42 abundant TFs, targeted in part and 221 control brain-specific genes, which were not used in the present analysis, as well as sets of synthetic control genes including SIRVs, Sequins and a partial set of the ERCC spikes-in, comprising 56 of the 92 ERCCs. Target regions were trimmed to remove any potential unwanted overlap with non-target genes, highly expressed RNA repeat or pseudogenes. A full description of the probe design can be found in (Mercer et al. 2014; Curion
Probe synthesis was performed by Roche NimbleGen. Target capture was performed as previously described (Curion et al. 2020) using the protocol from (Mercer et al. 2014) with slight modifications. Briefly, 850 ng of pooled Smart-seq2 libraries were utilised for capture hybridization along with Cot1 and blocking oligos (xGen® NXT Universal Blockers, IDT). Hybridisation was performed for 3 days. Post-capture LMPCR was performed for 12 cycles as per SeqCap RNA Enrichment System User’s Guide V1.0 (Roche) with KAPA Taq and Roche post-cap LMPCR primers, except that PCR input was 17ul of resuspended capture beads. QC of captured libraries was performed by Qubit (Thermo Fisher Scientific) and Tapestation (Agilent) to measure post-capture library concentration, yield and size distribution. Successful capture enrichment was confirmed by qPCR (QuantStudio 6, Thermo Fisher Scientific). Libraries were sequenced at the Wellcome Trust Centre for Human Genetics (WTCHG) (Oxford, UK).

scRNA-seq analysis

The scRNA-seq libraries were analysed similarly to (Volpato et al. 2018), except that we used more recent references genomes hg38.p10 with GENCODE Release 26 annotation. The average size of pre-capture libraries was 1.5 million read pairs. The contamination with ribosomal rRNA was removed with the Sortmerna package (Kopylova et al. 2012). The reads were mapped to the hg38 genome using the STAR aligner (Dobin et al. 2013), as before. The average proportion of uniquely mapped reads was 59%. Gene counts for the mapped reads were obtained using the FeatureCounts function of the Subread package (Liao et al. 2013). For the analysis we retained only genes expressed with mean counts > 1 and filtered out the low quality cells, based on pre-capture libraries, as before (Volpato et al. 2018). Briefly, we filtered out the cells expressing less than 2000 genes; having initial library size less than 0.5 million of mapped read pairs; having low complexity (200 most expressed genes representing more than 50% of all counts) and cells with low endogenous RNA (ERCC spikes-in representing more than 14% of all counts). 279 cells passed quality control and were used in the analysis. The proportion of reads assigned to the capture TFs was calculated as the sum of TF-mapped counts divided by library size. Increasing the expression threshold for gene expression (e.g., to mean counts 4) preserves the post-capture TF enrichment. Post-capture gene-wise enrichment was calculated using CPM normalized counts. For each gene the enrichment was a ratio between the average CPMs in the post- and pre-capture libraries, plotted in log2 scale by adding a pseudocount of 1.
To validate scCapture-seq using a different dataset, we also captured our TF panel in 126 out of 191 intestinal stromal cells from UC patients, which passed the quality control. The scRNA-seq was made using the 10x droplet-based protocol (Kinchen et al. 2018). TF capture was performed and the libraries processed as above. Post-capture libraries were sequenced on 1 lane of a HiSeq 4000 at 75 bp paired end. The total number of the targeted TFs, which were expressed with non-zero mean counts was increased post-capture from 530 to 631 TFs.

NeuroGWAS Capture

To validate scCapture-seq on a different capture panel, we performed capture using the previously designed NeuroGWAS (NG) panel targeting transcripts implicated in neurological diseases and traits (Curion et al. 2020). The NG capture was performed on Smart-seq2 libraries from a separate pool of 376 iPSC-derived cortical neurons compared to the TF capture. Capture hybridisation and sequencing was performed as per the TF cortical neuron capture. 359 out of 376 cells passed the quality control and libraries were processed as above. 71 out of 101 NG genes were expressed pre-capture, and 87 genes were expressed post-capture.

Clustering and differential expression analysis

Most of the initial steps of the analysis, including clustering and differential expression were done in the Seurat version 3 package (Stuart et al. 2019). Heat maps were based on differential expression (FDR <= 0.05) in each cluster against all other cells. Volcano plots were based on differential expression between the N1 and N2 neurons or between G cells from Lab D and Lab E. Pre-capture N1 and N2 neurons were defined as cells of clusters 3 and 4 (N1 group) or cells of cluster 1 (N2 group).

Using unsupervised hierarchical clustering on the expression profiles we identified several clusters of cells within the pre- and post-capture data. To classify these cell clusters accordingly to cell types we further compared the cluster DEGs with the cell type markers known from the reference datasets. As the reference datasets we used three sets of RNA-seq data containing purified cortical neurons, astrocytes, microglia, endothelial cells, and oligodendrocytes (van de Leemput et al. 2014; Darmanis et al. 2015; Song et al. 2017).
The cell type markers in (Darmanis et al. 2015) and (Song et al. 2017) datasets were determined by Seurat as being differentially expressed between the cell types in each dataset (FDR ≤ 0.05). We also extended the set of neuronal markers by adding the markers from (van de Leemput et al. 2014) (reported with FDR ≤ 0.05), after the removal of genes highly expressed in other cell types, based on (Zhang et al. 2014) expression data (using 2-fold cut off). We combined all targeted TF markers and used only unique markers for each cell type. There were in total 131 targeted TF markers present in the reference datasets, with 53 of them being present amongst 155 post-capture DEGs (Supplemental Fig. S2b,e) and 34 of the marker TFs being present amongst 129 pre-capture TF DEGs (Supplemental Fig. S3d). The significance of the identified TF markers in pre- and post-capture datasets was assessed by comparing the sets of TF markers in each dataset to the sets of the referenced TF markers using a hypergeometric test (Supplemental Table S3).

**Gene regulatory network analysis**

The co-expression GRNs were constructed with the bigScale package using default parameters, by retaining only significant correlations with absolute value of Pearson coefficient > 0.8 (Iacono et al. 2019). The pre-capture, whole-transcriptome GRN consisted of 4286 nodes (genes) and 18851 edges/connections between co-expressed genes. We also built the post-capture imputed GRN by extending the pre-capture libraries by the expression of 731 capture TFs re-scaled for the differences in the sizes of pre- and post-capture libraries. The imputed GRN had 6365 nodes and 39101 edges. For the plotting we used only the cluster-specific GRNs, including DEGs that are upregulated in ≥ 50% of the cluster cells (FDR ≤ 0.05). These were defined as genes differentially expressed amongst the pre-capture or imputed GRN genes either against all other clusters or against another single cluster. In case of differential expression in several comparisons between different clusters, DEGs were assigned to the cluster with the highest fold change. We defined cluster-specific TFs as those which were differentially expressed amongst the capture TFs in ≥ 50% of cluster cells. We retained only subnetworks with 3 or more nodes. To assess the significance of GRNs enrichment in capture TFs we used the network enrichment analysis, NEA(Alexeyenko et al. 2012), which estimates the level of connectivity of a gene set (TFs) to the rest of the GRN by generating permuted networks with preserved degree distribution.

**Further GRN analysis (TF PPI and TF targets)**
To explore the enrichment of the cluster subnetworks with TF PPIs, we used the “Transcription factor PPI” database of the Enrichr web server (Kuleshov et al. 2016). For this analysis the cluster subnetworks were defined as those including the cluster-associated genes on the plotted GRNs: DEGs and captured TFs with ≥ 20% of their neighbours being the cluster DEGs. We plotted the top TF PPI terms of targeted TFs, having false discovery rate (FDR) ≤ 0.05 for the imputed GRN (and respective terms in pre-capture GRN). The significance of the difference between the number of genes in the reported pre- and post-capture TF PPIs was assessed with exact Wilcoxon-signed-rank test. The lists of potential TF targets were combined from 2 sources. Firstly, we collected TFs and their targets from neuron and astrocyte networks ((Marbach et al. 2016); http://regulatorycircuits.org/), in which the tissue-specific gene-regulatory networks had been inferred by combining transcription factor sequence motifs with activity data for promoters and enhancers from the FANTOM5 project. To reduce the number of potentially false positive targets, we used the interactions with high evidence scores with edge weights ≥ 0.1 (Misselbeck et al. 2019). Secondly, we predicted TF-target links specific to early stages in neuronal maturation from postmortem human brains (Colantuoni et al. 2011). We downloaded processed BrainCloud gene expression data from GEO [GSE30272] and selected 70 samples from fetal to early post-natal stages (< 1 y.o.). On these we built co-expression networks with TFs as hub nodes by using ARACNE (Lachmann et al. 2016) with default parameters and selected significant TF-target links at P value < 10^{-7}. The significance of the enrichment of post-capture GRN in TF targets was assessed with the Fisher’s exact test using the 2×2 contingency table.

**Pseudotime Analysis**

The pseudotime trajectories were estimated with the Destiny package (Angerer et al. 2016), on log_{2}-transformed counts after adding 1 count and scaling for the size factor for each library. Cells were ordered by pseudotime using the slingshot package, specifying the initial cluster through inspection of TFs differentially expressed in each cluster with relation to known maturation stages (Street et al. 2018).

**Immunocytochemistry**

iPSC-derived cultures were grown on coverslips and fixed in 4 % paraformaldehyde/4 % sucrose (w/v) in phosphate-buffered saline (PBS) solution. Cells were permeabilized in PBS with 0.4 % v/v Triton-X for 2×7 min...
at room temperature (RT), then blocked in PBS with 10 % v/v goat serum for 2 h at RT. GAD65 primary antibody (Chemicon int. Inc; MAB351), diluted in PBS with 5 % v/v goat serum (1:250), was applied for 2 h at RT. Goat anti-mouse Alexa Fluor 488 secondary antibody (Life Technologies; A21131), diluted in PBS with 5 % v/v goat serum (1:1000), was applied for 1 h at RT. Coverslips were washed 3 times for 5 min with PBS after each antibody application. Coverslips were then incubated with DAPI (Thermo Fisher Scientific; 1:5000 in PBS) for 5 min before being washed in PBS and briefly in dH2O, then mounted using Prolong Diamond Anti-Fade Mounting Solution (Thermo Fisher Scientific). Images were acquired on an Olympus BX40 Epifluorescence microscope using HCImage imaging software and processed using ImageJ.

**Patch clamp electrophysiology**

To record spontaneous synaptic currents, iPSC-derived cultures grown on coverslips were bathed in a Tyrode’s solution (140 mM NaCl, 5 mM KCl, 2 mM CaCl, 2 mM MgCl, 10 mM HEPES, 10 mM glucose; pH 7.36, osmolarity 290 mOsm and maintained at 30 °C), visualised under an upright microscope (Olympus BX51WI), and targeted for whole cell patch clamp recording with glass pipettes (tip resistance 5-10 MΩ) that had been pulled from standard wall borosilicate capillaries (OD 1.2 mm, ID 0.69 mm with filament; Warner Instruments). Pipettes were filled with a caesium gluconate solution (140 mM CsGlu, 6 mM NaCl, 1 mM EGTA, 10 mM HEPES, 4 mM MgATP, 0.4 mM Na3GTP). Signals were recorded at 10 kHz using a CVB-7 headstage and Multiclamp 700B amplifier controlled via Clampex (Molecular Devices), and subsequently analysed using Clampfit (Molecular Devices). Each cell was voltage clamped at the reversal potential for glutamatergic currents (0 mV) and GABA-A receptor mediated spontaneous currents were recorded over a period of 4 min.

**Data access**

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; [https://www.ncbi.nlm.nih.gov/geo/](https://www.ncbi.nlm.nih.gov/geo/)) under accession numbers GSE157835, GSE168590 and GSE168626 combined into the reference Series GSE168634.

**Disclosure declaration**
Acknowledgments

This work was supported by a grant SFB646 from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking (IM2PACT grant no. 807015 and AIMS-2-TRIALS grant agreement no 777394); from National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); from BBSRC project BB/S007938/1; from ERC Grant Agreement 617670 and the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115439); from an Australian National Health and Medical Research Council (NHMRC) Early Career Fellowship (APP1072662 to M.B.C); from an EMBO Long Term Fellowship (ALTF 864-2013 to M.B.C); from Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS. AEH was supported by an NIHR Clinical Lectureship and Wellcome Trust Research Training Fellowship (100643/Z/12/Z). Core funding to the Wellcome Centre for Human Genetics was provided by the Wellcome Trust (awards 090532/Z/09/Z and 203141/Z/16/Z).

Author contributions: Z.C. and R.B. conceived and led this work. Z.C, A.H, M.B.C, R.B and F.C designed the capture experiments. M.B.C designed capture pools with assistance from A.H. M.B.C performed scCapture-seq. A.P performed the analysis with assistance from A.H and F.C. E.S.W, S.B and S.E.N performed immunostaining, histochemistry and patch clamp electrophysiology. C.J.A supervised the molecular biology and patch-clamp experiments. V.V. performed the maturation analysis on the bulk data. C.W. supported the maturation and network analysis. A.P wrote the manuscript with Z.C, A.H, M.B.C and R.B. All authors commented on and approved the final manuscript.

References


### Figure legends

**Figure 1.** scCapture-seq improves mapping quality and identification of TFs. **a,** Schematic of the experimental design. **b,c.** Post-capture increase in the proportion of reads assigned to all transcripts (a) and to the captured TFs (b). **d,** Increase in the number of TFs detected per cell post-capture.

**Figure 2.** UMAPs of pre- and post-capture cells show distinct cellular clusters. UMAPs of pre - (a) and post- (b) captured clusters were constructed using Seurat package. **c,** Post-capture cells are coloured by pre-capture clusters.

**Figure 3.** TF scCapture-seq observations are corroborated by immunostaining and cell physiology. **a,b,** A subpopulation of iPSC-derived cortical neurons were immunopositive for the inhibitory interneuron marker, GAD65, and represented a small subset of DAPI-positive nuclei. **c,** Population data showing proportion of GAD65-positive cells (n = 12 FOVs from 3 differentiations). **d,** iPSC-derived cortical neurons were targeted for whole cell patch clamp recordings. **e,** A subset of neurons exhibited spontaneous outward synaptic currents, consistent with the pre-synaptic release of GABA from interneurons (top), which could be blocked with the GABA-A receptor antagonist, picrotoxin (100 μM; bottom). **f,** Population data showing proportion of neurons receiving GABAergic synaptic input (n = 14).

**Figure 4.** TF scCapture-seq improves gene regulatory networks. The networks of co-expressed pre-capture genes with (b) or without (a) imputed post-capture TFs. Only genes, which are differentially expressed between 3 cell groups are shown. The TF nodes are labelled, with node sizes reflecting their degree on the whole network. TF nodes, which are differentially expressed post-capture are shown as embedded pie charts, with colours corresponding to the numbers of DEGs neighbours for each node, from each of 3 cell types. TFs, that are not connected to cluster DEGs are grey. TFs which are also pre-capture DEGs are shown by squares (TFs, which were detected as DEGs in the pre-capture libraries are indicated by brighter colours). **c,**
Distribution of network degree between nodes in pre-capture and post-capture imputed networks. d, Total numbers of genes (DEGs+TFs), TFs and TF targets present on the pre-capture and post-capture-imputed networks. e, Enrichment of the cluster subnetworks in TF PPIs (TF networks), for pre-capture (PRE) and imputed (POST) networks. The circle sizes show the number of genes in each TF PPI term.

Figure 5. TF scCapture-seq improves analysis of cell differentiation trajectories. Diffusion maps of pre-capture (a,c,e) and post-capture (b,d,f) cells built using the Destiny package (Angerer et al. 2016) with pseudotime direction determined using Slingshot (Street et al. 2018). The cells are coloured by clusters (a,b) or pseudotime (c,d). e, f, Log2 expression of representative genes for each cell group on pseudotime space for pre- (e) and post-capture (f).

Figure 6. TF scCapture-seq reveals differentially expressed genes underlying inter-lab variability. a, Volcano plot of all pre-capture DEGs between N1 and N2 neurons (based on pre-capture gene expression in cells of cluster 1 vs cells of clusters 3 and 4). b, Volcano plot of all post-capture DEGs between N1 and N2 neurons (based on post-capture gene expression in cells of clusters N1 vs N2). Significant DEGs are labelled (FDR ≤ 0.05, natural log fold changes ≥ 0.5). c,d, Heatmaps of post-capture expression of the key TF DEGs with different temporal kinetics between the labs. Each heatmap includes only cells from the respective lab: lab D (c) or lab E (d). The cells are ordered by pseudotime (based on all cells, as in Fig. 6a,b), including cells of the earlier G cluster. Expression is log2-normalized and centred.