Intercellular extrachromosomal DNA copy number heterogeneity drives cancer cell state diversity

Maja C Stöber[1,2,3], Rocío Chamorro González[5], Lotte Brückner[5], Thomas Conrad[1], Nadine Wittstruck[1,5], Annabell Szymansky[5], Angelika Eggert[5], Johannes H Schulte[5], Richard P Koche[9], Anton G Henssen[1,5,6,8,10*], Roland F Schwarz[4,6,1*], Kerstin Haase[5,7*]

[1]Berlin Institute for Medical Systems Biology at the Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany.
[2]Institute of Pathology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany.
[3]Humboldt-Universita t zu Berlin, Faculty of Life Science, 10099 Berlin, Germany.
[4]Institute for Computational Cancer Biology, Center for Integrated Oncology (CIO), Cancer Research Center Cologne Essen (CCCE), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany.
[5]Department of Pediatric Oncology/Hematology, Charité-Universitätsmedizin Berlin, Germany.
[6]BIFOLD - Berlin Institute for the Foundations of Learning and Data, Berlin, Germany.
[7]German Cancer Consortium (DKTK), partner site Berlin, and German Cancer Research Center (DKFZ), Heidelberg, Germany.
[8]Berlin Institute of Health, 10178 Berlin, Germany.
[9]Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, New York, USA.
[10]Experimental and Clinical Research Center (ECRC) of the MDC and Charité Berlin, Berlin, Germany.

Correspondence should be addressed to henssenlab@gmail.com, roland.schwarz@uni-koeln.de, kerstin.haase@charite.de

Abstract

Neuroblastoma is characterised by extensive inter- and intra-tumour genetic heterogeneity and varying clinical outcomes. One possible driver for this heterogeneity are extrachromosomal DNAs (ecDNA), which segregate independently to the daughter cells during cell division and can lead to rapid amplification of oncogenes. While ecDNA-mediated oncogene amplification has been shown to be associated with poor prognosis in many cancer entities, the effects of ecDNA copy number heterogeneity on intermediate phenotypes are still poorly understood.

Here, we leverage DNA and RNA sequencing data from the same single cells in cell lines and neuroblastoma patients to investigate these effects. We utilise ecDNA amplicon structures to determine precise ecDNA copy numbers and reveal extensive intercellular ecDNA copy number heterogeneity. We further provide direct evidence for the effects of this heterogeneity on gene expression of cargo genes, including *MYCN* and its downstream targets, and the overall transcriptional state of neuroblastoma cells.

These results highlight the potential for rapid adaptability of cellular states within a tumour cell population mediated by ecDNA copy number, emphasising the need for ecDNA-specific treatment strategies to tackle tumour formation and adaptation.

Keywords: extrachromosomal DNA, tumour heterogeneity, cell state diversity, copy number, neuroblastoma, single-cell RNA sequencing

Introduction

Paediatric neuroblastoma is a genetically heterogeneous tumour demonstrating a spectrum of different clinical outcomes (1,2). It is characterised by relatively few somatic nucleotide variants (SNVs) and known driver events, but considerable chromosomal instability and

somatic copy number alterations (SCNAs) (3,4). One key genetic alteration is frequent amplification of the *MYCN* oncogene, associated with unfavourable outcome and aggressive disease, which often occurs in the form of a high number of copies of one or multiple extrachromosomal DNA (ecDNA) amplicons (5,6).

EcDNAs form as the result of DNA damage, in particular double-strand breaks (7), which may occur on their own or as part of larger catastrophic events including chromothripsis (8). Despite lacking centromeres, ecDNAs can be maintained in circularised form in the nucleus, where they replicate proportionally with the chromosomes during S-phase and subsequently segregate independently and randomly to daughter cells upon cell division. When paired with a distinctive selective advantage, as in the case of *MYCN*, these random segregation patterns can lead to a rapid increase in the number of oncogene copies in the tumour cell population.

While ecDNA in cancer was first described over 50 years ago (9), its high prevalence amongst tumour types and the crucial role it plays in oncogene amplification and tumour evolution was only recently revisited (6,10–19). Importantly, ecDNA not only serves as a substrate for oncogene amplification and copy number heterogeneity, but also drives high oncogene expression by increasing copy number dosage, enhancing chromatin accessibility and by optimising enhancer topology (15,18,19).

These transcriptional effects possibly contribute to providing the tumour with increased plastic potential to evade therapeutic selection pressures and rapidly adapt to changing environmental conditions. Recent studies have leveraged fluorescence in-situ hybridisation (FISH) to visualise this increased genetic plasticity within individual tumours, where the number of ecDNA copies varies substantially between cell populations and clones (20). However, it is unclear if and to what extent these varying copy numbers affect cell states and influence cellular phenotypes within individual patients.

We here use a combination of previously published single-cell transcriptome data of neuroblastoma patients and newly derived single-cell genome-and-transcriptome (G&T) (21) and ecDNA-and-transcriptome (scEC&T) (22) sequencing of neuroblastoma cell lines and patients to address this issue. We demonstrate substantial transcriptional heterogeneity within individual neuroblastoma patients and provide evidence that ecDNA copy number variability is causally linked to these transcriptional changes. We believe that understanding the precise role that ecDNA plays in generating intra-tumour heterogeneity will not only enhance our understanding of cancer evolution as a whole, but will further inform our treatment strategies.

Results

Intra-patient *MYCN* expression heterogeneity is linked to distinct transcriptional states in *MYCN*-amplified primary neuroblastomas

To assess the extent of transcriptional heterogeneity between individual cells in neuroblastoma, we first analysed gene expression data of twelve *MYCN*-amplified primary neuroblastoma samples using 10x single-nuclei RNA sequencing [Figure 1a]. To this end we combined samples collected locally at the Charité university hospital Berlin (N=4) with two published cohorts from the University Hospital of Cologne ((23) , N=4) and St. Jude's Hospital Memphis ((https://scpca.alexslemonade.org/), N=4) [Table S1, Additional File 2].

We annotated cell types by combining principal component analysis (PCA) with canonical marker gene expression (Methods) obtained from (23) and quantified endothelial cells, immune cells, mesenchymal cells and tumour cells for all patients. Samples across the cohort showed an overall high tumour cell content (average 86%, +/- 21), in line with previous findings (23,24). Most samples consisted of a substantial proportion of immune cells (average 5%, +/- 9), and varying degrees of endothelial (average 4%, +/- 3) and

mesenchymal cells (average 4%, +/- 4) [Table S2, Additional File 2]. Transcriptional profiles were visually inspected using UMAP per patient, which confirmed separation of cell types into distinct clusters [Figure S1a, Additional File 1], and all non-tumour cells were excluded for downstream analyses.

To obtain an in-depth characterisation of the transcriptional landscape of *MYCN*-amplified neuroblastomas and to investigate its heterogeneity, we next identified transcriptional programs (modules) for each patient using non-negative matrix factorisation (cNMF (25), Methods). We chose an optimal number of modules per patient based on a trade-off between module stability and reconstruction error (Methods) and identified 106 transcriptional programs across the cohort (mean 9 [6 - 12]). To investigate commonalities between patients, we performed pairwise Pearson correlation analysis of all modules followed by hierarchical clustering [Figure 1c] and identified 3 meta modules which were further split into 10 sub modules (Methods, [Figure S1b, Additional File 1]). Thirty-one modules without significant correlation to at least 50% of other modules were considered uncommon and removed from downstream analyses. Average gene activity scores for each meta and submodule followed by gene set enrichment analysis (GSEA) revealed high activity of genes involved in cell cycle progression and cell division for Meta Module 1 (e.g., KIF18B, ASPM, KIF14), in line with recent findings in other cancer entities (26). In particular, submodules of this cell cycle meta module showed enrichment of replication (S1) and cell division (S8). Meta module 2 was strongly enriched for genes involved in ribosome biogenesis and the third meta module contained genes associated with cell-cell interactions (e.g., CNTN5, TENM2, CTNNA2). The submodules of the ribosome meta module showed enrichment of genes involved in translation (S2), post-transcriptional regulation (S6) and cellular response to stress (S10). The cell-cell interaction meta module was divided into submodules associated with neuronal differentiation (S3), sensory perception (S4), regulation of cell size (S5), axonogenesis (S7) and synaptic signalling (S9) [Figure 1d; Table S3, Additional File 2].

Recent large-scale efforts into transcriptional cell states pan-cancer have identified cell-cycle related modules, but not any associated with ribosome biogenesis or cell-cell interactions (26). We thus hypothesised that the latter modules might be specific to neuroblastoma biology and indicative of *MYCN*-mediated upregulation of ribosome biogenesis (27,28) and downregulation of neurogenesis (29). To assess this, we correlated *MYCN* expression levels with module activity for all patients and observed higher average correlation scores with modules grouped into the ribosomal biogenesis meta module 2, as opposed to cell cycle or cell-cell interaction [Figure 1f]. The direct association between *MYCN* expression and ribosomal biogenesis activity was further confirmed by visual inspection of UMAPs with meta module activity and *MYCN* expression overlaid respectively [Figure 1b, e].

Interestingly, within the cell cycle meta module we found the sub module associated with replication pathways and the G1/S transition pathway (S1) to be positively correlated with *MYCN* expression, while the sub module associated with cell division (S8) was negatively correlated with *MYCN* [Figure S1c, Additional File 1]. Investigating canonically expressed cell cycle marker genes for each cell (30), we found cells with high activity of the cell cycle meta module 1 to be predominantly in G2M and S phase [Figure S1d, Additional File 1] in line with the role of *MYCN* in cellular proliferation (31). *MYCN* expression was further significantly associated with cell cycle phase in 7 out of 12 samples. In the remaining 5 samples the number of detected features and reads was significantly lower on average to those samples with association, suggesting technical rather than biological effects as a potential cause for this lack of association [Figure S1e, Additional File 1].

Taken together, we observe substantial transcriptional heterogeneity and distinct transcriptional states of cells within individual patients directly associated with and potentially causally linked to heterogeneous *MYCN* expression levels.

Intra-patient *MYCN* expression heterogeneity is biologically functional and consistent between patients

To confirm that the observed *MYCN* expression heterogeneity is biologically functional, we next grouped cancer cells into discrete groups with high (MYCN-high), intermediate (MYCN-med) and low (MYCN-low) *MYCN* expression levels based on the top and bottom 30% expression quantiles per patient. Differential gene expression analysis between MYCN-high and MYCN-low cells [Figure 2a, b] showed an average *MYCN* log2 fold change of 1.613 (1.138 - 2.132). Next, we ranked all genes according to their fold change and tested whether known MYCN target genes (32) were enriched in this ordered list using GSEA, which was the case in 11 out of 12 patients. This consistent differential activity of MYCN target genes indicates that observed *MYCN* expression differences may affect expression of downstream MYCN targets [Figure 2c inset] and strongly suggests that the observed *MYCN* variability is biologically functional. Additionally, the normalised enrichment score (NES) of MYCN target genes was significantly correlated with the difference in gene expression between MYCN-high and MYCN-low cells (Pearson correlation, $p = 0.0043$), indicating a direct connection between the degree of *MYCN* expression variability and the observed differences in downstream MYCN target activity [Figure 2c].

To identify additional differences between MYCN-high and MYCN-low cells, we performed GSEA on gene ontology (GO) biological processes and identified a set of 38 pathways that were recurrently enriched in every single patient and positively associated with *MYCN* expression. These 38 pathways include ribosome biogenesis, RNA catabolic processes, protein targeting, peptide biosynthetic and viral processes [Figure 2d,e; Table S4, Additional File 2; Table S5, Additional File 3]. To investigate how these 38 recurrent pathways relate to the transcriptional cell states identified above, we performed a meta pathway enrichment analysis (Methods) [Figure 2f]. Briefly, all GO terms were ranked according to their averaged NES in each meta module and this ranked list was tested for enrichment for each of the 38

original pathways. The ribosome meta module 2 and its sub modules showed a strong positive association with all 38 pathways similar to MYCN-high cells [Figure 2e,f], whereas the cell-cell interaction module 3 and cell-cycle module 1 showed a strong negative association.

We further investigated whether MYCN-high and -low cells expressed signatures of mesenchymal and adrenergic differentiation states (33,34). Overall, all samples in the 3 cohorts primarily expressed the adrenergic signature. We found cells with high *MYCN* expression to show lower expression of adrenergic features but also lower expression of mesenchymal features than cells with low *MYCN* expression [Figure S1f, Additional File 1]. In conclusion, we do not find any evidence for adrenergic to mesenchymal state transition driven by *MYCN* expression within individual patients.

The enriched pathways demonstrate biological functionality of the observed *MYCN* expression heterogeneity on the single-cell level and its consistency across a cohort of *MYCN*-amplified primary neuroblastomas. We describe its association with changes in ribosomal activity and cell-cell interaction and to a weaker degree with cell cycle transition, potentially indirectly through co-activation of other cancer pathways (26).

MYCN-amplified neuroblastoma cells show high inter-cellular ecDNA copy number heterogeneity

One potential cause of the observed *MYCN* gene expression heterogeneity are variations in copy number, possibly driven by ecDNA (6,20,35). Unequal mitotic segregation of ecDNA can lead to high inter-cellular copy number heterogeneity, but the extent of this heterogeneity within patients remains unclear.

We therefore performed genome and transcriptome (G&T) sequencing of one primary neuroblastoma (N=96 cells) and two neuroblastoma cell lines CHP212 (N=96 cells) and

TR14 (N=192 cells)), which are known to harbour ecDNA-linked *MYCN* amplifications. Additionally, we performed single-cell scEC&T-seq (22) on the same patient sample (N=96 cells) and cell lines (CHP212 (N=170 cells) and TR14 (N=42 cells)), to confirm the presence of ecDNA and to determine ecDNA-amplified regions in the genome [Figure 3a]. We hypothesised that ecDNA-amplified regions show patient- or cell-line specific amplification and expression patterns and that ecDNA copy number variation is the main contributor to *MYCN* expression heterogeneity.

Previous investigations into the ecDNA amplicon structures in CHP212 and TR14 (18) and patient 8 (22) revealed that the CHP212 cell line contains one single circular amplicon of size 1.7Mb containing genes *LPIN1*, *TRIB2*, *DDX1* and *MYCN*. In contrast, TR14 contains three different circular amplicons 710, 475 and 1,000 kbp in size respectively, harbouring together over 29 genes of which we only consider those containing the known neuroblastoma oncogenes *MYCN*, *CDK4* and *MDM2* [Figure 3b]. The ecDNA amplicon structure in patient 8 is 500 kbp long and only contains *MYCN* [Figure S2b, Additional File 1]. The varying amplicon structures were also clearly visible from overall read coverage in DNA sequencing [Figure 3a, track “DNA”].

We next determined global copy number profiles for each single cell from G&T sequencing using Ginkgo (36), and then refined those copy number profiles by leveraging previously reconstructed precise ecDNA breakpoints (Methods) (18,22). Genomic regions that contained these oncogenes on ecDNA showed extensive copy number heterogeneity across cells within a single cell line and patient, whereas regions that were not on ecDNA did not [Figure 4a]. The *MYCN* locus in CHP212 showed on average a copy number of 50 (range 3 - 353), for the *MYCN* locus in TR14 the average copy number was estimated at 105 (range 6 - 852). The copy number of the *MYCN* amplicon locus in patient 8 was on average 191 (range 5 - 916) [Figure S2c, Additional File 1]. For a full overview of the copy number states and expression levels for all genes considered, please refer to [Table S6,S7, Additional File 2].

To confirm the presence of ecDNA and to validate the copy number estimates, we performed FISH by staining centromeres and genomic regions containing *MYCN*, *CDK4* and *MDM2* [Figure S2a, Additional File 1] (Methods). FISH in metaphase spreads confirmed the presence of ecDNA in all cases, and copy number estimates from FISH in interphase cells showed distributions similar to those obtained from sequencing [Figure 4a].

In TR14, the *MYCN* locus is present in two different ecDNA amplicons, which raises the question whether one amplicon contributes more copies to the total *MYCN* copy number than the other. We thus estimated the fraction of copies contributed per amplicon by leveraging additional non-overlapping loci on the amplicon and distributing the *MYCN* copy number accordingly (Methods). The amplicon containing only *MYCN* was substantially larger and contributed more copies than the amplicon containing both *MYCN* and *CDK4*. However, when comparing the number of *MYCN*-only amplicons to the largest amplicon in this cell line carrying *MDM2*, we found two thirds of the cells to harbour more ecDNA containing *MYCN* than *MDM2*, suggesting that amplicon size alone does not equate to ecDNA copy number.

We next investigated whether the observed copy number heterogeneity is reflected on the transcriptional level by leveraging RNA expression readouts from the same single-cells. Due to the cell lines' unique amplicon structures, different genes can be amplified on ecDNA. As expected, we observed increased gene expression levels in genes present on the ecDNA including *MYCN*, *DDX1*, *TRIB2* and *LPIN1* in CHP212 and *MYCN*, *CDK4* and *MDM2* in TR14. In contrast, genes not present on their respective amplicons showed only low overall levels of expression. For an example, consider *MDM2*, *CDK4* in the CHP212 cell line in [Figure 3b, track “RNA”]. Thus, significant inter-cellular ecDNA copy number heterogeneity occurs in neuroblastoma cells.

Inter-cellular ecDNA copy number heterogeneity drives transcriptional states in neuroblastoma cells

Pan-cancer analyses have recently confirmed that copy number is a main driver of aberrant gene expression across human cancers (37). Since ecDNA presence can lead to exceptionally high copy number levels (10,16), we leveraged the combined genome and transcriptome data from G&T and scEC&T sequencing to investigate whether the observed inter-cellular ecDNA copy number heterogeneity is reflected in the transcriptome.

Overall we found strong positive correlations between ecDNA copy number and mRNA expression across all genes contained on the ecDNA amplicons [Figure 4b; Figure S3a,b, Additional File 1]. Gene expression measured in TPM increased linearly with ecDNA copy number, explaining gene expression variability on average by 42% (median: 47, range 2.5 - 76) in CHP212 and by 25% (median: 22, range 0.2 - 66) in TR14. Interestingly, effect sizes differed significantly between genes on amplicons, but also between cell lines for the same gene [Table S8, Additional File 2]. This sample- and gene-specific dosage effect of ecDNA copy number suggests other regulatory mechanisms to be involved in gene expression heterogeneity such as chromatin conformation, enhancer hijacking and ecDNA hub formation (15,18,38).

To ascertain that this dosage effect is truly driven by ecDNA, we exploited the linear relationship of ecDNA copy number and gene expression to build a gene-wise sample-specific linear model from the G&T-seq data [Table S8, Additional File 2]. We then predicted copy number from gene expression in cells sequenced with scEC&T-seq, for which absolute copy number measurements are not readily available, and observed a strong correlation of the predicted copy number with the number of ecDNA reads in a region covering that gene [Figure S3c, Additional File 1]. This strongly suggests that the observed gene dosage is

298 indeed caused by ecDNA copy number which ultimately drives the increase in gene
299 expression in the analysed samples.
300
301 To confirm that the observed expression differences found in the high-throughput 10X
302 patient cohort might be plausibly linked to ecDNA, we repeated the above analysis on the
303 cell lines by considering the transcriptomes from both G&T and scEC&T data. We once
304 again stratified cells into *MYCN* high and low expressing cells (30% quantile) and tested for
305 differential gene expression between both groups. In CHP212, *MYCN*, *LPIN1*, *DDX1* and
306 *TRIB2* were consistently identified, in line with CHP212 containing a single ecDNA
307 containing those genes [Figure 3b]. Interestingly, in TR14 only *MYCN* itself and one
308 additional, non-amplified gene (*E2F1*), were identified, likely due to the relative
309 overabundance of the ecDNA amplicon containing only *MYCN* compared to the two
310 alternative circle structures present in the cell line [Figure 3b; Table S9,S10, Additional File
311 2].
312
313 As previously, we ranked the genes based on their expression difference and then
314 performed GSEA for GO biological processes. Consistent with our observations in the high-
315 throughput patient data we observed elevated MYCN target gene expression in MYCN-high
316 cells [Figure 4c; Figure S3d, Additional File 1] which supports our previous result of
317 functional *MYCN* heterogeneity in patients. Among the 38 identified pathways enriched in
318 MYCN-high cells in patients, 5 and 17 pathways were also significantly positively enriched in
319 CHP212 and TR14 respectively [Figure 4d; Table S11, Additional File 2] including the
320 ribosome biogenesis pathway.
321
322 In summary, we observed similar transcriptional patterns in the patient cohort as compared
323 to the cell lines for which we have established a clear link to ecDNA presence.

Conclusions/Discussion

The role of ecDNA in the development of malignant phenotypes has been explored in recent studies which uncovered ecDNA-associated poorer survival and treatment resistance (13,15,39). We here use genomic and transcriptomic information from the same single cells to link *MYCN*-amplifications on ecDNA to downstream transcriptional effects and cell states. We show that ecDNA-mediated intercellular heterogeneity of *MYCN* expression within patients creates various co-existing cellular subpopulations with differing transcriptional states. We demonstrate changes in key pathways including ribosome biogenesis and cell-cell interaction, a potential substrate for rapid adaptation to environmental changes including treatment [Figure 4e].

Our characterization of transcriptional programs in *MYCN*-amplified neuroblastoma revealed 3 recurring meta pathways across 12 patients, which are associated with cell cycle, ribosome biogenesis and cell-cell interaction. While the cell cycle module was also found in recent investigations pan-cancer (26), the lack of ribosome biogenesis in this study might indicate that it is neuroblastoma specific. We demonstrated functional intra-patient *MYCN* expression heterogeneity across the cohort leading to upregulation of ribosome biogenesis and deregulation of neurogenesis genes within individual patients, effects that were previously only described in bulk between patients or cell lines with varying MYCN expression (32), (27,28)(29). Future work will need to elaborate to what extent this heterogeneity can be found in non-*MYCN*-amplified neuroblastoma or in tumours harbouring other forms of amplification, such as homogeneously staining regions.

Surprisingly, not all individuals showed significant associations between *MYCN* expression levels and cell cycle phase, although it has been shown that *MYCN* amplification is associated with the cells ability to escape G1 phase (40,41). This might be explained by the

varying degrees of MYCN expression heterogeneity in our cohort, where in some patients phenotypic effects might be weaker and remain undetected.

To investigate the role of ecDNA in the observed transcriptional heterogeneity, we inferred ecDNA amplicon-specific copy number from single-cell DNAseq data. While FISH followed by semi-automated counting of fluorescent markers remains the gold standard for ecDNA detection, the technique is limited by the 2D nature of the images and can underestimate ecDNA copy number due to stacking of cells. We show that single-cell DNA sequencing is sufficiently accurate to recapitulate amplicon boundaries and that, depending on the amplicon architecture, accurate ecDNA copy numbers can be derived from read counts by combining general copy number calling methods (36) with a custom inference algorithm. However, naturally, such efforts are dependent on the quality of the output of the copy number calling algorithm.

Another possible source of noise is the integration of different sequencing technologies in our cohort, in particular single-nuclei sequencing in patients with single-cell sequencing in cell lines. While both approaches were found to be comparable with similar sensitivity (42–44), single-nuclei sequencing can be prone to a higher gene dropout rate, which might affect the size of the discovered gene sets. However, we also found a generally good agreement between approaches and sequencing technologies in this study.

In conclusion, we were able to associate cell state heterogeneity in *MYCN*-amplified neuroblastomas with ecDNA copy number heterogeneity, implying that the rapid evolutionary dynamics associated with ecDNA (20) have the potential to also enable rapid phenotypic adaptation potentially within a single cell division cycle. One important question is thereby whether the relationship between the number of ecDNA copies and the transcriptional effects and its function are linear, and if and where there is an upper limit to the fitness advantage accrued through ecDNA accumulation. Arguably, the replicative and metabolic

burden inferred by excessive ecDNA copy numbers will likely lead to diminishing returns in terms of clonal fitness beyond a certain level. However, in our study we observed largely linear relationships between ecDNA copy number and transcriptomic output within the observed copy number range. Additionally, we could show that increases in MYCN target gene expression activity are linearly correlated with *MYCN* expression fold change increase, suggesting that additional ecDNA copies continue to linearly affect oncogene function within the range of copy numbers observed in real tumours and cell lines. Additional experiments will need to investigate whether this linear increase directly translates to an increase in biological function, for example by increasing cell growth and proliferation through upregulation of ribosome biogenesis.

Treatment strategies targeting downstream effects of ecDNA-mediated pathways have been shown to lead to therapy resistance or recurrence after the treatment ended (14), likely because of rapid re-emergence of cells with high ecDNA copy number. Investigating the ecDNA evolution and associated cellular states during and after treatment could potentially uncover new treatment strategies.

Figures

Figure 1: Cellular state heterogeneity in *MYCN*-amplified neuroblastomas

A: Schematic of available data and sample preparation of 12 *MYCN*-amplified neuroblastoma patients with single-nuclei RNA-seq (10X genomics) and workflow; **B**: UMAP of 4,641 single-nuclei of patient 1, tumour cells coloured by *MYCN* expression level; **C**: Heatmap of Pearson correlation coefficients of TPM Z-scores of patient derived modules from non-negative matrix factorization, rows coloured by meta and sub modules, columns coloured by patient of origin; **D**: Heatmap of average TPM Z-score across sub modules of meta module defining genes; **E**: UMAPs of patient 1 coloured according to corresponding meta module activity; **F**: Correlation of *MYCN* expression and patient derived modules, grouped by meta modules

Figure 2: Functional *MYCN* expression heterogeneity in *MYCN*-amplified neuroblastoma

A: *MYCN* expression distribution of patient 1 coloured by *MYCN* groups; **B**: Workflow schematic of DEG and GSEA **C**: Correlation of *MYCN* expression difference between MYCN-high and MYCN-low cells and normalised enrichment score of MYCN target genes per patient, colours represent the negative logarithmic adjusted p-value of the enrichment, Pearson correlation coefficient and p-value are given as inset, inset: GSEA of MYCN target genes, genes decreasingly ordered by logarithmic fold change derived from differential gene expression analysis between MYCN-high and MYCN-low cells in patient 1; **D**: Barplot of the number of pathways recurrently positive (red) or negative (green) enriched in the respective number of patients; **E**: Network of 38 recurring pathways enriched in cells with high *MYCN* expression across all 12 patients, edges refer to similarity of underlying gene set over a threshold, line type corresponds to kappa score referring to the number of overlapping genes in the gene set, colours highlight manual set summary term, some labels omitted for better readability, full list can be accessed in [Table S5, Additional File 3]; **F**: Meta pathway enrichment of 38 GO-Terms in ranked average NES list of meta modules and sub modules

Figure 3: EcDNA amplicon structures in neuroblastoma cell lines

A: Two MYCN-amplified neuroblastoma cell lines, TR14 and CHP212, with G&T-seq and scEC&T-seq; B: Mean scDNA-seq read coverage of selected regions on chromosomes 2 and 12; amplicon boundaries from scEC-seq (red); schematic of ecDNA amplicons in CHP212 and TR14; distribution of gene expression (TPM) for CHP212 and TR14 cells of *LPIN1*, *TRIB2*, *DDX1*, *MYCN*, *CDK4* and *MDM2*.

Figure 4: EcDNA copy number heterogeneity in neuroblastoma cell lines

A: Distribution of ecDNA amplicon copy number adapted from Ginkgo copy number profiles (500kb bin size) from single-cell whole genome sequencing (dark) and distribution of foci

counts from FISH (light) for *MYCN* in CHP212 and *MYCN*, *CDK4* and *MDM2* in TR14); **B**: Correlation of gene expression and copy number of *MYCN* in CHP212, Pearson correlation coefficient and p-value are given as inset; **C**: GSEA of MYCN target genes, genes decreasingly ordered by logarithmic fold change derived from differential gene expression analysis between MYCN-high and MYCN-low cells in CHP212; **D**: NES of significantly enriched recurring pathways in patients per cell line; **E**: Schematic of summarising results

Methods

Cell culture

Human cancer cell line CHP212 was obtained from the American Type Culture Collection (ATCC; Manassas, VA, USA) and cancer cell line TR14 was kindly provided by J. J. Molenaar (Princess Máxima Center for Pediatric Oncology, Utrecht, Netherlands). Cells were tested for *Mycoplasma sp.* contamination with a Lonza MycoAlert system (Lonza) and absence of contamination was confirmed biweekly. STR genotyping (Genetica DNA Laboratories and IDEXX BioResearch) was performed to confirm the identity of both cell lines. For cell culture, we used RPMI-1640 medium (Thermo Fisher Scientific) supplemented with 1% penicillin, streptomycin, and 10% FCS. Cell viability was assessed with 0.02% trypan blue (Thermo Fisher Scientific) mixed in a 1:1 ratio, and counted with a BioRad TC20 cell counter.

Patient samples and clinical data access

This study comprised the analyses of tumour and blood samples of patients diagnosed with neuroblastoma between 1991 and 2016. Specimens and clinical data were archived and made available by Charité-Universitätsmedizin Berlin or the National Neuroblastoma Biobank and Neuroblastoma Trial Registry (University Children's Hospital Cologne) of the GPOH. The *MYCN* gene copy number was determined as a routine diagnostic method using

457 FISH. DNA and total RNA were isolated from tumour samples with at least 60% tumour cell
458 content as evaluated by a pathologist.

459 Preparation of Metaphase spreads and FISH

460 Cells were cultured in a 15 cm dish and grown to 80% confluency. Metaphase arrest was
461 performed by adding KaryoMAX™ Colcemid™ (10 µL/mL, Gibco) and incubating for 1-2
462 hours. Afterwards, we washed the cells with PBS, trypsinized and centrifuged at 200 g for 10
463 min. We slowly added a total of 10 mL of 0.075 M KCl preheated at 37 °C, one mL at a time
464 and vortexing at maximum speed in between. Cells were then incubated for 20 min at 37 °C.
465 For cell fixation, we added 5 mL of ice-cold 3:1 MeOH/acetic acid (kept at -20 °C), one mL at
466 a time and resuspending the cells by flicking the tube. We centrifuged the sample at 200 g
467 for 5 min. We repeated this step of addition of the fixate followed by centrifugation four times.
468 Finally, two drops of cells within 200 µL of MeOH/acetic acid were added onto prewarmed
469 slides from a height of 15cm and slides were incubated overnight. We fixed the slides in
470 MeOH/acetic acid for 10 min at -20 °C and washed them in PBS for 5 min at room
471 temperature (RT). We incubated the slides in pepsin solution (10 µL pepsin (1 g / 50 mL) in
472 0.001N HCl) at 37 °C for 10 min and washed in 0.5x SSC buffer for 5 min. Dehydration of
473 the slides was performed by 3-minutes washes in 70%, 90% and 100% cold ethanol (stored
474 at -20 °C). After drying, we stained the slides with 10 µL of Vysis LSI N-MYC
475 SpectrumGreen/CEP 2 SpectrumOrange Probes (Abbott), ZytoLight ® Spec CDK4/CEN12
476 Dual Color Probe (ZytoVision) or ZytoLight ® SPEC MDM2/CEN 12 Dual Color Probe
477 (Zytovision), covered with a coverslip and sealed with rubber cement. The probes were
478 denatured by incubation at 72 °C for 5 min in a Thermobrite (Abbott) followed by overnight
479 incubation at 37°C. We washed the slides for 5 min in 2× SSC/0.1% IGEPAL at RT followed
480 by a 3-minutes wash at 60 °C in 0.4× SSC/0.3% IGEPAL (Sigma-Aldrich Inc.), and an
481 additional 3-minutes wash in 2× SSC/0.1% IGEPAL at RT. After drying, we used 12 µL
482 Hoechst 33342 (10 µM, Thermo Fisher Scientific) to stain the slides for 10 min, followed by a

wash with PBS for 5 min. Once the slides were completely dried, a coverslip was mounted and sealed with nail polish. Images were taken using a Leica SP5 Confocal microscope.

Interphase FISH

TR14 cells for interphase FISH were grown in 8-chamber slides (Thermo Scientific™ Nunc™ Lab-Tek™) to 80 % confluence. Wells were fixed in MeOH/acetic acid for 20 min at -20 °C followed by a wash of the slide in PBS for 5 min at room temperature (RT). The wells were removed and digestion of the slides was done in Pepsin solution (0.001 N HCl) with the addition of 10 µl pepsin (1 gr/50 mL) at 37 degrees for 10 min. Slides were washed in 0.5x SSC for 5 min and dehydrated by washing in 70 %, 90 % and 100 % cold ethanol stored at -20 °C (3min each). Dried slides were stained with either a 5 µl of Vysis LSI N-MYC SpectrumGreen/CEP 2 SpectrumOrange Probes (Abbott), ZytoLight ® Spec CDK4/CEN12 Dual Color Probe (ZytoVision) or ZytoLight ® SPEC MDM2/CEN 12 Dual Color Probe (Zytovision), covered with a coverslip and sealed with rubber cement. Denaturing occurred in a Thermobrite (Abbott) for 5min at 72 °C followed by 37 °C overnight. The slides were washed for 5 min at RT within 2× SSC/0.1 % IGEPAL, followed by 3 min at 60 in 0.4× SSC/0.3 % IGEPAL (Sigma-Aldrich Inc.) and further 3 min in 2× SSC/0.1 % IGEPAL at RT. Dried slides were stained with 12 µl Hoechst 33342 (10 µM, Thermo Fisher) for 10 min and washed with PBS for 5 min. After drying, a coverslip was mounted on the slide and sealed with nail polish. Images were taken using a Leica SP5 Confocal microscope and analysed using the FIJI find maxima function.

Nuclei isolation

For nuclei isolation, tissue samples were added in 1mL of ice-cold EZ PREP buffer (Sigma) and homogenised using a pre-cooled glass dounce tissue homogenizer (Wheaton). We used ten strokes with the loose pestle followed by 5 strokes with the tight pestle for adequate tissue homogenization. The sample was kept on ice at all times during homogenization to avoid heat generation caused by friction. After homogenization, we filtered the sample using

a BD Falcon tube with a 35µm cell strainer cap (Becton Dickinson). To estimate the number of intact nuclei, we stained with 0.02% trypan Blue (Thermo Fisher Scientific) in a 1:1 ratio.

Fluorescence-activated cell sorting (FACS)

One to ten million neuroblastoma cells were stained with Propidium Iodide (PI, Thermo Fisher Scientific) in 1× PBS, and viable cells selected based on the forward and side scattering properties as well as PI staining. Nuclei suspensions were stained with DAPI (Thermo Fisher Scientific, final concentration 2 µM). For plate-based single-cell sequencing, viable cells were sorted using a FACSAria Fusion flow cytometer (Biosciences) into 2.5 µL of RLT Plus buffer (Qiagen) in low binding 96-well plates (4titude) sealed with foil (4titude) and stored at −80 °C until processing. For droplet-based single-nuclei RNA-seq, DAPI-positive nuclei were sorted using a FACSAria Fusion flow cytometer (Biosciences) into 20 µL of 4% (w/vol) Bovine Serum Albumin (BSA; Sigma) in 1× PBS, supplemented with 2 µL of RNAse-In (40 U/µL; Life Technologies) and 2 µL of SUPERase-In (20 U/µL; Life Technologies).

Droplet-Based snRNA-seq

Droplet-based single-nuclei RNA-seq was performed using the 10x Genomics Chromium Single Cell 3' Kit (v.3.1) following the manufacturer's protocol (45). For single nuclei gel bead-in-emulsions (GEMs) generation, we aimed for a target output of 10,000 nuclei for each sample. The amplified cDNA and final libraries were evaluated on a 4200 Tapestation (Agilent Technologies) using the HS-D5000 and HS-D1000 High Sensitivity DNA kits (Agilent Technologies), respectively. snRNA-seq libraries were sequenced on an Illumina NovaSeq 6000.

G&T-seq and scEC&T-seq

For plate-based single-cell sequencing, physical separation of genomic DNA and mRNA, and cDNA generation was performed as described in the G&T-seq protocol by Macaulay et al. (21). For G&T-seq, single-cell's gDNA was purified using 0.8× AMPure XP beads (Beckman Coulter) and genomic DNA amplification was carried out using the PicoPLEX Single Cell WGA kit v3 (Takara) and following the manufacturer's instructions. For scEC&T-seq, the purified gDNA was subjected to exonuclease digestion and rolling-circle amplification as previously described (22). All single-cell libraries were prepared using the NEBNext Ultra II FS kit (New England Biolabs) following the manufacturer's instructions but using one-fourth volumes. Unique dual index primer pairs (New England Biolabs) were used to barcode single-cell libraries. Pooled libraries were sequenced on a HiSeq 4000 instrument (Illumina) or a NovaSeq 6000 instrument with 2× 150bp paired-end reads for genomic DNA and circular DNA libraries and 2× 75 bp paired-end reads for cDNA libraries.

Single-nuclei RNA-seq processing

10x Genomics Cell Ranger v.5.0.1 was used to quantify the sequencing reads against the human genome build 38 (hg38), distinguish cells from the background and generate count tables of unique molecular identifiers (UMIs) for each gene per cell. Intronic counts were included.

Single-cell DNA-seq and RNA-seq processing

Reads sequenced from the genomic DNA libraries were trimmed using Trim Galore (version 0.6.4) (46) and mapped to the human genome build 19 (hg19). Alignment was performed with bwa mem (version 0.7.17) (47).

Hisat2 (version 2.2.1) (48) was used to align the RNAseq data obtained from Smart-Seq2 (49) against a transcriptome reference created from hg19 and ENCODE annotation v19 (50). Afterwards genes and isoforms were quantified using rsem (version 1.3.1) (51) with a single cell prior.

Single-cell/nuclei RNA-seq analysis

The following data analyses on count matrices from single-cell/nuclei RNA-seq were performed using the R package Seurat v4.1.0 (52).

Quality control

For data generated using the 10X single-nuclei technology, nuclei with fewer than 1000 counts, 300 distinct features or more than 2.5% of reads mapping to mitochondrial genes were omitted. Sequencing libraries generated with Smart-seq2 (21,22,49) from patients were filtered by omitting nuclei with fewer than 2500 distinct features or more than 1.5% of reads mapping to mitochondrial genes. Sequencing libraries generated with Smart-seq2 from cell lines were filtered by excluding cells with fewer than 5000 distinct features or more than 15% of reads mapping to mitochondrial genes.

The R package DoubletFinder v2.0.3 (53) was used to detect and filter doublets in 10X single-nuclei samples. Default settings were used and 7.5% doublet rate was estimated based on the number of recovered cells.

Genes present in fewer than five cells were excluded and analysis was restricted to protein-coding genes.

Normalisation of RNA

10X single-nuclei data was normalised using the Seurat function ‘NormalizeData’ accounting for sequencing depth, scaling counts to 10,000 and adding a pseudocount of one before natural-log transformation. Genes were scaled using the Seurat function ‘ScaleData’ with mean of 0 and standard deviation of 1 (default).

Smart-seq2 data was normalised using transcripts per million (TPM), accounting for gene length and total read count in each cell. For downstream analyses a pseudocount of one was added and then natural-log transformed.

Feature selection and dimension reduction

The Seurat function 'FindVariableGenes' was used to find the top 2000 most variable genes in each patient and cell line individually. Principal component analysis was performed on most variable genes and the first 20 components were used to generate the clustering ('FindClusters') and the uniform manifold approximation and projection (UMAP) embeddings (resolution of 0.5).

Module Scores

To determine the cell cycle phase for each cell, module scores for S-phase and G2M-phase were estimated from gene sets (30) using the Seurat function 'CellCycleScoring'. Module scores for mesenchymal and adrenergic state were calculated from published gene sets (33,34) using the Seurat function 'AddModuleScore'.

Cell type annotation

Cell types were annotated per cluster and sample by using marker genes and cell type annotation curated from (23).

Differential gene expression and gene set enrichment analysis

For cells sequenced using the 10X single-nuclei technology, tumour cells were identified and cells without measured *MYCN* expression were removed.

Remaining nuclei in each sample were ranked by their *MYCN* expression level and grouped by assigning the top 30 percent of cells with highest expression levels the label 'MYCN-high' and bottom 30 percent of cells with lowest expression the label 'MYCN-low'. All other cells were annotated as 'MYCN-med' corresponding to intermediate expression levels. The cell line samples were grouped in the same manner.

Differential expression analysis was performed between MYCN-high and MYCN-low cells in each sample and cell line individually using the Seurat function 'FindMarkers' without

606 logarithmic fold change threshold and a minimum of 5% presence of a feature in the sample
607 of only regarding protein-coding genes.
608 For GSEA, genes were ranked by their logarithmic fold change in decreasing order. The
609 enrichment score of MYCN target genes (32) were calculated using the R package fgsea
610 v1.18 (54). Unsupervised gene set enrichment of all biological processes in the gene
611 ontology terms was performed using the R package clusterProfiler v4.0.5 (55) function
612 'gseGO' with a gene set size between 3 and 800 genes and p-values were corrected using
613 BH. The network of recurrent significant enriched pathways was built using the Add-on
614 ClueGO v.2.5.9 in Cytoscape v.3.9.1 (56,57).

615 **Non negative matrix factorisation and module scores**

616 Transcriptional profiles (modules) for each high-throughput patient sample were determined
617 by non-negative matrix factorization (NMF) using cNMF v1.4 (25). The input matrix was
618 restricted to only contain tumour cells and protein-coding genes. The number of modules k
619 for each sample was determined by running the 'cnmf prepare' command with variable k
620 equals 5 through 15. The resulting stability and error plots were used as guidance as
621 described by Kotliar et al., mostly choosing the most stable number of modules. Each
622 module activity matrix was normalised, so that the sum for each cell equals 1.
623 Pairwise Pearson correlation of module TPM gene score (further as gene score) was
624 performed to determine similar modules. Modules that showed less than 50% significant
625 correlation ($p<0.05$) with other modules were excluded. The remaining modules were
626 grouped using hierarchical clustering and the number of meta modules was determined by
627 comparing the heights in the corresponding dendrogram, by choosing the maximum height.
628 The number of submodules was chosen such that each meta module is divided into at least
629 2 groups and the height in the dendrogram is the largest under this assumption.
630 Functional association of meta modules and sub modules was determined using the top 10
631 genes with the highest gene score in each module and ranking those genes by their
632 frequency among the modules classified as the corresponding meta and sub module. The

top 50 genes were evaluated using g:profiler (58) and STRING (59). In addition, GSEA of all GO-biological processes was performed in each module and the most frequent pathways with a significant positive NES were evaluated.

For meta module representation in UMAP space, the module activity was determined by the sample specific module activity corresponding to the meta module, in case multiple sample modules refer to the same meta module, the sum of module activity is displayed.

The meta pathway analysis is performed for each meta and sub module separately on the ranked list of pathways based on the average NES across sample modules in the respective meta and sub module and uses the set of previously described recurrent significant pathways as pathway test set.

Single-cell DNA-seq analysis

The copy number profiles from cells sequenced with G&T-seq were determined using Ginkgo (36) on the DNA data with bin size 500 kB for CHP212 and TR14 cells and 250 kB for the patient sample. EcDNA amplicon specific copy number was estimated from the raw Ginkgo output (Normalised read counts) by leveraging the bins that overlap amplicon boundaries. Amplicon boundaries were obtained from previous publications (18,22) and recapitulated in the DNA data. For each cell a step function was determined based on the raw Ginkgo output and the Ginkgo copy number. Then the step function was applied to the average read count in the overlapping bins.

For the TR14 *MYCN* and *CDK4* amplicon an additional step was included, because of their overlapping region. The percentage of contributing normalised read count of each amplicon to the overlapping region was estimated by averaging only unique amplicon bins and dividing the normalised read count of the unique *MYCN* amplicon by the sum of the unique *MYCN* and *CDK4* amplicon. The normalised read count in the overlapping region was then split up with respect to the contributing percentage and was further used to average over the raw data of the bins overlapping the amplicon regions.

659 Correlation of genomic and transcriptomic content

660 A sample specific linear model was built for each gene present on an ecDNA amplicon using
661 the lm function in R. The models were built on the G&T-seq data using the gene expression
662 from RNA-seq and the respective amplicon copy number determined as described above.
663 The scEC&T-seq data was used to correlate the gene expression with extrachromosomal
664 (ec) content. Gene specific ec content was determined by binning the genome into 1MB
665 segments, summing up their reads from EC-seq and overlapping the segment boundaries
666 with the gene location. The copy number was estimated using gene expression and applying
667 the gene and sample specific linear model described above.

668 **Declarations**

669 **Ethics approval and consent to participate**

670 Patients were registered and treated according to the trial protocols of the German Society
671 of Pediatric Oncology and Hematology (GPOH). This study was conducted in accordance
672 with the World Medical Association Declaration of Helsinki (2013) and good clinical practice;
673 informed consent was obtained from all patients or their guardians. The collection and use of
674 patient specimens was approved by the institutional review boards of Charité-
675 Universitätsmedizin Berlin and the Medical Faculty, University of Cologne.

676 **Consent for publication**

677 Not applicable

678 **Availability of data and materials**

679 The dataset of the Berlin cohort generated and analysed during the current study will be
680 available in the EGA repository. The other datasets analysed during this study are included
681 in Janksy et al. (23) and at https://scpca.alexslemonade.org/.
682 Code will be made publicly available on zenodo.

Competing interests

AGH and RK are co-founders of AMZL therapeutics. The remaining authors have no competing interests to declare.

Acknowledgements and Funding

AGH is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 398299703 and by the Deutsche Krebshilfe (German Cancer Aid) Mildred Scheel Professorship program – 70114107. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 949172). This project was supported by Cancer Research UK and the National Institute of Health (398299703, the eDynamic Cancer Grand Challenge). This project was supported by the Berlin Institute of Health (BIH). Computation has been performed on the HPC for Research cluster of the Berlin Institute of Health. MCS is funded by the Deutsche Forschungsgemeinschaft (DFG) via the graduate programme CompCancer (RTG2424). RFS is a Professor at the Cancer Research Center Cologne Essen (CCCE) funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia. This work was partially funded by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref 01IS18037A). RCG is supported by a fellowship from the "la Caixa" Foundation (ID 100010434). The fellowship code is LCF/BQ/EU20/11810051.

Authors' contributions

AGH, KH and RFS contributed to the study and analysis design and supervised the project. MCS and KH performed the bioinformatic analyses of single-cell sequencing data and data analysis. MCS, AGH, KH and RFS wrote the manuscript. LB and NW performed FISH experiments and analysed the data. RCG, TC and NW performed sequencing experiments and contributed to quality control. RK contributed to study design and analysis of the

scEC&T data. AE, JHS and AS contributed clinical specimens. All authors approved the final version of the manuscript.

References

1. Mosse YP, Diskin SJ, Wasserman N, Rinaldi K, Attiyeh EF, Cole K, et al. Neuroblastomas have distinct genomic DNA profiles that predict clinical phenotype and regional gene expression. Genes Chromosomes Cancer. 2007 Oct;46(10):936–49.
2. Janoueix-Lerosey I, Schleiermacher G, Michels E, Mosseri V, Ribeiro A, Lequin D, et al. Overall Genomic Pattern Is a Predictor of Outcome in Neuroblastoma. J Clin Oncol. 2009 Mar;27(7):1026–33.
3. Eleveld TF, Oldridge DA, Bernard V, Koster J, Daage LC, Diskin SJ, et al. Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. Nat Genet. 2015 Aug;47(8):864–71.
4. Schramm A, Köster J, Assenov Y, Althoff K, Peifer M, Mahlow E, et al. Mutational dynamics between primary and relapse neuroblastomas. Nat Genet. 2015 Aug;47(8):872–7.
5. Brodeur GM, Seeger RC, Schwab M, Varmus HE, Bishop JM. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. Science. 1984 Jun 8;224(4653):1121–4.
6. Koche RP, Rodriguez-Fos E, Helmsauer K, Burkert M, MacArthur IC, Maag J, et al. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nat Genet. 2020 Jan;52(1):29–34.
7. Gibaud A, Vogt N, Hadj-Hamou NS, Meyniel JP, Hupé P, Debatisse M, et al. Extrachromosomal amplification mechanisms in a glioma with amplified sequences from multiple chromosome loci. Hum Mol Genet. 2010 Apr 1;19(7):1276–85.
8. Vogt N, Gibaud A, Lemoine F, de la Grange P, Debatisse M, Malfoy B. Amplicon rearrangements during the extrachromosomal and intrachromosomal amplification process in a glioma. Nucleic Acids Res. 2014 Dec 1;42(21):13194–205.

9. Cox D, Yuncken C, Spriggs AI. MINUTE CHROMATIN BODIES IN MALIGNANT TUMOURS OF CHILDHOOD. Lancet. 1965 Oct;1(7402):55–8.
10. deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat Genet. 2018 May;50(5):708–17.
11. Verhaak RGW, Bafna V, Mischel PS. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat Rev Cancer. 2019 May;19(5):283–8.
12. Kim H, Zheng S, Amini SS, Virk SM, Mikkelsen T, Brat DJ, et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. Genome Res. 2015 Mar;25(3):316–27.
13. Kim H, Nguyen NP, Turner K, Wu S, Gujar AD, Luebeck J, et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat Genet [Internet]. 2020; Available from: https://doi.org/10.1038/s41588-020-0678-2
14. Nathanson DA, Gini B, Mottahedeh J, Visnyei K, Koga T, Gomez G, et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. Science. 2014 Mar;343(6166):72–6.
15. Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J, et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature. 2019 Nov;575(7784):699–703.
16. Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 2017 Feb;543(7643):122–5.
17. Deshpande V, Luebeck J, Nguyen NPD, Bakhtiari M, Turner KM, Schwab R, et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. Nat Commun. 2019;10(1):392.
18. Helmsauer K, Valieva ME, Ali S, González RC, Schöpflin R, Röefzaad C, et al. Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in

763 neuroblastoma. Nat Commun. 2020;11(1):1–12.
764 19. Morton AR, Dogan-Artun N, Faber ZJ, MacLeod G, Bartels CF, Piazza MS, et al.
765 Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. Cell.
766 2019;179(6):1330-1341.e13.
767 20. Lange JT, Rose JC, Chen CY, Pichugin Y, Xie L, Tang J, et al. The evolutionary
768 dynamics of extrachromosomal DNA in human cancers. Nat Genet. 2022 Sep 19;1–7.
769 21. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel
770 sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015
771 Jun;12(6):519–22.
772 22. Rocío Chamorro González, Thomas Conrad, Maja Stöber, Robin Xu, Kerstin Haase,
773 Richard Koche, et al. Parallel sequencing of extrachromosomal circular DNAs and
774 transcriptomes in single cancer cells. Rev.
775 23. Jansky S, Sharma AK, Körber V, Quintero A, Toprak UH, Wecht EM, et al. Single-cell
776 transcriptomic analyses provide insights into the developmental origins of
777 neuroblastoma. Nat Genet. 2021 May;53(5):683–93.
778 24. Dong R, Yang R, Zhan Y, Lai HD, Ye CJ, Yao XY, et al. Single-Cell Characterization of
779 Malignant Phenotypes and Developmental Trajectories of Adrenal Neuroblastoma.
780 Cancer Cell [Internet]. 2020; Available from:
781 http://www.sciencedirect.com/science/article/pii/S1535610820304256
782 25. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, et al. Identifying gene
783 expression programs of cell-type identity and cellular activity with single-cell RNA-Seq.
784 Valencia A, Barkai N, Mereu E, Göttgens B, editors. eLife. 2019 Jul 8;8:e43803.
785 26. Barkley D, Moncada R, Pour M, Liberman DA, Dryg I, Werba G, et al. Recurrence of
786 cancer cell states across diverse tumors and their interactions with the microenvironment
787 [Internet]. bioRxiv; 2021 [cited 2022 Sep 1]. p. 2021.12.20.473565. Available from:
788 https://www.biorxiv.org/content/10.1101/2021.12.20.473565v1
789 27. Boon K, Caron HN, van Asperen R, Valentijn L, Hermus MC, van Sluis P, et al. N-myc
790 enhances the expression of a large set of genes functioning in ribosome biogenesis and

protein synthesis. EMBO J. 2001 Mar 15;20(6):1383–93.

28. Hald ØH, Olsen L, Gallo-Oller G, Elfman LHM, Løkke C, Kogner P, et al. Inhibitors of ribosome biogenesis repress the growth of MYCN-amplified neuroblastoma. Oncogene. 2019 Apr;38(15):2800–13.

29. Dzieran J, Rodriguez Garcia A, Westermark UK, Henley AB, Eyre Sánchez E, Träger C, et al. MYCN-amplified neuroblastoma maintains an aggressive and undifferentiated phenotype by deregulation of estrogen and NGF signaling. Proc Natl Acad Sci. 2018 Feb 6;115(6):E1229–38.

30. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016 Apr 8;352(6282):189–96.

31. Galderisi U, Di Bernardo G, Cipollaro M, Peluso G, Cascino A, Cotrufo R, et al. Differentiation and apoptosis of neuroblastoma cells: Role of N-myc gene product. J Cell Biochem. 1999;73(1):97–105.

32. Westermann F, Muth D, Benner A, Bauer T, Henrich KO, Oberthuer A, et al. Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. Genome Biol. 2008;9(10):R150.

33. Boeva V, Louis-Brennetot C, Peltier A, Durand S, Pierre-Eugène C, Raynal V, et al. Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries. Nat Genet. 2017 Sep;49(9):1408–13.

34. van Groningen T, Koster J, Valentijn LJ, Zwijnenburg DA, Akogul N, Hasselt NE, et al. Neuroblastoma is composed of two super-enhancer-associated differentiation states. Nat Genet. 2017 Aug;49(8):1261–6.

35. Yi E, Gujar AD, Guthrie M, Kim H, Zhao D, Johnson KC, et al. Live-Cell Imaging Shows Uneven Segregation of Extrachromosomal DNA Elements and Transcriptionally Active Extrachromosomal DNA Hubs in Cancer. Cancer Discov. 2022 Feb;12(2):468–83.

36. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, et al. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods. 2015

819 Nov;12(11):1058–60.
820 37. Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, Kahles A, et al. Genomic
821 basis for RNA alterations in cancer. Nature. 2020 Feb;578(7793):129–36.
822 38. Hung KL, Yost KE, Xie L, Wu S, Lange JT, Duffy CV, et al. EcDNA hubs drive
823 cooperative intermolecular oncogene expression [Internet]. Cold Spring Harbor
824 Laboratory. 2020. Available from:
825 https://www.biorxiv.org/content/10.1101/2020.11.19.390278v1
826 39. Zhou YH, Chen Y, Hu Y, Yu L, Tran K, Giedzinski E, et al. The role of EGFR double
827 minutes in modulating the response of malignant gliomas to radiotherapy. Oncotarget.
828 2017 Sep 8;8(46):80853–68.
829 40. Peirce SK, Findley HW. High level MycN expression in non-MYCN amplified
830 neuroblastoma is induced by the combination treatment nutlin-3 and doxorubicin and
831 enhances chemosensitivity. Oncol Rep. 2009 Dec;22(6):1443–9.
832 41. Ryl T, Kuchen EE, Bell E, Shao C, Flórez AF, Mönke G, et al. Cell-Cycle Position of
833 Single MYC-Driven Cancer Cells Dictates Their Susceptibility to a Chemotherapeutic
834 Drug. Cell Syst. 2017 Sep;5(3):237-250.e8.
835 42. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al.
836 Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat
837 Biotechnol. 2020 Jun;38(6):737–46.
838 43. Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. Single-
839 nucleus and single-cell transcriptomes compared in matched cortical cell types. PLOS
840 ONE. 2018 Dec 26;13(12):e0209648.
841 44. Lake BB, Codeluppi S, Yung YC, Gao D, Chun J, Kharchenko PV, et al. A comparative
842 strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted
843 cell-type expression from nuclear RNA. Sci Rep. 2017 Jul 20;7(1):6031.
844 45. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively
845 parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Apr;8(1):14049.
846 46. Krueger F, James F, Ewels P, Afyounian E, Schuster-Boeckler B.

FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo [Internet]. Zenodo; 2021 [cited 2022 Dec 19]. Available from: https://zenodo.org/record/5127899

47. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv; 2013 [cited 2022 Dec 19]. Available from: http://arxiv.org/abs/1303.3997
48. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019 Aug;37(8):907–15.
49. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013 Nov;10(11):1096–8.
50. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018 Jan 4;46(D1):D794–801.
51. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12(1):323.
52. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.
53. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019 Apr 24;8(4):329-337.e4.
54. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis [Internet]. bioRxiv; 2021 [cited 2022 Sep 1]. p. 060012. Available from: https://www.biorxiv.org/content/10.1101/060012v3
55. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021 Aug 28;2(3):100141.
56. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504.
57. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a

Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009 Apr 15;25(8):1091–3.

58. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019 Jul 2;47(W1):W191–8.

59. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607–13.

Additional Files

Additional File 1

Word document (.docx)

Supplementary Figures

Supplementary Figure1a: UMAPs per patient, cell type annotation and MYCN level

Supplementary Figure1b: Barplot of heights in dendrogram

Supplementary Figure1c: Boxplot of correlation between MYCN expression and module activity grouped by submodule

Supplementary Figure1d: UMAP of patient1 coloured by cell cycle phase

Supplementary Figure1e: Barplot of MYCN group per sample coloured by cell cycle phase

Supplementary Figure1f: Boxplot of MES and ADRN score for each patient

Supplementary Figure2a: FISH images uncombined (interphase and metaphase)

Supplementary Figure2b: Patient amplicon

Supplementary Figure2c: Patient copy number count

Supplementary Figure2d: Patient MYCN CN vs expression plot

Supplementary Figure3a/b: Correlation copy number gene expression for all genes on amplicon in CHP212 and TR14

Supplementary Figure3c: Correlation EC-Seq reads and gene expression

903 Supplementary Figure3d: GSEA of MYCN target genes for TR14 cells

904

905 Additional File 2

906 Excel spreadsheet (.xlsx)

907 Supplementary Tables

908 Table S1: Patient meta data including QC statistics

909 Table S2: Meta data per cell including cell cycle phase, cell type, mesenchymal and

910 adrenergic score

911 Table S3: List of top 10 gene scores per NMF module

912 Table S4: DEG between MYCN-high and MYCN-low per patient

913 Table S6: Cell line copy number and gene expression of oncogenes

914 Table S7: Patient 8 GnT-seq copy number and gene expression of *MYCN*

915 Table S8: Linear model of oncogenes in CHP212 and TR14

916 Table S9: DEG between MYCN-high and MYCN-low in CHP212

917 Table S10: DEG between MYCN-high and MYCN-low in TR14

918 Table S11: GSEA results of 38 recurring pathways in CHP212 and TR14

919

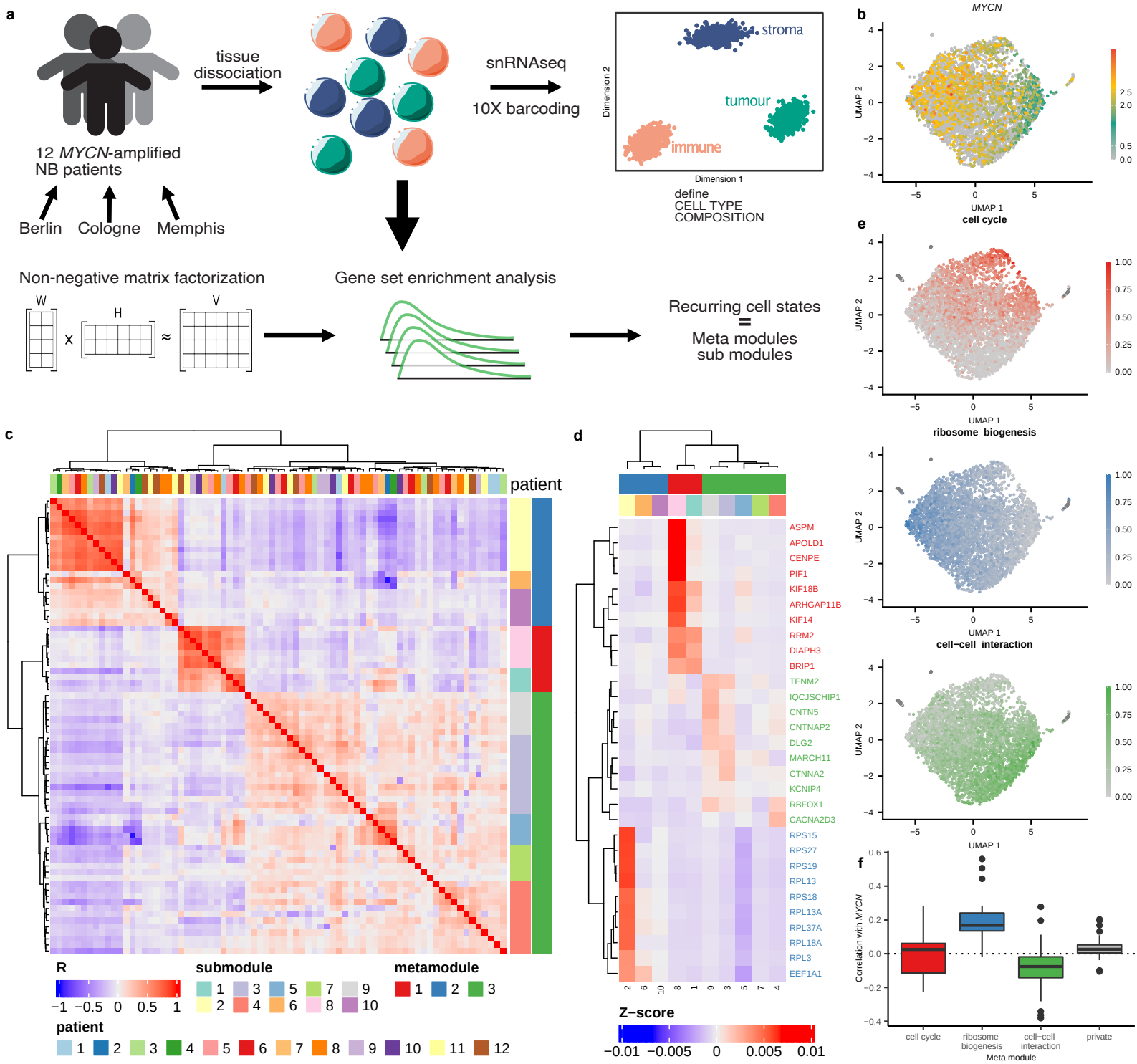
920 Additional File 3

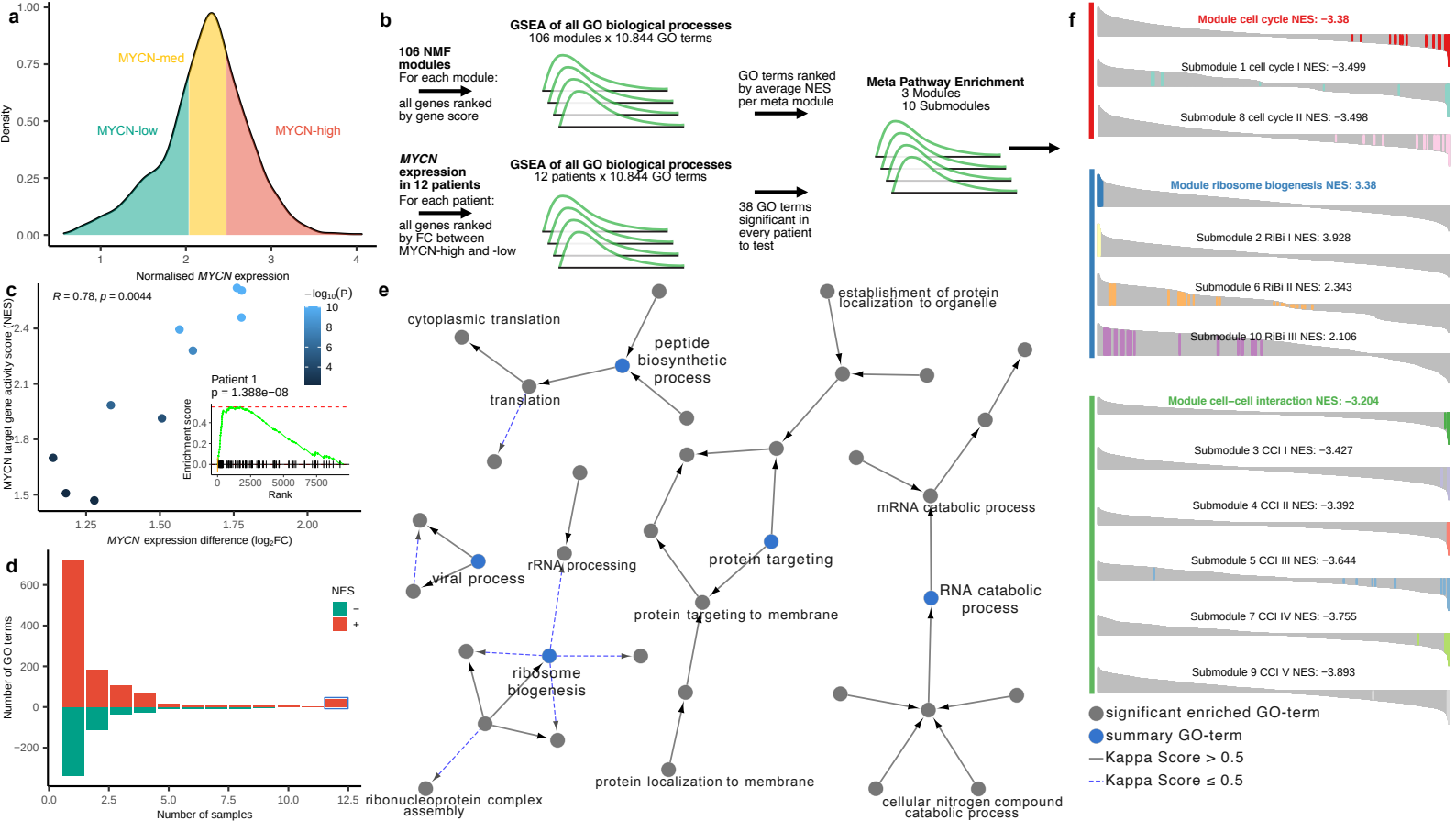
921 Excel spreadsheet (.xlsx)

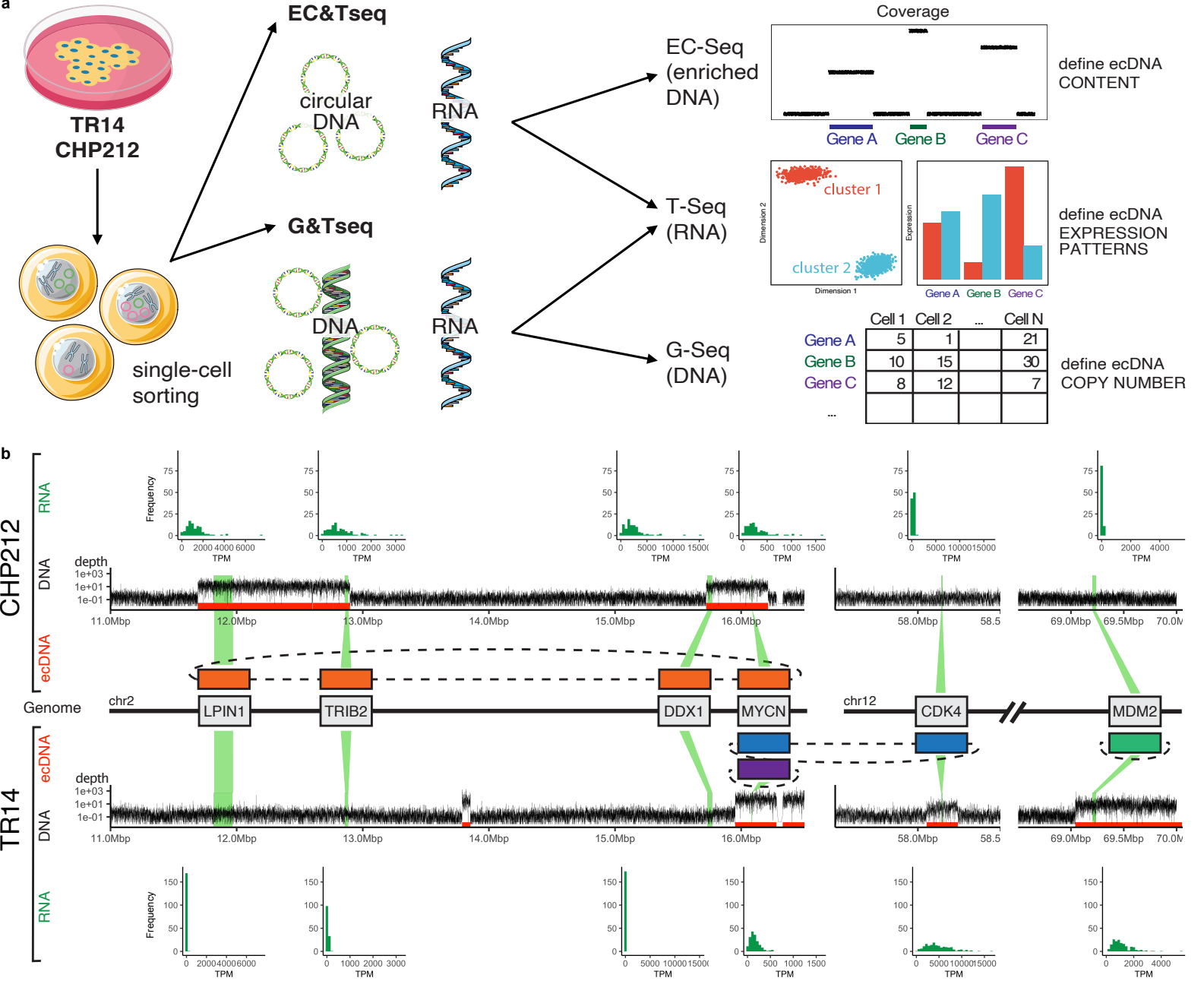
922 Supplementary Table

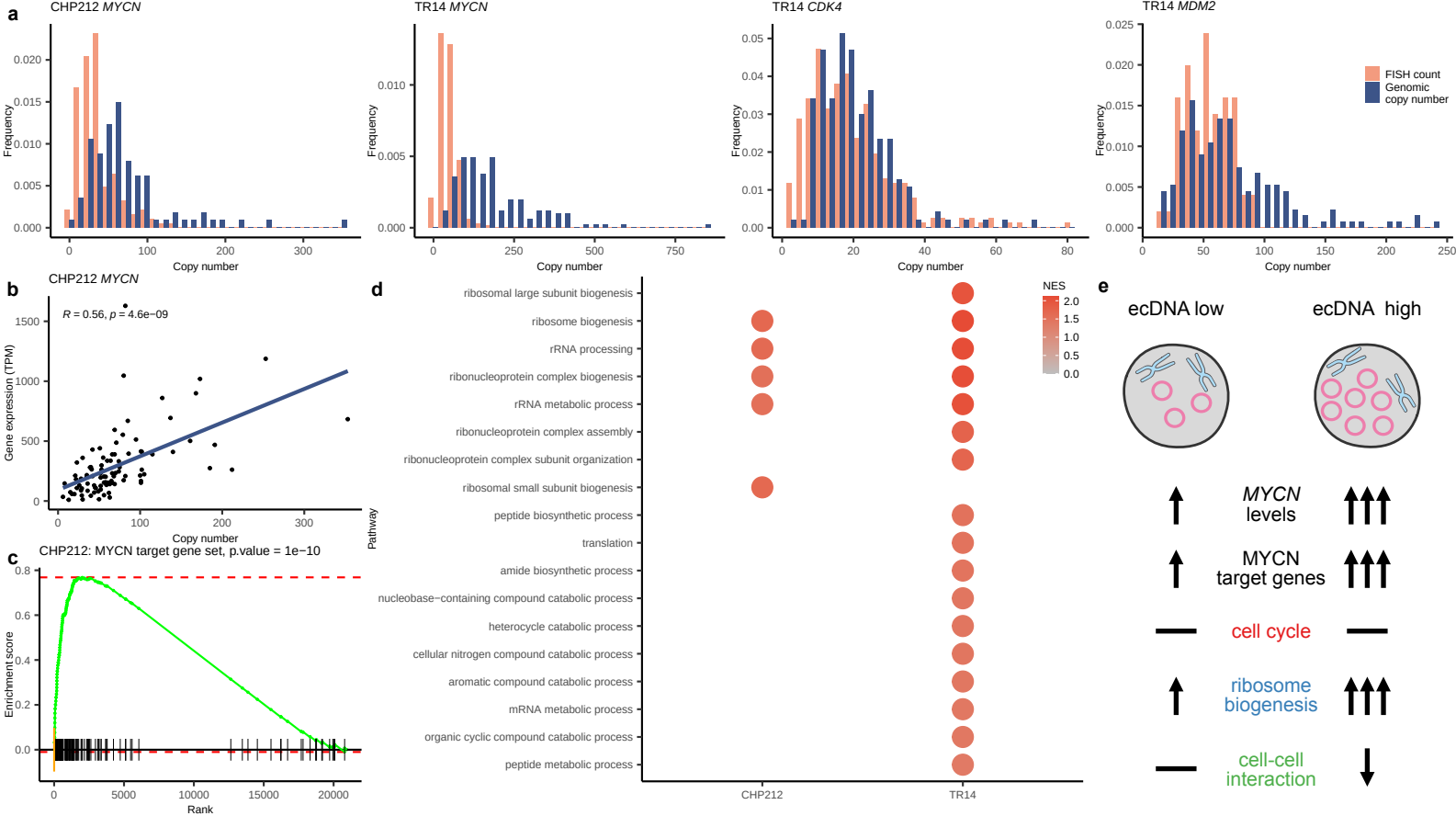
923 Table S5: GSEA results for patients

924


a
12 MYCN-amplified NB patients
Berlin
Cologne
Memphis
tissue dissociation
snRNAseq
10X barcoding
stroma
tumour
immune
Dimension 1
Dimension 2
define CELL TYPE COMPOSITION
Non-negative matrix factorization
W
H
V
Gene set enrichment analysis
Recurring cell states = Meta modules sub modules
b
MYCN
UMAP 1
UMAP 2
e
cell cycle
ribosome biogenesis
cell−cell interaction
c
patient
R
submodule
metamodule
patient
d
ASPM
APOLD1
CENPE
PIF1
KIF18B
ARHGAP11B
KIF14
RRM2
DIAPH3
BRIP1
TENM2
IQCJSCHIP1
CNTN5
CNTNAP2
DLG2
MARCH11
CTNNA2
KCNIP4
RBFOX1
CACNA2D3
RPS15
RPS27
RPS19
RPL13
RPS18
RPL13A
RPL37A
RPL18A
RPL3
EEF1A1
Z−score
f
Correlation with MYCN
cell cycle
ribosome biogenesis
cell−cell interaction
private
Meta module


a
MYCN-med
MYCN-low
MYCN-high
Density
Normalised MYCN expression
b
GSEA of all GO biological processes
106 modules x 10.844 GO terms
106 NMF modules
For each module:
all genes ranked by gene score
MYCN expression in 12 patients
For each patient:
all genes ranked by FC between MYCN-high and -low
GSEA of all GO biological processes
12 patients x 10.844 GO terms
GO terms ranked by average NES per meta module
38 GO terms significant in every patient to test
Meta Pathway Enrichment
3 Modules
10 Submodules
c
R = 0.78, p = 0.0044
−log10(P)
Patient 1
p = 1.388e−08
Enrichment score
Rank
MYCN target gene activity score (NES)
MYCN expression difference (log2FC)
d
NES
Number of GO terms
Number of samples
e
cytoplasmic translation
peptide biosynthetic process
translation
establishment of protein localization to organelle
mRNA catabolic process
viral process
rRNA processing
protein targeting
protein targeting to membrane
RNA catabolic process
ribosome biogenesis
protein localization to membrane
ribonucleoprotein complex assembly
cellular nitrogen compound catabolic process
f
Module cell cycle NES: −3.38
Submodule 1 cell cycle I NES: −3.499
Submodule 8 cell cycle II NES: −3.498
Module ribosome biogenesis NES: 3.38
Submodule 2 RiBi I NES: 3.928
Submodule 6 RiBi II NES: 2.343
Submodule 10 RiBi III NES: 2.106
Module cell−cell interaction NES: −3.204
Submodule 3 CCI I NES: −3.427
Submodule 4 CCI II NES: −3.392
Submodule 5 CCI III NES: −3.644
Submodule 7 CCI IV NES: −3.755
Submodule 9 CCI V NES: −3.893
significant enriched GO-term
summary GO-term
Kappa Score > 0.5
Kappa Score ≤ 0.5


a
TR14
CHP212
single-cell sorting
EC&Tseq
circular DNA
RNA
G&Tseq
DNA
RNA
EC-Seq (enriched DNA)
T-Seq (RNA)
G-Seq (DNA)
Coverage
Gene A
Gene B
Gene C
define ecDNA CONTENT
cluster 1
cluster 2
Dimension 1
Dimension 2
Expression
define ecDNA EXPRESSION PATTERNS
Cell 1
Cell 2
Cell N
Gene A 5 1 21
Gene B 10 15 30
Gene C 8 12 7
define ecDNA COPY NUMBER
b
CHP212
TR14
RNA
DNA
ecDNA
Frequency
TPM
depth
1e+03
1e+01
1e-01
11.0Mbp
12.0Mbp
13.0Mbp
14.0Mbp
15.0Mbp
16.0Mbp
58.0Mbp
69.0Mbp
69.5Mbp
70.0Mbp
Genome
chr2
chr12
LPIN1
TRIB2
DDX1
MYCN
CDK4
MDM2


a
CHP212 MYCN
TR14 MYCN
TR14 CDK4
TR14 MDM2
Frequency
Copy number
FISH count
Genomic copy number
b
CHP212 MYCN
R = 0.56, p = 4.6e−09
Gene expression (TPM)
Copy number
c
CHP212: MYCN target gene set, p.value = 1e−10
Enrichment score
Rank
d
ribosomal large subunit biogenesis
ribosome biogenesis
rRNA processing
ribonucleoprotein complex biogenesis
rRNA metabolic process
ribonucleoprotein complex assembly
ribonucleoprotein complex subunit organization
ribosomal small subunit biogenesis
peptide biosynthetic process
translation
amide biosynthetic process
nucleobase−containing compound catabolic process
heterocycle catabolic process
cellular nitrogen compound catabolic process
aromatic compound catabolic process
mRNA metabolic process
organic cyclic compound catabolic process
peptide metabolic process
Pathway
NES
CHP212
TR14
e
ecDNA low
ecDNA high
MYCN levels
MYCN target genes
cell cycle
ribosome biogenesis
cell-cell interaction