sciCNV: High-throughput paired profiling of transcriptomes and DNA copy number variations at single cell resolution

Authors: Ali Mahdipour-Shirayeh, Natalie Erdmann, Chungyee Leung-Hagesteijn and Rodger E. Tiedemann*

Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

Departments of Medicine and Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

*Senior and corresponding author: Rodger E. Tiedemann: rodger.tiedemann@uhn.ca

Key words: single cell RNA sequencing (scRNA-seq), copy number variation (CNV), normalization, multiple myeloma, RTAM, sciCNV, multi-omics

SUMMARY

Chromosome copy number variations (CNVs) are a near-universal feature of cancer however their effects on cellular function are incompletely understood. Single cell RNA sequencing (scRNA-seq) can reveal cellular gene expression however cannot directly link this to CNVs. Here we report new normalization methods (RTAM1 and -2) for scRNA-seq that improve gene expression alignment between cells, enhancing gene expression comparisons and the application of scRNA-seq to CNV detection. We also report sciCNV, a pipeline for inferring CNVs from RTAM-normalized data. Together, these tools provide dual profiling of transcriptomes and CNVs at single-cell resolution, enabling exploration of the effects of cancer CNVs on cellular programs. We apply these tools to multiple myeloma (MM) and examine the cellular effects of cancer CNVs +8q. Consistent with prior reports, MM cells with +8q22-24 upregulate MYC, MYC-target genes, mRNA processing and protein synthesis, verifying the approach. Overall, we provide new tools for scRNA-seq that enable matched profiling of the CNV landscape and transcriptome of single cells, facilitate deconstruction of the effects of cancer CNVs on cellular reprogramming within single samples.

INTRODUCTION

Aneuploidy and focal CNVs are a pervasive feature of cancer[1]. However, the specific influence of many cancer CNVs on the transcription programs and molecular processes within cells remains only partially understood[2,3]. While may studies have sought to integrate cancer CNVs with gene expression effects by examining differences across tumor samples, such studies are often impeded by the extensive genetic heterogeneity that exists between tumors, confounding the attribution of transcriptional variances to single genetic loci. As a consequence, integrative inter-tumor studies commonly require the profiling of large numbers of samples to minimize confounding variations, and their application to less common cancer types or CNVs may be problematic.

To assess the transcriptional effects of specific CNVs, examination of the intra-clonal effects of emergent CNV at single cell resolution would have the advantage that subclonal cells with and without CNV are otherwise virtually isogenic, and that gene expression changes could be more readily attributed to the CNV. Therefore, simultaneous sequencing of single cell genomes and transcriptomes should in theory enable direct pairing of CNVs with their transcriptional outcomes with high precision. Notably, sequencing of both genomic DNA and RNA within single cells has been reported[4-6]. Prior techniques, however, provided profiling of only a few cells and thus afford only a limited view of the heterogeneity within any cancer. For the identification of CNV function, sequencing of much larger numbers of cells is required to capture intra-tumor subclones and to adequately profile subclonal populations for the gene expression pattern that aligns with an individual CNV.

Whilst scRNA-seq can reveal the transcription state of single cells, it cannot directly relate this to DNA lesions. CNVs can be inferred from scRNA-seq, which could thus be leveraged to provide both layers of omics information within individual cells. However, previously reported approaches[7-9] reveal constraints imposed by the sparsity of single-cell data and have not reported quantitative measures of accuracy. In particular, inconsistencies in the detection of lowly-expressed genes within single cells causes stochastic noise that influences transcriptome

distribution and interferes with RNA-based CNV detection. Normalization is thus critical for accurate scRNA-seq interpretation[10-14] and for secondary CNV detection.

Here we report RTAM1 and RTAM2 normalization methods for scRNA-seq that improve gene expression alignment between cells and thus enhance the sensitivity of scRNA-seq for the detection of small expression changes arising from gene copy number differences. We also report sciCNV, a new pipeline for inferring CNVs in single cells from RTAM-normalized scRNA-seq data. Together, these methods enable high-throughput paired profiling of both DNA copy number and RNA in the same cells, facilitating direct intra-clonal examination of the effects of cancer CNVs on gene expression programs at a cellular level.

RESULTS

Enhanced single-cell RNA-seq normalization methods: RTAM1 and -2

Single-cell RNA-seq enables gene expression comparisons between cells. However, the accuracy of these comparisons depends critically upon data normalization. Normalization is essential to convert raw RNA transcript data from single cells into gene expression values that are objectively comparable between cells, eradicating biases related to cellular sequencing depth, transcriptome size, technical factors or batch effect. As the best methods for normalizing scRNA-seq remain controversial, each with its own drawbacks, we developed RTAM1 and -2 (described in the **online methods** and **supplementary figures S1-3**) and compared the RTAM methods with other normalization strategies currently in use.

To compare the methods for their control of systemic and stochastic variations between cells due to size or sequence depth, we generated scRNA-seq data for normal and malignant cells belonging uniformly to the B cell lineage (n>15,000 cells) (**figure 1a**). We examined cells with a similar ancestry in order to minimize confounding variations between them due to lineage specification. However, we deliberately generated data from a mix of both small quiescent B cells and large proliferating plasma cells to ensure that the normalization methods would be challenged by cells embodying a full spectrum of sizes and transcriptional activities. The cells

were isolated from human bone marrow samples by FACS and were profiled using the 10X Genomics single cell RNA-seq library kit. Cell- and gene-specific transcripts were enumerated using barcoded unique molecular identifiers (UMI).

The raw scRNA-seq data from one of the test samples is depicted in **figure 1b**. As shown, the distributions of transcript numbers per gene varied significantly from cell to cell, predominantly reflecting differences in the cellular transcriptome sizes and demonstrating a clear need for normalization. The test samples were next normalized using either TPM[15], SCRAN[11], SCONE[12] or Seurat's SCTransform function[16] (**figure 1b** and **supplementary figures S4-S20**). To compare the alignments of the normalized transcriptomes, we examined the mean and median expression in each cell of a curated list of housekeeping genes (HKG) known to be broadly expressed with low variation[9]. We also examined the average expression in each cell of all of the ubiquitously-expressed genes (UEG) detected in >95% of the cells in the sample. For each sample tested, the UEG represent the largest possible set of genes that are commonly expressed across the test cells. Whilst the expression of any individual gene is expected to vary between cells for both biological and technical reasons, the average expression per cell of a large set of ubiquitous genes should be similar, particularly amongst cells of the same lineage, and its variance between cells provides a metric of normalization effectiveness.

As shown in **figures 1c, 2** and **supplementary figures S4-20**, TPM, which normalizes cellular transcriptomes primarily by their total transcript count, produced a very large variance in the average expression of HKG or UEG between individual cells, suggesting significant limitations for scRNA-seq application. By comparison, SCRAN and SCONE produced superior alignment of gene expression averages across cells. However, SCONE, which produced the better alignment, achieved this only by implementing quantile normalization – exchanging the actual distribution of transcript counts in each cell for a standardized distribution – which caused a significant loss of inter-cellular variation, particularly in key highly-expressed genes. The expression of IGH or IGL genes, for example, a critical feature of plasma cells, was reduced by quantile normalization into a virtual constant across cells (**supplementary figure S21**).

111 As each of these scRNA-seq normalization methods has limitations, we developed RTAM1 and -
112 2. The RTAM methods originate from a consideration of the strengths and weaknesses of
113 scRNA-seq. Whereas lowly expressed genes are detected within single cells with low resolution
114 (due to low integer transcript counts) and show significant stochastic variation, highly expressed
115 genes are robustly detected and show finer quantisation of variation relative to intensity. RTAM1
116 and -2 thus utilize highly-expressed genes, whose expression is resolved with greater accuracy,
117 to align cellular transcriptomes. Genes are ranked in each cell by their expression and the
118 summed intensities of the top-ranked genes is standardized in log-space using unique non-linear
119 cell- and gene-specific adjustments of gene expression determined either by cellular gene
120 expression rank (RTAM1) or by gene expression intensity (RTAM2) (see **methods**).
121
122 Importantly, compared to TPM, SCTransform or SCRAN, both RTAM1 and RTAM2 reduce the
123 cell-to-cell variance in the average (median or mean) expression of HKG and UEG sets (**figure 2**
124 and **supplementary figures S4-S20**). The coefficients of variation (CV) produced by each
125 normalization method for the "average" expression of HKG or UEG in individual cells is shown
126 in **figure 2**, for 3 independent patient samples. As shown, RTAM1 and RTAM2 reduce
127 variations in the average gene expression of single cells, irrespective of whether this average
128 expression is calculated by any of 3 different methods. By design, the RTAM methods also
129 standardize the average expression of highly-expressed genes, and thus overall these methods
130 produce superior alignments of cellular transcriptomes and of gene expression between cells. At
131 the same time, both RTAM1 and RTAM2 maintain the original variability observed between
132 cells in the expression of individual highly-transcribed genes, unlike the quantile normalization
133 implemented by SCONE (**supplementary figure S21**). Overall, therefore, the RTAM methods
134 represent useful new strategies for normalizing scRNA-seq data that can enhance the accuracy of
135 gene expression comparisons between cells.
136
137 **Single-cell inferred chromosomal copy number variation: sciCNV**
138 We next sought to develop a method for detecting single-cell chromosomal CNV from scRNA-
139 seq, leveraging the enhanced normalization provided by RTAM to increase the sensitivity of
140 single-cell transcriptomics for CNV detection. To optimize DNA copy number estimates from
141 gene expression, and to mitigate against data sparsity in single cells, we developed a two-

142 pronged approach, called sciCNV (described in the **methods**). Briefly, for each cell RTAM-
143 normalized gene expression was aligned with that of control diploid cells to develop expression
144 disparity scores, which were averaged in a moving window across the genome. The average gene
145 expression of the control cells was weighted by the probability of gene detection, enhancing its
146 comparison with single cell data, where signal '*dropout*' was common for many genes. In a
147 parallel method, the expression disparity values for each gene were exchanged for binary +1 or -
148 1 values, which were summed cumulatively in each cell as a function of genomic location; the
149 gradient of this function yielded a second estimate of the single cell CNV profile that was
150 sensitive to small concordant expression variations in contiguous genes and that was insensitive
151 to large single-gene variations. The CNV estimates of the two methods were then combined by
152 their geometric mean.
153
154 **Figure 3** shows sciCNV applied to scRNA-seq data from primary multiple myeloma (MM) cells.
155 Significantly, the CNV profile of a single MM cell, inferred from its RNA, closely resembles the
156 average CNV profile of $>10^4$ tumor bulk cells, derived from whole exome DNA sequencing
157 (WES) of the same MM sample (R^2=0.72) (**figure 3a,b**). In addition, the CNV predictions
158 produced from a single cell by sciCNV also aligned with CNV estimates provided by FISH
159 (**figure 3c**). Furthermore, examination of >1,700 individual plasma cells from the same MM
160 patient biopsy using sciCNV revealed that the tumor clone CNVs were robustly detected in each
161 of the MM cells (**supplementary figure S22a**), despite biological and technical variations
162 between them; and were not detected in normal plasma cells (NPC). Thus, sciCNV can utilize
163 scRNA-seq data to reveal CNVs in single cells. Moreover, it can distinguish cancer cells and
164 normal cells on the basis of their CNV profile (**supplementary figure S22b,c**).
165
166 To further examine the accuracy and consistency of sciCNV in detecting tumor CNVs in single
167 cells, we next compared the sciCNV profiles of thousands of individual plasma cells from 4 MM
168 samples with the corresponding tumor CNV profiles generated from DNA sequencing of $>10^6$
169 tumor cells (**figure 4**). Remarkably, the single cell CNV profiles calculated by sciCNV from
170 RNA substantively recapitulated the complex CNV landscape of each tumor clone, defined by
171 tumor bulk DNA sequencing. Thus, for the MM199 sample, gains within 1q, 7q, 8q, 9, 17q, 18q,
172 19p and Xq, and losses within 1p, 2q, 4p, 6p, 6q, 13, 14q, 16p, 17p, 17q, and Xp were all

173 identified by sciCNV at a single cell level, and these corresponded to CNVs identified by tumor
174 bulk WES (**figure 4a**). Similarly, for sample MM241, trisomies of chromsomes 5, 7 and 19 and
175 monosomy of chromosome 13 were all identified by sciCNV in single MM cells, matching the
176 WES-defined CNV profile of pooled tumor bulk cells (**figure 4b**). Furthermore, sciCNV also
177 identified a subpopulation of MM241 cells with subclonal gain of 1q, and the existence of this
178 subclone appeared to be borne out by the tumor DNA WES profile which also identified a 1q-
179 arm copy gain with an average CNV peak height of approximately 35%. Similarly, sciCNV
180 profiling of copy number variations in individual cells from sample MM244, developed using
181 RNA-seq data, identified clonal gains within chromosome 7 and clonal losses within
182 chromosomes 1p, 10p, 13, 14q, 16q and 17p that were also identified by DNA sequencing of
183 pooled tumor bulk cells from the same patient; and additionally identified subclonal loss of the
184 distal arms of 1p and 2q, and subclonal gains of 11q and 19p (in subpopulations of cells) that
185 were also detected at subclonal CNV levels (0.15-0.50) by WES (**figure 4c**). Finally, sciCNV
186 identified clonal gains of 1q and 6p and clonal deletions within 1p, 2p, 6q, 13, 14q, 18 and 22 in
187 single cells from MM254 that were all also identified on tumor bulk DNA sequencing (**figure**
188 **4d**). Thus, these data demonstrate that the genomic CNV landscape of individual tumor cells can
189 largely be inferred from scRNA-seq transcriptome data by sciCNV, with results that show
190 reasonable correlation to DNA-based CNV profiling of pooled tumor bulk cells, although the
191 authors acknowledge that using RNA-based approaches small CNV regions may be missed.
192
193 Notably, several alternative methods exist for inferring CNVs from scRNA-seq[8,9,17,18]. Of these,
194 inferCNV[17] has been broadly used to calculate and visualize CNVs at single cell resolution.
195 Next, therefore, we compared our RTAM2-sciCNV pipeline with inferCNV, applying both
196 methods to the same set of normal and MM cells shown in figure 4 and comparing their single
197 cell outputs with the CNV profiles obtained from tumor bulk WES (**supplementary figures**
198 **S23-26**). As shown, sciCNV more reliably predicted tumor CNVs that were identified by WES,
199 and did so more consistently across individual cells. For example, for MM199 cells sciCNV
200 more readily detected gains of the distal arm of 1q, 3q, 7q, 9q and within 17q and 18q than
201 inferCNV (**supplementary figures S23)**. At the same time sciCNV generally produced less false
202 positive CNV noise than inferCNV. Thus, for MM241 both methods detected subclonal gain of
203 1q and trisomies of 5,7 and 19, and deletion of 13 and 14q, but inferCNV also reported a number

204 of false positive deletions, particularly in chromosomes 2 and 4 (**supplementary figures S24)**.
205 Overall, therefore, whilst both methods provide CNV predictions that are substantially correct, in
206 general sciCNV offered a closer correlation with WES-based CNV findings.
207
208 **Dissecting the effects of CNVs on gene expression within single cells: the transcriptional**
209 **consequences of +8q22-24 in MM**
210 Simultaneous profiling of both DNA copy number and RNA in the same cell should enable
211 examination of the effect of a CNV on transcriptional programs. To test this, we used sciCNV to
212 study the effect of gain of the distal arm of 8q on cellular processes. We examined +8q as this is
213 one of the most recurrent abnormalities in human cancer[2,17] and is known to target MYC[18],
214 providing a benchmark result for testing our method.
215
216 Using sciCNV, sample MM199 was found to contain cells with subclonal gain of chromosome
217 8q22-24 (**figure 5a**). Importantly, MM199 also contained intra-clonal isogenic cells without +8q.
218 To facilitate gene set enrichment analyses (GSEA)[19] of these sibling cells, with and without +8q,
219 the subclones were next subsampled to obtain subpopulations of cells with matching
220 transcriptome depth (**figure 5b**). This eliminated biases in cell sequencing depths or in
221 transcriptome sizes between the subclones, and thus balanced the likelihood of gene-set detection
222 between subclones, at least in the absence of biological enrichment of specific genes. The gene
223 expression of the matched intra-clonal subpopulations, representing isogenic sibling cells with
224 and without +8q22-24, with equivalent transcriptome sizes, were then compared by GSEA using
225 RTAM2-normalized data. From an analysis of 215 cytoband gene sets (comprising genes defined
226 by chromosome location), +8q cells were strongly enriched for gene-sets located at 8q22-24,
227 with striking statistical confidence ($p=0.000$, $q=0.000$, FWER=0.000), compared to cells without
228 +8q (**figure 5d** and **supplemental figure S27).** In contrast, no other genomic regions were
229 significantly enriched in either subclone. Thus, sciCNV accurately resolved single MM cells into
230 intra-tumor subclones, isolating +8q22-24 as a unique variation distinguishing these.
231
232 We next used GSEA to explore the influence of +8q22-24 on cellular programming. Notably, the
233 primary MM cells from MM199 identified by sciCNV as having +8q22-24 showed increased
234 MYC expression ($p<0.05$) compared to sibling cells without +8q, as expected (**figure 5c**).

235 Furthermore, the +8q cells also showed broadly increased expression of hallmark MYC target
236 genes (p=0.000, q=0.000, FWER=0.000). Finally, from an analysis of 3303 curated gene sets,
237 primary MM199 cells with +8q22-24 also showed upregulation of gene-sets encoding the
238 machinery of mRNA translation and protein synthesis, including specifically genes involved in
239 3'UTR-mediated mRNA translation regulation (enrichment rank 5/3303), ribosome biogenesis
240 (rank 4/3303) and peptide chain elongation (rank 3/3303)(**figure 5d, supplemental figure S27**).
241 Conspicuously, these transcriptional effects of +8q22-24 in MM cells were remarkably close to
242 those of +8q23-24 in breast cancer (Nikolsky_Breast_Cancer_8q23-24 amplicon, FWER
243 p=0.000, enrichment rank 1/3303) (**figure 5d, supplemental figure S27**). Thus, overall our
244 analyses indicate that +8q22-24 in primary MM cells induces upregulation of MYC, MYC-target
245 genes and transcriptional programs that promote the upregulation of mRNA translation and
246 peptide synthesis. Analogous gene expression changes appear to occur across malignancies
247 including MM and breast cancer. Strikingly, we were able to identify these transcriptional
248 consequences of +8q22-24 using scRNA-seq analysis of a single MM patient sample. By
249 comparison, the Nikolsky_Breast_Cancer_8q23-24 amplicon signature was derived from the
250 combined genomic analyses of 191 breast cancer samples[20].
251
252 The cellular re-programming induced by +8q22-24 might be expected to promote an increase in
253 gene expression and in cell mass. Notably, however, in the MM sample examined the mTOR-
254 interacting gene, DEPTOR, encoded at 8q24, was also upregulated in +8q cells (**figure 5c**).
255 Upregulation of DEPTOR may limit cell size[21] and counter any tendency to increased cell mass
256 driven by MYC and ribosome biogenesis. Indeed, from our examination of +8q at a single cell
257 level we observed that the transcriptome sizes of +8q cells were in fact mildly reduced,
258 compared to sibling cells without the CNV ($p<0.001$) (**figure 5c**). Examination of cellular
259 transcriptome size is a capability unique to scRNA-seq but is generally not feasible from bulk
260 tumor gene expression studies. Thus, from single-cell analyses of +8q22-24 in primary MM cells
261 we find that copy number gain of this region acts to boost MYC, MYC target gene expression
262 and protein synthesis capacity (ribosomes, translation) without increasing cellular transcriptome
263 size. As the transcription of MYC-target genes is relatively increased, this likely leads to
264 enhanced expression of MYC-target genes as proteins. However, the increase in protein
265 synthesis capacity with reduced transcriptome size may also serve to improve the dynamics of

266 protein production by cells and may thus reduce the lag-time between gene expression and
267 proteome change, potentially enhancing cellular adaptability. This represents a unique
268 perspective on the function of +8q22-24 which is offered only by a multi-omic analysis of single
269 cells.
270
271 **DISCUSSION**
272 CNVs are critical drivers of cancer biology yet their specific effects on cellular processes often
273 remain only partially characterized. Here, we report bioinformatic methods for the dual profiling
274 of DNA copy number and RNA within the same cells, using scRNA-seq data, which enable
275 direct examination of the effect of intra-clonal CNVs on gene expression. We describe RTAM
276 normalization, which provides enhanced transcriptome alignment to improve gene expression
277 comparisons between cells, and sciCNV, which provides robust RNA-based CNV profiling at
278 single cell resolution. In order to demonstrate and validate these new tools, we examine the
279 transcriptional effects of copy number gain of chromosome region 8q22-24, which represents
280 one of the most common CNV in human cancer. From an examination of intra-clonal
281 subpopulations of cells with and without 8q gain we show using only a single MM patient
282 sample that +8q22-24 induces expression of MYC and MYC target genes, as expected,
283 demonstrating that intra-clonal CNVs can be linked to transcriptional consequences using only
284 scRNA-seq alone. This process can be compared to historical studies of cancer CNVs, in which
285 DNA and RNA profiling of hundreds of tumor samples must be performed to achieve similar
286 results. Moreover, single cell approaches potentially provide higher precision results due to the
287 ability to use intra-clonal isogenic cells as controls. Tumor registry studies, in contrast, are
288 encumbered by the immense genomic heterogeneity the exists between tumors, which persists
289 even when tumors are grouped into cohorts.
290
291 Although +8q22-24 upregulates the expression of a broad spectrum of MYC target genes, using
292 single cell profiling we demonstrate that the transcriptomes of MM cells with +8q are in fact
293 smaller than those of cells lacking +8q, at least in the sample examined by us, likely as a result of
294 DEPTOR overexpression from 8q24. At the same time, we find that +8q22-24 cells in primary
295 MM consistently upregulate gene sets involved in mRNA translation, ribosomal biogenesis and
296 peptide elongation. Thus +8q22-24 enhances protein synthesis capacity, without increasing

297 transcriptome size. This likely improves the dynamics of protein expression responses to gene
298 expression; and may enhance the malleability of MM cells with +8q24 to environmental
299 challenges.
300
301 Study of the effects of CNVs on cellular programming via single-cell transcriptomics offers a
302 number of advantages. As intra-clonal cells that diverge at a single CNV are virtually isogenic,
303 divergences in their gene expression can be precisely attributed to the CNV. Furthermore, as
304 cells with and without the CNV are derived from the same sample, differences in their gene
305 expression due to the microenvironment, clinical factors or sample processing are minimized. By
306 comparison, tumor cohort studies must instead rely upon the identification of correlations
307 between CNVs and gene expression across unrelated samples and suffer from the substantial
308 confounding genetic and clinical heterogeneity that exists between samples. As a result, the
309 effects of many cancer CNVs on cellular processes remain only partially characterized.
310
311 Defining the molecular mechanisms by which oncogenic chromosomal events influence cellular
312 behaviour is a critical step in understanding cancer pathogenesis. Indeed, it is essential
313 knowledge that underlies the development of therapeutic strategies to address the foundations of
314 malignant disease. Auspiciously, single cell multi omics offers the promise of a new era in
315 cancer genomics, in which the multi-faceted effects of CNVs on gene expression and cancer cell
316 programming can now be determined with precision.
317
318
319 **METHODS**
320
321 **Samples**
322 Bone marrow samples were collected from MM patients in heparinized tubes in compliance with
323 the Declaration of Helsinki and a protocol approved by the University Health Network Research
324 Ethics Board.
325
326 **Cell isolation**

327 Marrow cells were separated by Ficoll-Plaque (600g x 30 minutes) and treated with ACK red
328 blood cell lysis buffer (Sigma-Aldrich) for 2 minutes. Cells were then stained with fluorescently-
329 labelled antibodies (Biolegend) targeting CD2 (APC-cy7, RPA-2.10), CD3 (APC-cy7, HIT3a),
330 CD4 (APC-cy7, A161A1), CD11b (PEcy7, CBRM1/5), CD14 (PE-cy7, M5E2), CD34 (PE-
331 Dazzle, 581), CD235a (PE-cy7, H1264), CD38 (APC, HB-7), CD138 (FITC, MI15), BCMA
332 (PE, 19F2,) and CD20 (Pacific blue, 2H7) for 30 minutes in PBS 10% FBS 5mM EDTA at 4°C.
333 B cells and plasma cells were enriched on a MoFlo Astrio (Beckman Coulter) by excluding
334 duplets, non-viable cells (binding 7AAD) and cells expressing non-B lineage markers (CD2,
335 CD3, CD4, CD11b, CD124, CD235a, CD34), and by gating cells with lymphoid FSC-SSC
336 characteristics and expressing CD38, CD138 or CD20, as described[23] and illustrated in
337 **Supplemental Fig. S29**.
338
339 **Single cell RNA-sequencing**
340 Sequencing libraries were prepared from enriched cells using the 10X Genomics Chromium
341 Single Cell Library Kit chemistry v2 and Chromium Single Cell Controller; and were sequenced
342 on an Illumina HiSeq 2500 using paired-ends, targeting >50K mean reads per cell, >80%
343 sequencing saturation and phred-Q30 quality scores >80. Reads were de-multiplexed using
344 Illumina bcl2fastq, mapped to GRCh38 and were then consolidated into unique transcripts per
345 gene per cell using Cell Ranger 2.1.0 (10X Genomics). Poor quality cells with high
346 mitochondrial transcript content were excluded if this represented >5% of total cellular
347 transcripts or was >3 median absolute deviations from the median.
348
349 **Non-RTAM scRNA-seq normalization**
350 Normalization and data analysis were performed with R 3.6.1. scRNA-seq expression matrices
351 were normalized using RSEM's transcript per million (TPM) approach[15], SCRAN[11], SCONE[12]
352 or Suerat's (v3.1) SCTransform[16] function, employing *scran*, *scone* or *seurat V3.1* R packages.
353 To facilitate SCONE's quantile normalization, we limited the input matrix to only those genes
354 that had ≥1 read in ≥1 cell. When applying Seurat's SCTransform we used the 'corrected UMI'
355 count provided by this function. To compare the output data from each method on similar scales
356 we applied $\log_2(.+1)$ transformation as required.
357

358 **RTAM scRNA-seq normalization**

359 ***RTAM1:*** The normalized gene expression $GE_{i,j}$ of gene i in cell j was calculated from the raw
360 transcript count $T_{i,j}$, which was first normalized to the total transcript count per cell and then
361 scaled to a standard transcriptome library size (10^5) and $\log_2$ transformed, mirroring the TPM
362 approach[15], yielding library-normalized values $L_{i,j}$:

363

364 $$L_{i,j} = log_2\left[\left(T_{i,j} \times \frac{10^5}{\sum_{i=o}^{i=Nj} T_{i,j}}\right) + 1\right]$$

365

366 N_j is the number of expressed genes (with ≥ 1 transcript) detected in cell j. Genes with non-zero
367 values were then assigned a rank in each cell $R_{i,j}$ according to descending $L_{i,j}$. Where multiple
368 genes had the same $L_{i,j}$ these were assigned an equal rank equivalent to the mean of the values
369 had these genes instead been assigned sequential ranks. M_j, the sum of $L_{i,j}$ of the top-ranked N_T
370 genes was calculated for each cell (from genes satisfying $1 \leq R_{i,j} \leq N_T$), where N_T was the number
371 of highly-expressed genes used to normalize the dataset.

372 $$M_j = \sum_{R_{i,j}=1}^{R_{i,j}=N_T} L_{i,j}$$

373

374 M_k, a target constant for correcting M_j values, was calculated as the mean of the initial M_j of all
375 cells. The adjustment variable, $h_{i,j}$, for gene i in cell j, was derived such that it adjusts the initial
376 M_j for each cell to the target M_k by differentially adjusting cellular gene expression on a $\log_2$
377 scale according to $R_{i,j}$, and was calculated in the form:

378 $$h_{i,j} = C_j \cdot log_2[R_{i,j} + 1]$$

379

380 where C_j is a constant for each cell and equals:

381

382 $$C_j = \frac{M_k - M_j}{\left(N_T + \frac{3}{2}\right) \cdot \log_2\left(N_T + \frac{3}{2}\right) - \left(\frac{N_T}{\log_e 2}\right) - \frac{3}{2} \cdot \log_2(\frac{3}{2})}$$

383

384

385 Normalized $GE_{i,j}$ was calculated as: $GE_{i,j} = L_{i,j} + h_{i,j}$. Or overall,

386 $$GE_{i,j} = log_2\left[\left(T_{i,j} \times \frac{10^5}{\sum_{i=o}^{i=Nj} T_{i,j}}\right) + 1\right] + log_2[R_{i,j} + 1].\left[\frac{M_k - M_j}{\left(N_T + \frac{3}{2}\right) \cdot \log_2\left(N_T + \frac{3}{2}\right) - \left(\frac{N_T}{\log_e 2}\right) - \frac{3}{2} \cdot \log_2(\frac{3}{2})}\right]$$

387 ***RTAM2:*** N_j, $L_{i,j}$, $R_{i,j}$, M_j and M_k were calculated as for RTAM1. $G_{i,j}$ was defined as

388 $$G_{i,j} = T_{i,j} \times \frac{10^5}{\sum_{i=o}^{i=Nj} T_{i,j}}$$

389 and reflects the transcript count $T_{i,j}$ of gene i in cell j, corrected by scaling to a standard library
390 size (10^5). The RTAM2 adjustment variable, $h_{i,j}$, once again was derived to correct cellular M_j to
391 the constant M_k by adjusting gene expression on a non-linear $\log_2$ scale. However, the RTAM2
392 adjustment differs from RTAM1 in that the per-gene adjustment is derived from the gene's
393 expression intensity rather than its expression rank and was calculated as:
394

395 $$h_{i,j} = (M_k - M_j) \cdot \frac{[G_{1,j} - G_{i,j} + 1]}{\sum_{p=1}^{p=N_T}[G_{1,j} - G_{p,j} + 1]}$$

396

397 RTAM2 normalized gene expression, $GE_{i,j}$, was calculated as: $GE_{i,j} = L_{i,j} + h_{i,j}$.
398 Or,

399 $$GE_{i,j} = log_2[G_{i,j} + 1] + (M_k - M_j) \cdot \frac{[G_{1,j} - G_{i,j} + 1]}{\sum_{p=1}^{p=N_T}[G_{1,j} - G_{p,j} + 1]}$$

400
401 The derivation of RTAM1 and RTAM2 functions are described in the supplemental materials.
402

403 **sciCNV**

404 Single-cell CNV were inferred from the RNA counts of test cells by comparing these normal
405 control cells of the same lineage. In brief, test and control cells were co-normalized using
406 RTAM2. The combined $GE_{i,j}$ matrix was then truncated to informative genes (N_{CNV}) by
407 excluding genes expressed in <2% cells; and was re-sorted by gene location. *i was* redefined by
408 the rank location of informative genes, numbered from chromosomes 1→Y, p-arm→q-arm, i =
409 1, 2, 3, ⋯, N_{CNV}. To mitigate against data sparsity within individual cells, a moving average of
410 gene expression (maGE_i) was calculated across chromosome segments extending $\pm n/2$ genes
411 from gene i, where $n = N_{CNV}/(S\times50)$. The variable S defined the sharpness (default=1.0) of the
412 analysis resolution and was adjustable to offset sample data sparsity. Data from control cells was

merged into a single '*average cell*' with the expression of each gene calculated from its average across all control cells, including zero values, thus integrating the probability of gene detection. A moving average of gene expression across chromosome segments extending $\pm n/2$ genes from gene i was calculated for the merged control cell, $maGE_{i,control}$, mirroring that of test cells. A weighted disparity score, $W_{i,j}$, comparing gene expression in test cell j and the merged control cells was calculated as:

$$W_{i,j} = \frac{maGE_{i,j} - maGE_{i,control}}{maGE_{i,j} + maGE_{i,control} + \alpha}$$

To avoid division by zero, $\alpha = 0.00001$. $W_{i,j}$ was centred on zero in each cell by subtracting its median, and then linearized by the equation:

$$W_{i,j}^{lin} = 3.3222(W_{i,j})^4 + 5.6399(W_{i,j})^3 + 4.2189(W_{i,j})^2 + 3.8956W_{i,j}$$

$W_{i,j}^{lin}$ provided a first-pass estimate of single cell CNVs. To further enhance CNV detection, we deployed a second method of assessing gene expression disparity that was sensitive to small concordant expression variations in contiguous genes and insensitive to large single-gene variations. A moving average (*ma*) of gene expression, $U_{i,j}$, across chromosome segments was calculated for both test and control cells using the $W_{i,j}$ function above but employing a smaller window extending from $i \pm n/(5\times S)$ genes. $U_{i,j}$ was centered on zero in each cell by subtracting its median and then positive/negative values were digitized into +1/-1 values respectively. The digital values were summed cumulatively across the genome as a function of i, generating a digital moving score (*DMS)*.

$$maDMS_{i,j} = \begin{cases} maDMS_{i-1,j} + 1, & U_{i,j} > 0, \\ maDMS_{i-1,j} - 1, & U_{i,j} < 0, \\ maDMS_{i-1,j}, & U_{i,j} = 0. \end{cases}$$

439 A second digital moving score, $V_{i,j}$, was calculated as for $U_{i,j}$ but employed gene expression
440 values without average smoothing in order to obtain enhanced resolution for small CNV. The
441 non-moving average DMS (*nmaDMS*) was then calculated as:
442

443 $$nmaDMS_{i,j} = \begin{cases} nmaDMS_{i-1,j} + 1 - d_j, & V_{i,j} > 0, \\ nmaDMS_{i-1,j} - 1 - d_j, & V_{i,j} < 0, \\ nmaDMS_{i-1,j} - d_j, & V_{i,j} = 0. \end{cases}$$

444
445 The *nmaDMS* function in each cell *j* was adjusted to sum cumulatively to zero at the last gene
446 (it's distal terminus) by cumulatively subtracting d_j at each gene, where:
447

448 $$d_j = \frac{\left|\{V_{i,j} : V_{i,j} > 0\}\right| - \left|\{V_{i,j} : V_{i,j} < 0\}\right|}{|N_{CNV}|}$$

449
450 nma*DMS* was scaled to ma*DMS* by the ratio of their respective ranges, ($maDMS_{MAX}$ -
451 $maDMS_{MIN}$) / (nmaDMS_{MAX} – $nmaDMS_{MIN}$). maDMS and nmaDMS were then combined by
452 weighted average to provide the merged function, $DMS_{i,j}$, integrating both smooth low-resolution
453 and noisy high-resolution data. The gradient of this function, $dDMS_{i,j}/di$, measured over
454 chromosome segments $i \pm n/2$, provided a second measure of CNV.
455
456 The $W_{i,j}^{lin}$ and $dDMS_{i,j}$ CNV methods were combined by geometric mean. When the two methods
457 yielded conflicting polarity results, $sciCNV_{i,j} \rightarrow 0$. Using the mean $sciCNV_{i,j}$ result from all
458 clonal test cells a scale factor, *s*, was manually derived to ensure that CNV values optimally
459 centered on integers. For example, for MM samples, *s* was calculated so that clone mean sciCNV
460 for chromosome 13 deletion scaled to -1. *s* was then uniformly applied to all cells:
461

462 $$sciCNV_{i,j} = s.\sqrt[2]{W_{i,j}^{lin}.dDMS_{i,j}}$$

463
464 Heatmaps were generated using heatmap.3 R package. Cells were clustered iteratively to identify
465 subclones, using gene expression-based tSNE clustering and sciCNV-based Pearson correlation
466 coefficients. After subclones were identified, supervised clustering was applied. Stochastic noise

467 was reduced by standardizing each sciCNV profile against the median of its nearest neighbours.
468 To further reduce low-level sciCNV noise centered around zero, $|sciCNV_{i,j}|$ values less than
469 threshold, *T*, were set to zero. For tumors in which chromosome gains and losses were
470 substantially imbalanced (with respect to the numbers of genes involved) an optional sciCNV
471 baseline correction, *b,* was applied to correct the zero-set point and improve CNV detection. The
472 method is described further in the **supplemental materials**.
473
474 **Whole exome sequencing**
475 DNA was isolated from tumor plasma cells using a Micro kit (Qiagen) and then sheared into
476 150-200bp fragments using a Covaris E220. Exome libraries were prepared using a Hyper Prep
477 Kit (KAPA Biosystems), pre-capture PCR and SureSelectXT Human All Exon V5+UTR baits
478 (Agilent). Post-capture fragments were amplified by PCR, purified using AMPure-XP beads
479 (Beckman Coulter Genomics) and sequenced using paired-ends on an Illumina HiSeq. FASTQ
480 files were generated with Illumina bcl2fastq and reads were aligned to GRCh38 using BWA
481 v0.7.8[24] and SAMTOOLS v1.2[22].
482
483 **Correlation of sciCNV and WES-derived CNV**
484 MM199 scRNA-seq data was RTAM2 normalized and sciCNV profiles were calculated (S=1.4,
485 b=0.1) without a noise filter. MM199 WES was performed using 200ng DNA (>30,000
486 equivalent exomes) extracted from 1.9×10^6 pooled cells. WES-derived CNVs were identified
487 using DNAcopy R package v1.54.0. CNV estimates from the two methods were smoothed using
488 windows of identical genomic size and the datasets were paired by nearest genomic location.
489
490 **Fluorescence in-situ hybridization (FISH)**
491 Cells on cytospin slides were hybridized with Vysis LSI IGH/CCND1 (14q32:11q13),
492 IGH/FGFR3(14q32:4p16), D5S23, D5S721 SGn/CEP9 SA/CEP15 probes (Abbott Molecular) or
493 Cytocell Aquarius CKS1B/CDKN2C probes according to the manufacturer's instructions. At
494 least 100 cells were scored for each.
495

496 **Tumor CNV score**

497 For each sample, we identified cells with clonal CNVs by sciCNV and using these derived a
498 tumor-specific average sciCNV profile. The sciCNV profiles of all individual cells were then
499 assessed against the tumor average. The tumor score (TS) for cell j was summed cumulatively
500 across the genome as:

501 $$TS_j = \sum_{i=1}^{|\mathcal{G}|} TS_{i,j}$$

502 where $\mathcal{G}$ is the set of all genes in sciCNV matrix and

503 $$TS_{i,j} = \begin{cases} +\sqrt{sciCNV_{i,j}}, & sciCNV_{i,j} \times \mathbb{E}[sciCNV_i] > 0, \\ -\sqrt{sciCNV_{i,j}}, & sciCNV_{i,j} \times \mathbb{E}[sciCNV_i] > 0, \\ -\sqrt{sciCNV_{i,j}}, & \mathbb{E}[sciCNV_i] = 0. \end{cases}$$

504 $\mathbb{E}[sciCNV_i]$ is the average tumor cell $sciCNV_i$ for gene i.

505

506 **Immunoglobulin isotype score**

507 To evaluate if plasma cells belonged to the tumor clone, we scored their immunoglobulin heavy
508 and light chain (IGH, IGK, IGL) gene expression for restriction to the tumor isotype. Using
509 RTAM2 data, a threshold was identified for IGHM, IGHD, IGHG1/2, IGHA1/2, IGHE and for
510 IGKC and IGL(C2-7) which segregated non-expressing and expressing cells. The
511 immunoglobulin isotype score was calculated for each cell by summing the supra-threshold
512 expression of tumor immunoglobulin genes and subtracting the supra-threshold expression of
513 non-tumor immunoglobulin genes.

514

515 **Gene-set enrichment analysis**

516 Intra-clonal cells with divergent CNV were segregated by sciCNV. Paired subpopulations of
517 cells that diverged at a single CNV were then subsampled for cells that were matched for
518 normalized cellular transcriptome size. The matched cells were then compared using gene-set
519 enrichment analysis (GSEA v3.0)[19] of RTAM2-normalized transcriptomes.

520

521 **DATA & CODE AVAILABILITY**

522 The raw and processed scRNA-seq and WES data are deposited at NCBI GEO, GSE141299. The
523 microarray data (Affymetrix U133Plus2.0) for MM tumors with 1q-FISH status was obtained
524 from the NCBI GEO, GSE2658[22]. The code used in this study is available at
525 https://www.github.com/TiedemannLab/sciCNV.
526

REFERENCES

527 1 Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human
528 cancers. *Nature* **463**, 899-905, doi:10.1038/nature08822 (2010).
529 2 Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids*
530 *Res* **47**, D941-D947, doi:10.1093/nar/gky1015 (2019).
531 3 Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer
532 Aneuploidy. *Cancer Cell* **33**, 676-689 e673, doi:10.1016/j.ccell.2018.03.007 (2018).
533 4 Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated
534 genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33**, 285-289,
535 doi:10.1038/nbt.3129 (2015).
536 5 Han, K. Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of genomic DNA
537 and total RNA from single cells. *Genome Res* **28**, 75-87, doi:10.1101/gr.223263.117
538 (2018).
539 6 Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and
540 transcriptomes. *Nat Methods* **12**, 519-522, doi:10.1038/nmeth.3370 (2015).
541 7 Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele
542 analysis of single-cell RNA-seq data. *Genome Res* **28**, 1217-1227,
543 doi:10.1101/gr.228080.117 (2018).
544 8 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary
545 glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).
546 9 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-
547 cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501 (2016).
548 10 Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell
549 RNA sequencing data: challenges and opportunities. *Nat Methods* **14**, 565-571,
550 doi:10.1038/nmeth.4292 (2017).
551 11 Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA
552 sequencing data with many zero counts. *Genome Biol* **17**, 75, doi:10.1186/s13059-016-
553 0947-7 (2016).
554 12 Cole, M. B. *et al.* Performance Assessment and Selection of Normalization Procedures
555 for Single-Cell RNA-Seq. *bioRxiv*, doi:doi: https://doi.org/10.1101/235382 (2017).
556 13 Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in
557 single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145, doi:10.1038/nrg3833 (2015).
558 14 Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A Systematic Evaluation of
559 Single Cell RNA-Seq Analysis Pipelines: Library preparation and normalisation methods
560 have the biggest impact on the performance of scRNA-seq studies. *bioRxiv*, 583013,
561 doi:10.1101/583013 (2019).
562 15 Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression
563 estimation with read mapping uncertainty. *Bioinformatics* **26**, 493-500,
564 doi:10.1093/bioinformatics/btp692 (2010).

565 16 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-
566 seq data using regularized negative binomial regression. *Genome Biol* **20**, 296,
567 doi:10.1186/s13059-019-1874-1 (2019).
568 17 Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with
569 markers of immune evasion and with reduced response to immunotherapy. *Science* **355**,
570 doi:10.1126/science.aaf8399 (2017).
571 18 Misund, K. *et al.* MYC dysregulation in the progression of multiple myeloma. *Leukemia*,
572 doi:10.1038/s41375-019-0543-4 (2019).
573 19 Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation
574 are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-273,
575 doi:10.1038/ng1180 (2003).
576 20 Nikolsky, Y. *et al.* Genome-wide functional synergy between amplified and mutated
577 genes in human breast cancer. *Cancer Res* **68**, 9532-9540, doi:10.1158/0008-5472.CAN-
578 08-3082 (2008).
579 21 Peterson, T. R. *et al.* DEPTOR is an mTOR inhibitor frequently overexpressed in multiple
580 myeloma cells and required for their survival. *Cell* **137**, 873-886,
581 doi:10.1016/j.cell.2009.03.046 (2009).
582 22 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
583 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

584 **SUPPLEMENTARY INFORMATION:**

585 Supplementary methods, notes and figures can be found on-line.

586

587 **ACKNOWLEDGEMENTS:**

588 The authors thank the patients and physicians who made this study possible. They also thank N.
589 Winegarden, N. Khuu and G. Basi in the Princess Margaret Genomics Facility and Z. Lu in the
590 Princess Margaret Bioinformatics Core for technical assistance; and Drs. Gary Bader and Caleb
591 Stein for independent review and comments. This work was supported by funding from The
592 Princess Margaret Cancer Centre Foundation, the Terry Fox Foundation and the Canadian
593 Cancer Society Research Institute.

594

595 **Author Contribution**

596 A.M-S performed research and analyzed data. N.E., C.L-H. FISH, FACS and whole exome
597 sequencing, respectively. R.E.T. designed research, analyzed data and wrote the paper.

598

599 **Competing interests**

600 The authors declare that they have no competing interests.

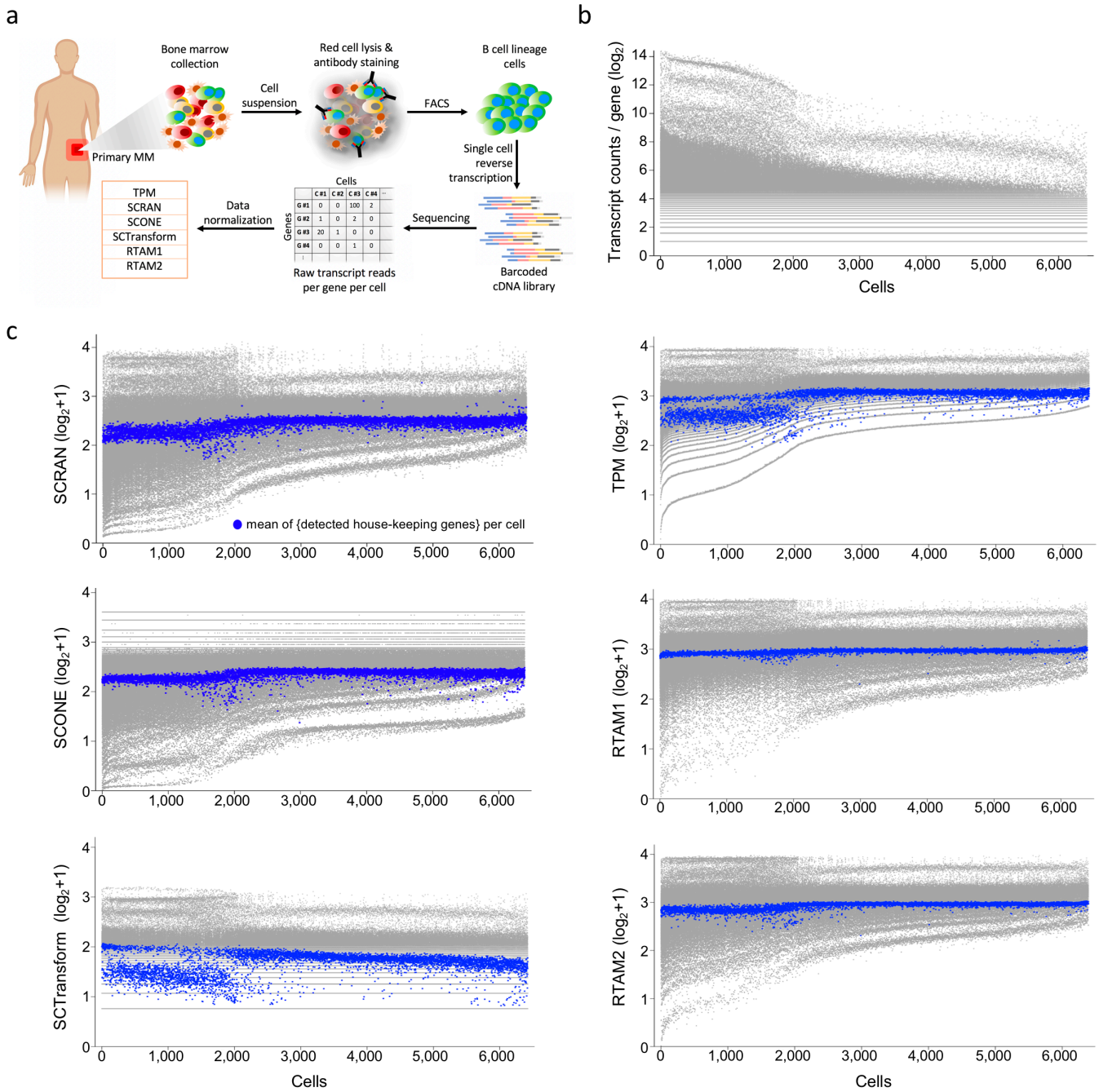

Figure 1. Comparison scRNA-seq normalization strategies, including RTAM1 and RTAM2 methods.

a. Overview of workflow showing bone marrow sampling, FACS sorting of B cell lineage cells, scRNA-seq and data normalization. MM, multiple myeloma. FACS, fluorescence activated cell sorting.

b. Plot of scRNA-seq data from >6,000 B cells and plasma cells, isolated from the bone marrow of a MM patient, depicting the raw (pre-normalized) transcript counts per gene per cell. Each dot represents an integer transcript count for one or more genes in a single cell; cells (columns) are ranked from left to right by their total transcript count.

c. The same data is shown following normalization using TPM, SCRAN, SCONE, SCTransform, RTAM1 or RTAM2 methods (and following log transformation). To compare the methods, the mean expression of a curated set of house-keeping genes (HKG) is plotted in each cell as a blue dot.

See also **supplemental figures S4-S20**.

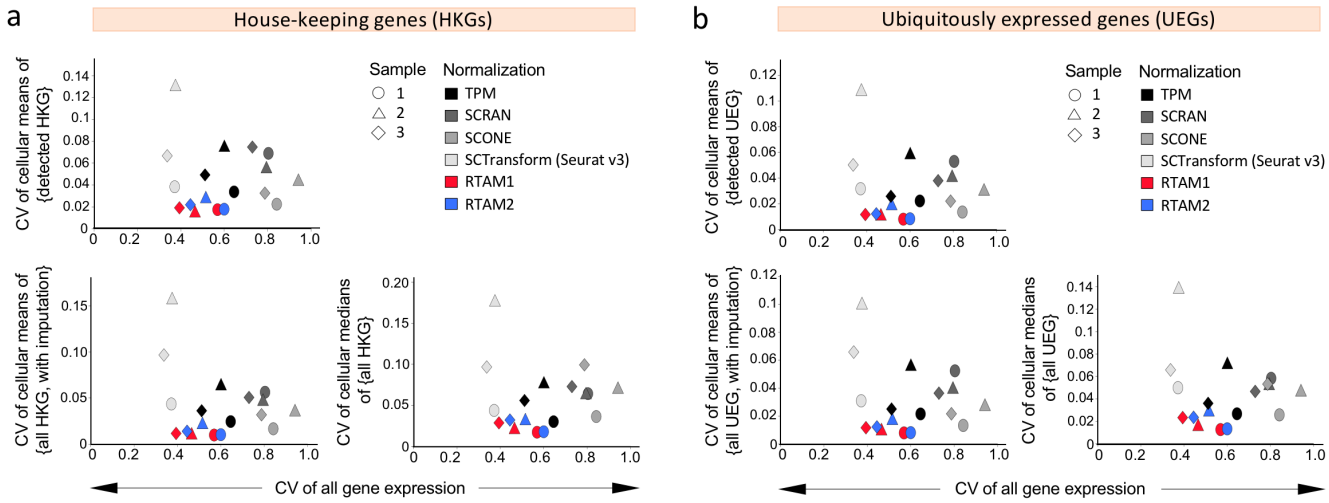

Figure 2. RTAM scRNA-seq normalization reduces cell to cell variation in average expression of house-keeping genes (HKG) and ubiquitously-expressed genes (UEG).

Alignment of single cell transcriptomes is compared using 6 different methods for scRNA-seq normalization.

a. The coefficient of variation (CV) across cells in the average expression of HKGs within each cell is shown for 3 scRNA-seq samples (containing >15,000 cells), normalized by 6 different methods. To avoid bias in the analysis, average expression was alternatively calculated as either the mean of the detected HKG (top panel), the mean of all HKG [with imputation of null dropout values] (bottom left panel), or the median of all HKG without imputation (right panel). As the various normalization methods expand or compress the distribution of the overall gene expression data to different extents, the CV of HKG averages is plotted against the CV of expression of all genes.

b. The coefficient of variation (CV) for all genes ubiquitously expressed (UEGs)(in >95% cells) was determined for the same patient samples. Once again, the average expression of UEG in each cell was alternatively calculated as either the mean of the detected UEG (top panel), the mean of all UEG imputing null dropouts (bottom left panel), or the median of all UEG without imputation (right panel).

A lower CV in the mean expression of HKG/UEG implies better alignment of average gene expression across cells. As the normalization methods expand or compress the distribution of the overall gene expression data to different extents, the CV of HKG/UEG averages (y-axis) is plotted against the CV of expression of all genes (x-axis).

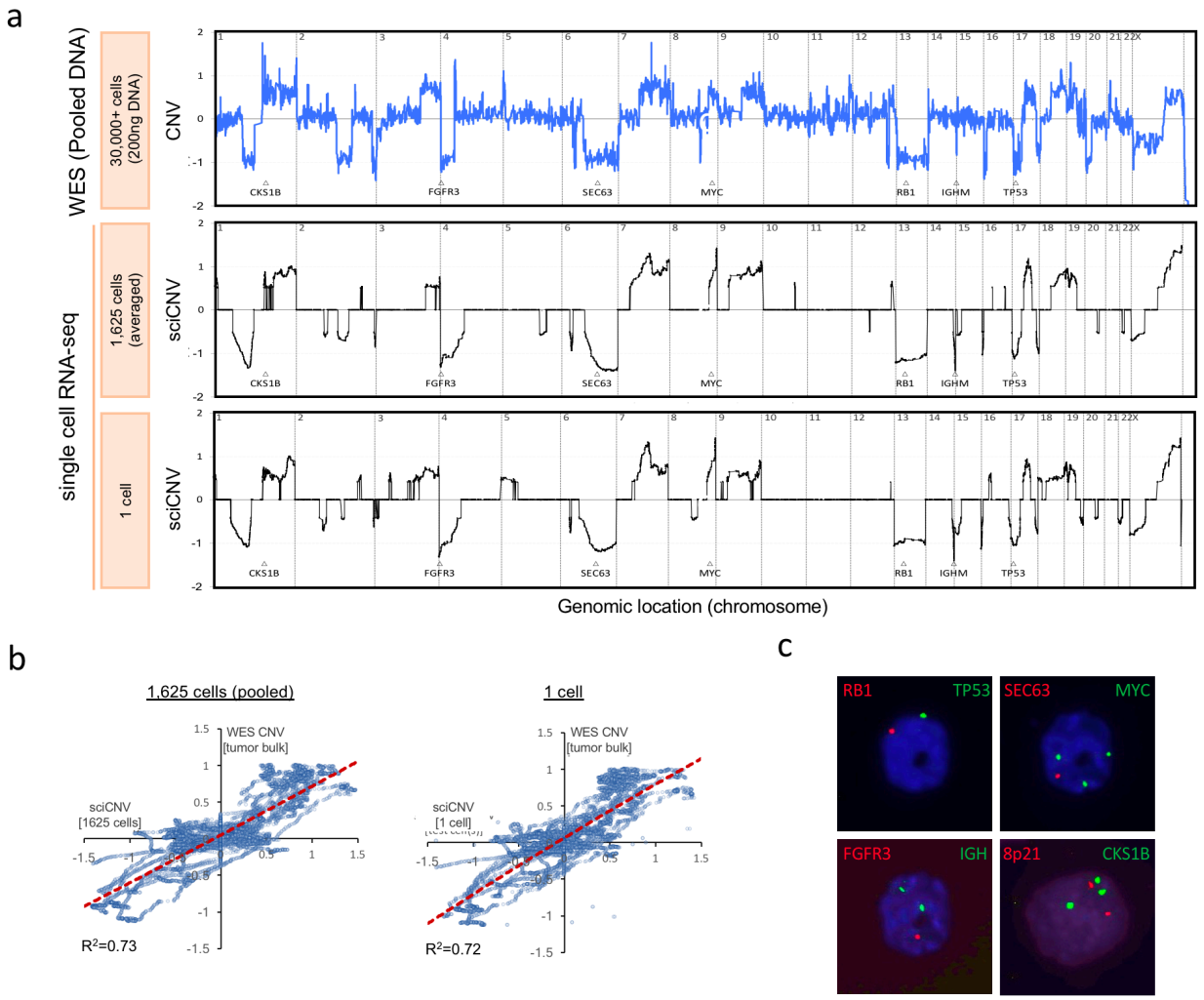

Figure 3. sciCNV can predict DNA copy number variations (CNV) in single cells from scRNA-seq data.

a. The CNV profiles of 1,625 pooled tumor cells (middle panel) or of a single tumor cell (lower panel) from sample MM199 were inferred from scRNA-seq using RTAM2/sciCNV and are shown compared with the bulk tumor CNV profile (upper panel), which was obtained by whole exome sequencing (WES) of DNA purified from 1.9×10^6 cells. Cells were isolated from bone marrow by FACS.

b. The scRNA-seq-based sciCNV profiles from (**a**) are correlated with the tumor bulk CNV profile derived from WES.

48 **c.** FISH verification of sciCNV copy number predictions, for the genes highlighted in **a.**,
49 showing 3 copies of CKS1B (1q21), 1 copy of FGFR3 (4p16), 1 copy of SEC63 (6q21), 2 copies
50 of PNOC (8p21), 3 copies of MYC (8q24), 1 copy of RB1 (13q14,) and 1 copy of TP53 (17P13),
51 as predicted. Brightness and contrast were adjusted during figure construction to enhance probe
52 visualization.

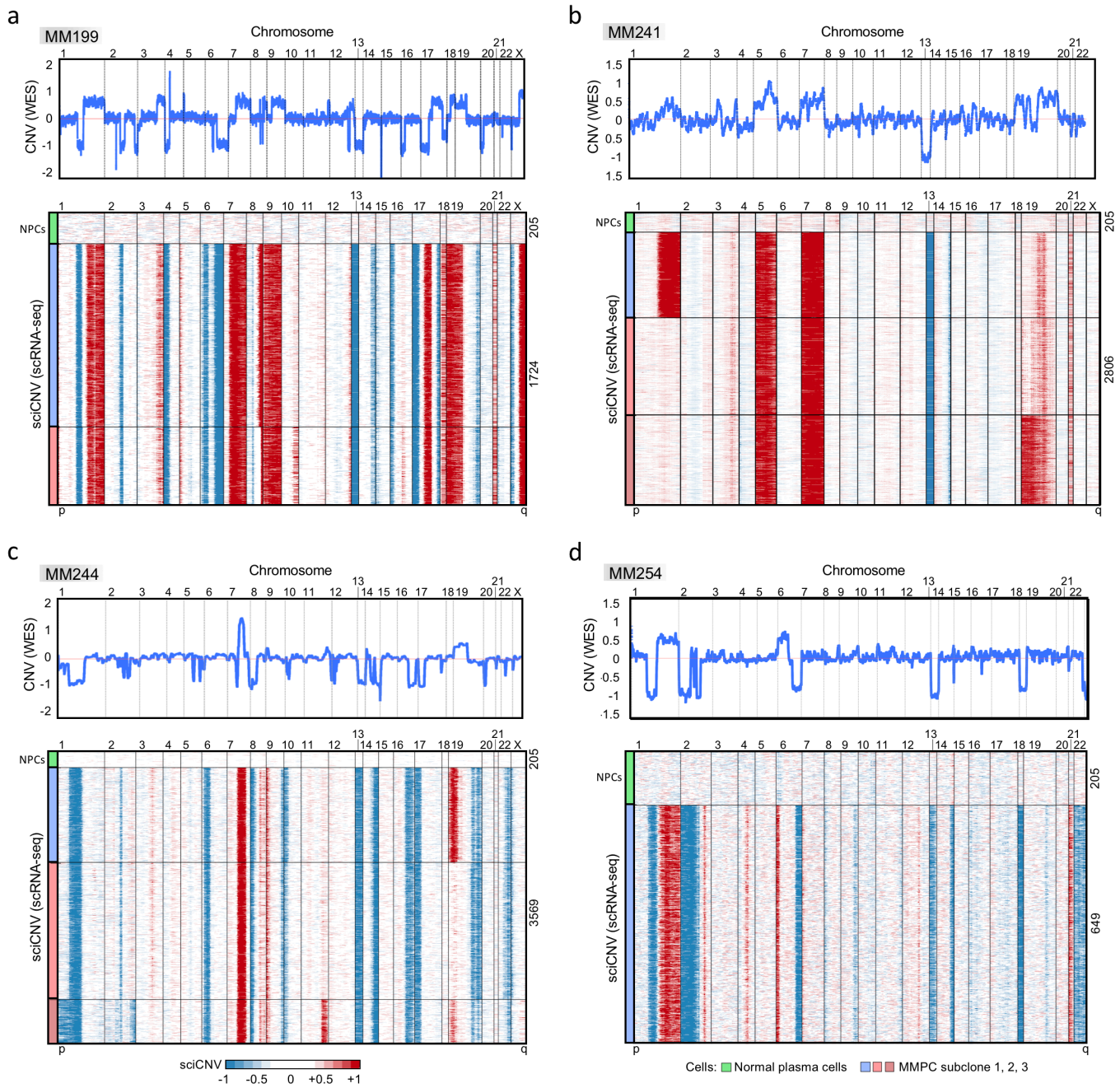

Figure 4. Using scRNA-seq data, sciCNV consistently infers single cell CNVs across thousands of cells that correlate with bulk tumor DNA-based CNV profiling

a.-d. The CNV profiles of thousands of individual plasma cells from MM patient samples MM199 (n=1724)(a), MM241(n=2806) (b), MM244 (n=3569)(c) and MM254 (n=649)(d) were inferred from scRNA-seq using RTAM2/sciCNV (lower panels) and are shown compared to the bulk tumor CNV profiles generated by DNA whole exome sequencing (upper panels). The sciCNV profiles of normal plasma cells (NPCs)(n=205) are shown as additional controls. For the

61 heatmaps revealing CNV profiles, at single cell resolution, the tumor cells in each sample are
62 grouped into inferred subclones (color bars at left) distinguished by divergent CNV.

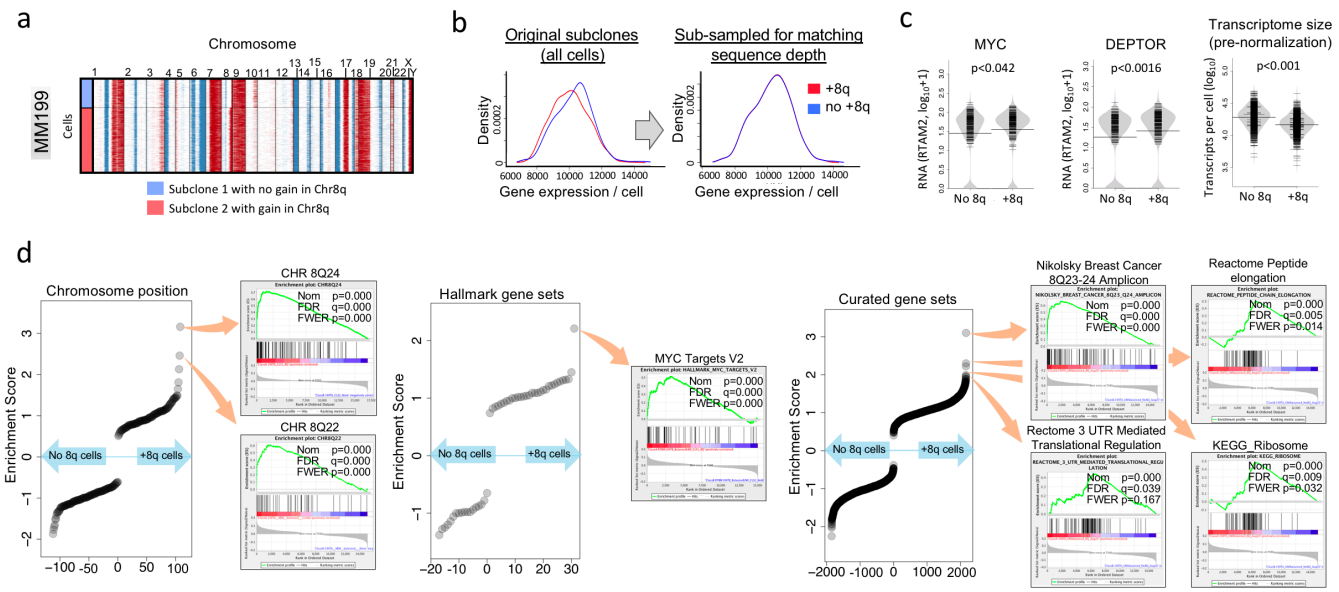

Figure 5. The transcriptional effects of +8q22-24 in MM cells, revealed by dual DNA CNV and RNA profiling of single cells using scRNA-seq and sciCNV.

a. Heatmaps showing the sciCNV profiles of MM199 cells, showing intra-clonal subpopulations (subclones) that diverge at +8q.

b. The transcriptome size distribution for cells in each subclones is shown. The original subclones (left) were sampled for subpopulations of cells with comparable transcriptome sizes (right panel), which were then compared in subsequent analyses.

c. Plots showing the mRNA expression of MYC or DEPTOR genes, located on chromosome 8q24, in MM199 cells by +8q status. Expression data is compressed by normalization (RTAM2) and log scale. The pre-normalized transcriptome sizes of subclonal MM cells from MM199 are also shown, demonstrating fewer RNA transcripts in cells with +8q. P-values were calculated by t-test.

d. Results of gene set enrichment analysis (GSEA) performed on MM199 intra-clonal subpopulations, comparing cells with or without +8q. GSEA of chromosome position gene sets (n=215) showed enrichment of chromosome 8q22-24 gene sets in cells identified by sciCNV at single cell resolution as containing +8q (left panels). Key results from GSEA of hallmark (n=49) and curated (n=3303) gene sets are shown in the middle and right panels; demonstrating upregulation of MYC target genes in MM199 +8q cells, and enriched expression of the gene sets: Nikolsky_Breast_Cancer_8Q23-24_Amplicon, KEGG_Ribosome, Reactome_Peptide Elongation and Reactome_3'_UTR_mediated translation_regulation.