Simultaneous Profiling of DNA Copy Number Variations and Transcriptional Programs in Single Cells using RNA-seq

Authors: Ali Madipour-Shirayeh[1], Natalie Erdmann[1], Chungyee Leung-Hagesteijn[1], Paola Neri[2], Ines Tagoug[1], Rodger E. Tiedemann[1,3*]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

[2]Arnie Charbonneau Cancer Institute, University of Calgary, Alberta, Canada

[3]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, multi-omics, multiple myeloma, RTAM, sciCNV

SUMMARY

Chromosome copy number variations (CNVs) are a near-universal feature of cancer however their specific effects on cellular function are poorly understood. Single-cell RNA sequencing (scRNA-seq) can reveal cellular gene expression however cannot directly link this to CNVs. Here we report scRNA-seq normalization methods that improve gene expression alignment between cells, increasing the sensitivity of scRNA-seq for CNV detection. We also report sciCNV, a tool for inferring CNVs from scRNA-seq. Together, these tools enable dual profiling of DNA and RNA in single cells. We apply these techniques to multiple myeloma (MM) and examine the cellular effects of cancer CNVs +8q23-24 and +1q21-44. Primary MM cells with +8q23-24 upregulate MYC, MYC-target genes, mRNA processing and protein synthesis; but also upregulate DEPTOR and have smaller transcriptomes. MM cells with +1q21-44 instead reconfigure translation and suppress unfolded protein stress whilst increasing proliferation, oxidative phosphorylation and MCL1. Overall, we provide tools that can enhance the analysis of scRNA-seq and help reveal the effects of cancer CNVs on cellular reprogramming.

Genomic CNVs are a pervasive feature of cancer. Copy number gains on chromosome arms 8q, 1q, 3q and 5p are amongst the most common karyotype abnormalities in human cancer, yet the action of these and other CNVs on the molecular processes within cancer cells remains poorly understood[1,2].

ScRNA-seq can reveal the transcription state of single cells, however it cannot directly relate this to DNA lesions. Although physical sequencing of both DNA and RNA within single cells has been reported[3-5], and should enable pairing of CNVs with their transcriptional outcomes, existing techniques provide profiling of only a few cells and thus afford only a limited view of the genomic and transcriptional heterogeneity within any cancer. Furthermore, while CNVs and gene expression can be profiled in separate populations of cells and computationally integrated[6], this may not recapitulate the biological state of individual cells.

DNA 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. 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 scRNA-seq normalization methods that reduce the influence of noise from lowly-expressed genes on single-cell transcriptome scale. These methods improve gene expression comparisons 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 tool for inferring CNVs in single cells from scRNA-seq. Together, these methods enable high-throughput profiling of both DNA copy number and RNA in the same cell, facilitating direct 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 such comparisons depends critically upon data normalization. As the best methods for normalizing scRNA-seq remain controversial, 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 cells belonging uniformly to the B cell lineage (n>15,000) (**figure 1a**). We examined a single lineage in order to minimize confounding biological variation between cells due to their ancestry. However, we deliberately generated data from a mix of both small quiescent B cells and large transformed 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 MM patient 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 three initial test samples is depicted in **figure 1b**. As shown, the distributions of transcript counts per gene varied significantly from cell to cell, reflecting differences in their cellular transcriptome sizes and demonstrating a clear need for normalization. The 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, 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 cells, suggesting significant limitations for scRNA-seq application. By comparison, SCRAN and SCONE produced superior alignments 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 loss of inter-cellular variation, particularly in highly-expressed genes. The expression of IGH or IGL genes, for example, a critical feature of plasma cells, was reduced by SCONE's quantile normalization into a virtual constant across cells (**supplementary figure S21**).

As each of these scRNA-seq normalization methods has limitations, we developed RTAM1 and -2. The RTAM approach originates from a consideration of the strengths and weaknesses of scRNA-seq. Whereas lowly expressed genes are detected within single cells with low resolution (due to low integer transcript counts) and show significant stochastic variation, highly expressed genes are robustly detected and show finer quantisation of variation relative to intensity. RTAM thus utilizes highly-expressed genes, whose expression is resolved with greater accuracy, to align cellular transcriptomes. Genes are ranked in each cell by their expression and the summed intensities of the top-ranked genes is standardized in log-space using unique non-linear cell- and gene-specific adjustments of gene expression determined either by cellular gene expression rank (RTAM1) or by gene expression intensity (RTAM2) (see **methods**).

Importantly, compared to TPM, SCTransform or SCRAN, both RTAM1 and RTAM2 reduce the cell-to-cell variance in the average (median or mean) expression of HKG and UEG sets (**figure 1c** and **supplementary figures S4-S20**). The coefficients of variation (CV) produced by each normalization method for the "average" expression of HKG or UEG in individual cells is shown in **figure 1d** and **supplemental figure S21a**, for 3 independent patient samples. As shown,

RTAM1 (red) and RTAM2 (blue) reduce variations in the average gene expression of single cells, even when this average expression is calculated by 3 different methods. By design, the RTAM methods also standardize the average expression of highly-expressed genes, and thus overall these methods produce superior alignments of cellular transcriptomes and of gene expression between cells. At the same time, both RTAM1 and RTAM2 maintain the original variability observed between cells in the expression of individual highly-transcribed genes, unlike the quantile normalization implemented by SCONE (**supplementary figure S21b-d**). Overall, therefore, the RTAM methods represent useful new strategies for normalizing scRNA-seq data that can enhance the accuracy of gene expression comparisons between cells.

Single-cell inferred chromosomal copy number variation: sciCNV

We next sought to develop a method for detecting single-cell chromosomal CNV from scRNA-seq, leveraging the enhanced normalization provided by RTAM to increase the sensitivity of single-cell transcriptomics for CNV detection. To optimize DNA copy number estimates from gene expression, and to mitigate against data sparsity in single cells, we developed a two-pronged approach, called sciCNV (described in the **supplemental methods**). Briefly, RTAM-normalized gene expression data from single cells was aligned with matching data from pooled control cells to develop expression disparity scores, which were averaged in a moving window defined by genomic location. Gene expression in the control cells was weighted according to the probability of gene detection, enhancing the comparison with single cell data, where signal dropout was common for many genes. In a parallel method, the expression disparity values were exchanged for binary values, which were summed cumulatively as a function of genomic location; the gradient of this function yielded a second estimate of CNV that was sensitive to small concordant expression variations in contiguous genes and that was insensitive to large single-gene variations. The CNV estimates of the two methods were combined by their geometric mean.

Figure 2 shows sciCNV applied to scRNA-seq data from primary MM cells. Significantly, the CNV profile of a single cell, inferred from its RNA, closely resembles the average CNV profile of $>10^4$ tumor bulk cells, derived from whole exome DNA sequencing (WES) (R^2=0.72) (**figure 2a-b**). The CNV predictions produced from a single cell by sciCNV were also validated at key

locations by FISH (**figure 2c**). Furthermore, examination of >1700 plasma cells from the same MM patient biopsy using sciCNV revealed that the tumor-specific CNV were robustly detected in all of the MMPC (**figure 2d**), despite biological and technical variations between the cells; and were not detected in normal plasma cells (NPC). Thus, sciCNV can utilize scRNA-seq to reveal CNVs in single cells. Moreover, it can distinguish cancer cells and normal cells on the basis of their CNV profile (**figures 2e-f**).

Identification of subclones and intra-clonal evolution using scRNA-seq

The detection of CNV with single cells from scRNA-seq data enables the identification of subclones and examination of intra-clonal evolution. Using scRNA-seq, RTAM2 and sciCNV we readily detected up to 7 subclones in primary MM samples comprising <4000 cells (**figure 3a-b**) and identified an average of 2-3 subclones per sample. Examination of the sciCNV profiles of the individual MM cells yielded evidence of both branching and linear intra-clonal evolution (**figure 3c-d**). In some tumors, marked divergence of two subclones from an inferred ancestral cell was evident, as in **figure 3a, c**; however, in the majority of MM samples examined the subclones diverged at only one or two loci.

Dissecting the effects of CNVs on gene expression: +8q23-24 in MM

Simultaneous profiling of both DNA copy number and RNA in the same cell should enable examination of the influence of CNVs on transcriptional programs. To explore this, we used sciCNV to screen MM patient bone marrow samples for tumor cells with +8q24. We sought to examine +8q24 as this is one of the most recurrent abnormalities in human cancer[1,17] and is known to target MYC[18], providing a benchmark for our analyses.

Using sciCNV, primary MM samples MM199 and MM244 were both found to contain subclonal gains of chromosome 8 encompassing 8q23-24 (**figure 3e**). Both samples also contained closely-related isogenic subclones without +8q. To facilitate gene set enrichment analyses (GSEA)[19] of the intra-tumor subclone pairs, these subclones were next subsampled to yield cellular subpopulations with matching transcriptome depth (**figure 3f**). This prevented subclone biases in total cellular gene expression from influencing specific gene-set detection. The gene expression of the intra-clonal subpopulations, representing isogenic cells with and without +8q23-24, with

matched transcriptome sizes, were then compared by GSEA using RTAM2-normalized data. From an analysis of 215 gene sets defined by chromosome location, +8q cells in both samples were strongly enriched for the gene-sets located at 8q23-24, with striking statistical confidence (p=0.000, q=0.000, FWER=0.000), compared to cells without +8q (**supplemental figure S22**). In contrast, no other genomic regions were significantly enriched. Thus, sciCNV accurately resolved single MM cells into intra-tumor subclones, isolating +8q23-24 as a unique variation distinguishing these.

We next used GSEA to explore the influence of +8q23-24 on cellular programming. As expected, +8q cells from both MM199 and MM244 samples showed increased MYC expression ($p<0.05$) compared to sibling cells without +8q (**figure 3g**). Surprisingly, however, only +8q cells from MM199 showed a broad increase in MYC target genes (p=0.000, q=0.000, FWER=0.000). Canonical MYC signature genes were not upregulated in MM244 +8q cells (p=0.767, FWER=1.0)(**figure 3h**). Nevertheless, from an analysis of 3303 curated gene sets, +8q cells from both MM199 and MM244 tumors showed similar upregulation of gene-sets encoding the machinery of mRNA translation and protein synthesis, including specifically genes involved in 3'UTR-mediated mRNA translation regulation (enrichment rank 5/3303 in MM199 and 9/3303 in MM244), ribosome biogenesis (enrichment rank 4/3303 in both) and peptide chain elongation (enrichment rank 3/3303 and 6/3303)(**figure 3h, supplemental figure S22**), potentially representing a more restricted MYC response. Conspicuously, these transcriptional effects of +8q23-24 in MM cells were remarkably close to those of +8q23-24 in breast cancer (FWER p=0.000, enrichment rank 1-2/3303 in both tumors), and this similarity was strong even when MYC hallmark genes were not increased (**figure 3h, supplemental figure S22**). Thus, +8q23-24 induces analogous gene expression changes across malignancies; and these analogous effects are not dependent on broadly-defined MYC-target genes but instead map to the specific upregulation of mRNA translation and protein synthesis.

The cellular re-programming induced by +8q23-24 might be expected to promote significant increases in gene expression and in cell mass. Notably, however, in the MM samples examined the mTOR-interacting gene, DEPTOR, located at 8q24, was also upregulated in +8q cells (**figure 3g**), and likely serves to counter increases in cell size, as previously reported[20]. Indeed, from our

examination of +8q at a single cell level we uniquely observed that the transcriptome sizes of +8q cells were in fact mildly reduced, compared to sibling cells without the CNV ($p<0.001$)(**figure 3i**). Thus, from a single-cell analysis of +8q23-24 it appears that this CNV acts to boost protein synthesis capacity (ribosomes, translation) without increasing cellular transcriptome size. Ultimately this may lead to enhanced expression of MYC-target genes as proteins in some cancers, but may also serve more broadly to improve the dynamics of protein synthesis and reduce the lag-time required to respond to gene expression changes, potentially enhancing cellular adaptability.

The effects of +1q on MM cells

Like +8q23-24, gain of chromosome 1q is highly recurrent in human cancer and is present in >30% of clinical tumors[1,17] Although rare in MM precursor disease, the prevalence of +1q increases significantly in symptomatic MM, more so than any other copy number gain.[18,21] In newly diagnosed MM, +1q is found in 35% of cases and is associated with poor prognosis.[22-29] Despite this, the effects of +1q on cancer cell biology remain poorly understood.

To examine the cellular effects of +1q, we screened MM patient bone marrows (n=30) by scRNA-seq and RTAM2/sciCNV, and identified ten tumors with +1q (**figure 4a**), including three (MM241, MM244 and MM379) containing synchronous subclones with and without the CNV (**figure 4b**). Although these tumor samples contained 2-6 subclones by sciCNV profiling, the subclones were only partially segregated by expression-based clustering (**supplementary figure S23**).

By GSEA, +1q cells in MM241 showed significant enrichment for all 10 chromosome position gene-sets located at 1q21-1q44 (p=0.000, FDR q<0.005, FWER p=0.000-0.058), while MM244 and MM379 +1q cells were correspondingly enriched for gene-sets located at 1q23-1q32 (p=0.000, $q\leq0.004$, FWER≤0.019) or 1q22-1q42 ($p\leq0.004$, $q\leq0.03$, FWER≤0.024; 1q23 FWER=0.359)(**Figure 5a-b** and **supplementary figures S24-26**). No other genomic regions were significantly enriched, confirming that the intra-clonal +1q subpopulations identified by sciCNV were uniquely divergent at this locus alone.

We next examined the influence of +1q on transcriptional programs in MM241, MM244 and MM379. Remarkably, the +1q cells in all three tumors showed significant reductions in the unfolded protein response (UPR) compared to their sibling cells lacking +1q ($p<0.003$, FDR≤0.015, FWER≤0.028), suggesting that +1q acts consistently in MM to reduce endoplasmic reticulum (ER) stress (**figure 5b** and **supplementary figures 27-29**). This effect of +1q on the UPR has not previously been reported, though is likely highly advantageous to MM cells, which are professional secretor cells burdened by high proteotoxic stress. In MM241, with the largest +1q CNV, UPR genes EIF4EBP1, EIF4A2, DDIT4, ATF4, ERN1, XBP1 and CEBPB were amongst the genes most downregulated in +1q cells (**figure 5c**). In contrast, ATF6, UAP1 and PSMD4 were incongruously upregulated, likely as result of their location within the 1q gain. With respect to mechanism, we observed that the 1q24 gene EEF1AKNMT, which selectively enhances protein translation in a codon-specific manner[30] to support oncogenic growth[31], was increased in all three +1q subclones, as was TIPRL, which regulates the mTORC1 pathway by inhibiting PP2A and sustaining phosphorylation of EIF4EBP1 and RPS6KB1. In contrast, EIF4A1 or EIF4A2, which jointly promote EIF4E-dependent translation (ET), were reduced, as was the ET-repressor EIF4EBP1 (**figure 5d**). Thus +1q induces complex alterations of translation and of the mTORC pathway that likely influence misfolded protein load. Expression of UAP1 and/or COPA from 1q23 may further alleviate ER stress[32,33].

Additional +1q effects were observed. Mitochondrial oxidative phosphorylation (OxPhos) and reactive-oxygen gene sets were enriched in MM241 +1q cells, likely driven by the increased expression of COX20, NDUFS2, SDHC, MRPS14 and MRPS21 from 1q21-44 (**figure 5b-c**). However, similar metabolic signatures were not observed in MM244 or MM379, perhaps because MRPS21 (1q21.2) falls outside of the +1q CNV in these later samples, or because enhanced NF-κB signaling may also be required for OxPhos augmentation[34] and was observed only in the MM241 subclone (**supplementary figure S27**), associated with TNFRSF13B over-expression (**figure 5c**).

Both MM244 and MM379 also showed significant enrichment of E2F, G2M and mitosis programs in +1q cells ($p=0.000$, FDR=0.000, FWER≤0.001) (**figure 5b**) and small increases in cycling cells in G2/M (**figure 5e**), consistent with increased proliferation. However, no increase

in proliferation was observed in MM241 +1q cells, indicating that 1q-induced proliferation requires a permissive cellular context. Although CKS1B has been proposed to be mechanistic in +1q-induced proliferation[22,35], we observed no increase in CKS1B in two of the three +1q subclones examined (**figure 5f**), indicating that alternative mechanisms likely drive cell cycling. Overexpression of EEF1AKNMT[31], increased oxidative phosphorylation and reductions in the UPR, may instead contribute to the enhanced proliferation of +1q cells.

MCL1, a critical anti-apoptosis gene for MMPC[36,37] located at 1q21.2, was also increased 1.45-fold ($p<10^{-9}$) in +1q cells from MM241 (**figure 5f**) in direct proportion to 1q copy number. MCL1 was not however upregulated in either MM244 or MM379, whose 1q gains narrowly excluded the MCL1 locus. Increased MCL1 and apoptotic threshold thus represents an additional function of +1q that may further increase cancer cell aggressiveness.

A summary of these cellular effects of +1q21-44 in MM is shown in **figure 6a**.

Comparison of intra-tumor and inter-tumor CNV studies

We next compared our intra-tumor studies (**figure 6b)** with a traditional inter-tumor study designed to identify the biological role of +1q (**figure 6c**). To perform the inter-tumor study, we examined microarray data from a large published series of MM tumor samples (n=532) characterized by +1q FISH[22] (**supplementary Figures S30-32**). As expected, the MM samples with 1q21 gain by FISH showed enrichment by GSEA for chromosomal position gene sets located at 1q21-44. However, the same samples also showed enrichment for gene-sets located on chromosome 1p22, 13q22, 11q13, 11q22, 5q14, 8q24 and Xq28, compared to tumors without +1q, undermining the value of this cohort for isolating gene expression changes attributable to +1q (**figure 6c**). The samples defined by +1q FISH were also biased towards distinctive MM subtypes, as the +1q cohort included more tumors with t(4;14) while the control samples included more tumors with t(11;14) or hyperdiploidy. Consequently, the utility of these cohorts for the isolation effects specifically attributable to +1q was undermined. GSEA of the cohorts yielded an overabundance of putative +1q-associations whose attribution to +1q or to confounding CNVs or biases in MM subtype was unclear (**figure 6c**).

Conspicuously, both intra- and inter-tumor studies identified the UPR as a significant +1q co-variant in MM. Strikingly, however, the direction of association differed between the studies, suggesting an error in one of the approaches. Notably, whereas dual profiling of DNA and RNA in single cells enables direct matching of a CNV with its effects on gene expression (**figure 6d**), inter-tumor studies must instead infer associations between CNVs and gene expression from their correlation across unrelated tumors, which can lead to erroneous conclusions as demonstrated in **figure 6e**. Thus, single cell studies of intra-tumor heterogeneity can better isolate CNV-specific effects than traditional multi-tumor bulk profiling studies and may reveal the cellular effects of CNVs with greater accuracy.

DISCUSSION

CNVs are critical drivers of cancer biology yet their specific effects on cellular processes remain poorly understood. Here, we report the dual profiling of DNA copy number and RNA within the same cells, using scRNA-seq, and leverage this to explore the effect of CNVs on gene expression. To capture intra-tumor heterogeneity, we profile the RNA and CNVs or thousands of cells per sample. Using these new techniques, we examine the transcriptional effects of copy number gains of chromosome regions 8q23-24 and 1q21-44, representing two of the most common CNVs in human cancer. We show that these lesions induce critical reprogramming of cancer cells that can explain their influence on clinical disease.

Chromosome +1q is the most common adverse CNV in MM. We demonstrate that +1q causes multiple effects on MM cells including a reduction in the unfolded protein response, which likely results from 1q-associated reconfiguration of translation and from changes in the mTOR pathway. In addition, we demonstrate that primary MM cells with +1q show enhanced oncogenic growth, oxidative phosphorylation and MCL1 expression. Significantly, these specific reprogramming effects may explain the inferior disease control achieved by MM patients with tumors harboring this abnormality, following standard of care therapies[22,26-28,35,38,39]. Thus, the suppression of unfolded protein stress in +1q MM cells may counteract the activity of proteasome inhibitors[26-28], which induce cytotoxicity via ER stress[40,41]. Similarly, the upregulation of MCL1 in cells with +1q21 may counteract treatment-induced apoptosis. And

cellular proliferation, which may be induced by 1q-mediated upregulation of EEF1AKNMT, or by UPR reduction, may further contribute to early disease recurrence.

We demonstrate that the transcriptional effects of +8q23-24 are remarkably similar in MM and breast cancer (FWER p=0.000), irrespective of whether or not hallmark MYC target genes are increased (**figure 3h**). Although +8q23-24 can upregulate the expression of a broad spectrum of MYC target genes, we demonstrate that the transcriptomes of MM cells with +8q are in fact smaller than those of cells lacking +8q, at least in the samples examined by us. Significantly, we demonstrate that a consistent function of +8q23-24 is the upregulation of gene sets involved in mRNA translation, ribosomal biogenesis and peptide elongation. Thus +8q23-24 selectively enhances protein synthesis capacity, without increasing transcriptome size. We propose that this may improve the dynamics of proteome reconfiguration following gene expression changes; and that this may enhance the malleability of cancer cells to environmental challenges.

We show here that the study of CNVs via single-cell transcriptomics offers a number of advantages. As intra-clonal cells that diverge at a single CNV are virtually isogenic, any consistent divergence in their gene expression can be precisely matched to the subclonal CNV. Furthermore, as the test and control cells are present within the same sample, differences in gene expression due to the microenvironment, clinical factors or due to sample processing are minimized. Inter-tumor cohort studies instead rely upon the identification of correlations between CNVs and gene expression across unrelated samples, and suffer from the substantial additional genetic and clinical heterogeneity that exists between samples. As a result of these limitations, the effects of most cancer CNVs on gene expression remain poorly understood. Fortunately, the compelling benefits of intra-clonal studies suggest that a new era of cancer genomics is emerging in which the precise effects of all cancer CNVs on cellular programming can be determined at the single-cell level. This important knowledge is critical for understanding cancer and for advancing therapeutic strategies that seek to address the foundations of this disease.

REFERENCES

1 Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947, doi:10.1093/nar/gky1015 (2019).

2 Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676-689 e673, doi:10.1016/j.ccell.2018.03.007 (2018).

3 Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* **33**, 285-289, doi:10.1038/nbt.3129 (2015).

4 Han, K. Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res* **28**, 75-87, doi:10.1101/gr.223263.117 (2018).

5 Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519-522, doi:10.1038/nmeth.3370 (2015).

6 Campbell, K. R. *et al.* clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol* **20**, 54, doi:10.1186/s13059-019-1645-z (2019).

7 Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res* **28**, 1217-1227, doi:10.1101/gr.228080.117 (2018).

8 Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401, doi:10.1126/science.1254257 (2014).

9 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501 (2016).

10 Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* **14**, 565-571, doi:10.1038/nmeth.4292 (2017).

11 Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* **17**, 75, doi:10.1186/s13059-016-0947-7 (2016).

12 Cole, M. B. *et al.* Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *bioRxiv*, doi:doi: https://doi.org/10.1101/235382 (2017).

13 Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145, doi:10.1038/nrg3833 (2015).

14 Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines: Library preparation and normalisation methods have the biggest impact on the performance of scRNA-seq studies. *bioRxiv*, 583013, doi:10.1101/583013 (2019).

15 Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493-500, doi:10.1093/bioinformatics/btp692 (2010).

16 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296, doi:10.1186/s13059-019-1874-1 (2019).

17 Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, doi:10.1126/science.aaf8399 (2017).

18 Misund, K. *et al.* MYC dysregulation in the progression of multiple myeloma. *Leukemia*, doi:10.1038/s41375-019-0543-4 (2019).

19 Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-273, doi:10.1038/ng1180 (2003).

20 Peterson, T. R. *et al.* DEPTOR is an mTOR inhibitor frequently overexpressed in multiple myeloma cells and required for their survival. *Cell* **137**, 873-886, doi:10.1016/j.cell.2009.03.046 (2009).

21 Mikulasova, A. *et al.* Genomewide profiling of copy-number alteration in monoclonal gammopathy of undetermined significance. *Eur J Haematol* **97**, 568-575, doi:10.1111/ejh.12774 (2016).

22 Shaughnessy, J. D., Jr. *et al.* A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276-2284, doi:10.1182/blood-2006-07-038430 (2007).

23 Avet-Loiseau, H. *et al.* Long-term analysis of the IFM 99 trials for myeloma: cytogenetic abnormalities [t(4;14), del(17p), 1q gains] play a major role in defining long-term survival. *J Clin Oncol* **30**, 1949-1952, doi:10.1200/JCO.2011.36.5726 (2012).

24 Neben, K. *et al.* Progression in smoldering myeloma is independently determined by the chromosomal abnormalities del(17p), t(4;14), gain 1q, hyperdiploidy, and tumor load. *J Clin Oncol* **31**, 4325-4332, doi:10.1200/JCO.2012.48.4923 (2013).

25 Walker, B. A. *et al.* A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood* **116**, e56-65, doi:10.1182/blood-2010-04-279596 (2010).

26 An, G. *et al.* Chromosome 1q21 gains confer inferior outcomes in multiple myeloma treated with bortezomib but copy number variation and percentage of plasma cells involved have no additional prognostic value. *Haematologica* **99**, 353-359, doi:10.3324/haematol.2013.088211 (2014).

27 Yu, W. *et al.* The amplification of 1q21 is an adverse prognostic factor in patients with multiple myeloma in a Chinese population. *Onco Targets Ther* **9**, 295-302, doi:10.2147/OTT.S95381 (2016).

28 Mai, E. K. *et al.* Phase III trial of bortezomib, cyclophosphamide and dexamethasone (VCD) versus bortezomib, doxorubicin and dexamethasone (PAd) in newly diagnosed myeloma. *Leukemia* **29**, 1721-1729, doi:10.1038/leu.2015.80 (2015).

29 Hanamura, I. *et al.* Frequent gain of chromosome band 1q21 in plasma-cell dyscrasias detected by fluorescence in situ hybridization: incidence increases from MGUS to relapsed myeloma and is related to prognosis and disease progression following tandem stem-cell transplantation. *Blood* **108**, 1724-1732, doi:10.1182/blood-2006-03-009910 (2006).

30 Jakobsson, M. E. *et al.* The dual methyltransferase METTL13 targets N terminus and Lys55 of eEF1A and modulates codon-specific translation rates. *Nat Commun* **9**, 3411, doi:10.1038/s41467-018-05646-y (2018).

31 Liu, S. *et al.* METTL13 Methylation of eEF1A Increases Translational Output to Promote Tumorigenesis. *Cell* **176**, 491-504 e421, doi:10.1016/j.cell.2018.11.038 (2019).

32 Itkonen, H. M. *et al.* UAP1 is overexpressed in prostate cancer and is protective against inhibitors of N-linked glycosylation. *Oncogene* **34**, 3744-3750, doi:10.1038/onc.2014.307 (2015).

33 Watkin, L. B. *et al.* COPA mutations impair ER-Golgi transport and cause hereditary autoimmune-mediated lung disease and arthritis. *Nature Genetics* **47**, 654 (2015).

34 Mauro, C. *et al.* NF-kappaB controls energy homeostasis and metabolic adaptation by upregulating mitochondrial respiration. *Nat Cell Biol* **13**, 1272-1279, doi:10.1038/ncb2324 (2011).

35 Zhan, F. *et al.* CKS1B, overexpressed in aggressive disease, regulates multiple myeloma growth and survival through SKP2- and p27Kip1-dependent and -independent mechanisms. *Blood* **109**, 4995-5001, doi:10.1182/blood-2006-07-038703 (2007).

36 Tiedemann, R. E. *et al.* Identification of molecular vulnerabilities in human multiple myeloma cells by RNA interference lethality screening of the druggable genome. *Cancer Res* **72**, 757-768, doi:0008-5472.CAN-11-2781 [pii]. (2012).

37 Derenne, S. *et al.* Antisense strategy shows that Mcl-1 rather than Bcl-2 or Bcl-x(L) is an essential survival protein of human myeloma cells. *Blood* **100**, 194-199 (2002).

38 Soriano, G. P. *et al.* Proteasome inhibitor-adapted myeloma cells are largely independent from proteasome activity and show complex proteomic changes, in particular in redox and energy metabolism. *Leukemia* **30**, 2198-2207, doi:10.1038/leu.2016.102 (2016).

39 Morales, A. A. *et al.* Distribution of Bim determines Mcl-1 dependence or codependence with Bcl-xL/Bcl-2 in Mcl-1-expressing myeloma cells. *Blood* **118**, 1329-1339, doi:10.1182/blood-2011-01-327197 (2011).

40 Obeng, E. A. *et al.* Proteasome inhibitors induce a terminal unfolded protein response in multiple myeloma cells. *Blood* **107**, 4907-4916, doi:10.1182/blood-2005-08-3531 (2006).

41 Bianchi, G. *et al.* The proteasome load versus capacity balance determines apoptotic sensitivity of multiple myeloma cells to proteasome inhibition. *Blood* **113**, 3040-3049, doi:10.1182/blood-2008-08-172734 (2009).

42 Ledergor, G. *et al.* Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat Med* **24**, 1867-1876, doi:10.1038/s41591-018-0269-2 (2018).

43 Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108**, 2020-2028, doi:10.1182/blood-2005-11-013458 (2006).

SUPPLEMENTARY INFORMATION:

Methods and supplementary figures can be found on-line.

ACKNOWLEDGEMENTS:

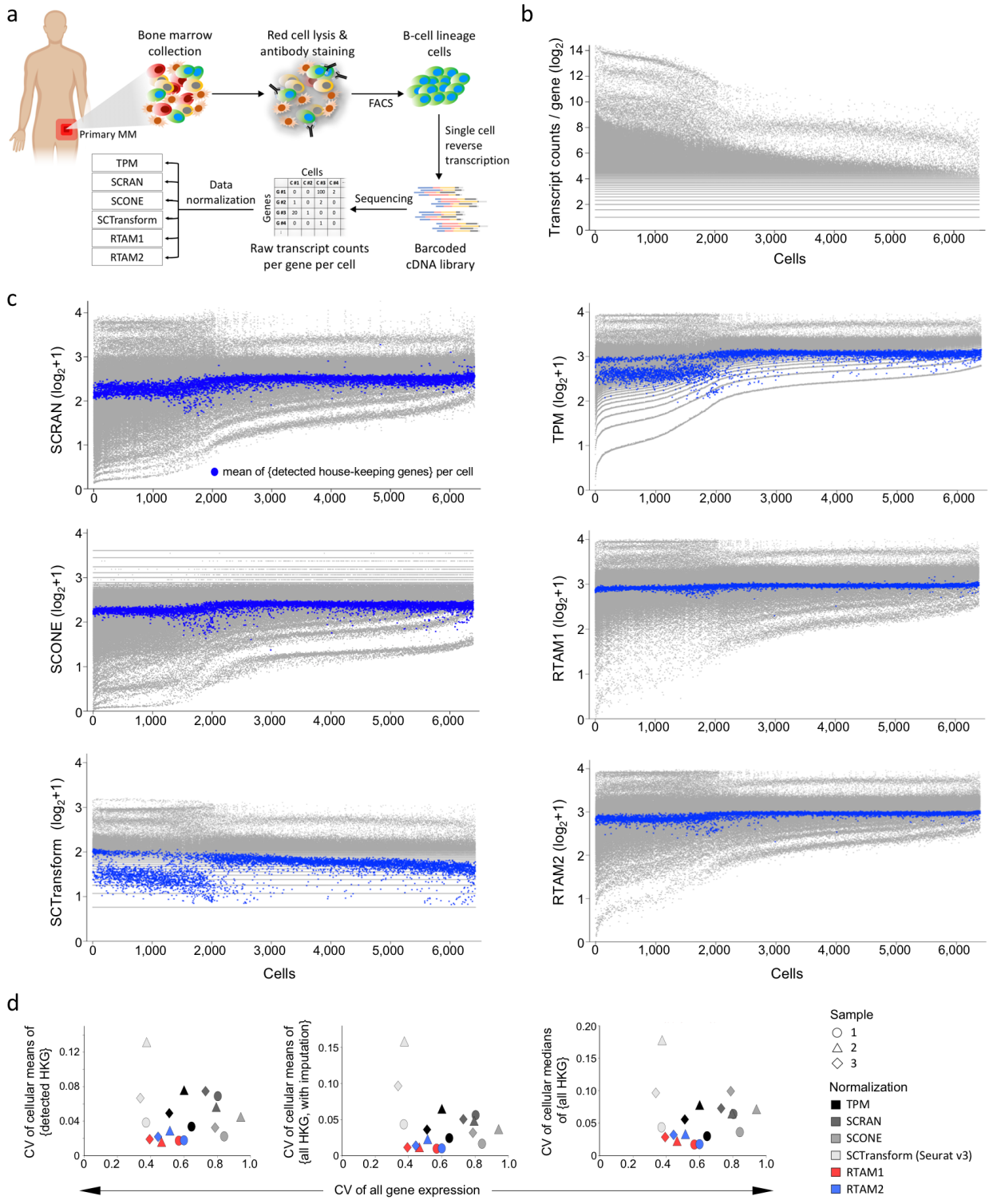
The authors thank the patients, their families and the physicians who made this study possible. They also thank N. Winegarden, N. Khuu and G. Basi in the Princess Margaret Genomics Facility and Z. Lu in the Princess Margaret Bioinformatics Core for technical assistance; and Drs. Gary Bader and Caleb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

Author Contribution

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

Competing interests

The authors declare that they have no competing interests.



1 **Figure 1. Comparison of scRNA-seq normalization strategies**

2 **a.** Overview of workflow. MM, multiple myeloma. FACS, fluorescence activated cell sorting. **b.**

3 Plot of scRNA-seq data from >6,000 cells of B cell lineage, 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 (blue) of a curated set of house-keeping genes (HKG) is plotted in each cell, omitting genes with zero values due to non-expression or detection "drop-out". **d.** The coefficient of variation (CV) across cells in the average expression of HKGs within each cell is shown for 3 patient samples containing >15,000 cells. The average HKG expression in each cell was calculated in 3 different ways as either the mean of the detected HKG (left panel), the mean of all HKG [with imputation of null "dropout" values] (middle 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.

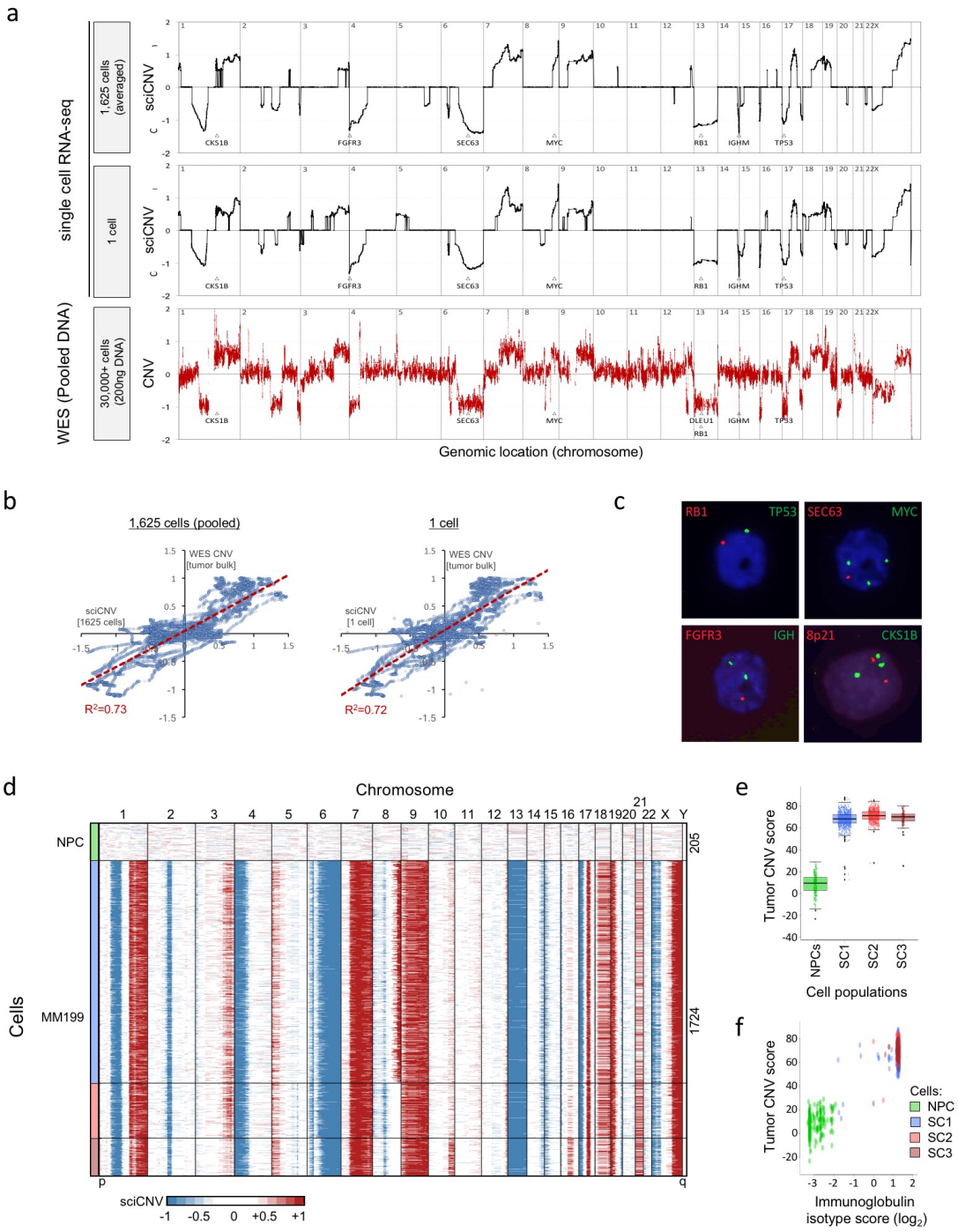


Figure 2. Single cell inferred chromosomal copy number variation (sciCNV)

a. The inferred CNV profiles of 1,625 pooled MM cells (top panel) or of a single MM cell (middle panel) were calculated from scRNA-seq data using RTAM2/sciCNV and are shown compared with the CNV profile of bulk tumor cells (lower panel), which was determined by whole exome sequencing (WES) of 200ng DNA (representing >30,000 complete exomes)

purified from $1.9x10^6$ cells. Cells were isolated from bone marrow by FACS. **b.** Correlation of the scRNA-seq sciCNV profiles from (a) with the tumor bulk CNV profile derived from WES. For the correlation, the CNV results from sciCNV and WES were paired by genomic location and averaged over similar chromosomal segment lengths; sciCNV results were generated without a noise cut-off filter. **c.** FISH was also used to verify sciCNV–derived copy number predictions, focusing on the genes highlighted in **a.**; this showed 3 copies of CKS1B (1q21), 1 copy of FGFR3 (4p16), 1 copy of SEC63 (6q21), 2 copies of PNOC (8p21), 3 copies of MYC (8q24), 1 copy of RB1 (13q14,) and 1 copy of TP53 (17P13) in accordance with sciCNV predictions derived from the RNA of a single cell. Brightness and contrast were adjusted during figure construction to enhance probe visualization. **d.** Heatmap showing chromosome copy number gains (red) and losses (blue) in individual multiple myeloma plasma cells (MMPC, n=1724), inferred from scRNA-seq using sciCNV. The MMPC are grouped into subclones (coloured bars at left) and their CNVs are compared with that of normal plasma cells (NPC, n=205, green bar) from a control sample. **e.** Identification of malignant cells using scRNA-seq and sciCNV. The tumor plasma cells in subclones (SC) 1-3 in **d.** were distinguishable from NPC on the basis of the similarity of their individual sciCNV profiles to the mean tumor clone sciCNV profile, calculated as a 'tumor CNV score'. **f.** Validation of cancer cell identification by the tumor CNV score. The tumor CNV scores for single cells are shown plotted again a cellular immunoglobulin-isotype score, derived to distinguish cells expressing immunoglobulin of the tumor clone isotype from polyclonal cells expressing other isotypes. Virtually all cells with a high tumor CNV score also expressed immunoglobulin of the tumor isotype. Whereas immunoglobulin restriction is only informative for lymphoid malignancies the tumor CNV score can be applied to all tumor types.

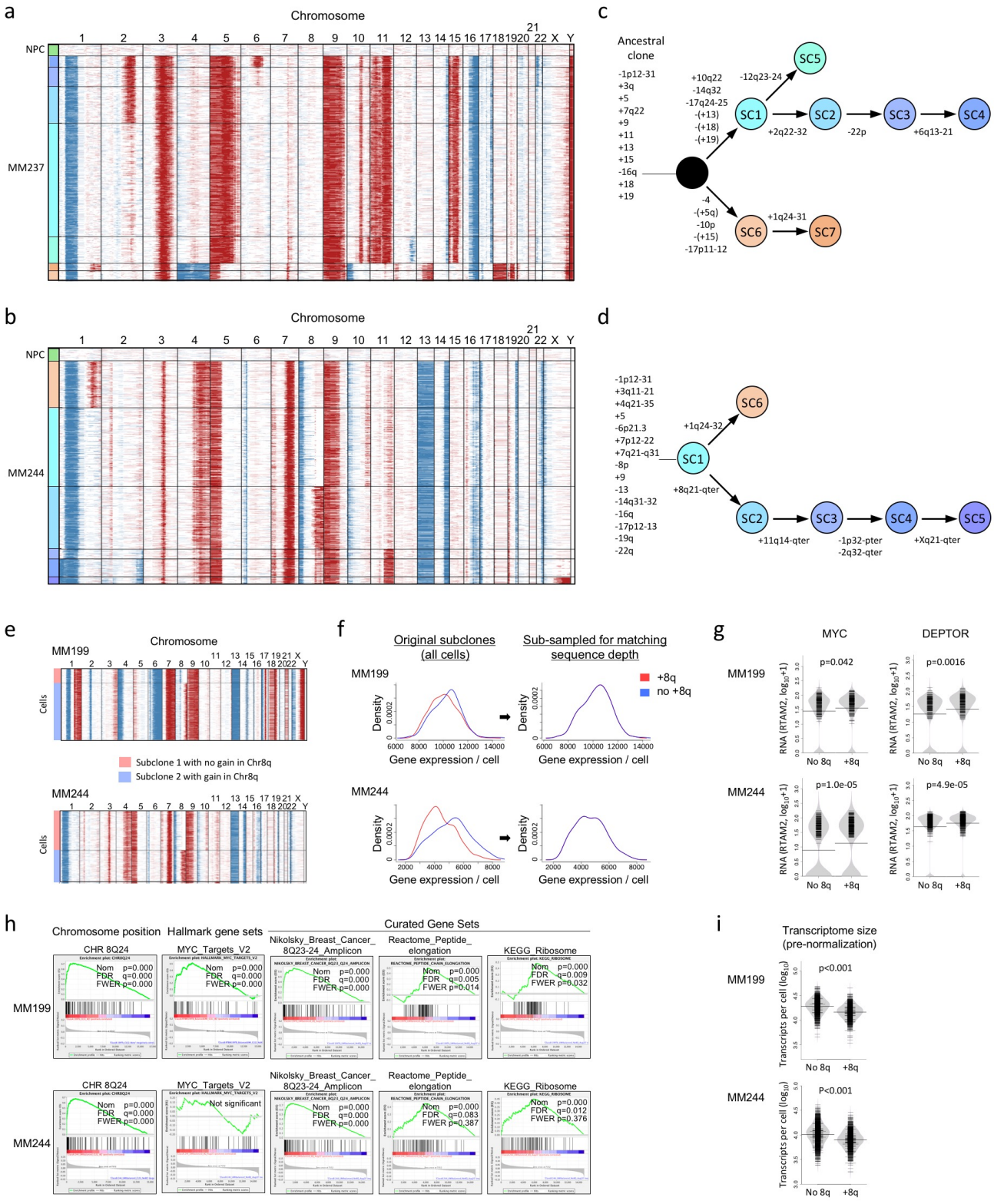
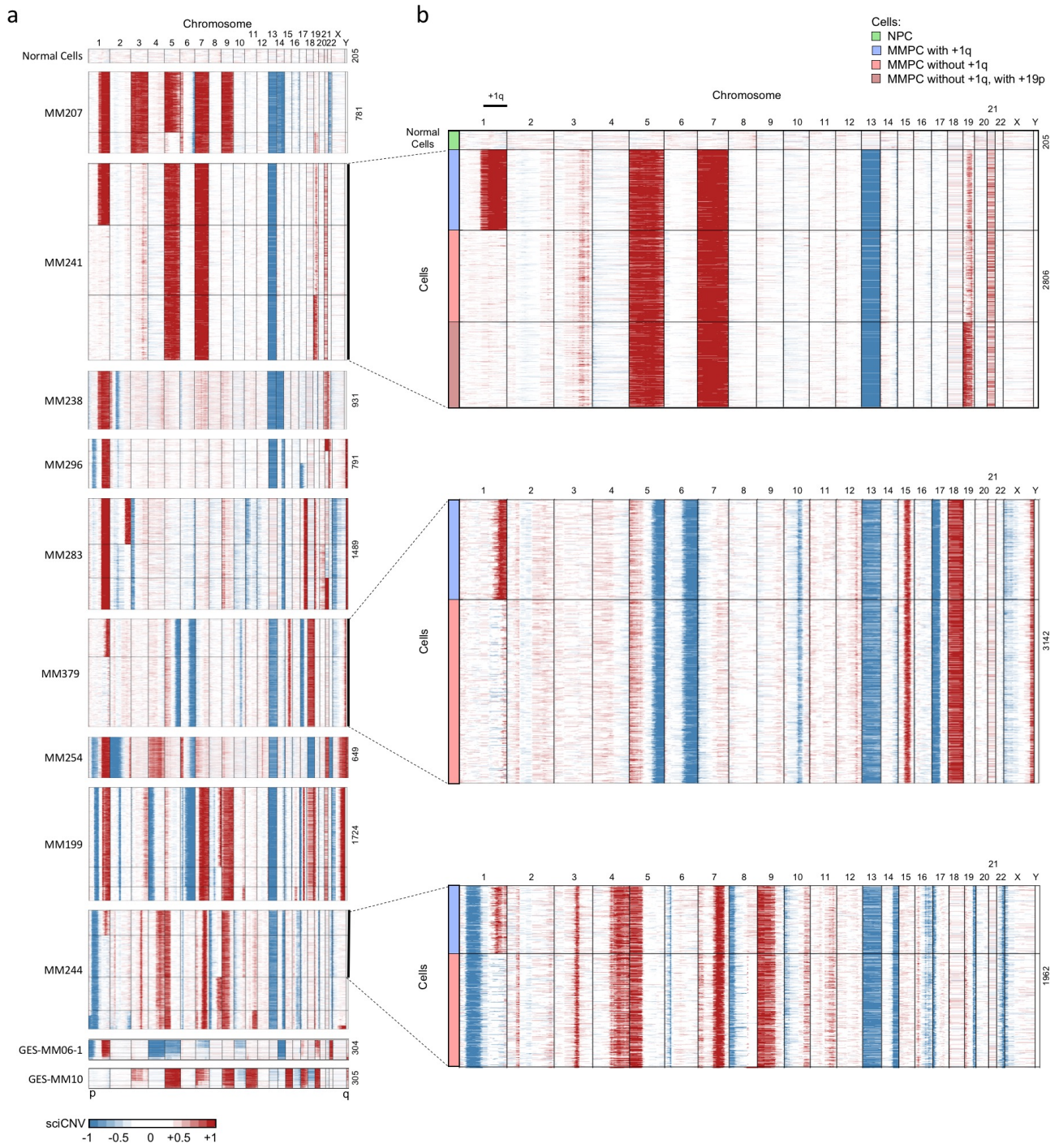


Figure 3. Examination of subclones with 8q gain at single cell resolution using sciCNV.

a. and **b.** The sciCNV profiles of plasma cells from multiple myeloma patient bone marrow samples MM237 (**a**) and MM244 (**b**) were calculated using scRNA-seq and are shown compared

to normal plasma cells (NPC). The tumor cells in each sample are grouped into subclones (colour bars at left) distinguished by divergent CNV. **c.** and **d.** Possible evolutionary paths for the subclones detected in MM237 and MM244, revealing branching and linear intra-clonal evolution. Subclones (SC) are represented by coloured circles corresponding to the colour bars in **a.** and **b.**

e. Heatmaps showing the sciCNV profiles of near isogenic subclone cells in MM199 and MM244 that that diverge at +8q. **f.** The distribution of total gene expression per cell (normalized transcriptome size) for the subclones shown in **e**. The subclones were sampled for subpopulations of cells with matching transcriptome sizes (right panel), which were then compared in subsequent panels (**g.-i.**). **g.** Bean plots showing the mRNA expression (RTAM2) of MYC or DEPTOR genes, located on chromosome 8q24, in transcriptome size-matched subpopulations from MM199 or MM244, by +8q status. Expression is plotted on $\log_{10}$ scale. P-values were calculated by t-test. **h.** Results of gene set enrichment analysis (GSEA) performed on subpopulations of MM199 and MM244 cells, comparing cells with or without +8q. The analysis of chromosome position gene sets (n=215) shows highly-significant enrichment of gene sets located on chromosome 8q23-24 in the populations of cells identified at single cell resolution as containing +8q by sciCNV (left panels). Key results of GSEA for hallmark (n=49) and curated (n=3303) gene sets are shown in the middle and right panels, demonstrating broad upregulation of MYC target genes in MM199 +8q cells, but not in MM244 +8q cells, and upregulation of ribosome and peptide_elongation signatures in +8q cells from both tumors; the expression changes attributable to +8q in MM cells from both tumors strongly resemble those found in breast cancer cells with an 8q23-24 amplicon. **i.** Bean plots showing the pre-normalized transcriptome sizes of subclonal MM cells from MM199 or MM244, demonstrating slightly fewer RNA transcripts in cells with +8q. P-values were calculated by t-test.



73 **Figure 4. sciCNV profiles of MM samples with chromosome 1q gain.**

74 **a.** The CNV profiles of MM cells (n=16,299) from 10 MM tumor samples with chromosome 1q
75 gain, inferred from scRNA-seq by sciCNV. The profiles of normal plasma cells (n=205) are
76 shown at the top. The number of cells in each sample is shown at the right. Samples containing
77 >3,000 cells are scaled by 0.5x for figure construction. Samples GES-MM06-1 and GES-MM10

78 at the bottom were characterized by MARS-seq[42]; single cell CNV predictions are shown here
79 calculated on the MARS-seq data using sciCNV.
80 **b.** Magnified view of MM241, MM379 and MM244, which were identified by sciCNV as
81 containing sibling subclones with and without gain of chromosome 1q (blue and red bars at left).

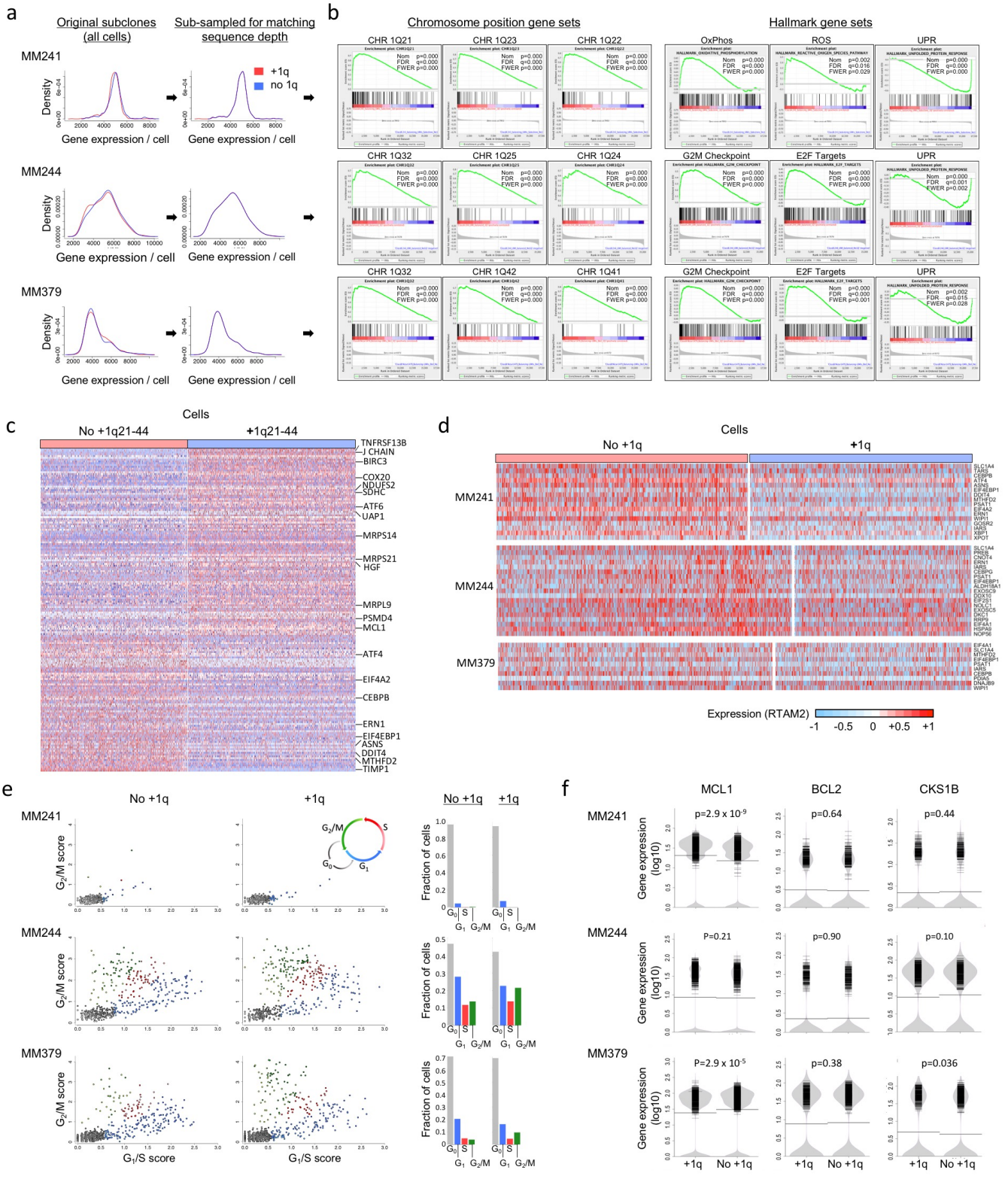


Figure 5. The effects of +1q on cellular programs in primary MM cells

a. The total gene expression (mRNA per cell) of cells in MM241, MM244 and MM379 with or without +1q (left column), demonstrating that +1q has no effect on transcriptome size (unlike

+8q). The subclones were nevertheless subsampled for subpopulations that were matched for transcriptome depth (right column), which were examined in the subsequent studies. **b.** Results of GSEA comparing sciCNV-resolved primary MM cells with or without +1q using RTAM2-normalized transcriptomics data. Analysis of chromosome position gene sets (n=215) revealed highly-significant enrichment for 1q gene sets in the cells identified individually as containing +1q by sciCNV (left panels). GSEA results for hallmark gene sets are shown at the right. G2M, E2F, oxidative phosphorylation (OxPhos) and reactive oxygen species (ROS) gene sets were variably enriched in subclones with +1q, while the UPR was decreased in all subclones with +1q. **c.** Heatmap depicting the differential gene expression of MM cells with or without +1q from sample MM241 (which contains a full-length +1q21-44 CNV). Columns represent cells and rows represent genes. **d.** Heatmaps showing the differential expression of UPR genes in cells with or without +1q, for MM241, MM244 and MM379 patient samples. **e.** Cell cycle phase of matched primary MM cells from the 3 patient samples, comparing cells with or without +1q. Cells were assigned to a cell cycle phase (colour-coded as per the legend) and plotted according to their relative expression of gene sets associated with G1/S and G2/M. The fraction of cells in each phase according to +1q status is summarized by histogram (right). **f.** Bean plots depicting the relative expression of MCL1 (located at 1q21.2) and CKS1B (1q21.3) in cells with or without +1q. The expression of BCL2 (located at 18q21) is shown as a control. Expression is plotted on a $\log_{10}$ scale. P-values were calculated by t-test.

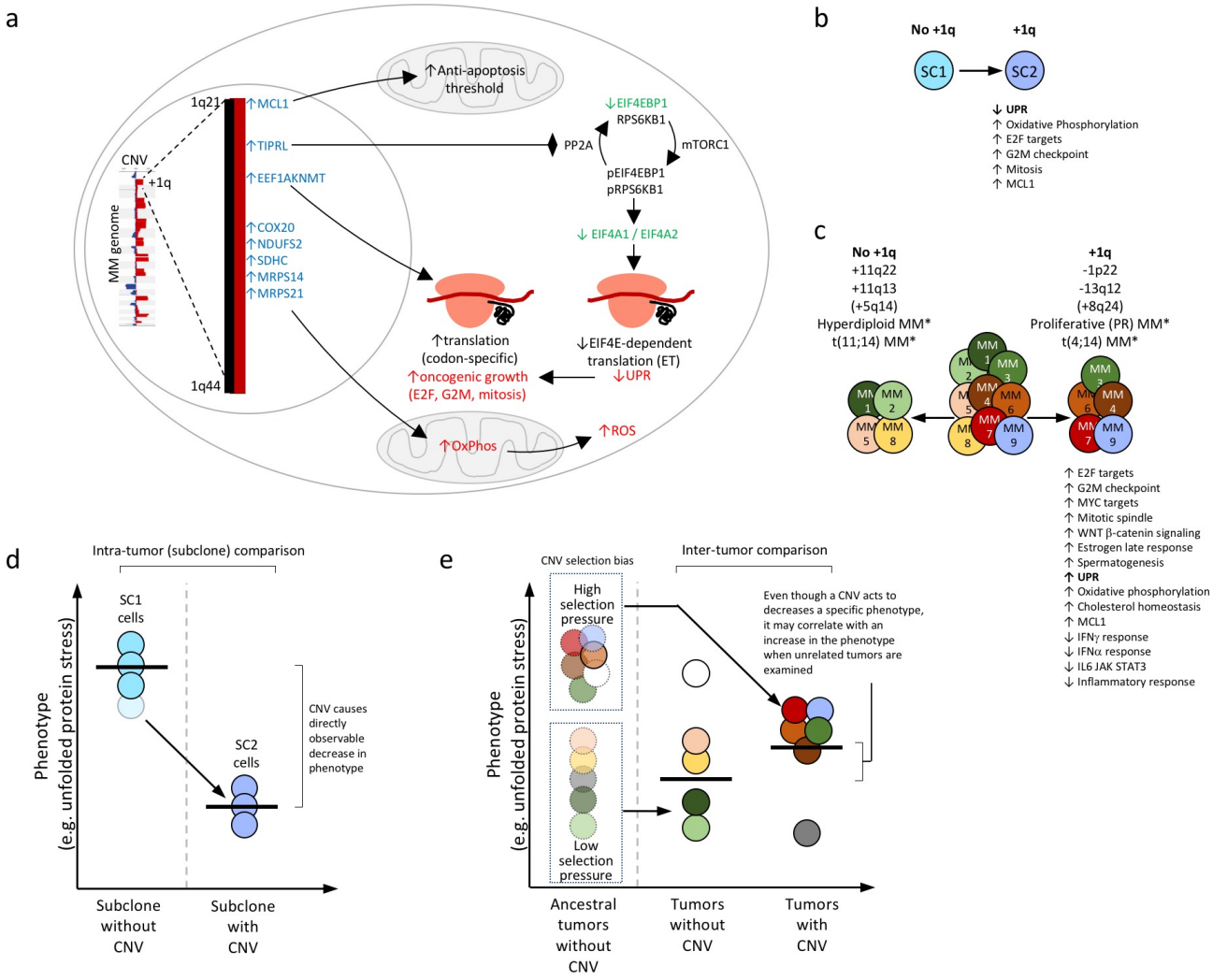


Figure 6. The effect of +1q on cellular programs in MM and comparison of intra-tumor and inter-tumor studies.

a. Summary of the influences of +1q21-44 on MM cell biology, as determined by scRNA-seq. Genes located on 1q that are increased in +1q cells (blue) are linked to downstream subcellular programs that altered by +1q (red) via intermediate genes that also show altered expression in +1q cells (green).

b. and **c.** Comparison of intra-tumor and inter-tumor studies to determine the effects of +1q in MM. **b.** The results of intra-tumor GSEA of MM241, MM244 and MM379 are summarized. Amongst chromosome position gene sets, only 1q gene sets were enriched in +1q cells. The hallmark gene sets that were significantly co-modulated are shown. **c**. The results of an inter-tumor analysis addressing the same biological question are summarized. MM tumor cohorts from GSE2658[22] (n=532 samples) were defined by the presence or absence of +1q by FISH.

Chromosome position gene-sets that were significantly divergent between the cohorts by GSEA are listed above the graphic. Gene sets with nominal p-value<0.05 but FWER p-value>0.05 are bracketed. Although large numbers of tumor samples were grouped specifically according to their 1q status, additional genomic heterogeneity persists between the cohorts. Biases in MM genetic subtypes(*), correlating with +1q status, were also observed, as reported[22,43]. Hallmark gene sets that were divergent between the cohorts are listed below the schema. MCL1 expression was analyzed at single gene level.

d. Illustration of an intra-tumor analysis of sibling subclones. The effect of a divergent CNV on transcriptional programs can be directly assessed. The subclones are otherwise isogenic, reducing the influence of confounding genetic variations, and are derived from the same sample, minimizing the influence of confounding variations due to sample processing, batch effect or recent patient treatment.

e. Illustration of an inter-tumor analysis in which the influence of a CNV on cell phenotype is examined, highlighting potential biases. In the example, the CNV does not occur randomly but is preferentially selected for by tumors experiencing a specific stressor (left column). The gene expression of the tumor cohorts that do or do not develop the CNV are therefore not identical at baseline. Although the CNV may act to reduce the stressor, it's occurrence may appear to correlate with increased rather than with decreased stress, or may fail to correlate.