nature biotechnology

Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells

Daniel Ramsköld[1,2,7], Shujun Luo[3,7], Yu-Chieh Wang[4], Robin Li[3], Qiaolin Deng[1], Omid R Faridani[1], Gregory A Daniels[5], Irina Khrebtukova[3], Jeanne F Loring[4], Louise C Laurent[6], Gary P Schroth[3] & Rickard Sandberg[1,2]

Genome-wide transcriptome analyses are routinely used to monitor tissue-, disease- and cell type–specific gene expression, but it has been technically challenging to generate expression profiles from single cells. Here we describe a robust mRNA-Seq protocol (Smart-Seq) that is applicable down to single cell levels. Compared with existing methods, Smart-Seq has improved read coverage across transcripts, which enhances detailed analyses of alternative transcript isoforms and identification of single-nucleotide polymorphisms. We determined the sensitivity and quantitative accuracy of Smart-Seq for single-cell transcriptomics by evaluating it on total RNA dilution series. We found that although gene expression estimates from single cells have increased noise, hundreds of differentially expressed genes could be identified using few cells per cell type. Applying Smart-Seq to circulating tumor cells from melanomas, we identified distinct gene expression patterns, including candidate biomarkers for melanoma circulating tumor cells. Our protocol will be useful for addressing fundamental biological problems requiring genome-wide transcriptome profiling in rare cells.

Analyses of transcriptomes through massively parallel sequencing of cDNAs (mRNA-Seq) generates millions of short sequence fragments that can be analyzed to accurately quantify expression levels[1], assemble new transcripts[2,3] and investigate alternate RNA processing[4,5]. These techniques have been consistently pushed toward development of methods that require lower starting amounts of RNA, ideally as small as single cells. A protocol initially developed for single-cell microarray studies[6] has been adapted for mRNA-Seq and used to generate transcriptome data for individual mouse oocytes and early embryonic cells[7,8]. Using the method, thousands of genes expressed in mouse oocytes had been detected, and it yielded increased sensitivity compared with microarrays[7]. However, this first single-cell mRNA-Seq experiment lacked technical controls, making it impossible to distinguish biological variation between different cells from the technical variation that is intrinsic to cDNA amplification protocols when starting with small amounts of RNA. Therefore, the question remained whether single-cell transcriptomes faithfully represent the RNA population before amplification and how technical variation limits the power to find differences in expression. This initial mRNA-Seq method also preferentially amplified the 3′ ends of mRNAs, and hence the data could only be used to identify distal splicing events. Recently, a method for multiplexed single-cell RNA-Seq has been introduced that quantifies transcripts through reads mapping to mRNA 5′ ends[9]. Neither of these methods generates read coverage across full transcripts. As most mammalian multi-exon genes are subject to alternative RNA processing[4,5], there is a need for a single-cell transcriptome method that can be used to both quantify gene expression and provide the coverage for efficient detection of transcript variants and alleles.

Here we introduce a single-cell RNA-sequencing protocol with markedly improved transcriptome coverage, which samples cDNAs from more than just the ends of mRNAs. Using this protocol, we sequenced the mRNAs from many individual mammalian cells, as well as well-defined dilution series of purified total RNAs, to comprehensively assess how sensitivity, variability and detection of differing expression vary with different amounts of starting material. Our results demonstrate the power of single-cell RNA-Seq for both transcriptional and post-transcriptional studies, and provide valuable insights into the design of experiments that start from few or single cells. To demonstrate the biological importance of this method, we applied this assay to putative circulating tumor cells (CTCs) captured from the blood of a melanoma patient to demonstrate how Smart-Seq enables high-quality transcriptome mapping in individual, clinically important cells.

RESULTS

Efficient and robust single-cell RNA sequencing

For Smart-Seq, first we lysed each cell in hypotonic solution and converted poly(A)$^+$ RNA to full-length cDNA using oligo(dT) priming and SMART template switching technology, followed by 12–18 cycles of PCR preamplification of cDNA. We used the amplified cDNA to construct standard Illumina sequencing libraries using either Covaris shearing followed by ligation of adaptors (PE)

[1]Ludwig Institute for Cancer Research, Stockholm, Sweden. [2]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. [3]Illumina, Inc., Hayward, California, USA. [4]Department of Chemical Physiology, Center for Regenerative Medicine, The Scripps Research Institute, San Diego, La Jolla, California, USA. [5]Rebecca and John Moores Cancer Center, San Diego, La Jolla, California, USA. [6]Department of Reproductive Medicine, University of California, San Diego, La Jolla, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to R.S. (rickard.sandberg@ki.se).

Received 14 November 2011; accepted 22 May 2012; published online 22 July 2012; doi:10.1038/nbt.2282

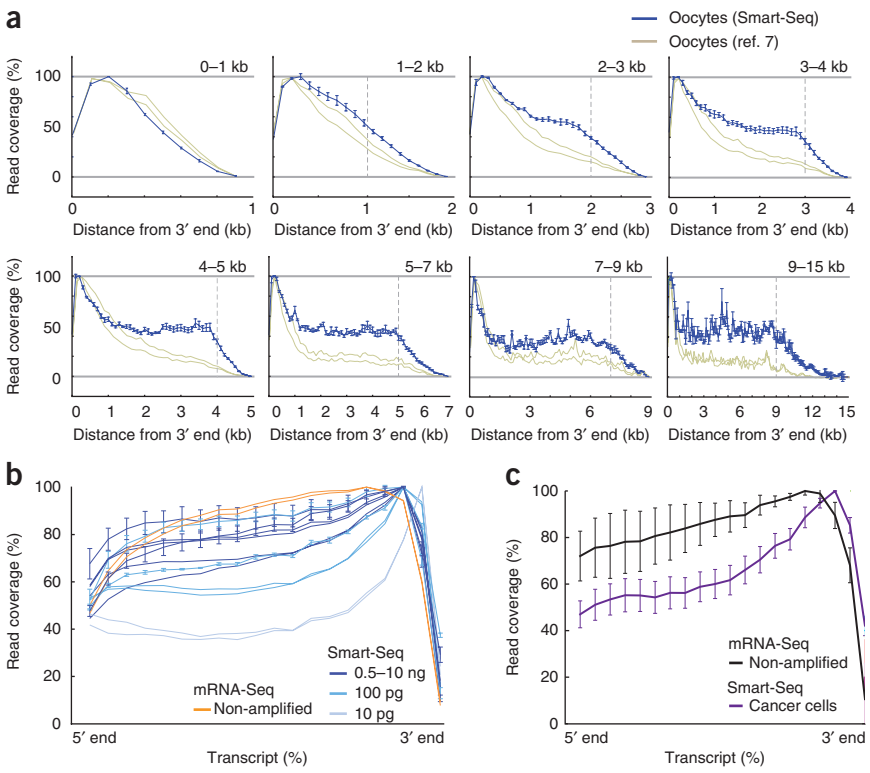

Figure 1 Smart-Seq read coverage across transcripts. (**a**) Comparison of read coverage over transcripts for Smart-Seq–analyzed mouse oocytes ($n = 3$) and previously published mouse oocyte transcriptome data (ref. 7; $n = 2$). Transcripts were grouped according to annotated lengths and analyzed separately, with the transcript length ranges indicated (top right). We display the read coverage over the transcripts as a distance from the 3′ end, with the vertical dashed gray line showing the length of the shortest included transcripts after which a decline in read coverage is expected. Error bars represent s.d. among biological replicates. (**b**) Mean read coverage over transcripts for Smart-Seq data generated from diluted amounts of mouse brain RNA. Independent dilution series (including data from different laboratories) are shown as separate data sets, and sample numbers are listed from uppermost line down. For comparison, we included data from standard mRNA-Seq on 100 ng of mouse brain RNA (non-amplified). Errors bars, s.d. ($n = 5$, 3, 4 and 4 for lines top to bottom). (**c**) Read coverage (as in **b**) for 12 individual human cells of prostate and bladder cancer line origin, analyzed using Smart-Seq (cancer cells; $n = 12$) and for prostate cell line LNCaP analyzed with standard mRNA-Seq (non-amplified; $n = 4$). Error bars, s.d.

or Tn5-mediated 'tagmentation' using the Nextera technology (Tn5). Both of these library preparation methods enable random shotgun sequencing of cDNAs (**Supplementary Fig. 1**). We generated Smart-Seq libraries from 42 individual human or mouse cells, and in addition we generated 64 libraries from dilution series of total RNA derived from human brain (16 samples), mouse brain (28 samples) and universal human reference RNA (UHRR, 20 samples). We sequenced each sequencing library on the Illumina platform, typically generating >20 million uniquely mapping reads (**Supplementary Table 1**). For comparison, we also made several standard mRNA-Seq libraries from 100 ng to a few micrograms of total RNA.

Smart-Seq improves coverage across transcripts

In previous single-cell mRNA-sequencing studies[7,8], the data suffered from a pronounced 3′-end bias that limited analysis across full-length transcripts. We sequenced single-cell transcriptomes from mouse oocytes to enable a direct comparison with published mouse oocyte single-cell data[7]. Analyses of read coverage across transcripts demonstrated that Smart-Seq has considerably improved full-length coverage of all transcripts longer than 1 kb (**Fig. 1a** and **Supplementary Fig. 2a–h**). Smart-Seq analyses of mouse brain RNA at different dilutions showed that even better coverage was obtained with increased starting amounts, with nanogram dilutions reaching close to the coverage observed using standard mRNA-Seq from 100 ng to 1 µg total RNA (**Fig. 1b**). From only 10 pg input amounts (the estimated amount of RNA in a small eukaryotic cell, **Supplementary Table 2**), we achieved close to 40% coverage at the 5′ end. Analyses of single-cell transcriptomes from cancer cell lines (four cells each from LNCaP, PC3 and T24) obtained equally good read coverage (**Fig. 1c**) and, indeed, for 25% of all expressed, multi-exon genes our read coverage enabled full-length transcript reconstruction (**Supplementary Fig. 3**). We conclude that Smart-Seq has substantially improved read coverage compared with previous single-cell transcriptome methods.

Quantitative assessment of single-cell transcriptomics

Analyses of gene expression from millions of cells using mRNA-Seq is highly reproducible and has low technical variation[1,4]. To our knowledge, no single-cell mRNA-Seq study has measured the technical variation intrinsic to the cDNA pre-amplification components of single-cell methods. We therefore diluted microgram amounts of reference total RNA down to nano- and picogram levels and applied Smart-Seq to assess sensitivity, technical variability and detection of differentially expressed transcripts of Smart-Seq on low amounts of total RNA. For comparison, we generated standard mRNA-Seq libraries from 100 ng to microgram amounts of reference total RNA.

First, we addressed the sensitivity of the method in detecting transcripts expressed at different levels. Starting with 10 ng or 1 ng of total RNA, we found no or minimal decline in sensitivity compared with standard mRNA-Seq. However, lowering the starting amounts to single-cell levels decreased the detection rate of less abundant transcripts (**Fig. 2a**). Analyses of the 12 cancer cell line cells (four cells each from the LNCaP, PC3 and T24 lines) showed that ~76% of transcripts expressed at 10 RPKM (reads per kilobase exon model and million mappable reads), which roughly equals the median expression for detected transcripts, were reproducibly detected in all single-cell profiles (**Fig. 2b**). We found that the sensitivity of gene detection for the individual cancer cells was similar to that obtained with ~20 pg of starting total RNA (**Fig. 2b**), with ~8,000 genes detected per cell and increasing with the number of analyzed cells (**Supplementary Fig. 4a**) Furthermore, we observed that the starting amount of total RNA had a larger impact on sensitivity than the number of PCR cycles used (**Supplementary Fig. 5**) and that the sequence depth had little effect on transcript detection at levels above a million uniquely mapping reads per cell, with expression levels stabilizing after 3 million uniquely mapped reads (**Supplementary Fig. 4c,d**). Comparisons of Smart-Seq and previous mouse oocyte data[7] demonstrated similar sensitivity (**Supplementary Fig. 2i,j**). We conclude that transcript

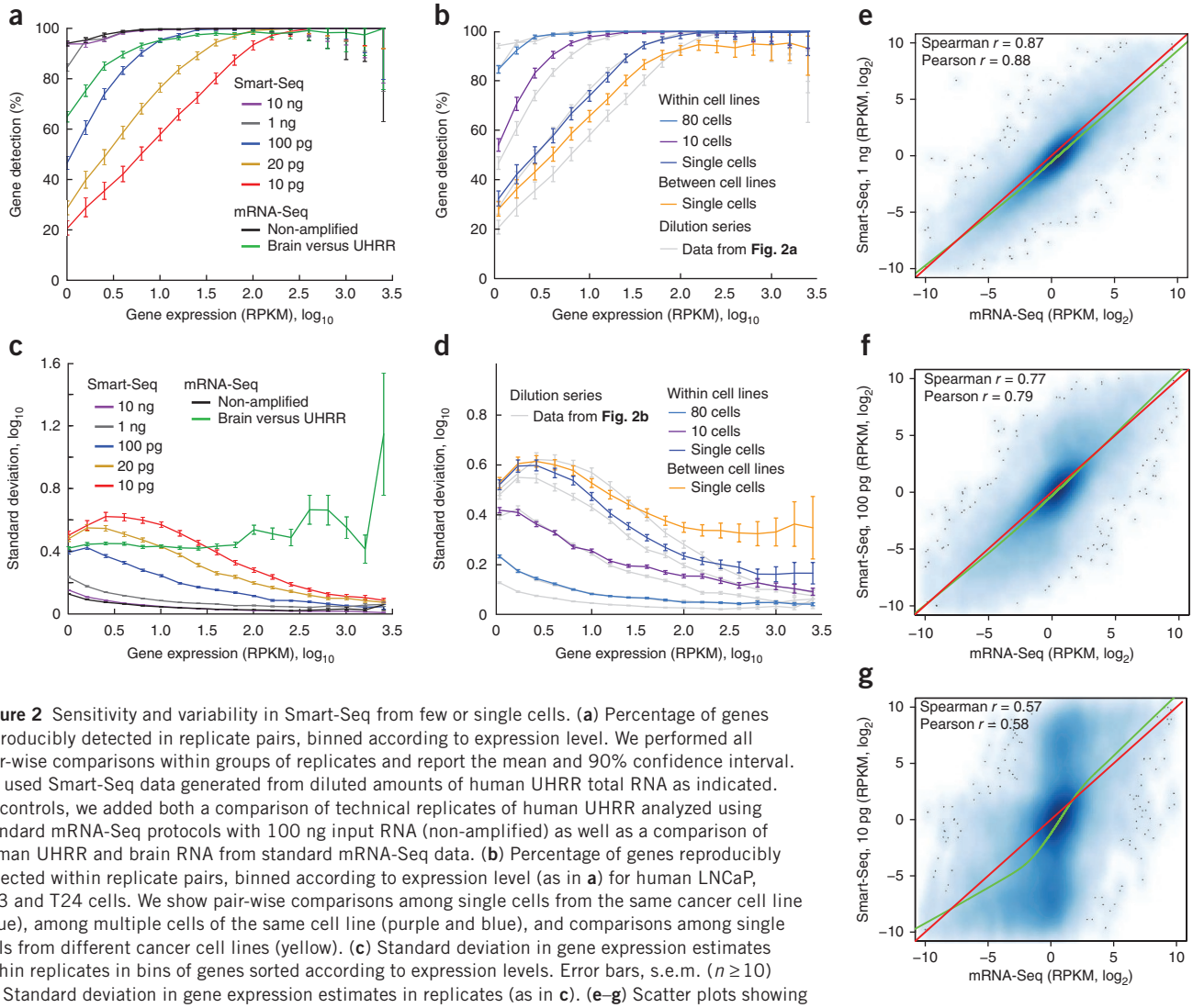

Figure 2 Sensitivity and variability in Smart-Seq from few or single cells. (**a**) Percentage of genes reproducibly detected in replicate pairs, binned according to expression level. We performed all pair-wise comparisons within groups of replicates and report the mean and 90% confidence interval. We used Smart-Seq data generated from diluted amounts of human UHRR total RNA as indicated. As controls, we added both a comparison of technical replicates of human UHRR analyzed using standard mRNA-Seq protocols with 100 ng input RNA (non-amplified) as well as a comparison of human UHRR and brain RNA from standard mRNA-Seq data. (**b**) Percentage of genes reproducibly detected within replicate pairs, binned according to expression level (as in **a**) for human LNCaP, PC3 and T24 cells. We show pair-wise comparisons among single cells from the same cancer cell line (blue), among multiple cells of the same cell line (purple and blue), and comparisons among single cells from different cancer cell lines (yellow). (**c**) Standard deviation in gene expression estimates within replicates in bins of genes sorted according to expression levels. Error bars, s.e.m. ($n \geq 10$) (**d**) Standard deviation in gene expression estimates in replicates (as in **c**). (**e**–**g**) Scatter plots showing the relative differences between human UHRR and brain gene expression levels estimated from standard mRNA-Seq data on 100 ng input RNA (*x* axis) and Smart-Seq generated data (*y* axis) starting from 1 ng total RNA (**e**), 100 pg total RNA (**f**) and 10 pg total RNA (**g**). Correlation coefficients computed from $\log_2$ transformed relative gene expression profiles, together with nonlinear loess regression curves (green) and $y = x$ lines (red).

detection sensitivity is affected by limiting starting amounts of RNA that lead to random loss of low-abundance transcripts, but still the majority of low-abundance and the vast majority of highly expressed transcripts are reliably detected even in single cells.

Second, we determined the reproducibility in expression levels generated from diluted RNA and individual cells. Comparison of Smart-Seq and previous mouse oocyte data[7] demonstrated improved estimation of expression with Smart-Seq (lower variability in data from oocyte to oocyte) across the whole range of expression levels (**Supplementary Fig. 2k**). Correlation analyses between technical replicates of diluted RNA showed increasing concordance with larger amounts of RNA. Comparing data from the single cells against the RNA-dilution data, we observed higher correlations (Pearson correlations of 0.75–0.85) among individual cells of the same type than among dilution replicates at 10 pg (Pearson correlations of 0.65–0.75) (**Supplementary Fig. 6**). As variability in measurements of expression depends on transcript expression levels, we computed the variability as a function of the expression level (**Fig. 2c**,**d**). This analysis showed that Smart-Seq on 10 ng total RNA had the same technical variability as standard mRNA-Seq and that Smart-Seq on 1 ng total RNA showed only a modest increase in technical noise (**Fig. 2c**). When lowering input amounts down to picogram levels, there was a clear increase in technical variability, particularly for less abundantly expressed transcripts (**Fig. 2c**). We compared technical variability at picogram levels of total RNA to the biological variation found in comparisons of human brain samples and UHRR using standard mRNA-Seq (**Fig. 2c**). Notably, analyses of gene-expression variation between individual cancer cells of different origin revealed extensive biological variation in highly expressed genes (**Fig. 2d**).

Finally, we assessed whether single-cell expression profiles from preamplified material were representative of the original expression profiles. Comparing relative gene expression levels (UHRR minus brain) estimated using standard mRNA-Seq to those estimated from Smart-Seq with different amounts of input RNA, we again found a high concordance (**Fig. 2e**–**g**). Starting with 1 ng or 100 pg total RNA, the relative expression in Smart-Seq and standard mRNA-Seq, respectively, had Spearman correlations of 0.87 and 0.77 (**Fig. 2e**,**f**). Comparisons with 10 pg input RNA showed overall good correlation (**Fig. 2g**) but identified two populations of transcripts with distorted expression in Smart-Seq data from either human brain sample or UHRR, reflecting

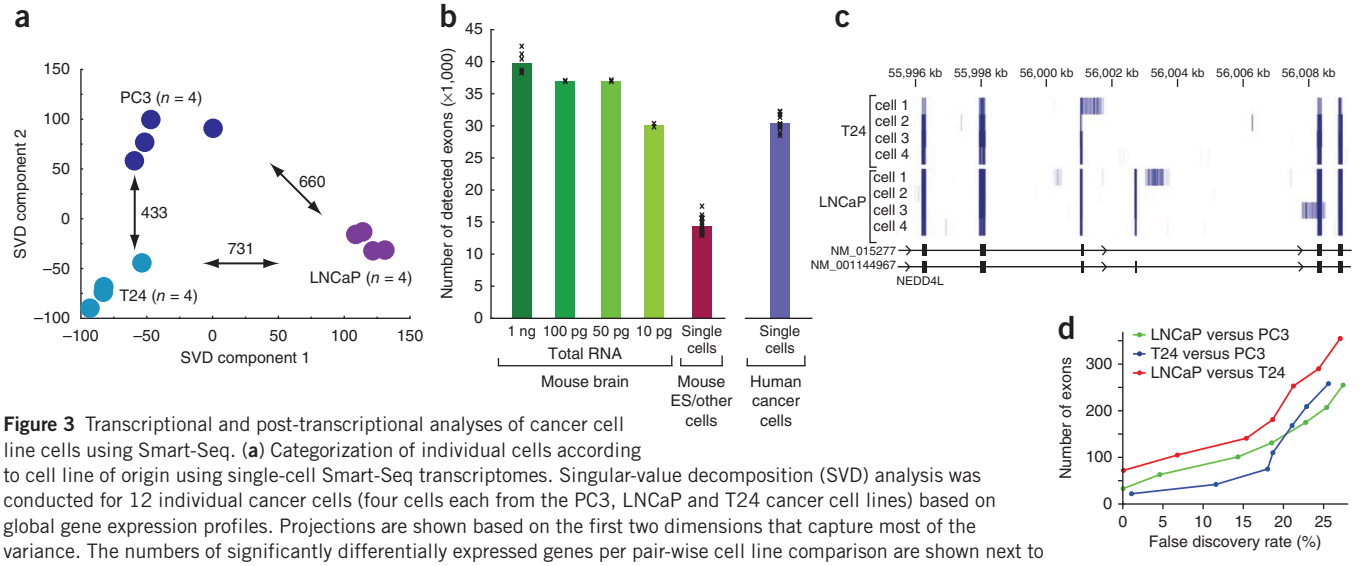

Figure 3 Transcriptional and post-transcriptional analyses of cancer cell line cells using Smart-Seq. (**a**) Categorization of individual cells according to cell line of origin using single-cell Smart-Seq transcriptomes. Singular-value decomposition (SVD) analysis was conducted for 12 individual cancer cells (four cells each from the PC3, LNCaP and T24 cancer cell lines) based on global gene expression profiles. Projections are shown based on the first two dimensions that capture most of the variance. The numbers of significantly differentially expressed genes per pair-wise cell line comparison are shown next to the arrows ($P < 0.05$, 1-way ANOVA and Tukey post-hoc test). (**b**) Mean number of exons with sufficient read coverage for MISO analyses of exon inclusion levels in sequence-depth matched single-cell mRNA-Seq data. Smart-Seq data from diluted mouse brain RNA (green) compared with previously published mouse ESCs[8] and ESC-derived cells (red) and 12 Smart-Seq–analyzed individual prostate and bladder cell line cells (purple). Individual RNA or cell measurements are plotted. (**c**) Single-cell Smart-Seq reads mapping to a portion of the *NEDD4L* gene locus from four individual T24 and LNCaP cells. Read coverage is shown as a heatmap with darker blue indicating higher read coverage. (**d**) Number of differentially included exons identified among the PC3, LNCaP and T24 cell lines from single-cell Smart-Seq analysis on four cells per cell line as a function of estimated false discovery rate.

stochastic losses, mostly of low-abundance transcripts when starting with such minute RNA amounts (**Fig. 2a**,**g**). Preamplification of cDNA could also lead to disproportionate amplification of short transcripts, but we found no systematic bias (**Supplementary Fig. 7**). A previous microarray study had analyzed PCR-amplified cDNA (from picogram starting amounts) and found the transcriptome overall preserved but skewed[10]. Our data from 1 ng and 100 pg total RNA showed no skewing, that is, the loess slopes estimated from the data approximated 1 (**Fig. 2e**–**g**). Together, these results demonstrated that transcriptome analyses from few or single cells, in general, preserved relative differences in expression for detected transcripts.

Transcriptional and post-transcriptional differences

Having demonstrated the improved performance of Smart-Seq on low amounts of RNA compared with previously published methods, we focused our analyses on single-cell transcriptomes from prostate (PC3 and LNCaP) and bladder (T24) cancer cell line cells. The global gene expression of 12 individual cells (four from each cell line) clustered according to cell line of origin and we identified hundreds of differentially expressed genes among the three cell lines (**Fig. 3a**; $q < 0.05$ ANOVA; $P < 0.05$ post-hoc test).

The pronounced 3′-end bias of previous single-cell mRNA-Seq studies has hampered the ability to identify alternative splicing differences in single cells. We used the Bayesian mixture of isoforms framework (MISO)[11] to infer exon inclusion levels for known alternatively spliced exons in the 12 individual cells. The improved read coverage with Smart-Seq resulted in a twofold increase in the number of potential alternatively spliced exons that could be assessed, compared to previously published single-cell mRNA-Seq data (**Fig. 3b**), substantially improving our ability to detect alternative splicing. Cell type–specific alternative splicing could be inferred from single-cell transcriptomes, as seen in read coverage across the differentially included exon 13 of the *NEDD4L* gene (**Fig. 3c**). This exon was frequently included in LNCaP cells (93% mean inclusion level) but was included at much lower levels in T24 cells (15% mean inclusion levels) whereas low expression of *NEDD4L* in PC3 cells precluded inclusion level estimation. In this comparison of three cancer cell lines, we found 100 exons with differential exon inclusion levels among the three cell lines, with a less than 1% false discovery rate (**Fig. 3d** and **Supplementary Table 3**). We conclude that Smart-Seq considerably improves our ability to detect alternative RNA processing in single cells.

Analyses of circulating tumor cell transcriptomes

Having demonstrated that Smart-Seq generates quantitative and reproducible single-cell transcriptomes, we asked whether global transcriptome analyses of putative CTCs could reveal their tumor of origin and provide data to support the use of this method for unbiased cancer-specific biomarker identification. To this end, we generated transcriptomes from six single NG2$^+$ putative melanoma CTCs isolated from peripheral blood drawn from a patient with recurrent melanoma using immunomagnetic purification with a MagSweeper instrument (Illumina)[12]. For comparison, we also generated Smart-Seq libraries from single cells derived from primary melanocytes ($n = 2$), melanoma cancer cell line (SKMEL5, $n = 4$ and UACC257, $n = 3$) cells and from human embryonic stem cells (ESCs, $n = 8$). As the NG2$^+$ putative CTCs were isolated from blood, it was important to compare them to blood cells. The putative CTCs were distinct from lymphoma cell lines (BL41 and BJAB)[13] and immune tissues (lymphnode and white blood cell samples), as well as embryonic stem cells, and instead were highly similar to primary melanocytes and melanoma cell line cells. Unsupervised hierarchical clustering and correlation analyses of gene expression levels showed a clear clustering of cells according to cell type of origin (**Fig. 4a** and **Supplementary Fig. 8**), and separation from the human brain RNA samples that were previously analyzed with Smart-Seq or mRNA-Seq (data not shown). Additional support for the melanocytic origin of the putative melanoma CTCs came from analyses of melanocyte lineage–specific markers, as all NG2$^+$ cells expressed high mRNAs levels for *MLANA*[14], *TYR*[15] and the melanocyte specific m-form of *MITF*[16] but not immune markers such as *PTPRC* (**Fig. 4b**), in contrast to peripheral blood lymphocytes (**Supplementary Fig. 9**). Furthermore, NG2$^+$ cells expressed high levels of melanoma-associated genes (based on our unbiased selection of the 100 transcripts most strongly associated

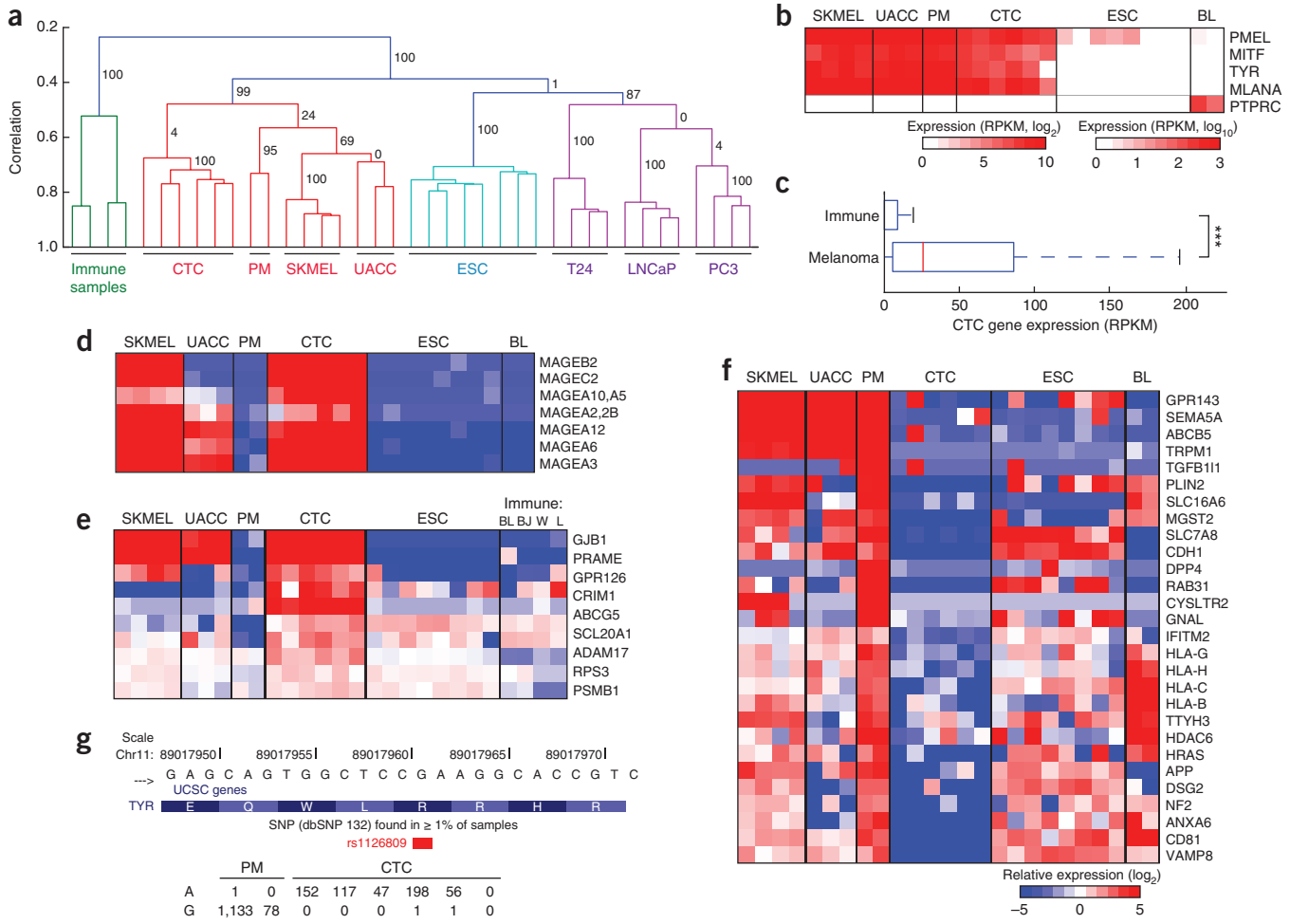

Figure 4 Single-cell transcriptomes of circulating tumor cells. (**a**) Hierarchical clustering of human samples based on gene expression of highly expressed genes (>100 RPKM). Coloring indicates high-order clusters and the confidence in clusters are indicated with bootstrap values (percentage). Samples analyzed include human immune samples (Burkitt's lymphoma cell lines BL41 and BJAB, and white blood cells and lymph node samples) and cells from putative melanoma CTCs (CTC), primary melanocytes (PM), melanoma cell lines SKMEL5 (SKMEL) and UACC257 (UACC), prostate cancer cell lines (LNCaP and PC3), bladder cancer cell line (T24) and human embryonic stem cells (ESC). (**b**) Expression of melanocyte makers (PMEL, MITF, TYR and MLANA) and immune marker PTPRC in single-cell transcriptomes from **a** with Burkitt's lymphoma cell lines BL41 and BJAB (BL). (**c**) Gene expression levels in CTCs for an unbiased set of 100 immune and melanoma markers. (**d**–**f**) Heatmaps showing relative expression of melanoma associated tumor antigens (**d**), upregulated plasma-membrane proteins (**e**), and downregulated plasma-membrane proteins (**f**) in single-cell transcriptomes as in **b** with the addition of more immune samples (W, white blood cells; L, lymph node). (**g**) Number of reads from individual PMs and putative CTCs that support the reference (G) or risk (A) allele for the melanoma-associated SNP (rs1126809).

with melanoma; see Online Methods), but not immune cell–associated genes selected in a similar manner (**Fig. 4c**, $P < 3.7 \times 10^{-15}$, Wilcoxon rank sum test). Thus, both their global transcriptomes and expression patterns of melanoma-associated transcripts clearly support a melanoma CTC identity for the NG2$^+$ cells.

We next investigated whether the NG2$^+$ putative CTCs showed signs of originating from a melanoma tumor. Comparison of their gene expression profiles with those of individual primary melanocytes identified 289 genes with significantly ($q < 0.05$ ANOVA; $P < 0.05$ post-hoc test) higher expression in the putative CTCs than the primary melanocytes, and 436 genes with significantly ($q < 0.05$ ANOVA; $P < 0.05$ post-hoc test) lower levels (**Supplementary Table 4**). The upregulated genes were significantly (Benjamini-Hochberg adjusted $P < 0.05$) enriched for melanoma-associated antigens (**Fig. 4d** and **Supplementary Fig. 10**) that have been repeatedly found to be upregulated in cancer[17], mitotic cell cycle genes and additional categories (**Supplementary Table 5**). Downregulated genes were enriched for regulators of cell death and MHC class I genes. Notably, the preferentially expressed antigen in melanoma (PRAME) was highly expressed in NG2$^+$ cells, which together with elevated expression of known melanoma tumor antigens, provides strong support for the conclusion that the NG2$^+$ cells were CTCs that originated from a melanoma.

In recent years, there has been a strong interest in identifying CTCs from different tumors using the a priori assumption that plasma-membrane proteins would be good diagnostic biomarkers. We used the CTC transcriptome analysis to screen for membrane proteins selectively expressed in melanoma-derived CTCs compared to primary melanocytes and immune cells. We identified nine upregulated plasma membrane–associated transcripts in the CTCs compared to primary melanocytes ($q < 0.05$ ANOVA; $P < 0.05$ post-hoc test), many of which are not expressed in immune cells and have not been previously associated with melanomas (**Fig. 4e**). Similarly, screening for loss of expression of plasma-membrane proteins identified 37 genes with significantly ($q < 0.05$ ANOVA; $P < 0.05$ post-hoc test) lower expression

in the CTCs than primary melanocytes (**Fig. 4f**). Of note, epithelial Cadherin 1 (CDH1) showed no expression in the CTCs, and loss of CDH1 is thought to contribute to cancer progression by increasing proliferation, invasion and metastasis[18]. We also found downregulation of genes associated with the escape from immune surveillance, including five HLA genes (**Fig. 4f**), and *TRPM1*, suggesting that these gene expression changes might enable the CTCs to escape from immune surveillance. Notably, low expression of TRPM1 has been shown to correlate with melanoma aggressiveness and metastasis[19]. Future studies of these membrane proteins will likely enhance our understanding of CTC migration and invasiveness, and these results highlight the utility of studying single CTC cells with RNA-Seq.

Lastly, we investigated whether Smart-Seq transcriptome data could be mined for single-nucleotide polymorphisms (SNPs) and other genetic variants associated with melanomas or other cancers. With the improved read coverage provided by the Smart-Seq method, we identified 4,312 high-confidence genomic sites with support for an alternative allele in at least two CTCs, whereas genotype calls only supported by a single cell showed an excess of previously unidentified, likely artifactual, sites (**Supplementary Fig. 11**) together with a smaller subset (9%) of A-to-G RNA editing sites (data not shown). Ninety-two percent of the high-confidence sites coincided with documented SNPs, for example, the melanoma-associated SNP in the *TYR* gene (rs1126809)[20] (**Fig. 4g**). We conclude that Smart-Seq enables screening for SNPs and mutations in transcribed regions using only few cells.

DISCUSSION

Generating high-coverage transcriptomes from single cells and small numbers of cells will have many applications for studying rare cells; such cells can be either individually picked or identified through cell sorting or laser-capture techniques. Our results showed that using Smart-Seq on 10 ng of total RNA was practically indistinguishable from a standard mRNA-Seq, whereas starting with 1 ng (corresponding roughly to 50–100 cells) showed only a minor (less than twofold) increase in expression-level variability. Therefore, this method could be applied to studies on homogeneous cell populations available in quantities of tens to hundreds of cells.

However, many biologically and clinically important cell types exist in rare quantities and often in heterogeneous milieus, which necessitates single-cell approaches. Smart-Seq generates robust and quantitative transcriptome data from single cells. We found hundreds of differentially expressed genes using only a few individual cells per cell type; for example, comparing only two primary melanocytes to six melanoma CTCs identified biologically meaningful differences. Even sequencing of a single cell yielded useful information, as we, in each cell, detected most of the genes active in a culture of LNCaP cells.

Smart-Seq is a robust method for single-cell RNA-Seq with improved read coverage across transcripts, which enables more detailed analyses of alternative splicing. Based on our CTC transcriptome results, single-cell analyses using Smart-Seq are also highly informative for identifying candidate biomarkers, SNPs and mutations. In conclusion, data sets obtained with the Smart-Seq protocol provide improved representation of the transcriptomes of individual cells, which should be useful for both basic and clinical studies.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. Gene Expression Omnibus: GSE38495 (sequencing read data).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Burge and G. Winberg for critical reading of the manuscript, T. Juarez and J. Cotton at the University of California San Diego for their help in Internal Review Board protocol preparation and aquisition of clinical samples, A.A. Talasaz and G. Cann for assistance with the Magsweeper, members of the Science for Life laboratory (Stockholm) for assistance with MiSeq sequencer. Y.-C.W. was supported by a fellowship from the Marie Mayer Foundation. L.C.L. was supported by US National Institutes of Health (NIH) K12HD001259. J.F.L. was supported by NIH R33MH87925 and California Institute for Regenerative Medicine (CL1-00502, RT1-01108, TR1-01250, and RN2-00931). R.S. was supported by European Research Council (starting grant 243066), Swedish Research Council (2008-4562), Foundation for Strategic Research (FFL4) and Åke Wiberg Foundation (756194131).

AUTHOR CONTRIBUTIONS

D.R. designed and performed the computational analyses of sequencing reads, prepared figures, tables and methods, and contributed manuscript text. S.L. and R.L. developed protocols and created libraries. I.K. and S.L. did primary data analysis. Y.-C.W., G.A.D. and J.F.L. prepared melanoma circulating tumor cells, melanocytes and melanoma cell line cells. O.R.F. and Q.D. contributed additional sequencing libraries. L.C.L. and G.P.S. contributed to study design and manuscript text. R.S. designed the study and prepared the manuscript, with input from other authors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/nbt.2282.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Mortazavi, A., Williams, B., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
2. Guttman, M. *et al.* Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
3. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
4. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
5. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
6. Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* **34**, e42 (2006).
7. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
8. Tang, F. *et al.* Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468–478 (2010).
9. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
10. Iscove, N.N. *et al.* Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat. Biotechnol.* **20**, 940–943 (2002).
11. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
12. Talasaz, A.H. *et al.* Isolating highly enriched populations of circulating epithelial cells and other rare cells from blood using a magnetic sweeper device. *Proc. Natl. Acad. Sci. USA* **106**, 3970–3975 (2009).
13. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **3**, 74–79 (2011).
14. Jungbluth, A.A. *et al.* Expression of melanocyte-associated markers gp-100 and Melan-A/MART-1 in angiomyolipomas. An immunohistochemical and rt-PCR analysis. *Virchows Arch.* **434**, 429–435 (1999).
15. Tomita, Y., Montague, P.M. & Hearing, V.J. Anti-T4-tyrosinase monoclonal antibodies–specific markers for pigmented melanocytes. *J. Invest. Dermatol.* **85**, 426–430 (1985).
16. Fang, D. & Setaluri, V. Role of microphthalmia transcription factor in regulation of melanocyte differentiation marker TRP-1. *Biochem. Biophys. Res. Commun.* **256**, 657–663 (1999).
17. Chomez, P. *et al.* An overview of the MAGE gene family with the identification of all human members of the family. *Cancer Res.* **61**, 5544–5551 (2001).
18. Tang, A. *et al.* E-cadherin is the major mediator of human melanocyte adhesion to keratinocytes in vitro. *J. Cell Sci.* **107**, 983–992 (1994).
19. Duncan, L.M. *et al.* Down-regulation of the novel gene melastatin correlates with potential for melanoma metastasis. *Cancer Res.* **58**, 1515–1520 (1998).
20. Gudbjartsson, D.F. *et al.* ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. *Nat. Genet.* **40**, 886–891 (2008).

ONLINE METHODS

Generation and amplification of Smart-Seq cDNA. The Smart-Seq cDNA generation and amplification methods developed for this manuscript have recently become available in a kit marketed by Clontech called the SMARTer Ultra Low RNA Kit for Illumina sequencing. Although all the libraries in this manuscript were generated before the kit became commercially available, our protocol is reflected in the detailed instructions for generating cDNA from cell(s) or 100 pg–10 ng of total RNA that is now included in the manual for this kit. For single cell applications, each cell (or control RNA) was added in max 1 µl of media to 4 µl of hypotonic lysis buffer consisting of 0.2% Triton X-100 and 2 U/µl of ribonuclease (RNase) inhibitors (Clontech, 2313B) in RNase free water. The deposition of an intact cell in the hypotonic lysis buffer leads to immediate lysis and stabilization of the RNA through RNase inhibitors. Then, poly(A)$^+$ RNA was reverse-transcribed through tailed oligo(dT) priming using the CDS primer (5′–AAGCAGTGGTATCAACGCAGAGTACT(30)VN–3′, where V represents A, C or G) directly in total RNA or a whole cell lysate using Moloney murine leukemia virus reverse transcriptase (MMLV RT). The first-strand cDNA generation was carried out with the addition of 5× First Strand Buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl and 30 mM $MgCl_2$), dithiothreitol (100 mM), dNTP mix (10 mM), RNAse inihibitor, oligos (CDS primer and SMARTer II A oligo) and SmartScribe Reverse Transcriptase in a total volume of 10 µl (see Clontech manual for details). Once the reverse transcription reaction reaches the 5′ end of an RNA molecule, the terminal transferase activity of MMLV adds a few nontemplated C nucleotides to the 3′ end of the cDNA. The carefully designed SMARTer II A oligo (5′-AAGCAGTGGTATCAACGCAGAGTACATrGrGrG-3′, where r indicate ribonucleotide bases) then base-pairs with these additional C nucleotides, creating an extended template. The reverse transcriptase then switches templates and continues transcribing to the end of the oligonucleotide. The resulting full-length cDNA contains the complete 5′ end of the mRNA as well as an anchor sequence that serves as a universal priming site for second-strand synthesis. The cDNA was then amplified using 12 cycles for 1 ng of total RNA, 15 cycles for 100 pg of total RNA, and 18 cycles for 10 pg total RNA or from single cells. The exact number of cycles used for each dilution replicate or single-cell is detailed in **Supplementary Table 1**. The PCR was performed in 50 µl reaction volumes with Advantage 2 PCR Buffer (Clontech), dNTP mix, PCR primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′), Advantage 2 Polymerase Mix (Clontech) and Nuclease-Free water, resulting in a few nanograms of amplified cDNA. The length distribution of amplified cDNA was monitored using High Sensitivity kits on a Bioanalyzer (Agilent), expecting a distinct peak around 500–5,000 bp (although lengths of mRNAs differ between cell types).

Construction and sequencing of Smart-Seq sequencing libraries. Amplified cDNA (~5 ng cDNA) was used to construct Illumina sequencing libraries using either Illumina's Ultra Low Input mRNA-Seq Guide (the 'PE' protocol) or a modification of Epicentre's Nextera DNA sample preparation protocol (the 'Tn5' protocol). With the PE protocol, the amplified cDNA was fragmented using a Covaris acoustic shearing instrument. The resulting fragments were end-repaired, followed by the addition of a single A base, ligation to Illumina PE adaptors, and then amplification in 12–18 cycles of PCR (depending on starting amounts of RNA, see **Supplementary Table 1** for detailed instructions of all libraries generated). With the Tn5 protocol, the amplified cDNA was 'tagmentated' at 55 °C for 5 min in a 20-µl reaction with 0.25 µl of transposase and 4 µl of 5× HMW Nextera reaction buffer. We added 35 µl of PB to the tagmentation reaction mix to strip the transposase off the DNA, and the tagmentated DNA was purified with 88 µl of SPRI XP beads (sample to beads ratio of 1:1.6). Purified DNA was then amplified by nine cycles of standard Nextera PCR. Library quality was confirmed using DNA 1000 kits on a Bioanalyzer (Agilent), and the libraries were then sequenced on either Illumina's HiSeq 2000, GAIIx or MiSeq instruments, and all clusters that passed filter were exported into fastq files. Details on the sequence depth, sequencing platform and library construction method for each dilution replicate and single cell are included in **Supplementary Table 1**. All data shown in the figures of this manuscript were generated using the PE protocol unless otherwise specified in the figure legend.

Construction and sequencing of standard mRNA-Seq libraries. We generated mRNA-Seq transcriptome data following the Illumina mRNA-Seq kit from 100 ng and 1 µg of total RNA, as detailed in **Supplementary Table 1**.

Isolation of individual CTCs from peripheral blood. Ten milliliters of peripheral blood was collected from a male patient with recurrent, metastatic melanoma using K2 EDTA blood collection tubes (Becton Dickinson). Melanoma CTCs were collected under UCSD IRB #101330, 'Detection and Molecular Characterization of Circulating Melanoma Cells'. The blood sample was processed within 3 h of collection. The erythrocytes in 4.5 ml of the blood sample were lysed with BD Pharm Lyse lysing solution (Becton Dickinson) for 10 min at room temperature. The nucleated cells were pelleted, resuspended in HBSS containing 1% BSA and 5 mM EDTA, pelleted, resuspended in 1 ml of HBSS containing 1% BSA and 5 mM EDTA. The nucleated cells were stained with Phycoerythrin-conjugated anti-human CD45 IgG to label leukocytes. The cells were subsequently reacted with biotinylated anti-human CSPG4 (also known as NG2) mouse IgG at 4 °C for 2 h, washed with HBSS, and reacted with streptavidin-conjugated MG980A magnetic beads at 4 °C for 2 h. The cells were captured based on magnetic sweeping to harvest the beads from cell suspension using the MagSweeper instrument (Illumina) as previously described[12]. The collected cells were stained with 5 µg/ml Calcein AM (Life Technologies) in HBSS for 20 min to identify viable cells. Manual picking of viable cells showing desired Calcein-positive/CD45-negative/bead-attached profile was performed to isolate cells for molecular profiling. The individual cells were placed into 2.5 µl of Superblock (Thermo Scientific) containing 4,000 unit/ml RNase inhibitor (New England Biolabs) and stored at −80 °C until preparation of Smart-Seq libraries.

Isolation of mouse oocytes and human lymphocytes. MII oocytes were isolated from 4-week old CAST/EiJ female mice. Mice were superovulated by injection of 5 IU PMSG, followed by injection of 5 IU of hCG 48 h later. MII oocytes were isolated 14–15 h after hCG treatment by dissection of the ampulla of the oviduct and cumulus cells were removed by hyaluronidase digestion. Single oocytes were manually picked, lysed in dilution buffer, and cDNA constructed as described above. Peripheral blood lymphocytes from healthy human volunteers were isolated on Ficoll gradients using LymphoPrep (Fresenius Kabi, Norway). Individual cells were manually picked into lysis buffer and cDNA constructed as described above.

Alignment of short reads to genome and transcriptome. Reads were independently aligned using Bowtie[21] against the respective genome assembly (hg19 or mm9) and transcriptome sequences (Ensembl, human and mouse annotations were downloaded 16 May 2011 and 13 December 2010, respectively). Transcriptome mapped reads were converted from transcriptome coordinates to genomic coordinates and thereafter compared with the genome mapped reads to identify reads that map to a unique genomic location. This procedure ensured that mapped reads were unique across both the genome and transcriptome, while allowing for reads to map to different transcripts of the same gene in the initial transcriptome mapping. The uniquely mapped reads were converted to binary BAM files using Samtools[22]. The resulting transcriptome data were visualized using the Integrated Genome Viewer (IGV, Broad Institute) using the histogram visualization for **Supplementary Figure 3** and heatmap visualization for **Figure 3c**.

Expression level estimation and technical comparisons of sensitivity and variation. Gene expression levels for Refseq transcripts were summarized as RPKM values and read counts using rpkmforgenes[23]. RefSeq annotations for human and mouse were downloaded on the 31 August 2011 and 13 December 2010, respectively. RPKM calculations only considered uniquely mappable positions for transcript length normalizations using the ENCODE Mappability track (wgEncodeCrgMapabilityAlign50mer.bigWig) for human and in-house–computed uniqueness files for mouse. Overlapping RefSeq transcripts were collapsed giving one expression value per gene locus. Only 10 million randomly selected mapped reads were used per sample to compare sensitivity and variation in gene and exon levels. Samples with fewer than 10 million uniquely mappable reads (a few ESCs[8]) were therefore discarded from analyses. Samples with 20 pg of total RNA (used in **Fig. 2b**,**d**) were simulated by using 5 million reads each of two different 10 pg samples. Analyses of gene detection (**Fig. 2a**,**b** and **Supplementary Fig. 4b**,**c**) were calculated over pairs of technical replicates or individual cells. Genes were binned by the highest expression level of the two samples, and was considered detected if it had an

RPKM above 0.1 in both samples. The mean for all possible pairs of technical replicates within a group was used together with standard deviation using the adjusted Wald method. Analyses of variation (**Fig. 2b**,**d**) were also calculated on pairs of samples, binning genes by the mean of log expression, excluding genes below 0.1 RPKM in either sample. As gene expression levels across single cells are often log normally distributed[24], we calculated absolute difference in $\log_{10}$ expression values and s.d. by multiplying mean variation in a bin with 0.886. Scatter plots were generated in R using smoothScatter (geneplotter package) and loess nonlinear regression using the graphics package. Pearson and Spearman correlations were computed using absolute or relative expression levels as $\log_2$ RPKM values. We included publicly available human UHRR, brain and LNCaP data for comparison[4,25–27]. To analyze how sensitivity and variation improve with a larger numbers of cells, we used Smart-Seq data generated from 10 LNCaP cells (**Supplementary Table 1**). To obtain estimates for the effect of using larger numbers of cells (used in **Fig. 2b**,**d**), we created two combined samples using 25, 10 and 3 cell samples from picked LNCaP cells, 25, 10 and 5 cell samples from LNCaP cells spiked into healthy donor's blood and isolated using the EPCAM marker, and 2 single-cell LNCaP samples, achieving a total of 80 cells per each of the two sample pools. These were sequence-depth matched to 10 millions reads, by using 125,000 random reads from single-cell samples, 375,000 from 3-cell samples and so on.

Analyses of read coverage across transcriptome. The read coverage analyses were based on human and mouse RefSeq transcripts. Reads were mapped to RefSeq transcripts directly rather than to the genome, using Bowtie allowing for up to 10 hits per read. Each transcript was divided into 40 equally sized bins, and the number of reads was counted for each bin and gene. The read count per bin for each gene was divided by total read count for that gene before the bins for all the different genes were summed up. The calculated read coverage per bin was later normalized through the division by the bin with the largest read coverage. The mean and s.d. over replicates were shown in **Figure 1** and **Supplementary Figure 2**, including all transcripts with at least ten mapped reads. Analyses of full-length transcript reconstructions were based on RefSeq annotations, and we defined full-length reconstructed genes as those for which we obtained correct exon-intron structure throughout all annotated exons of at least one isoform. We limited the analyses to expressed (≥0.1 RPKM) and multi-exon (≥2 exons) genes.

Singular value decomposition. The global transcript expression values for cancer cells were analyzed using singular value decomposition (SVD) to determine the fundamental patterns in the transcriptomes. The expression levels in RPKM were normalized to unit length and the SVD computed using SVDMAN[28]. Each cell was then projected onto the two strongest SVD components to visualize the overall similarity in gene expression (**Fig. 3a**).

Analyses of differential expression. One-way analysis of variance (ANOVA) was performed on expression levels (RPKM, log2) followed by Tukey post-hoc test in R/Bioconductor. Only genes significant after multiple testing corrections (5% FDR, Benjamini-Hochberg) were evaluated with post-hoc test ($P < 0.05$). Lists of significantly differently expressed genes are available in **Supplementary Table 4** for CTC, primary melanocyte and melanoma cell line comparisons, and in **Figure 3a** for comparisons between prostate and bladder cancer cell line cells.

Selection of marker genes for melanoma and immune cells. To identify the 100 transcripts most strongly associated with melanoma and immune cells, respectively, we initially calculated the difference in mean gene expression between melanoma samples[29] and a combination of monocytes[30], T cells[31], white blood cells and lymph node samples (**Fig. 4a**). The differences were divided by the highest expression value in any of the samples, to avoid differences driven by outlier expression values in one replicate only. We ranked genes according to this metric and selected the 100 strongest markers for the melanoma and for the immune cell combination. We then evaluated the mean expression values of each gene in the individual putative CTCs. To include the monocyte SAGE data, we converted 1.5 RPM to 1 RPKM, assuming an average transcripts length of 1.5 kb (ref. 23).

Detection of alternatively spliced exons. We analyzed exon inclusion levels for a collection of alternatively skipped exons previously identified from EST and cDNA data[4]. We used the mixture of isoforms (MISO) framework[11] to calculate exon inclusion levels with confidence intervals. We used the default MISO settings, which require at least 20 reads mapping to the alternative exon or the immediate upstream or downstream exon or exon-exon junctions between them. For a fair assessment of read coverage across exons (**Fig. 3b**), we matched the sequence depth by randomly sampling 10 million uniquely mapped reads per sample.

Hierarchical clustering analyses. Genes with average expression above 20 RPKM (3,690 genes) were clustered by Spearman correlation and complete linkage using python scipy (hcluster). To evaluate the significance (or robustness) of each branchpoint, we generated thousand bootstrap gene set replicates that were independently clustered, and from these we counted the percentage of times each branch was recovered.

Analyses of differential exon inclusion. To find significant differences in inclusion levels of alternative exons we applied a *t*-test with variance shrinkage, known to counteract false positives in microarray analyses[32]. A variance was calculated for each alternative exon based on the exon inclusion levels across biological replicates. For each sample group (cell line) the 90th percentile of the variation was included in the variance term ((90th percentile variation + gene variation)/2) when calculating the *t* statistic. The null distribution of the *t* statistic was calculated by shuffling the sample labels (cell-to-cell line mapping) repeatedly and for each shuffle compute the *t* statistics, thus allowing the conversion of *t* statistics to *P* values for the cancer-cell comparison. To estimate false discovery rates, the sample groups were randomly split in half and combined with half from the other sample group, and the number of significant exons was counted using the *t* statistics introduced above (repeatedly, to vary the random splitting of sample groups). The false discovery rate was then estimated as the number of significant exons in random shuffles divided by the number of significant events with correct sample groups. The numbers of significant exons at different false discovery rates are presented in **Figure 3d**.

SNP and mutation detection. CTC RNA-Seq Fastq files were mapped to transcriptome (Ensembl, annotations downloaded 16 May 2011) and genome with BWA[33], allowing for no indels and removing multi-mapping reads. Samtools rmdup[22] was used to filter PCR duplicates, and BAM files were reordered by Picard (http://picard.sourceforge.net/). Variant sites were called by the Genome Analysis Toolkit[34] jointly on reads from all six CTC samples, with a quality score threshold for sites of 500 and requiring detection in two or more samples (see **Supplementary Fig. 11** for more detailed information on varying these threshold). We limited the analyses to transcribed regions using RefSeq gene models, and the last 35 base pairs of transcripts were not considered to remove false positives arising from mapping of reads with partial poly(A) tail. Analyses of overlap with known SNPs were based on dbSNP build 132 (ref. 35).

21. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
23. Ramsköld, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLOS Comput. Biol.* **5**, e1000598 (2009).
24. Bengtsson, M., Ståhlberg, A., Rorsman, P. & Kubista, M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* **15**, 1388–1392 (2005).
25. Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* **38**, 4570–4578 (2010).
26. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
27. Sam, L.T. *et al.* A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS ONE* **6**, e17305 (2011).
28. Wall, M.E., Dyck, P.A. & Brettin, T.S. SVDMAN–singular value decomposition analysis of microarray data. *Bioinformatics* **17**, 566–568 (2001).
29. Berger, M.F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.* **20**, 413–427 (2010).

doi:10.1038/nbt.2282

30. Zawada, A.M. *et al.* SuperSAGE evidence for CD14.CD16+ monocytes as a third monocyte subset. *Blood* **118**, e50–e61 (2011).
31. Bernstein, B.E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
32. Allison, D.B., Cui, X., Page, G.P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **7**, 55–65 (2006).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
34. McKenna, A. *et al.* The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
35. Sherry, S.T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).