Single-cell Transcriptomics Uncovers Distinct Molecular Signatures of Stem Cells in Chronic Myeloid Leukemia

Giustacchini, A.,[1,2*] Thongjuea, S.,[1,2*] Barkas, N.,[1,2] Woll, P.,[2] Povinelli, B.,[1,2] Booth, C.,[1,2] Sopp, P.,[1] Norfo, R.,[1,2] Rodriguez-Meira , A., [1,2] Ashley, N., [1,2] Jamieson, L.,[1,2] Vyas, P.,[1] Anderson, K.,[3] Segerstolpe, Å.,[4, 5] Qian, H., [6] Olsson-Strömberg, U.,[7] Mustjoki, S.,[8] Sandberg, R.,[4, 9] Jacobsen, S.E.W.[1,2,4,6,10+] and Mead, A.J.[1,2,11+]

1. MRC Molecular Hematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom
2. Haemopoietic Stem Cell Laboratory, Weatherall Institute for Molecular Medicine, University of Oxford, Oxford, United Kingdom
3. Department of Cellular Therapy, Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway
4. Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden
5. Integrated Cardio Metabolic Center (ICMC), Karolinska Institutet, Huddinge, Sweden
6. Department of Medicine, Center for Hematology and Regenerative Medicine, Karolinska Institutet, Stockholm, Sweden
7. Department of Hematology, Uppsala University Hospital, Uppsala, Sweden
8. Hematology Research Unit Helsinki, Department of Clinical Chemistry and Hematology, University of Helsinki and Helsinki University Hospital Comprehensive Cancer Center, Helsinki, Finland
9. Ludwig Institute for Cancer Research, Stockholm, Sweden
10. Karolinska University Hospital, Stockholm, Sweden
11. NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

* or + These authors contributed equally to this work

Corresponding authors: Adam J. Mead (adam.mead@imm.ox.ac.uk) and Sten Eirik W. Jacobsen (sten.eirik.jacobsen@ki.se or sten.jacobsen@imm.ox.ac.uk)

Abstract

Recent advances in single-cell transcriptomics are ideally placed to unravel intratumoral heterogeneity and selective resistance of cancer stem cell (SC) subpopulations to molecularly targeted cancer therapies. However, current single-cell RNA-sequencing approaches lack the sensitivity required to reliably detect somatic mutations. We developed a method combining high-sensitivity mutation detection with whole-transcriptome analysis of the same single-cell. We applied this technique to analyze over 2000 SCs from chronic myeloid leukemia (CML) patients throughout the disease course, revealing heterogeneity of CML-SCs, including the identification of a subgroup of CML-SCs with a distinct molecular signature that selectively persisted during prolonged therapy. Analysis of non-leukemic SCs from CML patients also provided new insights into cell-extrinsic disruption of hematopoiesis in CML associated with clinical outcome. Furthermore, we used this single-cell approach to identify a blast crisis specific SC population which was also present in a subclone of CML-SCs during chronic phase in a patient who subsequently developed blast crisis. This approach, which might be broadly applied to any malignancy, illustrates how single-cell analysis can identify subpopulations of therapy-resistant SCs that are not apparent through cell-population analysis.

Introduction

Molecularly targeted therapies for cancer frequently induce impressive remissions, however, complete disease elimination remains rare, and patients remain at risk of disease relapse. At a cellular level this is likely to reflect intratumoral heterogeneity, with differential response to treatment in distinct tumor subpopulations[1]. This phenomenon relates to the proposed hierarchical organization of some tumors, with only rare “cancer stem cells” (CSCs) being capable of tumor propagation.[2-4] There is now ample evidence for the existence of such rare CSCs in some tumors, subsets of which are therapy-resistant and persist during remission.[2-4] However, studies characterizing CSCs during remission are lacking, reflecting in part that these residual CSCs are typically rare and outnumbered by their normal tissue counterparts from which they cannot easily be separated.[5,6]

Advances in single-cell gene expression techniques offer great potential to study the CSC heterogeneity that might underlie therapy resistance.[1,7-9] However, to date, the application of single-cell RNA-sequencing in cancer has been relatively limited in patients achieving remission following therapy,[1,7-12] partly because detection of somatic mutations is grossly underappreciated using current techniques.[11] This primarily relates to poor coverage in the RNA sequencing reads from single cells across the specific mutated region of a gene due to both technical dropouts and stochastic gene expression in individual cells.[8] Thus, it is difficult to simultaneously apply single-cell transcriptome analysis with highly sensitive detection of specific mutations; the latter is essential to reliably distinguish normal cells from somatically mutated cells which form part of the malignant clone. This is of particular importance when analyzing CSC during remission, when malignant cells are rare and may largely share transcriptomic features with normal tissue counterparts.

Chronic Myeloid Leukemia (CML) is a paradigm for molecularly targeted therapy and an ideal disease to explore the cellular basis of selective resistance to targeted therapy.[13,14] CML is less genetically complex than most cancers and is defined by presence of the BCR-ABL fusion gene, the product of which is the target of tyrosine kinase inhibitor (TKI) treatments which have dramatically improved outcomes for this disease.[15] However, chronic-phase CML (CP-CML) is propagated by rare CML stem cells (CML-SCs) that are selectively resistant to TKI therapy and are incompletely eradicated in most patients,[16,17] leading to frequent relapse following treatment discontinuation.[18] CML-SCs reside in the same phenotypic compartment as their normal hematopoietic stem cell (HSC) counterparts, and both express a $CD34^{+}CD38^{-}$ surface phenotype[5,6]. Techniques to selectively analyze BCR-ABL^{+} SCs throughout the disease course are not currently available. It therefore remains to be established whether therapy-resistant CML-SCs following TKI therapy represent the stochastic persistence of heterogeneous CML-SCs, a selective persistence of a pre-existing distinct therapy-resistant CML-SC subset, or a resistant CML-SC with novel properties that evolved as a result of the

therapeutic selection process.

In addition, there is ample evidence in hematological malignancies that dysregulated hematopoiesis occurs as much through extrinsic disruption of the normal-HSC compartment as through intrinsic expansion of the leukemic clones[19-21]. For example, recent evidence from mouse models supports involvement of non-clonal BCR-ABL$^-$ SCs in the CML disease phenotype[22,23]. However, in the absence of single-cell analysis enabling separation of BCR-ABL$^-$ and BCR-ABL$^+$ SCs within individual patients, it remains unclear to what degree disruption of BCR-ABL$^-$ SCs occurs in CML patients and how disruption of the non-clonal SC compartment might correlate with response to treatment.

Herein, we developed a new protocol integrating fluorescence activated cell sorting (FACS), high sensitivity single-cell mutation detection and single-cell RNA-sequencing. We apply this method to characterize distinct molecular signatures of SC subpopulations in human CML samples from diagnosis through remission and disease progression.

Results

Combined single-cell mutation detection and transcriptomics

Presence of the BCR-ABL fusion gene remains the only unequivocal marker of CML-SCs and we therefore first sought to determine the sensitivity of BCR-ABL detection using Smart-seq2, a commonly used single-cell RNA-sequencing approach,[8,24,25] by analysing the BCR-ABL$^+$ K562 cell-line.[26] BCR-ABL transcripts were not detected in as many as 18/24 cells (75%; Fig. 1a), despite generation of satisfactory cDNA libraries as determined through bioanalyser analysis of the size and concentration of amplified cDNA libraries (Supplementary Fig. 1a). We obtained a similar result using a commercial nanofluidic platform[27] (Supplementary Fig. 1b) and across a range of other myeloid leukemia mutation hotspots (Supplementary Fig. 1c), validating that current single-cell RNA-sequencing techniques do not enable sensitive

mutation detection.[10]

To improve sensitivity of BCR-ABL detection, we developed a BCR-ABL targeted Smart-seq2 protocol (BCR-ABL tSS2) (Supplementary Fig. 2a-d). By multiplexing BCR-ABL specific primers at the reverse transcription and amplification steps, BCR-ABL-detection was improved to 100% of K562 cells in plate-based (Fig. 1b) or microfluidic-based platforms (Supplementary Fig. 1d). Importantly, there was no evidence of bias caused by BCR-ABL tSS2 in relation to library quality (Supplementary Fig. 2d), with good correlation between level of expression of 14,240 RefSeq genes (Fig. 1c) generated by Smart-seq2 or BCR-ABL tSS2; these samples also did not show separate clustering (Fig. 1d; Supplementary Fig. 3). BCR-ABL plasmid "spike-in" experiments demonstrated sensitivity to detect single molecules of BCR-ABL with expected Poisson distribution. Importantly, no BCR-ABL amplification was observed from any negative control cells in this (Fig. 1e), or in any subsequent experiment (n=232 cells). This BCR-ABL tSS2 method therefore allows highly specific, sensitive and quantitative BCR-ABL detection with parallel unbiased whole transcriptome analysis from the same single-cell.

Single-cell RNA-sequencing and BCR-ABL detection in CML Stem Cells

As human HSCs are small and highly quiescent cells compared to K562 cells, we first analyzed 232 $Lin^-CD34^+CD38^-$ BM cells from five healthy human donors (normal-HSCs) using BCR-ABL tSS2. Satisfactory cDNA libraries were generated (Supplementary Fig. 4a), with a plateau in the numbers of genes detected above $1x10^6$ mapped reads/cell (Fig. 2a). With an average sequencing depth of $3.4x10^6$ mapped reads, a mean of 3,445 genes were detected in each cell (Supplementary Fig. 4b). 12,018 genes were detected (RPKM ≥1) in single-cell ensembles (sequencing data from all 232 cells was pooled *in silico*) correlating well with cell-population data (Supplementary Fig. 4c) and with sensitivity to detect low level expressed transcripts (Supplementary Fig. 4d), in line with previous reports.[28] Human HSCs clustered separately and were more heterogeneous than K562 cells (Fig. 2b; Supplementary Fig. 4e). Importantly, independently-processed cells from five

different donors clustered together, illustrating stability of the data across independent experiments (Fig. 2b).

We next analyzed 40 $Lin^{-}CD34^{+}CD38^{-}$ SCs from a CML patient in hematologic remission following 3 months of TKI therapy (OX1407; Supplementary Table 1). BCR-ABL was detected in 17/40 cells (43%) by BCR-ABL tSS2 and in 7/20 cells (35%; P=0.8) by single-cell fluorescent-in-situ-hybridization. We detected 12,499 genes in data ensembles, which correlated well with bulk analysis data (Fig. 2c). Comparison of $BCR\text{-}ABL^{+}$ and $BCR\text{-}ABL^{-}$ SCs identified genes showing differential expression (Fig. 2d-e). Level of expression correlated well between single-cell RNA-sequencing and QPCR data (Fig. 2f-h; Supplementary Fig. 5a-c). Together, these data provide proof of principle that BCR-ABL tSS2 can be applied to detect distinct gene expression in $BCR\text{-}ABL^{+}$ versus $BCR\text{-}ABL^{-}$ SCs from the same patient during TKI treatment.

Single-cell RNA-sequencing of CML-SCs at diagnosis

We next used BCR-ABL tSS2 to process 2070 $Lin^{-}CD34^{+}CD38^{-}$ BM SCs from diagnosis samples from 20 patients with CP-CML (Supplementary Table 1). Two of these CP patients developed early progression to blast crisis (BC) and these cells were removed from the current analysis and analyzed in later experiments (Figure 6). As previously reported[29], although the progenitor compartment was disrupted in CML patients, the HSC-containing $Lin^{-}CD34^{+}CD38^{-}$ compartment was relatively intact phenotypically (Supplementary Fig. 6). As expected[5,6,30], frequency of $BCR\text{-}ABL^{+}$ SCs was variable (median: 69%, 9%-94%; Supplementary Table 1).

We selected 854 CP-CML-SCs (477 $BCR\text{-}ABL^{+}$ and 377 $BCR\text{-}ABL^{-}$) for sequencing and detected a mean of 3591 genes/cell (Supplementary Fig. 7a). Read depth and mapped reads/cell were not different between normal-HSCs (n=232), $BCR\text{-}ABL^{-}$ SCs and $BCR\text{-}ABL^{+}$ SCs (Supplementary Fig. 7b,c). Expression of housekeeping genes e.g. *B2M* was also comparable in the three groups (Supplementary Fig. 7d). In contrast, whilst the mean number of

genes detected was comparable between normal-HSCs (n=3,445) and BCR-ABL$^-$ SCs (n=3,409), a significantly higher number of genes was detected in BCR-ABL$^+$ SCs (n=3,735, P=1.67e-06, Supplementary Fig. 7e). This correlated with BCR-ABL driven proliferation, as markedly increased proliferation gene expression (Fig. 3a) and reduced quiescence-associated gene expression (Fig. 3b) was observed in BCR-ABL$^+$ SCs in comparison with normal-HSCs. In contrast, BCR-ABL$^-$ SCs showed similar proliferation (Supplementary Fig. 7f) and quiescence-associated gene expression (Supplementary Fig. 7g) as normal-HSCs. Consequently, co-expression of G2M-associated genes was selectively increased in BCR-ABL$^+$ SCs (Supplementary Fig. 7h).

t-Distributed Stochastic Neighbor Embedding (tSNE) analysis using 8,589 highly-variable genes revealed distinct clustering of normal-HSCs, BCR-ABL$^+$ and BCR-ABL$^-$ SCs (Fig. 3c). Differentially expressed genes between normal-HSCs, BCR-ABL$^+$ and BCR-ABL$^-$ SCs included many that were previously implicated in CML pathogenesis (Supplementary Fig. 8a; Supplementary Table 2) but also a number of novel candidate genes of interest such as *RXFP1, RAB31, SRSF2* and *LGALS1* (Supplementary Fig. 8b; Supplementary Table 2). *In silico* generation of cell-ensemble data demonstrated that very few of these differentially expressed genes would have been revealed without single-cell analysis (Supplementary Fig. 8c). Using the top 245 differentially expressed genes, BCR-ABL$^+$ cells clustered separately from BCR-ABL$^-$ SCs, importantly without evidence of major patient-specific clustering (Fig. 3d), reflecting consistency of aberrant gene expression in BCR-ABL$^+$ SCs across different patients (Supplementary Fig. 9a,b).

Comparison of BCR-ABL$^+$ SCs versus normal-HSCs and/or BCR-ABL$^-$ SCs showed expected enrichment in BCR-ABL$^+$ SCs for the large majority of established CML stem/progenitor gene-sets (Supplementary Tables 3 and 4; Supplementary Fig. 10). Analysis using unbiased gene-sets (Supplementary Table 5; Fig. 3e), uncovered multiple gene-sets selectively enriched in BCR-ABL$^+$ SCs (e.g. overexpression of *MTORC*, E2F-targets, G2M-checkpoint,

oxidative phosphorylation and glycolysis associated gene expression; Supplementary Table 5; Fig. 3e), none of which showed enrichment through *in silico* bulk analysis of the same dataset.

Importantly, our single-cell approach also uniquely allowed analysis of BCR-ABL$^-$ SCs within the same patients, of relevance for recent evidence that the microenvironment is disrupted in CML mouse models[22,23]. IL6-associated gene expression and downstream mediators such as STAT5a were indeed significantly enriched in BCR-ABL$^-$ SCs in comparison with normal HSCs (Fig 3e, Supplementary Fig. 10 and Supplementary Tables 4 and 5). Furthermore, other inflammation associated gene expression, including TGFβ and TNFα pathways were also markedly enriched in BCR-ABL$^-$ SCs in comparison with normal HSCs (Fig. 3e). Inflammation is an important suppressor of HSC function[31,32], including TGFβ and TNFα which are notably both cell-extrinsic suppressors of HSCs[33,34].

Single-cell RNA-sequencing of CML-SCs predicts TKI response

Next, to establish the potential clinical utility of CML-SC single-cell gene expression signatures, in line with current guidelines,[35] we stratified patients with sufficient response data available as good (n=11) or poor (n=5) responders on the basis of subsequent achievement of a major molecular response (MMR) to TKI, defined as a BCR-ABL transcript level <0.1% (Supplementary Table 1). There was no significant difference in the frequency of BCR-ABL$^+$ SCs between good (61%) and poor (58%) responders (P=0.7). While BCR-ABL$^+$ SCs at diagnosis did not clearly cluster according to response category (Fig.4a), BCR-ABL$^-$ SCs from poor-responder patients showed highly distinct clustering using 5,611 highly-variable genes (Fig. 4b). Notably, in all 5 CML patients failing to achieve MMR, the frequency of BCR-ABL$^-$ SCs contained within the poor-responder cluster were higher than for all 11 patients that achieved MMR, in 4 cases with virtually all BCR-ABL$^-$ SCs falling within the poor-responder cluster (Fig. 4c). The five patients with >10% of BCR-ABL$^-$ SCs falling within the poor-responder cluster had a markedly inferior likelihood of achieving MMR (Fig. 4d, P<0.01).

GSEA also showed enrichment at diagnosis of signaling pathways, inflammation, TGFβ and TNFα associated gene expression in BCR-ABL$^-$ SCs from poor as compared to good responders (Fig. 4e and Supplementary Table 6). In contrast, both BCR-ABL$^-$ and BCR-ABL$^+$ SCs from good responders showed enrichment of MYC, E2F and G2M checkpoint gene expression, associated with increased proliferation (Fig. 4e and Supplementary Table 6). These data demonstrate that BCR-ABL$^+$ as well as BCR-ABL$^-$ SCs in poor responding patients are already at diagnosis expressing more quiescence associated genes than in patients who will later achieve MMR; as this was observed for both BCR-ABL$^+$ as well as BCR-ABL$^-$ SCs, this may reflect differences in cell-extrinsic, microenvironmental factors in good versus poor responders.

As poor responders showed upregulation of TGFβ and TNFα associated gene expression, combined with a highly quiescent CML-SC signature, we reasoned that TGFβ and TNFα might promote quiescence in the CML-SC compartment and thereby confer TKI resistance. We therefore cultured single normal-HSCs and CML-SCs *in vitro* with or without TGFβ or TNFα and tracked the time taken for the SCs to divide. TNFα promoted quiescence of both CML-SCs and normal-HSCs (Supplementary Fig.11). Notably, TGFβ more strongly impacted on the rate of cell division of CML-SCs compared to normal-HSCs (Supplementary Fig.11). Together, these data highlight the power of single-cell (unlike bulk) RNA-sequencing of CML-SCs at diagnosis to reveal gene expression patterns in leukemic as well as non-leukemic SCs within the same patient.

Characterisation of quiescent CML-SCs persisting during TKI therapy

We next analyzed 19 patients who had already commenced TKI therapy and had achieved at least a hematological remission (normalization of blood counts) and with additional cytogenetic response in most patients (Supplementary Table 1). In 11 of these patients, paired diagnosis and follow-up BM samples were available following either 3 or 6 months of TKI (Table 1).

In follow-up samples, the percent of BCR-ABL$^+$ SCs (median: 9%, 0%-82%) was lower than in diagnosis samples from the same patient (P=0.0001; Supplementary Table 1). From a total of 3,306 cells processed, we selected 245 BCR-ABL$^+$ SCs and 420 BCR-ABL$^-$ SCs for single-cell sequencing. Notably, unlike in diagnosis samples, the average number of genes detected in each cell was similar between BCR-ABL$^+$ (n=3,284) and BCR-ABL$^-$ SCs (n=3,196) in follow-up samples.

Using the top 500 genes informative for distinguishing normal-HSCs from BCR-ABL$^+$ SCs at diagnosis and during remission (Supplementary table 7), tSNE analysis revealed two distinct clusters of remission BCR-ABL$^+$ SCs (group-A and group-B; Fig. 5a). Group-A remission BCR-ABL$^+$ SCs were enriched for quiescence and HSC-associated gene expression whereas group-B showed enrichment of MYC, E2F and proliferation-associated gene sets (Fig. 5b and Supplementary Table 8). Group-A cells were progressively enriched with more prolonged TKI treatment, accounting for 43% of BCR-ABL$^+$ SCs at 3 months and 84% at ≥1year (P<0.01; Fig. 5c). This enrichment for group-A cells was even more striking when only including patients subsequently achieving MMR with 65% and 91% of BCR-ABL$^+$ SCs falling within group-A at 3 months and 1 year following initiation of TKI treatment respectively (P<0.01; Fig. 5d; Supplementary Fig. 12a). The only exceptions were one patient who temporarily interrupted TKI therapy and a patient failing to achieve therapeutic imatinib levels, in both cases showing predominantly group-B SCs at 3 months (Supplementary Fig. 12b). This supports the concept that an excess of group-B cells during TKI therapy identifies patients with inadequate BCR-ABL inhibition. We also noted that in 15 of 18 (83%) diagnosis samples a minority of BCR-ABL$^+$ SCs clustered within group-A (26% of all diagnosis BCR-ABL$^+$ SCs; Fig. 5a, c and d and Supplementary Fig. 12), although the frequency of group A cells at diagnosis did not correlate with response to TKI in this small cohort of patients. Together, these data suggest that prolonged TKI treatment results in the selective persistence of a distinct and highly quiescent BCR-ABL$^+$ CML-SC subset (group-A) already present at diagnosis, rather than a stochastic persistence of heterogeneous

CML-SCs or a resistant CML-SC with novel properties. To better understand the selective persistence of quiescent CML-SCs during long-term TKI treatment we therefore focused subsequent analysis on remission group-A cells.

Most group-A BCR-ABL$^+$ SCs clustered separately from normal-HSCs (Fig. 5a); we detected 1,086 differentially expressed genes in group-A remission BCR-ABL$^+$ SCs in comparison with normal-HSCs (Supplementary Table 9). We also detected 1681 and 1348 differentially expressed genes in group-A remission BCR-ABL$^+$ SCs in comparison with group-B remission BCR-ABL$^+$ SCs and BCR-ABL$^+$ SCs at diagnosis respectively (Supplementary Table 9). In comparison with normal HSCs, Group-A BCR-ABL$^+$ SCs showed enrichment of TGFβ, TNFα via NFKB and IL6-JAK-STAT associated gene expression whereas E2F, G2M checkpoint and MYC associated gene expression was enriched in normal-HSCs (Fig. 5e and Supplementary Fig. 13 and Supplementary Table 10). Similar findings were obtained by comparing group-A remission cells with BCR-ABL$^-$ SCs during TKI treatment (Supplementary Fig. 13). These findings support that group A remission CML-SCs, characterized by marked quiescence-associated gene expression, selectively evade eradication by TKI. These cells show more quiescence-associated gene expression than normal HSCs or BCR-ABL$^-$ SCs during remission; likely because the latter are intrinsically much less sensitive to TKIs due to absence of BCR-ABL expression. TGFβ and TNFα via NFKB associated gene expression was progressively more enriched within remission group-A BCR-ABL$^+$ SCs during the course of TKI treatment (Fig. 5f), supporting that these pathways may be important to sustain this resistant and quiescent CML-SC population during TKI treatment. Remission group-A BCR-ABL$^+$ SCs, also showed overexpression of Wnt/β-Catenin pathway genes (*GAS2* and *CTNNB1*), the TGF-β pathway gene *SKIL*, regulators of NF-κB (*NFKB1A* and *p62/SQSTM*), the hypoxia factors *HIF1A* and the WT1 partner *WTAP* as well as downregulation of the chemokine receptor *CXCR4* and the transcription factor *FOS* in comparison with normal HSCs (Fig. 5g and Supplementary Table 9). This single-cell analysis provides insight into

pathways that may be involved in promoting the selective persistence of distinct BCR-ABL$^+$ SCs following TKI treatment.

Analysis of CML-SC heterogeneity during blast crisis

We next analyzed three patients with lymphoid (n=2) or myeloid (n=1) BC transformation of CML (Supplementary Table 1), to explore the possibility that single-cell sequencing of BCR-ABL$^+$ CML-SCs could already in CP predict a subsequent BC transformation. At the time of BC, tSNE analysis of CML-SCs revealed a separate cluster of BCR-ABL$^+$SCs, clearly distinct from both normal-HSCs, BCR-ABL$^+$SCs from 18 CP-CML patients at diagnosis and K562 cells (Fig. 6a). Notably, myeloid and lymphoid blast crisis BCR-ABL$^+$SCs clustered together. Comparison of gene expression of the BC and CP BCR-ABL$^+$SC clusters revealed 1,166 differentially expressed genes (Fig. 6b and Supplementary Table 11), including overexpression of HGF[36] and reduced expression of the Wnt pathway negative-regulator EAF2[37].

Two of the patients who developed BC following TKI initiation also had samples available from diagnosis, when the patients presented in CP (pre-BC) 12 months and 3 months before transformation to myeloid and lymphoid BC, respectively (Supplementary Table 1). All pre-BC cells from the patient transforming to myeloid-BC 12 months later clustered with other CP-CML SCs (CP-CML cluster, Supplementary Fig. 14a). However, the pre-BC SCs from the patient who 3 months later developed lymphoid BC fell into 2 distinct groups, one clustering close to the BC-SCs (BC cluster, n= 124), but notably with a minority (n=8) clustering separately from the BC cluster within the CP-CML cluster (Fig. 6c), providing direct evidence of evolution from CP to BC within the SC compartment of this patient, before any clinical or morphological evidence of development of BC. In further support of this, the pre-BC single BCR-ABL$^+$SCs cells falling within the BC cluster showed aberrant co-expression of myeloid and lymphoid genes in comparison with normal HSCs or CP-CML-SCs, as did cells within the BC cluster from all the 3 investigated BC patients, whereas none of the pre-BC BCR-ABL$^+$SCs cells clustering with the CP CML-SCs showed this aberrant co-expression pattern (Fig. 6d) with

validation of a number of aberrant expressed genes by single-cell QPCR (Supplementary Fig. 14b). Moreover, index-sorting analysis (allowing specific FACS data of individual cells to be linked with gene expression data from the same cell) of the rare pre-BC cells in the CP-SC cluster showed that they all resided within the normal $Lin^-CD34^+CD38^-CD90^+CD45RA^-$ HSC compartment, whereas in contrast 62 of 68 of the pre-BC SCs falling in the BC-SC cluster had a distinct $Lin^-CD34^+CD38^-CD90^-CD45RA^+$ phenotype (Fig. 6e). Indeed, in contrast to CP-CML patients (Supplementary Fig. 6), all BC patients analyzed showed a marked expansion of $Lin\text{-}CD34^+CD38^-CD90^-CD45RA^+$ lymphoid-primed multipotent progenitor (LMPP) like cells (Supplementary Fig. 15), a population previously implicated to propagate acute leukemia[38].

Finally, to explore a possible genetic basis for clonal evolution within the pre-BC SCs, we carried out exome sequencing of the patient with early lymphoid BC, which revealed a somatic *RUNX1* mutation (c.G521A, Supplementary Fig. 16a). In order to track acquisition of the *RUNX1* mutation within the BCR-ABL^+ SC compartment, we carried out parallel targeted amplification of both BCR-ABL and the *RUNX1* mutation. All 4 pre-BC cells falling in the CP-SC cluster, were *RUNX1* wild-type. In contrast, all *RUNX1* mutated pre-BC SCs (n=43) were found within the BC-SC cluster ($P<0.01$). This distinct distribution of the *RUNX1* mutation was confirmed by single-cell QPCR (Supplementary Fig. 16b,c). Furthermore, differentially expressed genes between the pre-BC CP-SC and BC-SC clusters were typically RUNX1 target genes (Fig. 6f). These findings are consistent with acquisition of a *RUNX1* mutation as a key genomic event occurring during pre-BC, driving subsequent BC-transformation at least in this one patient, with expansion of lympho-myeloid transcriptionally primed LMPP-like SCs preceding the clinical BC. This is also consistent with our observed expansion of this distinct SC population in both myeloid and lymphoid BC patients (Supplementary Fig. 15). These data illustrate how integrated single-cell gene expression, mutational profiling and index sorting can be used to unravel CSC heterogeneity and reveal insights that may help predict and understand subsequent disease progression.

Discussion

Single-cell gene expression approaches offer great promise to explore the cellular heterogeneity that might underlie therapy resistance and disease progression in cancer,[1-3,8,10,11,16,17], not-the-least in rare CSC populations, of crucial importance as therapeutic elimination of all CSCs is not only required but might also be sufficient to cure cancers[3]. However, lack of coverage in the RNA-sequencing data has precluded parallel mutation analysis[10,11], representing a major limitation with current techniques.

We used CML as the disease model for single-cell CSC analysis as the identity of the CSC-compartment is well-established[39], and rare CML-SCs persisting during therapy remain a key challenge[16]. Although certain cell surface markers have been proposed to allow for selective enrichment of CML-SCs[40-42], they are not reproducible across all patients nor do they allow effective purification of BCR-ABL$^+$ CML-SCs during remission. In reality, the presence of BCR-ABL remains the only unequivocal marker of CML-SC. Therefore, we herein established a method for single-cell RNA-sequencing with markedly improved sensitivity for BCR-ABL detection compared to standard techniques. This new technique uniquely allowed us to selectively analyse aberrant gene expression in BCR-ABL$^-$ SCs at diagnosis, of relevance in view of recent findings that cell-extrinsic factors disrupt normal stem/progenitor cells in CML mouse models, and other hematological malignancies[21-23]. Our analysis revealed marked dysregulation of TGFβ and TNFα pathways in BCR-ABL$^-$ (as well as BCR-ABL$^+$) SCs, associated with increased SC quiescence. Moreover, we uncovered heterogeneity of BCR-ABL$^-$ SCs in CML patients, with a distinct cluster of BCR-ABL$^-$ SCs dominating already at diagnosis in patients who later failed to achieve MMR on TKI treatment. Indeed, elevated serum levels of TNFα and TGFα also correlates with poor treatment-response in CML.[43] Further validation studies in larger patient cohorts will be required to determine whether gene expression signatures of BCR-ABL$^-$ SCs might have utility as a clinically predictive biomarker. Furthermore, targeting inflammatory pathways such as TGFβ and TNFα might also be of therapeutic value, by reducing microenvironment-induced quiescence of CML-SCs, although further

476 preclinical evidence of the feasibility of such an approach is needed before
477 this could be taken forward into a clinical trial setting.
478
479 Our single-cell method also provided a unique opportunity to assess rare
480 BCR-ABL$^+$ SCs persisting during TKI-induced remission[16,17]. It was not
481 possible to analyse resistant CML-SCs in patients who had already achieved
482 deep molecular remissions due to very low frequency of BCR-ABL$^+$ SCs in
483 these patients. However, analysis of samples from patients established on
484 TKI, including serial samples and patients on long term TKI (>1 year),
485 identified a distinct subpopulation of highly quiescent BCR-ABL$^+$ SCs, already
486 present at diagnosis, that is markedly selected for during otherwise clinically
487 effective TKI treatment. Quiescence is a hallmark of many normal SCs,
488 including HSCs, conferring selective resistance to therapeutic targeting[44,45].
489 Crucially, our data using a whole-transcriptome approach, support that TKI-
490 resistant CML-SCs are transcriptionally distinct from quiescent normal-HSCs,
491 with dysregulation of specific genes and pathways (TGFβ, TNFα, JAK/STAT,
492 *CTNNB1*, *NFKB1A*) that might be selectively targeted in CML-SCs. Another
493 recent study applied a single-cell targeted gene expression analysis of BCR-
494 ABL$^+$ CML-SCs[46], rather than unbiased single-cell global RNA-sequencing.
495 While the much more restricted gene expression analysis was focused at
496 looking at the heterogeneity of lineage programs in BCR-ABL$^+$ stem cells and
497 improved strategies for prospective purification of CML-SCs, also the findings
498 in those studies supported a TKI-induced enrichment of quiescent BCR-ABL$^+$
499 stem cells, although only investigated following short-term TKI treatment.
500
501 CML is an ideal tractable disease model to apply this single-cell technique
502 due to its relative genomic simplicity[15], however, a number of our findings may
503 also be more generally applicable to other malignant disease. For example,
504 although limited by relatively small numbers of blast crisis patient samples
505 available, our analysis of BC-CML patients support that a single-cell approach
506 may prove powerful towards predicting imminent disease progression in CSC
507 populations. Specifically, our ability to detect RUNX1 mutations in distinct BC
508 CML-SCs subclones shows how a single cell approach can help to unravel
509 the mechanisms underlying clonal progression associated with certain

mutations at the CSC level. However, further work is required to determine feasibility of applying this new method to detect a range of other mutations and a number of possible limitations in the approach need to be considered: Some tumors are characterized by exceedingly complex clonal heterogeneity. It is likely that there will be a limitation in relation to the number of mutations that could be simultaneously detected by targeted amplification in individual cells before this impacts on the complexity of the RNA-seq library generated, although this remains to be determined. Our technique also relies on expression of the mutation of interest, and with increasing interest in mutations in the non-coding space[47], further modifications to this approach will be required, for example to allow parallel gDNA analysis. Furthermore, in order to obtain a high level of sensitivity for BCR-ABL detection, the amplicon size used in this study was short and did not encompass the kinase domain of ABL. Longer BCR-ABL amplicons were less efficient at BCR-ABL detection (data not shown). We were therefore unable to detect presence of kinase domain mutations in individual cells, of relevance for TKI resistance[48]. Further modification to our technique will be required to detect multiple, distantly located mutations occurring in *CIS* within the same allele.

In summary, we present a novel method allowing simultaneous single-cell RNA-sequencing and high-sensitivity targeted mutation detection. We demonstrate how this technique can be applied to unravel heterogeneity in clonal CSCs as well as in co-existing and frequently suppressed normal SCs, to provide novel insights into cellular and molecular mechanisms of therapy resistance and clonal evolution. In principle, this approach can be applied across a broad range of clonal disorders. Although considerable technical challenges remain in relation to standardization of single cell genomics techniques, we anticipate that the next few years will see major inroads towards clinical application of this powerful new technology.

Accession codes

Gene Expression Omnibus (GEO) accession code GSE76312.

Data Availability Statement

Single-cell RNA sequencing data are available at NCBI's Gene Expression Omnibus (GEO) data repository with the accession code GSE76312.

Acknowledgments

This work was funded by a Medical Research Council Senior Clinical Fellowship (MR/L006340/1), MRC Confidence in Concept award (MC_PC_13073) and Rosetrees Trust award (M435) to A.J.M., the MRC Molecular Haematology Unit core award (A.J.M. and S.E.J.; MC_UU_12009/5), a MRC programme grant to SEJ (G0801073), an international recruitment award from the Swedish Research Council (SEJ), and grants from the Tobias Foundation (SEJ) and the Center for Innovative Medicine (CIMED) at the Karolinska Institute (SEJ). This work was also supported by The MRC funded Oxford Consortium for Single-cell Biology (MR/M00919X/1) and the Oxford NIHR Biomedical Centre based at Oxford University Hospitals NHS Trust and University of Oxford (131/030). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or the NIH. The work was also supported by an educational grant from Novartis. The authors acknowledge the contributions of the WIMM Flow Cytometry Facility, supported by the MRC HIU; MRC MHU (MC_UU_12009); NIHR Oxford BRC and John Fell Fund (131/030 and 101/517), the EPA fund (CF182 and CF170) and by the WIMM Strategic Alliance awards G0902418 and MC_UU_12025. N.A. was supported by the Oxford-Wellcome Trust Institutional Strategic Support Fund. S.M. is supported by the Finnish Cancer Institute and the Finnish Cancer Organizations.

Author Contribution

A.G. designed, performed and analyzed experiments and contributed to writing the manuscript. S.T. designed and performed bioinformatic analyses and contributed to writing the manuscript. N.B. and B.P. performed analyses of RNA-sequencing and QPCR results. P.W. and P.S. were involved in FACS

analysis/sorting. R.N., A.R.M., C.B., L.J. performed experiments. N.A. maintained single-cell facility infrastructure. P.V., S.M. and H.Q. provided infrastructure for sample banking and provided input on experimental design and analysis. K.A. performed FISH experiments. A.S. was involved in RNA-sequencing experiments. S.U. collected clinical information. R.S. provided input on RNA-sequencing experiments. A.J.M. and S.E.W.J. conceived and supervised the project, designed and analyzed experiments and wrote the manuscript.

Competing Financial Interests

A.J.M. has received honoraria and research funding from Novartis.

References

1. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell* **27**, 15-26 (2015).
2. Tehranchi, R.*, et al.* Persistent malignant stem cells in del(5q) myelodysplasia in remission. *The New England journal of medicine* **363**, 1025-1037 (2010).
3. Magee, J.A., Piskounova, E. & Morrison, S.J. Cancer stem cells: impact, heterogeneity, and uncertainty. *Cancer cell* **21**, 283-296 (2012).
4. Woll, P.S.*, et al.* Myelodysplastic syndromes are propagated by rare and distinct human cancer stem cells in vivo. *Cancer cell* **25**, 794-808 (2014).
5. Sloma, I.*, et al.* Genotypic and functional diversity of phenotypically defined primitive hematopoietic cells in patients with chronic myeloid leukemia. *Exp Hematol* **41**, 837-847 (2013).
6. Mustjoki, S.*, et al.* Impact of malignant stem cell burden on therapy outcome in newly diagnosed chronic myeloid leukemia patients. *Leukemia* **27**, 1520-1526 (2013).
7. Alizadeh, A.A.*, et al.* Toward understanding and exploiting tumor heterogeneity. *Nature medicine* **21**, 846-853 (2015).
8. Wills, Q.F. & Mead, A.J. Application of single-cell genomics in cancer: promise and challenges. *Human molecular genetics* (2015).
9. Wang, Y. & Navin, N.E. Advances and applications of single-cell sequencing technologies. *Molecular cell* **58**, 598-609 (2015).
10. Patel, A.P.*, et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396-1401 (2014).
11. Miyamoto, D.T.*, et al.* RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* **349**, 1351-1356 (2015).

12. Tirosh, I.*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
13. Druker, B.J.*, et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nature medicine* **2**, 561-566 (1996).
14. Goldman, J.M. & Melo, J.V. Targeting the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *The New England journal of medicine* **344**, 1084-1086 (2001).
15. Longo, D.L. Imatinib Changed Everything. *The New England journal of medicine* **376**, 982-983 (2017).
16. Gallipoli, P., Abraham, S.A. & Holyoake, T.L. Hurdles toward a cure for CML: the CML stem cell. *Hematology/oncology clinics of North America* **25**, 951-966, v (2011).
17. Chu, S.*, et al.* Persistence of leukemia stem cells in chronic myelogenous leukemia patients in prolonged remission with imatinib treatment. *Blood* **118**, 5565-5572 (2011).
18. Mahon, F.X.*, et al.* Discontinuation of imatinib in patients with chronic myeloid leukaemia who have maintained complete molecular remission for at least 2 years: the prospective, multicentre Stop Imatinib (STIM) trial. *Lancet Oncol* **11**, 1029-1035 (2010).
19. Schepers, K.*, et al.* Myeloproliferative neoplasia remodels the endosteal bone marrow niche into a self-reinforcing leukemic niche. *Cell Stem Cell* **13**, 285-299 (2013).
20. Colmone, A.*, et al.* Leukemic cells create bone marrow niches that disrupt the behavior of normal hematopoietic progenitor cells. *Science* **322**, 1861-1865 (2008).
21. Schepers, K., Campbell, T.B. & Passegue, E. Normal and leukemic stem cell niches: insights and therapeutic opportunities. *Cell Stem Cell* **16**, 254-267 (2015).
22. Welner, R.S.*, et al.* Treatment of chronic myelogenous leukemia by blocking cytokine alterations found in normal stem and progenitor cells. *Cancer cell* **27**, 671-681 (2015).
23. Reynaud, D.*, et al.* IL-6 controls leukemic multipotent progenitor cell fate and contributes to chronic myelogenous leukemia development. *Cancer cell* **20**, 661-673 (2011).
24. Picelli, S.*, et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* **9**, 171-181 (2014).
25. Wilson, N.K.*, et al.* Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell* **16**, 712-724 (2015).
26. Wu, S.Q.*, et al.* Extensive amplification of bcr/abl fusion genes clustered on three marker chromosomes in human leukemic cell line K-562. *Leukemia* **9**, 858-862 (1995).
27. Wu, A.R.*, et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* **11**, 41-46 (2014).
28. Islam, S.*, et al.* Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nature protocols* **7**, 813-828 (2012).
29. Bruns, I.*, et al.* The hematopoietic stem cell in chronic phase CML is characterized by a transcriptional profile resembling normal myeloid

progenitor cells and reflecting loss of quiescence. *Leukemia* **23**, 892-899 (2009).
30. Miyawaki, K.*, et al.* The Expansion Of CML Clones Initiates At The CMP Stage, and Is Associated With The Down-Regulation Of IRF8 and GFI1. *Blood* **122**(2013).
31. Schuettpelz, L.G. & Link, D.C. Regulation of hematopoietic stem cell activity by inflammation. *Front Immunol* **4**, 204 (2013).
32. King, K.Y. & Goodell, M.A. Inflammatory modulation of HSCs: viewing the HSC as a foundation for the immune response. *Nat Rev Immunol* **11**, 685-692 (2011).
33. Pronk, C.J., Veiby, O.P., Bryder, D. & Jacobsen, S.E. Tumor necrosis factor restricts hematopoietic stem cell activity in mice: involvement of two distinct receptors. *The Journal of experimental medicine* **208**, 1563-1570 (2011).
34. Cashman, J.D., Eaves, A.C., Raines, E.W., Ross, R. & Eaves, C.J. Mechanisms that regulate the cell cycle status of very primitive hematopoietic cells in long-term human marrow cultures. I. Stimulatory role of a variety of mesenchymal cell activators and inhibitory role of TGF-beta. *Blood* **75**, 96-101 (1990).
35. Baccarani, M.*, et al.* European LeukemiaNet recommendations for the management of chronic myeloid leukemia: 2013. *Blood* **122**, 872-884 (2013).
36. Kentsis, A.*, et al.* Autocrine activation of the MET receptor tyrosine kinase in acute myeloid leukemia. *Nature medicine* **18**, 1118-1122 (2012).
37. Liu, J.X.*, et al.* Eaf1 and Eaf2 negatively regulate canonical Wnt/beta-catenin signaling. *Development* **140**, 1067-1078 (2013).
38. Goardon, N.*, et al.* Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer cell* **19**, 138-152 (2011).
39. Jamieson, C.H. Chronic myeloid leukemia stem cells. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, 436-442 (2008).
40. Zhao, K.*, et al.* IL1RAP as a surface marker for leukemia stem cells is related to clinical phase of chronic myeloid leukemia patients. *Int J Clin Exp Med* **7**, 4787-4798 (2014).
41. Herrmann, H.*, et al.* Dipeptidylpeptidase IV (CD26) defines leukemic stem cells (LSC) in chronic myeloid leukemia. *Blood* **123**, 3951-3962 (2014).
42. Gerber, J.M.*, et al.* Genome-wide comparison of the transcriptomes of highly enriched normal and chronic myeloid leukemia stem and progenitor cell populations. *Oncotarget* **4**, 715-728 (2013).
43. Nievergall, E.*, et al.* TGF-alpha and IL-6 plasma levels selectively identify CML patients who fail to achieve an early molecular response or progress in the first year of therapy. *Leukemia* **30**, 1263-1272 (2016).
44. Trumpp, A., Essers, M. & Wilson, A. Awakening dormant haematopoietic stem cells. *Nat Rev Immunol* **10**, 201-209 (2010).
45. Clevers, H. The cancer stem cell: premises, promises and challenges. *Nature medicine* **17**, 313-319 (2011).
46. Warfvinge, R.*, et al.* Single-cell molecular analysis defines therapy response and immunophenotype of stem cell subpopulations in CML. *Blood* (2017).

47. Mansour, M.R., *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373-1377 (2014).
48. Soverini, S., De Benedittis, C., Mancini, M. & Martinelli, G. Present and future of molecular monitoring in chronic myeloid leukaemia. *Br J Haematol* **173**, 337-349 (2016).

Figure Legends

Figure 1. High sensitivity single-cell detection of BCR-ABL with parallel unbiased whole transcriptome analysis. (**a, b**) Detection of BCR-ABL and GAPDH by QPCR in libraries from single K562 cells processed by Smart-seq2 (**a**), or BCR-ABL tSS2 (**b**). Values shown are the gene expression levels relative to the limit of detection (LOD), indicated by the dashed horizontal line. The box plot shows median and quartile values and whiskers show outlier values within 1.5 interquartile range of the quartiles. Numbers below each plot show the frequency of cells showing expression above the LOD. (**c**) Correlation of expression data for 14,240 RefSeq genes generated from K562 single cells using Smart-seq2 (n=38) or BCR-ABL tSS2 (n=38). (**d**) RNA-sequencing results from single K562 cells processed with Smart-seq2 (blue n=38) or by BCR-ABL tSS2 (red n=38) shown by t-Distributed Stochastic Neighbor Embedding (tSNE) using 3,368 highly variable genes (see methods). (**e**) Dot plot illustrating the sensitivity to detect specific copy numbers of BCR-ABL spiked-in before BCR-ABL tSS2 amplification of single BM cells from a healthy donor. The Y-axis indicates the gene expression level of BCR-ABL relative to the LOD. The X-axis indicates the absolute number of copies of BCR-ABL expected to be present in each reaction, calculated using a commercial standard. The table above shows the numbers of wells that would be expected to contain at least one copy of BCR-ABL by Poisson distribution and the actual frequency of amplification following BCR-ABL tSS2.

Figure 2. Single-cell whole transcriptome analysis and BCR-ABL detection in single CML stem cells. (**a**) Box plot illustrating the number of genes detected (RPKM ≥1) in relation to depth of sequencing in normal-HSC samples (shown

as million reads/cell). (**b**) RNA-sequencing results from single K562 cells processed by BCR-ABL tSS2 (purple n=38) and normal-HSCs (n=232) shown by t-Distributed Stochastic Neighbor Embedding (tSNE) using 7,428 highly-variable genes. (**c**) Correlation between the merged data from 40 single-cells from CML patient OX1407 ("ensemble") and the bulk (100 cells sorted together) RNA-sequencing measurement of gene expression from the same patient. The ensemble was created by computationally pooling all the reads obtained from the 40 single $Lin^-CD34^+CD38^-$ cells from patient OX1407. Some of the genes shown in panel **f** are highlighted. (**d**) Correlation between the levels of gene expression of BCR-ABL^+ ensemble and BCR-ABL^- ensemble data. Some of the genes shown in panel **f** are highlighted. (**e**) Heat map illustrating the hierarchical clustering of BCR-ABL^+ SCs (red, n=17) or BCR-ABL^- SCs (Blue, n=23) showing the top 75 differentially expressed genes. (**f**) Correlation of log2(FC) by RNA-sequencing (y-axis) and by QPCR (x-axis) between BCR-ABL^+ and BCR-ABL^- SCs for selected genes. Differentially expressed genes (red dots) were selected by setting a fold change cutoff >8 and selected 12 genes of potential biologic interest. Non-differentially expressed genes (grey dots, n=12) were selected as housekeeping genes or relevant genes for the cell type analyzed. (**g, h**) Beeswarm plots for 6 of the 12 selected differentially expressed genes between BCR-ABL^+ and BCR-ABL^- SCs showing RNA-sequencing (**g**) and QPCR (**h**) data. Numbers of cells analyzed and numbers showing amplification for the selected gene are shown below the plot. Nonparametric Wilcoxon test p-values are shown on top of each bar graph. Fisher's exact test p-values are shown below the graph. The average gene expression levels are indicated by red squares, the median and quartiles of gene expression levels are represented by the boxes. The dashed lines represent the LOD.

Figure 3. Single-cell RNA-sequencing reveals distinct molecular signatures of BCR-ABL^+ CML-SCs at diagnosis. (**a, b**) Gene-set enrichment analysis on 477 BCR-ABL^+ single-cells from 18 chronic phase CML patients at diagnosis versus 232 normal-HSCs from 5 normal donors. Gene-sets shown are (**a**) cell proliferation and (**b**) quiescence associated genes. (**c**) tSNE visualization of single normal-HSCs (Grey circles; n=232), BCR-ABL^- SCs (blue diamonds;

n=377) and BCR-ABL$^+$ SCs (red triangles; n=477) using 8,589 highly-variable genes. (**d**) Hierarchical clustering analysis of the same 1086 cells. The heatmap is built using Pearson correlation generated using the top 245 differentially expressed genes. The horizontal color bar on top of the heatmap indicates the sample from which each single SC was purified (upper bar, individual color for each patient) and the cell ID (lower bar): normal-HSCs (black), BCR-ABL$^-$ SCs (blue) and BCR-ABL$^+$ SCs (red). (**e**) GSEA of unbiased HALLMARK gene-sets for 1) Normal-HSCs (n=6) vs BCR-ABL$^+$ SCs (n=18) as an *in silico* bulk analysis; 2) Single-cell analysis of normal-HSCs (n=232) vs BCR-ABL$^-$ SCs (n=377); 3) Single-cell analysis of normal-HSCs (n=232) vs BCR-ABL$^+$ SCs (n=477); 4) Single-cell analysis of BCR-ABL$^-$ SCs (n=377) vs BCR-ABL$^+$ SCs (n=477). A false discovery rate (FDR) cut-off of 0.25 was used.

Figure 4. Single-cell RNA-sequencing of SCs in CML patients at diagnosis predicts molecular response to TKI. (**a**) tSNE visualization of single BCR-ABL$^+$ SCs (from 16 CP-CML patients with molecular follow-up data available, n=436) using 5,011 highly-variable genes. Color indicates if cells were isolated from good responders (n=11 patients achieving MMR, blue) or poor responders (n=5 patients not achieving MMR, red). (**b**) tSNE visualization of single BCR-ABL$^-$ SCs from 16 patients with molecular follow-up data available (n=356) using 5,611 highly-variable genes. Color indicates if cells were isolated from good responders (11 patients achieving MMR, blue) or poor responders (5 patients not achieving MMR, red). (**c**) The dot plot shows the proportion (%) of BCR-ABL$^-$ SCs falling in the “poor responders” cluster for individual patients (n=16, in red patients with >10% of cells in poor responder cluster, in blue patients with <10% of cells in poor responder cluster; squares represent patients failing to achieve MMR and circles patients who achieved MMR). (**d**) Kaplan Meier curves showing time for MMR achievement for patients with >10% (red, n=5) or <10% (blue, n=11;) of BCR-ABL$^-$ SCs falling in the poor responder cluster. P-value represents the logrank test. (**e**) GSEA of unbiased HALLMARK gene sets comparing BCR-ABL$^-$ SCs (n=356) and BCR-ABL$^+$ SCs (n=436) from good (n=11) and poor (n=5) TKI responders.

Figure 5. Single-cell analysis reveals distinct molecular signatures of quiescent CML-SCs persisting during TKI therapy. (**a**) tSNE visualization of single normal-HSCs (black circles; n=232, 5 donors), BCR-ABL$^+$ SCs from patients at diagnosis (grey circles; n=477, 18 donors) and BCR-ABL$^+$ SCs from patients at remission (light-blue diamonds and dark-blue triangles, n=245, 16 donors). Remission BCR-ABL$^+$ SCs clustering closer to normal-HSCs (light-blue diamonds, n=122) are defined as group-A BCR-ABL$^+$ SCs, while those cells clustering with most diagnostic BCR-ABL$^+$ SCs are defined as group-B (dark-blue triangles, n=123). (**b**) Gene-set enrichment analysis of group-A vs group-B BCR-ABL$^+$ SCs at remission (n=122 and n=123, respectively). Gene-sets shown are cell proliferation, quiescence and HSC-associated genes. (**c**) Bar graph showing the proportion (%) of group-A BCR-ABL$^+$ SCs and group-B BCR-ABL$^+$ SCs for all patients analyzed at diagnosis (n=18), at 3 months (n=11) and more than 1 year (n=4) after TKI initiation. Chi-square and Fisher's exact test $p<0.01$ for comparison of diagnosis versus 1-year samples. (**d**) The bar graph shows same results as in panel **c** but only for patients eventually achieving MMR, samples taken at diagnosis (n=11), at 3 months (n=6) and more than 1 year (n=2) after TKI initiation. Chi-square and Fisher's exact test $p<0.01$ for comparison of diagnosis versus 1-year samples. (**e**) Gene-set enrichment analysis of TNFα, TGFβ and IL6-JAK-STAT pathways comparing group-A BCR-ABL$^+$ SCs at remission (n=122) vs normal-HSCs (n=232). (**f**) Gene-set enrichment analysis of TNF-α and TGF β pathways performed on normal-HSCs (n=232) vs group-A BCR-ABL$^+$ SCs at 3 months (n=72), 6 months (n=24) and over 1 year after TKI initiation (n=27). (**g**) Beeswarm plots for 10 selected differentially expressed genes between normal-HSCs (black; n=232, 5 donors), BCR-ABL$^+$ SCs from patients at diagnosis (red; n=477, 18 donors) and BCR-ABL$^+$ SCs from patients at remission group-A (light blue, n=122, 16 donors). Numbers of cells analyzed and numbers showing amplification for the selected gene are shown below the plot. The average gene expression levels are indicated by red squares, the boxes represent the median and quartiles of gene expression levels. Nonparametric Wilcoxon test p-values are shown on top of each bar graph. Fisher's exact test p-values are shown below the graph.

Figure 6. Single-cell RNA-sequencing reveals heterogeneity of CML stem cells associated with disease progression in CML. (**a**) tSNE visualization of single normal-HSCs from 5 donors (grey circles; n=232), BCR-ABL$^+$ SCs from 18 CP-CML patients (red triangles; n=477) BCR-ABL$^+$ SCs from 3 patients at the time of BC (CML1931, light blue squares, n=85; CML1266, purple squares, n=63; CML1203, pink squares, n=7 and K562 cells (brown circles, n=53). The tSNE has been generated using 207 differentially expressed genes as described in the methods. (**b**) The heatmap shows the top 40 genes differentially expressed between BCR-ABL$^+$ SCs falling in the CP-CML cluster (n=477) and BCR-ABL$^+$ SCs falling in the BC cluster (n=155). The bar above the heatmap indicates CP-CML cluster in red, BC-CML cluster in purple. (**c**) tSNE visualization as shown in panel **a** but with BC (light-blue squares, n=85) and pre-BC (orange diamonds, n=132) cells from patient 1931 highlighted. Arrow indicates 8 pre-BC cells clustering separately from remaining pre-BC cells and together with BCR-ABL$^+$ CP-SCs. (**d**) Heatmap of log2(RPKM) of selected lymphoid and myeloid genes in BCR-ABL$^+$ CML-SCs at BC (n=155, 3 donors), pre-BC (n=185, 2 donors with SCs from patient OX1931 annotated according to those falling within the CP-CML cluster in yellow or BC-CML cluster in orange) and CP-CML SCs at diagnosis (n=477, 18 donors) and normal-HSCs (n=232, 5 donors), showing aberrant co-expression of lymphoid and myeloid genes in SCs falling within the BC cluster. (**e**) Dot plot showing index sort results corresponding to individual BCR-ABL$^+$ SCs from pre-BCs OX1931. The color and shape of the dots indicate if the SC clustered with CP-CML BCR-ABL$^+$ SCs (blue triangles) or with BC-CML BCR-ABL$^+$ SCs (red circles) according to the RNA-seq results presented in the tSNE analysis in panel **a**. The value is expressed as fluorescent intensity for CD90 and CD45RA antigens (y and x axis, respectively). (**f**) Histogram (left panel) shows the frequency of differentially expressed genes between pre-BC OX1931 BCR-ABL$^+$ SCs clustering with CP-CML or BC-CML BCR-ABL$^+$ SCs (Y axis) with respect to the distance (Kb) of RUNX1 binding sites from the respective transcription start site (TSS; X axis). The box plot (right panel) shows the fraction of RUNX1 binding sites/window found in the genes that are differentially expressed between the CP-CML and the BC-CML SC clusters

(red box) versus those found in background genes (grey box). P value =0.0018 by Wilcoxon test.

Online Methods

Cell Lines

Authenticated K562 and mycoplasma negative (Chronic Myeloid Leukemia, human cell line) were obtained from ATCC and grown in IMDM, 10% Fetal Bovine Serum (FBS).

Samples and bone marrow mononuclear cells processing

CML patients included in the study and their clinical details are listed in Supplementary Table 1. Patients provided written informed consent in accordance with the Declaration of Helsinki for sample collection and use in research under Oxford University ethics committee approval (MREC 06/Q1606/110). Bone marrow (BM) mononuclear cells (MNCs) were isolated using Ficoll density gradient. Cryopreserved BM MNCs were thawed and processed for flow cytometry analysis as previously described[4].

FACS staining and single-cell sorting

All Fluorescence Activated Cell Sorting (FACS) experiments included single-color stained CompBeads (BD Biosciences) and fluorescent-minus-one (FMO) controls. Live cells were selected based on their non-permeability and subsequent lack of fluorescence associated with 7AAD or DAPI. The combination of monoclonal antibodies used to identify hematopoietic stem and progenitor cell populations was previously described[4] and is listed below. The cocktail of lineage markers used (Lin) was: CD2, CD3, CD4, CD7, CD8a, CD10, CD11b, CD14, CD19, CD20, CD56, CD235ab. Single cells were isolated from BM samples of healthy controls or CML patients. Single cells were FACS sorted as Lin-$CD34^{+}CD38^{-}$. For some experiments, index sort data of the mean fluorescence intensities (MFI) of CD90, CD45RA and CD123 were also recorded for each individual cell isolated.

Anti-human Antibodies

Antigen	**Clone**	**Conjugate**	**Company**
CD34	8G12	APC	Biolegend
CD38	HIT2	PETXR	Life Technologies
CD90	5E10	PE	Biolegend
CD45RA	MEM56	FITC	Life Technologies
CD123	6H6	PECy7	Biolegend
CD2	RPA-2.10	PECy5	Biolegend
CD3	HIT3a	PECy5	Biolegend
CD4	RPA-T4	PECy5	Biolegend
CD7	CD7-6B7	PECy5	Biolegend
CD8a	RPA-T8	PECy5	Biolegend
CD10	HI10a	PECy5	Biolegend

CD11b	ICRF44	PECy5	Biolegend
CD14	RMO52	PECy5	Biolegend
CD19	HIB19	PECy5	Biolegend
CD20	2H7	PECy5	Biolegend
CD56	B159	PECy5	BD
CD235ab	HIR2	PECy5	Biolegend

Single-cell sorting was performed on FACS ARIA II, FACS ARIA III or FACS ARIA Fusion (Becton Dickinson) directly into 96-well plates (PCR micro plate Thermo-Fast 96 well, semi-skirted). To check the correct alignment of the sorter, BD FACS™ Accudrop Beads (BD Biosciences) were deposited initially onto the lid or film cover of a setup plate. After this sort, 50 beads were sorted into several wells of a clean PCR plate where it was checked that the beads formed a discrete drop in the center of the bottom of the well. If any splashing was noticed on the sides of the well, the alignment was adjusted. Single 488-Flow-Check Fluorospheres (Beckman Coulter) were then deposited into each well of a flat bottomed 96 well tissue culture plate and single-cell mode sorting was verified by checking the presence of 1 fluorosphere/well using a conventional fluorescence microscope. The investigators were not blinded when performing this and following steps of the experiments. Experiments were not randomized.

Short-term culture from single cells for first division measurement.

Single $Lin^-CD34^+CD38^-$ cells from normal BM donors or from CML patients were sorted using FACSAriaIII into 60-well Terasaki plates containing 25µl of Stemspan SFEM (Stemcell Technologies) medium supplemented with 10% BIT 9500 serum substitute (Stemcell Technologies), 2 mM L-Glutamine (P A A Laboratories), 10-4 M 2-mercaptoethanol (Sigma), 100 U/mL penicillin/streptomycin (PAA), 100 ng/mL rhSCF (Amgen), 100 ng/mL rhFLT3-ligand (FL; Immunex), 50 ng/mL rhTPO (Peprotech), 10 ng/mL rhIL-3 (Peprotech), 10 ng/mL rhG-CSF (Amgen), 10 ng/mL rhIL-6 (Peprotech). rhTNFα (Miltenyi Biotec) or rhTGFβ (Miltenyi Biotec) were added to the

culture at 20 ng/ml as indicated. Single cells were scored microscopically for number of cells that had reached time of first division after 96 hours of culture.

Fluorescence in situ hybridization (FISH)

For interphase FISH, $Lin^-CD34^+CD38^-$ cells were cytocentrifuged onto slides and hybridized with the LSI BCR/ABL Dual Color, Dual Fusion Translocation Probe (Abbot Molecular) spanning the ABL1 and BCR respective breakpoints involved in the t(9:22) translocation (ABL1: 9q34, BCR: 22q11.2). Fluorescence images were obtained with the use of fluorescence microscopy. In nuclei from normal cells lacking the t(9:22) translocation the probe hybridizing to ABL1 region appears as 2 orange signals while the probe hybridizing to BCR region appears as 2 green signals. Nuclei containing a balanced t(9;22) will display one orange and one green signal from the normal 9 and 22 chromosomes and two orange/green (yellow) fusion signals, one each from the derivative 9 and 22 chromosomes.

Generation of single-cell cDNA libraries using Smart-seq2 protocol

Single K562 or $Lin^-CD34^+CD38^-$ cells were FACS sorted into 96-well plates (Thermo) containing 4 µL of a lysis mix including oligo dT (Biomers), RNAse inhibitor (Takara), dNTPs mix (Fermentas) at concentrations described in the original Smart-seq2 protocol and listed in the lysis mix box below[24]. ERCC spikes (AMbion) were pre-diluted to 1:400,000 from stock concentration and added to the lysis mix at a final dilution of 1:40,000,000. ERCC spikes were not included in the analysis of patient samples as a number of samples were analyzed before ERCC spikes were routinely included in the reaction. Retrotranscription and PCR amplification steps were performed following the Smart-seq2 protocol using reagents concentrations optimized for small cells (see box RT mix and PCR mix for individual reagents concentration). The thermal conditions for RT and PCR reactions were according to the original Smart-seq2 protocol. The number of cycles used for PCR amplification was 22. After PCR amplification, cDNA libraries from single-cells were purified

using Ampure XP magnetic beads according to the manufacturer instructions in a ratio of 0.8 to 1 with cDNA. After purification the libraries were resuspended in 17.5 µL of buffer EB (Qiagen) and stored at -20 °C. Quality and concentration of the cDNA libraries generated was assessed using High Sensitivity Bioanalyzer (Agilent).

Generation of single-cell cDNA libraries using BCR-ABL tSS2 protocol

BCR-ABL targeted amplification Smart-seq2 protocol (tSS2) was implemented during both RT and PCR steps of Smart-seq2 as described in Supplementary Figure 2. During RT and PCR-amplification a pair of primers recognizing the BCR and the ABL portion of the fusion transcripts (sequences indicated in Supplementary Fig. 2b) were added to the RT and PCR mixes respectively at the concentrations indicated in Supplementary Figure 2c in the condition 6. The primers pair was designed to give rise to a PCR amplicon of 505 bp for BCR-ABL e14a2 transcript and 430 for BCR-ABL e13a2 transcript.

Lysis Mix reagents

Reagent	Volume for 1 cell (µL)
0.4% Triton X + RNAse Inhibitor (1:20)	2
dNTPS (10 mM)	1
Oligo dT (10 µM)	1
ERCC (pre-diluted 1:400,000)	0.1
TOTAL	4 (4.1 with ERCC)

RT mix reagents

Reagent	Volume for 1 cell (µL)
Superscript II first strand buffer (5x)	2
DTT (100 mM)	0.5
Betaine (5 M)	2
MgCl2 (1 M)	0.1
RNAse Inhibitor (40 U/µL)	0.25
TSO (100 µM)	0.1
BCR-ABL Primer set #1 F+R (200 µM)	0.07
Superscript II (200 U/µL)	0.25
Water	0.33
TOTAL	5.6

PCR mix reagents

Reagent	Volume for 1 cell (µL)
KAPA Hifi HS Ready Mix (2x)	12.5
ISPCR oligo (10 µM)	0.125
BCR-ABL Primer set#2 F+R (20 µM)	0.07
Water	2.305
TOTAL	15

Generation of single-cell cDNA libraries using BCR-ABL targeted amplification protocol with C1 microfluidic platform

K562 cells were captured on a large-sized (17-25 µm cell diameter**)** C_1™ Single-Cell Auto Prep IFC for mRNA Sequencing (Fluidigm) using the Fluidigm C1 system. Cells were loaded onto the chip at a concentration of ~250K cells/mL and imaged by phase-contrast microscopy to check single-cell per capture site. Cells were lysed and cDNA prepared on the C1 Fluidigm chip according to manufacturer's protocol, using SMARTer Ultra Low RNA kit for Illumina (Clontech). BCR-ABL targeted amplification in the C1 setting was performed using the modified C1 PCR MIX protocol described in the table below. BCR-ABL Taqman assay (20X) was included in the C1 PCR MIX at a final dilution of 1:495. Any other step in the C1 protocol was performed following manufacturer's indications.

Modified C1 PCR MIX for BCR-ABL targeted amplification

Reagent	Volume (µL)
PCR Water (Advantage 2 Kit)	59.5
10X Advantage 2 PCR Buffer (Advantage 2 Kit)	10
50X dNTP Mix (Advantage 2 Kit)	4
IS PCR primer (Clontech SMARTer)	4
50X Advantage 2 Polymerase Mix (Advantage 2 Kit)	4
C1 Loading Reagent (Fluidigm)	4.5
BCR-ABL Taqman assay pre-diluted 1:22 (Hs03024541_ft)	4

Illumina library preparation and sequencing

1.25 µL of cDNA was used for tagmentation reaction carried out with Nextera XT DNA Sample Preparation kit (Illumina) according to manufacturer's instruction but using one-fourth of the volumes. Purification of the product was done with a 1:1 ratio of AMPure XP beads with a final elution in 17.5 µl in resuspension buffer provided from the Nextera kit. Samples were loaded on a High-Sensitivity DNA chip to check the size and quality of the indexed library while the concentration was measured with Qubit High-Sensitivity DNA kit (Invitrogen). BCR-ABL^{+} or BCR-ABL^{-} SCs eligible for sequencing were selected based on the quality of their indexed cDNA libraries (size 400-900bp; concentration > 4 ng/ml). The number of BCR-ABL^{+} or BCR-ABL^{-} SCs to be sequenced per patient was determined by availability of SCs from each sample and space available per flow cell to ensure sufficient depth of sequencing. Libraries were pooled to a final concentration ranging between 3 and 10 nM and were sequenced with Illumina HiSeq 2000 and Illumina Hiseq 4000 (51 bp single-end read) at The Wellcome Trust Centre for Human Genetics in Oxford.

BCR-ABL genotyping of single-cell cDNA libraries

BCR-ABL genotyping of cDNA libraries from single-cells was performed using QPCR reaction in a 384-well plate (Roche, Lightcycler). QPCR was performed in duplicate using 1.5 µL of the cDNA library for each reaction. The expression of *BCR-ABL* and *GAPDH* were measured using the following Taqman FAM-MGB assays *BCR-ABL*: Hs03024541_ft and *GAPDH*: Hs02758991_g1 (Life Technologies). The reactions were performed using a minimum of 60 cycles of amplification. We used QPCR and not raw sequencing reads to genotype cells for presence of BCR-ABL due to the low coverage of BCR-ABL in the sequencing data.

Exome Sequencing.

Genomic DNA was extracted from unfractionated BM MNCs from patient OX1931 at both pre-BC and BC stages using QIAamp DNA Blood Mini Kit (Qiagen) according to manufacturer's instructions. Exome capture was performed from GATC Biotech, using INVIEW Human Exome Library

preparation Enrichment with SureSelectXT Human All Exon Kit for Illumina Paired-End Sequencing (Read length: 2 x 125 bp). The number of PCR cycles performed for the amplification of the adaptor-ligated library was 5. The number of cycles used for the post- hybridization captured library amplification step was 12. The enriched exome fragments were pooled and paired-end sequenced on a HiSeq 2000 platform (Illumina). From this we obtained >60x on target coverage for the majority of positions for each of the samples.

Assessment of BCR-ABL tSS2 sensitivity using plasmid spike-in

BCR-ABL breakpoint region (e14a2) was PCR-amplified from cDNA of K562 cells using specific BCR-ABL primer set #1 described in Supplementary Figure 2b. The resulting PCR-amplicon was Sanger sequenced before being cloned into the pcr™- Blunt II-TOPO® vector using Zero Blunt® TOPO PCR Cloning Kit (Thermo Fisher). Correct size of the BCR-ABL insert was verified by PCR. Concentration of the resulting BCR-ABL plasmid was measured using Qubit (Invitrogen) and the absolute number of plasmid copies/µL was calculated. Several plasmid pre-dilutions were produced in order to be able to spike in the retro-transcription reaction of single BCR-ABL¯ SCs (HSCs from a normal donor), the desired amount of plasmid copies (1, 2, 5,10, 20, 50, 100, 1000) always at a volume of 1 µL. The PCR step was performed according to the standard BCR-ABL tSS2 protocol. Quantification of absolute number of BCR-ABL amplified copies after BCR-ABL tSS2 reaction was carried out by QPCR (Roche, Lightcycler) using a commercial BCR-ABL standard curve as a reference (Ipsogen BCR-ABL1 Mbcr, Qiagen).

RT mix reagents for BCR-ABL plasmid standard curve

Reagent	Volume for 1 cell (µL)
Superscript II first strand buffer (5x)	2
DTT (100 mM)	0.5
Betaine (5 M)	2
MgCl2 (1 M)	0.1
RNAse Inhibitor (40 U/µL)	0.25
TSO (100 µM)	0.1
BCR-ABL Primer set #1 (200 µM)	0.07
Superscript II (200 U/µL)	0.25

Plasmid (serial pre-dilutions)	1
Water	0.33
TOTAL	6.6

RUNX1 c.G521A detection with single-cell QPCR

RUNX1 c.G521A mutation detected in patient OX1931 by exome sequencing was PCR amplified and validated by Sanger sequencing. (Fw:GGCTGGCAATGATGAAAACT and Rev:CAATGGATCCCAGGTATTGG). A SNP genotyping Taqman assay specific for RUNX1 c.G521A was designed using the Custom Assay Design tool (ThermoFisher) and validated on positive (OX1931) and negative controls.

Single-cell gene expression analysis

For single-cell gene expression analysis, single-cells isolated by FACS were collected in each well of a 96-well plate containing 5 µl Cells Direct One-Step qRT–PCR (Invitrogen) mix and pre-amplified as previously described[49]. Pre-amplified samples were diluted 1:5 with TE before analysis of gene expression analysis on either a Fluidigm 96.96 or 192.24 Dynamic array using gene-specific Taqman assays (Life technologies). No template and no reverse transcriptase were included as negative controls.

Taqman assays used for Dynamic Array

Gene Symbol	**Taqman Assay ID**
ABL1	Hs00245443_m1
ATG3	Hs00223937_m1
B2M	Hs00984230_m1
BCL2	Hs00608023_m1
BCR	Hs00244731_m1
BCR-ABL	Hs03024541_ft
BLNK	Hs00179459_m1
CD33	Hs01076281_m1
CD34	Hs00990732_m1
CD79A	Hs00998119_m1
CD79B	Hs01058826_g1

CD164	Hs00174789_m1
CDK6	Hs01026371_m1
CKLF	Hs03047057_s1
CLU	Hs00156548_m1
CSF1R	Hs00911250_m1
CTNNB1	Hs00355049_m1
CXCR4	Hs00607978_s1
DNTT	Hs00172743_m1
FCER1A	Hs00758600_m1
GAPDH	Hs02758991_g1
GAS2	Hs01086684_m1
GOLGA8A	Hs01104342_m1
HPRT	Hs02800695_m1
HSP90A1	Hs03043878_g1
IFITM1	Hs00705137_s1
IGF1R	Hs00609566_m1
IGJ	Hs00376160_m1
ITGA6	Hs01041011_m1
MEIS1	Hs01017441_m1
MLLT3	Hs00180312_m1
MMRN1	Hs00201182_m1
MPL	Hs00180489_m1
MZB1	Hs00414907_m1
PTRF	Hs00396859_m1
RGS2	Hs01009070_g1
RXFP1	Hs01073141_m1
SAT1	Hs00161511_m1
SELL	Hs00174151_m1
SELP	Hs00927900_m1
SOD2	Hs00167309_m1
TESPA1	Hs00207702_m1
VWF	Hs01109446_m1

Analysis of quantitative PCR single-cell gene expression data

We calculated ΔCt values, which are relative to the mean expression level of two housekeeping genes (B2M and GAPDH). As previously described[25,50], Ct values were subtracted from the limit of detection (CT=30) followed by subtraction of the mean Ct value of housekeeping genes for each cell. Ct=40 was used for the comparative analysis of the detection of BCR-ABL and GAPDH in K562 cells between Smart-seq2 and BCR-ABL tSS2 protocols. Cells not expressing at the 15th percentile of all genes, or two housekeeping genes were removed from the analysis. Analysis of differential gene expression between BCR-ABL^{+} and BCR-ABL^{-} SCs was performed using the

Wilcoxon test and Fisher's exact test to compare expression level and expression frequency respectively.

Analysis of single-cell RNA sequencing

Short reads (51-bp) were aligned to the human genome (GRCh37 assembly (hg19)) using Tophat[51] with a supplied set of known RefSeq transcripts as the input. The mapping parameters '-g 1' was used to allow one alignment to the reference for a given read. Expression values were quantified as read per kilobase of transcript length per million mapped reads (RPKM) based on the RefSeq gene model using the rpkmforgenes[52]. As previously demonstrated the reliable classification of cell types at a sequencing depth of 50,000 reads per cell,[53,54] we used cells with higher than 50,000 mapping reads and 1,000 detected genes (RPKM ≥1) for the downstream analysis. We used the genes that were highly expressed in more than 50% of each population of cells to identify the candidate outliers based on gene expression level, similar to the method previously described in Singular™ from Fluidigm, using the standard method for the outlier detection[55]. The modified Z-scores were calculated using the formula 0.6745(xi - x̃)/MAD; MAD denoting the median absolute deviation and x̃ denoting the median. Cells with the absolute modified Z-score greater than 3 were considered as candidate outliers (28 out of 2,287 cells), and these cells were monitored during the analysis. We found that excluding or including them in our analysis did not have any significant impact on the results.

Analysis of the effective sequencing depth

To examine the effective sequencing depth, we selected 12 normal-HSCs with a sequencing depth larger than 6 million mapped reads. We randomly sampled reads in the range of 0.1 to 6 million mapped reads and calculated a number of detected genes with RPKM ≥1 in each category. We observed that detected number of genes plateaued at a sequencing depth of beyond 1 million mapped reads per cell (Fig. 2a).

T-distributed stochastic neighbor embedding (tSNE) analysis

As previously described, the advantage of using t-distributed stochastic neighbor embedding (tSNE) over a traditional principal component analysis is to visualize the projection of high dimensional single-cell gene expression data into a low dimensional space.[25,56,57] We selected genes expressed in ≥10 cells with a coefficient of variation score (CV), "standard deviation/mean", ≥1 and the summed up of genes expression values in log2 scale ≥1 for the tSNE analysis. We normalized the RPKM values into log2(RPKM) scale and set up the limit of detection at 1 RPKM. Log2 scale of genes expressed < 1 RPKM was set up to 0. Possible batches from processing samples in different dates were removed from expression values using the function "removeBatchEffect" in Limma package[58]. We then downloaded the tSNE software from https://lvdmaaten.github.io/tsne/ to perform the analysis using the Matlab implementation with "initial dims=20" and "perplexity=20" parameters.

To identify variable genes, similarly to a previously described approach[56], we fitted a simple noise model using the lowess model of mean expression level and the coefficient of variation (CV) to estimate the high variable genes from each type of cells. The lowess model predicted 3,368, 5,611, and 5,011 and 5,522 genes from K562, BCR-ABL$^-$ and BCR-ABL$^+$ SCs (from diagnosis), and normal-HSCs that show high variation compared to the whole genes set with mean of expression log2(RPKM) higher than 0. We next used these genes for the tSNE analysis, and compared the tSNE results to the previous tSNE results of different gene sets. We found the same pattern of clustering, suggesting reproducibility of our results. We then selected 3,368, 7,428 (combined variable genes from K562 and normal-HSCs) and 8,589 (combined variable genes from BCR-ABL$^-$ SCs, BCR-ABL$^+$ SCs and normal-HSCs) genes to generate Figure 1d, Figures 2b and 3c respectively. 5,011 and 5,611 variable genes in BCR-ABL$^+$ and BCR-ABL$^-$ SCs were used to generate Figure 4a and 4b for the good and poor responder classification.

For the tSNE analysis of samples following TKI therapy, we performed the random forests analysis of normal-HSCs, BCR-ABL$^+$SCs (diagnosis), and BCR-ABL$^+$SCs (remission) cells using the "randomForest" package in R (ntree parameter = 2,000). We obtained top 500 important genes, measured by the Gini index (Supplementary Table 7). These genes were used for distinguishing normal-HSCs from BCR-ABL$^+$SCs at diagnosis and during remission. We next used this gene set for the tSNE analysis of the remission cells (Fig. 5a). We applied K-means clustering (k=3) based on tSNE analysis results (from dimensions 1 and 2) to assign remission cells to group A and group B (Fig. 5a).

For the tSNE analysis of normal-HSCs, K562, and BCR-ABL$^+$SCs from diagnosis, pre-blast crisis and blast crisis samples, we obtained combined differentially expressed genes from the multiple ways comparison. 207 genes shown to be differentially expressed between BCR-ABL$^+$SCs from 18 chronic-phase CML patients (n=477), BCR-ABL$^+$SCs at BC (n=148), BCR-ABL$^+$SCs at pre-BC (n=185), and normal-HSCs (n=232). We next used this gene set for the tSNE to generate Figures 6a, 6c, and Supplementary Figure 14a. We note that pre-BC cells were involved in the tSNE analysis but were not shown in the Figure 6a.

Cell to cell variation analysis

To analyze the variation within K562 and normal-HSCs, the Pearson correlation was calculated based on log2(RPKM) expression values among the cells of each group using the same set of genes from the tSNE analysis. Kolmogorov–Smirnov test was used to test the difference of correlation score distribution (Supplementary Fig. 4e).

Differentially expressed gene analysis

Differentially expressed gene analysis was performed using the nonparametric Wilcoxon test on log2(RPKM) expression values for the

comparison of expression level and Fisher's exact test for the comparison of expressing cell frequency. P-values generated from both tests were then combined using Fisher's method and were adjusted using Benjamini-Hochberg (BH). Differentially expressed genes were selected based on the absolute log2 fold change ≥1 and the adjusted p-value < 0.05. Selected genes were subjected to the hierarchical clustering analysis using Pearson correlation as a distance with the complete clustering method performed in R with the "pheatmap" function. Beeswarm plots from selected genes were generated using the "beeswarm" package in R. We determined the top 100 differentially expressed genes ranked by adjusted p-values from normal-HSCs against BCR-ABL$^-$ and BCR-ABL$^+$ SCs, and BCR-ABL$^-$ against BCR-ABL$^+$ SCs at diagnosis. We then used the 245 unique genes from this analysis to make the heatmap of Figure 3d.

Gene Set Enrichment Analysis (GSEA)

GSEA[59] was performed using GSEA software (http://www.broadinstitute.org/gsea) with permutation on the phenotype, 1000 permutations, and default values for other parameters. Gene sets used in this study were selected from the CML, proliferation, quiescence, and HSC related pathways shown in Supplementary Table 3 and the MSigDB hallmark gene sets (Supplementary Table 5 and http://www.broadinstitute.org/gsea/msigdb/collections.jsp).

Comparison of bulk and single-cells analysis

To compare differentially expressed genes identified between analysis performed at the bulk level and at the single-cell level, in *silico* bulk data were generated by generating a data "ensemble", by combining mapped reads per gene for all cells, for 5 normal donors (5 replicates from the ensemble of single cells from 5 donors plus 1 set of normal-HSCs from a sixth donor, that were isolated as a bulk population of 100 cells rather than as single-cells). 18 replicates were generated from the ensemble of single-cells from each CML patient. DESeq2[60] was then performed from the raw read count of the

ensemble to get differentially expressed genes. We then performed differentially expressed genes analysis for the single-cell analysis as described above using 232 single normal-HSCs against 477 BCR-ABL$^+$ SCs and 377 BCR-ABL$^-$ SCs. To make a comparison of differentially expressed genes, we applied the same cutoff (adjusted p-value < 0.05 and the absolute log2 fold change ≥0.5) to get the number of differentially expressed genes from both bulk and single-cells analyses.

Co-expression analysis of G2M, lymphoid and myeloid genes

We selected a gene set from gene ontology "G2M transition of mitotic cell cycle (GO:0000086)" from "amigo2.berkeleybop.org/amigo/term/GO:0000086" to analyze co-expression. Gene would be called "present" when the quantile normalized RPKM value ≥1 and "absent" when RPKM values <1. We counted the frequency of genes that expressed in the same cells (Supplementary Fig. 7h). We calculated the co-expression frequency of random genes set excluding cell cycle related genes as the background. Kolmogorov–Smirnov test was used to test the difference of correlation score distribution. For analysis of co-expression of lymphoid and myeloid genes, we selected known lymphoid and myeloid signature genes to show the co-expression analysis in the CP-CML, BC-CML, and normal-HSC clusters. The heatmap was generated using the log2(RPKM) with the pheatmap function in R (Fig. 6d).

ChIP-Seq analysis

RUNX1 and IgG control ChIP-Seq data (CD34$^+$ HSPCs)[61] were downloaded from the GEO database (GSE45144). Raw reads were mapped to the human genome (GRCh37 assembly (hg19)) using bowtie2[62] with default parameters. The peak calling was performed by MACS2[63] using default parameters with IgG ChIP-Seq data as a control. 8,706 RUNX1 binding sites were identified with 5% FDR. We next calculated the distribution of distances between RUNX1 binding sites and transcription start sites (TSS) of differentially expressed genes that are between the CP-CML and the BC-CML clusters. We further analyzed the fraction of RUNX1 binding sites/window (ranging

from ±0.5 to ±10kb windows around a TSS of a gene) found in the differentially expressed genes in comparison to randomly selected 2,000 background genes (Fig. 6f).

Code availability

R and MATLAB scripts used for data analyses are available on request.

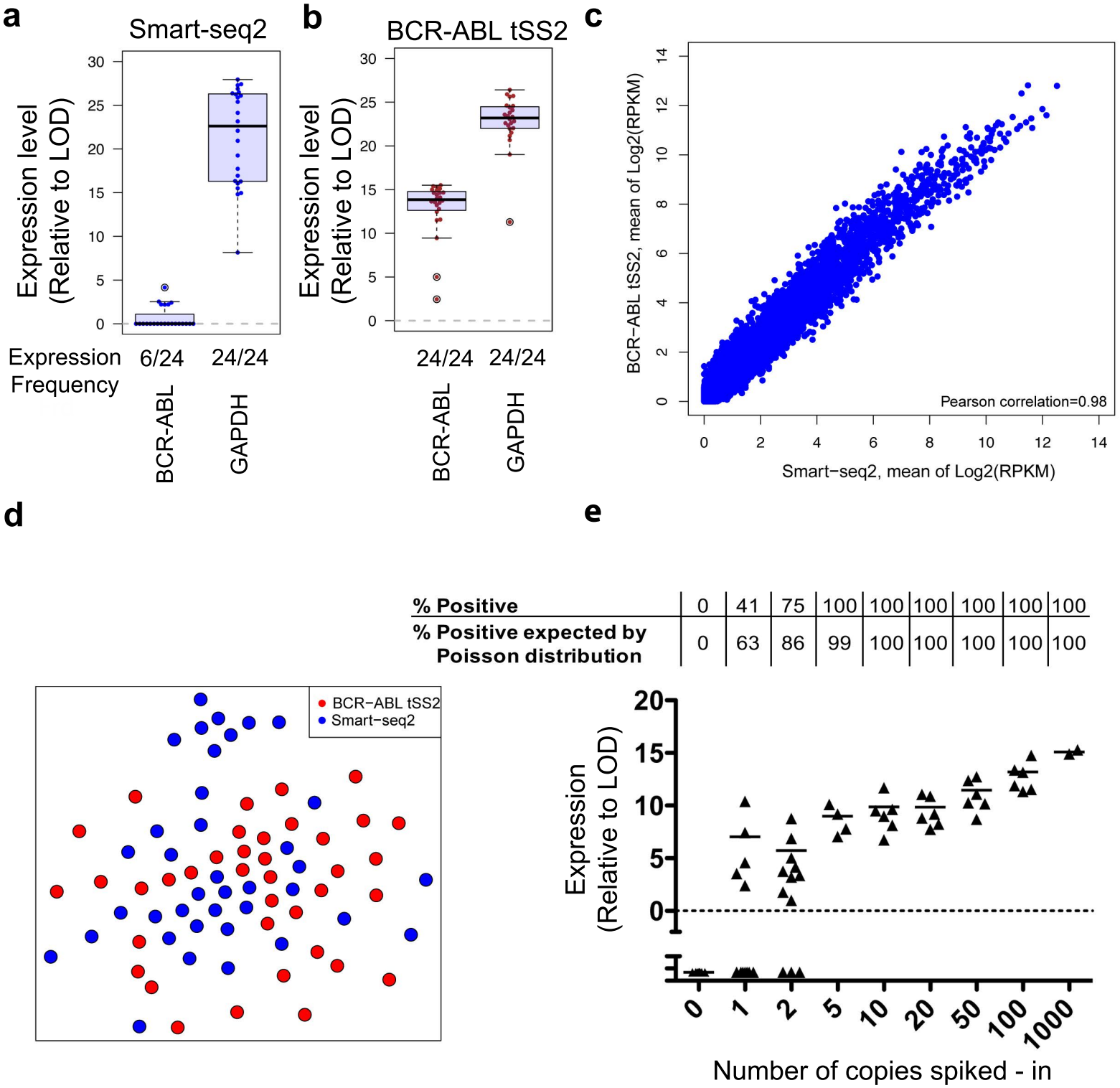
Statistical analysis

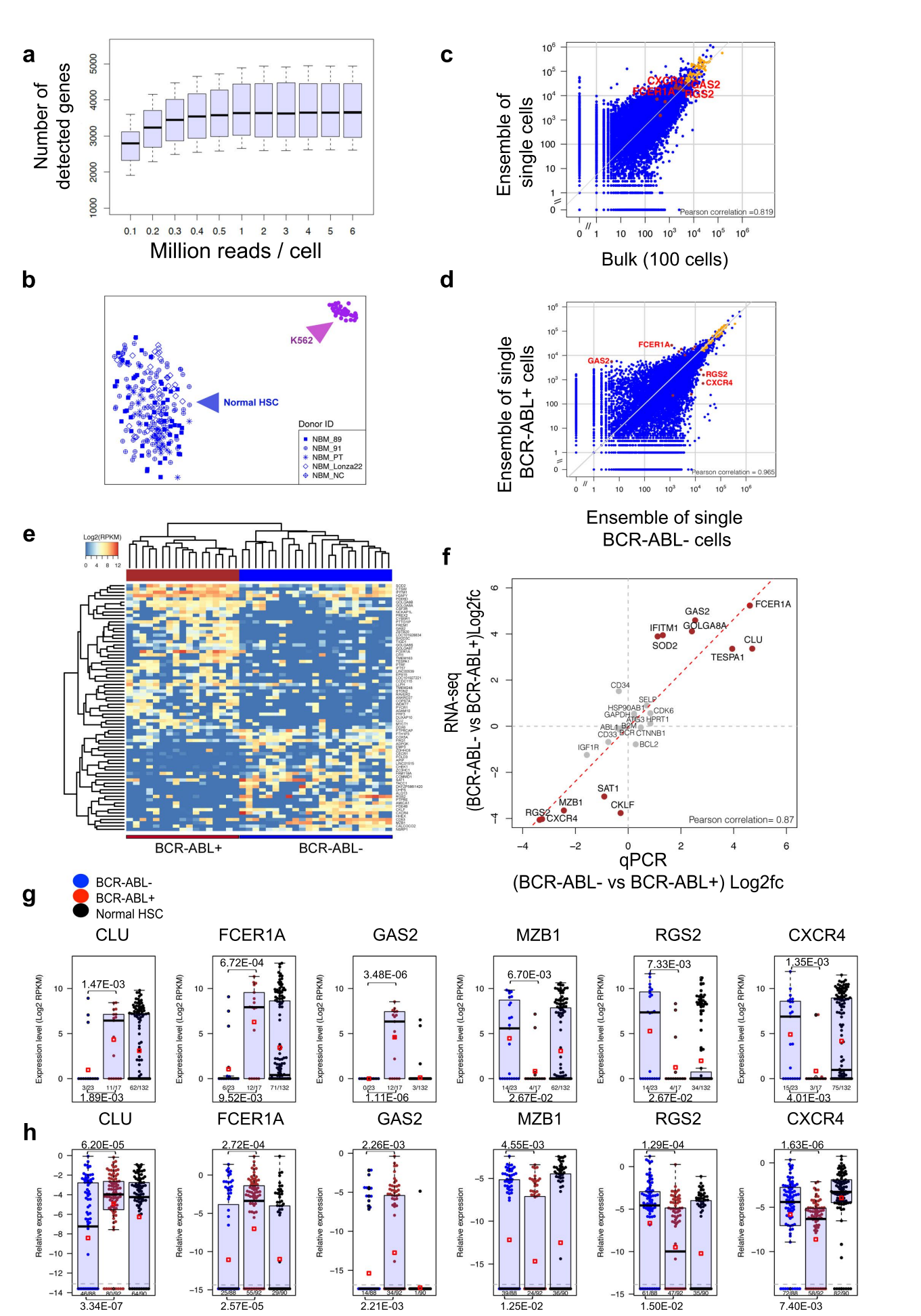
All statistical analyses were performed in R and GraphPad Prism 6 (GraphPad Software, San Diego, CA). For single-cell expression levels, nonparametric Wilcoxon test was used, and Fisher's exact test was used to compare expression frequencies at the single cell level between defined populations. No statistical method was used to predetermine sample size, and experiments were not randomized. The Investigators were not blinded to allocation during experiments or outcome assessment.

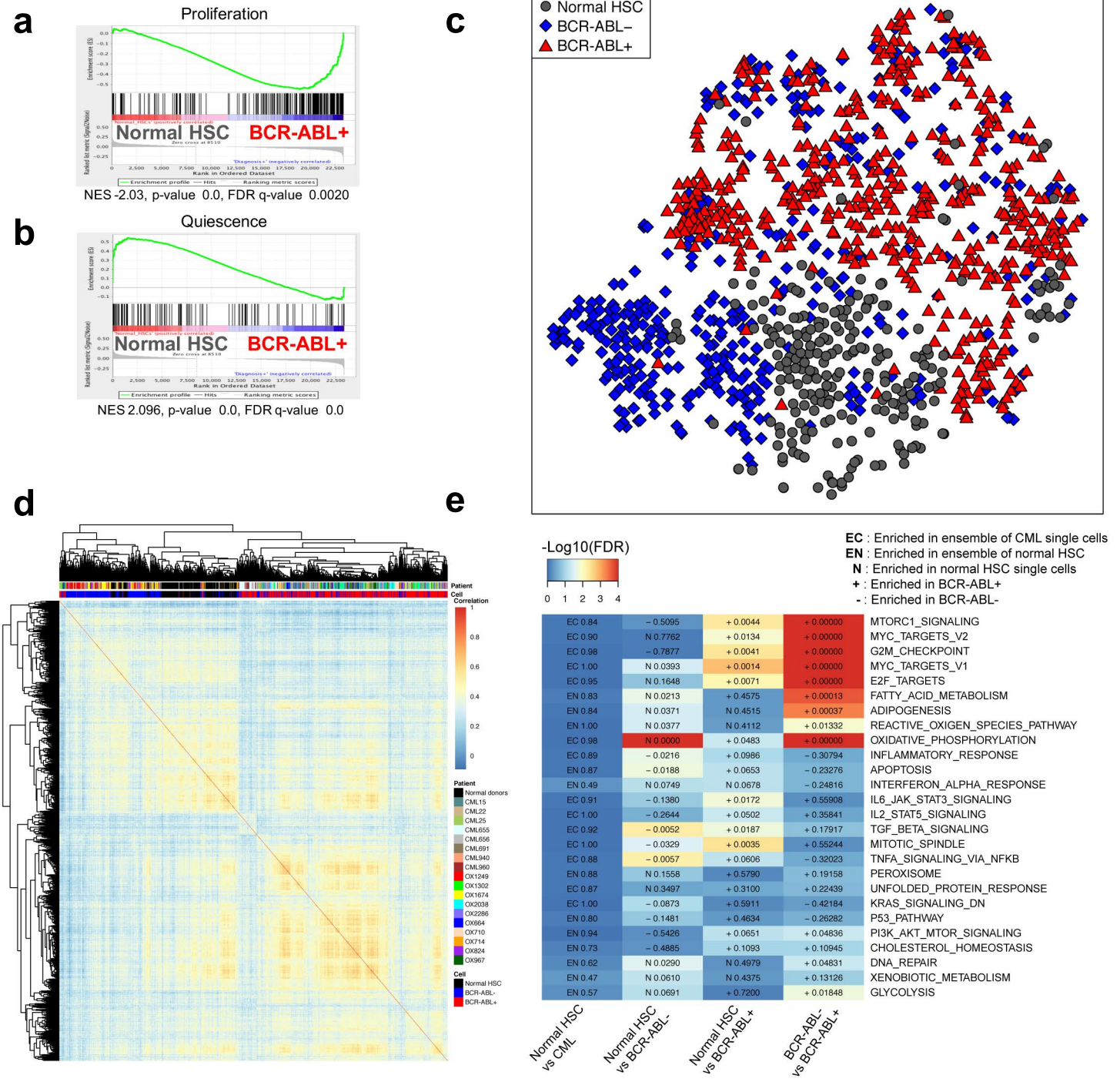
Methods-only References

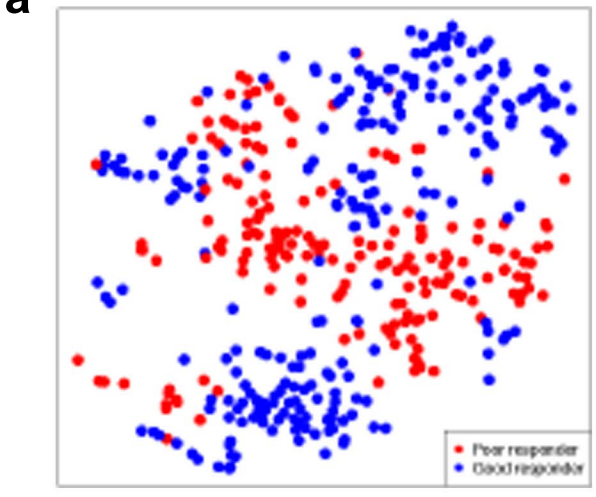
49. Sanjuan-Pla, A.*, et al.* Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature* **502**, 232-236 (2013).
50. Guo, G.*, et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* **18**, 675-685 (2010).
51. Kim, D.*, et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
52. Ramskold, 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).
53. Pollen, A.A.*, et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* **32**, 1053-+ (2014).
54. Streets, A.M. & Huang, Y.Y. How deep is enough in single-cell RNA-seq? *Nat Biotechnol* **32**, 1005-1006 (2014).
55. Iglewicz, B. & Hoaglin, D.C. *How to detect and handle outliers*, (ASQC Quality Press, Milwaukee, Wis., 1993).
56. Zeisel, A.*, et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).

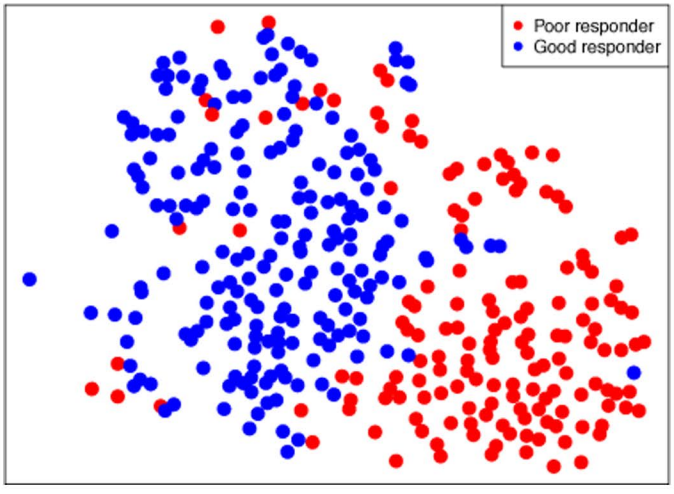
57. Saadatpour, A., Guo, G., Orkin, S.H. & Yuan, G.C. Characterizing heterogeneity in leukemic cells using single-cell gene expression analysis. *Genome Biol* **15**, 525 (2014).
58. Ritchie, M.E.*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
59. Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
60. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
61. Beck, D.*, et al.* Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood* **122**, e12-22 (2013).
62. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
63. Zhang, Y.*, et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137 (2008).

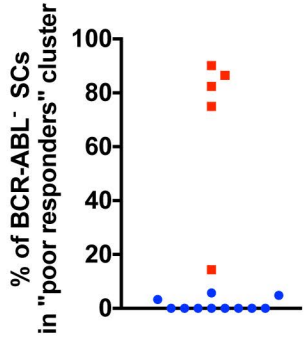

a
Smart-seq2
Expression level (Relative to LOD)
Expression Frequency
6/24
24/24
BCR-ABL
GAPDH
b
BCR-ABL tSS2
Expression level (Relative to LOD)
24/24
24/24
BCR-ABL
GAPDH
c
BCR−ABL tSS2, mean of Log2(RPKM)
Smart−seq2, mean of Log2(RPKM)
Pearson correlation=0.98
d
BCR−ABL tSS2
Smart−seq2
e
% Positive 0 41 75 100 100 100 100 100 100
% Positive expected by Poisson distribution 0 63 86 99 100 100 100 100 100
Expression (Relative to LOD)
Number of copies spiked - in

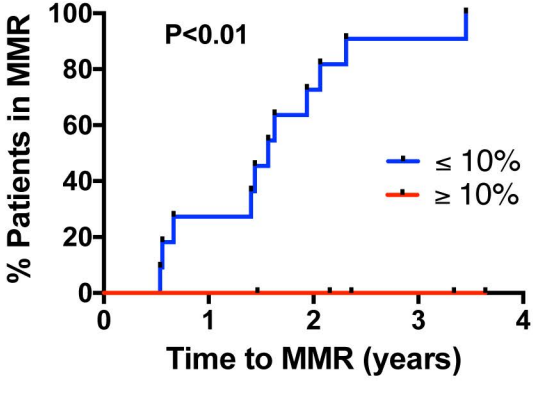

a
Number of detected genes
Million reads / cell
b
K562
Normal HSC
Donor ID
NBM_89
NBM_91
NBM_PT
NBM_Lonza22
NBM_NC
c
Ensemble of single cells
Bulk (100 cells)
Pearson correlation =0.819
d
Ensemble of single BCR-ABL+ cells
Ensemble of single BCR-ABL- cells
Pearson correlation = 0.965
e
Log2(RPKM)
BCR-ABL+
BCR-ABL-
f
RNA-seq (BCR-ABL- vs BCR-ABL+)Log2fc
qPCR (BCR-ABL- vs BCR-ABL+) Log2fc
Pearson correlation= 0.87
g
BCR-ABL-
BCR-ABL+
Normal HSC
CLU
FCER1A
GAS2
MZB1
RGS2
CXCR4
Expression level (Log2 RPKM)
h
Relative expression

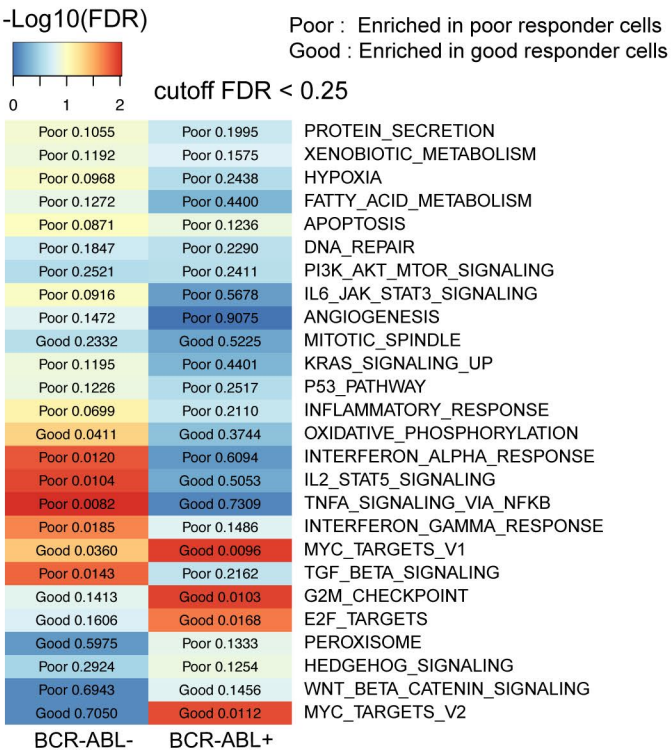

a
Proliferation
Normal HSC
BCR-ABL+
NES -2.03, p-value 0.0, FDR q-value 0.0020
b
Quiescence
Normal HSC
BCR-ABL+
NES 2.096, p-value 0.0, FDR q-value 0.0
c
Normal HSC
BCR-ABL−
BCR-ABL+
d
Patient
Cell
Correlation
Patient
Normal donors
CML15
CML22
CML25
CML655
CML656
CML691
CML940
CML960
OX1249
OX1302
OX1674
OX2038
OX2286
OX664
OX710
OX714
OX824
OX967
Cell
Normal HSC
BCR-ABL-
BCR-ABL+
e
-Log10(FDR)
0 1 2 3 4
EC : Enriched in ensemble of CML single cells
EN : Enriched in ensemble of normal HSC
N : Enriched in normal HSC single cells
+ : Enriched in BCR-ABL+
- : Enriched in BCR-ABL-
EC 0.84 − 0.5095 + 0.0044 + 0.00000 MTORC1_SIGNALING
EC 0.90 N 0.7762 + 0.0134 + 0.00000 MYC_TARGETS_V2
EC 0.98 − 0.7877 + 0.0041 + 0.00000 G2M_CHECKPOINT
EC 1.00 N 0.0393 + 0.0014 + 0.00000 MYC_TARGETS_V1
EC 0.95 N 0.1648 + 0.0071 + 0.00000 E2F_TARGETS
EN 0.83 N 0.0213 + 0.4575 + 0.00013 FATTY_ACID_METABOLISM
EN 0.84 N 0.0371 N 0.4515 + 0.00037 ADIPOGENESIS
EN 1.00 N 0.0377 N 0.4112 + 0.01332 REACTIVE_OXIGEN_SPECIES_PATHWAY
EC 0.98 N 0.0000 + 0.0483 + 0.00000 OXIDATIVE_PHOSPHORYLATION
EC 0.89 − 0.0216 + 0.0986 − 0.30794 INFLAMMATORY_RESPONSE
EN 0.87 − 0.0188 + 0.0653 − 0.23276 APOPTOSIS
EN 0.49 N 0.0749 N 0.0678 − 0.24816 INTERFERON_ALPHA_RESPONSE
EC 0.91 − 0.1380 + 0.0172 + 0.55908 IL6_JAK_STAT3_SIGNALING
EC 1.00 − 0.2644 + 0.0502 + 0.35841 IL2_STAT5_SIGNALING
EC 0.92 − 0.0052 + 0.0187 + 0.17917 TGF_BETA_SIGNALING
EC 1.00 − 0.0329 + 0.0035 + 0.55244 MITOTIC_SPINDLE
EC 0.88 − 0.0057 + 0.0606 − 0.32023 TNFA_SIGNALING_VIA_NFKB
EN 0.88 N 0.1558 + 0.5790 + 0.19158 PEROXISOME
EC 0.87 N 0.3497 + 0.3100 + 0.22439 UNFOLDED_PROTEIN_RESPONSE
EC 1.00 − 0.0873 + 0.5911 − 0.42184 KRAS_SIGNALING_DN
EN 0.80 − 0.1481 + 0.4634 − 0.26282 P53_PATHWAY
EN 0.94 − 0.5426 + 0.0651 + 0.04836 PI3K_AKT_MTOR_SIGNALING
EN 0.73 − 0.4885 + 0.1093 + 0.10945 CHOLESTEROL_HOMEOSTASIS
EN 0.62 N 0.0290 N 0.4979 + 0.04831 DNA_REPAIR
EN 0.47 N 0.0610 N 0.4375 + 0.13126 XENOBIOTIC_METABOLISM
EN 0.57 N 0.0691 + 0.7200 + 0.01848 GLYCOLYSIS
Normal HSC vs CML
Normal HSC vs BCR-ABL-
Normal HSC vs BCR-ABL+
BCR-ABL- vs BCR-ABL+

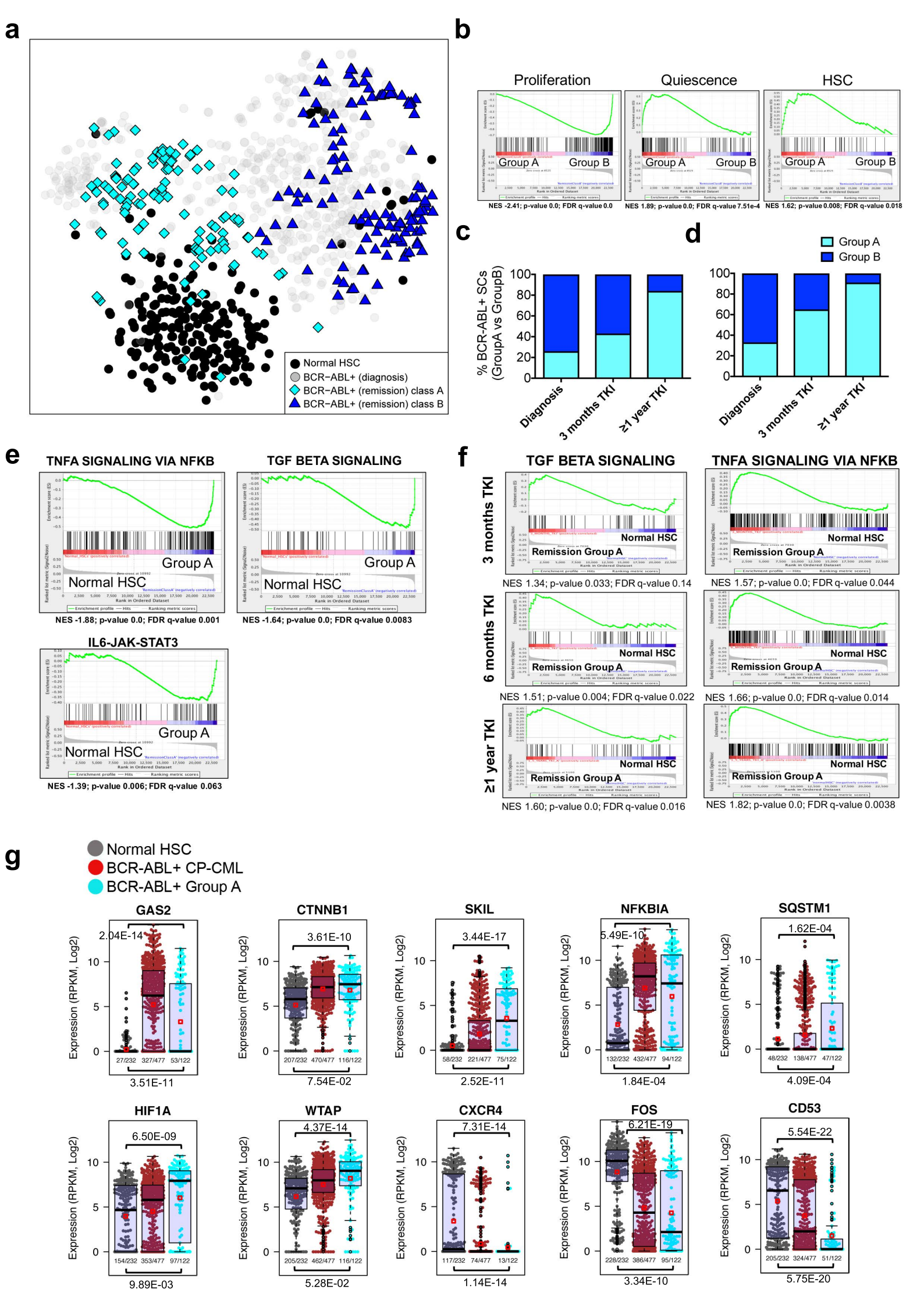

Poor responder
Good responder

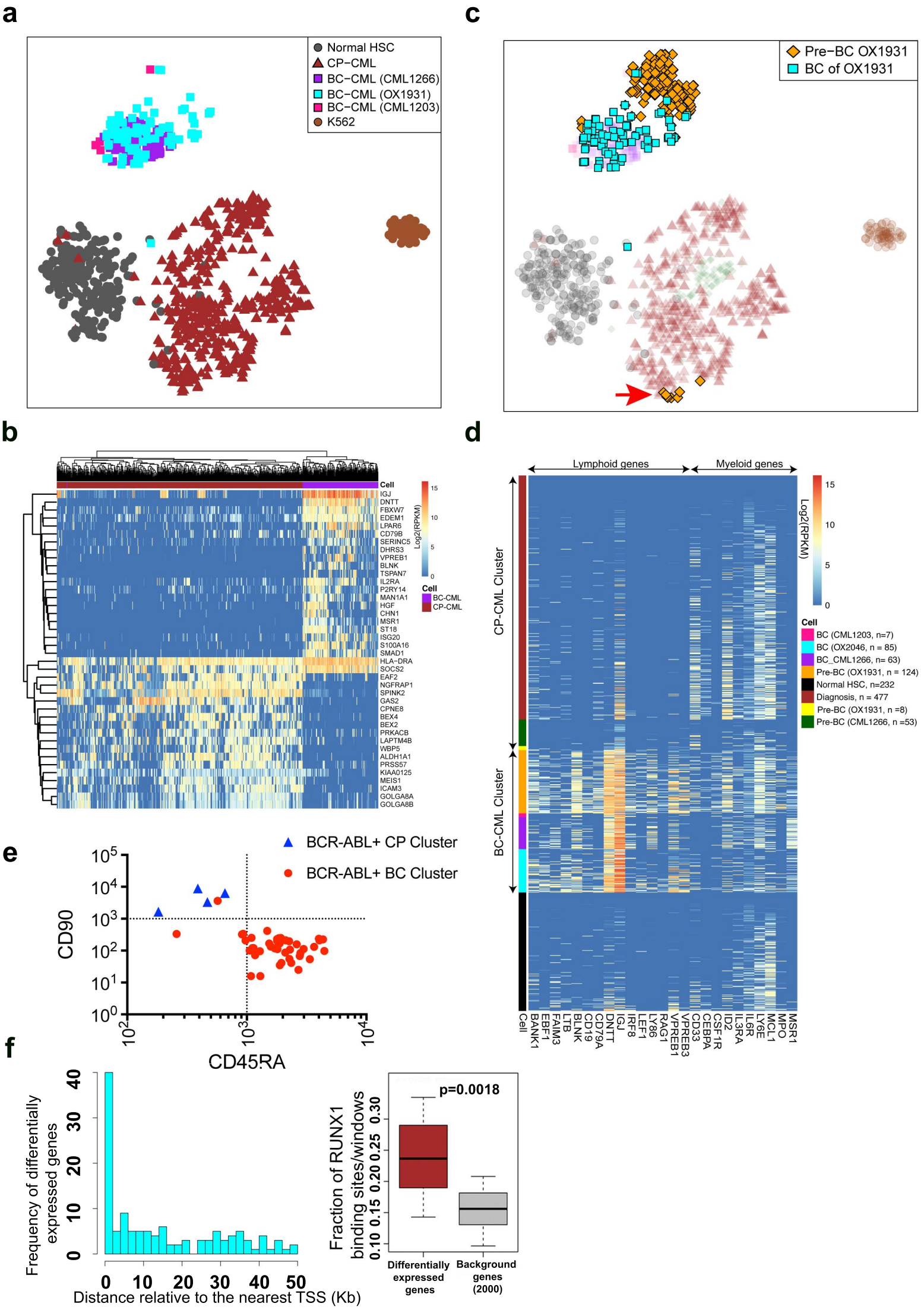

Poor responder
Good responder


% of BCR-ABL- SCs in "poor responders" cluster
100
80
60
40
20
0


P<0.01
% Patients in MMR
100
80
60
40
20
0
0
1
2
3
4
Time to MMR (years)
≤ 10%
≥ 10%


-Log10(FDR)
0
1
2
Poor : Enriched in poor responder cells
Good : Enriched in good responder cells
cutoff FDR < 0.25
Poor 0.1055
Poor 0.1995
PROTEIN_SECRETION
Poor 0.1192
Poor 0.1575
XENOBIOTIC_METABOLISM
Poor 0.0968
Poor 0.2438
HYPOXIA
Poor 0.1272
Poor 0.4400
FATTY_ACID_METABOLISM
Poor 0.0871
Poor 0.1236
APOPTOSIS
Poor 0.1847
Poor 0.2290
DNA_REPAIR
Poor 0.2521
Poor 0.2411
PI3K_AKT_MTOR_SIGNALING
Poor 0.0916
Poor 0.5678
IL6_JAK_STAT3_SIGNALING
Poor 0.1472
Poor 0.9075
ANGIOGENESIS
Good 0.2332
Good 0.5225
MITOTIC_SPINDLE
Poor 0.1195
Poor 0.4401
KRAS_SIGNALING_UP
Poor 0.1226
Poor 0.2517
P53_PATHWAY
Poor 0.0699
Poor 0.2110
INFLAMMATORY_RESPONSE
Good 0.0411
Good 0.3744
OXIDATIVE_PHOSPHORYLATION
Poor 0.0120
Poor 0.6094
INTERFERON_ALPHA_RESPONSE
Poor 0.0104
Good 0.5053
IL2_STAT5_SIGNALING
Poor 0.0082
Good 0.7309
TNFA_SIGNALING_VIA_NFKB
Poor 0.0185
Poor 0.1486
INTERFERON_GAMMA_RESPONSE
Good 0.0360
Good 0.0096
MYC_TARGETS_V1
Poor 0.0143
Poor 0.2162
TGF_BETA_SIGNALING
Good 0.1413
Good 0.0103
G2M_CHECKPOINT
Good 0.1606
Good 0.0168
E2F_TARGETS
Good 0.5975
Poor 0.1333
PEROXISOME
Poor 0.2924
Poor 0.1254
HEDGEHOG_SIGNALING
Poor 0.6943
Good 0.1456
WNT_BETA_CATENIN_SIGNALING
Good 0.7050
Good 0.0112
MYC_TARGETS_V2
BCR-ABL-
BCR-ABL+


a
Normal HSC
BCR-ABL+ (diagnosis)
BCR-ABL+ (remission) class A
BCR-ABL+ (remission) class B
b
Proliferation
Quiescence
HSC
Group A
Group B
NES -2.41; p-value 0.0; FDR q-value 0.0
NES 1.89; p-value 0.0; FDR q-value 7.51e-4
NES 1.62; p-value 0.008; FDR q-value 0.018
c
% BCR-ABL+ SCs (GroupA vs GroupB)
Diagnosis
3 months TKI
≥1 year TKI
d
Group A
Group B
e
TNFA SIGNALING VIA NFKB
TGF BETA SIGNALING
IL6-JAK-STAT3
Group A
Normal HSC
NES -1.88; p-value 0.0; FDR q-value 0.001
NES -1.64; p-value 0.0; FDR q-value 0.0083
NES -1.39; p-value 0.006; FDR q-value 0.063
f
TGF BETA SIGNALING
TNFA SIGNALING VIA NFKB
3 months TKI
6 months TKI
≥1 year TKI
Normal HSC
Remission Group A
NES 1.34; p-value 0.033; FDR q-value 0.14
NES 1.57; p-value 0.0; FDR q-value 0.044
NES 1.51; p-value 0.004; FDR q-value 0.022
NES 1.66; p-value 0.0; FDR q-value 0.014
NES 1.60; p-value 0.0; FDR q-value 0.016
NES 1.82; p-value 0.0; FDR q-value 0.0038
g
Normal HSC
BCR-ABL+ CP-CML
BCR-ABL+ Group A
GAS2
2.04E-14
27/232 327/477 53/122
3.51E-11
CTNNB1
3.61E-10
207/232 470/477 116/122
7.54E-02
SKIL
3.44E-17
58/232 221/477 75/122
2.52E-11
NFKBIA
5.49E-10
132/232 432/477 94/122
1.84E-04
SQSTM1
1.62E-04
48/232 138/477 47/122
4.09E-04
HIF1A
6.50E-09
154/232 353/477 97/122
9.89E-03
WTAP
4.37E-14
205/232 462/477 116/122
5.28E-02
CXCR4
7.31E-14
117/232 74/477 13/122
1.14E-14
FOS
6.21E-19
228/232 386/477 95/122
3.34E-10
CD53
5.54E-22
205/232 324/477 51/122
5.75E-20
Expression (RPKM, Log2)


a
Normal HSC
CP-CML
BC-CML (CML1266)
BC-CML (OX1931)
BC-CML (CML1203)
K562
c
Pre-BC OX1931
BC of OX1931
b
Cell
IGJ
DNTT
FBXW7
EDEM1
LPAR6
CD79B
SERINC5
DHRS3
VPREB1
BLNK
TSPAN7
IL2RA
P2RY14
MAN1A1
HGF
CHN1
MSR1
ST18
ISG20
S100A16
SMAD1
HLA-DRA
SOCS2
EAF2
NGFRAP1
SPINK2
GAS2
CPNE8
BEX4
BEX2
PRKACB
LAPTM4B
WBP5
ALDH1A1
PRSS57
KIAA0125
MEIS1
ICAM3
GOLGA8A
GOLGA8B
Log2(RPKM)
15
10
5
0
Cell
BC-CML
CP-CML
d
Lymphoid genes
Myeloid genes
CP-CML Cluster
BC-CML Cluster
Log2(RPKM)
15
10
5
0
Cell
BC (CML1203, n=7)
BC (OX2046, n = 85)
BC_CML1266, n= 63)
Pre-BC (OX1931, n = 124)
Normal HSC, n=232
Diagnosis, n = 477
Pre-BC (OX1931, n =8)
Pre-BC (CML1266, n =53)
Cell
BANK1
EBF1
FAIM3
LTB
BLNK
CD19
CD79A
DNTT
IGJ
IRF8
LEF1
LY86
RAG1
VPREB1
VPREB3
CD33
CEBPA
CSF-1R
ID2
IL3RA
IL6R
LY6E
MCL1
MPO
MSR1
e
BCR-ABL+ CP Cluster
BCR-ABL+ BC Cluster
CD90
CD45RA
f
Frequency of differentially expressed genes
Distance relative to the nearest TSS (Kb)
Fraction of RUNX1 binding sites/windows
p=0.0018
Differentially expressed genes
Background genes (2000)