Cell

Single-Cell Exome Sequencing and Monoclonal Evolution of a *JAK2*-Negative Myeloproliferative Neoplasm

Yong Hou,[1,2,3,11] Luting Song,[1,4,5,6,11] Ping Zhu,[7,11] Bo Zhang,[1,11] Ye Tao,[1,11] Xun Xu,[1] Fuqiang Li,[1] Kui Wu,[1] Jie Liang,[1] Di Shao,[1] Hanjie Wu,[1] Xiaofei Ye,[1] Chen Ye,[1] Renhua Wu,[1] Min Jian,[1] Yan Chen,[7] Wei Xie,[1,3] Ruren Zhang,[1,3] Lei Chen,[1,4,5,6] Xin Liu,[1] Xiaotian Yao,[1] Hancheng Zheng,[1] Chang Yu,[1] Qibin Li,[1] Zhuolin Gong,[1] Mao Mao,[8] Xu Yang,[1] Lin Yang,[1] Jingxiang Li,[1] Wen Wang,[5] Zuhong Lu,[2,3] Ning Gu,[2,3] Goodman Laurie,[1] Lars Bolund,[1] Karsten Kristiansen,[1,9] Jian Wang,[1] Huanming Yang,[1] Yingrui Li,[1,*] Xiuqing Zhang,[1,*] and Jun Wang[1,9,10,*]

[1]BGI-Shenzhen, Shenzhen, 518083, China
[2]State Key Laboratory of Bioelectronics
[3]School of Biological Science and Medical Engineering
Southeast University, Nanjing 210096, China
[4]College of Life Sciences, Wuhan University, Wuhan 430072, China
[5]CAS-Max Planck Junior Research Group, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (CAS), Kunming, Yunnan 650223, China
[6]Graduate School of Chinese Academy of Sciences, Beijing 100049, China
[7]Department of Hematology, Peking University First Hospital, Beijing 100034, China
[8]Pfizer Inc., San Diego, CA 92121, USA
[9]Department of Biology
[10]The Novo Nordisk Foundation Center for Basic Metabolic Research
University of Copenhagen, DK-1165 Copenhagen, Denmark
[11]These authors contributed equally to this work
*Correspondence: wangj@genomics.org.cn (J.W.), zhangxq@genomics.org.cn (X.Z.), liyr@genomics.org.cn (Y.L.)
DOI 10.1016/j.cell.2012.02.028

SUMMARY

Tumor heterogeneity presents a challenge for inferring clonal evolution and driver gene identification. Here, we describe a method for analyzing the cancer genome at a single-cell nucleotide level. To perform our analyses, we first devised and validated a high-throughput whole-genome single-cell sequencing method using two lymphoblastoid cell line single cells. We then carried out whole-exome single-cell sequencing of 90 cells from a *JAK2*-negative myeloproliferative neoplasm patient. The sequencing data from 58 cells passed our quality control criteria, and these data indicated that this neoplasm represented a monoclonal evolution. We further identified essential thrombocythemia (ET)-related candidate mutations such as *SESN2* and *NTRK1*, which may be involved in neoplasm progression. This pilot study allowed the initial characterization of the disease-related genetic architecture at the single-cell nucleotide level. Further, we established a single-cell sequencing method that opens the way for detailed analyses of a variety of tumor types, including those with high genetic complex between patients.

INTRODUCTION

Tumor evolution is an important area in cancer research because information on fundamental genetic changes that occur as a tumor develops is essential for effective diagnosis, prognosis, and therapy (Cairns, 1975; Nowell, 1976). However, characterizing the underpinnings of this process remains difficult. Current theories propose that neoplasms arise either from monoclonal or polyclonal somatic mutant cells, and there is evidence that supports both a gradual change and an instantaneous change in the genome to promote progression (Stephens et al., 2011; Visvader, 2011). The heterogeneous nature of tumors, however, makes it difficult for researchers to analyze the intratumoral genetic structure and identify key changes during neoplasm progression. Without additional cell sorting experiments, hematopoietic tumors, in particular, are very heterogeneous, making it especially difficult to identify genetic mutations that have a major impact on cancer development.

Myeloproliferative neoplasms are hematopoietic tumors. They originate from genetic variations occurring in hematopoietic stem cells or progenitors, which lead to abnormal differentiation and myelopoiesis. A typical myeloproliferative neoplasm is essential thrombocythemia (ET). ET is characterized by malignant myeloid and a sustained proliferation of megakaryocytes, which leads to an increasing number of circulating platelets, typically in excess of 600 × 10^9/l. ET affects about 2 out of 100,000 adults per year, with the incidence rate increasing in

recent years (Mesa et al., 1999). ET is also a slowly progressing neoplasm, with ~50% of patients being asymptomatic and the remainder presenting vasomotor, thrombotic, or hemorrhagic disturbances (Tefferi, 2001).

Previous studies on ET have provided evidence supporting both a monoclonal and a polyclonal origin for ET initiation and proliferation (Tefferi, 2010). Approximately 55% of ET patients harbor *JAK2* mutations. However, JAK2 and other rare mutations identified can be found in both ET and other types of myeloproliferative neoplasms. Thus, none of these mutations provide unique markers to infer the evolution history of ET, nor could they be traced back to a common originating clone (Tefferi, 2010). One approach to revealing the underlying genetic mechanisms of ET is to assess mutations within the individual cancer cells, which would circumvent issues of tumor heterogeneity.

Recent work on breast cancer has demonstrated the potential applicability of single-nucleus sequencing technology for characterizing tumor evolution (Navin et al., 2011). Specifically, Navin et al. (2011) investigated copy number variation (CNV) in single tumor cells using DOP WGA followed by DNA sequencing to determine cell population structure and tumor evolution patterns in a single breast tumor. This study provided an important breakthrough for research of tumor evolution and offered a way to assess the genetic details of tumor structure. This method, however, is not suitable for assessing the genetic characteristics of single tumor cells at a single-nucleotide resolution, and, at this stage, it cannot provide high genome coverage; thus, this approach does not allow the detection of the single-nucleotide changes that commonly underlie tumor development. The use of MDA for WGA analysis, however, allows greater resolution and genome coverage due to the MDA products being a higher molecular weight DNA (average length > 10 kb), which results in significantly higher genome recovery (Dean et al., 2002b). Thus, this technique, as we demonstrate here, can allow for whole-genome sequencing, with high accuracy at the nucleotide level.

Here, we present a high-throughput single-cell sequencing method to analyze tumor evolution in cancers. We use this technique to perform single-cell genetic analysis of a *JAK2*-negative ET patient. We design and test our single-cell sequencing method using two single cells from a lymphoblastoid cell line and evaluated its whole-genome recovery, amplification uniformity, sensitivity, and specificity. We then perform whole-exome sequencing of 90 single cells from a typical *JAK2*-negative ET patient and obtain a comprehensive genetic landscape of ET. Analysis of the somatic mutant allele frequency spectrum (SMAFS) reveals that this ET neoplasm was likely of monoclonal origin, and statistical analyses of the genes carrying mutations provide a list of candidate genes that might be involved in neoplasm progression.

RESULTS

Whole-Genome Single-Cell Sequencing of Cells Derived from a Previously Sequenced Individual

To develop and test a method for carrying out whole-genome single-cell sequencing at a single-nucleotide level, we used cells that were derived from an individual whose genome had been previously sequenced. This genome would serve as a whole-tissue genome sequence control to assess our method. We therefore constructed a lymphoblastoid cell line from lymphoblast cells previously obtained at BGI from the individual (YH) who provided DNA for the first Asian diploid genome sequence (Wang et al., 2008). Under an inverted microscope and through cascade dilution, we randomly extracted two single cells (hereafter referred to as YH-1 and YH-2) from the YH lymphoblastoid cell line using standard manipulation methods (Spits et al., 2006) (Figure S1A available online). We then carried out whole-genome amplification (WGA) based on multiple displacement amplification (MDA), which uses the Φ29 enzyme to amplify DNA in a linear process (Dean et al., 2002a), on the DNA from each single cell. The amplicon showed a peak size at ~23 kb, which is much longer than the degenerate oligonucleotide-primed (DOP; randomly amplify the whole genome with degenerated PCR primer) PCR (<1 kb) from the same samples (Figures S1B and S1C) and from others reports (Navin et al., 2011). We used DNA fluorometry-based quantitation to select products that were quantitation qualified (see Experimental Procedures). These were assayed for genomic integrity using ten housekeeping gene PCR tests for each single cell. The products that passed this filter underwent massively parallel whole-genome single-cell sequencing using paired-end 100-bp reads and ~350 bp size inserts. We uniquely mapped and aligned 42.27 Gb (YH-1) and 47.65 Gb (YH-2) of high-quality sequences to the human reference genome (Hg18) with the SOAP program (Li et al., 2009b) and obtained 97.25% of bases with 15.88 × mean fold coverage of the whole genome of YH-1 and 95.64% of bases with 17.90 × mean fold coverage of the whole genome of YH-2 (Table 1). We also carried out massively parallel whole-genome sequencing on a multicell sample from the YH lymphoblastoid cell line as a control and obtained 99.91% of bases with ~18 × mean fold coverage of the whole genome (comparable with the single-cell data) (Table 1). For each sample, using only the reads that could be uniquely mapped to the reference genome, we identified single-nucleotide polymorphisms (SNPs) with SOAPsnp (Li et al., 2009b).

For standard genome sequencing in which researchers are trying to identify mutations involved in disease, 30× is the typical sequencing depth. However, here we are using the sequencing data of two identical YH cells to assess the fidelity of our amplification method (i.e., allele dropout (ADO) and the false discovery rate), not to identify mutations. Thus, a mean sequencing depth of 15×–18× is sufficient, given that the sequence for these cells is already known.

Single-Cell Sequencing Data Are of High Sensitivity and Have a Distinct Genome Distribution from Tissue Sequencing

To evaluate the whole-genome recovery, amplification uniformity, and accuracy of single-cell sequencing, we used a Circos map (Krzywinski et al., 2009) to graphically assess the whole-genome coverage and characteristics of our single-cell sequencing data by looking at the sequence recovery and distribution uniformity on the whole genome (Figure 1). The sequence from both single cells covered more than 90% of the human reference genome (Figures 1C and 1D), and additional

Table 1. Single-Cell Sequencing Data Generation and Evaluation

	Whole Genome		Coding Region				Data Evaluation	
Samples	mean depth	% of bases at ≥ 1 depth	mean depth	% of bases at ≥ 1 depth	% of bases at ≥ 15 depth	% of bases at ≥ 18 depth	FD#/#	ADO #/#
YH-1	15.88	97.25	18.95	94.05	18.09	13.96	2/99,152 (2.02×10^{-5})	18,807/246,314 (7.64%)
YH-2	17.9	95.64	19.13	91.74	35.43	29.50	3/99,152 (3.03×10^{-5})	36,523/246,314 (14.8%)
Control	18.04	99.91	20.53	97.30	62.80	47.20	–	–

(FD #/#) False discovery site number/high-confidence homozygous control subsets number. High-confidence homozygous control subsets were defined as sites that are consistent between control (subset with quality score of 99) and the first Asian diploid genome study (Illumina 1M genotyping and Illumina sequencing). (ADO #/#) ADO site number/high-confidence heterozygous control subsets number. High-confidence heterozygous control subsets were defined as the sites that are consistent in the first Asian diploid genome study between Illumina 1M genotyping and Illumina sequencing.

analysis showed that single-cell sequencing retrieved > 95% of the bases in the reference genome at a 15× sequencing depth (Figure S1D), which indicated high genome coverage sensitivity. The cumulative distribution of the sequencing fold coverage across the coding regions showed that ~25% of the bases were covered by 18× or more in the single-cell sequencing data, which is 50% less than that of the multicell (unamplified) control (red and blue in Figure S1E). The lower coverage was expected, as there was likely bias in the amplification process.

We also looked at the distribution of the sequence data across the genome to determine whether there were genomic regions that had a specific impact on the WGA process. The data distribution exhibited a correlation with the GC content distribution in some regions of the genomes (Figures 1B–1D). Accordingly, we examined how the single-cell sequencing reads were distributed across the genome relative to the GC content, and the distribution pattern of the reads showed that the GC content effected the even distribution of amplification products from single-cell WGA, with regions of extreme GC content showing lower amplification efficiency (Figure S1F). The median GC content in places with 0 coverage (i.e., amplification failure) in gene-coding sequence regions and the whole genome was 60.12% and 49.40%, respectively. These percentages were higher (p value < 0.001, Student's t test) than the average 41% GC content of human reference genome (Figure S1G). Similar assessments of read distribution coverage in repeat regions or at different chromosomal locations, however, showed no significant correlation (Figures S1H–S1I). These data indicated that amplification efficiency was primarily dependent on GC content.

Allele Dropout and False Discovery Affect the Sensitivity of Single-Cell Sequencing

To detect the specificity and fidelity of the genomic sequence from the single cells, we first evaluated the allele dropout (ADO), which shows whether nonamplification occurred in one of the alleles present in a heterozygous sample, as loss of heterozygous sites would lead to calling false negatives in the single-cell sequencing data. We calculated the false negative ratio per sequencing depth using the multicell sample sequence as a control and determined the ADO per cell as the median of false negative ratios (Figure S2A). The average ADO ratio for all single cells was 11% (Table 1), which is comparable to that of previous analyses (Spits et al., 2006), indicating that the single-cell sequences are of standard quality. As an additional means to determine the specificity and fidelity of single-cell sequencing, we evaluated the false discovery rate in the single-cell sequencing data. The false discovery ratio is defined as a false discovery (FD) heterozygous site in a homozygous sample, which might arise due to amplification, hybridization, or sequencing errors. Taking the multicell sample sequence as control, we found that only two~three bases in the single-cell sequencing were discrepant within a subset of 99,152 high-confidence homozygous background sites, indicating that our single-cell sequencing had an extremely low error rate. The average false discovery ratio of our single-cell sequencing was ~2.52×10^{-5} (Table 1), which was similar to that of the multicell sample sequencing using the same sequencing platform, according to a previous report (Bentley et al., 2008). As YH was a male, we further assessed our data quality by determining the false heterozygous allele rate across the X chromosome between each single cell and the multicell sample sequencing data (Figure S2B). We also examined the sequencing data of the mitochondrial DNA from the two YH single cells and found no false discovery site, further indicating the low amplification error of our methodology.

Allele Dropout and False Discovery Show No Bias Relative to Genomic Region or Base Type

To assess the impact of ADO and false discovery artifacts in our sequence, we analyzed the distribution of these artifacts relative to genomic region and base type. We analyzed the distribution across chromosome 1, and, with the exception of regions near the telomeres and centromere, there was an even distribution of the ADO and false discovery bases (Figure 2A), indicating that the artifacts occurred randomly. Additionally, we built a high-resolution map of the artifact distribution to assess whether there was a prevalence of common artifacts between two single cells (Figure 2A) and found very few that were in common. These analyses also provided a method by which to detect true nucleotide changes from amplification and sequencing artifacts when mutations are present in a substantial number of individual cells.

We next evaluated the per base characteristics of these artifacts. We determined whether the ADOs were random with regard to different bases by calculating the number of ADO for

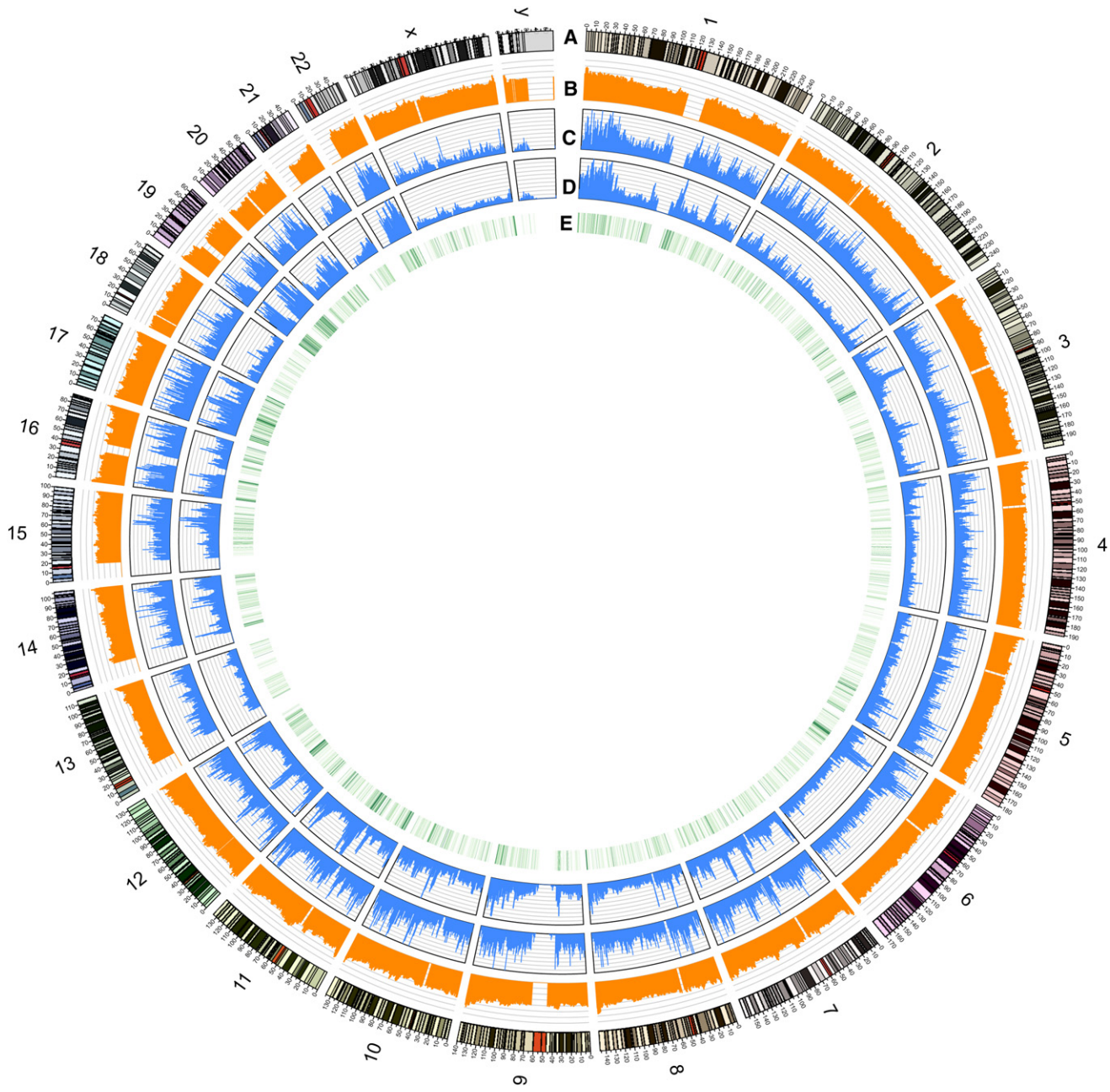

Figure 1. Graphic Representation of the Whole Genome of Two Single YH Cells

(A) Karyotype of the human reference genome (Hg18).

(B) GC content distribution of the reference genome (height of orange rectangles ranges from 0%–70%, bin = 1 Mb). (C) Whole-genome coverage of YH-2 (height of blue rectangles ranges from 0×–40×, bin = 1 Mb).

(D) Whole-genome coverage of YH-1 (height of blue rectangles ranges from 0×–40×, bin = 1 Mb).

(E) Gene density across the reference genome (Hg18) (gradually changing green represents from 0 to 30 genes per 100 kb).

See also Figure S1.

each of the four base types of YH-1 and YH-2 in heterozygous SNPs of the multicell sequence data (Figure 2B). We observed no significant difference (Fisher's exact test, $p = 0.3$) between the ADO numbers for the four base types. We also evaluated the "mutation change" base distribution spectrum of the false discovery sites. Unlike the ADO, the false discovery bases showed a preference for C:G to T:A changes (Fisher's exact test, $p < 0.01$) (Figure 2C). This same pattern has been observed previously in other tumor sequencing data (Greenman et al., 2007) and is also fit with the expectation that transitions occur

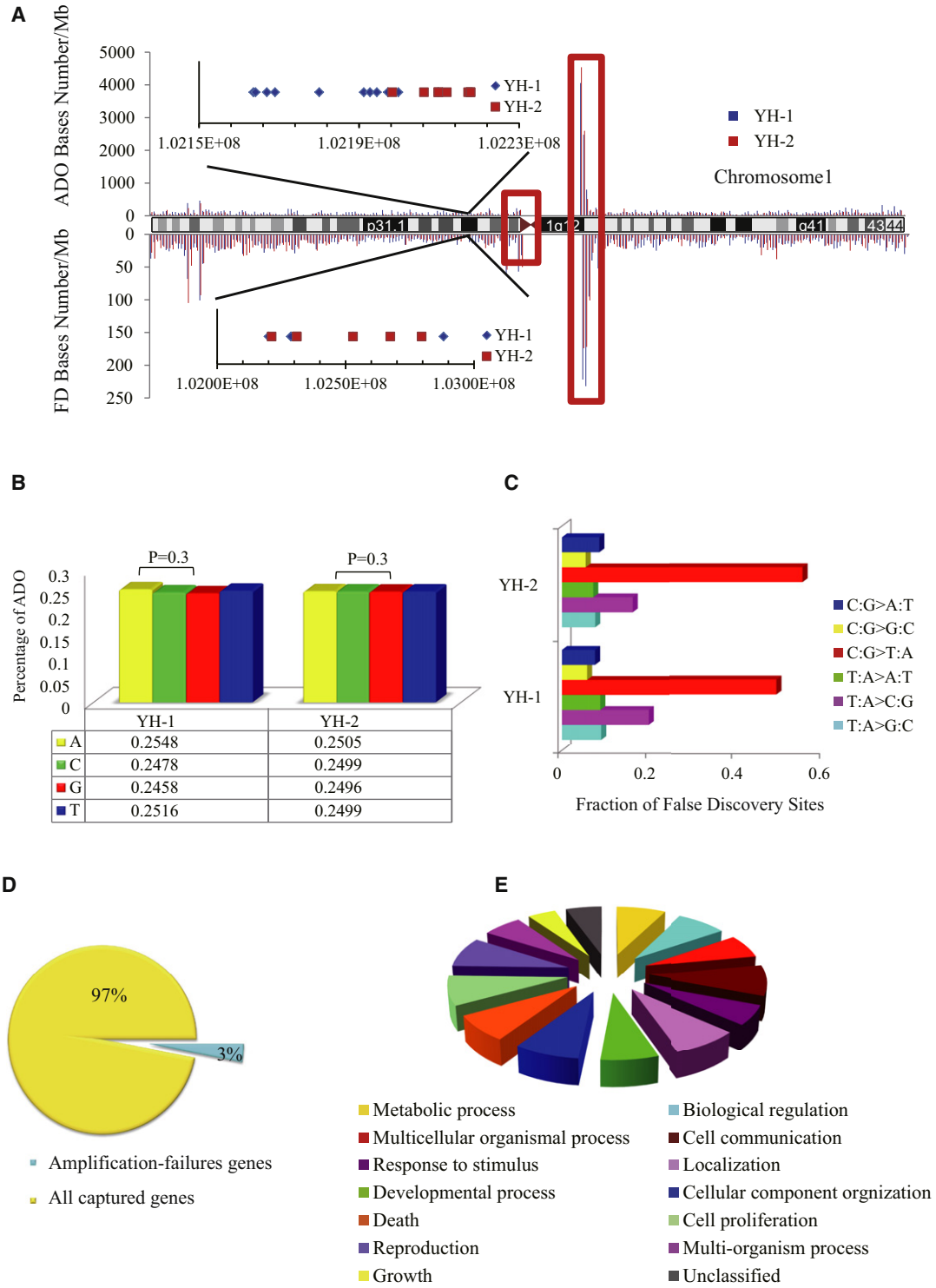

Figure 2. Characteristics of Artifacts in the Two YH Cells Induced by Single-Cell Sequencing

(A) Distribution of ADO and false discovery bases across chromosome 1.

(B) Evaluation of ADO on the four different base types.

(C) Evaluation of false discovery on different types of base changes.

(D) The fraction of genes that failed to amplify in all captured genes.

(E) Relative distribution of amplification failure ratio of biological categories of genes.

See also Figure S2.

Table 2. Whole-Exome Sequencing Data Production of the ET Mixed Multicell Sample and the Oral Mucosal Epithelium Mixed Multicell Sample

	Sample	Mean Fold Coverage	Fraction of Targets Covered by at Least 1×	Fraction of Targets Covered by at Least 20 ×
Tissue sequencing	LN-T1	57.62	0.97	0.58
	LC-T1	91.12	0.99	0.88

Whole exome that we captured was 36,184,808 bp. LN-T1 is the oral mucosal epithelium mixed multicell sample, and LC-T1 is the ET mixed multicell sample.

more readily than transversions during cell division. These data indicated that the presence of artifacts will have very limited artifactual impact on the mutation pattern that we would see in our ET by single-cell sequencing. Additionally, our analyses here allowed us a means to detect and remove artifacts that do occur and thus improve the quality of our mutation calling quality, and this was incorporated into our SNP calling method described below in our ET single-cell sequencing.

WGA Errors Do Not Selectively Affect Genes with Specific Biological Functions

We further assessed whether the genomic regions that were not sufficiently covered in our single-cell sequencing data were likely to influence our biological analysis. Here, we selected genes from areas in which amplification failed and identified 640 genes in which a total of 643 amplification failure exons (Figure 2D). Gene ontology (GO) mapping analysis showed that there was no enrichment with regard to any specific biological function for any of these genes (Figures 2E and S2C). The correlation coefficient (R^2) of each proportion of biological process category between amplification failure genes and all captured genes was 0.99 (Figure S2D), showing that these data were highly consistent and had no systematic bias in gene amplification failure. Thus, it is unlikely that WGA led to selective loss of genes involved in specific biological pathways, and thus it would have limited influence in subsequent biological analyses. Taking together all of the above error bias analyses, these results indicated that our single-cell sequencing strategy is of high sensitivity and specificity and can be used to carry out accurate genetic analyses.

Single-Cell Exome Sequencing of a *JAK2*-Negative ET Patient

Having established the single-cell sequencing technology as a qualified means to explore genetic changes at a single-cell nucleotide level, we used this method to investigate intercellular genomic heterogeneity and tumor evolution in ET. We applied our single-cell sequencing method to individual cells from a previously healthy 58-year-old Chinese male ET patient without a family history of leukemia. This patient is a typical *JAK2*V617F wild-type (validated by PCR-Sanger sequencing) ET patient with more than 80% abnormal myeloid cells within his bone marrow (Figure S3A), from whom we could infer neoplasm evolution and could potentially also obtain novel genetic information on this type of cancer in this individual (see Extended Experimental Procedures for detailed clinical report).

Using sufficient dispersion and cascade dilution of cells, we randomly selected 82 cells from a sampling of fresh bone marrow and 8 cells from a sampling of normal oral mucosal epithelium by micromanipulation, as described above. We sequenced matched tissue samples from fresh bone marrow and normal oral mucosal epithelium. Here, to provide a more cost-effective means of analysis and because our interest was to look at genes involved in cancer development, we used a standard exome sequencing strategy instead of whole-genome sequencing to assess genetic variation within the coding regions of the single cells.

In the ET single-cell study, on average, we sequenced the 37 Mb targeted exome regions of each cell (90 cells total) to a mean depth of 30× (Table S1). We evaluated the whole-exome single-cell sequencing and defined the genetic characteristics of ET using the bioinformatics analysis pipeline shown in Figure S3B. We filtered out the cells that had coverage < 70%, as the information from these is more likely to be negatively influenced by amplification/capture errors. The cells from the mature oral mucosal epithelium all had average target base coverage < 70%; thus, the 58 cells that we used for further analysis were all from the bone marrow (Table S1). In all, we obtained 58 single cells, with an average of ~70% of target bases at ≥ 5 depth. This coverage is sufficient for confident population variant calling when multiple cells carry the same variant (Table S1) (Li et al., 2010; Yi et al., 2010). For instance, if five of the cells contain the same variant at a specific site, then having a 5-fold coverage for each cell is the equivalent of calling the variant of this site at ~25-fold coverage, and this coverage is sufficient to call mutational variants when mathematical models (e.g., Bayesian estimation, maximum likelihood estimation) is incorporated into the analysis (Li et al., 2010; Yi et al., 2010).

For the corresponding tissue sequencing control, we sequenced the 37 Mb targeted exome regions of ET-mixed multicell sample (more than 1×10^6 cells) to a mean depth of 91.12× and the oral mucosal epithelium mixed multicell sample (more than 1×10^6 cells) to a mean depth of 57.62× (Table 2). Thus, we obtained more than 88% of target bases at ≥ 20× of the ET mixed multicell sample.

Population Somatic Mutation Calling and Validation of This Patient

To carry out a tumor cell population variant analysis, we needed to detect SNPs within all of our samples from which to create an overall genotype spectrum for the tumor tissue. We first used the unique reads from each sample to detect SNPs in the exome regions using SOAPsnp and assembled a consensus sequence for each individual cell and the two tissue samples. We then grouped these to define a cell population genotype, as is used in population genetic analysis (Behar et al., 2010; Yi et al., 2010).

To determine the accuracy of the whole-exome single-cell sequencing, we determined the ADO and false discovery ratio as described previously (see also Table S2). The average ADO of single-cell sequencing of ET was determined by the ADO of the 58 ET cells in the heterozygous SNPs in both tissue sequencing of the normal and ET samples using dbSNP as the

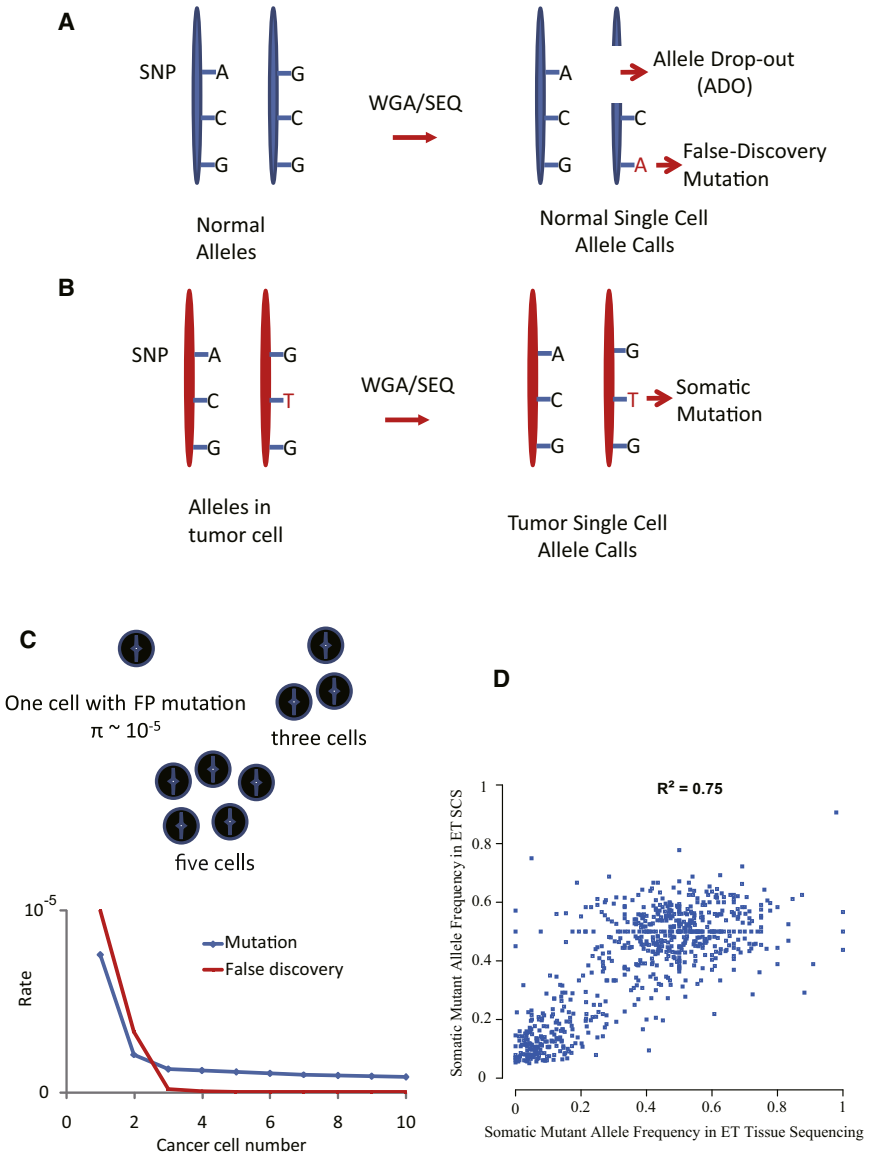

Figure 3. Schematic of the Somatic Mutation Calling Method in the Single-Cell Population and the Somatic Mutant Allele Frequency Correlation between Single-Cell Data and the Tissue of ET

(A) A normal pair of chromosomes is shown with an A/G SNP and two homozygous sites. Following WGA, PCR, and sequencing, a genotype is assembled from the reads. In the vicinity of the SNP, the region surrounding the A allele failed to amplify, and the result is an allele dropout (ADO) identifiable by a homozygous call in a single normal cell and a heterozygous call in the normal bulk tissue. The "A" in red represents a false discovery mutation in the normal single cell. The false discovery ratio can be estimated from these sites in normal single cells as compared to the normal bulk DNA.

(B) The same region of the tumor pair of chromosomes is shown with a putative somatic mutation detected in the genome of bulk tumor DNA as well as the genome of single tumor cell.

(C) Shown are ten hypothetical single tumor cells with putative somatic mutations. A mutation found in only a single cell is more likely to be a false discovery and is removed from consideration. Assuming a false discovery rate (π) = 10^{-5}, the probability of finding the same mutation in a third or fourth single cell is extremely low and also significantly lower than mutation, as depicted in the graph.

(D) Correlation between single-cell sequencing and tissue sequencing. The somatic mutant allele frequency in ET single-cell sequencing is indicated as an unfolded (that is, each genotype is taken as two alleles for a diploid-genome cell) site frequency of mutant alleles. The somatic mutant allele frequency in ET tissue sequencing is indicated as a read frequency of mutant alleles. R^2 here refers to the square of the correlation coefficient.

See also Figure S3 and Tables S1, S2, and S3.

background control. We obtained an average ADO of 43.09%. The false discovery ratio was 6.04×10^{-5}, which was determined using eight oral mucosal cells compared to 4,923,547 homozygous sites in oral mucosal tissue sequencing. The higher ADO and false discovery ratio observed here, as compared to those we obtained from the whole-genome sequencing of YH-1 and YH-2, may be due to differences in cell status and cell pattern and to loss of heterozygosity during the exome capture step.

To detect high-confidence point somatic mutations and circumvent the influence of amplification artifacts among the single-cell population, we defined a true mutation only if it occurred in a specified number of cancer cells (Figures 3A–3C). We carried out an extremely rigorous binomial test based on false discovery ratio, qualified ET cell number, and whole-exome length size to eliminate false somatic mutations that might have been caused by random single-cell sequencing errors. We also required that the somatic mutant loci be present in at least five ET cells and be homozygous normal in the oral mucosal epithelium tissue sequence data. We identified a total of 712 point somatic mutations in the exomes and flanking regions (within 100 bp) (Tables S3 and S4). We further assessed the coverage of each identified mutation, and 100% of the mutant alleles were covered by 10× or more reads, which is sufficient for our mutation calling by binomial test and also more rigorous than previous exome sequencing (≥8×) (Li et al., 2009a). Analysis of the base mutation type revealed that the majority of mutations were base substitutions between C:G and T:A (Figure S4A) in both the single-cell sequencing and tissue sequences, which is concordant with that seen in other cancers (Greenman et al., 2007).

We validated the somatic mutations from the single-cell sequencing in two ways. First, we compared all of the single-cell sequencing somatic mutations to those identified in the ET tissue sequencing. The correlation coefficient (R^2) of the somatic

mutation allele frequency between single-cell sequencing and tissue sequencing was 0.75, which indicated a highly consistent frequency (Figure 3D). Second, we validated the identified somatic mutations by randomly selecting 30 somatic mutations and assessed their presence in 52 randomly selected cells using PCR-Sanger sequencing; 90% (27 out of 30) were present in the PCR data.

The Use of 58 ET Cells Is Suitable for Identifying the Main Clonal Makeup of the ET Sample

Before carrying out additional genetic analyses on the identified mutations, we first determined whether 58 cells were sufficient to provide sufficient information on the genetic changes that occurred during ET development. The most important prerequisite of our evaluation is that we had randomly selected single cells from the bone marrow sample of the ET patient. To assess this, we examined the number of somatic mutations observed using an increasing number of cells. As expected, the number of somatic mutations increases with the use of data from greater cell numbers, but the number of somatic mutations reaches a plateau of cumulative somatic mutations at 25 single cells (Figure S4B). Statistic analysis showed that sequencing more cells would almost not increase the number of somatic mutations called from the cell population (cell number from 25 to 58, kappa square test, $p = 0.57$). This indicated that we have covered the majority of the somatic mutations within the ET sample; thus, this sample size is sufficient for identifying the main clonal architecture of this neoplasm.

Although it is possible that we might observe clones that were more highly selected, our single-cell exome sequencing data is comparable with the tissue exome data with regard to the capability of detecting different frequencies of mutations from different clones. Additionally, given that the sensitivity of exome capture in tissue sequencing is about 95%, the tissue sequence data might contain 5% of the heterozygous sites that are lost (Figure S4C). However, for single-cell sequencing, the likelihood of the same sites being lost in different cells by chance is extremely low. These results were also consistent with the observations of sequencing errors in multicell sequencing and of the ADO and amplification artifacts in single-cell sequencing (Figures S4D–S4F).

Population Analysis Indicated that None of the Sequenced ET Cells Were Normal Cells

Given the potential that normal cells could have been collected by chance during our isolation of ET cells, we determined whether there were some normal bone marrow cells among the 58 cells from the neoplasm. We performed a principle component analysis (PCA) (Jolliffe, 2002) based on the somatic mutations (Figure 4A) and found that the first component (eigenvector 1) showed that the ET cells were clearly different from the oral epithelium (tissue sequencing), indicating that no obviously normal cells were mixed with the ET cells. We also observed that, on each eigenvector, the ET cells were also separated from each other, indicating that the ET cells have substantial genetic diversity. However, we did not detect any significant subclusters among the ET cells on each eigenvector, indicating that the ET cells may be of monoclonal origin (discussed below).

The *JAK2*-Negative ET Patient Harbors a Distinct Set of Mutations

To confirm that this ET patient did not carry mutations in genes previously reported in ET studies and that this form of ET was truly *JAK2* negative, we first tested whether any of our identified somatic mutations were present in reported ET-mutated genes: *TET2*, *MPL*, *ASXL1*, *CBL*, *IDH*, and *IKZF1*. All of these genes were appropriately covered by our sequencing reads, but none had mutations (mutant reads in more than two of the single cells). Further screening of the reads mapping across *JAK2* exon 12 and exon 14, *MPL* exon 10 in tissue sequencing data of ET, and in the control data indicated that the previously reported hot spots were not altered in our ET patient.

Copy number alterations are rare in essential thrombocythemia (Kawamata et al., 2008; Stegelmann et al., 2010). We performed loss-of-heterozygous (LOH) analysis across the whole exome by comparing the heterozygous rate between the paired tissue sequencing data. There were no obvious LOH changes around these hot spots across the genome (Fisher's exact test, $p > 0.05$; Figure S4G). These data all confirmed that this patient was truly *JAK2* negative and further demonstrated that an elucidation of the genetic underpinnings of this type of ET required further genetic analyses to identify mutations that are important for cancer development in this patient.

The Population Genetic Pattern Indicates Monoclonal Origin of ET Cells

Our PCA analysis provided the first indication that this ET likely was of monoclonal origin. To make full use of our single-cell sequencing data to reconstruct the intratumoral genomic architecture and infer the possible developmental pattern, we examined the somatic mutant allele frequency spectrum (SMAFS) of our ET cells. It is important to note that, if the sequence coverage was insufficient to determine the genotype with certainty for each individual cell, the population genetic inferences based on called (inferred) somatic mutations would potentially lead to serious biases and possibly false inferences. We therefore used a Bayesian estimation-based method to call the site frequency spectrum to calculate the somatic allele frequency at each site for all ET cells (Yi et al., 2010). Given the randomness of ADO for each base type, the SMAFS is likely to be an unbiased site frequency spectrum. No significant selection signal was observed (Figure 4B), as there were no more nonsynonymous SNPs than synonymous SNPs at the lowest or highest somatic mutant allele frequencies. The SMAFS also showed an apparent increase in the number of sites at the ~50% somatic mutant allele frequency, both for synonymous and nonsynonymous somatic mutations. This indicates that all of the single cells contain specific heterozygous somatic mutant alleles or that half of the single cells contain homozygous somatic mutant alleles (but this is extremely unlikely). Thus, this patient's ET most likely originated from one clone or expanded from a single clone that had gained sufficient growth advantage after neoplasm expansion during progression to become the primary initiating clone. However, this does not rule out the possibility that cells of other clonal origin are present in an extremely low number.

To provide further genetic evidence of the proposed monoclonal origin, we calculated the likelihood of these ET cells

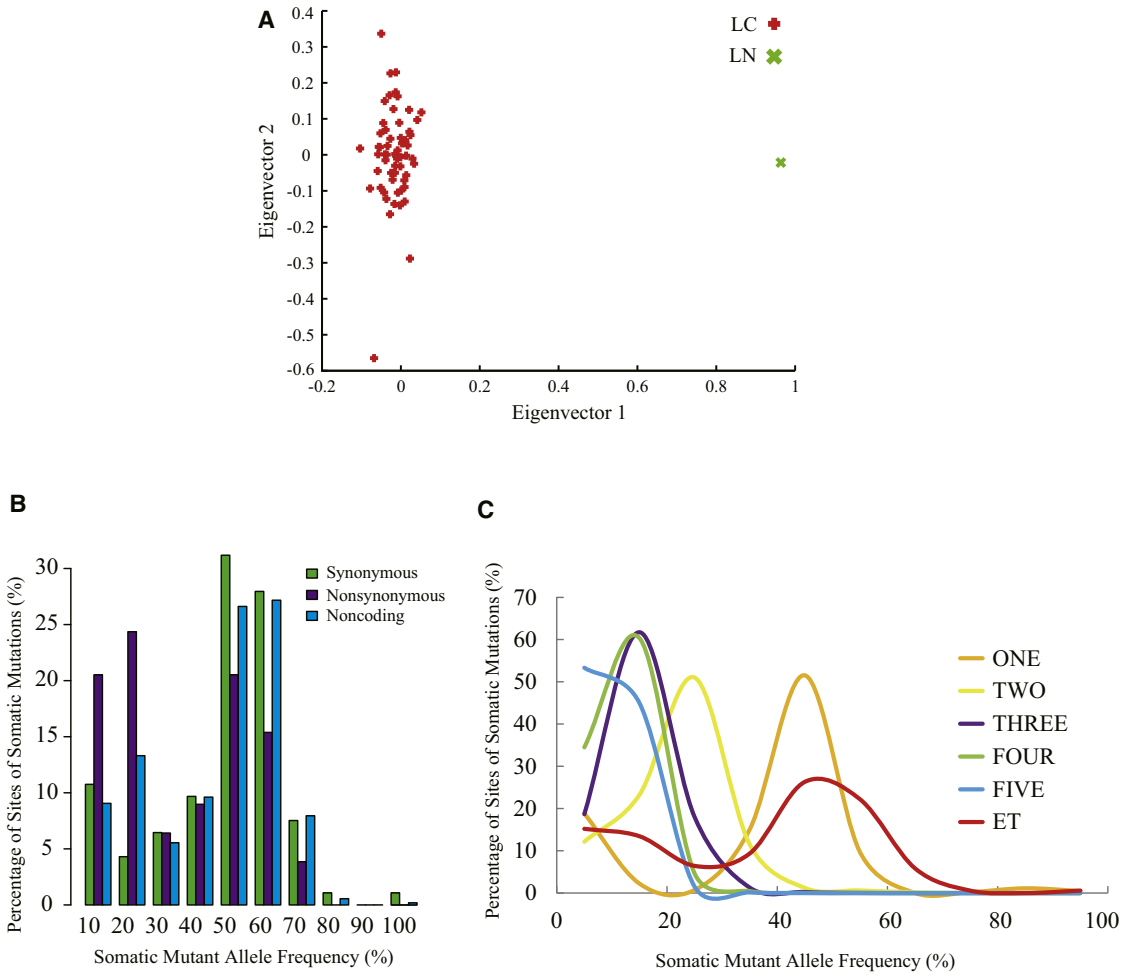

Figure 4. Genetic Characteristics of the ET Patient

(A) Principle component analysis (PCA) of the mutations in the ET cells (red) and the matched normal tissue (green). (B) Somatic mutant allele frequency spectrum (SMAFS) of ET. The SMAFS of ET was calculated for each mutation site of the ET by taking the oral mucosal epithelium mix as a control. The mutations are indicated as synonymous (green), nonsynonymous (purple), and noncoding (blue) mutations.

(C) Monoclonal and polyclonal evolution simulation of ET. The SMAFS of simulated one (orange), two (yellow), three (purple), four (green), and five (blue) clonal evolutions was calculated and compared with the SMAFS of ET (red) in the coding regions.

See also Figure S4 and Table S4.

having monoclonal evolution by carrying out an in silico clonal modeling simulation. For the simulation, we used a modified mathematical model with fit parameters from several previous reports (Abramson and Melton, 2000; Haeno et al., 2009; Lynch and Conery, 2003; Yachida et al., 2010) (see Extended Experimental Procedures for details). To simplify the evolution model, we only calculated the SMAFS of clones originating and proliferating from 1, 2, 3, 4, and 5 clones by the average of 100 times simulation. We started our simulation with the mutations obtained during each cell cycle across the human whole-exome region, which was comparable with our observations for single-ET cell sequencing. When comparing these simulated models with that observed in our ET patient, we found that the single-clone simulation (monoclonal evolution) was most similar to the SMAFS distribution that we obtained (Figure 4C). Taken together, these data indicated that a monoclonal evolution pattern of this neoplasm was most likely. This pattern might also be observed under conditions in which there was a brief polyclonal origin of the neoplasm but in which a specific clone subsequently had a very strong growth advantage over all other mutated clones; this would also produce a monoclonal evolution pattern and would provide the same genetic characteristics of monoclonal evolution.

Mutated Genes and Their Potential Roles in ET Monoclonal Progression

To identify key genes underlying ET monoclonal initiation and progression, we carried out a series of analyses as outlined in Figure S3C. Here, we assessed the 171 somatic mutations out of the total 712 somatic mutations (Table S4A) that were present in the coding regions that have a higher likelihood of having a functional impact. Of these coding somatic mutations,

Table 3. Key Genes and Their Known Biological Functions in ET Neoplasm Progression

Gene Name	Mutation Type	Amino Acid Changes	Functional Data
SESN2	Missense	P87S	*SESN2* encoded a member of the sestrin family of SESN1-related proteins and was an antioxidant activated by p53. Mutation in *SESN2* may lead to DNA damage and genetic instability.
ST13	Nonsense	Q349*	*ST13* encodes an Hsc70-interacting protein that controls the activity of regulatory proteins such as steroid receptors and regulators of proliferation or apoptosis. A mutation in *ST13* may contribute to loss of apoptotic control and may lead to abnormal proliferation.
NTRK1	Missense	N323S	A known oncogene; a mutation in *NTRK1* may contribute to sustained angiogenesis and cell proliferation.
ABCB5	Missense	G365V	Upregulation of *ABCB5* has been shown to be responsible for multidrug resistance in several cancers.
FRG1	Missense	C205Y	May be involved in pre-mRNA splicing.
ASNS	Missense	D118V	Is an asparagine synthetase.
TOP1MT	Missense	S479L	Acts as a DNA topoisomerase important during mitochondrial DNA synthesis.
DNAJC17	Missense	A292P	Is a DnaJ homolog subfamily C member 17

78 were nonsynonymous and present in a total of 71 genes (Table S4B).

We determined which of these 71 genes were likely to contain protein-damaging mutations or had been identified previously as associated with cancer. Using SIFT (Kumar et al., 2009), we found that 15 genes were likely to contain protein-damaging mutations (including protein truncation). We also found that three additional genes of the 71 (*MLL3*, *NTRK1*, and *PDE4DIP*) were present in the COSMIC Cancer Gene Census database (Futreal et al., 2004).

We further assessed the relative likelihood of these 18 genes being important to ET development using a modified Poisson model, whereby we proposed that driver genes were likely to contain significantly more nonsynonymous mutations than background mutations (Youn and Simon, 2011). Such driver gene prediction methods have previously been used in analyses of a variety of tumor types (Carter et al., 2009; Youn and Simon, 2011) and have been based on the mutation frequency in different samples from different patients. Of our 18 candidate genes, we identified 8 genes that had a significantly higher prevalence of protein-function-alternative somatic mutations, with a Q score $\geq$ 1.0 ($\leq$10% false discovery rate) (Table 3). Thus, these genes would have the highest likelihood of being involved in ET initiation and/or progression. Of interest, among these eight genes, four (*SESN2*, *ST13*, *DNAJC17*, and *TOP1MT*) had Q scores that were significantly higher than all other candidate genes (Figure 5). We propose that these genes may be of interest for future biological investigation in relation to ET.

DISCUSSION

Here, we developed a robust MDA-based single-cell sequencing method by carrying out whole-genome single-cell sequencing of two lymphoblastoid cell line cells. The data analysis indicated that it provides a higher sensitivity (more than 95% of bases with $\sim$18 × mean fold coverage of a whole genome of a single cell) than does DOP-PCR-based single-cell sequencing and a comparable specificity (false discovery rate: $2.52 \times 10^{-5} \sim 6.04 \times 10^{-5}$) with other amplification-based methods.

We next applied this single-cell sequencing method to a typical myeloproliferative neoplasm, ET. Based on the population genetic analyses of 58 qualified single cell exomes, we used somatic allele frequency data to demonstrate that this ET was likely of monoclonal origin and identified several genes that may play roles in ET neoplasm initiation and progression. These analyses demonstrate the ability of the method to begin characterizing the genetic architecture of the neoplasm, its clonal evolution, and candidate driver genes.

In particular, as *SESN2* is known to be involved in DNA damage and genetic instability (Sablina et al., 2005), change in *SESN2* function could lead to more rapid accumulation of additional somatic mutations. In addition, our identification of a mutation in the *NTRK1* gene is especially intriguing, given that it is a tyrosine kinase receptor that functions in a similar biological pathway as *JAK2*, the gene that has been found to be the most commonly mutated gene in ET. Because this patient was *JAK2* mutation negative, it raises the possibility that interference of this type of biological process through a variety of genetic mutations may be important for ET progression. Other genes like the tumor suppressor *ST13* (Yi et al., 2010) and the oncogene *ABCB5* (Frank and Frank, 2009) may also be involved in ET progression in concert with *ASNS*, *DNAJC17*, *TOP1MT*, and *FRG1*, which may be of interest for future biological investigation as well.

Overall, in addition to shedding light on ET development, our study demonstrates that single-cell sequencing enables the detection of numerous point mutations and that whole-exome sequencing of single tumor cells provides an excellent tool for assessing the complexity of small genetic changes in a variety of tumors tumor at a greater resolution (Xu et al., 2012, this issue of *Cell*). Accompanying single-cell sequencing with SMAFS and matched clone-evolution simulation analyses also enables researchers to define whether a tumor arises from a single or polyclone evolution process. Our method further opens the

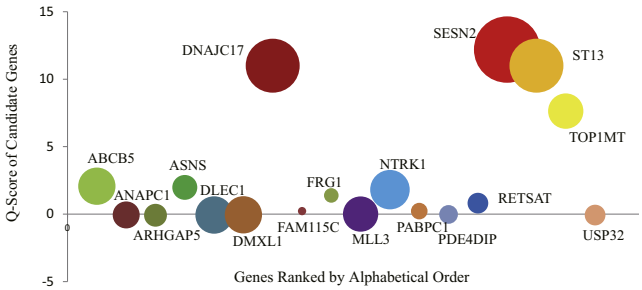

Figure 5. Key Gene Identification of the ET Patient

The driver gene prediction analysis of the 18 ET candidate genes is indicated as Q score. The vertical axis is the Q score, and the circle size (diameter) indicates the cell mutation frequency.

way for carrying out studies on a variety of complex diseases and biological processes that would benefit from the advantages of single-nucleotide resolution from individuals cells. This would be especially useful for diseases that develop additional and multiple mutations over time, which alter disease progression and would therefore impact treatment strategies.

EXPERIMENTAL PROCEDURES

Sample Collection

YH lymphoblastoid cell line, which was constructed from the healthy individual who provided DNA for the first Asian genome sequence (http://yh.genomics.org.cn/), was obtained from BGI. For ET patient, fresh bone marrow cells and normal oral mucous epithelium were obtained from a 58-year-old Chinese *JAK2* V617F-wild-type ET male patient. The platelet count was 600 × 10^9/l, whereas the white cell count and the hemoglobin level were normal. Informed written consent was obtained from the study participant. The studies were conducted in accordance with the Declaration of Helsinki II and were approved by the local Ethical Committees.

Collection, Lysis, and WGA of Single Cells

Every step during the experiment was reduced to strict minimum. With sufficient dispersion and cascade dilution of cells, single cells were randomly isolated from bloods or digested (collagenase I and IV) tissues into PCR-ready tubes using an inverted microscope and a mouth-controlled, fine hand-drawn microcapillary pipetting system made in house. The single-cell isolation was visually confirmed by microscopy and documented as micrographs. The cells were washed three times using the elution buffer. WGA of single cells was performed with REPLI-g Mini Kit (QIAGEN, Inc.) according to the instructions of the manufacturer, along with a no cell reaction as a negative control and a reaction of human tissue genomic DNA as positive control.

Quantitation and Genome Integrity Assessment of the WGA Products

The DNA concentration of the WGA products was measured with Quant-iT assays (Invitrogen Life Science, Inc.) according to the manufacturer's instruction. Then, genomic integrity of the qualified products (>60 ng/ml) was assayed by PCR amplification of ten housekeeping genes representing ten genes interspersed across ten different chromosomes. WGA products of best performance in relation to housekeeping PCR (≥8/10) and Qubit (>60 ng/μl) were selected for downstream experiments. Every step was performed along with a sample of human tissue genomic DNA as positive and a no template reaction as negative control, respectively.

Sequencing of Two Lymphoblastoid Cell Line Cells

For single cells from YH lymphoblastoid cell line, 2 μg genomic DNA from the selected WGA products was used to construct ~350 bp sequencing libraries and then subjected to massively parallel sequencing using the HiSeq2000 (Illumina, Inc.), with the paired-end 100 bp read option, according to the manufacturer's instructions.

Whole-Exome Capture and Sequencing of Samples from the ET Patient

For matched tissues from the ET patient, high molecular weight genomic DNA was extracted from freshly frozen bone marrow cells and normal oral mucous epithelium. For each DNA sample (both single cells and matched tissues) from the ET patient, whole-exome capture was accomplished based on liquid phase hybridization of 2 μg sonicated genomic DNA to the bait cRNA library that was synthesized on magnetic beads using SureSelect Human All Exon kit (Agilent Technology, Inc.), according to the manufacturer's protocol. The captured targets were subjected to massively parallel sequencing using HiSeq2000 with the paired-end 100 bp read option, according to the manufacturer's instructions.

Public Data Set Access

Human (*Homo sapiens*) reference genome sequence (Hg18) and its annotation files (dbSNP v128) were downloaded from University of California Santa Cruz Genome Bioinformatics (http://genome.ucsc.edu/). The YH reference genome and the YHSNPs files were downloaded from the First Asian Diploid Genome database (http://yh.genomics.org.cn/). The target region files of exome capture were downloaded from the Agilent website (http://www.genomics.agilent.com).

Reference Guide Genome Assembly and SNP Calling

SOAPaligner/SOAP2 version 2.20 was used to align all sequencing reads to the Hg18 reference genome with a maximum of two mismatches and nongap parameters. For YH single cells and matched tissue, all reads mapping uniquely to the whole genome were selected for SNP calling; for the ET patient, all reads mapping uniquely to exome regions and 100 bp flanking regions were selected for SNP calling. SOAPsnp version 1.03 was used for calculating the likelihoods of genotypes for each cell and matched tissue (see Extended Experimental Procedures).

Data Evaluation by ADO and False Discovery Ratio

With the final SNPs, the ADO was defined as the random nonamplification of one of the alleles present in a heterozygous sample. The false discovery ratio was defined as a false heterozygous site in a homozygous sample. The evaluation of ADO and false discovery ratio were performed comparing the single-cell data and the tissue sequencing control data (see Extended Experimental Procedures).

Randomness Evaluation of ADO and False Discovery Ratio

We evaluated randomness of ADO by calculating the counts of ADO of the four base types in YH-1 and YH-2 among the previously noted high-confidence heterozygous SNPs. We evaluated the randomness of false discovery ratio by separating the artifacts into six different mutation types. The numbers of different base types of ADO were normalized. We used Fisher' exact test to calculate the p value of the variation between base types.

High-Confidence Somatic Mutation Identification and Experiment Verification

To eliminate random errors induced by single-cell sequencing, we created a binomial test to detect high-confidence point somatic mutations. The putative somatic mutations were filtered by the following criteria: (1) the oral mucosal epithelium tissue sequence was normal homozygous for the site, and (2) the mutation was present in at least five ET cells among the total 58 qualified ET cells, with criterion (2) set using a binomial test with false discovery ratio, qualified ET cell number, and whole-exome length size to eliminate random errors.

$$P(i) = C_n^i p^i (1-p)^{(n-i)}$$

$$P(i) \cdot S < 1$$

wherein i is the cell number of mutants of a specific mutation, n is the qualified ET cell number, p is the false discovery ratio, $P(i)$ is the probability under

binomial distribution, and S is the whole-exome length size after which criterion (2) was set according to the minimum cell number that fulfilled the above equation. High-confidence somatic mutations in ET cells were then randomly selected for PCR-Sanger experiment verification.

Principal Component Analysis

To identify the most variable factors in classifying subgroups among single cancer cells, we utilized an R package, pcaMethods v1.12.0 (http://rss.acs.unt.edu/Rdoc/library/pcaMethods/), in performing the PCA based on the genotyping result at all somatic mutation sites on each single cell. Missing values were automatically estimated by probabilistic method within the R package.

Correlation of Somatic Mutant Allele Frequency between Single-Cell Sequencing and Tissue Sequencing

Correlation coefficient of determination is a goodness-of-fit measure for models based on the proportion of explained variance. The somatic mutant allele frequency in ET single-cell sequencing was indicated as the unfolded (that is, each genotype is taken as two alleles for a diploid-genome cell) site frequency of mutant alleles, and the somatic mutant allele frequency in ET tissue sequencing was indicated as read frequency of mutant alleles.

Somatic Mutant Allele Frequency Spectrum Analysis

To take uncertainty in genotype calling and allele frequency estimation into account, instead of utilizing a Bayesian estimation-based method called site frequency spectrum (SFS) on population individuals (Yi et al., 2010), we used this SFS to calculate the somatic mutant allele frequency in each SM site of all single cells, based on the same methods of estimating allele frequencies from reads in one site and additional estimating of sample allele frequencies.

In Silico Clonal Modeling Simulation

In our simulation, we used a modified mathematical model with fit parameters from several previous reports (Abramson and Melton, 2000; Haeno et al., 2009; Lynch and Conery, 2003; Yachida et al., 2010) (see Extended Experimental Procedures). Using fit parameters, we calculated the SMAFS of monoclonal evolution and polyclonal evolution (originating and proliferating from 2, 3, 4, and 5 clones) by the average of 100 times simulation, respectively.

Driver Gene Predication

We used a modified Poisson model that hypothesized that driver genes lean to contain significantly more nonsynonymous mutations than the background mutations (Youn and Simon, 2011) (see Extended Experimental Procedures).

ACCESSION NUMBERS

All sequencing data from this study are deposited in NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRA050202.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and four tables and can be found with this article online at doi:10.1016/j.cell.2012.02.028.

ACKNOWLEDGMENTS

This work was supported by a National Basic Research Program of China (973 program number 2011CB809202;2011CB809203), the Chinese 863 program (numbers 2009AA022707 and 2012AA02A201), the Shenzhen Municipal Government of China (grant ZYC201005250020A), the Key Laboratory Project Supported by Shenzhen City (grants CX B200903110066A and CXB201108250096A), and Shenzhen Key Laboratory of Gene Bank for National Life Science. This project was also supported by grants from the Innovative Research Team Project of Guangdong and the Guangdong Enterprise Key Laboratory of Human Disease Genomics. We also acknowledge the Ole Rømer grant from the Danish Natural Science Research Council, the Danish National Research Foundation, the National Natural Science Foundation of China, and funds from the Shenzhen Municipal Government and the Local Government of Yantian District of Shenzhen.

Received: October 14, 2011
Revised: December 16, 2011
Accepted: February 15, 2012
Published: March 1, 2012

REFERENCES

Abramson, N., and Melton, B. (2000). Leukocytosis: basics of clinical assessment. Am. Fam. Physician *62*, 2053–2060.

Behar, D.M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., et al. (2010). The genome-wide structure of the Jewish people. Nature *466*, 238–242.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Cairns, J. (1975). Mutation selection and the natural history of cancer. Nature *255*, 197–200.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. *69*, 6660–6667.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002a). Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA *99*, 5261–5266.

Dean, F.B., Hosono, S., Fang, L.H., Wu, X.H., Faruqi, A.F., Bray-Ward, P., Sun, Z.Y., Zong, Q.L., Du, Y.F., Du, J., et al. (2002b). Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA *99*, 5261–5266.

Frank, N.Y., and Frank, M.H. (2009). ABCB5 gene amplification in human leukemia cells. Leuk. Res. *33*, 1303–1305.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Haeno, H., Levine, R.L., Gilliland, D.G., and Michor, F. (2009). A progenitor cell origin of myeloid malignancies. Proc. Natl. Acad. Sci. USA *106*, 16616–16621.

Jolliffe, I.T. (2002). Principal component analysis, Second Edition (New York: Springer).

Kawamata, N., Ogawa, S., Yamamoto, G., Lehmann, S., Levine, R.L., Pikman, Y., Nannya, Y., Sanada, M., Miller, C.W., Gilliland, D.G., and Koeffler, H.P. (2008). Genetic profiling of myeloproliferative disorders by single-nucleotide polymorphism oligonucleotide microarray. Exp. Hematol. *36*, 1471–1479.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009a). SNP detection for massively parallel whole-genome resequencing. Genome Res. *19*, 1124–1132.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.

Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T., et al. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat. Genet. *42*, 969–972.

Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404.

Mesa, R.A., Silverstein, M.N., Jacobsen, S.J., Wollan, P.C., and Tefferi, A. (1999). Population-based incidence and survival figures in essential thrombocythemia and agnogenic myeloid metaplasia: an Olmsted County Study, 1976–1995. Am. J. Hematol. *61*, 10–15.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

Sablina, A.A., Budanov, A.V., Ilyinskaya, G.V., Agapova, L.S., Kravchenko, J.E., and Chumakov, P.M. (2005). The antioxidant function of the p53 tumor suppressor. Nat. Med. *11*, 1306–1313.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. (2006). Whole-genome multiple displacement amplification from single cells. Nat. Protoc. *1*, 1965–1970.

Stegelmann, F., Bullinger, L., Griesshammer, M., Holzmann, K., Habdank, M., Kuhn, S., Maile, C., Schauer, S., Döhner, H., and Döhner, K. (2010). High-resolution single-nucleotide polymorphism array-profiling in myeloproliferative neoplasms identifies novel genomic aberrations. Haematologica *95*, 666–669.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.

Tefferi, A. (2001). Recent progress in the pathogenesis and management of essential thrombocythemia. Leuk. Res. *25*, 369–377.

Tefferi, A. (2010). Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms: JAK2, MPL, TET2, ASXL1, CBL, IDH and IKZF1. Leukemia *24*, 1128–1138.

Visvader, J.E. (2011). Cells of origin in cancer. Nature *469*, 314–322.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., et al. (2008). The diploid genome sequence of an Asian individual. Nature *456*, 60–65.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Li, F., Tsang, S., Wu, K., Wu, H., He, W., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, this issue, 886–895.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.

Youn, A., and Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics *27*, 175–181.