Supplementary Information
for
The landscape of somatic mutations in protein coding genes in apparently healthy human tissues

Vinod K Yadav, James DeGregori, Subhajyoti De

Supplementary Text:
• Analysis of somatic variant calls
• Clonal composition of peripheral blood.
• Modeling of hematopoiesis
• Modeling of developmental hierarchies in tissues

Supplementary Figures:
• Supplementary Figure 1: Assessment of quality of the somatic variant calls
• Supplementary Figure 2: Detecting somatic mutations in RNAseq data
• Supplementary Figure 3: Chromosomal locations of somatic, exonic SNVs and detection of outliers
• Supplementary Figure 4: Variation in ADMB in blood between different cohorts.
• Supplementary Figure 5: Comparing tissue level burden of mutations with that in individual clones.
• Supplementary Figure 6: Proportion of C:G>A:T substitutions in lowly and highly expressed genes.
• Supplementary Figure 7: Network representation of genes and enriched pathways.
• Supplementary Figure 8: Proportion of deleterious and benign missense mutations.
• Supplementary Figure 9: Correlation of allelic frequency of potentially deleterious mutations in exome-seq and RNA-seq data from tissue samples
• Supplementary Figure 10: Comparison of missense, nonsense and silent mutation allelic frequency.
• Supplementary Figure 11: Effects of replication rate on survivability of HSC clones with time.
• Supplementary Figure 12: Effects of differentiation rate on survivability of HSC clones with time.
• Supplementary Figure 13: Survivability of HSC clones with time in mice.
• Supplementary Figure 14: Effects of different combinations of replication rates, death rates, and differentiation rates on survivability of clones in a tissue with time.

Supplementary Tables:
• Supplementary Table 1: Summary of the donors.
• Supplementary Table 2: The TCR scores for different tissue and cell types.
Supplementary Text

Analysis of somatic variant calls: We took multiple measures to estimate the confidence in the reported somatic variant calls, after accounting for possible sources of errors due to misclassified germ line heterozygous SNPs, contamination due to tumor-derived DNA, and sequencing error.

Assessment of misclassified germ line heterozygous SNPs: A heterozygous germ line SNP is present in the DNA derived from all somatic tissue types, and therefore should be detectable in exome-seq data from the paired solid tissue (NT) and blood (NB) at sufficient depth of coverage. However, if the variant is detected in exome-seq data from only one tissue (by chance, or due to systematic allele-specific sequencing biases), it could be misclassified as a somatic variant. To assess these possibilities, we performed a systematic analysis. We randomly selected 32 samples (16 pairs) across different tissue type (4 from HNSC, 3 from THCA, 2 LIHC, 3 from LUAD, 4 from STAD and 16 corresponding blood samples). In total, 24,662 heterozygous germ line SNPs were detected in the Affymetrix Genome-Wide Human SNP Array 6.0, of which 23,140 variant sites were covered at 25X or higher depth in exome-seq data from paired tissues for these samples. The frequency distribution of the alternate allele in the exome-seq data closely followed a beta-binomial distribution with a mean of 0.459, and in >99.6% of the cases the alternate allele frequency was within the range of [0.2, 0.8]. None of the variants was present in exome-seq data from one tissue but undetectable in the other at 25X coverage. In 23,117 cases the germ line variant was detectable in both tissues, and in remaining 23 cases it was absent in both tissues (indicating potential erroneous call in the SNP array analysis). Additionally, mutation detection across tissue type was uniformed and we also not found in batch effect on mutation detection or mutation allele frequency in sequencing data. This analysis indicates that it is highly unlikely that misclassified germ line heterozygous SNPs would contaminate our catalog of somatic mutations.

We also fitted a beta-binomial distribution to the alternate allele frequency distribution by maximum likelihood method, and estimated the model parameters \(a, b\). We then calculated the probability of observing no alternate allele (the number of success, \(k=0\) at different depth of coverage (number of trials \(n=\{25, 50, 75, 100\}\)) by chance alone. In all the cases, we found that the corresponding probability was small (\(p\)-value <1.0E-05), restating that it is unlikely that misclassified germ line heterozygous SNPs would bias our catalog of mutations.

Assessment of allelic coverage biases: We also probed for allele-specific coverage biases due to sequence protocol at the site of heterozygous SNPs. We compared allele frequency distribution of different combinations of base substitutions (A-C, A-G, A-T, C-G, C-T, G-T) for the germ line SNPs that were detectable in both tissues. We anticipated that allele-specific coverage biases due to sequence protocol would lead to systematic skewing of allelic frequency. We found no systematic change in mean allele frequency of any nucleotide over other nucleotide in sequencing data. Additionally, substitution pattern identified in sequencing was consistent with SNP array. Therefore, there was no systematic allele-specific coverage bias in our dataset.

Assessment of tumor DNA contamination: Contamination of tumor variants (e.g. due to circulating cell-free tumor DNA in blood, or contamination during sample preparation) can potentially contribute to false positive somatic variant calls in non-malignant tissues. To test the possibilities, we compared with the catalog of somatic mutations in the normal blood (identified by comparing with the normal solid tissue), with the somatic variant calls in the matched tumor samples (identified by comparing with the same normal solid tissue as well). If the somatic variant calls in the normal blood were due to tumor DNA contamination, we anticipate detecting the same variant in the matched tumor samples as well. Our analysis indicated that none of the somatic mutations in normal blood in our dataset was due to tumor DNA contamination.

An equivalent analysis for solid tissue is conceptually more challenging. A mutation absent in blood, but present in solid tissue and tumor could arise due to occurrence of the shared mutation early during development before the onset of malignancy. Nevertheless, only a minor subset of the somatic mutations in solid tissues was shared with the matched tumor samples, as expected.

Assessment of sequencing error: Sequencing error could also result in erroneous variant calls. Since we did not have access to the actual samples, Sanger sequencing to confirm the biological variants was not an option. At this end, we calculated the probability of sequencing error using base quality, coverage, and other information. We called the variants using VarScan2 (1) with base quality >20, base coverage >25X, variant allele supported by >10% reads and at least 3 reads. All the somatic mutations in our catalog were deemed ‘high confidence calls’ by VarScan2.

In addition, we calculated the probability of sequencing error theoretically from base quality scores etc using published approach (2). At this end, we can estimate the \(p\)-value for the null hypothesis that \(r\) reads supporting alternate base calls at a position with a coverage of \(n\) at that position, represent sequencing errors:
\[ P(\geq s \text{ erroneous calls} \mid Q_1, Q_2, \ldots, Q_n) \]
\[ \text{or} \ P(\geq s \text{ erroneous calls} \mid p_1, p_2, \ldots, p_n) \]
such that \( Q_1, Q_2, \ldots, Q_n \) represent the base quality values of the \( n \) base calls, and \( p_1, p_2, \ldots, p_n \) represent the corresponding base calling error probability, such that \( p_i = 10^{Q_i/10} \). If we assume that base calls from non-duplicate reads are independent, \( P \) can be computed using the distribution of the sum of \( n \) independent Bernoulli random variables with success probabilities \( \{p_i; i=(1, \ldots, n)\} \). As shown by Bansal (2), the probability that the sum \( X = X_1 + X_2 + \cdots + X_n \) of \( n \) independent Bernoulli random variables deviate from its mean \( \mu = \sum_i p_i \) can be approximated using the Chernoff bound (3):

\[ P(X > (1 + \delta)\mu) < \left[ \frac{e^\delta}{(1 + \delta)^{(1+\delta)}} \right]^\mu \]

where we can express \( r = (1 + \delta)\mu \). Based on that estimate, the somatic variants in our catalog were significantly unlikely (FDR adjusted \( p \)-value <0.05) to be due to sequencing error, which is consistent with the sequencing error estimates independently calculated by VarScan2. In fact, a vast majority of the somatic mutations were called with even higher confidence.

Strand bias can provide indirect assessment of the quality of variant calls. However, true variants can also sometimes show significant strand bias. For instance, for the known heterozygous SNPs, the strand bias in the exome-data does not mean sequencing error, and instead provide a null distribution of strand bias for true biological variants. We evaluated strand bias using FisherStrand, as implemented in the GATK (4) for the somatic mutations in our dataset and also the heterozygous SNPs detected using exome-seq and SNP arrays. We found that the extent of strand bias (distribution of FisherStrand \( p \)-values) was comparable between somatic mutations and heterozygous SNPs (Supplementary Figure 1).

**Clonality in blood:** In the process of hematopoiesis ~1011 blood cells are produced daily, which originate from a pool of approximately 11,000 HSCs in adults (5). While a single HSC clone is sufficient to maintain hematopoiesis over the lifetime of an individual (6), peripheral blood is typically oligoclonal, and bulk of it is derived from a small number of actively cycling HSCs. In adult human ~1000 HSCs are thought to contribute to hematopoiesis (7,8), albeit relative contribution of individual clones in peripheral blood cells is skewed (Figure 1). Blood sequencing of an old woman indicated that bulk of the cells in her peripheral blood was derived from only 2 genetically distinct clones (9). Barcode sequencing in mice (post bone marrow transplantation) showed that 5 – 20 clones represent 80% of the granulocytes (or T-cell) (10). They also found that the number of granulocytes derived from individual clones varied by orders of magnitude during 6-24 weeks post-transplantation. These indicate that at any point of time, majority of the blood cells trace their root back to a small number of actively cycling HSCs.

**Modeling of hematopoiesis:** In addition to the model of hematopoiesis developed by Abkowitz et al. (5) (as described in the main text) we used an alternate model developed by Busch et al. (11) derived from unperturbed hematopoiesis in mice. According to the model,

\[ \text{HSC} \rightarrow \text{ST.HSC} \rightarrow \text{MMP} \rightarrow (\text{CLP}, \text{CMP}) \rightarrow \text{differeniated, peripheral blood cells} \]

The number of HSCs in adult \( \approx 1.7 \times 10^4 \) which remains reasonably constant throughout adulthood, i.e.

\[ \frac{dN_{\text{HSC}}}{dt} = \beta_{\text{HSC}} N_{\text{HSC}} - \alpha_{\text{HSC}} N_{\text{HSC}} = 0 \]

where \( \alpha, \beta \) are differentiation and net proliferation rate of long-term HSCs, respectively. The best fitted values for the parameters are \( \alpha = 0.009 \text{ d}^{-1} \), C.I. [0.006, 0.012] and \( \beta = 0.009 \text{ d}^{-1} \), C.I. [0.006, 0.012]. Short term HSCs (ST.HSCs), multi-potent progenitors (MMPs), common lymphoid and myeloid progenitors (CLPs and CMPs, respectively) are short lived. In the steady state,

\[ \frac{dN_{\text{ST.HSC}}}{dt} = \frac{dN_{\text{MMP}}}{dt} = \frac{dN_{\text{CLP}}}{dt} = \frac{dN_{\text{CMP}}}{dt} = 0 \]

The full models are available elsewhere. In any case, we used the framework to calculate the probability that a somatic mutation can survive in the HSC population in mice for some time. We found that mutation acquired in a single HSC can survive for 1 month, 3 months, 6 months, 1 year, and 2 years, with probability 0.79, 0.55, 0.39, 0.24, and 0.13, respectively, and importantly 10% decrease in net proliferation rate of an HSC clone carrying a somatic mutation did not reduce the survivability of that clone in absolute time (Supplementary Figure 11). We suspect that this is perhaps due to slow cell division rate in HSC pool. We note that, owing to difference in life-span between mice and human, 1 year for mouse would be comparable to several decades in human, and in that regard, the long term probability of survival of the HSC clone in human (e.g. over decades)
and mice (i.e. over years) are similar. Furthermore, in both cases, the effect of change in proliferation rate (i.e. birth rate, death rate, or both) had only minor effects.

**Modeling of developmental hierarchies in tissues:**

Most of the tissues have a hierarchical developmental architecture comprising of multiple developmental compartments where each compartment signifies a cell-type at a specific state of differentiation. This architecture can be exemplified by blood, as well as skin, and colonic crypt; and perhaps many other tissues follow the same, general pattern (12). Based on a published model (12), the system comprising of k+1 compartments can be represented using a set of differential equations.

\[
\frac{dN_i}{dt} = - (2v_i + 2\lambda_i - 1) \lambda_i N_i + 2v_{i-1}\lambda_{i-1}N_{i-1}
\]

where, \(N_i\), \(\lambda_i\), \(a_i\), and \(v_i\) are the number of cells, replication rate, death rate, and differentiation rate for \(i\)-th compartment \((1 \leq i \leq k)\). The model assumes that at the stem cell level in which cell differentiation, cell death and self renewal are balanced such that the average number of cells remains constant.

For a well-studied system such as hematopoiesis, the number of compartments and the model parameters for the compartments were previously estimated, which reduce the model to that used in Figure 4. For most of the other tissue types, experimentally estimated data for the number of compartments and the model parameters for the compartments are not readily available. Therefore, we used a wide range of parameter-values and their combinations for replication \((\lambda)\), differentiation \((v)\), and death rates \((a)\), which are at par with reasonable approximations published. Assuming that a mutation occurred in the stem-cell compartment \((k=0)\), we used these combinations of parameter values over a broad range to estimate the theoretical probability of probability of survival of a clone as a function of time in a tissue (Supplementary Figure 14). In general, we found that high differentiation rate \((v)\) considerably reduces the probability. Even then a clone with deleterious mutation can survive in the tissue compartment with high probability (>20%) for several years (or more).

We note that cells in the Intermediate developmental states \((k > 0)\) can also acquire somatic mutations, but such clones would typically be short-lived. We also note that, the models provide the best approximations based on current understanding, and more accurate understanding of the biology of tissue developments and experimental estimation of appropriate parameters would be very helpful.

**References:**

1. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L. and Wilson, R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22, 568-576.

2. Bansal, V. (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, 26, i318-324.

3. Chernoff, H. (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.*, 23, 493-507.

4. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20, 1297-1303.

5. Abkowitz, J.L., Catlin, S.N., McCallie, M.T. and Guttorp, P. (2002) Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. *Blood*, 100, 2665-2667.

6. Jordan, C.T. and Lemischka, I.R. (1990) Clonal and systemic analysis of long-term hematopoiesis in the mouse. *Genes & development*, 4, 220-232.

7. Catlin, S.N., Guttorp, P. and Abkowitz, J.L. (2005) The kinetics of clonal dominance in myeloproliferative disorders. *Blood*, 106, 2688-2692.

8. Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J. *et al.* (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 150, 264-278.

9. Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B. *et al.* (2014) Somatic mutations found in the healthy blood compartment of a 115-year-old woman demonstrate oligoclonal hematopoiesis. *Genome research*, 24, 733-742.

10. Verovskaya, E., Broekhuis, M.J., Zwart, E., Riitsema, M., van Os, R., de Haan, G. and Bystricky, L.V. (2013) Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood*, 122, 523-532.

11. Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S.M., Reth, M., Hofer, T. and Rodewald, H.R. (2015) Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518, 542-546.

12. Werner, B., Dingli, D., Lenaerts, T., Pacheco, J.M. and Traulsen, A. (2011) Dynamics of mutant cells in hierarchical organized tissues. *PLoS computational biology*, 7, e1002290.
Supplementary Figure 1: A) In total, 24,662 heterozygous germ line SNPs were detected in the Affymetrix Genome-Wide Human SNP Array 6.0, of which 23,140 variant sites were covered at 25X or higher depth in exome-seq data from paired tissues for these samples. In >99.6% of the cases the alternate allele frequency was within the range of [0.2, 0.8]. None of the variants was present in exome-seq data from one tissue but undetectable in the other at 25X coverage. In 23,117 cases the germ line variant was detectable in both tissues, and in remaining 23 cases it was absent in both tissues (indicating potential erroneous call in the SNP array analysis). B) Heterozygous germ line mutation allele frequency ratio was presented as A/(A+B), where A and B were the two alleles (e.g. A/T). This ratio was calculated for different substitutions using exome-seq data at the sites of germ line heterozygous SNPs and no obvious difference was detected for any particular nucleotide over other. Additionally, substitution pattern identified in sequencing was consistent with SNP array. Taken together, this data indicates that there was no systematic allele-specific coverage bias in our dataset. C) At the sites of heterozygous SNPs, the substitution pattern identified in sequencing was consistent with SNP arrays. D) Distribution of ratio between forward and total aligned reads for somatic and germline mutations.
Supplementary Figure 2

Supplementary Figure 2: Proportion of somatic mutations in exome seq data from normal somatic tissues that were found in the RNA-seq data from the same patients at different RNA-seq coverage. We obtained matched RNAseq data (bam files) for 23 blood samples from the Cancer Genomics Hub (also see Supplementary Table 1). Next, for each position of predicted somatic mutation in the DNA samples, we asked whether the matched RNA sample has sufficient depth of coverage (>10X, >20X .. >100X), and if so, whether the alternate allele could be detected with confidence by variant callers in the RNA sample as well. For each depth of coverage cut-off, we calculated the number of sites of somatic mutations that had similar or higher coverage in the RNASEq data (N), and the number of sites at which the alternate allele was detected with confidence (r), and reported r/N as the proportion validated. Therefore, the reported rate of overlap is an under-estimate since we had less power to detect low allele frequency mutations in lowly expressed genes. Even then the proportions of somatic mutations that were also supported by transcriptomic data (which could be considered as additional evidence) was comparable or higher relative to that reported in recent studies.

Supplementary Figure 3

Supplementary Figure 3: Genome-wide location of the somatic, exonic mutations in the apparently normal tissue samples. The outliers are marked with red arrow.
Supplementary Figure 4: Variation in the AMDR in the normal blood samples from different TCGA cohorts. Washington University Genome Sequencing Center (WUGSC) sequenced BRCA Samples, Baylor College of Medicine (BCM) sequenced LIHC samples and broad institute (BI) sequenced BLCA, HNSC, LUAD, LUSC, PRAD, STAD and THCA samples.

Supplementary Figure 5: A) A schematic representation showing that mutations detected (at certain allele frequency cutoff) in blood, which is typically polyclonal, can come from one or more major clones. Individual hematopoietic stem/progenitor cells from adults typically harbor 2 – 10 exonic SNVs (Welch et al. Cell, 2012). It is consistent with our observation based on mixed populations of cells in peripheral blood (exonic SNV: median: 8, 25 percentile: 4; and 75 percentile: 16 at allele frequency threshold >0.1).
Supplementary Figure 6

Supplementary Figure 6: Left panels) Proportion of C:G>A:T substitutions in lowly and highly expressed genes, based on expression data in whole blood and CD34+ hematopoietic stem cells. Right panels) Transcriptional strand bias, as estimated by the proportion of G>T/C>A substitution on the transcribed strand for lowly and highly expressed genes. Highly expressed genes have proportionally more C>A substitution than G>T substitution on the transcribed strand than lowly expressed genes. Different expression cutoffs were used.

Supplementary Figure 7

Supplementary Figure 7: Pathways that are enriched for potentially deleterious mutations in genes that also have top 50 percentile expression in blood.
Supplementary Figure 8 (A) Frequency and (B) proportion of potentially damaging and benign missense mutations in blood. Expression was based on whole blood (top 25 percentile), and we got comparable results using alternative expression threshold, or expression data from CD34+ HSCs.

Supplementary Figure 9: Scatterplot showing allelic frequency distributions of potentially deleterious mutations in paired exome-seq and RNA-seq data in apparently normal blood samples.
Supplementary Figure 10: Distribution of allelic frequency of missense, nonsense and silent mutations.

Supplementary Figure 11: Graph showing the probability of survival of an HSC clone as a function of time. Black, green, and orange curves show the probability for wild type HSCs, and those whose replication rate ($\lambda$) is 20% higher, and 20% lower, respectively. $\alpha/\lambda$ and $\nu/\lambda$ remained unchanged.

Supplementary Figure 12: Graph showing the probability of survival of an HSC clone as a function of time. Black, green, and orange curves show the probability for wild type HSCs, and those whose differentiation rate ($\nu$) is 20% higher, and 20% lower, respectively. $\lambda$ and $\alpha$ remained unchanged.
Supplementary Figure 13: Graph showing the probability of survival of an HSC clone as a function of time in mice. Black and orange curves show the probability for wild type HSCs, and those whose proliferation rate is 10% lower, respectively. Note that, owing to difference in life-span between mice and human, 1 year for mouse would be comparable to several decades in human. Also see Supplementary Text.

Supplementary Figure 14: Graph showing the probability of survival of a clone as a function of time in a tissue. Different colored curves show the probability for different combinations of $\lambda$, $\alpha$, and $\nu$ values (unit: per year). In this model, tissues with proportionally high differentiation rate ($\nu/\lambda$) have relatively rapid decrease in probability of the survival of the mutant clone as a function of time. Within a tissue stem cell niche, the relative effect of a mutation on the $\lambda$, $\alpha$, and $\nu$ values are qualitatively similar to that reported in Figure 4 and Supplementary Figures 11-12.
### Supplementary Table 1: Summary of the donors.

| Sample ID     | Gender | Age (Years) | Sequencing Center | Sample ID     | Gender | Age (Years) | Sequencing Center |
|---------------|--------|-------------|-------------------|---------------|--------|-------------|-------------------|
| TCGA-DD-A113  | FEMALE | 55          | BCM               | TCGA-43-6771*| MALE   | 85          | BI                |
| TCGA-DD-D135  | MALE   | 53          | BCM               | TCGA-43-6773*| MALE   | 76          | BI                |
| TCGA-DD-D119  | MALE   | 40          | BCM               | TCGA-56-8625 | FEMALE | 66          | BI                |
| TCGA-DD-D11A  | MALE   | 67          | BCM               | TCGA-77-6843 | MALE   | 74          | BI                |
| TCGA-DD-D11E  | MALE   | 72          | BCM               | TCGA-77-6844 | MALE   | 74          | BI                |
| TCGA-BH-A0AY  | FEMALE | 62          | WUGSC             | TCGA-77-6845 | MALE   | 69          | BI                |
| TCGA-BH-A0B7  | FEMALE | 42          | WUGSC             | TCGA-43-5670 | MALE   | 70          | BI                |
| TCGA-BH-A0BC  | FEMALE | 60          | WUGSC             | TCGA-56-7579 | MALE   | 62          | BI                |
| TCGA-BH-A0DE  | FEMALE | 62          | WUGSC             | TCGA-56-7582 | MALE   | 84          | BI                |
| TCGA-BH-A0DL  | FEMALE | 64          | WUGSC             | TCGA-43-7658 | FEMALE | 75          | BI                |
| TCGA-E2-A14Y  | FEMALE | 35          | WUGSC             | TCGA-90-7767 | MALE   | 56          | BI                |
| TCGA-E2-A15A  | FEMALE | 43          | WUGSC             | TCGA-56-7822 | MALE   | 76          | BI                |
| TCGA-GC-A3BM* | MALE   | 70          | BI                | TCGA-G9-6342*| MALE   | 61          | BI                |
| TCGA-GC-A3WC* | FEMALE | 80          | BI                | TCGA-G9-6364 | MALE   | 72          | BI                |
| TCGA-GD-A2C5* | FEMALE | 53          | BI                | TCGA-G9-6373 | MALE   | 68          | BI                |
| TCGA-GD-A3OP* | FEMALE | 84          | BI                | TCGA-G9-6494 | MALE   | 66          | BI                |
| TCGA-GD-A3OQ* | MALE   | 48          | BI                | TCGA-HC-7740*| MALE   | 59          | BI                |
| TCGA-CV-5432  | MALE   | 68          | BI                | TCGA-HC-7752*| MALE   | 61          | BI                |
| TCGA-CV-5435  | MALE   | 57          | BI                | TCGA-HC-7820 | MALE   | 72          | BI                |
| TCGA-CV-5436  | MALE   | 65          | BI                | TCGA-BR-6452 | FEMALE | 78          | BI                |
| TCGA-CV-6933* | MALE   | 53          | BI                | TCGA-BR-6456 | FEMALE | 74          | BI                |
| TCGA-CV-6938* | MALE   | 87          | BI                | TCGA-BR-6564 | FEMALE | 46          | BI                |
| TCGA-CV-6954  | MALE   | 59          | BI                | TCGA-BR-6709 | FEMALE | 57          | BI                |
| TCGA-CV-6960* | MALE   | 49          | BI                | TCGA-BJ-A28R*| FEMALE | 38          | BI                |
| TCGA-CV-6962* | MALE   | 66          | BI                | TCGA-BJ-A28T | FEMALE | 34          | BI                |
| TCGA-CV-7089  | MALE   | 74          | BI                | TCGA-BJ-A28W*| FEMALE | 32          | BI                |
| TCGA-44-2662* | MALE   | 65          | BI                | TCGA-BJ-A290*| MALE   | 70          | BI                |
| TCGA-44-2665* | FEMALE | 55          | BI                | TCGA-BJ-A2N7 | FEMALE | 30          | BI                |
| TCGA-44-2666  | MALE   | 43          | BI                | TCGA-BJ-A2N8*| FEMALE | 30          | BI                |
| TCGA-44-5643  | MALE   | 53          | BI                | TCGA-44-7669 | MALE   | 59          | BI                |
| TCGA-44-6778* | MALE   | 59          | BI                | TCGA-44-7670 | FEMALE | 47          | BI                |
| TCGA-44-7660  | MALE   | 72          | BI                | TCGA-44-7672 | FEMALE | 52          | BI                |
| TCGA-44-7661  | FEMALE | 69          | BI                | TCGA-55-7570 | MALE   | 60          | BI                |
| TCGA-44-7662  | MALE   | 61          | BI                | TCGA-55-7576 | MALE   | 54          | BI                |
| TCGA-GC-A6I3* | NA     | NA          | BI                | TCGA-43-6647*| FEMALE | 69          | BI                |

Note: Sample ID with asterisk (*) have matched RNA-seq data, which we used for validation purpose.
Supplementary Table 2: The TCR scores for different tissue and cell types.

| Tissue                              | TCR score | Tissue                        | Score |
|-------------------------------------|-----------|-------------------------------|-------|
| Lymph node                          | 2.308     | Adrenal gland                 | 1.387 |
| Cardiac myocytes                    | 2.293     | Whole brain                   | 1.384 |
| BM-CD33+Myeloid                     | 2.197     | Bronchial epithelial cells    | 1.375 |
| Heart                               | 2.162     | Skeletal muscle               | 1.374 |
| Trachea                             | 2.135     | Thalamus                      | 1.37  |
| BM-CD34+                            | 1.893     | Amygdala                      | 1.358 |
| Globus pallidus                     | 1.846     | Liver                         | 1.349 |
| Adipocyte                           | 1.793     | Spinal cord                   | 1.348 |
| Dorsal root ganglion                | 1.786     | Adrenal cortex                | 1.329 |
| PB-CD14+Monocytes                   | 1.741     | Testis seminiferous tubule    | 1.321 |
| Appendix                            | 1.713     | Atrioventricular node         | 1.299 |
| Lung                                | 1.699     | PB-CD8+T Cells                | 1.297 |
| Tonsil                              | 1.697     | Cingulate cortex              | 1.284 |
| Salivary gland                      | 1.682     | Pancreas                      | 1.284 |
| Thymus                              | 1.663     | Cerebellum peduncles          | 1.262 |
| Tongue                              | 1.661     | Placenta                      | 1.254 |
| Uterus                              | 1.639     | Pons                          | 1.237 |
| Pancreatic islets                   | 1.618     | Caudate nucleus               | 1.228 |
| Thyroid                             | 1.606     | Olfactory bulb                | 1.223 |
| PB-BDCA4+Dendritic cells            | 1.585     | Prefrontal cortex             | 1.216 |
| Testis interstitial                 | 1.565     | Hypothalamus                  | 1.215 |
| Whole Blood                         | 1.541     | 721_B lymphoblasts            | 1.181 |
| PB-CD19+B cells                    | 1.53      | Temporal lobe                 | 1.159 |
| BM-CD71+Early erythroid            | 1.512     | Cerebellum                    | 1.148 |
| Ciliary ganglion                    | 1.51      | Testis germ cell              | 1.147 |
| Bone marrow                         | 1.502     | PB-CD56+NK cells              | 1.135 |
| Fetal lung                          | 1.484     | Fetal brain                   | 1.134 |
| PB-CD4+T cells                     | 1.465     | Fetal thyroid                 | 1.131 |
| Superior cervical ganglion          | 1.447     | Parietal lobe                 | 1.126 |
| Skin                                | 1.446     | Testis                        | 1.101 |
| Trigeminal ganglion                 | 1.445     | Prostate                      | 0.982 |
| Testis leydig cell                  | 1.441     | Medulla oblongata             | 0.971 |
| Fetal liver                         | 1.423     | Pituitary                     | 0.968 |
| Smooth muscle                       | 1.422     | Occipital lobe                | 0.906 |
| Ovary                               | 1.387     | Kidney                        | 0.842 |