Open access • Posted Content • DOI:10.1101/836841

Nascent RNA sequencing of peripheral blood leukocytes reveal gene expression diversity — Source link

Samantha Sae-Young Kim, Alexis Dziubek, Seungha Alisa Lee, Hojoong Kwak

Institutions: Cornell University

Published on: 09 Nov 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: MRNA Sequencing, Chromatin, Transcription (biology), RNA and Gene expression

Nascent RNA sequencing of peripheral blood leukocytes reveal gene expression diversity

[1]Samantha Sae-Young Kim, [1,2]Alexis Dziubek, [1,2]Seungha Alisa Lee, and [1]Hojoong Kwak

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, 14853, USA
[2]Program in Genetics, Genomics, and Development, Cornell University, Ithaca, NY, 14853, USA

Nuclear Run-On sequencing is a powerful method to measure transcription with high resolution, sensitivity, and directional information, which provides alternative perspective from existing methods such as chromatin immunoprecipitation or mRNA sequencing. Current form of Nuclear Run-On assays such as Precision Run-On sequencing (PRO-seq) involves multiple RNA chemistry steps including RNA end repairs and ligations. These have limited the widespread use of PRO-seq by requiring robust RNA handling skills and multiple days of effort. To solve this, we developed an ultrashort PRO-seq (uPRO) method that requires minimal steps. In uPRO, the requirement of only two reactions - RNA adaptor ligation and template switch reverse transcription - reduced the procedure into less than a single day. Using uPRO, we generated genome-wide transcription profiles of human haploid cell lines (HAP1) and peripheral blood samples combined with Chromatin Run-On sequencing (pChRO). Blood cell handling procedure is dramatically reduced using pChRO directly on crude chromatin preparations, and enables utilizing archived specimens. As a result, we identified individual differences in the transcriptional profiles of human whole blood from a small volume (~1 ml). We also generated blood cell type specific transcription data, and deconvoluted the nucleated blood cell compositions by modeling to the reference datasets. Overall, uPRO and pChRO provided a powerful platform to identify differentially expressed genes between individuals with minimal sample requirements.

Introduction

Analyzing and measuring the density of RNA polymerase in the genome enables us to see a glimpse of transcription both qualitatively and quantitatively[1]. These series of transcription, which are reactions from regulatory switches, come together to show specific genes that respond to specific signals. Identification of these genes and further analysis help to better grasp the mechanisms that explain their regulation. Not only that, the ability to quantify RNA polymerase density is crucial in breaking apart and understanding the regulatory steps involved in transcription[2].

Several other regions in the genome besides the protein-coding region are transcribed: enhancers, upstream divergent regions, and regions downstream of mRNA poly A sites. Short unstable RNAs called enhancer RNAs (eRNAs) are produced from enhancers and do not code for proteins[3] but rather identifiers of active transcription regulation[4]. Differential regulation of enhancer-mediated transcription is involved in several diseases[5]. Analysis of this differential regulation is crucial in discovering transcriptional changes from nutritional, environmental, and developmental factors. However, using RNA-seq to sequence total RNA is not efficient enough to detect such unstable RNAs.

Although there are several methods that have been proposed to enrich and sequence nascent RNAs attached to RNA polymerase, they depend on purification of insoluble chromatin[6] or are built upon immunoprecipitation of RNA polymerase[7-9]. These imply that the methods currently available highly rely on antibody specificity or the efficiency of chromatin fractionation alone. Nuclear Run-On (NRO) based methods use nascent RNAs elongated by polymerases with nucleotide analogs and can accurately map the polymerases as well as their start sites[10-12]. Nascent RNAs are selectively labeled by nucleotide analogs using the endogenous activity of RNA polymerase. These analogs serve as affinity purification tags, providing highly specific enrichment of the nascent RNA over other forms of RNA[11]. In addition, the direction of transcription is unambiguously identified through the directional sequencing of RNA.

Therefore, NRO methods are not only useful to analyze gene expression, but also to access the activities of regulatory elements and enhancers by capturing the noncoding RNAs simultaneously[3,12,13].
A larger scale analysis would be able to provide several advantages: 1) identification of context specific transcription and regulatory landscape, such as in human population or disease samples[14,15]; 2) comparison of genotype variations and transcriptional variation to identify complex transcription related quantitative trait loci (QTLs)[15]; 3) high statistical power to identify disease-specific transcription profiles from patient samples.

Use of unsupervised machine learning could allow novel discoveries in gene expression and regulatory element architecture, but requires a large number of training set data. The advantages in NRO methods to identify both gene expression and regulatory activity with high sensitivity and specificity make them a powerful platform to accumulate large scale databases.

However, practical limitations exist such as feasibility of the method and accessibility of the samples. The Precision Run-On sequencing (PRO-seq) method requires multiple days of hands-on procedure and robust RNA handling skills. Additionally, Nuclear Run-On requires the isolation of nuclei from intact cells, which is often a challenge for in vivo samples[12]. We previously introduced a Chromatin Run-On method (ChRO-seq) that uses insoluble chromatin isolates[14]. We demonstrated that the chromatin fraction contains enzymatically active RNA polymerases that are suitable for nuclear run-ons. Therefore, a combination of shortened PRO-seq procedure coupled to ChRO-seq in easily accessible samples would provide a powerful strategy to explore gene expression and regulatory landscape profiles in patients of specific diseases.

Here we present an ultrashort PRO-seq procedure (uPRO) coupled to peripheral blood Chromatin Run-On (pChRO) that takes less than a day to produce an Illumina compatible sequencing library. We demonstrate that uPRO provides nascent RNA data with a quality comparable to the conventional PRO-seq method. We used

uPRO to explore the transcriptional landscape of human haploid cell line HAP1. This cell line was specifically selected for its haploid characteristics which made CRISPR genome editing more efficient[16,17].

We applied pChRO on whole peripheral blood samples from different individuals and were able to calculate blood cell type compositions including Peripheral Blood Mononuclear Cells (PBMC) and Polymorphonuclear Leukocytes (PMNL). Not only that, we were able to measure and analyze inter-individual differences which allowed us to evaluate gene expression diversity that is unaffected by cell composition variations that we have observed. pChRO was also used on PBMC and PMNL cells separated which were experimentally isolated from whole blood samples.

Materials and Methods

Materials

HAP1 cells were maintained in IMDM media with 10% FBS and 1% penicillin-streptomycin. Only cells less than passage 10, mostly between 4-6 were used.

Chromatin preparation from human whole blood

1 ml of frozen blood sample is thawed and lysed in 10 ml of the NUN buffer (0.3M NaCl, 1M Urea, 1% NP-40, 20mM HEPES, pH 7.5, 7.5mM MgCl2, 0.2mM EDTA, 1x protease inhibitor cocktail, 1 mM DTT, 4 u/ml RNase inhibitor) with gentle mixing. Blood chromatin is pelleted by centrifugation at 15,000 g for 20 min, 4°C, and resuspended in Wash buffer (50 mM Tris pH 7.5). After brief centrifugation, the pellet is washed once more in Buffer D (50 mM Tris-HCl, pH 8.0, 25% glycerol, 5 mM Mg Acetate, 0.1 mM EDTA, 5 mM DTT), then homogenized using short sonication cycles in 50 µl of Buffer D.

Isolation of Peripheral Blood Mononuclear Cells (PBMC) and Polymorphonuclear Leukocytes (PMNL)

5-10 ml of peripheral whole blood is sampled from brachial veins. A final concentration of 1.5 mM EDTA is added to the whole blood to prevent clotting. To isolate PBMC and PMNL from the peripheral blood, 3 ml of Poly Cell Separation Media, 4 ml of Human Cell Separation Media, and 5 ml of blood are carefully layered consecutively with minimal mixing. Separation occurs with a centrifugation of 450-500 g for 30-35 minutes at 18-22°C. 3 interface layers should be visible. The top interface is PBMC and the middle interface including the layer right below is PMN. 5ml of each PBMC and PMN layers are collected in 1.7 ml microcentrifuge tubes. PBMC and PMN are pelleted by centrifugation at 4,000 for 4 min, 4°C, and washed in PBS. Centrifugation and wash is repeated once more. After brief centrifugation, the pellet is washed in Buffer D (50 mM Tris-HCl, pH 8.0, 25% glycerol, 5 mM Mg Acetate, 0.1 mM EDTA, 5 mM DTT), then resuspended in 50 µl of Buffer D.

uPRO library preparation

Chromatin or cells were incubated in the nuclear run-on reaction condition (5 mM Tris-HCl pH 8.0, 2.5 mM $MgCl_2$, 0.5 mM DTT, 150 mM KCl, 0.5% Sarkosyl, 0.4 units / µl of RNase inhibitor) with biotin-NTPs and rNTPs supplied (18.75 µM rATP, 18.75 µM rGTP, 1.875 µM biotin-11-CTP, 1.875 µM biotin-11-UTP for uPRO; 18.75 µM rATP, 18.75 µM rGTP, 18.75 µM rUTP, 0.75 µM CTP, 7.5 µM biotin-11-CTP for pChRO) for 5 min at 37°C. Run-On RNA was extracted using TRIzol, and fragmented under 0.2 N NaOH for 15 min on ice. Fragmented RNA was neutralized, and buffer exchanged by passing through P-30 columns (Biorad). 3′ RNA adaptor (/5Phos/NNNNNNNNGAUCGUCGGACUGUAGAACUCUGAAC/3InvdT/) is ligated at 5 µM concentration for 1 hours at room temperature using T4 RNA ligase (NEB), followed by 2 consecutive streptavidin bead bindings and extractions. Extracted RNA is converted to cDNA using template switch reverse transcription with 1 µM RP1-short RT primer (GTTCAGAGTTCTACAGTCCGA), 3.75 µM RTP-Template Switch Oligo (GCCTTGGCACCCGAGAATTCCArGrGrG), 1x Template Switch Enzyme and Buffer (NEB) at 42°C for 30 min. After a SPRI bead clean-up, the cDNA is PCR amplified using primers compatible with Illumina Small RNA sequencing. The whole procedure takes ~6 hours with ~3.5 hours of hands on time.

Sequencing data processing

Illumina sequencing data is processed using a PRO-seq specialized analysis toolkit STOAT (https://github.com/sl2665/stoat). STOAT automatically converts a raw read fastq file to unique molecular identifier (UMI) sorted alignment file (bam) and genome wide read count coverage files (bedgraph) on both strands, and gene expression tables in promoter proximal, gene body and exon regions in one step. To map transcriptional regulatory elements (TREs) de novo, we used dREG (https://github.com/Danko-Lab/dREG)[13]. Read counts on de novo gene annotations were made using the BEDtools suite (https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html)

Differential expression analysis and Distal Enhancer (DE)-gene interaction analysis

We used the raw read count tables generated by STOAT on reference gene annotations or on dREG identified TREs. The raw read count tables were processed through DESeq2[34] and selected differentially expressed genes or TREs with FDR < 0.05. To identify DE-gene pairs, we search for the nearest transcription start sites (TSS) from each dREG TRE entry. The nearst TSS was paired to the TRE, then the location of the TRE relative to the gene TSS was categorized into one of promoter (PRM), gene-body (GB) or distal-enhancer (DE) categories. Only the DE and gene TSS pairs were used in the distant interaction analysis.

Cell type signature gene selection and decomposition analysis

Gene body uPRO data from PBMC and PMNL are normalized to reads per million mapped reads per kilobase (RPKM). PBMC and PMNL signature genes were selected by $\log_2$ ratio between the two, using ±4 (16 fold difference) as a cutoff. Gene ontology analysis of the signature genes were performed using the PANTHER geneontology analysis[35]. Cell type fractions were calculated from the relative expression levels of each signature gene subset relative to the reference data, assuming reference peripheral leukocyte compositions of PBMC = 0.35 and PMNL = 0.65, B = 0.07, T = 0.21, and CD4 = 0.06[25].

Results

Comparison between PRO-seq and uPRO procedures

Compared to PRO-seq, uPRO requires less RNA chemistry and handling steps (**Fig 1A**)[12]. In PRO-seq, Nuclear Run-On (NRO) using biotin-NTPs is performed on isolated nuclei. In uPRO, the NRO reaction is performed directly on washed cells or resuspended chromatin isolates. After the NRO reaction, the biotin-labeled nascent RNA is fragmented and the buffer is exchanged to remove excess biotin-NTPs and salts. In uPRO, 3' RNA adaptor ligation takes place for 1 hour before biotin-RNA enrichment as opposed to the 6 hour - overnight ligation after biotin-RNA enrichment in PRO-seq. This change greatly shortens the amount of time spent. 2 consecutive streptavidin bead binding then take place and an extraction is performed to enrich biotin-labeled nascent RNA. Rather than having 3 streptavidin affinity purifications throughout the procedure, we found that 2 consecutive affinity purifications were sufficient to remove potential adaptor dimers and unlabeled endogenous RNAs. In PRO-seq, RNA extraction from the beads includes multiple ethanol precipitations which often serves as a point of failure and loss of RNA materials. In uPRO, we replaced it with direct buffer exchange between the consecutive affinity purifications and a column based RNA purification to further shorten time and improve RNA yield.

PRO-seq requires two 5' RNA end pair chemistries: de-capping and phosphorylation to modify the 5' ends to become acceptor sites for RNA adaptor ligation. In uPRO, we proceed directly to reverse transcription using template switching to produce cDNA and add 5' adaptor sequence at the same time[28]. The cDNA product is processed through SPRI bead clean-up steps which removes short unused excess adaptors and primers. This serves as an additional enrichment step that reduces unwanted adaptor dimer products. As a result, the amount of amplified product in uPRO is more predictable than the conventional PRO-seq. Not only that, uPRO does not require test amplifications or polyacrylamide gel electrophoresis (PAGE) purifications (**Fig 1B**). The relatively small amount of adaptor dimers are usually negligible for the Illumina sequencing (**Fig 1C**). Overall,

uPRO may take as short as 6 hours to complete, compared to the 4-day conventional PRO-seq procedure (**Fig 1A**).

Global transcriptional landscape of the human haploid cell line HAP1

We applied uPRO and PRO-seq on the human haploid HAP1 cell line derived from human myeloid leukemia cells[16], as well as one of the widely used cancer cell lines HeLa. Haploid cell lines provide an advantage over other diploid or multiploid cell lines since only one allele of the gene or elements needs to be modified. This is a critical advantage in large scale genetic screening using genome editing technologies such as TALE or CRISPR. Therefore, a HAP1 transcriptional landscape profile will serve as a useful baseline dataset[18]. We also used the PRO-seq data from human embryonic kidney HEK293 cells and human lymphoblastoid cell lines (LCLs) previously presented in our publications[19,15]. Overall, uPRO shows transcription profiles to be equivalent to PRO-seq results in HAP1 as well as HeLa cells. For example, both uPRO and PRO-seq show highly consistent transcription profiles at TAL1 gene, a HAP1 specific gene that is often up-regulated in T cell origin leukemias (**Fig 2A**)[20]. Both the sense and antisense strand transcription patterns are efficiently captured. An adjacent gene STIL is expressed in both HAP1 and HEK293 cells, and the expression pattern between uPRO and PRO-seq is in high agreement.

As a quantitative measure of uPRO's reproducibility of previously presented PRO-seq data, we compared promoter proximal and gene body read counts. Promoter proximal regions are defined as ± 500 bp from the transcription start sites and reflect the amount of RNA polymerases that are paused. Gene body regions reflect the amount of actively transcribing RNA polymerases which represents overall gene expression levels. Correlation coefficients between regions are greater than 0.9 between uPRO and PRO-seq in promoter proximal regions and 0.97-0.98 in gene body regions in HAP1 and HeLa cells (**Fig 2B, 2C**). These correlation coefficients between uPRO and PRO-seq are slightly less than between PRO-seq replicates in HEK293 cells (**Fig 2D**), but still demonstrates that uPRO quantification is a reasonably close estimate of PRO-seq quantification, in particular on the gene bodies. When we included other blood cell derived LCLs (GM18520, GM19222) from different individuals in the hierarchical clustering analysis, uPRO and PRO-seq results cluster together within the same cell lines and clustering isn't affected by the method. This demonstrates the robustness of uPRO in agreement with PRO-seq (**Fig 2E**). In particular, HAP1 cells cluster together with other aggressively growing transformed cell lines (HEK293 and HeLa), compared to less aggressively immortalized cell lines with normal karyotypes (LCLs)

Identification of HAP1 specific genes and regulatory elements

Nascent RNA sequencing can be used to measure both gene expression level and activity of regulatory elements such as enhancers. We used a machine learning tool called dREG to identify these regulatory elements from bidirectional patterns of transcription from nascent RNA sequencing data. Using dREG, we identified 68,896 annotated genes and 29,461 transcriptional regulatory elements which included both promoters (n = 14,980) and enhancers (n = 14,481)from our collection of HAP1, HeLa, and HEK293 cell data. Of the enhancers, 11,212 were distal enhancers (DEs) located upstream of the promoters.

We used uPRO and PRO-seq data to identify differentially expressed genes in HAP1 cells compared to HeLa or HEK293 cells using the gene body expression levels. 8,565 genes are differentially expressed in HAP1 (**Fig 3A**), showing that more than 10% of the genes are involved in cell type specification in HAP1 cells. 3,469 dREG promoters (23%) are differentially expressed (1,968 up-regulated) in HAP1 cells (**Fig 2B**), indicating that a large fraction of active promoters are cell type specifically regulated. Likewise, 2,644 distal enhancers (24%) are differentially expressed (1,557 up-regulated) (**Fig 3C**) in HAP1 cells.

To determine if these dREG-identified DEs in HAP1 cells are associated with target gene expression, we tested the correlation between nascent RNA expression levels at DEs and the gene bodies of their nearest TSSs. Only the upstream intergenic DEs are tested to remove confounding effects from the gene body

transcription of other genes. When we compared the expression level of the elements normalized by the HEK293 cell levels, we observed significant correlation of expression levels between the DEs and their nearest target genes (**Fig 3D**). This indicates that the DEs and their paired genes are co-regulated and that the DEs are likely regulators. We further categorized the DE-gene pairs dependent on their distances and saw overall decay in correlation over the distance (**Fig 3E**). This distance dependent decay trend is more apparent in HeLa cells than in HAP1 cells. While we cannot completely rule out the effect of technical variabilities, this result may suggest the presence of differential genome organization between haploid HAP1 and hyper-triploid HeLa cells.

We also investigated whether specific transcription factor binding sites are enriched in HAP1 specific dREG elements. While we did not observe statistically strong enrichment, CEBPA sites are marginally enriched (1.3 fold) in HAP1 specific elements which agrees with other reports that CEBP transcription factors are involved in hematological malignancies.

Chromatin Run-On from peripheral blood samples (pChRO)

While cell lines are reliable sources for transcriptional profiling, there are limited resources to established cell lines. Using a large scale study will increase the analytical power in discovering inter-individual or disease associated gene expression differences. But it is not always feasible to generate cell lines (primary cells or iPSCs) from clinical subjects at a large scale. One of the most accessible clinical specimens is peripheral blood. Gene expression analysis performed directly on blood samples may provide a feasible approach in large scale studies. However, blood plasma has high-concentrations of nucleases that degrade the quality of RNA. Additionally, over-abundance of globulin mRNAs from red blood cells (RBCs) complicate RNA expression profiling experiments. Isolated leukocytes can be used for RNA expression of other transcriptional assays such as ChIP or chromatin accessibility assays, but the isolation step itself can add variability and be laborious to implement at a large scale.

Chromatin Run-On (ChRO) is an alternative Nuclear Run-On (NRO) based assay that uses precipitated chromatin isolates that contain active RNA polymerases. ChRO-seq is able to successfully generate nascent RNA data from cryopreserved archived solid tissue specimen, despite that RNA is degraded over time from these harsh conditions, because RNA polymerases can remain actively engaged. With length extension of the nascent RNAs in NRO, RNA polymerase levels and positions can be mapped even under severe RNA degradation conditions. We applied this strategy to human peripheral blood samples. Since leukocytes unlike RBCs contain a nucleus, ChRO-seq results will reflect the transcriptional landscape of peripheral leukocytes. We were able to successfully generate ChRO-seq libraries from just 1 ml of peripheral blood samples that did not undergo any special treatment but simple storage in -20°C after sampling (pChRO).

The resulting pChRO data shows high correlation in gene body read counts (0.97 - 0.98) within technical three replicates (**Fig 4A**). We also compared pChRO data from a different individual and observed slightly less but correlated gene body levels (0.95 - 0.96). The pChRO profile is reproducible between different individuals and within cell types. For example, the expression of IKZF1 gene, a known regulator of lymphocyte differentiation, is consistently high in the pChRO data from both individuals but is not expressed in HeLa cells (**Fig 4B**). On the other hand, the FIGNL1 gene is only expressed in HeLa cells but not in either of the pChRO data from the two individual samples. When we tested the gene body correlation between the pChRO replicates and individuals against HeLa and LCL samples, we saw clear clustering of pChRO data among the pChRO replicates (**Fig 4C**). The pChRO data reflecting peripheral leukocytes correlate relatively better with LCLs, B cell derived cell lines, than HeLa cells indicating that the cell type lineage is still preserved in immortalized LCLs.

Leukocyte decomposition between Peripheral Blood Mononuclear Cells (PBMCs) and Polymorphonuclear leukocytes (PMNLs)

One of the challenges in identifying differentially expressed genes and regulatory elements from primary cells is cellular heterogeneity. If there are factors that influence cell type compositions, this will lead to false identification of cell type specific genes as differentially expressed genes. Peripheral leukocytes are composed of PBMCs that include B, T lymphocytes, and monocytes; PMNLs are considered as the granulocyte population composed of neutrophils, eosinophils, and basophils. The nuclear morphology and gene expression profiles are known to be different between the two, but there has been no systematic comparative transcription analysis. In particular, the gene expression profiles of PMNLs are not as extensively studied as PBMCs.

To address this, we isolated PBMCs and PMNLs fresh from peripheral blood samples and applied uPRO on the isolated cells. The majority of the transcription profile appears to be similar between PBMC, PMNL, and whole leukocyte. However, we found genes that are differentially and almost exclusively expressed in one cell type versus the other (**Fig 5A**). For example, the ALPL gene is exclusively expressed in PMNLs but not in PBMCs. Therefore, since the level of ALPL expression is specific to PMNL cells, its expression level should also be reflected from the amount PMNL cells in the whole blood.

On the other hand, a nearby gene USP48 is more highly expressed in PBMCs compared to PMNLs (**Fig 5A**). Quantitative assessment of these exclusively expressed signature genes should allow us to precisely estimate the PBMC and PMNL ratio in the whole blood.

To identify PMNL or PBMC exclusive genes, we calculated the ratio between PMNL and PBMC normalized gene body read counts. We identified 157 PMNL and 429 PBMC signature genes that have at least 16 fold expression differences (**Fig 5B**). Gene ontology analysis of these signature genes were consistent with the expectation. For example, PMNL signature genes are enriched with granulocyte activation (GO:0036230), neutrophil involved pathways (GO:0002446, 0002283, 0043312, 0042119), and cell motility/migration (GO:0040011, 0048870, 0050900) (**Fig 5C**). These pathways are consistent with the function of granulocytes and neutrophil innate immune responses. On the other hand, PBMC signature genes are enriched with adaptive immune response (GO:0002250), receptor mediated immune signaling (GO:0050851, 0002768, 0002429), T cell pathways (GO:0050852, 0050853, 0045058), and B cell pathway (GO:0050853) (**Fig 5C**). This is consistent with the fact that PBMCs are mostly composed of T cells and B cells.

Interestingly, we found that many of the top PBMC signature genes are among the ZNF subfamily genes. These genes are clustered on chr19 q13.31 ~1 megabase region (chr19: 43,700,000 - 44,700,000). We found that the whole 1 megabase region is repressed in PMNLs but is expressed in PBMCs (**Fig 5D**). This variation in transcription could potentially lead to a mis-interpretation that a factor affecting the PBMC and PMNL ratio may appear to influence the large range repression of this ZNF cluster. This case illustrates the importance of deconvoluting cell type heterogeneity in peripheral blood gene expression profiling.

Decomposition of leukocyte subtype fractions from pChRO and reference PBMC/PMNL data

For exclusively expressed signature genes, the ratio between normalized pChRO data and the signature cell type will reflect the relative cell fraction in the whole blood leukocyte population. Since PMNLs normally compose 65% of the leukocyte population, we should expect to see the pChRO data recapitulate the PMNL profile more than the PBMC profile. However, PBMC profiles are closer to the whole blood pChRO profile and $\log_2$ fold difference is closer to 0 after normalizing to the pChRO data (**Fig 5E**). This indicates that PBMCs are transcriptionally much more active than PMNLs; Even though PMNLs compose greater fraction of the cells, PBMCs override the transcriptional profiles. We calculated the relative transcriptional activities and PBMCs are on average ~ 4.1 fold transcriptionally more active than PMNLs. Regardless, the expression difference between the two cell types is pronouncedly distinct in the signature genes (**Fig 5E**).

To calculate the PBMC and PMNL compositions, we calculated the average expression level of signature genes relative to the reference pChRO data. This average is proportional to the relative fraction of the

signature cell type. From the pChRO profiles of the individuals 1 and 2, Individual 2 appears to have a slightly higher overall expression level of PMNL signature genes than Individual 1 (**Fig 5F**). We were able to calculate the cell subtype fractions from these signature gene averages (**Fig 5G**). Individual 2 has marginally but significantly higher levels of PMNL fractions while the two technical replicates of Individual 1 showed overlapping error margins.

Cell type decomposition using comparative analysis between mRNA expression data and pChRO

Whereas the distinction between PBMC and PMNL is the biggest portion of peripheral leukocyte fractions, subpopulation compositions within PBMC can also influence the overall pChRO profile. Since PBMCs have much higher transcriptional activity than PMNL, PBMC subpopulations such as B cell and T cell subtypes can also influence the overall pChRO expression profile. To further investigate this possibility and subclassify PBMC populations, we performed a comparative analysis between existing RNA expression data in PBMC subtypes[26] and our pChRO data.

We first compared our uPRO data from PBMCs and PMNLs to a microarray data in all leukocytes: PBMC, B cell, T cells, and CD8+ T cells. Direct comparison between nascent RNA sequencing and mRNA microarray results can be affected by a lot of variables which does not make it feasible. Instead, we normalized the cell type specific uPRO data by all leukocyte pChRO profiles and microarray subtype data by all leukocyte results. We saw that there is significant global correlation between uPRO PBMC and microarray PBMC when normalized by all leukocyte data (**Fig 6A**). In addition, microarray data in other PBMC subtypes such as B cells and CD4+ T cells are correlated with the uPRO PBMC data. On the other hand, uPRO PMNL data did not show any significant correlation with any other microarray-based cell subtype data. Normalizing the microarray data by cell types other than all leukocytes made the correlations disappear, indicating that the correlation between uPRO PBMC and microarray PBMC cell subtypes are specific to the cell subtype and proper normalization (**Fig 6A**).

After confirming that the nascent RNA sequencing data is in agreement with the existing mRNA expression data, we compared the signature gene lists from both data sets. The PBMC signature genes from the uPRO data are driven mostly by strong depletion in the PMNL population (**Fig 6B**). According to the mRNA data, enrichment of the signature genes in PBMC subpopulations was variable. However, we identified groups of the uPRO PBMC cluster genes that appear more specific to B cell or T cell subpopulations. On the other hand, PMNL signature genes that are depleted in PBMC cells also appear depleted in the mRNA microarray data (**Fig 6B**).

Conversely, the mRNA microarray signature genes show expression level differences in the uPRO data (**Fig 6C**). The mRNA data did not include PMNL cell isolates as an exclusive comparison but rather used the subtractive enrichment or depletion in PBMC over all leukocyte as a PBMC marker (labeled LYMPHS) or PMNL marker (labeled GRANS). Therefore the PBMC(LYMPHS, green) signature gene expression was not as prominent in the uPRO PBMC data (pPBMC). However other signature genes in B, T, and CD8 cells are more strongly enriched (**Fig 6C**). Conversely, the PMNL (GRANS, blue) signature gene expression is markedly reduced in PBMC showing the reproducibility of mRNA microarray signature genes in uPRO data.

Finally, we compared the individual pChRO data to the mRNA microarray signature genes (**Fig 6D**). Consistent with the uPRO signature genes (**Fig 5G**), we saw overall slight increase in PMNL (GRANS) signature genes in Individual 2 which indicates an increase in the PMNL subpopulation (**Fig 6D**). Overall, these results demonstrate that nascent RNA sequencing and mRNA microarray are in good agreement and can be used together to further deconvolute specific cell subtypes from the pChRO data.

Discussion

Nascent RNA sequencing is a powerful method that can map and measure gene expression and the activities of the transcriptional regulatory elements such as enhancers at the same time. Existing methods are considered difficult and pose practical limitations for large scale applications. Our new uPRO approach, coupled to peripheral blood ChRO-seq is expected to lower the barrier and facilitate the production of nascent RNA data in larger scales.

Using human cell lines, we demonstrated that uPRO data quality is comparable to conventional PRO-seq, but can be processed much faster. Removal and shortening of many critical enzymatic steps allowed a much efficient use of library preparation times. High correlation of the gene body read counts between uPRO and PRO-seq reassures that uPRO is equivalent to PRO-seq in terms of gene expression analysis. Correlation in the promoter proximal read counts were slightly lower, suggesting that there may be method specific biases in collect read counts from a shorter range region near the 5' end of the gene. One of the potential concerns was the use of template switch reverse transcription, since it has been reported that template switch is more efficient on 5' capped RNA ends than other forms of 5' ends[27,28]. While we observed minor differences in the uPRO peak patterns, we did not see systematic evidence that 5' capped ends are more enriched. In fact, the degree of promoter proximal correlation is similar to or better than previous reports of correlation between PRO-seq and PRO-cap, which use similar approaches with different adaptor ligation strategies[29].

We were further able to map and quantify the expression of cell type specific genes and regulatory elements. We further observed that the cell type specific distal enhancer transcription (eRNA transcription) correlated with the putative target gene in the vicinity. The strength of correlation was negatively associated with the distance as expected. But the degree of distance dependence was variable between the cell lines, polyploid HeLa cells[30] appear to have more rapid decay than the haploid HAP1 cells within 1 Mb regions, and HAP1 cells appear to have more long-range interactions. Although we will need to rule out potential biases, but this observation on potential association between ploidy and distance is somewhat consistent with a study reporting fewer short range interactions (< 1 Mb) in Arabidopsis tetraploid cells compared to diploid cells[31].

Peripheral blood is one of most accessible specimens, and can be a valuable resource to obtain large scale data. However, there are often limitations due to the presence of plasma nucleases and the requirement to isolate leukocyte from massively abundant red blood cells. Our pChRO procedure is best optimized for this purpose requiring no sample pretreatment and only very simple preparation steps. In turn, we successfully generated whole blood leukocytes as well as leukocyte subpopulation data. In particular, polymorphonuclear leukocytes (PMNLs) including neutrophils are known to be transcriptionally inert, and we have quantitatively analyzed the global transcriptional activity. However, we identified PMNL specific expressed genes as well as the large domain repression as seen in the ZNF cluster. The large domain repression may be related to the segmentation of the nuclei in neutrophils and would provide further insight in linking gross nuclear morphology with molecular events taking place at these domains[32,33].

Our pChRO data demonstrated high level of consistency between technical replicates, and showed potential to identify differentially expressed genes between different individuals. We also demonstrated that cell subtype fractions can be calculated and provide important check-points in discovering true differentially expressed genes. Furthermore, in addition to identifying differential gene expressions and mapping regulatory elements, these peripheral leukocyte pChRO data may provide further information on personalized health conditions, and may open up a new revenue in the studies of genome-wide transcription.

Acknowledgements

We thank the members of the Kwak lab and John Lis lab in Cornell University for the scientific and technical discussions and reagent sharing.

Author Contributions

SSYK, and AD performed the molecular experiments. SSYK produced most of the data used in this manuscript. AD carried out a major fraction of the uPRO method optimization. SSYK completed the development of uPRO and pChRO methods. SAL and HK performed the computational analysis. The manuscript was written by SSYK and HK.

Data depository

GSE103719, GSE110638

Figure legends

Fig 1. Schematics of the uPRO procedure
A. Comparison between conventional PRO-seq and uPRO procedures. Adapted from Mahat et al[12]
B. Polyacrylamide gel electrophoresis of PRO-seq and uPRO libraries.
C. Capillary electrophoresis trace (BioAnalyzer) of a representative uPRO library. LM: lower marker (50 bp), UM: upper marker (5,000 bp)

Fig 2. Comparison between uPRO and PRO-seq.
A. Browser view of a representative loci showing cell type specific genes. Red: plus strand gene or PRO-seq track, blue: minus strand gene or PRO-seq track.
B. Correlation scatterplots between uPRO and PRO-seq in the promoter proximal and gene body regions in HAP1 cells. x- and y-axes: $\log_{10}$ reads per kilobase per million mapped reads (RPKM). Number on lower right corner is the Pearson's correlation coefficient.
C. Correlation scatterplots between uPRO and PRO-seq in HeLa cells.
D. Correlation scatterplots between PRO-seq biological replicates in HEK293 cells.
E. Correlation heatmaps among uPRO and PRO-seq samples from various cell types. Color scale bar represents Pearson correlation coefficient of the $\log_{10}$ RPKM reads.

Fig 3. Identification of cell type specific genes and non-coding RNA expression
A. Heatmap of HAP1 specific genes. Color label represents $\log_2$ fold difference from the mean of all cell types. Column labels suffixes are .u: uPRO, .pro: PRO-seq, 1: PRO-seq replicate 1, 2: PRO-seq replicate 2.
B. Heatmap of HAP1 specific dREG promoters. See panel A for description.
C. Heatmap of HAP1 specific dREG enhancers. See panel A for description.
D. Correlation scatterplots between Distal Enhancer (DE) and paired target gene expression. x- and y-axes are $\log_{10}$ RPKM. Lower right corner: Pearson's correlation coefficient.
E. Correlation-distance plot. DE-gene pairs are classified into 0-1, 1-2, 2-4, 4-8, 8-16, 16-32, 32-64, >64 kb bins, and Pearson's correlation coefficient between the $\log_{10}$ RPKMs of DE and gene body levels are plotted.

Fig 4. pChRO and peripheral leukocyte transcriptome profiling
A. Correlation scatterplots of the pChRO gene body data (3 technical replicates of Individual 1 + Individual 2). X- and y-axes: $\log_{10}$ RPKM.
B. Browser view of an example site. Red: plus strand gene or PRO-seq track, blue: minus strand gene or PRO-seq track.
C. Correlation heatmaps among uPRO and PRO-seq samples from various cell types. Color scale bar represents Pearson correlation coefficient of the $\log_{10}$ RPKM reads. Column labels suffixes are .u:uPRO, .pro:PRO-seq, 1.1: Individual 1 replicate 1, 1.2: Individual 1 replicate 2, 1.3: Individual 1 replicate 3, 2: Individual 2. GM18520 & GM19222: lymphoblastoid cell lines (LCL).

Fig 5. Decomposition of heterogeneous blood cell types between PBMC and PMNL
A. Browser view of an example exclusively expressed site. Red: plus strand gene or PRO-seq track, blue: minus strand gene or PRO-seq track.
B. PMNL/PBMMC ratio of all genes. Blue: PMNL signature genes, red: PBMC signature genes. Signature gene cut-off: greater than 16 fold difference.
C. Gene ontology (GO) catagories of PMNL and PBMC signature genes. Top 25 are shown
D. Browser view of the ZNF cluster at chr19q13.31. Red: plus strand gene or PRO-seq track, blue: minus strand gene or PRO-seq track.
E. Heatmap of signature gene expression in PMNL and PBMC uPRO gene body normalized to whole leukocyte pChRO levels. Color scale bar: $\log_2$ fold difference.
F. Heatmap of signature gene expression in pChRO individual samples. Thin ribbons on the right represent the average fold difference of the signature gene group in the corresponding individual. Color scale bar: $\log_2$ fold difference.
G. Estimated cell type fractions calculated from signature gene expression levels. Error bars: standard error of the mean.

Fig 6. Decomposition of sPBMC subtypes and identification of differentially expressed genes
A. Correlation heatmap between PRO-seq and mRNA micrarray data from different peripheral leukocyte subpopulations. Color scale bar: Pearson correlation coefficient of the $\log_2$ read count/micrarray levels normalized to the reference cell type indicated on the denominator.
B. Heatmap of PRO-seq PBMC and PMNL signature gene expressions in PRO-seq or mRNA micrarray data in different subpopulations. Column prefixes: p-PRO-seq, r-mRNA microarray. Color scale bar: $\log_2$ fold difference.
C. Heatmap of microarray leukocyte subtype (B, T, CD8+ T, LYMPHS = PBMC, GRANS = PMNL). signature gene expressions in PRO-seq or mRNA micrarray data. Column prefixes: p-PRO-seq, r-mRNA microarray. Color scale bar: $\log_2$ fold difference.
D. Heatmap of microarray signature gene expression (B, T, CD8+ T, LYMPHS = PBMC, GRANS = PMNL) in pChRO individual samples. Thin ribbons on the right represent the average fold difference of the signature gene group in the corresponding individual. Color scale bar: $\log_2$ fold difference.

Figures

Fig 1.

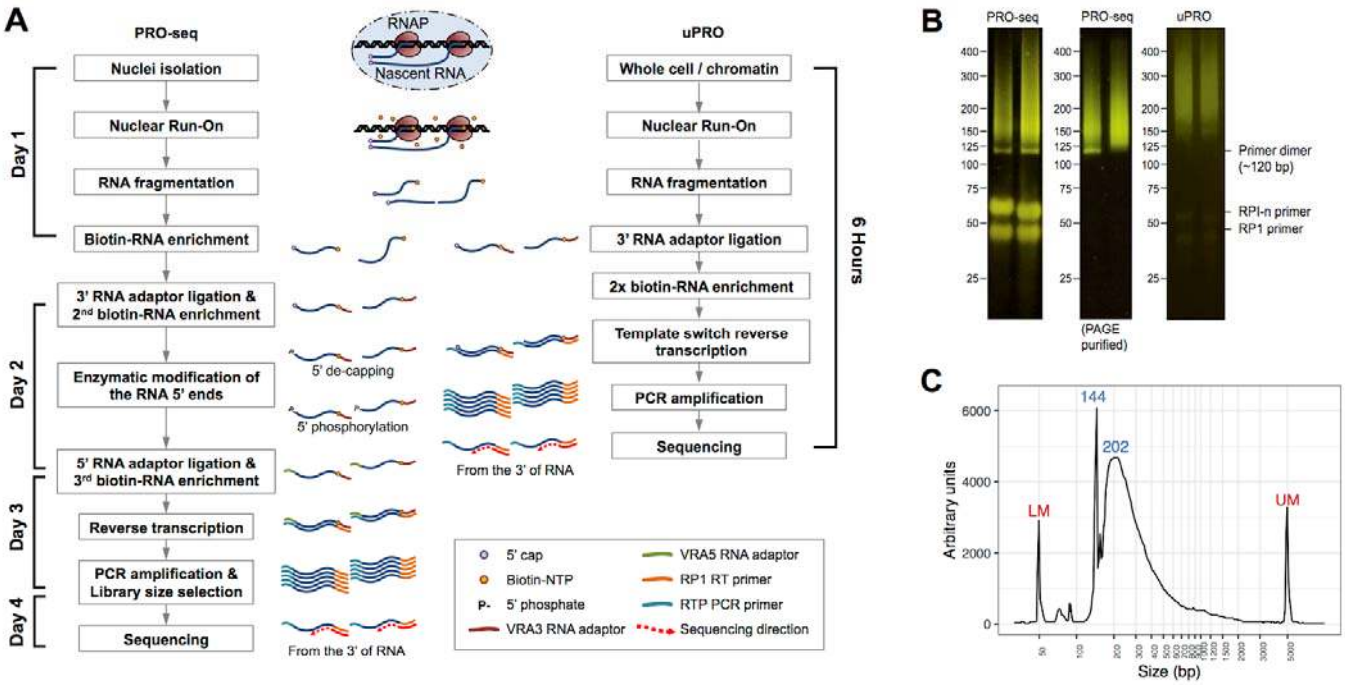


Fig 2

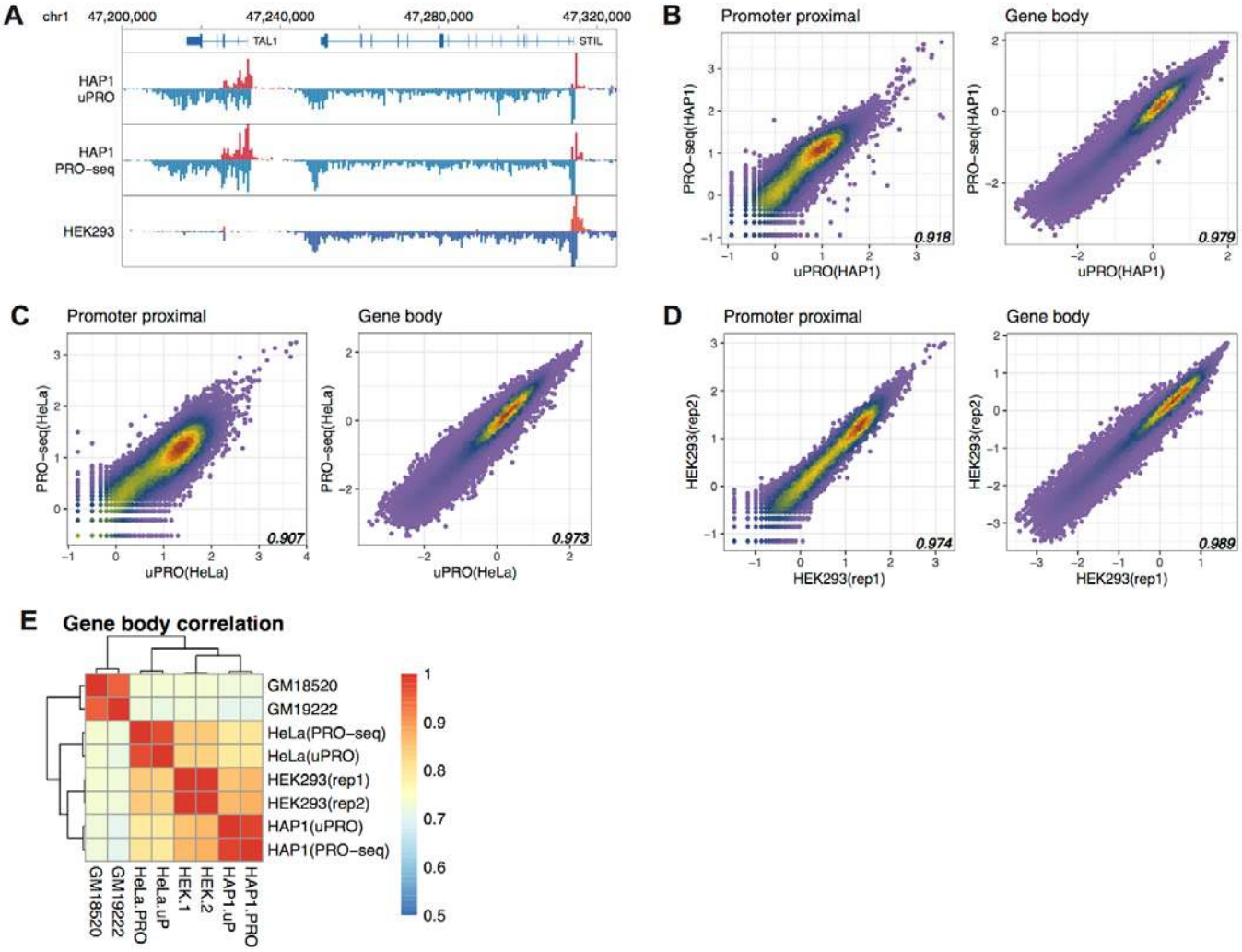


Fig 3

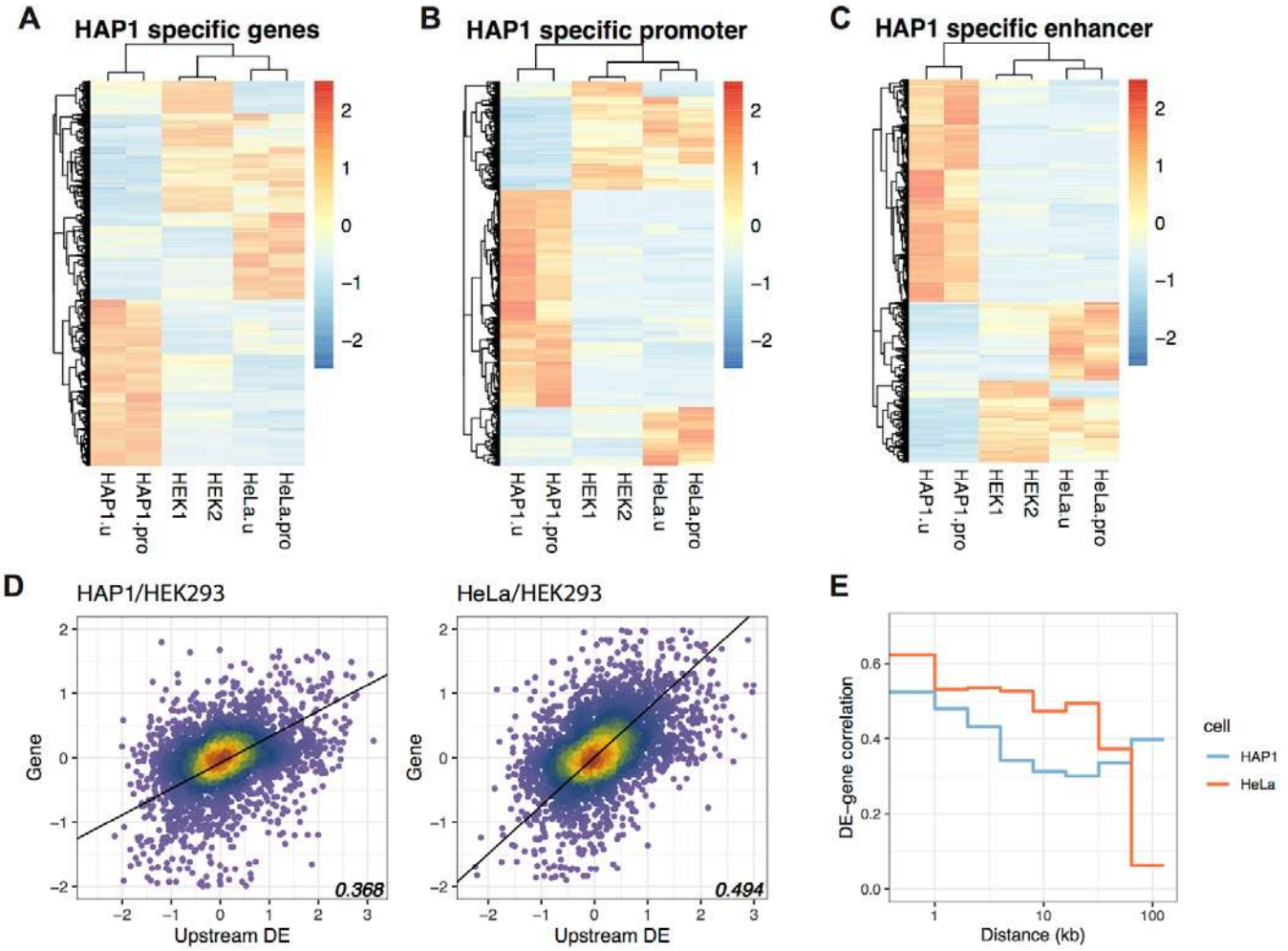


Fig 4.

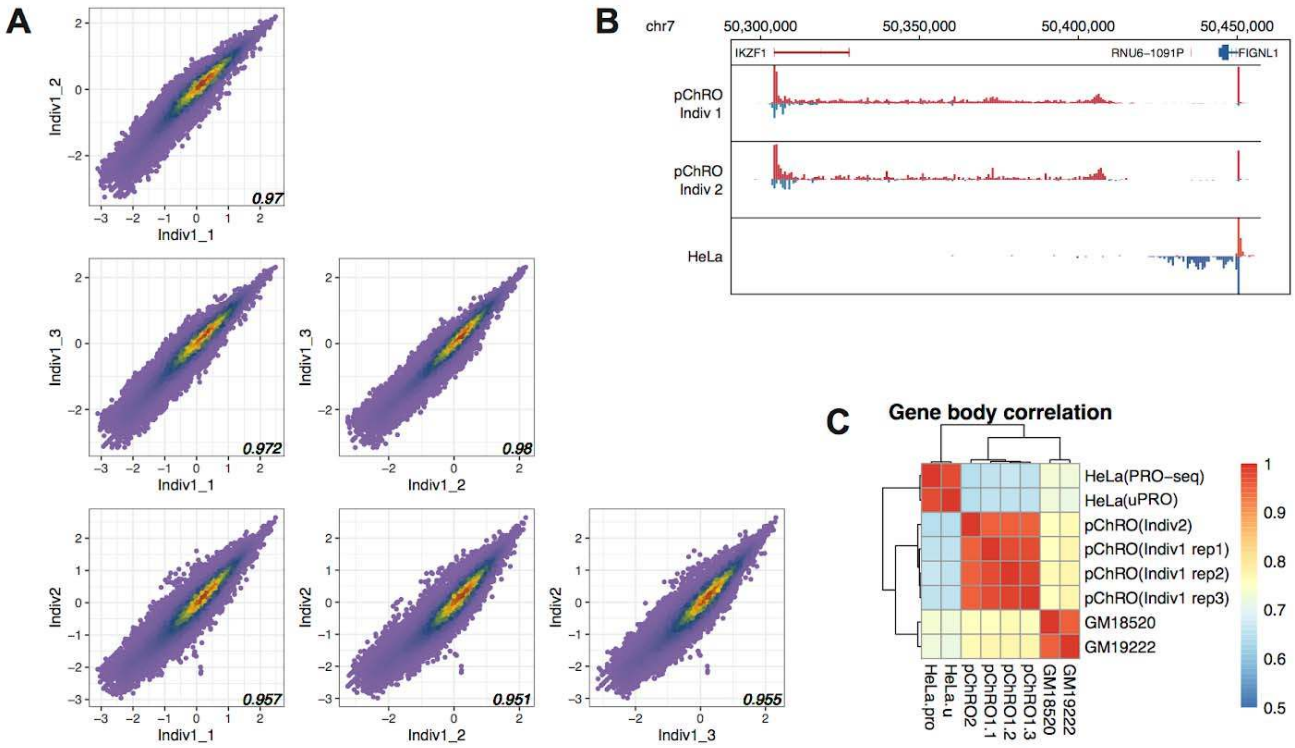


Fig 5.

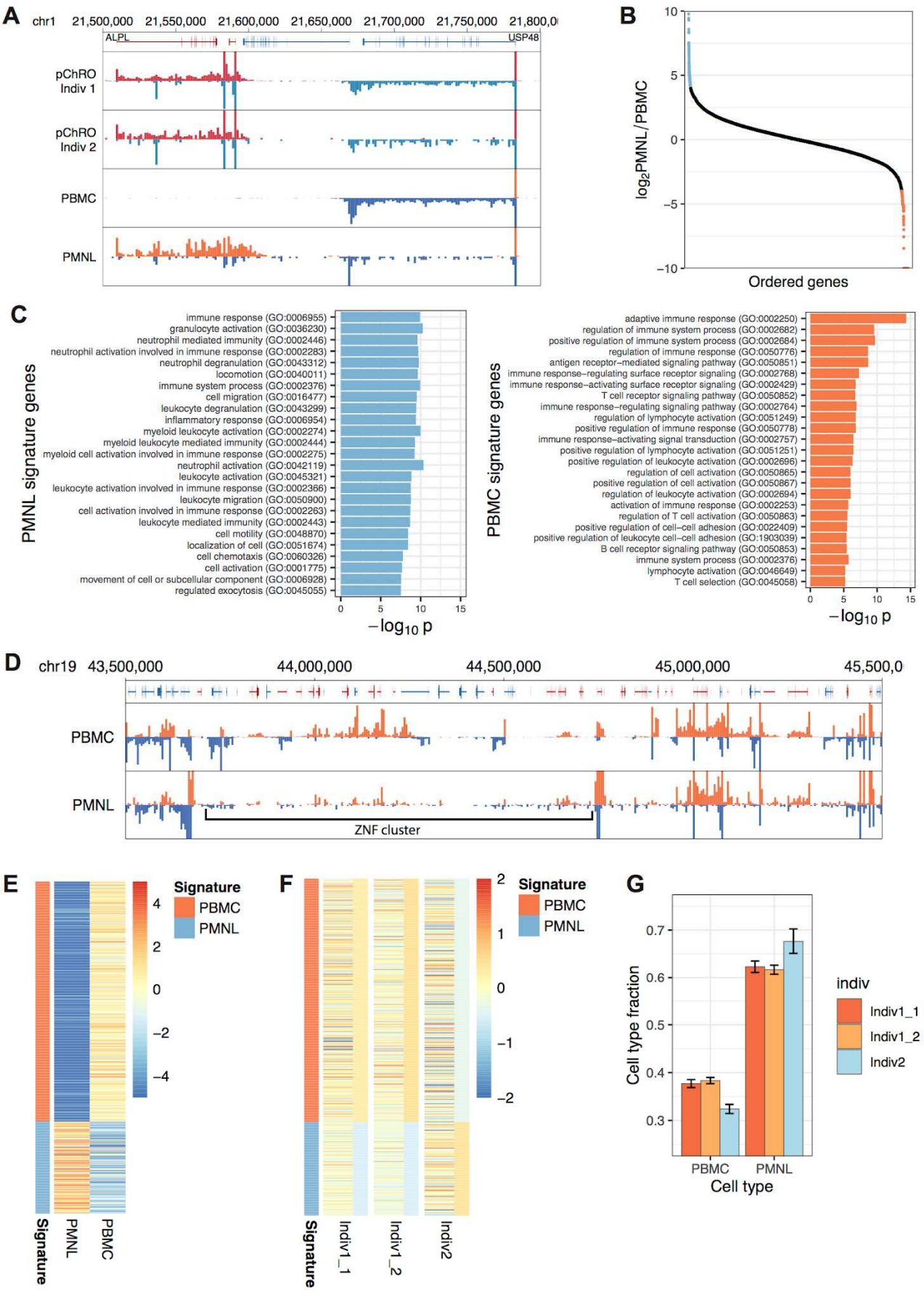
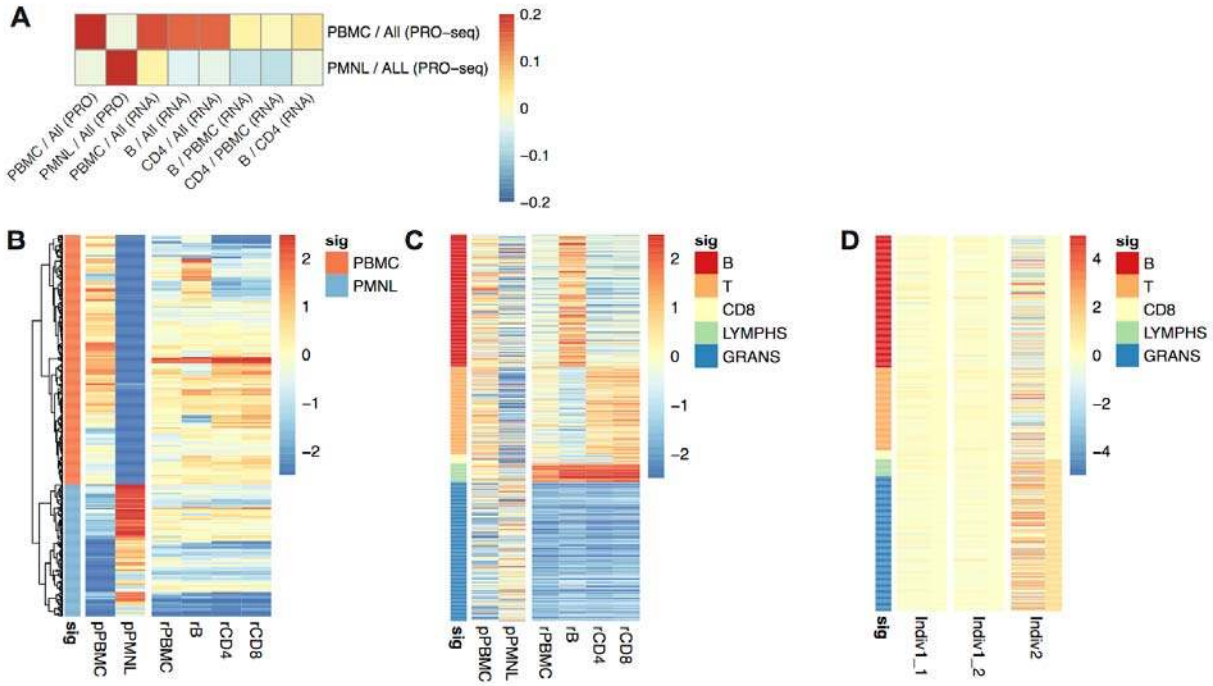


Fig 6



References

1. Fuda, N. J., Ardehali, M. B., & Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, *461*(7261), 186–192. doi: 10.1038/nature08449
2. Adelman, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics*, *13*(10), 720–731. doi: 10.1038/nrg3293
3. Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320. doi: 10.1038/ng.3142
4. Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, *16*(3), 144–154. doi: 10.1038/nrm3949
5. Vahedi, G., Kanno, Y., Furumoto, Y., Jiang, K., Parker, S. C. J., Erdos, M. R., … O'Shea, J. J. (2015). Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature*, *520*(7548), 558–562. doi: 10.1038/nature14154
6. Weber, C. M., Ramachandran, S., & Henikoff, S. (2014). Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase. *Molecular Cell*, *53*(5), 819–830. doi: 10.1016/j.molcel.2014.02.014
7. Churchman, L. S., & Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, *469*(7330), 368–373. doi: 10.1038/nature09652
8. Larson, M. H., Mooney, R. A., Peters, J. M., Windgassen, T., Nayak, D., Gross, C. A., … Weissman, J. S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, *344*(6187), 1042–1047. doi: 10.1126/science.1251871
9. Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., … Proudfoot, N. J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, *161*(3), 526–540. doi: 10.1016/j.cell.2015.03.027
10. Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, *322*(5909), 1845–1848. doi: 10.1126/science.1162228

11. Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science*, *339*(6122), 950–953. doi: 10.1126/science.1229386

12. Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., … Lis, J. T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature Protocols*, *11*(8), 1455–1476. doi: 10.1038/nprot.2016.086

13. Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., … Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. *Nature Methods*, *12*(5), 433–438. doi: 10.1038/nmeth.3329

14. Chu, T., Rice, E. J., Booth, G. T., Salamanca, H. H., Wang, Z., Core, L. J., … Danko, C. G. (2018). Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nature Genetics*, *50*(11), 1553–1564. doi: 10.1038/s41588-018-0244-3

15. Kristjánsdóttir, K., Kwak, Y., Tippens, N. D., Lis, J. T., Kang, H. M., & Kwak, H. (2018). Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. doi: 10.1101/426908

16. Kotecki, M. (1999). Isolation and Characterization of a Near-Haploid Human Cell Line. *Experimental Cell Research*, *252*(2), 273–280. doi: 10.1006/excr.1999.4656

17. Carette, J. E., Guimaraes, C. P., Varadarajan, M., Park, A. S., Wuethrich, I., Godarova, A., … Brummelkamp, T. R. (2009). Haploid Genetic Screens in Human Cells Identify Host Factors Used by Pathogens. *Science*, *326*(5957), 1231–1235. doi: 10.1126/science.1178955

18. Gowen, B. G., Chim, B., Marceau, C. D., Greene, T. T., Burr, P., Gonzalez, J. R., … Raulet, D. H. (2015). A forward genetic screen reveals novel independent regulators of ULBP1, an activating ligand for natural killer cells. *ELife*, *4*. doi: 10.7554/elife.08474

19. Woo, Y. M., Kwak, Y., Namkoong, S., Kristjánsdóttir, K., Lee, S. H., Lee, J. H., & Kwak, H. (2018). TED-Seq Identifies the Dynamics of Poly(A) Length during ER Stress. *Cell Reports*, *24*(13). doi: 10.1016/j.celrep.2018.08.084

20. Cheng JT, Hsu HL, Hwang LY, Baer R. (1993). Products of the TAL1 oncogene: basic helix-loop-helix proteins phosphorylated at serine residues. *Oncogene*. 1993;8: 677-683.

21. Core, L. J., Waterfall, J. J., Gilchrist, D. A., Fargo, D. C., Kwak, H., Adelman, K., & Lis, J. T. (2012). Defining the Status of RNA Polymerase at Promoters. *Cell Reports*, *2*(4), 1025–1035. doi: 10.1016/j.celrep.2012.08.034

22. Leroy, H., Roumier, C., Huyghe, P., Biggio, V., Fenaux, P., & Preudhomme, C. (2005). CEBPA point mutations in hematological malignancies. *Leukemia*, *19*(3), 329–334. doi: 10.1038/sj.leu.2403614

23. The Hd Ipsc Consortium. (2012). Induced Pluripotent Stem Cells from Patients with Huntingtons Disease Show CAG-Repeat-Expansion-Associated Phenotypes. *Cell Stem Cell*, *11*(2), 264–278. doi: 10.1016/j.stem.2012.04.027

24. Kim, J., Sif, S., Jones, B., Jackson, A., Koipally, J., Heller, E., … Georgopoulos, K. (1999). Ikaros DNA-Binding Proteins Direct Formation of Chromatin Remodeling Complexes in Lymphocytes. *Immunity*, *10*(3), 345–355. doi: 10.1016/s1074-7613(00)80034-5

25. Teetson, W., Cartwright, C., Dreiling, B. J., & Steinberg, M. H. (1983). The Leukocyte Composition of Peripheral Blood Buffy Coat. *American Journal of Clinical Pathology*, *79*(4), 500–501. doi: 10.1093/ajcp/79.4.500

26. Palmer, C., Diehn, M., Alizadeh, A. A., & Brown, P. O. (2006). Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics*, *7*(1), 115. doi: 10.1186/1471-2164-7-115

27. Schmidt, W. (1999). CapSelect: a highly sensitive method for 5 CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Research*, *27*(21). doi: 10.1093/nar/27.21.e31

28. Zhu, Y., Machleder, E., Chenchik, A., Li, R., & Siebert, P. (2001). Reverse Transcriptase Template Switching: A SMART™ Approach for Full-Length cDNA Library Construction. *BioTechniques*, *30*(4), 892–897. doi: 10.2144/01304pf02

29. Tome, J. M., Tippens, N. D., & Lis, J. T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nature Genetics*, *50*(11), 1533–1541. doi: 10.1038/s41588-018-0234-5

30. Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., … Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, *500*(7461), 207–211. doi: 10.1038/nature12064

31. Zhang, H., Zheng, R., Wang, Y., Zhang, Y., Hong, P., Fang, Y., … Fang, Y. (2019). The effects of Arabidopsis genome duplication on the chromatin organization and transcriptional regulation. *Nucleic Acids Research*, *47*(15), 7857–7869. doi: 10.1093/nar/gkz511

32. Zhu, Y., Gong, K., Denholtz, M., Chandra, V., Kamps, M. P., Alber, F., & Murre, C. (2017). Comprehensive characterization of neutrophil genome topology. *Genes & Development*, *31*(2), 141–153. doi: 10.1101/gad.293910.116

33. Azzalin, C. M., Reichenbach, P., Khoriauli, L., Giulotto, E., & Lingner, J. (2007). Telomeric Repeat Containing RNA and RNA Surveillance Factors at Mammalian Chromosome Ends. *Science*, *318*(5851), 798–801. doi: 10.1126/science.1147182

34. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). doi: 10.1186/s13059-014-0550-8

35. Mi, H., Muruganujan, A., Ebert, D., Huang, X., Thomas, P.D. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research, 47(D1). Doi: 10.1093/nar/gky1038