A biological-computational human cell lineage discovery platform based on duplex molecular inversion probes

Liming Tao[1#], Ofir Raz[1#], Zipora Marx[1], Tamir Biezuner[1], Shiran Amir[1], Lilach Milo[1], Rivka Adar[1], Amos Onn[1], Noa Chapal-Ilani[1], Veronika Berman[1], Ron Levy[1] , Barak Oron[1], Ruth Halaban[2], Ehud Shapiro[1*]

1. Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 761001, Israel

2. Department of Dermatology, Yale University School of Medicine, New Haven, Connecticut 06520-8059, USA

* To whom correspondence should be addressed. Tel: +972-8-934-4506; Fax: +972-8-934-1746; Email: ehud.shapiro@weizmann.ac.il

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors".

ABSTRACT

Short Tandem Repeats (STRs) are highly mutable genomic elements composed of repetitive short motifs, widely distributed within the human genome, and as such are the most promising source for somatic genomic variations. We present an affordable and scalable cell lineage reconstruction platform that combines customizable duplex Molecular Inversion Probes (MIPs), high throughput targeted sequencing and tailored analysis, all integrated in a bioinformatics Database Management System (DBMS). By applying this platform to a benchmark of *ex vivo* lineage samples, we demonstrate efficient acquisition of tens of thousands of targets in single-cell whole-genome amplified DNA and the discovery of lineage relations among these cells with superior accuracy. We then reconstruct a cell lineage tree of healthy and metastatic cells from a melanoma patient, supporting the hypothesis of clonal metastases and demonstrating that a naïve panel targeting STR somatic mutations in single cells can outperform a cancer specific SNP panel in reconstruction accuracy.

INTRODUCTION

Cell lineage aims to uncover the developmental history of organism cells back to their cell of origin. Many fundamental open questions in human biology and medicine can be answered if the structure and dynamics of the human cell lineage tree could be uncovered(Shapiro et al. 2013). Yet *Caenorhabditis elegans* is still the only organism with a known cell lineage tree, derived from visual observation of its developmental process (Sulston and Horvitz 1977; Sulston et al. 1983).

Recently, the potential of CRISPR-Cas9 genome-editing based cell lineage systems have been demonstrated on complex organisms such as *C. elegans* and zebrafish (Frieda et al. 2017; Kalhor et al.

2017; Schmidt et al. 2017). Human cell lineages, on the other hand, could only be reconstructed based on somatic mutations that occur naturally during cell divisions, such as L1 retro transposition event, Copy Number Variants (CNV), Single Nucleotide Variant (SNV) and Short Tandem repeats (STRs) (Frumkin et al. 2005; Frumkin et al. 2008; Wasserstrom et al. 2008; Evrony et al. 2012; Behjati et al. 2014; Evrony et al. 2015; Lodato et al. 2015; Mann et al. 2016). STRs, also known as microsatellites, are among the largest contributors of *de novo* mutations and are highly abundant (Woodworth et al. 2017). The STR mutation rate corresponds to known features in the human reference genome such as the type of the repeating unit and number of repeats (Willems et al. 2014). These aspects make STRs a promising mutational source, so that by targeting a selection of the most mutable loci, the human cell lineage can be unraveled.

Existing methods for accurate genotyping of STRs remain limited in their throughput and single cell (SC) compatibility (Supplemental Table1). The most suitable method for cell lineage reconstruction to date, previously developed in our lab, uses an Access Array (AA, Fluidigm) based, ~2,000 multiplexed PCRs paired with Illumina Next Generation Sequencing (NGS). However, it suffers from expensive initial setup that limits its scalability. Further scaling would also require much larger multiplex groups, which are prone to failure; or more AA chips, which are costly (Biezuner et al. 2016).

Molecular Inversion Probes (MIPs, also termed Padlock Probes) are single strand DNA molecules composed of two targeting arms and a linker between the two arms and have the potential for much higher targeting throughput and better specificity compared to PCR multiplexing. The concept of MIPs was initially published by Ulf Landegren's group(Nilsson et al. 1994). Since then, this technology has been developed in two directions: *in situ* molecular detections, including specific RNAs, pathogenic somatic mutations *etc*(Larsson et al. 2010; Ke et al. 2013; Schneider and Meier 2017) and targeted enrichment of SNVs or longer targets. Shen *et al*.(Shen et al. 2011; Shen et al. 2013) have developed single strand long MIPs (~325 bp) and succeeded in capturing 500~600bp targets in exons; they further developed the first duplex MIPs pipeline which simplified the process of MIPs creation. The limitation of this procedure is the necessity to build the MIPs one by one for each target, which makes the procedure expensive and time consuming for large panels. Recent improvements in error-rate and quantity of oligo-pool synthesis enabled the parallel synthesis of oligonucleotides longer than 150nt in vast amounts (Kosuri and Church 2014). Yoon *et al.* described short duplex MIPs generated by microarray for SNP targeting in exons (Yoon et al. 2015). Individually synthesized single strand MIPs have been shown to successfully capture tri- and hexa-nucleotide STRs in a scale of 102 loci at once, *A. thaliana*(Carlson et al. 2015). Based on this knowledge, we developed our strategy for high throughput Illumina-NGS-compatible, STR targeting MIPs, which precursors could be synthesized on massively parallel microarray. Using the biochemical pipeline published by Shen *et al*.[25] as a guideline, we developed a duplex-MIPs protocol, paired with an in-house Laboratory Information Management System (LIMS) and STR-aware analysis pipeline. This platform enables to capture tens of thousands of STR loci and conduct tailored analysis with adaptive error correction at high efficiency and low costs.

RESULTS

The workflow of the biological-computational cell lineage discovery platform

Highly mutable STR targets, together with other patient and disease specific areas of interest, were selected for targeting by specific MIPs. MIP precursors were synthesized on a microarray (Custom Array) and processed to final formation. Following a series of biochemical reactions, we finalize with a ready-to-run sequencing libraries. After sequencing, subsequent bioinformatics analysis processed the sequencing data enabling STR-aware mapping for reconstruction of the cell lineage tree. As depicted in Figure1a, the 150bp precursors are composed of a pair of universal adaptors, two-3bp Unique Molecular Identifier (UMI), two target specific arms and one Illumina sequencing compatible spacer. Once synthesized, the duplex MIPs precursors undergo PCR amplification, MlyI digestion, purification and quality control to reach their final duplex MIP structure (Figure1b). Introduction of the duplex MIPs to the template DNA consists of a series of reactions: hybridization, gap filling, ligation, and digestion (Figure1c). Sequencing library is prepared using barcoding PCR on each well in the plate followed by pooling, Illumina NGS run and analysis for cell lineage reconstruction (Figure1c). A sanity check was carried out by clustering single cells from four individuals based on their STR genotypes generated by our platform. The reconstructed lineage tree cluster all the samples into four groups in full agreement with their individual origins and using naïve analysis (Supplemental Figure 25).

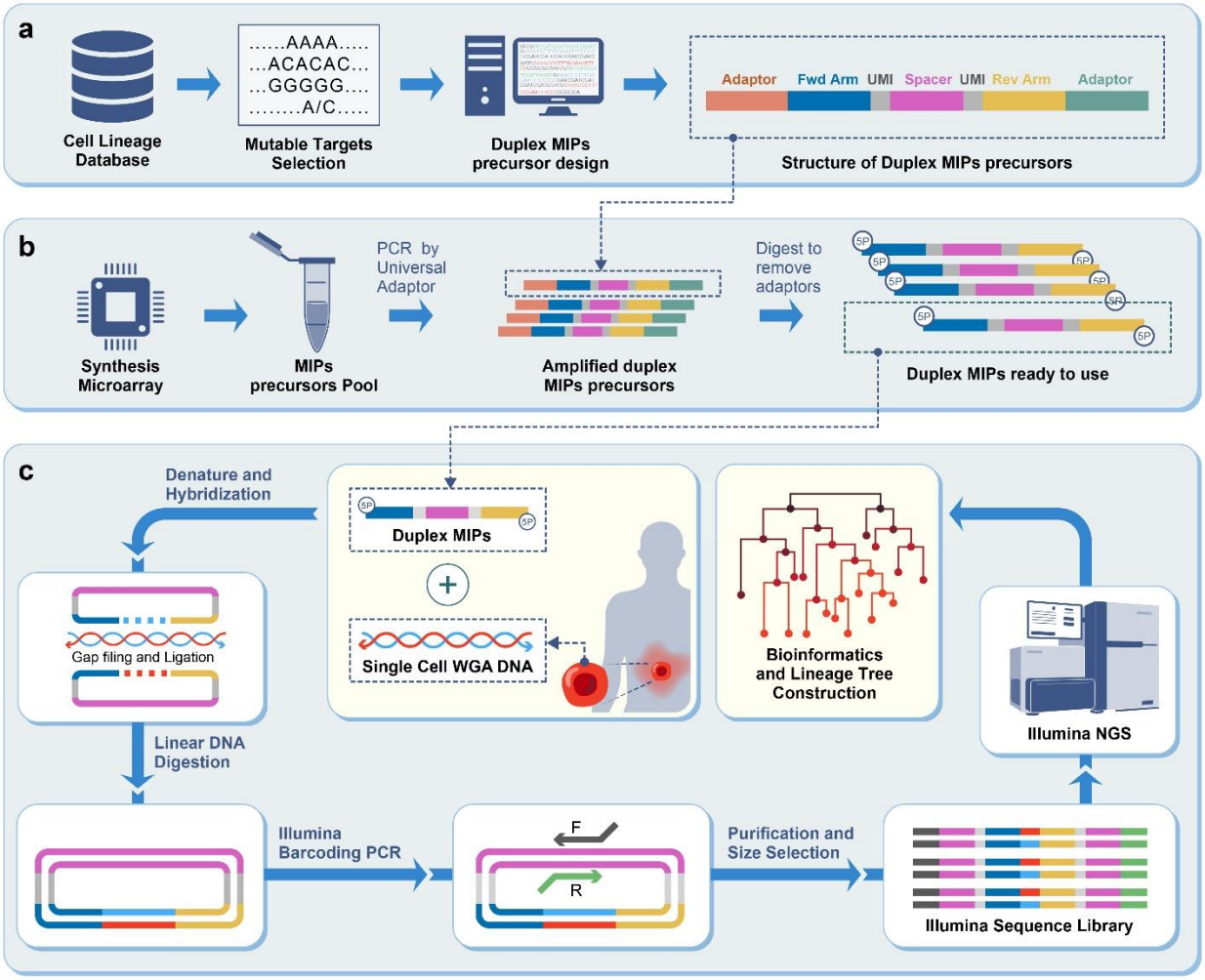

Figure1. Duplex MIPs based cell lineage workflow (a) Design of duplex MIPs precursor: Desired targets are selected from our cell lineage database and precursors are designed. (b) Duplex MIPs preparation: duplex MIPs precursors are synthesized on microarray collected and amplified by PCR as a pool. PCR product is then digested to remove the universal adaptors (red and green); the digested product is purified and diluted to obtain active duplex MIPs. (c) Duplex MIPs and template DNA are mixed together, the targeting arms (blue and yellow) anneal to the flanking regions of the targets and the MIPs are then circularized by gap filling with DNA polymerase and ligase. Linear DNA, including excess MIPs and template DNA, is digested by exonucleases and an Illumina sequencing library is generated by adding adaptors and barcodes using PCR for each sample separately. Libraries are pooled and sequenced by Illumina NGS platform, followed by analysis of the raw reads to detect mutations. This mutation information is then used to infer the cell lineage tree.

Integrated bioinformatics Database Management System (DBMS)

We designed and implemented a scalable architecture of Cell Lineage Discovery Workflow DBMS for collaborative cell lineage discovery. The DBMS supports (i) Data storage and labeling that allows access

to all workflow data, including the data of each donor (anonymized), sample, cell, reagent type and physical location, measurement, sequencing run and analysis steps. (ii) The application of any registered algorithm on any stored data. (iii) Tracking of all workflow protocols, algorithms, sequencing and data in all analysis steps.

The design, as outlined in Figure 2, dissects both the biological and the computational workflow into atomic objects that are documented and referenced in every stored experiment. The full Entity Relation Diagram (ERD) of this cell-lineage database structure is shown in Supplemental Figure1.

Samples are documented using the web based graphical user interface, from individual to SC DNA resolution. Cell IDs are assigned to the documented SCs in the process and spreadsheet reports can be produced and filtered using the webserver. The processes such as target enrichment protocols, library preparation and NGS outputs are documented for all cells (Figure 2a). The NGS raw data (BCL files) is uploaded to the system and demultiplexed according to the documented sample information. Followed by a merger step, the raw reads for each sample can be processed by either of the two STR-aware alignment pipelines developed in-house. The alignment is implemented for parallel execution on a computing cluster using the Dask-Distributed package, resulting in the creation of millions of histograms depicting the STR length distribution for each cell-locus combination. The analysis is designed with fail-safe points at every step so that any underlying malfunction can be recovered, if needed, from prior computation step. Those histograms are then processed by the genotyping module employing our STR amplification stutter model (BioRxiv: https://www.biorxiv.org/content/early/2016/07/21/065110). Individual genotyping results are aggregated for haplotyping, harnessing a population wide perspective to discern the two underlying alleles for each locus (Figure 2c).

Bi-allelic cases are treated as two distinct mono-allelic loci, allowing for integration with unaware phylogenetic reconstruction tools. The system accommodates multiple distance based phylogenetic reconstruction algorithms as well as other hierarchical clustering approaches such as triplets based reconstruction(Sevillya et al. 2016). Reconstructed phylogenetic trees can be further processed by three in-house adaptations of plot tools as well as multiple methods for clustering significance assessment.

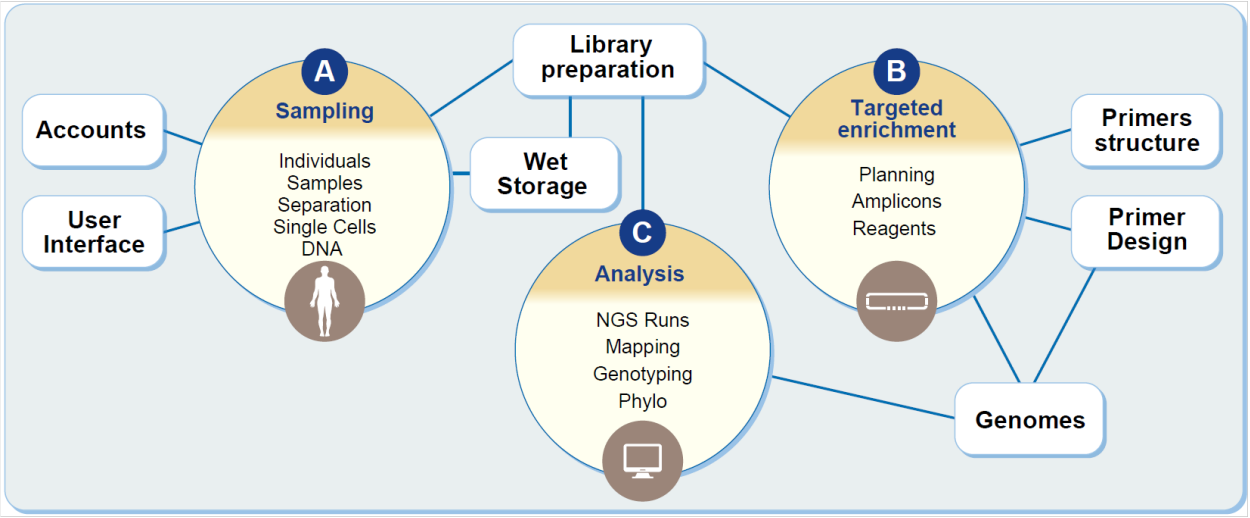

Figure 2. Outline of the Workflow Database and Management System | a. Sampling - sampling documentation from patient to DNA, paired with User Interface for viewing, searching and documenting sampling components. b. Targeted Enrichment – documenting target selection and probe design. c. Analysis – steps from NGS raw data to Tree across multiple tools, versions and parameters. Paired with the Dask-Distributed package for computing clusters.

Utilizing duplex MIPs protocol to accurately reconstruct an ex vivo cell lineage tree

In order to provide a proof-of-concept of the duplex MIPs protocol, we measured the reconstruction accuracies of both the AA based platform(Biezuner et al. 2016) and the duplex MIPs platform (detailed in online methods) in the context of the previously developed DU145 *ex vivo* benchmark tree(Biezuner et al. 2016). In short, a SC from the DU145 human male prostate cancer cell line has been cultured to generate a nine clonal generations DU145 *ex vivo* tree. Each generation is composed of 12 to 15 cell divisions. SCs were sampled from all the generations, and were annotated with their coordinates in the lineage tree. WGA products from SCs were prepared and served as a benchmark for multiple cell lineage reconstruction targeted enrichment protocols. For the duplex MIPs protocol, we used duplex MIPs panel “OM6” (Supplemental File1). We demonstrate an improved lineage reconstruction accuracy for the duplex MIPs protocol over AA (Figure 3e, f).

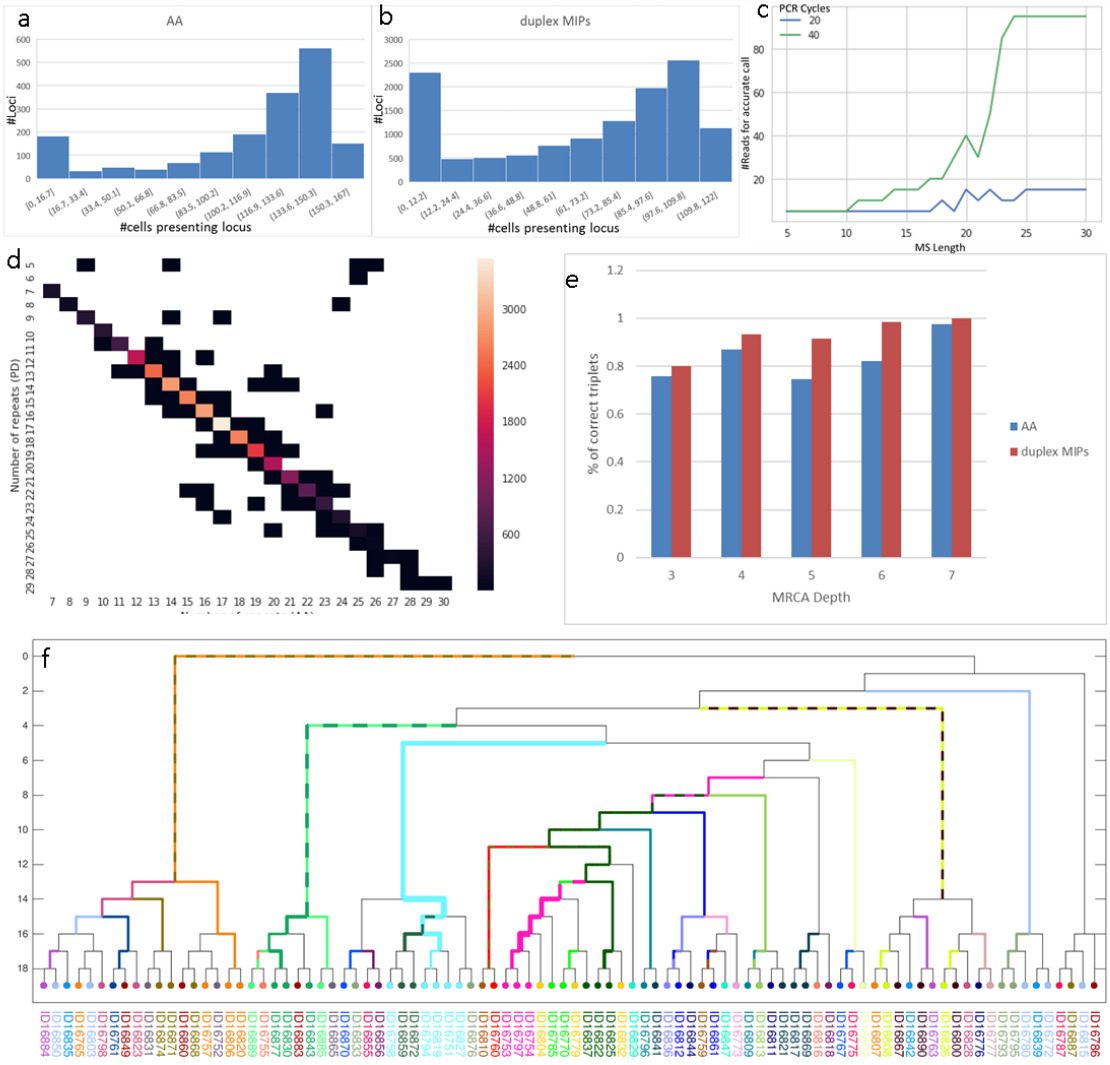

Figure 3. DU145 ex vivo performance comparison between Duplex MIPS pipeline and Access Array pipeline. (a, b) STR coverage comparison between Access Array and Duplex MIPs. For each locus in the attempted panel, we counted the number of cells, which presented in it with a coverage of over 30X. (c) Simulation analysis for the minimal number of reads required for accurate (less than 1 mistake in 1000 attempts) genotyping of AC microsatellite given two PCR steps (AA protocol, estimated 40 amplification cycles) and a single PCR step (duplex MIPs protocol, estimated 20 amplification cycles). (d) Genotyping correlation, AA vs Duplex MIPs. (e) Percentage of correct triplets as a function of the distance between the two Most Recent Common Ancestor (MRCA) within the triplet (higher is better). (f) The ex vivo lineage tree as reconstructed by duplex MIPs.

Cell lineage reconstruction of a melanoma patient (YUCLAT)

To demonstrate the potential of duplex MIPs SC lineage reconstruction of real human clinical case, SCs were obtained from a melanoma patient described as YUCLAT in Krauthammer *et al*(Krauthammer et al. 2015). In total, five different groups of SCs were collected: normal peripheral blood lymphocytes, metastasis groups 1, 2, 3 collected from the scapula and metastasis group 4 from the axilla. The cells were processed using our duplex MIPs platform using duplex MIPs panel “OM7” (Figure 4, Supplemental File2). The reconstructed STR lineage tree demonstrates an effective *in vivo* separation, validating both the expected grouping suggested by the samples origin as well as the clonality of melanoma metastases (Figure 4a,b). The PBL cell group, metastasis groups 2, 3, and 4 are separated using only STR signatures. Additional validation comes from the SNP based reconstruction (Figure 4c,d) that displays similar clustering of the PBL group and metastasis 4 group. In STR lineage tree, normal PBL cell group is clustered separately from all cancer groups, all metastasis groups are clustered in correspondence with their spatial origin, except for metastasis group 1, which has only three cells in this experiment. The four outliers in the STR based clustering are also outliers in SNP clusters. This reconstruction was performed without filtering any samples or loci.

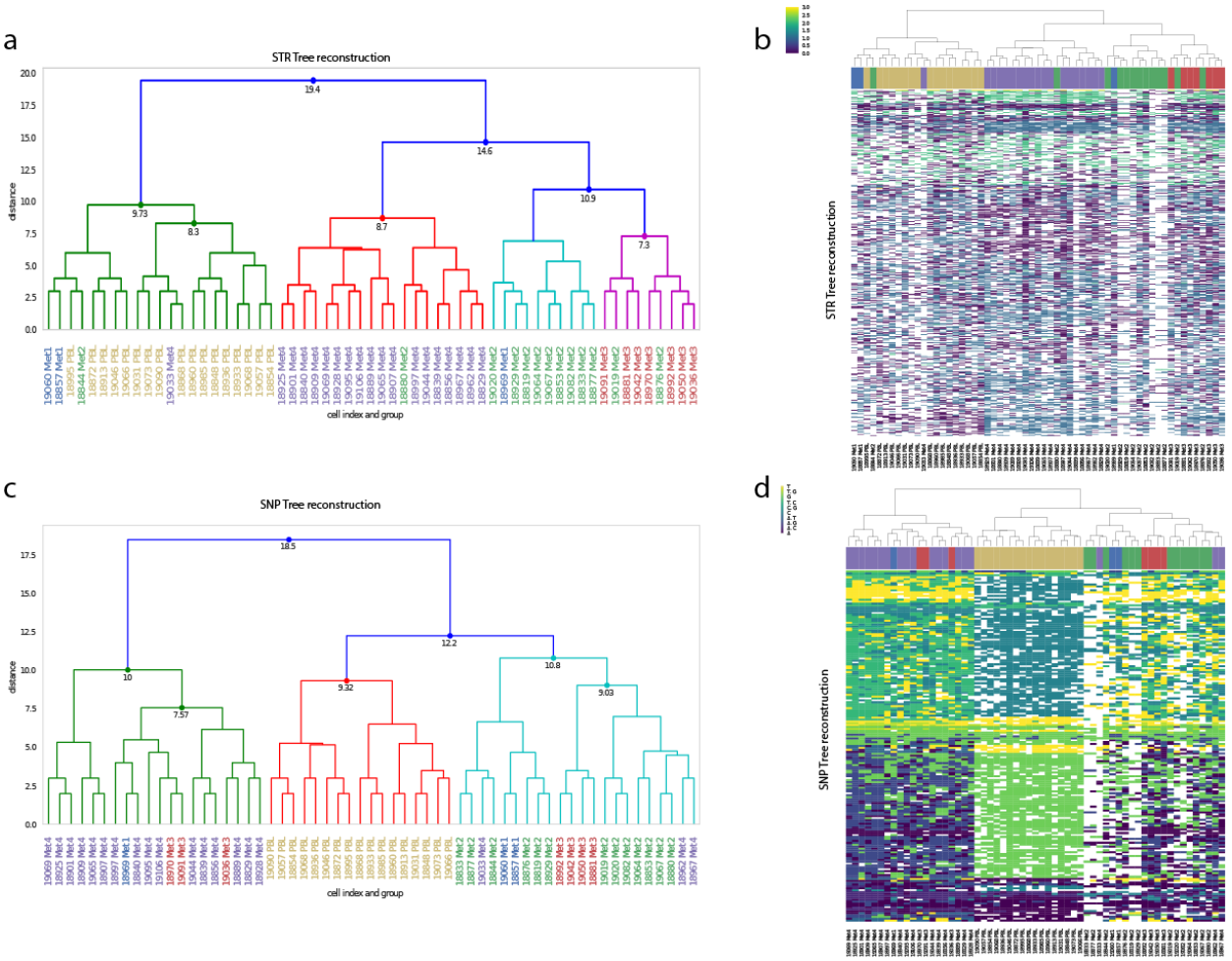

Figure 4. Cell lineage reconstruction of melanoma metastases and peripheral blood lymphocytes (PBL) from the same patient (YUCLAT) | (a) STR based reconstruction dendrogram of YUCLAT (b) Combined heatmap and dendrogram of YUCLAT based on STR (same dendrogram as a), (c) SNP based reconstruction of YUCLAT (d) Combined heatmap and dendrogram of YUCLAT based on SNP (same dendrogram as c). Sample groups color index: blue represents Met1, green for Met2, red for Met3, purple for Met4 and ochre for the PBL group.

Comparison between duplex MIPs pipeline and Access Array pipeline

The cost between duplex MIPs pipeline and AA pipeline differs in two main aspects: the initial, one-time, synthesis of MIP precursors (Figure 1) or primers; and the per-reaction consumables such as reagents or microfluidics chip. While the initial synthesis cost remains affordable for both systems at the 2000 loci scale, when scaling up to 100,000 loci, the cost of primer synthesis for AA increases to roughly 1.8M$, while the cost of duplex MIPs synthesis remains reasonable, at around 12K$. Once synthesized, due to the ability to PCR amplify the precursors, sufficient duplex MIPs can be produced for millions of cells. The cost of reagents and consumables in targeting reactions does not change much as the duplex MIPs pipeline's scales; but for the AA pipeline, several chips are required to capture more loci, increasing the cost

significantly (Supplemental Figure2). In total, the cost per locus per cell in duplex MIPs platform is reduced as much as ~8 times (Supplemental Table2&3).

The duplex MIPs pipeline uses a single PCR phase rather than two in the AA pipeline, resulting in 15 to 20 less cycles as inferred by the STR genotyping algorithm (Supplemental Figure3). This translates to less artificial noise introduced into the STR repeats during the target enrichment process, which in turn allows a reduction in sequencing costs by enabling confident genotyping with lower sequencing coverage (Figure 3c). Thus, STR genotyping was performed using 5X as the minimal coverage threshold in duplex MIPs pipeline, compared 30X for AA pipeline. The STR genotyping results of the same SC WGA DNA produced by duplex MIPs pipeline were compared with their genotypes reported by Biezuner *et al*(Biezuner et al. 2016). We found that the genotypes from two pipelines matched in 99.2% of the cases (Figure 3d).

Sequencing coverage distribution across panels was also very similar between the AA and duplex MIPs pipelines. For each locus in the attempted panel (AA or Duplex MIPS, Figures 3a and 3b, respectively), the cells in which it was found with a coverage of at least 30X were counted. Since the cells in both experiments are the same, similar distributions of sequencing depths indicate similar loci retrieving performance, i.e. the dropout ratio, of duplex MIPs is not impaired by the ~10X scale up in the number of targets.

DISCUSSION

Several protocols for STR genotyping or target sequencing were developed (Supplemental Table1), most of them designed for bulk DNA samples. However, a scalable and affordable high throughput method for SC lineage reconstruction from WGA DNA is still lacking. During the past decade, two STR target enrichment protocols for SC whole genome amplified DNA were developed in our lab. The first was targeting 128 STRs in 4X multiplex PCR, genotyping was analyzed by fragment length distribution using capillary electrophoresis(CE)(Frumkin et al. 2005). To overcome the low throughput property of CE, the second protocol was developed with up to 50X multiplex PCR and subsequent genotyping using NGS sequencing(Biezuner et al. 2016). This protocol could target over 2000 STRs for 48 samples on one Access Array (AA) chip. Yet this protocol was too expensive to scale up for more than100,000 STRs due to the accumulating cost of primer synthesis and AA chips(Biezuner et al. 2016). Stemming from a published MIP protocol designed for complex targets[25], we developed a cost efficient pipeline for STRs target sequencing based on duplex MIPs and demonstrated it in scales of 12K and 50K targets (Supplemental Table4). Since the duplex MIPs could be generated efficiently by microarray based synthesis for as little as 2,200$ for 12,000 probes or 11,000$ for 100,000 probes, while maintaining a fixed capture reaction cost with the increment of targeting panel, the cost and scalability of STR target enrichment is significantly improved (Supplemental Figure2).

Furthermore, the amplification cycles during sequencing library construction were reduced by 15~20 cycles compared to the AA pipeline. This allows for accurate STR genotyping using as little as 5X reads.

Currently, with a 12K panel, 150~200 cells can be sequenced in one NextSeq run as demonstrated with the *ex vivo* lineage reconstruction benchmark (Figure 3). We also demonstrated the feasibility to combine two independent duplex MIPs panels as one panel working in one single reaction (Supplemental Figure4). This allows us to design, purchase and test our panels incrementally and flexibly. Thus, customized panels could be created and optimized by combining several of them. Six base pairs of Unique Molecular Identifier(UMI) are implemented in the duplex MIPs structure, creating 4096 different UMI combinations that can be used to detect capture events for current and future applications (Supplemental Figure5). We note that due to the low coverage requirements of our pipeline (minimum of 5 reads) collapsing reads by their UMI content can be neglected. The whole workflow takes around 5 days to run, from hybridization to data analysis, with roughly 3-hour hands on time (Supplemental Figure6; detailed description of the duplex MIPs protocol in the online methods section). With more biochemical calibrations, bioinformatics improvements in the duplex MIP arm design process and sequencing analysis, future plans are to fit more SCs into a single sequencing run, and to increase the targets scale to 100K targets, allowing for cell lineage reconstruction of more challenging cells with low mutation rate (Biezuner et al. 2016).

As previously reported by Biezuner *et al*(Biezuner et al. 2016), the DU145 cell line used in the *ex vivo* experiment carries various chromosomal aberrations, while the X chromosome remains mostly haploid for most loci, other chromosome were found to be entirely or partially polyploidy. While the AA platform was designed mostly for targeting loci on the X chromosome, the duplex MIPs platform is distributed more evenly in terms of loci selection across chromosomes. As an outcome, the scale advantage of duplex MIPs is mostly lost on the *ex vivo* benchmark as recently duplicated genomic regions are often highly similar and impossible to haplotype following WGA. Despite the significant loss of genomic loci for this reason, superior reconstruction resolution was demonstrated using the duplex MIPs system comparing to AA.

Another case for comparison is the YUCLAT melanoma cancer patient (Figure4). While previously a separation of Metastasis group number 4 from healthy PBL was demonstrated(Biezuner et al. 2016), here we showed much finer separation among the different metastasis groups on top of the separation of the PBL group. Metastasis groups 1,2,3 were collected from scapula metastases and metastasis 4 was collected from the axilla. This topology is evident in the resulting cell lineage tree, supporting the clonality of melanoma metastases and the spatial progression of the disease. Part of the duplex MIPs panel applied on the YUCLAT samples targets cancer specific mutations and hotspots, providing an independent source of variability that supports the STR based results.

As the reconstruction of a perfect lineage tree remains a grand challenge, several issues could be addressed in the near future. The artificial noise introduced during whole genome amplification can be reduced by: (1) nano liter scale reactions such as WGA on droplet or Fluidigm chip(Gole et al. 2013; Fu et al. 2015), (2) reduced amplification time and (3) lower cell lysis temperature(Dong et al. 2017). The artificial noise introduced in the target enrichment process can be reduced by use less barcoding PCR cycles. An improved SC WGA kits can help increase the uniformity of amplification across the genome. Several steps in the biochemical pipeline could also be further improved such as the duplex MIPs/Template ratio,

hybridization temperature for different GC content targets, extra purification steps to remove closed MIPs and dimers, bead enrichment, *etc*. More mutable mono-repeat STRs can be included in the panel to improve performance in cases with fewer cell divisions and lower mutation rates. Duplex MIPs targeting joint SNV-STRs regions within one amplicon can be designed to aid bi-allelic STR genotyping, and improved haplotyping strategies can be implemented to harness this additional information(Willems et al. 2017). A more careful loci selection might also allow higher success rates, thus reducing the minimal sequencing coverage. The STR selection strategy can also improve by filtering out close bi-allelic loci like (AC)X14, (AC)X15. To do this, a more accurate STR annotated human reference genome may be needed(Willems et al. 2014).

In summary, we have developed an easily initiated, scalable, cost effective cell lineage discovery platform integrated in a bioinformatics Database Management System. This platform features efficient synthesis of duplex MIPs based high throughput targeting sequencing technologies, adaptive error correction, tailored sequencing analysis and lineage reconstruction modules. It supports quick development iterations with customizable targets integration including STRs, SNVs, *etc*. By applying this platform to various types of human cells, we demonstrated that tens-of-thousands hyper-mutable STR targets in SC whole genome amplified DNA could be efficiently acquired and lineage relations among these cells can be discovered.

With the advancement of SC analysis methodologies, the cost of human cell lineage tracing can be further reduced and the accuracy and resolution can be improved. The cell states, SC spatial information and phenotypes can be integrated into the cell lineage tree. After exporting the duplex MIPs protocol to multiple labs worldwide, we are planning to integrate our databases and analysis pipelines into the Human Cell Atlas cloud-based analysis platform. This will facilitate our collaborators to conduct their own experiments with our protocol and analyze their data efficiently. Together with other SC methods, our platform will help study the development of human in both health and disease status.

MATERIAL AND METHODS

Melanoma patient single cells WGA DNA

YUCLAT metastatic melanoma cells and Peripheral Blood Lymphocytes (PBL) were collected from a 64-yr-old male patient by the Tissue Resource Core of the Yale SPORE in Skin Cancer with the participant's signed informed consent according to Health Insurance Portability and Accountability Act (HIPAA) regulations with a Human Investigative Committee protocol as described.(Krauthammer et al. 2015) Single cell WGA DNA was prepared in our previous work(Biezuner et al. 2016).

DU145 *ex vivo* tree

Briefly, a single DU145 human male prostate cancer cell line was cultured to generate a nine-level ex vivo tree assisted by CellCelector (ALS). Single cells were sampled nine times every 12~ 15 cell divisions. The single cell WGA DNA of this *ex vivo* cells were prepared with a modified RepliG Mini protocol as described previously in our previous work(Biezuner et al. 2016).

KOD hot start real time PCR Mix 5X (KOD 5X MIX)

First, SYBR 1:100 was prepared by mixing 10 µl from stock SYBR green I (Lonza, 10,000X) and 990 µL Dimethyl Sulfoxide (DMSO) (Sigma). The KOD 5X Mix was prepared in a final concentration of 5X KOD Buffer (Merck); 7.5mM $MgSO_4$ (Merck); 1mM dNTP each (Bioline); 0.1U/µl KOD Enzyme (Merck); 0.5X SYBR green I (Lonza).

Database Management System (DBMS)

Our computational workflow can be divided into two parts: (1) Sample documentation from Individual to NGS barcoded sample that later annotates the reconstructed cell lineage trees; (2)Analysis workflow that starts with raw genomic sequencing data and ends with cell lineage trees, including mutation analysis and lineage tree reconstruction (Supplemental Figure10).

All reported coordinates are based on the hg19 genome build. STR loci are identified from the human reference genome using the tool Phobos (http://www.rub.de/ecoevo/cm/cm_phobos.htm) and primers are designed for each STR loci using Primer3**(Untergasser et al. 2012)**. The resulting STR-bearing amplicons are filtered for patterns that can conflict with the biochemical pipeline and prioritized based on their STR repeat unit, repeat number and genomic coordinates. Data is stored server-side in a MariaDB database. Web pages are served by Gunicorn, a Python WSGI HTTP Server for UNIX, through NGINX acting as a reverse proxy. Behind those, a Django webserver retrieves and process the data. The Django web framework is being used both through classical http interfaces, exposing a Dojo based GUI for samples documentation and management (Figure 2b, Supplemental Figure 11-23) and as an ORM in the data analysis processes. Data analysis is further mediated by Dask.distributed (Figure 2c), a lightweight library for distributed computing in Python.

Target specific duplex MIPs design and preparation

Hyper-mutable STRs selection criteria: Several types of STRs were chosen based on the hg19 reverence human genome annotation. AC-type STRs longer than 10 repeats, AG-type STRs longer than 10 repeats, A-type STRs longer than 6 repeats and G-type STR longer than six repeats were selected. SNVs targets were chosen based on cancer related highly mutable regions or known cancer associated regions.

Amplicon criteria: Amplicons contains TTAA sequence were ruled out to fit Ampli1 WGA kit.

Amplicon size was designed around ~150bp.

Primer-3 based python script was used to design the targeting primers. Only top-scored primers were chosen as candidates for duplex MIPs precursor design. From these candidates, several more filters were used in the designing: both forward primer and reverse primer shall be unique across the human genome; precursors with MlyI digestion recognition site shall be ruled out; the total length of the precursors shall not be longer than 150bp, which is the maximum limit of the oligonucleotides synthesis provider.

Duplex MIPs structure:

Adaptors: Mly1_F: 5'TATGAGTGTG**GAGTC**GTTGC3`; Mly1_R:5'GCTTCCTGAT**GAGTC**CGATG3`

FW_PRIMER_SEQUENCE and RV_PRIMER_SEQUENCE are sequence of primers designed by primer3 for each target.

Full structure formula:

5`[Mly1_F]+[FW_PRIMER_SEQUENCE]+[NNN]+[AGATCGGAAGAGCACACGTCTGAACTCTTTCCCTACACGACGCTCTTCCGATCT]+[NNN]+[Reverse Complement(RV_PRIMER_SEQUENCE)]+[Reverse complement(Mly1_R)]-3`

Duplex MIPS precursors sequence example:

5`TATGAGTGTGGAGTCGTTGCTACTTGGTGGCTAATTCAGCAGGNNNAGATCGGAAGAGCACACG
TCTGAACTCTTTCCCTACACGACGCTCTTCCGATCTNNNTTGCAAGCTCCCTCTGAAAAGTTCATC
GGACTCATCAGGAAGC3`

Ex-vivo reconstruction parameters for Figure3

A new mapping strategy was used for duplex MIPs sequencing data which improved computing efficiency. Reads were aligned against a custom reference genome of all possible STR variations in the panel. References sequences for an STR locus is showed as an example (Supplemental Figure 24). AA data(Biezuner et al. 2016) was genotyped with a minimal coverage of 30X reads, a confidence threshold of 0.05 (correlation above 0.95) between the measured histogram and the reported model. Duplex MIPs data was genotyped with a minimal coverage of 5X and the same confidence threshold as AA. In both attempts, low coverage samples and loci were filtered out, leaving the top 75% of the cells and loci as input. In both attempts, the reconstruction was performed using the Neighbor Joining algorithm with the absolute distance function.

YUCLAT reconstruction parameters for Figure4

The data was genotyped with a minimal coverage of 10X for targets on the X chromosome and 30X for non-X loci, a confidence threshold of 0.01 (correlation above 0.99) between the measured histogram and the simulated model. The reconstruction was performed by applying the TMC algorithm on the leaves triplets space(Sevillya et al. 2016).

Duplex MIPs Generation for OM6 and OM7

(A). PreAmp PCR on oligonucleotides pool:

Oligonucleotides pool of OM6 or OM7 (order details in Supplemental File1&2) received from provider (Custom Array, Inc.) was diluted to 1ng/µl according the conc. provided as PCR template. PreAmp PCR primers (OM4_Mly_F:5`GTCTATGAGTGTGGAGTCGTTGC3`;OM4_Mly_R:5`CTAGCTTCCTGATGAGTCCGATG3`) were designed to fit the universal adaptors. 45ul PCR was prepared with a final concentration of 0.2 ng/µl template, 0.3 pmol/µl OM4_Mly_F, 0.3 pmol/µl OM4_Mly_R, 1X KOD MIX. PCR was performed in the LightCycler 480 (LC480, Roche) with 95 ºC, 2 min denature step; 18 cycles of 95 ºC 20 sec, 60 ºC 10 sec, 70ºC 5 sec; then 70 ºC elongation step; keep at 4ºC. PreAmp PCR product was purified by MinElute PCR purification kit (Qiagen), its concentration was measured by Qubit dsDNA HS Assay Kit (Life Technologies). The purified PreAmp PCR product was diluted to 1ng/µl as the template for next step, the production PCR.

(B). Production PCR (48 reactions)

48 reactions of 45ul production PCR in a 96 well plate (Roche) were prepared with a final concentration of 0.2 ng/µl template, 0.3 pmol/µl OM4_Mly_F, 0.3 pmol/µl OM4_Mly_R, 1X KOD MIX. PCR was performed in the LightCycler 480 (LC480, Roche) with 95 ºC, 2 min denature step; 12 cycles of 95 ºC 20 sec, 60 ºC 10 sec, 70ºC 5 sec; then 70 ºC elongation step; keep at 4ºC. Every four wells of PCR product were merged together and purified by one MinElute column according to the manufacturer's protocol. Elution was in 45µl DDW, and all products were pooled; 1µl of the pool was used to check the concentration by NanoDrop spectrophotometers (Thermo Scientific) according to the manufacturer's protocol. The pool was diluted to ~30ng/µl based on measured concentration. 20µl of sample was kept for quality control; the rest was processed to next step.

(C). 84 ul of the diluted DNA from last step was digested in a 100ul reaction with a final concentration of 0.6 U/µl MlyI (NEB);1X NEB Smarter Buffer. The mixture was incubated in Biometra T3 thermal cycler (Biometra, 3100-810-13) at 37 °C overnight (12 hours), then deactivated at 80 °C for 20min and finally kept at 4 °C. Digested DNA was cleaned by MinElute. If more than one PCR was performed, all elution samples were merged into one tube. Concentration was measured using by Qubit dsDNA HS (High Sensitivity) assay kit according to the manufacturer's protocol.

(D). Tape Station size check

The cleaned digested product was the final duplex MIPs. Its size (~105bp) was measured together with an undigested sample from step (C) (~150bp) by Tape Station (Agilent) (Supplemental Figure7).

Optional: If minor peaks appeared after digestion, BluePippin (3% Gel Cassettes, 105bp tight mode) can be used to further purify the product.

(E). Duplex MIPs working solution preparation

Based on length of 105 bp and the measured concentration, the final duplex MIPs was diluted to 5.8ng/µl working solution, equivalent to 80nM (80fmol/µl). Adjusted to 8nM by 1:10 dilution where needed. Working solutions were stored in -20°C freezer.

The list of all major reagents mentioned above is in Supplemental Table5.

Calibration of duplex MIPs pipeline

Three key steps, hybridization, gap-filling, digestion in the MIPs pipeline were calibrated. Hybridization was tested in 2, 4 and 18 hours; gap filling was tested in 1, 2 and 4 hours; and the digestion was tested in 1 and 2 hours. This was performed in an all-by-all fashion, amounting to 18 combinations, each assessed by downstream sequencing. 200 ng HeLa genomic DNA (NEB) and 80nM duplex MIPs (OM6, Supplemental File1) were used for all reactions. Among all the conditions, the protocol with 18 hours hybridization, 4 hours of gap-filling and 1 hour of digestion proved best. Sequenced at a depth of 10~15X, we found that ~83% of the resulting reads successfully mapped to the designed targets and a similar percentage of the 12K unique targets was obtained. Thus, the protocol 18-4-1 was chosen as our standard protocol (Supplemental Table6).

To decide the proper range of library size, several ranges for BluePippin (Sage Science) size selection were chosen for comparison: 240-340bp, 270-310bp, and 300bp based on the designed amplicon size

distribution. The sequencing result of seven single cell whole genome amplified samples and one bulk DNA sample, both 300 and 270-310 selection ranges had slightly better success rate (2~3% more) compared to 240-340 range; But 240-340 size selection range was standing out with more loci captured (8~15%). Therefore, 240-340 size selection range was chosen for later experiments (Supplemental Table7).

Whole Genome Amplification (WGA) is a necessary prior step for most single cell genomics studies, however the stochastic nature of WGA protocols can pose challenges for downstream analysis(Huang et al. 2015). Working under the assumption that limiting concentration of probes can help normalize underlining WGA biases, we set to further calibrate the ratio between duplex MIPs concentration and template DNA amount. Combinations of both variables were explored in logarithmic steps between 0.01ng to 2000 ng template Hela DNA and 0.08ng to 80nM MIPs in this calibration. Samples with less than 8000 reads in total or under 4000 unique loci were regarded as outliers due to low coverage and dropped from the calibration experiment. Concentrations of 8~80nM duplex MIPs probes paired with 250 to 500 ng input DNA were the most robust range of capture efficiency for our pipeline (Supplemental Figure8). Considered the yield of single cell WGA reaction, we decided to use 8nM MIPs and 250-500 ng single cell WGA DNA as the template DNA for our standard protocol. The final calibrated protocol described in details is in the methods section.

Duplex MIPs-based targeted enrichment pipeline

(A). Hybridization

Single cell WGA product concentration is generally 100-200 ng/µl. 200~500 ng single cell WGA DNA (~2µl) was used as template for each reaction. Reaction mix was prepared in a 10ul reaction with final concentration of ~20 ng/µl Single Cell WGA DNA; 8fmol/µl duplex MIPs; 1X Ampligase Buffer; 0.9 M betaine.For a big batch experiment, hybridization mix was prepared based on the above table without DNA according to the sample numbers. 8µl hybridization mix was distributed to a 96-well plate, 2µl DNA or DDW was added to each well and mixed by liquid handling system (Evoware, Tecan) or manually.

The reaction plate was put into a PCR machine with 100°C lid temperature, and then heated to 98°C for 3 minutes, followed by a gradual decrease in temperature of 0.01°C per second to 56°C and incubated at 56°C for 17 hours.

Optional: if your PCR machine could not decrease as slow as 0.01°C/second, alternative strategy could be applied: the reaction plates was heated to 98°C and kept for 3 minutes, decreased by 0.1°C every cycle as slow as possible and was kept 15 second at this temperature. Cycling until 56°C and incubated at 56°C for 17 hours.

(B). Gap filling

The gap filling mix was prepared half an hour before hybridization finished with a final contration of 0.3 mM each dNTP; 2 mM NAD; 1.1 M betaine; 1X Ampligase buffer; 0. 5U/µl Ampligase and 0.08 U/µl Phusion in a total volume of 10 µl and kept at 56°C on the heat block.

The reaction plate was transferred from the PCR machine to a 56°C heat block when the hybridization step finished. 10µl of gap filling mix was added to each well, carefully mixed by pipette, sealed tightly and quickly

and put it back to the PCR machine for 56°C incubation for 4 hour, then 68°C for 20 minutes and 4°C until next step.

Optional: After the gap filling, the reaction plate can be stored at 4°C fridge for up to two days.

(C). Digestion of linear DNA:

The digestion mix with a final concentration of 3.5 U/µl exo I,18 U/µl exo III, 4 U/µl T7, 0.4 U/µl exo exo T, 3 U/µl RecJf, 0.2 U/µl lambda exo was prepared 15 min before gap filling step finished.

The reaction plate was taken off from PCR machine, the cover was carefully removed. 2µl of the digestion mix was added to each well and mixed. The reaction plate was spin down and sealed, then incubated at 37°C for 60 minutes, 80°C for 10 minutes and 95°C for 5 minutes.

Pause Point: the reactions can be stored at -20°C for months after the digestion step.

Sequencing library preparation

(A). Sample specific barcoding PCR

The barcoding PCR was using the same dual-index barcoding primers designed and published by Biezuner *et al.* (Biezuner et al. 2016) with a modification of the PCR enzyme to NEBNext Ultra II Q5. Here just showed the structure of the dual-index Illumina barcoding primers used in the experiments, details of their sequences could be found in the supplemental information by Biezuner *et al.* (Biezuner et al. 2016).

i5-index-primer: AATGATACGGCGACCACCGAGATCTACAC[i5-8bp- index]ACACTCTTTCCCTACACGACGCTCTTCCG;

I7-index-primer: CAAGCAGAAGACGGCATACGAGAT[i7-8bp-index]GTGACTGGAGTTCAGACGTGTGCTCTTCCG;

2µl product from last step was amplified in a 20 µl PCR reaction with a final concentration of 0.5 pmol/µl each unique barcoding pair of dual-index Illumina primers for each sample; 1X NEBNext Ultra II Q5 Master Mix; 0.5X SYBER Green. The PCR is performed in Roch480 with program:

98 ºC, 30 sec denature; 5 cycles of 98 ºC for 10 sec,56 ºC for 30 sec and 65 ºC for 45 sec; then 15 cycles of 98 ºC for 10 sec and 65 ºC for 75 sec; then 65 ºC for 5 min elongation; hold at 4 ºC.

(B). Diagnosis Sequencing Sample pooling and Purification

Barcoding PCR product was cleaned in 96-well plate by 0.8x AMPure XP SPRI magnetic beads (Beckman Coulter) according to manufactory's manual by Tecan liquid handling system, eluted in 40µl DDW. Equal volume of purified samples was pooled by Echo (Echo550, Labcyte). The pool was concentrated by MinElute according to manufactory into 35µl DDW.

(C). Size Selection for Diagnosis Sequencing

3ul of the concentrated pool was kept for later quality control. 30µl of the concentrated pool was used on a 2% V1 cassette BluePippin (Sage Science) with setting range 240-340bp according to manufactory's protocol. The size selected elution was collected and cleaned by MinElute into 15µl DDW. The concentration was measured by Qubit dsDNA HS (High Sensitivity) assay kit. The size distribution of the concentrated pool before and after BluePippin was measured by Tape Station dsDNA chip (Supplemental Figure9). The size selected pool with a single peak around 300bp was used to prepare 12µl of 4nM (4fmol/µl) library for Illumina NGS calculated based on the measured concentration and the average size on Tape Station.

(D). Diagnosis Sequencing

10pM library was sequenced on MiSeq 1M nano or 4M micro flow cell with 151x2 pair-end run parameters according to manufactory manual, the default sequencing primers were used.

(E). Based on the diagnosis sequencing result, the volume for each sample to equalize the reads was calculated to create production sequencing Echo pooling table. According to this table, the purified samples from step (B) were pooled by Echo550 and concentrated by into 35µl DDW. The pool was used to prepare the production-sequencing library similar process in step (C).

(F). Production sequencing

1.8~2.2pM library was sequenced on NextSeq500 flow cell with 151x2 pair-end run parameters according to manufactory manual, the default sequencing primers were used.

(Optional) If the production sequencing did not generate enough reads for some samples, another round of NextSeq could be conducted using the same library to get more reads.

AVAILABILITY

All the codes of cell lineage analysis tools mentioned in the text are freely available under the GNU General Public License v2.0 at https://github.com/ofirr/clineage

ACCESSION NUMBERS

Sequencing data has been deposited with ArrayExpress access number E-MTAB-6411.

SUPPLEMENTARY DATA

Supplementary Data are available at Genome Research online.

ACKNOWLEDGMENTS

Liming Tao is partially supported by VATAT postdoctoral fellowship from Israel's Council for Higher Education Planning and Budgeting Committee. Ehud Shapiro is the incumbent of The Harry Weinrebe Professorial Chair of Computer Science and Biology.

FUNDING

This research was supported by the following foundations: the European Union FP7-ERC-AdG (European Research Council, 233047); The EU-H2020-ERC-AdG (European Research Council, 670535); the Deutsche Forschungsgemeinschaft (DFG, 611042); the Israeli Science Foundation (ISF, P14587); the Israeli Science Foundation-BROAD (ISF, P15439); the National Institutes of Health (VUMC 38347); the

National Cancer Institute Fund (P50 CA121974, R. Halaban, PI); and the Kenneth and Sally Leafman Appelbaum Discovery Fund.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

REFERENCES

Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, Martincorena I, Petljak M, Alexandrov LB, Gundem G et al. 2014. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**: 422-425.

Biezuner T, Spiro A, Raz O, Amir S, Milo L, Adar R, Chapal-Ilani N, Berman V, Fried Y, Ainbinder E et al. 2016. A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res* **26**: 1588-1599.

Carlson KD, Sudmant PH, Press MO, Eichler EE, Shendure J, Queitsch C. 2015. MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Res* **25**: 750-761.

Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, Vijg J. 2017. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods* **14**: 491-493.

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A et al. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**: 483-496.

Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai XY, Yang LX, Haseley P, Lehmann HS, Park PJ et al. 2015. Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. *Neuron* **85**: 49-59.

Frieda KL, Linton JM, Hormoz S, Choi J, Chow KK, Singer ZS, Budde MW, Elowitz MB, Cai L. 2017. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**: 107-111.

Frumkin D, Wasserstrom A, Itzkovitz S, Stern T, Harmelin A, Eilam R, Rechavi G, Shapiro E. 2008. Cell lineage analysis of a mouse tumor. *Cancer Res* **68**: 5924-5931.

Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. 2005. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput Biol* **1**: e50.

Fu Y, Li C, Lu S, Zhou W, Tang F, Xie XS, Huang Y. 2015. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proc Natl Acad Sci U S A* **112**: 11923-11928.

Gole J, Gore A, Richards A, Chiu YJ, Fung HL, Bushman D, Chiang HI, Chun J, Lo YH, Zhang K. 2013. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat Biotechnol* **31**: 1126-+.

Huang L, Ma F, Chapman A, Lu S, Xie XS. 2015. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annual Review of Genomics and Human Genetics* **16**: 79-102.

Kalhor R, Mali P, Church GM. 2017. Rapidly evolving homing CRISPR barcodes. *Nat Methods* **14**: 195-200.

Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wahlby C, Nilsson M. 2013. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* **10**: 857-860.

Kosuri S, Church GM. 2014. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* **11**: 499-507.

Krauthammer M, Kong Y, Bacchiocchi A, Evans P, Pornputtapong N, Wu C, McCusker JP, Ma S, Cheng E, Straub R et al. 2015. Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat Genet* **47**: 996-1002.

Larsson C, Grundberg I, Soderberg O, Nilsson M. 2010. In situ detection and genotyping of individual mRNA molecules. *Nat Methods* **7**: 395-397.

Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Lee S, Chittenden TW, D'Gama AM, Cai XY et al. 2015. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**: 94-98.

Mann KM, Newberg JY, Black MA, Jones DJ, Amaya-Manzanares F, Guzman-Rojas L, Kodama T, Ward JM, Rust AG, van der Weyden L et al. 2016. Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq. *Nat Biotechnol* **34**: 962-972.

Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary BP, Landegren U. 1994. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**: 2085-2088.

Schmidt ST, Zimmerman SM, Wang J, Kim SK, Quake SR. 2017. Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding. *ACS Synth Biol* **6**: 936-942.

Schneider N, Meier M. 2017. Efficient in situ detection of mRNAs using the Chlorella virus DNA ligase for padlock probe ligation. *RNA* **23**: 250-256.

Sevillya G, Frenkel Z, Snir S, Paradis E. 2016. Triplet MaxCut: a new toolkit for rooted supertree. *Methods in Ecology and Evolution* **7**: 1359-1365.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**: 618-630.

Shen P, Wang W, Chi AK, Fan Y, Davis RW, Scharfe C. 2013. Multiplex target capture with double-stranded DNA probes. *Genome Med* **5**: 50.

Shen P, Wang W, Krishnakumar S, Palm C, Chi AK, Enns GM, Davis RW, Speed TP, Mindrinos MN, Scharfe C. 2011. High-quality DNA sequence capture of 524 disease candidate genes. *Proc Natl Acad Sci U S A* **108**: 6549-6554.

Sulston JE, Horvitz HR. 1977. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev Biol* **56**: 110-156.

Sulston JE, Schierenberg E, White JG, Thomson JN. 1983. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**: 64-119.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**: e115.

Wasserstrom A, Adar R, Shefer G, Frumkin D, Itzkovitz S, Stern T, Shur I, Zangi L, Kaplan S, Harmelin A et al. 2008. Reconstruction of cell lineage trees in mice. *PLoS One* **3**: e1939.

Willems T, Gymrek M, Highnam G, Genomes Project C, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res* **24**: 1894-1904.

Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. 2017. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* **14**: 590-592.

Woodworth MB, Girskis KM, Walsh CA. 2017. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nature Reviews Genetics* **18**: 230-244.

Yoon JK, Ahn J, Kim HS, Han SM, Jang H, Lee MG, Lee JH, Bang D. 2015. microDuMIP: target-enrichment technique for microarray-based duplex molecular inversion probes. *Nucleic Acids Res* **43**: e28.