A duplex MIPs-based biological-computational cell lineage discovery platform

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], Ehud Shapiro[1*]

1. Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 761001, Israel
# Equal Contribution.
* Corresponding Author: ehud.shapiro@weizmann.ac.il

[ABSTRACT]

Somatic mutations can reveal cell lineages during development, health, ageing and disease; however, the rare nature of such mutations makes identifying them time and resource consuming. Short Tandem Repeats (STRs) are highly mutable genomic elements composed of repetitive short motifs, widely distributed within the human's genome and as such, they are the most promising source for genomic variance among the cells of an individual. Earlier we presented a single-cell mutation discovery platform utilizing chip-based multiplex PCR targeting thousands of STRs. Here we present a significant improvement over the platform that combines efficient synthesis of duplex Molecular Inversion Probes (MIPs), high throughput targeting sequencing technologies together with tailored analysis and adaptive error correction all within an integrated bioinformatics Database Management System (DBMS). By applying this platform to various types of human cells, we demonstrated efficient acquisition of tens-of-thousands targets in single-cell whole-genome amplified DNA and discovered lineage relations among these cells.

[INTRODUCTION]

Cell lineaging aims to uncover the developmental history of organism cells back to their cell of origin. Caenorhabditis elegans is the only organism with a known cell lineage tree, derived from visual observation of its developmental process[1, 2]. 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[3].

Recently, the potential of CRISPR-Cas9 genome-editing systems based cell lineaging of model organisms has been demonstrated on cell lines and zebrafish[4-7]. Human lineages, on the other hand, could only be reconstructed utilizing somatic mutations that occur naturally during cell divisions, such as L1 retro transposition event, Copy Number Variants (CNV), Single Nucleotide Variant (SNV) and STRs (also termed as microsatellites)[8-15]. Among these, STRs are highly mutable and abundant[16, 17]. These features make STRs a promising mutational source for human cell lineaging.

Massive parallel target enrichment method for accurately genotyping STRs is not quite developed yet (Sup Table1). Our lab has developed a generic cell lineage analysis platform based on ~2,000 multiplexed PCRs targeting enriched pipeline enabled by Access Array (AA, Fluidigm) and Illumina Next Generation Sequencing (NGS) technologies. However, the initial setup is very expensive as synthesis of over ~2000 long PCR primers is required. Although PCR primers are a onetime purchase, scaling it up to tens of thousands of targets, the cost of primer synthesis alone becomes too high to consider. It will also require much larger multiplex groups which are prone to failure or more AA chips which cost also accumulates[18].

To overcome these limitations, we turned to Molecular Inversion Probes (MIPs, also termed Padlock Probes), which have the potential of much higher targeting throughput and better specificity. MIPs are single strand DNA molecules composed of two targeting arms and a linker between the two arms. The concept of MIPs was initially published by Ulf Landegren's group[19]. Since then, this technology was developed in two directions. The first was in situ molecular detections, including detecting specific RNAs, pathogenic somatic mutations etc[20-22]. The second is target enrichment of SNPs or bigger targets like exons. Single strand MIPs targeting SNV and exon built from a mixture of unique oligonucleotides synthesized using massively parallel microarray has been developed[23]. Shen *et al*. have developed single strand long MIPs (~325 bp) and succeeded in targeting 500~600bp targets in exons, they further upgraded to the first duplex MIPs pipeline[24, 25], 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. Recently, Yoon *et al* described short duplex MIPs generated by microarray targeting SNPs in exons[26]. Single strand MIPs built one by one has been shown successfully targeting 102x tri- and hexa-nucleotide STRs in *A. thaliana*[27].

Thus previous work suggests that short duplex MIPs could be synthesized efficiently on a chip[23] and MIPs could detect STRs in a small scale[27]. Based on these publications, we developed our strategy to design high throughput standard-Illumina-NGS-compatible MIPs targeting STR regions, whose precursors could be synthesized on massively parallel microarray. To simplify our design further, we modified based on the biochemical pipeline developed by Shen *et al*. [25] to make our MIPs working in a duplex style. Following several round calibrations, we developed our duplex MIPs protocol. A duplex-MIPs-dedicated, STR-aware computational pipeline together with a Laboratory Information Management System (LIMS) software for

sample, data management, tracking, and production of lineage trees have been developed and integrated into one bioinformatics DBMS. This system enabled us to capture tens of thousands of STRs together with SNPs and conduct tailored analysis with adaptive error correction with high efficiency at a low cost.
In summary, we developed a duplex MIPs-based biological-computational cell lineage discovery platform, which integrated samples and data tracking, targets selection, MIPs design, NGS sequencing data management, tailored analysis and cell lineage reconstruction.

[RESULTS]

The workflow of duplex MIPs-based biological-computational cell lineage discovery platform

From our cell lineage database, highly mutable STR targets and other demanded targets were selected. The specific MIPs precursors were designed. The 150bp precursor is composed of a pair of universal adaptor, two 3bp unique molecular identifier (UMI), two target specific arms, one illumina sequencing compatible spacer (Figure1A). 4^6=4096 different UMI combinations can be used to detect capture events (Supplemental Figure2).

The MIPs precursors were synthesized on a microarray (Custom Array). Duplex MIPs were created from the synthesized oligoes by PCR amplification and MlyI digestion. As a quality control step of the duplex MIPs preparation, the sizes of precursors and duplex MIPs were both measured by Tape Station (Supplemental Figure1)**.** Purified by MinElute (Qiagen), the duplex MIPs were diluted to a working conc. 80nM or 8nM (Figure1B).

Template DNA/WGA product were mixed together with duplex MIPs in separate wells. Then, a series of molecular biological reactions were applied: hybridization, gap filling, ligation, and digestion. Then, barcoding PCR was applied to each well using unique Illumina barcoding primers, to generate a sequencing library. Libraries were pooled and sequenced by the Illumina NGS platform; the raw reads were analyzed by our bioinformatics pipeline to detect mutations. The processes and sequencing results were documented into the cell lineage database. Somatic mutations were used to infer the cell lineage tree (Figure 1C). The whole workflow took around 5 days from hybridization to data analysis, with roughly 3-hour hands on time (Supplemental Figure3; detailed description of the duplex MIPs protocol in the methods section).

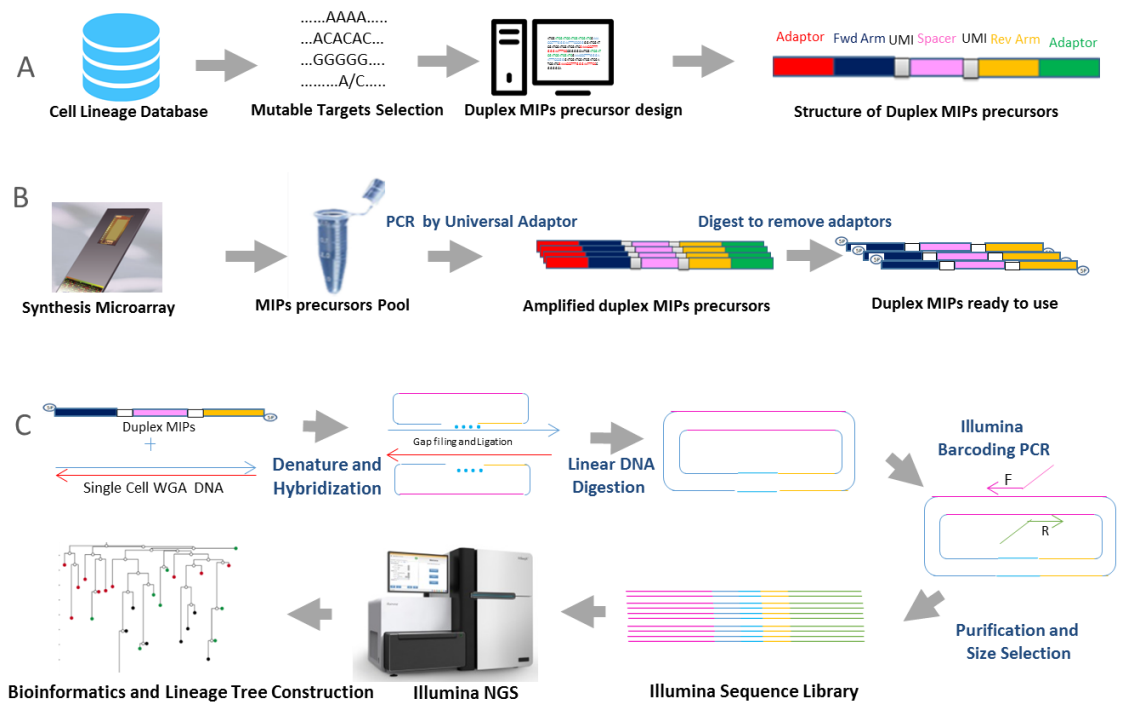

Figure1. Duplex MIPs based cell lineage platform workflow (A) Design of duplex MIPs precursor: Desired targets were selected from our cell lineage database; precursors were designed in the structure shown. (B) Duplex MIPs preparation: duplex MIPs precursors were synthesized on microarray and collected; then amplified by PCR as a pool with universal primers. PCR product was then digested to remove universal adaptors (red and green colors); the digested product was purified and diluted to obtain duplex MIPs that ready for use. (C) Duplex MIPs and template DNA were mixed. The targeting arms (blue and yellow) were annealed to the flanking regions of the targets. The MIPs were circularized by gap filling with DNA polymerase and ligase; linear DNA, including excess MIPs and template DNA, was removed by exonucleases. An Illumina sequencing library was generated by adding adaptors and barcodes using PCR, for each sample separately. Libraries were pooled and sequenced by the Illumina NGS platform and raw reads were analyzed to detect mutations. Mutation information was then used to infer the cell lineage tree.

Calibration of duplex MIPs pipeline

Three major steps, hybridization, gap-filling, digestion in the MIPs capture pipeline were calibrated in 18 different conditions. Hybridization was tested in 2, 4 and 18 hours; gap filling was tested in 1, 2 and 4 hours; while the digestion was tested in 1 and 2 hours. In the calibration experiment, 200 ng HeLa genomic DNA (NEB) and 80nM duplex MIPs were used for all reactions (corresponding to a ratio ~70:1 duplex MIPs to DNA template). The sequencing results of these conditions were used to evaluate the protocol. Among all the conditions, the protocol 18-4-1 showed the best success rate, ~83.5% and most loci captured, over 10,000. Thus, the protocol 18-4-1 was chosen as our standard protocol (Supplemental Table3).

To decide the proper range of library size, several ranges for BluePippin (Sage Science) size selection were chosen for comparison: 240-340bp, 270-310 bp, and 300 bp 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 compared to 240-340 range; But 240-340 size selection range was with much more loci captured. Therefore, 240-340 size selection range was chosen for later experiments (Supplemental Table4).

To further calibrate the impact of ratio between duplex MIPs concentration and template DNA amount, a systematic evaluation of the ratio was carried out. 0 to 2000 ng template Hela DNA, 0 to 80nM MIPs were tested in this calibration (Supplemental Table5 & Supplemental Table6). Samples with more than 8000 reads and over 4000 captured loci were chosen for efficiency comparison. 8 to 80nM duplex MIPs probes with 250 to 500 ng input DNA were the most robust range of capture efficiency for our pipeline (Supplemental Figure 4). Considered the yield of single cell WGA reaction, we decided to use 8nM MIPs and 250 to 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.

Integrated bioinformatics Workflow Database Management System (DBMS) that enables efficient data management and reproducibility of analysis

A scalable architecture of Cell Lineage Discovery Workflow DBMS for collaborative cell lineage discovery was designed and implemented. The DBMS support (i) Computer storage and access to all workflow data, including data of each donor (anonymized), sample, cell, reagent type and location, measurement, sequencing run, analysis steps and algorithm runs. (ii) The application of any registered algorithm on any stored data. (iii) Version tracking of all workflow protocols and algorithms, and recording for each measurement, sequencing and algorithm-produced data which workflow version was used in its production. The design, as outlined in Figure 2, dissects both the biological and the computational workflow into atomic objects that are documented and referenced in all of experiments.

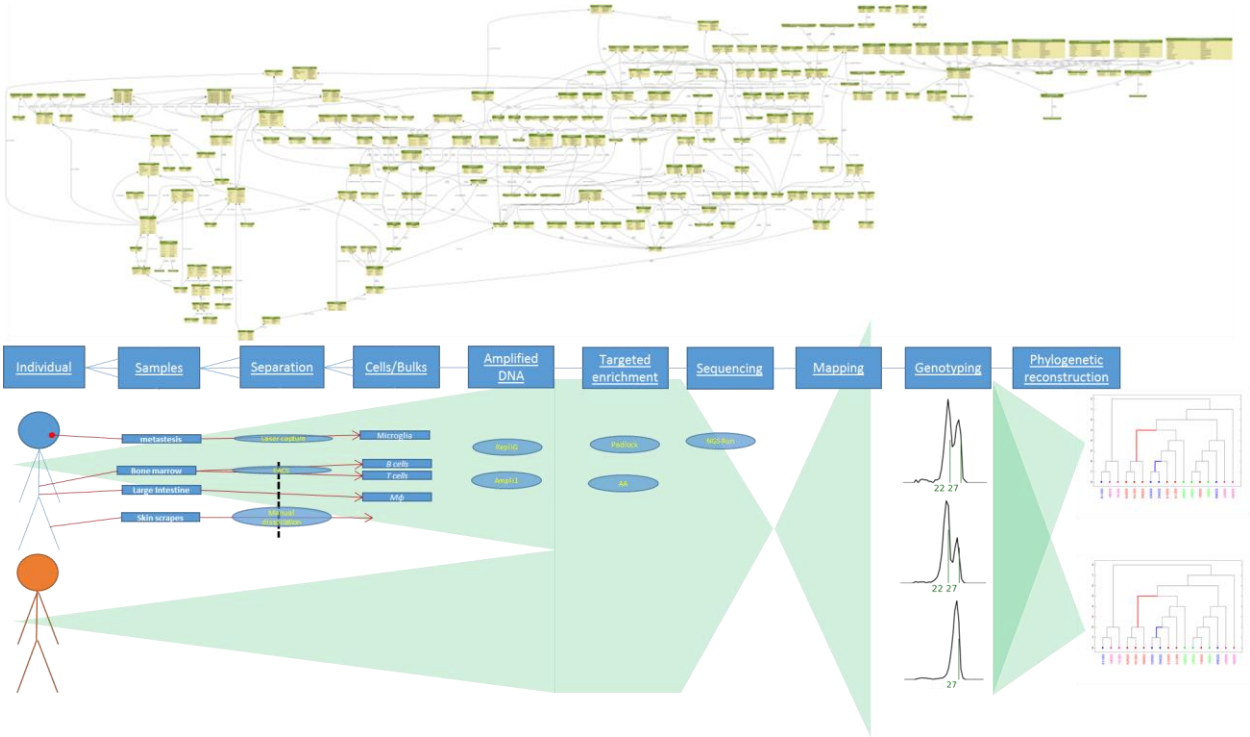

Figure 2. Outline of the Workflow Database and Management System | Top, Entity relation diagram (ERD) of the database structure. Bottom, schematic of the workflow from patient to tree through sampling documentation, library preparation, genotyping and phylogenetic analysis.

Comparison between duplex MIPs pipeline and Access Array pipeline

The cost between duplex MIPs pipeline and AA pipeline mainly differs in two aspects: the synthesis of oligoes or primers, reagents or chip for capture reactions per cell. As shown in Supplemental Figure4, the cost for synthesis was similar at 2000 loci scale, but when scaled up to 100,000 loci, the AA increased to 1,800,000$ roughly, while the duplex MIPs cost around 12,000$. The cost of reagents does not change much with scale up for duplex MIPs pipeline. For capture reaction per cell, more AA chips may be needed when scaled up, which increased the cost significantly. The initial cost of duplex MIPs pipeline was ~10,000$ for 12K loci panel (NGS run cost not included); capture reaction per cell was ~5.33$ (Supplemental Table 8).

The duplex MIPs pipeline uses a single PCR phase comparing to two in the AA pipeline[28]. A 15~20 cycles decrees inferred by the STR genotyping algorithm was shown in our pipeline (Figure 3C). This means less artificial noise was introduced into the STR repeats in the process of target enrichment that in turn translates into saving sequencing costs by allowing confident genotyping using less reads (Figure 3C).

DU145 *ex vivo* tree has been generated in our previous work that demonstrated cell lineage analysis using ~2000 MS loci[18]. Briefly, a single DU145 human male prostate cancer cell was cultured to generate a 9 generations *ex vivo* tree, each generation is of 12-15 cell divisions. Single cells were sampled in multiple stages of the tree and were subjected to WGA. The quality of the resulting genotypes was asserted by targeting the same single cell WGA DNA from *ex vivo* tree through the duplex MIPs pipeline[18]. The genotyping results reported by Biezuner *et al*[18] match those from our protocol in 99.2% of the cases. Our results demonstrate an improvement over the results reported by Biezuner *et al* using the AA chips[18] (Figure 3D). Duplex MIPs genotyping was performed with 5X as the minimal coverage against 30X reported for AA, further establishing the former introduces less noise.

Sequencing coverage distribution across panels was equivalent between Access Array and duplex MIPs. For each locus in the attempted panel (AA, Figure 3A or Duplex MIPs, Figure 3B), we counted the number of cells which present it with a coverage of over 30X. As the cells involved are actually the same, a similar distribution of sequencing depths indicated the loci retrieving performance of Duplex MIPs was not impaired by the ~10X scale up in the number of targets.

Finally, we compared the lineage reconstruction accuracy of AA and Duplex MIPs given the known lineage benchmark DU145 *ex vivo* tree, demonstrating improved lineage reconstruction accuracy (Figure 3E). Both reconstruction attempts were performed under naïve and equal parameters (detailed in methods section), the reconstructed lineage tree is shown in Figure 3F.

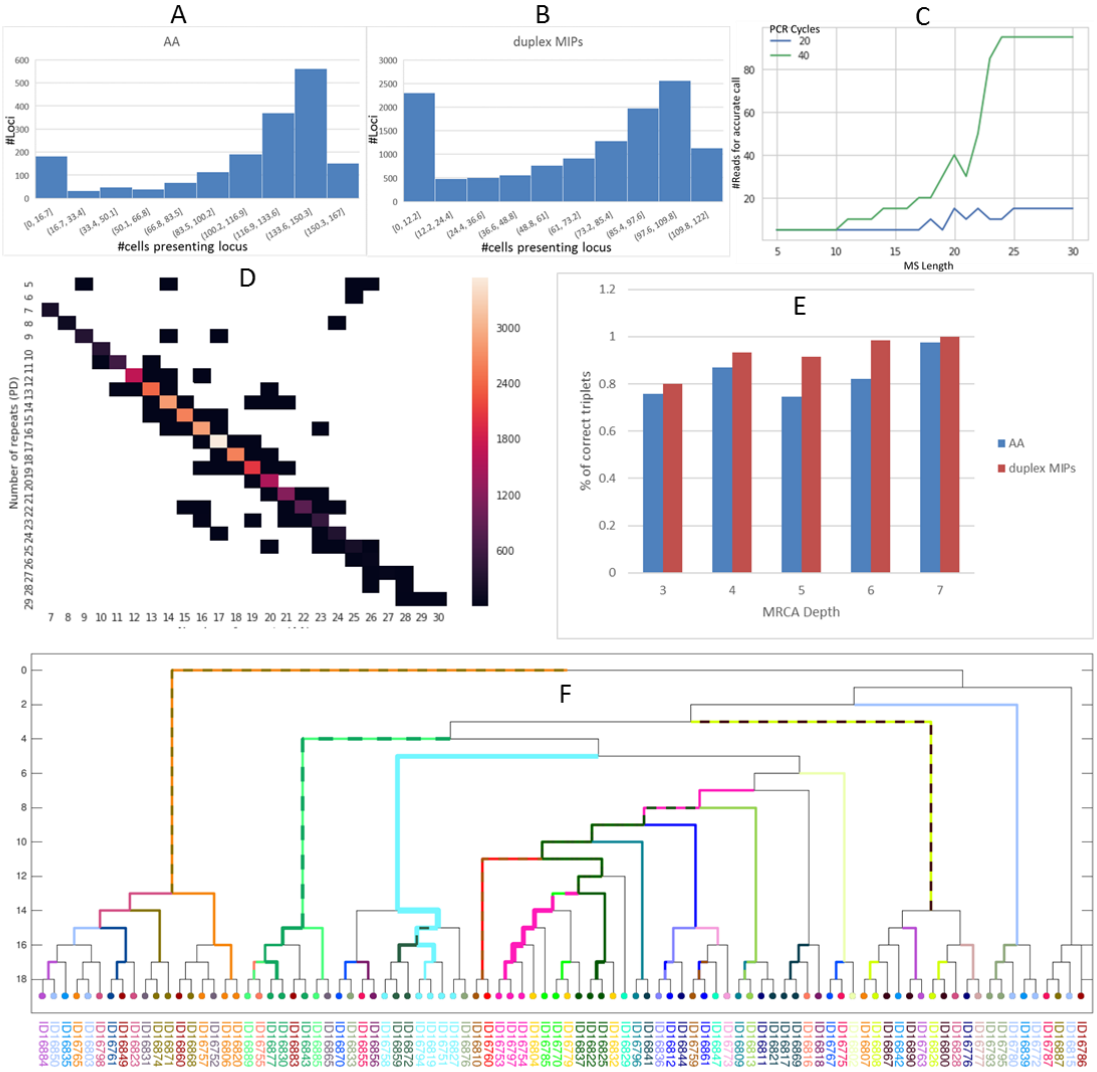

Figure3. *ex vivo* reconstruction and comparison between duplex MIPS pipeline and Access Array pipeline (A, B) STR signal comparison between Access Array and duplex MIPs. For each locus in the attempted panel, we count the number of cells, which present it with a coverage of over 30X. (C) Simulation analysis for the minimal number of reads required for accurate (less than 1 mistake in a 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 length between the two MRCAs of the triplet (higher is better). (F) Reconstructed Ex-Vivo lineage tree by duplex MIPs.

Cell lineage reconstruction of a melanoma patient (YUCLAT)

Single cell WGA DNA was obtained from both metastases cancer cells and normal peripheral blood lymphocytes (PBL) donated by a single melanoma patient as described in YUCLAT Krauthammer *et al*[29]. Cells were then processed using our platform for their cell lineage tree reconstruction (Figure 4). The reconstructed tree demonstrated an effective *in vivo* separation, validating both the expected grouping suggested by the samples origin as well as the SNP based clustering. Furthermore, a promising symbiosis of SNP and STR data was suggested by the additional clustering revealed in Figure 4C.

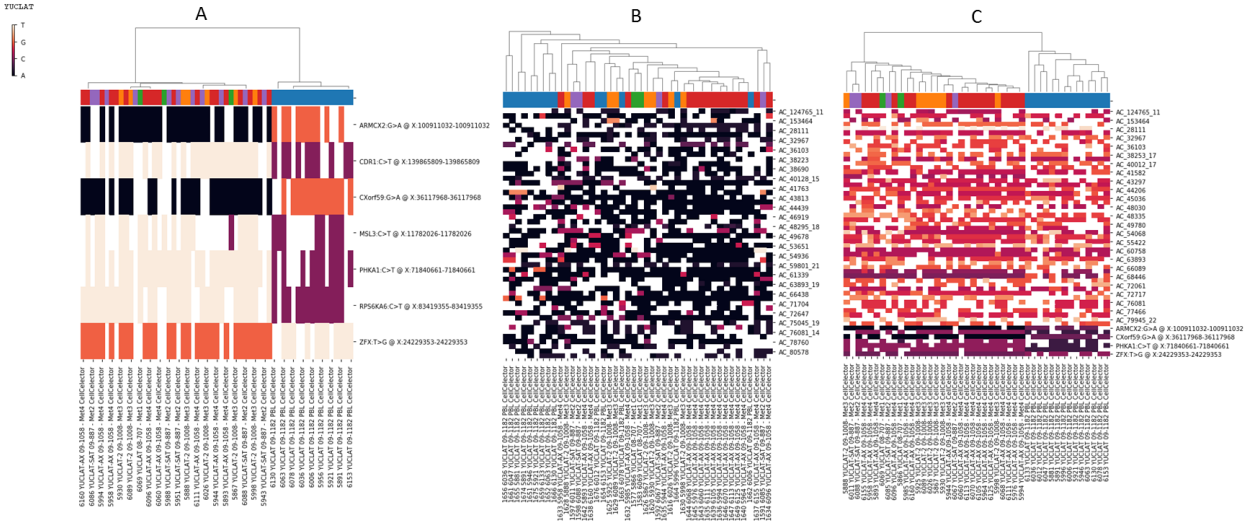

Figure 4. Cell lineage reconstruction of melanoma and normal lymphocytes from the same patient (YUCLAT) | (A,B) Combined heatmap and dendrograms, top: phylogenetic dendrogram that best mediates the clustering constraints imposed by the signal presented in the heatmaps (white indicated missing data, colors represent ACTG bases in A and STR repeat number differences in B, C). Between the heatmap and the dendrogram, a color-coded labels band highlights the original cells grouping labels. SNP based reconstruction shows clustering of healthy PBL from other Melanoma metastases (B) STR based reconstruction agrees with A and adds a cluster of Metastasis4 (C) Combined SNP and STR reconstruction agrees with both A and B and suggests a cluster of Metastasis3.

[DISCUSSION]

Several protocols for STRs genotyping or target sequencing were developed (Supplemental Table1), most of them designed for bulk DNA samples. Although they suit their needs well, a scalable high throughput method for cell lineage reconstruction from whole genome amplified DNA is still lacking. During the past decade, two STRs target enrichment platforms for single cell whole genome amplified DNA have been development in our lab: the first one was targeting 128 STRs based on 4X multiplex PCR, genotyping them by their fragment length detected by capillary electrophoresis[14]. To overcome the limitation from the low throughput of capillary electrophoresis, the second generation, NGS based protocol was developed with up to 50X multiplex PCR on AA chips[18]. This protocol could target over 2000 STRs for 48 samples on one AA chip. This Access Array based protocol was also too expensive to scale to target more than 100,000 STRs due to the accumulating cost of primer synthesis and AA chips[18]. Based on the MIP biochemical pipeline published by Shen *et al.*[25], we have developed a modified protocol: based on a microarray synthesized oligo pool, we generated a duplex MIPs based STRs target sequencing platform, which was demonstrated in 4,000 (data not shown) and 12K targets scale. Since the duplex MIPs could be generated from massive array based synthesis cheaply as an initial one- time purchase which is 10K$ for 100K and 2K$ for 12K, and the capture reaction cost does not change with the increment of targeting panel, the cost and scalability of STRs target enrichment could be significantly improved (Supplemental Figure6). The amplification cycles during sequencing library construction were reduced by 15~20 cycles compared to AA chip based pipeline, this allows us to obtain accurate STR genotyping with less reads as amplification based *in vitro* mutations are reduced. Currently, with a 12K panel, 150~200 cells can be sequenced in one NextSeq run. With more biochemical calibrations and improvement of the bioinformatics MIP arm design and sequencing analysis, future plans are to fit more single cells into a single sequencing run, and to increase the targets scale to 100K targets, thus improve the resolution of the tree[18].

Reconstructing a perfect lineage tree remains challenging. However, several issues we could make our effort to address. The artificial noise during both whole genome amplification can be reduced by Nano liter scale reaction WGA volume on droplet or Fluidigm chip[30, 31]; reduce amplification time and lower cell lysis temperature[32]. The artificial noise introduced by target enrichment process can also be reduced by use less barcoding PCR cycles. We could also use an improved single cell WGA kit to increase the uniformity of amplification. To aid bi-allelic or multi-allelic loci genotyping, duplex MIPs targeting SNV-STRs region within one amplicon can be designed. The STRs 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[33].

In summary, we have developed an easily initiated, scalable, cost effective platform for massive parallel STRs mutation discovery 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 integrations including STRs, SNVs *etc*. on demand. By applying this platform to various types of human cells, we demonstrated capabilities that tens-of-thousands hyper-mutable STR targets in single cell whole genome amplified DNA could be acquired efficiently and discovered lineage relations among these cells.

With the advancement of single cell analysis methodologies, the cost of human cell lineage tracing can be reduced and the accuracy and resolution can be improved. What's more, the cell states, single cell spatial information, genomic functions could be integrated into the cell lineage tree. Together, this will help us better understand the development of human in both health and disease status.

[**ACKNOWLEDGMENTS**]

We thank our collaborator Ruth Halaban from the Department of Dermatology, Yale University School of Medicine, New Haven, Connecticut for providing us with SC genomic DNA samples, thus supporting the individual's cell lineage tree experiment (Figure 4). 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); and the Kenneth and Sally Leafman Appelbaum Discovery Fund. Ehud Shapiro is the incumbent of The Harry Weinrebe Professorial Chair of Computer Science and Biology.

[REFERENCES]

1. Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**, 64-119 (1983).
2. Sulston, J.E. & Horvitz, H.R. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev Biol* **56**, 110-156 (1977).
3. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**, 618-630 (2013).
4. Jan Philipp Junker, B.S., Josi Peterson-Maduro, Anna Alemany, Bo Hu, Maria Florescu, Alexander van Oudenaarden Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars.
5. Frieda, K.L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111 (2017).
6. Kalhor, R., Mali, P. & Church, G.M. Rapidly evolving homing CRISPR barcodes. *Nat Methods* **14**, 195-200 (2017).
7. Schmidt, S.T., Zimmerman, S.M., Wang, J., Kim, S.K. & Quake, S.R. Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding. *ACS Synth Biol* **6**, 936-942 (2017).
8. Evrony, G.D. et al. Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. *Neuron* **85**, 49-59 (2015).
9. Evrony, G.D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496 (2012).
10. Mann, K.M. et al. Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq. *Nat Biotechnol* **34**, 962-972 (2016).
11. Lodato, M.A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94-98 (2015).
12. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422-425 (2014).
13. Frumkin, D. et al. Cell lineage analysis of a mouse tumor. *Cancer Res* **68**, 5924-5931 (2008).

14. Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. & Shapiro, E. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput Biol* **1**, e50 (2005).
15. Wasserstrom, A. et al. Reconstruction of cell lineage trees in mice. *PLoS One* **3**, e1939 (2008).
16. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**, 435-445 (2004).
17. Press, M.O., Carlson, K.D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet* **30**, 504-512 (2014).
18. Biezuner, T. et al. A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res* **26**, 1588-1599 (2016).
19. Nilsson, M. et al. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085-2088 (1994).
20. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* **10**, 857-860 (2013).
21. Schneider, N. & Meier, M. Efficient in situ detection of mRNAs using the Chlorella virus DNA ligase for padlock probe ligation. *RNA* **23**, 250-256 (2017).
22. Larsson, C., Grundberg, I., Soderberg, O. & Nilsson, M. In situ detection and genotyping of individual mRNA molecules. *Nat Methods* **7**, 395-397 (2010).
23. Porreca, G.J. et al. Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-936 (2007).
24. Shen, P. et al. Multiplex target capture with double-stranded DNA probes. *Genome Med* **5**, 50 (2013).
25. Shen, P. et al. High-quality DNA sequence capture of 524 disease candidate genes. *Proc Natl Acad Sci U S A* **108**, 6549-6554 (2011).
26. Yoon, J.K. et al. microDuMIP: target-enrichment technique for microarray-based duplex molecular inversion probes. *Nucleic Acids Res* **43**, e28 (2015).
27. Carlson, K.D. et al. MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Res* **25**, 750-761 (2015).
28. Raz, O. et al. Short Tandem Repeat stutter model inferred from direct measurement of in vitro stutter noise. *bioRxiv* (2016).
29. Krauthammer, M. et al. Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat Genet* **47**, 996-1002 (2015).
30. Gole, J. et al. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.* **31**, 1126-+ (2013).
31. Fu, Y. et al. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proc Natl Acad Sci U S A* **112**, 11923-11928 (2015).
32. Dong, X. et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods* **14**, 491-493 (2017).
33. Willems, T. et al. The landscape of human STR variation. *Genome Res* **24**, 1894-1904 (2014).