MICROFLUIDIC FRACTIONATION AND ANALYSIS OF CYTOPLASMIC VERSUS NUCLEAR NUCLEIC ACIDS IN SINGLE CELLS

A DISSERTATION

SUBMITTED TO THE DEPARTMENT

OF MECHANICAL ENGINEERING

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Denitsa Milanova

December, 2015

This dissertation is online at: http://purl.stanford.edu/vm277bt4701

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Juan Santiago, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Stephen Quake

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Michael Snyder, PhD

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

ABSTRACT

Single-cell gene expression studies have matured and are now widely used to uncover cell-to-cell variability, transcript processing and inhomogeneous response to external stimuli. Despite great progress, significant challenges remain including understanding fundamental questions of human genome transcription, gene expression ranges, localization, and processing destinies of RNAs. Eukaryotic cells make different types of primary and processed transcripts, which are either exclusively found in particular sub-cellular compartments or dispersed throughout the whole cell. These sub-cellular localizations of RNAs are weakly understood at a single-cell level, and are important for fully realizing gene functions.

Understanding sub-cellular transcript localization is particularly important for studying the process of splicing, a pathway during which introns of pre-messenger RNA's are excised and exons are stitched together to form mature mRNA transcript. We here introduce a single-cell-isotachophoresis (sc-ITP) method to analyze co-transcriptional and alternative splicing in reference lymphoblastoid cell line (LCL-Snyder), and two sub-lines of CML cells (K562-ATCC and K562-ENCODE). The method is unique in that we physically separate the contents of cell nucleus from those of the cytoplasm and analyze independently. It allows for rapid, electric-field-based selective lysis of cytoplasmic membrane (leaving nucleus intact); and then purification and simultaneous fractionation of total RNA in cytosol (cyt-RNA) and total RNA in the nucleus (nuc-RNA) from single cells with no intra-compartment cross-contamination.

We then use our system and state-of-the-art NGS technologies to perform single-cell nuclear and cytoplasmic RNA-seq to study fundamental questions of human genome transcription, differential gene expression, localization, and processing destinies of RNAs for various cell types, disease, and differentiation states. We evaluate the

distribution of whole transcriptome gene features and gene expression of precursor and processed transcripts and find that there are considerable differences in expression levels across sub-cellular compartments and among individual cells. We then evaluate cell-to-cell variability of the highly expressed GAPDH housekeeping gene in alternative splicing between the lymphoblastoid and leukaemia cells using sequencing-specific probes and RT-qPCR. The data suggest evidence of significant distinction in splicing patterns. By analyzing nuclear and cytoplasmic compartments of a single cell individually at a whole transcriptome level, we achieve an unprecedented precision in splicing quantification. Together, our results describe an experimental method for single-cell fractionation and an analytical tool for gene and isoform expression analysis in rare cell types, cell differentiation and disease states.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Appendices

1 BACKGROUND

NEXT-GENERATION SEQUENCING TECHNOLOGIES

Several section of this background chapter are based on a review article by Thomas P. Niedringhaus, Denitsa Milanova, Matthew B. Kerby, Michael P. Snyder, and Annelise E. Barron, named "*Landscape of Next-Generation Sequencing Technologies*" in Analytical Chemistry [1]. The content presented here strongly emphasizes the contributions of Denitsa Milanova, and is presented here with minor modifications.

1.1 Introduction

DNA sequencing is in the throes of a disruptive technological shift marked by dramatic throughput increases, a precipitously dropping per-base cost of raw sequence, and an accompanying requirement for substantial investment in large capital equipment in order to "play". Investigations that were, for most, unreachable luxuries just a few years ago - individual genome sequencing,

metagenomics studies, and the sequencing of myriad organisms of interest - are being increasingly enabled, at a rapid pace. This review concentrates on the technology behind the third- and fourth-generation sequencing methods: their challenges and current limitations, and their tantalizing promise.

First-generation sequencing encompasses the chain termination method pioneered by Sanger and Coulson [2] in 1975 or the chemical method of Maxam and Gilbert in 1976-1977 [3]. In 1977, Sanger sequenced the first genome, bacteriophage ΦX 174, which is 5,375 bases in length [4]. These methods and their early history [5] have been reviewed in detail previously [6]. Four-color fluorescent Sanger sequencing, where each color corresponds to one of the four DNA bases, is the method used by the automated capillary electrophoresis (CE) systems marketed by Applied Biosystems Inc., now integrated into Life Technologies, and by Beckman Coulter Inc. [7]. The first composite human genome sequence, reported in 2001, was obtained largely using CE, at great cost and with intense human effort over more than a decade [8, 9]. While the genome reported in 2001 was a work in progress, the availability of an ever-improving "reference" genome is the basis an on-going transformation of biological science, and remains fundamental to investigations of genotype-phenotype relationships. Considering reports that have appeared (and not appeared) in the literature to date, it could well be that medically meaningful (actionable) insights into complex diseases will require additional types of "personal" genomic data, for instance, tissue-specific mRNA expression profiling and mRNA sequencing, individualized analysis of gene regulatory regions, epigenetic profiling, and high-quality, long-range chromosome mapping to catalog significant deletions, insertions, rearrangements, *etc*. Correlation of

such integrated genomic datasets with comprehensive medical histories for hundreds or thousands of individuals may be what it takes to reach an era of personalized medicine [10-12]. Large-scale sequencing centers are now completing the conversion to next-generation sequencers; the Joint Genome Institute (JGI) has retired all of their Sanger sequencing instruments [13]. At the other extreme, until small-scale next-generation sequencers can outperform CE on a cost per accurate base called as well as read length, CE systems will likely remain in heavy use for benchtop-scale, targeted sequencing for directed investigations such as quantitative gene expression, biomarker identification and pathway analysis.

1.2 2nd generation sequencing

Several reviews of what were first called "next-generation" or more precisely, second-generation sequencing technologies have appeared [5, 14-16]. We propose to classify the second-generation technologies as a combination of a synchronized reagent wash of NTPs with a synchronized optical detection method. However, this definition is not rigid, as several real-time synthesis strategies, which comprise third-generation technologies, also rely on optical detection, with Pacific Biosciences' single DNA polymerase sequencing method being a prime example. Second-generation technologies rely upon sequencing by ligation or sequencing by synthesis, including pyrosequencing and reversible chain termination. Commercially available instruments from Roche, Illumina, Helicos, and Life Technologies deliver several Gbp of DNA sequence per week in the form of short contiguous fragments, or reads. A review of second-generation methods based on

sequencing by synthesis, in which a polymerase or ligase controls the biochemistry, details the challenges and advantages to these types of enzymatic approaches [17].

1.3 2nd generation costs

Over the last few years, companies marketing second-generation sequencers have competed to demonstrate their increasingly cost-effective approaches to generating an assembled complete human genome, relying on the known reference genome. Compared to the costs of generating the draft of J. Craig Venter's genome with ABI's Sanger-CE instruments [18], Roche's 454 Genome Sequencer FLX [19], Illumina's Genome Analyzer [20-22], and Helicos' Heliscope [23] have decreased the cost of obtaining raw sequence by roughly one, two, and three orders of magnitude, respectively. In all of these reports, only the costs of consumables and reagents were taken into account, however. These new "massively parallel" sequencing instruments require a concomitantly massive investment in capital equipment, since many of these high-throughput instruments are priced between $500K and $1M each. The labor costs to operate the equipment and the informatics cost for reassembly of the sequence should be factored into the overall sequencing cost. As of this writing, Illumina dominates the market with 60% [24] of the second-generation sequencer installations, while Life Technologies' SOLiD system and Roche split nearly all of the remaining market at 19% each. Illumina's whole-genome sequencing service will sequence a human genome for $19,500 [25] - a great deal less than Illumina's reagent costs of $250,000 needed to sequencing a

complete human genome (or $0.0002 per sequenced base) in 2008 [20] and even greater than the cost back in 1996, when first-generation sequencing cost $1 for each sequenced (finished) base. To reduce costs, Illumina, which uses reversible terminator-based sequencing by synthesis chemistry, recently launched the smaller, less expensive, MiSeq platform, which promises over 1 Gb of 150 bp reads in 27 hours. This more compact system is specifically designed to challenge CE-based sequencing for common experiments such as clone verification, amplicon sequencing and small genome sequencing. On the larger scale, Life Technologies 5500xl series instruments, which use sequencing by ligation chemistry, can collect up to 30 Gb per day over 7 days of operation. For the benchtop market, Ion Torrent, a division of Life Technologies, is developing a third-generation solution, and has recently launched the Personal Gene Machine (PGM) and the Ion Express OneTouch template preparation system [26]. The Roche 454 relies on pyrosequencing to detect single base extensions from beads using a luciferase-based method, refined for synchronized DNA sequencing in 1996 [27]. The light-emitting pyrosequencing method, which does not use multiple fluorophores, does not require lasers or expensive optical filters, greatly reducing the cost of the equipment. The Roche 454 GLX Flex Titanium series, a $500k instrument, reportedly can generate 400-600 million high-quality base calls per day. New development aims to raise the read length to 800+ base calls [28]. The $100K 454 GS Junior launched in 2009 and also targeted for benchtop research, produces 35Mb in 10 hours, with 400 base pair reads. “Benchtop” NGS technology development, which squarely challenges first-generation Sanger CE sequencing

[28], seeks to achieve a drastic decrease in cost, physical size, and complexity while continuously increasing throughput, read length, and read accuracy.

In an effort to illustrate the true cost of complete genome sequencing, the National Human Genome Research Institute (NHGRI) has compiled data from their sequencing centers to appropriately estimate the overall costs of sequencing a human genome [29]. Their calculations take into account labor, three-year amortization of sequencing instruments, data processing, and sample preparation. Figure 1-1 illustrates the cost associated with sequencing a human-sized haploid genome (3,000 Mb) over time since the initial draft of the human genome was published in 2001. The dramatic drop in cost seen in 2008 is the result of transitioning from first-generation Sanger CE sequencing to second-generation platforms installed in sequencing centers (*i.e.*, 454, Illumina, and SOLiD). The second-generation technologies yield lower contiguous read lengths and require greater genome coverage for assembly; however, their high throughput reduces consumable costs and the number of sequencing runs.

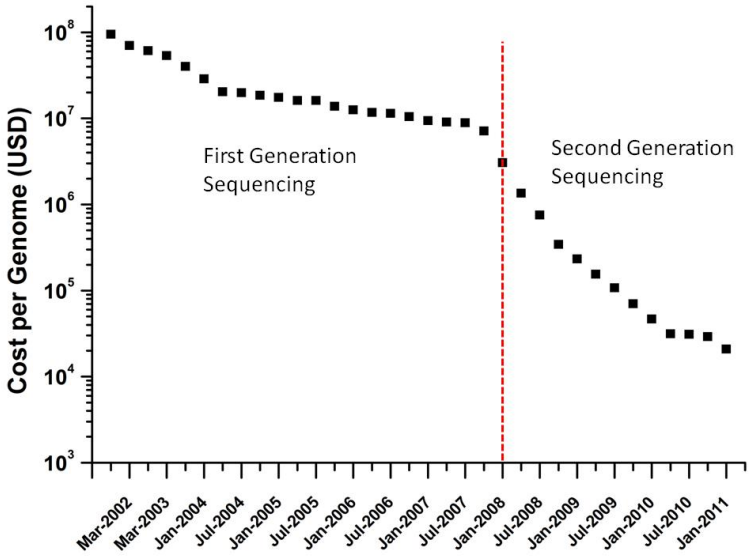


Figure 1-1. *Estimated cost required to sequence a complete human genome based on data generated from NHGRI-funded large-scale DNA sequencing centers [29].*

Technology development costs and data analysis costs are omitted from these sequencing cost calculations. In general, these costs are much higher for less established second- and third-generation sequencing technologies. For instance, the data depicted in Figure 1-1 produced by second-generation sequencing technologies (after 2008) are the result of re-sequencing efforts in which a reference human genome was used to guide the reassembly process. The practicality and cost associated with the sequencing and *de novo* assembly of a human genome using only second- or third-generation technologies is difficult to assess at this time, given that *de novo* sequencing has only been accomplished using Sanger-based CE [30]. It appears that the greatest cost barrier is the complex hardware required for the achievement of precisely aligned optical detection and downstream data processing.

1.4 Third-Generation sequencing technologies

With the final goal of bringing the cost of a human genome to under $1000, NIH/NHGRI has funded several groups developing alternative approaches to improving second-generation technologies, as well as novel approaches to sequencing that include the use of scanning TEM, FRET, single-molecule detection, and protein nanopores. Two of the leading third-generation sequencing technologies (Pacific Biosciences and Complete Genomics) still rely on optical detection of fluorescent events, but aim to increase sequencing speed and throughput. Ion Torrent's technology, on the other hand, uses ion-sensitive field effect transistors (ISFETs) to eliminate the need for optical detection of sequencing events. Nanopore technologies, such as Oxford Nanopore, also aim to remove optics as well as the need for DNA amplification in their sequencing design by measuring changes in conductivity across a nanopore. Non-optical TEM approaches used by Halcyon Molecular and ZS Genetics require million-dollar capital equipment and, to date, have limited throughput, yet in principle could give the sequence of thousands of contiguous bases. Finally, new methods involving optical methods are being developed that allow for previously unattainable levels of long range mapping, which is essential for accurate assembly of individual human genomes and cancer genomes. We now examine these third- and next-generation technologies in detail and outline the advantages and disadvantages of each technique.

1.5 Single-Molecule Sequencing

1.5.1 Pacific Biosciences

Pacific Biosciences (PacBio) has led the charge to develop a reliable third-generation sequencing platform based on a real-time, single-molecule sequencing technology. Their process directly measures DNA polymerase incorporation of fluorescently labeled nucleotides onto a complementary sequencing template. At the heart of this technology is a dense array of zero-mode waveguide (ZMW) nanostructures that allow for optical interrogation of single fluorescent molecules. While ZMW structures have been used in the past to differentiate single fluorescent molecules from substantially large bulk concentrations [31-35], they have not been used in a highly parallel fashion. To address this issue and increase throughput, PacBio developed a method to efficiently pack ZMW nanostructures onto a surface using electron beam lithography and ultraviolet photolithography [36] as well as a highly parallel confocal imaging system that permits high sensitivity and resolution of fluorescent molecules in each of the ZMW nanostructures [37]. Specialized heavy concrete bases are employed to maintain optical confocality.

Once the ZMW array fabrication and detection scheme was established, the major technical hurdles for this technology came in the form of immobilizing a single functioning DNA polymerase at the bottom of each ZMW, which can process fluorescently labelled nucleotide substrates. This was accomplished in two steps. First, a set of fluorescently labelled deoxyribonucleoside pentaphosphate (dN5Ps) substrates were synthesized so that each base is spectrally differentiable without

decreasing the processivity of the DNA polymerase [38]. Second, surface treatment of the ZMW nanostructure was needed to selectively localize the DNA polymerase. The ZMW array is comprised of a fused silica bottom layer and an aluminum top layer in which all ZMW nanostructures are defined. By derivatizing the aluminum surface with polyvinylphosphonic acid (PVPA), protein adsorption to the aluminum layer was significantly decreased without compromising protein adsorption to the bottom glass layer [39]. Combining these chemical modifications with the highly parallel ZMW array, PacBio was able to demonstrate a single-molecule real-time (SMRT) sequencing technique that generates long read lengths (on the order of 1000 bases) and a four-color sequencing trace [40]. A limit to throughput was imposed by the stochastic nature of immobilizing DNA polymerases at the bottom of each ZMW. In the published study, roughly one-third of the ZMWs in the array contained a single DNA polymerase and had the capacity to generate full-length sequencing reads. Figure 1-2 depicts the four-color SMRT sequencing strategy employed in this important article.

Following the proof-of-concept of SMRT sequencing study, PacBio streamlined the sequencing template construction by creating what they call a SMRTbell template [41]. The SMRTbell template allows consecutive sequencing of both the sense and antisense strand of a double-stranded DNA fragment by ligating universal hairpin loops to the ends of the fragment. Sample preparation time is decreased since template amplification is not needed and DNA fragments over a broad size range can be used to generate SMRTbell templates. In the end, the use of the SMRTbell template increases the accuracy of sequencing and SNP detection.

PacBio now offers a commercially available sequencing instrument, the PacBio RS system. Consumables for this instrument include single-use ZMW arrays (called SMRT Cells) that contain 150,000 ZMWs and kits for SMRTbell template preparation. Recently, the PacBio RS Sequencer was used for the rapid genotyping of five *Vibrio cholerae* strains to determine the source of a cholera outbreak in Haiti [42]. Average read lengths for the five strains ranged from ~ 700 to 1,000 bases, while the average depth of coverage ranged from 28 to 60, and the mean single-pass accuracy ranged from 81% to 83%. For three of the strains, read lengths approaching 3,000 bases were reported for a small percentage of the sequencing runs.

Besides sequencing, other applications are being developed using the SMRT detection technology. PacBio enhanced the robustness of genetic information generated by their single-molecule detection assay by correlating polymerase kinetic data to DNA methylation patterns during DNA sequencing [43]. The ability to sequence strands of mRNA at the level of codon resolution has been proven by simply substituting the DNA polymerase at the bottom of each ZMW with a ribosome and monitoring incorporation of fluorescently labelling transfer RNAs (tRNAs) [44].

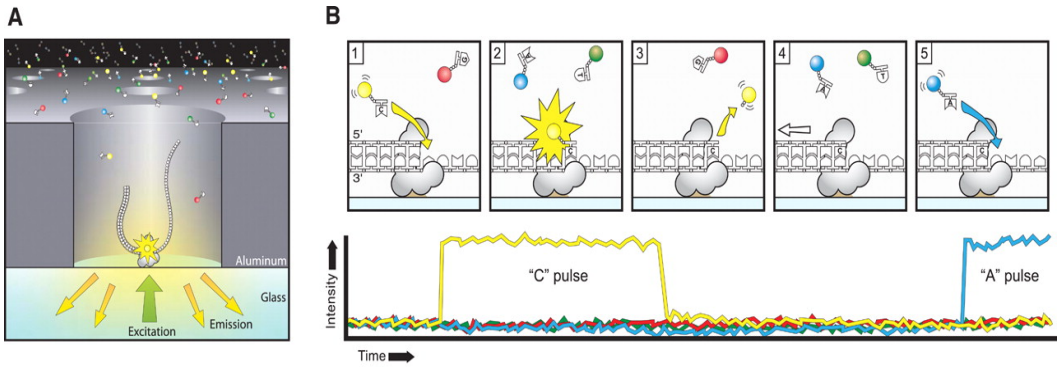


Figure 1-2. *Schematic of PacBio's real-time single molecule sequencing. (a) The side view of a single ZMW nanostrucute containing a single DNA polymerase (Phi29) bound to the bottom glass surface. The ZMW and the confocal imagining system allow fluorescence detection only at the bottom surface of each ZMW. (b) Representation of fluorescently labelled nucleotide substrate incorporation on to a sequencing template. The corresponding temporal fluorescence detection with respect to each of the five incorporation steps is shown below. Image reproduced with permission [40].*

1.6 Sequencing by Ligation

1.6.1 Complete Genomics

The basis for Complete Genomics' sequencing platform is centered on a hybridization and ligation method. While sequencing by hybridization and ligation has been around for some time [45-50], the sample preparation and nanoarray platform developed by Complete Genomics is novel [51]. Sequencing fragments are prepared by sonication of genomic DNA followed by a series of repeated adapter site insertions, template circularization, and restriction enzyme scission. In the end, circularized sequencing fragments on the order of 400 bases are generated,

each containing four distinct adapter sites. Circularized fragments are amplified by two orders of magnitude using Φ29 polymerase. Each amplified product of a circularized fragment is called a DNA nanoball (DNB). DNBs are selectively attached to hexamethyldisilizane (HMDS) coated silicon chip that is photolithographically patterned with aminosilane active sites. Figure 1-3.a illustrates the DNB array design.

The use of the DNBs coupled with the highly patterned array offers several advantages. The production of DNBs increases signal intensity by simply increasing the number of hybridization sites available for probing. Also, the size of the DNB is on the same length scale as the active site or "sticky" spot patterned on the chip, which results in attachment of one DNB per site. Since the active sites are spaced roughly 1 μm apart, up to three billion DNB can be fixed to a 1 inch by 3 inch silicon chip [52]. In addition to increasing the number of sequencing fragments per chip, the length scales of the size and spacing of the DNBs maximizes the pixel use in the detector. This highly efficient approach to generating a hybridization array results in decreased reagent costs and increased throughput compared to other second generation DNA sequencing arrays that have been used to sequence complete human genomes [20, 23, 53].

Once the DNB array chip is generated, a library of forty common probes is used in combination with standard anchors and extended anchors to perform an unchained hybridization and ligation assay. The forty common probes consist of two subsets: probes that interrogate 5' of the distinct adapter site in the DNB and probes that interrogate 3' of the distinct adapter site in the DNB. In each subset there are five sets of four common probes; each probe 9 bases in length. Each set corresponds to

positions 1 to 5 bases away from the distinct adapter sites in the sequencing substrate and within each set there are four distinct markers corresponding to each base. The standard anchors bind directly to the 5' or 3' end of the adapter site on the DNB and allow for hybridization and ligation of the common probes. The extended anchor scheme consists of ligation of a pair of oligo anchors (degenerate and standard) to expand the hybridized anchor region 5 bases beyond the adapter sites in the DNB and into the sequencing region. This combinatorial probe-anchor ligation (cPAL) method developed by Complete Genomics extends read lengths from 5 bases to 10 bases and results in a total of 62 to 70 bases sequenced per DNB. A schematic demonstrating both the standard anchor scheme and the extended anchor scheme is shown in Figure 1-3.b.

Each hybridization and ligation cycle is followed by fluorescent imaging of the DNB spotted chip and subsequently regeneration of the DNBs with a formamide solution. This cycle is repeated until the entire combinatorial library of probes and anchors is examined. This formula of using unchained reads and regeneration of the sequencing fragment reduces reagent consumption and eliminates potential accumulation errors that can arise in other sequencing technologies that require close to completion of each sequencing reaction [20, 53, 54].

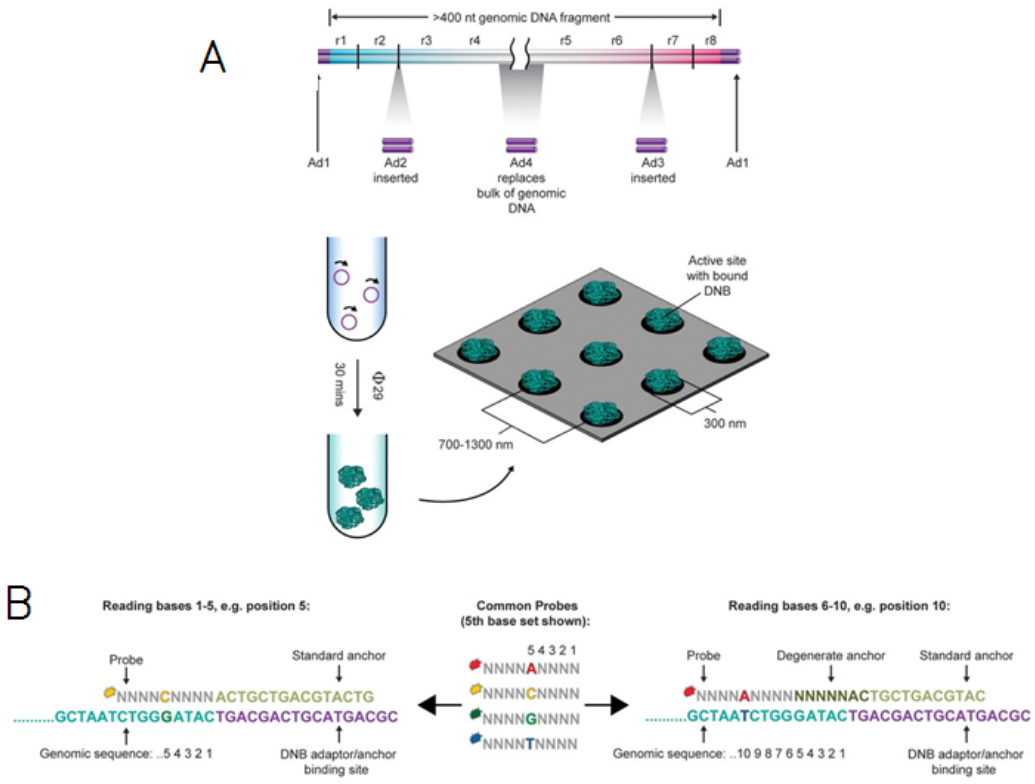


Figure 1-3. *Schematic of Complete Genomics' DNB array generation and cPAL technology. (a) Design of sequencing fragments, subsequent DNB synthesis, and dimensions of the patterned nanoarray used to localize DNBs illustrate the DNB array formation. (b) Illustration of sequencing with a set of common probes corresponding to 5 bases from the distinct adapter site. Both standard and extended anchor schemes are shown. Image reproduced with permission [51].*

Complete Genomics showcased their DNB array and cPAL technology by re-sequencing three human genomes, and reported an average reagent cost of $4,400 per genome [51]. The three-genome samples sequenced in this study (NA07022, NA19240, and NA20431) were then compared to previous sequence data [55, 56]. The average coverage of these samples ranged from 45X to 87X and the percent of genome identified ranged from 86% to 95%. While this technology greatly

increases throughput compared to Sanger/CE and second-generation sequencing technologies, there are several drawbacks to Complete Genomics' approach. First, the construction of circular sequencing fragments results in an underrepresentation of certain genome regions, which leads to partial genome assembly downstream. Also, the size of the circular sequencing fragments (~400 bases) as well as the very short read lengths (~10 bases) prevents complete and accurate genome assembly, given that these fragments are shorter than a number of the long repetitive regions.

Just five months after Complete Genomics' proof-of-concept study was published, the first externally published application of Complete Genomics sequencing technology was released. A group at the Institute for Systems Biology in Seattle, Washington (USA) studied the genetic differences in a human family of four [57]. In the study, whole-genome sequencing was used to determine four candidate genes responsible for two rare Mendelian disorders: Miller syndrome and primary ciliary dyskinesia. The subjects were a set of parents and their two children who both suffer from the disorders. This study highlighted the benefits of whole genome sequencing within a family when determining Mendelian traits. The ability to recognize inheritance patterns greatly reduced the genetic search space for recessive disorders and increased the sequencing accuracy. In the end, sequencing the entire family instead of just the two siblings affected by Miller syndrome and primary ciliary dyskinesia greatly decreased the number of false positive gene candidates, which ultimately reduced the number gene candidates from 34 to just four.

Just one month later, the second externally published application of Complete Genomics' sequencing technology was released by a research group at Genentech

[58]. The study compared the genome of primary lung tumor cells to that of adjacent normal tissue obtained from a 51-year-old Caucasian male who reported a heavy 15-year smoking history. By comparing the complete genomes of different tissue samples, over 50,000 single-base variations were identified and 530 previously reported single-base mutations were confirmed. Consequently, the importance of complete cancer genome analysis in understanding cancer evolution and treatment was brought to light due to the large number of single nucleotide mutations located outside of oncogenes as well as chromosomal structural variations found in the primary lung tumor.

A third application of the high-throughput cPAL method developed by Complete Genomics was published by a research group from the University of Texas Southwestern Medical Center in Dallas, Texas (USA) [59]. This group used whole genome sequencing to diagnose a hypercholesterolemic 11-month-old girl with sitosterolemia after a series of blood tests and selective genetic sequencing were unable to confer a reasonable diagnosis. The gene and the subsequent mutations responsible for the sitosterolemia phenotype were determined after comparison of the patient's genome to a collection of reference genomes. Ultimately, it was determined that the patient failed the standard blood test due to low levels of plant sterols that were the result of a heavy diet of breast milk. This study illustrated the importance of whole genome sequencing in effectively diagnosing a disease in the presence of complex environmental factors that can influence standard assays.

1.7 Sequencing by Synthesis

The idea of sequencing by synthesis has been around for some time and is the basis for several second-generation sequencing technologies including Roche's 454 sequencing platform and Illumina's line of sequencing systems. 454's pyrosequencing method, which uses an enzyme cascade to produce light from a pyrophosphate released during nucleotide incorporation, was first piloted in the late 1980's and developed for DNA sequencing in the mid-1990's [27, 60-64]. Illumina's fluorescently labeled sequencing by synthesis technique employs fluorescently labeled nucleotides with reversible termination chemistry and modified polymerases for improved incorporation of nucleotide analogues [20]. These sequencing by synthesis methods increased throughput compared to first-generation sequencing methods, however optical imaging is needed to detect each sequencing step. Since an intricate optics system can increase the overall cost of a sequencing system, the next logical advancement in the sequencing field has been to abandon the use of optics for a less expensive approach to detection.

Taking this into account, research done in the Pease and Davis labs at Stanford University evolved from earlier pyrosequencing technology by proposing a new method of detection to measure temperature or pH change in microstructures [65-67]. Since both changes are by-products of nucleotide incorporation in a DNA polymerization reaction, the need for optical detection of light produced by the luciferase enzyme was eliminated. Like pyrosequencing, this thermo-sequencing method requires sequential cycles in which one of the four nucleotides is introduced to the system, followed by measurement of nucleotide incorporation by

heat detection. Between each cycle, the system is regenerated by thorough washing of reaction wells to minimize residual NTPs, and therefore reduce error accumulation. This innovative detection scheme led to the start-up of the company Genapsys from the Stanford Genome Technology Center. Thermal detection has an inherent advantage over pH detection, in that temperature can be reset quickly by conduction from a cooling block, while hydrogen ions must be washed away. Ion Torrent, a startup recently acquired by Life Technologies [68], has made significant progress in bringing to market a next-generation sequencing system that utilizes pH changes to detect base incorporation events.

1.7.1 Ion Torrent

According to Ion Torrent's patent applications (US2009/0026082 A1 and US2009/0127589 A1), field-effect transistors (FETs) are used to measure a change in pH in a microwell structure (see Figure 1-3). To increase throughput, the Ion Torrent sequencing chip makes use of a highly dense microwell array in which each well acts as an individual DNA polymerization reaction chamber containing a DNA polymerase and a sequencing fragment. Just below this layer of microwells is an ion-sensitive layer followed by a sub-layer composed of a highly dense FET array aligned with the microwell array. Following the pyrosequencing scheme, sequential cycling of the four nucleotides into the microwells enables primary sequence resolution since the FET detector senses the change in pH created during nucleotide incorporation and converts this signal to a recordable voltage change. Since the change in voltage scales with the number of nucleotides incorporated at each step, Ion Torrent's sequencing chip has an inherent capacity to call repeats.

At present, Ion Torrent offers the one-time-use Ion 314 sequencing chip, however within the next year they are scheduled to release their second- and third-generation chips: the Ion 316 sequencing chip and the Ion 318 sequencing chip (Ion Torrent Application Note, Spring 2011). The 1.2 million microwells on the Ion 314 chip generates roughly 10 Mb of sequence information with average read lengths on the order of a 100 bases. To further increase throughput, the Ion 316 chip and the Ion 318 chip are being built with 6.2 million and 11.1 million microwells, respectively. The expectations for the Ion 318 chip are to produce 1 Gb of sequencing data with average read lengths of 200 bases or higher. Ultimately, Ion Torrent seeks to "democratize" sequencing by offering the first reasonably priced (~ $50K) benchtop-scale, high-throughput sequencing machine [69].

While this newly developed method of ion sensing-based sequencing by synthesis offers great potential to reduce the cost of sequencing, there are several limitations with regards to sequencing complete genomes. Currently, the short read lengths place a large burden on the reassembly process and limit the assembly of *de novo* sequencing projects due to an inability to read through long repetitive regions in the genome. Also, due to the sequential nature of this sequencing by synthesis method, error accumulation can occur if reaction wells are not properly purged between reaction steps. Finally, as for pyrosequencing in the previous generation, sequencing through smaller repetitive regions of the same nucleotide (homopolymer regions) on the order of 5 to 10 bases can prove challenging. Ion Torrent has reported sequencing accuracy data in which an *E. coli* DH10B sample was sequenced and homopolymer regions were analyzed (Ion Torrent Application

Note, Spring 2011). The sequencing accuracy for a 5-mer-homopolymer region was shown to be around 97.5%, however it was difficult to tell the size of sample set from which these data were generated. Also, accuracy data for homopolymer lengths greater than 5 bases were not reported.

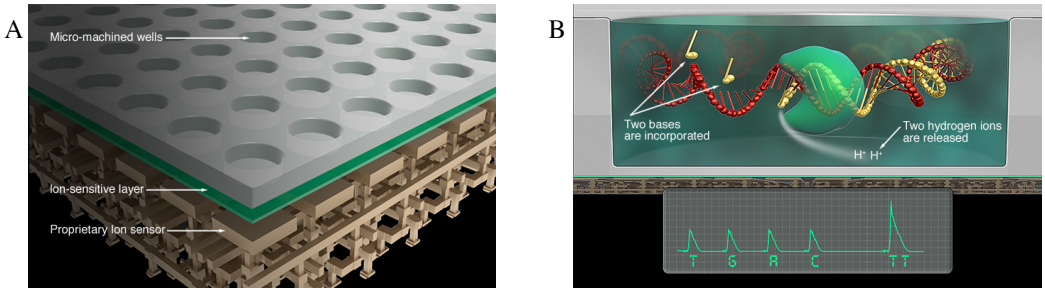


Figure 1-4. *Layout of Ion Torrent's semiconductor sequencing chip technology. (a) A layer-by-layer view of the chip revealing the structural design. The top layer contains the individual DNA polymerization reaction wells and the bottom two layers comprise the FET ion sensor. Each well has a corresponding FET detector that identifies a change in pH. (b) A side view of an individual reaction well depicting DNA polymerase incorporation of a repeat of two TTP nucleotides on a sequencing fragment. The hydrogen ions released during this process are detected by the FET below. Image reproduced with permission from Ion Torrent.*

1.8 Nanopore Sequencing Technologies

A fundamentally different class of sequencing technology is based on nanopore structures, described in prior reviews by Branton [70] and Bayley [71]. Individual base detection was envisioned to be possible through the measurement of

conductivity either across or through a membrane, via a nanoscale pore. These nanopores consist of an orifice slightly larger than the width a double-stranded DNA molecule, which is 4 nm, where DNA is threaded through the pore. The chemical differences of each base would result, in theory, in detectably altered current flow through the pore. Theoretically, nanopores could also be designed to measure tunnelling current across the pore as bases, each with a distinct tunnelling potential, could be read. The nanopore approach, while still in development, remains an interesting potential fourth-generation technology. This "fourth-generation" moniker is suggested, since optical detection is eliminated along with the requirement for synchronous reagent wash steps.

Nanopore technologies may be broadly categorized into two types, biological and solid state. The protein alpha hemolysin, which natively bridges cellular membranes causing lysis, was first used as a model biological nanopore. The protein was inserted into a biolayer membrane separating two chambers while sensitive electronics measured the blockade current, which changed as DNA molecules moved through the pore. However, chemical and physical similarities between the four nucleotides made the sequence much more difficult to read than envisioned. Further, sufficient reduction of electronic noise remains a constant challenge, which is achievable in part by slowing the rate of DNA translocation. Recently, Oxford Nanopore and several other academic groups working on this concept have made progress toward addressing these challenges.

The second class is based on the use of nanopores fabricated mechanically in silicon or other derivative. By transferring to a manufactured material, Oxford Nanopore's difficulties with membrane stability, protein positioning, number, or

order could be sidestepped. For example, Nabsys created a system using a silicon wafer drilled with nanopores using a focused ion beam (FIB), which detects differences in blockade current as single-stranded DNA bound with specific primers passes through the pore. IBM created a more complex device that aims to actively pause DNA translocation and interrogate each base for tunneling current during the pause step. The technology for both of these nanopore types is presented below.

John Kasianowicz and colleagues [72] were the first to show the translocation of polynucleotides (poly[U]) through a *Staphylococcal* α-hemolysin (αHL) biological nanopore suspended in a lipid bilayer, using ionic current blockage method. The authors predicted that single nucleotides could be discriminated as long as: 1. each nucleotide produces a unique signal signature; 2. the nanopore possesses proper aperture geometry to accommodate one nucleotide at a time; 3. the current measurements have sufficient resolution to detect the rate of strand translocation; 4. the fragment should translocate in a single direction when potential is applied; and 5. the nanopore/supporting membrane assembly should be sufficiently robust. All of the biological and synthetic nanopores have barrels of ~ 5 nm (which is considerably longer than the base-to-base distance of 3.4 Å) in thickness and accommodate ~ 10-15 nucleotides at a time. It is, therefore, impossible to achieve single-base resolution using blockage current measurements. In addition, the average rate at which a polymer typically translocates through a nanopore is on the order of 1 nucleotide/μsec (*i.e.*, on the order of MHz detection), which is too fast to resolve. The nucleotide strand should be slowed down to ~ 1 nucleotide/msec to allow for a pA-current signal at 120-150 mV applied potential [73]. Furthermore,

the translocation of a polymer strand should be uniform between two events. The time distribution of two processes (capture, entry, and translocation) is non-Poisson and often differs by an order of magnitude. This means that two molecules pass through a nanopore at considerably different rates and the slower one could be missed or misinterpreted. Andre Marziali and co-workers at UBC [74, 75] used force spectroscopy to study these events through single-molecule bond characterization. The non-uniform kinetics of DNA passage through an αHL nanopore is attributed to weak binding of DNA to amino acid residues of the protein nanopore [76].

Because of these challenges with ionic current measurements (the current created by the flow of ions through the nanopore), researchers have looked at other measurement schemes such as the detection of tunneling current and capacitance changes (1-5A). In transverse tunneling current scheme, electrodes are positioned at the pore opening and the signal is detected from sub-nanometer probes [77]. In capacitance measurements, voltage is detected across a metal oxide-silicon layered structure. The voltage signal is induced across the capacitor by the passage of charged nucleotides in longitudinal direction [78]. A different readout approach is optical detection (1-5B). A typical optical recognition of nucleotides is essentially executed in two steps. First, each base (A, C, G, or T) in the target sequence is converted into a sequence of oligonucleotides, which are then hybridized to two-color molecular beacons (with fluorophores attached) [79]. Because the four nucleotides (A, C, G, or T) have to be determined, the two fluorescent probes are coupled in pairs to uniquely define each base. For example, if the two probes are A and B, the four unique permutations will be AA, AB, BA, and BB. As the

hybridized DNA strand is threaded through the nanopore, the fluorescent tag is stripped off from its quencher and an optical signal is detected. Both protein [80] and solid-state nanopores can be used [79, 81]. Detailed electronic measurement schemes and optical readout methods have been reviewed in more detail in previously published papers [70, 82, 83].

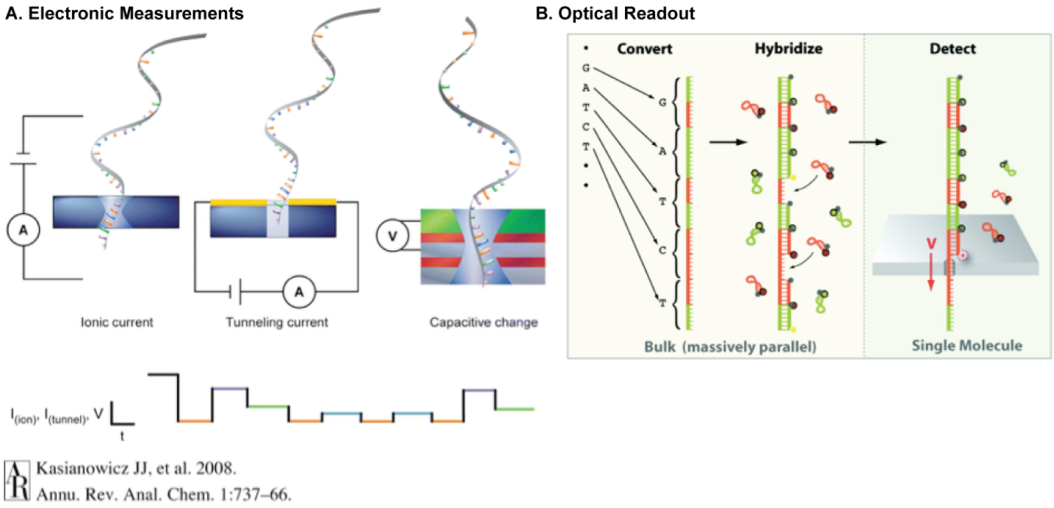


Figure 1-5. *Electronic measurements and optical readout. (a) Nanopore DNA sequencing electronic schemes. Signal is obtained through ionic current [72], tunneling current [77], and voltage difference [78] measurements. Each of the schemes must produce a characteristic magnitude of the signal, such that the four DNA bases could be distinguished. (Part A is reprinted with permission. [82]) (b) DNA sequencing through optical readout [81]. Each nucleotide from the target sequence is converted to a known oligonucleotide sequence, which is subsequently hybridized with molecular beacons. In the detection step, the hybridized DNA strand is threaded through a nanopore in such a way that molecular beacons are released. (Part b is reprinted with permission. [81])*

In a 2008 review article [70], Daniel Branton and colleagues discussed the nanopore sequencing development and the prospect of low sample preparation cost at high throughput. They estimated that purified genomic DNA sufficient for sequencing (~10^8 copies or 700 μg) could be extracted and purified from blood at a cost of less than $40/sample using commercial kits. All existing sequencing techniques require breaking the DNA into small fragments of ~100 bps, and sequencing those chunks multiple times to find overlapping regions, so that they can be reassembled together. Because one of most appealing advantages of nanopores is achieving long read lengths, the genomic assembly process should be considerably simplified. In practice, only the DNA shearing, which occurs during pipetting in the sample preparation step, may limit the read length. For example, Meller and Branton [84] demonstrated that 25 kb ssDNA could be threaded through a biological nanopore, and 5.4 kb ssDNA translocated through a solid-state nanopore. In addition, several groups have shown very high throughput of small oligonucleotides (~5.8 oligomers (sec μM)$^{-1}$) [84], and native ssDNA and dsDNA (~3-10-kb at 10-20 nM concentration) [85, 86].

1.8.1 Protein Nanopore Sequencing

Oxford Nanopore Technologies, formerly Oxford Nanolabs, together with leading academic collaborators, has addressed some of the aforementioned technological challenges and implemented the nanopore technology in a commercial product (GridION system). Oxford Nanopore, founded by Prof. Hagan Bayley at University of Oxford, aimed at commercializing the research work on biological nanopores coming out of his laboratory. The company works in collaboration with

Professors Daniel Branton, George Church and Jene Golovchenko at Harvard; David Deamer and Mark Akeson at UCSC, and John Kasianowicz at NIST.

In a recent editorial article in The Economist [87], Gordon Sanghera, chief execute officer of Oxford Nanopore, announced that the company is preparing to launch the GridION system for direct single-molecule analysis which would adopt exonuclease sequencing. The system is based on "lab on a chip" technology and integrates multiple electronic cartridges into a rack-like device. A single protein nanopore is integrated in a lipid bilayer across the top of a microwell, equipped with electrodes. Multiple microwells are incorporated onto an array chip and each cartridge holds a single chip with integrated fluidics and electronics for sample preparation, detection and analysis. The sample is introduced into the cartridge, which is then inserted in an instrument called a GridION node. Each node can be used separately or in a cluster and all nodes communicate with each other and with user's network and storage system in real time. Although the main application of the platform is sequencing of DNA, it can be adapted (by proper modification of the αHL nanopore) for the detection of proteins and small molecules.

Oxford Nanopore's first-generation systems utilize the heptameric protein α-hemolysin (αHL) (Figure 1-6.a) [71, 88-90]. αHL is secreted from bacteria, providing low-cost production of these robust bionanopores. Oxford Nanopore is working towards commercialization of two types of sequencing methodologies: exonuclease sequencing and strand sequencing. In the exonuclease method [91], a cyclodextrin adapter molecule (Figure 1-6.b) is bound to the inside of a protein nanopore, and serves as a DNA binding site. The nanopore is additionally coupled with an exonuclease [92], a processive enzyme that cleaves individual bases from

the DNA strand and allows for accurate detection as DNA bases pass through and interact with the cyclodextrin (Figure 1-6.c). Progressive enzymes positioned on the top of the nanopore regulate the translocation rate of the DNA strand by slowing down (to the order of msec) the intrinsic electrophoretic motion (to the order of μsec) [92]. Essentially, one nucleotide passes through the nanopore approximately every 20 msec, which is slow enough for accurate detection. The four nucleotides produce different magnitudes of current disruption (Figure 1-6.d) and, therefore, the determination of DNA sequence is possible.

Oxford Nanopore is also working towards the development of a strand sequencing technology, in which a single-stranded DNA fragment is passed through the pore and identification of single bases is achieved as they pass through the pore (Figure 1-6.e) [93]. This method is potentially faster and more accurate than exonuclease sequencing. Because all nucleotides are attached to each other, there is no chance of reading them in the wrong order, however the challenge lies in achieving the accurate identification of the individual bases as they pass through the nanopore.

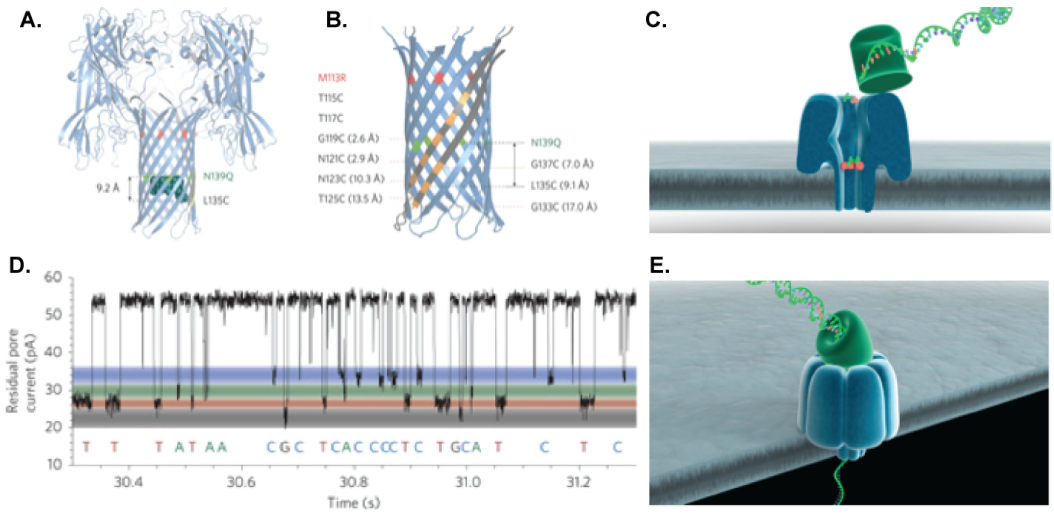


Figure 1-6. *The biological nanopore scheme employed by Oxford Nanopore. (a) Schematic of αHL protein nanopore mutant WT-(M113R/N139Q)$_6$ (M113R/N139Q/L135C)$_1$. The cartoon picture shows the positions of the cyclodextrin (at residue 135) and glutamines (at residue 139). [90] (b) A detailed view of the β barrel of the mutant nanopore shows the locations of the arginines (at residue 113) and the cysteines. The mutants are listed to the left of the figure using standard single-letter amino-acid codes. [90] (c) Exonuclease sequencing: A processive enzyme is attached to the top of the nanopore. Its function is to cleave single nucleotides from the target DNA strand and pass them through the nanopore. (Image obtained from Oxford Nanopore Technologies, Ltd with permission.) (d) Residual current-vs-time signal trace from WT-(M113R/N139Q)$_6$(M113R/N139Q/L135C)$_1$-am$_6$amDP$_1$βCD pore. The trace shows a clear discrimination between single bases (dGMP, dTMP, dAMP and dCMP). The width of each colored band is three standard deviations from the mean of the signal, fitted to a Gaussian. [90] (e) Strand sequencing: ssDNA is threaded through a protein nanopore and individual bases are identified, as the strand*

remains intact. (Image obtained from Oxford Nanopore Technologies, Ltd with permission).

1.8.2 Solid-State Nanopore Sequencing

Although the αHL heptamer pores are rather robust [94], the lipid bilayers on which biopores are suspended are unstable and hard to manipulate. Solid-state or man-made nanopores are considered to be next-generation nanopore technology, because they bypass the use of organic supports and are thus, in principle, more stable. Also solid-state nanopores could be multiplexed to work in parallel on a single device, which is a challenge for biological nanopores. Artificial pores are fabricated in solid-state materials such as silicon nitride, silicon or metal oxides, and more recently graphene. Graphene is a new, single-atom thick material, which is known to be the thinnest possible membrane. A group at the University of Pennsylvania led by Marija Drndic [95] presented translocation measurements of DNA through graphene nanopores, which comprised 1-5 nm thick membranes with 5-10 nm diameter pores (Figure 1-7.a). In another publication [96], researchers in the Golovchenko lab at Harvard showed that a graphene sheet can be used as a membrane material that holds a solid state nanopore and separates two chambers of ionic solution (Figure 1-7.b).

IBM currently develops a novel approach to DNA sequencing through artificial nanopores in solid-state material (specifically a metal-dielectric layered structure), in collaboration with 454 Life Sciences (currently Roche). The idea originated with systems biologist Gustavo Stolovitzky and electrical engineer Stanislav Polonsky at IBM in 2006. Three-nm artificial nanopores are fabricated by e-beam drilling in 10 nm thick membranes made of titanium nitride, separated by insulating layers of

silica. As the DNA strand is drawn through the nanopore, electrical field across the metal layers can be flipped, also referred to as the ratchet effect, resulting in immobilization and, subsequently, in principle, controlled motion of the DNA strand (Figure 1-7.c). Alternation of the electric field can be potentially beneficial for improving of sequencing accuracy. Two possibilities for detecting the signal are capacitance or ionic current measurements (similar to Oxford Nanopore detection, except that the DNA strand will remain intact). To obtain a strong enough signal, the DNA needs to be trapped for interrogation only for a millisecond. Most of the work reported by the IBM group has been numerical, so far, through MD simulations [97, 98]. Although 5 to 7 years of further development is expected to be required [28] before any commercial release, the idea of electronic detection combined with easy sample preparation offers the exciting potential for very cheap sequence readout.

Despite the challenges of achieving single-base resolution by means of current blockage measurements through a man-made nanopore, a number of groups have easily distinguished a translocation of ssDNA from dsDNA in a nanopore wide enough to accommodate the double-strand [86, 99, 100]. Because coarse-grained resolution is easily achieved, researchers started looking into the creation of *de novo* sequencing techniques, by attaching hybridization probes to DNA fragments. Recently, Balagurusamy *et al.* [99] experimentally showed a translocation and successful electrical detection of two consecutive 12-mer long double strands through a nanopore drilled in a 20 nm-thick silicon nitride membrane. Another solid-state nanopore study [100] reported the detection of dsDNA hybridized with peptide nucleic acid (PNA) probes threaded through a sub-5 nm nanopore in 30-

nm thick membrane. These studies are promising for the realization of the sequencing by hybridization [101] (SBH) through nanopores, also known as hybridization-assisted nanopore sequencing [102] (HANS) method. This technology has been licensed by NABsys which is a DNA sequencing start-up company founded by Brown physics professor Sean Ling in 2005. The company is working towards the development and commercialization of a computer chip to "electronically read" DNA. In practice, 6mer hybridization probes will be bound to 100-kb genomic fragments, which are then driven electrophoretically through a solid-state nanopore, creating a current signal (Figure 7.d). Based on this current tracing, the positioning of the probes and thus, the sequence of small fragments are determined. The process is done in parallel with an entire library of probes, which will in principle allow the assembly of the whole genome length and readout. The company promises an eventual four order of magnitude decrease in the cost of whole genome sequencing. However, an open question for the HANS technology is the sufficient resolution for accurate readout of the electronic signal.

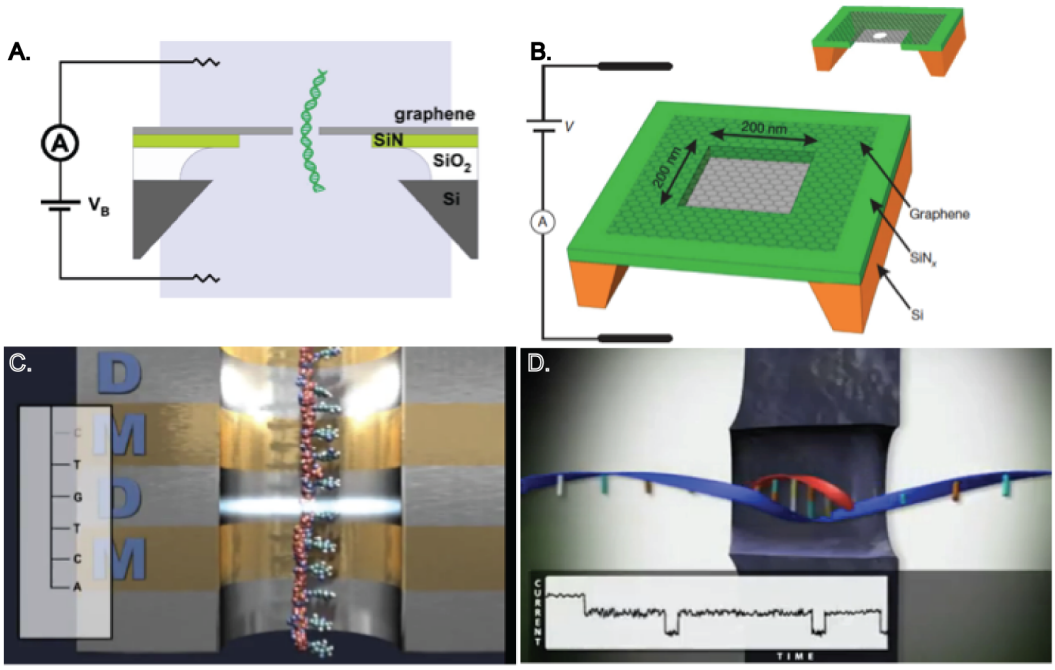


Figure 1-7. *Several synthetic nanopore sequencing device designs. (a) The device consists of 1-5 nm thick graphene membrane, which is suspended in a Si chip coated with 5 μm SiO_2 layer. It is placed in a PDMS cell with microfluidic channels on both sides of the chip [95]. (b) A nanopore (shown in the inset to the figure) is drilled through a graphene membrane, which is suspended in SiN_x across a Si frame. The graphene membrane separates two ionic solutions and is in contact with Ag/AgCl electrodes [96]. (c) IBM DNA transistor setup. A nanometer sized pore is fabricated by using an electron beam. Electric field is created between the gated regions allowing for charge trapping. Hence, a DNA molecule is immobilized and its translocation is slowed providing enough time for measurement of individual bases. The substrate is composed of metal and dielectric regions, labeled with ''M'' and ''D'', respectively. (Image obtained from IBM with permission). (d) HANS method adopted by NABsys for electronic readout of DNA fragments through solid-state nanopores. 6-mer probes are hybridized to ssDNA fragments and current-verses-time trace is detected. This results in a small*

areas of known sequencing (because of base-pairing), which are then lined up to create a map for the genome. The process is done in parallel for an entire library of probes and the whole genome length is mapped. (Image obtained from NABsys, Inc. with permission).

1.9 Long read DNA extension methods

While short-read methods that rely on DNA fragments that are less than 400 bases long constitute the bulk of the current DNA sequencing technologies, several different, new approaches aim to sequence DNA up to several megabases in length. Recent reports have highlighted the limitations of short-read technologies for genome assembly of prokaryotes [103]. Mapping of more extended DNA regions can provide data on the number of repeats, deletions, insertions and transpositions that are unobtainable with any of the currently available short-read methods.

1.9.1 Final Assembly by Optical Mapping

At the University of Wisconsin-Madison, Prof. David C. Schwartz and colleagues have created the only system (Optical Mapping) available to date with the capacity to strategically guide, validate, and complete the sequence assembly of whole, complex genomes. The Optical Mapping System constructs genome-wide, long-range, ordered restriction maps from large datasets comprising 5,000 - 2,000,000 individual genomic DNA molecules (~500 kb), “bar-coded” by restriction digestion and directly imaged by fluorescence microscopy. This highly automated

system is the first single-molecule platform proven capable of whole genome analysis [104]. The Optical Mapping System boasts computational tools that include alignment capabilities [105], which position nascent sequence assemblies onto *de novo* optical maps [106] spanning entire genomes. Such alignments place orphan sequence assemblies, order and orient scaffolds and contigs, size gaps, and reveal assembly errors, in addition to accurate accounting of chromosome number and sizes. Early applications of Optical Mapping have centered on bacterial [107] and lower eukaryotic genomes [108, 109]; however, more recently, Optical Mapping analysis has successfully guided the assembly and validation of complex genomes that included rice [110] and maize [111-114], which is the most complex genome ever sequenced. Because very large ~500 kb genomic DNA molecules are analyzed, complex genomic regions near centromeres, or those rife with segmental duplications become measurable to reveal new structural variants not approachable by sequencing. This advantage allows discovery of many new structural variants as insertions, or complex rearrangements within human genomes [115, 116] that confound sequencing approaches and portend significant analytical approaches for dissecting breakpoints and rearrangements in cancer genomes (unpublished, Schwartz, *et al.*).

The Schwartz laboratory has further advanced genome mapping approaches through the addition of sequence reads to long double-stranded molecules and the development of the Nanocoding System [117]. Nanocoding uses genomic analytes and within a single reaction mix, nicking restriction enzymes selectively clip only one strand of the double stranded DNA, at cognate recognition sites. Newly created nicks are then tagged by polymerase-mediated nick translation using

fluorochrome labeled nucleotides. Unique single-molecule barcodes emerge because the end products are full-length dsDNA molecules precisely decorated by fluorochromes at each enzyme recognition site. Decorated DNA molecules are loaded into a microfluidic chip to flow into channels on the order of 50 μm in size. These channels are bisected at a 45° angle by nanofluidic channels 1 μm wide and 100 nm deep. The combination of the microfluidic-nanofluidic channel angle and the nanofluidic channel width significantly reduce the entropic penalty required to fully stretch the DNA from a coiled form, while low ionic strength buffers greatly enhance molecular stretching within the nanoslits. Once stretched in the channel, fluorescence imaging [FRET: Fluorescence Resonance Energy Transfer: intercalated YOYO-1 dye (Donor) and Alexa Fluor 647 (acceptor)] and machine vision identify the locations of covalently incorporated fluorochromes for the construction of single-molecule barcodes that are assembled into genome-wide physical maps.

A second company using nanofluidics, BioNanomatrix, was established with technology licensed from Princeton University. Their design also uses nanofluidic channels to stretch the DNA with a modified channel entrance design. The channels are on the order of 100 nm or less in width as well as in depth. To overcome the entropic barrier to entry, the channel transitions from a microns deep to nanometers deep using a lithography pattern to introduce a pillar type pattern that gradually forces the DNA to uncoil and extend into the nanochannels for imaging [56-62]. These chips may also contain constrictions to force the DNA through a narrow gap. The BioNanomatrix chip has been used with formamide and controlled localized heating in the presence of YOYO-1 to partially denature the

DNA and infer a sequence from a pattern [118]. A second technique identified landmarks on λ-DNA using a nicking enzyme to displace a recognition site on, backfill with nucleotides, and subsequently hybridize the displaced strand with a fluorophore-labeled probe [119]. A camera and imaging software were used for analysis. Of the 300 molecules imaged in 30 seconds, 85% of the two-targeted sites were properly labeled.

1.9.2 Non-optical, stretched DNA molecule methods

The methods discussed here still stretch DNA over a surface, which is probed to read each individual base, but eschews the use of a video camera completely, for atomic imaging methods. Halcyon Molecular is a fourth-generation technology that relies on rapid-scan tunneling electron microscope (TEM) method [63]. Individual DNA bases are labeled with distinct heavy atoms to differentiate between each base as described generally here [120]. ZS Genetics, where ZS stands for "Zero Science", has also pursued a TEM method, but has yet to publish detailed methods or results. A Scanning Tunneling Microscopy (STM) was reportedly used to identify guanine from non-guanine bases [121]. STM measures the density of electron flow through a scanning tip. Although 140 bases were read and compared to a reference gene sequence, a number of limitations, most notably speed, currently prevents commercial viability. A recent review [122] describes these techniques in greater detail.

CONCLUDING REMARKS

Technology and funding in the field of novel DNA sequencing technologies have been growing at a rate never before seen. As discussed in this review, there has been a proliferation of vastly different approaches to DNA sequencing, across all generations of the newer technologies. Each technique has its own advantages and limitations; so, ultimately the specific genotyping application must be evaluated in order to choose the appropriate sequencing platform. While second- and third-generation platforms boast considerable throughput, Sanger-based CE sequencing is still the gold standard for ultra-high-accuracy sequencing, and is the only technique that has so far provided both *de novo* sequencing and the *de novo* assembly of a human genome. In order to gain wide-spread recognition as the front-running next-generation sequencing technology, one of the second- or third-generation platforms must provide a side-by-side study with a first-generation CE-based platform and quantitatively compare the outcomes of the sequencing and assembly of the same *de novo* sample. This will provide concrete evidence of the true cost of *de novo* sequencing and will serve as the jumping off point from which current and future researchers can make decisions on how to tackle the next wave of human genome sequencing projects, or the de novo sequencing of similarly sized complex genomes. Currently, based on the current limitations of technologies, it appears that several of these technologies must be used in tandem to achieve the benefits of high-throughput, accuracy, long contiguous read lengths, and long-range mapping that would be needed to catalog the complete such a complex genome, *de novo*.

Table 1-1. *Summary of first and second generation sequencing technologies*

Generation	Company	Platform Name	Method of Sequencing	Method of Detection	Approx. Read Length (bases)	Advantages	Relative Limitations
First	ABI/Life technologies	3130xL -3730xL	CE-Sanger	Fluorescence/Optical	600-1000	Long read lengths; high single-pass accuracy; good ability to call repeats and homopolymer regions	Low throughput; high cost of Sanger sample preparation makes massively parallel sequencing prohibitive
First	Beckman	GeXP Genetic Analysis System	CE-Sanger	Fluorescence/Optical	600-1000	Long read lengths; high single-pass accuracy; can call repeats/homopolymer regions; scales down well	Low throughput; relatively high per-sample cost of Sanger sample preparation
Second	Roche/454	Genome Sequencer FLX System	Pyrosequencing	Optical	230-400	Longest read lengths among second generation; high throughput compared to first-generation sequencing	Challenging sample prep; Difficulty reading through repetitive/homopolymer regions; sequential reagent washing gives steady accumulation of errors; expensive instrument ($500K)

Table 1-2. *Summary of first and second generation sequencing technologies.*

Second	Illumina	HiSeq 2000/ miSeq	Reversible terminator sequencing by synthesis	Fluorescence/Optical	2 x 150	Very high throughput	Expensive instrument; significant cost of data reduction and analysis (~$650K)
Second	ABI/SOLiD	5500xl SOLiD System	Sequencing by ligation	Fluorescence/Optical	25-35	Very high throughput; lowest reagent cost needed to reassemble a human genome among the widely accepted second generation platforms (Illumina, 454, SOLiD)	Long sequencing runs (days); short reads increase cost and difficulty of data analysis and genome assembly; high instrument cost (~$700K)
Second	Helicos	HeliScope	Single-molecule sequencing by synthesis	Fluorescence/Optical	25-30	High throughput; single-molecule nature of technology unique among second-gen platforms	Short reads increasing the costs and reduce quality of genome assembly; very costly instrument (~$1M)

Table 1-3. *Summary of next-generation sequencing technologies.*

Generation	Company	Platform Name	Method of Sequencing	Method of Detection	Read Length (bases)	Advantages	Limitations
Third	Pacific Biosciences	PacBio RS	Real-time, single-molecule DNA sequencing	Fluorescence/Optical	~ 1000	Long average read lengths; decreased sequencing time compared to first-gen platforms; no amplification of sequencing fragments; longest individual reads approach 3,000 bases	Inefficient loading of DNA polymerase in ZMWs; low single-pass accuracy (81-83%); degradation of the polymerase in ZMWs; overall, high cost per base (expensive instrument)
Third	Complete Genomics	In-house lab-built instrumentation	Combinatorial probe-anchor hybridization and ligation (cPAL)	Fluorescence/Optical	10	Highest (claimed) throughput of third-gen platforms; lowest reagent cost for reassembling a human genome of all sequencing technologies; each sequencing step is independent, minimizing accumulation of errors	Short read lengths; template preparation prevents sequencing through long repetitive regions; labor intensive sample preparation; no commercially available instrument
Third	Ion Torrent/Life Technologies	Personal Genome Machine (PGM)	Sequencing by synthesis	Change in pH detected by Ion-	100-200	Direct measurement of nucleobase incorporation	Sequential washing steps can lead to accumulation of

Table 1-4. *Summary of next-generation sequencing technologies.*

Third	Ion Torrent/Life Technologies	Personal Genome Machine (PGM) sequencer	Sequencing by synthesis	Change in pH detected by Ion-Sensitive Field Effect Transistors (ISFETs)	100-200	Direct measurement of nucleobase incorporation events; DNA synthesis reaction operates under natural conditions (no need for modified DNA bases)	Sequential washing steps can lead to accumulation of errors; potential difficulties in reading through highly repetitive or homopolymer regions of the genome
Fourth	Oxford Nanopore	gridION	Nanopore exonuclease or DNA strand migration	Current	Not yet quantified	Potential for long read lengths; low cost of αHL nanopore production; no fluorescent labeling or optics necessary	Cleaved nucleotides may be read in the wrong order; difficult to fabricate a device with multiple parallel pores.

2 ELECTROOSMOTIC MOBILITY

EFFECT OF POLYVINYLPYRROLIDONE (PVP) ON THE ELECTROOSMOTIC MOBILITY OF WET-ETCHED GLASS MICROCHANNELS

Several sections of this chapter are based on an article by Denitsa Milanova, Robert D. Chambers, Supreet S. Bahga, and Juan G. Santiago, named "*Effect of polyvinylpyrrolidone (PVP) on the electroosmotic mobility of wet-etched glass microchannels*" in Electrophoresis [123], and are reproduced here with minor modifications.

We present an experimental study on the effect of polymer polyvinylpyrrolidone (PVP) on electroosmotic flow (EOF) mobility of microchannels wet etched into optical white soda lime glass, also known as Crown glass. We performed experiments to evaluate the effect of polyvinylpyrrolidone (PVP) concentration and pH on EOF mobility. We used on-chip capillary zone electrophoresis and a

neutral fluorescent dye as a passive marker to quantify the electroosmotic flow. We performed experiments under controlled conditions by varying pH from 5.2 and 10.3 and concentration of PVP from 0 to 2.0% w/w at constant ionic strength (30 mM). Our experiments show that PVP at concentrations of 1.0% or above very effectively suppress EOF at low pH (6.6). At high pH of 10.3, PVP has a much weaker suppressing effect on EOF and increasing its concentration above about 0.5% showed negligible effect on EOF mobility. Lastly, we briefly discuss the effects of pH on using PVP as an adsorbed coating. Our experiments provide useful guidelines on choosing correct pH and concentration of PVP for effective EOF suppression in glass channels.

2.1 Introduction

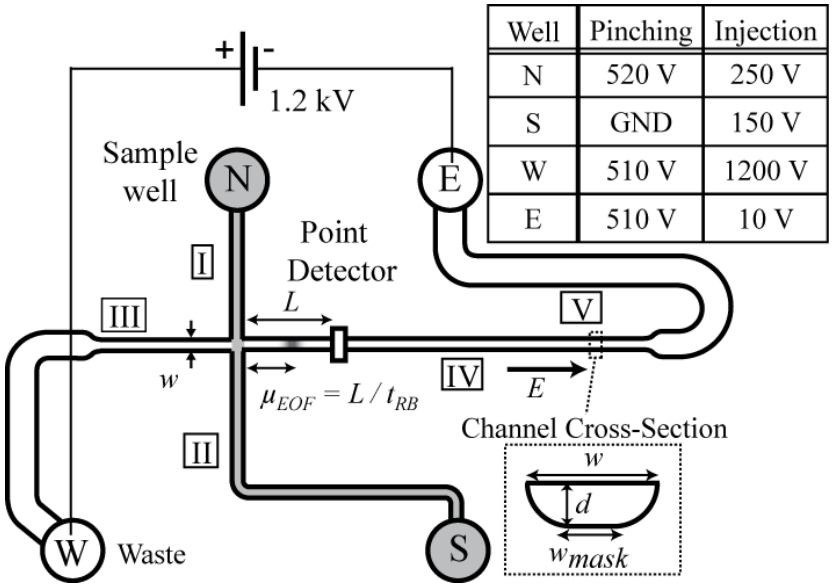
Polyvinylpyrrolidone (PVP) is used in many applications to suppress protein adsorption [124], extract analytes [125], or separate biomolecules [126]. PVP is also used as a simultaneous dynamic coating and sieving matrix for DNA separations [127, 128]. Most relevant here, PVP is used to control and suppress electroosmotic flow (EOF) in electrophoresis based techniques, including capillary zone electrophoresis (CZE) [129, 130] and isotachophoresis (ITP) [131, 132]. Improved coating capacity has been reported for PVP used as an outer layer wall coating for EOF suppression [133]. For example, a glass surface can be silanized and PVP subsequently adsorbed as an upper polymer layer by the hydrogen abstraction method [134]. Further, PVP has been used as a dynamic coating for

composite glass-and-polydimethylsiloxane (PDMS) [135] or glass-on-silicon (substrate) chips [136] for PCR systems.

We here focus on using PVP's use as a dynamic surface passivation additive. Dynamic coating (including the additive in the buffer of interest) is simple and cost-effective; thus attractive across a broad range of applications. Further, aqueous PVP solutions have relatively low viscosity [137] and, therefore, facilitating buffer loading into capillaries and on-chip. To our knowledge, the only study to systematically evaluate PVP as a dynamic coating for EOF suppression was that of Kaneta *et al.* [129] who employed it for EOF suppression in a standard capillary zone electrophoresis (CZE) experiment in silica capillaries. Kaneta *et al.* [129] reported a 10-fold reduction of EOF at 1.0% concentration of PVP compared to untreated silica capillary. They varied PVP concentration, PVP molecular weight, and pH, and studied EOF mobility in the presence of sodium dodecyl sulfate. Despite its popularity and common use as an EOF suppressant in on-chip capillary electrophoresis, we know of no systematic experimental study on the effects of PVP concentration and pH on the glass surfaces of microchannels. Here we report EOF mobility measurements taken in microchannels etched in white soda lime at varying PVP concentrations and pH values. We used a neutral marker [138] (Rhodamine B fluorescent dye) and straightforward electrokinetic injection experiments.

2.2 Materials and methods

We here only summarize our materials and methods while details can be found elsewhere in Milanova *et al.* [130]. We performed standard on-chip CZE experiments and used a neutral dye rhodamine B (RB) (Sigma-Aldrich, St. Louis, MO) at 200 μM to quantify EOF mobility. We explored the 16 background electrolyte chemistries listed in Table 2-1. We prepared buffer solutions of glycine (chemistries 1-4), tricine (5-8), MES (9-12), and acetic acid (13-16) titrated with sodium hydroxide to measured pH values of 10.3, 8.5, 6.6, and 5.2 respectively. We explored PVP (MW = 1,000,000, Polysciences Inc., Warrington, PA) concentrations of 0%, 0.5%, 1.0%, and 2.0% w/w of water. We here refer to PVP concentration using the usual designation as percentage concentration by weight (PVP density is 1.2 g/cm^3, so 1.0% PVP corresponds to 12 mg/ml at 25^{o}C). We used a Corning Pinnacle 542 pH/conductivity meter (Nova Analytics, Woburn, MA) for pH measurements. Predicted pH values were from the Peakmaster tool [139] and often differed by ~0.1-0.2 pH units from measured values, possibly due to the effect of ionic strength on acid dissociation constants [140] and experimental error.



Well	Pinching	Injection
N	520 V	250 V
S	GND	150 V
W	510 V	1200 V
E	510 V	10 V

Figure 2-1. *The experimental apparatus for capillary zone electrophoresis includes microfluidic chip, epifluorescence microscope, CCD camera, high voltage switching system, 1.2kV DC power supply, and a data acquisition system. We used a 10× objective (numerical aperture of 0.4) for all experiments. We used exposure times between 50 and 100 ms, depending on fluorescence signal strength. The chip was a cross type Caliper NS 95 with 12 μm etch depth and 10 μm mask width in the separation channel. Precise measurements of channel center contour lengths of various regions are: 5.0 mm (I), 16.3 mm (II), 8.4 mm (III - the separation channel), 16.1 mm (IV), and 4.3 mm (V). We used rhodamine B as a neutral dye loaded into the north reservoir (N). The inset table summarizes an empirically optimized voltage scheme for sample stream pinching and injection. Our main voltage (from W to E) was 1.2 kV, yielding an electric field of 29.4 V/cm in the separation channel, oriented left to right.*

Details of our optical and experimental setup are provided in Milanova *et al.* [130], and depicted here in Figure 2-1. Briefly, we performed measurements on an inverted epifluorescent microscope (IX70, Olympus, Hauppauge, NY) equipped with a mercury lamp, a U-MWIBA filter cube from Olympus (460-490 nm excitation, 515 nm emission) and a 10× (NA of 0.4) UPlanApo objective. We imaged a region only 1.5 mm into the separation channel (~1/10th of the separation channel length) as longer lengths results in overly weak signals at conditions of strongly suppressed EOF (migration times were as long as order 10^3 s). We analyzed images to quantify the migration time, t_{EOF}, for the maxima of the neutral dye peak. We quantified the local electric field, E, (~29.4 V/cm) in the separation channel by treating electrolyte filled channels as a network of resistors to compute the relation between channel geometries, applied potentials, and electric fields. For more details, see Section 3 of this document.

For uniform conductivity buffers with negligible pressure disturbances, EOF mobility μ_{EOF} can be expressed simply in terms of the EOF velocity V_{EOF} as

$$\mu_{EOF} = \frac{v_{EOF}}{E} = \frac{L}{Et_{EOF}} \qquad \textbf{2. 1}$$

where E is electric field, L is the travel distance (here L = 1.5 mm), and t_{EOF} is migration time. Our white soda lime glass microchips had a negative surface at all conditions, resulting in positive EOF in the direction of electric field.

The essence of the passive, neutral marker method is that the neutral fluorescent marker has negligible electrophoretic mobility and does not interact with the background electrolyte or channel walls. We observed no evidence (e.g., peak tailing, background fluorescence) of wall adsorption. Cationic RB fits these criteria

as it is suitably soluble, has a reported pK_a value of 3.22 [141], and is approximately neutral for our pH range of interest (5.2 < pH < 10.3).

We described our injection protocol in the Supplementary Information of [130]. Briefly, we used a "pinching" step to inject a finite amount of sample in the separation channel, and application of electric field in the separation channel included a "retraction" step to quickly interrupt sample flow into the channel. We show a chip schematic and summarize the two-stage voltage control scheme in Figure 2-1. Our chip was an off-the-shelf optical white soda lime glass microfluidic chip (model NS-95 from Caliper Life Sciences, Mountain View, CA), composed primarily of SiO_2 (69.5%), K_2O (8.3%), Na_2O (8.1%), CaO (7.1%), and several other minor oxide additives (McReynolds, R.J., Caliper Life Sciences, personal communication, June 19, 2012). Before introducing each new background electrolyte chemistry, we flushed the channels with 40 μL of 0.5 M NaOH for 10 min by applying vacuum to well labeled *S* in Figure 2-1, followed by deionized water for 5 min, then 100 mM HCl for 3 min, and then deionized water again for 3 min. Between repeats of the same buffer chemistries, we only rinsed with deionized water.

Table 2-1. *Details of background electrolyte buffer composition in our electroosmotic mobility study. In parenthesis, we list respectively species valence, absolute mobility as factors of 10^{-9} $m^2V^{-1}s^{-1}$, and acid dissociation constants (pK_a).*

	Measured, Predicted pH	C (acid) [mM]	C (base) [mM]	I [mM]
		Glycine (+1, 39.5, 2.32), (-1, 37.4, 9.78)	NaOH	
1	10.3, 10.11	40	30	30
2	9.8, 9.64	60	30	30

3	9.4, 9.16	120	30	30
		Tricine (-1, 26.6, 8.5)	NaOH	
4	8.5, 8.49	40	30	30
5	8.0, 8.00	60	30	30
6	7.6, 7.53	120	30	30
		HEPES (-1, 21.8, 7.5)	NaOH	
7	8.3, 7.84	40	30	30
8	7.7, 7.36	60	30	30
9	7.2, 6.88	120	30	30
		MES (-1, 26.8, 6.13)	NaOH	
10	6.6, 6.43	40	30	30
11	6.1, 5.95	60	30	30
12	5.7, 5.48	120	30	30
		Acetic Acid (-1, 42.4, 4.756)	NaOH	
13	5.2, 5.09	40	30	30
14	4.7, 4.62	60	30	30
15	4.2, 4.14	120	30	30

We note the ambient temperature for our experiments varied between 21 and 23°C. We report the observed experimental uncertainties from the mean for five realizations (using 95% confidence interval and the Student t-distribution). Uncertainties in the mean were about 2.9% for the 2.0% PVP and 7.0% for 0% PVP data. We verified that Joule heating effects were negligible (see Supplementary Information of [130]).

2.3 Experimental

We measured the electroosmotic flow in the presence of PVP for 16 background electrolyte chemistries. We explored PVP concentrations ranging from 0-2.0% w/w, keeping the background electrolyte ionic strength fixed at 30 mM. We note that RB visibly precipitates in background electrolytes with less than about 30 mM of ionic strength [130]. Figure 2-2 shows measured electroosmotic mobility on a logarithmic scale versus concentration, for measured pH values of 5.2, 6.6, 8.5, and 10.3. At low pH we observed a considerable reduction (110-fold drop to a value of 0.44×10^{-9} $m^2V^{-1}s^{-1}$) relative to the maximum value at pH = 10.3 and 0% PVP. However, we observed comparatively weaker effective EOF suppression at higher pH values. For example, adding 2.0% PVP concentration at pH = 10.3 causes a slight drop in EOF mobility from 48.2×10^{-9} to 29.9×10^{-9} $m^2V^{-1}s^{-1}$. Contrast this to the case of adding 2% PVP at pH = 5.2 which results in a drop from 35.5×10^{-9} to 0.44×10^{-9} $m^2V^{-1}s^{-1}$. Rhodamine B is weakly basic dye with $pK_{a,+1}$ of 3.2 [142, 143]. Therefore, at pH values of 6.6, 8.5, and 10.3, RB is less than 0.1% ionized and its electrophoretic mobility is negligible compared to the measured EOF mobilities measured at these pH values. At pH = 5.2, RB is 1% ionized, and has an effective mobility on the order of 0.1 x 10^{-9} $m^2V^{-1}s^{-1}$. This mobility magnitude is negligible compared to the measured EOF mobilities at pH 5.2 for PVP concentrations equal to 1.0% and higher. For the lowest EOF mobility data at pH = 5.2 and PVP concentrations of 1.0% and 2.0%, our reported values of EOF mobility may therefore be biased above the actual EOF mobility value due to the (cationic) electrophoretic mobility of RB. We therefore highlight these two data

points in Figure 2-2 using filled circles and caution that the corresponding values are correct only within an order of magnitude.

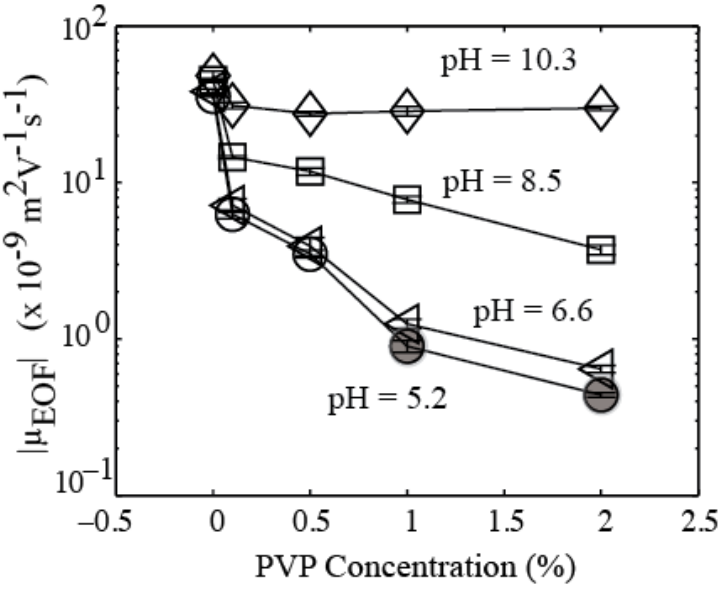


Figure 2-2. *Here we present electroosmotic mobility as deduced from motion of the neutral dye rhodamine B. We explored pH values of 5.2 (○), 6.6 (◁), 8.5 (□), and 10.3 (◇) and PVP concentrations ranging from 0 to 2.0% w/w. We placed the detector at L = 1.5 mm (see Fig. 1) in the separation channel. Data show EOF mobility decreases with increasing polymer concentration and decreasing pH. EOF mobility at pH 5.2 with 2.0% PVP is 0.44 × 10^{-9} m^2/Vs, more than 100-fold lower than the comparison case of equal pH but no polymer. We note that at pH of 5.2, RB is 1% ionized and has an effective mobility on order of 0.1 x 10^{-9} $m^2V^{-1}s^{-1}$. The electrophoretic mobilities of rhodamine B at pH 5.2 and PVP concentrations of 1.0% and 2.0% have comparable magnitudes, and so the EOF mobility data at these conditions is correct only within an order of magnitude.*

For all pH values and concentrations of PVP used in the current experiments, we observed no evidence of interaction of rhodamine B with channel walls. For example, we observed negligible increases in background fluorescence. Further, the rhodamine B peaks in electropherograms were symmetric. In Figure 2-3 we represent 16 representative electropherograms corresponding to each experimental condition shown in Figure 2-2.

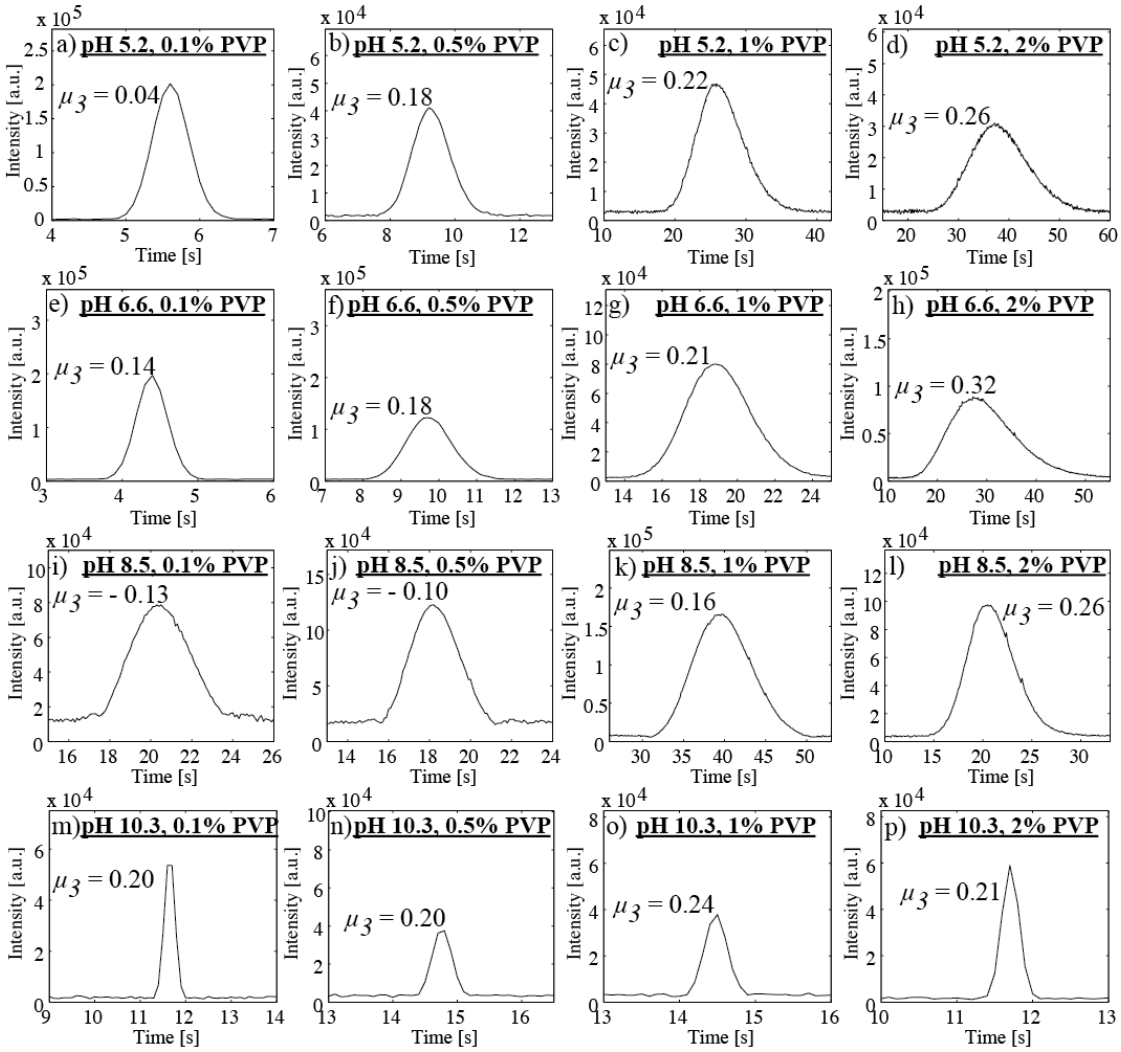


Figure 2-3. *Electropherograms for rhodamine B at pH values of 5.2, 6.6, 8.5, and 10.3 and at PVP concentrations of 0.1%, 0.5%, 1.0%, and 2.0%. Shown is the (third-moment) skewness, μ_3, for each peak. The majority of peak skewness*

magnitudes are between about 0.10-0.25, and the most asymmetric peak in our experiments has a skewness value of 0.32 (case h). These peak skewness values are significantly smaller than typical values of 1.00-2.00 for electropherogram peaks of species with significant wall adsorption/desorption [1,2].

Our data provide guidelines for designing electrophoresis experiments requiring precise control of EOF. Overall, we found that the largest reduction in EOF occurs when PVP concentration is varied from 0 to 1.0%. Above 1.0% PVP there is far less decrease in EOF mobility. For pH 10.3, electroosmotic flow mobility drops less than a factor of 2 between 0 and 0.1% and levels off above PVP concentrations of about 0.1%. The data of Fig. 2 for pH values lower than 10.3 show more significant decrease of electroosmotic mobility with increasing PVP concentration. Interestingly, in the lower pH range (of 6.6 and below), the EOF mobility drops by about 5 fold for PVP concentration of 0 to 0.1% and an additional 20 fold for 0.1-2.0%.

The observed strong dependence of suppression capacity of PVP on pH is very consistent with previous reports. For example, Kaneta *et al.* [129], Yu *et al.* [144], and Wang *et al*. [145] each reported strong pH-dependence of the degree of EOF mobility suppression in fused-silica channels using dynamic coatings in CZE experiments. Respectively, these studies used PVP (MW = 40,000, 360,000, and 1,000,000), poly(N-isopropylacrylamide) (PNIPAM) and P(VP-co-DMAEMA) copolymer coatings. A few of their observations are worth noting. Yu *et al.* [144] reported a 30-fold suppression of EOF at pH 5.84 *versus* a 10-fold suppression at pH 7.43. They attributed this to hydrogen bonding between the silanol groups of

the silica surface and the oxygen atoms of the carbonyl groups of PNIPAM. Both Kaneta *et al.* [129] and Wang *et al.* [145] showed that their coatings were stable only between pH of about 6 and 8; while we observe efficient and stable suppression at pH as low as 5.2. Further, Kaneta *et al.* [129] reported only a 10-fold reduction in EOF by PVP, while our optical white soda lime glass data show roughly 100-fold reduction. All three studies [129, 144, 145] reported a gradual reduction of EOF (and/or associated surface charge) with increasing polymer concentration, similar to our data for pH 8.5 and below. Wang *et al*. [145] noted a slight levelling off of EOF mobility above PVP concentrations of about ~1.25%, while Kaneta *et al.* [129] observed near uniform, monotonic EOF reduction at PVP concentrations larger than 1.0%.

We here provide evidence for evaluating adsorption/desorption of rhodamine B with channel walls. We quantify the relative asymmetry of rhodamine B peaks in all electropherograms by calculating their skewness (the third moment of distribution about the mean) for all pH and PVP concentration values. We apply the definition for skewness of a distribution:

$$\mu_3 = \frac{E(x-\mu)^3}{\sigma^3} \qquad \textbf{2. 2}$$

where μ is the mean of x (here, mobility), σ is the standard deviation of x, and $E(t)$ represents the expected value of the quantity t.

We present skewness values in Figure 2-3 and conclude minimal asymmetry of RB peaks, as the majority of the cases have values on order of 0.20, the worst being 0.32.

SECTION CONCLUSIONS

PVP as a dynamic surface coating for EOF reduction is attractive as it is a low viscosity polymer, which interacts with glass surfaces non-covalently. We have presented experimental data of electroosmotic mobility in the presence of a dynamic suppressing polymer PVP (MW=1,000,000) on channels wet etched into optical white soda lime glass. We performed on-chip CZE experiments at pH values of 5.2, 6.6, 8.5, and 10.3; and polymer concentrations between 0 and 2.0%. We used the fluorescent dye rhodamine B as a neutral marker. PVP is most effective for pH 5.2 and 6.6, moderately effective at pH 8.5, and least effective at pH 10.3. For example, at pH 5.2, 2.0% w/w PVP yields a 100-fold reduction in EOF mobility, but an equal concentration of PVP at pH 10.3 results in an EOF reduction of less than about 2 fold. We believe these data are useful and provide guidance in designing capillary electrophoresis experiments.

3 ELECTROPHORETIC MOBILITY

ELECTROPHORETIC MOBILITY MEASUREMENTS OF FLUORESCENT DYES USING ON-CHIP CAPILLARY ELECTROPHORESIS

Several sections of this chapter are based on an article by Denitsa Milanova, Robert D. Chambers, Supreet S. Bahga, and Juan G. Santiago, named "*Electrophoretic mobility measurements of fluorescent dyes using on-chip capillary electrophoresis*" in Electrophoresis [130], and are reproduced here with minor modifications.

We present an experimental study of the effect of pH, ionic strength, and concentrations of the EOF-suppressing polymer polyvinylpyrrolidone (PVP) on the electrophoretic mobilities of commonly used fluorescent dyes (fluorescein, Rhodamine 6G, and Alexa Fluor 488). We performed on-chip capillary zone electrophoresis experiments to directly quantify effective electrophoretic mobility.

We use Rhodamine B as a fluorescent neutral marker (to quantify electroosmotic flow) and CCD detection. We also report relevant acid dissociation constants and analyte diffusivities based on our absolute estimate (as per Nernst-Einstein diffusion). We perform well-controlled experiments in a pH range of 3 to 11 and ionic strengths ranging from 30 to 90 mM. We account for the influence of ionic strength on the electrophoretic transport of sample analytes through the Onsager and Fuoss theory extended for finite radii ions to obtain the absolute mobility of the fluorophores. Lastly, we briefly explore the effect of PVP on adsorption-desorption dynamics of all three analytes, with particular attention to cationic R6G.

3.1 Introduction

Fluorescent dyes are used in a wide range of applications including fluorescent probes [146]; fluorescent labels of nucleosides, nucleotides and nucleic acids [147]; biomolecule characterization [148]; and pH indicators [149]. Most relevant here, they are frequently used as markers or labels in a wide variety of electrophoresis techniques including isoelectric focusing (IEF) [150, 151], capillary zone electrophoresis (CZE) [152, 153] and isotachophoresis (ITP) [131, 154]. The latter techniques rely on electromigration, so accurate characterization of ion electrophoretic mobility is essential. Given ionization state information, the electrophoretic mobility of an ion can also be used to estimate diffusivity; for example, via the well known Nernst-Einstein relation [155]. Effective (observable) electrophoretic mobility depends on pH and ionic strength [156] and so systematic variation of these parameters is also important.

A straightforward approach to measure electrophoretic mobility of chemical species is by capillary zone electrophoresis (CZE) [157-162]. Effective electrophoretic mobility, the observable electrophoretic mobility of a dye, is measured by noting the time taken by an analyte peak to reach the detector (and correcting for electroosmotic flow). CE measurements of mobility are relatively simple (e.g., using a single, homogenous buffer chemistry) and robust to trace impurities. One challenge of applying CZE is quantifying low magnitude mobilities, which can take overly long to reach a detector (yielding low signal-to-noise ratio). The latter has been addressed by, for example, using pressure injection of the analyte and use of a neutral marker [163].

ITP offers an alternate method of quantifying mobilities [164-167]. ITP methods use a known leading electrolyte chemistry and focuses sample species into plateau mode (maximum, locally uniform concentration of the analyte) [164, 165] where purified analyte concentration (and zone order) can be related to analyte effective mobility. ITP is attractive as it can easily identify and quantify the mobility of multiple samples simultaneously, allows for low analyte concentrations (order of nmol), and can be robust to trace impurities [168, 169]. Hirokawa *et al*. [164, 165] and Pospichal *et al*. [166, 167] used ITP to quantify the absolute mobilities (mobility of chemical species when it is fully ionized under infinite dilution) and acid dissociation constant (pK_a) for a number of compounds with sufficiently high accuracy and good reproducibility.

In the current effort to directly quantify the mobility of fluorophores, we chose CZE over ITP as CZE offers more direct control of pH throughout the system. In

CZE, pH and ionic strength in the separation channel are uniform and determined directly by the background buffer chemistry, which can be quantified *ex situ*. Further, CZE is easily compatible with systems with unsuppressed electroosmotic flow, as CZE avoids non-uniform electric fields and non-uniform electroosmotic mobilities which can give rise to significant analyte zone dispersion [170, 171].

In the current chapter, we present measurements of absolute and effective electrophoretic mobilities, and diffusivities of three commonly used, namely fluorescent species fluorescein (anionic sodium fluorescein, FL), Rhodamine 6G (cationic Rhodamine 6G chloride, R6G), and Alexa Fluor 488 (anionic Alexa Fluor 488 succinimidyl ester, AF488). There are several references reporting mobility values of FL [172-174], but surprisingly we know of no quantitative study of the absolute mobility and pK_a of R6G or AF488. We also explore the effect of polyvinylpyrrolidone (PVP) concentration on the mobility of these three dyes. PVP is used commonly as dynamic wall coating for suppressing electroosmotic flow [175, 176] and yet we know of no such studies. We use on-chip CZE and CCD camera detection to quantify effective mobilities in a pH range of 3 to 11 and ionic strengths ranging from 30 to 90 mM. We use Rhodamine B (RB) as a neutral fluorescent tracer reference and to quantify electroosmotic flow mobilities. We analyze these data to report values of absolute mobility [177] for FL, R6G, and AF488. Where relevant, we experimentally quantify pK_a for effective mobility estimates. We account for and correct for the influence of ionic strength on all mobility measurements. This approach leverages the speed, low sample use, and relatively low cost of on-chip electrophoresis experiments. Our overall intent is to present a case study of how on-chip systems can be used to obtain accurate, highly

reproducible mobility measurements with minimal sample use; while also providing unique data for AF488 and R6G mobility in free solution and the effect of PVP on the mobilities of FL, AF488, and (most interestingly) R6G. When applicable, we highlight methods and issues relevant to leveraging on-chip systems to quantify ion mobilities.

3.2 Theory

We here review relevant electrophoretic mobility theory, which we used to interpret and standardize our measurements. The "actual mobili**Error! Not a valid link.**ty" μ_i of an ion is defined as the electrophoretic mobility of the molecule in its fully ionized state at a particular integer valence and at a particular finite ionic strength [177]. The degree of disassociation and the effective mobility of weak electrolytes depend on the pH of solution [177, 178]. Electrophoretic mobility of a partially ionized species in a solution is termed "effective mobility" [177, 178]. CZE experiments are typically performed in well-buffered solutions of known pH where weak electrolytes are often partially dissociated. Therefore, effective mobility is typically the empirically relevant, observable quantity. Effective mobility $\mu_{i,eff}$ is related to the degree of dissociation $g_{i,z}$ and actual mobility $\mu_{i,z}$ of species i and valence state z by [178],

$$\mu_{i,eff} = \sum_z \mu_i g_{i,z} \qquad \textbf{3. 1}$$

For example, the degree of dissociation and the effective mobility of a weak monovalent acid depends on the pH and acid dissociation constant pK_{-1} as,

$$\mu_{i,eff} = \mu_{i,-1} g_{i,-1} = \mu_{i,-1} \frac{1}{1+10^{pK_{-1}-pH}} \qquad \textbf{3. 2}$$

The effective mobility of a divalent acid depends on the dissociation level of the -1 and -2 valence states as,

$$\mu_{i,eff} = \mu_{i,-1} g_{i,-1} + \mu_{i,-2} g_{i,-2} = \frac{\mu_{i,-1} + \mu_{i,-2} 10^{pH-pK_{-2}}}{1+10^{pK_{-1}-pH}+10^{pH-pK_{-2}}}. \qquad \textbf{3. 3}$$

Here pK_n is the acid dissociation constant associated with valence state *n*. Persat *et al*. [178] review the topic of effective mobility of weak and strong electrolytes including pH and ionic strength effects. In this work, we measure $\mu_{i,eff}$ as a function of pH and for a range of ionic strengths and use this to quantify actual mobilities ($\mu_{i,z}$) and relevant acid dissociation constants (pK_n), as per relations (2) and (3). Ionic strength also influences the observable and actual mobilities of a species. Briefly, increasing ionic strength monotonically decreases (effective or actual) mobility, and the influence of ionic strength is stronger for higher valence values [140, 179]. We therefore correct our measurements for this effect in order to extract estimates for the fully ionized mobility of an isolated ion (i.e., in the limit of negligibly small ionic strength). The latter ideal quantity has been termed the ion absolute mobility, $\mu^0_{i,z}$, which is the quantity of most interest. Given estimates of $\mu^0_{i,z}$ and electrophoresis theory, we can predict a wide range of effective (observable) mobility values for wide ranges of pH and ionic strength.

Onsager and Fuoss [179] proposed a model of the ionic strength dependence of an ion's absolute mobility for an arbitrary mixture of species. However, the Onsager-Fuoss model treats the ions as point charges, and this limits its applicability to ionic strengths equal to or lower than order 1 mM. The Onsager-Fuoss model can

be extended to higher ionic strengths by, for example, including the finite ionic radius correction of Pitts [180]. This extended Onsager-Fuoss model for ionic strength dependence of mobility of *i*-th species in a mixture of *s* different species, can be written as [139, 140],

$$\mu_i = \mu_i^0 - (A\mu_i^0 + B)\frac{\sqrt{\Gamma}}{1+\frac{aD}{\sqrt{2}}\sqrt{\Gamma}} \qquad \textbf{3. 4}$$

$$A = z_i \frac{e^3}{12\pi}\sqrt{\frac{N_{AV}}{(\varepsilon kT)^3}}\sum_{n=0}^{\infty} C_n R_i^n,\ B = |z_i|\frac{e^2}{6\pi\eta}\sqrt{\frac{N_{AV}}{\varepsilon kT}}$$

$$D = \sqrt{\frac{2e^2 N_{AV}}{\varepsilon kT}}, \Gamma = \sum_{i=1}^{s}\Gamma_i,\ \Gamma_i = c_i z_i^2$$

Here z_i is the charge number of *i*-th ionic species, *e* the elementary charge, *k* the Boltzmann constant, N_{AV} the Avogadro constant, T the temperature, and Γ is twice the ionic strength I $(0.5\sum_i c_i z_i^2)$. The coefficients C_n and the vectors $\mathbf{R}^n = [R_1^n, R_2^n, ..., R_s^n]^T$ are given in [179]. In Eq. 3.4, *a* represents the mean distance of closest approach for the ions. For our calculations we chose a fixed value 1.5 $\text{mol}^{-1/2}\text{m}^{3/2}$ for *a* [140].

We here measure effective mobilities of fluorescent dyes for pH and ionic strengths ranging from 4-10 and 30-90 mM, respectively. We fit expressions (2) and (3) to the measured, effective mobilities for varying pH's to simultaneously determine the actual mobilities ($\mu_{i,z}$) and relevant dissociation constants (pK_a). We then correct the actual mobilities for finite ionic strength effects using the extended Onsager-Fuoss model, given by Eq. 3.4, to obtain estimates of the absolute

mobilities of ionic species (the ideal, fully-ionized mobility at infinite dilution) corresponding to each relevant dissociation level of our fluorescent dyes.

3.2.1 Estimation of effective mobility from CE experiments

We quantify the effective mobility of a species in the standard way, by applying an electric field and noting its migration time between a point of injection and a detector. The apparent mobility is calculated given migration time and electric field, as:

$$\mu_{i,app} = \frac{v_i}{E} = \frac{L}{Et_i} \qquad \textbf{3. 5}$$

where L is the length between the point of injection and detector, E the electric field, t_i the migration time and $v_i = L/t_i$ the electrophoretic velocity of species i. As usual, we obtain effective mobility from $\mu_{app,i}$ by accounting for EOF mobility, μ_{EOF}. We quantify μ_{EOF} by measuring the migration time t_{EOF} of a neutral species during the same experiment. The EOF mobility μ_{EOF} is obtained given the electroosmotic velocity v_{EOF} by,

$$\mu_{EOF} = \frac{v_{EOF}}{E} = \frac{L}{Et_{EOF}} \qquad \textbf{3. 6}$$

Combining these simple relations, we can write for the effective mobility

$$\mu_{i,eff} = \mu_{i,app} - \mu_{EOF} = \frac{L}{E}\left(\frac{1}{t_i} - \frac{1}{t_{EOF}}\right) \qquad \textbf{3. 7}$$

We used glass microchannels, where negative surface charge yields EOF in the direction of the electric field. Anions with electrophoretic mobility magnitudes lower than that of the EOF mobility therefore have net velocity directed towards

the negative electrode (cathode). This allows for having a single detector for both anions and cations on the cathode side.

Estimation of effective mobility using Eq. (3.7) requires accurate measurements of migration times t_i, t_{EOF}, E, and L. The chips we used have variable cross-sectional area channels (*c.f.* Figure 3-2), for which there are several choices in quantifying the local E in the separation channel section (which has locally uniform cross-section). For example, one method is to obtain precise estimates of local channel cross-sectional area A, applied current I, and the σ electrical conductivity of electrolyte solution to relate electric field as $E = I/(\sigma A)$. This requires current measurement of each individual run. Here, we chose to estimate E by using the analogy between electrokinetic chips of this type (for which electromigration current is dominant over diffusive and advection current components) and a simple resistor network.

3.2.2 Circuit Model Analogy

Figure 3-1 shows a representative resistance circuit model treating each channel region (e.g., associated with each channel width) as a separate resistor. Systems with relatively thin electric double layers, zero pressure gradients, ionic strengths of ~100 μM and higher, and uniform zeta potential obey a strong analogy with electrical circuits. We can then apply Kirchhoff's rules given by Eqs. 3.8 and 3.9.

$$\sum I_{Inter\,sec\,tion} = 0 \qquad \textbf{3. 8}$$

$$\sum V_{Loop} = 0 \qquad \textbf{3. 9}$$

In the case of high conductivity background electrolyte, we can approximate net current by electromigrational current by Eq. 3.10.

$$I_{net} \approx (\sigma A) E = \frac{(\Delta V)\sigma A}{L} = \frac{\Delta V}{R} \qquad \textbf{3. 10}$$

where R is the resistance of the channel, given by $R = L/(\sigma A)$, σ is the electrical conductivity of the buffer solution, A is the cross-sectional area of the channel, and ΔV is the voltage drop.

We model the chip channels as resistors in series (Figure 3-1a), simplify the electrical circuit (Figure 3-1b) and calculate the voltage drop V_c-V_1 in the separation channel using Eqs. 3.8-3.10, and 3.11, 3.12. We apply our best estimates of channel lengths and cross-sectional areas.

$$\frac{V_N - V_C}{R_N} - \frac{V_C - V_W}{R_W} - \frac{V_C - V_S}{R_S} - \frac{V_C - V_1}{R_C} = 0 \qquad \textbf{3. 11}$$

$$\frac{V_C - V_1}{R_C} = \frac{V_1 - V_E}{R_E} \qquad \textbf{3. 12}$$

There are seven unknown voltages and two equations, so we need specify four voltages. In our case those are the applied voltages at the four wells (V_N, V_S, V_E, V_W). We note that the voltage drop is independent of electrical conductivity, since it is constant and uniform and so drops out of Eqs. 3.11 and 3.12. Also, the estimates of V_1 and V_C turn out to be dependent only on area ratios and not absolute channel areas.

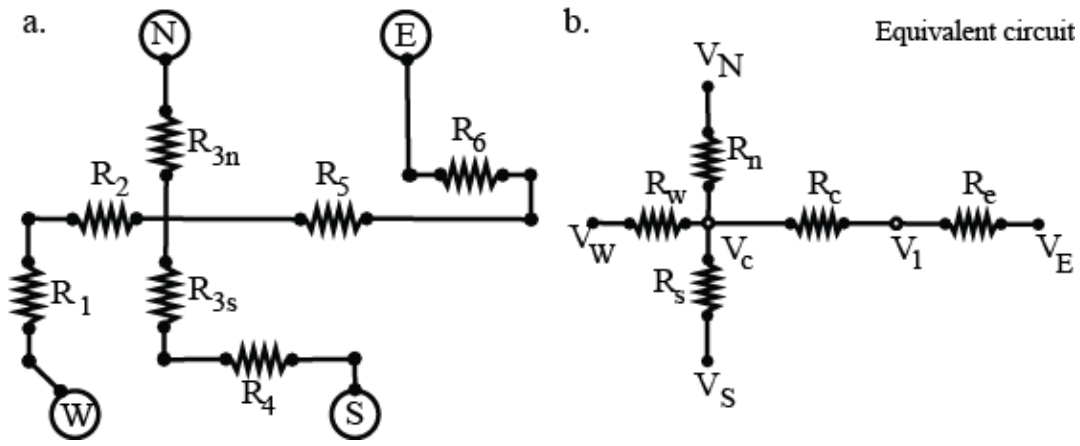


Figure 3-1. *(a) Representative cicuit treating each channel width as a separate resistor; (b) Equivalent resistance circuit model. The voltage drop in the separation channel is* V_C - V_1. *We estimate the electric field in the separation channel.*

3.2.3 Voltage Scheme for a Cross-Channel Sample Injection

We here present further details on the injection protocol of the CZE experiment. We empirically optimized the voltage scheme to suit our sample and background buffer of interest. As shown in the chip schematic of Fig. 1, we pulled the sample (consisting of one or more fluorophores and a neutral marker) into the chip by applying voltage (520 V) at N (north) well and ground S (south) well. The potential on the W (west) well was 510 V and on E (east) 510 V in this "pinching step". Next, we applied high potential difference between wells W (1200 V) and E (100 V) across the separation channel to inject the sample zone and initiate electrophoretic separation. During this step, we simultaneously applied potential on N (250 V) and S (150 V) to retract the supply (from N well) and waste stream (to S well) analyte regions from the separation channel. A qualitatively similar injection

protocol was described by Bharadwaj *et al*. [181] (see Figure 9, Table 1 and associated discussions of that reference).

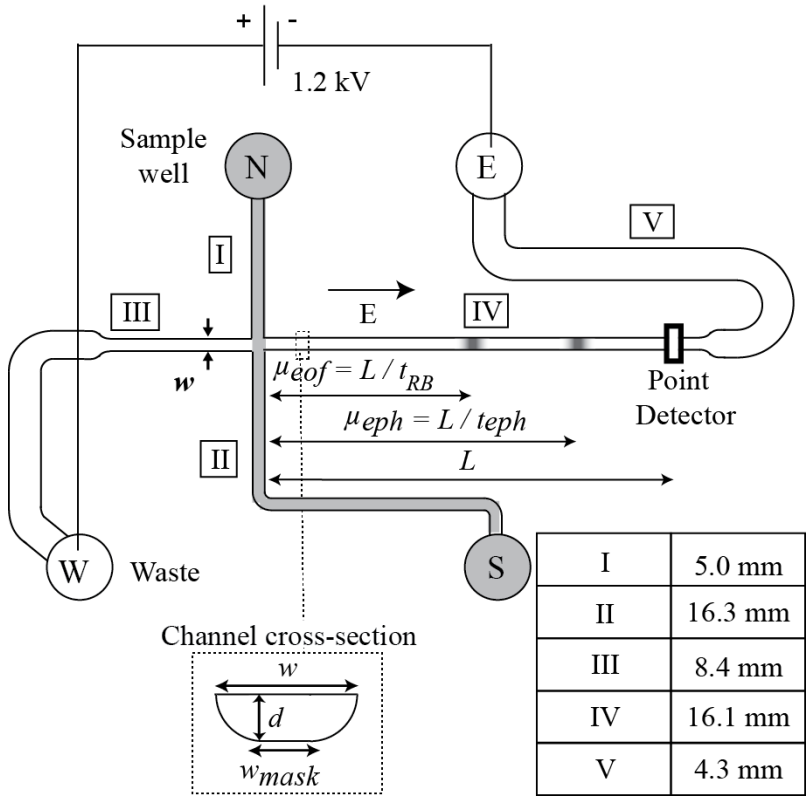


Figure 3-2. *The experimental apparatus for capillary electrophoresis includes microfluidic chip, epifluorescence microscope, CCD camera, high voltage switching system, 1.2kV DC power supply, and DAQ system. We used a 10× objective for all experiments. The exposure time varied between 50 and 100 ms depending on the signal strength. The chip used for all cases was a cross type Caliper NS 95 with 12 μm etch depth and 10 μm mask width in the separation channel. Precise measurements of channel center contour lengths of various regions (e.g., region IV, the separation channel) are provided in the inset table. We used either one or two analytes and a neutral dye (RB) loaded into the north reservoir. The electric field along the separation channel was 294 V/cm oriented from left to right.*

Briefly, our circuit model relates geometric channel parameters (channel lengths and cross-sectional area ratios) to compute the relation between channel geometries, applied potentials, and electric fields. The latter method is independent of the value of electrolyte conductivity and mapping system-wide electric fields also helps in optimizing injection protocols. We also performed 2D simulations (data not shown) of two-dimensional effects of electric fields (e.g., in channel turns) and concluded that such geometrical features have negligible effect on overall impedance (which is dominated by the channel curve centerline contour distances). Also, note that the molar concentration of our analyte fluorophores were in all cases ~3 orders of magnitude less than that of our background buffers (so they contributed negligibly to channel impedance estimates).

3.3 Materials and methods

3.3.1 Chemicals and Instrumentation

We performed controlled CE experiments in the pH range of 4.2 to 10.3 for determining effective mobility of FL, R6G, and AF488. We used anionic sodium-fluorescein (Molecular Probes, Eugene, OR) at 300 μM, anionic Alexa Fluor 488 succinimidyl ester at 150 μM (Molecular Probes, Eugene, OR) and cationic R6G (Acros Organics, Geel, Belgium) at 150 μM. R6G exists as two forms known respectively as dihydrorhodamine 6G and rhodamine 6G chloride. Dihydrorhodamine 6G is uncharged and non-fluorescent. Dihydrorhodamine 6G oxidizes to become the charged rhodamine 6G chloride which is fluorescent. We here studied rhodamine 6G chloride. We used RB dye (Sigma-Aldrich, St. Louis,

MO) at 200 μM to quantify EOF. RB has a reported pK_a value of 3.22 [141] and approximately neutral for our pH range of interest. We chose these dye concentrations to obtain electropherograms with comparable peak heights. We prepared buffer solutions of glycine, tricine, HEPES, MES, and acetic acid titrated with NaOH to pH's between 4.2 and 10.3. These 15 electrolyte chemistries are summarized in Table 3-1.

We diluted all stock buffer solutions with deionized ultrafiltered water (DIUF) (Fischer Scientific, Pittsburgh, PA). We used both PeakMaster [139] and SPRESSO [182] to aid in buffer design and analysis (both codes gave the same results). Predicted pH values often differed by ~0.1-0.2 pH units from measured values, possibly due to the effects of ionic strength [140]. Hence, we report both predicted and measured pH values for these buffers as determined using a Corning Pinnacle 542 pH/conductivity meter (Nova Analytics, Woburn, MA). Lastly, we explored the effect of 0.1% to 2% by weight PVP (Polysciences Inc., Warrington, PA) on R6G mobility.

We performed all assays on a commercial NS-95 borosilicate microchip purchased from Caliper Life Sciences (Mountain View, CA) with a simple cross pattern consisting of narrow and wide channel sections, as shown in Figure 3-2. The chip was wet etched and covered with a clear plate of the same material. Isotropically etched glass channels were 12 μm in depth, and 11 μm and 50 μm in mask width in the narrow and broad regions, respectively. The separation channel length was 16.1 mm.

Table 3-1. *Description of buffer solutions used to study pH effects. In parenthesis, we list respectively buffer valence, absolute mobility as* $10^{-9}\ m^2V^{-1}s^{-1}$*, and* pK_a*.*

	Measured, Predicted pH	C (acid) [mM]	C (base) [mM]	I [mM]
		Glycine (+1, 39.5, 2.32), (-1, 37.4, 9.78)	NaOH	
1	10.3, 10.11	40	30	30
2	9.8, 9.64	60	30	30
3	9.4, 9.16	120	30	30
		Tricine (-1, 26.6, 8.5)	NaOH	
4	8.5, 8.49	40	30	30
5	8.0, 8.00	60	30	30
6	7.6, 7.53	120	30	30
		HEPES (-1, 21.8, 7.5)	NaOH	
7	8.3, 7.84	40	30	30
8	7.7, 7.36	60	30	30
9	7.2, 6.88	120	30	30
		MES (-1, 26.8, 6.13)	NaOH	
10	6.6, 6.43	40	30	30
11	6.1, 5.95	60	30	30
12	5.7, 5.48	120	30	30
		Acetic Acid (-1, 42.4, 4.756)	NaOH	
13	5.2, 5.09	40	30	30
14	4.7, 4.62	60	30	30
15	4.2, 4.14	120	30	30

We imaged zones in the CZE experiment with an inverted epifluorescent microscope (IX70, Olympus, Hauppauge, NY) equipped with a mercury lamp, a

U-MWIBA filter-cube from Olympus (460-490 nm excitation, 515 nm emission) and a 10× (NA of 0.4) UPlanApo objective for fluorescence imaging. Images were captured using a 12 bit, 1300 by 1030 pixel array CCD camera (Coolsnap, Roper Scientific, Trenton, NJ), and with μ-Manager microscopy software (available for free at micro-manager.org). We performed post-processing of the data with custom MATLAB scripts. High voltage was applied at microchip wells using a computer-controlled Labsmith HVS-3000D (Livermore, CA) power supply and 10 mm lengths of 0.5 mm diameter platinum wire (Goodfellow, Oakdale, PA) soldered to high voltage leads.

3.3.2 Assay Protocols

We empirically optimized the voltage scheme for sample injection. The scheme uses a fairly standard pinching and "retraction" step, and is described in detail in the Supplementary Information document. The point of detection to measure elution time was typically placed 15 mm down the separation channel as shown in Figure 3-2. The only exception were experiments where we strongly suppressed EOF, where signal-to-noise ratio requirements compelled us to move it to only 1.5 mm from the injection region. For each run, we used a pipette to dispense 40 µl volumes of the dye concentrations described earlier into the sample reservoir (so we consumed 6-12 ng of fluorophore for each experiment). Between each run, we used a channel cleaning procedure similar to that of Chambers *et al*. [131]. To this end, we flushed the channel with 40 µL of 0.5 M NaOH for 10 min by applying vacuum to well the S in Figure 3-2, followed by deionized water for 5 min, 100 mM HCl for 3 min, and deionized water again for 3 min. Between each run, we found that flushing several times with deionized water was approximately

sufficient in refreshing the surface to its initial state (although we quantified EOF for each and every run).

3.4 Results and Discussion

3.4.1 Estimation of absolute mobility in CE experiments

As mentioned in Section 3.2, we measured migration times of both electrophoretic and neutral dyes to quantify effective mobility of species given apparent mobility, using Eq. (7). We then performed nonlinear regression best fits of expressions (2) and (3) to the measured effective mobility versus pH data to obtain actual mobility (fully ionized mobility at finite ionic strength) and pK_a's. We then correct the actual mobilities for finite ionic strength effects to obtain the absolute mobilities of fluorescent species. The absolute mobility and pK_a associated with each valence state can be interpreted as an estimated material property for the dye.

Figure 3-3 shows effective mobility measurements for FL, R6G, and AF488 versus pH at constant ion strength. Each data point is the mean of five realizations and the error bars denote a 95% confidence interval based on Student *t*-distribution. As shown, the effective mobility of univalent cationic R6G and anionic AF488 is constant within the pH range of study. We know of no source reporting pK_a's for R6G and AF488, and our experiments suggest these fluorophores have no pK_a in the 3-10 pH range. However, FL mobility initially increases with an increase in pH and subsequently plateaus at higher pH (~7-10). FL is a dibasic acid with a

dissociation constant of the mono-ion of ~6.8. Several groups report fluorescein pK_a's [141, 183] to be in the range of 2.1-2.2 (cation), 4.4 (neutral), and 6.7-6.8 (mono-anion). Due to its strong decrease in quantum yield at acid conditions, [183, 184] we were unable to obtain data for FL below pH 5 (quantum yield of FL is maximum near pH 8). In contrast, R6G and AF488 exhibited approximately uniform fluorescence within pH 4 to 10. Similar behavior for AF488 was observed by [185].

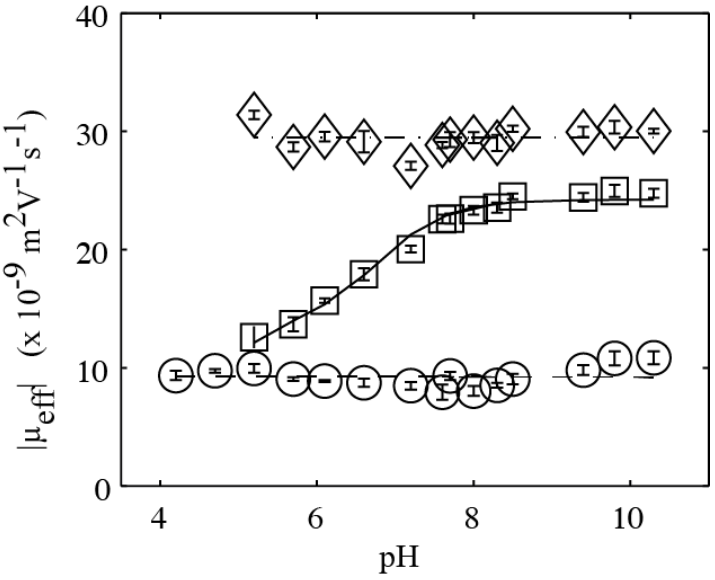


Figure 3-3. *Effective mobility data for Rhodamine 6G, Fluorescein, and Alexa Fluor 488 at 30 mM ionic strength and pH between ~4.2 and 10.4. Shown are experimental data for R6G (○), Fluorescein (□), and AF488 (◇). We show fits for effective mobility of R6G (- - -), fluorescein (—), and AF488 (–·–·–) 30 mM ionic strength. Fluorescein displays a pK_a at pH ~7. R6G and AF 488 seem to be fully ionized within the range. We performed a total of five repetitions for each case and show here the mean value. The error bars correspond to 95% confidence intervals on the means with N = 5 realizations at each pH. We least squares curve fit the data using effective mobility theory, including correcting for ionic strength based*

on an Onsager and Fuoss model with a Pitts correction [140]. For this theory, we assumed two pK_a values (4.45 and 6.8) reported in literature for FL, and use the fit to extract effective mobility data. FL has a third pKa (2.14), but this falls well outside the pH range of the experiments. From these data, we calculated absolute mobility values of 19×10^{-9} m^2/Vs and 36×10^{-9} m^2/Vs, corresponding to -1 and -2 valence states for Fluorescein. We did not observe pK_a's for AF 488 and R6G within this pH range. Their pH-averaged, absolute mobilities are 36×10^{-9} m^2/Vs and 14×10^{-9} m^2/Vs, respectively.

Next, we determine the absolute mobilities and relevant pK_a of FL, R6G, and AF488 from the experimental data. We summarize the values and relevant relations in Table 3.2.

We numerically calculated the following absolute mobilities for R6G and AF488: -14×10^{-9} $m^2V^{-1}s^{-1}$ and 36×10^{-9} $m^2V^{-1}s^{-1}$, corresponding to -1 and +1 valences, respectively. FL shows two absolute mobilities: 19×10^{-9} $m^2V^{-1}s^{-1}$ and 36×10^{-9} $m^2V^{-1}s^{-1}$, corresponding to -1 and -2 valence states. By comparison, the effective mobility value for fluorescein has been reported as $33.5 \pm 0.2 \times 10^{-9}$ $m^2V^{-1}s^{-1}$ in 1 mM Tris-HCl solution at pH 9.1 [174]. This value is consistent with our experimental data and other reported data at similar conditions [172, 173]. Our effective mobility curves for R6G and AF488 (eq. 4), and fluorescein (eq. 5) are shown in Figure 3-3 along with the experimental data for 30 mM ionic strength. Khurana *et al.* [186] report a pK_a of 7.5 for the related species dihydrorhodamine 6G but not for the species rhodamine 6G chloride of interest here; a value they obtained using an ARChem (Automated Reasoning in Chemistry) physicochemical

property calculator SPARC (**http://sparc.chem.uga.edu/sparc**). Duvvuri *et al.* [187] reported an experimental value of 7.5 for dihydrorhodamine 6G as well. Another group [188] found (experimentally) that the alkalinity of a rhodamine 6G species (the molecular structure was not specified) varied with light excitation and reported a pK_a value of 6.5. We know of no other reported values.

In Table 3-2, we also report diffusivities as per Nernst-Einstein relation $D_i = RT\mu_{i,eff}$, where we use the absolute mobility values (at infinite dilution). We calculated the diffusivities for FL, AF488, and R6G as 9.3×10^{-10}, 9.3×10^{-10}, and 3.6×10^{-10} m^2s^{-1}, respectively. Madge *et al*. [189] used fluorescence correlation spectroscopy (FCS) and reported D_{R6G} = 2.80 m^2/s at 22°C. Petrasek and Schwille [190] reported measured values of D_{R6G} = 4.26 m^2/s at 22.5°C using FCS. Mueller *et al.* [191] a value of D_{R6G} = 4.14 m^2/s at 25°C using multicolor dual focus FCS. Corrected for temperature effects (absolute viscosity and absolute temperature on diffusivity as per Einstein relation), these reported values are within about 14% and 8% of our measured value, respectively.

Table 3-2. *Absolute mobilities (i.e., fully-ionized value extrapolated to 0 ionic strength) and diffusivities based on these absolute mobility estimates (as per Nernst-Einstein diffusion) for Fluorescein, R6G, and AF488 (at 22°C), their pK_a's and prediction models. We report two absolute mobility values: our experimental values and values assuming a reference FL effective mobility [174] extrapolated to 22°C.*

Sample	$\mu_{abs,\,expt.,\,22°C}$ $\mu_{abs,\,ref.,\,22°C}$ [$\times 10^{-9}$ $m^2 V^{-1} s^{-1}$]	pK_a's	Relation for μ_{eff} [$\times 10^{-9}$ $m^2V^{-1}s^{-1}$]	D [$\times 10^{-10}$ $m^2 s^{-1}$]
Fluorescein	35.9 34.5	4.4 6.8	$\mu_{i,eff} = \frac{\mu^0_{i,-1} + \mu^0_{i,-2}10^{pH_i - pK_{i,-2}}}{1 + 10^{pK_{i,-1} - pH_i} + 10^{pH_i - pK_{i,-2}}}$ 9.3	
Alexa Fluor 488	36.0 36.1	—— [a)]	$\mu_{i,eff} = \frac{\mu^0_{i,-1} + \mu^0_{i,-2}10^{pH_i - pK_{i,-2}}}{1 + 10^{pK_{i,-1} - pH_i} + 10^{pH_i - pK_{i,-2}}}$ 9.3	
Rhodamine 6G	14.0 12.6	—— [b)]	$\mu_{i,eff} = \mu^0_{i,+1}\frac{1}{1 + 10^{pH_i - pK_{i,+1}}} \cong \mu^0_{i,+1}$ 3.6	

[a),b)] pK_a's for rhodamine 6G chloride and Alexa Fluor 488 are well outside the pH range used in these experiments. We know of no reported values in literature.

We note the ambient temperature for our experiments varied between 21 and 23°C; however, we consider a more conservative range of a 3°C variation. We estimate the maximum variations in mobility from the expected variation in dynamic viscosity of water for our aqueous solutions. A ±1.5°C variation in temperature results in about ±3% of absolute viscosity (using the viscosity versus temperature fit reported by [192]). Therefore, the absolute values of mobilities presented here varied with temperature by as much as ±2.4% for R6G and ±1.0% for FL and AF488. These estimated variations can be compared to the observed experimental uncertainties from the mean for five realizations (using 95% confidence interval

and the Student t-distribution). The latter uncertainties were 3.9% (R6G), 1.9% (FL), and 1.5% (AF488).

As a further comparison, we also report a second set of absolute mobility estimates based on a reference effective mobility value of 33.5 ± 0.2 × 10-9 m2/Vs for FL at 27°C, pH 9.1 (from [174]). In these additional mobility estimates, we first extrapolate this published FL value to 22°C using Walden's rule [177] and the aforementioned viscosity versus temperature fit by Touloukian et al. [192]. The extrapolation yields a new reference value of 32.4 ± 0.2 × 10-9 m2/Vs for FL at 22°C. We then assume this reference effective mobility at 22°C is correct and use it to normalize all of our effective mobility data for AF488 and R6G. To this end, we use the FL reference value to obtain new electric field estimates, construct mobility curves, and calculate the respective new absolute mobilities for FL, AF488, and R6G. In Table 3-2, we include these "normalized" absolute mobility estimates for FL, AF, and R6G at 22°C (the second value in each row of the mobility column). The absolute mobility values obtained directly from the experimental data agree well with mobilities normalized to this reference value.

3.4.2 Joule Heating

We note we explored the possible effect of Joule heating on our measured mobilities. For all of the chemistries explored we verified that Joule heating was insignificant by monitoring current versus voltage traces. The current versus applied voltage data was clearly linear over as much as twice the maximum voltages used in our experiments. This linearity implies negligible effects of Joule heating.

We here verified that Joule heating had negligible effect on our mobility measurements. In Figure 3-4 we show that the current versus voltage trace yields a linear relation. We used the highest conductivity buffer (90 mM NaOH, 180 mM Glycine) and applied voltages ranging from 250 to 2,000 V.

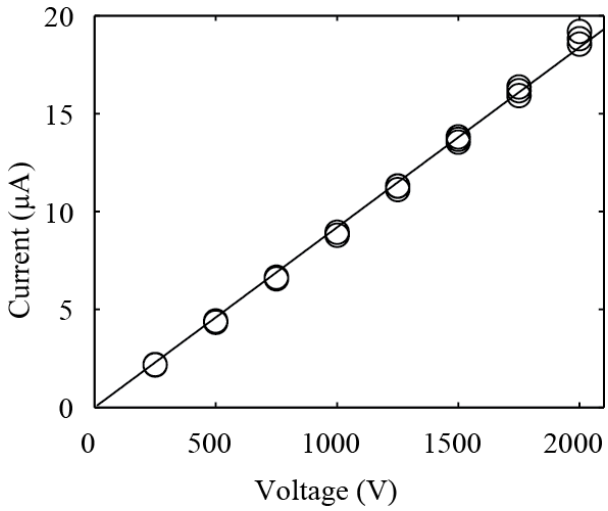


Figure 3-4. *Current-Voltage trace for 90 mM NaOH and 180 mM Glycine buffer. The current measurements were taken over 60 s and each run was repeated three times. We fit the data to a linear fit with regression coefficient value of R=0.997. The data verifies that Joule heating is insignificant in our experiments.*

Here we include a figure (Figure 3-4) of the current versus voltage trace for the highest conductivity buffer (90 mM NaOH, 180 mM Glycine) and applied voltages ranging from 250 to 2,000 V (which yielded a linear relation with a regression coefficient of R = 0.997).

3.4.3 Effect of Ionic Strength

In Figure 3-5 we show measurements of effective mobilities for FL and R6G at pH 9.4 and 7.2 (each) and eight ionic strengths in a range between 3 and 90 mM. FL is a divalent acid while R6G is monovalent, so the stronger dependence of FL to ionic strength is expected. FL mobility drops ~20-25% (depending on the pH) as ionic strength increases from 30 to 90 mM. On the other hand, R6G shows only weak dependence on ionic strength. Predictions based on extended Onsager and Fuoss model [139, 140] are shown as dashed curves, and these show fairly good agreement with our experimental data. We attempted but were unable to obtain accurate, meaningful data below ionic strengths of 30 mM. After a series of control experiments, we concluded that RB precipitates and forms observable aggregates below about 20 mM. Our observations suggest RB interacts strongly with the channel walls in this regime, strongly impeding (and biasing) our efforts to quantify EOF mobility. Such behavior has been reported for RB [193, 194].

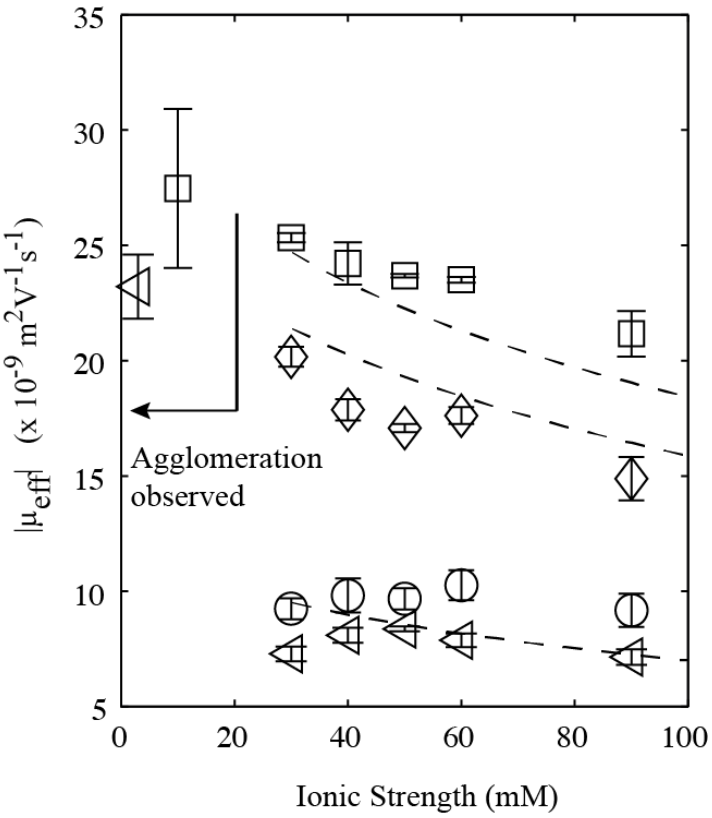


Figure 3-5. *Effective mobility data for R6G at pH 7.2 (◁), R6G at pH 9.4 (○), FL at pH 7.2 (◇), and FL at pH 9.4 (□) and numerical predictions (- - -). We based the numerical simulations leveraging the Onsager and Fuoss model and Spresso [140, 179]. The effective mobility for R6G approximately levels off at higher concentrations (>30 mM) and decreases only slightly with decreasing pH. (Below, we discuss R6G adsorption-desorption behavior and how this may affect results.) Fluorescein mobility decreases more drastically with ionic strength increase. FL mobility at pH 7.2 is lower than at pH 9.4, irrespective of ionic strength, consistent with the results in* **Figure 3-3**. *The data below ~20 mM for both R6G and FL are not representative of mobility data as we observed precipitation of the neutral marker RB in that regime. This precipitation impeded our ability to quantify EOF.*

3.4.4 Effects of polyvinylpyrrolidone on mobility

We also measured the effective mobilities of FL, R6G, and AF488 in the presence of the dynamic coating polyvinylpyrrolidone (PVP) polymer. We explored PVP concentrations from 0% to 2% and pH values of 5.2, 6.6, 8.5, and 10.3 with a fixed ionic strength of 30 mM. We found PVP changed the mobilities of FL and AF488 by amounts less than about our experimental uncertainty (approximately $\pm$ 1 $\times$ 10^{-9} $m^2V^{-1}s^{-1}$) through this PVP and pH range, and so these will not be discussed further here. However, PVP had a strong influence on the measurements of R6G mobilities. R6G is a cationic dye and as such more susceptible to wall interactions in our borosilicate glass channels. Figure 3-6 presents measurements of the effective mobility of R6G. As with the data discussed earlier, we used measurements of RB elution times to correct for the strong effects of EOF suppression by PVP. (We will present a study of the effects of PVP suppression of EOF in a future publication.) The data of Figure 3-6 shows R6G effective mobility mostly decreases with increasing PVP concentration. At and above pH 6.6, we see a monotonic decrease of mobility with increasing pH. The pH 5.2 data has the most pronounced decrease with decreasing PVP concentration. We hypothesize that R6G mobility varies due to interactions with channel walls. Analysis of the individual R6G peak shapes supports this hypothesis. Notably, high pH for both low and high PVP concentrations results in noticeable tailing of R6G peaks, suggesting wall adsorption/desorption type dynamics [195].

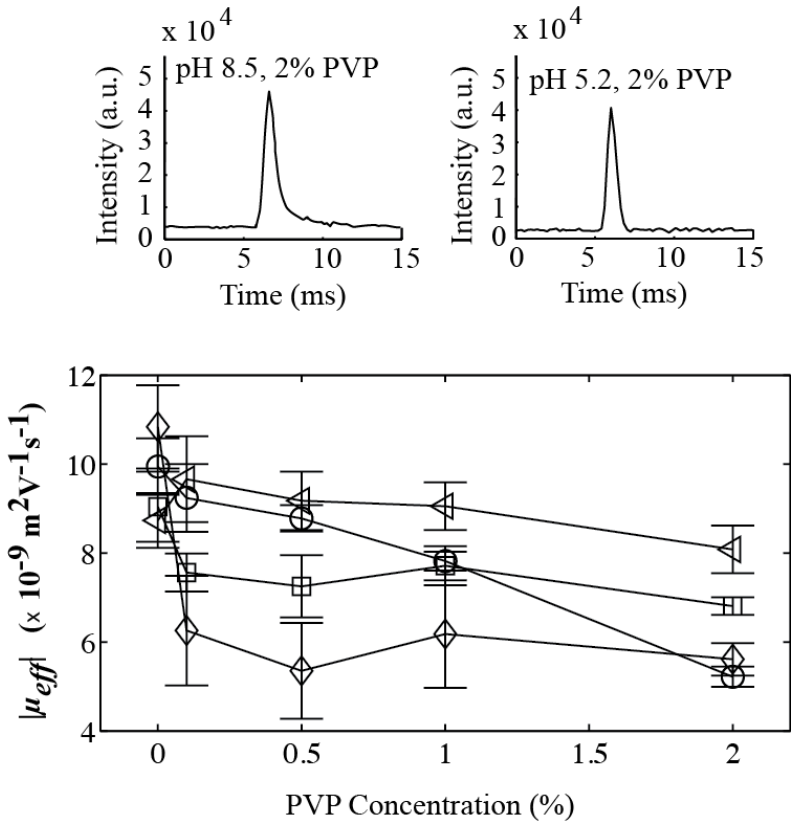


Figure 3-6. *(a) Effective mobility of R6G at 0%, 0.1%, 0.5%, 1% and 2% polyvinylpyrrolidone (PVP) for: pH 5.2 (○), 6.6 (◁), pH 8.5 (□), and 10.3 (◇). We show example electropherograms for R6G at pH 8.5 (b) and pH 5.2 (c) each with PVP concentration of 2%. These R6G mobility data correct for EOF using RB elution time measurements. Addition of PVP polymer decreased EOF significantly, so we placed the detection point 1.5 mm downstream of the channel intersection for enhanced signal-to-noise ratio. R6G shows no pK_a within the working range, so we hypothesize that its mobility varies with pH due to its interactions with the channel walls. The data with highest reproducibility were for pH of 5.2 and 6.6 data and high PVP concentration (1% and 2%). These cases exhibit no peak tailing which we attributed to adsorption/desorption phenomena. Electropherograms (b, c) show peak tailing at pH 8.5 with 2% PVP but no tailing for the same PVP concentration and pH 5.2.*

We show Figure 3-6.b and c as typical example data showing pronounced tailing for pH 8.5 but not pH 5.2. We note Hamai and Sasaki [196] also reported dispersion of R6G peaks due to polyvinyl sulfate (PVS). Their data and discussion suggest this is due to direct interactions between R6G and PVS. We did not observe such interactions. Instead, our observations show significant tailing of R6G peaks in the absence of PVS.

CONCLUDING REMARKS

We have presented experimental data of absolute and effective electrophoretic mobilities and diffusivity estimates for FL, R6G, and AF488. We performed on-chip capillary electrophoresis experiments for various pH's, ionic concentrations, and concentrations of the EOF-suppressing polymer PVP. We used RB as a neutral fluorescent marker to account for electroosmotic flow in each experiment. Experimentally, we observed that the mobility curve is nearly horizontal for both R6G and AF488, and has a sigmoid-like shape for FL. This behavior is consistent with a pK_a of ~6.8 for FL (within this pH range) and the absence of a pK_a in this range for both R6G and AF488. We accounted for and corrected for the influence of ionic strength on sample analytes. We demonstrated that analyte mobility decreases with increasing ionic strength. This effect is more pronounced for the divalent FL than for univalent R6G, as predicted by Onsager and Fuoss theory as extended by Pitts. Based on experimental data, we concluded that ionic strength should be at least 20 mM to prevent aggregation of the neutral marker RB. We pointed out that reduced adsorption is critical for clean and accurate separation. We further studied the effect of the EOF suppressant PVP on the mobilities of FL, AF488, and R6G. We found a negligible effect of PVP on FL and AF488. However, we found a strong PVP effect on R6G mobility, which we attribute to adsorption-desorption dynamics of the cationic R6G dye with our negatively charged channel walls. We found an addition of 2% PVP at low pH (~5.2) reduces electroosmotic flow more than 100× and R6G is well behaved. Adsorption-desorption of R6G is apparently very important at high pH, as evidenced by mobility trends and pronounced tailing of the signal peaks. Overall on-chip CZE

offers fairly rapid, highly reproducible mobility measurements which require very little sample use.

4 NUCLEAR VS CYTOSOLIC RNA-SEQ IN SINGLE CELLS

WHOLE GENOME RNA-SEQ ANALYSIS OF SINGLE-CELL SUBCELLULAR FRACTIONS

4.1 Introduction

RNA molecules can may undergo various processes, which results in the synthesis of multiple isoforms of each gene, and each gene has on average 10 to 12 isoforms. [197, 198] These processes may be errors in multi-exon genes such as exon rearrangements, removal/retention of introns, polyadenylation, or RNA editing. Alternative splicing is a post-transcriptional process and has implications in transcriptome variability [199] proteome diversity, and many human diseases. It is often associated with cancer, [200, 201] neurodegenerative brain diseases, [202] and even aging.[203]

Eukaryotic cells use the mechanism of splicing to make primary RNA transcripts, the process by which non-coding sequences (introns) are removed and the remaining coding ones (exons) are joined. While it is well established that RNA represents a more direct measure of the genetic information encoded by the genome, the processes of RNA processing, export, and localization within the cell remains poorly understood. Within the Encyclopedia of DNA Elements (ENCODE) project, Tilgner et al. [197] determined rates of splicing completion from bulk RNA-seq data in K562 cells. However, such genome-wide ENCODE measurements do not address cell-to-cell heterogeneity in any form.

Single cell gene expression studies uncover cell-to-cell variability in seemingly identical cells within populations. One source of variation underlying transcription is stochastic bursting [204], a dynamic fluctuation of gene expression. Because single cell analysis is not clouded by ensemble-averaging effects, it provides a precise measure of genomic and gene expression variations as a result of physiological processes from cell cycle states, and signaling to stress responses. More importantly, these heterogeneities may uncover important biological implications about cell function and even cell identity in complex samples. In addition to the stochastic nature of the transcription, heterogeneity in gene expression among cells is affected by cell cycle, microenvironment, and epigenetic states. Thus, single-cell transcriptional profiling is critical for gaining a deeper insight of heterogeneity. Furthermore, some cell types are rare and single-cell approaches become essential to their identification and characterization.

A particular challenge for current single-cell methods is synchronized analysis of nuclear vs. cytoplasmic contents. At the crossroad of this challenge is an efficient

single cell fractionation process. Specifically, the fractionation of the nucleus versus the cytosolic compartment in the same single cell with no cross-contamination. Current approaches for the analysis of single cell splicing leverage poly-A+ selection of mRNAs from a whole cell. [6] While this data is meaningful in that it shows single cell variability in alternative splicing and evidence for autosomal allelic exclusion, it does not distinguish between transcripts with introns retained and not-yet-spliced. The lack of methods to probe both polyadenylated-plus and polyadenylated-minus primary and processed transcripts within individual nuclei and their respective cytosols has prevented quantitative dissection of splicing patterns at sub-cellular level.

In this study, we build a novel microfluidic tool for the study and analysis of RNA transcripts from single cells. Our system leverages and complements existing, off-the-shelf next-generation sequencing (NGS) technologies and uniquely enables correlation of transcripts in cell nuclei to transcripts in cell cytosol. We use single-cell isotachophoresis (sc-ITP) to uniquely fractionate and extract nuclear versus cytoplasmic RNA. Then, we use our system and state-of-the-art NGS technologies to study fundamental questions of human genome transcription, differential gene expression, localization, and processing destinies of RNAs for various cell types, disease, and differentiation states. To this end, we have developed a method and device to isolate and trap single cells; performed rapid, electric-field-based selective lysis of cytoplasmic membrane (leaving nucleus intact); and then purified and simultaneous fractionated total RNA in cytosol (cyt-RNA) and total RNA in the nucleus (nuc-RNA) from single cells with no intra-compartment cross-contamination.

Our method leverages the ion displacement and focusing electrokinetics of isotachophoresis (ITP) for purification and focusing of RNA; separation of ITP-focused RNA from non-ITP-focused nucleus (containing nuc-RNA); and fractionation of cyt-RNA (ITP zone) and nuc-RNA (nucleus) into separate, recoverable outputs.

4.2 Methods

4.2.1 Gene Expression Analysis of Splicing Patterns in Sub-cellular Compartments of Single Cells

In this study, we describe a method for the fractionation of nuclear vs cytoplasmic RNAs in single cells. Briefly, our protocol is outlined in Figure 4-1 and involves cell capture (branch AB), electroporation (branch AE), fractionation (branch C'C for nuclear, and D'D for cytosolic). Sub-cellular fractions are recovered at respective outputs (C and D), as labeled in Figure 4-1.a.

After recovery, these fractions were analyzed by RT-qPCR and qPCR with sequence-specific (TaqMan) probes for U3 small nuclear RNA, unspliced mRNA (precursor), spliced mRNA (processed), and genomic- or extranuclear-DNA. To target unspliced mRNA (and not simply the excised intron), we used probes, which flank an exon-intron boundary. Similarly, to target spliced mRNAs, we designed our probes to cover an exon junction (Figure 4-1.b). For these gene expression experiments, we processed and analyzed 96 nuclei and 96 cytosolic samples from Snyder lymphoblastoid cells, and K562 chronic myelogenous leukemia cells (sub-lines from ATCC and ENCODE). We then developed measures for the Percent

Spliced Introns (in the nucleus) and Percent Retained Intron (in the cytosol). We analyzed U3-snRNA in K562 cells (from both sub-lines) and showed that it is present only in the nucleus (Figure 4-1.c).

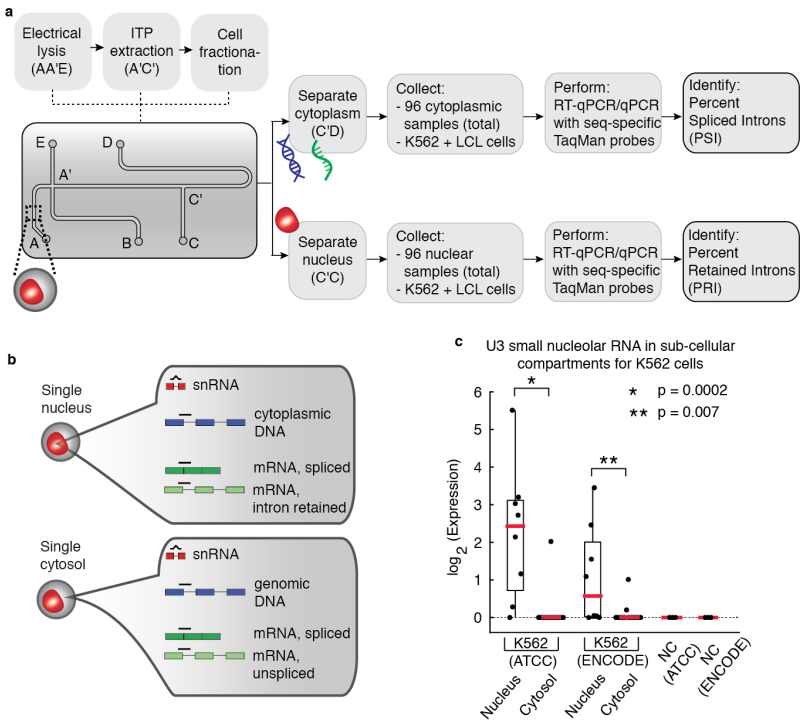


Figure 4-1. (a) *Workflow for the separation of RNAs localized in the nucleus and cytosol. (b) Direct one-step RT-qPCR and qPCR with sequence-specific probes (TaqMan) provide genome- and transcriptome-wide methods of gene-specific probing in sub-cellular compartments for precursor (unspliced), processed (spliced), small nucleolar RNAs and DNA. This development of an efficient fractionation protocol permits analysis of Percent Spliced Introns (PSI) and Percent Retained Introns (PRI) in the nucleus and cytosol, respectively. (c) Relative gene expression of U3 snRNA (an RNA-associated protein) localized exclusively in the cell nucleus of K562 cells.[4] Log2-transformed box-and-whisker plots show clear localization of U3 snRNA in the nucleus for two sublines of K562 cells.*

4.2.2 Single-cell electroporation and fractionation by sc-ITP

We have developed a method for fractionation of single cells in subcellular nuclear and cytoplasmic compartments, single-cell ITP sequencing (scITP). scITP is an electrokinetic technique which allows for the physical separation of total nuclear vs total cytoplasmic nucleic acids from single cells. In scITP (shown in Figure 4-2), a single cell is captured using a custom-made PDMS device from an aliquot of about 10 µl containing 5 cells/µl. The cytoplasmic membrane of each cell is electrically lysed, and the cyt-RNA thereby released. Cyt-RNA is rapidly purified and focused via ITP within the microfluidic channel. The nucleus lags behind, not focused in the ITP zone, enabling fractionation of total RNAs inside the nucleus vs total cytosol RNAs within the ITP zone downstream.

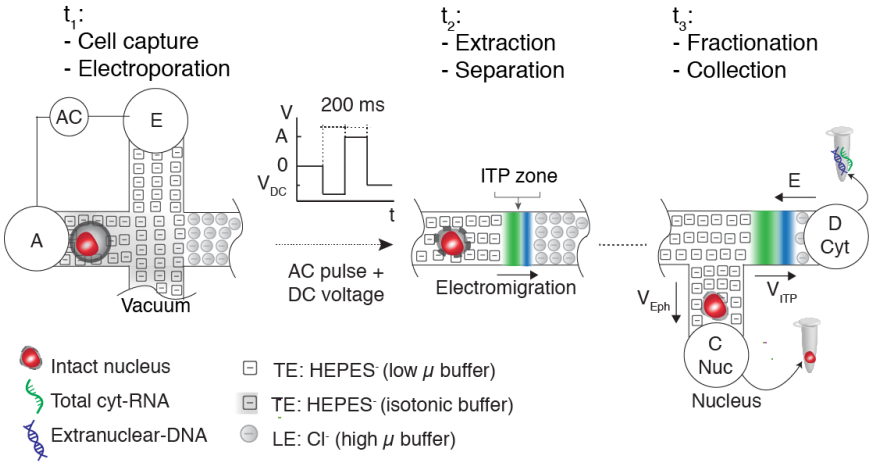


Figure 4-2. *Schematic of our published preliminary system for isolation and processing of cyt-RNA versus nucleus from single cells. Isolated single cells were electrically lysed with end-channel electrodes. Cyt-RNA was extracted from the lysed cell, purified, and focused into a discrete ITP zone within 1 s. The nucleus is not focused by ITP but conveniently follows the ITP zone at a slower drift velocity, enabling fractionation downstream. We control end-channel electrodes to divert*

the cyt-RNA focused zone and the nucleus to different respective outputs. We show cyt-RNA zone migrating through the T-junction region to the cyt-RNA reservoir.

4.2.3 Multiplex gene expression analysis in human leukemia K562 cell line

To evaluate extraction efficiency and sensitivity, we have completed a set of experiments demonstrating multiplexed gene expression from total cyt-RNA. For this, we used an off-the-shelf microfluidic chip (NS12A, Perkin Elmer) with manual isolation of single cell, and off-chip targeted pre-amplification of genes with varying expression levels. We used the K562 cell line and targeted one oncogene GATA1 (typically over-expressed in leukemia cells) and the housekeeping genes (GAPDH, Actin beta, HPRT1, and PPP1CB). Figure 4-3 summarizes the gene expression data of off-chip qPCR. To evaluate our RNA extraction efficiency, we spiked an external control of synthetic RNA (ERCC) at a copy number of 1500 copies per single cell, similar to the average expression of GAPDH.[205] GAPDH of housekeeping gene showed similar average *Ct* value to that of the external control, suggesting successful extraction of RNA from single cell. Slightly larger variation in GAPDH *Ct* reflects biological noise from the stochastic nature of gene expression. We explored also lower abundant genes such as PPP1CB and HPRT1 and successfully observed their repeated amplification.

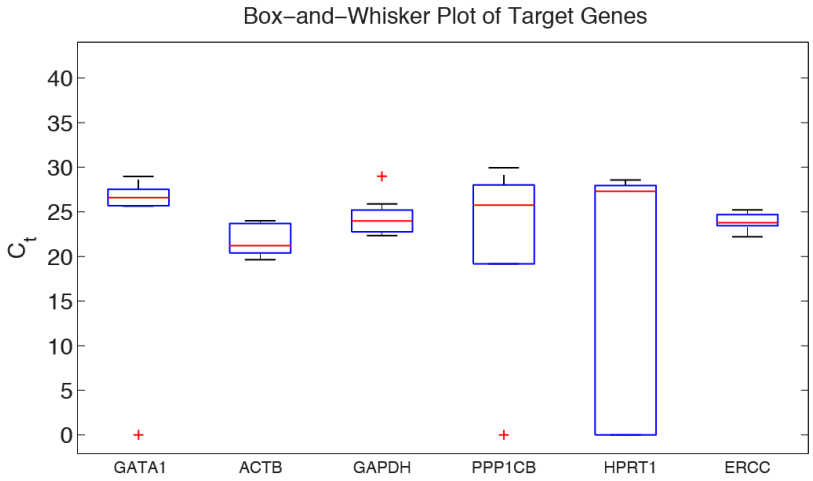


Figure 4-3. *Single-cell gene expression data for mature cytoplasmic mRNAs. Ct values from multiplexed qPCR analysis of 8 single cells. We used targeted pre-amplification and qPCR to quantify six genes of varying expression (GATA1, GAPDH, Actin beta, HPRT1, and PPP1CB). Horizontal (red) line shows the median value, the box, 25th and 75th percentile, and uncertain bars show one standard deviation of the underlying distribution (not confidence on the mean). Red crosses indicate data outliers.*

4.2.4 Single-cell nuclear vs. cytoplasmic RNA-seq

The outputs of the chip are corresponding to the extracted and purified cyt-RNA and nuc-RNA of each of the injected cells. We here used these aliquots for library prep in downstream NGS analyses.

We carried out the RNA-seq experiments of the whole transcriptome for the nuclear and cytosolic compartments by using the SMART-seq protocol (SMARTer Ultra Low RNA Kit for Illumina Sequencing by Clontech) and the Nextera XT kit for Illumina (c.f. Figure 4-4). We then pooled the libraries and sequenced the

fractions on a high-throughput sequencing machine (Illumina, HiSeq2500). Although we recovered total nuclear and total cytosolic RNAs including both poly-A+ and poly-A- transcripts, we selected only for the poly-A+ transcripts during the first-strand cDNA synthesis step in the SMARTer protocol.

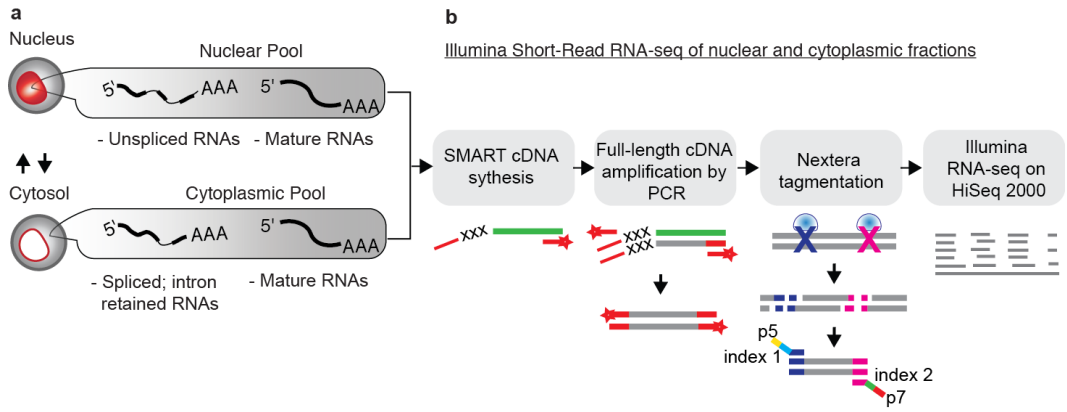


Figure 4-4. *(a) Schematics showing sub-cellular compartments and target transcripts. (b) Experimental workflow for nuclear and cytosolic pools using SMARTer cDNA prep and Nextera XT tagmentation protocol on Illumina platform.*

In this study, we generated a whole transcriptome map of RNA transcripts and their sub-cellular localizations for 12 (LCL, Snyder) lymphoblastoid cell fractions, and 22 K562 chronic myelogenous leukemia cell fractions, representing two Tier 1 ENCODE cell lines. To characterize the extent of fraction-to-fraction variability, we further analyzed two sublines of the K562 cells from ATCC and ENCODE stocks. This is an attractive model for the study of gene and isoform expression because export and processing of RNA in the nucleus and cytosol have been strongly linked to cell differentiation and cancer.

We sequenced all libraries to an average depth of 29.6 million reads per cell (Figure 4-5.a), determined the libraries which align at <20% to the human reference genome (hg19), and discarded from the data set. We then characterized the genomic alignment rates and the percentage of reads which mapped to multiple loci, and found the median genomic alignment rates to be 62.8% and 38.4% for the K562 and LCL fractions (Figure 4-5.b), respectively with 2.83% and 1.48% of duplicate reads (Figure 4-5.c).

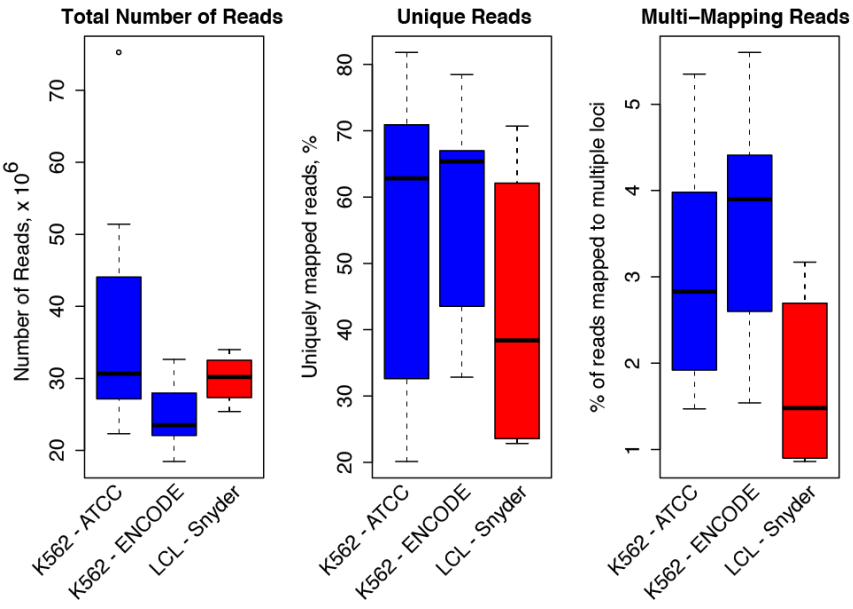


Figure 4-5. *Library quality metrics for K562 (source ATCC), K562 (source ENCODE), and LCL (source Snyder) cells. (a) Total number of sequencing reads. (b) Percetage of uniquely mapped reads to the hg19 human genome. (c) Percentage of duplicate reads mapped to multiple loci.*

4.2.5 Metrics for library quality

We next evaluate the uniformity of read coverage by examining if there is 3' to 5' gene bias. All cytoplasmic single cell fractions show little 3' bias (Figure 4-6), and only 3 nuclear fractions (denoted on Figure 4-6) have some non-uniformity. Although the starting pools for nuclear and cytoplasmic RNA-seq are polyadenylated (poly-A+) RNAs, these transcripts represent heterogeneous molecules at different processing stages. A common source of sequencing bias is the amount of intronic regions in immature polyadenylated (not-yet-spliced) transcripts, particularly in the nucleus and to a smaller extent in the cytosol. This is most common in long genes, which are highly expressed. The presence of introns in the sequencing pools affects the detection of gene and isoform expression levels and might cause false discovery of splice junctions.

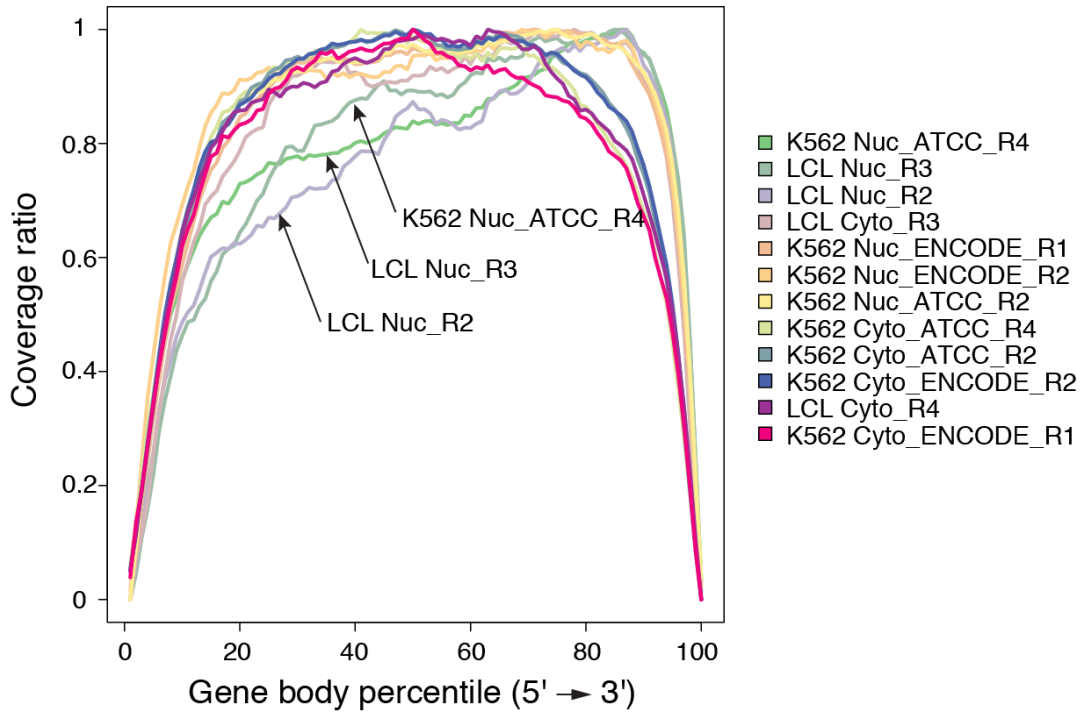


Figure 4-6. *Quality control for gene body coverage. Plot for normalized RNA-seq gene coverage from 5' to 3' end (left to right) for 12 selected sample fractions (6 cytoplasmic and 6 nuclear) calculated based on the Pearson's skewness coefficients. All factions are ranked by skewness of coverage, and samples with worst coverage are displayed on top of the figure legend. All fractions show little to no bias except for 3 nuclear cases (denoted with arrows).*

4.3 Discussion

4.3.1 Gene expression correlations

The gene expression correlations of replicates between single and multiple cells have higher correlations (Pearson's correlations of 0.7-0.74 for K562 cells and 0.49-0.54 for LCL cells; on log FPKM-scale, Figure 4-7.a,b, Figure 4-8.a,b, and

Figure 4-9a,b) compared to those between single fractions. We observe larger variations between individual fractions of same cell-to-cell nuclear and cytosolic compartments (Pearson's of 0.58-0.65 for K562 and 0.41-0.45 for LCL) (Figure 4-7.c,d, Figure 4-8.c,d, and Figure 4-9.c,d) and nuclear vs cytosolic compartments (Pearson's of 0.53-0.67 for K562 and 0.44-0.48 for LCL) of the same single cell (Figure 4-7.e,f, Figure 4-8.e,f, and Figure 4-9.e,f). While there are no substantial differences in the gene expression variations of same cell compartments and across cell compartments, there is larger variation for all fractions of the LCL cells compared to those of the K562 cells. Because lymphoblastoid cells inherently have smaller amounts of total RNA, we hypothesize that these variability differences are caused by sources of technical bias associated with the library preparation protocol. Despite the large cell-to-cell variation between whole cells of K562 and LCL cells (Pearson of 0.59 for 7 replicates), we find a tight correlation (Pearson of 0.8 for 7 replicates) between K562 ATCC and ENCODE whole cells Figure 4-10.a,b. To further verify that these variations are caused by true biological difference and not technical noise due to the small input RNA amounts, we show aggregate data of 20 fractions for nuclear K562 vs LCL (Pearson of 0.53; Figure 4-10.c), and cytoplasmic K562 vs LCL (Pearson of 0.53; Figure 4-10.d). To determine the statistical significance between gene expression data of two samples, we also compute the Spearman's rank coefficient which ranges form 0.47 to 0.67 for all single fractions.

To get whole cell gene assemblies, we first merged the individual assemblies corresponding to nuclear and cytoplasmic fractions of the same cell using Cuffmerge (an assembler which merges transfragments parsinoniusly, also part of

the Cufflinks package). For the gene epxpression data shown in Figure 4-7, Figure 4-8, Figure 4-9, and Figure 4-10, we calculated the differential expression in pairs of whole cells and sub-cellular fractions using Cuffdiff (a program, part of the Cufflinks package) and FPKM values with the assumption that the number of reads is proportional to gene abundance.

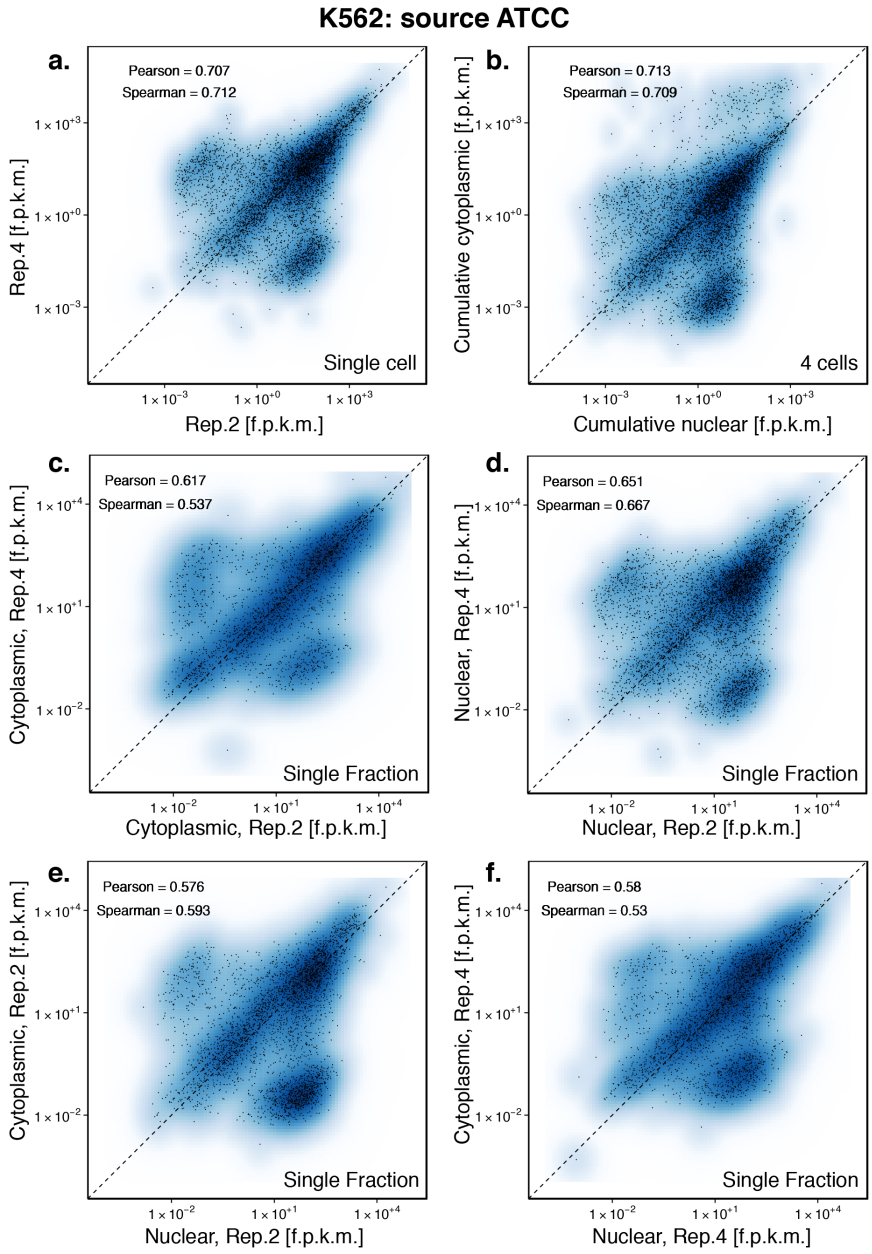


Figure 4-7. *Gene expression correlations for K562 (source ATCC) cells. We show the Pearson's and Spearman's coefficients of global gene expression for: (a) Single cell. (b) 4 cells; (c) cytoplasmic vs cytoplasmic. (d) nuclear vs nuclear, and (e,f). cytoplasmic vs. nuclear single fractions.*

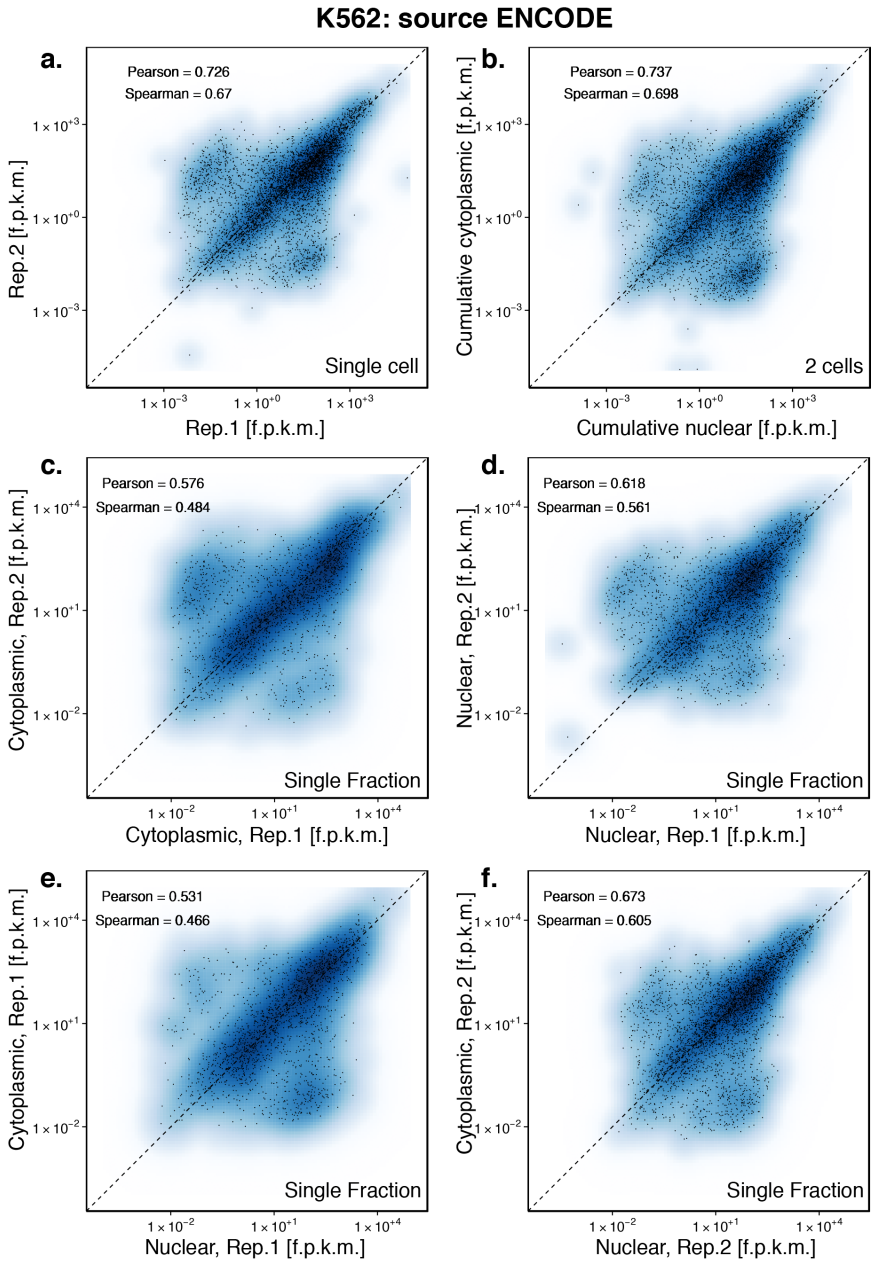


Figure 4-8. *Gene expression correlations for K562 (source ENCODE) cells. We show the Pearson's and Spearman's coefficients of global gene expression for: (a) Single cell. (b) 4 cells; (c) cytoplasmic vs cytoplasmic. (d) nuclear vs nuclear, and (e,f). cytoplasmic vs. nuclear single fractions.*

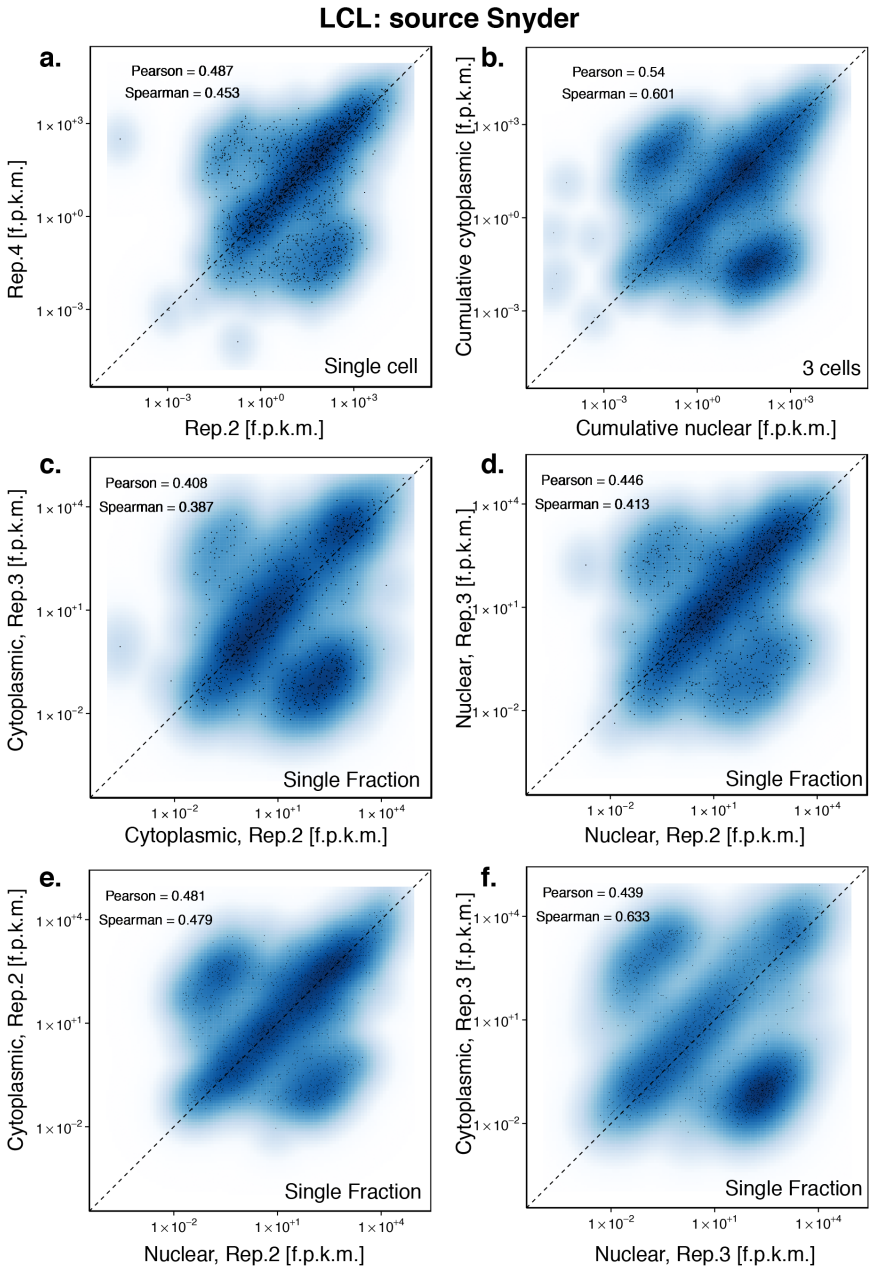


Figure 4-9. *Gene expression correlations for LCL (source Snyder) cells. We show the Pearson's and Spearman's coefficients of global gene expression for: (a) Single cell. (b) 4 cells; (c) cytoplasmic vs cytoplasmic. (d) nuclear vs nuclear, and (e,f). cytoplasmic vs. nuclear single fractions.*

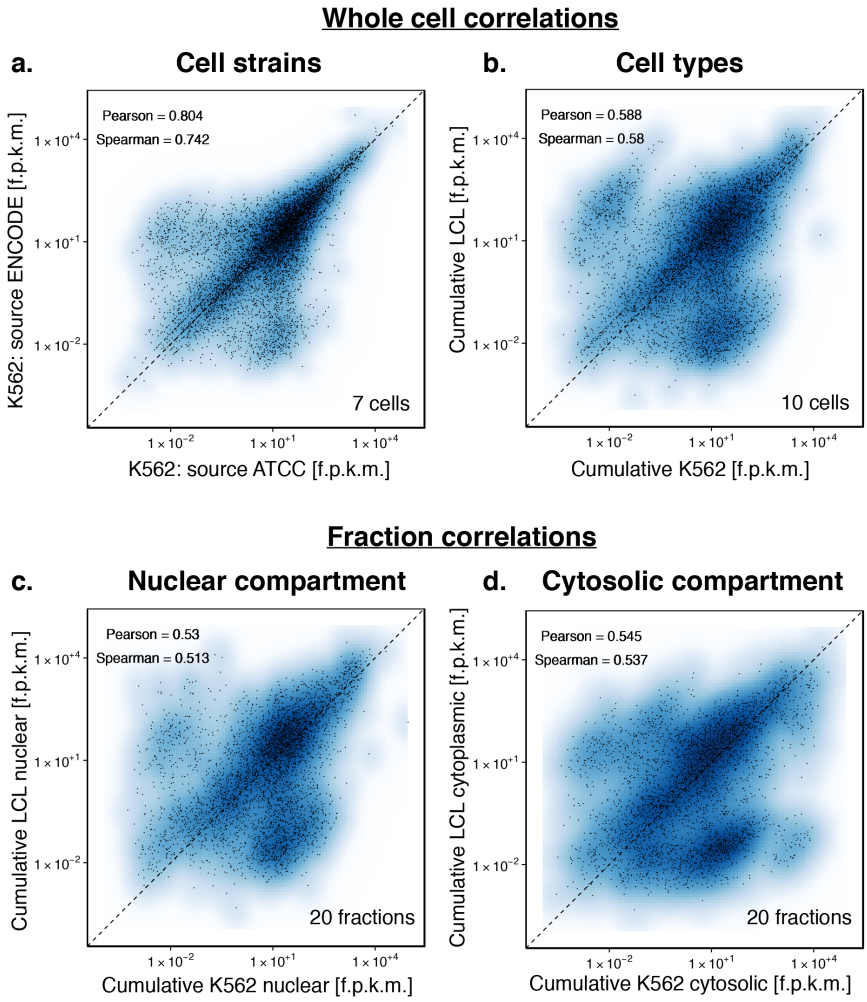


Figure 4-10. *Gene expression correlations for cell strains, cell types, and within compartments. We show the Pearson's and Spearman's coefficients of global gene expression for: (a) K562 (ENCODE) vs K562 (ATCC). (b) K562 vs LCL. (c) Cumulative nuclear K562 vs nuclear LCL. (d) Cumulative cytoplasmic K562 vs cytoplasmic LCL.*

4.3.2 Principal component analysis (PCA) and correlation matrix

To explore single cell and single fraction features of each cell type (K562 chronic myelogenous leukemia vs Snyder lymphoblastoid), we generated a spatial map for the most highly expressed genes across the data set. We retained only 15,413 genes, which were expressed in all cell fractions, and out of those, we selected 6966 that were expressed in at least 3 fractions at an expression of more than 1000. We then applied the Seurat (ref) method (an R package for single cell RNA-seq data) to identify the most variable genes (a total of 53) across all fractions. These genes have the highest z-score of variance over mean for a certain average expression and are placed into 20 bins (Figure 4-13). We used these high variability genes and linear principal component analysis to survey clustering of single cell nuclear and cytosolic fractions (Figure 4-13.a,b).

4.3.3 Gene density and transcriptome-wide variability

To demonstrate the variability of each RNA-seq data set, we calculated the squared coefficient of variation (squared-normalized standard deviation) for replicates of K562 and LCL cells for transcriptome-wide gene and isoform level distributions in each cell compartment (Figure 4-11). More generally, we observed greater fraction-to-fraction compartment variability for K562 cells across a wide range of transcript expressions. We explored both the gene and isoform expression in single cell's nucleus and cytosol and found that the variability of gene levels is larger in the cytosol compared to that in the nucleus, whereas the variability of isoforms remains constant across compartments. Similarly, studies of single cell expression and splicing distributions proposed that individual cells with several splice isoforms mostly expressed a single isoform, and thus show less variation in

isoform expression. Furthermore, we observe higher degree of heterogeneity of gene expression in the cytosol for highly abundant transcripts (calculated on average transcriptome-wide) across the two cell types. The subcellular gene density distribution is largely bimodal for the lymphoblastoid cells both in the nucleus and cytosol while that of K562 leukemia cells is more unimodal (Figure 4-12). We thus hypothesize that this gene distribution and bimodality might be related to functional response in immune response of the lymphoblast cells.

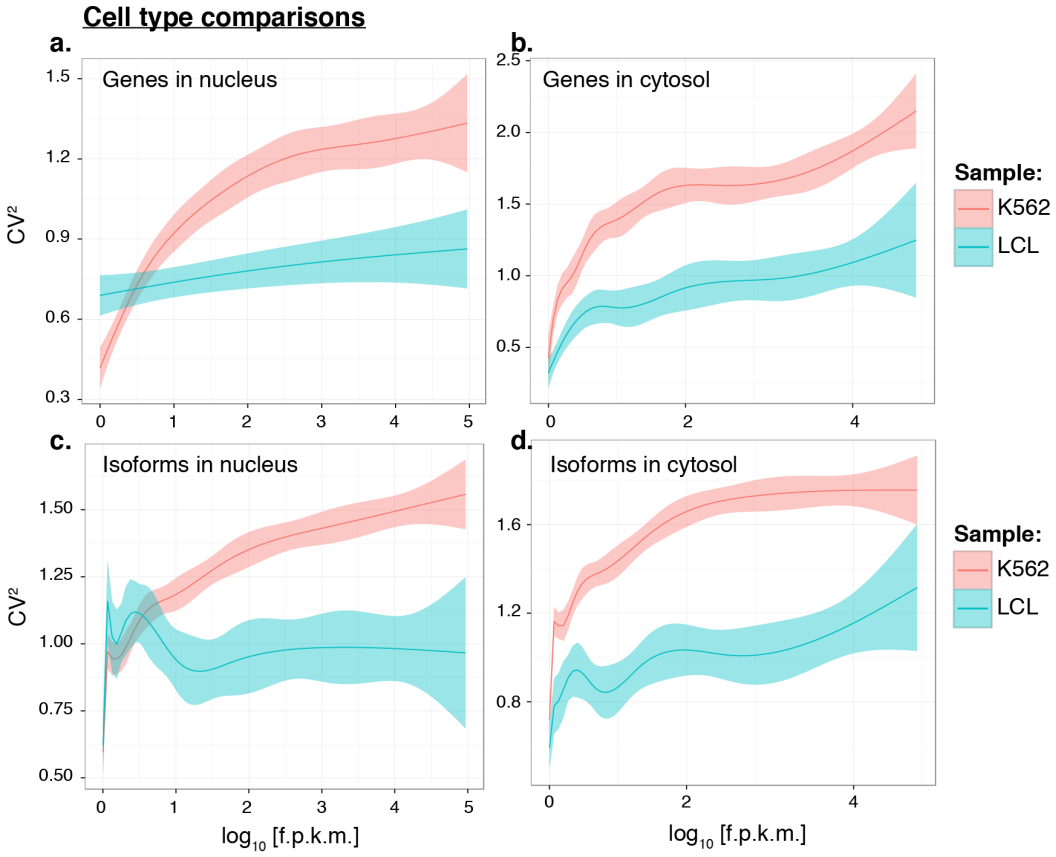


Figure 4-11. *The squared coeefficient of variation for transcript expression (in log10-transformed FPKM) of (a) genes in the nucleus, (b) genes in the cytosol, (c) isoforms in the nucleus, and (d) isoforms in the cytosol for RNA-seq data of K562 and LCL cells. The squared coefficient of variation is a metric for variability of each individual sample dissected for subcellular compartments.*

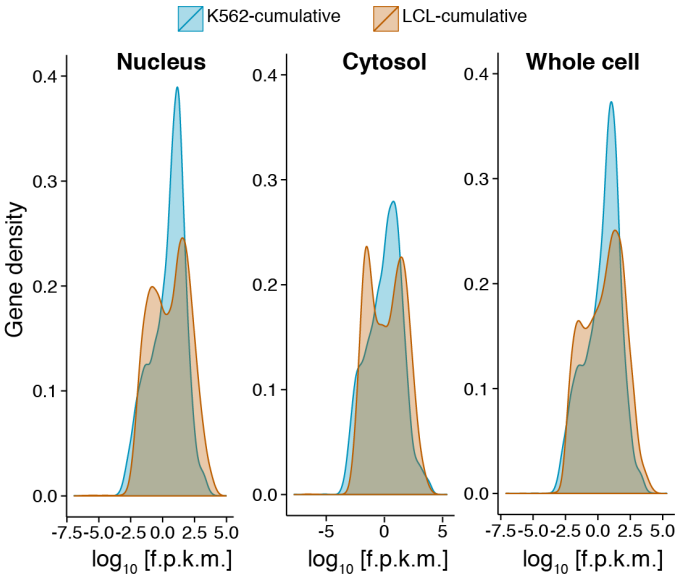


Figure 4-12. *Gene density plots for gene expression (in log10-transformed FPKM) in the nucleus, cytosol, and whole cell of individual cell types. The gene expression distribution for LCL cells is of bimodal character, whereas that of K562 cells is unimodal.*

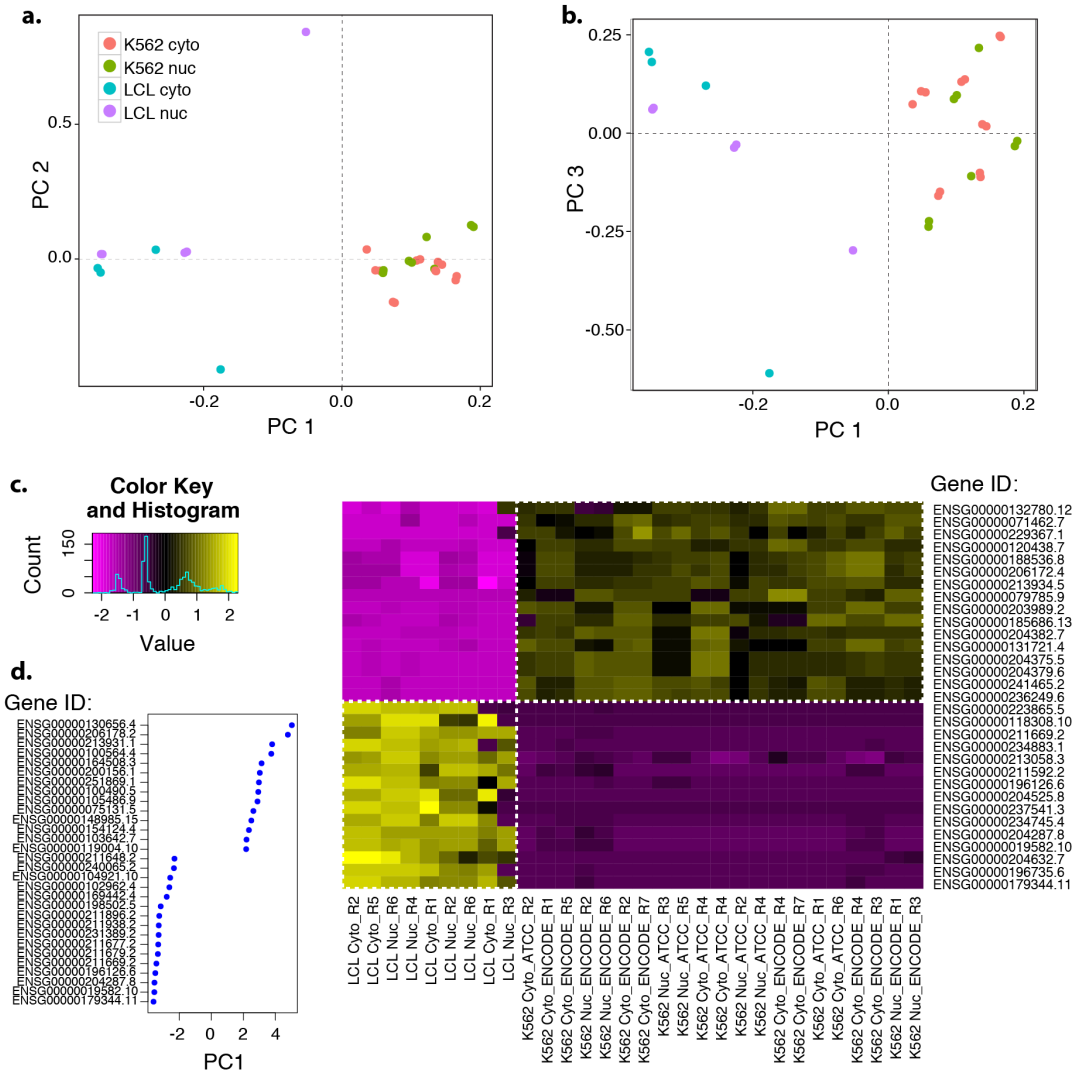


Figure 4-13. *Single-nuclear and single-cytoplasmic RNA-seq of K562 and LCL cells. (a,b) Principal component analysis (PCA) of 30 single cell nuclear and cytoplasmic fractions. Each cell population is based on the differentiation correlation with PC1, PC2 and PC1, PC3. (c) Hierarchical clustering of RNA-seq identifies the K562 and LCL cell populations. Each row represents a single-cell nuclear or cytoplasmic fraction and each column a gene (a total of 32). Cell fraction replicates and gene scores are arranged by PC score. (d) Ranked genes with respect to the first principal component (PC1).*

We observe that genes associated with high variability are cell type specific. Hierarchical clustering generated from the K562 ATCC, K562 ENCODE, and LCL Snyder shows cell type specific sets of genes highly associated with cell type (Figure 4-13.c,d). This analysis also shows that fractions from different cellular compartments cluster within their cell type, which suggests that scITP-seq is able to deconvolute samples mixtures of heterogeneous sub-cellular fractions. Systematic analysis of the z-score (number of standard deviations from the mean) reveals cell-type specific genes with highest variance. For example, we find that some of the most highly expressed genes in K562 cells are GAGE12I, NASP, PRAME, and DDX1, which are associated with tumorigenesis, cell growth and division, whereas genes such as IGLV3-10, HLA-DRB1, and HLA-B involved in immune function are prevalent in LCL cells. We also find that in K562 cells housekeeping genes (ACTB and GAPDH) are less variant than genes linked to the GATA transcription factors, and expression differences in GATA1 and PU.1 related transcription factors delineate subpopulations . In addition, we show that the gene expression of two transcription factors, NFKBIE and JUN is localized exclusively in the cell nuclear and cytoplasmic compartments, respectively. Previous reports suggest that these spatio-temporal fluctuations of transcription factors drive chromatin accessibility and regulate gene expression (Figure 4-14).

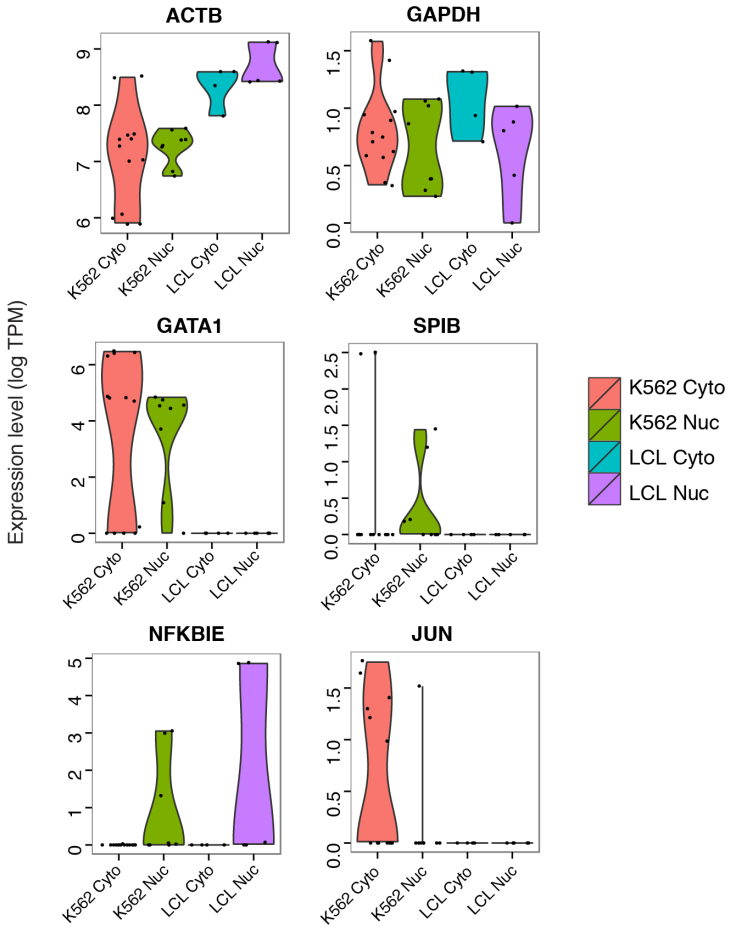


Figure 4-14. *Gene expression levels (in log10[TPM])) separated in nuclear and cytosolic compartments for housekeeping genes (ACTB and GAPDH), tumor-promoting genes (GATA1 and SPIB), and transcription factors (NFKBIE and JUN).*

Finally, we show that nuclear fractions of K562 cells obtained from ATCC and ENCODE have differences in GATA1 and SPIB (transcription factor related to PU.1) expression. These differences have been previously associated with subpopulations of erythrocyte and megakaryocyte cell lineages. Overall, these findings indicate that sub-cellular gene expression in the nucleus and cytosol is

cell-type and subpopulation specific, and largely associated with gene expression heterogeneity.

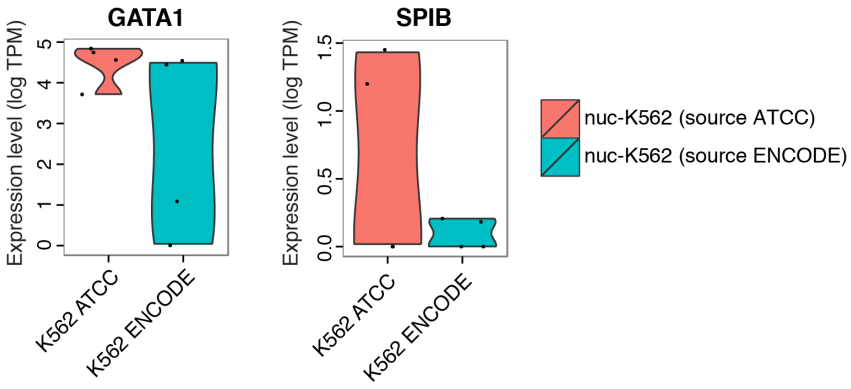


Figure 4-15. *Nuclear expression (in log10[TPM])) of GATA1 and SPIB factors for two subpopulation of K562 cells (ATCC and ENCODE). Gene expression levels reveal presence of subpopulations related to lineage differentiation within the K562 cells.*

4.3.4 Nuclear and cytosolic distribution of gene features

To quantify the relative enrichment of CDS exons, 5'UTR exons, 3' UTR exons, and introns in each fraction, we calculated the distribution count of these gene features in each subcellular fraction (shown in Table 4-1). We used the RSeQC package and the read_distribution.py script to assign the mapped reads to the proper gene region. We find that the percentage of introns is 8.1% vs 19.9% for the mean distribution in the cytosol and nucleus, respectively, while the CDS exons are nearly equally distributed at percentages of 44.7% (in cytosol) vs 40.1% (in nucleus). We also find that the 3'UTR exons are localized in the nucleus at 15.8% vs 10.8% in the cytosol. Finally, we observe that LCL cells have more nuclear intronic reads and thus unprocessed transcripts compared to K562 cells. These data corresponds well with their bimodal distribution of gene expression in

LCLs as shown in (Figure 4-16). As expected there are more reads mapping to introns in nuclear RNA even though we selected for poly-A+ transcripts during cDNA generation in the library preparation protocol. Because splicing is a predominantly co-transcription process localized in the nucleus, nuclear RNA is the preferred choice to study splicing events in single cells. On the other hand, cytoplasmic RNA is enriched for CDS exons in comparison to nuclear RNA, and thus analysis of cytoplasmic RNA enables a superior method for the study of mature protein coding genes and alternatively spliced genes.

Table 4-1. *Gene type for polyadenylated RNAs in K562 and LCL cell lines.*

Group	Bulk RNA	K562-Cyt-Mean	K562-Nuc-Mean	LCL-Cyt-Mean	LCL-Nuc-Mean	Fraction-Cyt-Mean	Fraction-Nuc-Mean
Introns	0.076	0.060	0.180	0.123	0.237	0.081	0.199
CDS_Exons	0.433	0.478	0.415	0.387	0.372	0.447	0.401
3'UTR_Exons	0.296	0.107	0.154	0.110	0.164	0.108	0.158
5'UTR_Exons	0.030	0.042	0.042	0.030	0.034	0.038	0.040

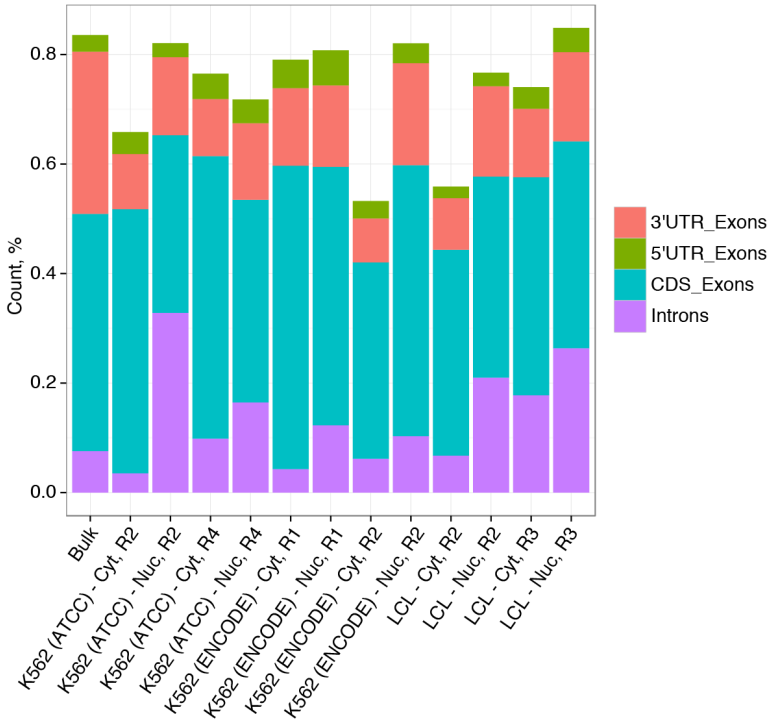


Figure 4-16. *Gene feature distribution of non-coding (3'UTR and 5'UTR exons, CDS exons, and introns) for bulk and single-cell nuclear and cytosolic replicates of K562 (ATCC), K562 (ENCODE), and LCL cells.*

We next show example genes of ACTB, GAPDH, and HBA1 at different maturation states for 3 replicates of nuclear and cytoplasmic K562 single cell fractions (Figure 4-17). Events of unspliced introns in the nucleus reflect transcripts, which have not been processed yet, while events of unspliced introns in the cytosol suggest intron retention. We then examined the isoforms of METTL5 and GATA1 to demonstrate instances of exon inclusion and intron retention, as revealed in the cytosolic RNA fractions (Figure 4-18).

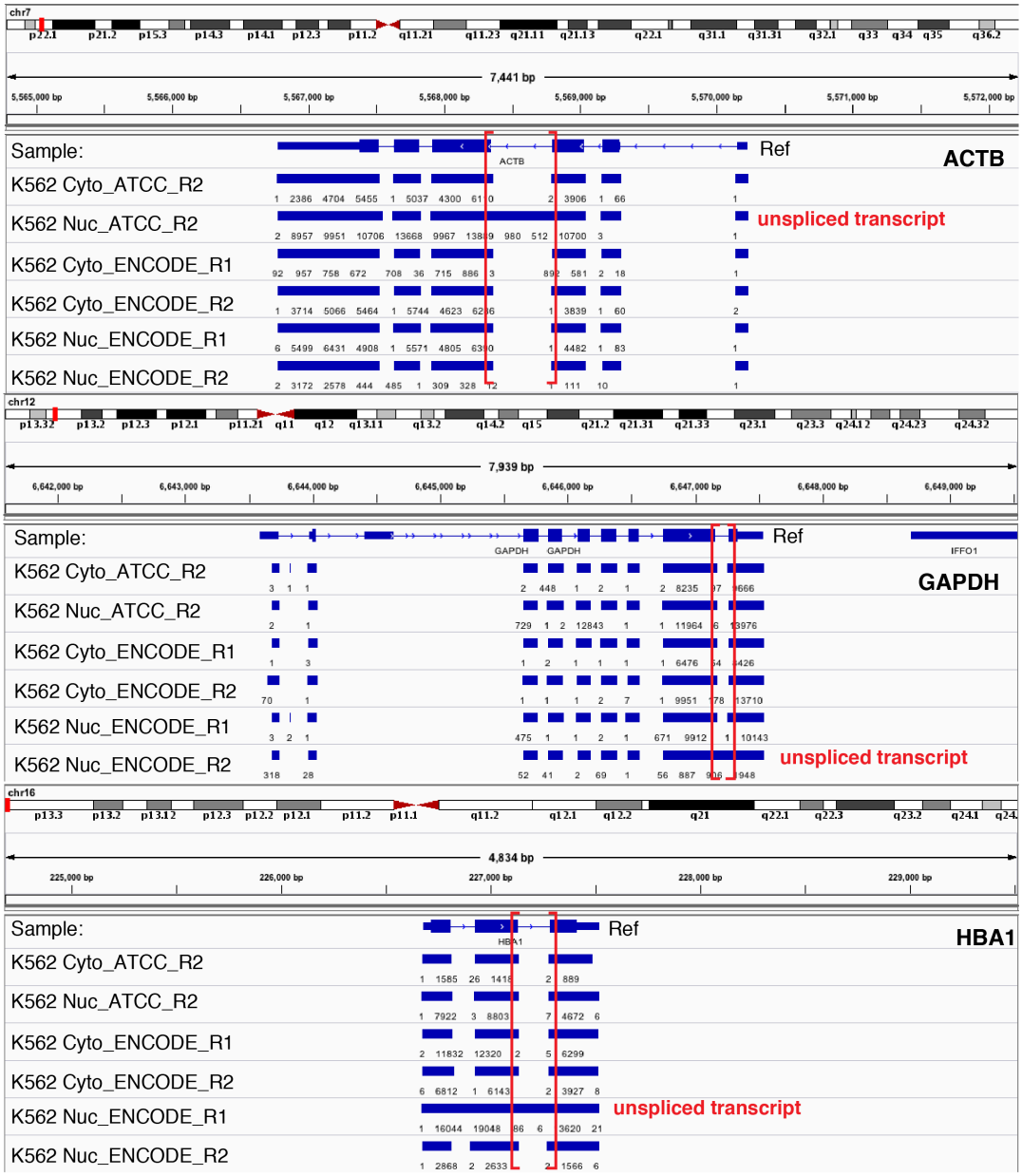


Figure 4-17. *Variation in nuclear and cytosolic RNA expression between single-cell fractions of 3 K562 cell replicates. IGV screenshots show read density of ACTB, GAPDH, and HBA1 genes. For each of these genes, we mark unspliced transcripts in the nucleus in red.*

Figure 4-18. *Variation in nuclear and cytosolic RNA expression between single-cell fractions of 3 K562 cell replicates. IGV screenshots show read densities of METTL5 and GATA1 genes. For each of these genes, together with unprocessed transcripts in the nucleus, we show alternatively spliced transcripts in the cytosol, marked in red. Specifically, we illustrate events of intron retention and exons inclusions for these example genes.*

4.3.5 Comparison against genome-wide long-read RNA-seq measurements of splicing completion

In our short-read RNA-seq experiments, we observe high integrity of the transcripts extracted by sc-ITP, which is essential for uniform full-length coverage and minimal 5' end bias. In Figure 4-6, for certain nuclear samples we observed

bias near the 5'-end of the transcripts. However, to enable accurate analysis of splice sites and alternative splicing, it is important that we ensure full coverage near the 5'-splice sites with minimal biases.

To investigate alternative splicing and characterize gene isoforms, we perform long-read sequencing of RNAs extracted from the cytosol (to exclude unspliced genes during transcription). Long-read RNA-seq methods are the "holy grail" of all sequencing technologies because they provide a comprehensive picture of exons/introns co-associations or exclusions, which are difficult to address using short-read technologies.[206, 207] Because splicing is predominantly a co-transcriptional process, [4] the cytosolic fraction is free of unspliced ("not-yet-processed") transcripts and is a "true" representation of alternatively spliced exons, introns, deletions, rearrangements, or other alternative splicing scenarios. To analyze these patterns, we use the SLR-RNA-seq method, which is based on the MOLECULO technology on bulk fractionated samples of K562 and GM12878 cells. For this project, we use this method to execute full transcriptome analysis on bulk samples which will serve as a comparison benchmark against our single-cell nuclear and cytosolic RT-qPCR data.

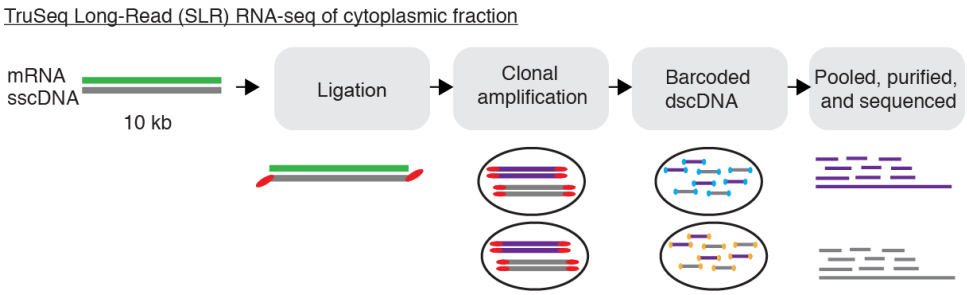


Figure 4-19. *Workflow for SLR-RNA-seq of the cell cytosolic content. Long read RNA-seq view allows for determining instances of intron retention and alternative splicing in mature transcript situations which cannot be correctly determined using traditional short-read RNA-seq.*

For the SLR-RNA-seq protocol, we prepared single-strand cDNA (sscDNA) that includes adapters containing PCR-primer sites at the beginning and the end of each cDNA molecule (c.f. Figure 4-19). Based on qPCR, approximately 1,000 such sscDNA-molecules are added into each well of a 384-well plate. In each well, sscDNA is amplified and the resulting double-stranded cDNA (dscDNA) molecules are fragmented and barcoded, so that each fragment can easily be assigned to its well. After sequencing using an Illumina HiSeq2000, 2x125bp paired-end reads are assembled into high coverage RNA-contigs in a well-specific fashion. For most genes, this minimizes the possibility of a non-identical molecule from the same locus interfering with the assembly (which we refer to as a "collision" of two non-identical transcripts of the same gene) – and fixed tags at each end of each transcript further reduce the possibility of interference from non-identical molecules, as long as they do not have identical transcript starts and ends. We have performed fairly extensive characterizations of our preliminary

fractionation using RT-qPCR and qPCR with sequence-specific probes for extracted nuc-mRNA, gDNA, cyt-mRNA, and one small nuclear RNA (snRNA) for benchmarking experiments. We describe these gene expression experiments and the comparisons to this bulk RNA-seq data in detail in Section 4.3.6 below.

4.3.6 Validation of splicing patterns of GAPDH gene in single-cell nuclear and cytosolic compartments via RT-qPCR

We further validated our approach by measuring the copy number (Figure 4-20.a) and gene expression (Figure 4-20.b) of a housekeeping gene (GAPDH) for processed (spliced) and not yet processed (unspliced) transcripts from a total of 192 outputs, representing 2 Tier 1 ENCODE cell lines (LCL-Snyder lymphoblastoid cells, and K562 chronic myelogenous leukemia cells), as well as K562 cells obtained from ATCC. The LCL-Snyder cells are the same cell type as GM 12878 but obtained from a different individual. We obtained and analyzed K562 cells from 2 sources, one from ATCC (American Type Culture Collection, Inc.) which we name sub-line K562-ATCC and one from stocks prepared and used in the ENCODE project, sub-line K562-ENCODE.

We focused our analysis on lymphoblastoid and chronic myelogenous leukemia cells because these cell types have been extensively analyzed and are the only two cell types for which bulk transcriptomic data sets are available within nuclear and cytosolic compartments. We explored the DNA content within each compartment of the three cell types by measuring the copy number of GAPDH (glyceraldehyde 3-phosphate dehydrogenase) gene (Figure 4-20.a). The LCL cells have a diploid genome and are not expected to show a copy number variation, unlike K562 cells, which are of triploid character and expected to show such variations. To quantify

these copy number variations across the two cell types, we first choose to analyze genomic-DNA (DNA localized in the nucleus) via a set of primers, which flank exon-intron boundary and one-step single-cell qPCR. We then calculated the copy number expression distribution in Log2-transformed space of the observed threshold cycle minus the threshold cycle for the background signal (set at Ct of 40). We found high abnormal variations in the copy number of K562 cells (sub-line ENCODE), while K562 (sub-line ATCC) cells showed only sporadic "burst-like" abnormal CNV's in some but not all biological replicates. In contrast, we find no abnormal CNV's for the lymphoblastoid cells. Finally, we looked for the presence of free DNA in the cytosolic compartments (Figure 4-20.b bottom) and discovered considerable DNA content (with a median of nearly 3 fold) in K562 (ENCODE) cells. LCL and K562 (ATCC) samples mostly did not show any extranuclear DNA content.

Extending this analysis to gene expression, we next assessed variations (Figure 4-20.b) among the expressions of "non-yet processed" and "mostly processed" transcripts of the GAPDH gene in each sub-cellular compartment for the LCL (Snyder), K562 (ATCC), and K562 (ENCODE) cells. Using one-step RT-qPCR protocol, we analyzed relative expression distributions of mature and precursor transcripts in nucleus and cytosol to find 2-to-4-fold variability range for spliced transcripts, and up to 8-fold for unspliced RNAs.

The gene expression distributions of processed transcripts in the nucleus and cytosol for the three samples showed significant statistical variations among the distribution means. Nuclear unspliced was the only compartment/probe pair in which there was no statistical significance in gene expression variation among the

lymphoblastoid and leukemia cells. We examined these statistical significances by one-way ANOVA and the non-parametric Kruskal-Wallis tests.

To evaluate the process of splicing at a single cell level in sub-cellular compartments, we introduce the Percent Spliced Introns (PSI) ratio, which is a measure of successful splicing completion of transcripts in the nucleus and the Percent Retained Introns (PRI), which is a measure of the retention of introns in transcripts already exported to the cytosol. Based on the relative expression of spliced and unspliced transcripts, we compute the PSI and PRI ratios, corresponding to the percentages of spliced and percentages of retained of introns at the interrogated exon-intron boundary for the GAPDH gene isoforms (in Figure 4-20.c). A PSI value of 1 (or 100%) indicates complete splicing, while a PRI value of 0 (0%) indicates no intron retention, and PSI = 1 – PRI. We also show the distributions of GAPDH PSI and PRI values for the lymphoblastoid and leukemia cells in Figure 4-20.d,e.

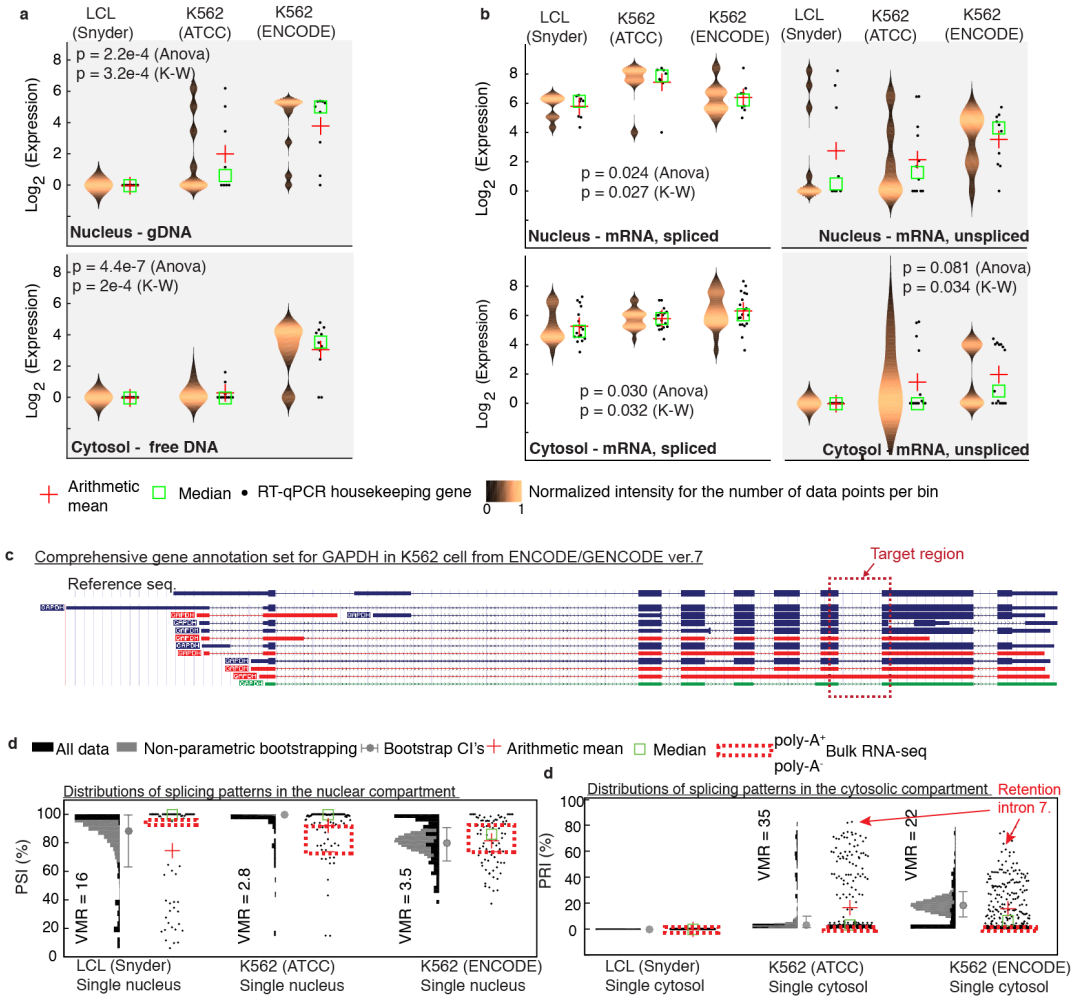


Figure 4-20. *Comparison of DNA and mRNA expression distributions for a housekeeping gene (GAPDH) between samples analyzed in the nuclear and cytosolic compartments (based solely on qPCR data). (a) Frequency distribution of gDNA vs extra-nuclear DNA amounts from single-cell qPCR shown as violin plots for LCL (Snyder), K562 (ATCC) and K562 (ENCODE) cell lines. The expression (vertical axis) is the Log2-transformed fold change over the background signal level for the data in each compartment. Shown are violin plots (left of each pair of columns) and corresponding raw data (right). Width of the violin plot indicates the frequency of expression level, whereas the color map is an indication of normalized intensity for the data points in each bin. We analyzed the variation*

among different sample means with ANOVA and a non-parametric test (Kruskal-Wallis), and determined statistical significance in the variation of sample means. (b) Frequency distribution of spliced- vs unspliced- mRNA amounts from single-cell RT-qPCR. Violin plots are presented as in (a). Based on ANOVA and Kruskal-Wallis tests, we determined statistical significance for the expression of spliced and unspliced genes in all compartments with the exception of nuclear unspliced. (c) Gene isoforms of GAPDH in K562 cells. We show the comprehensive annotation from ENCODE/GENCODE database (ver.7). Histogram (in black) and scatter plots of PSI (d) and PRI (e) values for total RNA transcripts localized in nucleus and cytosolic compartments. Bootstrapped data (shown in grey) and 95% CI intervals overlays the data plots. VMR values next to histogram plots represent Log2-transformed variance-to-mean ratios calculated for the data, and are measures of the dispersion underlying the distributions. Red squares form the upper- (poly-A+) and lower (poly-A-) bounds calculated from bulk long-read RNA-seq data for the target intron.

CONCLUDING REMARKS

The average mapping coverage of compartment-fractionated single cells is 62.8% and 38.4% for chronic myelogenous leukemia, K562 and lymphoblastoid, LCL cell lines, respectively. On average in the nucleus, for each cell line 19.9% of genome is covered by introns, and 40.1% by CDS exons, whereas in the cytosol, 8.1% is covered by introns, and 44.7% by CDS exons. When mapping reference bulk (whole cell, non-fractionated by compartment) RNA-seq data, we observe 7.6% coverage by introns and 40.3% coverage by exons. The consistently larger intron coverage in the nuclear compartments of all single cell fractions stems from not-yet-processed polyadenylated RNAs. In the nucleus, this figure defines a lower limit of intron coverage because we selected only for polyadenylated RNAs during the library preparation protocol. Other than the higher prevalence of introns in the nucleus, these estimates in the cytosol are in reasonable agreement with the bulk RNA-seq data.

In this study, we dissected single-cell transcriptomic heterogeneity and assessed the variability within each subcellular compartment (i.e. nucleus and cytosol) and across compartments. We identify genes associated with high variance in both compartments of each cell type. Furthermore, the expression levels and subcellular localization of lineage-specific genes suggest that single-cell fraction variance is an essential characteristic of different biological conditions. Together these reported data postulates a new biological hypothesis that that abnormal deviations in alternative splicing and gene isoforms are associated with transcriptome variation in both developing and mature cells, and these deviations are often associated with onset and progression of disease. Accurate quantification of alternative splicing is

possible only if the “not-yet-spliced” transcripts are physically removed from the pool of total transcripts. By individually analyzing nuclear and cytoplasmic compartments of a single cell at a whole transcriptome level, we achieve an unprecedented precision in single cell splicing quantification.

FUTURE STUDIES

Future technological improvements of our scITP-seq system and method will improve on RNA recovery, throughput, ease-of-use, and data analysis. To achieve higher throughput, we envision a novel parallel and highly automated system to fractionate single cells by compartment. The chip will have a 12.8 x 8.5 cm footprint and 24 output reservoirs compatible with robotic pipetting. A single aliquot of about 10 mL containing about 5 cells/mL will be dispensed into a single chip input using a standard micropipette. The system will drive flow from an input well to a waste well past individual, self-limiting (to single cell) cell traps. The cytoplasmic membrane of each cell will be electrically lysed, and the cyt-RNA thereby released. Cyt-RNA will be rapidly purified and focused via ITP within each of the 12 parallel channels. The nucleus will remain at this time trapped in the cell trap and not focused in ITP, enabling fractionation of total RNAs inside the nucleus vs total cytosol RNAs within the ITP zone downstream.

BIBLIOGRAPHY

1. Niedringhaus, T.P., et al., *Landscape of Next-Generation Sequencing Technologies.* Analytical Chemistry, 2011. **83**(12): p. 4327-4341.

2. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.

3. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A, 1977. **74**(2): p. 560-4.

4. Sanger, F., et al., *NUCLEOTIDE-SEQUENCE OF BACTERIOPHAGE PHICHI174 DNA.* Nature, 1977. **265**(5596): p. 687-695.

5. Ansorge, W.J., *Next-generation DNA sequencing techniques.* New Biotechnology, 2009. **25**(4): p. 195-203.

6. Franca, L.T., E. Carrilho, and T.B. Kist, *A review of DNA sequencing techniques.* Q Rev Biophys, 2002. **35**(2): p. 169-200.

7. Slatko, B.E., et al., *DNA sequencing by the dideoxy method.* Curr Protoc Mol Biol, 2001. **Chapter 7**: p. Unit7 4A.

8. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

9. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-+.

10. Conti, R., et al., Personalized Medicine and Genomics: Challenges and Opportunities in Assessing Effectiveness, Cost-Effectiveness, and Future Research Priorities. Medical Decision Making. **30**(3): p. 328-340.

11. Ginsburg, G.S., M.P. Donahue, and L.K. Newby, *Prospects for personalized cardiovascular medicine - The impact of genomics.* Journal of the American College of Cardiology, 2005. **46**(9): p. 1615-1627.

12. Ginsburg, G.S. and H.F. Willard, *Genomic and personalized medicine: foundations and applications.* Translational Research, 2009. **154**(6): p. 277-287.

13. Philippidis, A. After a Decade, JGI Retires the Last of Its Sanger Sequencers. GenomeWeb Daily News, 2010.

14. Hert, D.G., C.P. Fredlake, and A.E. Barron, Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. Electrophoresis, 2008. **29**(23): p. 4618-4626.

15. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.

16. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

17. Fuller, C.W., et al., *The challenges of sequencing by synthesis.* Nat Biotechnol, 2009. **27**(11): p. 1013-23.

18. Levy, S., et al., *The diploid genome sequence of an individual human.* Plos Biology, 2007. **5**(10): p. 2113-2144.

19. Wheeler, D.A., et al., The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008. **452**: p. 872-877.

20. Bentley, D.R., et al., Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 2008. **456**(7218): p. 53-59.

21. Ley, T.J., et al., DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature, 2008. **456**(7218): p. 66-72.

22. Wang, J., et al., *The diploid genome sequence of an Asian individual.* Nature, 2008. **456**(7218): p. 60-U1.

23. Pushkarev, D., N.F. Neff, and S.R. Quake, *Single-molecule sequencing of an individual human genome.* Nature Biotechnology, 2009. **27**(9): p. 847-U101.

24. Herper, M., *A First: Diagnosis By DNA*, in *Forbes*2010, Forbes.

25. Reporter, P., $19,500 Per Person, Group Rates Available!, in The Daily Scan2010, GenomeWeb.

26. Karow, J. Life Tech Streamlines Sample Prep for Ion Torrent; Moves Away from Traditional Emulsion PCR. In Sequence, 2011.

27. Ronaghi, M., et al., *Real-time DNA sequencing using detection of pyrophosphate release.* Analytical Biochemistry, 1996. **242**(1): p. 84-89.

28. Karow, J. For 454 Technology, Roche Bets on Specific Apps, Taking Share from Sanger. In Sequence, 2010.

29. Wetterstrand, K.A. *DNA Sequencing Costs: Dat from the NHGRI Large-Scale Genome Sequencing Program*. [cited 2011 April 18, 2011]; Available from: http://www.genome.gov/sequencingcosts.

30. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

31. Levene, M.J., et al., Zero-mode waveguides for single-molecule analysis at high concentrations. Science, 2003. **299**(5607): p. 682-686.

32. Rigneault, H., et al., Enhancement of single-molecule fluorescence detection in subwavelength apertures. Physical Review Letters, 2005. **95**(11).

33. Popov, E., et al., *Field enhancement in single subwavelength apertures.* Journal of the Optical Society of America a-Optics Image Science and Vision, 2006. **23**(9): p. 2342-2348.

34. Samiee, K.T., et al., lambda-repressor oligomerization kinetics at high concentrations using fluorescence correlation spectroscopy in zero-mode waveguides. Biophysical Journal, 2005. **88**(3): p. 2145-2153.

35. Samiee, K.T., et al., *Zero mode waveguides for single-molecule spectroscopy on lipid membranes.* Biophysical Journal, 2006. **90**(9): p. 3288-3299.

36. Foquet, M., et al., Improved fabrication of zero-mode waveguides for single-molecule detection. Journal of Applied Physics, 2008. **103**(3).

37. Lundquist, P.M., et al., *Parallel confocal detection of single molecules in real time.* Optics Letters, 2008. **33**(9): p. 1026-1028.

38. Korlach, J., et al., Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. Nucleosides Nucleotides & Nucleic Acids, 2008. **27**(9): p. 1072-1083.

39. Korlach, J., et al., Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(4): p. 1176-1181.

40. Eid, J., et al., *Real-Time DNA Sequencing from Single Polymerase Molecules.* Science, 2009. **323**(5910): p. 133-138.

41. Travers, K.J., et al., A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Research. **38**(15).

42. Chin, C.S., et al., *The Origin of the Haitian Cholera Outbreak Strain.* New England Journal of Medicine. **364**(1): p. 33-42.

43. Flusberg, B.A., et al., Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods. **7**(6): p. 461-U72.

44. Uemura, S., et al., Real-time tRNA transit on single translating ribosomes at codon resolution. Nature. **464**(7291): p. 1012-U73.

45. Strezoska, Z., et al., *DNA SEQUENCING BY HYBRIDIZATION - 100 BASES READ BY A NON-GEL-BASED METHOD.* Proceedings of the National Academy of Sciences of the United States of America, 1991. **88**(22): p. 10089-10093.

46. Drmanac, R., et al., SEQUENCING BY HYBRIDIZATION - TOWARDS AN AUTOMATED SEQUENCING OF ONE MILLION M13 CLONES ARRAYED ON MEMBRANES. Electrophoresis, 1992. **13**(8): p. 566-573.

47. Drmanac, R., et al., DNA-SEQUENCE DETERMINATION BY HYBRIDIZATION - A STRATEGY FOR EFFICIENT LARGE-SCALE SEQUENCING. Science, 1993. **260**(5114): p. 1649-1653.

48. Drmanac, S., et al., Accurate sequencing by hybridization for DNA diagnostics and individual genomics. Nature Biotechnology, 1998. **16**(1): p. 54-58.

49. Schirinzi, A., et al., Combinatorial sequencing-by-hybridization: Analysis of the NF1 gene. Genetic Testing, 2006. **10**(1): p. 8-17.

50. Shendure, J., et al., Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 2005. **309**(5741): p. 1728-1732.

51. Drmanac, R., et al., Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science. **327**(5961): p. 78-81.

52. Reid, C., *Complete Genomics Inc.* Future Oncology. **7**(2): p. 219-221.

53. McKernan, K.J., et al., Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research, 2009. **19**(9): p. 1527-1541.

54. Margulies, M., et al., Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005. **437**(7057): p. 376-380.

55. Li, J.B., et al., Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. Genome Research, 2009. **19**(9): p. 1606-1615.

56. A second generation human haplotype map of over 3.1 million SNPs. Nature, 2007. **449**(7164): p. 851-861.

57. Roach, J.C., et al., Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science. **328**(5978): p. 636-639.

58. Lee, W., et al., The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. **465**(7297): p. 473-477.

59. Oh, J.D., et al., *The complete genome sequence of a chronic atrophic gastritis Helicobacter pylori strain: Evolution during disease progression.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(26): p. 9999-10004.

60. Nyren, P., B. Pettersson, and M. Uhlen, SOLID-PHASE DNA MINISEQUENCING BY AN ENZYMATIC LUMINOMETRIC INORGANIC PYROPHOSPHATE DETECTION ASSAY. Analytical Biochemistry, 1993. **208**(1): p. 171-175.

61. Nyren, P., ENZYMATIC METHOD FOR CONTINUOUS MONITORING OF DNA-POLYMERASE-ACTIVITY. Analytical Biochemistry, 1987. **167**(2): p. 235-238.

62. Hyman, E.D., *A NEW METHOD OF SEQUENCING DNA*. Analytical Biochemistry, 1988. **174**(2): p. 423-436.

63. Ronaghi, M., M. Uhlen, and P. Nyren, *A sequencing method based on real-time pyrophosphate.* Science, 1998. **281**(5375): p. 363-+.

64. Gharizadeh, B., et al., Long-read pyrosequencing using pure 2 '-deoxyadenosine-5 '-O '-(1-thiotriphosphate) Sp-isomer. Analytical Biochemistry, 2002. **301**(1): p. 82-90.

65. Esfandyarpour, H., et al., *Picocalorimetric method for DNA sequencing.* Journal of Vacuum Science & Technology B, 2008. **26**(2): p. 661-665.

66. Esfandyarpour, H., et al., Structural optimization for heat detection of DNA thermosequencing platform using finite element analysis. Biomicrofluidics, 2008. **2**(2).

67. Esfandyarpour, H. and M. Ronaghi, Obtaining sequence information from single stranded DNA template comprises providing primer region of DNA template, placing multiple copies of template DNA, adding DNA polymerization mixture, and measuring temperature change, Univ Leland Stanford Junior (Strd).

68. *LIFE TECHNOLOGIES TO BUY ION TORRENT.* Chemical & Engineering News, 2010. **88**(34): p. 18-18.

69. Karow, J. *Ion Torrent Unveils New $50k Electronic Sequencer*. 2010; Available from: http://www.genomeweb.com/sequencing/ion-torrent-unveils-new-50k-electronic-sequencer.

70. Branton, D., et al., *The potential and challenges of nanopore sequencing.* Nature Biotechnology, 2008. **26**(10): p. 1146-1153.

71. Bayley, H., *Sequencing single molecules of DNA.* Current Opinion in Chemical Biology, 2006. **10**(6): p. 628-637.

72. Kasianowicz, J.J., et al., *Characterization of individual polynucleotide molecules using a membrane channel.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(24): p. 13770-13773.

73. Deamer, D.W. and D. Branton, *Characterization of nucleic acids by nanopore analysis.* Accounts of Chemical Research, 2002. **35**(10): p. 817-825.

74. Tabard-Cossa, V., et al., Single-Molecule Bonds Characterized by Solid-State Nanopore Force Spectroscopy. Acs Nano, 2009. **3**(10): p. 3009-3014.

75. Tropini, C. and A. Marziali, *Multi-nanopore force Spectroscopy for DNA analysis.* Biophysical Journal, 2007. **92**(5): p. 1632-1637.

76. Wiggin, M., et al., *Nonexponential Kinetics of DNA Escape from alpha-Hemolysin Nanopores.* Biophysical Journal, 2008. **95**(11): p. 5317-5323.

77. Zwolak, M. and M. Di Ventra, *Electronic signature of DNA nucleotides via transverse transport.* Nano Letters, 2005. **5**(3): p. 421-424.

78. Gracheva, M.E., A. Aksimentiev, and J.P. Leburton, *Electrical signatures of single-stranded DNA with single base mutations in a nanopore capacitor.* Nanotechnology, 2006. **17**(13): p. 3160-3165.

79. Soni, G.V. and A. Meller, *Progress toward ultrafast DNA Sequencing using solid-state nanopores.* Clinical Chemistry, 2007. **53**(11): p. 1996-2001.

80. Sauer-Budge, A.F., et al., *Unzipping kinetics of double-stranded DNA in a nanopore.* Physical Review Letters, 2003. **90**(23): p. 238101/1-238101/4.

81. McNally, B., et al., Optical Recognition of Converted DNA Nucleotides for Single-Molecule DNA Sequencing Using Nanopore Arrays. Nano Letters. **10**(6): p. 2237-2244.

82. Kasianowicz, J.J., et al., *Nanoscopic Porous Sensors.* Annual Review of Analytical Chemistry, 2008. **1**: p. 737-766.

83. Zwolak, M. and M. Di Ventra, *Colloquium: Physical approaches to DNA sequencing and detection.* Reviews of Modern Physics, 2008. **80**(1): p. 141-165.

84. Meller, A. and D. Branton, *Single molecule measurements of DNA transport through a nanopore.* Electrophoresis, 2002. **23**(16): p. 2583-2591.

85. Li, J.L., et al., DNA molecules and configurations in a solid-state nanopore microscope. Nature Materials, 2003. **2**(9): p. 611-615.

86. Fologea, D., et al., *Detecting single stranded DNA with a solid state nanopore.* Nano Letters, 2005. **5**(10): p. 1905-1909.

87. *Towards the 15-minute genome*, in *The Economist*2011, The Economist Newspaper Limited: London: from the print edition | Technology Quarterly.

88. Howorka, S., S. Cheley, and H. Bayley, *Sequence-specific detection of individual DNA strands using engineered nanopores.* Nature Biotechnology, 2001. **19**(7): p. 636-639.

89. Astier, Y., O. Braha, and H. Bayley, Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5 '-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. Journal of the American Chemical Society, 2006. **128**(5): p. 1705-1710.

90. Clarke, J., et al., Continuous base identification for single-molecule nanopore DNA sequencing. Nature Nanotechnology, 2009. **4**(4): p. 265-270.

91. Wu, H.C., et al., *Protein nanopores with covalently attached molecular adapters.* Journal of the American Chemical Society, 2007. **129**(51): p. 16142-16148.

92. Lieberman, K.R., et al., *Processive Replication of Single DNA Molecules in a Nanopore Catalyzed by phi29 DNA Polymerase.* Journal of the American Chemical Society. **132**(50): p. 17961-17972.

93. Stoddart, D., et al., Nucleobase Recognition in ssDNA at the Central Constriction of the alpha-Hemolysin Pore. Nano Letters. **10**(9): p. 3633-3637.

94. Kang, X.F., et al., *Single protein pores containing molecular adapters at high temperatures.* Angewandte Chemie-International Edition, 2005. **44**(10): p. 1495-1499.

95. Merchant, C.A., et al., *DNA Translocation through Graphene Nanopores.* Nano Letters. **10**(8): p. 2915-2921.

96. Garaj, S., et al., *Graphene as a subnanometre trans-electrode membrane.* Nature. **467**(7312): p. 190-U73.

97. Luan, B.Q., et al., Base-By-Base Ratcheting of Single Stranded DNA through a Solid-State Nanopore. Physical Review Letters. **104**(23): p. 4.

98. Polonsky, S., S. Rossnagel, and G. Stolovitzky, *Nanopore in metal-dielectric sandwich for DNA position control.* Applied Physics Letters, 2007. **91**(15): p. 3.

99. Balagurusamy, V.S.K., P. Weinger, and X.S. Ling, *Detection of DNA hybridizations using solid-state nanopores.* Nanotechnology. **21**(33): p. 9.

100. Singer, A., et al., Nanopore Based Sequence Specific Detection of Duplex DNA for Genomic Profiling. Nano Letters. **10**(2): p. 738-742.

101. Drmanac, R., et al., SEQUENCING OF MEGABASE PLUS DNA BY HYBRIDIZATION - THEORY OF THE METHOD. Genomics, 1989. **4**(2): p. 114-128.

102. Ling, X.S.B., B. Pertsinidis, A., Hybridization-assisted nanopore sequencing of nucleic acids, 2007.

103. Cahill, M.J., et al., Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. PLoS One, 2010. **5**(7): p. e11518.

104. Lin, J.Y., et al., *Whole-genome shotgun optical mapping of Deinococcus radiodurans.* Science, 1999. **285**(5433): p. 1558-1562.

105. Valouev, A., et al., *Alignment of optical maps.* Journal of Computational Biology, 2006. **13**(2): p. 442-462.

106. Valouev, A., et al., *An algorithm for assembly of ordered restriction maps from single DNA molecules.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(43): p. 15770-15775.

107. Zhou, S., et al., Single-molecule approach to bacterial genomic comparisons via optical mapping. Journal of Bacteriology, 2004. **186**(22): p. 7773-7782.

108. Armbrust, E.V., et al., The genome of the diatom Thalassiosira pseudonana: Ecology, evolution, and metabolism. Science, 2004. **306**(5693): p. 79-86.

109. Haas, B.J., et al., Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature, 2009. **461**(7262): p. 393-398.

110. Zhou, S., et al., *Validation of rice genome sequence by optical mapping.* BMC Genomics, 2007. **8**: p. 278.

111. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.

112. Wei, F., et al., *The physical and genetic framework of the maize B73 genome.* PLoS Genet, 2009. **5**(11): p. e1000715.

113. Wei, F., et al., Detailed analysis of a contiguous 22-Mb region of the maize genome. PLoS Genet, 2009. **5**(11): p. e1000728.

114. Zhou, S., et al., *A single molecule scaffold for the maize genome.* PLoS Genet, 2009. **5**(11): p. e1000711.

115. Kidd, J.M., et al., Mapping and sequencing of structural variation from eight human genomes. Nature, 2008. **453**(7191): p. 56-64.

116. Teague, B., et al., *High-resolution human genome structure by single-molecule analysis.* Proc Natl Acad Sci U S A, 2010. **107**(24): p. 10848-53.

117. Jo, K., T.M. Schramm, and D.C. Schwartz, *A single-molecule barcoding system using nanoslits for DNA analysis : nanocoding.* Methods Mol Biol, 2009. **544**: p. 29-42.

118. Reisner, W., et al., *Single-molecule denaturation mapping of DNA in nanofluidic channels.* Proc Natl Acad Sci U S A, 2010. **107**(30): p. 13294-9.

119. Das, S.K., et al., Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. Nucleic Acids Res, 2010. **38**(18): p. e177.

120. Nagayama, Method of determining base sequence of DNA or RNA using heavy element labeling and imaging by transmission electron microscopy, USPTO, Editor 2008: USA.

121. Tanaka, H. and T. Kawai, Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. Nat Nanotechnol, 2009. **4**(8): p. 518-22.

122. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.

123. Milanova, D., et al., Effect of PVP on the electroosmotic mobility of wet-etched glass microchannels. Electrophoresis, 2012. **33**(21): p. 3259-3262.

124. Robinson, S. and P.A. Williams, *Inhibition of protein adsorption onto silica by polyvinylpyrrolidone.* Langmuir, 2002. **18**(23): p. 8743-8748.

125. Kim, C.S., et al., A simple and rapid method for isolation of high quality genomic DNA from fruit trees and conifers using PVP. Nucleic Acids Research, 1997. **25**(5): p. 1085-1086.

126. Schutzner, W., et al., IMPROVED SEPARATION OF DIASTEREOMERIC DERIVATIVES OF ENANTIOMERS BY A PHYSICAL NETWORK OF LINEAR POLYVINYLPYRROLIDONE APPLIED AS PSEUDOPHASE IN CAPILLARY ZONE ELECTROPHORESIS. Electrophoresis, 1994. **15**(6): p. 769-773.

127. Munro, N.J., et al., Molecular diagnostics on microfabricated electrophoretic devices: From slab gel- to capillary- to microchip-based assays for T- and B-cell lymphoproliferative disorders. Clinical Chemistry, 1999. **45**(11): p. 1906-1917.

128. Gao, Q.F. and E.S. Yeung, A matrix for DNA separation: Genotyping and sequencing using poly(vinylpyrrolidone) solution in uncoated capillaries. Analytical Chemistry, 1998. **70**(7): p. 1382-1388.

129. Kaneta, T., et al., Suppression of electroosmotic flow and its application to determination of electrophoretic mobilities in a poly(vinylpyrrolidone)-coated capillary. Journal of Chromatography A, 2006. **1106**(1-2): p. 52-55.

130. Milanova, D., et al., Electrophoretic mobility measurements of fluorescent dyes using on-chip capillary electrophoresis. Electrophoresis, 2011. **32**(22): p. 3286-3294.

131. Chambers, R.D. and J.G. Santiago, Imaging and Quantification of Isotachophoresis Zones Using Nonfocusing Fluorescent Tracers. Analytical Chemistry, 2009. **81**(8): p. 3022-3028.

132. Madajova, V., E. Turcelova, and D. Kaniansky, INFLUENCE OF POLY(VINYLPYRROLIDONE) ON ISOTACHOPHORETIC SEPARATIONS OF INORGANIC ANIONS IN AQUEOUS-ELECTROLYTE SYSTEMS. Journal of Chromatography, 1992. **589**(1-2): p. 329-332.

133. Munro, N.J., A.F.R. Huhmer, and J.P. Landers, *Robust polymeric microchannel coatings for microchip-based analysis of neat PCR products.* Analytical Chemistry, 2001. **73**(8): p. 1784-1794.

134. Srinivasan, K., G. Pohl, and N. Avdalovic, Cross-linked polymer coatings for capillary electrophoresis and application to analysis of basic proteins, acidic proteins, and inorganic ions. Analytical Chemistry, 1997. **69**(14): p. 2798-2805.

135. Kim, J.A., et al., *Fabrication and characterization of a PDMS-glass hybrid continuous-flow PCR chip.* Biochemical Engineering Journal, 2006. **29**(1-2): p. 91-97.

136. Lou, X.J., et al., Increased amplification efficiency of microchip-based PCR by dynamic surface passivation. Biotechniques, 2004. **36**(2): p. 248-252.

137. Swei, J. and J.B. Talbot, *Viscosity correlation for aqueous polyvinylpyrrolidone (PVP) solutions.* Journal of Applied Polymer Science, 2003. **90**(4): p. 1153-1155.

138. Schrum, K.F., et al., Monitoring electroosmotic flow by periodic photobleaching of a dilute, neutral fluorophore. Analytical Chemistry, 2000. **72**(18): p. 4317-4321.

139. Jaros, M., et al., Optimization of background electrolytes for capillary electrophoresis: II. Computer simulation and comparison with experiments. Electrophoresis, 2002. **23**(16): p. 2667-2677.

140. Bahga, S., M. Bercovici, and J.G. Santiago, *Ionic strength effects on electrophoretic focusing and separations.* Electrophoresis, 2010. **31**: p. 910-919.

141. McHedlovpetrossyan, N.O., V.I. Kukhtik, and V.I. Alekseeva, IONIZATION AND TAUTOMERISM OF FLUORESCEIN, RHODAMINE-B, N,N-DIETHYLRHODOL AND RELATED DYES IN MIXED AND NONAQUEOUS SOLVENTS. Dyes and Pigments, 1994. **24**(1): p. 11-35.

142. Arbeloa, I.L. and P. Ruizojeda, *MOLECULAR-FORMS OF RHODAMINE-B.* Chemical Physics Letters, 1981. **79**(2): p. 347-350.

143. Mchedlov-Petrossyan, N.O., V.I. Kukhtik, and V.I. Alekseeva, IONIZATION AND TAUTOMERISM OF FLUORESCEIN, RHODAMINE-B, N,N-DIETHYLRHODOL AND RELATED DYES IN MIXED AND NONAQUEOUS SOLVENTS. Dyes and Pigments, 1994. **24**(1): p. 11-35.

144. Yu, S.B., et al., *pH effect on dynamic coating for capillary electrophoresis of DNA.* Analytical and Bioanalytical Chemistry, 2006. **385**(4): p. 730-736.

145. Wang, A.-J., et al., Noncovalent poly(1-vinylpyrrolidone)-based copolymer coating for the separation of basic proteins and lipoproteins by CE. Electrophoresis, 2009. **30**(22): p. 3939-3946.

146. Gryczynski, Z., I. Gryczynski, and J.R. Lakowicz, *Fluorescence-sensing methods*, in *Biophotonics, Pt A*. 2003, Academic Press Inc: San Diego. p. 44-75.

147. Kricka, L.J. and P. Fortina, Analytical Ancestry: "Firsts" in Fluorescent Labeling of Nucleosides, Nucleotides, and Nucleic Acids. Clinical Chemistry, 2009. **55**(4): p. 670-683.

148. Sameiro, M. and T. Goncalves, *Fluorescent Labeling of Biomolecules with Organic Probes.* Chemical Reviews, 2009. **109**(1): p. 190-212.

149. Opitz, N. and D.W. Lubbers, New fluorescence photometrical techniques for simultaneous and continuous measurements of ionic strength and hydrogen ion activities. Sensors and Actuators, 1983. **4**(3): p. 473-479.

150. Shimura, K., *Recent advances in IEF in capillary tubes and microchips.* Electrophoresis, 2009. **30**(1): p. 11-28.

151. Slais, K., et al. Fluorescein-based pl markers for capillary isoelectric focusing with laser-induced fluorescence detection. 2002. Wiley-V C H Verlag Gmbh.

152. Lacroix, M., et al., Laser-induced fluorescence detection schemes for the analysis of proteins and peptides using capillary electrophoresis. Electrophoresis, 2005. **26**(13): p. 2608-2621.

153. Moser, A.C. and D.S. Hage, Capillary electrophoresis-based immunoassays: Principles and quantitative applications. Electrophoresis, 2008. **29**(16): p. 3279-3295.

154. Bercovici, M., et al., Fluorescent Carrier Ampholytes Assay for Portable, Label-Free Detection of Chemical Toxins in Tap Water. Analytical Chemistry. **82**(5): p. 1858-1866.

155. Probstein, R.F., *Physicochemical Hydrodynamics: An Introduction*. 1994, New York: John Wiley & Sons, Inc.

156. See for a description of effective versus fully-ionized versus absolute mobility and their relations to ionic strength and pH.

157. Barron, D., E. Jimenez-Lozano, and J. Barbosa, *Prediction of electrophoretic behaviour of a series of quinolones in aqueous methanol.* Journal of Chromatography A, 2001. **919**(2): p. 395-406.

158. Beckers, J.L., F.M. Everaerts, and M.T. Ackermans, DETERMINATION OF ABSOLUTE MOBILITIES, PK VALUES AND SEPARATION NUMBERS BY CAPILLARY ZONE ELECTROPHORESIS - EFFECTIVE MOBILITY AS A PARAMETER FOR SCREENING. Journal of Chromatography, 1991. **537**(1-2): p. 407-428.

159. Canals, I., E. Bosch, and M. Roses, *Prediction of the separation of phenols by capillary zone electrophoresis.* Analytica Chimica Acta, 2002. **458**(2): p. 355-366.

160. Porras, S.P., M.L. Riekkola, and E. Kenndler, Capillary zone electrophoresis of basic analytes in methanol as non-aqueous solvent - Mobility and ionisation constant. Journal of Chromatography A, 2001. **905**(1-2): p. 259-268.

161. Vcelakova, K., et al., Determination of cationic mobilities and pK(a) values of 22 amino acids by capillary zone electrophoresis. Electrophoresis, 2004. **25**(2): p. 309-317.

162. Zuskova, I., et al., Determination of limiting mobilities and dissociation constants of 21 amino acids by capillary zone electrophoresis at very low pH. Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 2006. **841**(1-2): p. 129-134.

163. Williams, B.A. and C. Vigh, *Fast, accurate mobility determination method for capillary electrophoresis.* Analytical Chemistry, 1996. **68**(7): p. 1174-1180.

164. Hirokawa, T., T. Gojo, and Y. Kiso, ISOTACHOPHORETIC DETERMINATION OF MOBILITY AND PKA BY MEANS OF COMPUTER-SIMULATION .4. EVALUATION OF MO AND PKA OF 26 AMINO-ACIDS AND ASSESSMENT OF THE SEPARABILITY. Journal of Chromatography, 1986. **369**(1): p. 59-81.

165. Hirokawa, T., M. Nishino, and Y. Kiso, ISOTACHOPHORETIC DETERMINATION OF MOBILITY AND PKA BY MEANS OF COMPUTER-SIMULATION .2. EVALUATION OF MO AND PKA OF 65 ANIONS. Journal of Chromatography, 1982. **252**(DEC): p. 49-65.

166. Pospichal, J., M. Deml, and P. Bocek, DETERMINATION OF IONIC MOBILITIES AND DISSOCIATION-CONSTANTS OF MONO-VALENT ACIDS AND BASES BY MICROPREPARATIVE CAPILLARY ISOTACHOPHORESIS WITH OFF-LINE MEASUREMENT OF THE PH OF ZONES. Journal of Chromatography, 1987. **390**(1): p. 17-26.

167. Pospichal, J., et al., *DETERMINATION OF RELATIVE IONIC MOBILITIES BY CAPILLARY ISOTACHOPHORESIS.* Journal of Chromatography, 1985. **320**(1): p. 139-146.

168. Andersson, E.K.M., Impurities at a level of 0.01% in foscarnet sodium determined by capillary zone electrophoresis with indirect UV detection and sample self-stacking. Journal of Chromatography A, 1999. **846**(1-2): p. 245-253.

169. Urbanek, M., et al., Determination of trace cationic impurities in butylmethylimidazolium-based ionic liquids: From transient to comprehensive single-capillary counterflow isotachophoresis-zone electrophoresis. Electrophoresis, 2006. **27**(23): p. 4859-4871.

170. Garcia-Schwarz, G., et al., J. Fluid Mechanics, accepted.

171. Schonfeld, F., et al., Transition zone dynamics in combined isotachophoretic and electro-osmotic transport. Physics of Fluids, 2009. **21**(9).

172. Paul, P.H., M.G. Garguilo, and D.J. Rakestraw, *Imaging of pressure- and electrokinetically driven flows through open capillaries.* Analytical Chemistry, 1998. **70**(13): p. 2459-2467.

173. Shakalisava, Y., et al., *Versatile method for electroosmotic flow measurements in microchip electrophoresis.* Journal of Chromatography A, 2009. **1216**(6): p. 1030-1033.

174. Whang, C.W. and E.S. Yeung, *TEMPERATURE PROGRAMMING IN CAPILLARY ZONE ELECTROPHORESIS.* Analytical Chemistry, 1992. **64**(5): p. 502-506.

175. Kaniansky, D., M. Masar, and J. Bielcikova, Electroosmotic flow suppressing additives for capillary zone electrophoresis in a hydrodynamically closed separation system. Journal of Chromatography A, 1997. **792**(1-2): p. 483-494.

176. Schutzner, W., et al., Separation of diastereomers by capillary zone electrophoresis in free solution with polymer additive and organic solvent component Effect of pH and solvent composition. Journal of Chromatography A, 1996. **719**(2): p. 411-420.

177. Porras, S.P., M.L. Riekkola, and E. Kenndler, The principles of migration and dispersion in capillary zone electrophoresis in nonaqueous solvents. Electrophoresis, 2003. **24**(10): p. 1485-1498.

178. Persat, A., M.E. Suss, and J.G. Santiago, Basic principles of electrolyte chemistry for microfluidic electrokinetics. Part II: Coupling between ion mobility, electrolysis, and acid-base equilibria. Lab on a chip, 2009. **9**: p. 2454-2469.

179. Onsager, L. and R.M. Fuoss, Irreversible processes in electrolytes diffusion, conductance, and viscous flow in arbitrary mixtures of strong electrolytes. Journal of Physical Chemistry, 1932. **36**(7): p. 2689-2778.

180. Pitts, E., B.E. Tabor, and J. Daly, CONCENTRATION DEPENDENCE OF ELECTROLYTE CONDUCTANCE .2. COMPARISON OF EXPERIMENTAL DATA WITH FUOSS-ONSAGER AND PITTS TREATMENTS. Transactions of the Faraday Society, 1970. **66**(567): p. 693-&.

181. Bharadwaj, R., J.G. Santiago, and B. Mohammadi, *Design and optimization of on-chip capillary electrophoresis.* Electrophoresis, 2002. **23**(16): p. 2729-2744.

182. Bercovici, M., S.K. Lele, and J.G. Santiago, *Open source simulation tool for electrophoretic stacking, focusing, and separation.* Journal of Chromatography A, 2009. **1216**(6): p. 1008-1018.

183. Martin, M.M. and L. Lindqvist, *PH-DEPENDENCE OF FLUORESCEIN FLUORESCENCE.* Journal of Luminescence, 1975. **10**(6): p. 381-390.

184. Diehl, H. and R. Markuszewski, STUDIES ON FLUORESCEIN .7. THE FLUORESCENCE OF FLUORESCEIN AS A FUNCTION OF PH. Talanta, 1989. **36**(3): p. 416-418.

185. Panchuk-Voloshina, N., et al., *Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates.* Journal of Histochemistry & Cytochemistry, 1999. **47**(9): p. 1179-1188.

186. Khurana, T.K. and J.G. Santiago, Effects of carbon dioxide on peak mode isotachophoresis: Simultaneous preconcentration and separation. Lab on a chip, 2009. **9**(10): p. 1377-1384.

187. Duvvuri, M., et al., Weak base permeability characteristics influence the intracellular sequestration site in the multidrug-resistant human leukemic cell line HL-60. Journal of Biological Chemistry, 2004. **279**(31): p. 32367-32372.

188. Kuznetsov, R.T. and R.M. Fofonova, *Photostability of rhodamine 6G with respect to generating and spectral characteristics.* Journal of Applied Spectroscopy, 1984. **40**(4): p. 380-384384.

189. Magde, D., E.L. Elson, and W.W. Webb, *FLUORESCENCE CORRELATION SPECTROSCOPY .2. EXPERIMENTAL REALIZATION.* Biopolymers, 1974. **13**(1): p. 29-61.

190. Petrasek, Z. and P. Schwille, Precise measurement of diffusion coefficients using scanning fluorescence correlation spectroscopy. Biophysical Journal, 2008. **94**(4): p. 1437-1448.

191. Muller, C.B., et al., Precise measurement of diffusion by multi-color dual-focus fluorescence correlation spectroscopy. Epl, 2008. **83**(4): p. 5.

192. Touloukian, Y.S., Saxena, S.C., Hestermans, P., *Thermophysical Properties of Matter, the TPRC Data Series.* . Vol. Vol. 11 - Viscosity. 1975, New York: Plenum Publishing Corp.

193. Ilich, P., et al., *Direct observation of rhodamine dimer structures in water.* Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy, 1996. **52**(10): p. 1323-1330.

194. McHedlov-Petrosyan, N.O. and Y.V. Kholin, *Aggregation of rhodamine B in water.* Russian Journal of Applied Chemistry, 2004. **77**(3): p. 414-422.

195. Shariff, K. and S. Ghosal, Peak tailing in electrophoresis due to alteration of the wall charge by adsorbed analytes a Numerical simulations and asymptotic theory. Analytica Chimica Acta, 2004. **507**(1): p. 87-93.

196. Hamai, S. and K. Sasaki, Capillary electrophoretic investigations on the interactions between rhodamine 6G and poly(vinyl sulfate). Microchemical Journal, 2001. **69**(1): p. 27-36.

197. Tilgner, H., et al., Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Research, 2012. **22**(9): p. 1616-1625.

198. Dunham, I., et al., An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. **489**(7414): p. 57-74.

199. Kornblihtt, A.R., et al., *Alternative splicing: a pivotal step between eukaryotic transcription and translation.* Nature Reviews Molecular Cell Biology, 2013. **14**(3): p. 153-165.

200. Chen, J. and W.A. Weiss, Alternative splicing in cancer: implications for biology and therapy. Oncogene, 2015. **34**(1): p. 1-14.

201. Yoshida, K. and S. Ogawa, *Splicing factor mutations and cancer.* Wiley Interdisciplinary Reviews-Rna, 2014. **5**(4): p. 445-459.

202. Dredge, B.K., A.D. Polydorides, and R.B. Darnell, *The splice of life: Alternative splicing and neurological disease.* Nature Reviews Neuroscience, 2001. **2**(1): p. 43-50.

203. Harries, L.W., et al., Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. Aging Cell, 2011. **10**(5): p. 868-878.

204. Bengtsson, M., et al., Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Research, 2005. **15**(10): p. 1388-1392.

205. White, A.K., et al., *High-throughput microfluidic single-cell RT-qPCR.* Proc Natl Acad Sci U S A, 2011. **108**(34): p. 13999-4004.

206. Tilgner, H., et al., *Defining a personal, allele-specific, and single-molecule long-read transcriptome.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(27): p. 9869-9874.

207. Tilgner, H., et al., Accurate Identification and Analysis of Human mRNA Isoforms Using Deep Long Read Sequencing. G3-Genes Genomes Genetics, 2013. **3**(3): p. 387-397.

5 APPENDICES

APPENDIX A - CELL CULTURE

Harvesting cells

1) Heat PBS and cell media in 37˚C water bath for ~ 15 minutes.
2) Pipette 1.5mL of cells + media from culture into 5mL Falcon tube.
3) Centrifuge at 1000 rpm (~200g) for 3 minutes. Be sure to place a counter balance in centrifuge.
4) Aspirate all the media by moving glass 9" pipette along the side. A small volume of media left is fine.
5) Place 1mL of heated PBS above the cells (use new pipette).
6) Pipette up and down gentry to minimize the introduction of bubbles.
7) Centrifuge at 1000 rpm (~200g) for 3 minutes. Be sure to place a counter balance in centrifuge.
8) Aspirate all the PBS by moving glass 9" pipette along the side. A small volume of PBS left is fine.
9) Place 50µL of heated media above the cells (use new pipette).
10) Pipette gently up and down to minimize the introduction of bubbles.
11) Take 10µL and place in the InCyto C-Chip cell counter.
12) Target concentration 5 cells/µL, but can be slightly more. No more than 10-15 cells in the chip well.
13) Within 16 squares, count the cells. We aim for ~40-100 cells.
14) If there is density of cells is high, add more media, mix with pipette. Count again on a new side of chip. (We had to add 400 µL, then added additional 20-30 µL).

15) Place final cells including media mixture in 1mL vial.
16) Place 1mL vial in the incubator.
17) Once ready to begin experiments, take cells to experiment room and place in the incubator at 37˚C, protect with aluminum foil from light, place vial on vial holder (upright position).

APPENDIX B - MICROFLUIDIC FRACTIONATION PROTOCOL

Clean PDMS chips

1) Heat the readymade PDMS chips for 30 minutes at 302˚F (150˚C). In the mean time, harvest the cells.
2) After 30 minutes, place the heated chips along the edge of petri dish to cool down and not melt the petri dish.

Single Cell experiment

1) Clean entire working area with RNAse/DNAse, including pipettes.
2) Turn on the light source (no fluorescence needed)
    a) 20X objective
    b) Ph1, phase contrast objective, make sure the filter on the top is also Ph1.
3) Turn on the voltage source and vacuum, main switch on the power strip
4) NIS-ELEMENTS, camera software, click LIVE.
5) Align the chip so that when moving stage, West (W) to East (E) remains level.
6) Secure the chip in place with tape.
7) Chip Wash with NaOH 1M, HCl 1M, DI containing 0.1% Triton X-100
    a) Fill ALL wells with 40 µL of NaOH + Triton X-100
    b) Wash electrodes by dipping leads in the wells; use this time to get a feel for electrode placement.
    c) Remove electrodes, place to the side of chip with leads facing up.
    d) Aspirate the S and South East (SE) wells until empty

e) Vacuum from E and South East (SE) wells for 1 minute

f) Repeat using HCl and DI.

8) Prepare the sample solution with cells

a) 99 µL Sample Buffer (SS) + 1 µL Cells + Media

i) Can add 1µL more of cells if necessary.

ii) Note: Fresh solution should be made right before each experiment after which solution is discard as Biohazard Waste

b) Pipette up and down 4-5 times to mix. Slowly to make sure the cells are not lysed

c) Note: Sample Buffer has low conductivity and same osmolarity as cell environment.

9) 20 µL of LE 1 in N, W, S vacuum from E and SE for 1 min

10) Pipette the following in each well:

i) E: 10 µL of LE 2 (make sure no bubbles are created)

ii) SE: 10 µL of LE 2

iii) S: 10 µL of TE 1

iv) N: 15 µL of TE 1

v) W: 2 µL of SS + cells (place the cells in the well slowly and close to the channel inlet)

11) Vacuum Cell into channel

a) Vacuum control knob pointing to the N (Figure 2); aspirator vacuum ON

b) Place the two vacuum leads,

i) One lead in the S well

ii) Second lead on your right hand, thumb covering entrance for a finer vacuum control

c) Place the vacuum control knob pointing to the SE

d) Adjust the vacuum control knob and covering/uncovering second vacuum lead on your right hand; keep adjusting until one cells is in the vertical channel and relatively stays in place.

12) Turn vacuum off, by turning vacuum control knob to the S

13) Quickly, pipette 11 µL of TE 2 into W chip channel

a) Note, this may not be exactly 11 µL but as needed to keep the cell in place. Can add volume or take out volume until the cells is somewhat stationary or moving very slowly.

14) Place electrodes in the chip, secure in place but do not press to hard down since you can lose the cell in the channel. The S well on the chip does not have an electrode; all other wells have an electrode.

15) Run the Matlab program *kineticmeasure012015.m* ;

a) *Rename the text file to store current data.

16) Sequencer: *4electrodes.seq*; press A button

a) Make sure that yellow triangle is selected: Enables high voltage output

17) Follow the nucleus; press E button on the sequencer GUI once SE "T" cross section is visible on the microscope (Note: some cells move faster than others, this is due cell heterogeneity).

18) Keep track of the nucleus, observe the current vs time Matlab figure

a) Confirm that the current drops once the voltage is on; make sure to keep track of the nucleus at ALL times.

b) Keep the mouse on the E sequence ready to press when the nucleus approaches the T-junction and the SE channel.

19) Once current reaches steady state ~300 seconds.
    a) After successful fractionation, start looking at the current trace.
    b) Type 'q' on command line in Matlab to stop recording current data
    c) Turn off the voltage source
20) Remove the electrodes
21) Bend the tip of 2 μL pipette tip (change gloves)
22) Pipette 1 μL to obtain the nucleus in the SE chip well. Using the microscope, place tip covering the area where nucleus is.
    a) For sequencing: Pipette and additional 9 μL solution out of the SE chip well.
    b) For gene expression: Pipette only the nucleus, 1 μL
23) Place over dry ice.
24) Cytoplasm, pipette directly out of the E chip well
    a) For sequencing: Pipette 10 μL of cytoplasm solution out of the E chip well. No visual needed.
    b) For gene expression: Pipette 5 μL into each tubes, for a total volume of 10 μL; 2 tubes.
25) Place over dry ice.

APPENDIX C - VOLTAGE CONTROL

kineticmeasure012015.m

```
% trigger hardware and save ttl history
% using NI-DAQ USB-6009
%
clear all;
close all;
pause(0.1);
clc;
fclose ('all');

outputbias=5; % ttl voltage
plogtime=linspace(2,4,101);%;plot time(1)=0;

ptime=10.^plogtime;  % ttl timing is saved in file
ptime=cat(2,linspace(1,91,10),ptime);
dt=0.49;% ttl width
% file to save
FolderName='..\20141201\';
FileName=cat(2,FolderName,'20141201_exp1.txt');
%
ai=daq.createSession('ni');         % Set DAQ for Output
ao=daq.createSession('ni');         % Set DAQ for Input
ai.addAnalogInputChannel('Dev1',2,'Voltage'); % Set Input Channel
ai.Channels(1).Range=[-10 10]; % Set Range of Input Channel
ai.Channels(1).TerminalConfig='SingleEnded';
ao.addAnalogOutputChannel('Dev1',0,'Voltage');% Set Output Channel
%%
% Open output file
   WriteToFileFlag=1;
if WriteToFileFlag
   fid=fopen(FileName,'w');
```

```
    fprintf(fid,'%10s\t\t%10s\n','%Time [s]','TTL [V]');
end
icnt=0;
Tcnt=1;ttl=0;ptimelen=length(ptime);

tic                     % Acquire process start time
while Tcnt<=ptimelen
    icnt=icnt+1;
    data=ai.inputSingleScan();
    ttlsig(icnt)=data(:,1);
    time(icnt)=toc;        % get elapsed time from start

fprintf(fid,'%10.5f\t\t%10.5f\n',[time(icnt),ttlsig(icnt)]');
    timeshow;
    if ttl==0 & ptime(Tcnt)<=toc
        ttl=1;Tcnt=Tcnt+1;
        ao.outputSingleScan(outputbias);
    end
    if ttl & ptime(Tcnt-1)+dt <=toc
        ttl=0;
        ao.outputSingleScan(0);
    end
    drawnow
    keyIn = get(gcf, 'CurrentCharacter');
    if strcmp(keyIn,'q') || strcmp(keyIn,'b')
        break;
    end;
end

WriteToFileFlag=1;
if WriteToFileFlag
    fclose(fid);
end
```

APPENDIX D - STAR MAPPING AND CUFFLINKS

```
#!/bin/sh
# set the name of the job
#$ -N d1
# set the maximum memory usage (per slot)
#$ -l h_vmem=4.1G
# set the maximum run time
#$ -l h_rt=168:00:00
# send mail when job ends or aborts
#$ -m ea
# specify an email address
#$ -M milanova@stanford.edu
# check for errors in the job submission options
#$ -w e
# number of threads
#$ -pe shm 8
# running in the current working directory
#$ -cwd
# the directory where we want error and output messages
#$ -e /put/your/own/path/here///messages.error.d1
#$ -o /put/your/own/path/here///messages.out.d1
#$ -R y
### now the actual commands
## sourcing the .bashrc
source /home/htilgner/.bashrc
#
## making temporary directory
myTMPDIR=/put/your/own/path/here///HT.$$_tmp/
mkdir $myTMPDIR
echo "myTMPDIR="$myTMPDIR >> /put/your/own/path/here///REPORT.d1;
#
hostname >> /put/your/own/path/here///REPORT.d1;
```

```
pwd >> /put/your/own/path/here///REPORT.d1;
date >> /put/your/own/path/here///REPORT.d1;
cd /put/your/own/path/here//
n=0;
tot=`cat K562_RNA-seq_lib_ID.dataGuidePolished.txt_unix | wc -l`;
for i in `seq 1 $tot`; do
        ((n++));
        date >> /put/your/own/path/here///REPORT.d1;
        subdirname=`cat K562_RNA-seq_lib_ID.dataGuidePolished.txt_unix-2.1.corrected | awk -v i=$n '{if(NR==i){print $1;}}' `;
        echo "#### treating "$subdirname >> /put/your/own/path/here///REPORT.d1;
        mkdir $subdirname
        cd $subdirname
        file1=`cat ../K562_RNA-seq_lib_ID.dataGuidePolished.txt_unix-2.1.corrected | awk -v i=$n '{if(NR==i){print $2;}}' `;
        file2=`cat ../K562_RNA-seq_lib_ID.dataGuidePolished.txt_unix-2.1.corrected | awk -v i=$n '{if(NR==i){print $3;}}' `;
        echo -e  "### "$n"  "$file1"  "$file2 >> /put/your/own/path/here///REPORT.d1;
        echo -e  "### running STAR " >> /put/your/own/path/here///REPORT.d1;
        time /home/milanova/soft/bin/star-mapper/STAR_2.3.1l/STAR --readFilesIn $file1 $file2 --genomeDir /srv/gsfs0/projects/snyder/milanova/genomes/H.sapiens/golden_path_200902/wholeGenomeUnzipped/starIndex_gencode15/ --readFilesCommand zcat --outSAMunmapped Within --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 5 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 25 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --runThreadN 8 --outSAMstrandField intronMotif --outStd SAM | gzip -c > $myTMPDIR/tmp.STAR.sam.gz 2>> /put/your/own/path/here///REPORT.d1
        echo -e  "### sorting bam " >> /put/your/own/path/here///REPORT.d1;
        time zcat $myTMPDIR/tmp.STAR.sam.gz | samtools view -bS - | samtools sort -m 19G - $myTMPDIR/mapping 2>> /put/your/own/path/here///REPORT.d1
#### running cufflinks
        /home/milanova/soft/bin/cufflinks/cufflinks-2.1.1.Linux_x86_64/cufflinks -N mapping.bam -G /srv/gsfs0/projects/snyder/milanova/data/annotations/H.sapiens/hg19/gencode.v15.annotation.gtf.ExonLinesOnly.gtf.Unzipped --min-intron-length 25 -o ./ 2>> /put/your/own/path/here///REPORT.d1
```

```
        mv $myTMPDIR/* /put/your/own/path/here///$subdirname/
        cd ..
done
rmdir $myTMPDIR
#
```

APPENDIX E - CALCULATION OF GENE BODY COVERAGE

To evaluate read coverage over gene body, we used a Python script (geneBody_coverage.py, part of the RNA-seq quality control RSeQC package) to check uniformity of reads coverage and if there are any 5' or 3' biases. The program workflow is the following:

1 We start with inputting the bam file

2 The program calculates Pearson's skewness coefficients and ranks samples by skewness of the coverage.

RSeQC.sh

```
#!/bin/sh
#  RSeqQC.sh
#$ -N RSeQC.sh
#$ -l h_vmem=48G
# -l h_rt=167:00:00
#$ -o /srv/gsfs0/projects/snyder/milanova/logout_dup.txt
#$ -e /srv/gsfs0/projects/snyder/milanova/error_dup.txt
#$ -w e
#$ -cwd
module load python/2.7 r/3.2.2 gcc/5.2.0 rseqc/2.6.2 samtools
#read_distribution.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
LCL-Nuc-R3-Lane1/mapping.bam -r
/srv/gsfs0/projects/snyder/Damek/transcriptomes/hg19_RefSeq.bed


#FPKM_count.py -d '1++,1--,2+-,2-+' -r
/srv/gsfs0/projects/snyder/Damek/transcriptomes/hg19_RefSeq.bed
#read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
K562-ATC-Cyt-R4-Lane2/mapping.bam -o K562-ATC-Cyt-R4-Lane2
```

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Nuc-R4-Lane1/mapping.bam -o K562-ATC-Nuc-R4-Lane1

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Cyt-R1-Lane2/mapping.bam -o K562-code-Cyt-R1-Lane2

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Cyt-R2-Lane2/mapping.bam -o K562-code-Cyt-R2-Lane2

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Nuc-R1-Lane2/mapping.bam -o K562-code-Nuc-R1-Lane2

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Nuc-R2-Lane2/mapping.bam -o K562-code-Nuc-R2-Lane2

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Nuc-R2-Lane1/mapping.bam -o K562-ATC-Nuc-R2-Lane1

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Cyt-R2-Lane1/mapping.bam -o K562-ATC-Cyt-R2-Lane1

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Cyt-R2-Lane2/mapping.bam -o LCL-Cyt-R2-Lane2

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Cyt-R3-Lane1/mapping.bam -o LCL-Cyt-R3-Lane1

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Nuc-R2-Lane2/mapping.bam -o LCL-Nuc-R2-Lane2

#read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Nuc-R3-Lane1/mapping.bam -o LCL-Nuc-R3-Lane1

read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Cyt-R4-Lane2/mapping.bam -o K562-ATC-Cyt-R4-Lane2

read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Nuc-R4-Lane1/mapping.bam -o K562-ATC-Nuc-R4-Lane1

read_duplication.py -i /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Cyt-R1-Lane2/mapping.bam -o K562-code-Cyt-R1-Lane2

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
K562-code-Cyt-R2-Lane2/mapping.bam -o K562-code-Cyt-R2-Lane2

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
K562-code-Nuc-R1-Lane2/mapping.bam -o K562-code-Nuc-R1-Lane2

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
K562-code-Nuc-R2-Lane2/mapping.bam -o K562-code-Nuc-R2-Lane2

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
K562-ATC-Nuc-R2-Lane1/mapping.bam -o K562-ATC-Nuc-R2-Lane1

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
K562-ATC-Cyt-R2-Lane1/mapping.bam -o K562-ATC-Cyt-R2-Lane1

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
LCL-Cyt-R2-Lane2/mapping.bam -o LCL-Cyt-R2-Lane2

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
LCL-Cyt-R3-Lane1/mapping.bam -o LCL-Cyt-R3-Lane1

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
LCL-Nuc-R2-Lane2/mapping.bam -o LCL-Nuc-R2-Lane2

read_duplication.py -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
LCL-Nuc-R3-Lane1/mapping.bam -o LCL-Nuc-R3-Lane1

geneBody_coverage.py -r
/srv/gsfs0/projects/snyder/Damek/transcriptomes/hg19_RefSeq.bed -i
/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/
K562-ATC-Cyt-R4-
Lane2/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_R
NAseq_2015_10_v2/K562-ATC-Nuc-R4-
Lane1/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_R
NAseq_2015_10_v2/K562-code-Cyt-R1-
Lane2/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_R
NAseq_2015_10_v2/K562-code-Cyt-R2-
Lane2/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_R
NAseq_2015_10_v2/K562-code-Nuc-R1-
Lane2/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_R
NAseq_2015_10_v2/K562-code-Nuc-R2-
Lane2/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_R
NAseq_2015_10_v2/K562-ATC-Nuc-R2-
Lane1/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_R
NAseq_2015_10_v2/K562-ATC-Cyt-R2-

Lane1/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Cyt-R2-Lane2/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Cyt-R3-Lane1/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Nuc-R2-Lane2/mapping.bam,/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/LCL-Nuc-R3-Lane1/mapping.bam -o /srv/gsfs0/projects/snyder/milanova/output_gene_cov

APPENDIX F - DIFFERENTIAL GENE EXPRESSION ANALYSIS WITH CUFFMERGE AND CUFFDIFF

In this step we create the transcriptome assembly from the assembled transcripts generated from the Cufflinks output. We show example shell scripts for merging of bam files and calculation of differential gene expression between given single cell fractions and aggregated samples.

Cuffmerge:

```
#!/bin/sh
#  bammerge.sh
#$ -N bammerge.sh
#$ -l h_vmem=48G
#$ -pe shm 1
#$ -o /srv/gsfs0/projects/snyder/milanova/Samtools/logout.txt
#$ -e /srv/gsfs0/projects/snyder/milanova/Samtools/error.txt
#$ -w e
#$ -cwd
module load samtools
samtools merge out_K562all_5cell.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Cyt-R2-Lane1/mapping.bam
```

/srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Cyt-R3-Lane1/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Cyt-R4-Lane2/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Cyt-R1-Lane2/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Cyt-R2-Lane2/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Nuc-R2-Lane1/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Nuc-R3-Lane2/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-ATC-Nuc-R4-Lane1/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Nuc-R1-Lane2/mapping.bam /srv/gsfs0/projects/snyder/milanova/Cufflinks/nuc_vs_cyt_RNAseq_2015_10_v2/K562-code-Nuc-R2-Lane2/mapping.bam

Cuffdiff:

```
#$ -N cutadapt2.sh

#$ -l h_vmem=4G

#$ -pe shm 8

#$ -o /srv/gsfs0/projects/snyder/milanova/Cuffdiff/logout51.txt
```

```
#$ -e /srv/gsfs0/projects/snyder/milanova/Cuffdiff/error51.txt

#$ -cwd

module load python/2.7 cutadapt bowtie samtools tophat cufflinks

#outdir=/srv/gsfs0/projects/snyder/Damek/MS_Fib_R123

cuffdiff -p 8 -L allK562,allLCL -o out_K562_vs_LCL -b
/srv/gsfs0/projects/snyder/Damek/hg19/hg19.fa -u
/srv/gsfs0/projects/snyder/milanova/merge_out_all/merged.gtf
/srv/gsfs0/projects/snyder/milanova/Samtools/out_K562all_5cell.bam
/srv/gsfs0/projects/snyder/milanova/Samtools/out_LCL_3cell.bam
```

APPENDIX G - SCRIPTS FOR DATA POST-PROCESSING

5.1 R script for the generation of gene expression correlations

```
library(RSQLite)
 library(ggplot2)
 library(reshape2)
 library(plyr)
 library(fastcluster)
 library(rtracklayer)
 library(Gviz)
 library(BiocGenerics)
 library(Hmisc)
 library(cummeRbund)
 library(grid)
 library(cowplot)
 library(gridExtra)
 library(cowplot)
 ###################################
 ## f (scientific notation)
 ###################################
 fancy_scientific <- function(l) {
  # turn in to character string in scientific notation
  l <- format(l, scientific = TRUE)
  # quote the part before the exponent to keep all the digits
  l <- gsub("^(.*)e", "'\\1'e", l)
  # turn the 'e+' into plotmath format
  l <- gsub("e", "%*%10^", l)
  # return this as an expression
  parse(text=l)
 }
```

```
#####################################################################
##
  ### out_K562ATCC_cR2_cR4
#### change input file name
  Kat1.raw <-
read.table("/Users/milanova/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_cR2
_cR4/genes.fpkm_tracking",
                header = TRUE, sep = "\t", quote="\"", dec=".")
  Kat1.fpkm <- rbind(Kat1.raw[,10],Kat1.raw[,14])
  Kat1.fpkm <- t(Kat1.fpkm)
  Kat1.cleaned <- Kat1.raw[which(rowSums(Kat1.fpkm) > 0),]  #all data
## PLOT
  #### change aes names
  Kat1 <- ggplot(Kat1.raw) + aes(x=cR2_FPKM, y=cR4_FPKM) +
    scale_x_log10(limits= c(0.0001, 100000), labels=fancy_scientific) +
    scale_y_log10(limits= c(0.0001, 100000), labels=fancy_scientific) +
    stat_density2d(geom="tile", aes(fill=..density..^0.25, alpha=1),
contour=FALSE) +
    geom_point(size=0.3) +
    scale_alpha(guide = 'none') +
    stat_density2d(geom="tile", aes(fill=..density..^0.25,
alpha=ifelse(..density..^0.25<0.4,0,1)), contour=FALSE) +
    scale_fill_gradientn(colours = colorRampPalette(c("white", blues9))(256)) +
    theme(legend.position="none") +
    theme(panel.border = element_rect(colour = "black", fill=NA, size=1,
linetype="solid")) +
    geom_abline(position = "identity", intercept = 0, linetype= "dashed", size = 0.5)
+
    coord_fixed(1)+
  #### change aes names
    geom_point(aes(x=cR2_FPKM, y=cR4_FPKM),size=0.3) +
    theme(axis.title.x=element_blank(),
          axis.title.y=element_blank()) +
# put correlations
    geom_text(aes(x = 0.003, y = 50000,
```

```
        #### change input variable
        label=paste("Pearson =", round(cor(log10(cR2_FPKM+1),
        #### change input variable
        log10(cR4_FPKM+1), method='pearson'),3)), group=NULL,
parse=TRUE)) +
  geom_text(aes(x = 0.003, y = 10000,
        #### change input variable
        label=paste("Spearman =", round(cor(log10(cR2_FPKM+1),
        #### change input variable
        log10(cR4_FPKM+1), method='spearman'),3)), group=NULL,
parse=TRUE))
```

5.2 R script for the generation of data principal components

```
# load data

fpkm.data.raw    <-    read.table("/Users/milanova/Desktop/R-
scripts/combined_matrix 4.tab", header = TRUE, sep = "\t",
row.names=1)



#fpkm.data.selected <- rbind(fpkm.data.raw[,2:16])

#all data; data.frame

#fpkm.data.cleaned                                          <-
fpkm.data.raw[which(rowSums(fpkm.data.selected) > 0),]

#fpkm.data.matrix                                           <-
as.matrix(fpkm.data.cleaned[1:15429,2:16])



# transform data to log scale

fpkm.log=log(fpkm.data.raw+1)
```

```
corner(fpkm.log)

dim(fpkm.log)


data=new("seurat",raw.data=fpkm.log)

data=setup(data,project="NBT",min.cells = 3,

          names.field = 1,

          names.delim = "_",min.genes = 1000,is.expr=1)

# plot violin plots for ACTB, GAPDH, and GATA1

vlnPlot(data,c("ENSG00000075624.9","ENSG00000218582.2","ENSG0
0000102145.8","ENSG00000206172.4"))

vlnPlot(data,c("ENSG00000138382.9")) #METTL5

vlnPlot(data,c("ENSG00000179348.7")) #GATA2

# check fraction-to-fraction correlation

cellPlot(data,data@cell.names[1],data@cell.names[3],do.ident
= FALSE)

# select the top most highly expressed genes

data=mean.var.plot(data,y.cutoff = 2,x.low.cutoff = 2,fxn.x =
expMean,fxn.y = logVarDivMean)

length(data@var.genes)

# calculate the principal components and plot data wrt
PC1,PC2, and PC3

data=pca(data,do.print=FALSE)

pca.plot(data,1,2,pt.size = 4)

pca.plot(data,1,3,pt.size = 4)

####
```

```
# plot a heat map of all samples

data=project.pca(data,do.print=FALSE)

pcHeatmap(data,pc.use = 1,use.full = TRUE,do.balanced =
TRUE,remove.key = FALSE)

# visualize the first principal component

viz.pca(data,1)

data_matrix <-
matrix(c(mapping.1,mapping.11,mapping.10,mapping.9,mapping.4,
mapping.5,mapping.6,mapping,mapping.7,mapping.3,mapping.8,map
ping.2), byrow=T, ncol=100)

rowLabel <-
c("mapping.1","mapping.11","mapping.10","mapping.9","mapping.
4","mapping.5","mapping.6","mapping","mapping.7","mapping.3",
"mapping.8","mapping.2")


pdf("/srv/gsfs0/projects/snyder/milanova/output_gene_cov.gene
BodyCoverage.heatMap.pdf")

rc <- cm.colors(ncol(data_matrix))

heatmap(data_matrix, scale=c("none"),keep.dendro=F, labRow =
rowLabel ,Colv = NA,Rowv =
NA,labCol=NA,col=cm.colors(256),margins = c(6,
8),ColSideColors = rc,cexRow=1,cexCol=1,xlab="Gene body
percentile (5'->3')", add.expr=x_axis_expr <-
axis(side=1,at=c(1,10,20,30,40,50,60,70,80,90,100),labels=c("
1","10","20","30","40","50","60","70","80","90","100")))

dev.off()
```

5.3 R script for plotting the output for gene body coverage

```
pdf("/srv/gsfs0/projects/snyder/Damek/other/output_gene_cov.g
eneBodyCoverage.curves.pdf")

x=1:100

icolor =
colorRampPalette(c("#7fc97f","#beaed4","#fdc086","#ffff99","#
386cb0","#f0027f"))(12)

layout(matrix(c(1,1,1,2,1,1,1,2,1,1,1,2), 4, 4, byrow =
TRUE))

plot(x,mapping.1,type='l',xlab="Gene body percentile (5'-
>3')", ylab="Coverage",lwd=0.8,col=icolor[1])

lines(x,mapping.11,type='l',col=icolor[2])

lines(x,mapping.10,type='l',col=icolor[3])

lines(x,mapping.9,type='l',col=icolor[4])

lines(x,mapping.4,type='l',col=icolor[5])

lines(x,mapping.5,type='l',col=icolor[6])

lines(x,mapping.6,type='l',col=icolor[7])

lines(x,mapping,type='l',col=icolor[8])

lines(x,mapping.7,type='l',col=icolor[9])

lines(x,mapping.3,type='l',col=icolor[10])

lines(x,mapping.8,type='l',col=icolor[11])

lines(x,mapping.2,type='l',col=icolor[12])

par(mar=c(1,0,2,1))

plot.new()

legend(0,1,fill=icolor[1:12],legend=c('mapping.1','mapping.11
','mapping.10','mapping.9','mapping.4','mapping.5','mapping.6
','mapping','mapping.7','mapping.3','mapping.8','mapping.2'))

dev.off()
```

5.4 R script for data visualizations with the CummeRbund package

The CummeRbund package allows users to explore and plot data directly from the Cuffdiff output. We show an example script for basic data gene expression data exploration, gene + isoform expression correlations and statistical results.

```
library(RSQLite)
library(ggplot2)
library(reshape2)
library(plyr)
library(fastcluster)
library(rtracklayer)
library(Gviz)
library(BiocGenerics)
library(Hmisc)
library(cummeRbund)
library(grid)
library(cowplot)
library(gridExtra)
#####    READ FRACTION
#Compare single cyt reps
#ATCC
cuff_K562ATCC_cR1_cR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_cR1_cR2')
cuff_K562ATCC_cR2_cR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_cR2_cR3')
cuff_K562ATCC_cR2_cR4 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_cR2_cR4')
cuff_K562ATCC_cR3_cR4 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_cR3_cR4')
```

```
cuff_K562ATCC_nR2_nR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_nR2_nR3')

cuff_K562ATCC_nR2_nR4 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_nR2_nR4')

cuff_K562ATCC_nR3_nR4 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_nR3_nR4')

cuff_K562ATCC_nR2_cR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_nR2_cR2')

cuff_K562ATCC_nR3_cR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_nR3_cR3')

cuff_K562ATCC_nR4_cR4 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_nR4_cR4')

cuff_K562ATCC_cyt_nuc <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ATCC_cyt_nuc')

#ENCODE

cuff_K562ENCODE_cR1_cR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ENCODE_cR1_cR2')

cuff_K562ENCODE_cR1_cR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ENCODE_cR1_cR3')

cuff_K562ENCODE_cR2_cR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ENCODE_cR2_cR3')

cuff_K562ENCODE_nR1_nR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ENCODE_nR1_nR2')

cuff_K562ENCODE_nR1_cR1 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ENCODE_nR1_cR1')

cuff_K562ENCODE_nR2_cR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ENCODE_nR2_cR2')

cuff_K562ENCODE_cyt_nuc <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_K562ENCODE_cyt_nuc')

#LCL

cuff_LCL_cR1_cR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_cR1_cR2')

cuff_LCL_cR1_cR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_cR1_cR3')

cuff_LCL_cR2_cR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_cR2_cR3')

cuff_LCL_nR1_nR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_nR1_nR2')

cuff_LCL_nR1_nR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_nR1_nR3')
```

```
cuff_LCL_nR2_nR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_nR2_nR3')
cuff_LCL_nR1_cR1 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_nR1_cR1')
cuff_LCL_nR2_cR2 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_nR2_cR2')
cuff_LCL_nR3_cR3 <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_nR3_cR3')
cuff_LCL_cyt_nuc <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_LCL_cyt_nuc')
#across cell strain and type per compartment
cuff_cK562tot_cLCLtot <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_cK562tot_cLCLtot')
cuff_nK562tot_nLCLtot <-
readCufflinks('~/Desktop/Cuffdiffs_toR/Fraction/out_nK562tot_nLCLtot')
#####    CV2 CELL TYPE FIGURE
genes_f_nuc.scv <- fpkmSCVPlot(genes(cuff_nK562tot_nLCLtot))
genes_f_nuc.scv
genes_f_cyt.scv <- fpkmSCVPlot(genes(cuff_cK562tot_cLCLtot))
genes_f_cyt.scv
isoforms_f_cyt.scv <- fpkmSCVPlot(isoforms(cuff_cK562tot_cLCLtot))
isoforms_f_cyt.scv
isoforms_f_nuc.scv <- fpkmSCVPlot(isoforms(cuff_nK562tot_nLCLtot))
isoforms_f_nuc.scv
#####    CV2 COMPARTMENT FIGURE
genes_Kat.scv <- fpkmSCVPlot(genes(cuff_K562ATCC_cyt_nuc))
genes_Kat.scv
isoforms_Kat.scv <- fpkmSCVPlot(isoforms(cuff_K562ATCC_cyt_nuc))
isoforms_Kat.scv
genes_Ken.scv <- fpkmSCVPlot(genes(cuff_K562ENCODE_cyt_nuc))
genes_Ken.scv
isoforms_Ken.scv <- fpkmSCVPlot(isoforms(cuff_K562ENCODE_cyt_nuc))
isoforms_Ken.scv
genes_L.scv <- fpkmSCVPlot(genes(cuff_LCL_cyt_nuc))
genes_L.scv
isoforms_L.scv <- fpkmSCVPlot(isoforms(cuff_LCL_cyt_nuc))
```

```
isoforms_L.scv
#####    GENE DENSITY FIGURE
d1 <- csDensity(genes(cuff_nK562tot_nLCLtot)) +
  scale_y_continuous(limits= c(0, 0.4))
d2 <- csDensity(genes(cuff_cK562tot_cLCLtot)) +
  scale_y_continuous(limits= c(0, 0.4))
d3 <- csDensity(genes(cuff_K562_vs_LCL)) +
  scale_y_continuous(limits= c(0, 0.4))
blankPlot <- ggplot()+geom_blank(aes(1,1)) +
  cowplot::theme_nothing()
grid.arrange(d3,blankPlot,
             ncol=2, nrow=1)
```