Characterization of drug-induced splicing complexity in prostate cancer cell line using long read technology

Xintong Chen[1], Sander Houten[1], Kimaada Allette[1], Robert P.Sebra[1], Gustavo Stolovitzky*[1,2] and Bojan Losic*[1]

1. *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Ave, New York, NY 10029, USA*
2. *IBM Translational Systems Biology and Nanobiotechnology Research, Yorktown Heights, NY 10598, USA*

**Corresponding Authors: Email: bojan.losic@mssm.edu, gustavo@us.ibm.com*

Abstract

We characterize the transcriptional splicing landscape of a prostate cancer cell line treated with a previously identified synergistic drug combination. We use a combination of third generation long-read RNA sequencing technology and short-read RNAseq to create a high-fidelity map of expressed isoforms and fusions to quantify splicing events triggered by treatment. We find strong evidence for drug-induced, coherent splicing changes which disrupt the function of oncogenic proteins, and detect novel transcripts arising from previously unreported fusion events.

Keywords: Combination treatment; Long read sequencing; Alternative splicing; Cancer.

Introduction

Background

Prostate cancer is the second-most common cancer and has the third leading cancer mortality among men in the USA.[1] Major clinical interventions for prostate cancer include surgical procedure, radiation, androgen depletion treatment (ADT) and chemotherapy. As with other cancers, prognosis of prostate cancer varies largely depending on its molecular characteristics [2]. A number of large-scale collaborative efforts and crowd-sourcing initiatives have recently been used to profile genomic data on cancer systems perturbed by thousands of compounds to infer agents with curing potential. The Library of Integrated Network-Based Cellular Signatures (LINCS) Program, for example, provides a rich public source of gene expression data collected from cell lines with exposure to various compounds. These data allows researchers to gain mechanistic insight into the biological processes that are altered by different drugs in a given cellular context (cell line) [http://www.lincsproject.org/]. On the analytical side, the NCI-DREAM Drug Sensitivity Prediction Challenge, run in 2012, encouraged the development of algorithms to predict the sensitivity of cancer cell lines to a panel of drugs based on multi-omics

data, including gene expression, copy number variation, mutation and proteomics data [3]. The AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge, launched in 2015, is a similar international competition seeking for algorithms that accurately predict synergistic combination treatment based on gene expression data and multiple cancer cell lines [4]. Despite the importance of these and other efforts in probing the drug treated omics expression landscape, splicing modulation in treatment has remained largely unexplored.

Alternative splicing (AS) events are key generators of proteomic diversity. Yet the functional impact of alternative splicing is only now beginning to be systematically quantified, thanks, in part, to new technological advances that overcome the difficulties related to the read length of sequencing assays to detect isoforms. Indeed, the short-read length of second generation sequencing technology (usually 50bp or 100bp) directly leads to the key difficulty of unambiguously phasing isoforms and mapping highly repetitive sequence. Third generation sequencing platforms such as PacBio and Oxford Nanopore utilize long read sequences to help address this issue. Isoform sequencing (IsoSeq) is a recently developed PacBio assay which can directly sequence full-length transcript sequences. With the help of IsoSeq, an astonishing diversity of splicing events in various systems has started to emerge even in well-studied cancer cell-line systems such as MCF7[5]. A number of other analyses in other cellular contexts have also utilized the IsoSeq assay[6-11]. Given that alternative splicing is known to be tissue and condition dependent, we hypothesized that drug administration should also alter the splicing landscape of cells as part of multifaceted cellular response to stimuli which cannot be completely captured using standard RNAseq analysis.

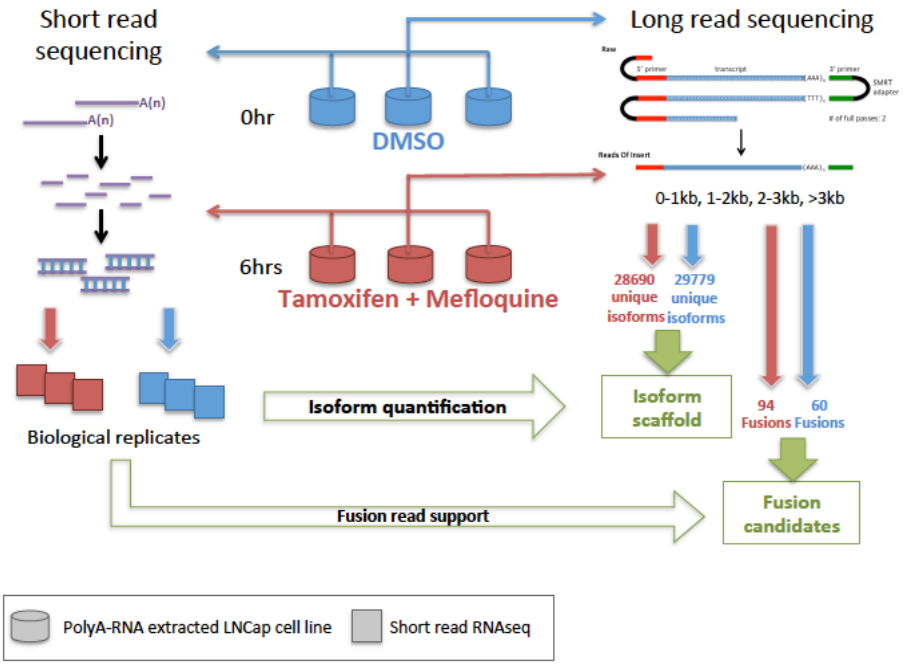

Figure1. A schematic of study design. Three biological replicates of short read RNAseq were generated for untreated (DMSO) and treated (TM) and used for subsequent analysis; long read sequencing were performed for untreated and treated samples. Both technologies were combined for isoform quantification and identification of fusion events.

Experimental design

We used LNCap prostate cell lines as our cancer model. After treatment with a previously identified synergistic combination treatment of Tamoxifen and Mefloquine(TM)[4, 12], we measured the LNCap cell viability as a function of time relative to treatment with DMSO and found that cell viability decreases to ~30% at times as early as 6h. As shown in Figure 1, we generated three biological replicates from a baseline condition with LNCap cells cultured in DMSO at 0hr; and a treated condition where cells were cultured with Tamoxifen and Mefloquine at 6hrs. We collected polyA enriched RNA from these six samples and performed short read RNAseq experiments. We also randomly selected one of the replicates from each condition to perform long read PacBio isoform sequencing (IsoSeq) (Figure1). To avoid loading bias against short transcripts, we performed size-selection for multiple bins (0-1kb, 1-2kb, 2-3kb and >3kb) using SageELF device for both samples. Purified SMRTbell libraries were sequenced on the PacBio RSII machine using P6-C4 chemistry with 8 SMRT cells.

Results

Bioinformatics pipeline for IsoSeq

We wrapped up a pipeline for IsoSeq analysis (Figure S1). In brief, by searching for sequencing adapters, reads of insert were classified as Circular Consensus Sequences (CCS) and non-CCS, and further classified as full length or not through searching of 5' and 3' primers and polyA signals. Next, we performed isoform-level-clustering (ICE) and Quiver polish on non-artifacts full length (FL) reads to improve error corrected consensus accuracy via SMRT Portal, yielding 30669 and 31095 FL consensus isoforms for baseline and treated conditions respectively, with expected accuracy greater than 0.99. All high quality FL transcripts were aligned to the hg19 genome using GMAP[13] with default parameters and collapsed to remove redundant sequences via the pbtranscript-TOFU package [https://github.com/PacificBiosciences/cDNA_primer/wiki] with minimum alignment accuracy of 0.99 and minimum coverage of 0.85. Transcripts sharing the exact same exons except those with an extended 5' end were collapsed into a single transcript. The unique transcripts were then aligned back using STAR[14] and compared with the gencode.v19 database[15] by MatchAnno [https://github.com/TomSkelly/MatchAnnot]. Junction modes were predicted for each transcript using Astalavista[16] through classifying splice sites of alternative splicing external(ASE), alternative splicing internal (ASI), splice sites that extend transcript structures(DSP) and any other differences in transcript structures not involving splice sites(VST). From there we computed open reading frames (ORFs) with at least 100 amino acids from all high quality FL unique transcripts and screened for homology to known proteins by

aligning to UniProtKB/Swiss-Prot protein database (BLASTP, e-value<1e-5) and applying hmmer (3.1b1)[17] to scan for protein domains (Pfam,E<10). Finally, TransDecoder (3.3.0) was used to leverage blast hit and detected domains to look for coding regions.

	#Unique isoform	Coding potential	Known isoforms	Novel isoforms			Others (unknown gene/not aligned)
				Exons match except size	Some exons match	Novel exons	
TM6hrs	28690	74.1% (21257)	31% (8894)	5%	46%	14%	4%
DMSO 0hr	29779	73.2% (21808)	36% (10720)	5%	40%	15%	4%

Table1. General statistics of PacBio FL isoforms detection.

Isoform complexity in LNCap

We detected 29779 and 28690 unique full-length (FL) isoforms from IsoSeq data of un-treated and treated LNCap cell line respectively (see methods), as shown in Table 1. Transcripts mapped to unknown gene locus or not aligned to the reference genome are marked as others and not included in the following comparison. For comparison, following the same procedure in the publicly available high quality deep sequenced (with 119 SMRT Cells) MCF7 IsoSeq data (2013 version) we detected 40447 unique FL isoforms. For all three datasets, the majority of the detected isoforms are novel: 60%, 65% and 71% for un-treated LNCap, treated LNCap and MCF7 respectively. We characterized the novel isoforms into three categories: 1) Novel isoforms with exons matching to exons one-for-one, but sizes of the internal exons may disagree; 2) Novel isoforms with only some of the exons matching annotated exons, and 3) Novel isoform overlap with annotated genes, but no exon matches annotation. The majority of detected novel isoforms fall into category (1) and (2), as shown in Table1. These "partially novel" isoforms are likely due to undocumented splicing patterns in current annotations, while only a small proportion of FL isoforms are completely novel (7%~15%).

These results imply that there is non-trivial information in the fine splitting of the short-read expression spectrum (over our long-read scaffold) across some genomic loci into distinct transcripts. In fact, the entropy S_i at a given gene locus i **for a given sample** is computed as the empirical Shannon entropy of the normalized, variance stabilized expression frequencies across all **informative** transcripts T_{ij} at that locus, namely the average log

$$S_i = -< \ln [P(T_{ij})] >_{P(T_{ij})} \quad (1)$$

such that $P(T_{ij})$ is the probability (normalized frequency) of j-th informative transcript T to appear at genomic locus i. The frequency bins are estimated using Sturge's rule [18]. A transcript is defined as **informative** if it is on average well expressed after being penalized by fitting a

global mean variance trend [19] of the short-read expression dataset. Indeed, this penalty is proportional to the **inverse variance** of the transcript expression, which increases as expression decreases due to amplification noise. We thus demand that $P(T_{ij})$ in equation (1) only includes transcripts that satisfy

$$E'_{ij} > \mu(E_{ij}\,\omega ij) \quad (2)$$

such that E_{ij} is the variance stabilized expression values, ωij is a quality factor proportional to the inverse count variance, and μ is mean of the distribution of $E_{ij}\,\omega ij$ at each locus i. Note that genomic loci with zero or only one informative transcript are naturally assigned an entropy of zero and $E_{ij}\,\omega ij$ tends to zero with low expression much faster than E_{ij}. Length-bias and sample variations are treated by dividing each S_i by the **number of transcripts** at the ith locus and then taking their **median across all samples**, i.e.

$$S' = median\,(\frac{S_i}{N}) \quad (3)$$

Figure 2A depicts the distribution of S' and its dependence on overall expression, from which it is clear that fine-splitting entropy per transcript is not a trivial artifact dominated by noisy, lowly expressed transcripts in outlier samples. It is also clear that while many genomic loci are well approximated by collapsing all transcripts to a single genic locus -- and thus have few informative distinct transcripts which corresponds to a low entropy (the 'zero mode') -- many others do not and likely imply nontrivial biological regulation.

To better characterize detected IsoSeq isoforms, we surveyed their coding potential. Through scanning of ORFs (>100 aa) in protein databases, protein homologies were then leveraged to maximize coding regions prediction sensitivity. We found 73.2% and 74.1% of detected isoforms were predicted to have coding potential which strongly suggests functional consequences of splicing events in the system. Taken together, novel splicing events leading to isoforms with coding potential are observed in LNCap cell lines with comparable statistics in external MCF7 IsoSeq data.

Treatment induced nontrivial splicing signals in LNCap cell line.

Only 10417 of the detected isoforms overlapped between treated (19362 treated-unique) and untreated (18273 untreated-unique) conditions. To analyze the functional impact of these isoforms, we perform functional annotation through classification of protein families and domains (see methods). In brief, we found 603 and 569 domains in DMSO and TM isoforms respectively, with 553 domains in common and recurrent (>2) condition-unique domains highlighted in text (Figure 2.B). We observed key oncogenic protein domains (families) frequently found in DMSO but not in treatment, such as histone deacetylase interacting, PI3-kinase, p85- and Ras binding

domains and DNA polymerase A/B families. Notably frizzled protein, which activates the Wnt pathway is ranked as the top DMSO-unique proteins[20]. A full list of condition specific domain annotation is shown in Table S1. We also observed a similar pattern through a different protein family prediction algorithm as shown in Figure S3. Thus we hypothesize that the decrease in survival under drug treatment may be implemented by breaking the oncogenic domains in LNCap by induction of targeted splicing events. For example, histone deacetylation (HDAC) has been found to play major role in prostate cancer progression making HDAC inhibitor a potential anti-tumor therapeutic target.[21, 22] In our IsoSeq data, HDAC interacting domain is only detected in isoforms in LNCap+DMSO but not the TM treated condition, while overall gene expression of *HDAC1* (encoding histone deacetylase family as a component of the histone deacetylase complex) is up-regulated in treatment of TM (fold change>1.68, FDR< 3.4e-07), indicating treatment induced splicing at the locus may shut down histone deacetylation regardless of increased gene expression.

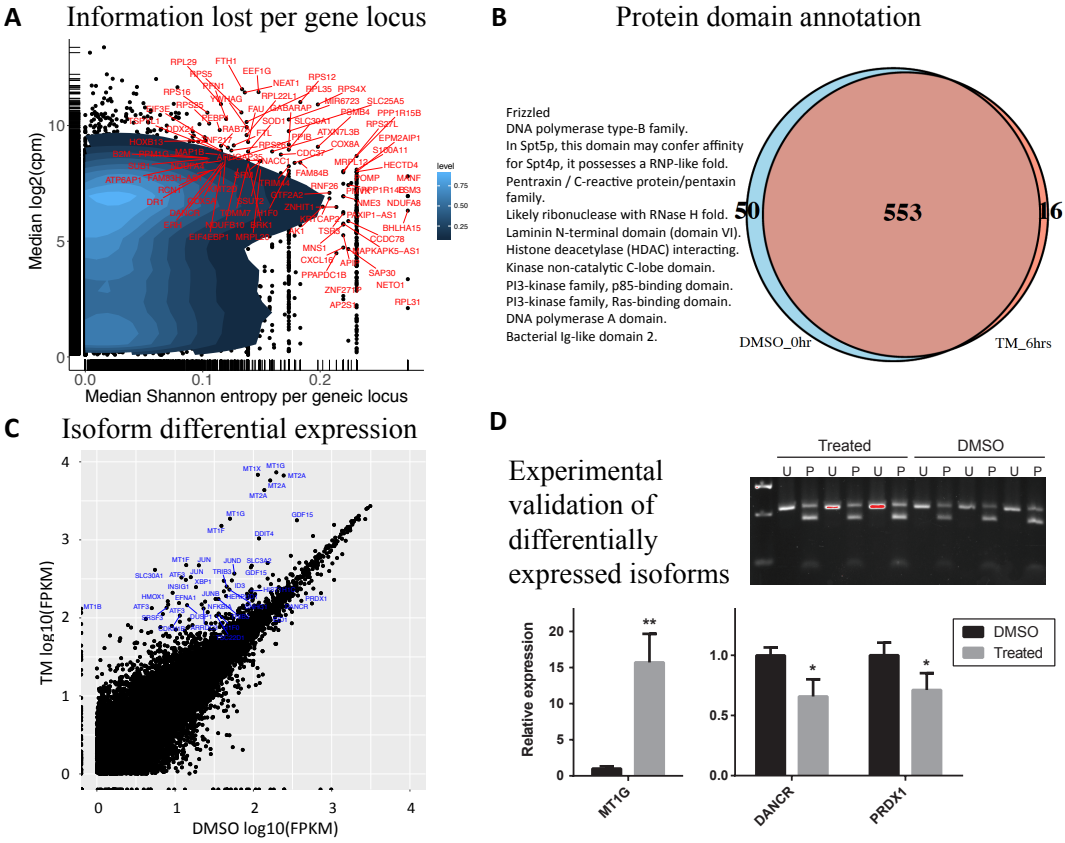

Figure2. A) Median normalized fine splitting entropy for each genomic locus is on x-axis with median expression on y-axis, outlier genes are labeled in red. B) Venn diagram of predicted protein domains in DMSO and TM. Frequently detected domains (frequency>=3) are highlighted in text. C) Plot of FPKM in log10 scale of DMSO (x-axis) VS FPKM in log10 scale of TM (y-axis), well expressed significant

differentially expressed isoforms (FDR<0.05, FPKM >1) are gene labeled in blue. D) Upper panel: MT1G PCR fragment (325bp) uncut (U) or digested with PstI (P). The intensity of the digested fragment is higher indicating that the NM_001301267 transcript is more prominent; lower panel: Quantitative PCR for MT1G (both isoforms), DANCR (NR_024031.2) and PRDX1 (NM_181697.2). Relative expression is calculated as 2 to the power of the Ct of RPLP0 - Ct of the tested gene. Average of the DMSO control group is set to 1.

We also find that the relative prevalence of key splicing modalities (exon skipping, intron retention, alt donor/receptor, etc.) among the detected isoforms shows no significant bias (Figure S2) although we do note that generally the baseline level of intron retention exceeds that found in the reference annotation [23]. We further summarize intron-retention events for known and novel isoforms (Table S2), partial novel isoforms has similar distribution as the known ones while novel isoforms with novel exons appear to be much fewer. Despite this lack of specificity of splicing modality, we conclude that treatment related isoforms are gaining and losing key functional protein domains previously identified to be important for cancer progression[21, 22, 24, 25].

Quantification and validation of IsoSeq FL isoforms using short read data

Although long read technology creates an integrated transcriptome scaffold, it is suboptimal for expression quantification essentially due to a relatively sparse and non-uniform coverage profile. Following previous work[5, 6], we used short read replicates to quantify IsoSeq isoforms (Figure 1). In brief, we generate an assembled GTF file summarizing all IsoSeq FL transcripts and their genomic information (hg19) and assign short reads from the 6 RNAseq samples to its features. Replicates per condition were grouped together to improve abundance estimates for isoform differential expression analysis between treatments and untreated. We detected 38 highly expressed isoforms up-regulated and 3 isoforms down-regulated in the treatment group relative to DMSO. The parent genes of these isoforms are known to be involved in cancer related pathways, such as *JUN*, *DDIT4*, *CDKN1B* and transcription factor *ATF3*. We also found a splice variant of a cell differentiation regulating long non-coding RNA: *DANCR* being down-regulated in treatment (Figure 2.C). We selected 3 differentially and well expressed isoforms (fpkm>100) for further experimental validation. The genes and variants of interest were first amplified by PCR and product sizes were estimated by agarose gel electrophoresis followed by Sanger sequencing to confirm correct amplification. The *MT1G* mRNA had two variants; NM_005950 and NM_001301267. The latter variant uses an alternate in-frame splice site in the 5' coding region and has a 3bp insert that introduces a PstI restriction site. A restriction digest revealed that most of the PCR product is digested by PstI indicating that NM_001301267 is the most prominent transcript (Figure 2.D), which was further confirmed by Sanger sequencing. Of the two tested

DANCR isoforms, we detected only NR_024031.2. For *PRDX1*, two variants were detected, but NM_181697.2 was most prominent. We performed quantitative PCR for *MT1G* (both isoforms), *DANCR* (NR_024031.2) and *PRDX1* (NM_181697.2). Treatment of cells increased *MT1G* mRNA expression, but decreased *DANCR* and *PRDX1* confirming the short read and IsoSeq data.

Fusion landscape revealed in treatment

Gene fusions have been found to play a major role in prostate tumorigenesis. We leveraged the long read lengths of the IsoSeq assay to create a detailed map to track the expression of fusion transcripts. Using HQ consensus reads we infer a set of fusion transcripts and further filter these candidates with stringent threshold of short read support (see Method), we detected 94 fusions (83 inter-chromosomal and 11 intra-chromosomal) and 60 fusions (55 inter-chromosomal and 5 intra-chromosomal) in treated and untreated, respectively (Figure S5). The number of fusion candidates found in treated condition more than in untreated with very few in common, suggest treatment may reveal the altered genome structure. We collected a list of 14 known LNCap fusions from literature[26-28](Table S4), and found 6 of them present in DMSO and 2 present in treated, in the set of IsoSeq fusion transcripts candidates. We also compare the long read detected fusions with short read fusion calls inferred from chimeric junctions. To our surprise, short read detects only 5 fusion candidates in untreated and only 3 fusions in treated. All short read called fusions in untreated cells are found in IsoSeq fusion transcripts before being filtered by short read support. Since any short read fusion caller must effectively exclude low-complexity regions in the genome to constrain false positives [https://github.com/LosicLab/starchip] this dramatically reduces our power in exploring full set of fusions since over 49% of human genome is composed of repetitive sequences[29, 30]. We find that our final set of well supported IsoSeq fusions preferentially originates from repeat enriched region (76 out of 94, 52 out of 60 coming from repetitive region for TM and DMSO respectively, $p < .01$) further reinforcing a key advantage of long read technology to complement current short read technology.

Discussion

The role of alternative splicing in the transcriptomic landscape in general and more so as a post-treatment cellular response is just starting to be explored. Our understanding of the splicing complexity is benefitting from the advances in long-read technology, which is allowing for the identification of highly homologous isoforms routinely and with high fidelity. Our results strongly suggest that there is ample biological and clinical relevance for transcript sensitive modulation in prostate cancer treatment and, we speculate, in many other disease systems in cancer and beyond.

To our knowledge, our work is the first effort to characterize the significance of treatment-induced alternative splicing in a cancer model using a de-novo assembled isoform reference via long read sequencing. Our results suggest treatments induce a large number of varied alternative splicing events that alter known oncogenic proteins. Crucially, although overall gene expression is *higher* in TM compared to DMSO, our functional analysis shows that the splice variants in fact lead to *fewer* functional protein products in treatment. Indeed, we observed more intron-retention in untreated cells and that treated cells preferentially splice in/out oncogenic domains. For example, we find that ER stress marker *JUN* is down-regulated in treatment (log2 fold change = -3.5; FDR < 4e-8), however we also find that treatment specifically activates certain domains within JUN such as leucine zipper domain. This key domain facilitates DNA binding and participation in dimerization [21,22], and we suspect forms a central element in cellular response to the treatment. Finally, our fusion analysis uncovers a number of novel fusion candidates, including treatment specific examples, which are currently undergoing validation. We hope to generalize this method to more cell lines/treatments to investigate if the observed splicing signal observed is a global mechanism for anti-cancer treatment and shed light on drug resistance study. In summary, it seems plausible that an entirely new layer of transcriptomic regulation in cancer-drug interactions is finally becoming amenable to systematic study and enabling novel pathway and target discovery.

Methods

Materials

Our LNCap cell line is from clone FGC (ATCC CRL-1740). The cells were plated at a density of 8,000 cells per well in a 96well plate (Greiner Cat. No. 655083) and placed in an incubator. After 24 hours, the plates were removed from the incubator and treated with drugs using the HP D300 Digital Dispenser. The cells were then collected at the targeted time point (0hr for control; 6hrs for after TM combination treatment) by removing the media and pipetting 150uL of Qiagen Buffer RLT into each well. The plates were then frozen and stored at -80C.

RNA Preparation and Illumina RNA sequencing

The Qiagen RNeasy 96 kit (Cat. No. 74181) was used to extract RNA, with the Hamilton ML STAR liquid handling machine equipped with a Vacuubrand 96 well plate vacuum manifold. A Sorvall HT six floors centrifuge was used to follow the vacuum/spin version of the RNeasy 96 kit protocol. The samples were treated with DNAse (Rnase-Free Dnase Set Qiagen Cat. No.79254) during RNA isolation. The RNA samples were then tested for yield and quality with the Bioanalyzer and the Agilent RNA 6000 Pico Kit. The TruSeq Stranded mRNA Library Prep Kit

(RS-122-2101/RS-122-2102) was then used to prepare the samples for 30 million reads of single end sequencing (100bp) with the Illumina HiSeq2500.

Quantitative PCR validation

We used 15ng of poly(A)+ RNA for cDNA synthesis with the SuperScript IV first strand synthesis system (Thermo Fisher Scientific) and random hexamers as primers. Quantitative PCR was performed using the ABI Prism 7900HT with Bio-rad iQ SYBR Green Mastermix. The primers used for these studies are designed using primer3 and listed in table S3.

Short read data analysis

Raw reads from Illumina sequencing were aligned to a hg19 genome with features from GTF file (hg19) downloaded from UCSC genome browser using STAR[14] with chimSegmentMin=15; chimJunctionOverhangMin=15, outSAMmapqUnique=60 and other parameters as default. The output chimeric junctions are used for short read fusion caller STARCHIP with >10 junction reads support and repeat region penalty as 0.5 [https://github.com/LosicLab/starchip].

Short read integration for isoform quantification

Raw reads from Illumina sequencing were aligned to a hg19 genome but with a scaffold generated from both IsoSeq samples using STAR (2.5.2b)[14] with parameters described in table S5. IsoSeq isoforms were quantified using Cufflinks (2.2.1)[31]. Aligned short reads of each sample are assigned to features of IsoSeq scaffold to estimate transcript abundance. Cuffdiff [31] were called to compare expression between treated and untreated condition with default parameters, three replicates are fed to Cuffdiff to increase statistical power.

Fusion transcript detection from IsoSeq

Search criteria for fusion transcripts included mapping to at least two distinct genomic loci and each mapped locus has to cover >10% of the transcript, with all mapped loci combined covering >99% of the entire transcript. The mapped loci must be at least 100kbp apart from each other. We further filter the fusion transcripts based on short read support. For each candidate transcript, we aligned short reads against it using bwa mem (0.7.15)[32] and require >90% of the transcript to have at least 40 read supports at each base position for all three replicate samples since the number of fusions start to saturate from 40 read support as shown in Figure S4.

Functional annotation of IsoSeq transcripts

InterProScan(5.15-54.0)[33] was used to infer isoform functions. These includes two steps: 1) Screening for protein domains/ functional motifs through PRINT [34] and SMART[35] with default setting for all unique FL isoforms. 2) Obtained domains were then mapped to GO terms

through InterPro[36]. We compared and visualized GO annotations for baseline and treated specific isoforms using WEGO[37].

Estimates of information volume of splicing events

To estimate the cumulative information loss of averaging all distinct transcripts (i.e. ignoring splicing) in a given experiment, it is straightforward to remove the correction due to locus length in Eq. (3) and sum up the sample-median loci entropies across loci to obtain a naive upper limit estimate of order **kilobit**. Simply summing Eq. (3) across all genomic loci instead reduces this by an order of magnitude to $O(100)\ bits$, and any co-splicing will also obviously reduce this estimate by reducing the number of independent splicing events. In the present work we see strong evidence of coherent splicing patterns but cannot estimate co-splicing with this number of samples. By comparison, however, gene *expression* states are often well approximated by a three-state system ($S \sim 1.09\ bits$) and we verify that this is the case of our data as well across samples. Thus, a naive upper limit estimate is such that $S_{splicing} \sim 10^2\, S_{expression}$. In the absence of a significant co-splicing effect this implies that splicing dynamics contain significantly more information than expression dynamics from the point of view of bulk RNA-seq.

Supplementary

All supplementary materials are hosted on https://chenx08.u.hpc.mssm.edu/PSB_CHEN/

Acknowledgements

We thank Ronald B. Realubit and Charles Karan from Columbia Genome Center for providing LNCap cells and performing Illumina sequencing.

References

1. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* Int J Cancer, 2015. **136**(5): p. E359-86.
2. Cancer Genome Atlas Research, N., *The Molecular Taxonomy of Primary Prostate Cancer.* Cell, 2015. **163**(4): p. 1011-25.
3. Costello, J.C., et al., *A community effort to assess and improve drug sensitivity prediction algorithms.* Nat Biotechnol, 2014. **32**(12): p. 1202-12.
4. Bansal, M., et al., *A community computational challenge to predict the activity of pairs of compounds.* Nat Biotechnol, 2014. **32**(12): p. 1213-22.
5. Weirather, J.L., et al., *Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing.* Nucleic Acids Res, 2015. **43**(18): p. e116.
6. Au, K.F., et al., *Characterization of the human ESC transcriptome by hybrid sequencing.* Proc Natl Acad Sci U S A, 2013. **110**(50): p. E4821-30.
7. Treutlein, B., et al., *Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing.* Proc Natl Acad Sci U S A, 2014. **111**(13): p. E1291-9.
8. Wang, B., et al., *Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing.* Nat Commun, 2016. **7**: p. 11708.
9. Abdel-Ghany, S.E., et al., *A survey of the sorghum transcriptome using single-molecule long reads.* Nat Commun, 2016. **7**: p. 11706.

10. Singh, N., et al., *IsoSeq analysis and functional annotation of the infratentorial ependymoma tumor tissue on PacBio RSII platform.* Meta Gene, 2016. **7**: p. 70-5.
11. Sharon, D., et al., *A single-molecule long-read survey of the human transcriptome.* Nat Biotechnol, 2013. **31**(11): p. 1009-14.
12. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.
13. Wu, T.D. and C.K. Watanabe, *GMAP: a genomic mapping and alignment program for mRNA and EST sequences.* Bioinformatics, 2005. **21**(9): p. 1859-75.
14. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
15. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
16. Foissac, S. and M. Sammeth, *ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W297-9.
17. Eddy, S.R., *A new generation of homology search tools based on probabilistic inference.* Genome Inform, 2009. **23**(1): p. 205-11.
18. Sturges, H.A., *The choice of a class interval Case I Computations involving a single Series.* Journal of the American Statistical Association, 1926. **21**: p. 65-66.
19. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.
20. Ueno, K., et al., *Frizzled homolog proteins, microRNAs and Wnt signaling in cancer.* Int J Cancer, 2013. **132**(8): p. 1731-40.
21. Abbas, A. and S. Gupta, *The role of histone deacetylases in prostate cancer.* Epigenetics, 2008. **3**(6): p. 300-9.
22. Halkidou, K., et al., *Upregulation and nuclear recruitment of HDAC1 in hormone refractory prostate cancer.* Prostate, 2004. **59**(2): p. 177-89.
23. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Res, 2012. **40**(Database issue): p. D130-5.
24. Ransone, L.J., et al., *Fos-Jun interaction: mutational analysis of the leucine zipper domain of both proteins.* Genes Dev, 1989. **3**(6): p. 770-81.
25. Kouzarides, T. and E. Ziff, *The role of the leucine zipper in the fos-jun interaction.* Nature, 1988. **336**(6200): p. 646-51.
26. McPherson, A., et al., *Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data.* Bioinformatics, 2011. **27**(11): p. 1481-8.
27. Maher, C.A., et al., *Chimeric transcript discovery by paired-end transcriptome sequencing.* Proc Natl Acad Sci U S A, 2009. **106**(30): p. 12353-8.
28. Maher, C.A., et al., *Transcriptome sequencing to detect gene fusions in cancer.* Nature, 2009. **458**(7234): p. 97-101.
29. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.
30. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
31. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
32. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
33. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.* Bioinformatics, 2014. **30**(9): p. 1236-40.
34. Attwood, T.K., et al., *The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012.* Database (Oxford), 2012. **2012**: p. bas019.
35. Schultz, J., et al., *SMART: a web-based tool for the study of genetically mobile domains.* Nucleic Acids Res, 2000. **28**(1): p. 231-4.
36. Finn, R.D., et al., *InterPro in 2017-beyond protein family and domain annotations.* Nucleic Acids Res, 2017. **45**(D1): p. D190-D199.
37. Ye, J., et al., *WEGO: a web tool for plotting GO annotations.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W293-7.