RNASeq

CENG 734

Advanced Topics in Bioinformatics

Fall 2011-2012

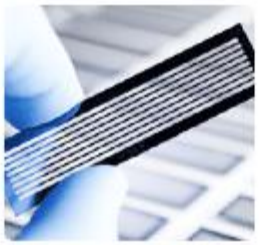
NGS

- Cheaper and massively parallel sequencing

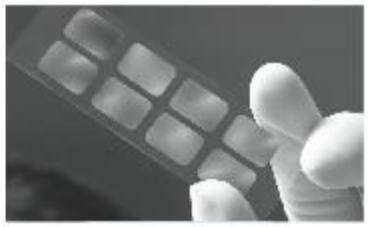
Capillary electrophoresis (Sanger)
Between 96 and 384 samples
(76 - 308 Kb/run)

454
400,000 samples
(120Mb / run)

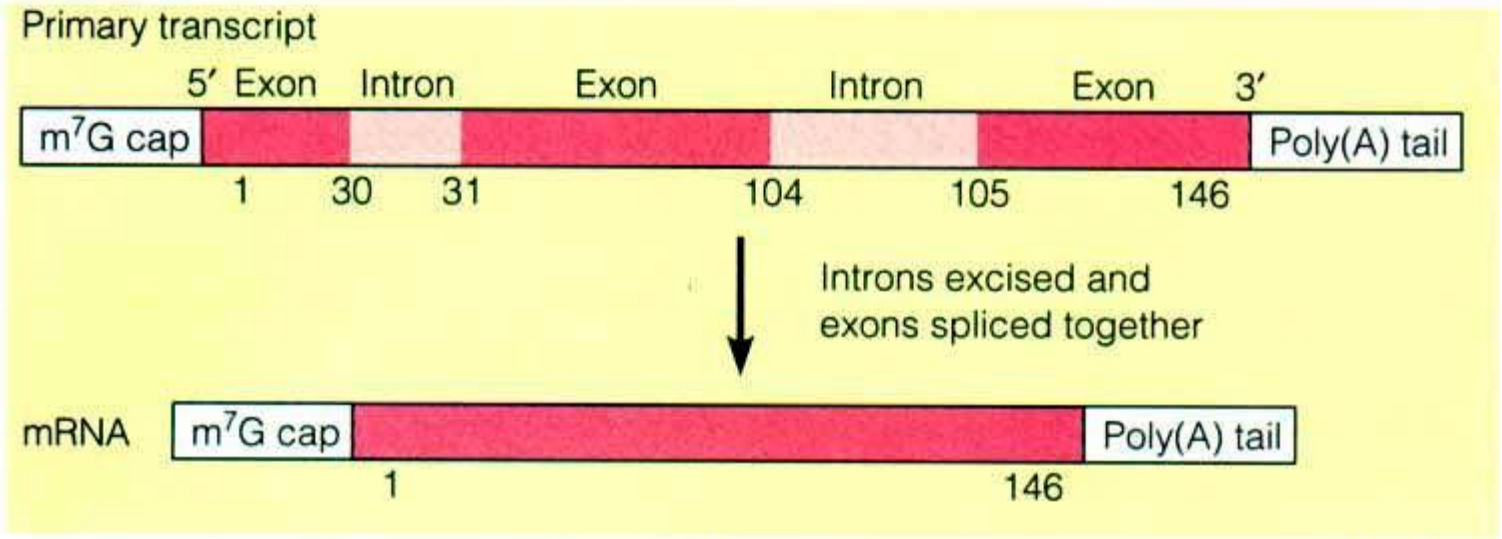
Solexa
32M samples
(2Gb / run)



Solid
40M samples
(2-6 Gb / run)

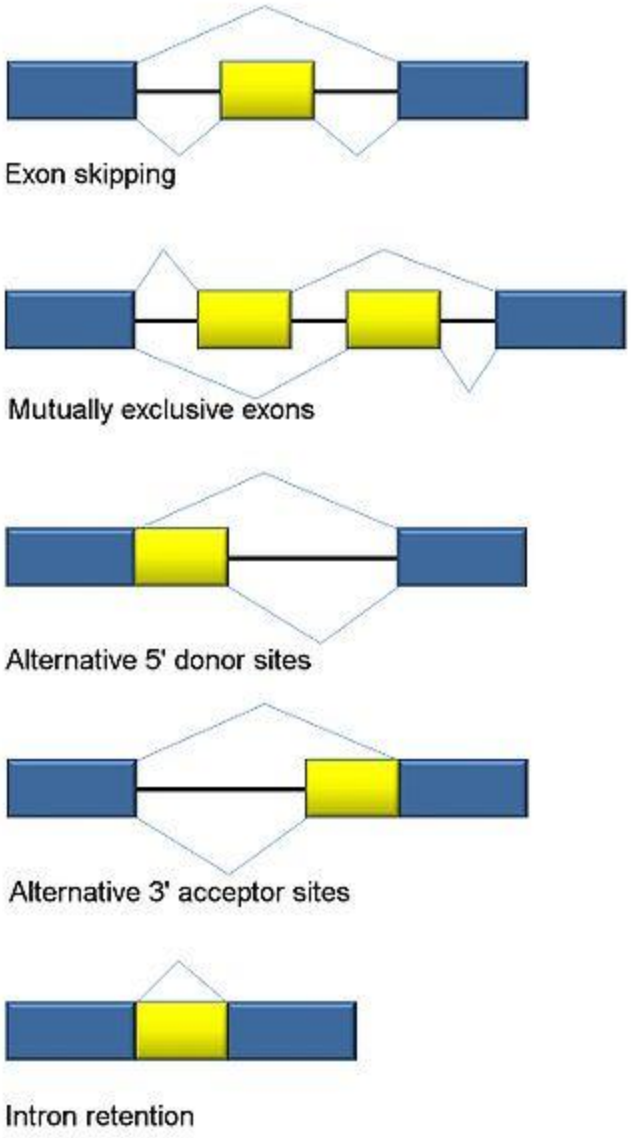


Gene Structure

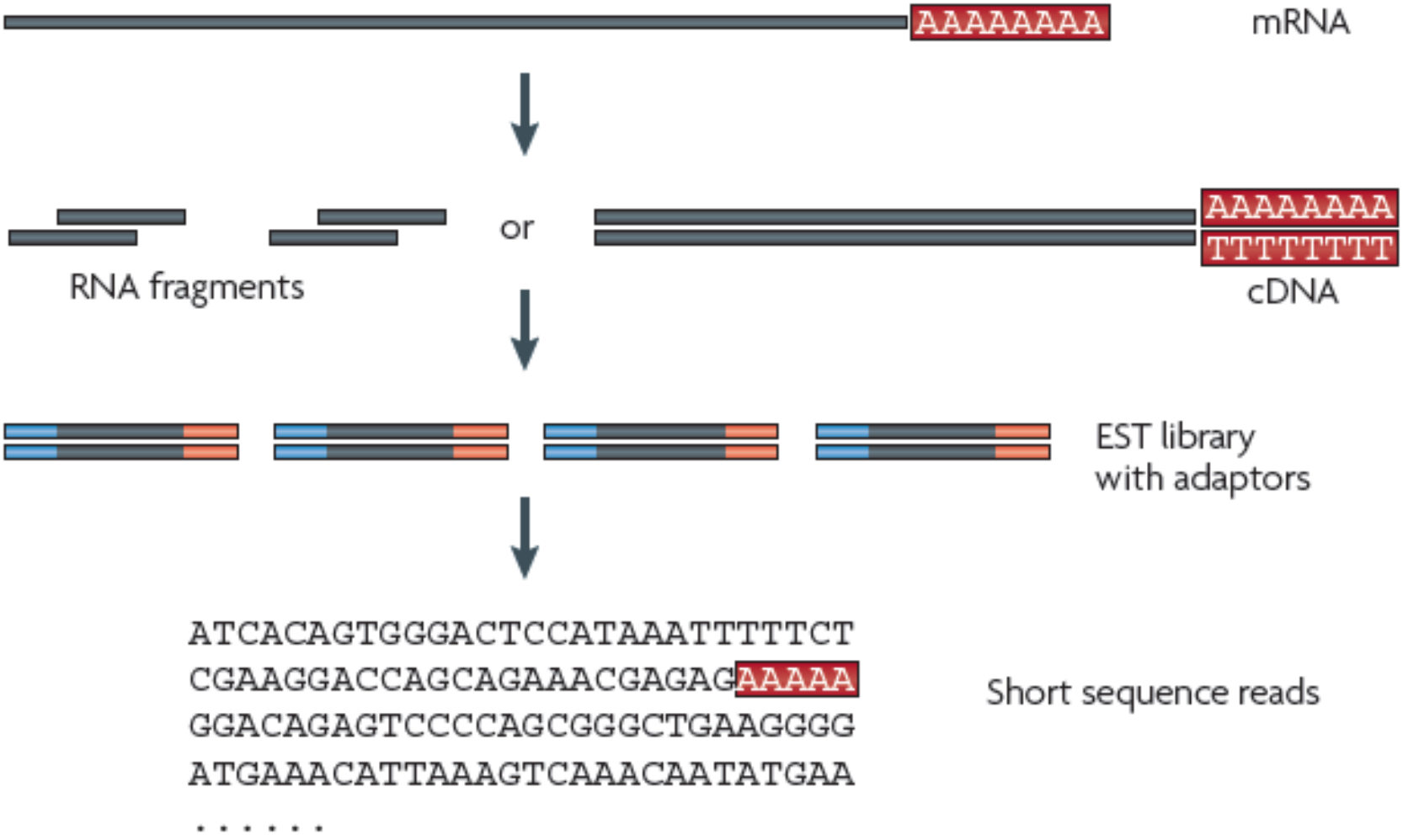


Isoforms

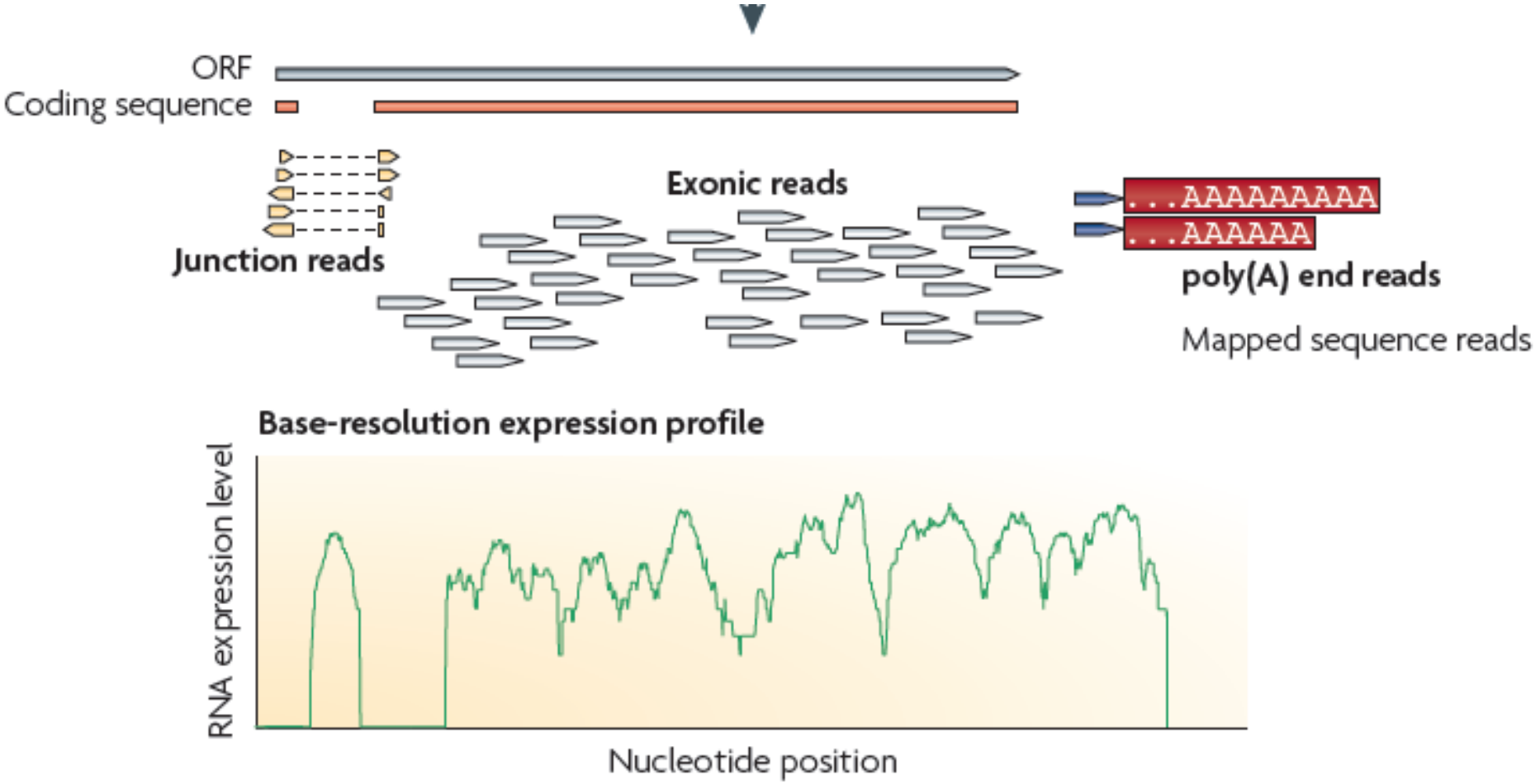
- Alternative mRNAs for the same gene
  - alternative 5’ (or 3’) splice sites
  - exon skipping
  - intron retention
  - mutually exclusive exons



RNASeq


AAAAAAAA
mRNA
or
AAAAAAAA
TTTTTTTT
cDNA
RNA fragments
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
.....
Short sequence reads

RNASeq


ORF
Coding sequence
Junction reads
Exonic reads
...AAAAAAAAA
...AAAAAA
poly(A) end reads
Mapped sequence reads
Base-resolution expression profile
RNA expression level
Nucleotide position

IsoLasso

- Given reads (single or paired-end reads) from an RNASeq experiment and a reference genome
  - Map reads to the genome
  - Infer isoforms
  - Infer their expression levels
- The predicted expression levels should be as close as to the number of reads (accuracy)
- The observed reads should be explainable by as few isoforms as possible (minimal interpretation)
- All the observed reads should be part of a predicted isoform (completeness)

Definitions

A gene sequence S of length n is an ordered character sequence $S = S_1 S_2 \cdots S_n$, $S_i \in \{A, T, G, C\}$. Define $B(n)$ as the set of binary vectors of length n. For a vector $b \in B(n)$, b_i indicates the ith element of vector b. For a subset $U \subset B(n)$, define $OR(U) = b \in B(n)$ with $b_i = 1$ iff there is an element $c \in U$ such that $c_i = 1$. For a binary vector $b \in B(n)$, define the start (or end) of b as the first (or last) non-zero index of b, and is denoted as $l(b)$ (or $u(b)$). Hence, each isoform on gene S could be represented as a binary vector $b \in B(n)$ with $b_i = 1$ iff the nucleotide S_i is included in this isoform. A single-end or paired-end read mapped to S could also be represented as an element $b \in B(n)$ with $b_i = 1$ iff this read contains S_i. A paired-end read is denoted as $p = (b^1, b^2)$, where b^1 and b^2 are the two mapped single-end reads, and $l(b^1) < l(b^2)$. Given a set of single-end or paired-end reads R, the coverage of S_i, or $cvg(S_i)$, is the number of reads b with $b_i = 1$.

Definitions

A single-end read b is *compatible* with an isoform t, denoted as $b \sim t$, iff $b_i = t_i$ for $l(b) \leq i \leq u(b)$. Similarly, a paired-end read $p = (b^1, b^2)$ is compatible with isoform t, denoted as $p \sim t$, iff $b^1 \sim t$ and $b^2 \sim t$. Given a set of single-end (or paired-end) reads R mapped to gene S, the *connectivity graph (CG)* [17] is a directed acyclic graph (DAG) $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ and $e = (v_i, v_j) \in E$ iff one of the following conditions is true:

Condition 1. There exists a single-end read or an end of some paired-end read $b \in R$ such that $b_i = 1$, $b_j = 1$, and $b_k = 0$, $\forall i < k < j$;

Condition 2. $cvg(S_i) > 0$, $cvg(S_j) > 0$, and $cvg(S_k) = 0$, $\forall i < k < j$.

Note that Condition 2 is designed to connect two mapped reads separated by a coverage gap. Based on the definition of CG, a path h in the CG could be readily treated as an isoform by defining the isoform t as $t_i = 1$ iff $v_i \in h$. Therefore, a read b is compatible with h (denoted as $b \sim h$) iff $b \sim t$. The isoform enumeration algorithm depicted in Algorithm 1 takes the connectivity graph as the input, and outputs a set of maximal candidate isoforms T. The algorithm consists of three phases, Enumeration, Filtration and Condensation. In the Enumeration phase, all maximal paths in the connectivity graph are enumerated. However, some of these isoforms are "infeasible" in the sense that they cannot be assembled from the mapped reads (see Figure 1 (right) for an example). In this case, the second phase (*i.e.*, the Filtration phase) is required to remove such isoforms. For each isoform t generated in the Enumeration phase, the Filtration phase first finds all reads that are compatible with t, and then checks if t can be assembled from these compatible reads (it replaces t otherwise). Finally, the Condensation phase removes all the isoforms that are not maximal candidates.

Isoform Enumeration

input : A CG $G = (V, E)$, and a set of mapped single-end or paired-end reads R
output: A set of isoforms T
begin
    Enumeration:
    $T \leftarrow \emptyset$
    for $v_j \in V$ *with* $indeg(v_j) = 0$ **do**
        Enumerate all possible maximal paths P that begin at v_j and end at some v_k with $outdeg(v_k) = 0$
        $T \leftarrow T \cup P$
    Filtration:
    for $t \in T$ **do**
        Let $t' = OR(\{b \in R | b \sim t\})$
        $T \leftarrow (T \backslash \{t\}) \cup \{t'\}$
    Condensation:
    for $t \in T$ **do**
        Let $R_t = \{b \in R|, b \sim t\}$
        for $t' \in T \backslash \{t\}$ **do**
            Let $R_{t'} = \{b \in R|, b \sim t'\}$
            if $R_t \subset R_{t'}$ **then**
                $T \leftarrow (T \backslash \{t\})$
end

Least Squares Regression

Multiple regression estimates the outcomes (dependent variables) which may be affected by more than one control parameter (independent variables) or there may be more than one control parameter being changed at the same time.

An example is the two independent variables x and y and one dependent variable z in the linear relationship case:

$$z = a + bx + cy$$

For a given data set (x_1, y_1, z_1), (x_2, y_2, z_2), ..., (x_n, y_n, z_n), where $n \geq 3$, the best fitting curve $f(x)$ has the least square error, i.e.,

$$\Pi = \sum_{i=1}^{n} [z_i - f(x_i, y_i)]^2 = \sum_{i=1}^{n} [z_i - (a + bx_i + cy_i)]^2 = \min.$$

Please note that a, b, and c are unknown coefficients while all x_i, y_i, and z_i are given. To obtain the least square error, the unknown coefficients a, b, and c must yield zero first derivatives.

LASSO

- Is a regularization technique if there are multiple solutions possible which yield a small squared error.
- LASSO constrains the L_1 norm of the inferred coefficient vector to be less then a defined threshold.

Mathematical Model of RNASeq

- The “segment” model
  - A gene is divided into a set of M segments where a segment is a continuous region uninterrupted by exon-intron boundaries
  - An isoform can be represented by a subset of segments
  - Hence given N isoforms, we can construct an NxM binary matrix to represent the isoform.

Mathematics Model of RNASeq

- The expected number of reads falling into the i^{th} segment is proportional to the length of the segment and the sum of all expression levels of all isoforms containing the i^{th} segment.

$$r_i = l_i \sum_{j=1}^{N} a_{ji} x_j$$

Least squares formulation

$$X^* = \underset{X}{\operatorname{argmin}} f(X) = \sum_{i=1}^{M} \left(\frac{r_i}{l_i} - \sum_{j=1}^{N} a_{ji} x_j \right)^2$$

LASSO approach

$$f(X) = \sum_{i=1}^{M} (\frac{r_i}{l_i} - \sum_{j=1}^{N} a_{ji} x_j)^2 + \lambda \sum_{j=1}^{N} |x_j|$$

QP Instance

$$\min f(X) = \sum_{i=1}^{M} \left(\frac{r_i}{l_i} - \sum_{j=1}^{N} a_{ji} x_j \right)^2$$

$$s.t. \qquad x_j \geq 0, \ 1 \leq j \leq N$$

$$\sum_{j=1}^{N} x_j \leq \gamma$$

How much to constrain?

$$i^* = \underset{1 \leq i \leq k}{\operatorname{argmin}} \{L_i : e_i \leq \beta * \min\{e_1, \ldots e_k\}\}$$

Completeness constraint

$$\min f(X) = \sum_{i=1}^{M} \left(\frac{r_i}{l_i} - \sum_{j=1}^{N} a_{ji} x_j\right)^2$$

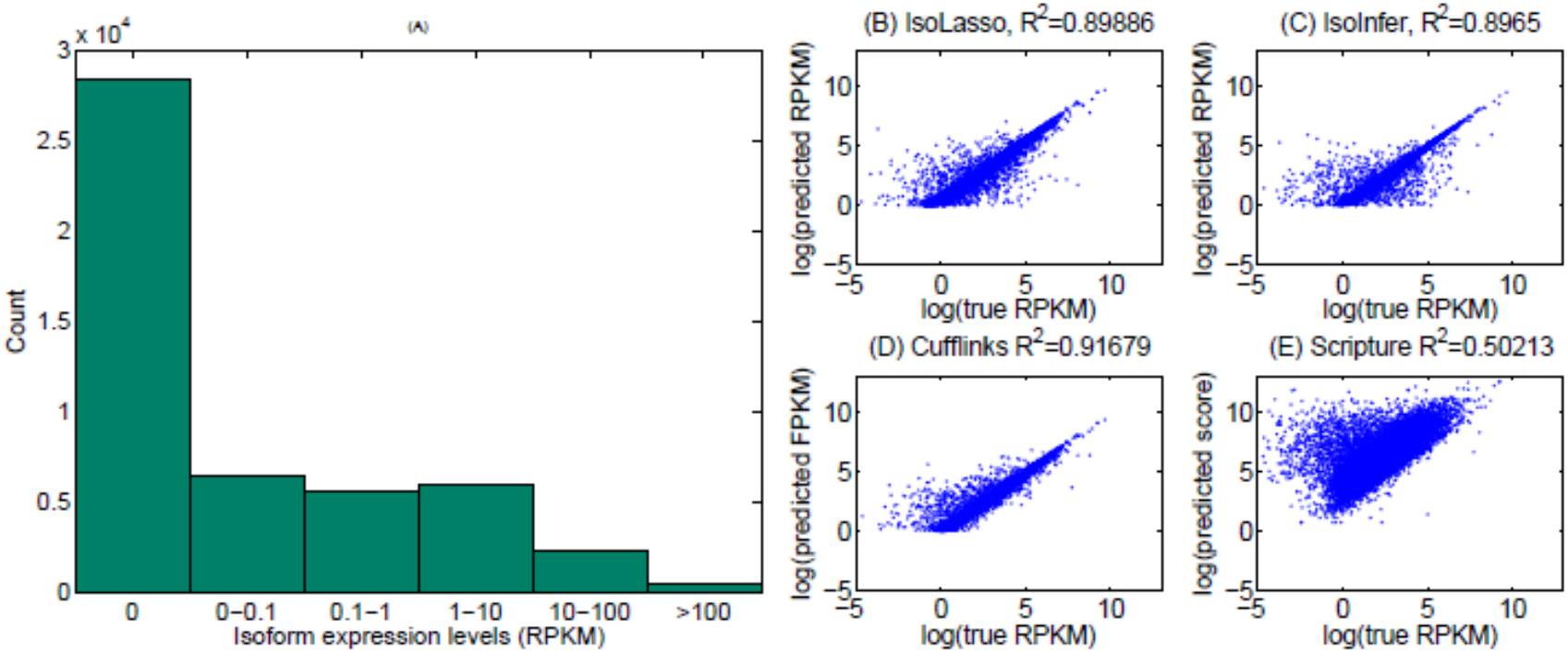
$$s.t. \qquad x_j \geq 0,\ 1 \leq j \leq N$$

$$\sum_{j=1}^{N} x_j \leq \lambda$$

$$\sum_{j=1}^{N} x_j a_{ji} \geq p, \text{ if segment } i \text{ has mapped reads}$$

$$\sum_{j=1}^{N} x_j a_{ji} a_{jk} \prod_{h=i+1}^{k-1} (1 - a_{jh}) \geq p, \text{ if the junction between segments } i \text{ and } k \text{ contains mapped reads}$$

Results


x 10^4
(A)
Count
Isoform expression levels (RPKM)
0
0–0.1
0.1–1
1–10
10–100
>100
(B) IsoLasso, R^2=0.89886
(C) IsoInfer, R^2=0.8965
(D) Cufflinks R^2=0.91679
(E) Scripture R^2=0.50213
log(predicted RPKM)
log(predicted FPKM)
log(predicted score)
log(true RPKM)

Results

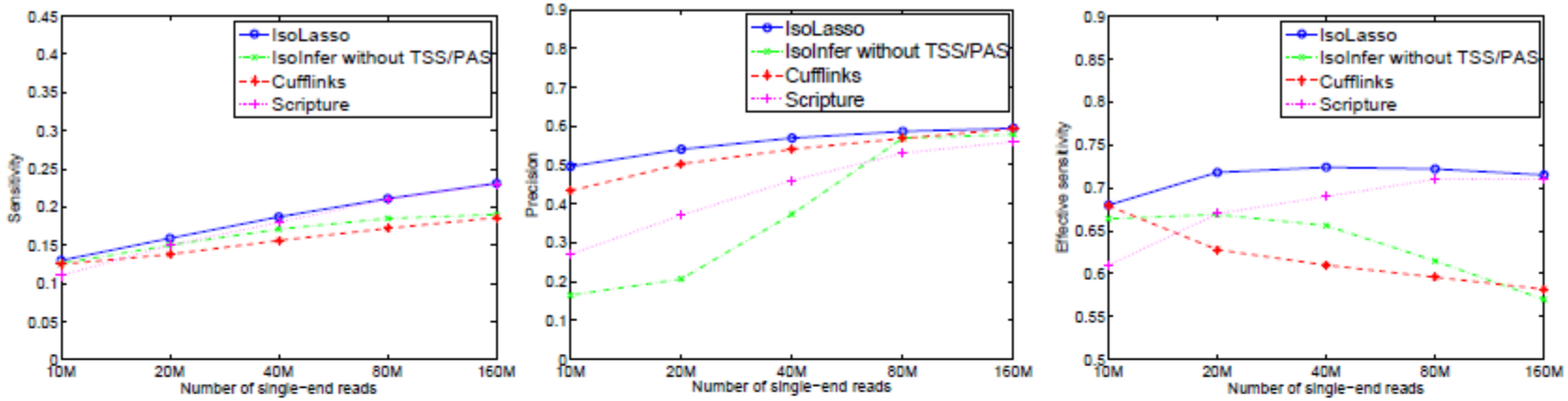


Fig. 3. Sensitivity (left), precision (middle) and effective sensitivity (right) on single-end reads.

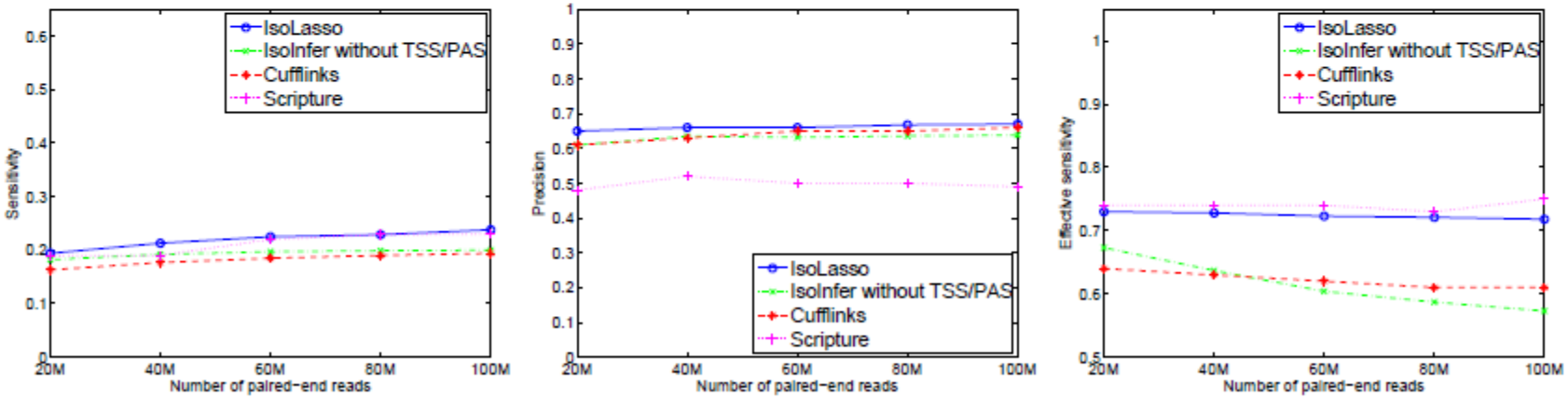
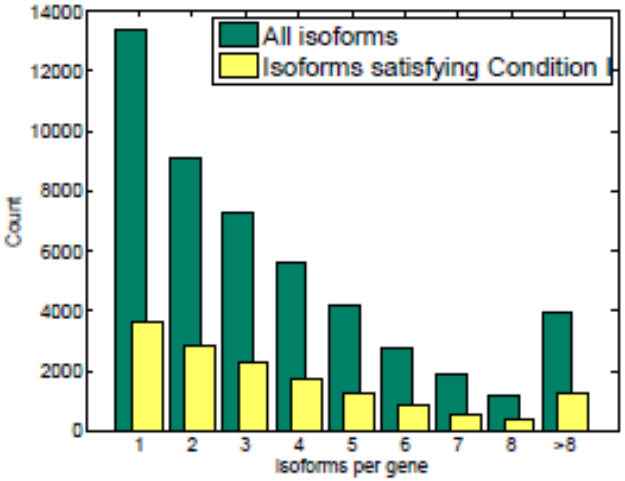
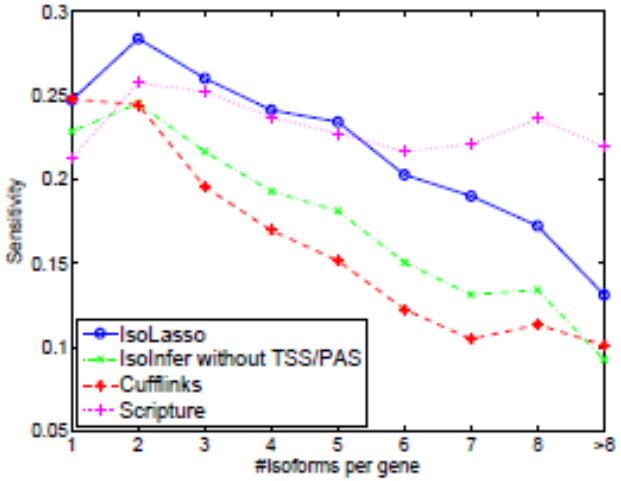
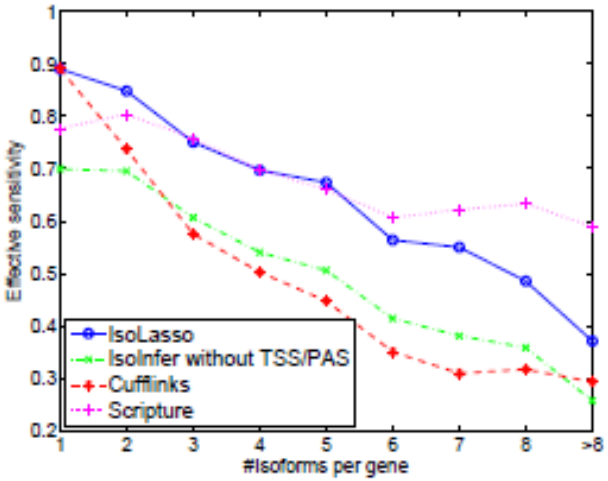


Fig. 4. Sensitivity (left), precision (middle) and effective sensitivity (right) on paired-end reads.

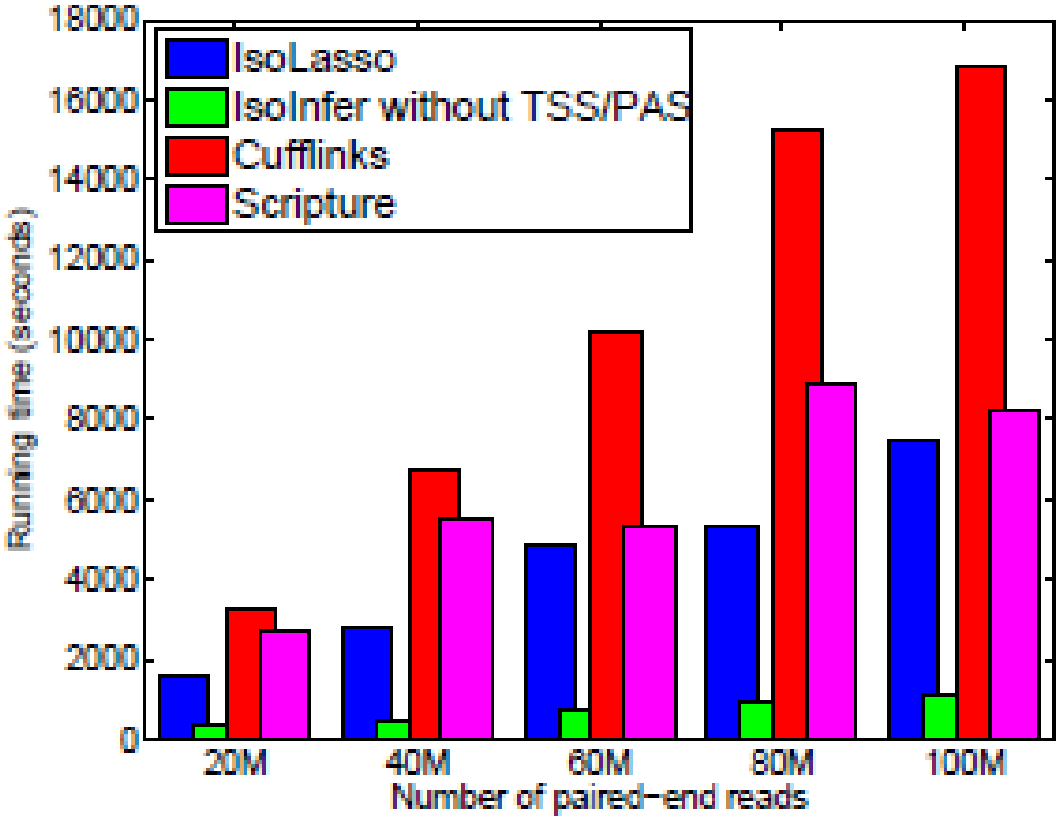
Results


All isoforms
Isoforms satisfying Condition I
Count
Isoforms per gene

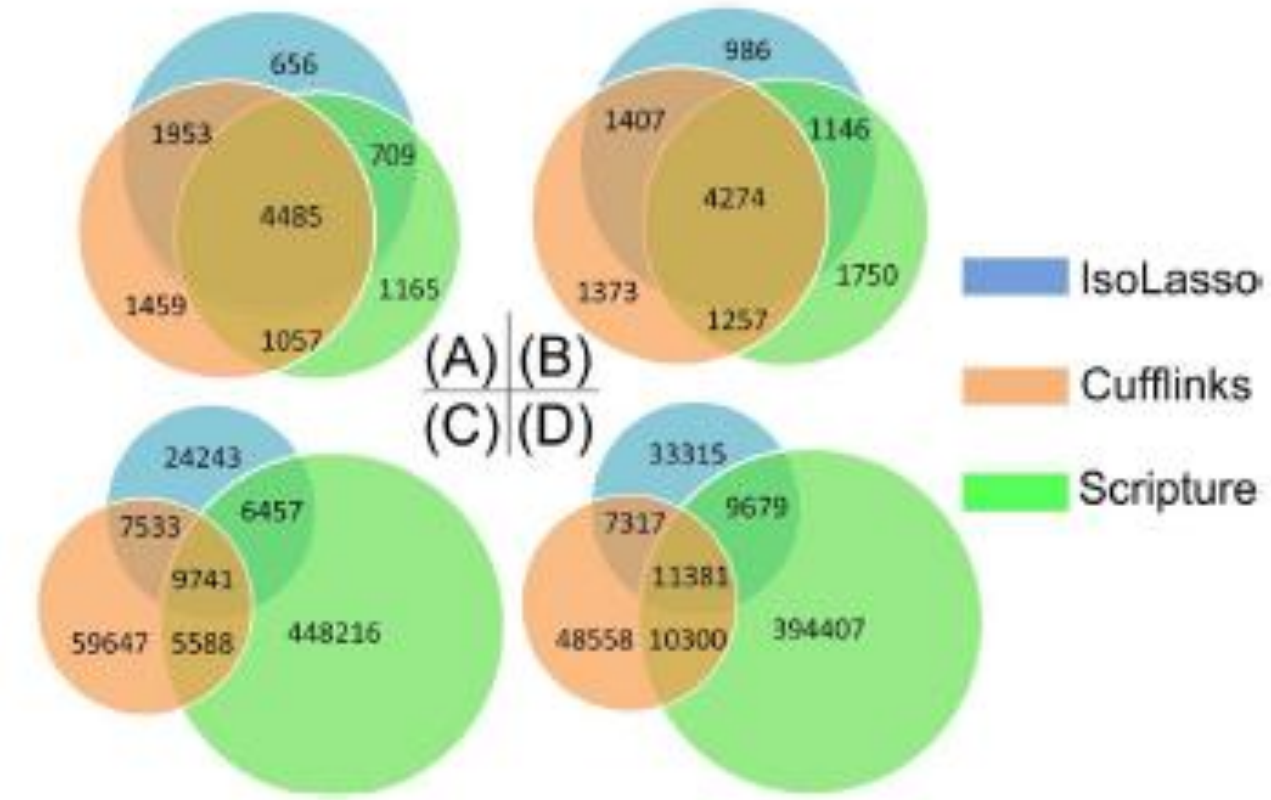

IsoLasso
IsoInfer without TSS/PAS
Cufflinks
Scripture
Sensitivity
#Isoforms per gene


IsoLasso
IsoInfer without TSS/PAS
Cufflinks
Scripture
Effective sensitivity
#Isoforms per gene

Results


IsoLasso
IsoInfer without TSS/PAS
Cufflinks
Scripture
Running time (seconds)
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20M
40M
60M
80M
100M
Number of paired−end reads

Diversity of Predicted Isoforms



Next week

- SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments by Ourfali et al. ISMB/ECCB 2007