Transcriptome analysis using Hybrid-Seq

Kin Fai Au

Department of Internal Medicine and Biostatistics, University of Iowa

THE UNIVERSITY OF IOWA
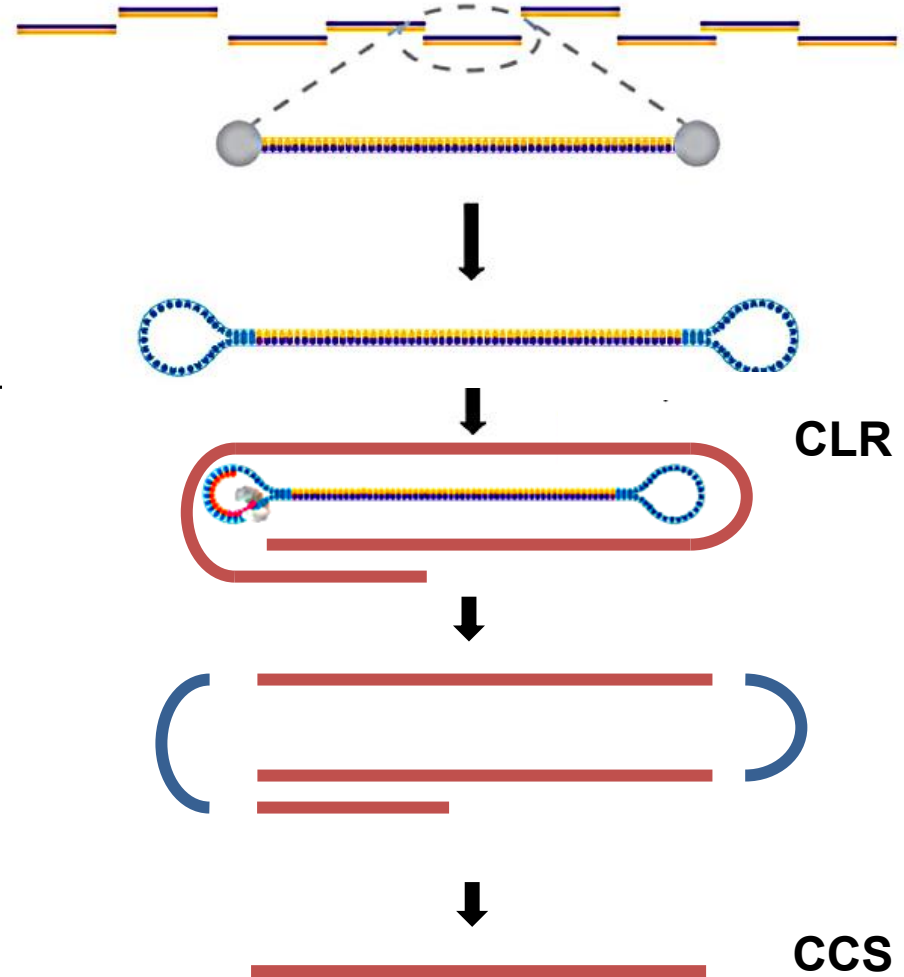
Accuracy VS read length

accuracy

# of passes

$$\text{\# of passes} = \frac{\text{Total length of CLR}}{\text{CCS read length}}$$

Movie time is fixed
(life time of polymerase is limited)
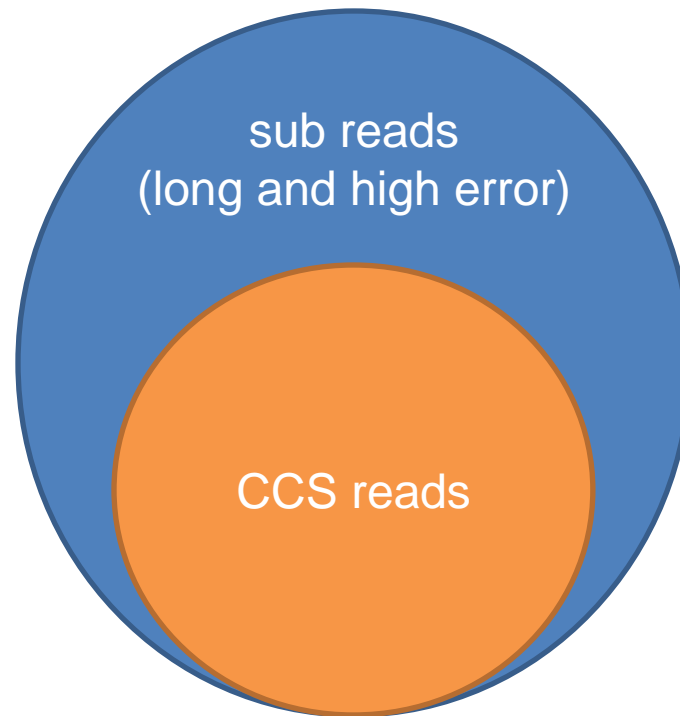
Total length of CLR is fixed

CCS read length

CLR

CCS

accuracy or **length**

Linkage
information

Data

sub reads
(long and high error)

CCS reads

Hybrid sequencing (Hybrid-Seq)

Second Generation Sequencing ➡ **accuracy and throughput**

Third Generation Sequencing ➡ **length**
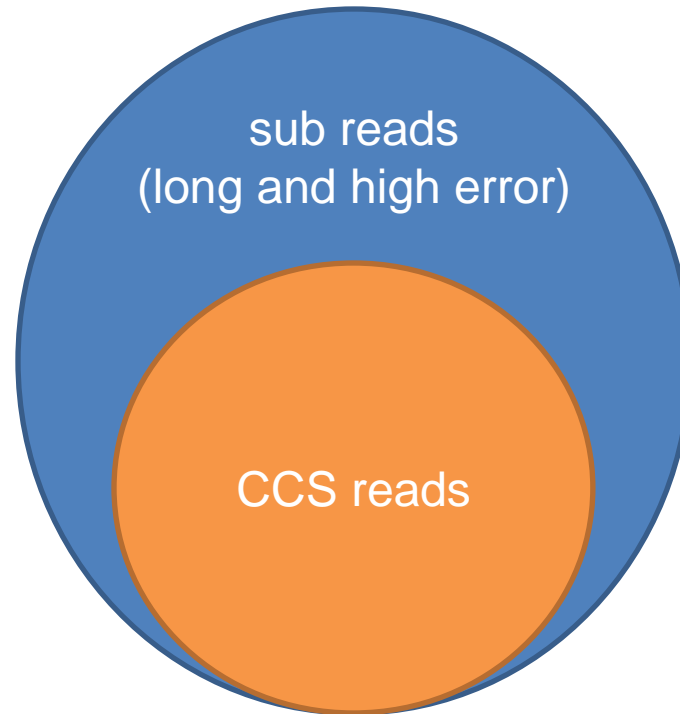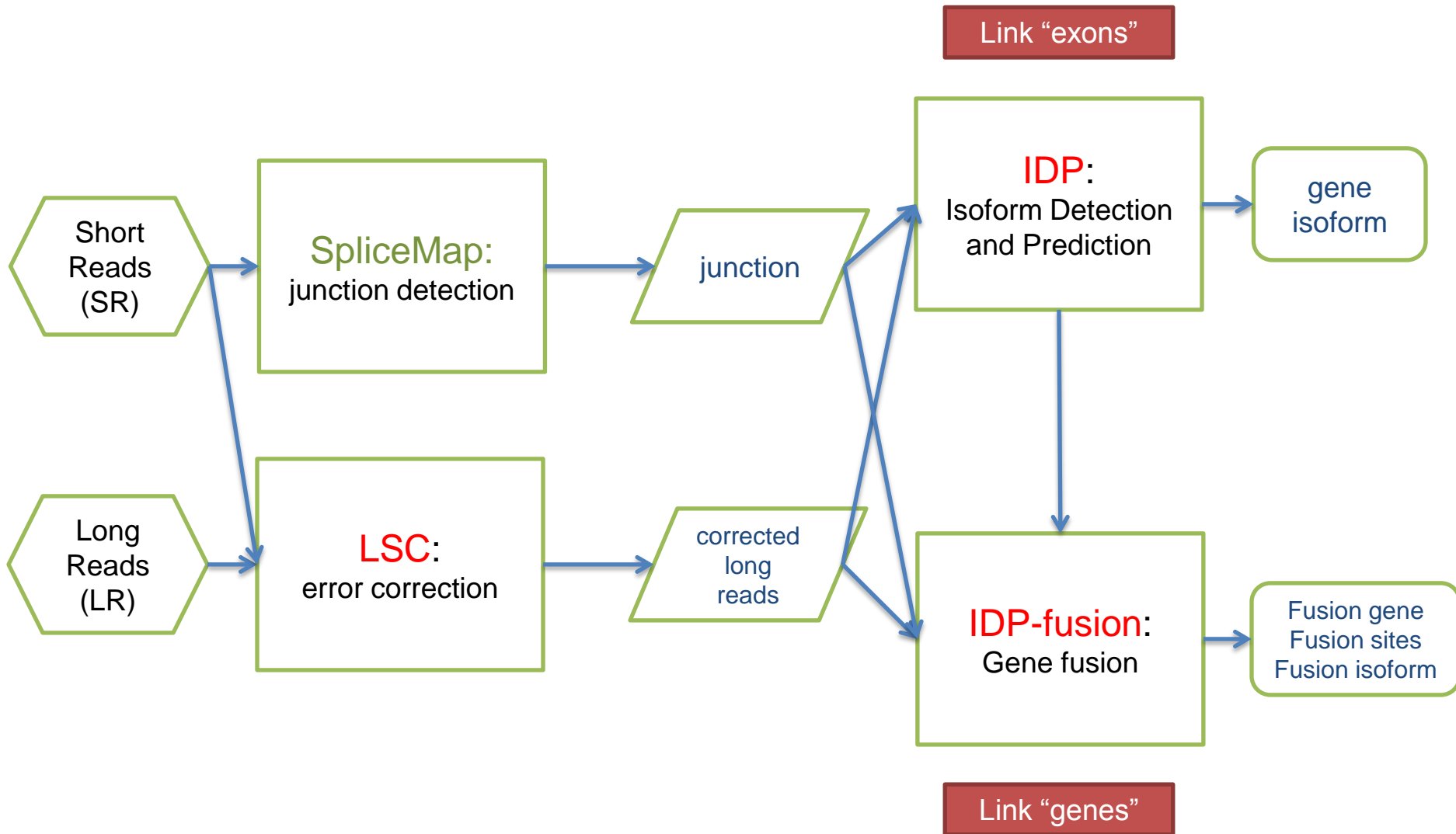
Hybrid-Seq

Cost

Short reads

sub reads
(long and high error)

CCS reads
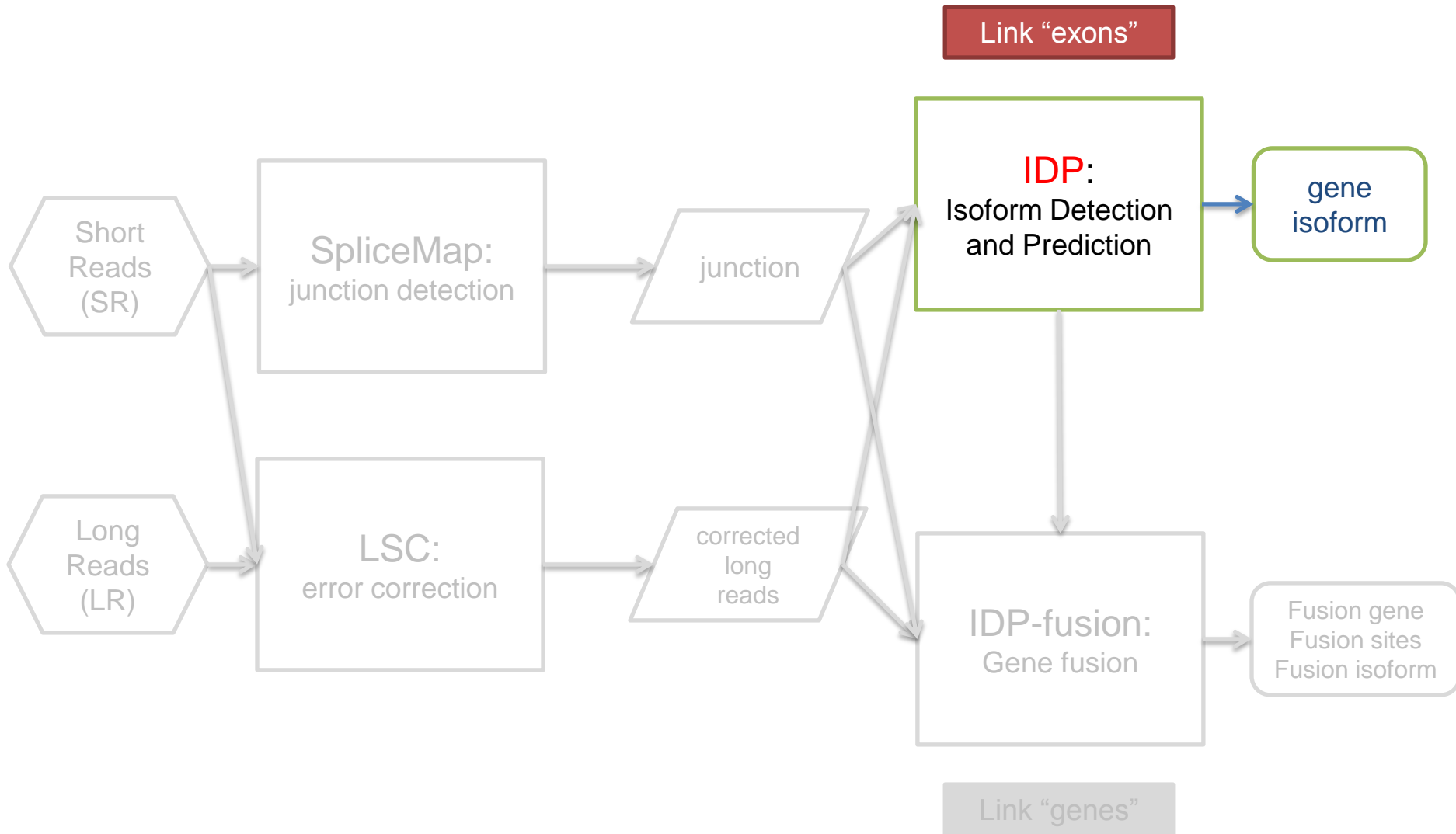
Hybrid-Seq pipeline

Hybrid-Seq pipeline

Hybrid-Seq pipeline

ROC

FPR = 0.05   FPR = 0.1

- IDP
- Cufflinks

Sensitivity

False Positive Rate (FPR)

Link "exons"

IDP:
Isoform Detection and Prediction

gene isoform

Identification rate VS Gene length

Number of genes (%)

— Long reads only
— IDP using long reads and short reads

gene length (bp)

nonredundant long reads

junctions detection from short reads

Junction 1   Junction 2   Junction 3   Junction 4   Junction 5   Junction 6   Junction 7

5' end   **3' end**

Isoform candidate construction

Fill the junctions in between long reads and 5'end/3'end
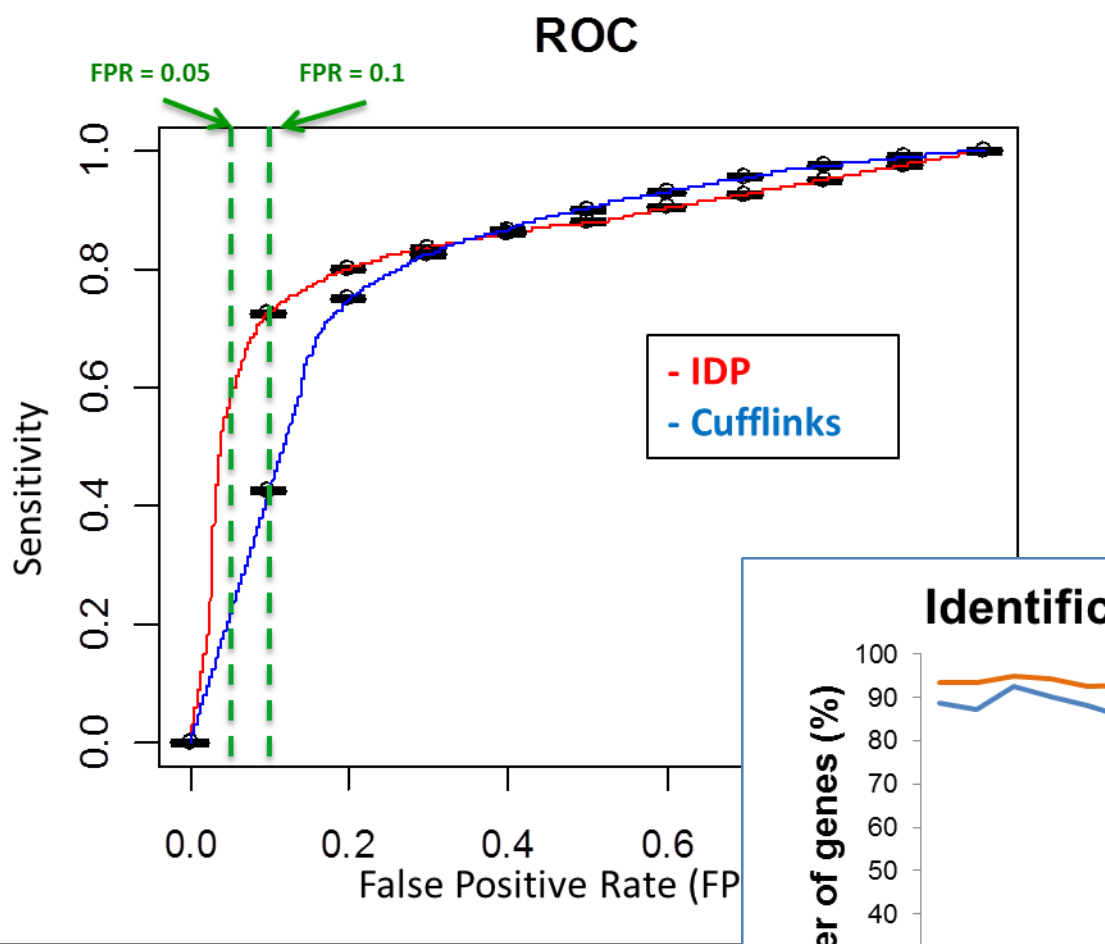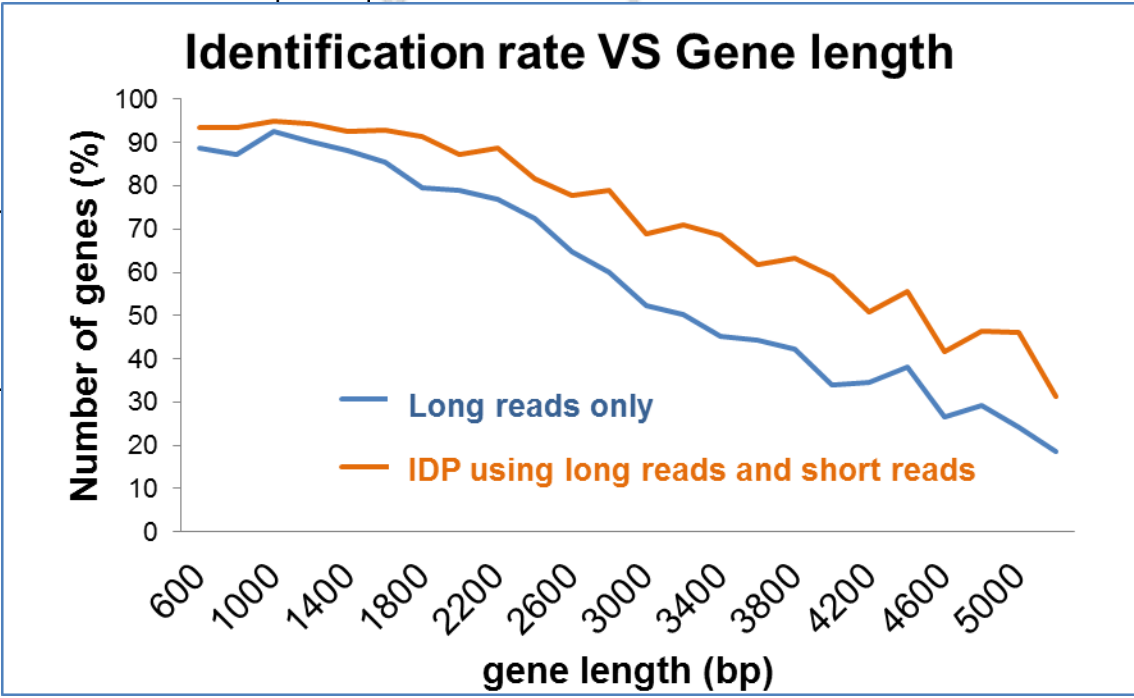
Hybrid-Seq pipeline
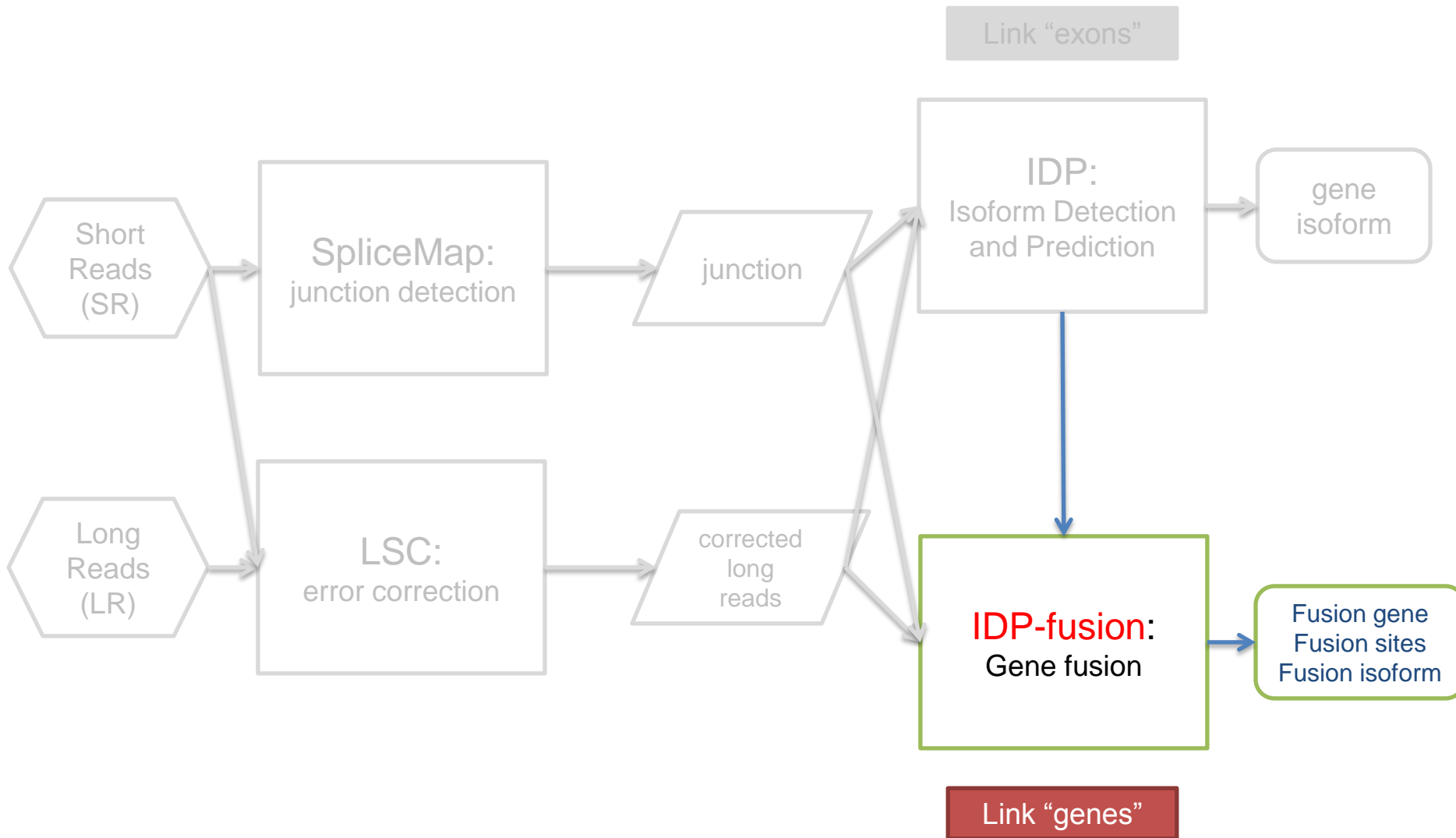
The default settings are label by "X"

Hybrid-Seq pipeline

Specificity of fusion gene detections

Link "exons"

IDP:
Isoform Detection
and Prediction
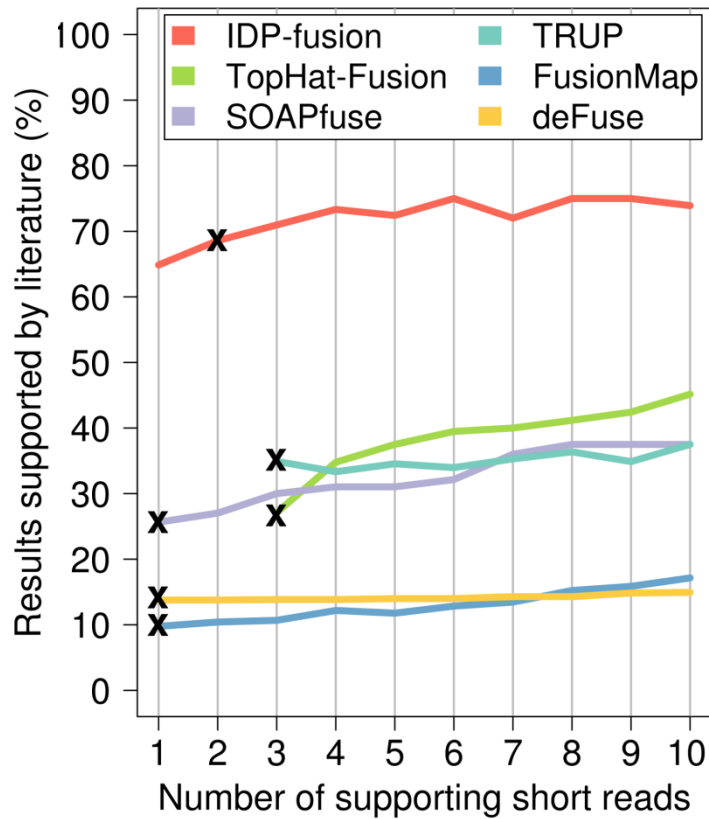
gene
isoform

Short
Reads
(SR)

Long
Reads
(LR)

IDP-fusion:
Gene fusion

Fusion gene
Fusion sites
Fusion isoform

Link "genes"
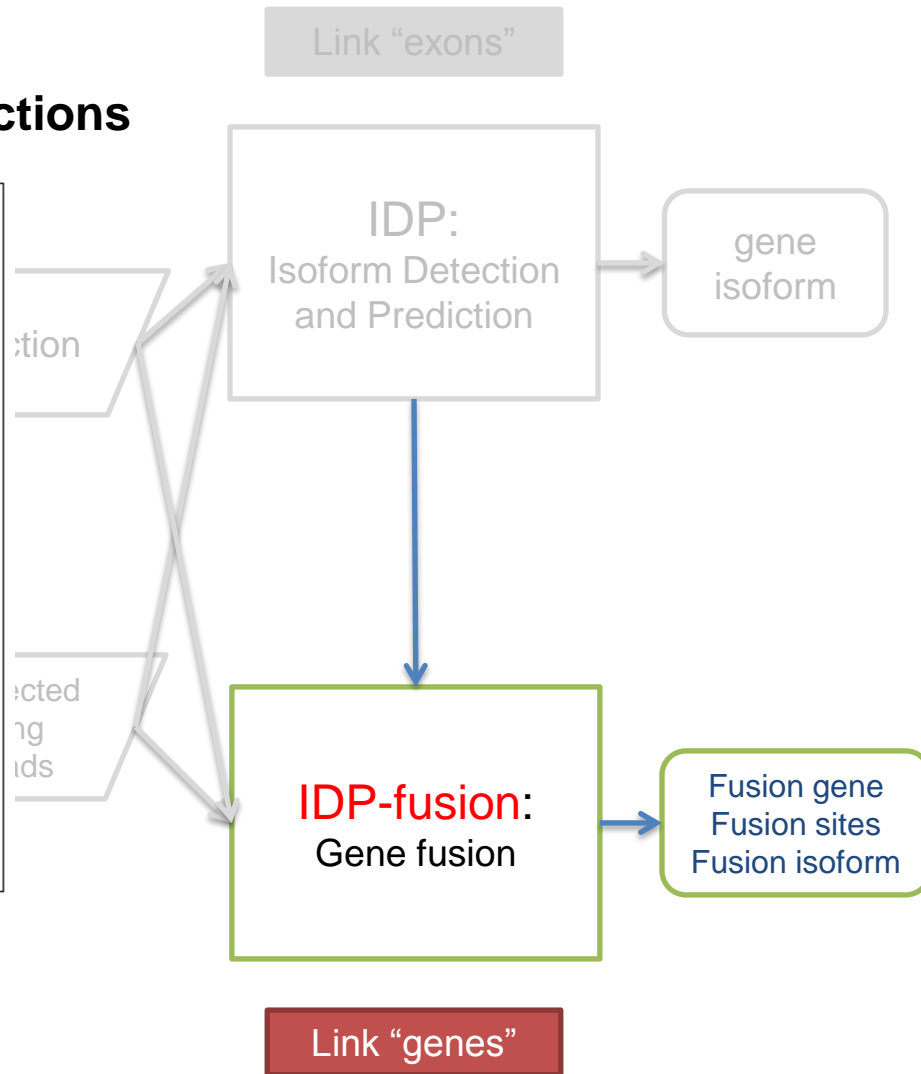
The default settings are label by "X"

Fusion gene detection by short reads

Gene A

Gene B

short reads (50-250bp)
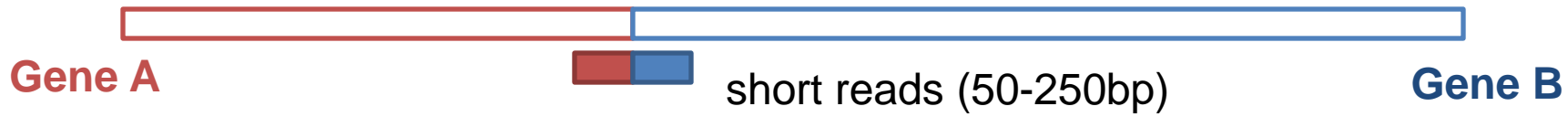
Too many pairs
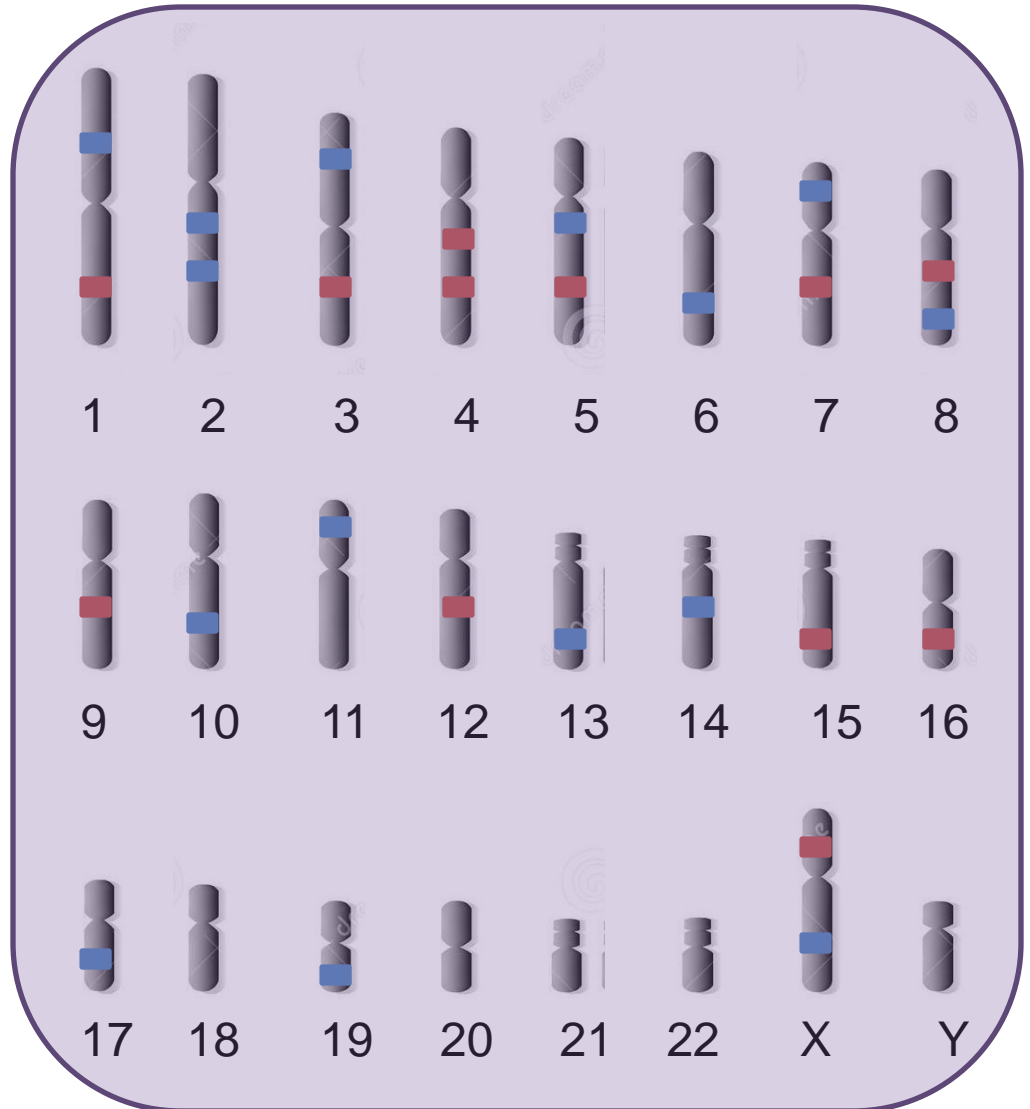
* multiple alignments

* mis-alignments

* intense computing

Fusion gene detection by long reads

Gene A

Gene B

long reads (e.g. 2000bp)

* multiple alignments
    -> unique alignment

* mis-alignments
    -> accurate alignment

* intense computing
    -> small search space

1 2 3 4 5 6 7 8

9 10 11 12 13 14 15 16

17 18 19 20 21 22 X Y
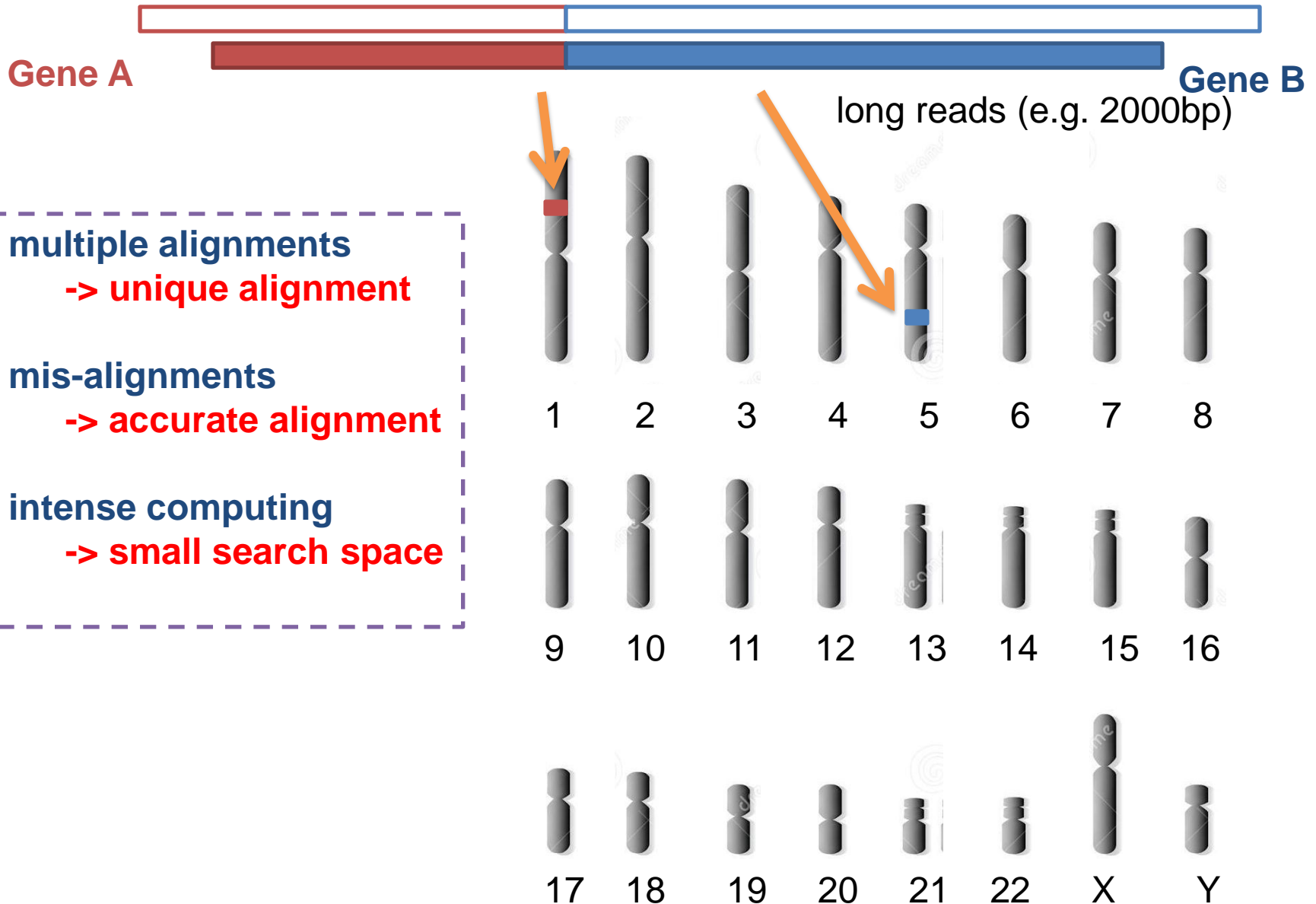
Hybrid-Seq pipeline
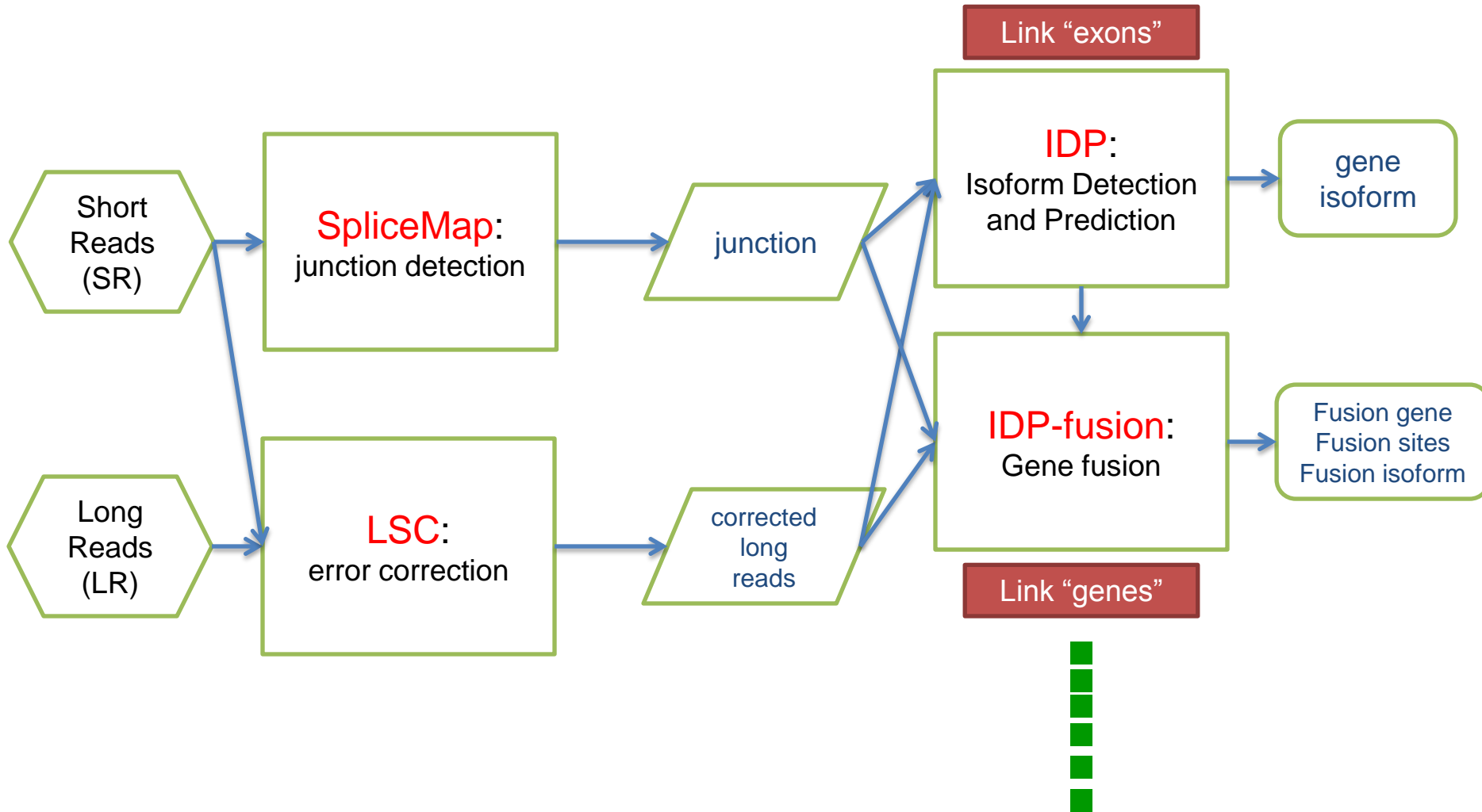
Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC			
IDP			
IDP-fusion			

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads		
IDP			
IDP-fusion			

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads	Improve correction	
IDP			
IDP-fusion			

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads	Improve correction	
IDP	Exon junction detection		
IDP-fusion			

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads	Improve correction	
IDP	Exon junction detection	1. sensitivity of junction detection 2. quantification	
IDP-fusion			

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads	Improve correction	
IDP	Exon junction detection	1. sensitivity of junction detection 2. quantification	1. detect exon linkage 2. isoform identification
IDP-fusion			

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads	Improve correction	
IDP	Exon junction detection	1. sensitivity of junction detection 2. quantification	1. detect exon linkage 2. isoform identification
IDP-fusion	Precise fusion site determination		

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads	Improve correction	
IDP	Exon junction detection	1. sensitivity of junction detection 2. quantification	1. detect exon linkage 2. isoform identification
IDP-fusion	Precise fusion site determination	1. sensitivity of junction detection 2. quantification	

Hybrid-Seq

	Short reads: accuracy	Short reads: throughput	Long reads: length
LSC	Correct error of long reads	Improve correction	
IDP	Exon junction detection	1. sensitivity of junction detection 2. quantification	1. detect exon linkage 2. isoform identification
IDP-fusion	Precise fusion site determination	1. sensitivity of junction detection 2. quantification	1. unique alignments for fusion gene detection 2. fusion isoform

Accuracy

Accuracy VS read length

accuracy

# of passes

$$\text{\# of passes} = \frac{\text{Total length of CLR}}{\text{CCS read length}}$$
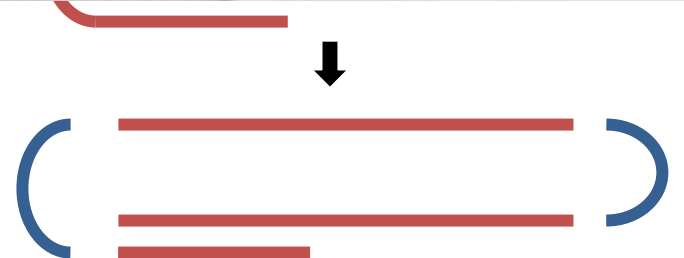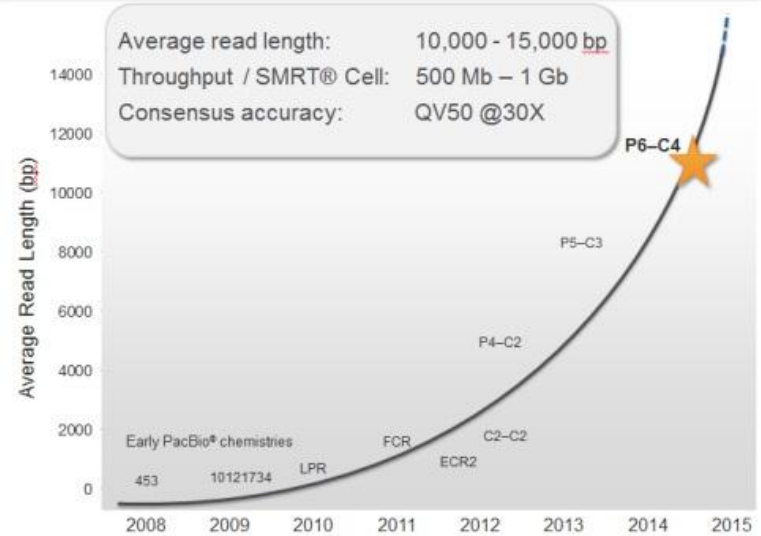
Movie time is fixed
life time of polymerase is limited

Total length of CLR is fixed

CCS read length

October 2014: P6-C5 Sequencing Chemistry Released

Average read length:	10,000 - 15,000 bp
Throughput / SMRT® Cell:	500 Mb – 1 Gb
Consensus accuracy:	QV50 @30X

P6–C4

P5–C3

P4–C2

Early PacBio® chemistries   FCR   C2–C2

453   10121734   LPR   ECR2

Average Read Length (bp)

14000
12000
10000
8000
6000
4000
2000
0

2008 2009 2010 2011 2012 2013 2014 2015

PACIFIC BIOSCIENCES®

CCS

Throughput

Throughput?
(Number of reads)

The upper limit is the number of ZMWs in a SMRT cell?

Thank you!