Computational Analysis of Next-Generation Sequencing Data

For Advancing Cancer Research

Andrea Sboner, Ph.D. May 23, 2012

Center For Health Informatics And Bioinformatic, New York University

Next-Generation Sequencing (NGS)

Next-Generation Sequencing (NGS)

- Paradigm shift in genomics and biology

Next-Generation Sequencing (NGS)

- Paradigm shift in genomics and biology
- Initially, whole genome sequencing

Next-Generation Sequencing (NGS)

- Paradigm shift in genomics and biology
- Initially, whole genome sequencing
- Currently, several applications of NGS:

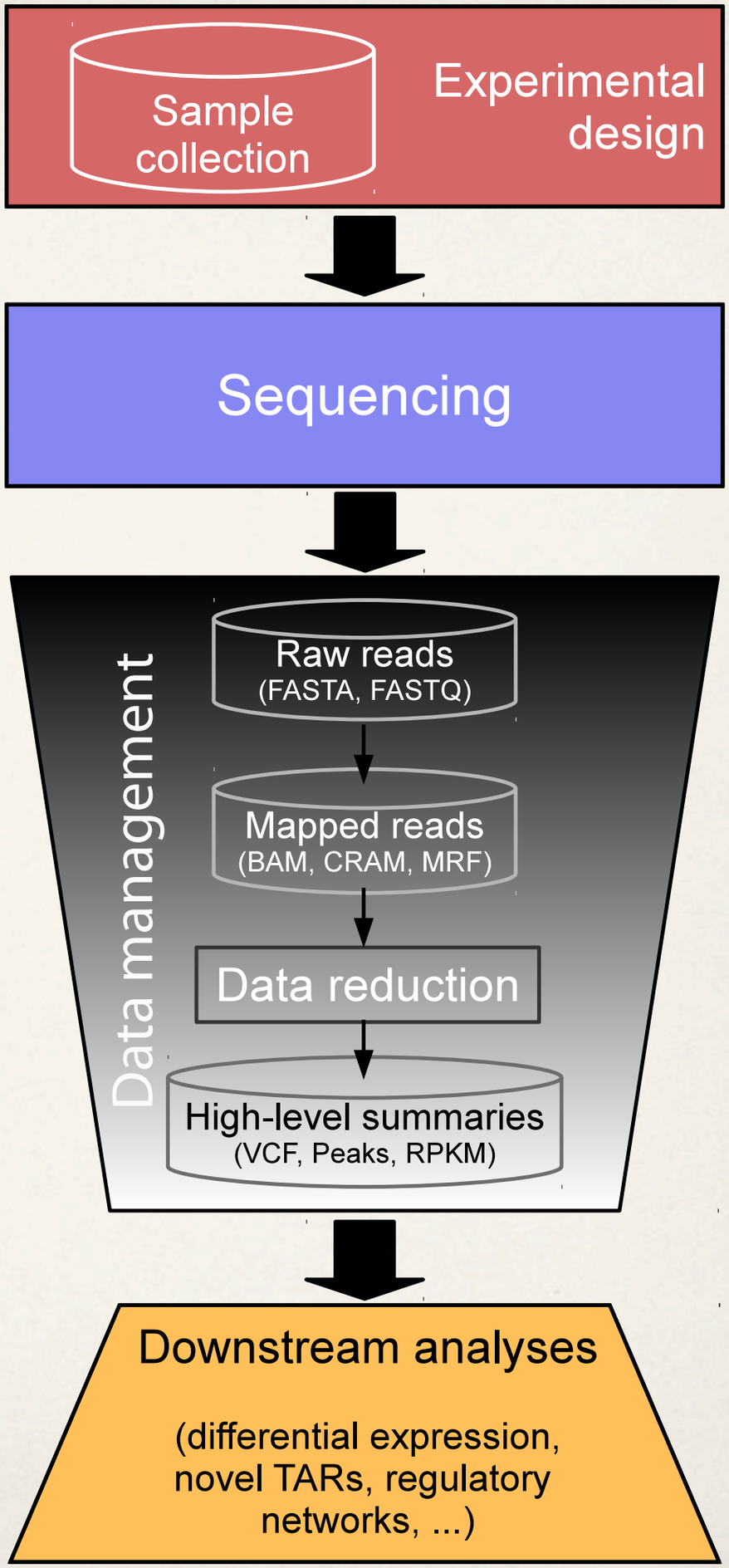
Next-Generation Sequencing (NGS)

- Paradigm shift in genomics and biology
- Initially, whole genome sequencing
- Currently, several applications of NGS:
  - Whole-genome and exome sequencing (WGS, WES); targeted re-sequencing

Next-Generation Sequencing (NGS)

- Paradigm shift in genomics and biology
- Initially, whole genome sequencing
- Currently, several applications of NGS:
  - Whole-genome and exome sequencing (WGS, WES); targeted re-sequencing
  - Functional genomics:
    - RNA-seq; CLIP-seq; ChIP-seq; chromosome conformation capture; bisulfite sequencing, etc.

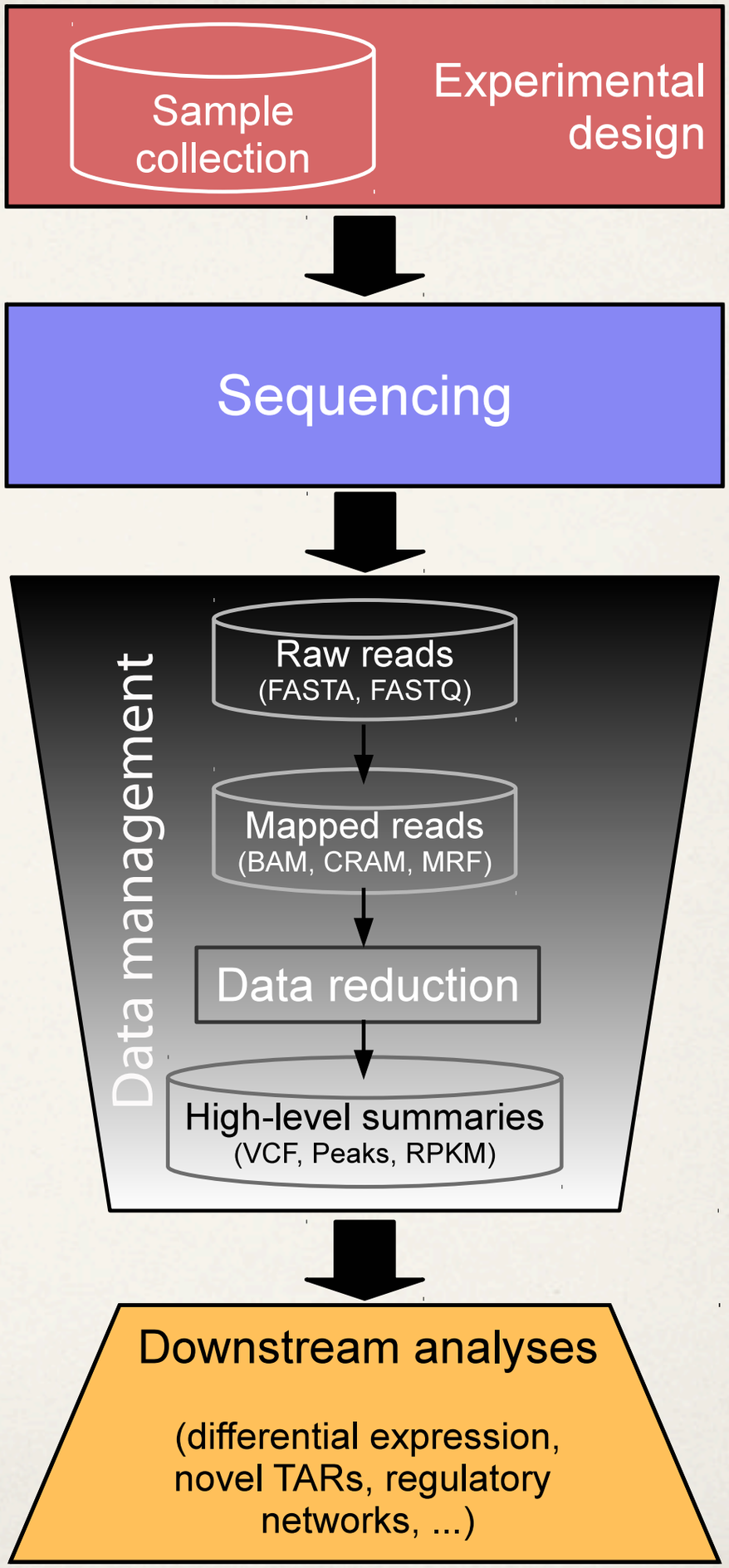
Schematic of a NGS experiment



Sboner, Andrea, et al. 2011. "The real cost of sequencing: higher than you think!" Genome Biology 12(8):25:125

Schematic of a NGS experiment

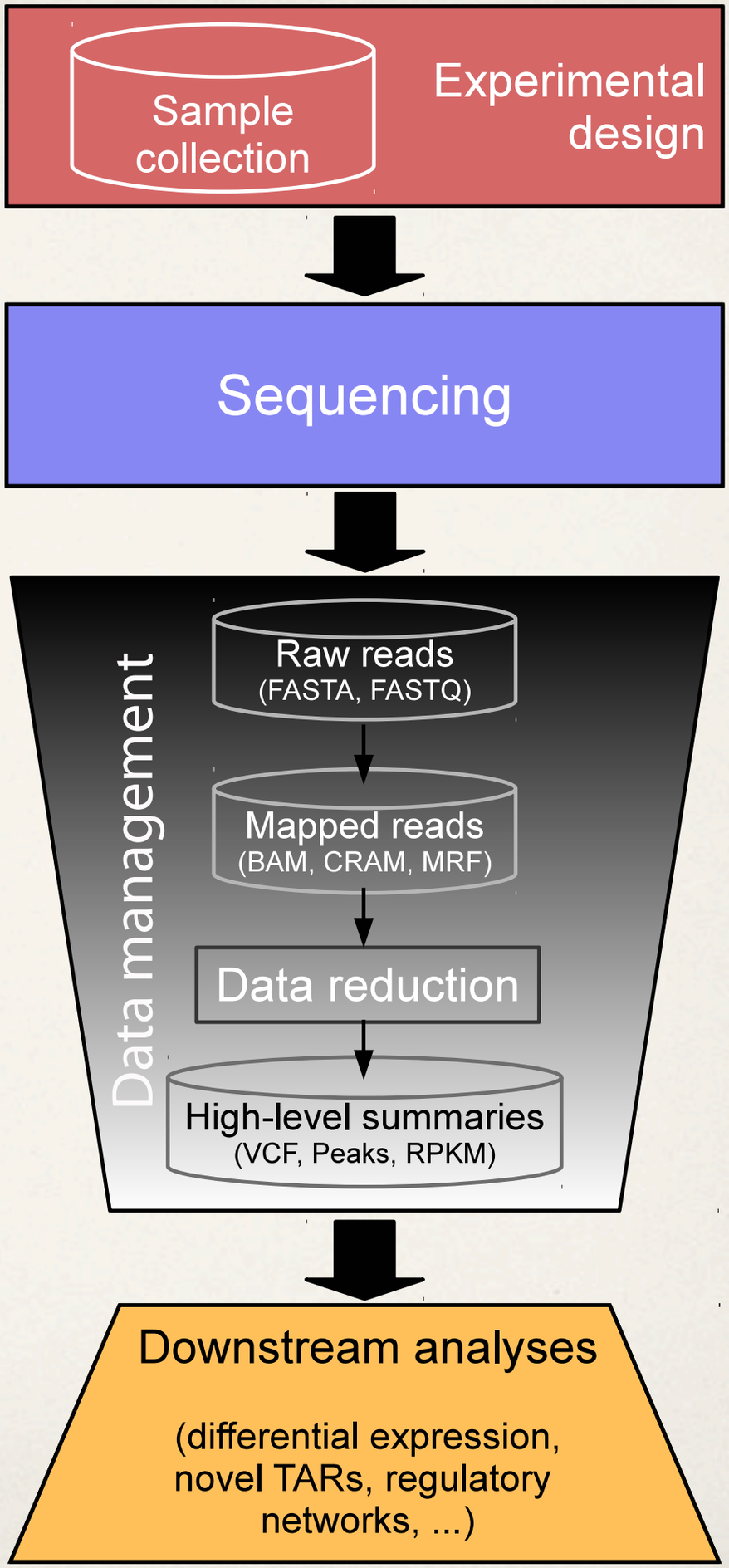
- Experimental design



Sboner, Andrea, et al. 2011. "The real cost of sequencing: higher than you think!" Genome Biology 12(8):25:125

Schematic of a NGS experiment

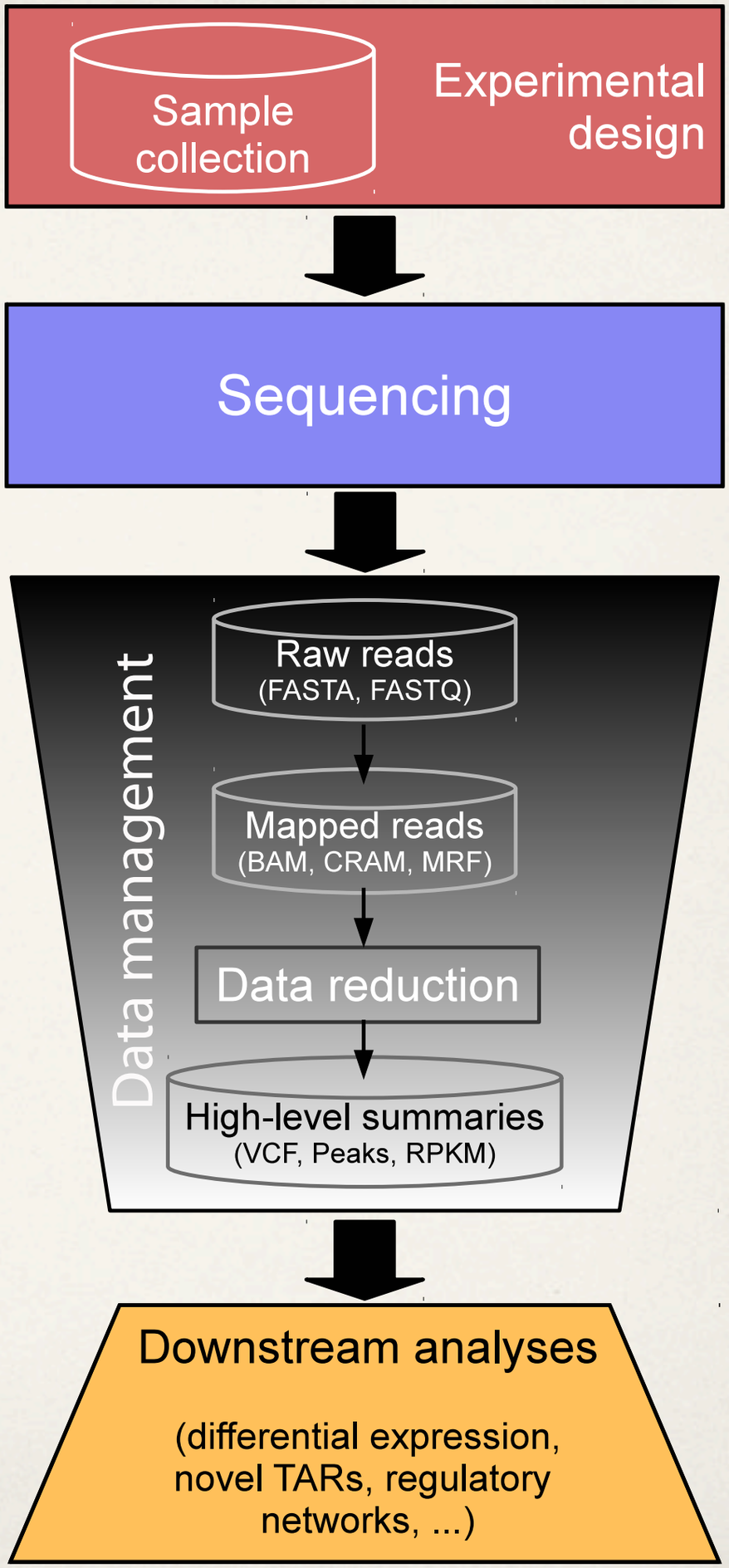
- Experimental design
- Sequencing



Sboner, Andrea, et al. 2011. "The real cost of sequencing: higher than you think!" Genome Biology 12(8):25:125

Schematic of a NGS experiment

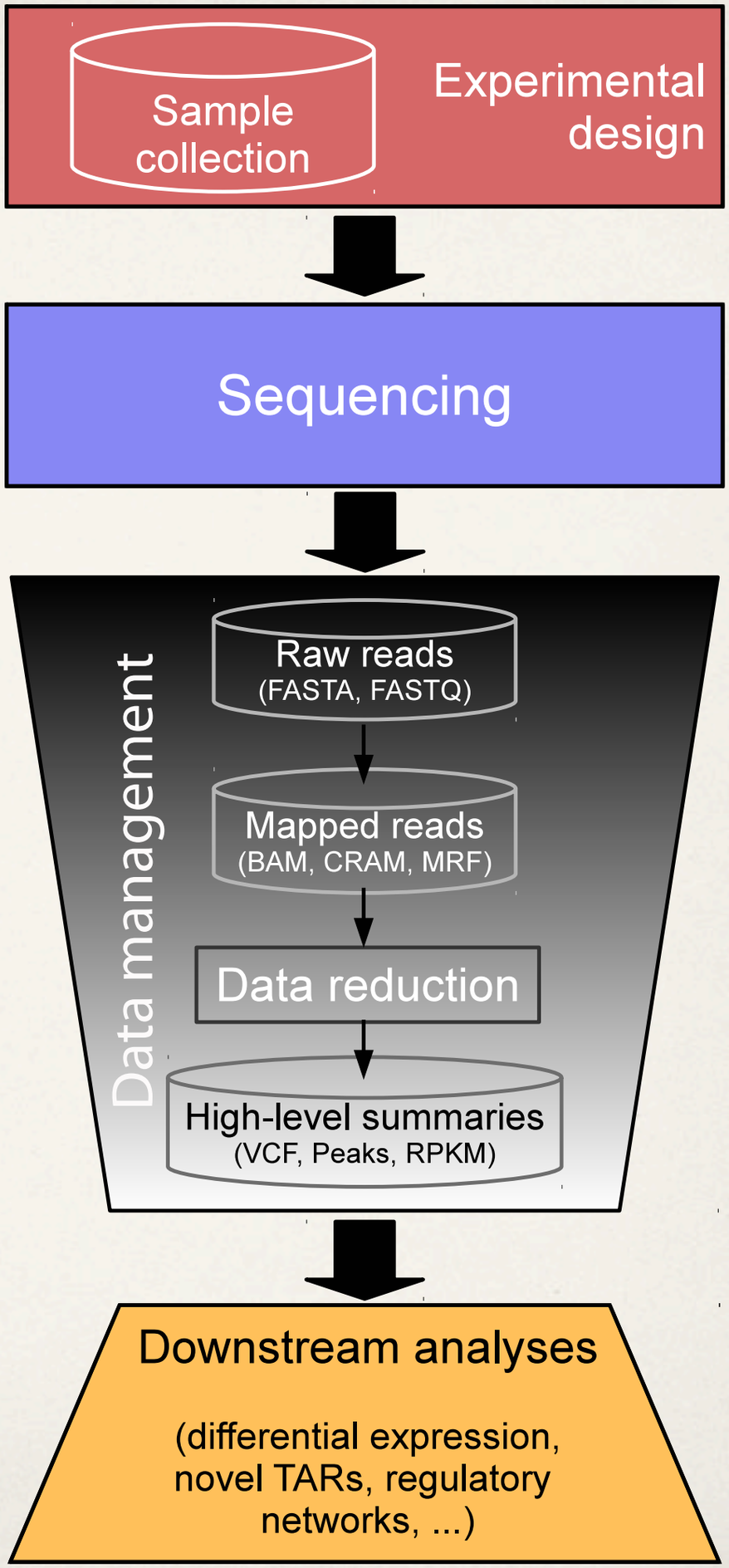
- Experimental design
- Sequencing
- **Data management**



Sboner, Andrea, et al. 2011. "The real cost of sequencing: higher than you think!" Genome Biology 12(8):25:125

Schematic of a NGS experiment

- Experimental design
- Sequencing
- **Data management**
- **Downstream analysis**



Sboner, Andrea, et al. 2011. "The real cost of sequencing: higher than you think!" Genome Biology 12(8):25:125

RNA-Seq

RNA-Seq

- Application of next-generation sequencing to the study of *transcriptomes*:

RNA-Seq

- Application of next-generation sequencing to the study of *transcriptomes*:
  - expression measurements:
    - *gene level*
    - *exon level*
    - *transcript level*

RNA-Seq

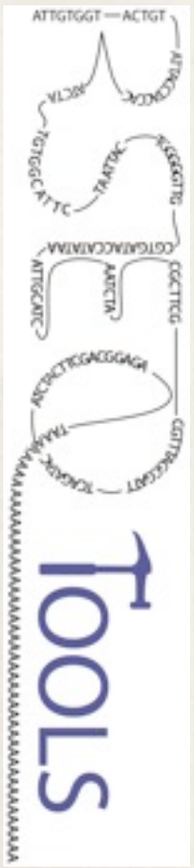
- Application of next-generation sequencing to the study of *transcriptomes*:
  - expression measurements:
    - *gene level*
    - *exon level*
    - *transcript level*
  - alternative splicing

RNA-Seq

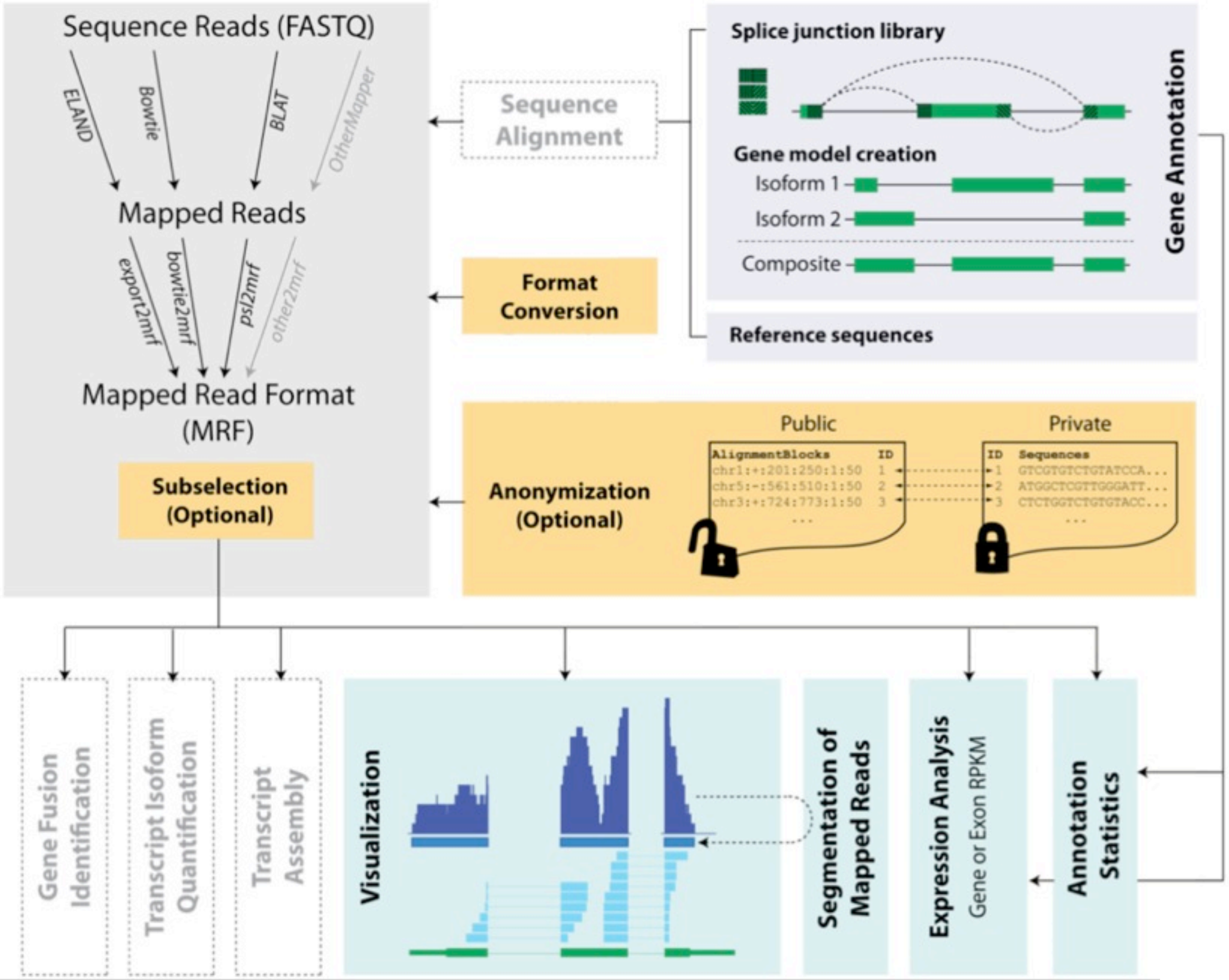
- Application of next-generation sequencing to the study of *transcriptomes*:
  - expression measurements:
    - *gene level*
    - *exon level*
    - *transcript level*
  - alternative splicing
  - transcript reconstruction

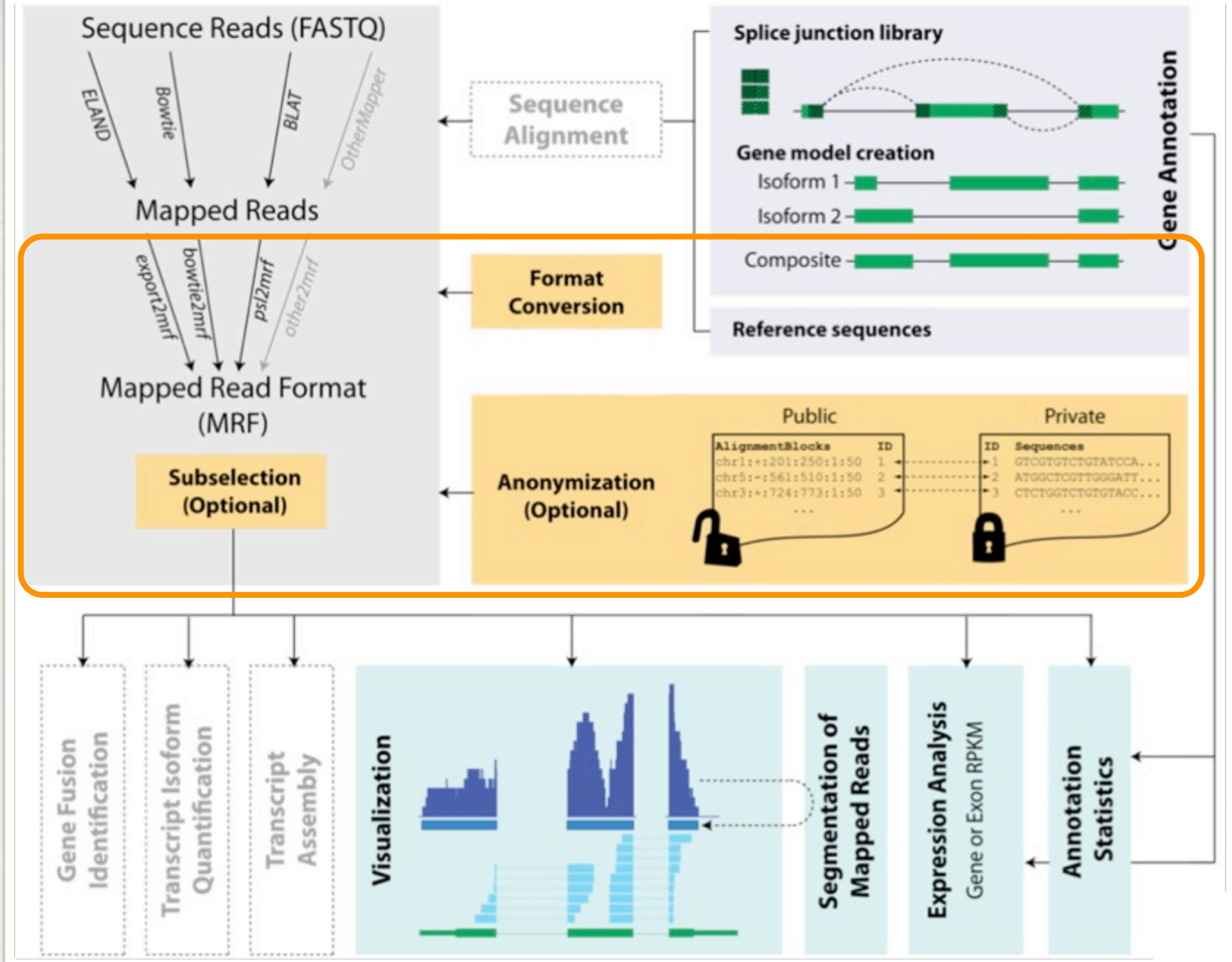
RNA-Seq

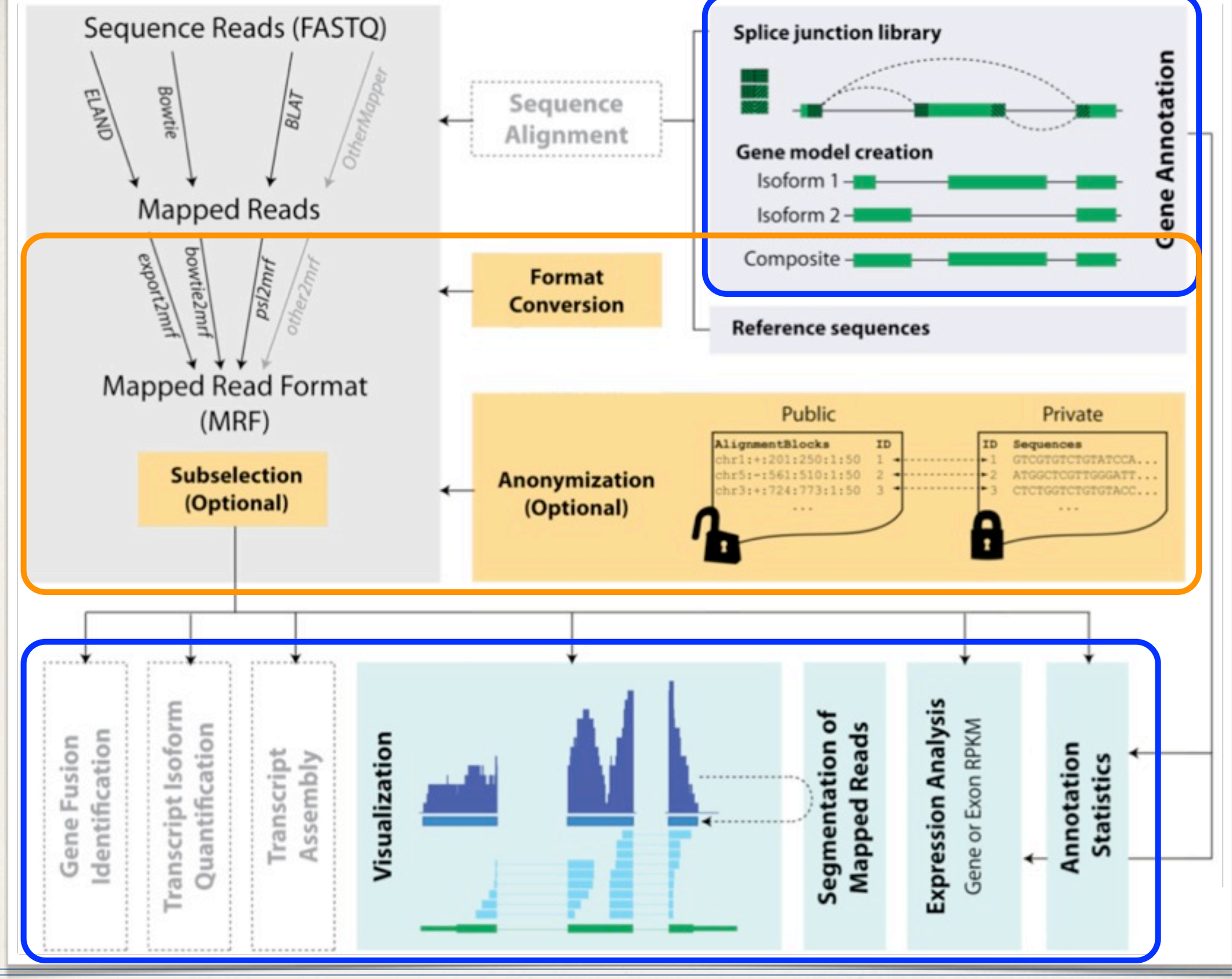
- Application of next-generation sequencing to the study of *transcriptomes*:
  - expression measurements:
    - *gene level*
    - *exon level*
    - *transcript level*
  - alternative splicing
  - transcript reconstruction
  - discovery of novel expressed regions
  - ...



Habegger,* Lukas, Andrea Sboner*, et al. "RSEQtools: a Modular Framework to Analyze RNA-Seq Data Using Compact, Anonymized Data Summaries." Bioinformatics 2011 27(2);281–3





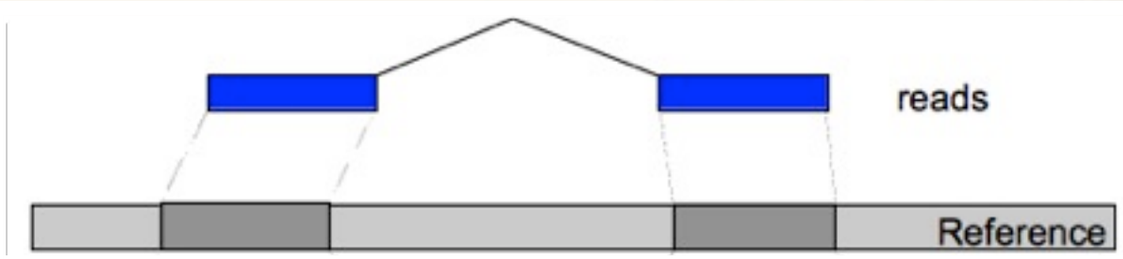


Mapped Read Format (MRF)

Habegger,* Lukas, Andrea Sboner*, et al. Bioinformatics 2011 27(2):281–3

Mapped Read Format (MRF)

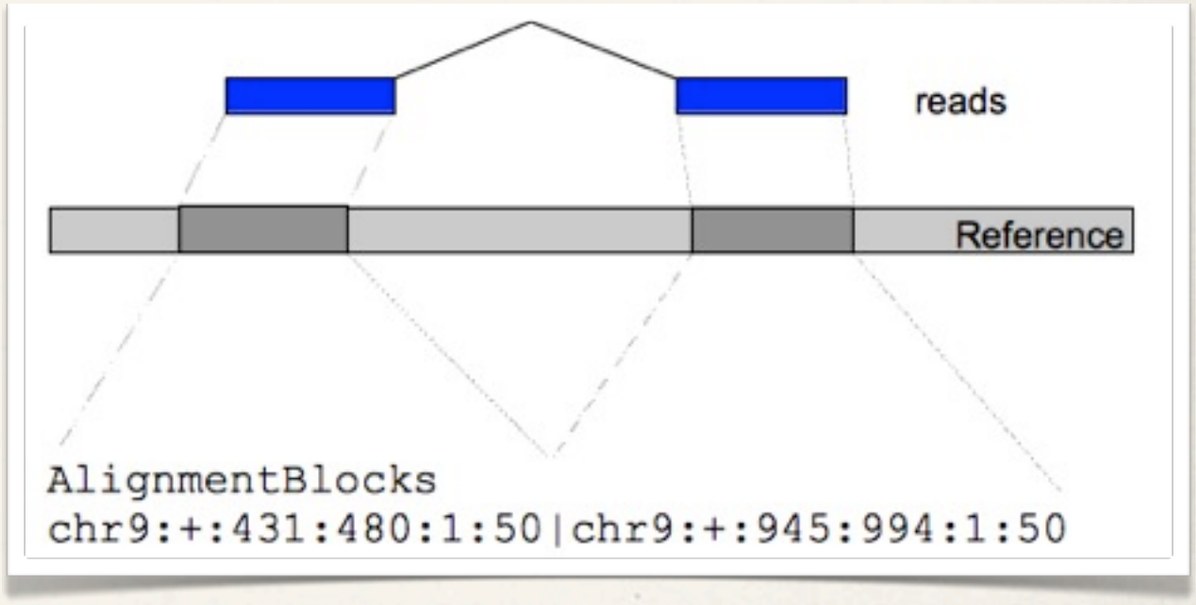
- As any next-generation sequencing project, large datasets are generated, along with issues in:
  - *storing*
  - *processing*
  - *sharing*



Habegger,* Lukas, Andrea Sboner*, et al. Bioinformatics 2011 27(2):281–3

Mapped Read Format (MRF)

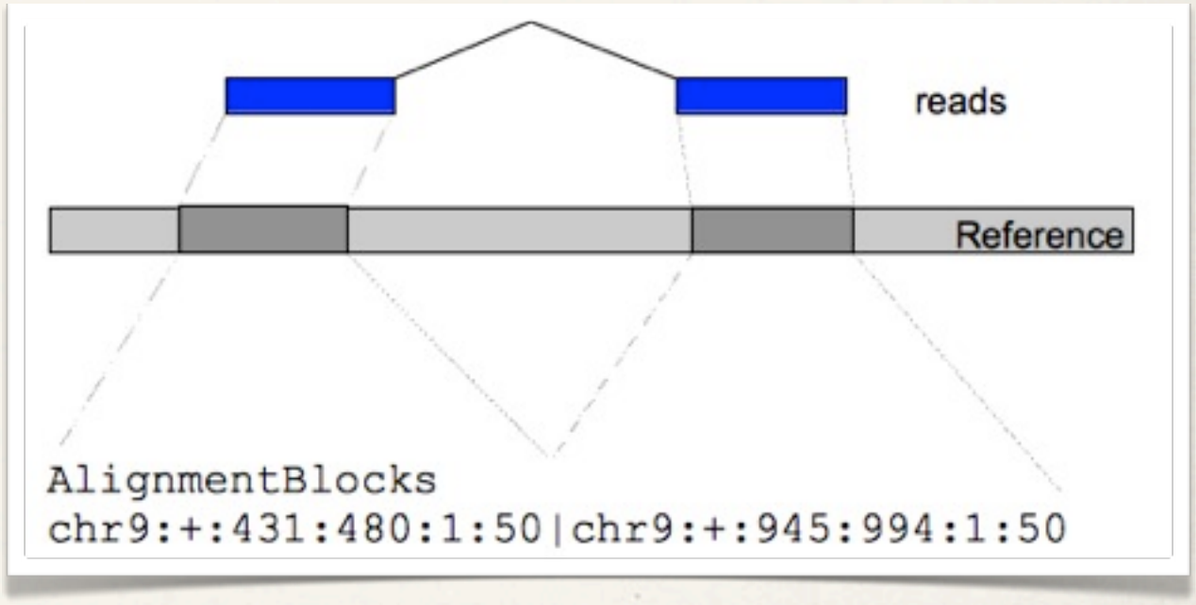
- As any next-generation sequencing project, large datasets are generated, along with issues in:
  - *storing*
  - *processing*
  - *sharing*



Habegger,* Lukas, Andrea Sboner*, et al. Bioinformatics 2011 27(2):281–3

Mapped Read Format (MRF)

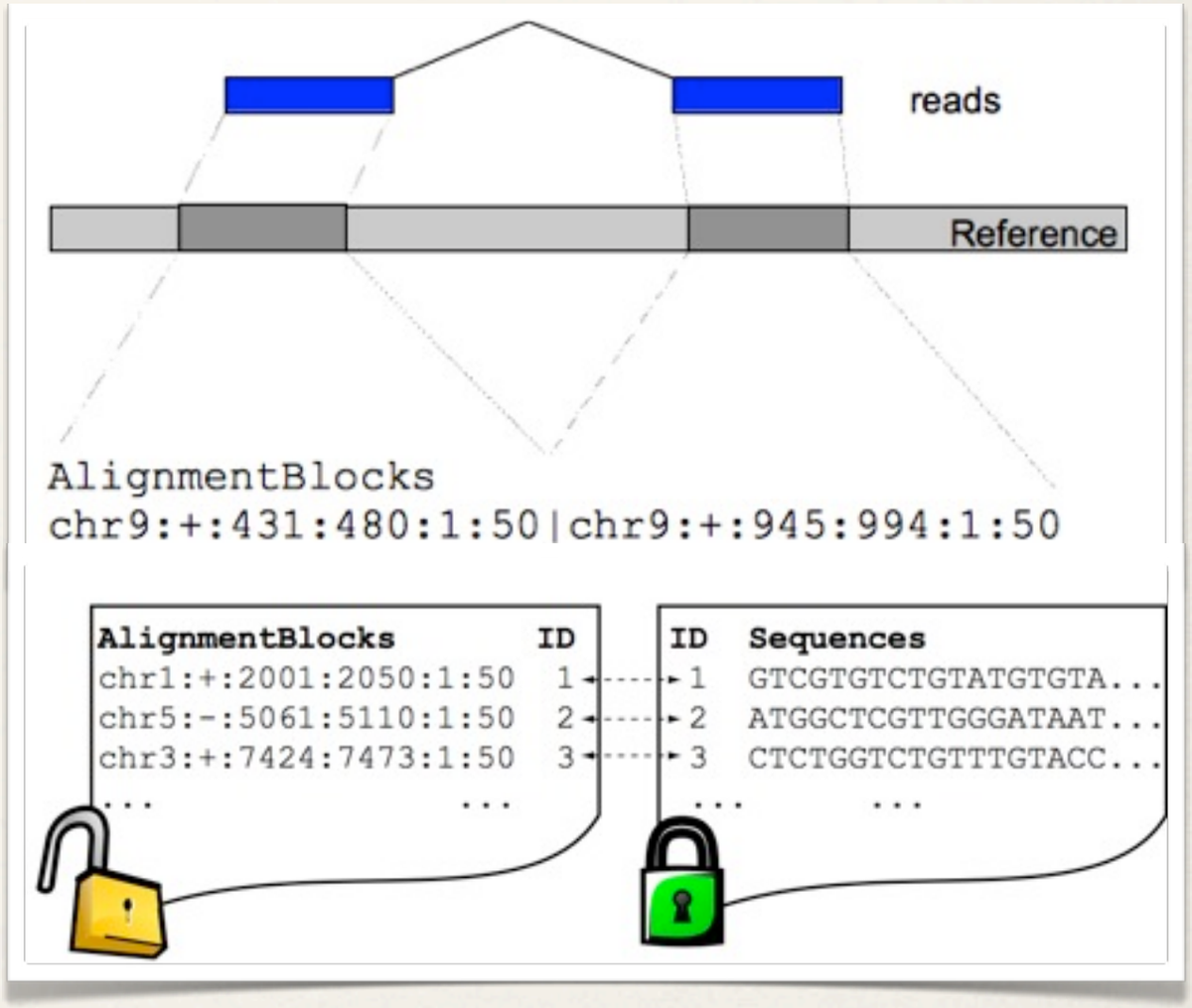
- As any next-generation sequencing project, large datasets are generated, along with issues in:
  - *storing*
  - *processing*
  - *sharing*
- Sequence information potentially includes data sufficient to identify and "genotype" an individual:
  - *privacy issues*



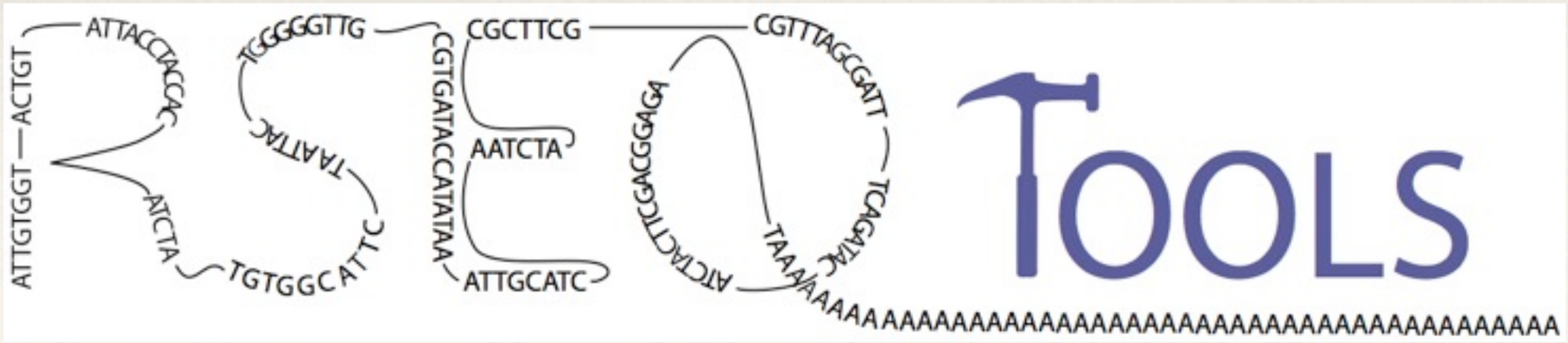
Habegger,* Lukas, Andrea Sboner*, et al. Bioinformatics 2011 27(2):281–3

Mapped Read Format (MRF)

- As any next-generation sequencing project, large datasets are generated, along with issues in:
  - *storing*
  - *processing*
  - *sharing*
- Sequence information potentially includes data sufficient to identify and "genotype" an individual:
  - *privacy issues*

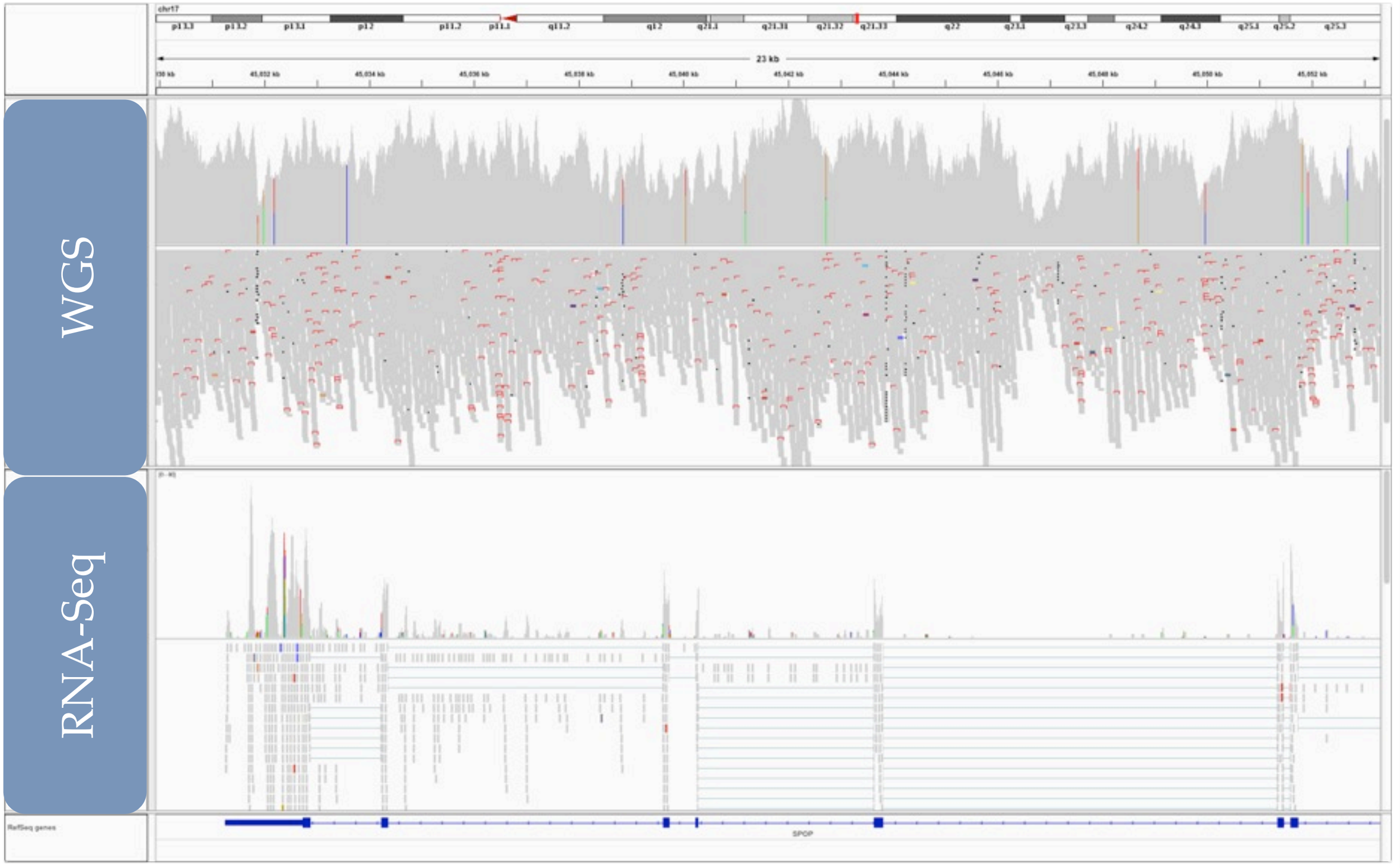


Habegger,* Lukas, Andrea Sboner*, et al. Bioinformatics 2011 27(2):281–3



Purpose	Program	Time to process	File Sizes (uncompressed)	Notes
Alignment + conversion	ELAND2	~1 day	Export: 4.2Gb	Processing of one flow cell (8 lanes)
	export2mrf	~2 minutes	MRF: 400Mb	Number of mapped reads: ~12M
Quantification	mrfQuantifier	45 seconds	Gene expression values: 3.5Mb	GENCODE composite gene models (~22,000)
Visualization	mrf2wig	~2 minutes	One WIG file per chromosome: 1Mb - 150Mb	Signal track of mapped reads normalized per million mapped reads
	mrf2gff	45 seconds	One GFF file per chromosome: 100Kb - 16Mb	To visualize splice junction reads

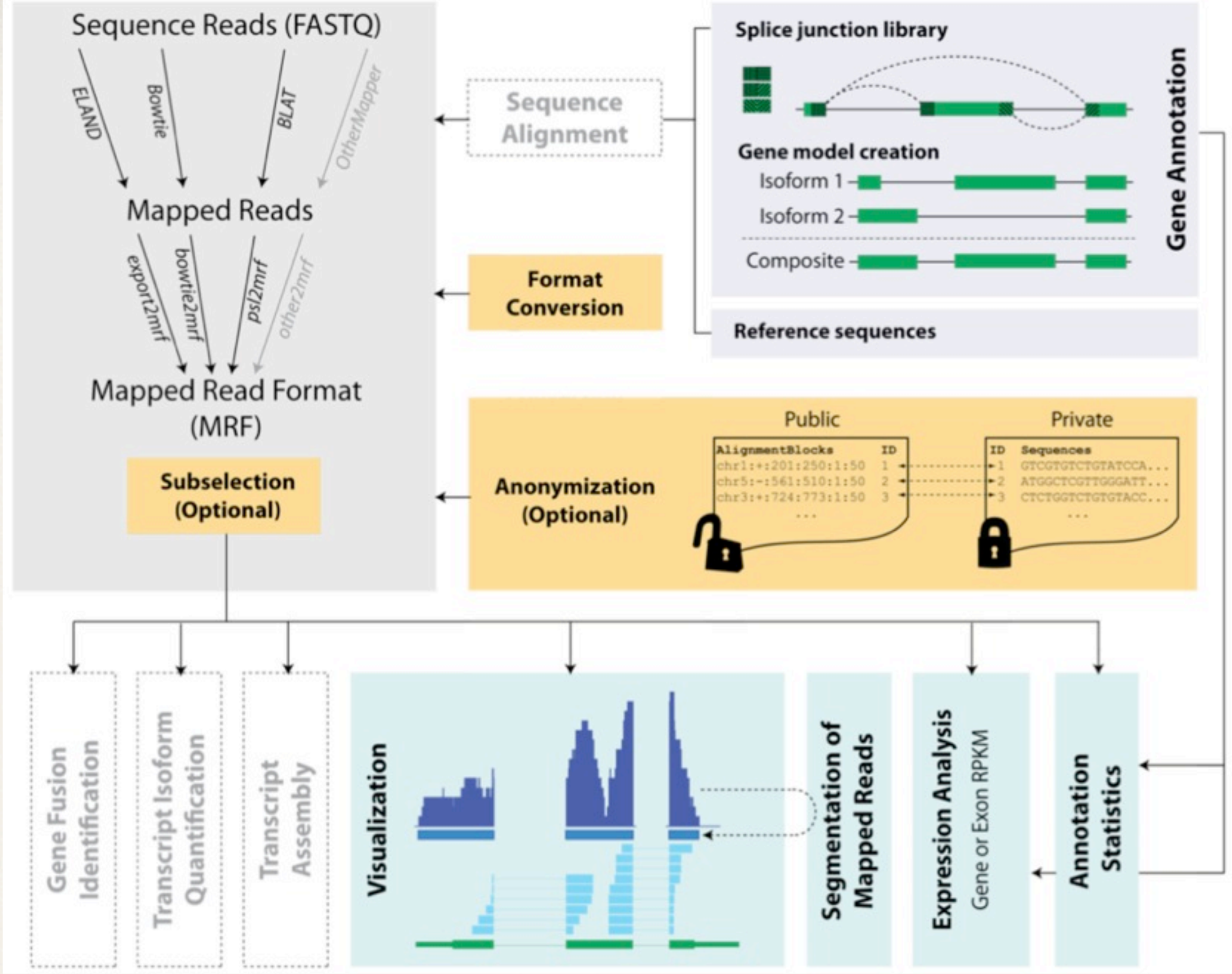
Habegger,* Lukas, Andrea Sboner*, et al. Bioinformatics 2011 27(2):281–3

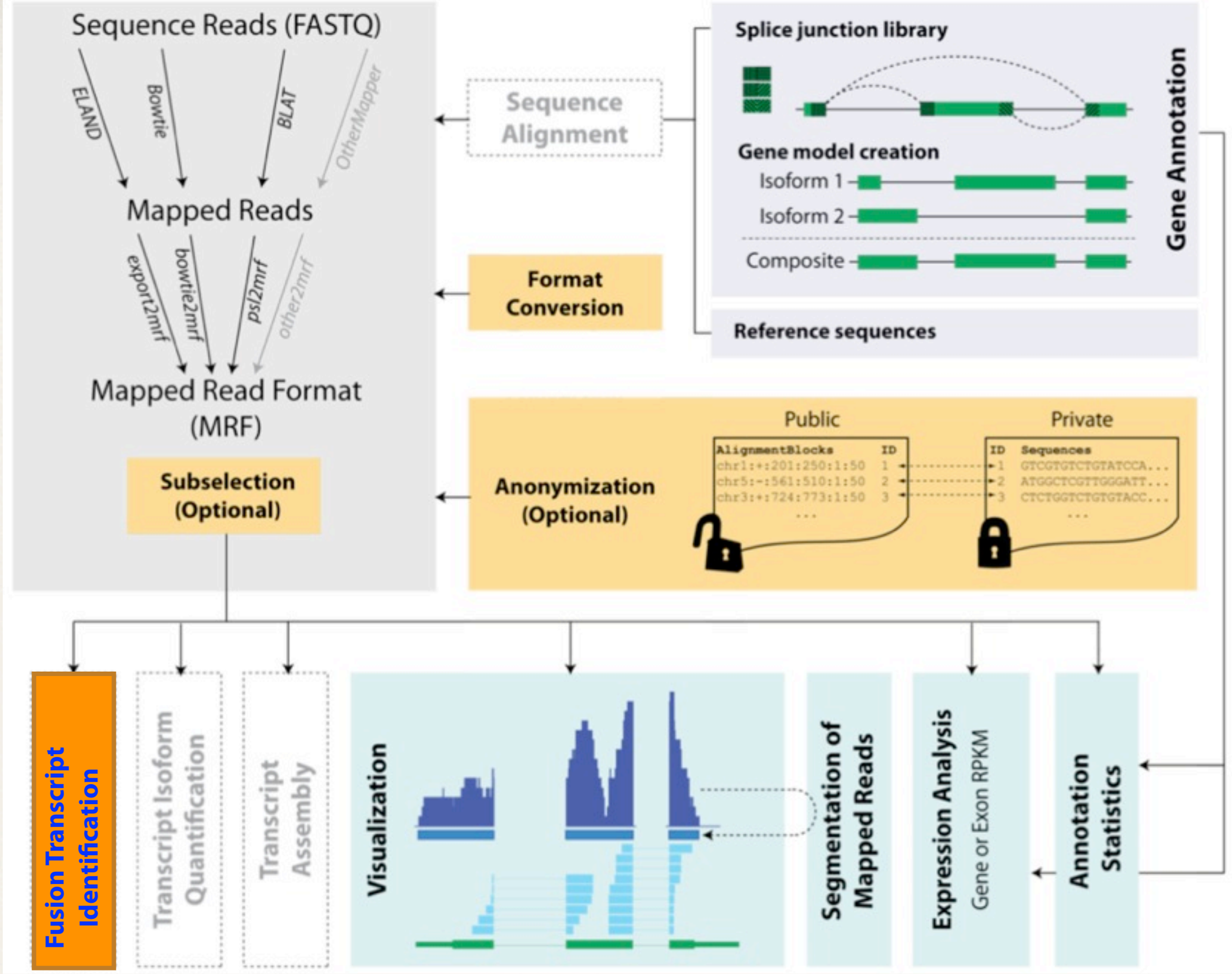


Visualizing mapped reads

e.g. via the Integrative Genome Viewer

Habegger,* Lukas, Andrea Sboner*, et al. Bioinformatics 2011 27(2):281–3





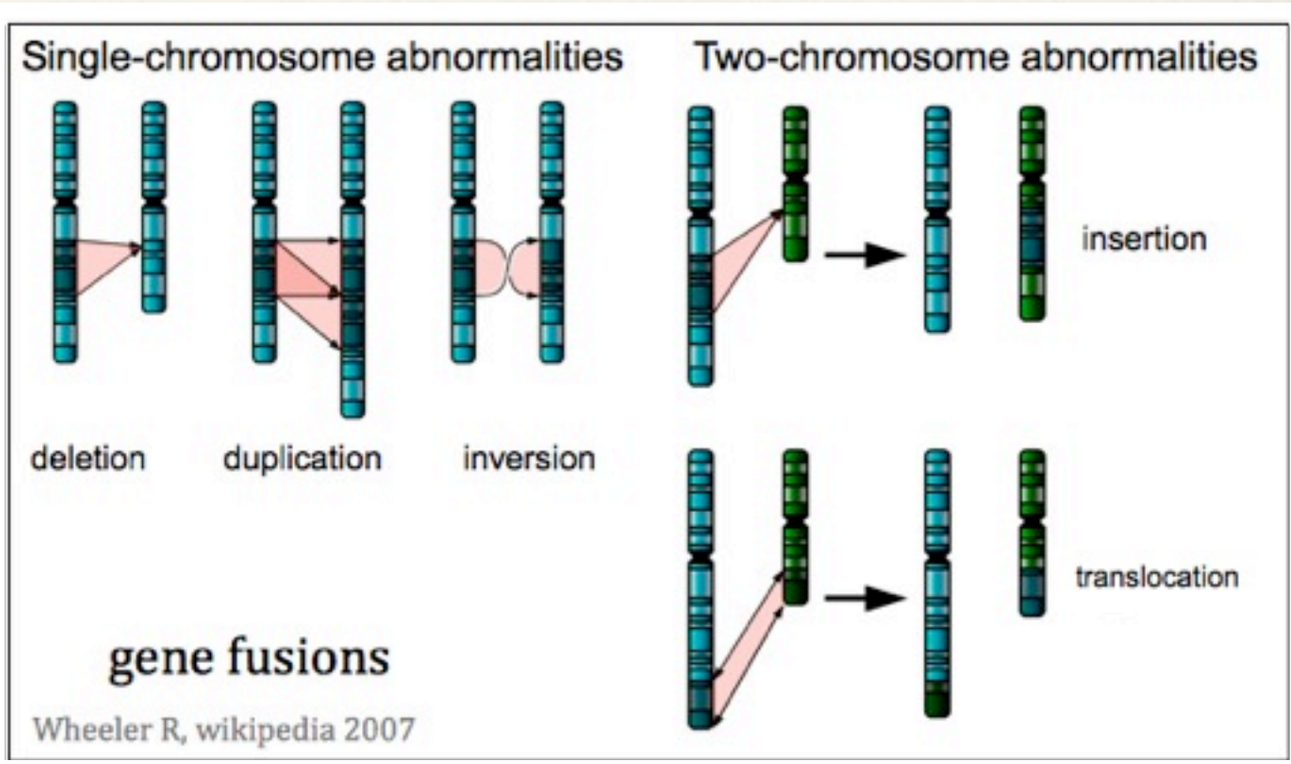
What are chimeric transcripts?

What are chimeric transcripts?

- Transcripts that are *not co-linear* in the genome space

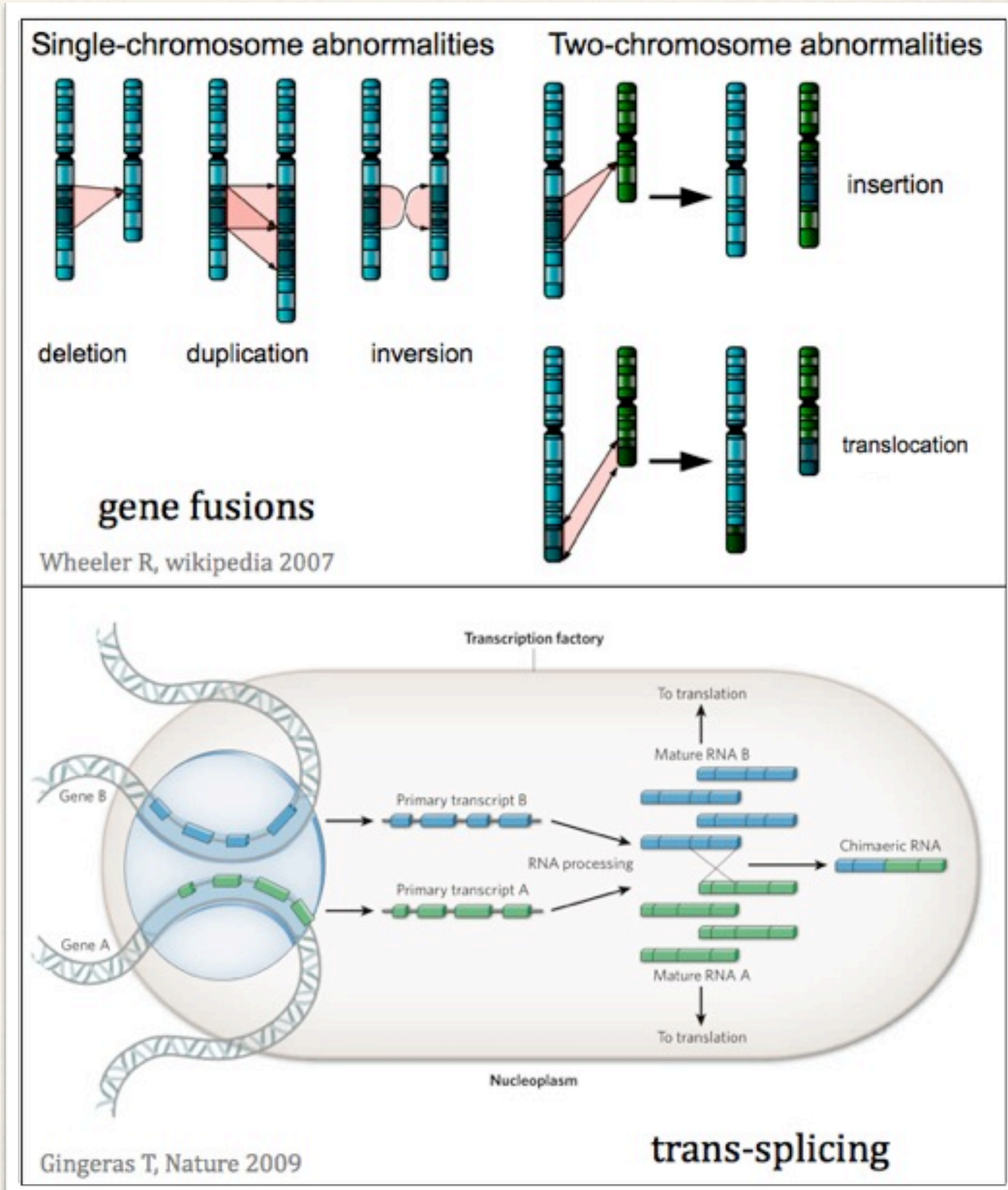
What are chimeric transcripts?

- Transcripts that are *not co-linear* in the genome space
- They can arise from:
  - genomic rearrangements, i.e. *gene fusions*

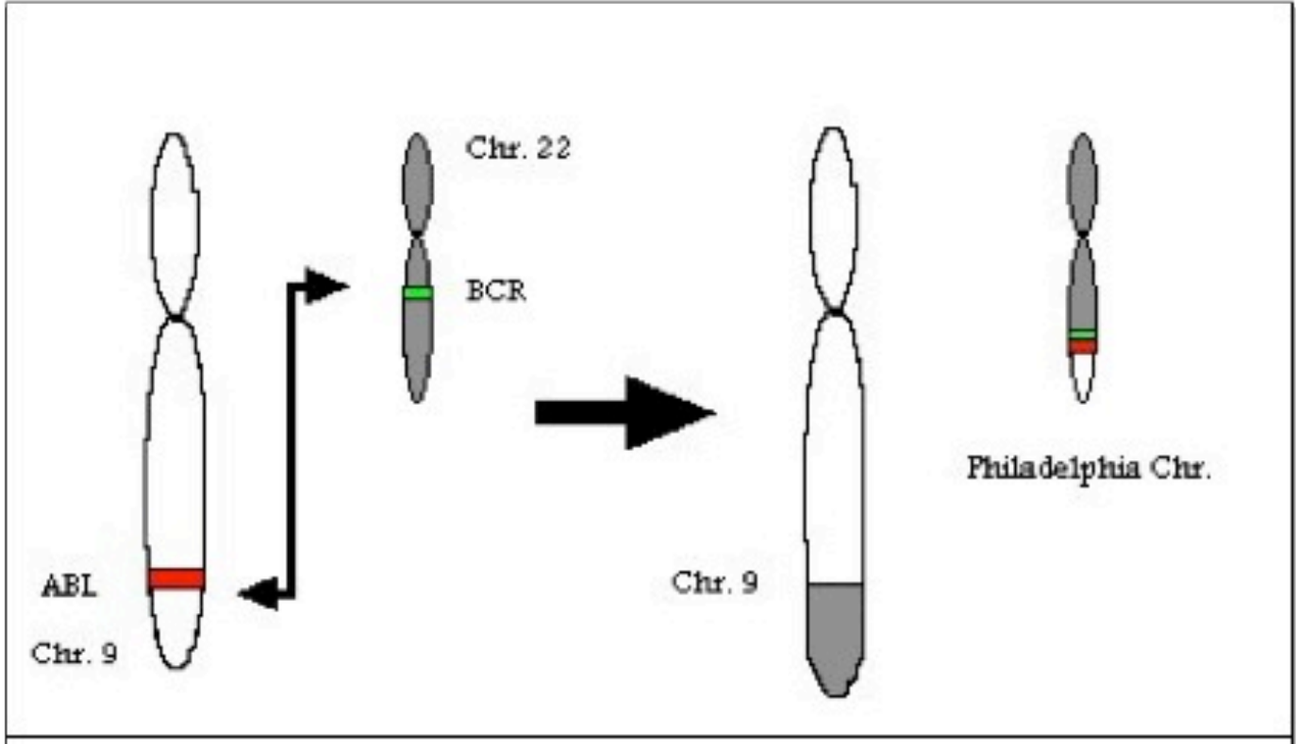


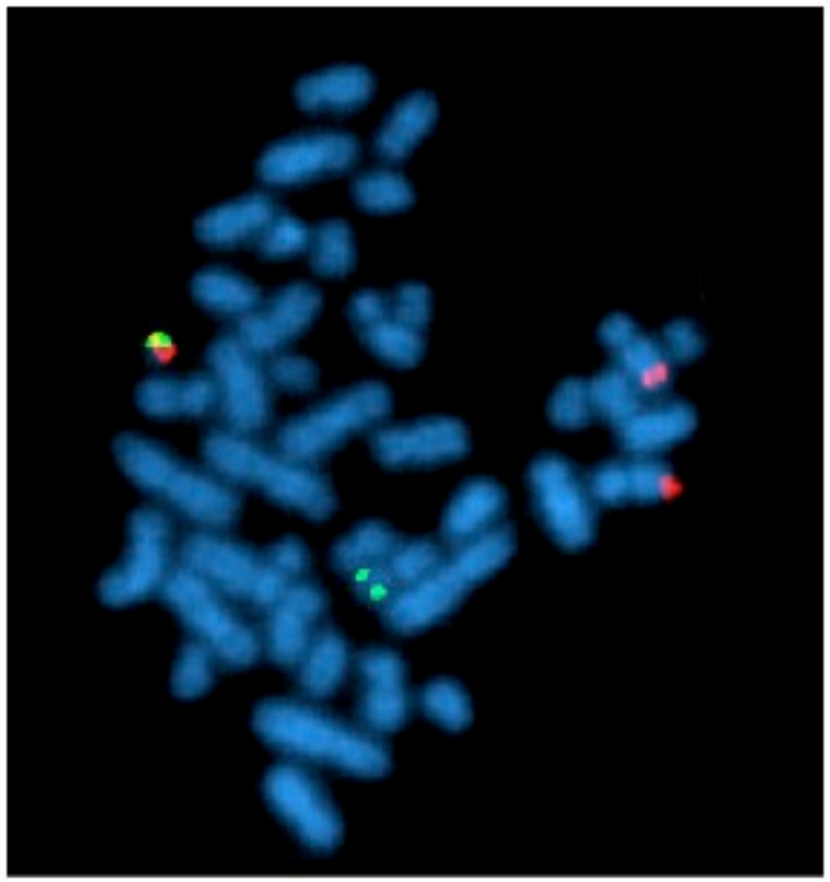
What are chimeric transcripts?

- Transcripts that are *not co-linear* in the genome space
- They can arise from:
  - genomic rearrangements, i.e. *gene fusions*
  - post-transcriptional events, i.e. *trans-splicing*



Why are they (gene fusions) important?



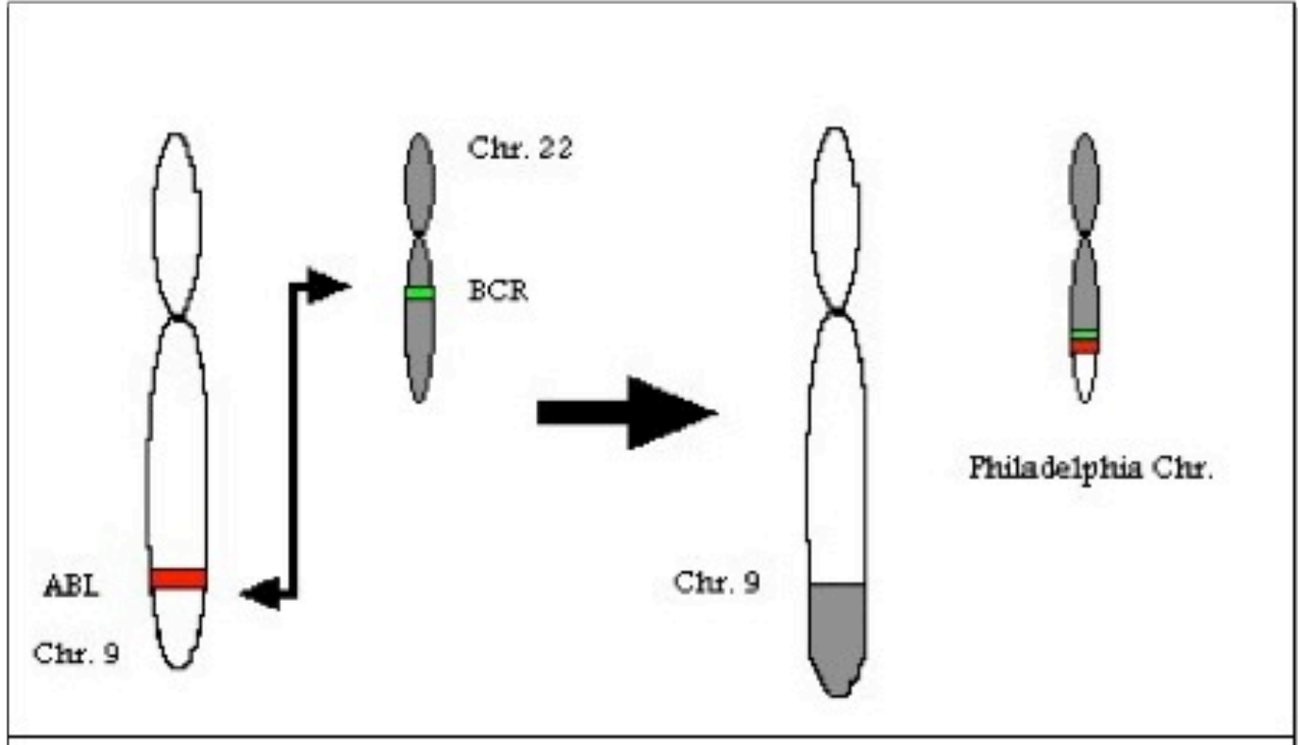


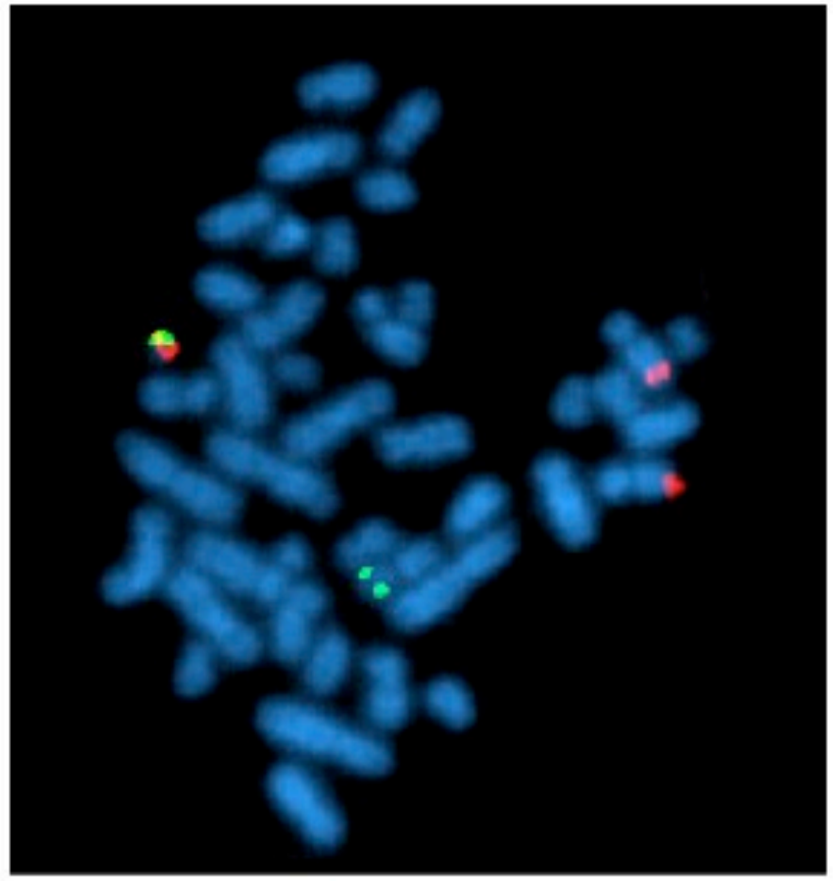
BCR-ABL gene fusion:
schematic and FISH

wikipedia 2005, 2007

Why are they (gene fusions) important?

- Fusion genes are often *oncogenes*
  - Ex: BCR-ABL1 (Philadelphia chromosome) in Chronic myelogenous leukemia (CML) and Acute Lymphoblastic leukemia (ALL) t(9;22)(q34;q11)



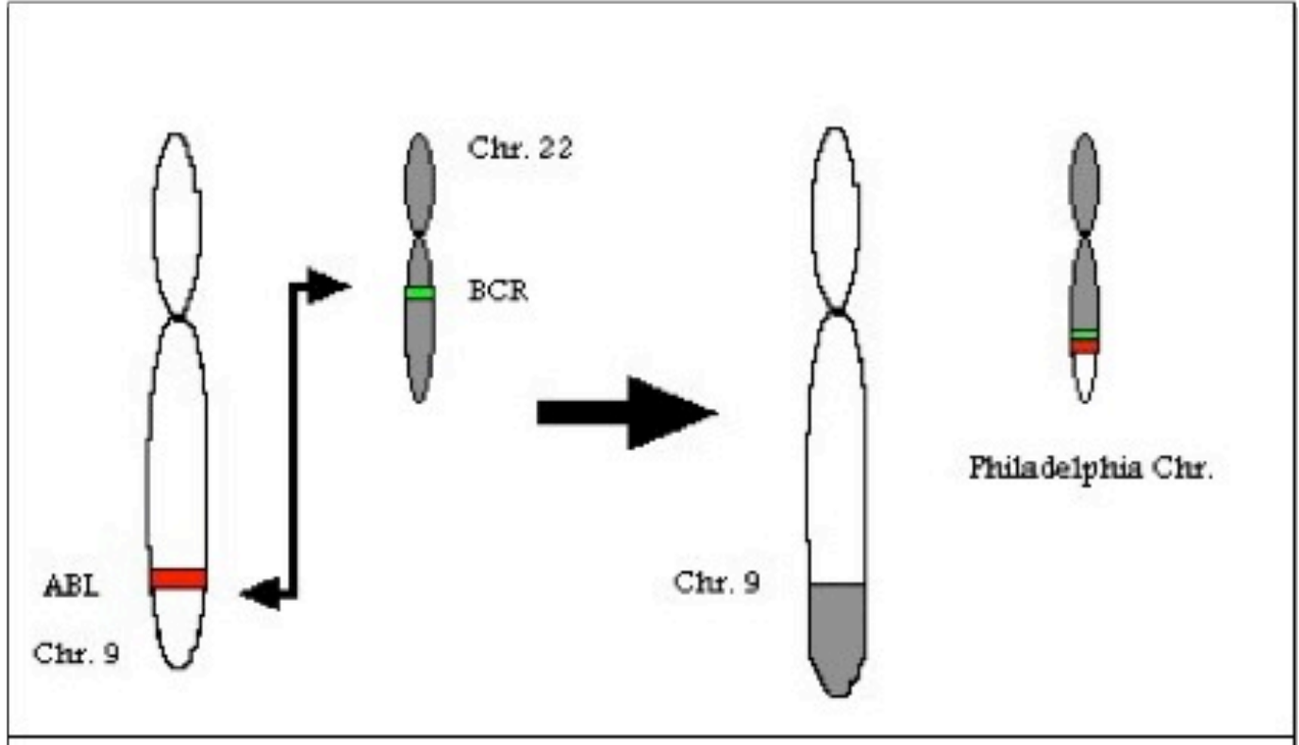


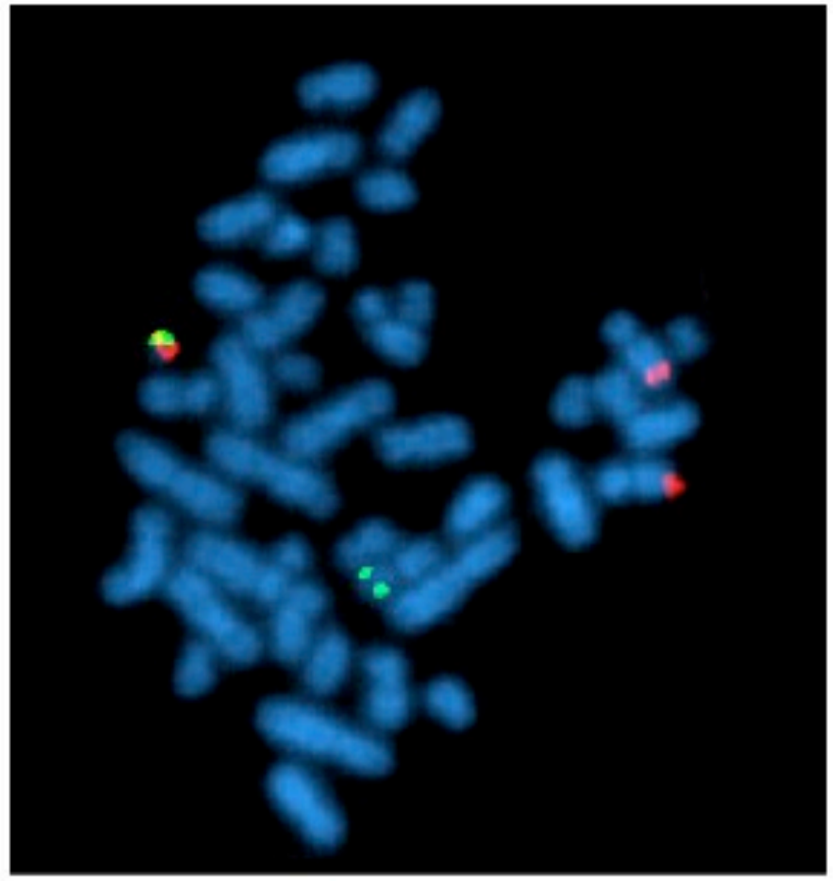
BCR-ABL gene fusion: schematic and FISH

wikipedia 2005, 2007

Why are they (gene fusions) important?

- Fusion genes are often *oncogenes*
  - Ex: BCR-ABL1 (Philadelphia chromosome) in Chronic myelogenous leukemia (CML) and Acute Lymphoblastic leukemia (ALL) t(9;22)(q34;q11)
- Fusion involving a proto-oncogene with a strong promoter resulting in *upregulation* (lymphomas)
  - Ex: (IgH locus)-MYC in Burkitt's lymphoma (cMYC over-expressed)

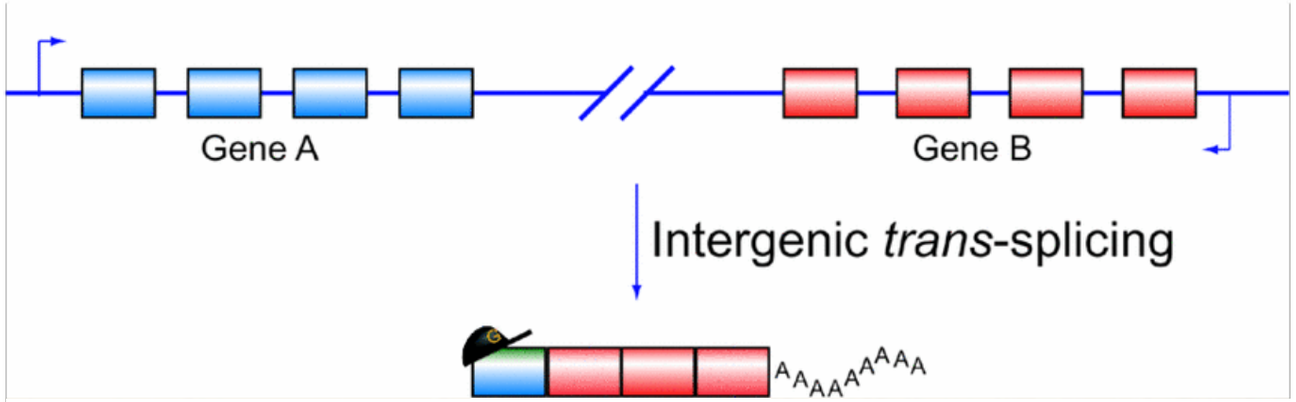




BCR-ABL gene fusion: schematic and FISH

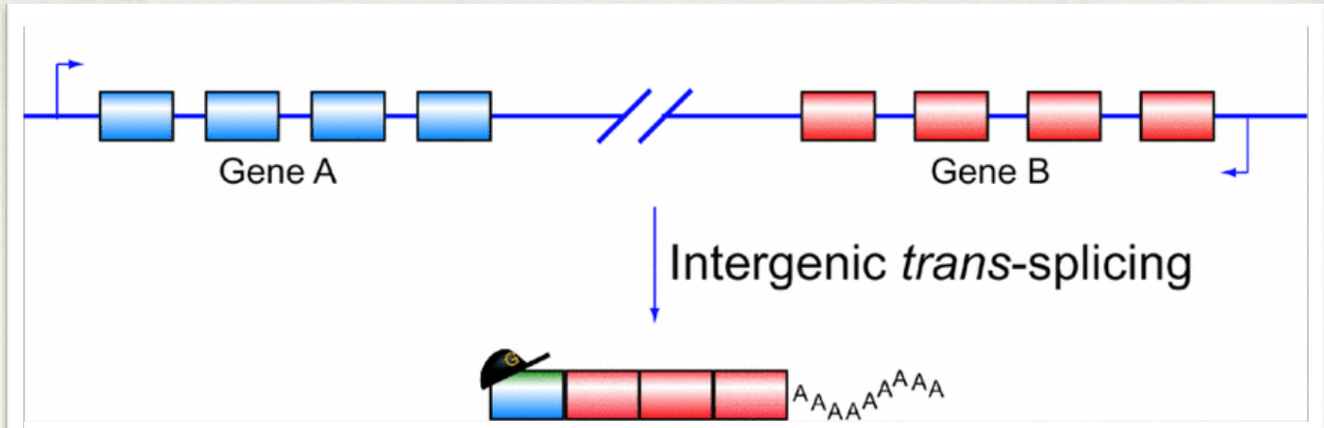
wikipedia 2005, 2007

Why are they (trans-splicing events) important?



Horiuchi, Takayuki, and Toshiro Aigaki. Biology of the Cell 98, no. 2 (January 9, 2012): 135–140.

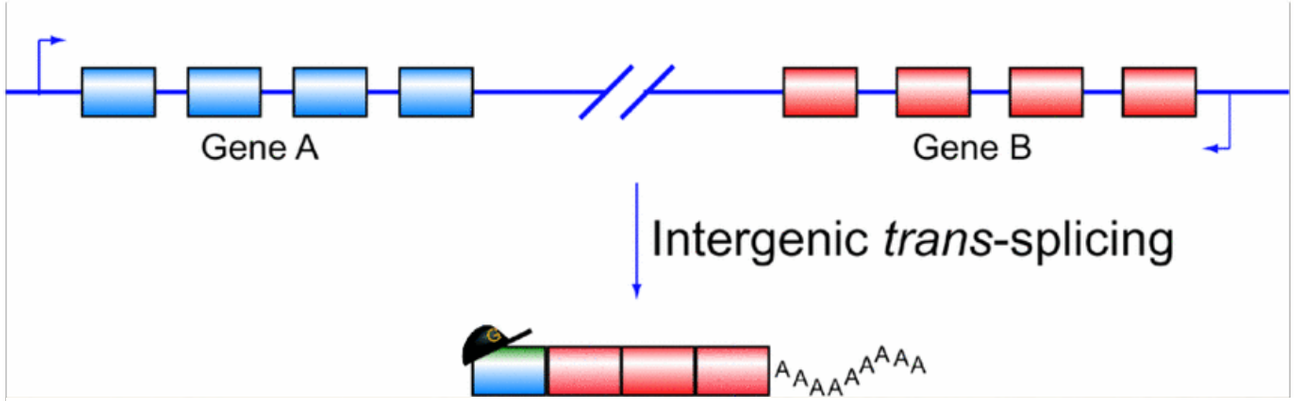
Why are they (trans-splicing events) important?



Horiuchi, Takayuki, and Toshiro Aigaki. Biology of the Cell 98, no. 2 (January 9, 2012): 135–140.

- Trans-splicing was initially found in lower eukariotes, such as trypanosomes and worms
  - Short sequences of nucleotides are trans-spliced to distant 5' of many protein coding genes

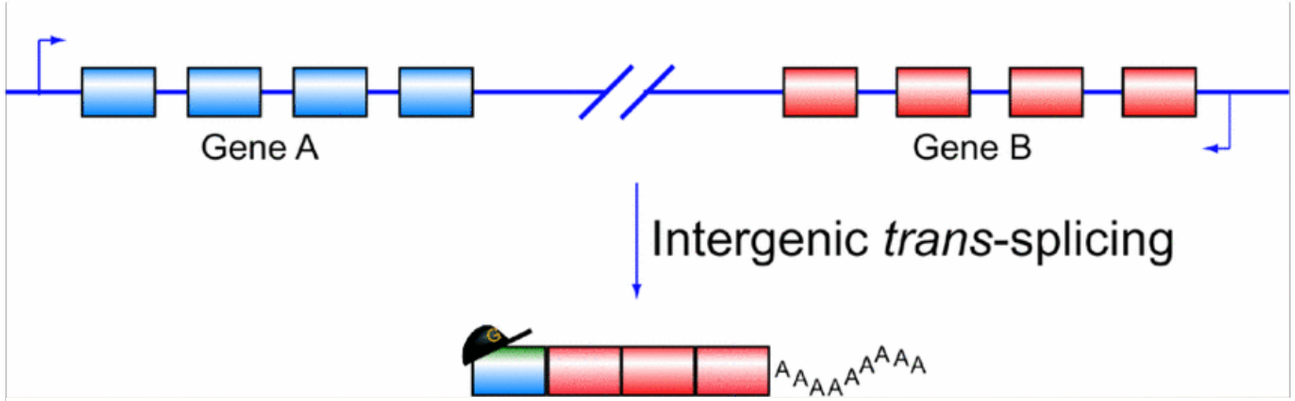
Why are they (trans-splicing events) important?



Horiuchi, Takayuki, and Toshiro Aigaki. Biology of the Cell 98, no. 2 (January 9, 2012): 135–140.

- Trans-splicing was initially found in lower eukariotes, such as trypanosomes and worms
  - Short sequences of nucleotides are trans-spliced to distant 5' of many protein coding genes
- Recently, they were found in mammalian cells:
  - JAZF1-SUZ12 in endometrial stroma cells (Li et al. Science 2008)
  - SLC45A3-ELK4 in prostate tissues (Rickman et al. Cancer Res 2009)

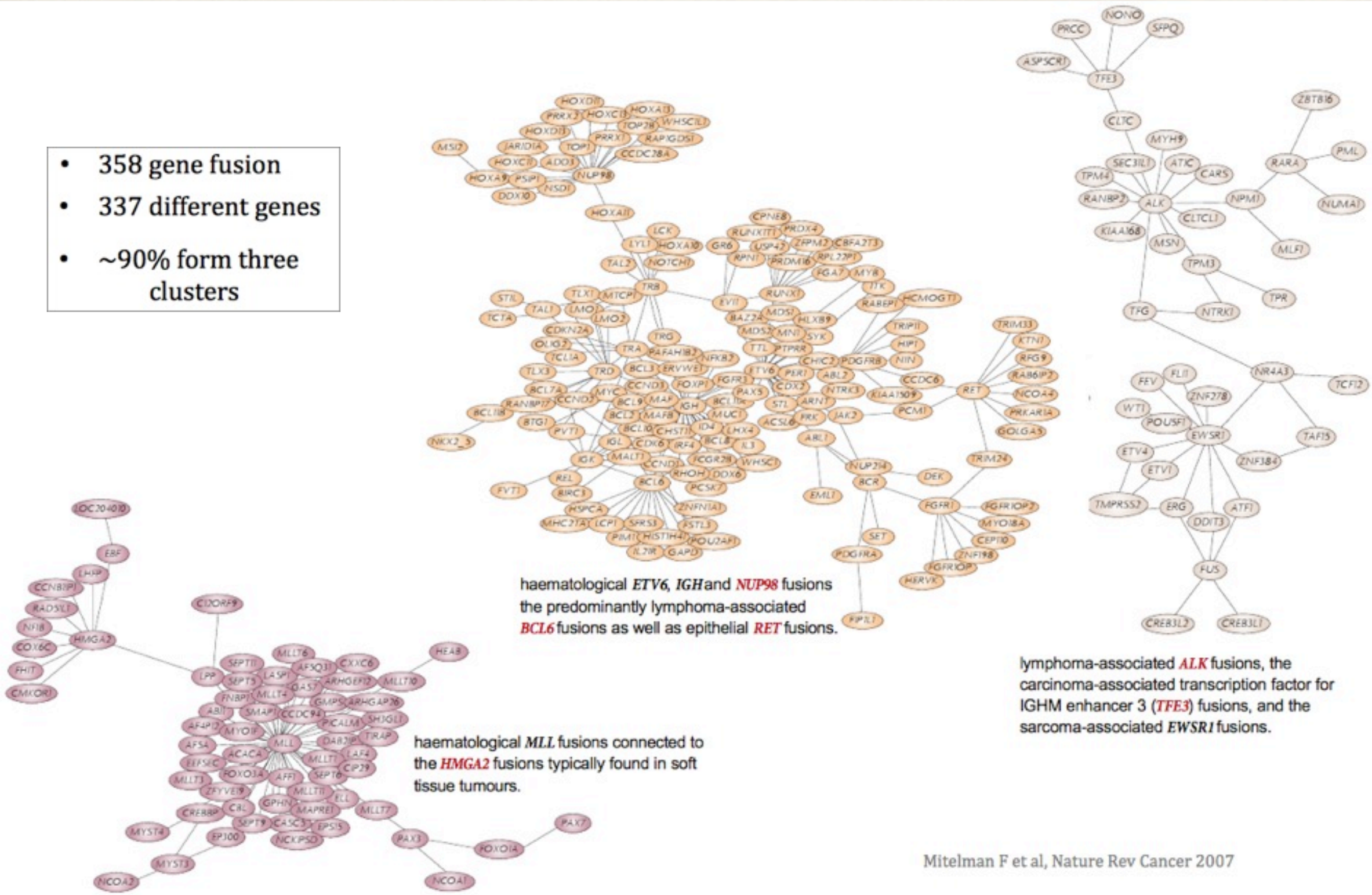
Why are they (trans-splicing events) important?



Horiuchi, Takayuki, and Toshiro Aigaki. Biology of the Cell 98, no. 2 (January 9, 2012): 135–140.

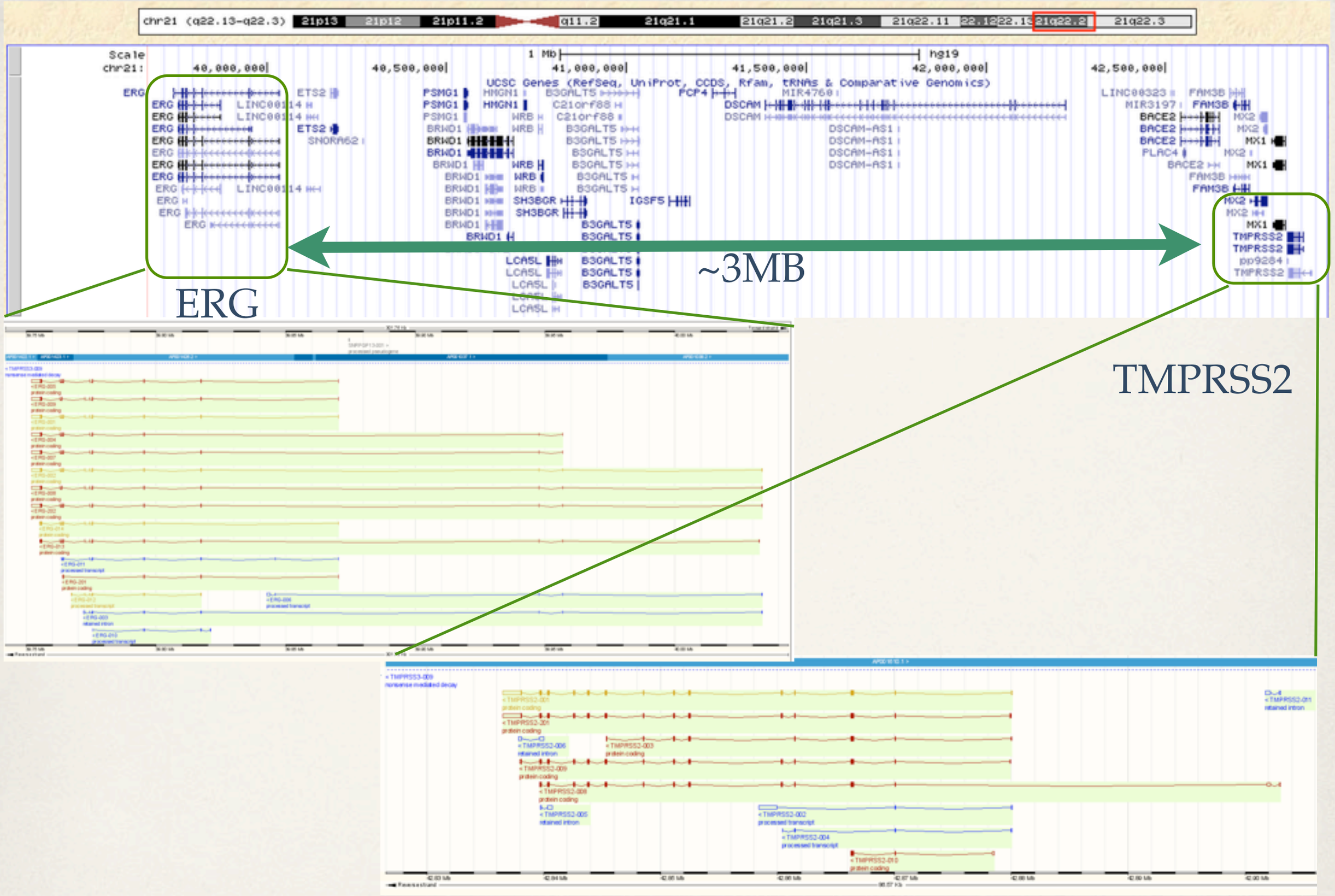
- Trans-splicing was initially found in lower eukariotes, such as trypanosomes and worms
  - Short sequences of nucleotides are trans-spliced to distant 5' of many protein coding genes
- Recently, they were found in mammalian cells:
  - JAZF1-SUZ12 in endometrial stroma cells (Li et al. Science 2008)
  - SLC45A3-ELK4 in prostate tissues (Rickman et al. Cancer Res 2009)
- 65% of protein-coding genes have distal 5' transcription start sites (ENCODE pilot)

- 358 gene fusion
- 337 different genes
- ~90% form three clusters



How many different gene fusions do we know?

What about prostate cancer?



What about prostate cancer?

Why are fusions important in prostate cancer?

Why are fusions important in prostate cancer?

- Prostate cancer is the most common tumor and second leading cause of death among men in the U.S.

Why are fusions important in prostate cancer?

- Prostate cancer is the most common tumor and second leading cause of death among men in the U.S.
- Prostate Specific Antigen (PSA) screening helped the early diagnosis of prostate cancer

Why are fusions important in prostate cancer?

- Prostate cancer is the most common tumor and second leading cause of death among men in the U.S.

- Prostate Specific Antigen (PSA) screening helped the early diagnosis of prostate cancer

- Most men have a slowly progressing tumor

Why are fusions important in prostate cancer?

- Prostate cancer is the most common tumor and second leading cause of death among men in the U.S.

- Prostate Specific Antigen (PSA) screening helped the early diagnosis of prostate cancer

- Most men have a slowly progressing tumor

- No clear mortality benefit from PSA screening

Why are fusions important in prostate cancer?

- Prostate cancer is the most common tumor and second leading cause of death among men in the U.S.

Mortality Results from a Randomized Prostate-Cancer Screening Trial

Andriole G et al. N Engl J Med 2009;360:1310-9.

- Prostate Specific Antigen (PSA) screening helped the early diagnosis of prostate cancer

EDITORIAL

The Prostate Cancer Muddle

Published: March 19, 2009 The New York Times

- Most men have a slowly progressing tumor

Screening and Prostate-Cancer Mortality in a Randomized European Study

Shroeder FH et al. N Engl J Med 2009;360:1320-8.

- No clear mortality benefit from PSA screening

Why are fusions important in prostate cancer?

- Prostate cancer is the most common tumor and second leading cause of death among men in the U.S.
- Prostate Specific Antigen (P… screening hel…
- …ost men have a slowly progressing tumor
- No clear mortality benefit from PSA screening

Mortality Results from a Randomized Prostate-Cancer Screening Trial

Andriole G et al. N Engl …

…e Prostate Cancer Muddle

Published: March 19, 2009 The New York Times

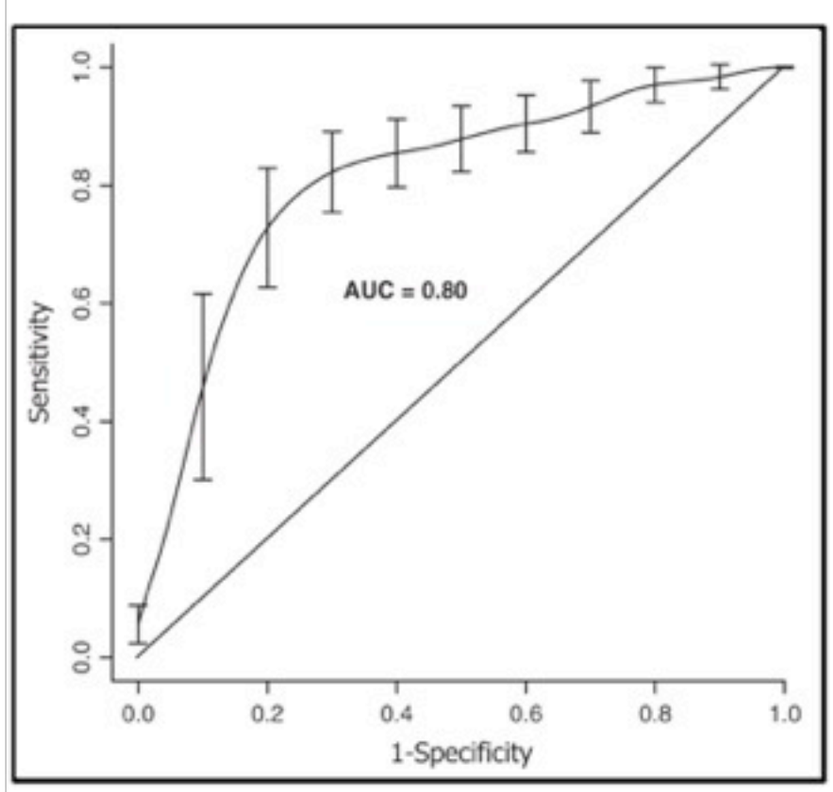
Screening and Prostate-Cancer Mortality in a Randomized European Study

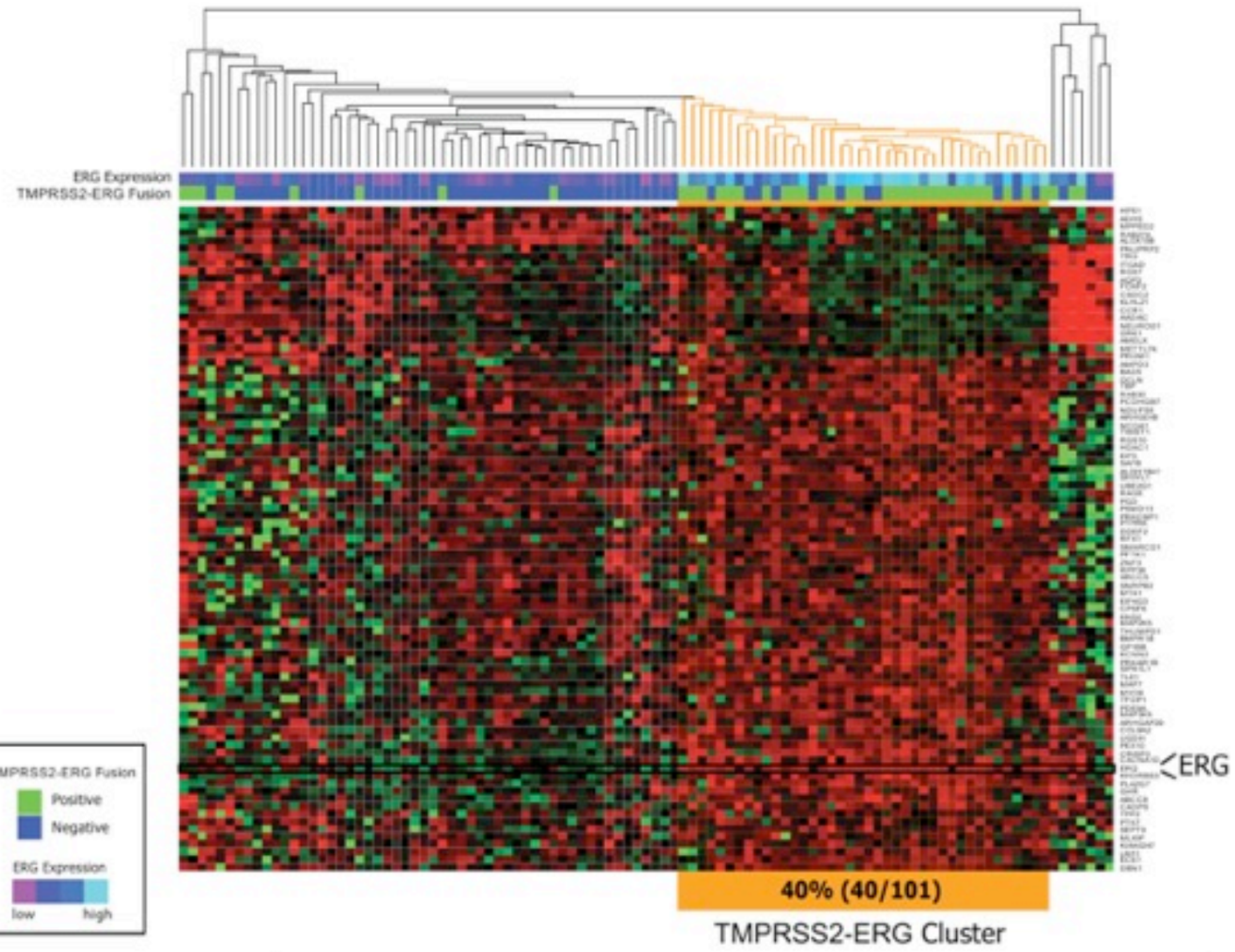
Shroeder FH et al. N Engl J Med 2009;360:1320-8.

Are fusions important for outcome prediction?

ERG rearranged cases as a subclass of prostate cancer?

ERG rearrangement status:
87-gene signature

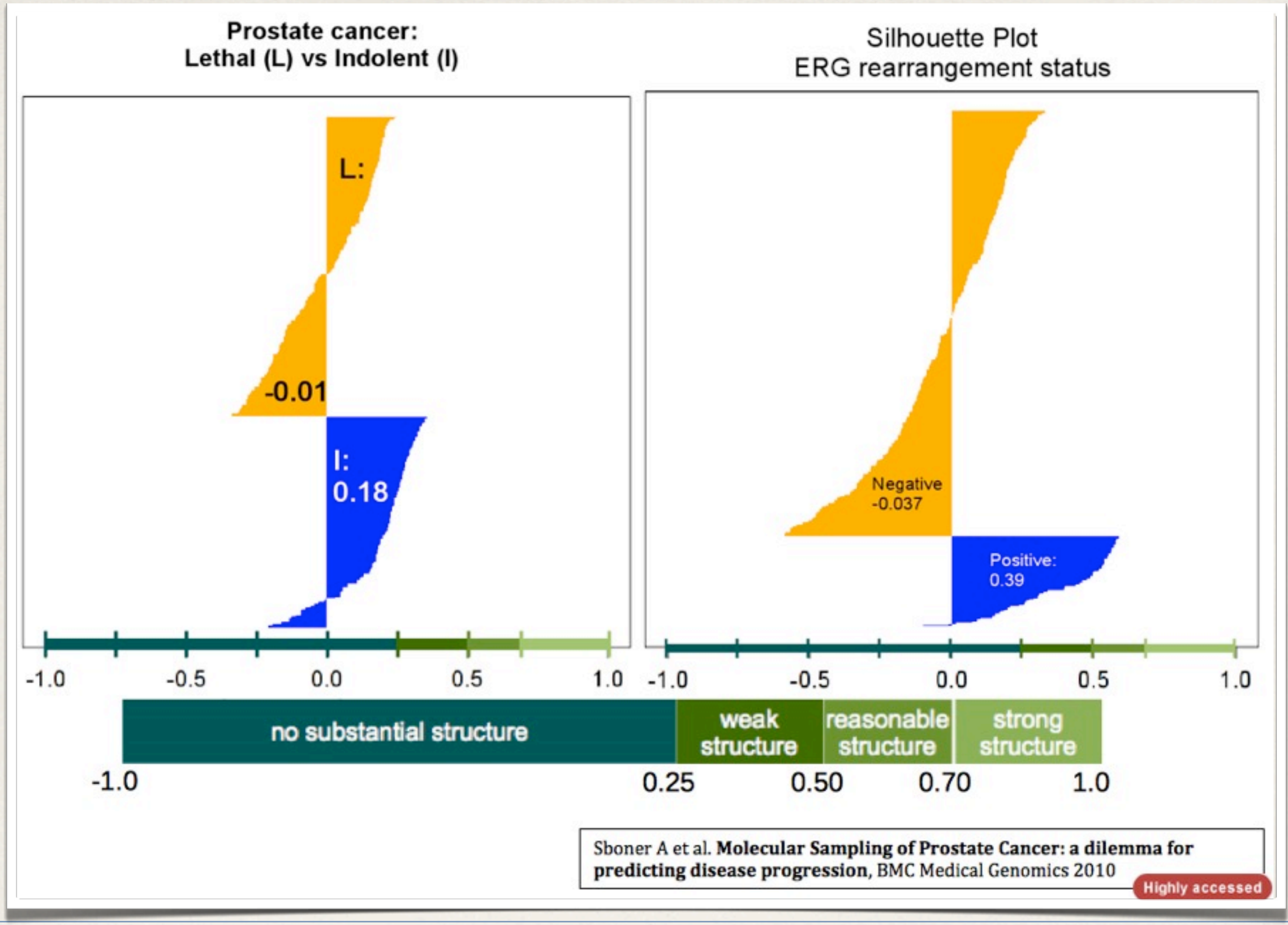




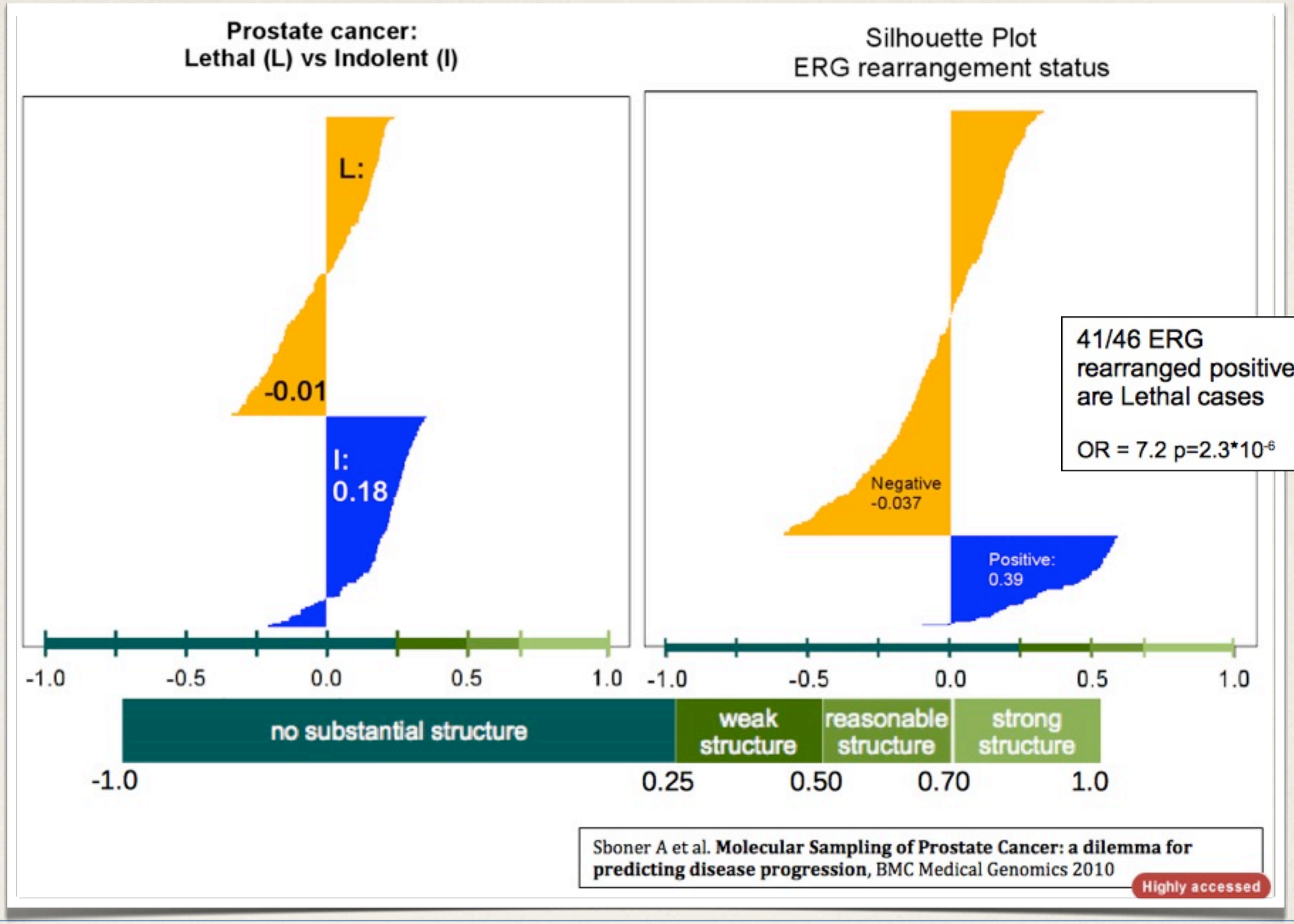
Estrogen-Dependent Signaling in a Molecularly Distinct Subclass of Aggressive Prostate Cancer

J Natl Cancer Inst 2008;100:815–825

ERG rearranged subclass is more homogeneous than lethal/indolent PCa



ERG rearranged subclass is more homogeneous than lethal/indolent PCa



ERG rearranged subclass is more homogeneous than lethal/indolent PCa

Are gene fusions key elements in prostate cancer?

- *Hypothesis I*: additional gene fusions may be present in prostate cancer

- *Hypothesis II*: can those fusions help better define the molecular landscape of disease development and progression?

Identification of gene fusions

Identification of gene fusions

- Traditional detection of fusion genes typically involves cytogenetic methods

Identification of gene fusions

- Traditional detection of fusion genes typically involves cytogenetic methods
- Some require a hypothesis about the genes involved in the fusion

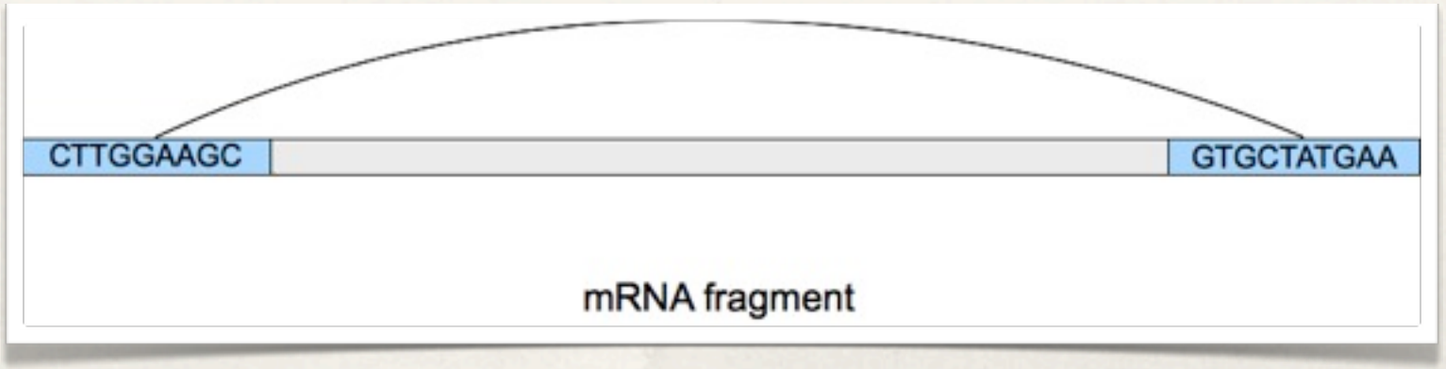
Identification of gene fusions

- Traditional detection of fusion genes typically involves cytogenetic methods

- Some require a hypothesis about the genes involved in the fusion

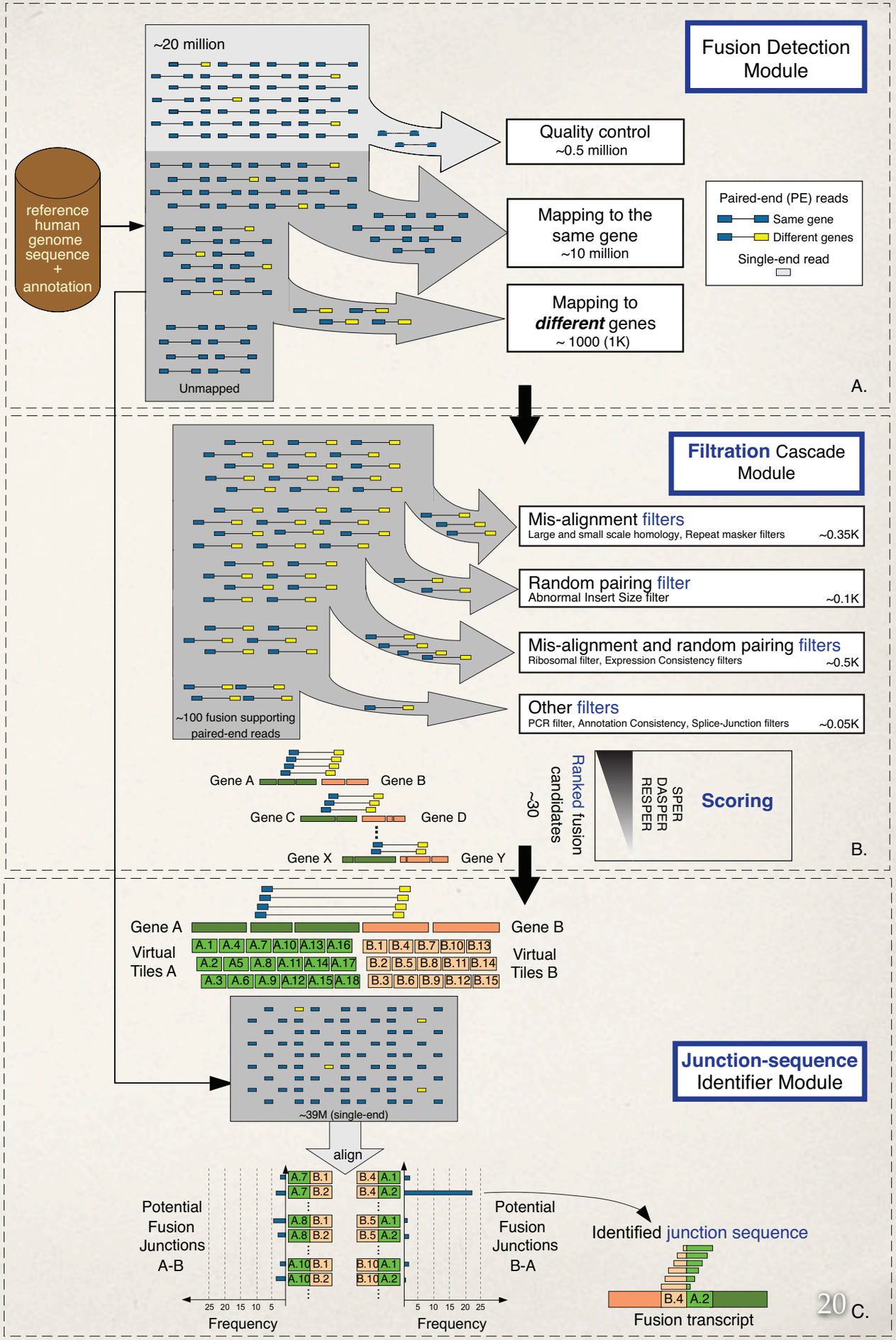
- Next-generation sequencing can address this challenge directly, especially with:

  - *Paired-end* RNA-Seq: keeping connectivity information



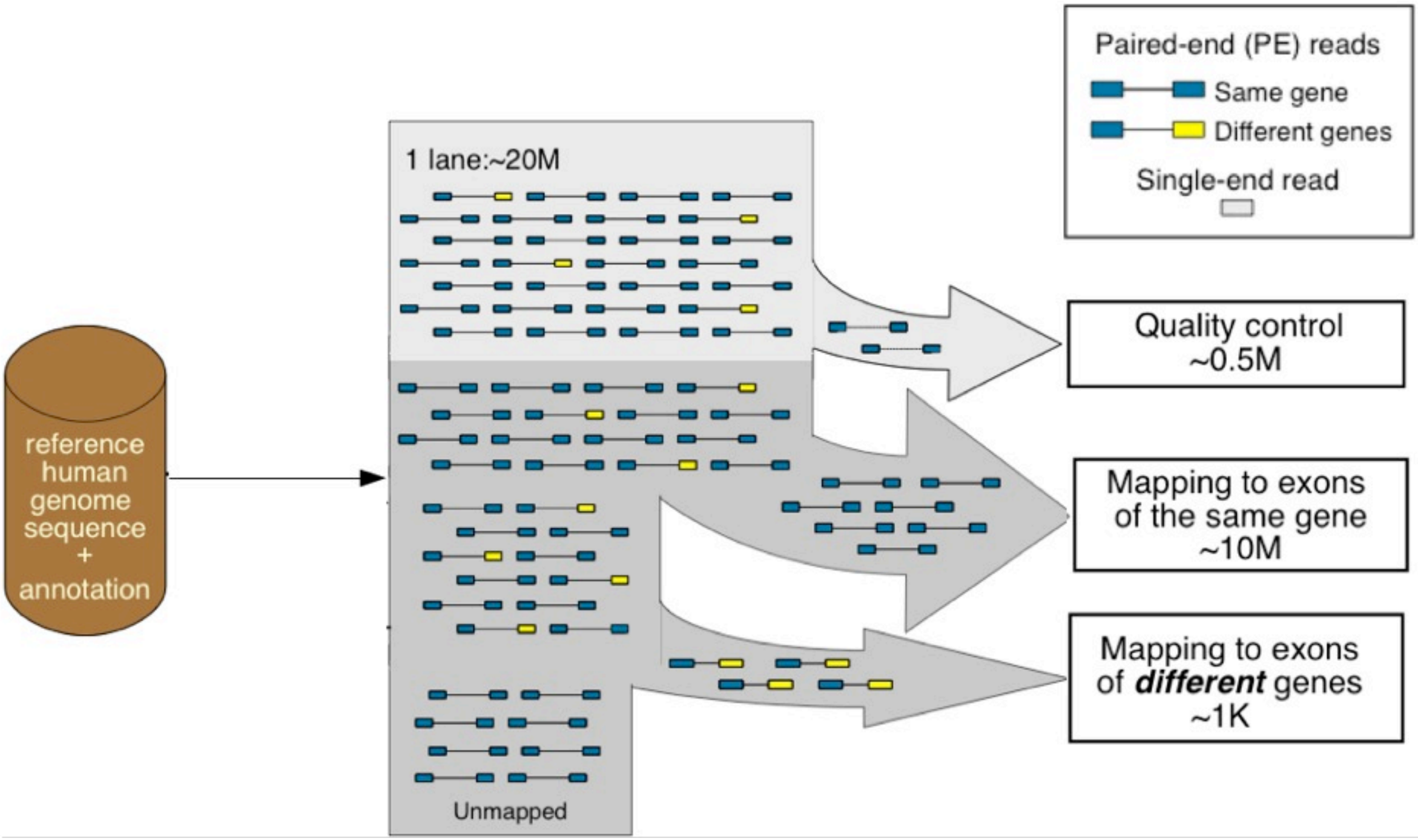
FusionSeq: a computational modular framework

- Fusion Detection Module
- Filtration Cascade Module
- Junction-sequence identifier module



Sboner, Andrea, et al. "FusionSeq: a Modular Framework for Finding Gene Fusions by Analyzing Paired-End RNA-Sequencing Data." Genome Biology 11, no. 10 (2010): R104.

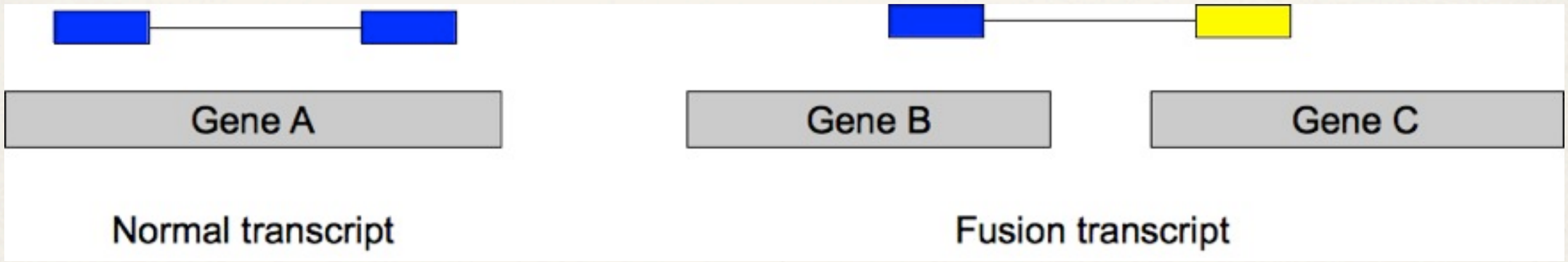
Highly accessed

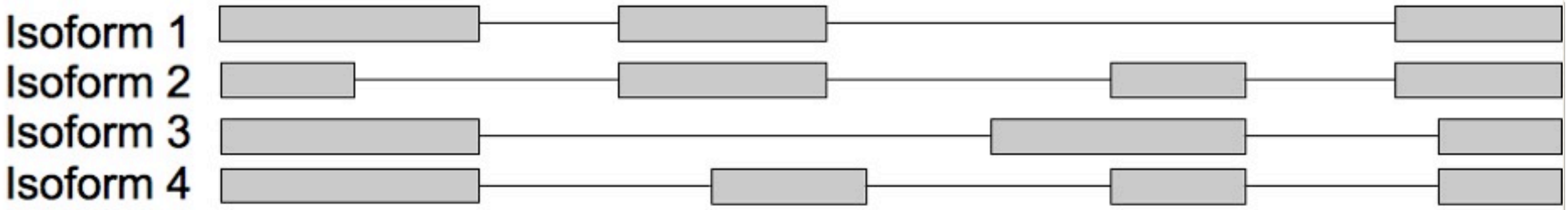


Fusion Detection Module

How to identify fusion transcripts?

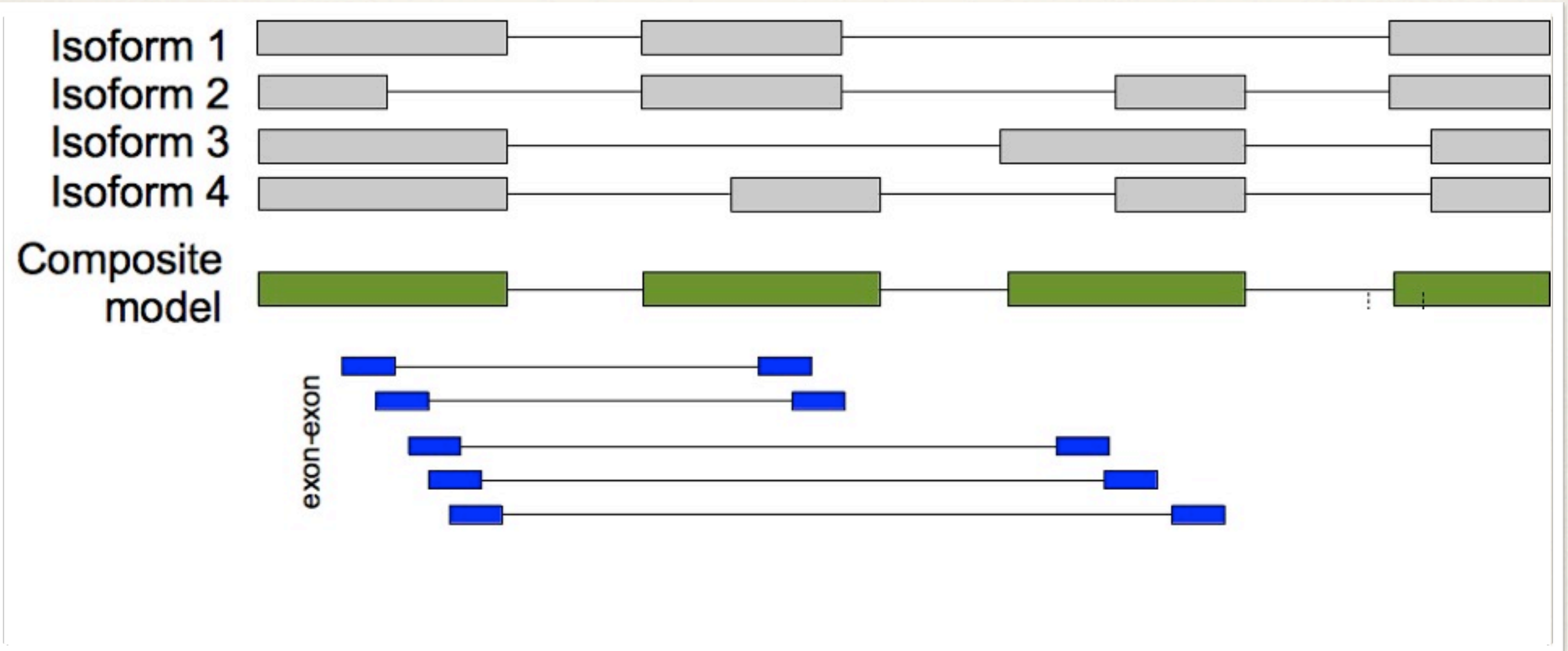
- *Straightforward*:
  - If the two ends map to different genes, then we have a potential fusion transcript



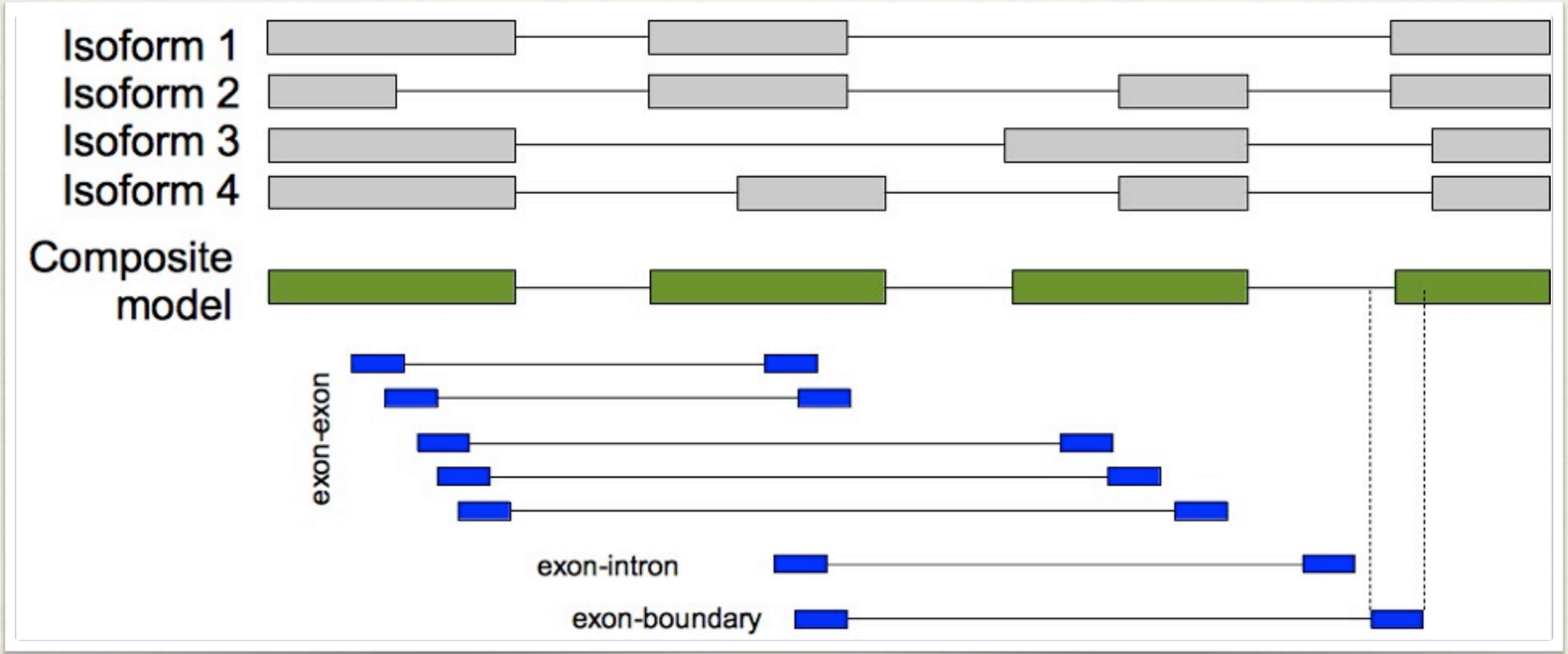


What about different isoforms?

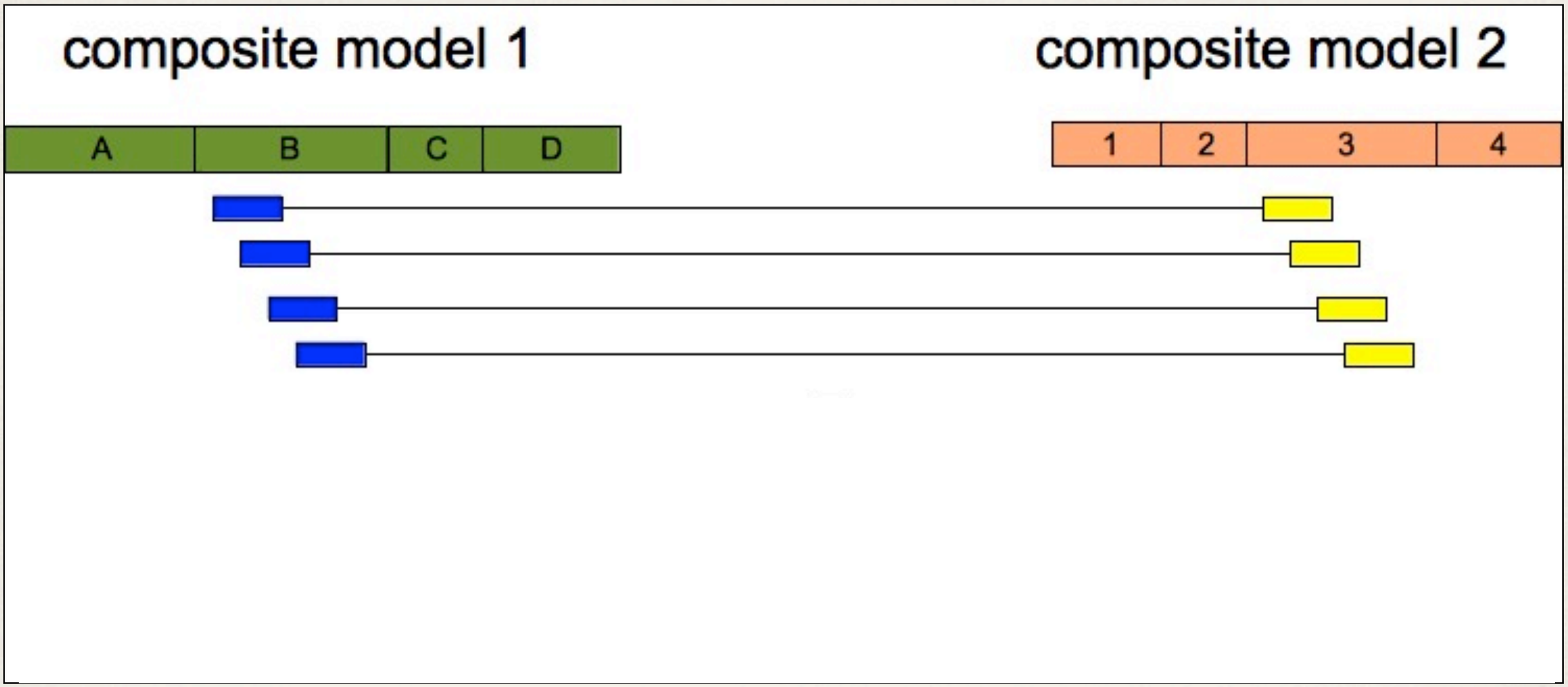
What about different isoforms?



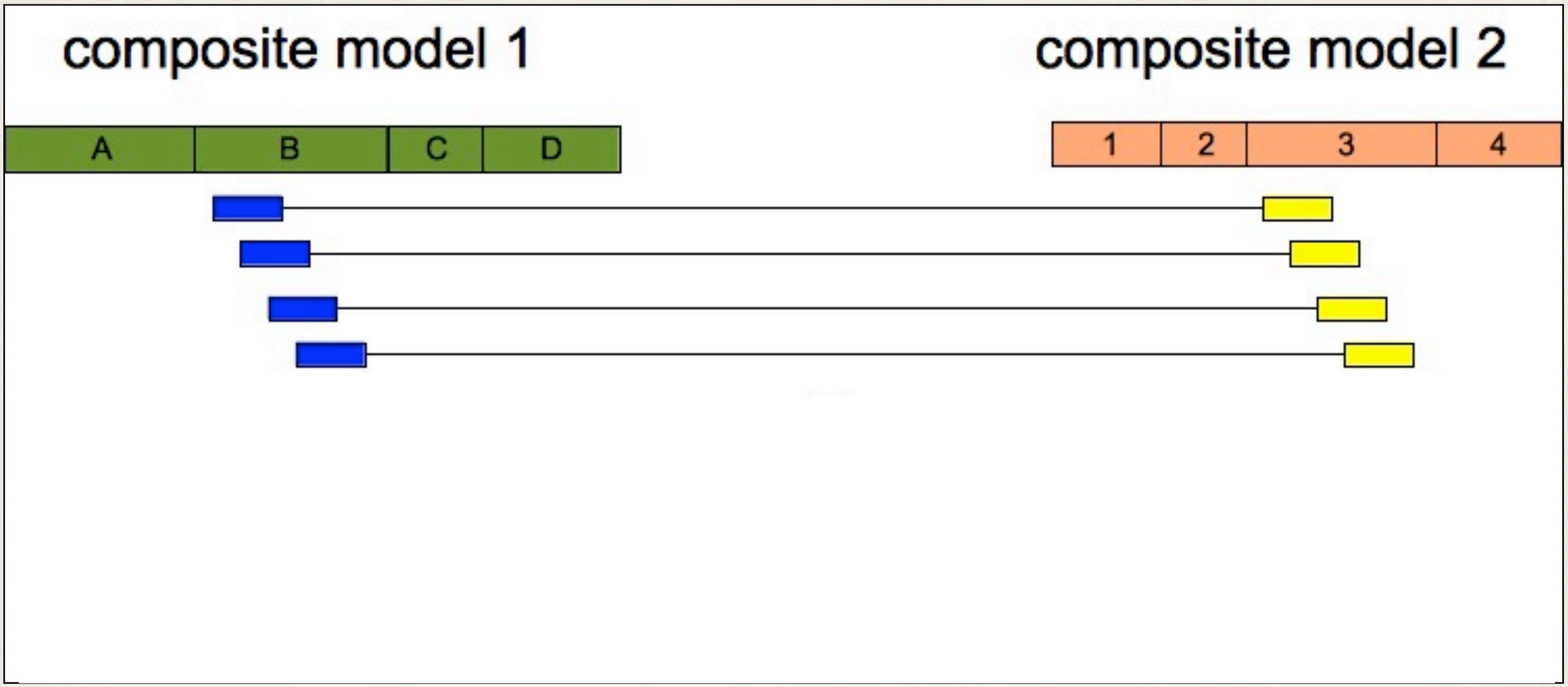
What about different isoforms?



What about different isoforms?

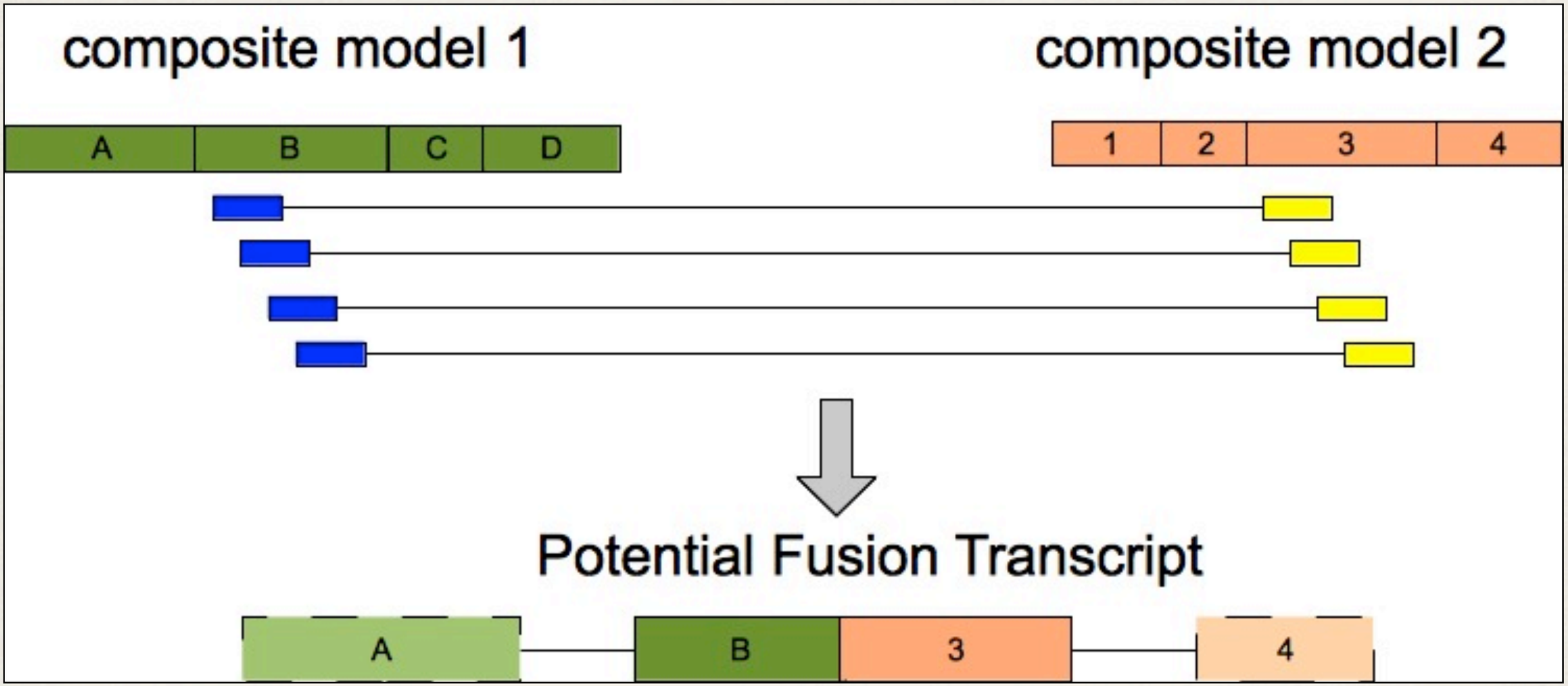


Finding fusion transcripts



- Each PE read can be assigned to one “gene”

Finding fusion transcripts



- Each PE read can be assigned to one “gene”
- *Potential Fusion Transcripts*: if pair belong to different genes

Finding fusion transcripts

Not an ideal world: sources of errors

Not an ideal world: sources of errors

- *Mis-alignments*
  - Base caller error
  - SNPs
  - RNA editing
  - Sequence similarity (paralogs, pseudogenes)

Not an ideal world: sources of errors

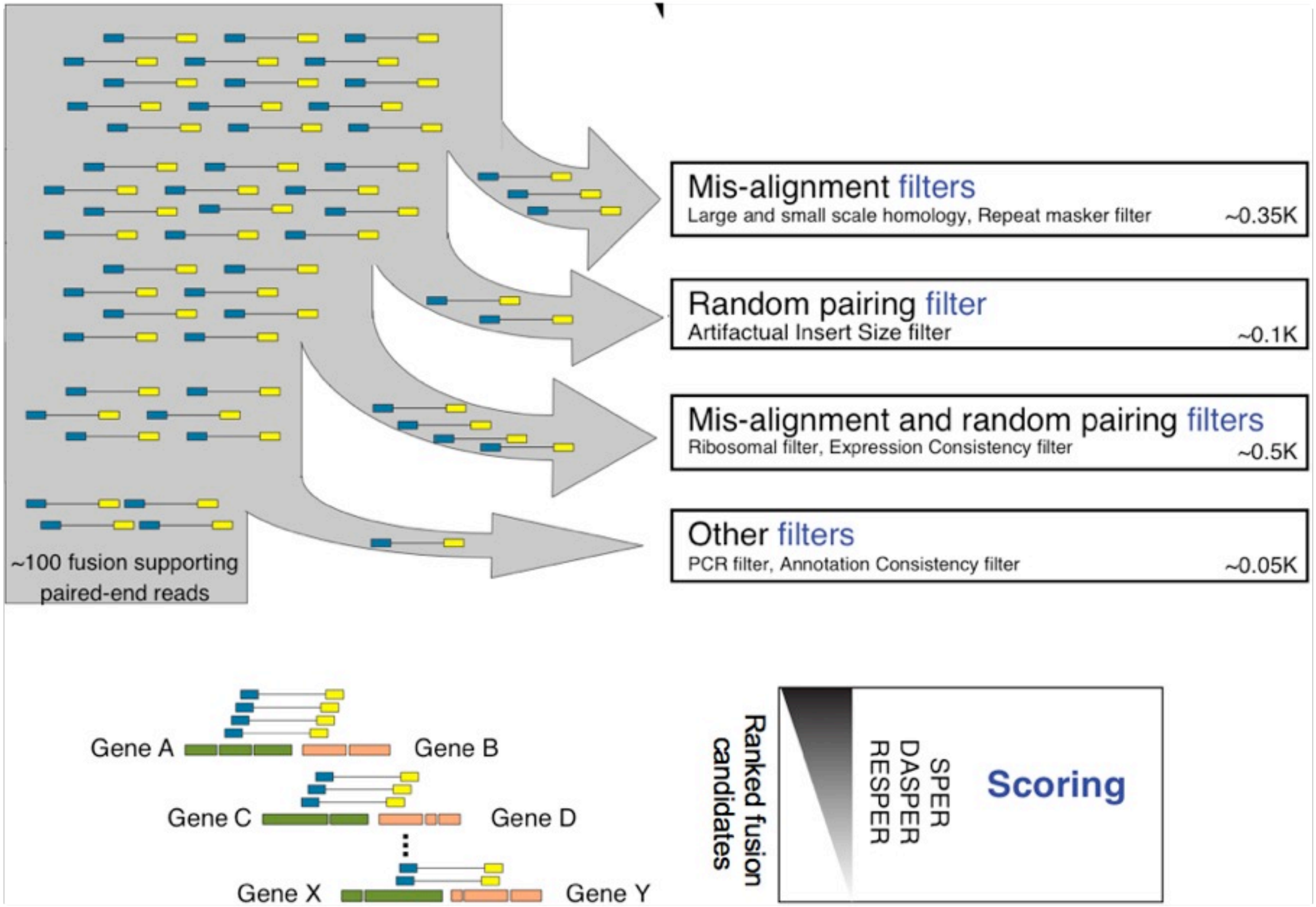
- *Mis-alignments*
  - Base caller error
  - SNPs
  - RNA editing
  - Sequence similarity (paralogs, pseudogenes)
- *Random pairing of transcript fragments*
  - Library preparation

Not an ideal world: sources of errors

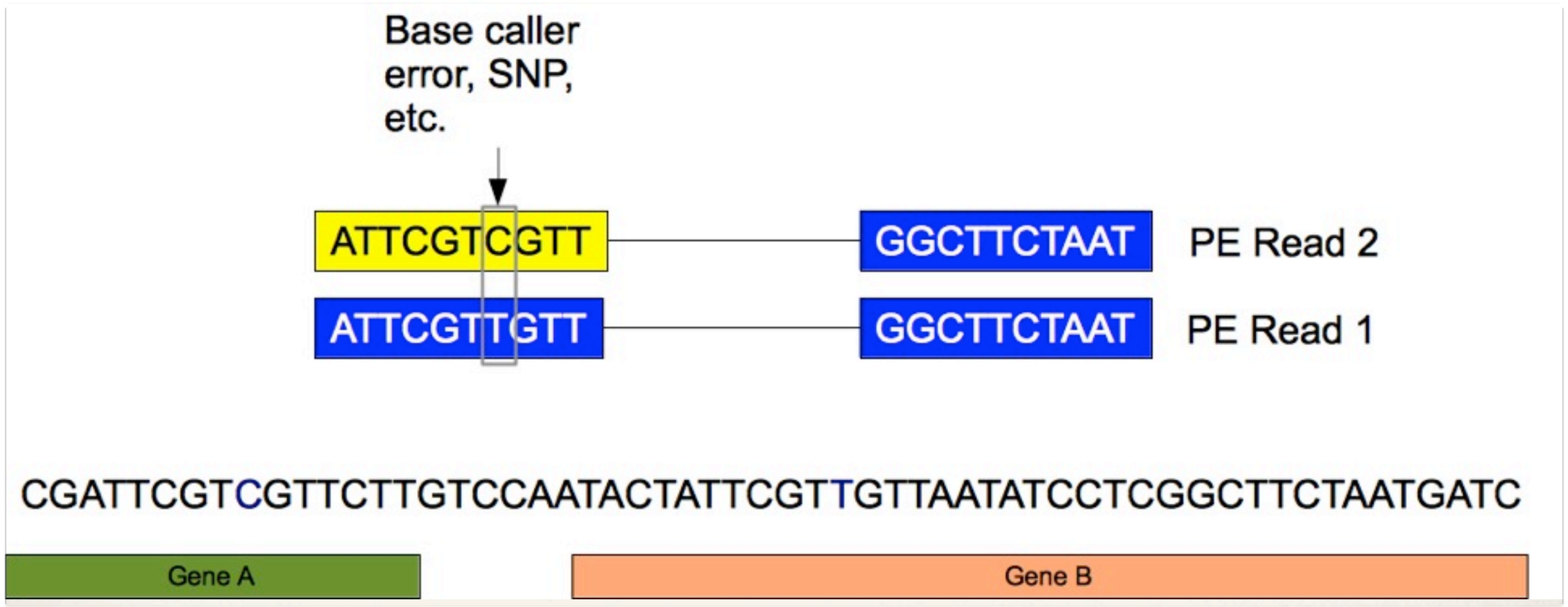
- *Mis-alignments*
  - Base caller error
  - SNPs
  - RNA editing
  - Sequence similarity (paralogs, pseudogenes)
- *Random pairing of transcript fragments*
  - Library preparation
- *Combination of mis-alignment and random pairing*

Not an ideal world: sources of errors

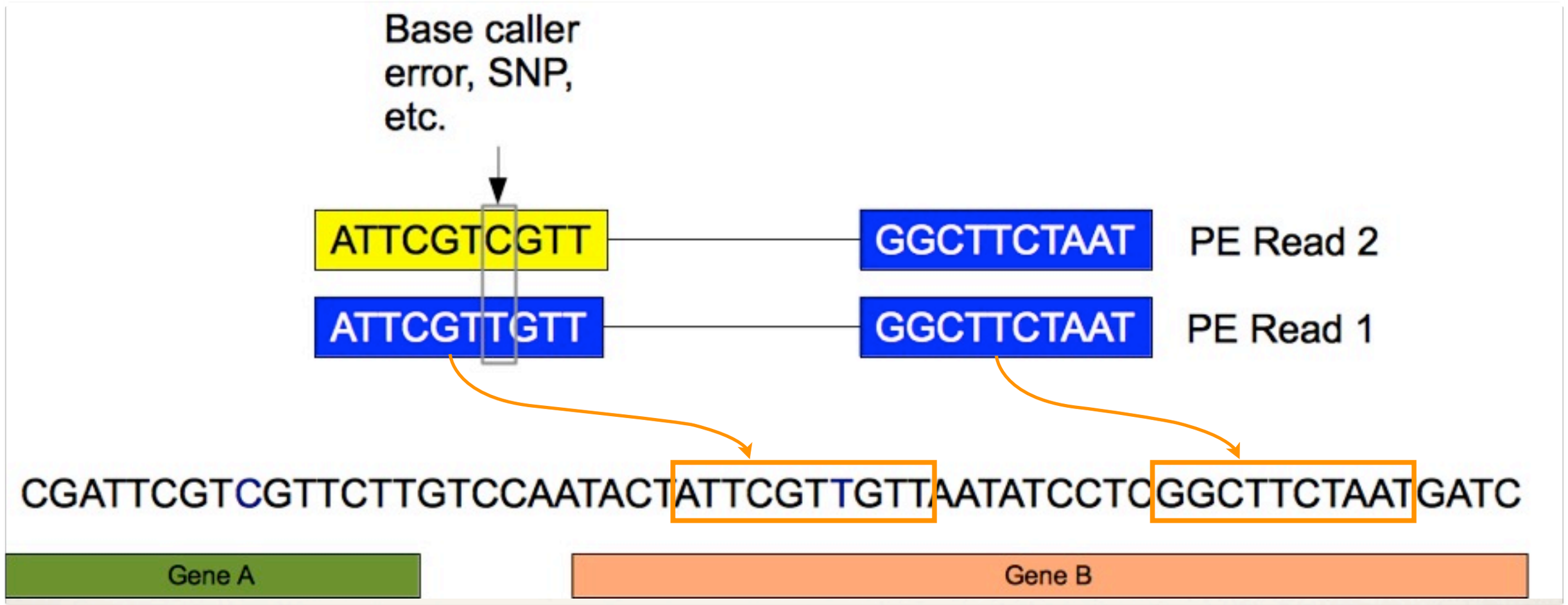
- *Mis-alignments*
  - Base caller error
  - SNPs
  - RNA editing
  - Sequence similarity (paralogs, pseudogenes)
- *Random pairing of transcript fragments*
  - Library preparation
- *Combination of mis-alignment and random pairing*
- *PCR amplification, gene annotation inconsistencies/incompleteness*



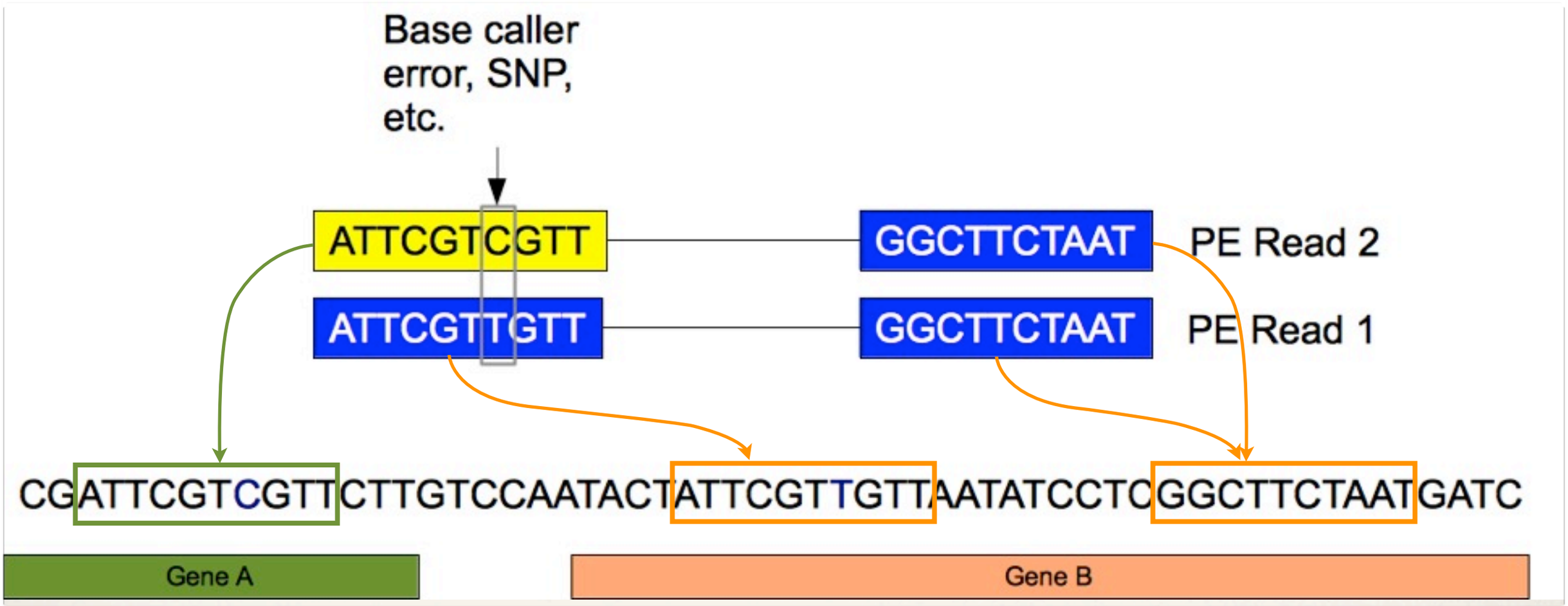
Filtration Cascade Module



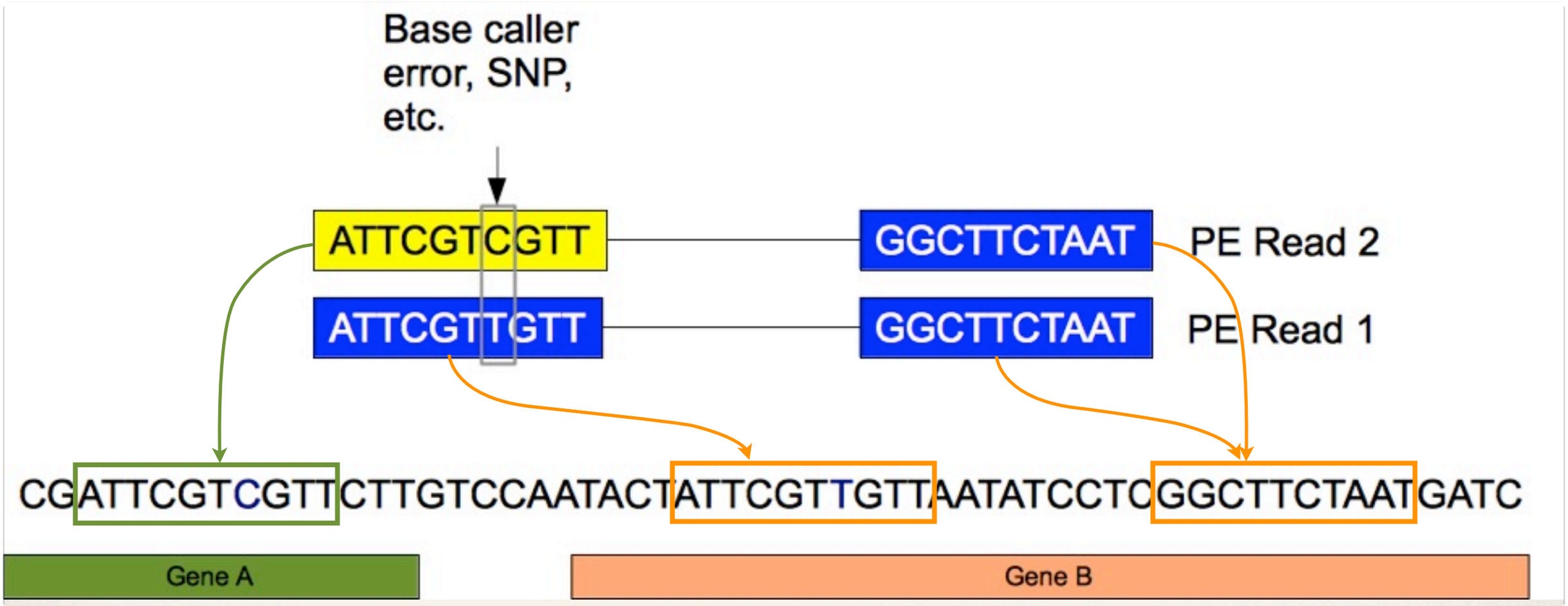
Mis-alignment



Mis-alignment

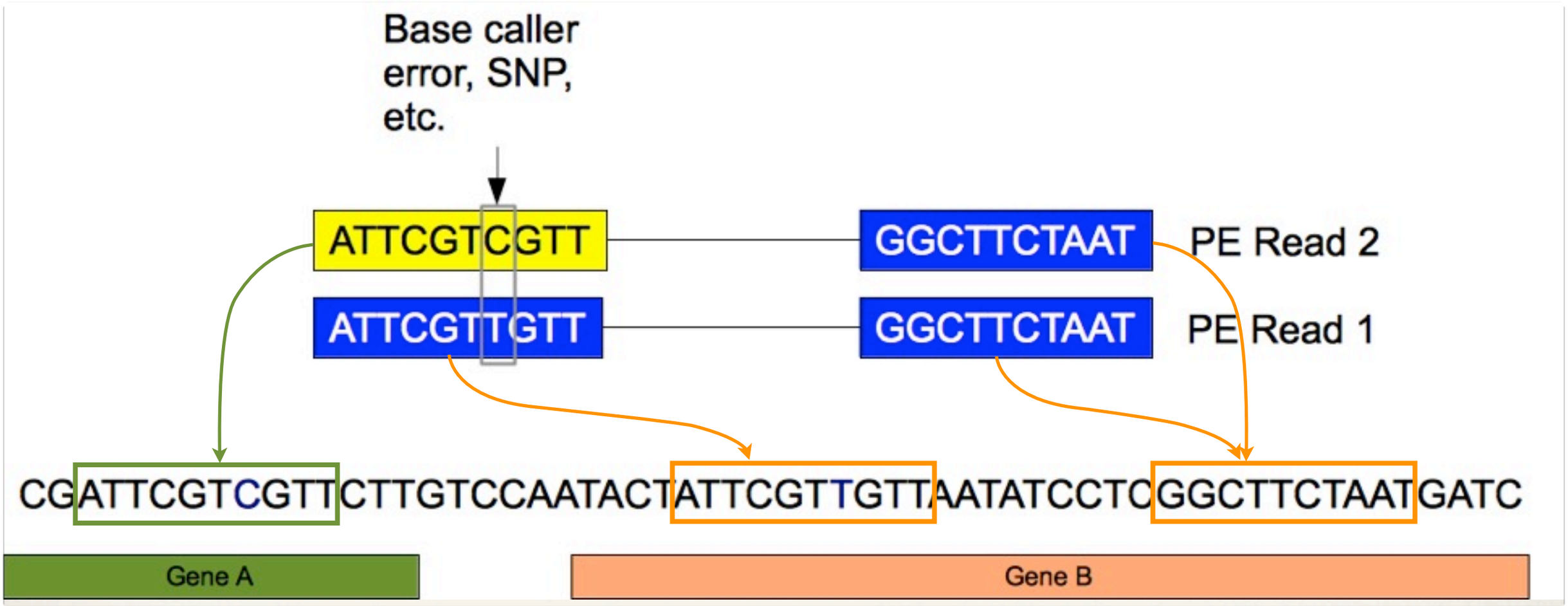


Mis-alignment



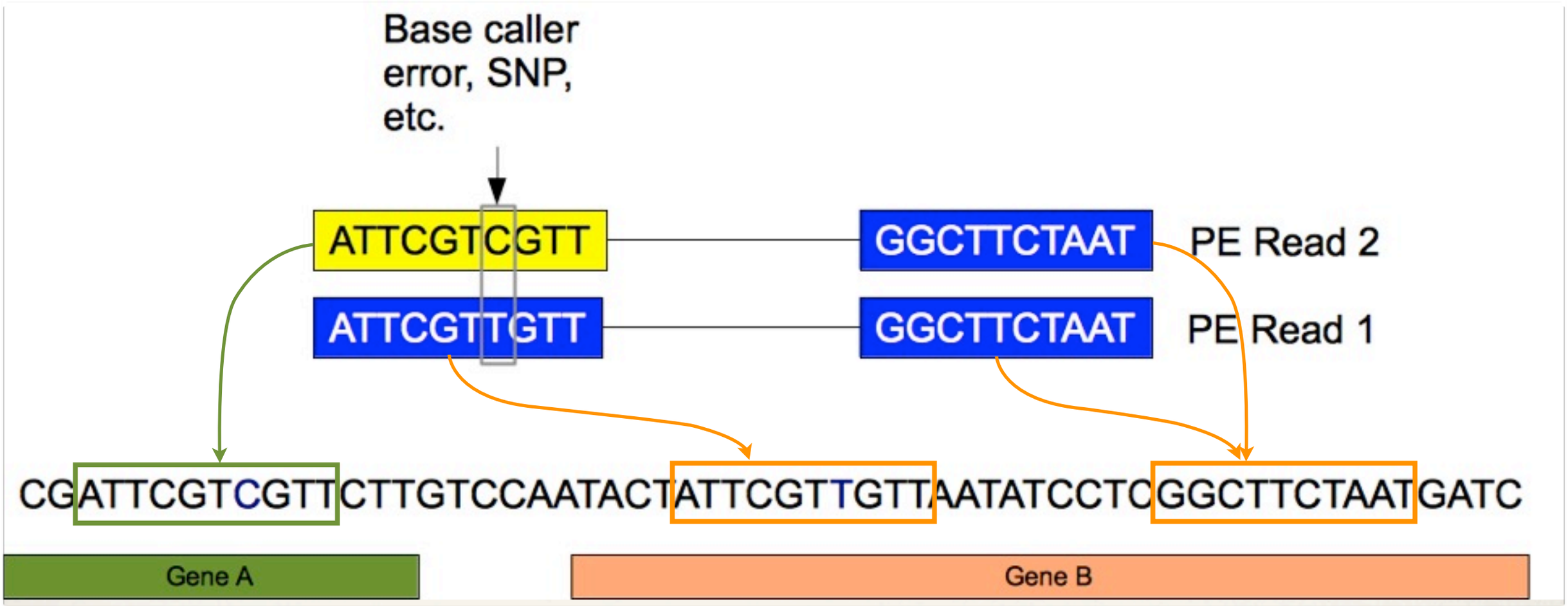
- *Large-scale*: similarity at the gene level → TreeFam to exclude paralogs

Mis-alignment



- *Large-scale*: similarity at the gene level → TreeFam to exclude paralogs
- *Small-scale*: similarity of smaller regions within the genes → 2-step approach using bowtie and BLAT

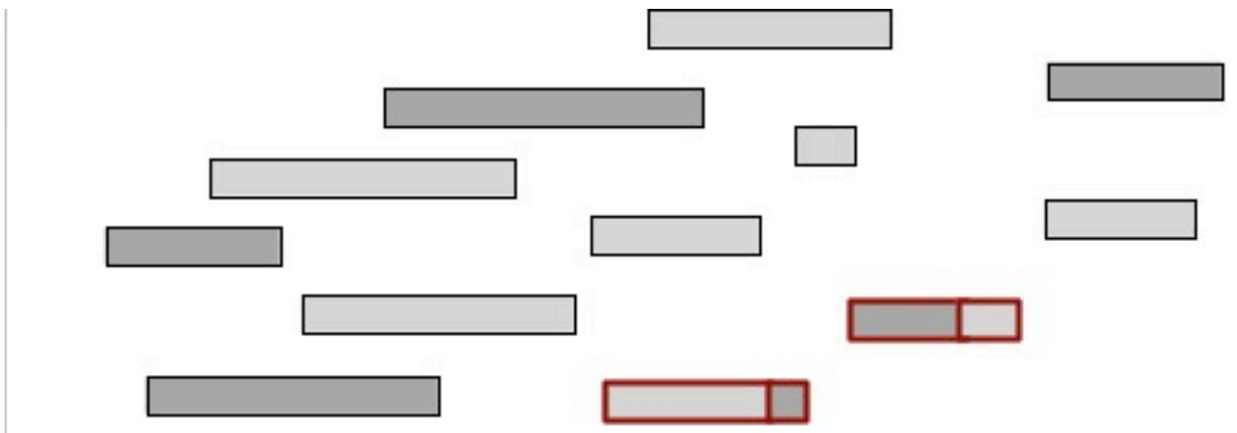
Mis-alignment



- *Large-scale*: similarity at the gene level → TreeFam to exclude paralogs
- *Small-scale*: similarity of smaller regions within the genes → 2-step approach using bowtie and BLAT
- *Repetitive region*: reads mapping to low-sequence complexity

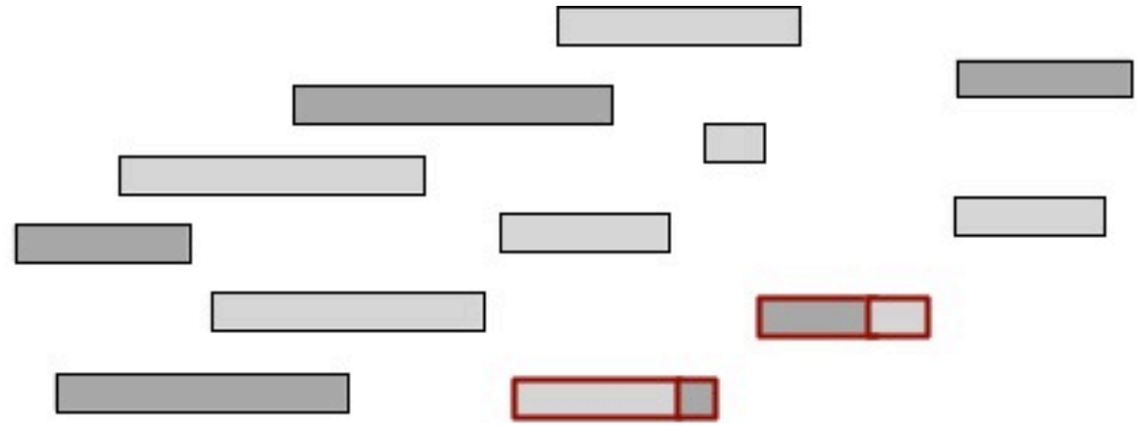
Mis-alignment

Random pairing of transcript fragments



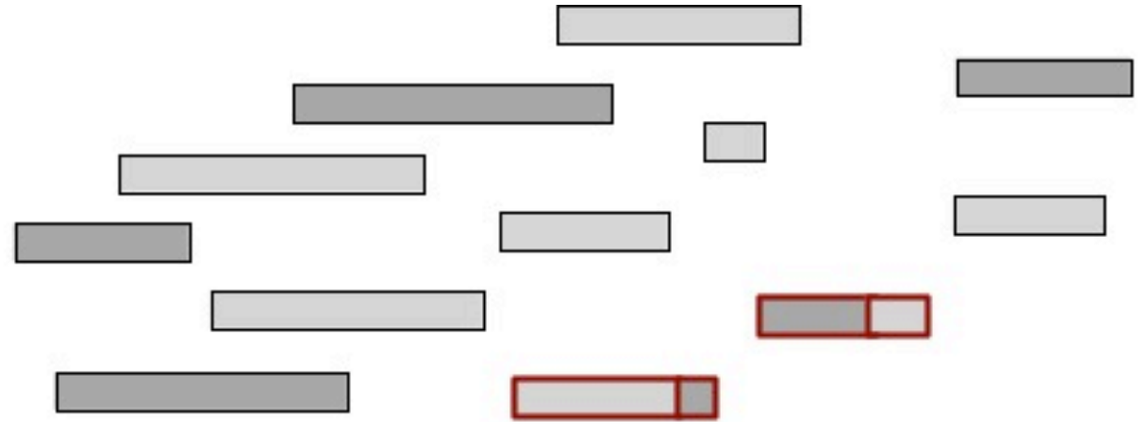
Random pairing of transcript fragments

- Inefficient A-tailing → random joining of transcript fragments → artificial chimeric transcripts



Random pairing of transcript fragments

- Inefficient A-tailing → random joining of transcript fragments → artificial chimeric transcripts
- How to remove them?

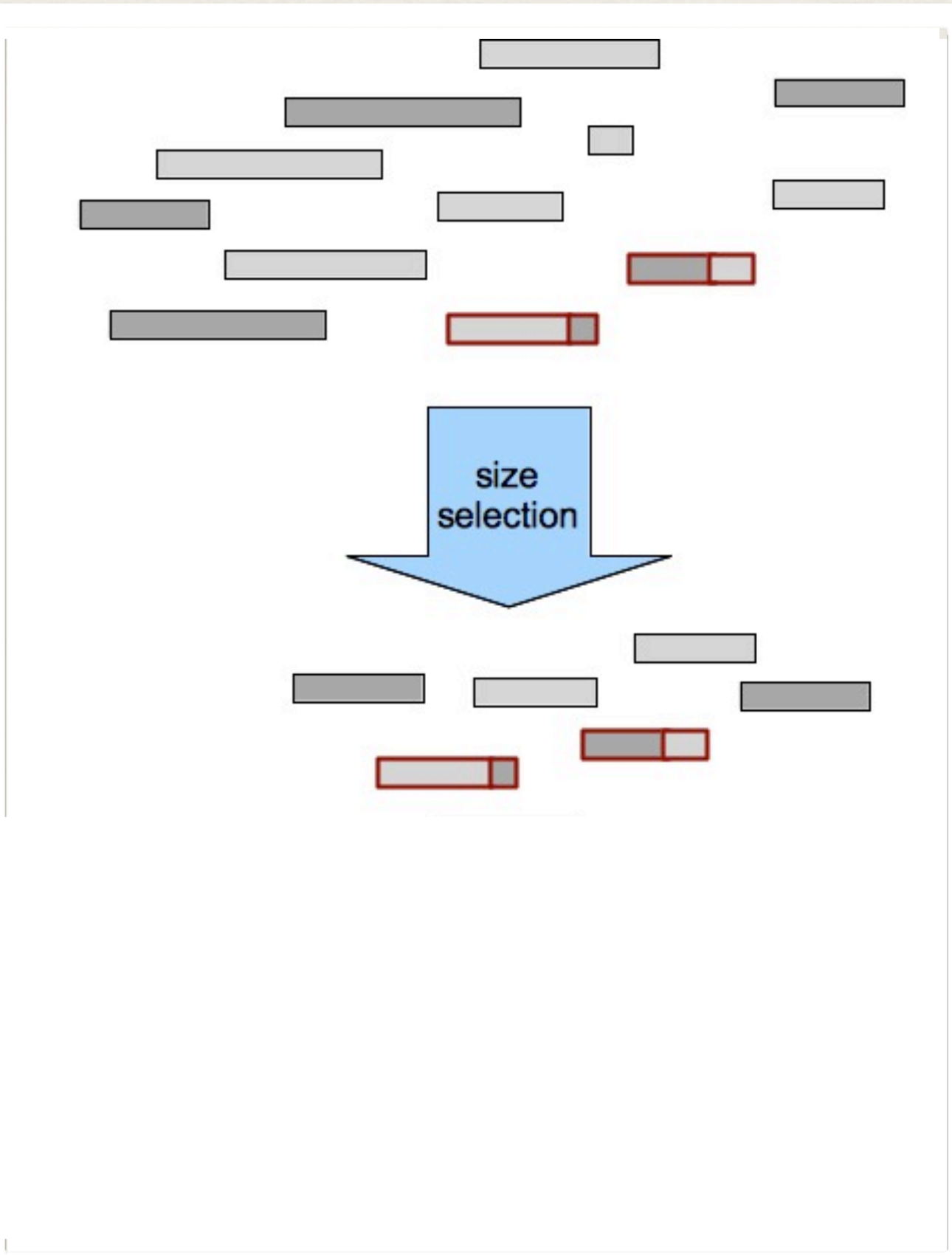


Random pairing of transcript fragments

- Inefficient A-tailing → random joining of transcript fragments → artificial chimeric transcripts

- How to remove them?

  - *Observation 1*: preparation protocol requires a size-selection step, i.e. all transcript fragments have similar size



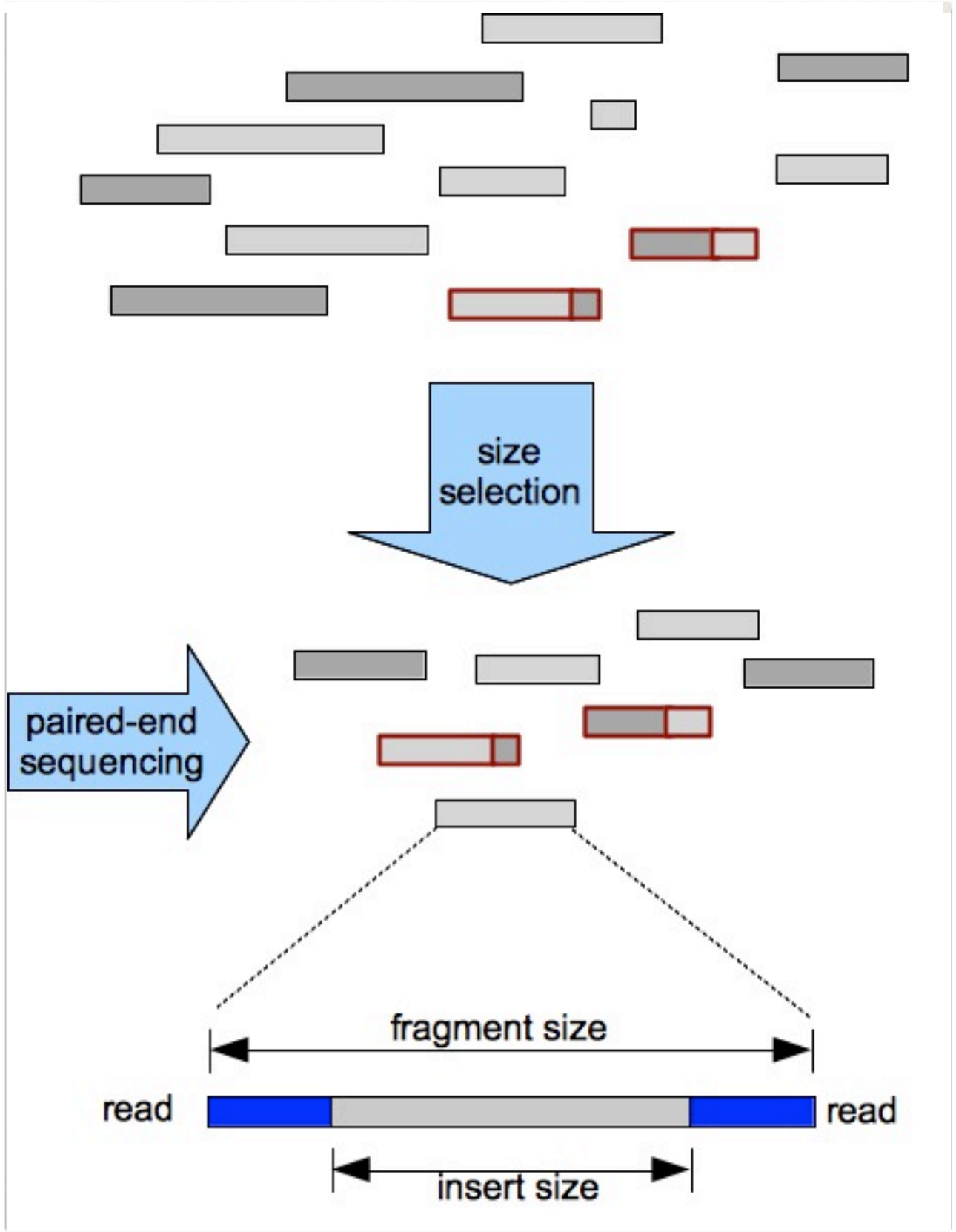
Random pairing of transcript fragments

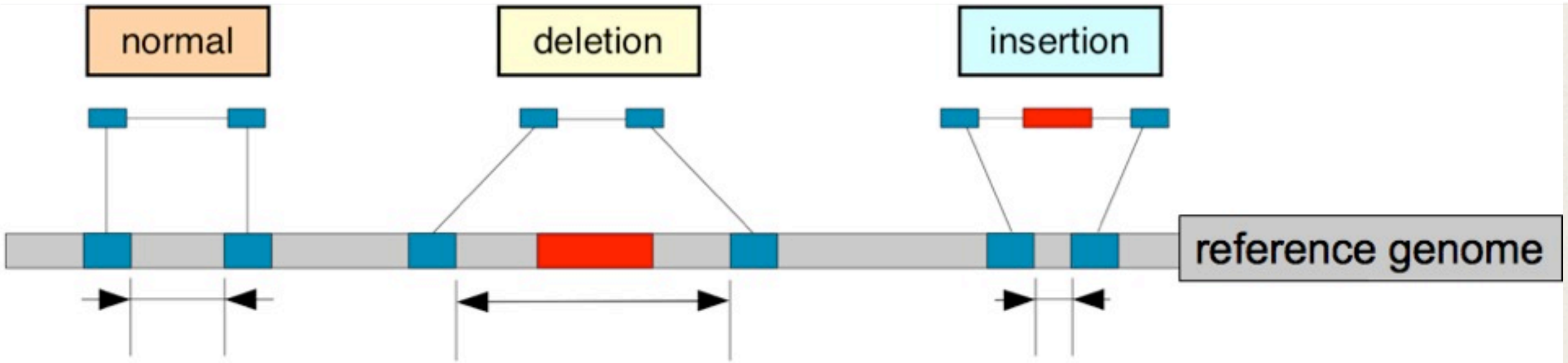
- Inefficient A-tailing → random joining of transcript fragments → artificial chimeric transcripts

- How to remove them?

  - *Observation 1*: preparation protocol requires a size-selection step, i.e. all transcript fragments have similar size

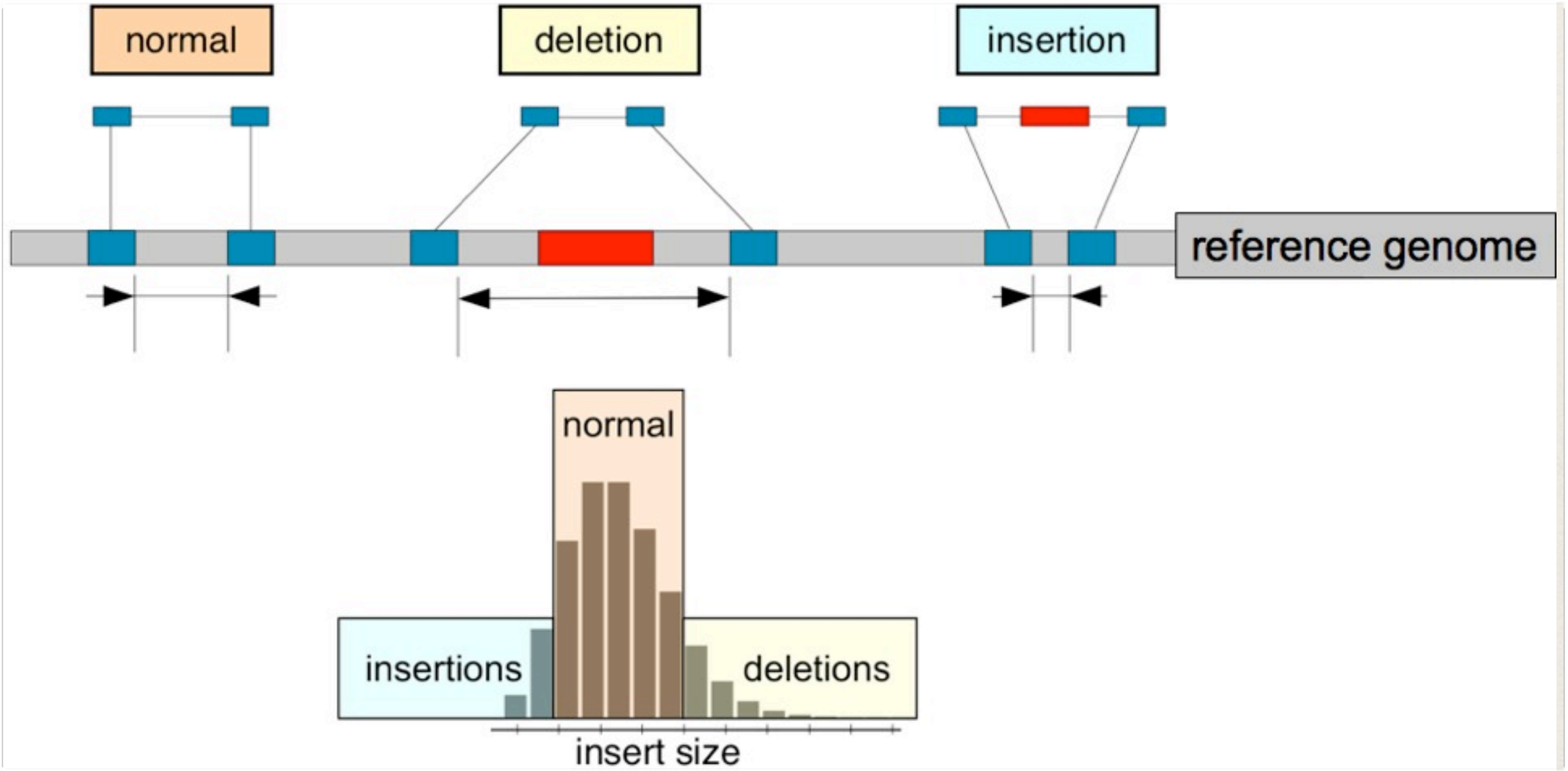
  - *Observation 2:* the insert-size distribution of the PE reads should be roughly the same for all PE reads





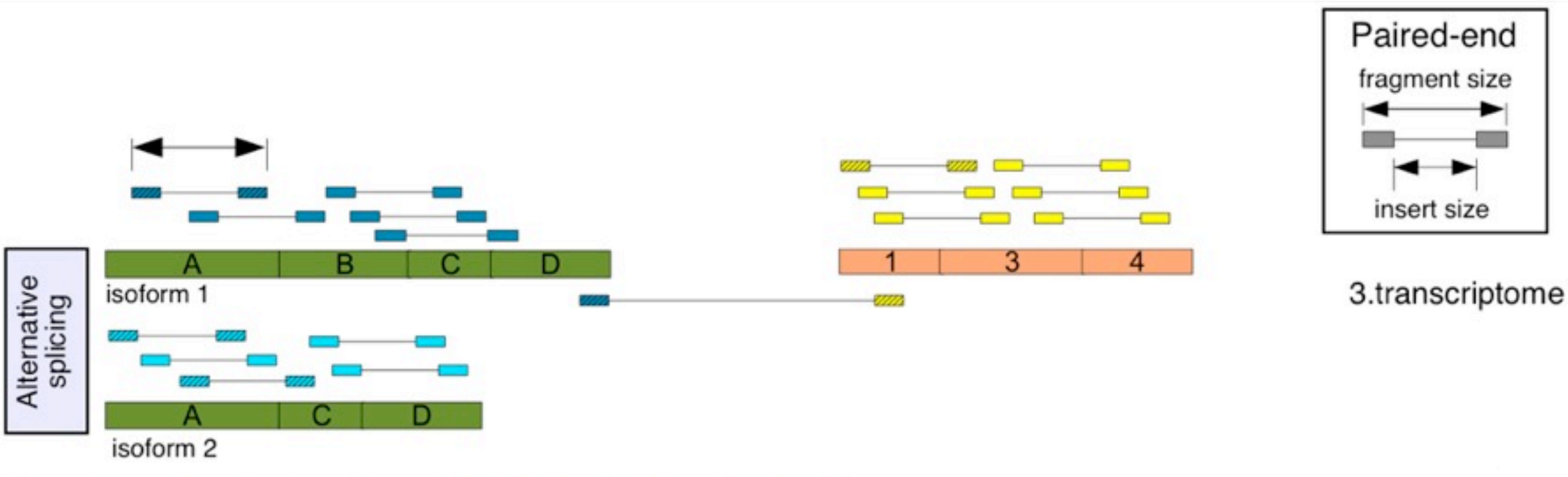
Insert-size distribution concept

Detecting structural variations (SVs)

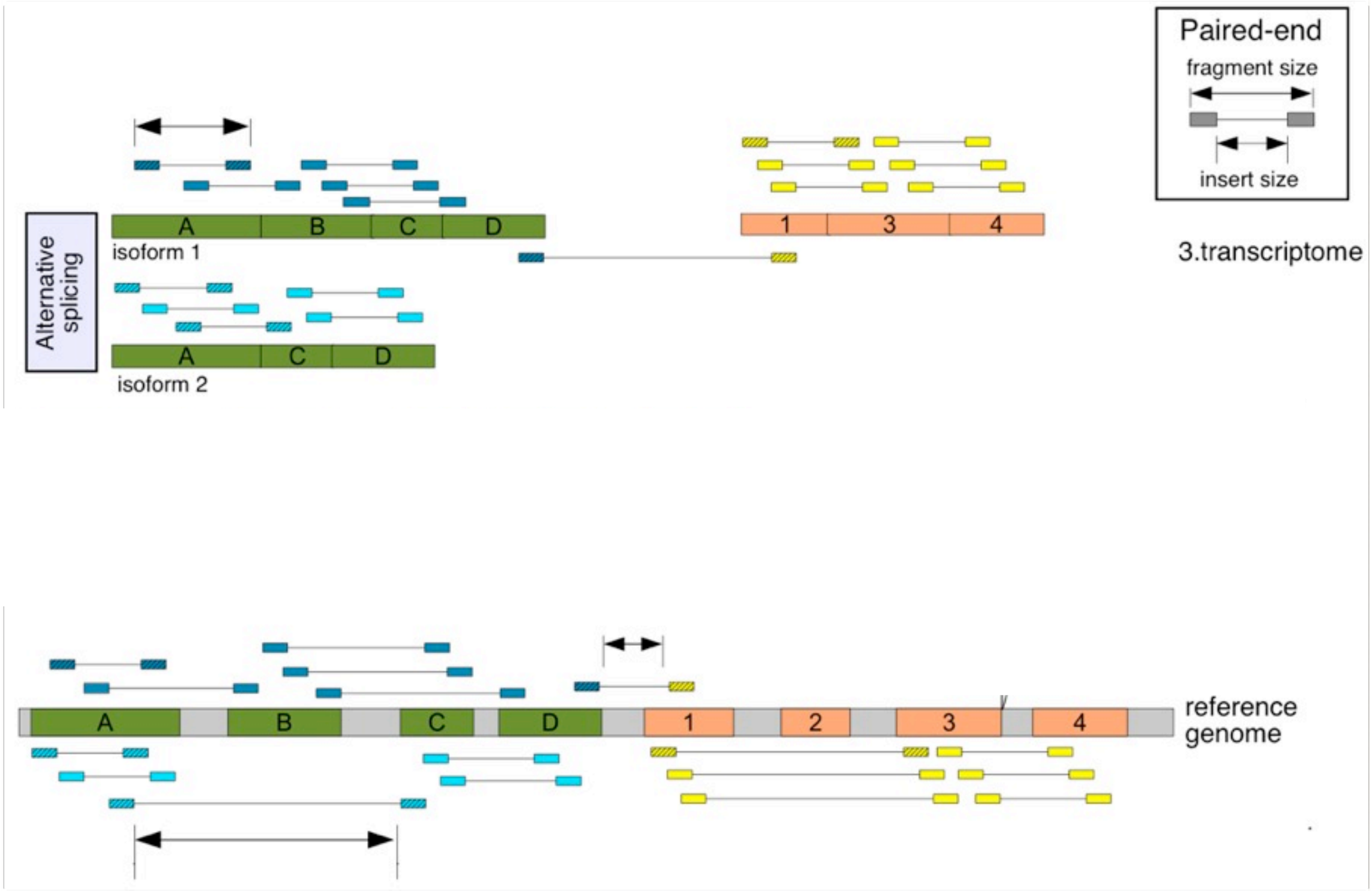


Insert-size distribution concept

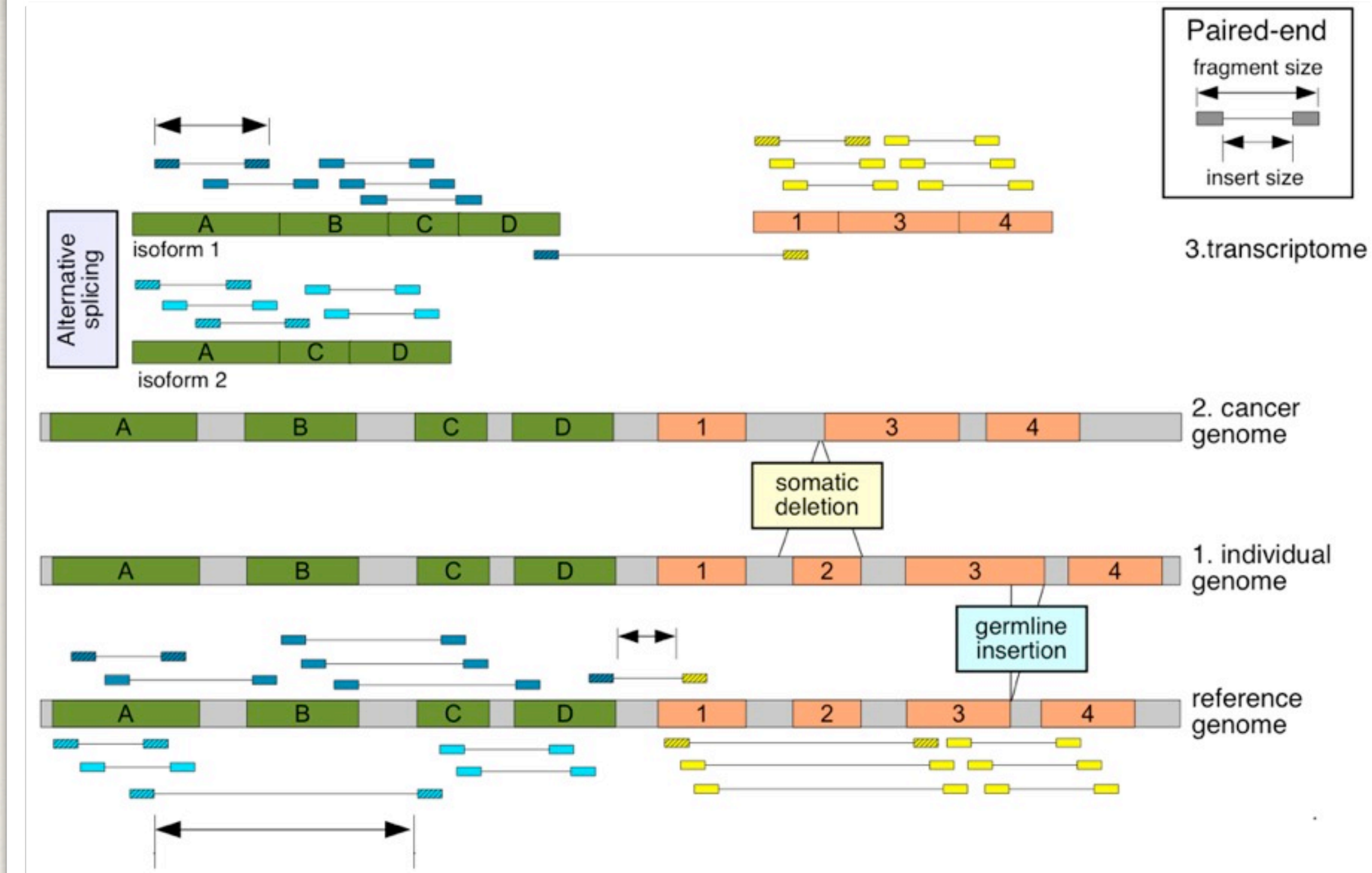
Detecting structural variations (SVs)



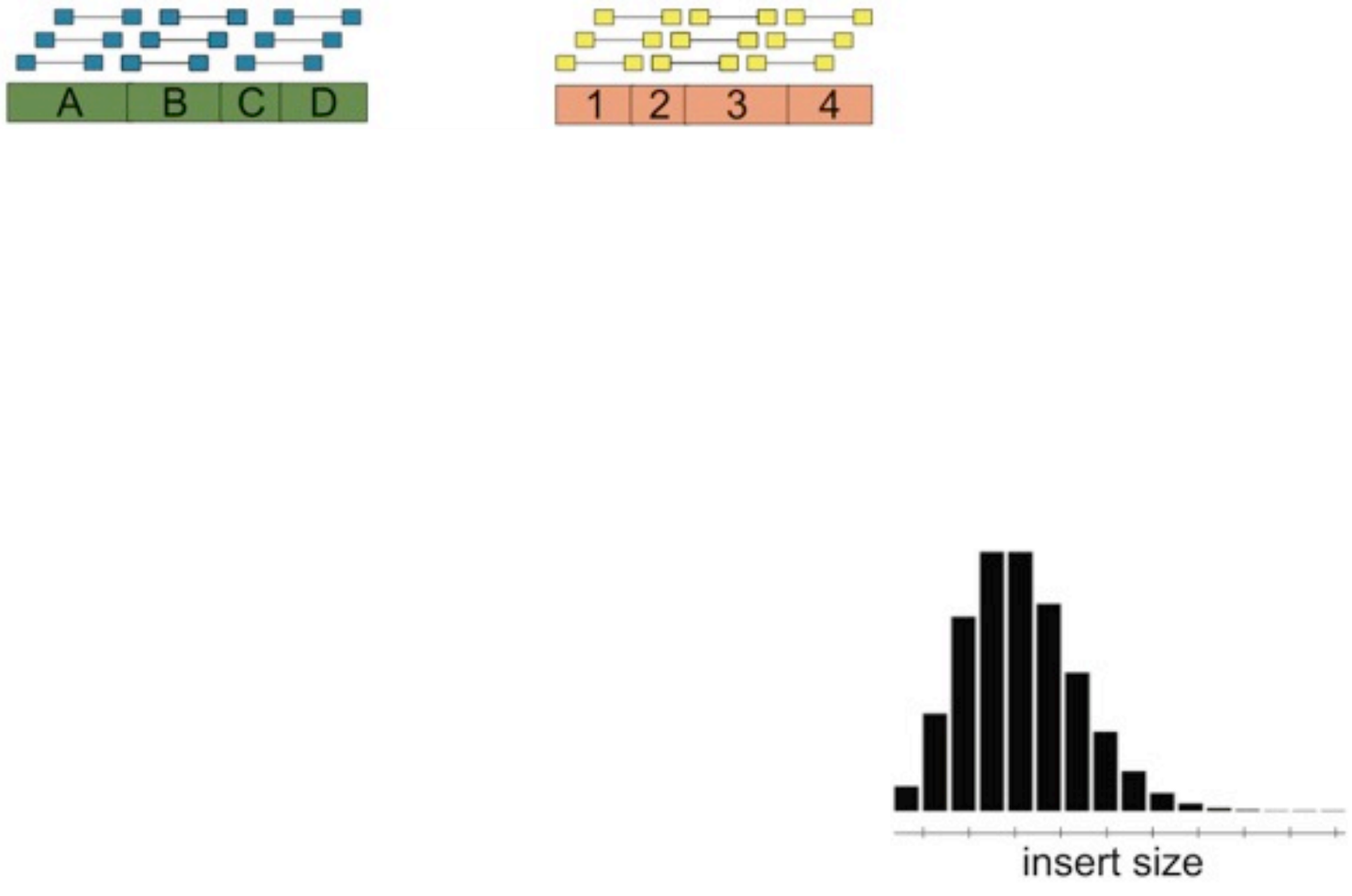
Complexity of applying the insert-size analysis to the transcriptome



Complexity of applying the insert-size analysis to the transcriptome

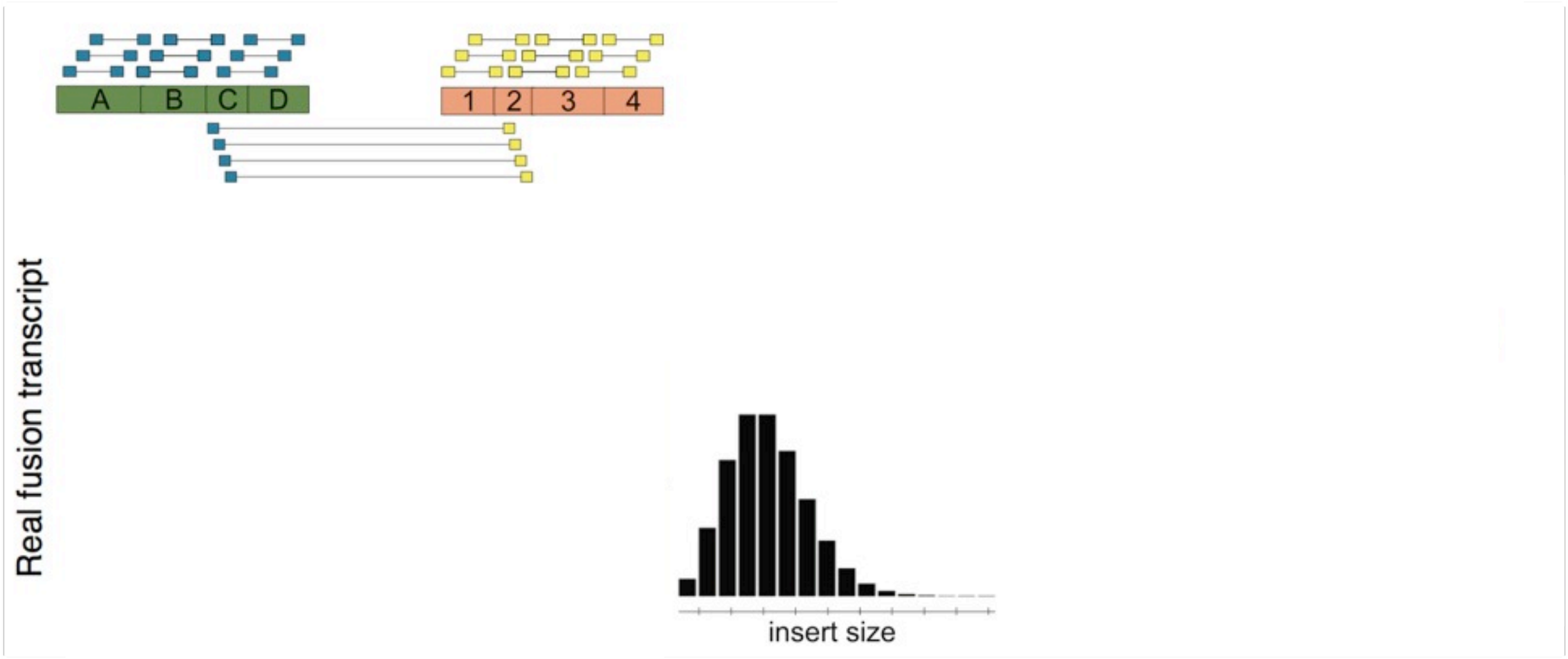


Complexity of applying the insert-size analysis to the transcriptome



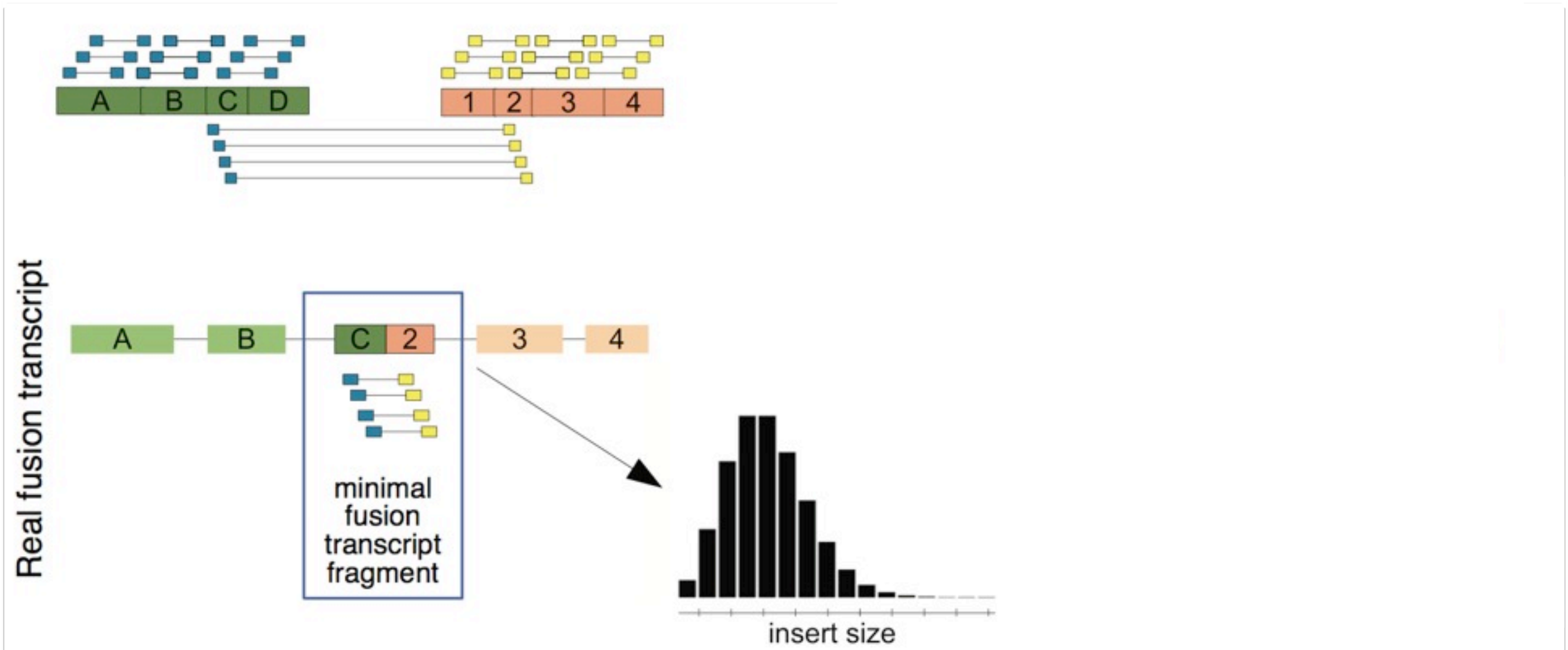
“Rescuing” the insert-size concept for transcriptome analysis

Minimal fusion transcript fragment



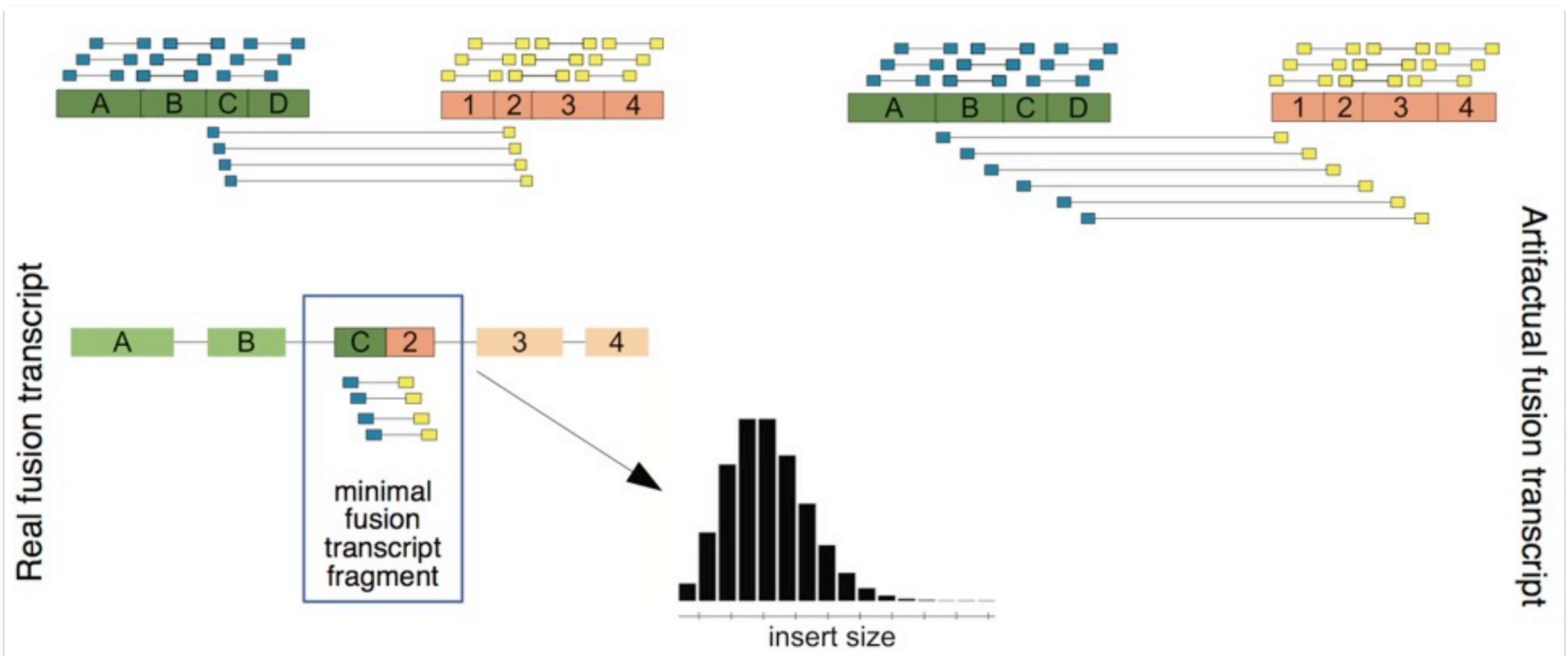
“Rescuing” the insert-size concept for transcriptome analysis

Minimal fusion transcript fragment



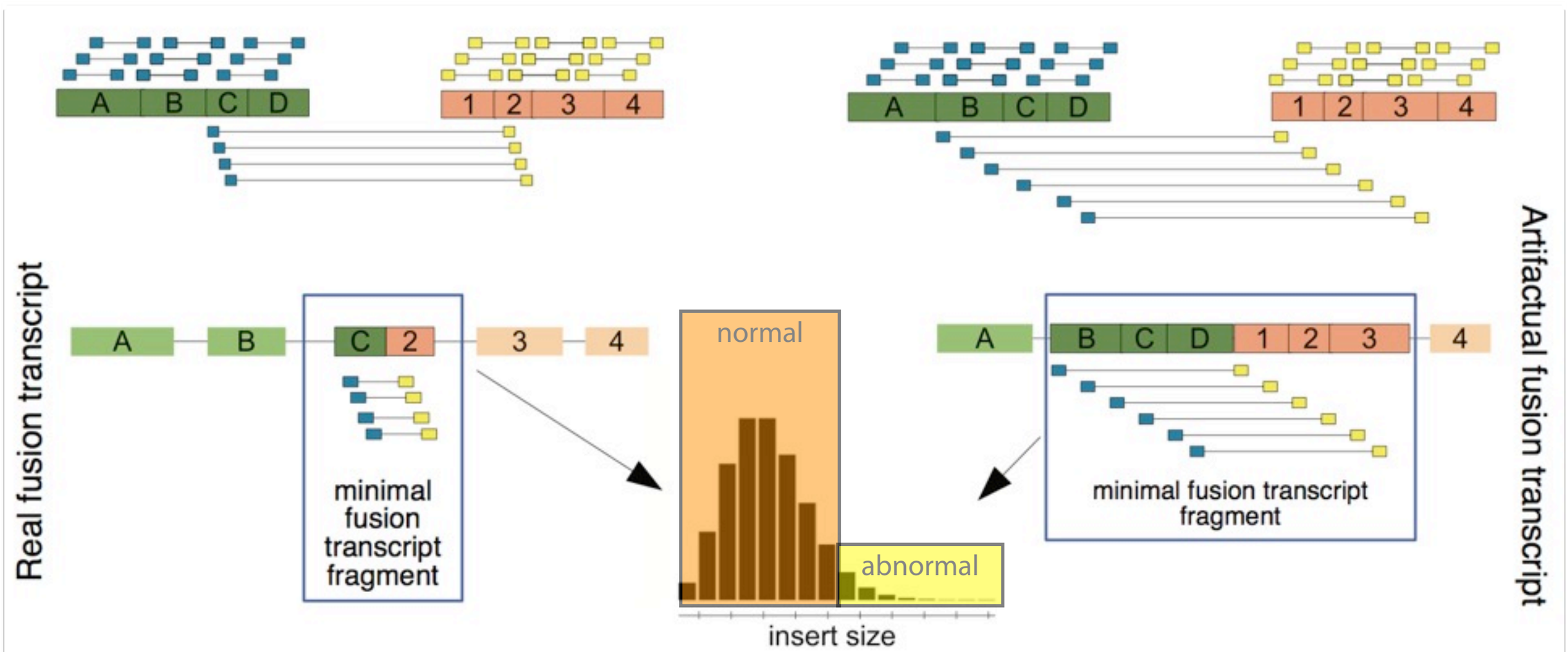
“Rescuing” the insert-size concept for transcriptome analysis

Minimal fusion transcript fragment



"Rescuing" the insert-size concept for transcriptome analysis

Minimal fusion transcript fragment

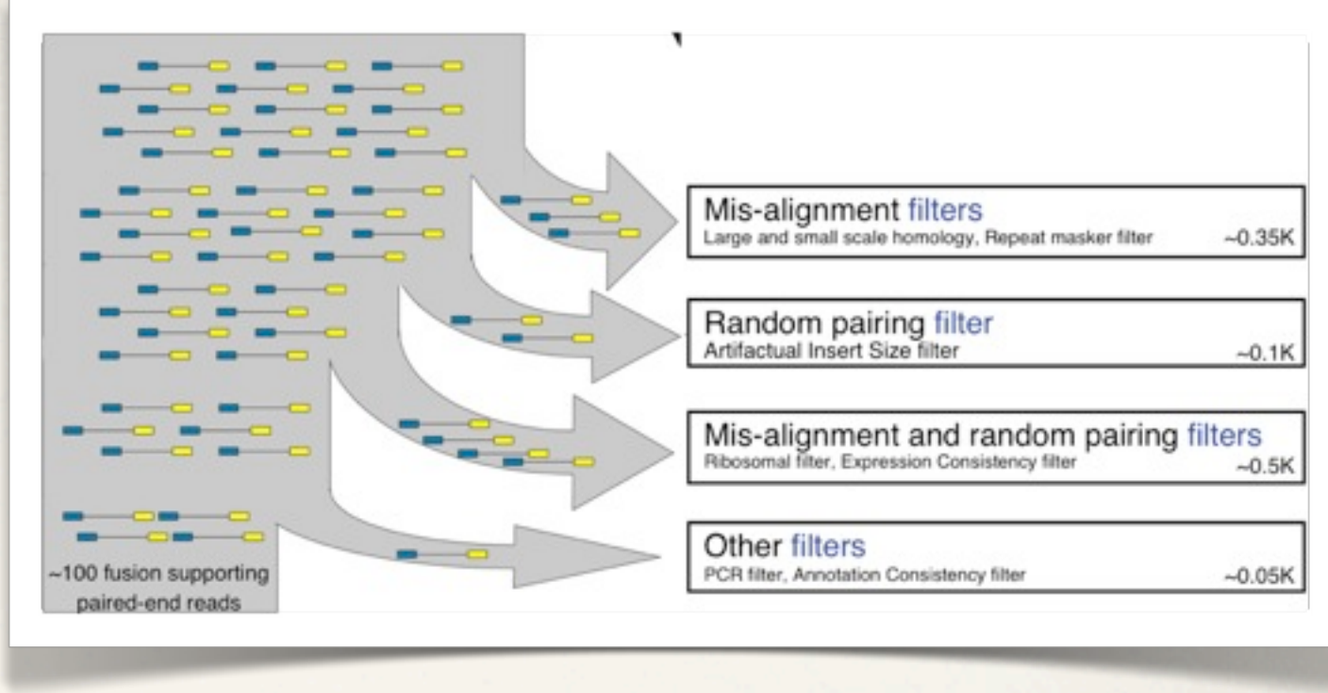


Permutation test: the probability that the inter-transcript insert-size distribution is compatible with the intra-transcript one.

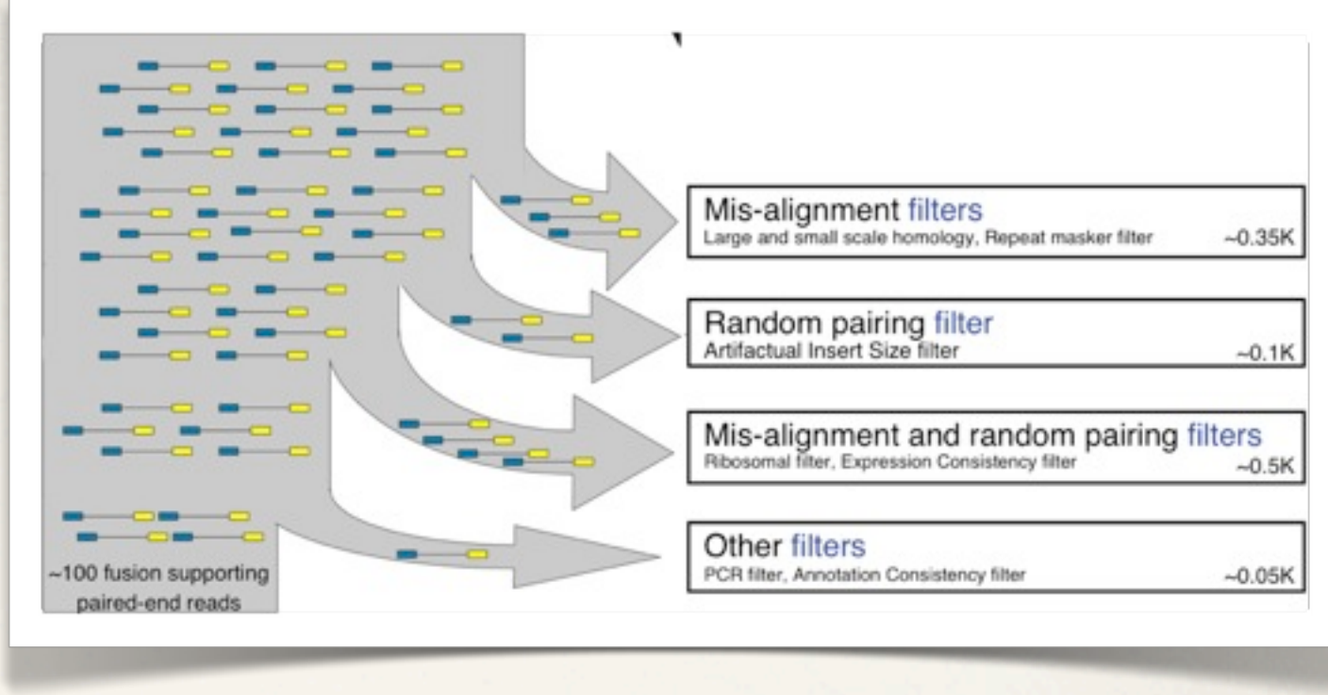
"Rescuing" the insert-size concept for transcriptome analysis

Minimal fusion transcript fragment

Mis-alignment and random pairing filters

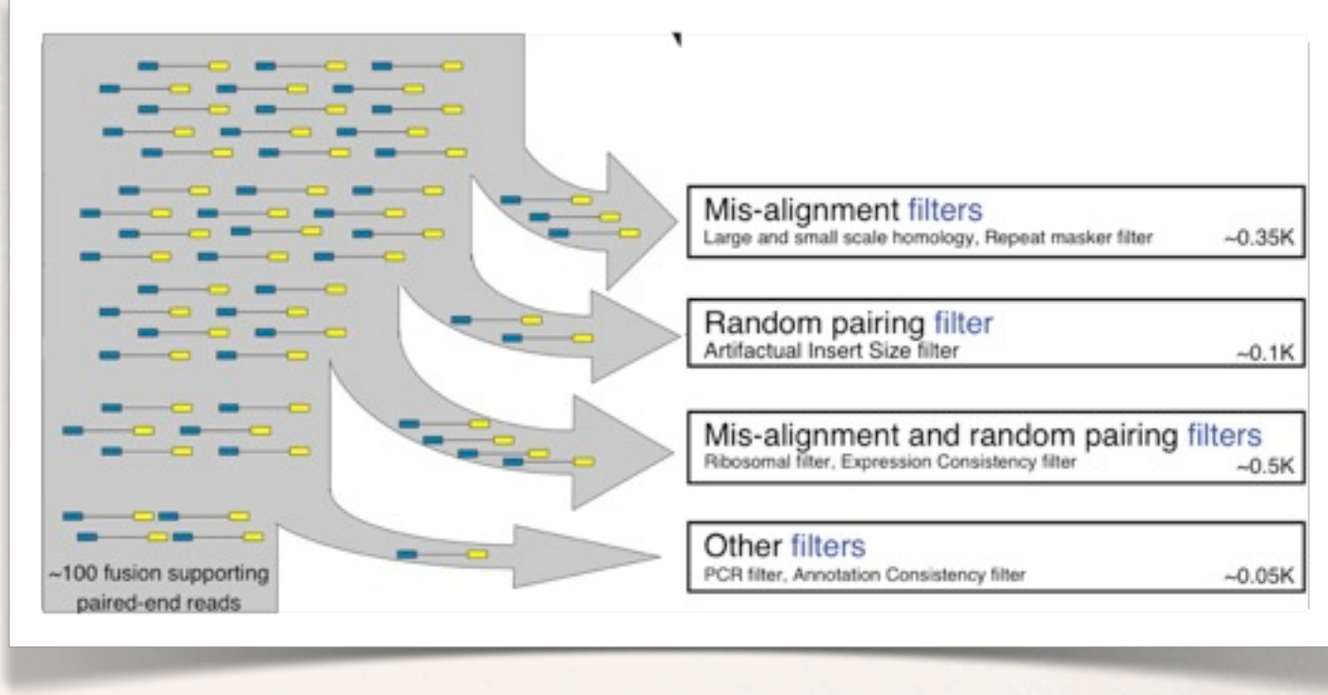


Mis-alignment and random pairing filters



- *Observation 1*: highly expressed genes are more likely to generate random chimeric transcripts

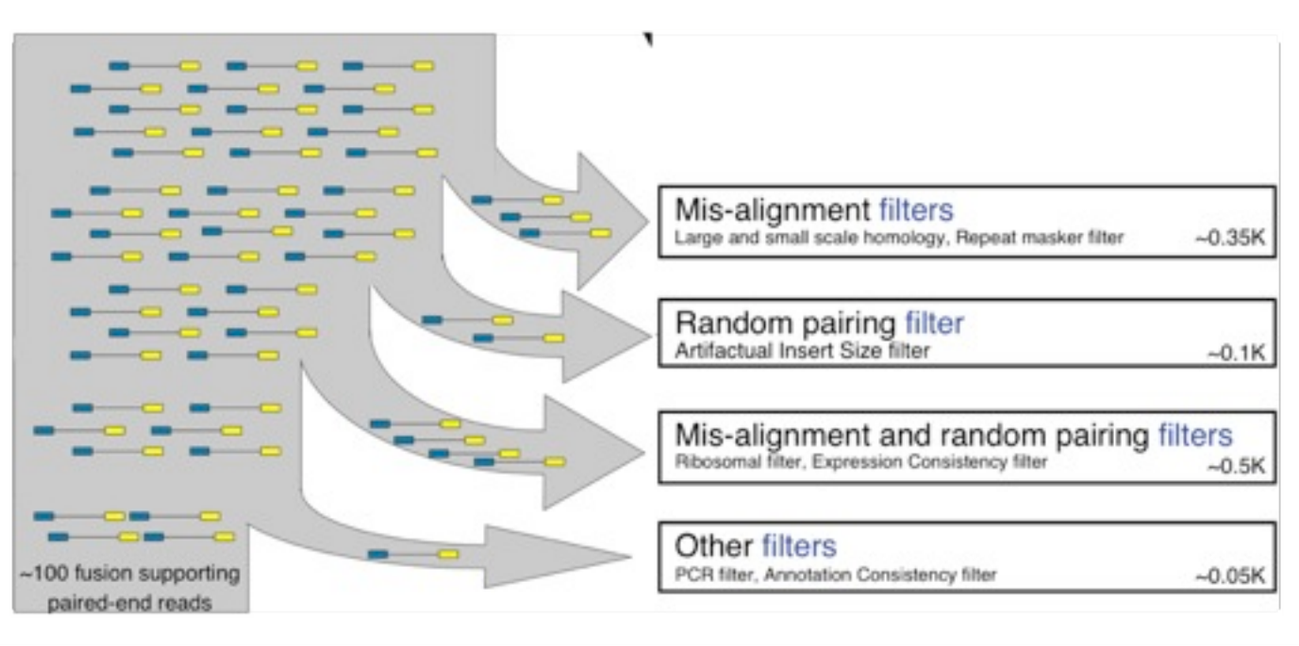
Mis-alignment and random pairing filters



- *Observation 1*: highly expressed genes are more likely to generate random chimeric transcripts

- *Observation 2*: some of the reads may be mis-aligned, complicating the identification of the genes involved in the random pairing

Mis-alignment and random pairing filters

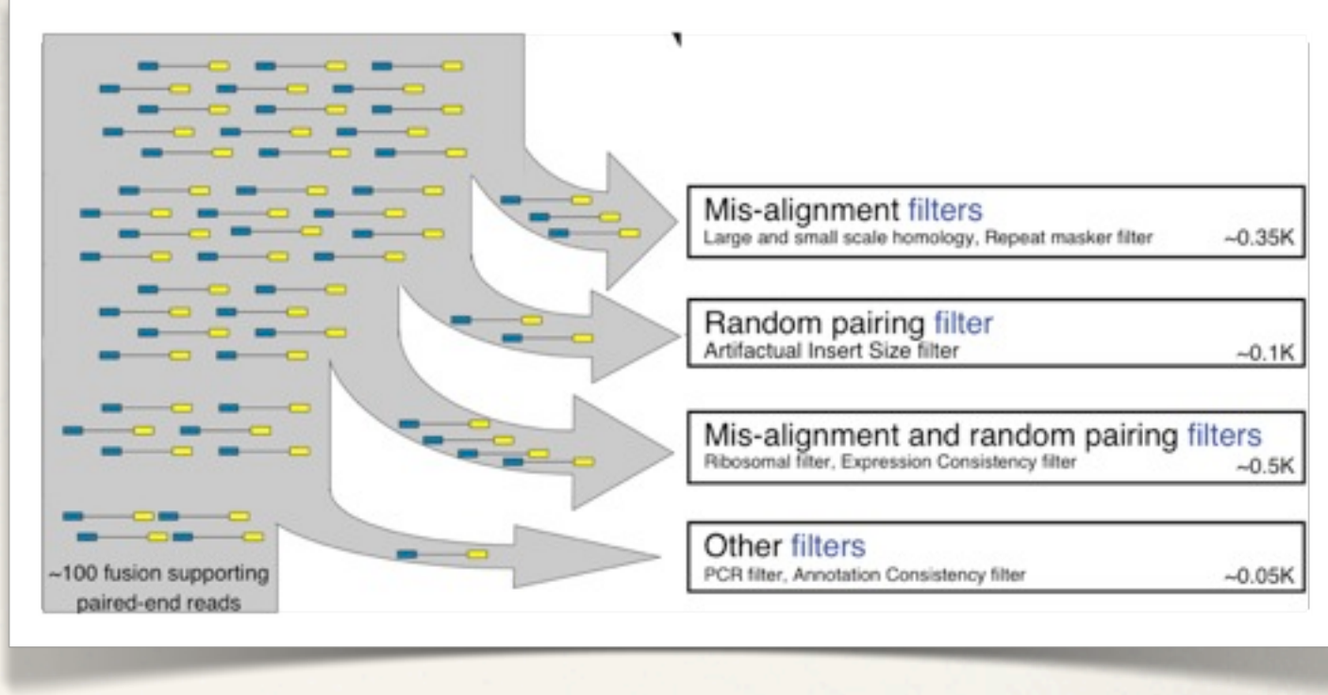


- *Observation 1*: highly expressed genes are more likely to generate random chimeric transcripts

- *Observation 2*: some of the reads may be mis-aligned, complicating the identification of the genes involved in the random pairing

- *Solution*:

Mis-alignment and random pairing filters



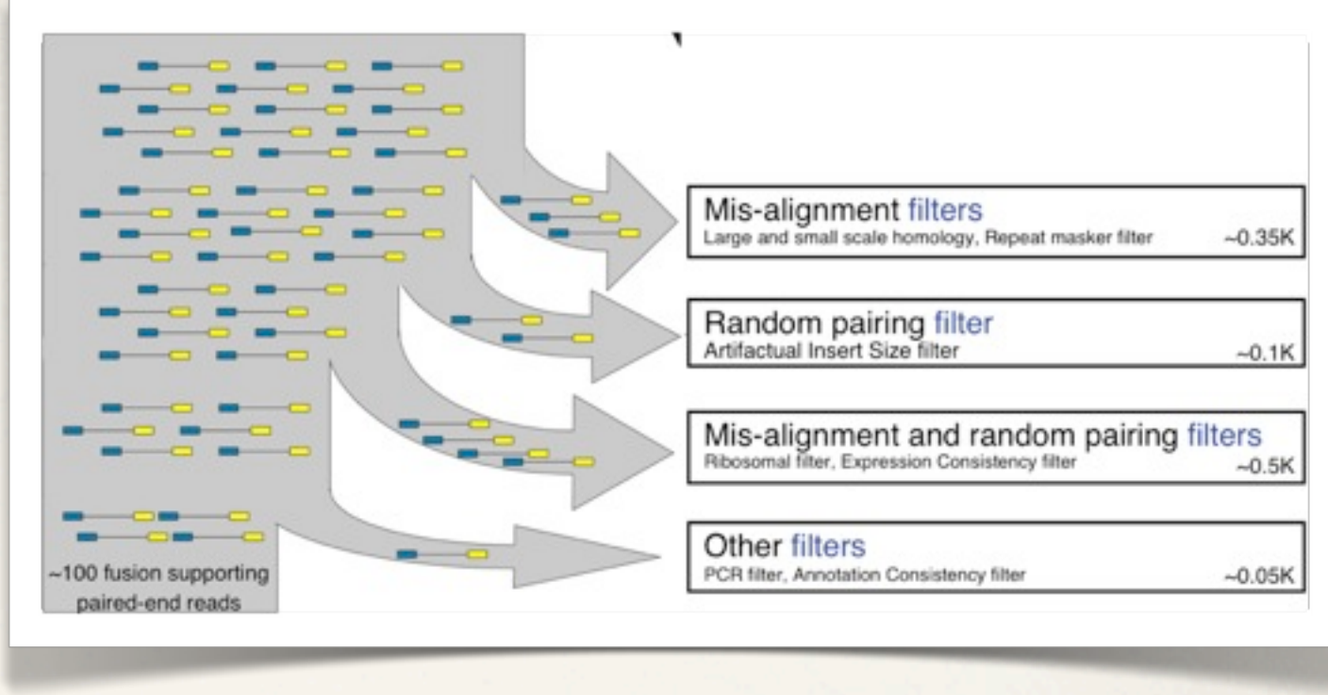
- *Observation 1*: highly expressed genes are more likely to generate random chimeric transcripts

- *Observation 2*: some of the reads may be mis-aligned, complicating the identification of the genes involved in the random pairing

- *Solution*:

  - *Ribosomal* filter (highly expressed genes)

Mis-alignment and random pairing filters



- *Observation 1*: highly expressed genes are more likely to generate random chimeric transcripts

- *Observation 2*: some of the reads may be mis-aligned, complicating the identification of the genes involved in the random pairing

- *Solution*:
  - *Ribosomal* filter (highly expressed genes)
  - *Expression consistency* filter

Classifying the candidates

- Fusion candidates are classified based on the location of the two genes on the genome:

- *Inter-chromosomal*: genes on different chromosomes;

- *Intra-chromosomal*: genes on the same chromosome and strand, with other genes between them;

- *Read-through*: genes of the same chromosome and strand, and without other genes between them;

- *Cis*: genes on the same chromosome, but in opposite strands.

Scoring the candidates

Scoring the candidates

- Supportive PE Reads per million mapped reads (*SPER*)
  - Normalized number of *inter*-transcript PE reads (m_i)

$$SPER_i = \frac{m_i}{N_{mapped}} \cdot 10^6$$

Scoring the candidates

- Supportive PE Reads per million mapped reads (*SPER*)
  - Normalized number of *inter*-transcript PE reads (m_i)

$$SPER_i = \frac{m_i}{N_{mapped}} \cdot 10^6$$

- How good is the observed SPER compared with the expected SPER?
  - Difference of observed SPER and analytically computed SPER (*DASPER*)
  - Ratio of observed SPER and empirically computed SPER (*RESPER*)

SPER expectations

Analytical

- Expected number of inter-transcript reads by chance (<m_{AB}>):

$$\langle SPER_i \rangle = \frac{\langle m_{AB} \rangle}{N_{mapped}} \cdot 10^6$$

- <m_{AB}> can be estimated from the joint probability:

$$\langle m_{AB} \rangle = P(A) \cdot P(B) \cdot N_{mapped} = \frac{m_A \cdot m_B}{N_{mapped}}$$

$$DASPER_i = SPER_i - \langle SPER_i \rangle$$

SPER expectations

Analytical

- Expected number of inter-transcript reads by chance (<m_{AB}>):

$$\langle SPER_i \rangle = \frac{\langle m_{AB} \rangle}{N_{mapped}} \cdot 10^6$$

- <m_{AB}> can be estimated from the joint probability:

$$\langle m_{AB} \rangle = P(A) \cdot P(B) \cdot N_{mapped} = \frac{m_A \cdot m_B}{N_{mapped}}$$

$$DASPER_i = SPER_i - \langle SPER_i \rangle$$

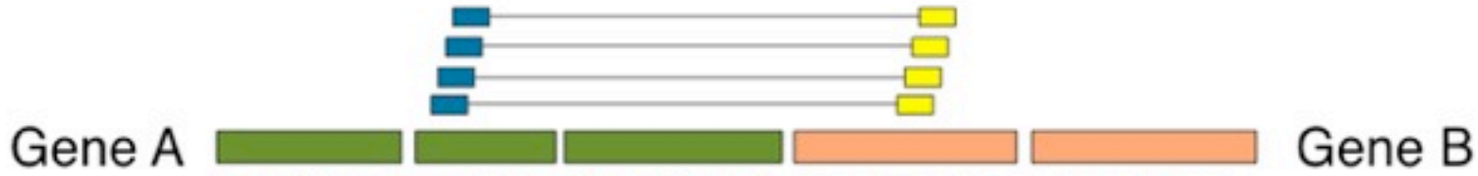
Empirical

- Average of the other candidates' SPER:

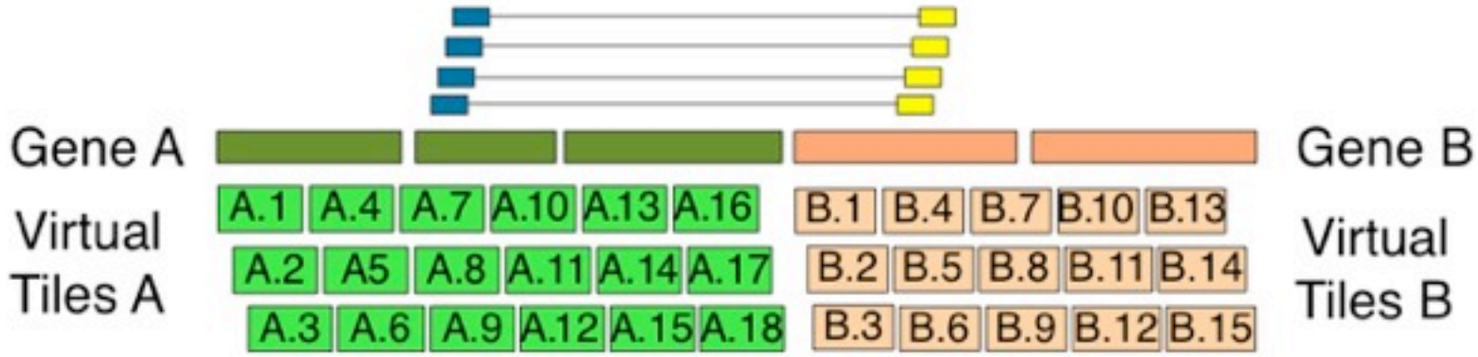
$$\overline{SPER} = \frac{1}{M} \cdot \sum_{j=1}^{M} SPER_j$$

- where M is the total number of fusion transcript candidates

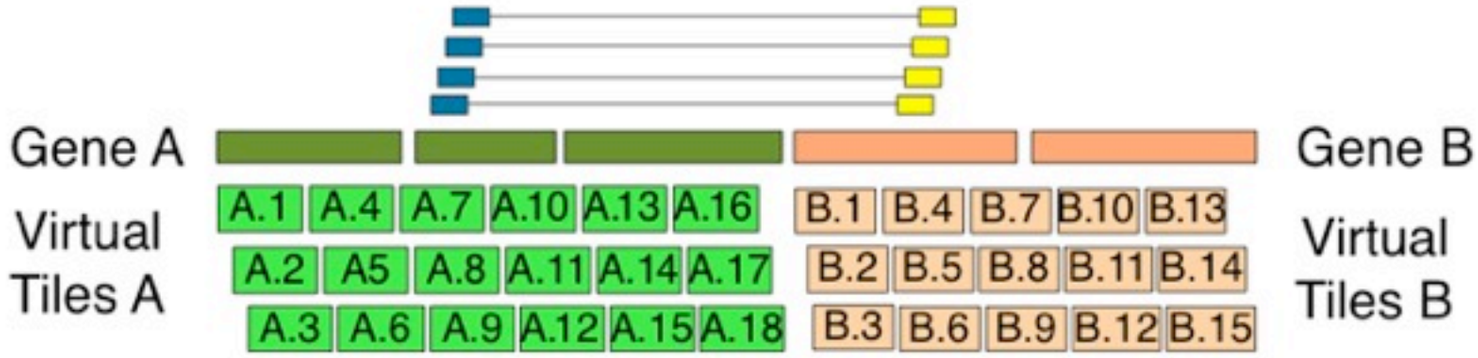
$$RESPER_i = \frac{SPER_i}{\overline{SPER}}$$

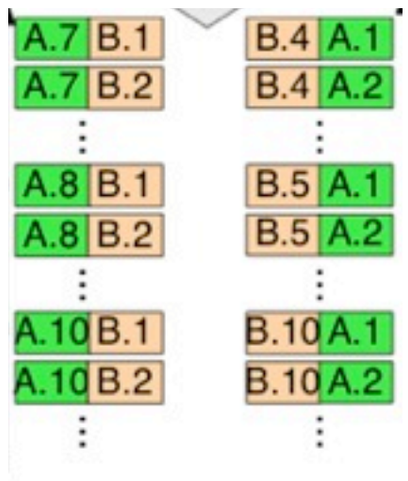


Junction-Sequence Identifier Module

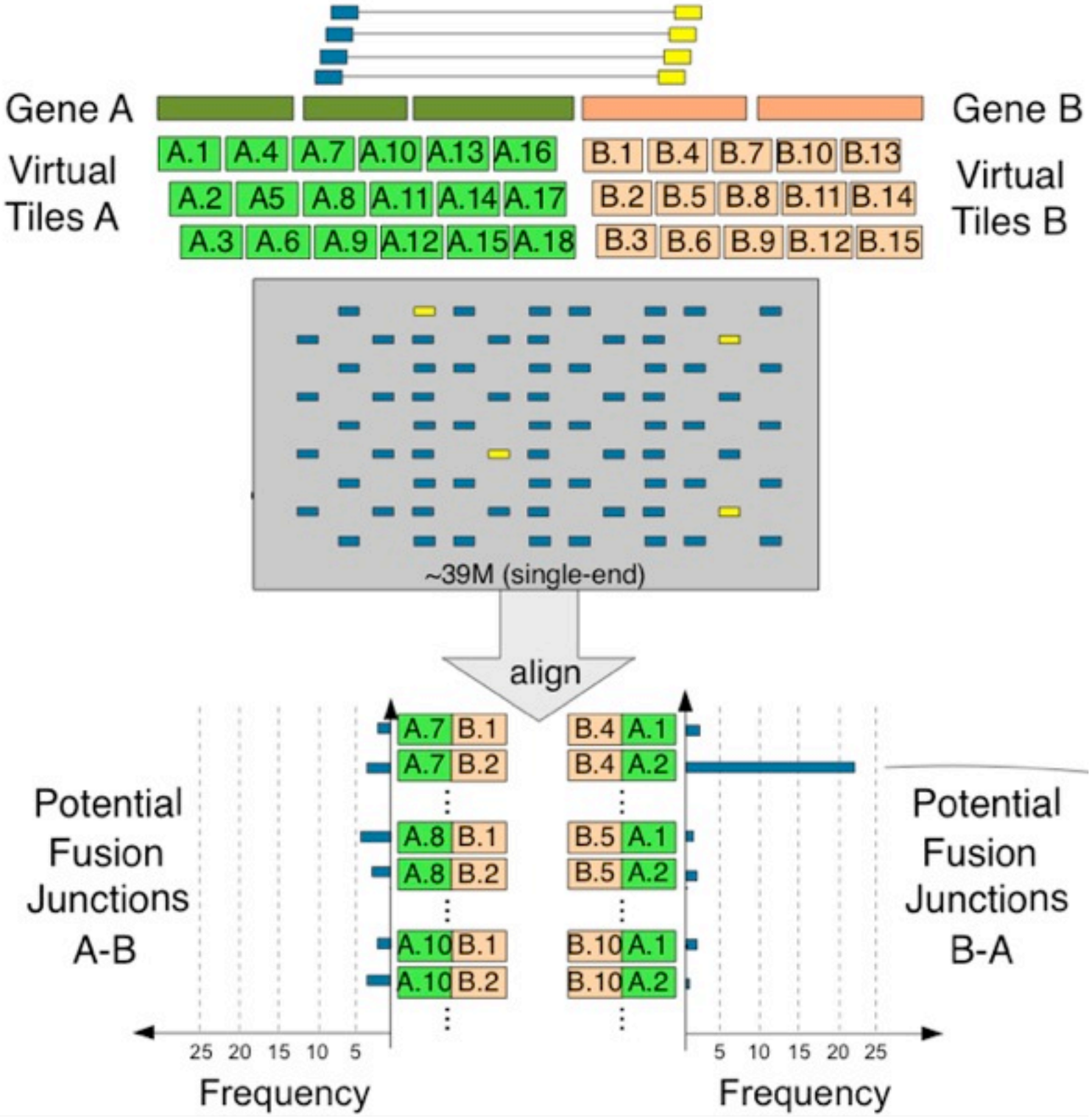


Junction-Sequence Identifier Module

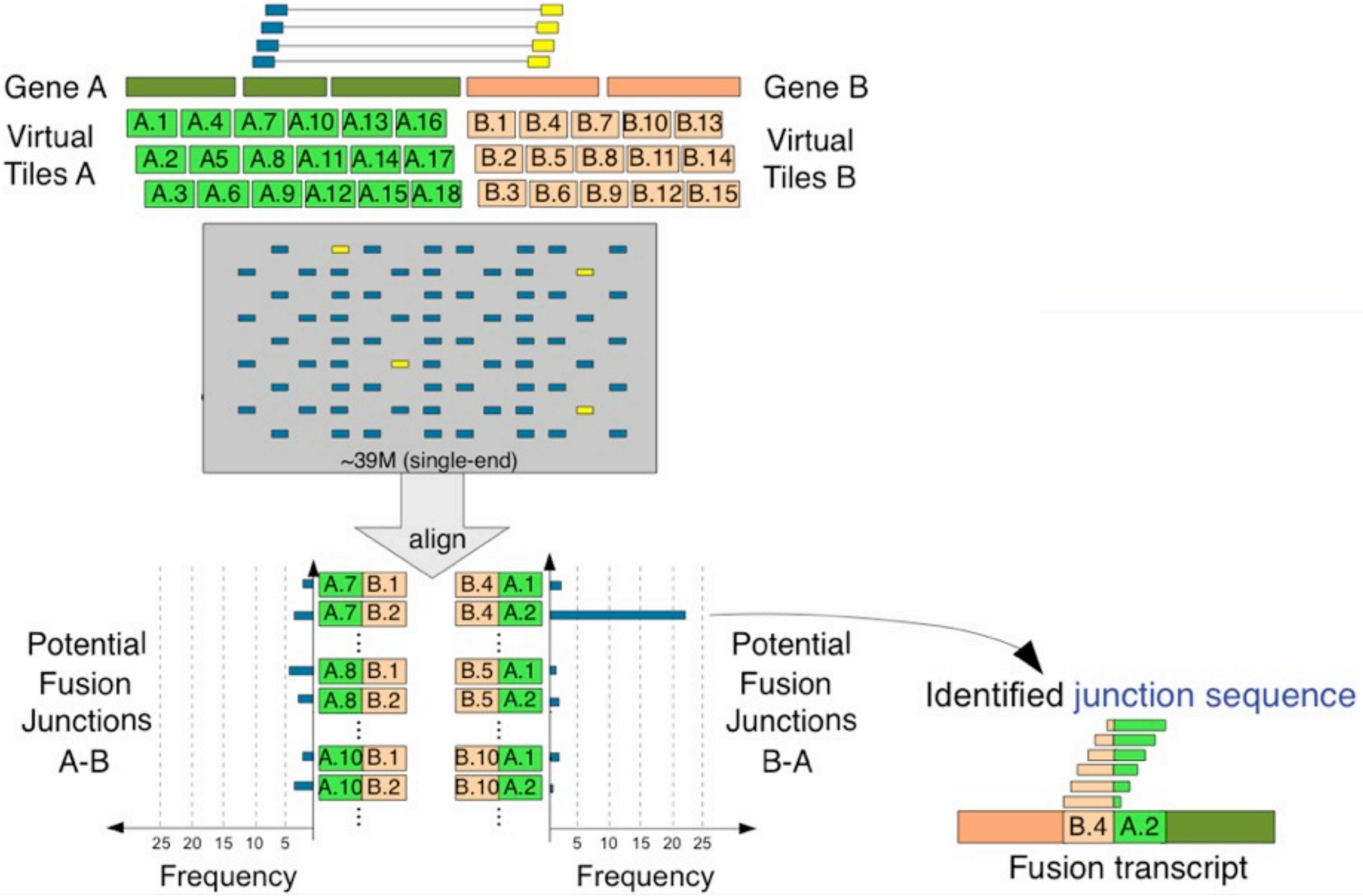




Junction-Sequence Identifier Module



Junction-Sequence Identifier Module



Junction-Sequence Identifier Module

Study design

- Well-characterized samples

Sample ID	Type (PCa=prostate cancer)	Known Fusion Type	Read size	Mapped PE reads
106_T	PCa	TMPRSS2-ERG	51	4,312,087
1700_D	PCa	TMPRSS2-ERG	51	6,805,776
580_B	PCa	TMPRSS2-ERG	36	5,211,234
99_T	PCa	NDRG1-ERG	36	1,515,444
2621_D	PCa	SLC45A3-ERG	54	11,899,985
1043_D	PCa	No known fusions	51	1,549,569
NCI-H660	PCa cell line	TMPRSS2-ERG	51	3,572,391
GM12878	Lymphoblastoid cell line	No known fusions	54	20,676,160

What is the effect of the filters?

- Total of 7,342 candidates

	Average number of identified candidate per sample	Average number of candidates per sample after filtering
Number	917.75	38
Range	451-1618	3-94

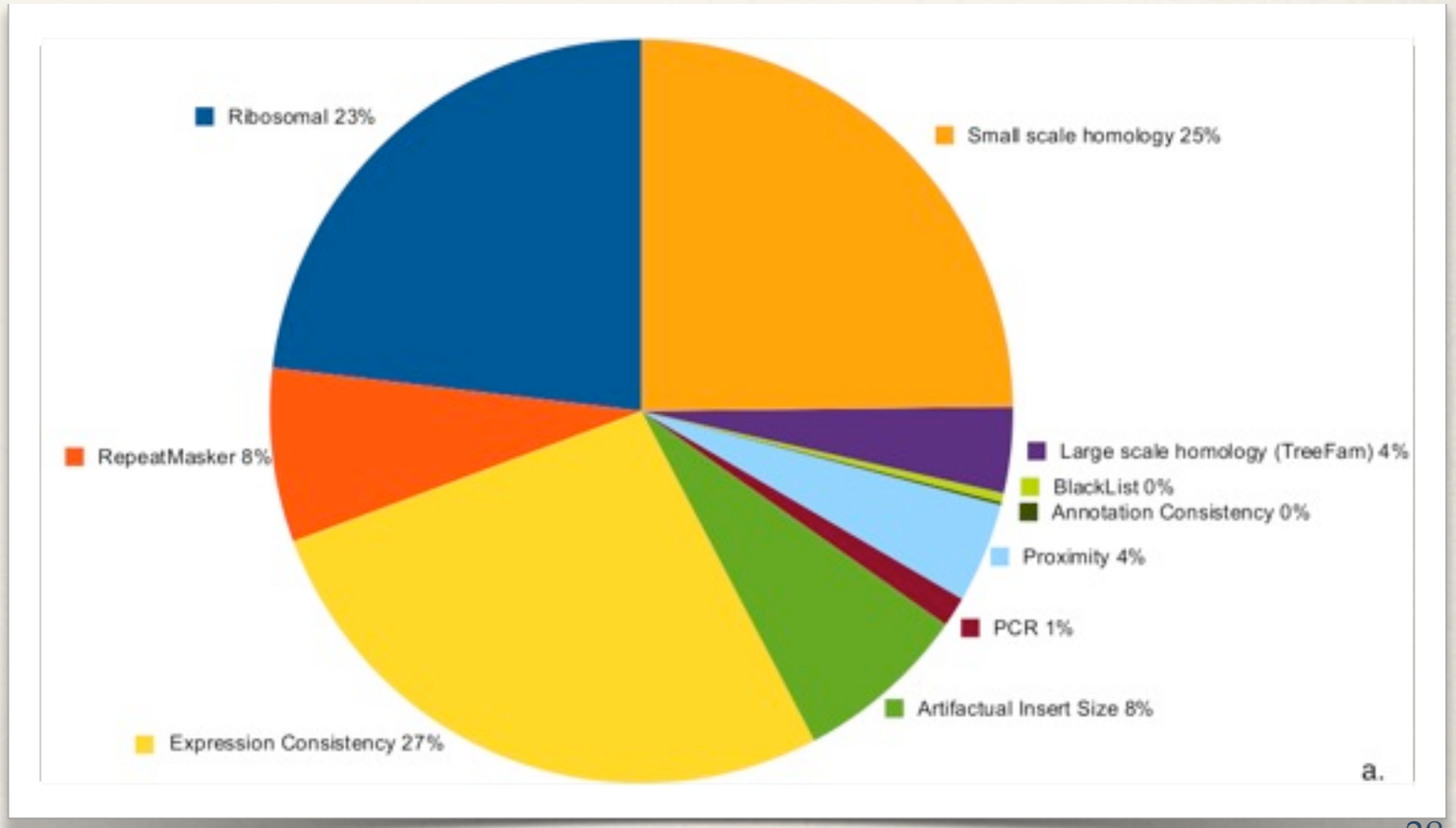
- 304 passed the filters, i.e. **96**% reduction

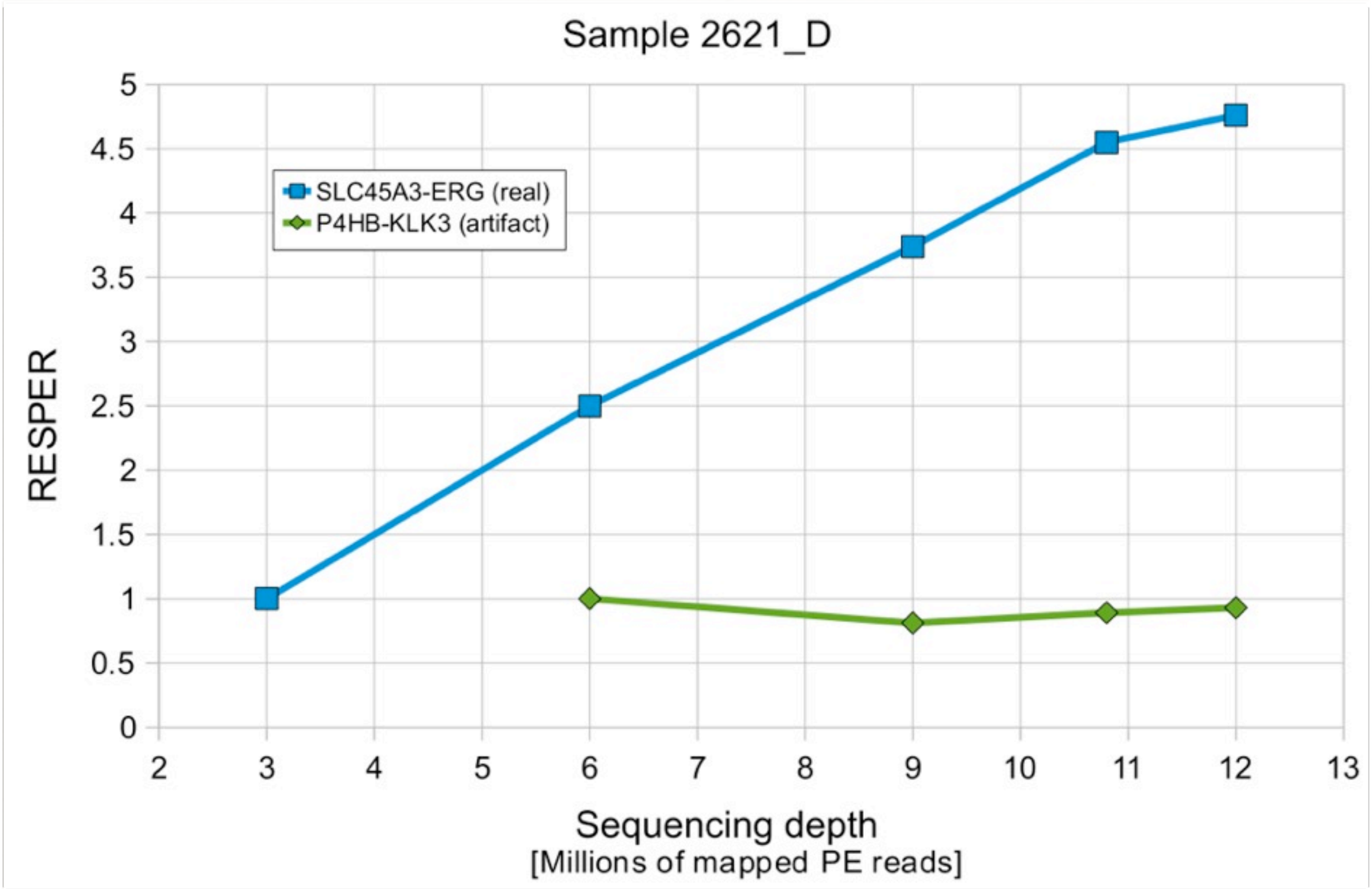
What is the effect of the filters?

	Average number of identified candidate per sample	Average number of candidates per sample after filtering
Number	917.75	38
Range	451-1618	3-94

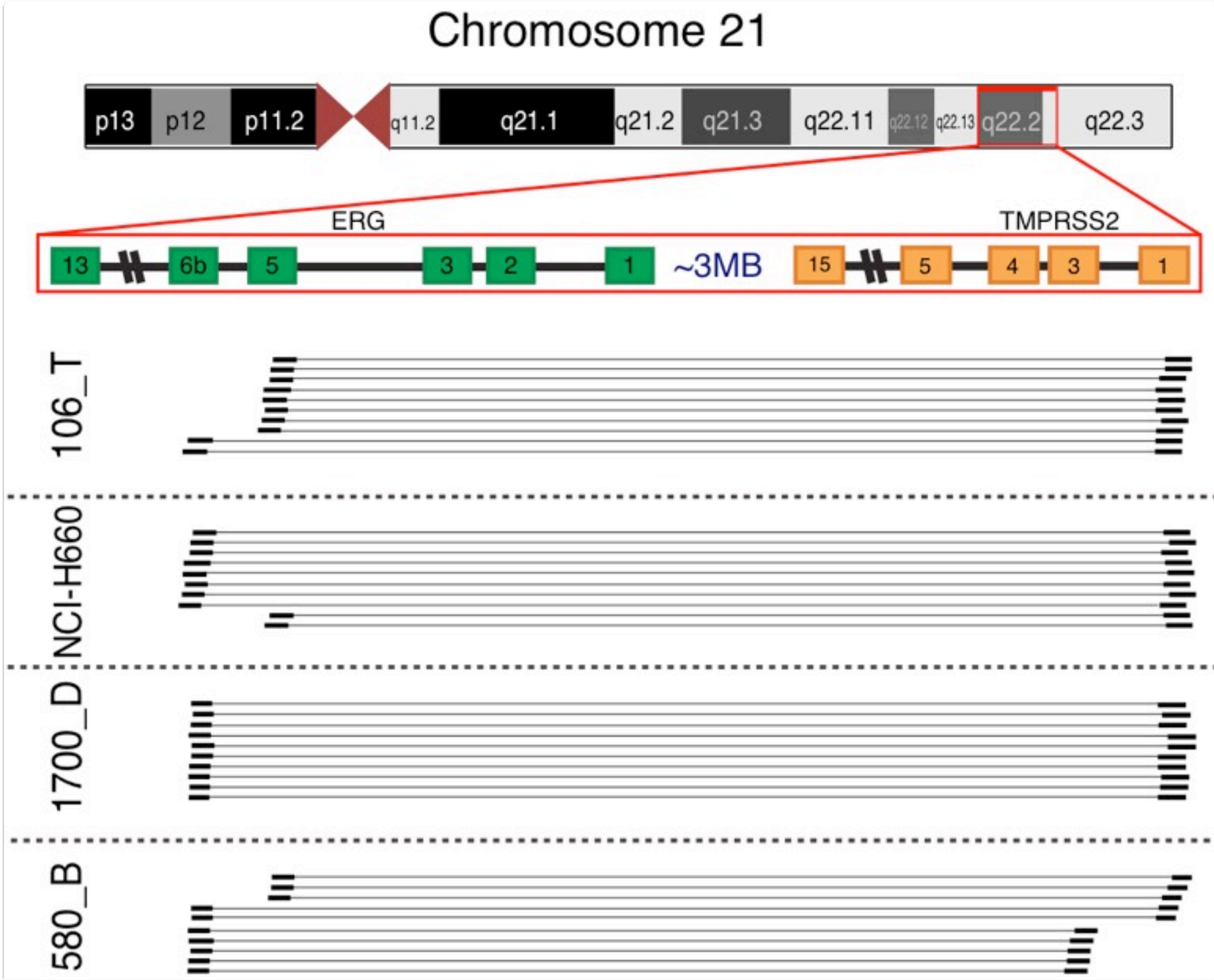
- Total of 7,342 candidates

- 304 passed the filters, i.e. **96**% reduction

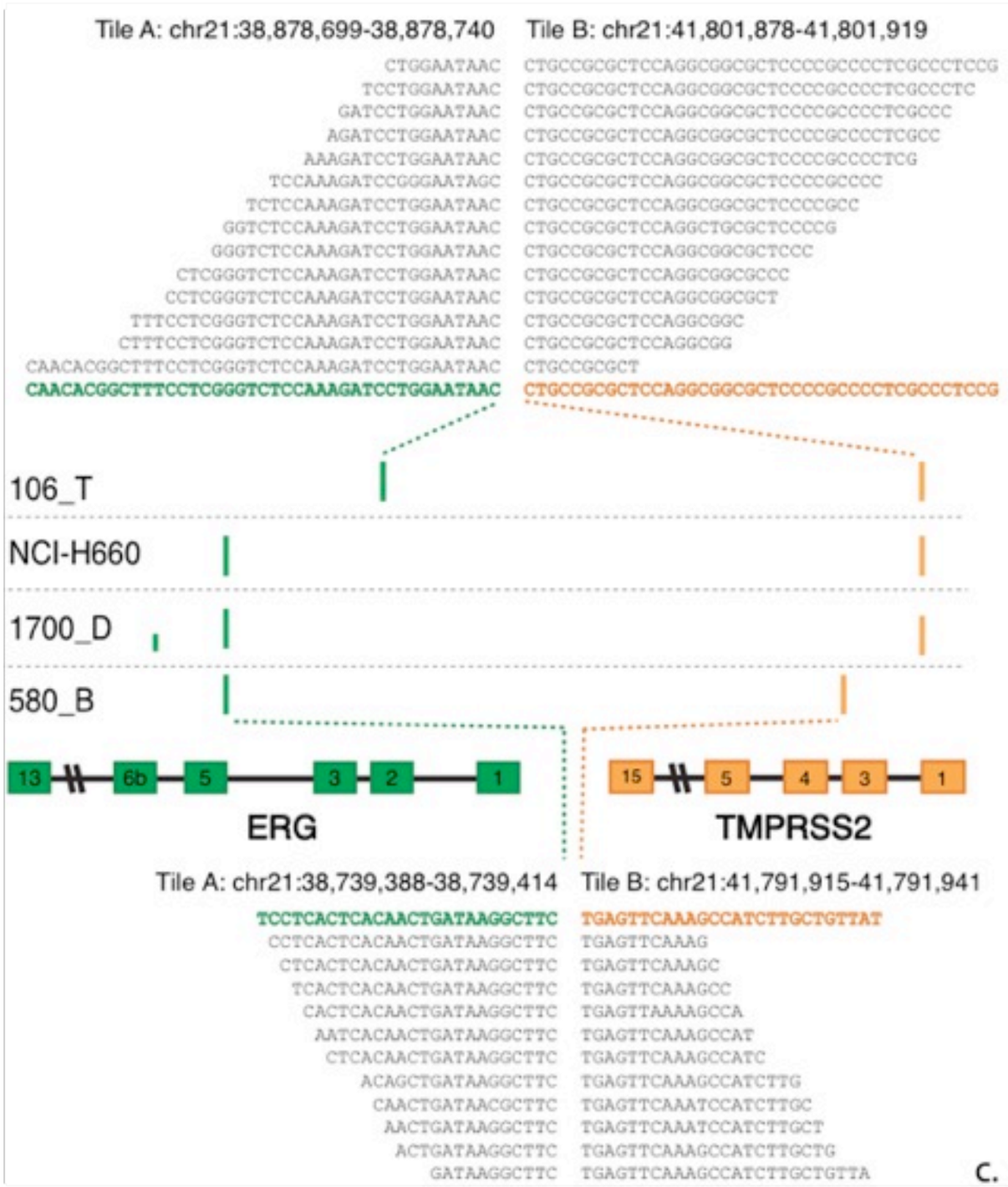




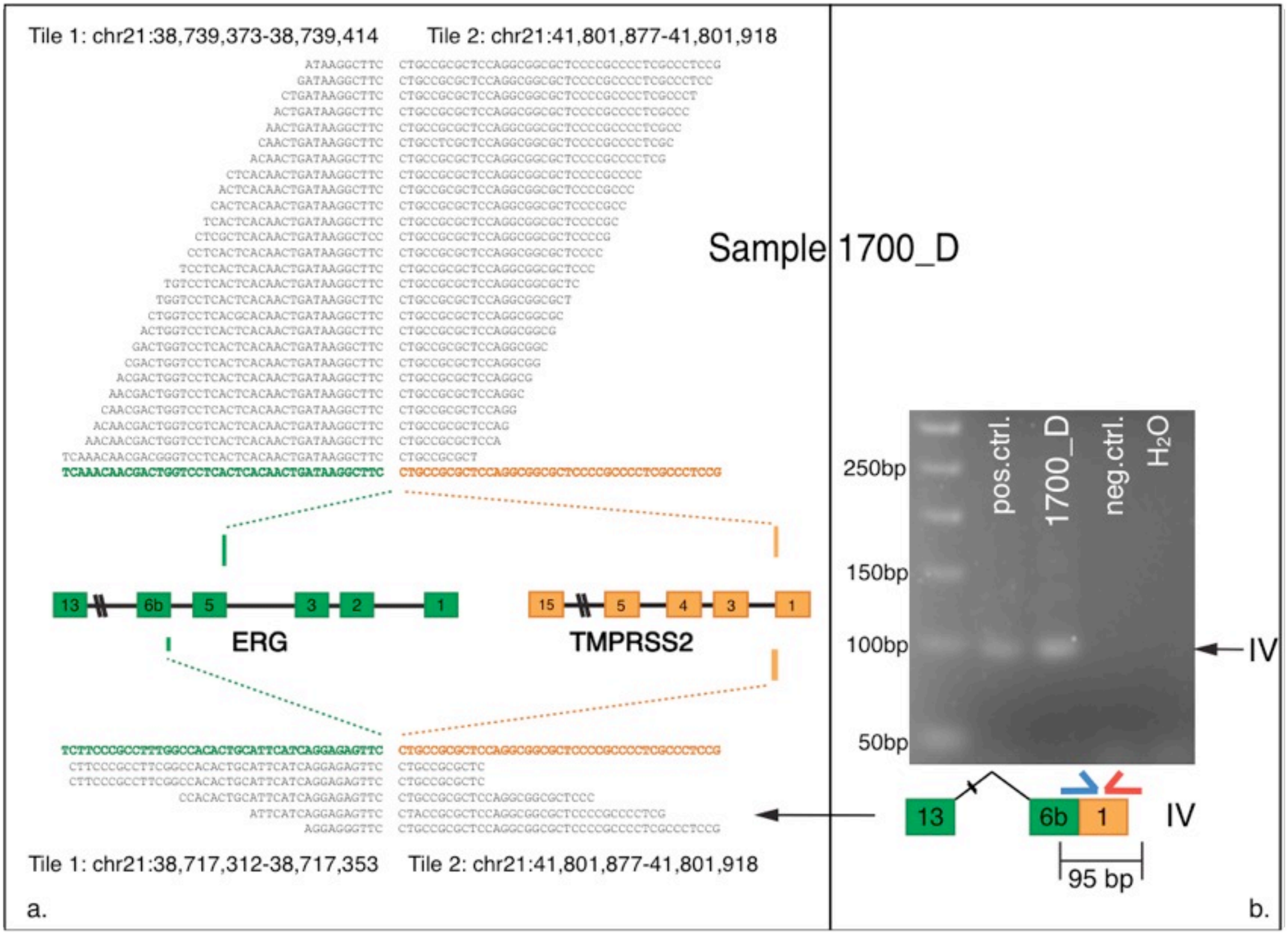
What is the effect of coverage?



TMPRSS2-ERG fusion positives



Junction-sequence identifier



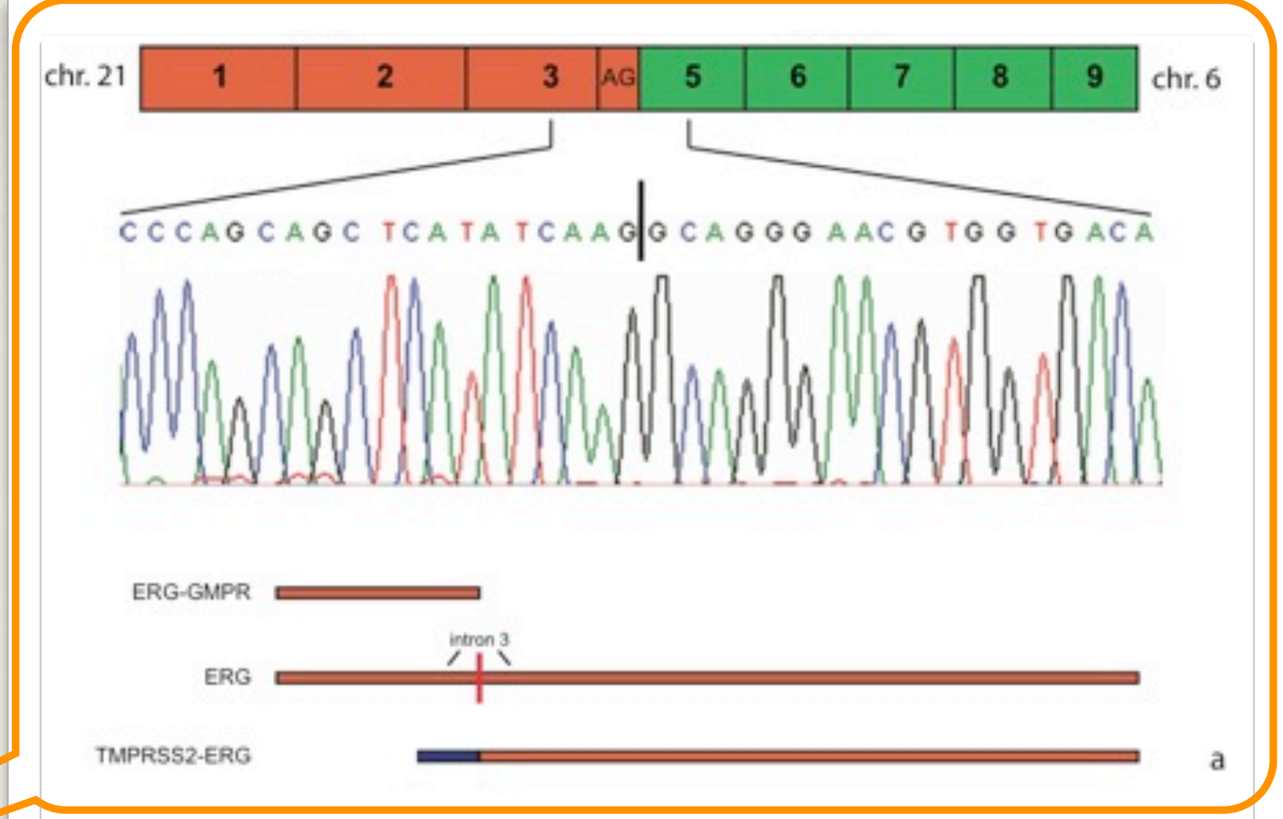
Second isoform

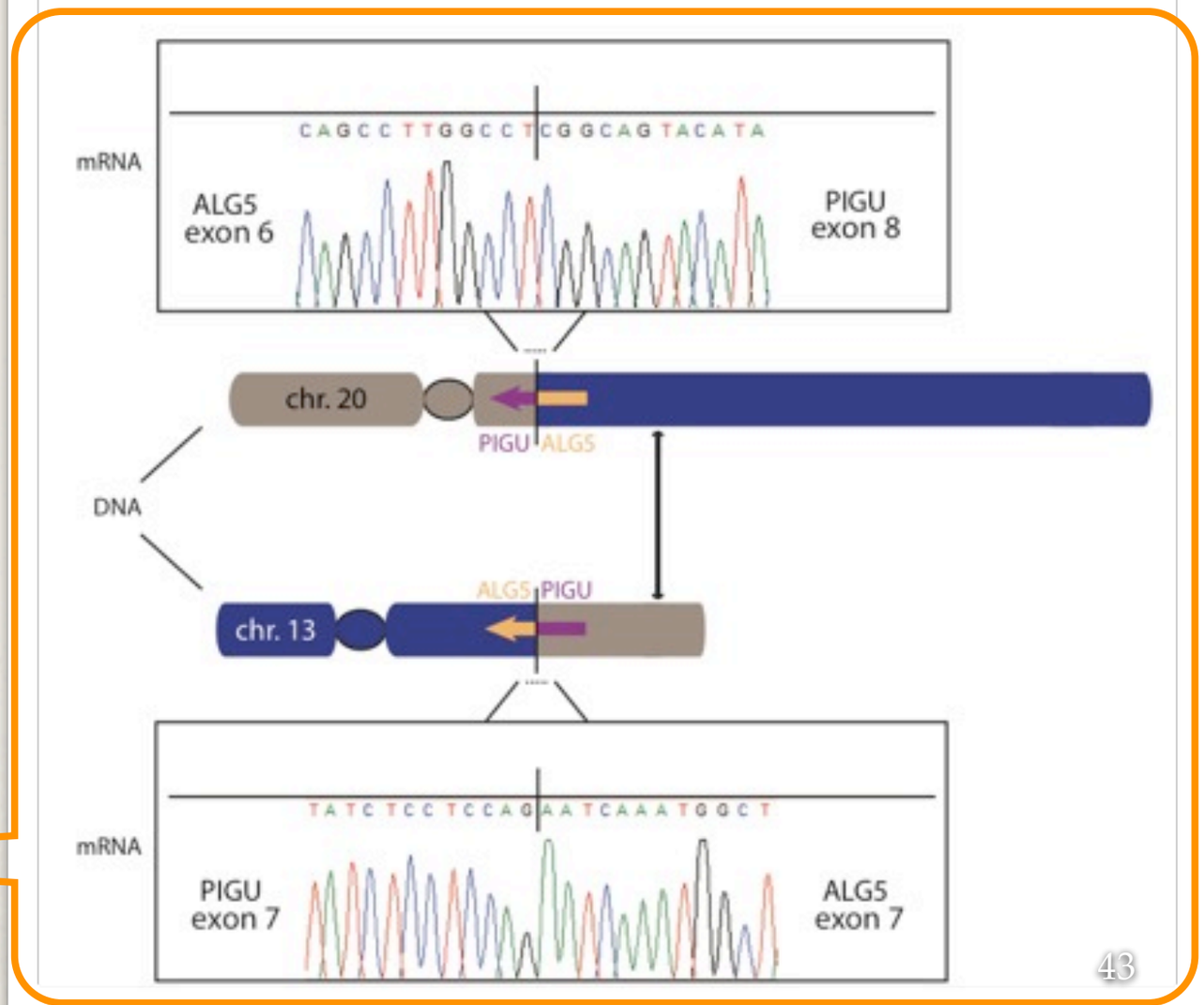
Top candidates

Type	Sample ID	Fusion Candidate	RESPER
intra	**580_B**	**TMPRSS2-ERG**	**14.31**
intra	**1700_D**	**TMPRSS2-ERG**	**8.79**
intra	**106_T**	**TMPRSS2-ERG**	**3.97**
inter	**2621_D**	**SLC45A3-ERG**	**3.56**
inter	**1700_D**	**ERG-GMPR**	**2.05**
read-through	*1700_D*	*SLC16A8-BAIAP2L2*	*1.93*
read-through	*106_T*	*AK094188-AK311452*	*1.9*
read-through	*1700_D*	*ZNF473-FLJ26850*	*1.58*
read-through	*580_B*	*ZNF577-FLJ26850*	*1.58*
read-through	*1043_D*	*ZNF577-ZNF649*	*1.55*
read-through	*1700_D*	*CAMTA2-INCA1*	*1.35*
inter	1700_D	HDAC5	1.29
read-through	*1043_D*	*FLJ00248-LRCH4*	*1.27*
read-through	*1700_D*	*VMAC-CAPS*	*1.17*
read-through	*106_T*	*FLJ00248-LRCH4*	*1.16*
cis	1043_D	AX747861-FLI1	1.13
read-through	*106_T*	*TAGLN-AK126420*	*1.07*
inter	**580_B**	**PIGU-ALG5**	**1.07**
inter	**99_T**	**NDRG1-ERG**	**1.02**

Top candidates

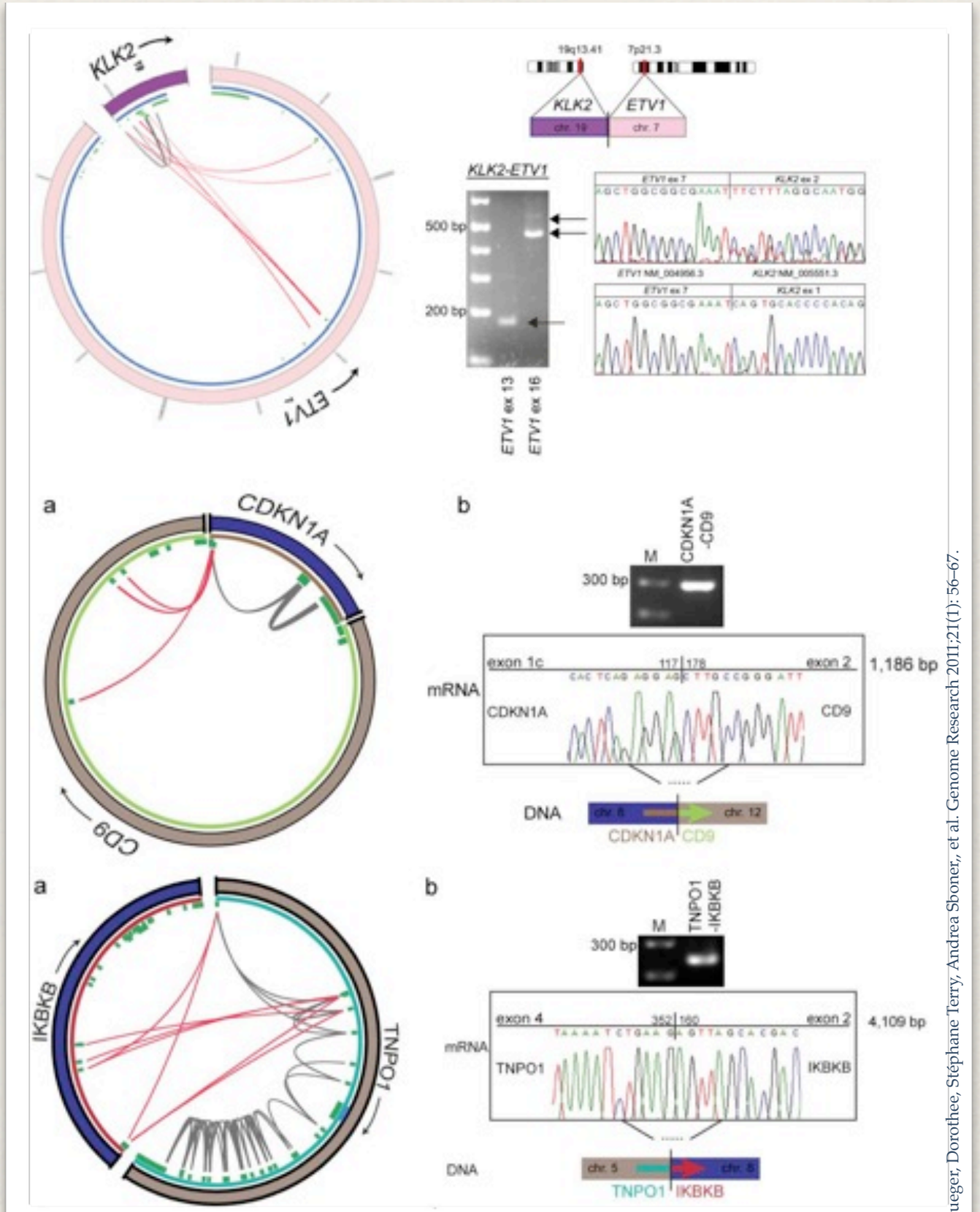
Type	Sample ID	Fusion Candidate	RESPER
intra	**580_B**	**TMPRSS2-ERG**	**14.31**
intra	**1700_D**	**TMPRSS2-ERG**	**8.79**
intra	**106_T**	**TMPRSS2-ERG**	**3.97**
inter	**2621_D**	**SLC45A3-ERG**	**3.56**
inter	**1700_D**	**ERG-GMPR**	**2.05**
read-through	*1700_D*	*SLC16A8-BAIAP2L2*	*1.93*
read-through	*106_T*	*AK094188-AK311452*	*1.9*
read-through	*1700_D*	*ZNF473-FLJ26850*	*1.58*
read-through	*580_B*	*ZNF577-FLJ26850*	*1.58*
read-through	*1043_D*	*ZNF577-ZNF649*	*1.55*
read-through	*1700_D*	*CAMTA2-INCA1*	*1.35*
inter	1700_D	HDAC5	1.29
read-through	*1043_D*	*FLJ00248-LRCH4*	*1.27*
read-through	*1700_D*	*VMAC-CAPS*	*1.17*
read-through	*106_T*	*FLJ00248-LRCH4*	*1.16*
cis	1043_D	AX747861-FLI1	1.13
read-through	*106_T*	*TAGLN-AK126420*	*1.07*
inter	**580_B**	**PIGU-ALG5**	**1.07**
inter	**99_T**	**NDRG1-ERG**	**1.02**



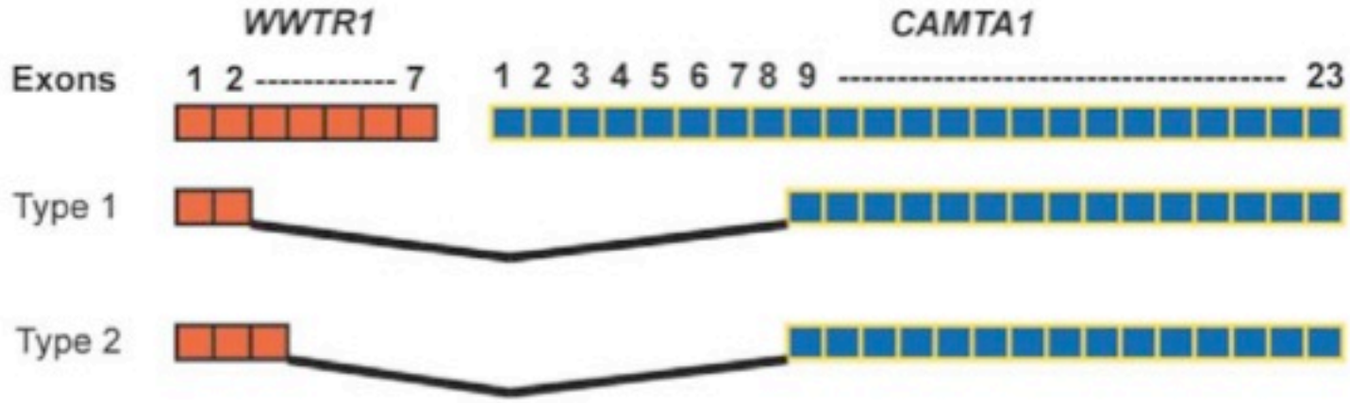


Other data sets?

- Larger prostate cancer data set
- Identified 7 novel fusion candidates
- Experimentally validated

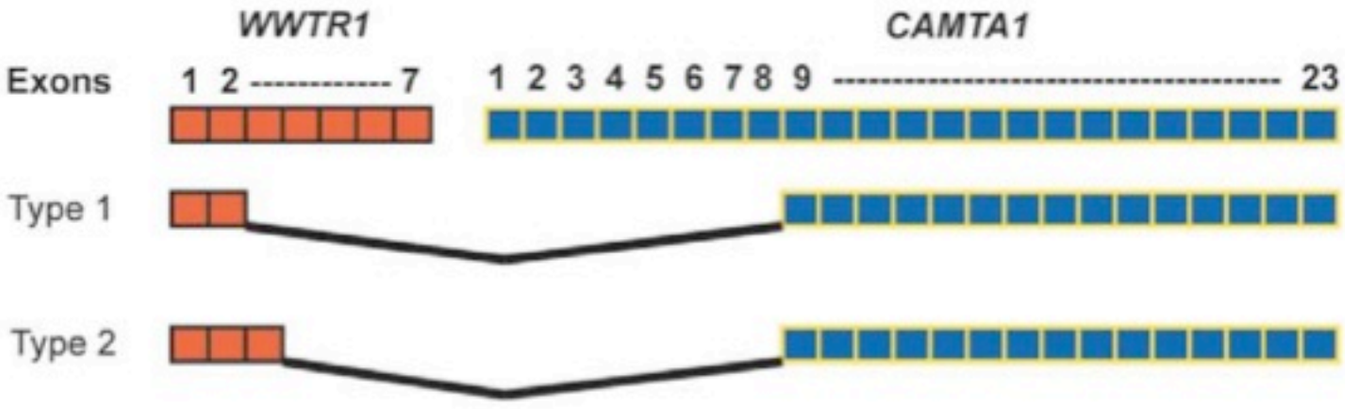


Pflueger, Dorothee, Stéphane Terry, Andrea Sboner,, et al. Genome Research 2011;21(1): 56–67.



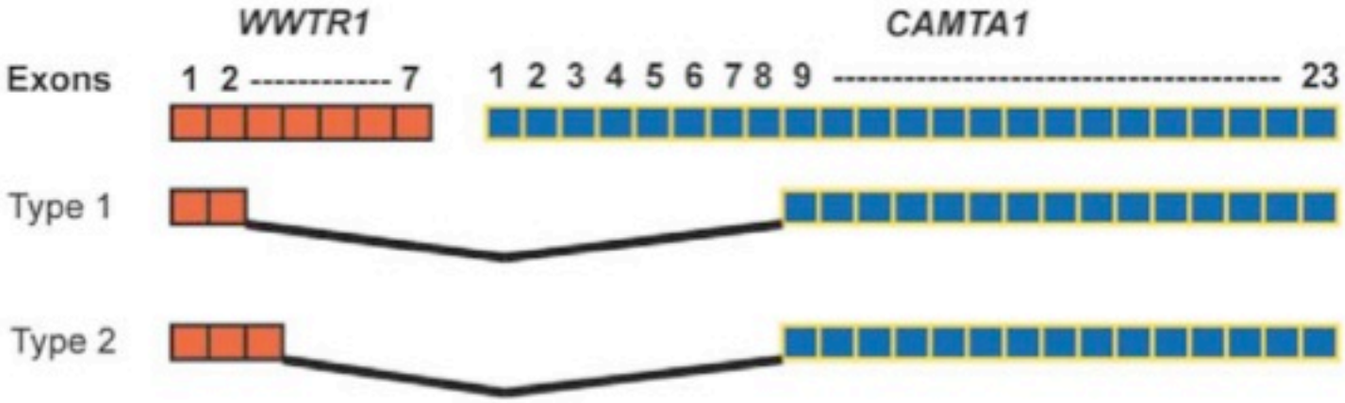
Novel *defining* fusion gene in epitheliod hemangioendothelioma

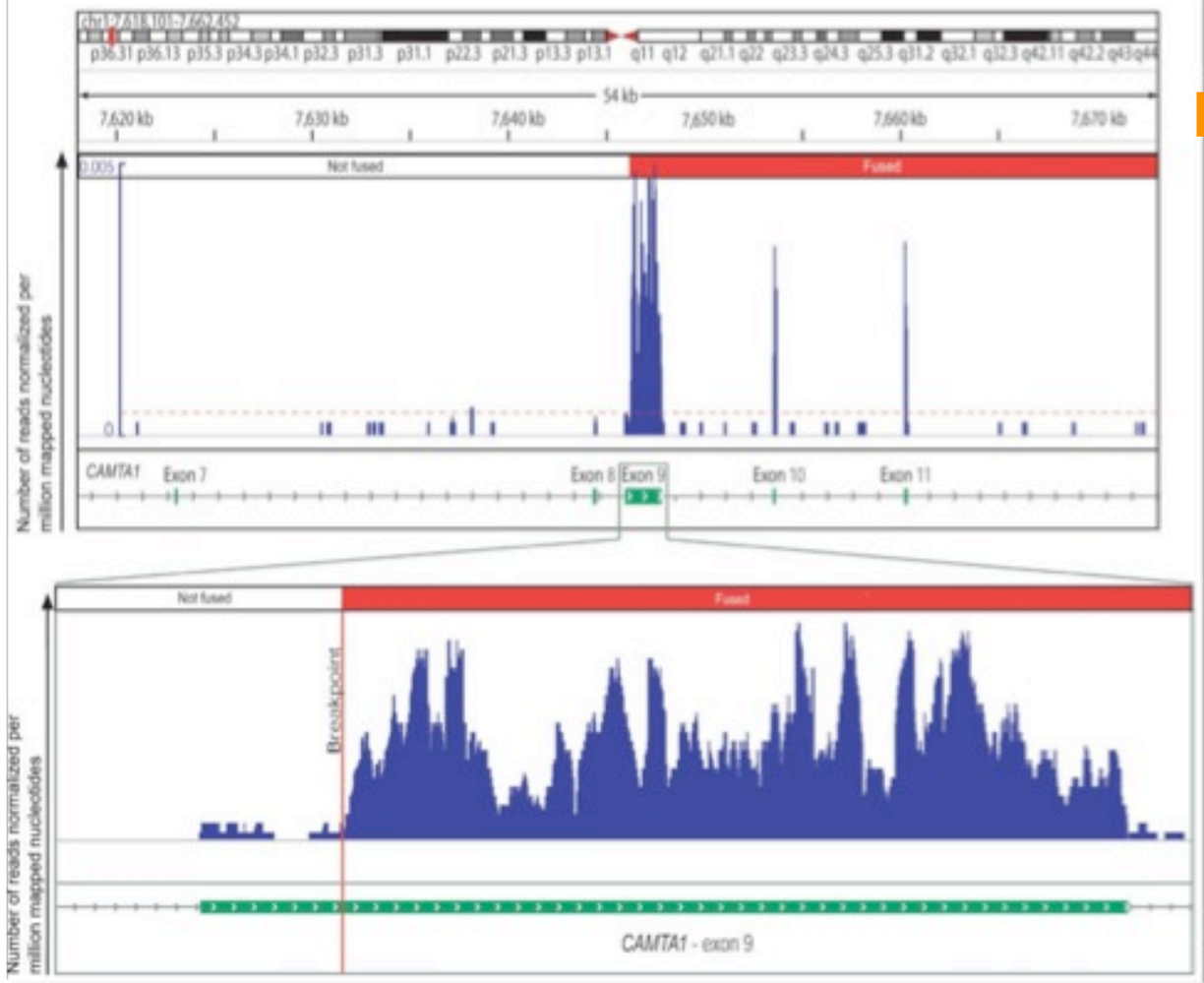
Tanas, Munir R, Andrea Sboner,et al. "Identification of a Disease-Defining Gene Fusion in Epithelioid Hemangioendothelioma." Science Translational Medicine 3, no. 98 (August 31, 2011): 98ra82–98ra82.



Novel *defining* fusion gene in epithelioid hemangioendothelioma

Tanas, Munir R, Andrea Sboner,et al. "Identification of a Disease-Defining Gene Fusion in Epithelioid Hemangioendothelioma." Science Translational Medicine 3, no. 98 (August 31, 2011): 98ra82–98ra82.



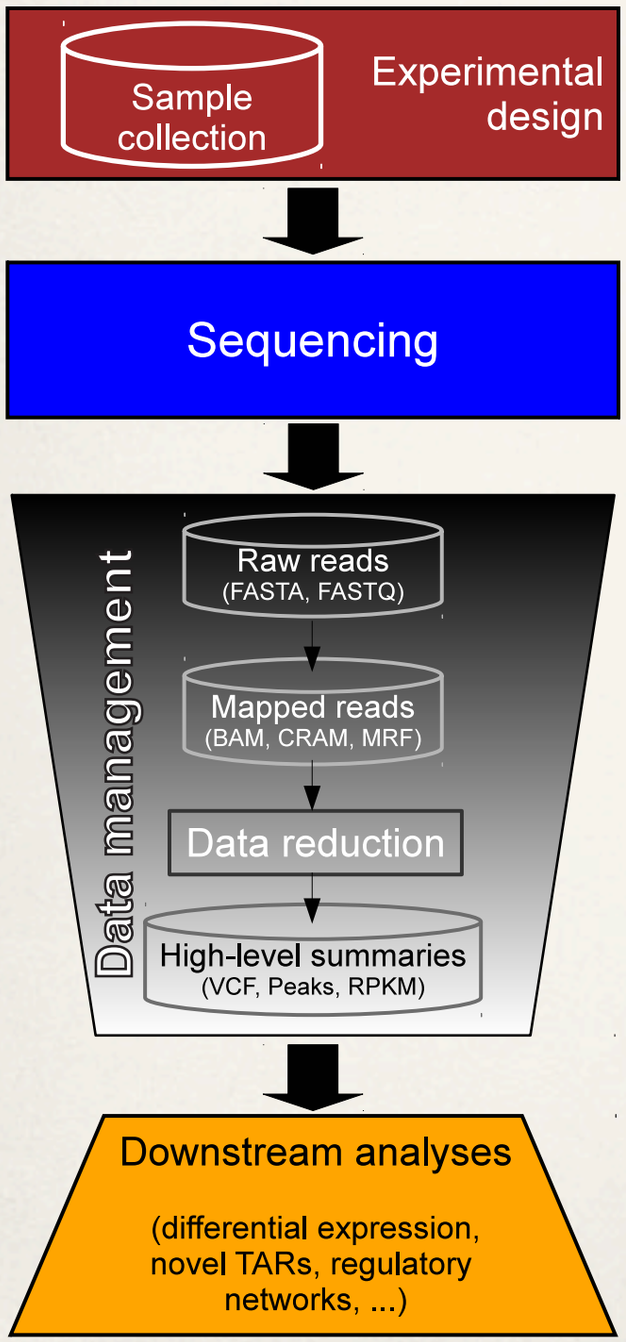


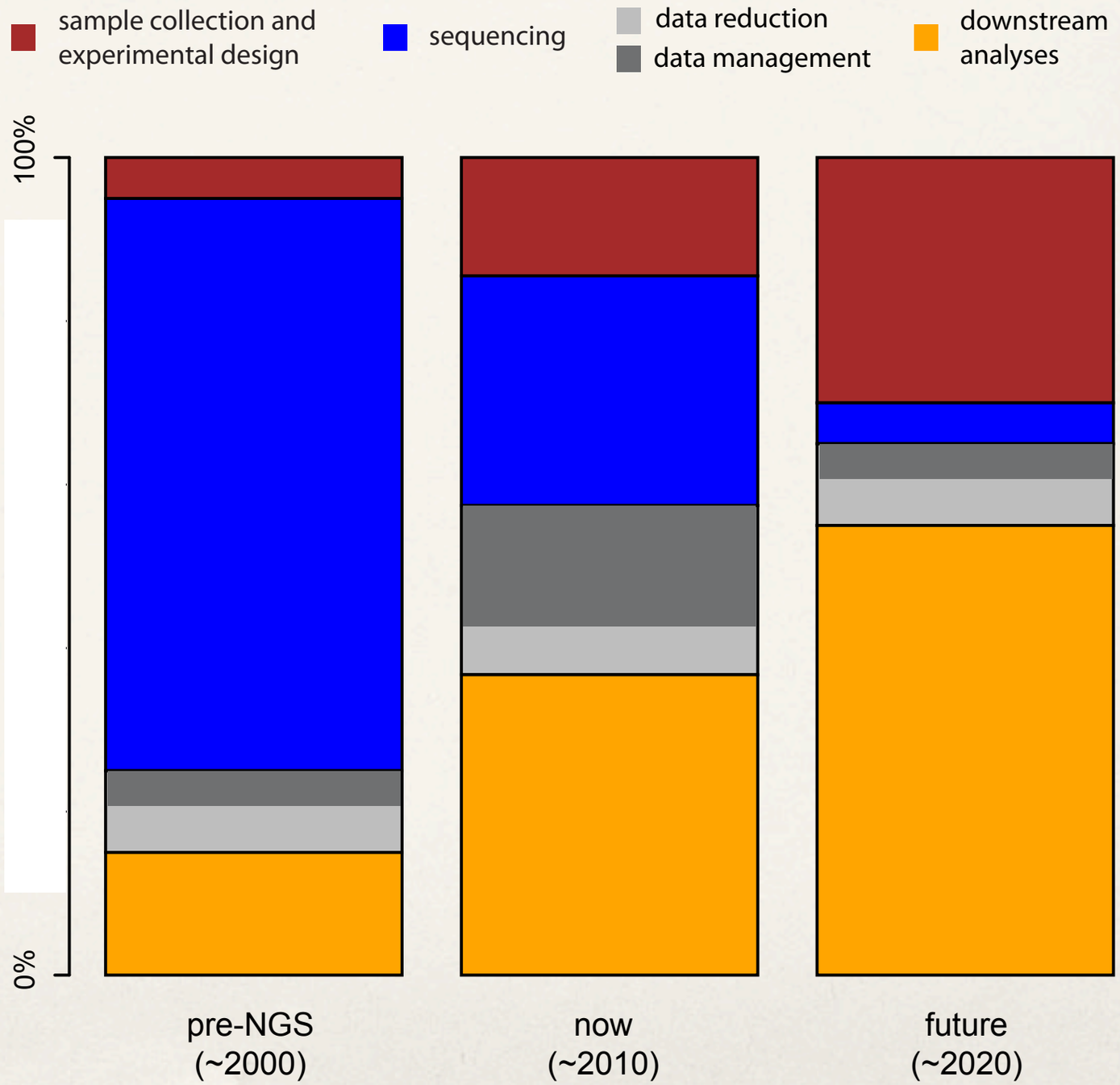
G	WWTR1		CAMTA1	
	Positive /total	%	Positive /total	%
Epithelioid hemangioendothelioma	42/47	89%	39/45	87%
Angiosarcoma, NOS	0/42	0%	0/39	0%
Epithelioid angiosarcoma	0/7	0%	0/7	0%
Intimal sarcoma	0/5	0%	0/3	0%
Kaposi's sarcoma	0/4	0%	0/4	0%
Malignant hemangioendothelioma, NOS	0/1	0%	0/1	0%
Retiform hemangioendothelioma	0/1	0%	0/1	0%
Kaposiform hemangioendothelioma	0/3	0%	0/2	0%
Epithelioid hemangioma	0/5	0%	0/4	0%
Arteriovenous malformation	0/2	0%	0/2	0%
Angiomatosis	0/1	0%	0/1	0%
Hemangioma, NOS	0/3	0%	0/3	0%
Capillary/pyogenic hemangioma	0/5	0%	0/5	0%
Cavernous hemangioma	0/5	0%	0/5	0%
Juvenile hemangioma	0/1	0%	0/1	0%
Spindle cell hemangioma	0/4	0%	0/4	0%
Synovial hemangioma	0/1	0%	0/1	0%
Intramuscular hemangioma	0/6	0%	0/5	0%
Littoral cell hemangioma	0/6	0%	0/2	0%
Malignant hemangiopericytoma	0/1	0%	0/1	0%
Hemangiopericytoma, NOS	0/1	0%	0/1	0%
Sinonasal hemangiopericytoma	0/1	0%	0/1	0%
Glomus tumor	0/1	0%	0/1	0%
Atypical glomus tumor	0/2	0%	0/2	0%
Lymphangioma	0/7	0%	0/7	0%
Lymphangioleiomyomatosis	0/1	0%	0/1	0%
Papillary endothelial hyperplasia	0/2	0%	0/2	0%
Total cases	165		151	

Tanas, Munir R, Andrea Sboner,et al. "Identification of a Disease-Defining Gene Fusion in Epithelioid Hemangioendothelioma." Science Translational Medicine 3, no. 98 (August 31, 2011): 98ra82–98ra82.

Novel *defining* fusion gene in epitheliod hemangioendothelioma

NGS future trends





Sboner, Andrea, et al. 2011. Genome Biology 12(8):25:125

Yale
Lukas Habegger Tara Gianoulis
Joel Rozowsky Ashish Agarwal
David Chen Jing Leng
Mark Gerstein Mike Snyder

Weill-Cornell Medical College
Dorothee Pflueger Stephane Terry
Francesca Demichelis Wasay Hussein
Naoki Kitabayashi Benjamin Moss
Mark Rubin

Cleveland Clinic
Tanas Munir
Brian P Rubin

Broad Institute
Michael F. Berger
Gad Getz Todd Golub
Levi A. Garraway

National Institutes of Health

National Cancer Institute
at the National Institutes of Health

genome.gov
National Human Genome Research Institute
National Institutes of Health

HHMI
HOWARD HUGHES MEDICAL INSTITUTE

U.S. Department of Defense

STARR CANCER
CONSORTIUM

DF/HCC
DANA-FARBER / HARVARD CANCER CENTER

THE KOHLBERG FOUNDATION

PROSTATE CANCER
FOUNDATION
Accelerating the world's most promising research

BREAST CANCER
BREAKTHROUGH

Acknowledgments