NATIONAL GENOMICS INFRASTRUCTURE

Transcriptome and isoform reconstruction with long reads

Adam Ameur, PhD
NGI, Uppsala Genome Center
adam.ameur@igp.uu.se

Enabler for Life Sciences

Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

National Genomics Infrastructure

NGI staff: 60 -70 FTE, including head of facility, lab research engineers, bioinformaticians, IT-experts, project coordinators.

UPPMAX/UPPNEX: Uppsala multidisciplinary center for advanced computational science, UPPNEX: UPPmax NEXt generation sequencing Cluster & Storage.

Enabler for Life Sciences

Karolinska Institutet

KTH

Stockholms universitet

UPPSALA UNIVERSITET

SciLifeLab

DNA sequencing at all scales

One of the most well-equipped NGS sites in Europe!

10	Illumina HiSeq Xten
17	Illumina HiSeq 2000/2500
3	Illumina MiSeq
1	Illumina NextSeq
2	**Life Technologies Ion Torrent**
6	**Life Technologies Ion Proton**
2	**Pacific Biosciences RSII**
2	**Sanger ABI3730**
1	Argus Whole Genome Map. Syst.
1	Oxford Nanopore MinIon

Enabler for Life Sciences

RNA-sequencing

with short reads

Enabler for Life Sciences

Karolinska Institutet

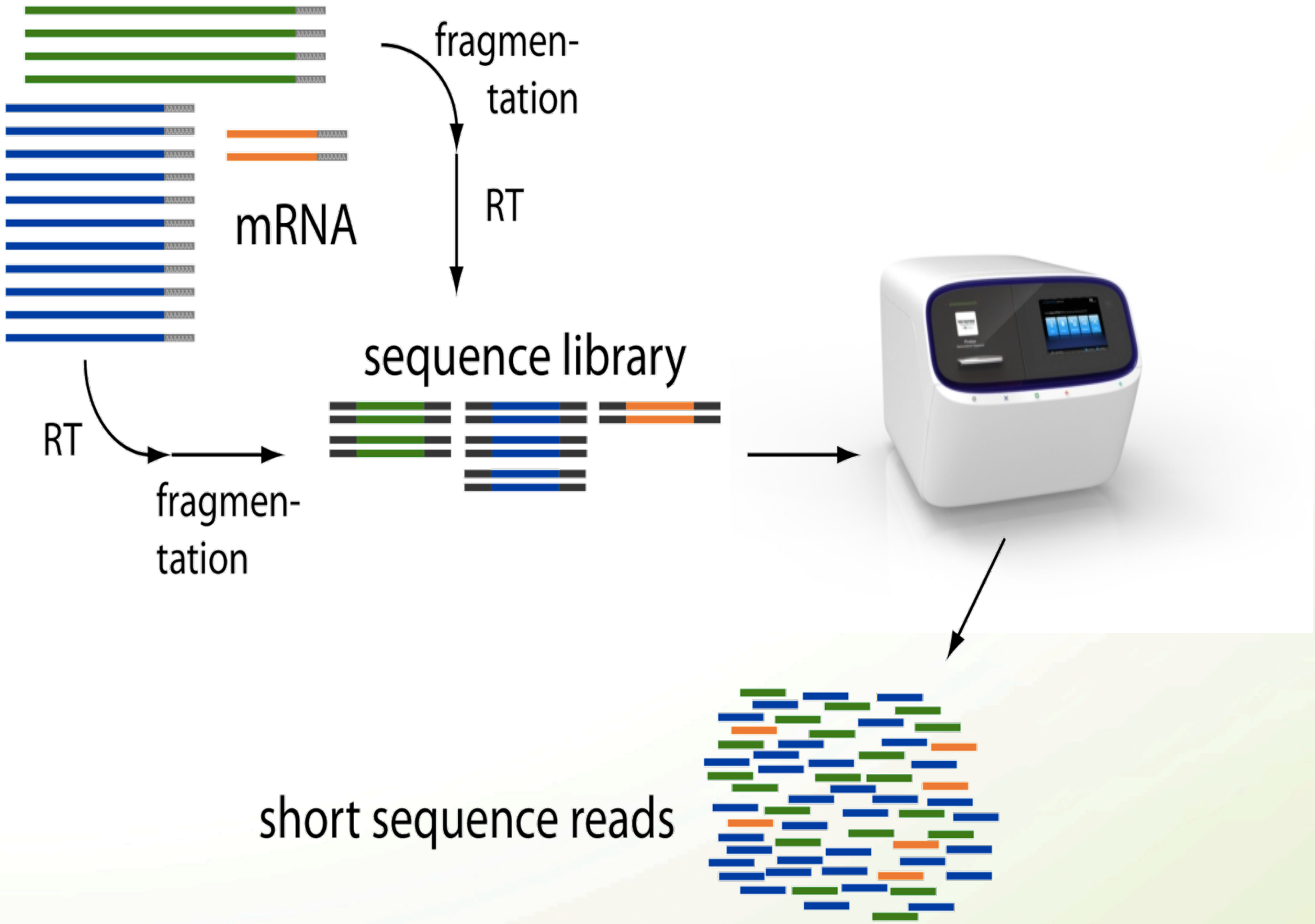
KTH

Stockholms universitet

UPPSALA UNIVERSITET

SciLifeLab

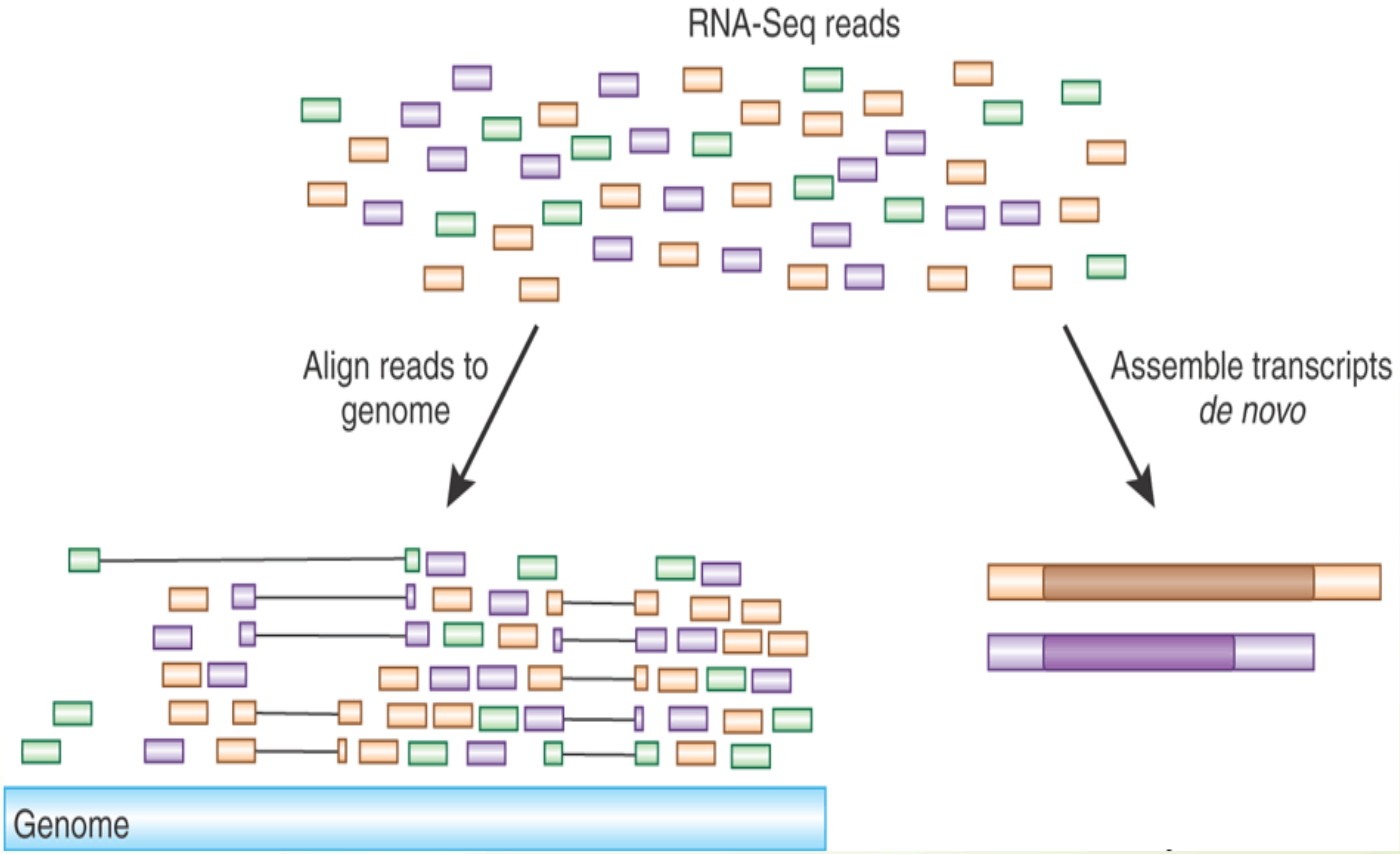
RNA-seq: standard procedure



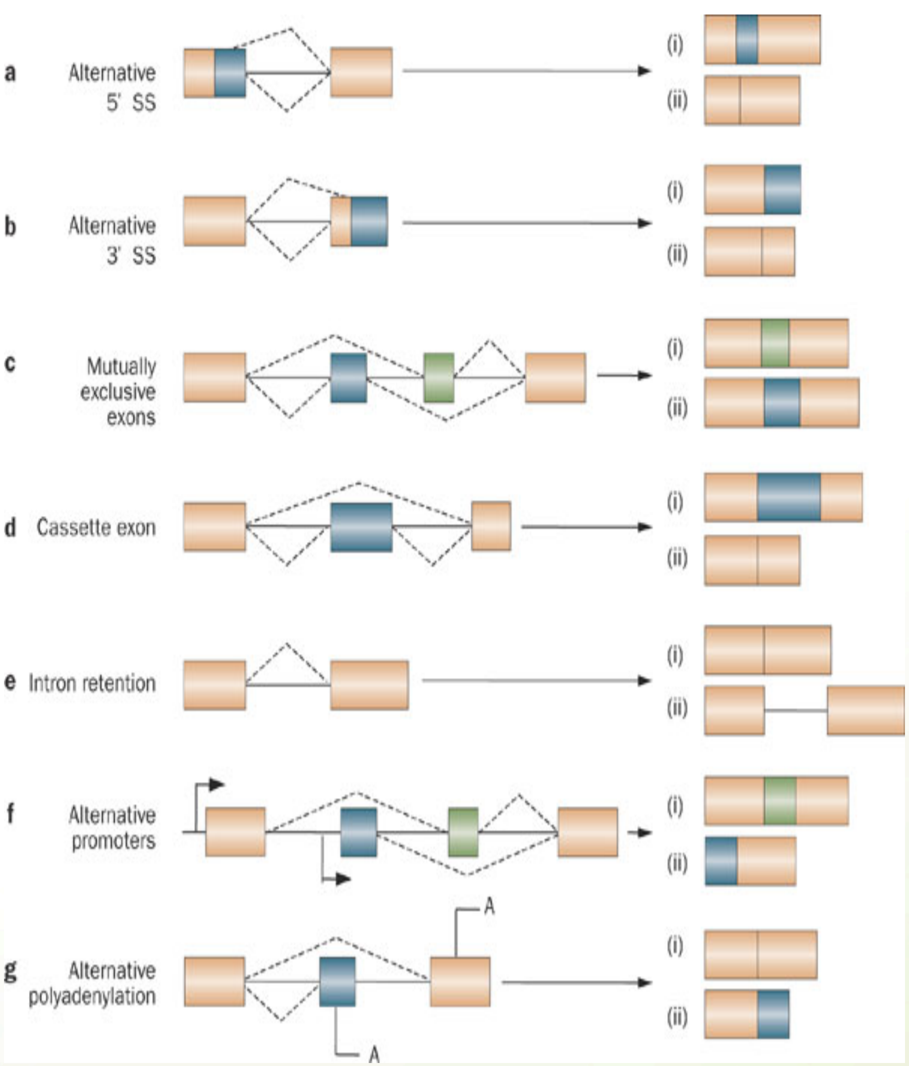
RNA-seq: the main question

What to do with this?

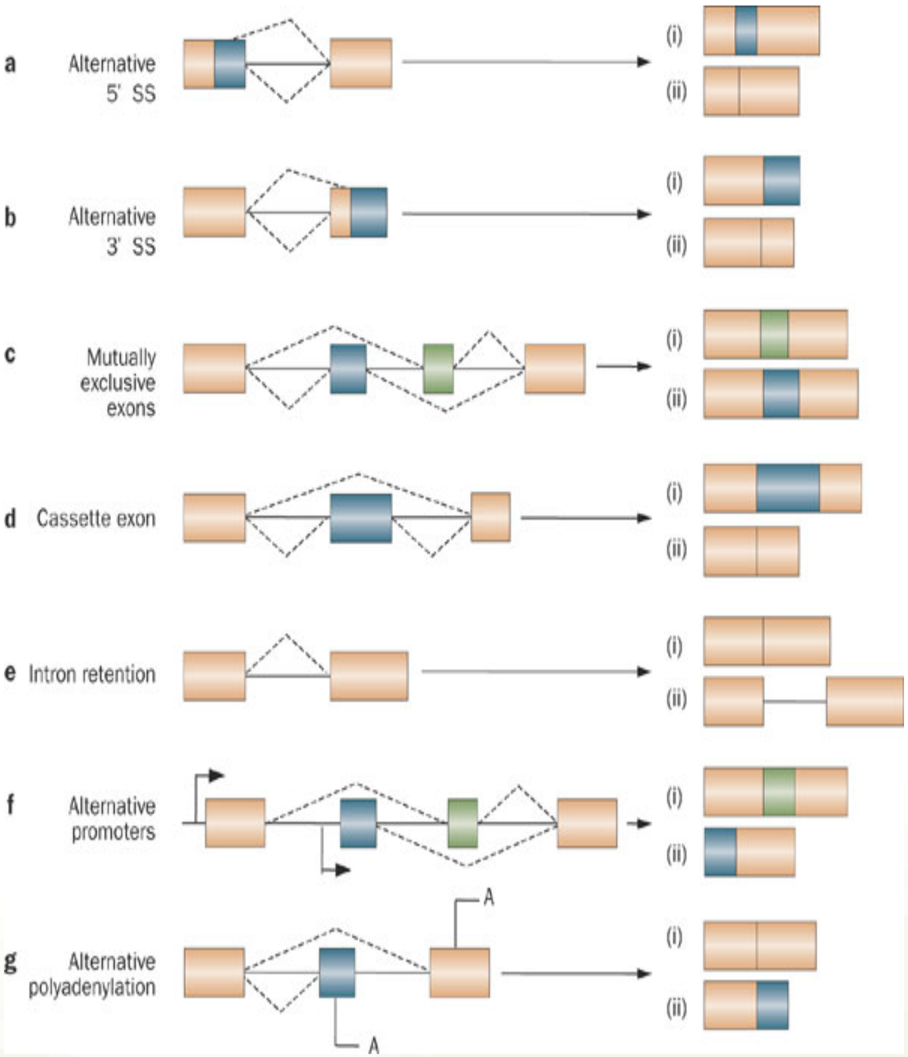
RNA-seq: analysis

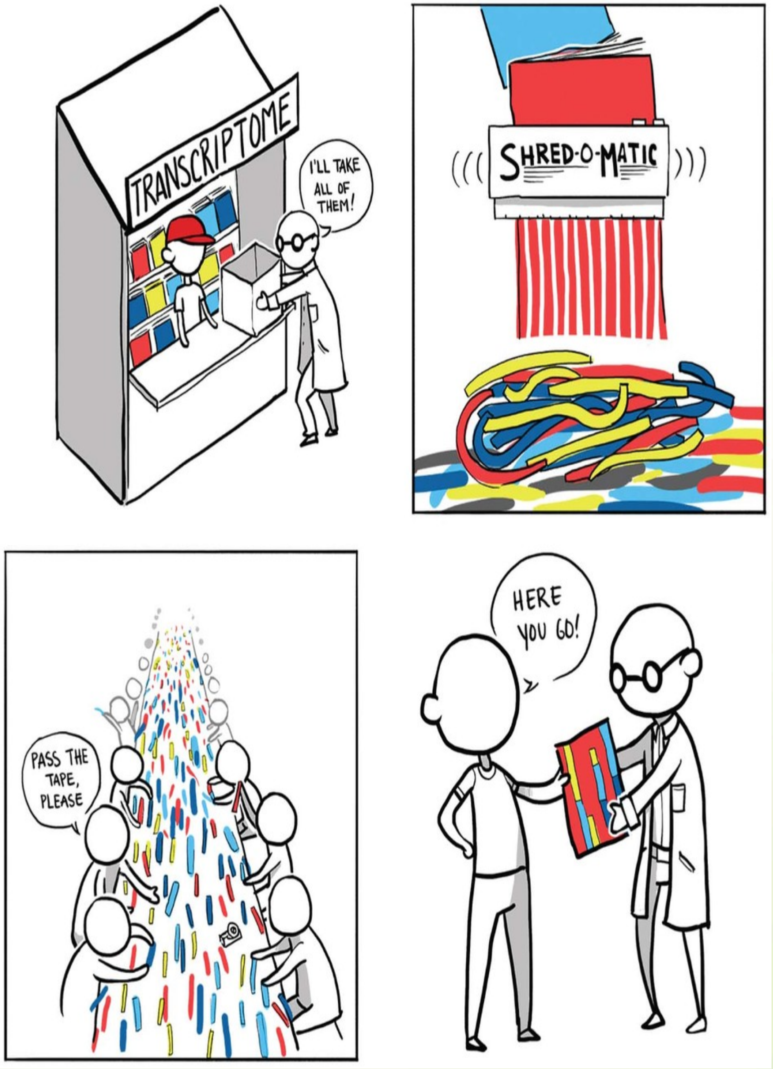


Complicating factor: alternative splicing

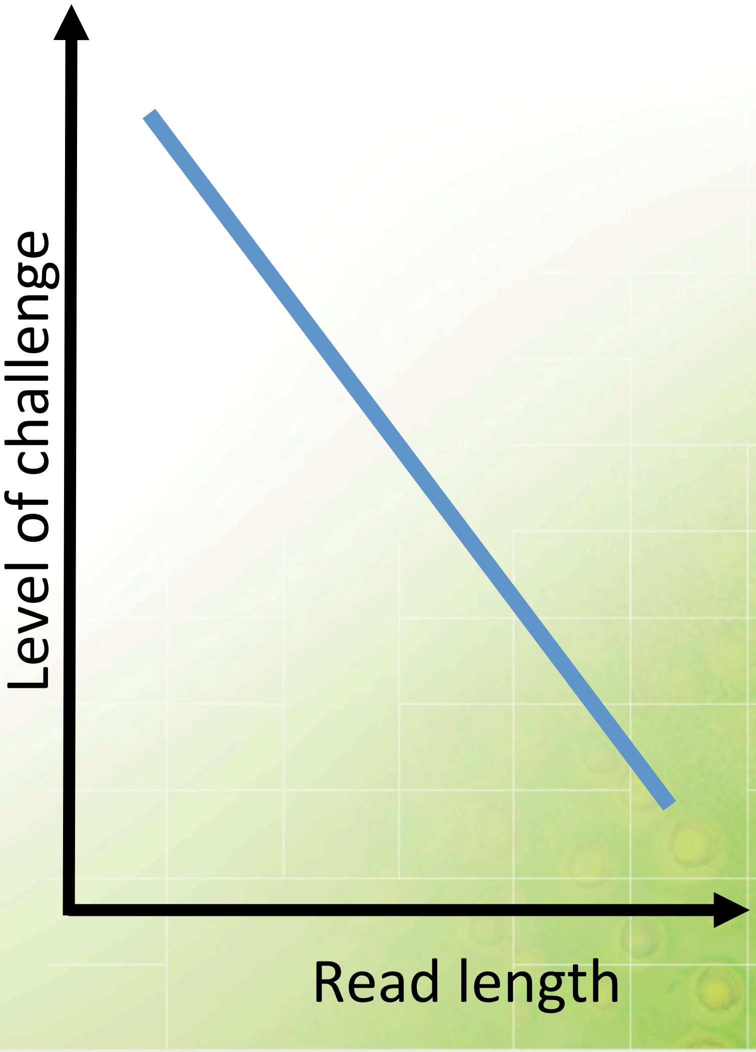


RNA-seq: problem with short reads





RNA-seq: problem with short reads



Enabler for Life Sciences

RNA-sequencing

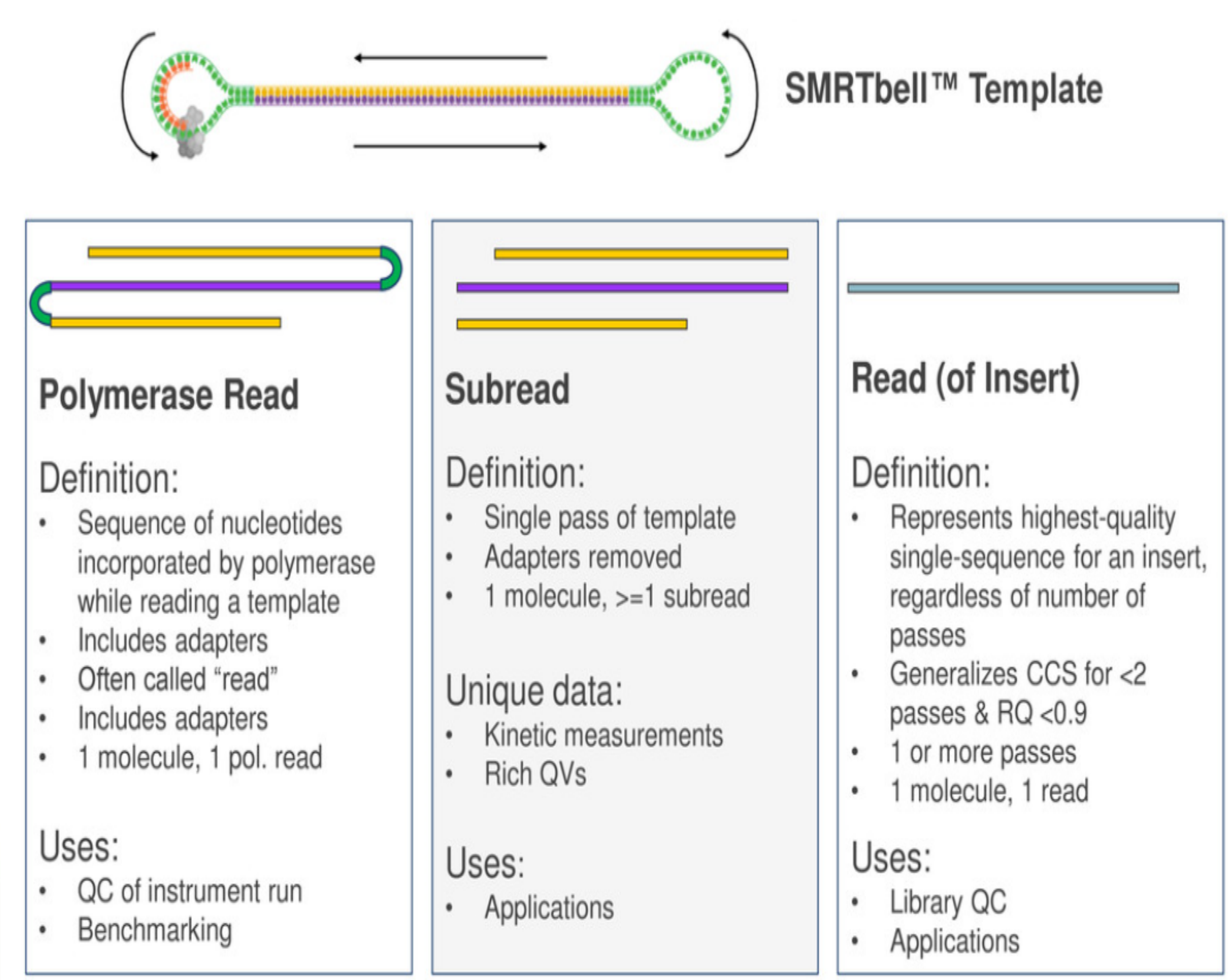
with very long reads!!!

Pacific Biosciences RS II

- Pacific Biosciences
  - Single molecule sequencing
  - Very long read lengths (up to 30 kb)
  - Rapid sequencing
  - Can detect base modifications (e.g. methylation)
  - Relatively low throughput

Pacific Biosciences RSII

PacBio – Sequencing Template



Polymerase Read

Definition:

- Sequence of nucleotides incorporated by polymerase while reading a template
- Includes adapters
- Often called "read"
- Includes adapters
- 1 molecule, 1 pol. read

Uses:

- QC of instrument run
- Benchmarking

Subread

Definition:

- Single pass of template
- Adapters removed
- 1 molecule, >=1 subread

Unique data:

- Kinetic measurements
- Rich QVs

Uses:

- Applications

Read (of Insert)

Definition:

- Represents highest-quality single-sequence for an insert, regardless of number of passes
- Generalizes CCS for <2 passes & RQ <0.9
- 1 or more passes
- 1 molecule, 1 read

Uses:

- Library QC
- Applications

PacBio output

- PacBio throughput

  ~ 500Mb-1Gb/SMRT cell

~1 bacterial genome

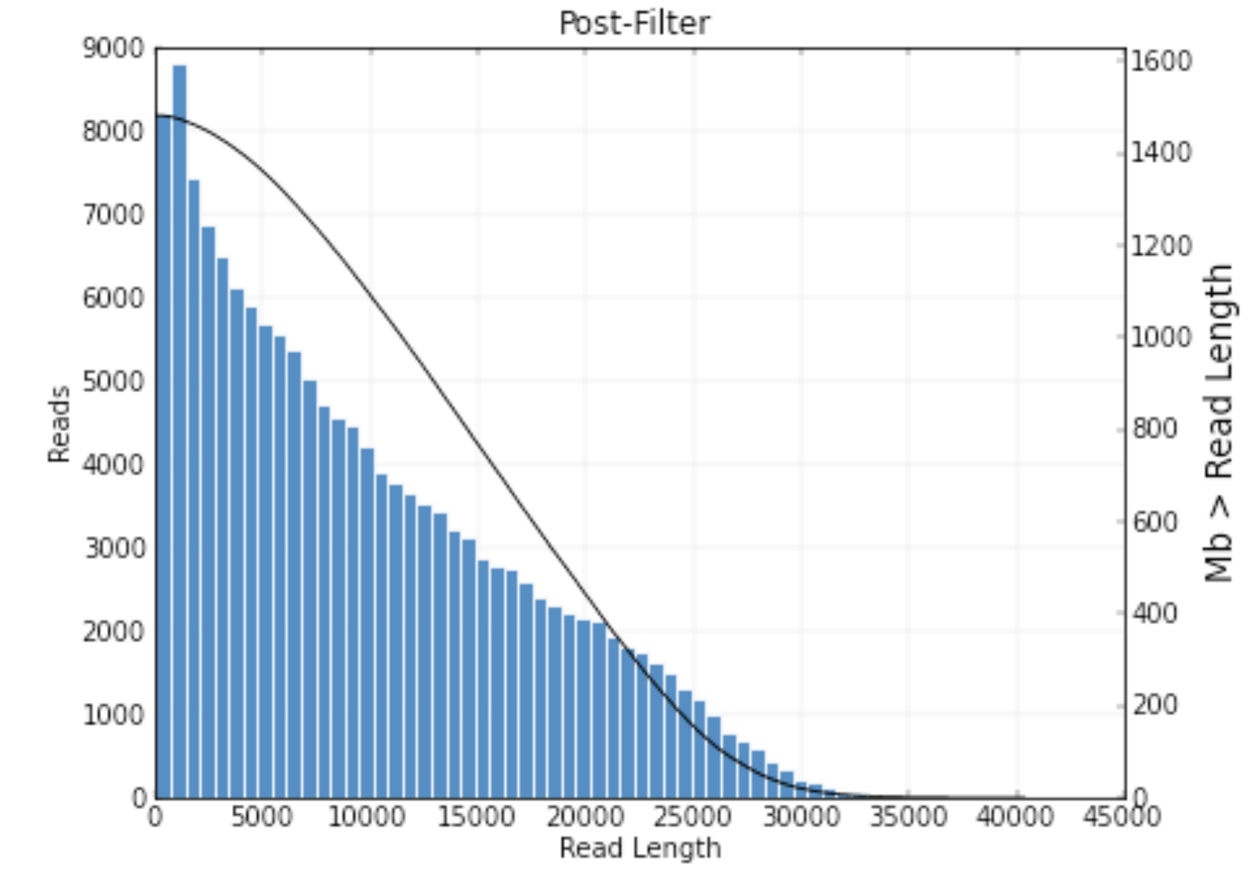
~1 bacterial transcriptome

1 human genome = 150 SMRT cells

- PacBio read lengths: 500bp-30kb

PacBio – Current read lengths

- >10kb average read lengths! (run from April 2014)

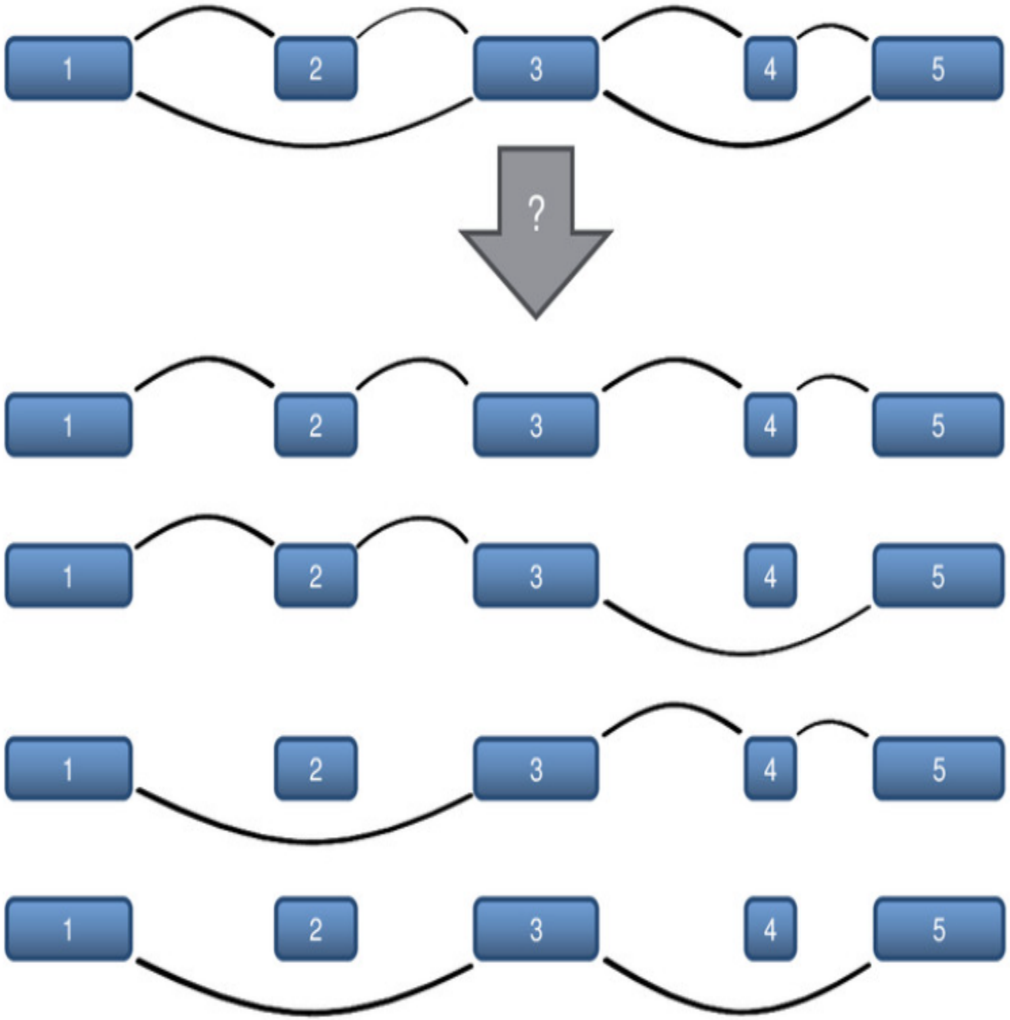


Iso-Seq: Full length RNA-seq on PacBio!

- Single molecule sequencing
  - One read – one transcript
- Transcript in full length
  - No assembly required
- No systematic bias
  - CG-rich, AT-rich, tandem repeats

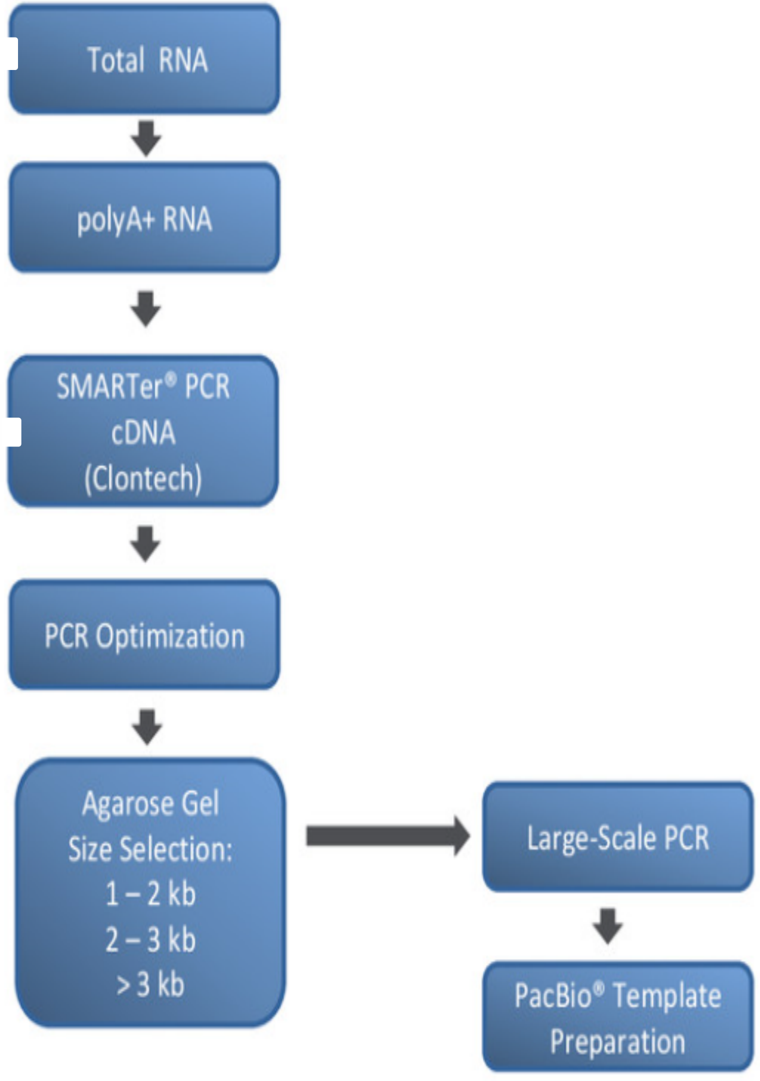
RNA-sequencing on PacBio

On average, 8 alt. isoforms per gene in human



Candidate space: 5.8 x 10^{76}

PacBio Iso-Seq - library preparation



Total RNA

polyA+ RNA

SMARTer® PCR cDNA (Clontech)

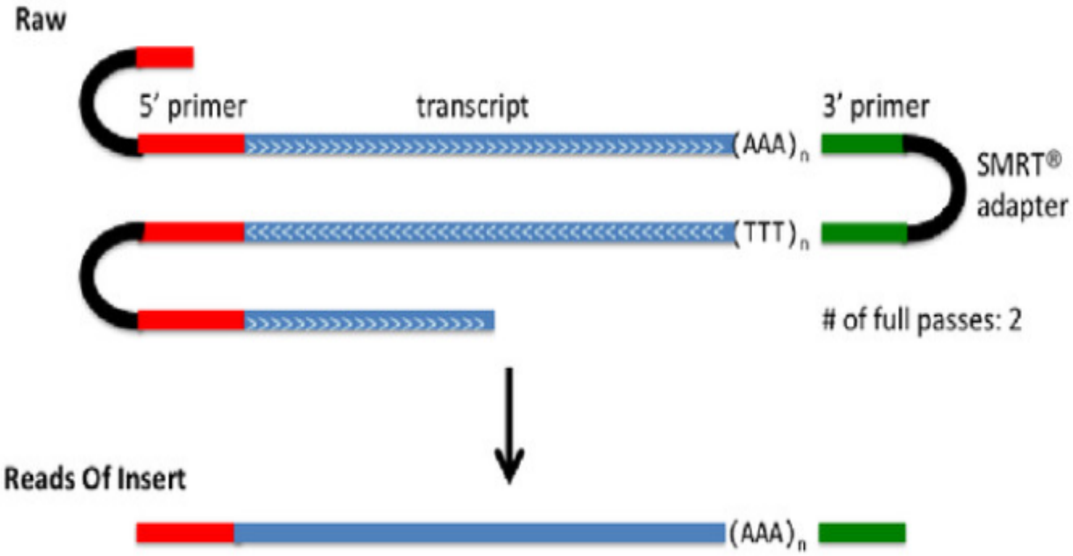
PCR Optimization

Agarose Gel Size Selection:
1 – 2 kb
2 – 3 kb
> 3 kb

Large-Scale PCR

PacBio® Template Preparation

PacBio Iso-Seq – reads of insert

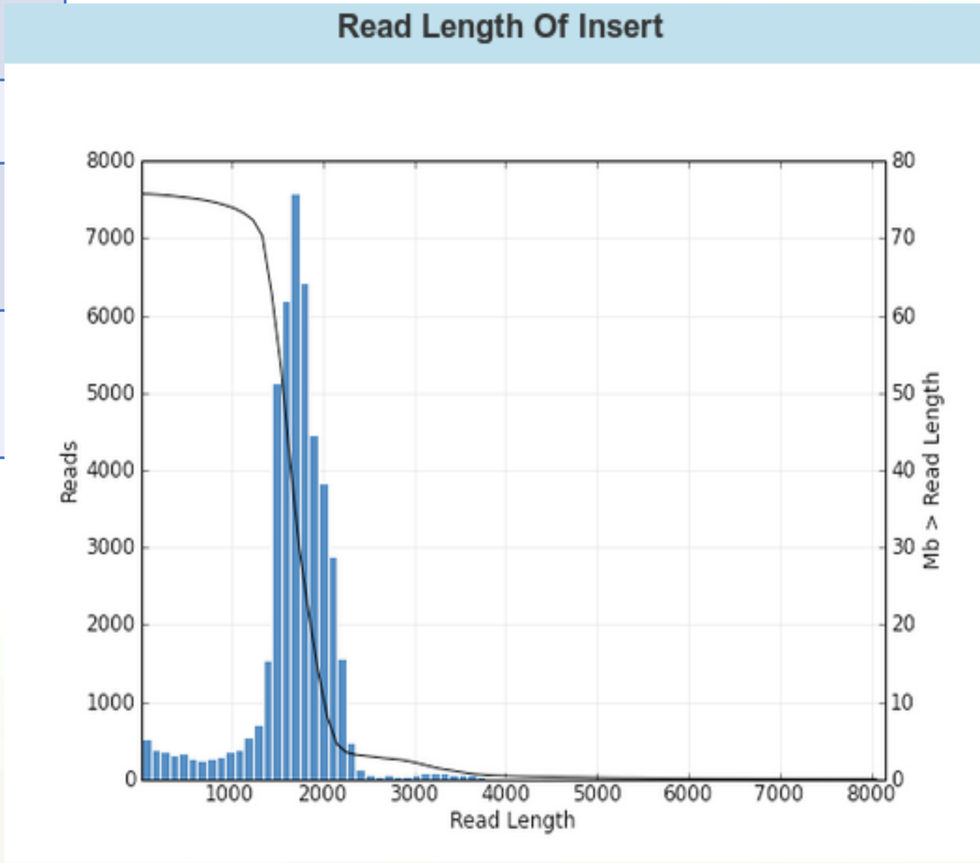


Full-Length = 5' primer seen, polyA tail seen, 3' primer seen

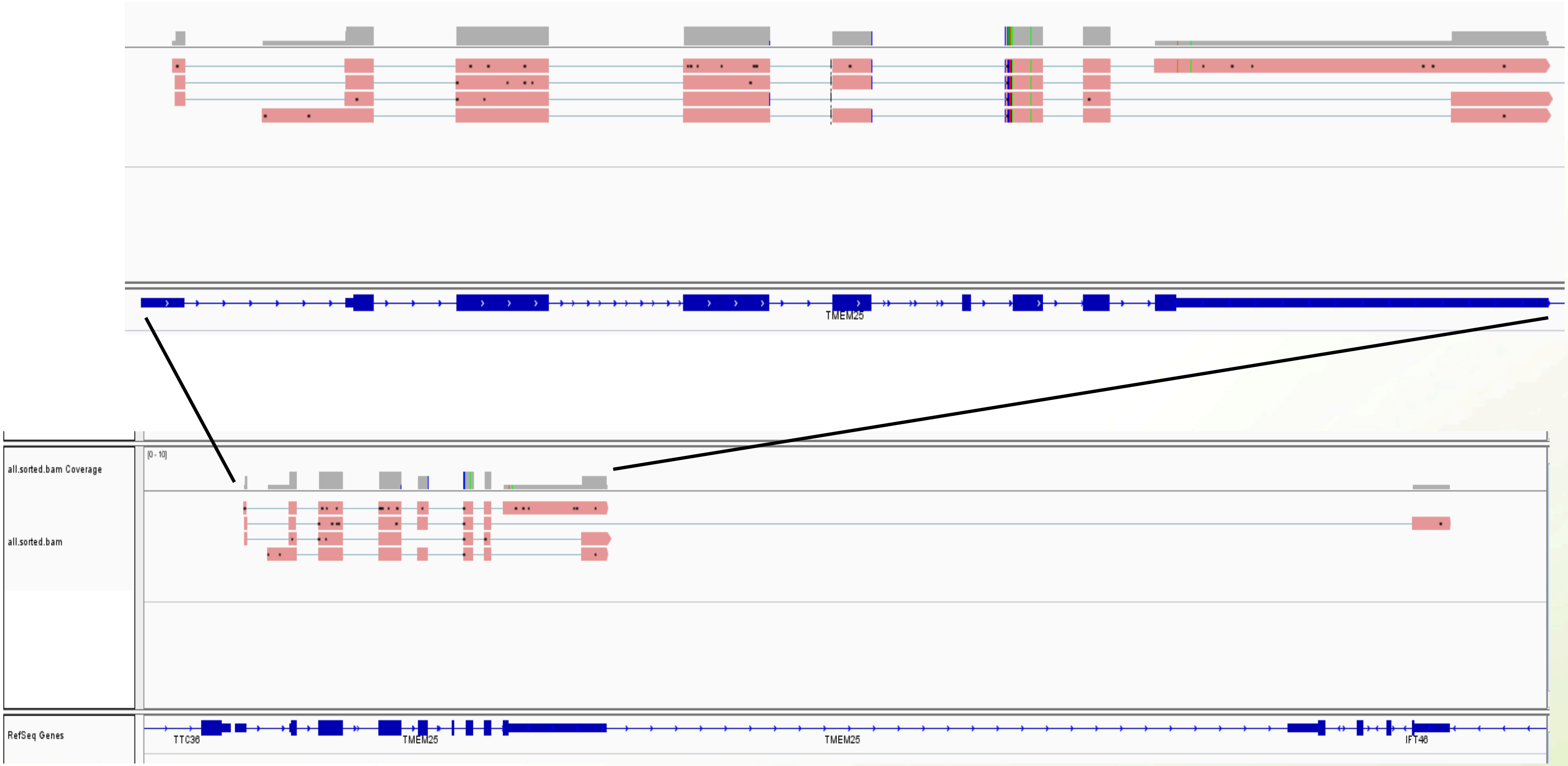
- Identify and remove primers and polyA/T tail
- Identify read strandedness

PacBio Iso-Seq: ROI of 1kb lib (2 cells)

Read Bases of Insert	78,108,189
Mean Read Length of Insert	1,687
Mean Number of Passes	**8.0**
Number of full-length non-chimeric reads	**35,467**
Average full-length non-chimeric read length	1,679

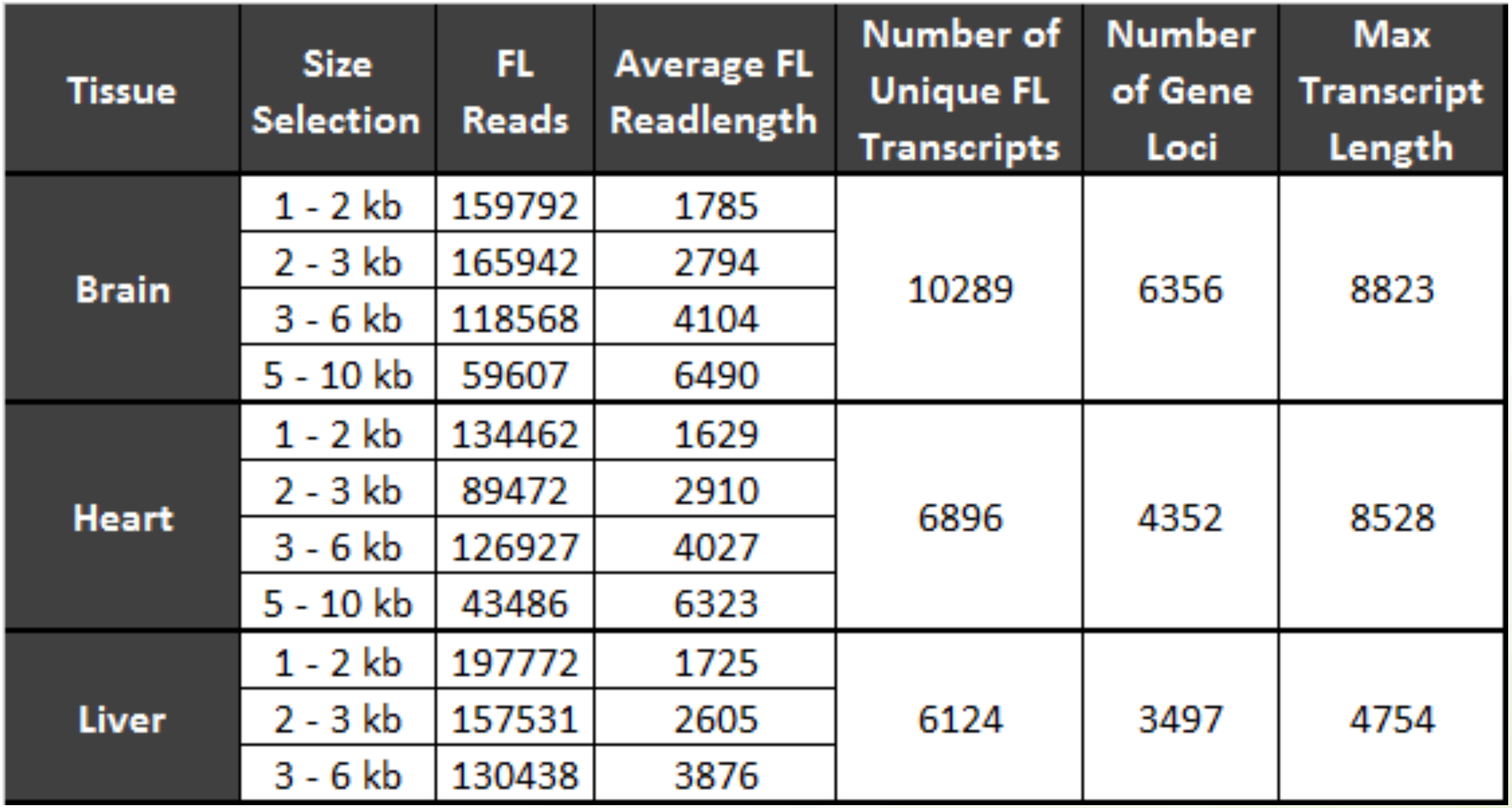


PacBio Iso-Seq: examples



GeneCards Summary for TMEM25 Gene:
TMEM25 is a protein-coding gene. Diseases associated with TMEM25 include breast cancer.

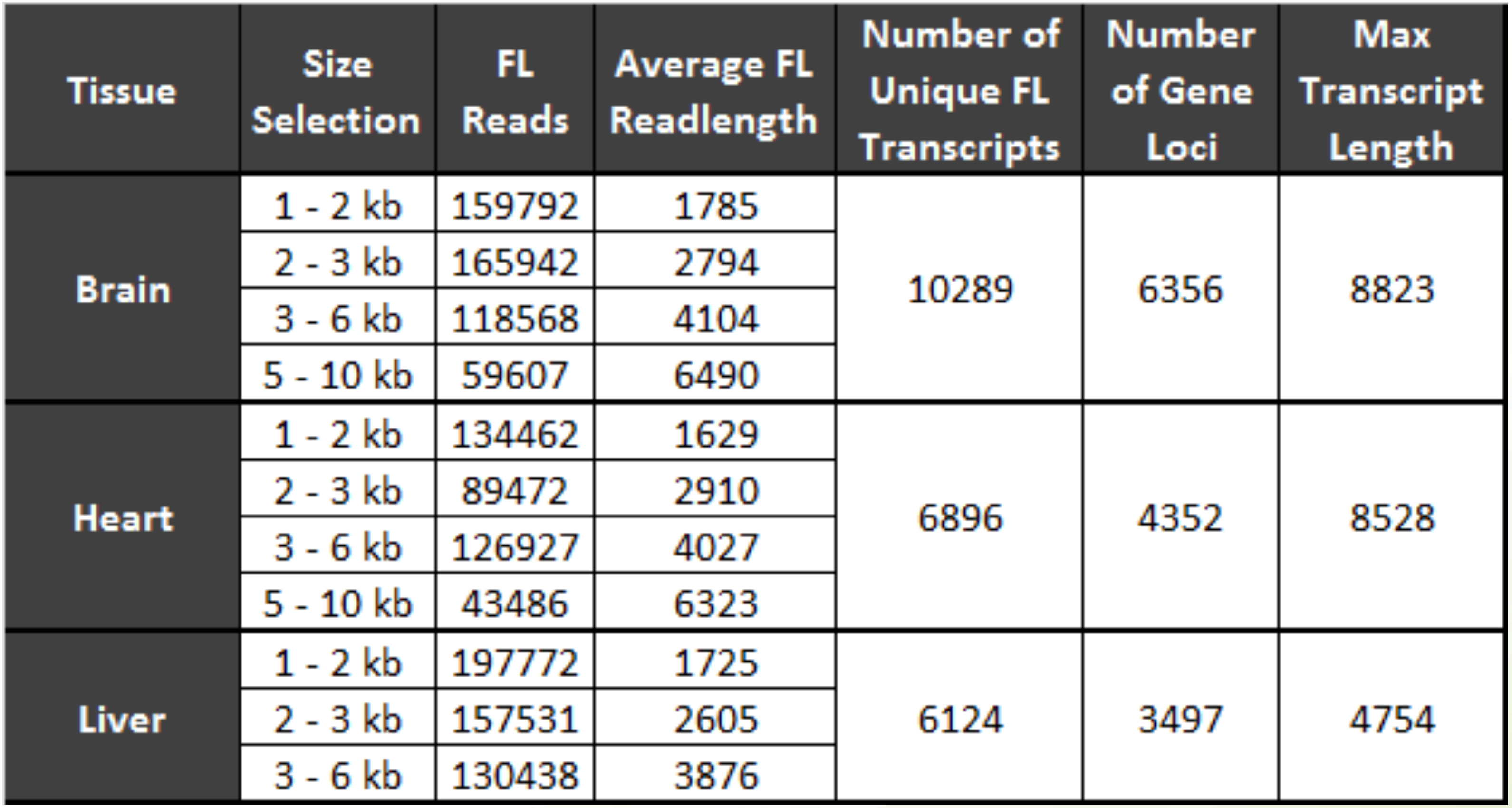
PacBio Iso-Seq: examples (from PacBio)



Tissue	Size Selection	FL Reads	Average FL Readlength	Number of Unique FL Transcripts	Number of Gene Loci	Max Transcript Length
Brain	1 - 2 kb	159792	1785	10289	6356	8823
	2 - 3 kb	165942	2794			
	3 - 6 kb	118568	4104			
	5 - 10 kb	59607	6490			
Heart	1 - 2 kb	134462	1629	6896	4352	8528
	2 - 3 kb	89472	2910			
	3 - 6 kb	126927	4027			
	5 - 10 kb	43486	6323			
Liver	1 - 2 kb	197772	1725	6124	3497	4754
	2 - 3 kb	157531	2605			
	3 - 6 kb	130438	3876			

http://blog.pacificbiosciences.com/2014/10/data-release-whole-human-transcriptome.html

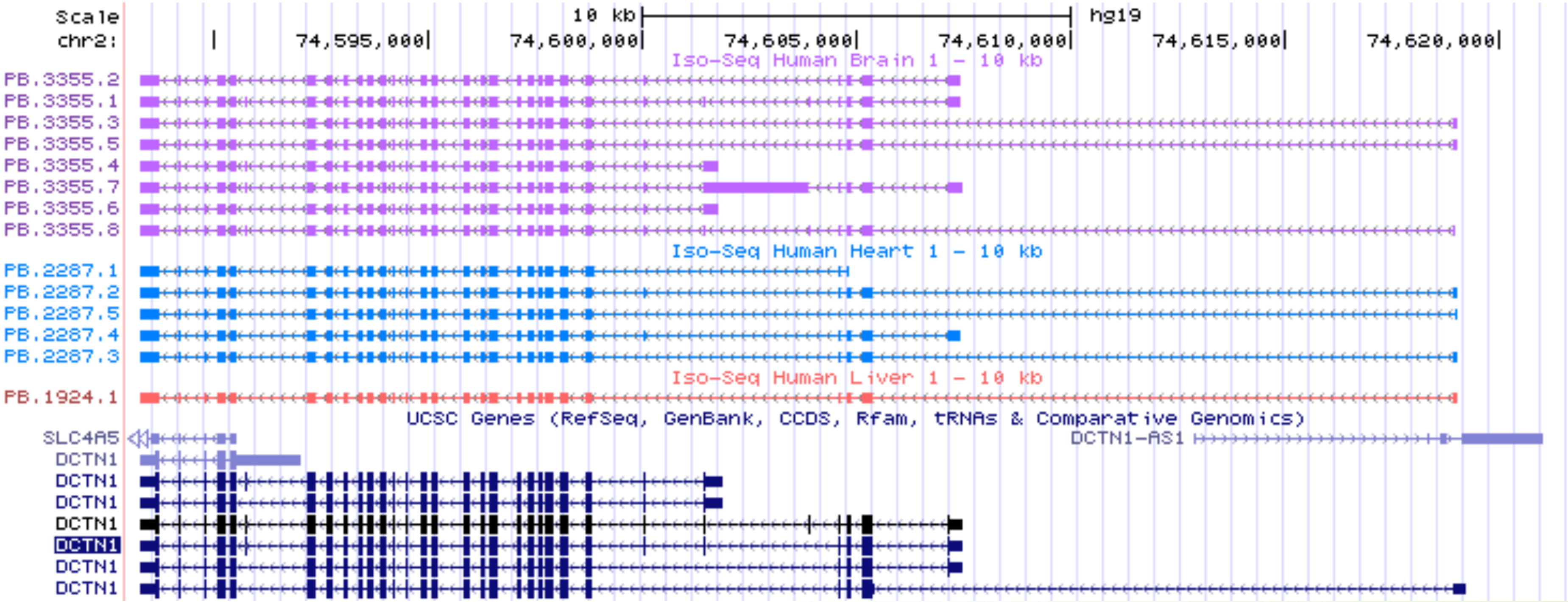
PacBio Iso-Seq: examples (from PacBio)



Tissue	Size Selection	FL Reads	Average FL Readlength	Number of Unique FL Transcripts	Number of Gene Loci	Max Transcript Length
Brain	1 - 2 kb	159792	1785	10289	6356	8823
	2 - 3 kb	165942	2794			
	3 - 6 kb	118568	4104			
	5 - 10 kb	59607	6490			
Heart	1 - 2 kb	134462	1629	6896	4352	8528
	2 - 3 kb	89472	2910			
	3 - 6 kb	126927	4027			
	5 - 10 kb	43486	6323			
Liver	1 - 2 kb	197772	1725	6124	3497	4754
	2 - 3 kb	157531	2605			
	3 - 6 kb	130438	3876			

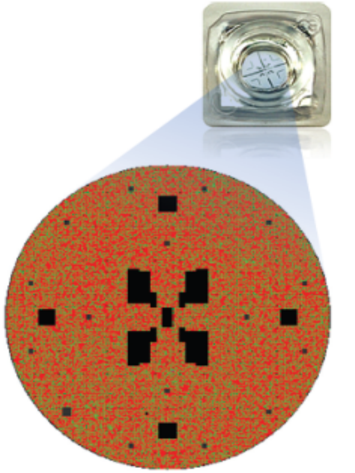
http://blog.pacificbiosciences.com/2014/10/data-release-whole-human-transcriptome.html

PacBio Iso-Seq: examples (from PacBio)



http://blog.pacificbiosciences.com/2014/10/data-release-whole-human-transcriptome.html

PacBio Iso-Seq: experimental design



So how many SMRT cells do I need?

Approximate scope guidance:

- **1 SMRT Cell**: targeted, gene-specific isoform characterization
- **1-8 SMRT Cells**: get a high-level overview of the transcriptome and isoforms of abundant transcripts
- **8-50 SMRT Cells**: get a detailed look at most transcripts and their isoforms
- **>50 SMRT Cells**: get a very thorough look at transcriptome with rare transcripts and rare isoforms or intermediates

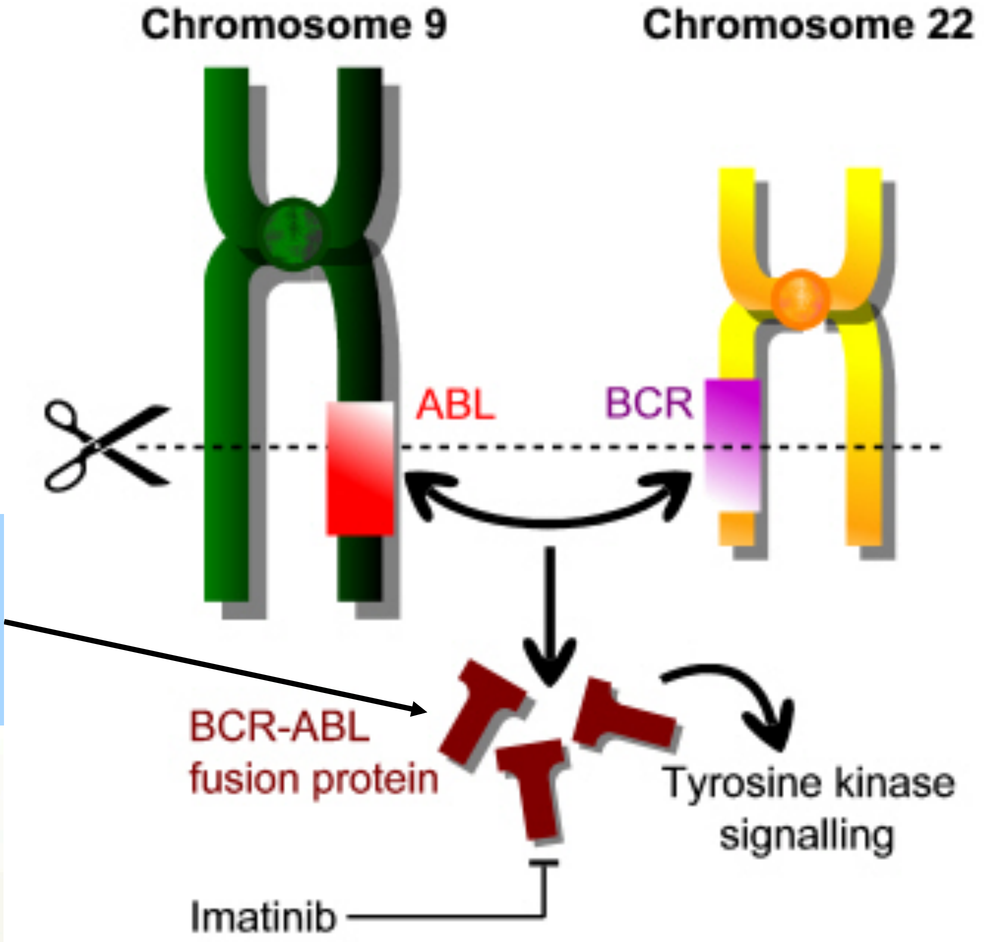
! Depends strongly on transcriptome complexity of the organism being studied !

Targeted RNA-sequencing

with very long reads!!!

Clinical project: Chronic Myeloid Leukemia

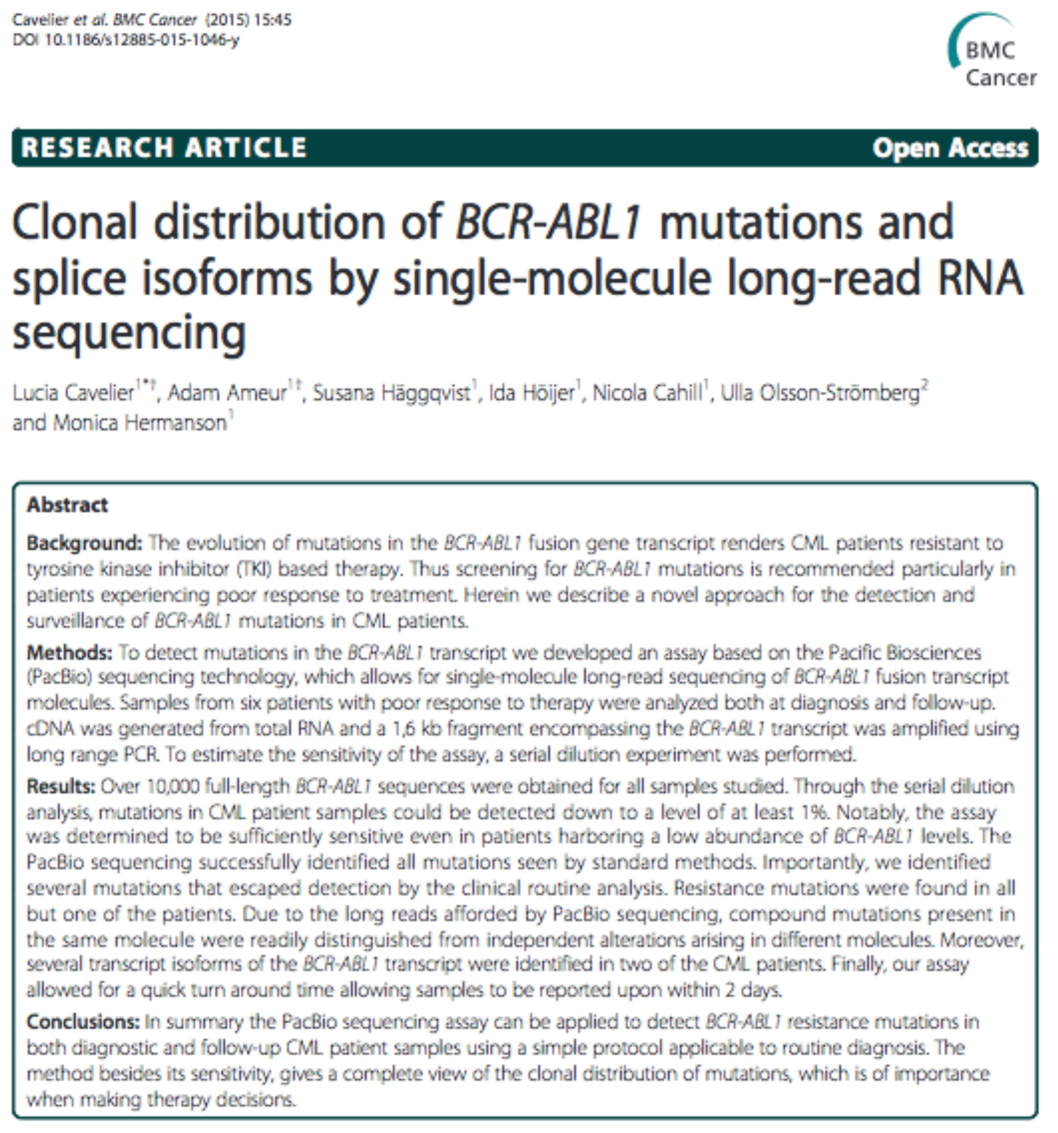
- BCR-ABL1 fusion protein – a CML drug target



The BCR-ABL1 fusion protein can acquire resistance mutations following drug treatment

www.cambridgemedicine.org/article/doi/10.7244/cmj-1355057881

BCR-ABL1 workflow – PacBio Sequencing



Cavelier *et al. BMC Cancer* (2015) 15:45
DOI 10.1186/s12885-015-1046-y

BMC Cancer

RESEARCH ARTICLE **Open Access**

Clonal distribution of *BCR-ABL1* mutations and splice isoforms by single-molecule long-read RNA sequencing

Lucia Cavelier[1*†], Adam Ameur[1†], Susana Häggqvist[1], Ida Höijer[1], Nicola Cahill[1], Ulla Olsson-Strömberg[2] and Monica Hermanson[1]

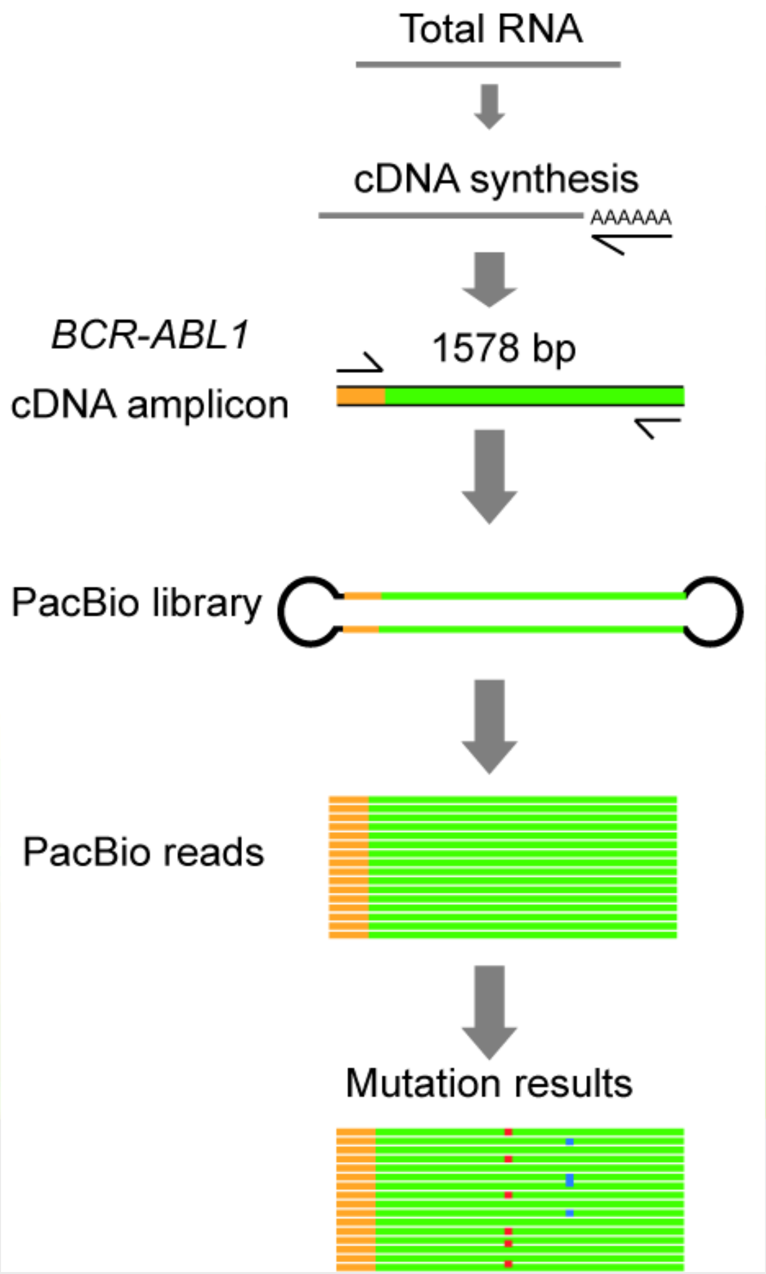
Abstract

Background: The evolution of mutations in the *BCR-ABL1* fusion gene transcript renders CML patients resistant to tyrosine kinase inhibitor (TKI) based therapy. Thus screening for *BCR-ABL1* mutations is recommended particularly in patients experiencing poor response to treatment. Herein we describe a novel approach for the detection and surveillance of *BCR-ABL1* mutations in CML patients.

Methods: To detect mutations in the *BCR-ABL1* transcript we developed an assay based on the Pacific Biosciences (PacBio) sequencing technology, which allows for single-molecule long-read sequencing of *BCR-ABL1* fusion transcript molecules. Samples from six patients with poor response to therapy were analyzed both at diagnosis and follow-up. cDNA was generated from total RNA and a 1,6 kb fragment encompassing the *BCR-ABL1* transcript was amplified using long range PCR. To estimate the sensitivity of the assay, a serial dilution experiment was performed.

Results: Over 10,000 full-length *BCR-ABL1* sequences were obtained for all samples studied. Through the serial dilution analysis, mutations in CML patient samples could be detected down to a level of at least 1%. Notably, the assay was determined to be sufficiently sensitive even in patients harboring a low abundance of *BCR-ABL1* levels. The PacBio sequencing successfully identified all mutations seen by standard methods. Importantly, we identified several mutations that escaped detection by the clinical routine analysis. Resistance mutations were found in all but one of the patients. Due to the long reads afforded by PacBio sequencing, compound mutations present in the same molecule were readily distinguished from independent alterations arising in different molecules. Moreover, several transcript isoforms of the *BCR-ABL1* transcript were identified in two of the CML patients. Finally, our assay allowed for a quick turn around time allowing samples to be reported upon within 2 days.

Conclusions: In summary the PacBio sequencing assay can be applied to detect *BCR-ABL1* resistance mutations in both diagnostic and follow-up CML patient samples using a simple protocol applicable to routine diagnosis. The method besides its sensitivity, gives a complete view of the clonal distribution of mutations, which is of importance when making therapy decisions.



Karolinska Institutet

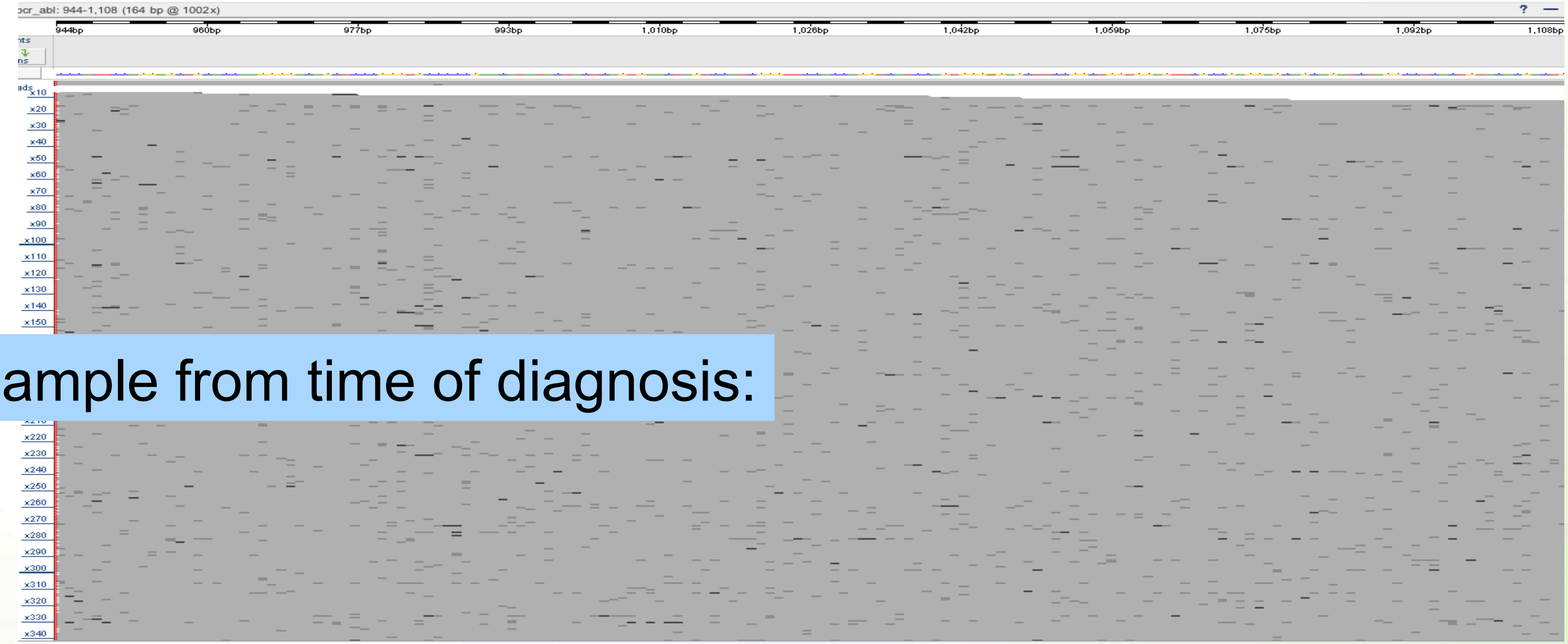
KTH VETENSKAP OCH KONST

Stockholms universitet

UPPSALA UNIVERSITET

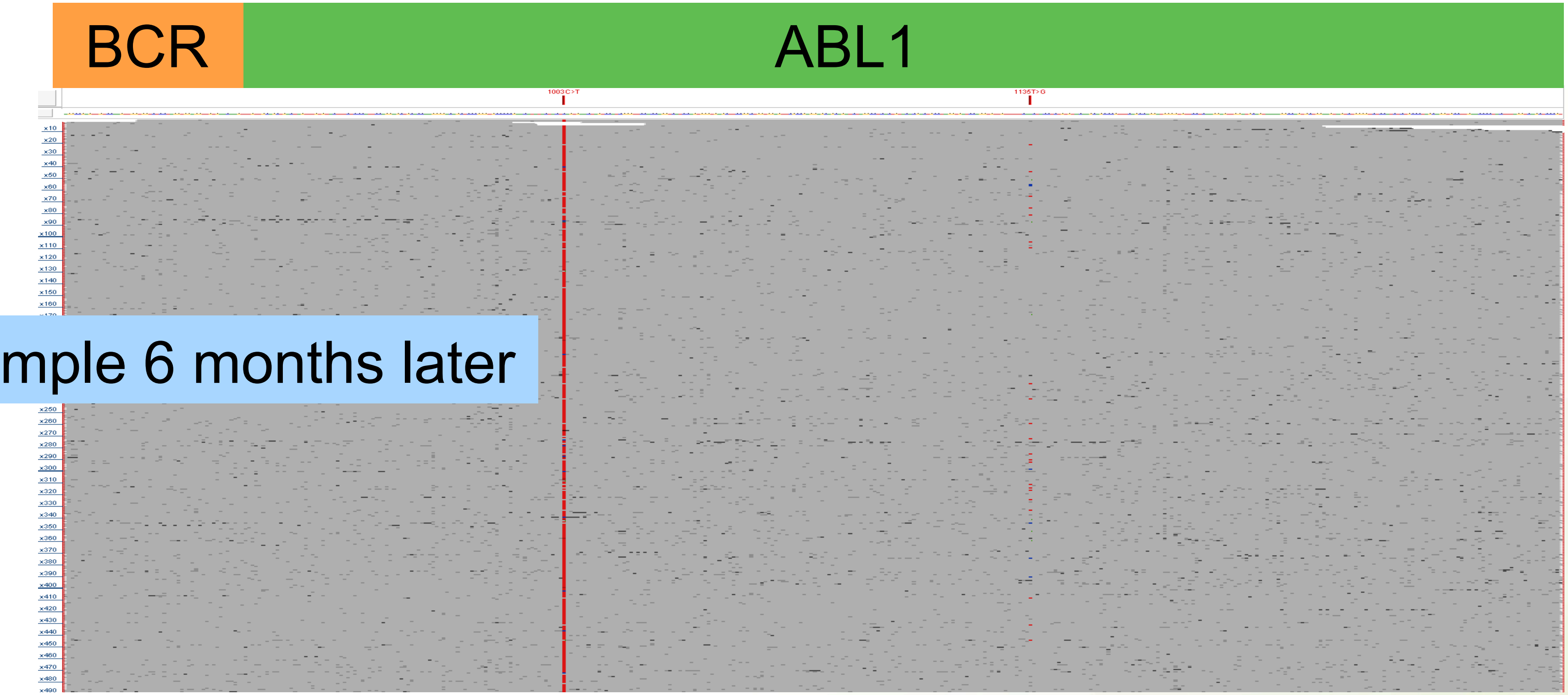
BCR-ABL1 mutations at diagnosis

PacBio sequencing generates ~10 000X coverage!


cr_abl: 944-1,108 (164 bp @ 1002x)

Sample from time of diagnosis:

BCR-ABL1 mutations in follow-up sample

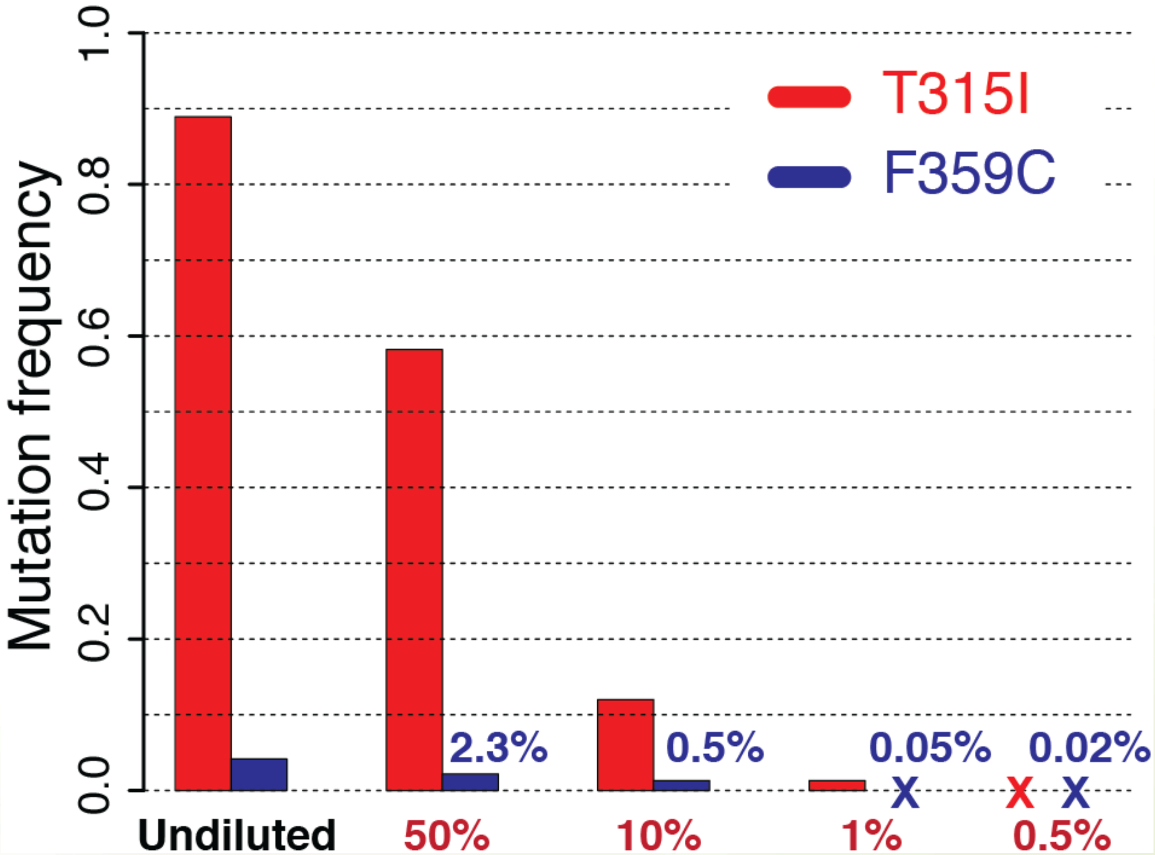


Sample 6 months later

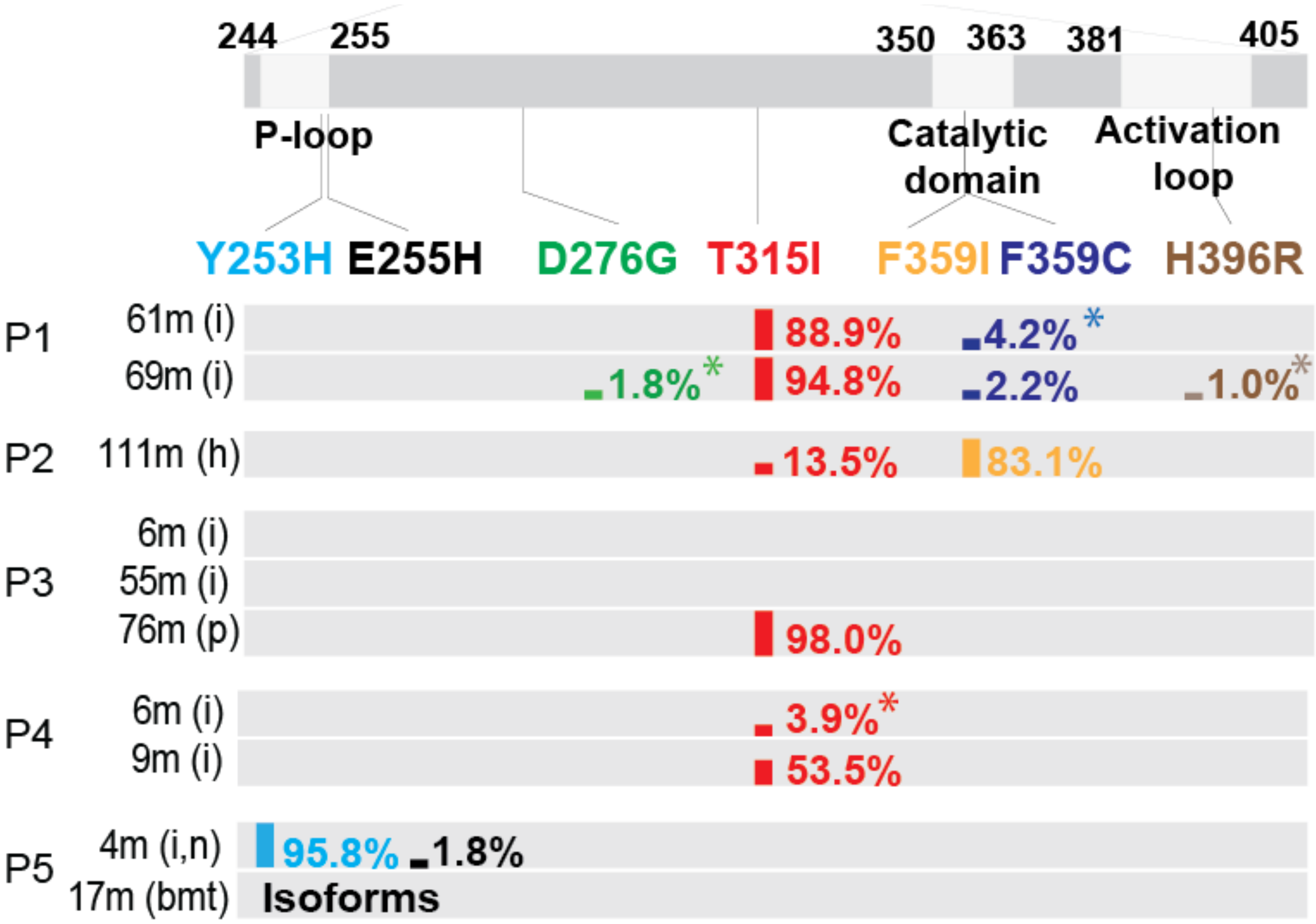
Mutations acquired in fusion transcript.
Might require treatment with alternative drug.

BCR-ABL1 dilution series results

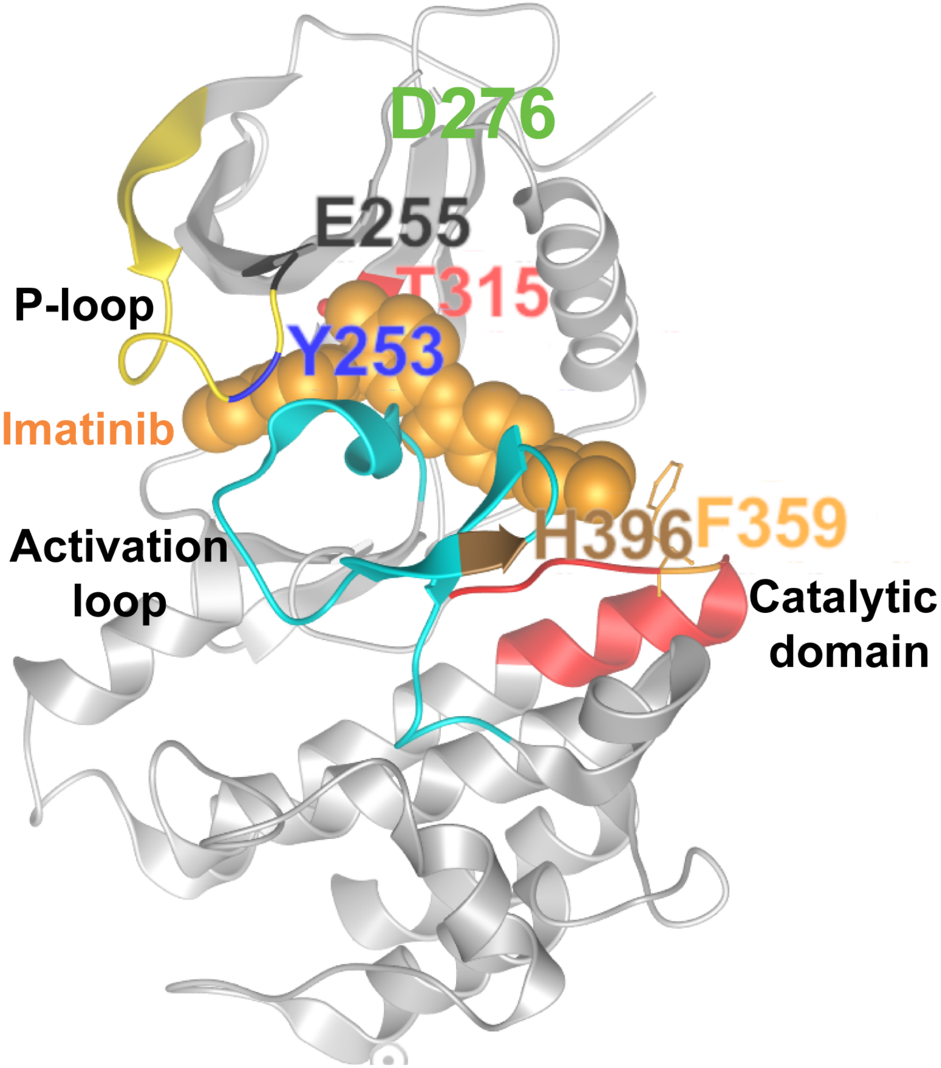
- Mutations down to 1% detected!



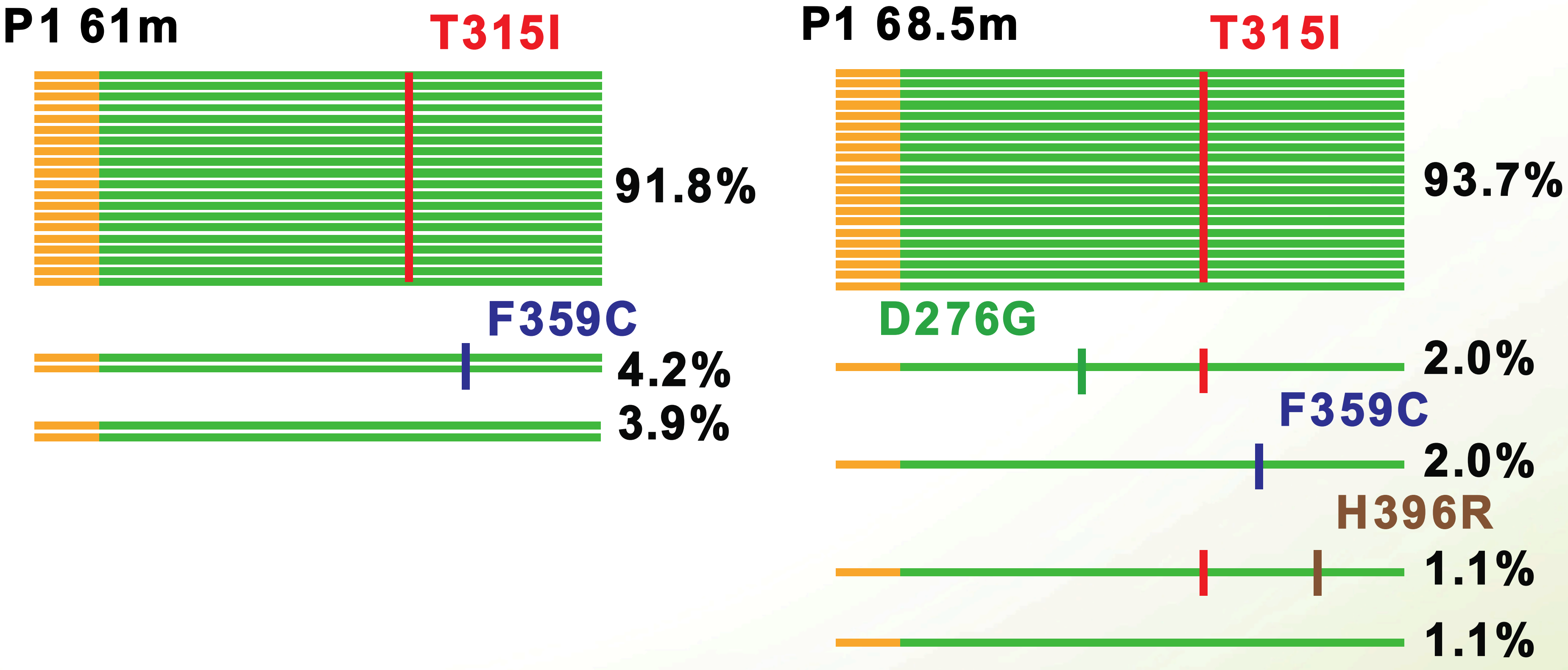
Summary of mutations in 5 CML patients



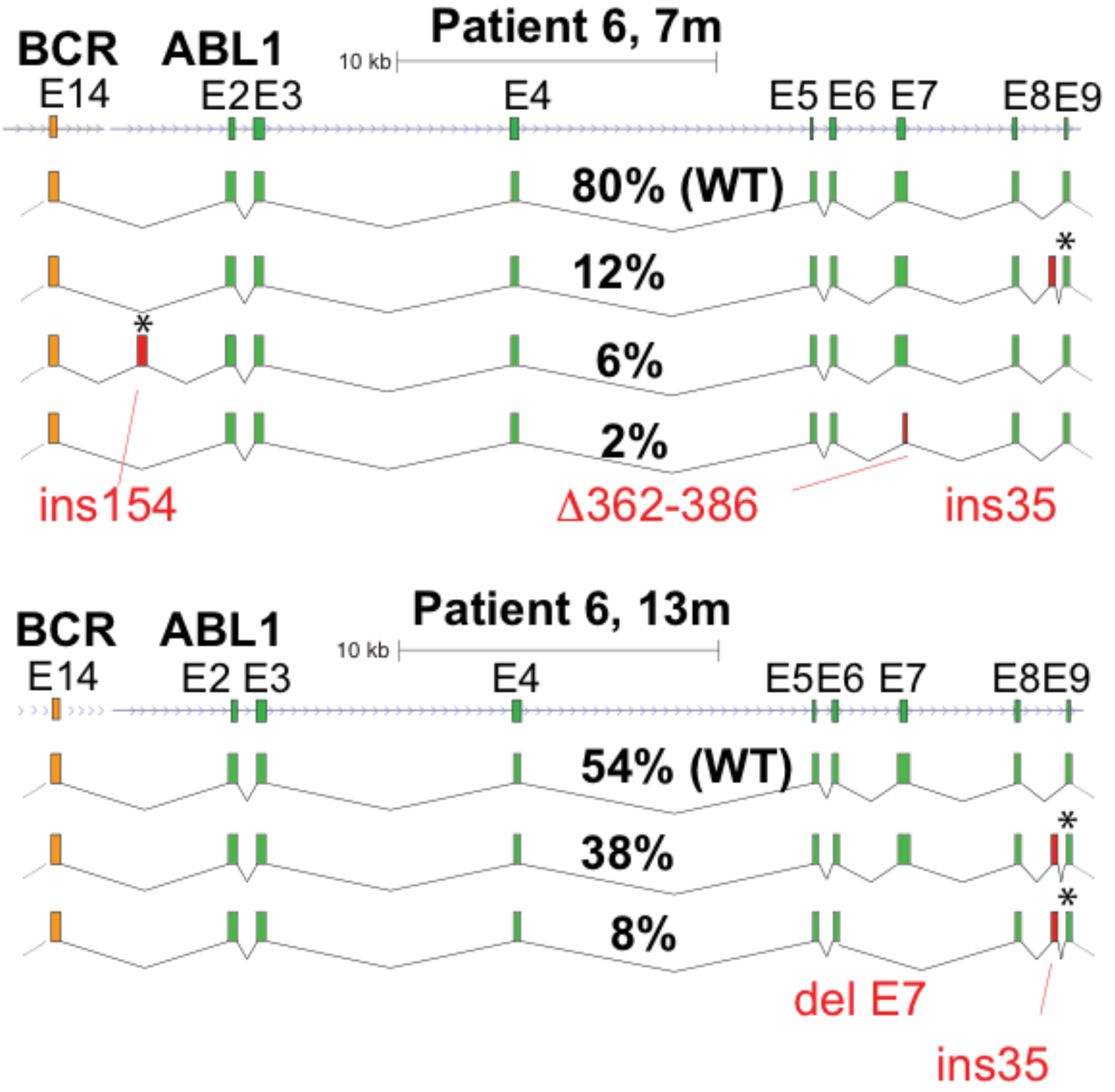
Mutations mapped to protein structure



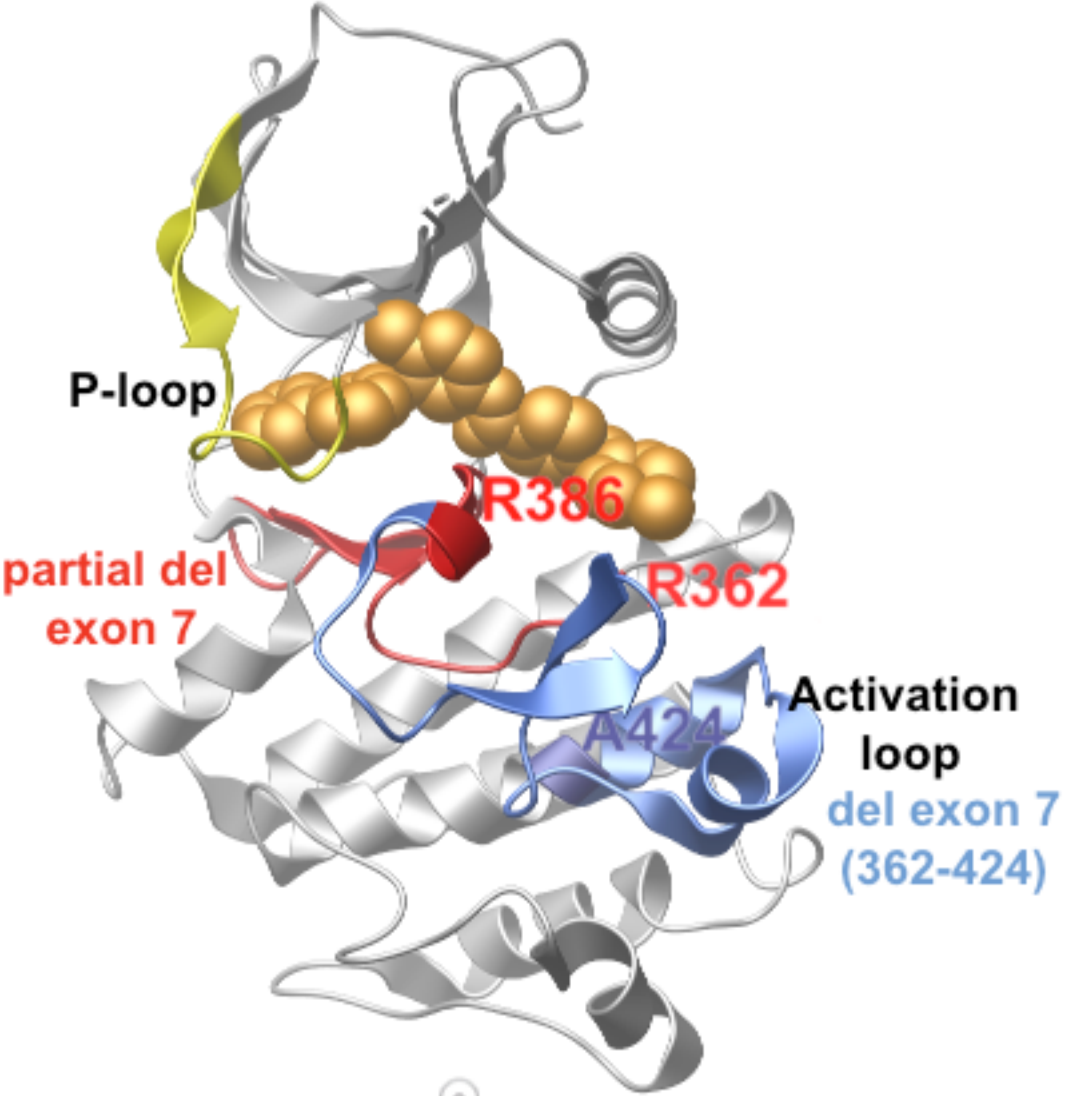
BCR-ABL1 - Compound mutations



BCR-ABL1 - Multiple isoforms in one individual!

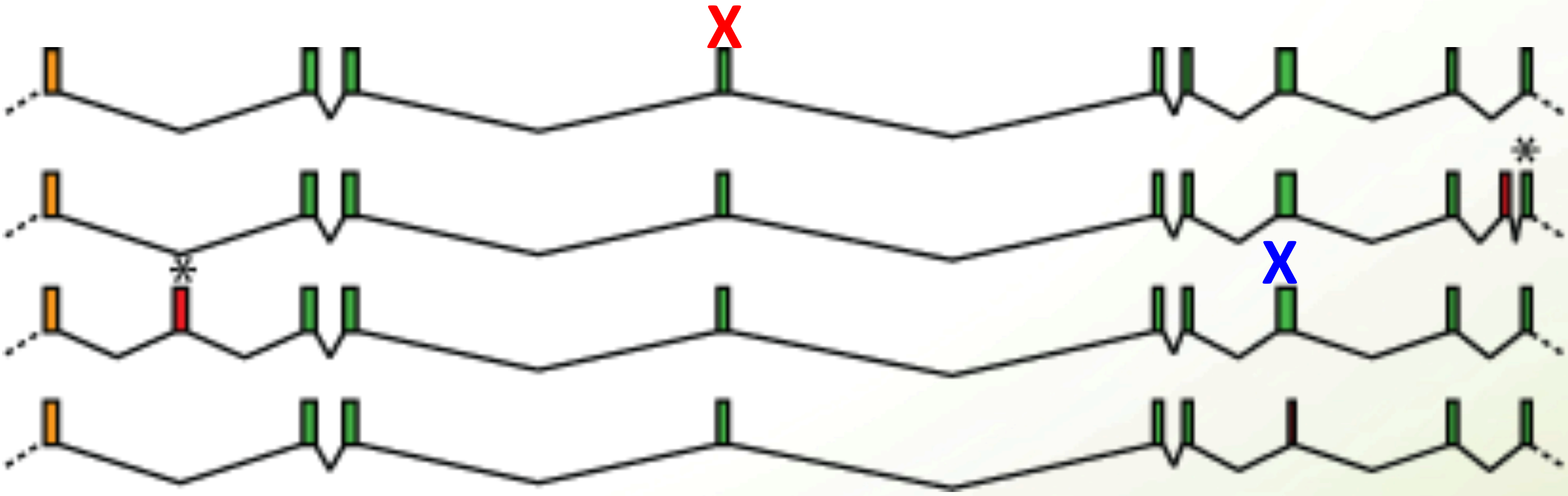


BCR-ABL1 – Isoforms and protein structure



Future bioinformatics challenge

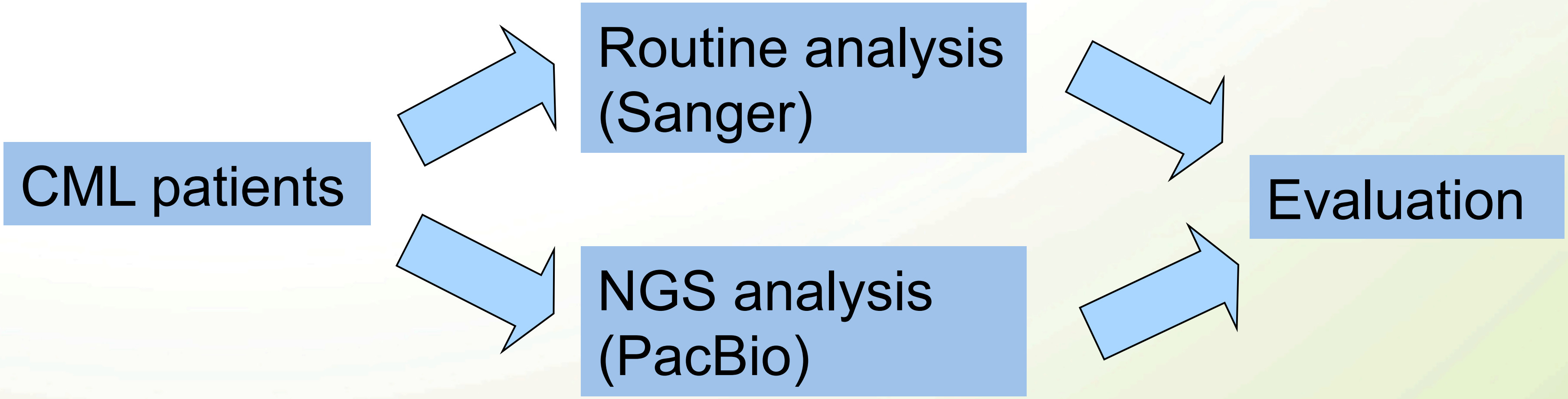
- How to find mutations within isoforms???



Conclusions and next steps

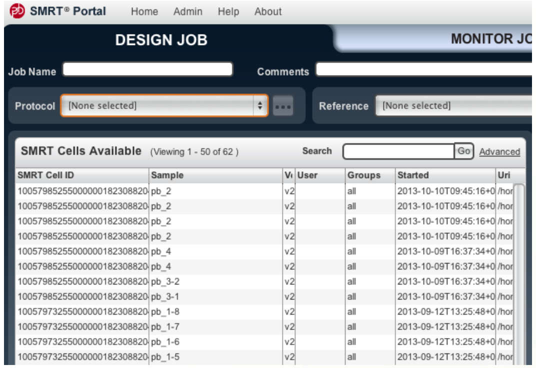
- Sensitive method for *BCR-ABL1* analysis!
  - Also for compound mutations and isoforms

- Method now used in clinical routine!
  - Patient samples coming to the clinic over a few months
  - Response time limit: 2 weeks



Our clinical diagnostics pipeline!

Step1. Create CCS reads



SMRT® Portal   Home   Admin   Help   About

DESIGN JOB   MONITOR JC

Job Name:

Comments:

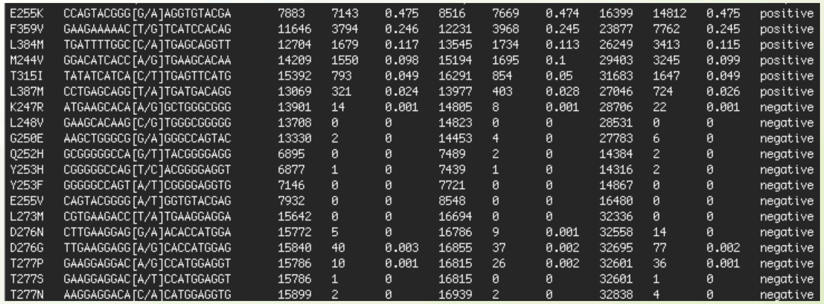
Protocol: [None selected]

Reference: [None selected]

SMRT Cells Available (Viewing 1 - 50 of 62)   Search [Go] Advanced

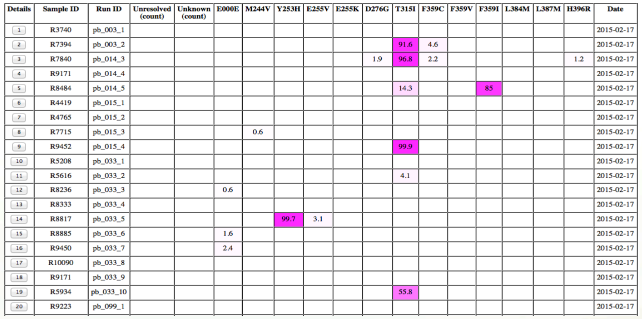
SMRT Cell ID	Sample	V	User	Groups	Started	Uri
10057985255000000182308820	pb_2	v2		all	2013-10-10T09:45:16+0	/hor
10057985255000000182308820	pb_2	v2		all	2013-10-10T09:45:16+0	/hor
10057985255000000182308820	pb_2	v2		all	2013-10-10T09:45:16+0	/hor
10057985255000000182308820	pb_2	v2		all	2013-10-10T09:45:16+0	/hor
10057985255000000182308820	pb_4	v2		all	2013-10-09T16:37:34+0	/hor
10057985255000000182308820	pb_4	v2		all	2013-10-09T16:37:34+0	/hor
10057985255000000182308820	pb_3-2	v2		all	2013-10-09T16:37:34+0	/hor
10057985255000000182308820	pb_3-1	v2		all	2013-10-09T16:37:34+0	/hor
10057973255000000182308820	pb_1-8	v2		all	2013-09-12T13:25:48+0	/hor
10057973255000000182308820	pb_1-7	v2		all	2013-09-12T13:25:48+0	/hor
10057973255000000182308820	pb_1-6	v2		all	2013-09-12T13:25:48+0	/hor
10057973255000000182308820	pb_1-5	v2		all	2013-09-12T13:25:48+0	/hor

Step2. Run mutation analysis



```
E255K   CCAGTACGGG[G/A]AGGTGTACGA    7883    7143   0.475   8516    7669   0.474   16399   14812   0.475   positive
F359V   GAAGAAAAAC[T/G]TCATCCACAG    11646   3794   0.246   12231   3968   0.245   23877   7762    0.245   positive
L384M   TGATTTTGGC[C/A]TGAGCAGGTT    12704   1679   0.117   13545   1734   0.113   26249   3413    0.115   positive
M244V   GGACATCACC[A/G]TGAAGCACAA    14209   1550   0.098   15194   1695   0.1     29403   3245    0.099   positive
T315I   TATATCATCA[C/T]TGAGTTCATG    15392   793    0.049   16291   854    0.05    31683   1647    0.049   positive
L387M   CCTGAGCAGG[T/A]TGATGACAGG    13069   321    0.024   13977   403    0.028   27046   724     0.026   positive
K247R   ATGAAGCACA[A/G]GCTGGGCGGG    13901   14     0.001   14805   8      0.001   28706   22      0.001   negative
L248V   GAAGCACAAG[C/G]TGGGCGGGGG    13708   0      0       14823   0      0       28531   0       0       negative
G250E   AAGCTGGGCG[G/A]GGGCCAGTAC    13330   2      0       14453   4      0       27783   6       0       negative
Q252H   GCGGGGGCCA[G/T]TACGGGGAGG    6895    0      0       7489    2      0       14384   2       0       negative
Y253H   CGGGGGCCAG[T/C]ACGGGGAGGT    6877    1      0       7439    1      0       14316   2       0       negative
Y253F   GGGGGCCAGT[A/T]CGGGGAGGTG    7146    0      0       7721    0      0       14867   0       0       negative
E255V   CAGTACGGGG[A/T]GGTGTACGAG    7932    0      0       8548    0      0       16480   0       0       negative
L273M   CGTGAAGACC[T/A]TGAAGGAGGA    15642   0      0       16694   0      0       32336   0       0       negative
D276N   CTTGAAGGAG[G/A]ACACCATGGA    15772   5      0       16786   9      0.001   32558   14      0       negative
D276G   TTGAAGGAGG[A/G]CACCATGGAG    15840   40     0.003   16855   37     0.002   32695   77      0.002   negative
T277P   GAAGGAGGAC[A/G]CCATGGAGGT    15786   10     0.001   16815   26     0.002   32601   36      0.001   negative
T277S   GAAGGAGGAC[A/T]CCATGGAGGT    15786   1      0       16815   0      0       32601   1       0       negative
T277N   AAGGAGGACA[C/A]CATGGAGGTG    15899   2      0       16939   2      0       32838   4       0       negative
```

Step3. Upload to result server



Details	Sample ID	Run ID	Unresolved (count)	Unknown (count)	E000E	M244V	Y253H	E255V	E255K	D276G	T315I	F359C	F359V	F359I	L384M	L387M	H396R	Date
1	R3740	pb_003_1																2015-02-17
2	R7394	pb_003_2									91.6	4.6						2015-02-17
3	R7840	pb_014_3								1.9	96.8	2.2					1.2	2015-02-17
4	R9171	pb_014_4																2015-02-17
5	R8484	pb_014_5									14.3			85				2015-02-17
6	R4419	pb_015_1																2015-02-17
7	R4765	pb_015_2																2015-02-17
8	R7715	pb_015_3				0.6												2015-02-17
9	R9452	pb_015_4									99.9							2015-02-17
10	R5208	pb_033_1																2015-02-17
11	R5616	pb_033_2									4.1							2015-02-17
12	R8236	pb_033_3			0.6													2015-02-17
13	R8333	pb_033_4																2015-02-17
14	R8817	pb_033_5					99.7	3.1										2015-02-17
15	R8885	pb_033_6			1.6													2015-02-17
16	R9450	pb_033_7			2.4													2015-02-17
17	R10090	pb_033_8																2015-02-17
18	R9171	pb_033_9																2015-02-17
19	R5934	pb_033_10									55.8							2015-02-17
20	R9223	pb_099_1																2015-02-17

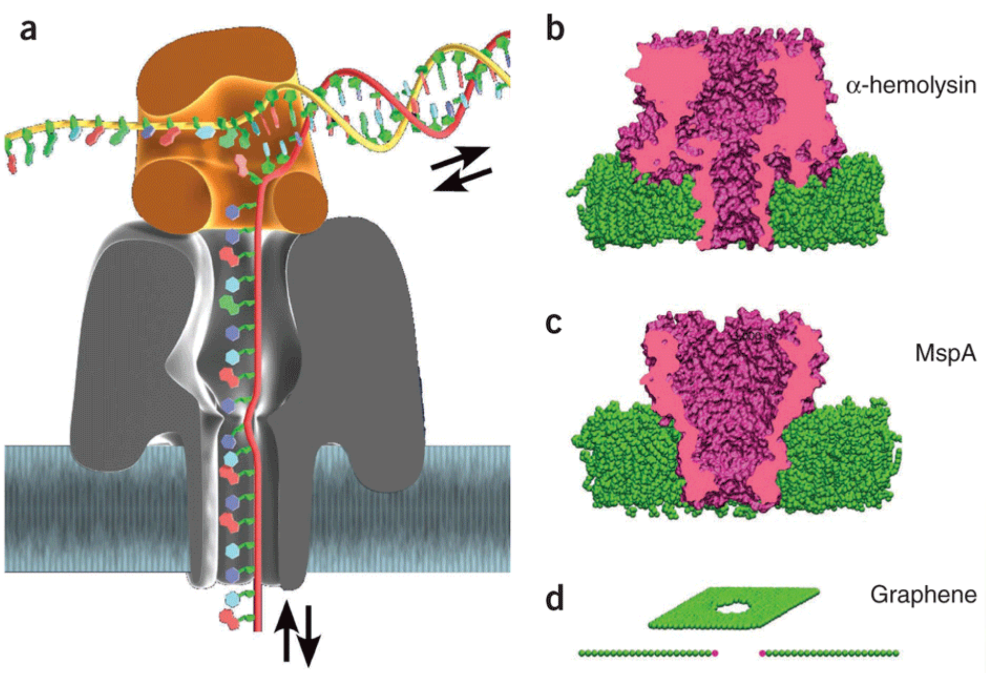
News and future directions (1)

Sequel - New PacBio instrument with higher throughput!

7x more data per SMRT cell!

News and future directions (2)

Nanopore technology - for direct RNA sequencing?



PromethION

Enables detection of modified RNA bases??