



RNA-Seq course in Uppsala

Transcriptome and isoform reconstruction using long reads

Adam Ameur
March 15, 2017



SciLifeLab
RNA-seq workshop 2014

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.

DNA sequencing at all scales



One of the most well-equipped NGS sites in Europe

10 HiSeq Xten, 17 HiSeq 2000/2500, 3 MiSeq, 1 NextSeq, 1 10X Genomics, 1 Ion PGM, 5 Ion Proton, 1 Ion S5XL, 2 PacBio RSII, 1 PacBio Sequel, 2 Sanger ABI3730, 1 BioNano Genomics Irys System, 1 Oxford Nanopore MinIon

Analysis cluster and storage of NGS data

~3 M cph/month on a dedicated cluster

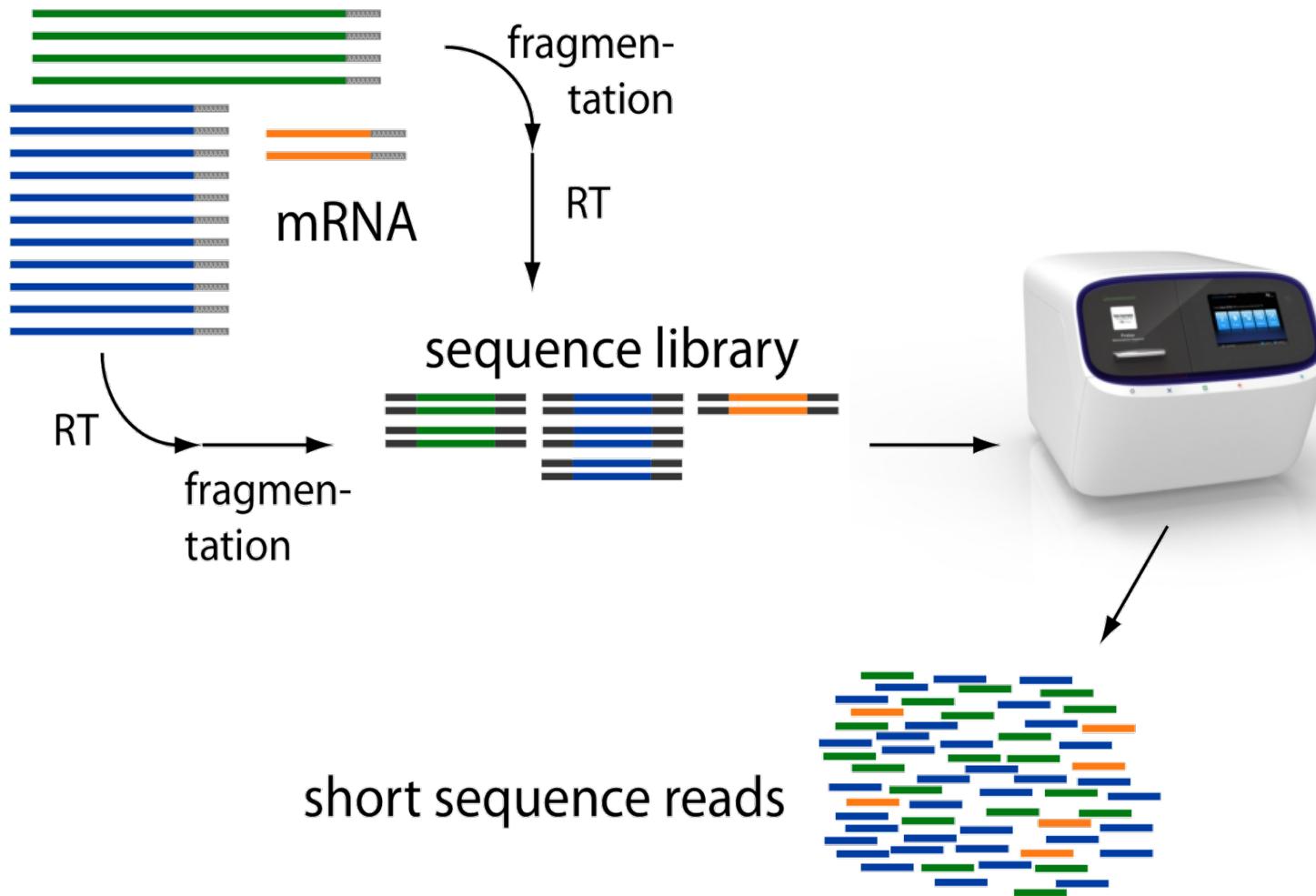
~7 PB storage. Long-term storage in archive



RNA-sequencing

with short reads

RNA-seq standard procedure

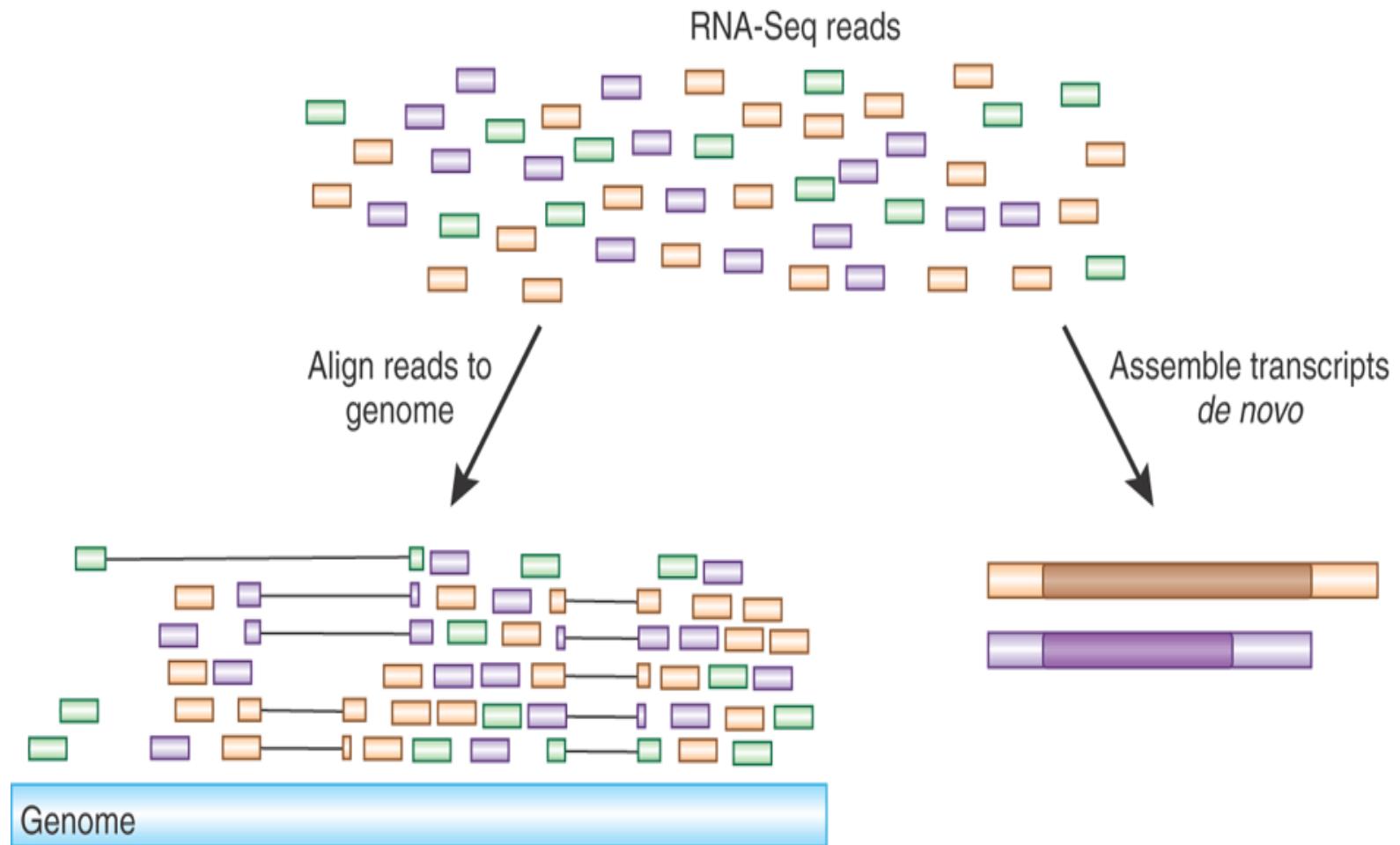


RNA-seq: the main question

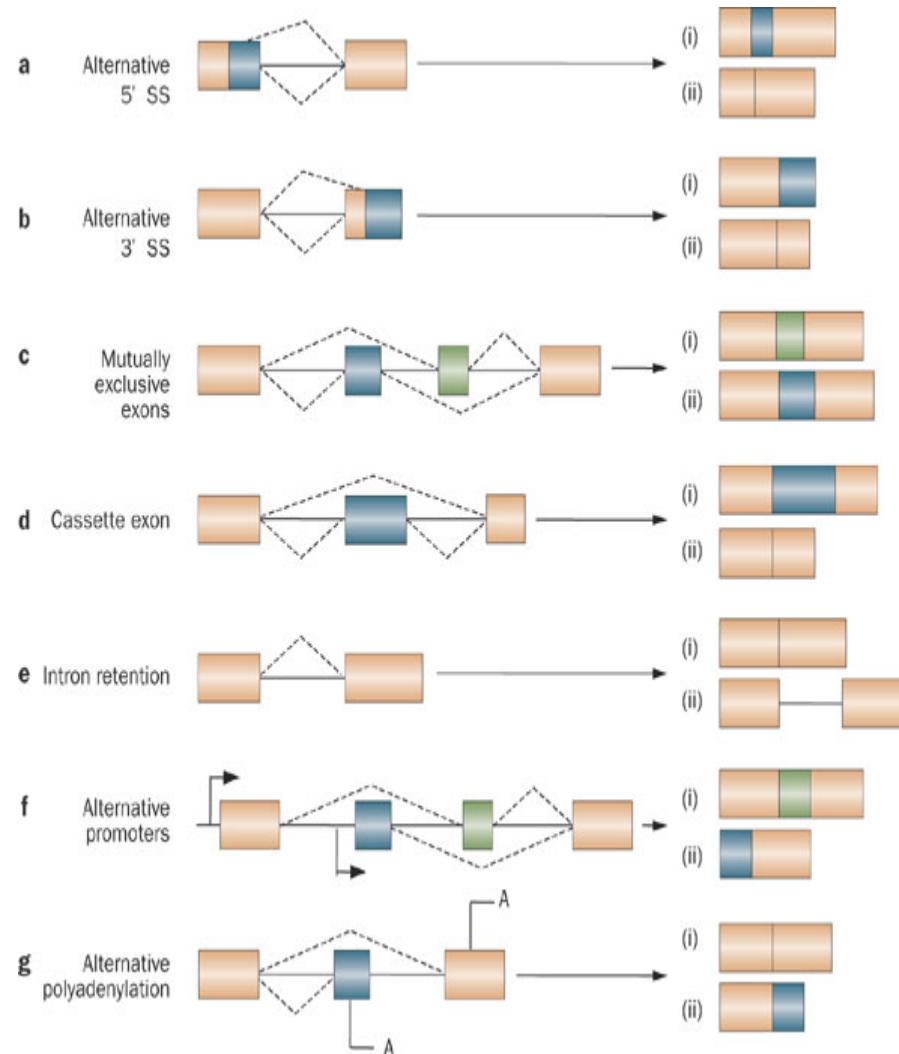
What to do with this?



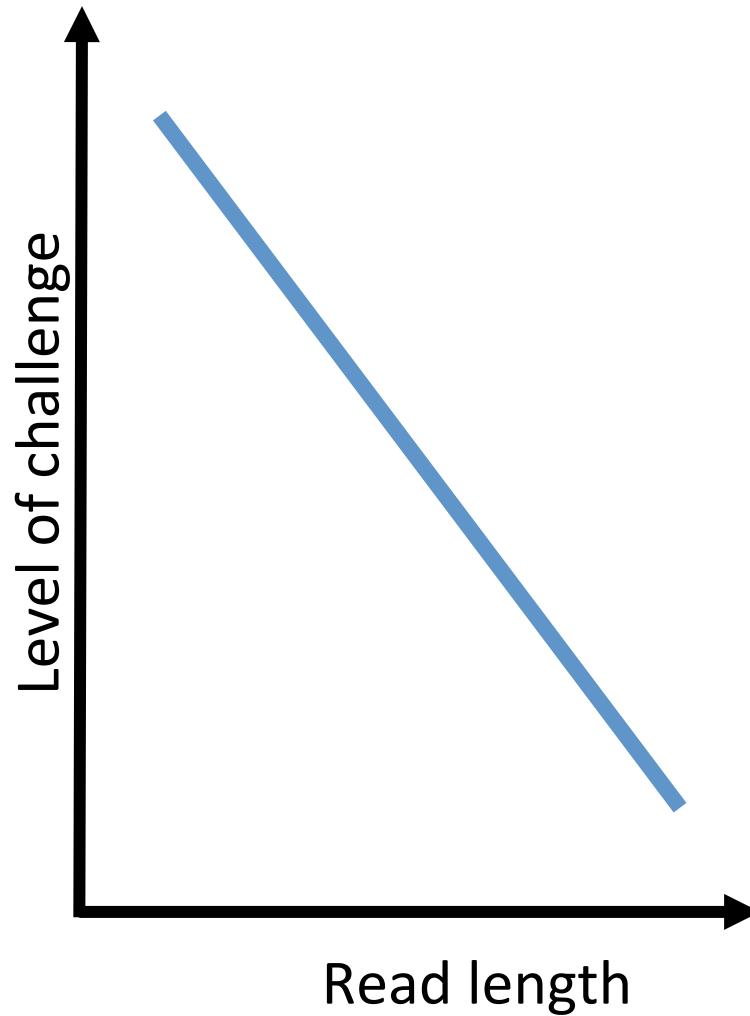
RNA-seq analysis



Complicating factor: alternative splicing



RNA-seq: problem with short reads



RNA-sequencing

with very long reads!!!

PacBio sequencing

- Long-read sequencing instrument
 - Single molecule sequencing
 - Very long read lengths (up to 30 kb or more)
 - Rapid sequencing
 - Can detect base modifications (e.g. methylation)
 - Two systems: RSII and Sequel

PacBio RSII



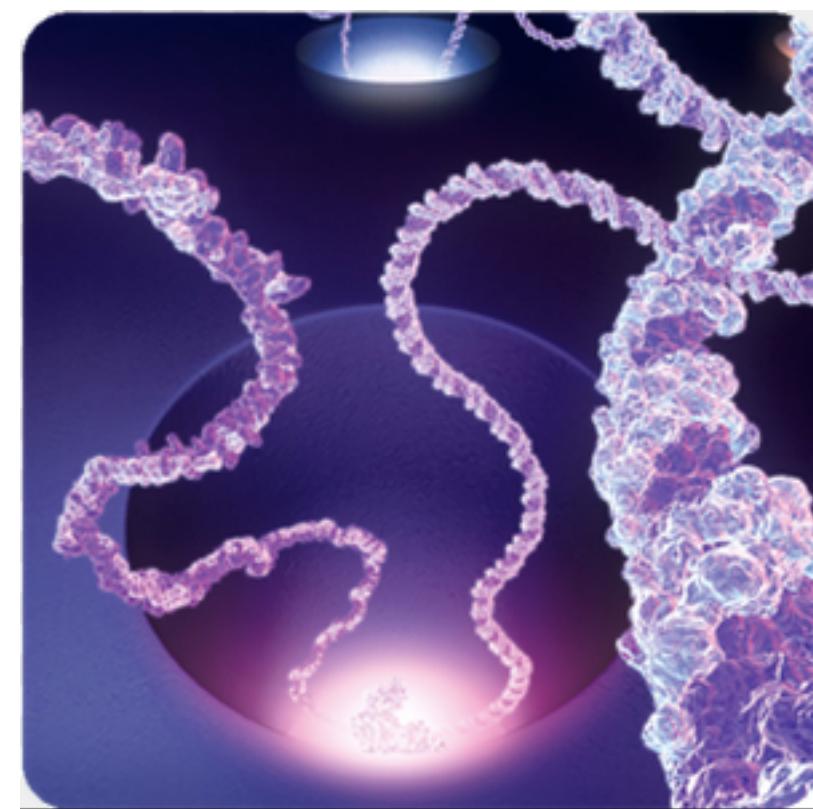
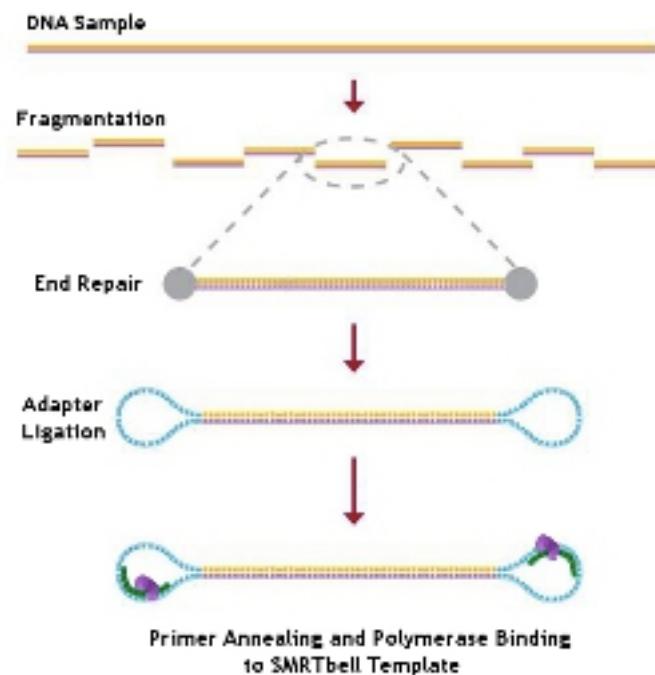
PacBio Sequel



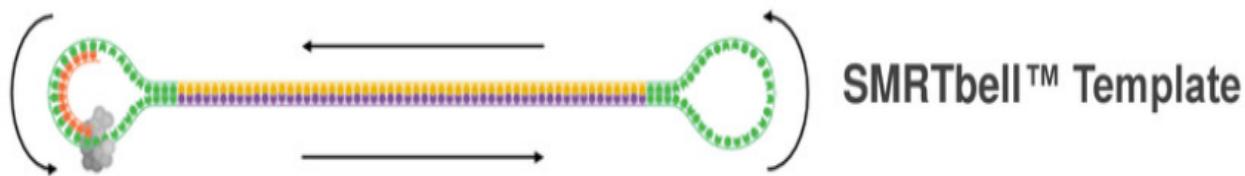
PacBio SMRT - technology



Single Molecule Real Time



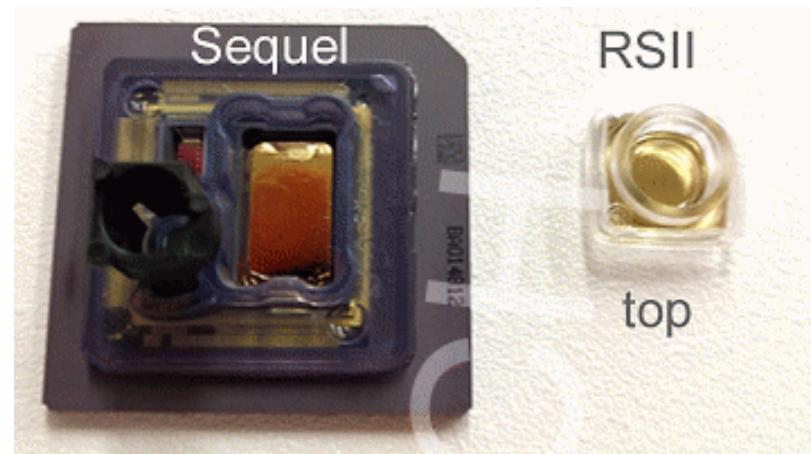
PacBio sequencing template



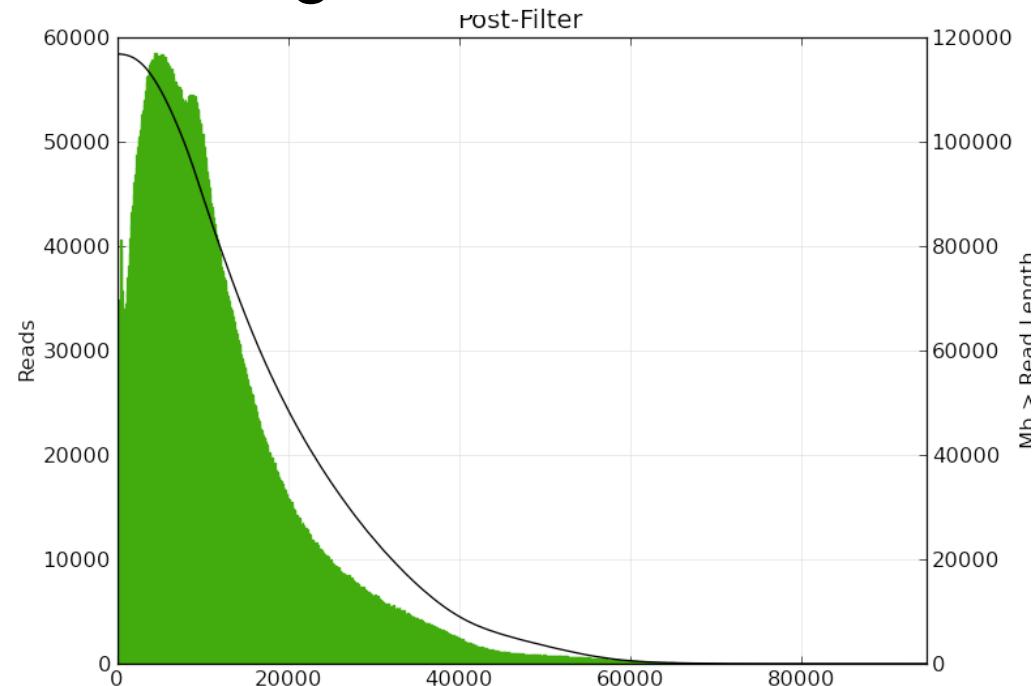
Polymerase Read	Subread	Read (of Insert)
Definition: <ul style="list-style-type: none">Sequence of nucleotides incorporated by polymerase while reading a templateIncludes adaptersOften called “read”Includes adapters1 molecule, 1 pol. read Uses: <ul style="list-style-type: none">QC of instrument runBenchmarking	Definition: <ul style="list-style-type: none">Single pass of templateAdapters removed1 molecule, ≥ 1 subread Unique data: <ul style="list-style-type: none">Kinetic measurementsRich QVs Uses: <ul style="list-style-type: none">Applications	Definition: <ul style="list-style-type: none">Represents highest-quality single-sequence for an insert, regardless of number of passesGeneralizes CCS for < 2 passes & RQ < 0.91 or more passes1 molecule, 1 read Uses: <ul style="list-style-type: none">Library QCApplications

PacBio throughput (spring 2017)

- RSII System
~ 50k reads/SMRT cell
- Sequel System
~ 300k reads/SMRT cell

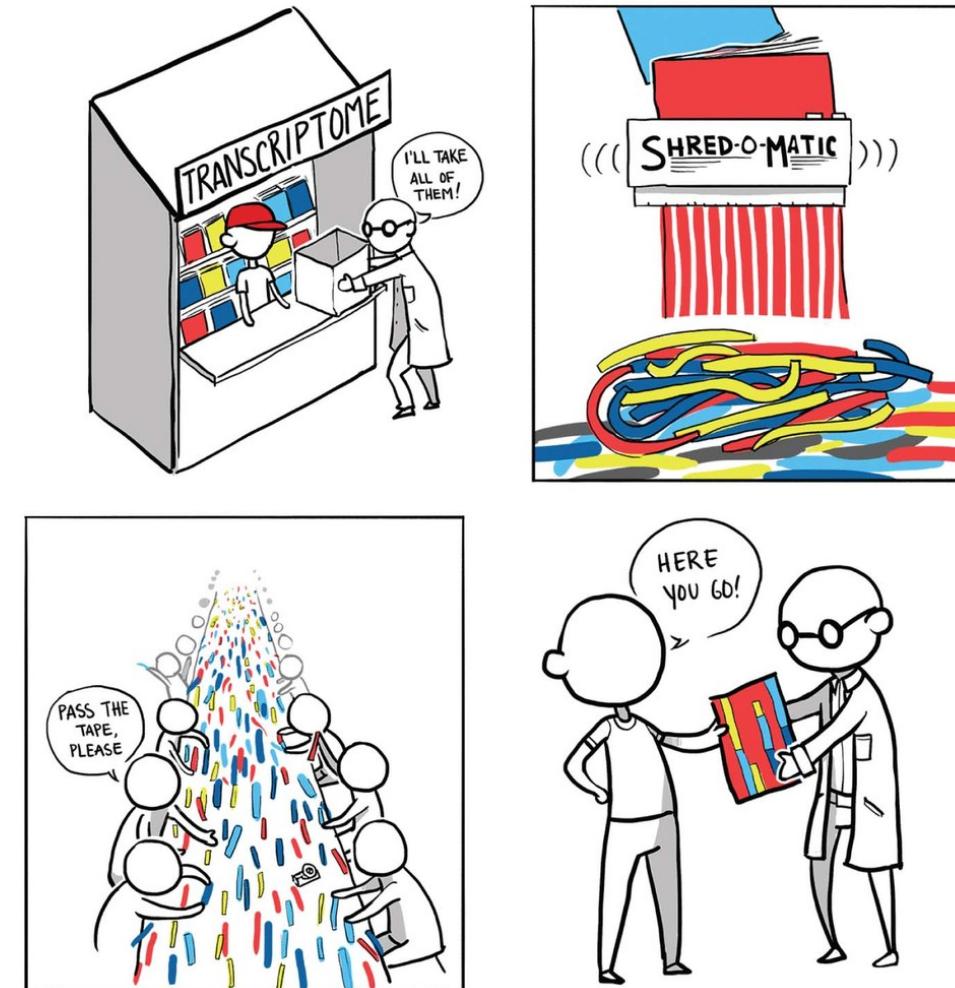


Polymerase read length:



PacBio's Iso-Seq Method

CURRENT STATE OF TRANSCRIPT ASSEMBLY



Abigail Yu

Figure 1 | Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder.

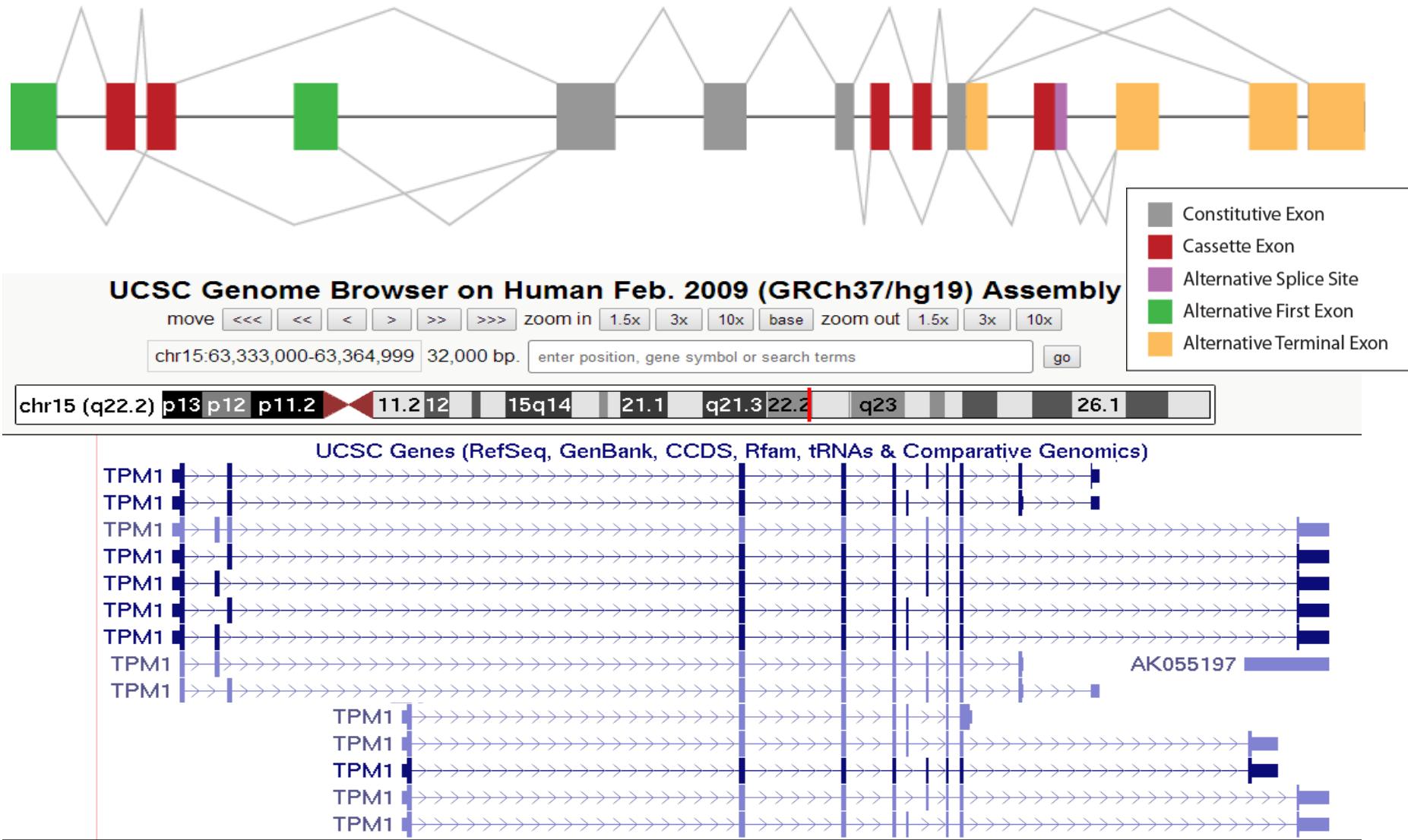
Ian Korf (2013) Genomics: the state of the art in RNA-seq analysis. *Nature Methods* 10: 1165–1166

"The way we do RNA-seq now is... you take the transcriptome, you **blow it up into pieces** and then you try to figure out **how they all go back together again...** If you think about it, it's kind of a **crazy way to do things.**"

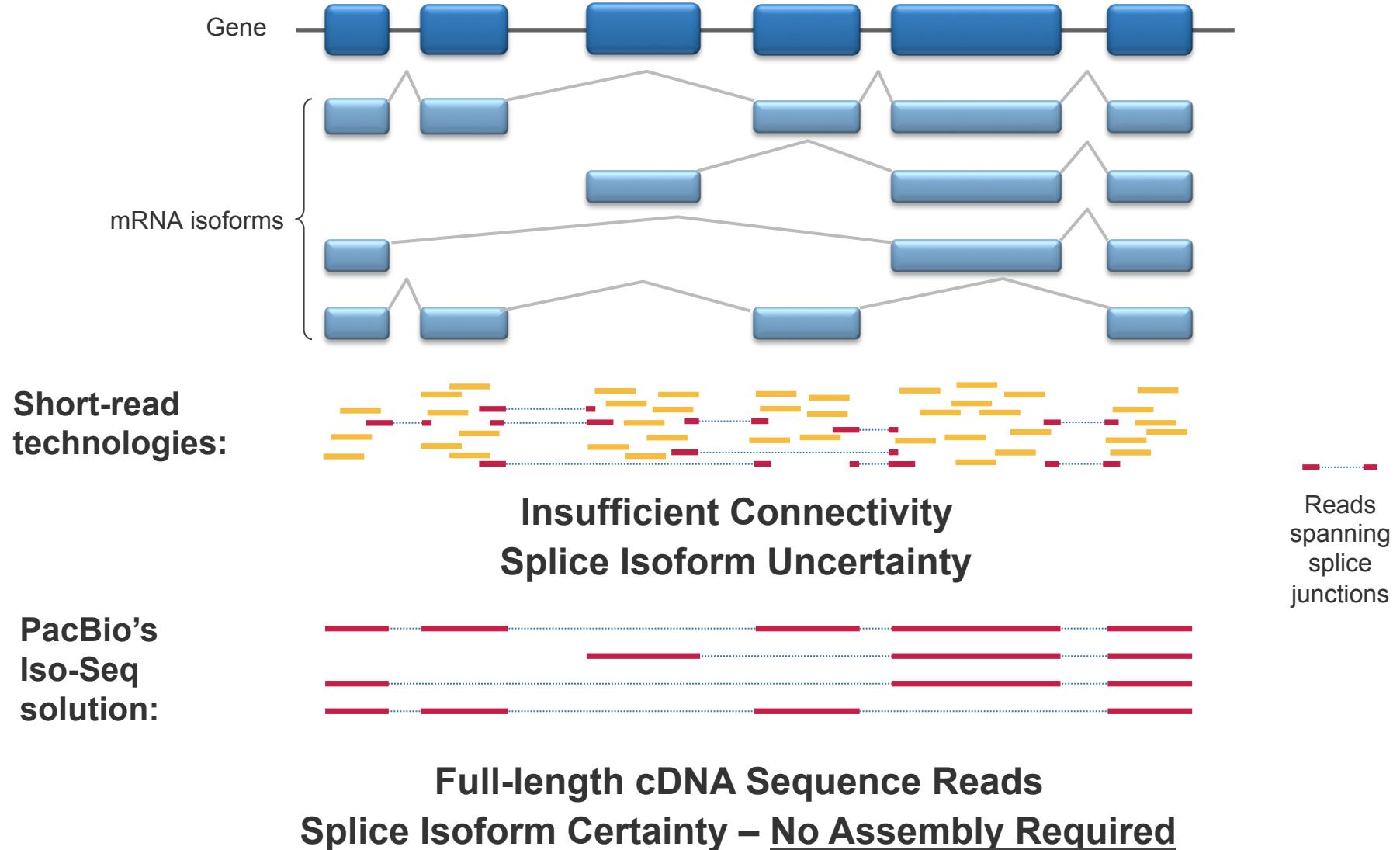
Michael Snyder
Stanford University

Tal Nawy (2013) End-to-end RNA sequencing,
Nature Methods 10: 1144–1145

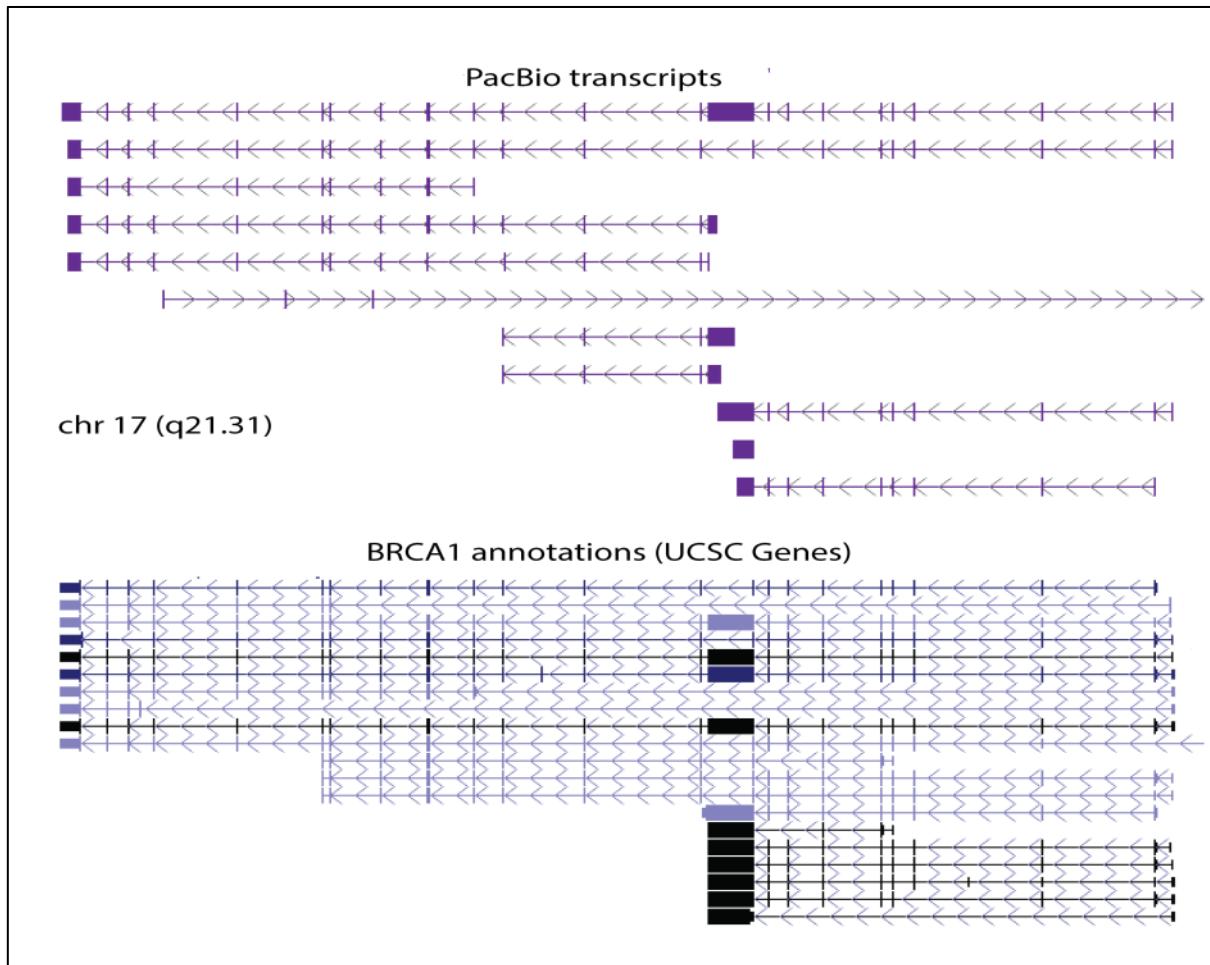
TRANSCRIPT DIVERSITY



DETERMINATION OF TRANSCRIPT ISOFORMS



BRCA1 ISOFORMS IN THE MCF-7 DATA



PacBio transcripts capture multiple isoforms of the BRCA1 gene, several of which are novel



PACBIO®

IsoSeq: Sample Preparation Workflow

RSII versus Sequel

CLONTECH SMARTER™ PCR CDNA SYNTHESIS KIT

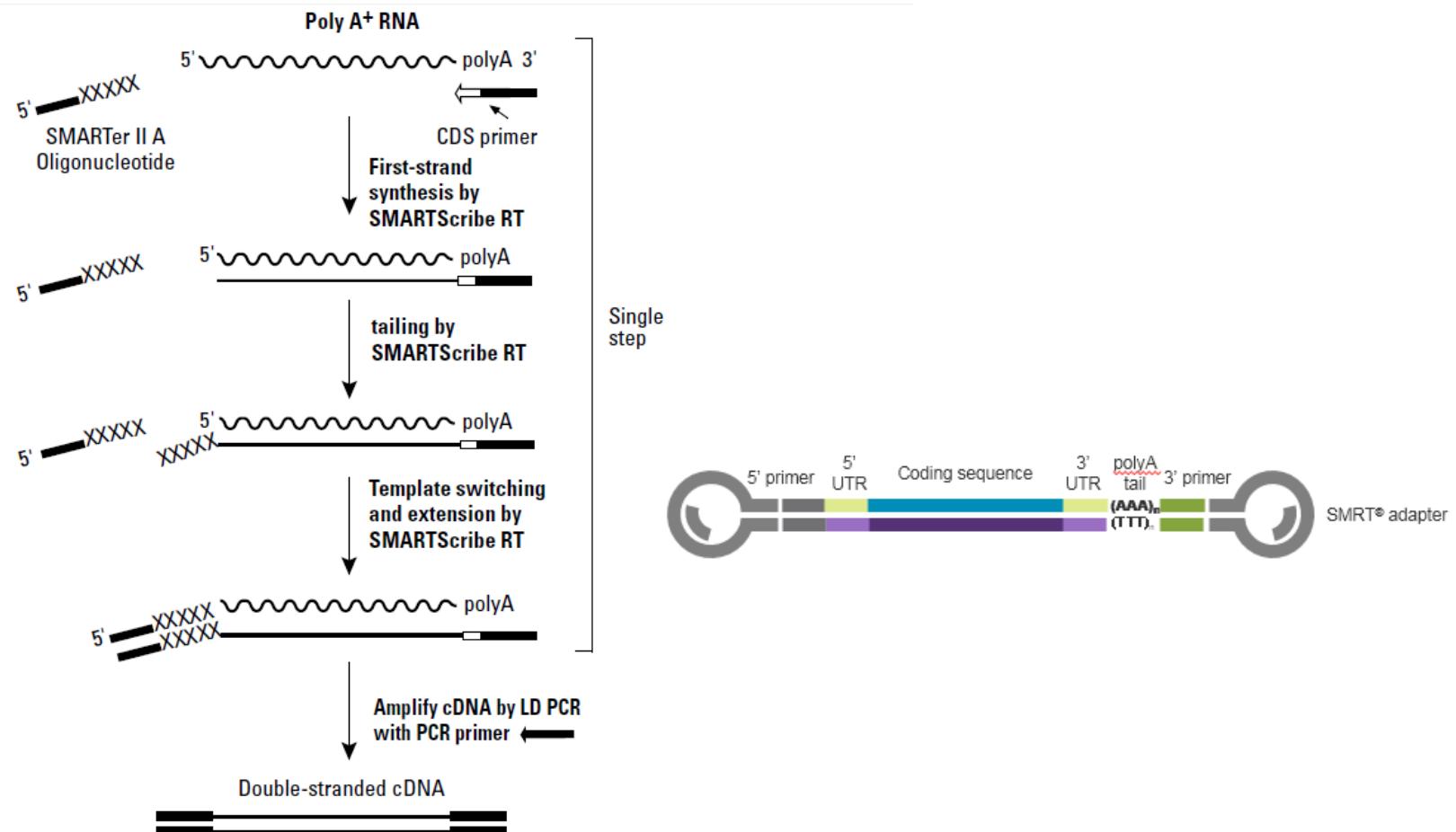
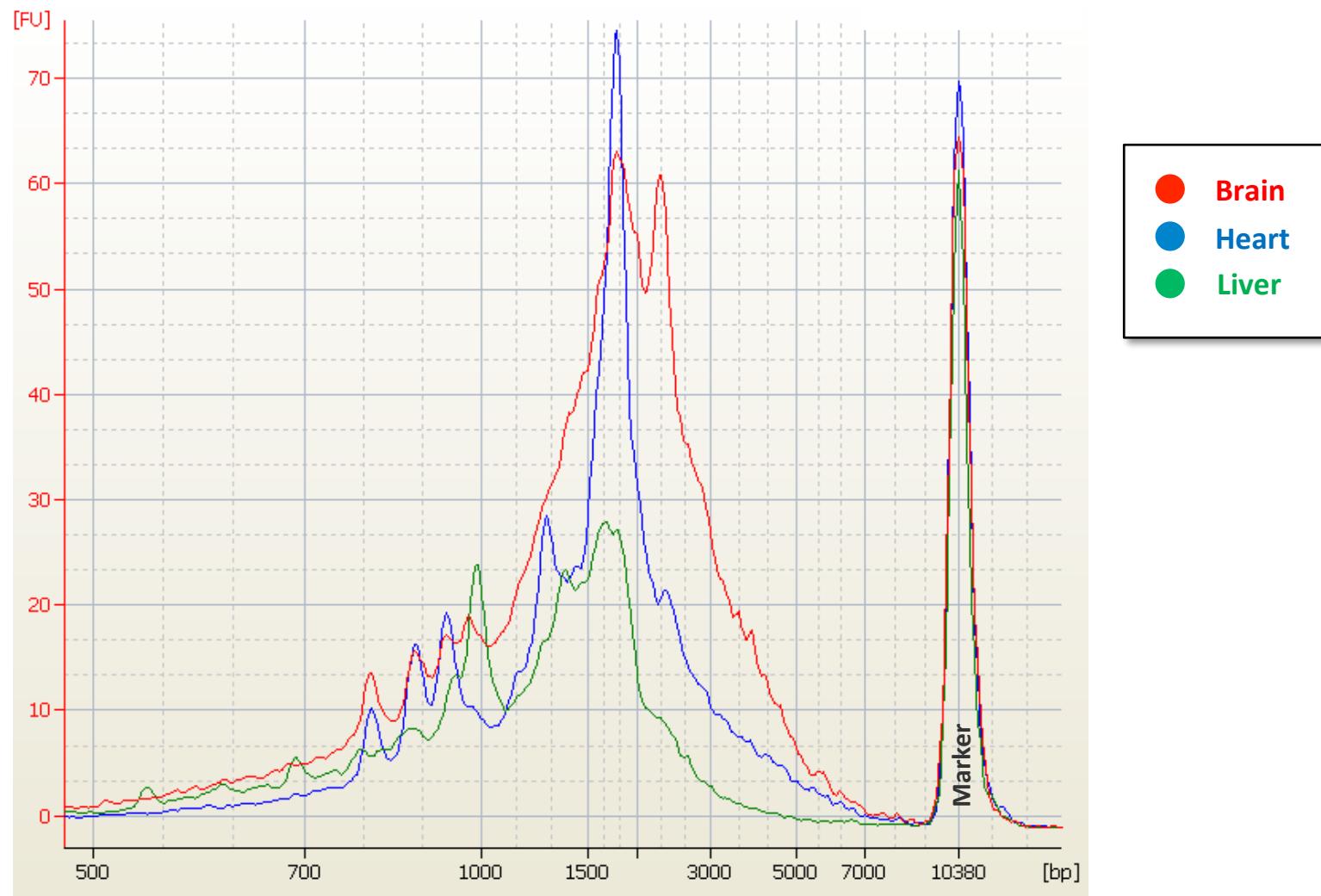


Figure 1. Flowchart of SMARTer cDNA synthesis. The SMARTer II A Oligonucleotide, 3' SMART CDS Primer II A, and 5' PCR Primer II A all contain a stretch of identical sequence (see Section I for sequence information).

SIZE DISTRIBUTION OF AMPLIFIED cDNA FROM MULTIPLE TISSUES



EXPERIMENTAL DESIGN CONSIDERATIONS

What are the goals of your application?

- Targeted or Full Transcriptome
- Alternative Splicing Analysis
- Gene Annotation

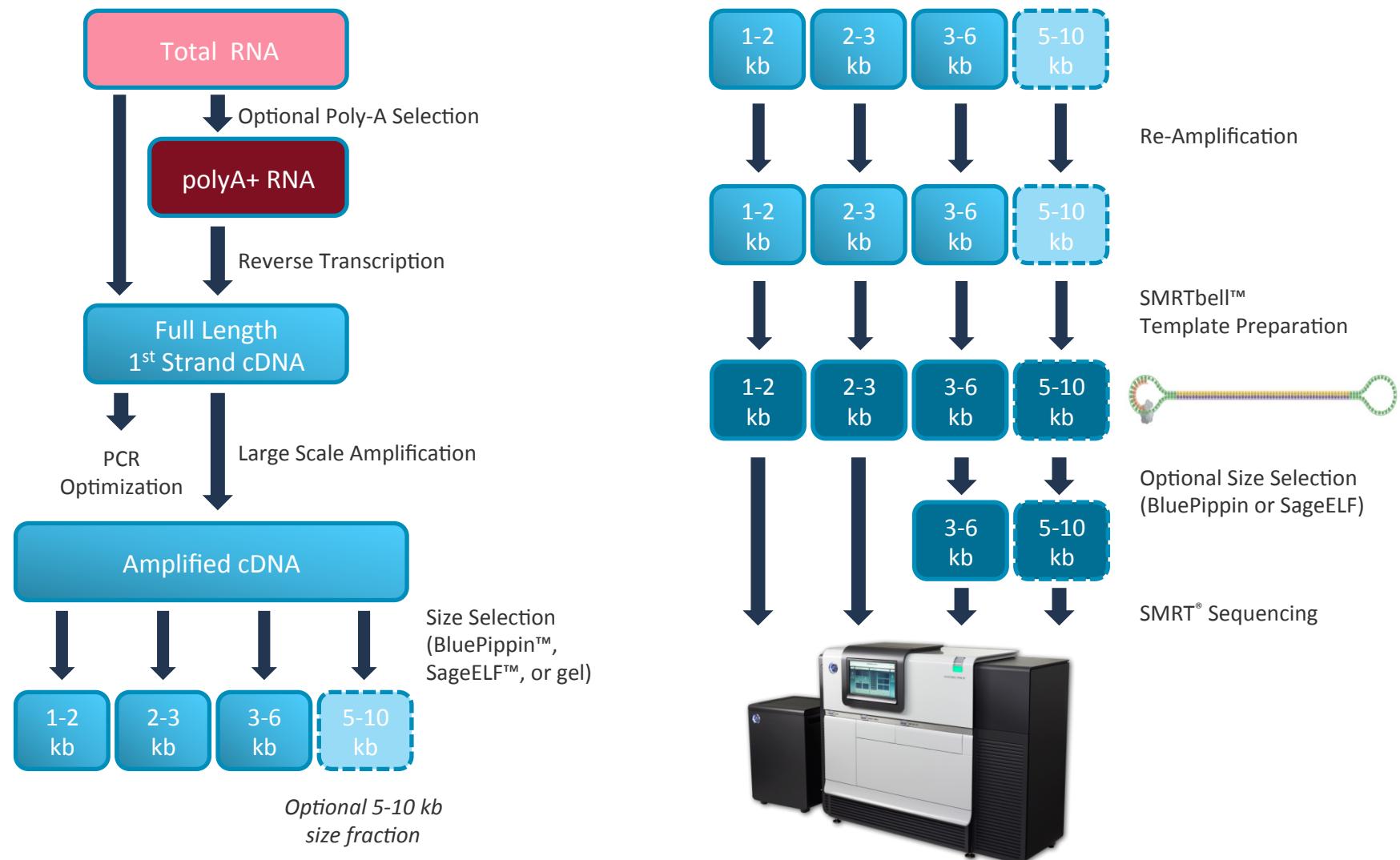
Is Size Selection needed? What size bins are required?

- Size Selection Yes/No
- Size selection via Agarose Gel or Sage BluePippin or SageELF System

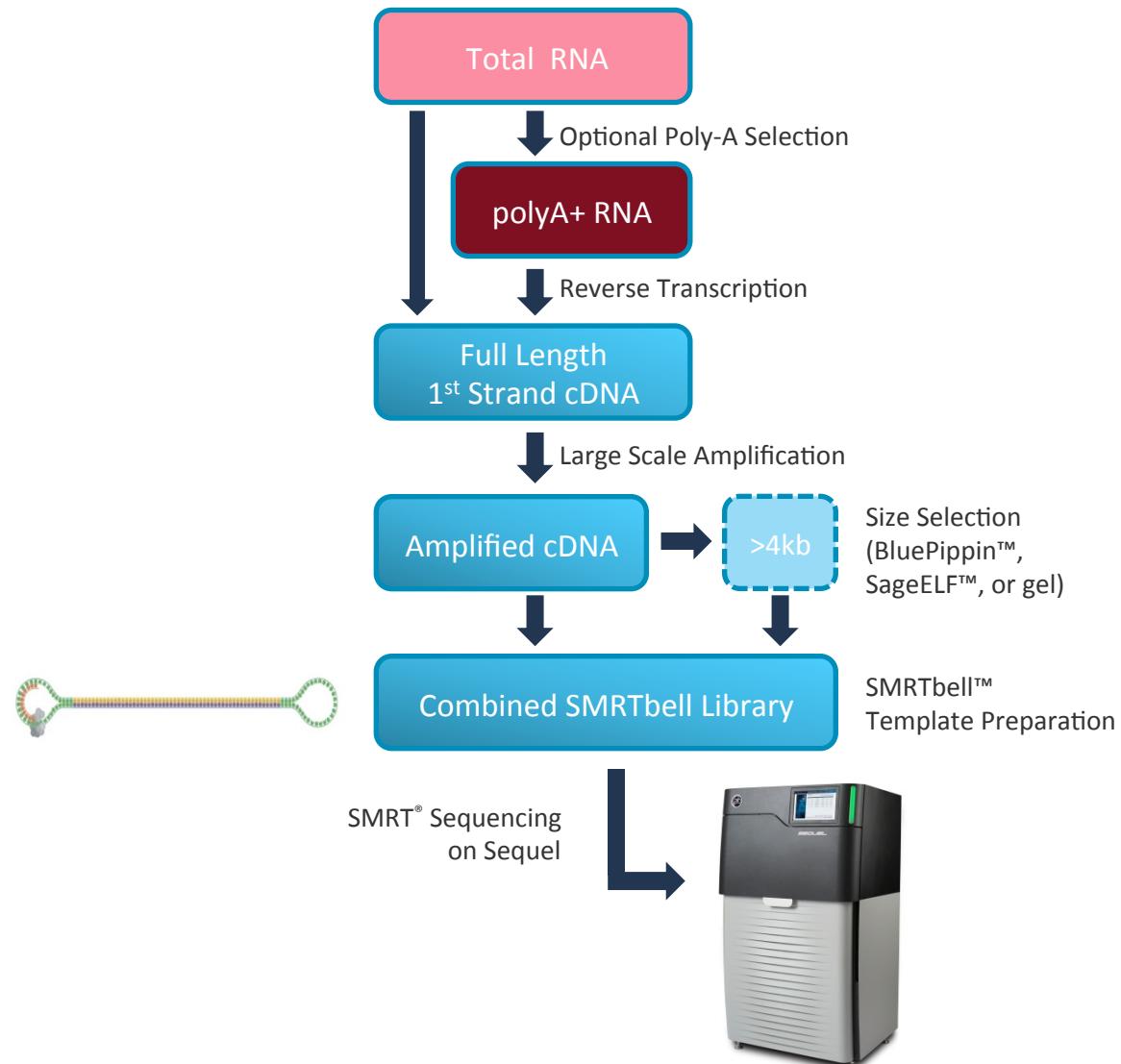
What are the estimated number of Full length transcripts, is this enough to answer my scientific question?

- **RS II:** ~20,000 to 25,000 full-length transcript sequences per SMRT Cell
- **Sequel:** ~100,000 to 150,000 FL transcript sequences per SMRT Cell
- Larger size fractions will have a lower percentage of FL reads

PREPARATION WORKFLOW FOR RSII



ISO-SEQ SAMPLE PREPARATION WORKFLOW FOR SEQUEL



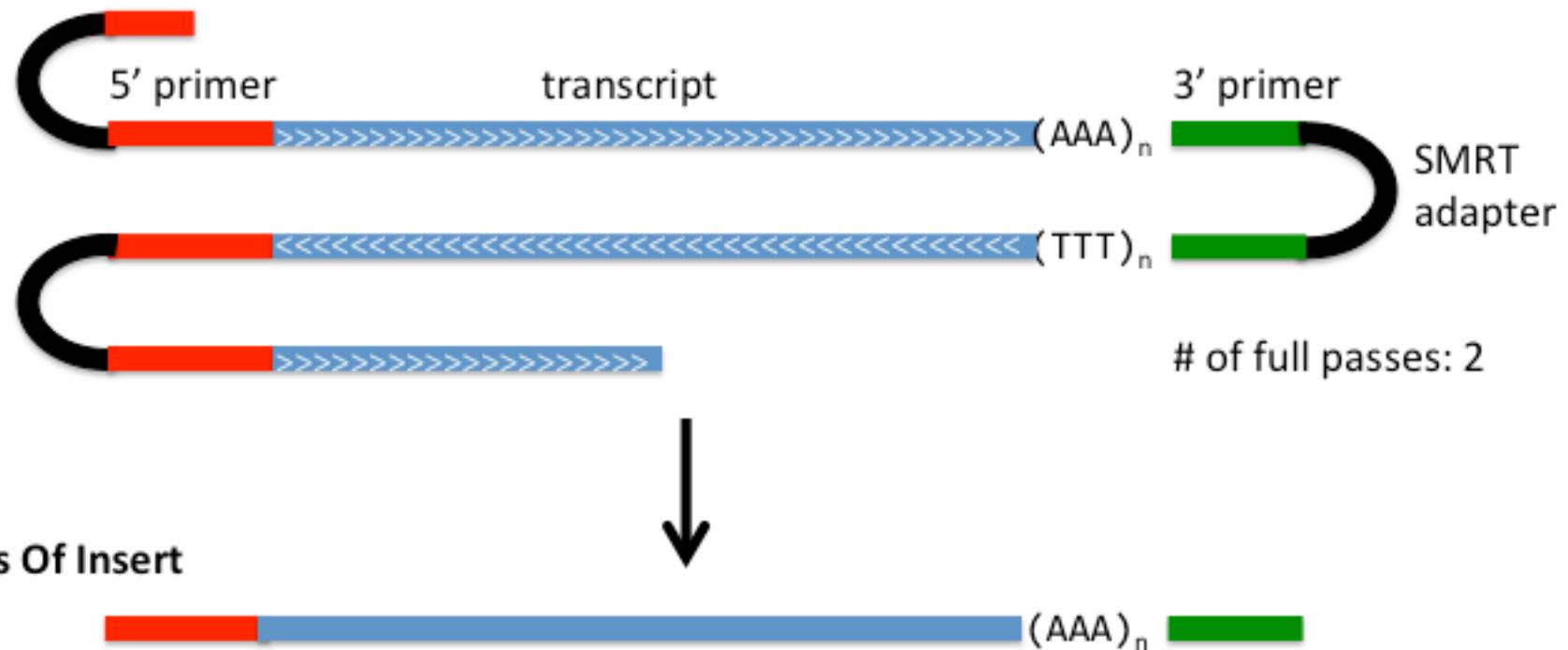
HOW MANY SMRT CELLS?

RS II SMRT Cells (per sample)	Sequel SMRT Cells (per sample)	Experimental Goals
1	<1	Targeted, gene-specific isoform characterization
1-8	1	General survey of full-length isoforms in a transcriptome (moderate to high expression levels) with or without size selection
12-16	1-2	A comprehensive survey of full-length isoforms in the transcriptome across 3-4 size fractions
>16	2+	Deep sequencing for comprehensive isoform discovery and identification of low abundance transcripts across 3-4 size fractions

Iso-seq data analysis

- Simple: Creates Reads of Inserts for FL transcripts!

Raw



Example of a recent Iso-Seq study



ARTICLE

Received 29 Oct 2015 | Accepted 20 Apr 2016 | Published 24 Jun 2016

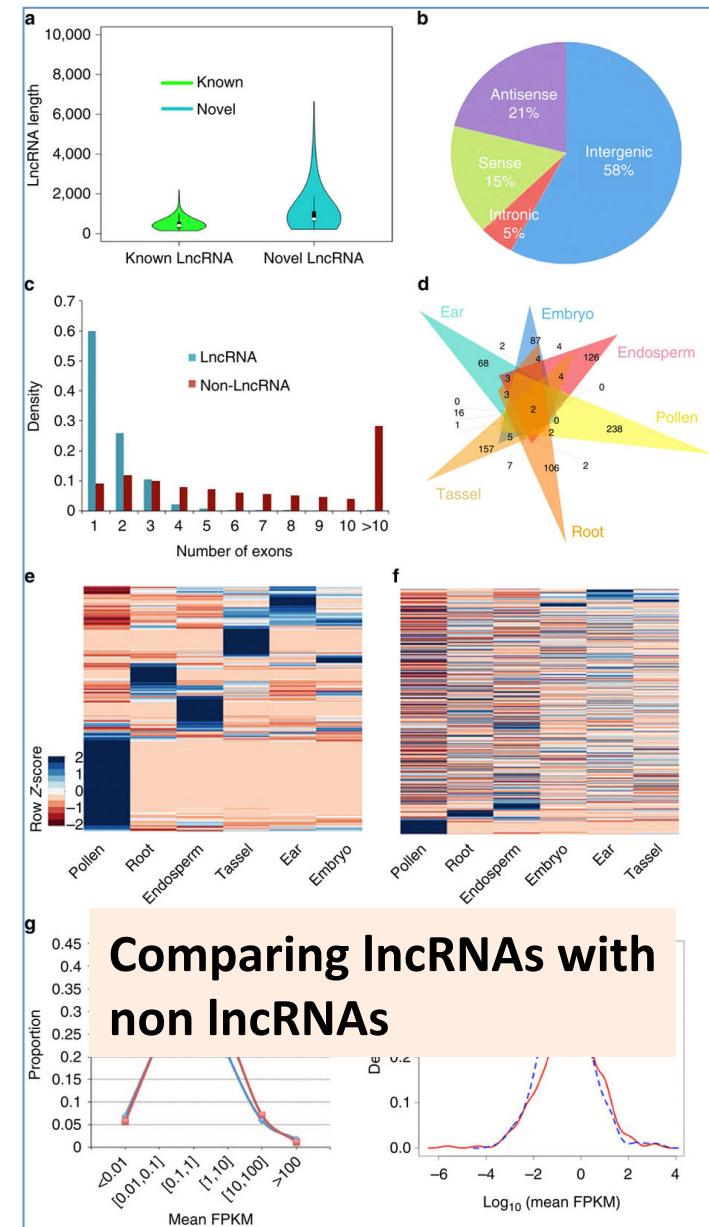
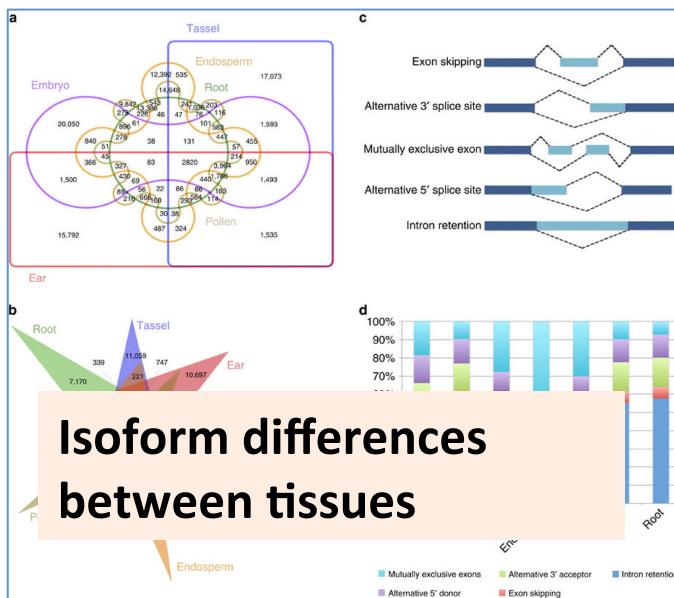
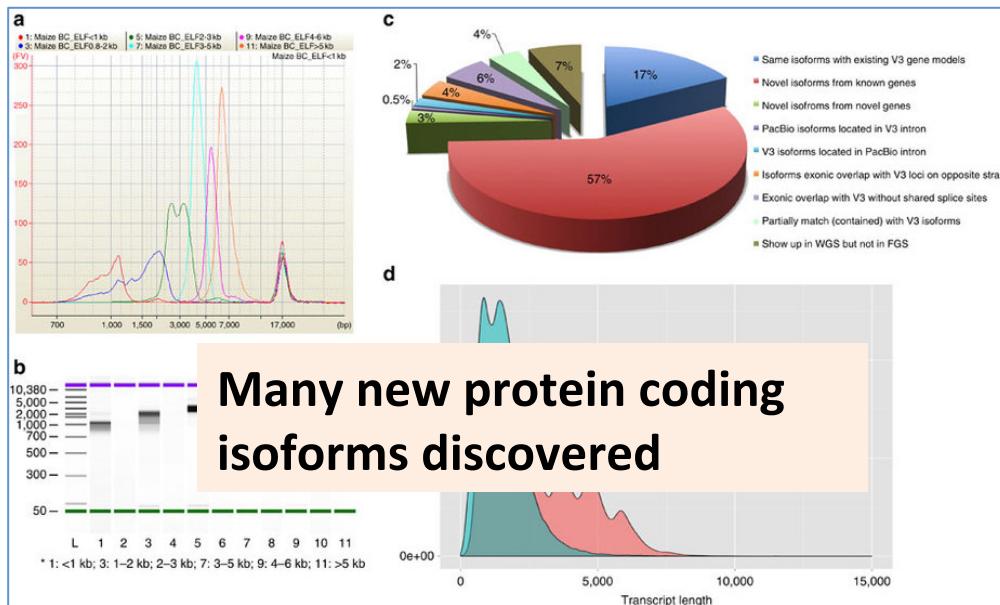
DOI: [10.1038/ncomms11708](https://doi.org/10.1038/ncomms11708)

OPEN

Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing

Bo Wang¹, Elizabeth Tseng², Michael Regulski¹, Tyson A. Clark², Ting Hon², Yinping Jiao¹, Zhenyuan Lu¹, Andrew Olson¹, Joshua C. Stein¹ & Doreen Ware^{1,3}

The complexity of the maize transcriptome

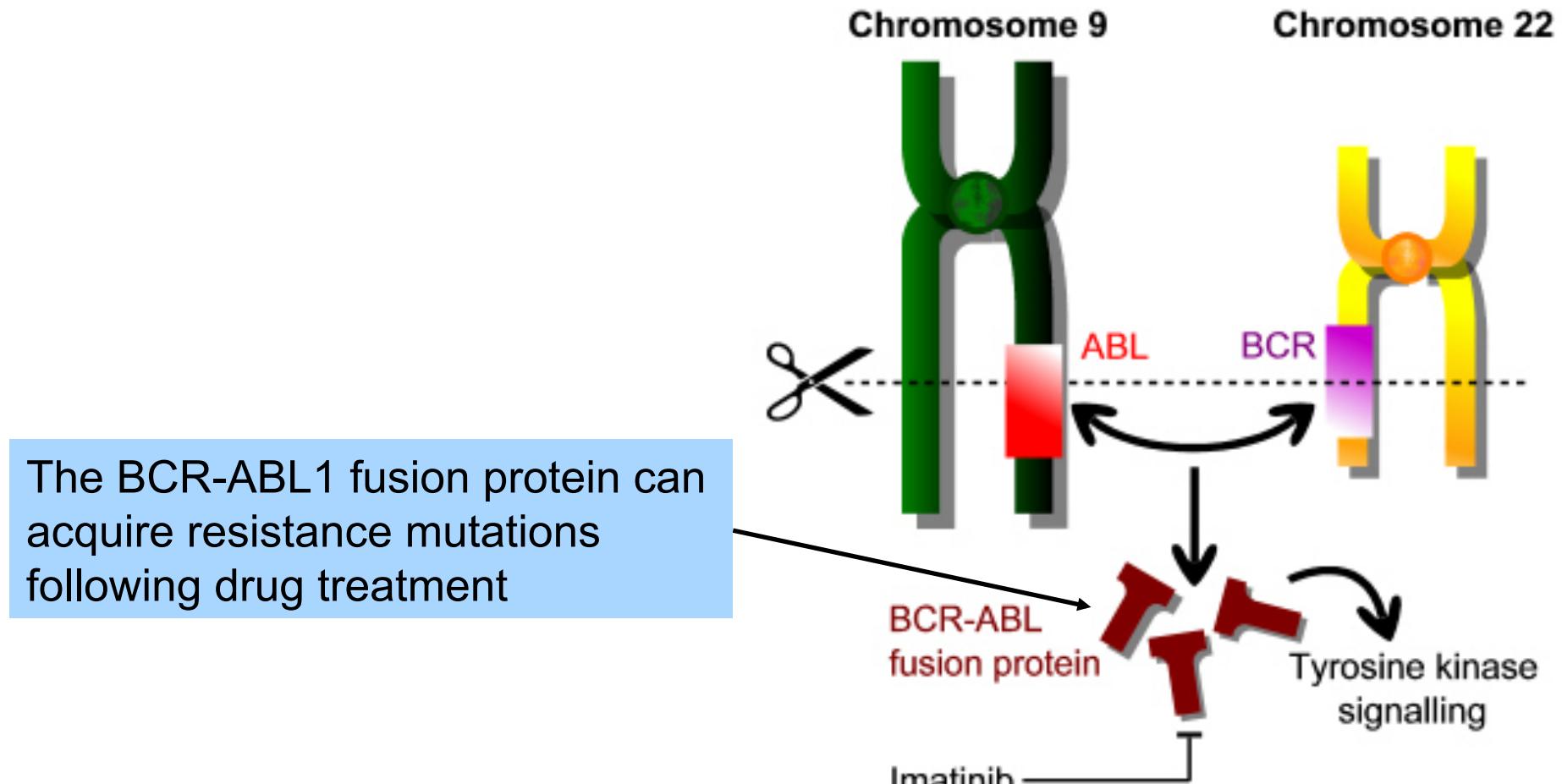


Targeted RNA-sequencing

with very long reads!!!

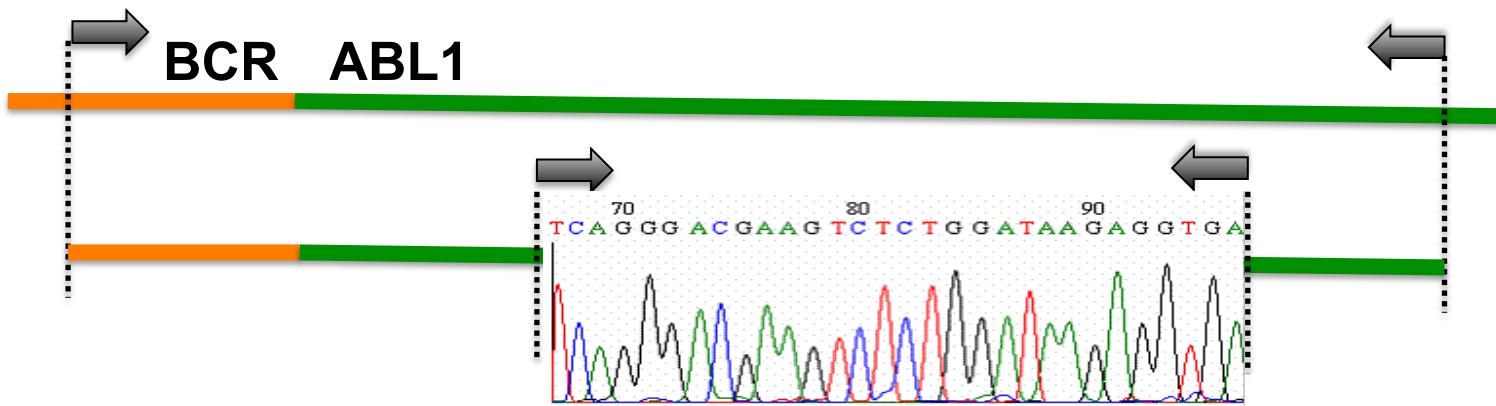
Clinical project: Chronic Myeloid Leukemia

- BCR-ABL1 fusion protein – a CML drug target



Traditional mutation screening in BCR-ABL1

Nested PCR and Sanger sequencing:



Limitations:

- Mutations at frequencies below 10-20% not seen
- Biases may be introduced by nested PCR
- Whole BCR-ABL1 fusion transcript not sequenced
- Clonal composition of mutations not determined

BCR-ABL1 workflow – PacBio Sequencing

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



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 Häggqvist¹, Ida Höijer¹, Nicola Cahill¹, Ulla Olsson-Strömberg² and Monica Hermanson¹

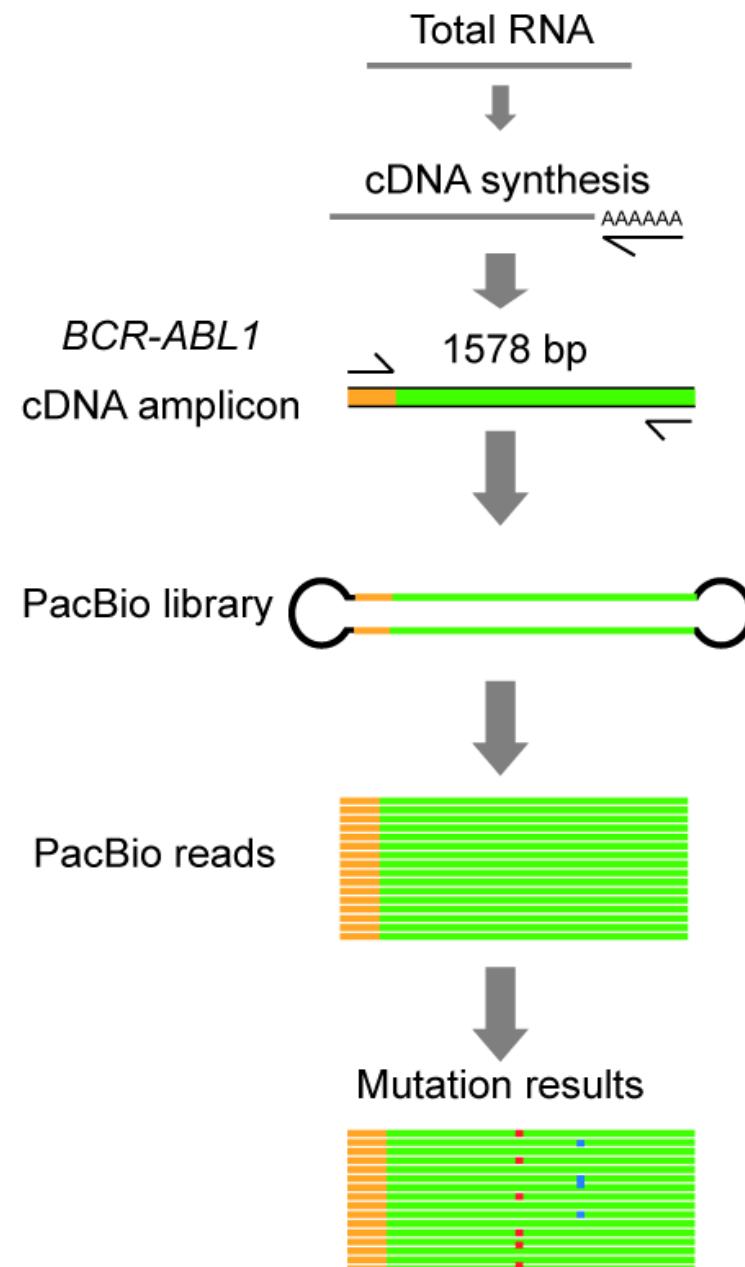
Abstract

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.

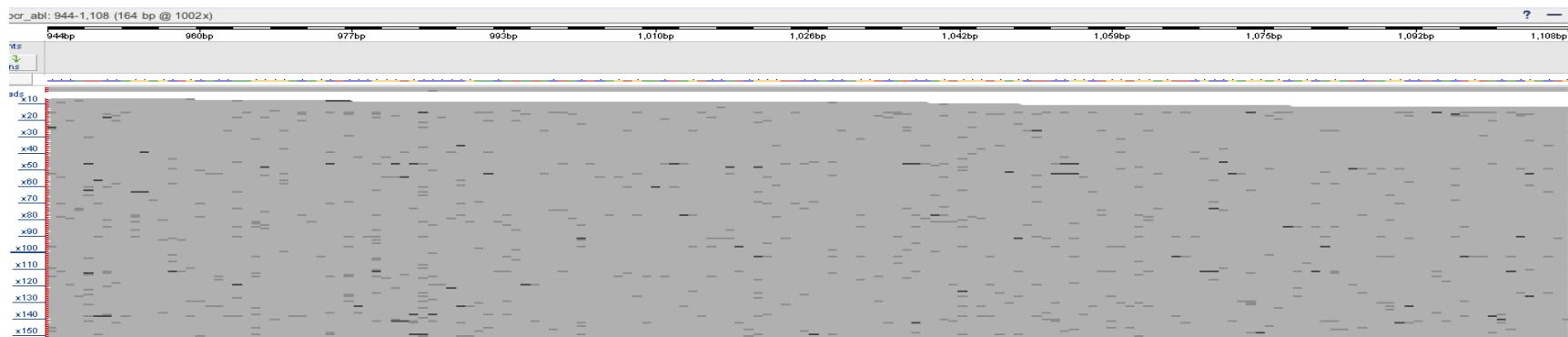
Results: Over 10,000 full-length *BCR-ABL1* sequences were obtained for all samples studied. Through the serial dilution analysis, mutations in CML patient samples could be detected down to a level of at least 1%. Notably, the assay was determined to be sufficiently sensitive even in patients harboring 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.



BCR-ABL1 mutations at diagnosis

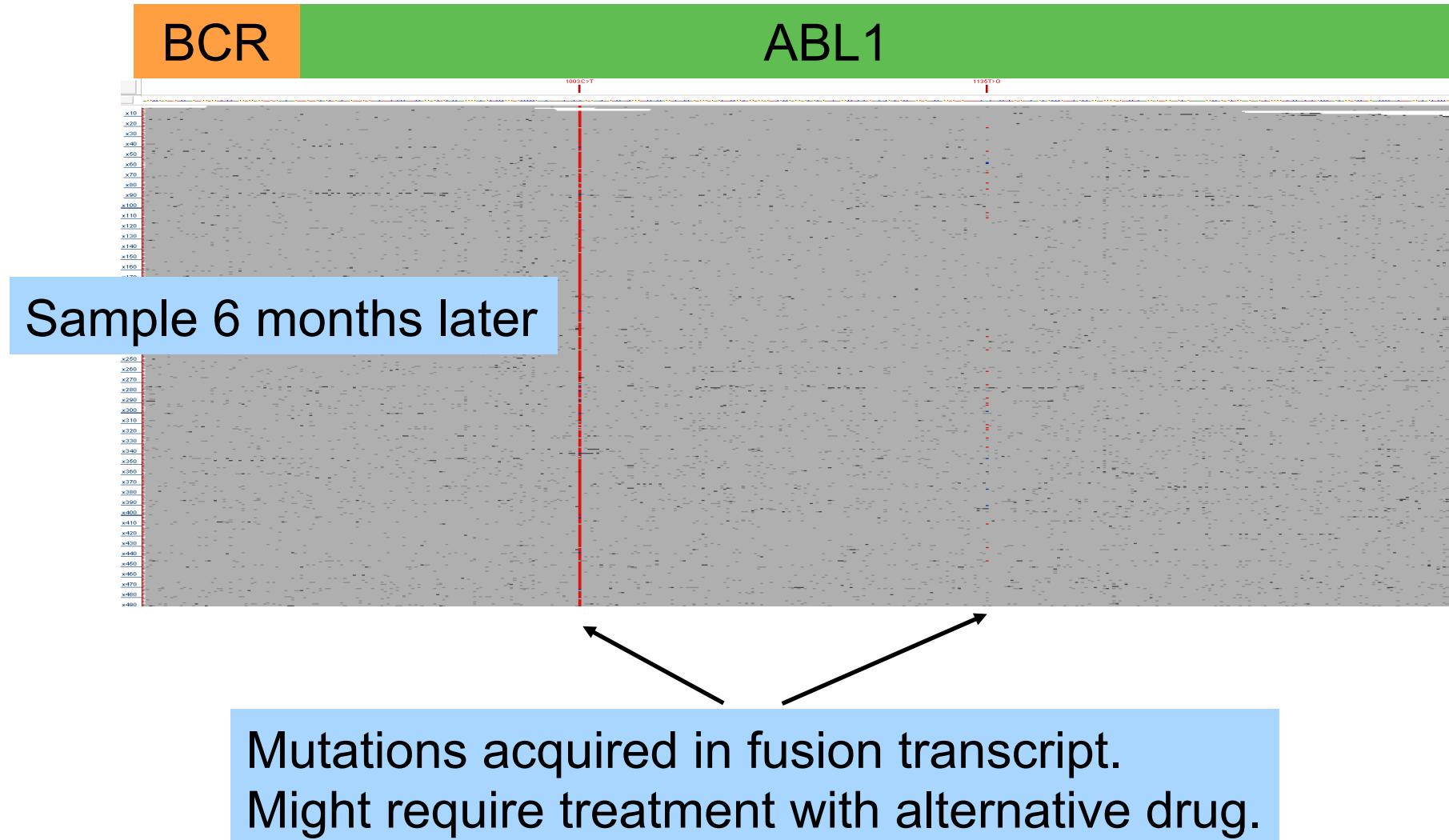
PacBio sequencing generates ~10 000X coverage!



Sample from time of diagnosis:

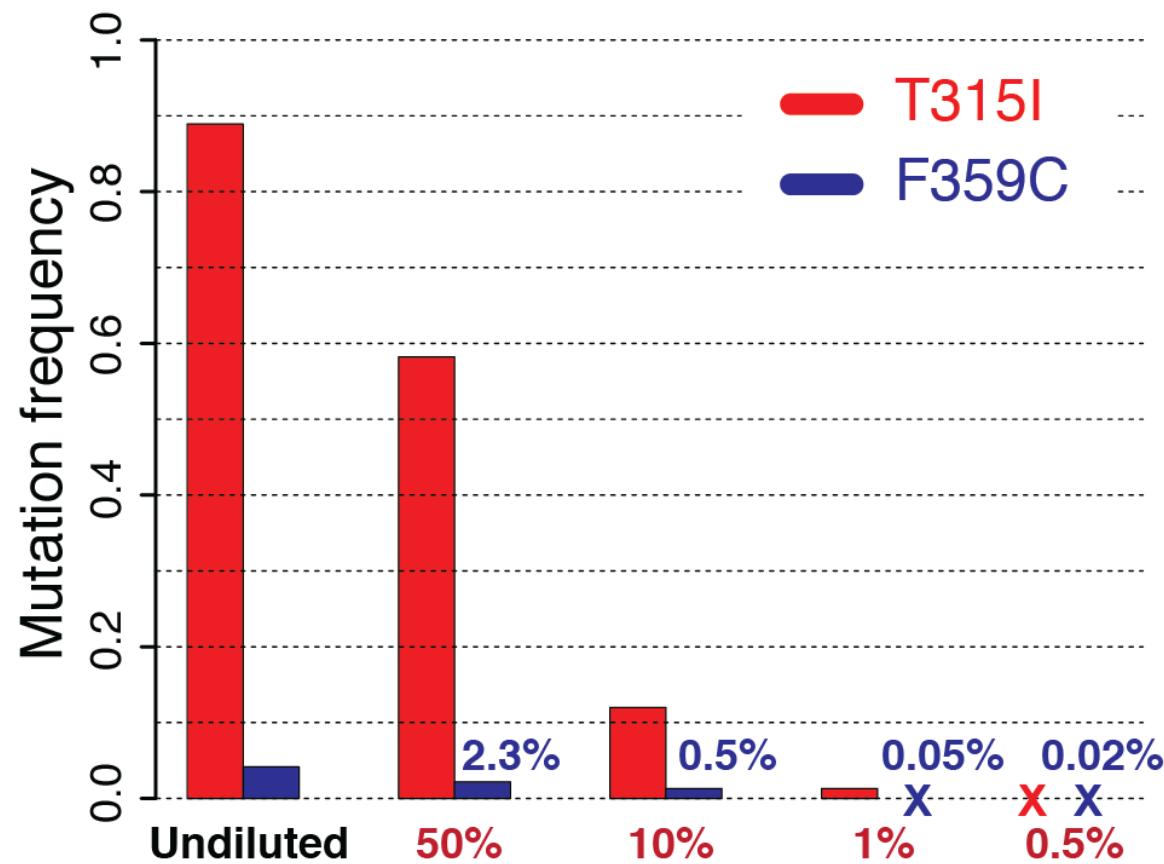


BCR-ABL1 mutations in follow-up sample

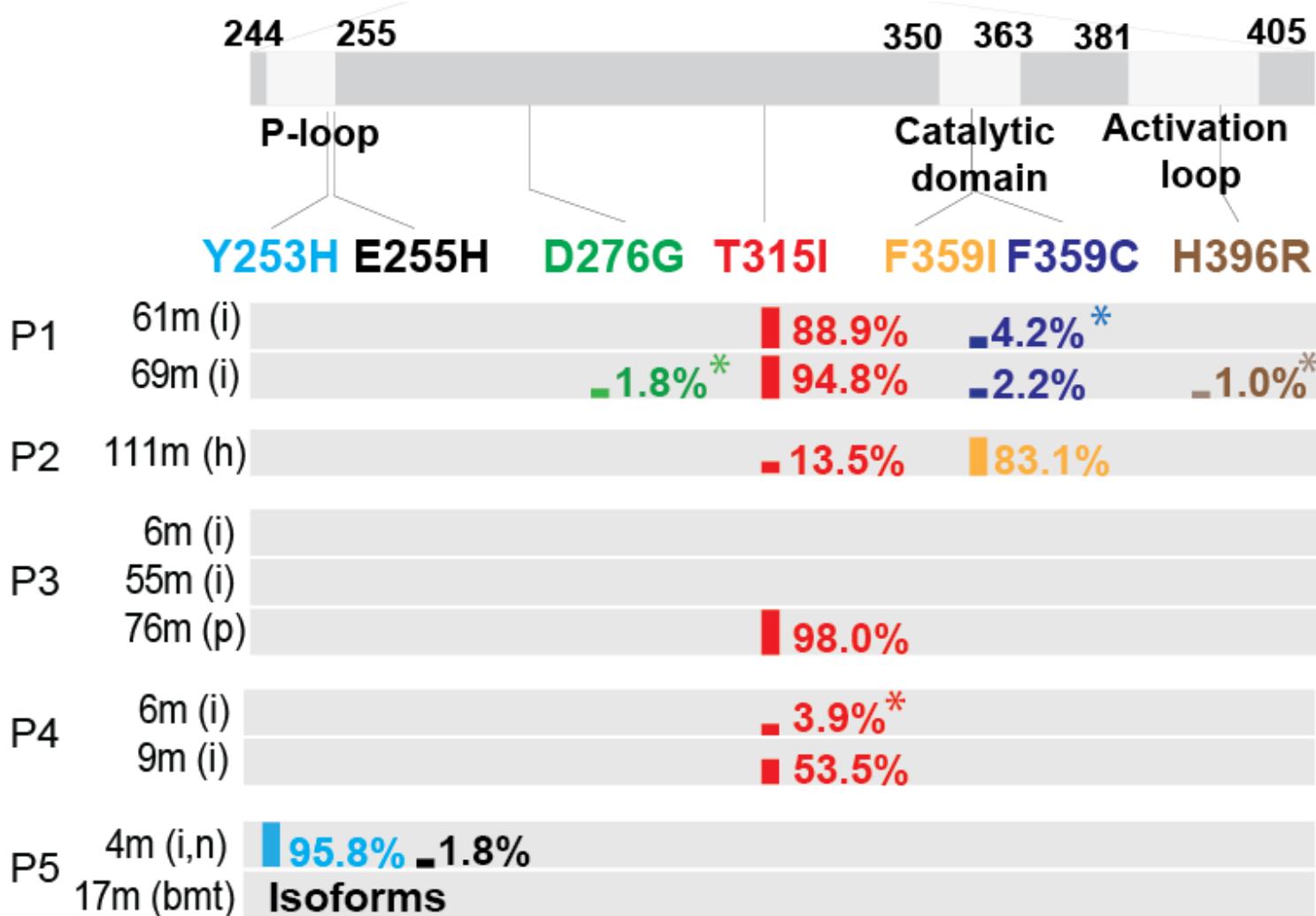


BCR-ABL1 dilution series results

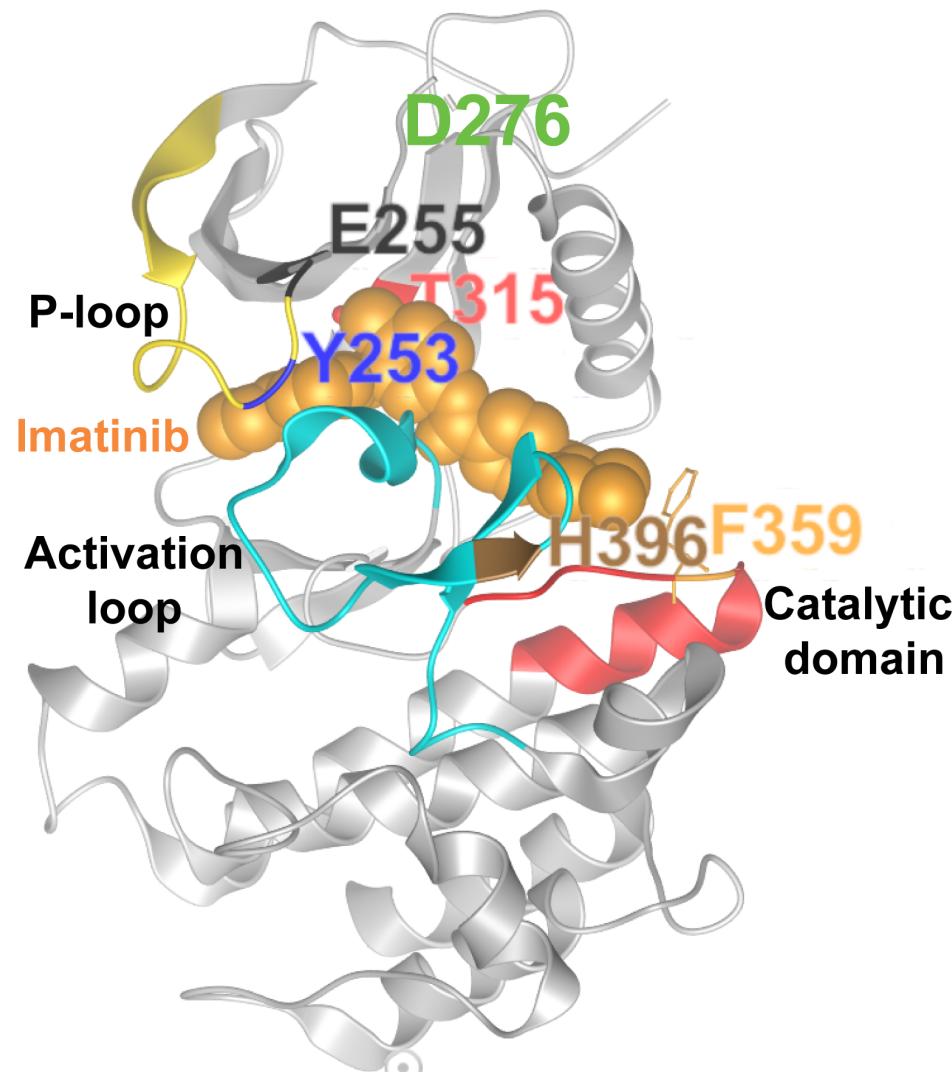
- Mutations down to 1% detected!



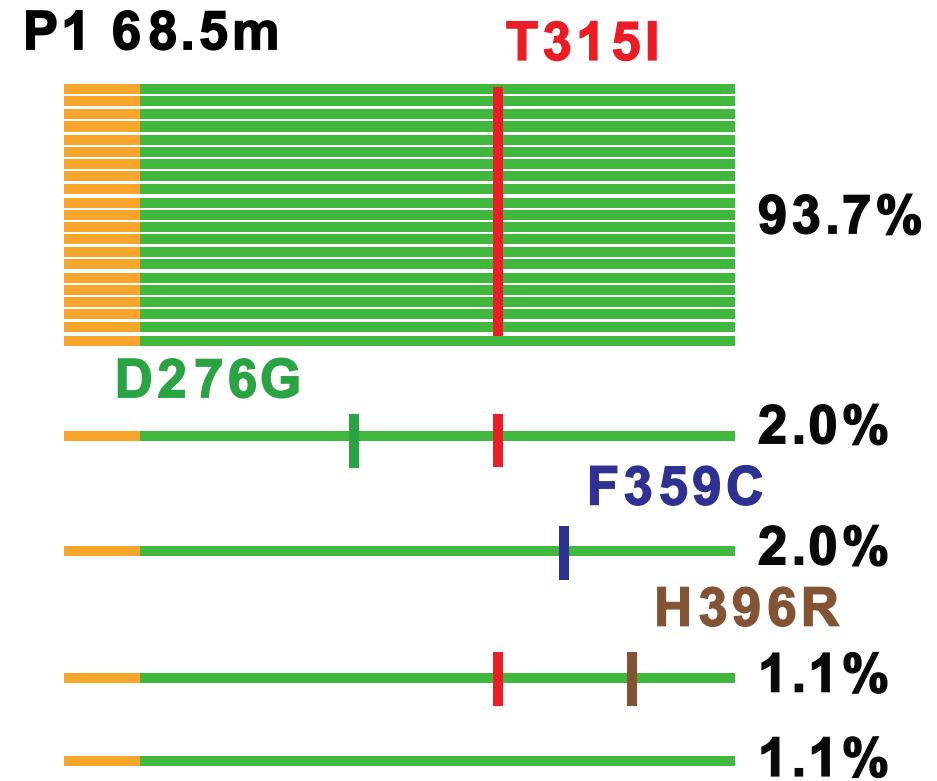
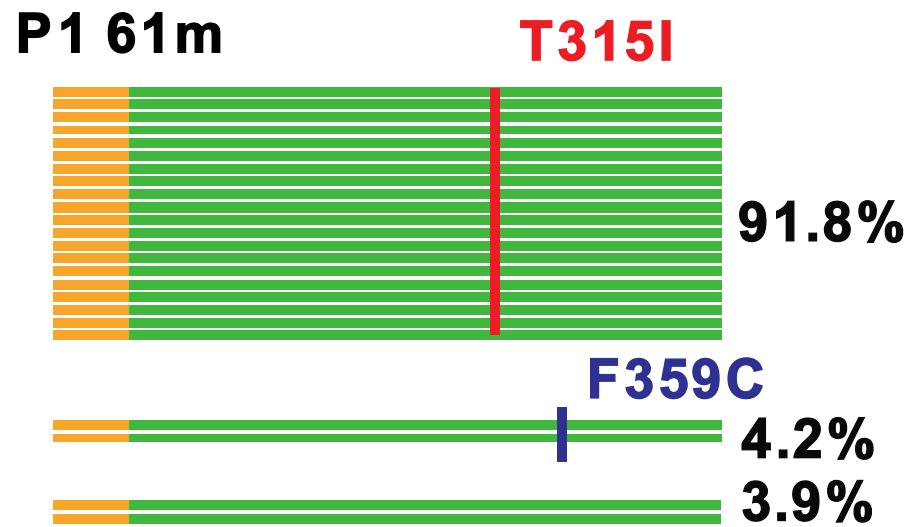
Summary of mutations in 5 CML patients



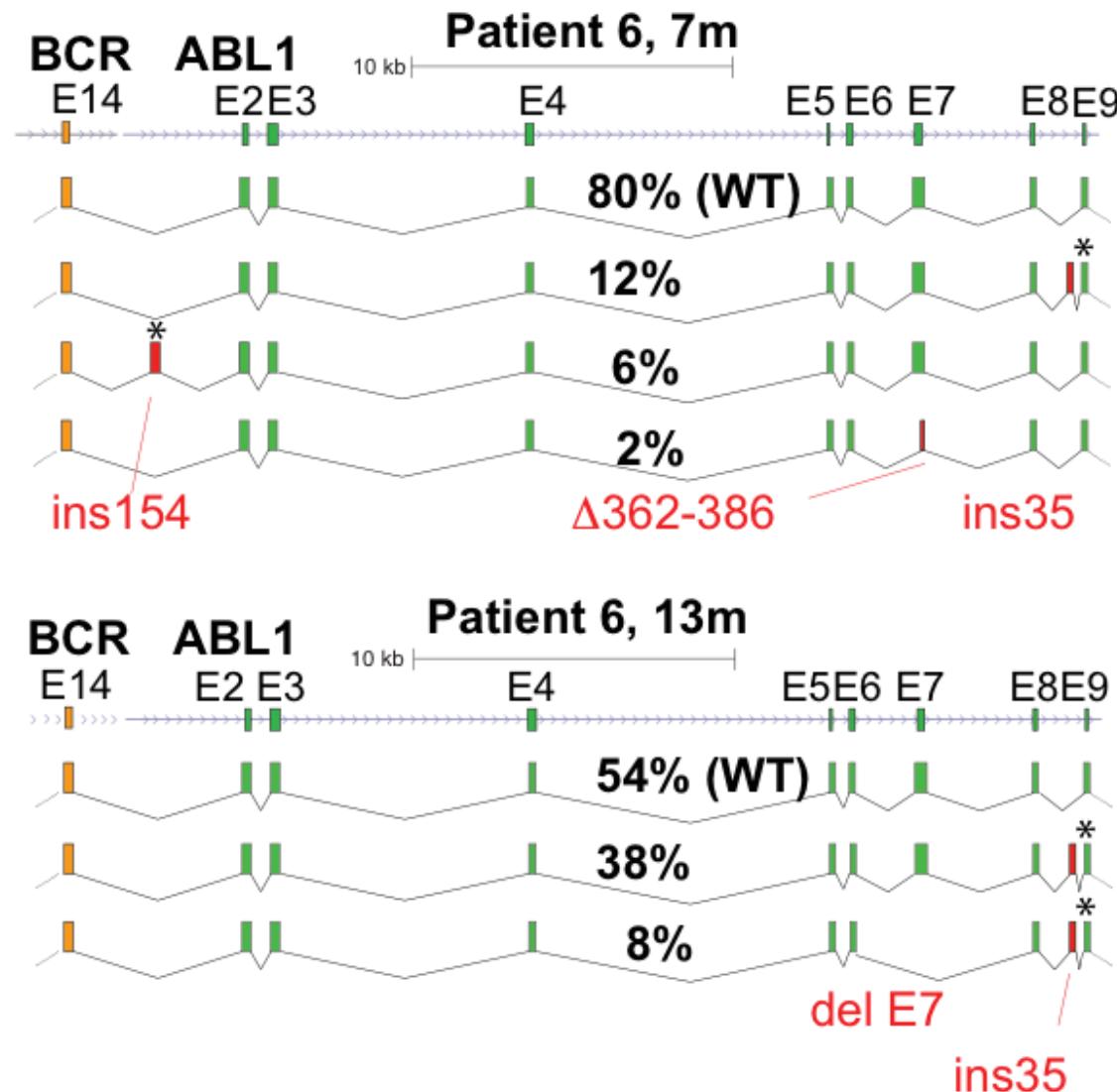
Mutations mapped to protein structure



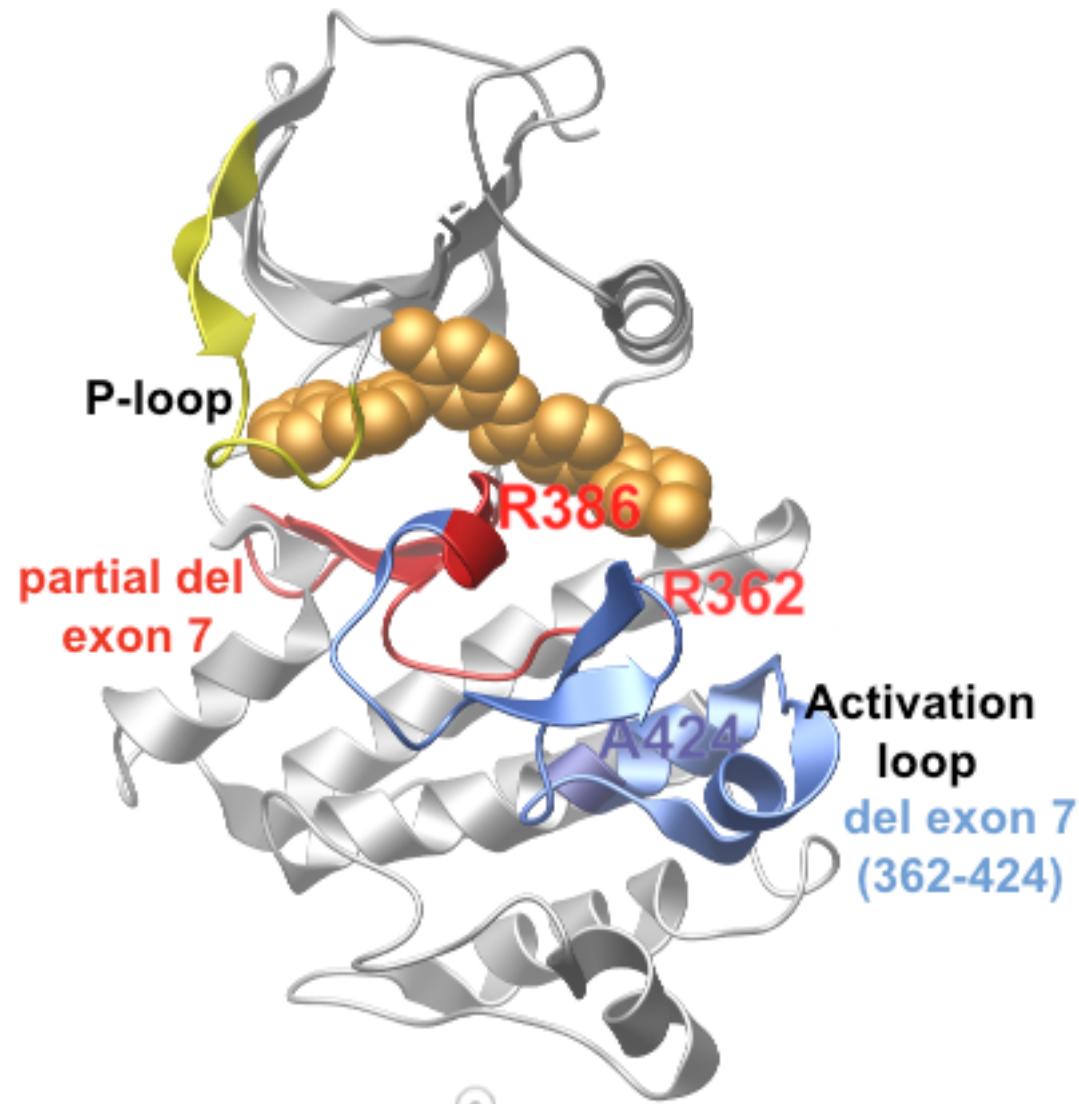
BCR-ABL1 - Compound mutations



BCR-ABL1 - Multiple isoforms in one individual!

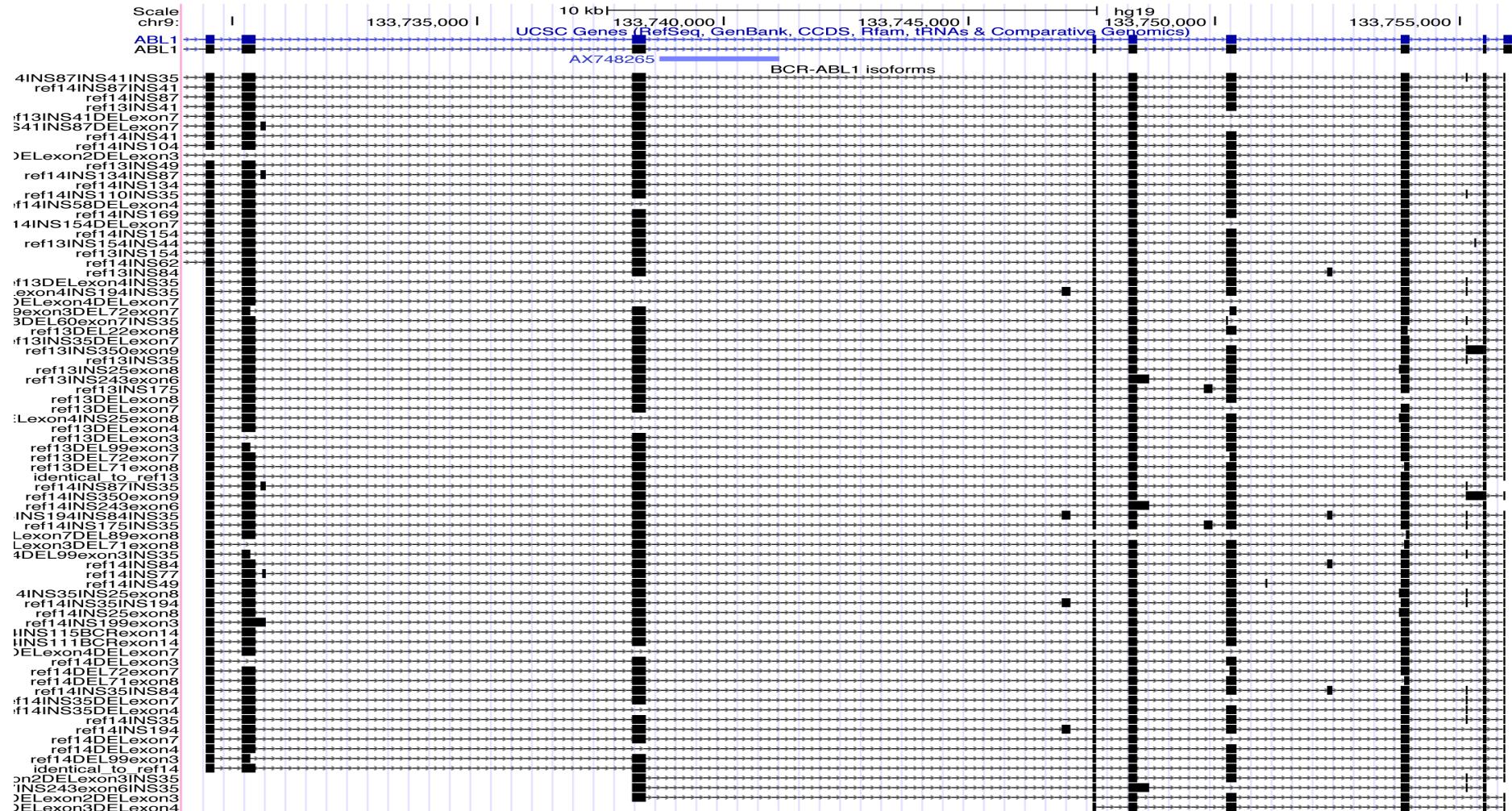


BCR-ABL1 – Isoforms and protein structure

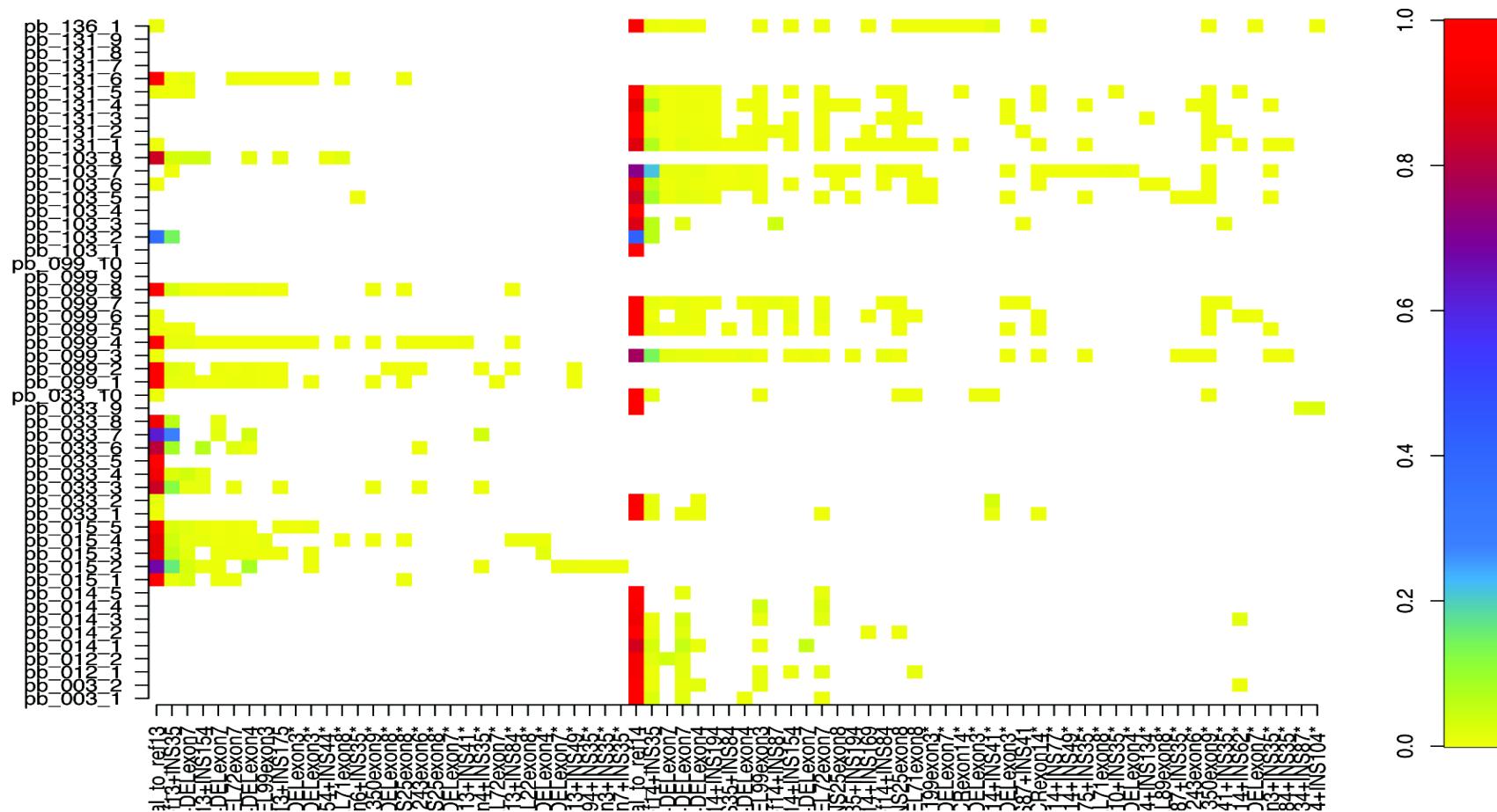


BCR-ABL1 splice isoforms

>100 different BCR-ABL1 isoforms identified!!!



Isoform expression levels



>20 isoforms found in some samples, most very low expressed!

Clinical Diagnosis of BCR-ABL1 mutations

Clinical Genetics



- Collection of samples
- Seq library preparation

Sequencing Facility



- SMRT sequencing
- CAVA analysis

IT developers



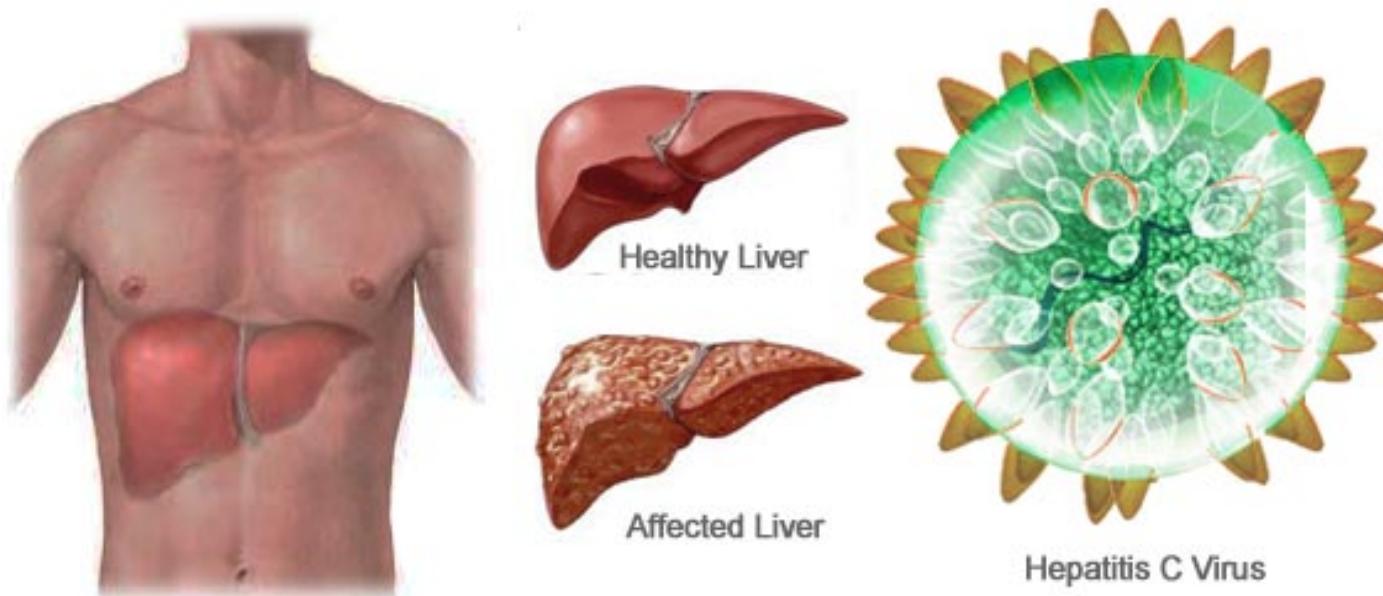
- Web server for results

- Ongoing routine service, 0-4 samples/week.
- Over 120 patient samples run so far
- 100% consistency with Sanger results

Web system for result sharing

Project II: Hepatitis C Virus Infection

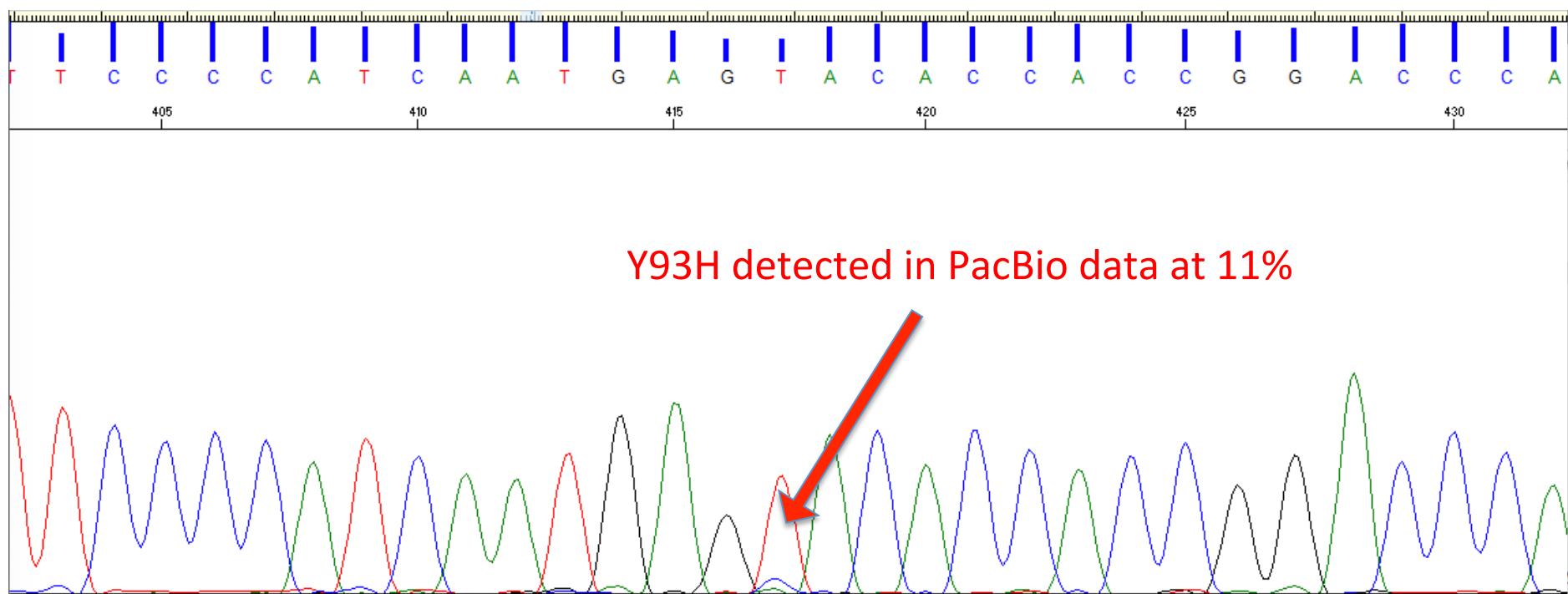
Infection of Hepatitis C (HCV) can cause liver disease



- Direct acting antiviral drugs (DAAs) target the Hepatitis C Virus
- Resistance development in response to DAA treatment
 - Depends on HCV genotype, resistance associated variants, etc...

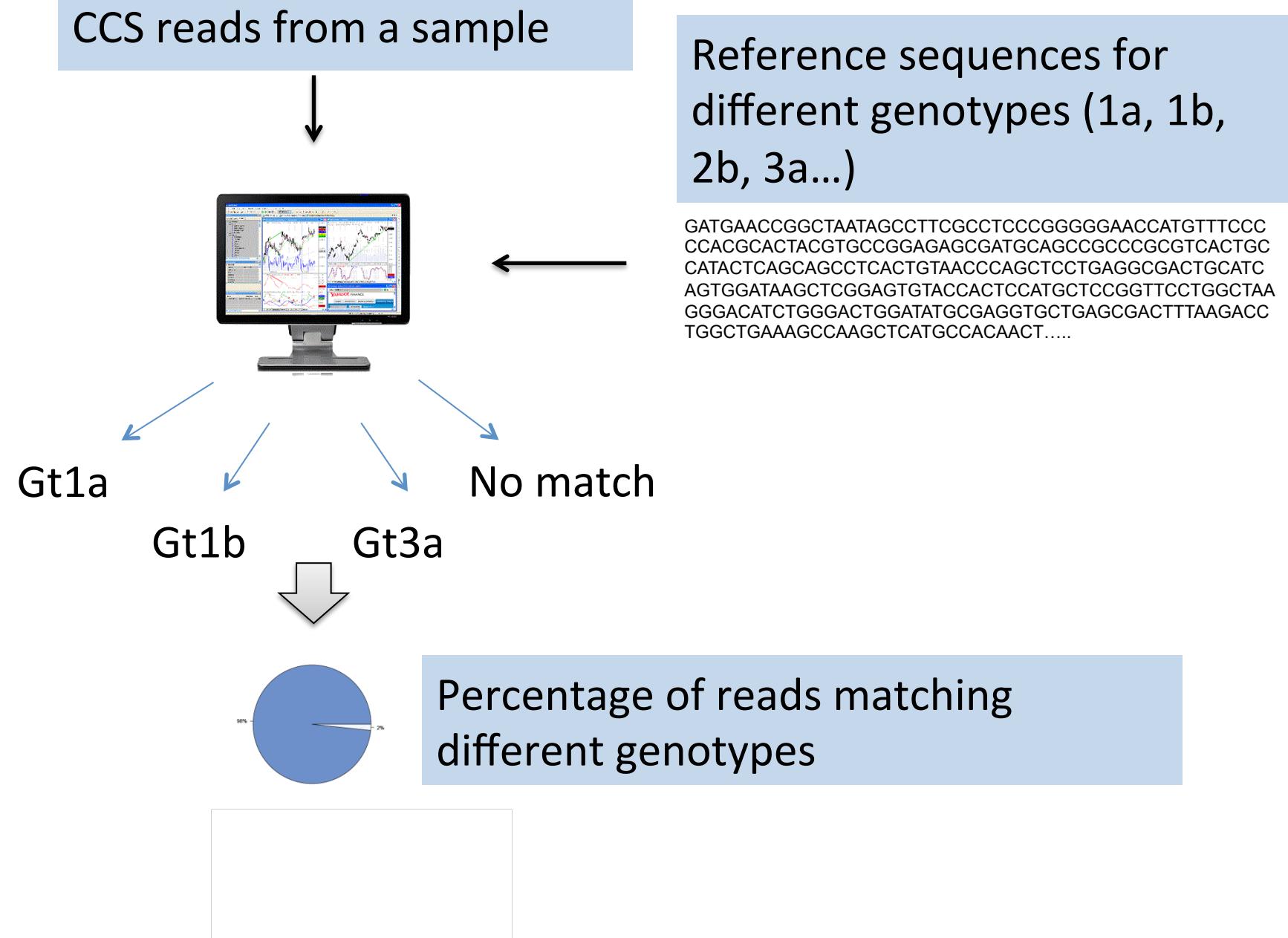
Results - low frequency mutations

- Example – We can see mutations that were missed by Sanger



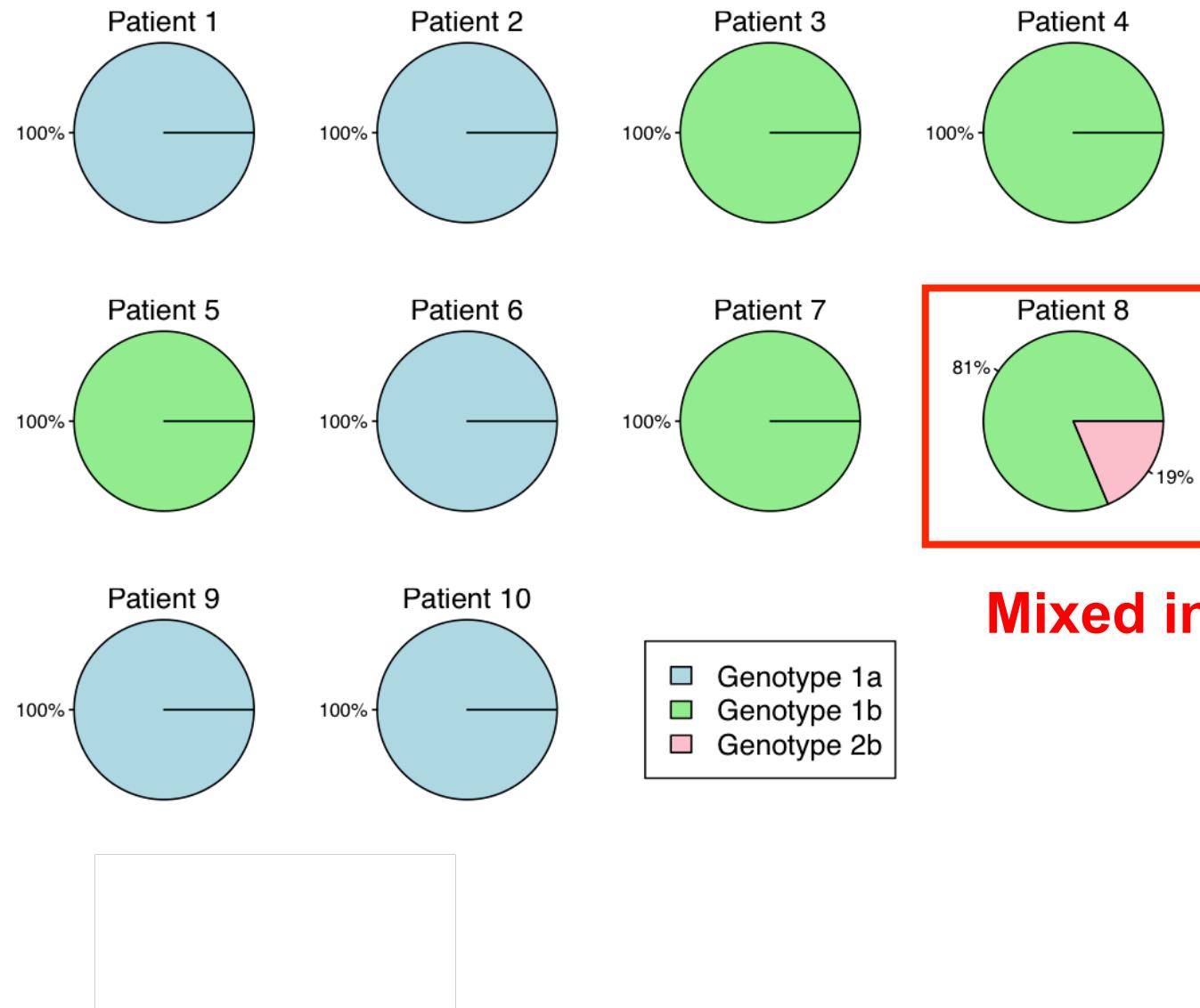
- Possible to detect developing mutations at an earlier stage!

HCV Genotyping by SMRT Sequencing

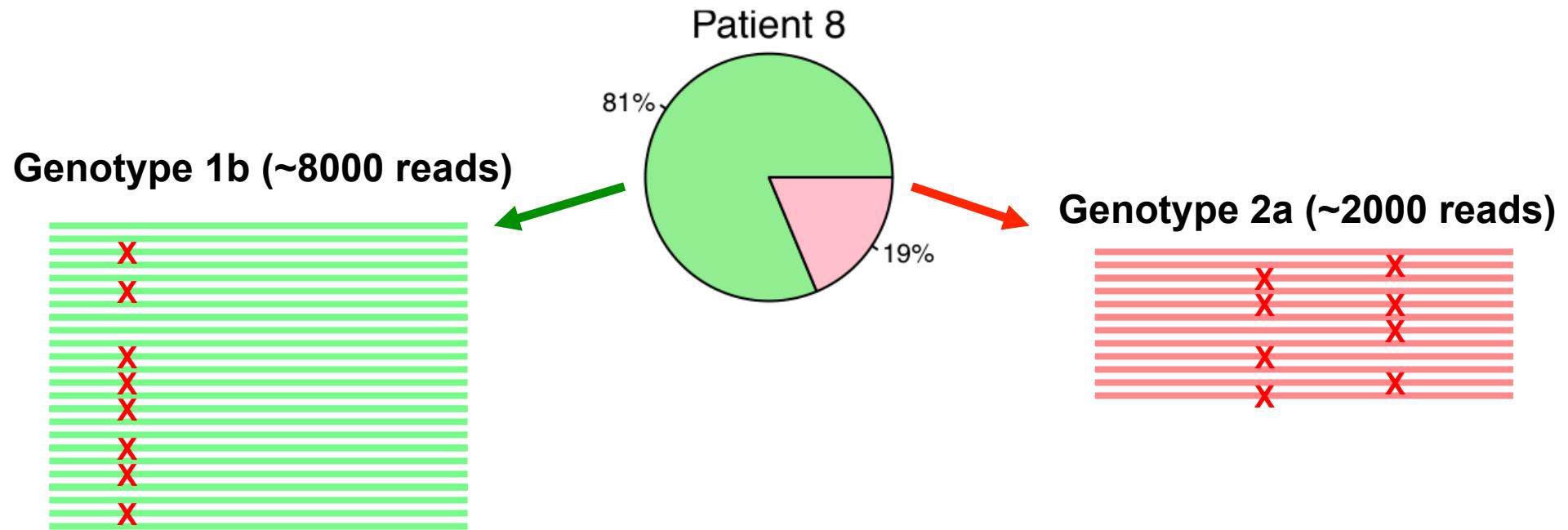


Genotyping of the Hepatitis C Virus

Distribution of reads in 10 patient samples (NS5B sequencing):



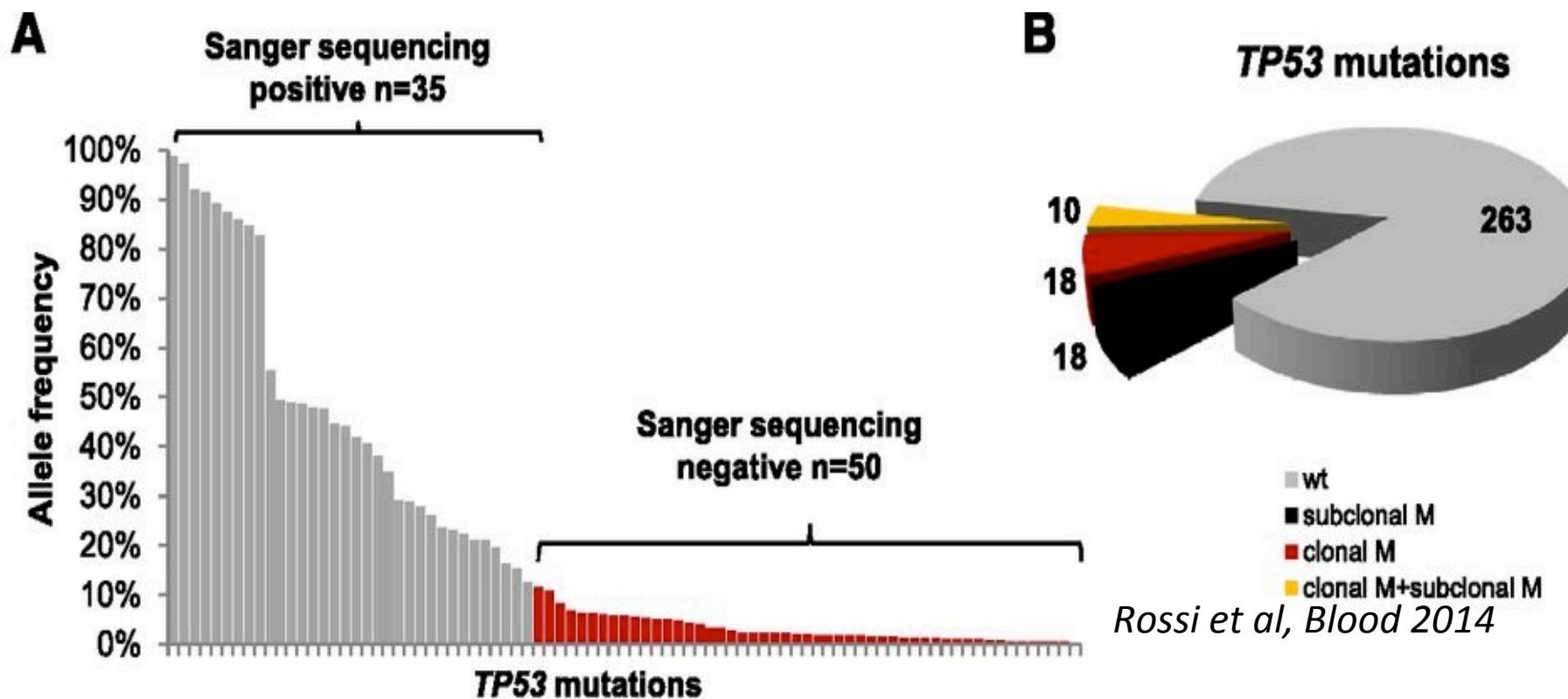
Detailed analyses of mixed HCV infections



- Reads from different HCV genotypes separated into groups
- Resistance mutations analyzed in each genotype!
- Ongoing work: Automation of genotype/mutation calling

Project III: Mutation screening of TP53

Identify low frequency mutations

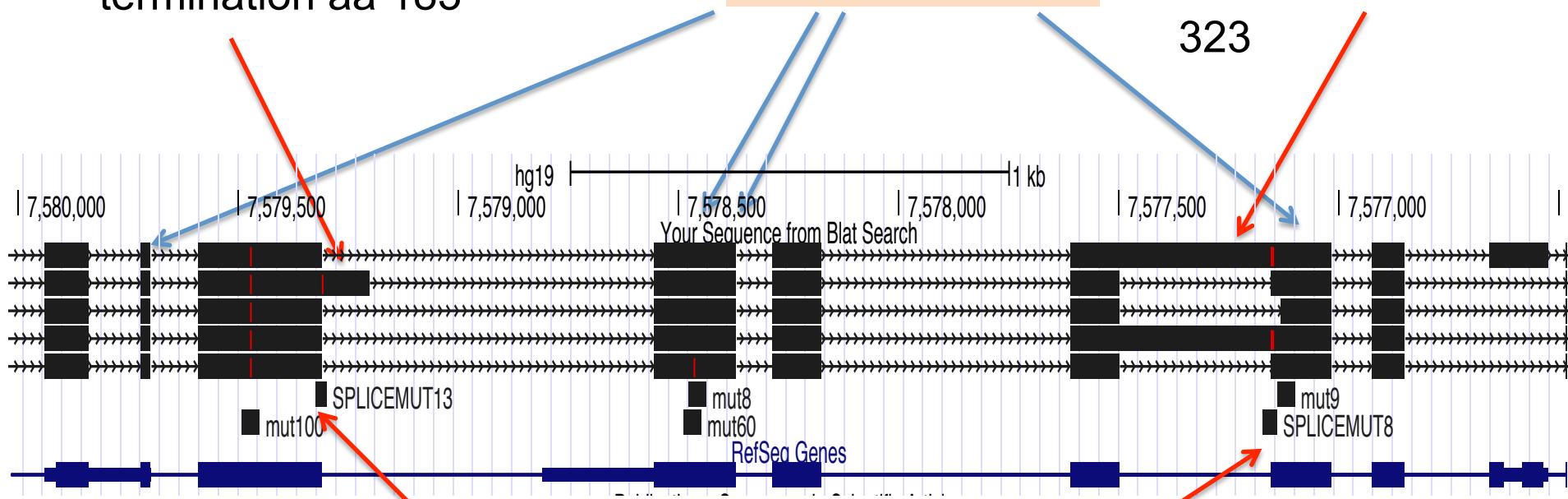


TP53 results – splice mutations and isoforms

Partial intron retention
and premature
termination aa 183

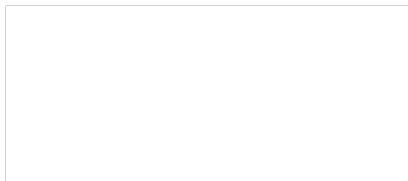
Other mutations

Intron retention
and premature
termination at aa
323



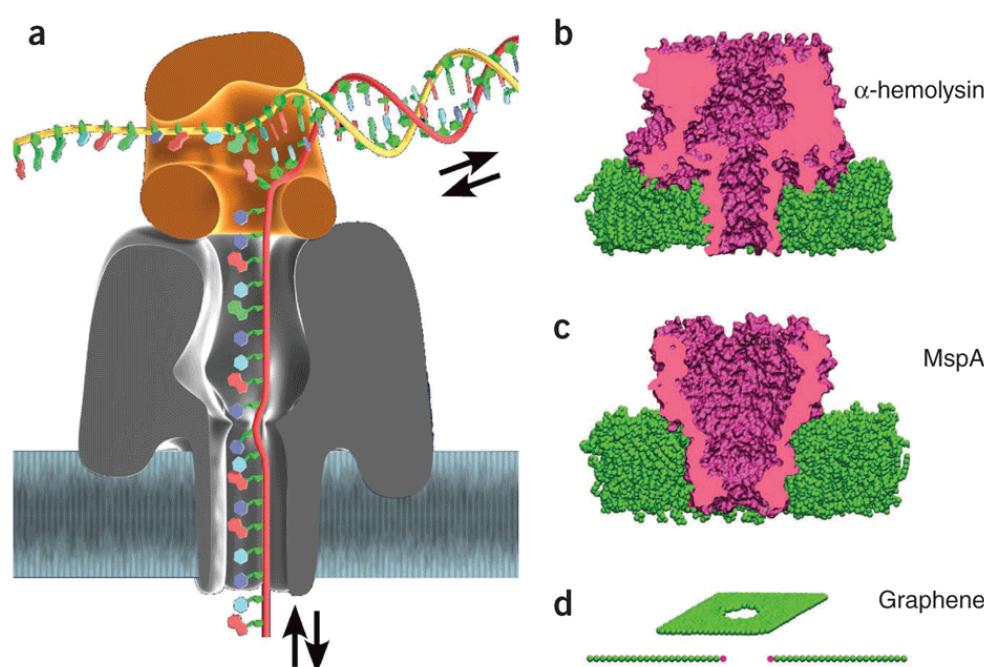
Splice site mutations

Are there other options?



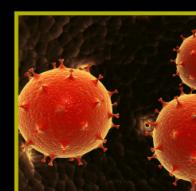
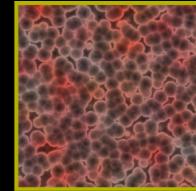
News and future directions

Nanopore technology - for direct RNA sequencing?



Enables detection of modified RNA bases??

What we sequence at NGI / SciLifeLab



Diabetes
Alzheimer's disease
Whole-genome sequencing
Gene therapy
Infection screen
Whole-transcriptome sequencing
Target sequencing
Cancer prognosis
Gene regulation
Crohn's disease
Genomics of ageing
Exome sequencing
Schizophrenia
Cancer diagnostics
Organ donor matching
Gut microflora
Gene fusions
RNA editing
HIV
HPV
HCV
Scoliosis
Immune response
Monogenic disorders
Sudden infant death
Cervical cancer
Lynch syndrom
Leukemia
Scoliosis
HLA typing
Dyslexia
MRSA / BRSA screen
Sudden cardiac arrest
Transcriptional regulation
Prenatal diagnostics
Muscle dystrophy
Individualised cancer therapy
and much more...

A stylized diagram of a human body showing internal organs and a network of yellow lines representing biological pathways or data flow.

THANK YOU