Human *de novo* genome sequencing with Pacific Biosciences platforms at the Institute for Genomic Medicine.

Sean D. McGrath, Amy Wetzel, Stephanie LaHaye, Catherine E. Cottrell, Richard K. Wilson and Vincent Magrini

Institute for Genomic Medicine
The Research Institute at Nationwide Children's Hospital

PacBio at IGM

Next Generation Sequencing (NGS) can be performed with short-read (50-300 bp) or long-read (>50 kb) technologies. Illumina dominates short-read technology and is used extensively at IGM (MiniSeq, MiSeq, HiSeq 2500, HiSeq 4000 & NovaSeq). Long-read technology is split between Oxford Nanopore and Pacific Biosciences (PacBio). IGM currently has two Pacific Biosciences Sequel sequencing systems that can sequence 1 million ZMW SMRT Cells for up to 20 hours per cell.

PacBio Glossary

Zero-mode waveguide (ZMW): The "well" of a PacBio SMRT Cell.
SMRT Cells: Consumable substrates comprising arrays of ZMW nanostructures.
Circular consensus sequencing (CCS) read: The consensus sequence resulting from alignment between subreads taken from a single ZMW. Generation of CCS reads using the CCS algorithm requires at least two full-pass subreads from the insert.
Polymerase read: A sequence of nucleotides incorporated by the DNA polymerase while reading a template, such as a circular SMRTbell template.
Subread: Each polymerase read is partitioned to form one or more subreads, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell template and no adapter sequences.
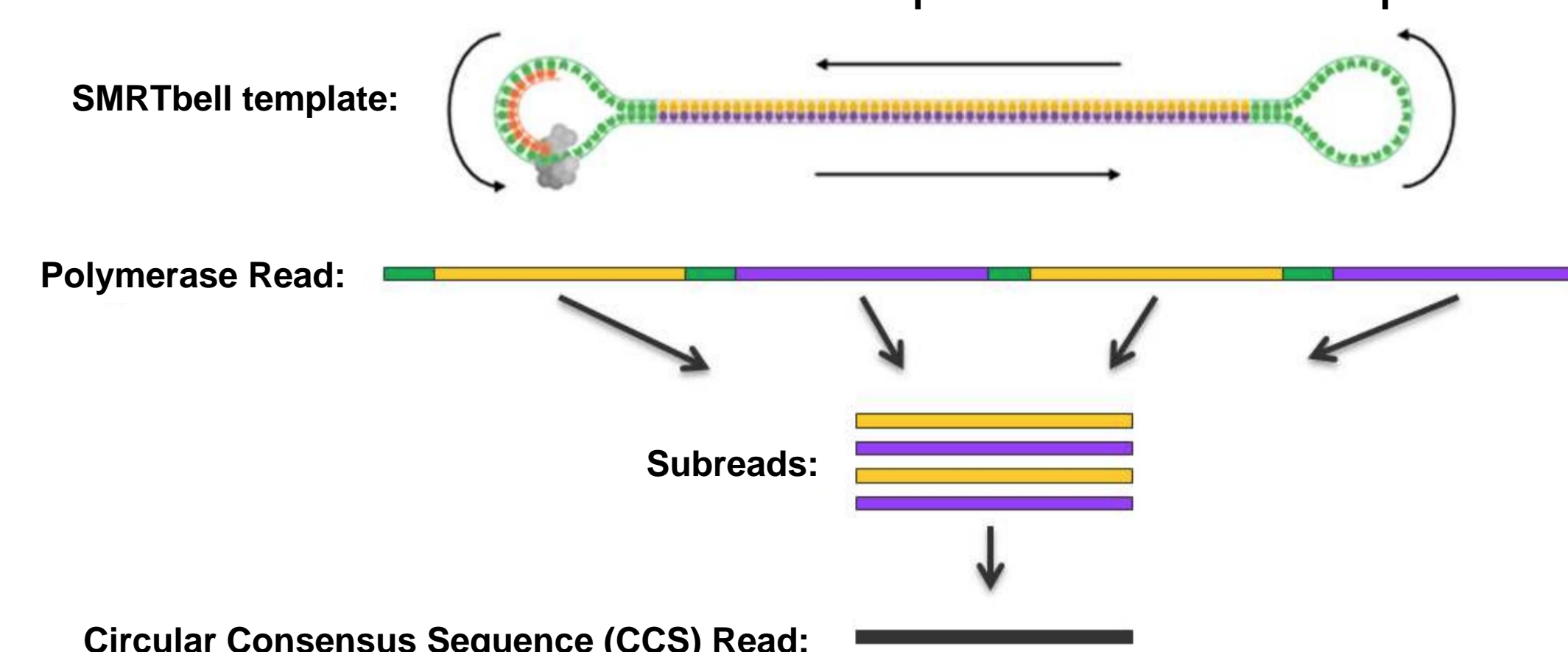
Current applications:
De novo Whole Genome Sequencing with >30 kb DNA
Human: Multiple SMRT cells to achieve 50-60x sequence coverage
Bacterial: Multiplex multiple samples in a single SMRT Cell
RNA-Seq aka PacBio Iso-Seq

Future applications:
Target enrichment
IDT xGen Lockdown with >5 kb DNA fragments
CRISPR/Cas9 "No Amp"
Whole Genome Sequencing by CCS
8 Million ZMW SMRT Cells
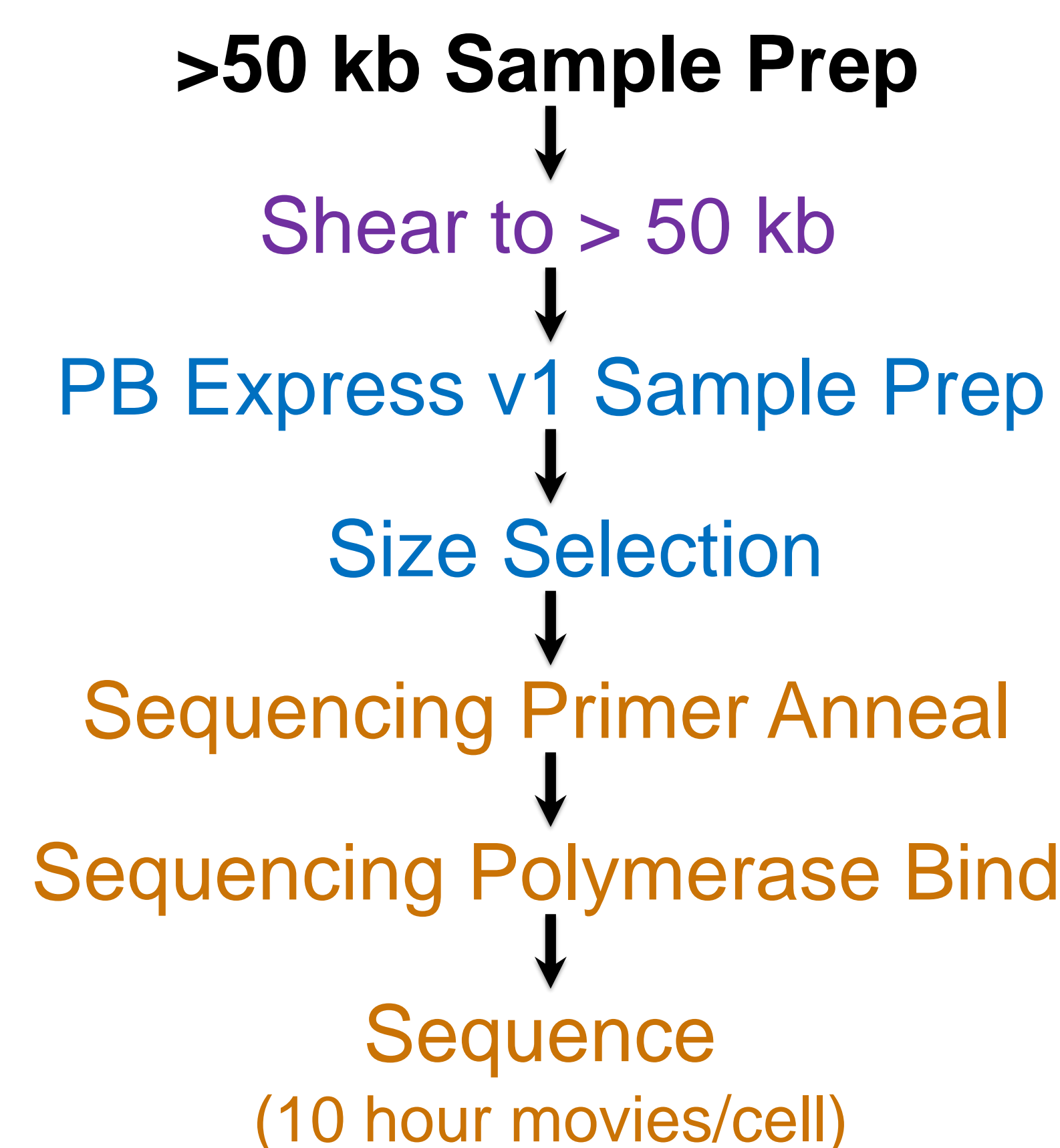
Human *de novo* Whole Genome Sequencing

>50 kb Sample Prep

Shear to > 50 kb
↓
PB Express v1 Sample Prep
↓
Size Selection
↓
Sequencing Primer Anneal
↓
Sequencing Polymerase Bind
↓
Sequence
(10 hour movies/cell)

Figure 1 - PacBio Express v1 Sample Prep Workflow

Table 1 - Human *de novo* Whole Genome Sequencing with PacBio long reads

Sample	# SMRT Cells	Total Bases (GB)	Bases/cell (GB)	Insert Length AVG	Insert N50	Binding version	Chemistry version
HG04217 (RS II)	284	157.28	0.55	11,473.13	NA	P6v2	4.0 v2
A673	26	178.40	6.86	13,256.59	22,942.71	2.00	2.10
HG04217 - All	62	334.87	5.40	11,397.14	20,180.56	2.00	2.10
HG04217 - Lib 7	44	234.03	5.32	12,296.40	21,520.83	2.00	2.10
HG02106	45	224.85	5.00	12,925.23	22,084.61	2.00	2.10
HG00268A	30	252.82	8.43	14,653.98	25,050.00	2.10	2.10
HG03125 - All	14	211.94	15.14	23,988.38	37,664.98	3.00	3.00

Continuing improvements in yield and read length when sequencing long-read libraries (>30 kb) PacBio SMRTbell libraries. The combination of Express v1 sample prep, v3 polymerase and v3 sequencing chemistry resulted in a 55.7% in sequence yield and a 61% increase in read length.
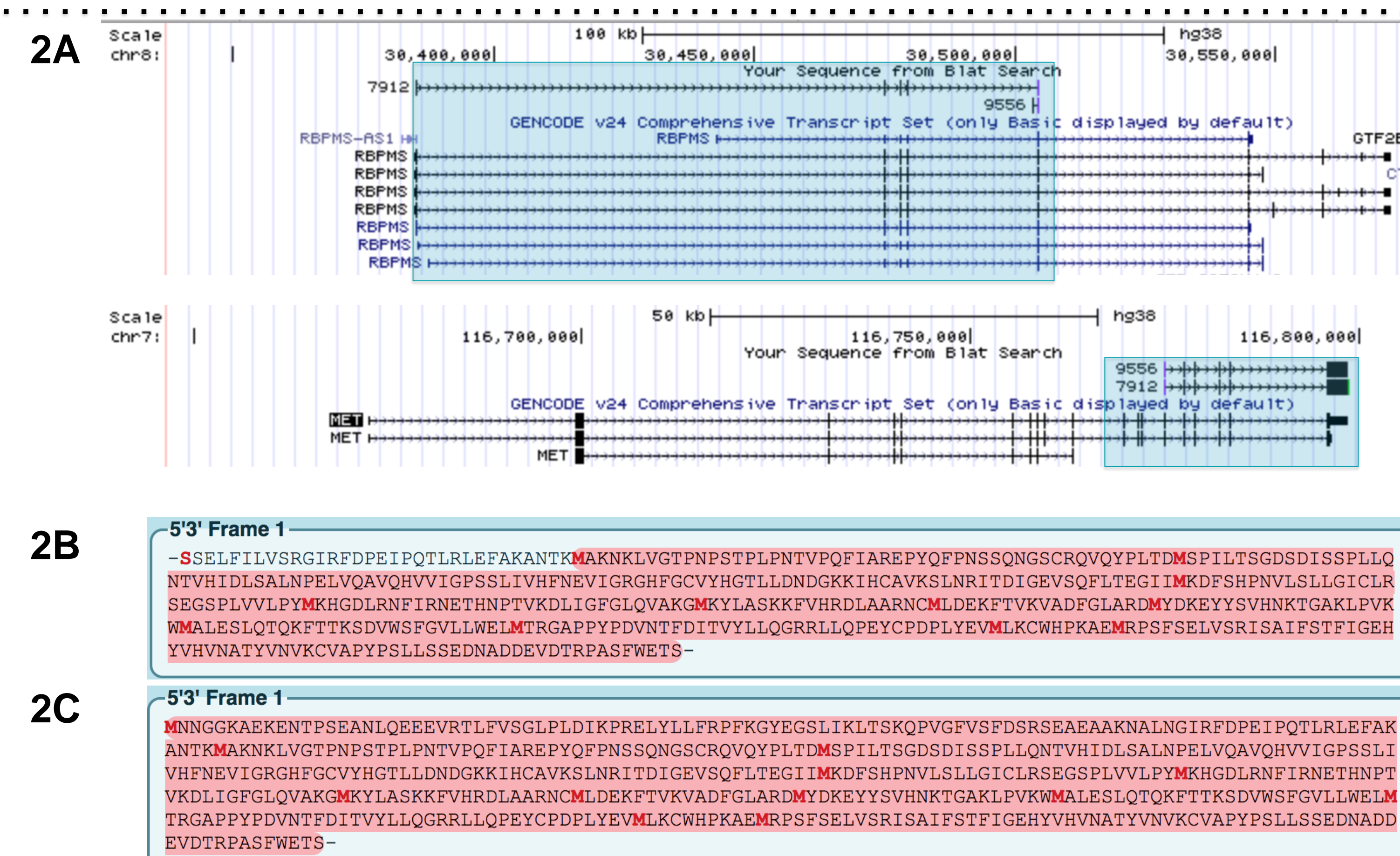
PacBio Iso-Seq

Figure 2 - PacBio Iso-Seq expressed fusion transcripts
2A: UCSC Browser view of two *RBPMS-MET* fusion transcripts
2B: *In silico* translation of fusion transcript 7912
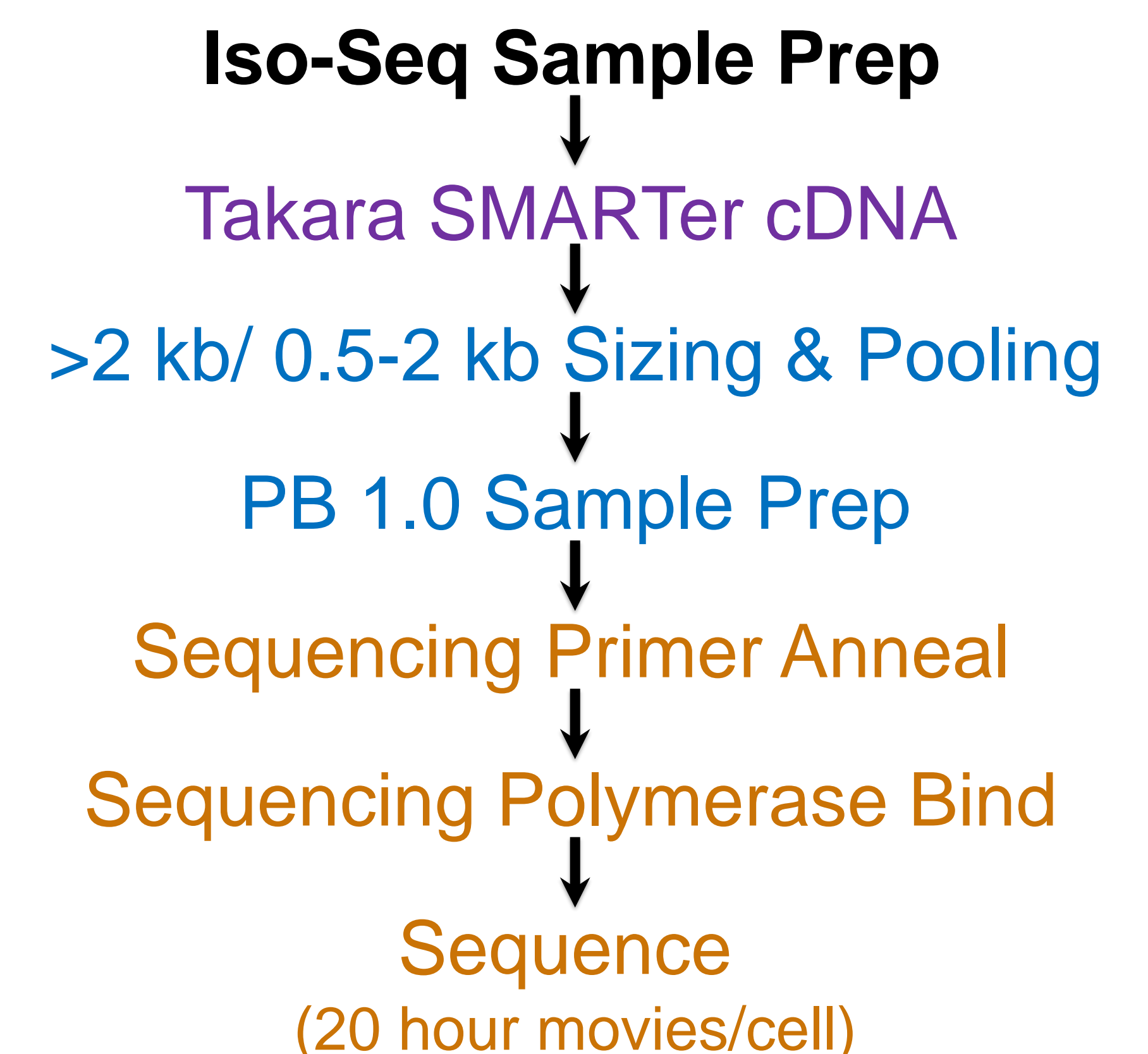2C: *In silico* translation of fusion transcript 9556

Iso-Seq Sample Prep

Takara SMARTer cDNA
↓
>2 kb/ 0.5-2 kb Sizing & Pooling
↓
PB 1.0 Sample Prep
↓
Sequencing Primer Anneal
↓
Sequencing Polymerase Bind
↓
Sequence
(20 hour movies/cell)

Figure 3 - PacBio Iso-Seq sample prep from 1 µg total RNA with RIN ≥ 7