

- How do long-read sequencing technologies work?
- When is long-read sequencing the right/wrong choice?
- Genomic and Transcriptomic applications

• How do long-read sequencing technologies work?

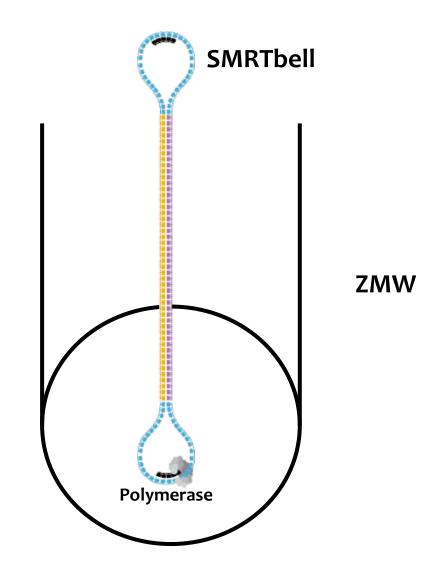
## PacBio SMRTbells

Polymerase binds to SMRTbell (<u>S</u>ingle-<u>M</u>olecule <u>R</u>eal-<u>T</u>ime), performs sequencing-by-synthesis inside ZMWs

Fluorophore emits light at nucleotide incorporation

Movie for each ZMW is parsed to produce read calls

• 16hr, 20hr, 30hr movies



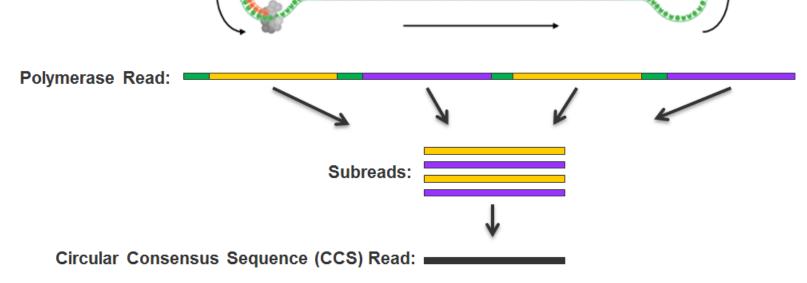
### **PacBio SMRTbells**

Circular Consensus Sequencing

Reads (**CCS Reads**) are produced when polymerase goes around SMRTbell ≥3 times

Can provide confidence for allele calling from single molecule, as a CCS read

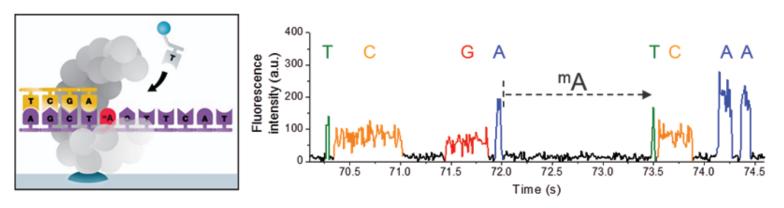
Large inserts (≥50 kbp) are unlikely to form CCS reads



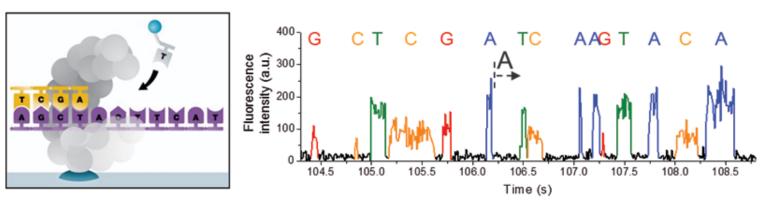
..........

# Detecting Base Modifications/Damage with PacBio SMRT bells

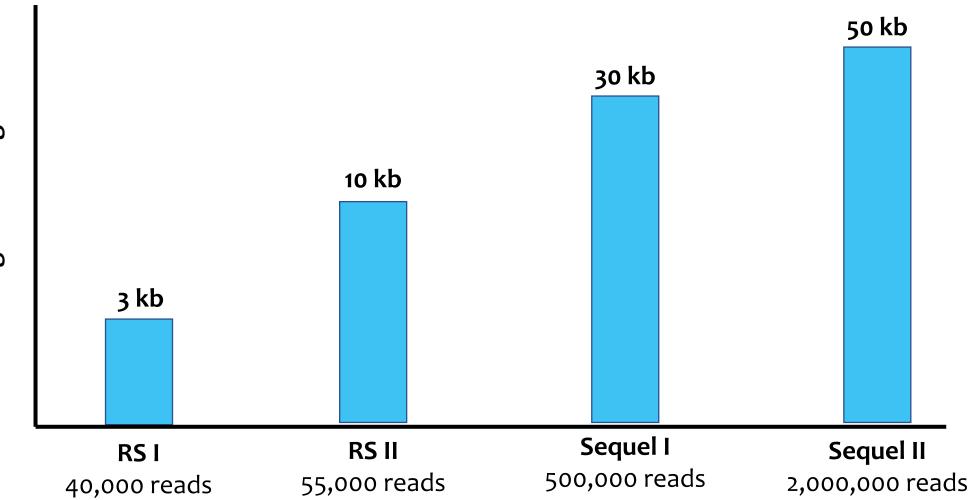
Base modifications impede polymerase processivity in a predictable manner



Can be measured with Inter-pulse Distance (IPD)



## PacBio read length is increasing



Average read length

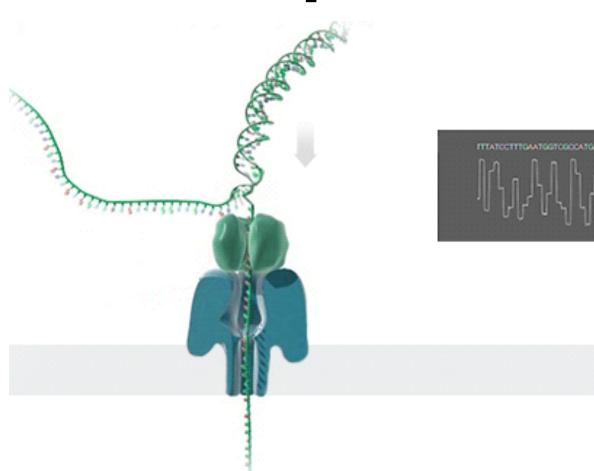
## **Oxford Nanopore**

*E. coli* channel protein embedded in membrane nanopore

Double-stranded DNA is unwound and fed through a channel

Change in voltage across membrane measured by flow of ions through channel

The extent to which **ssDNA blocks the flow of ions** is the output signal



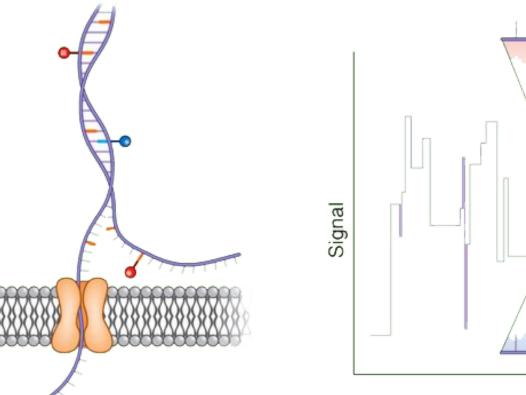
## **Oxford Nanopore**

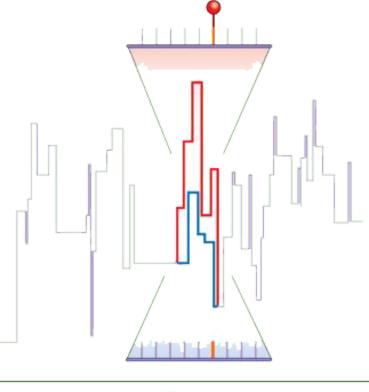
*E. coli* channel protein embedded in membrane nanopore

Double-stranded DNA is unwound and fed through a channel

Change in voltage across membrane measured by flow of ions through channel

The extent to which **ssDNA blocks the flow of ions** is the output signal

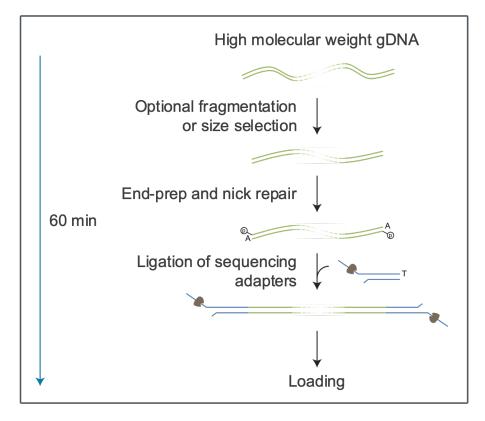




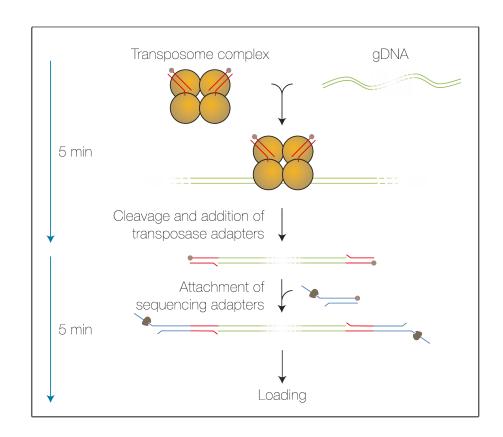


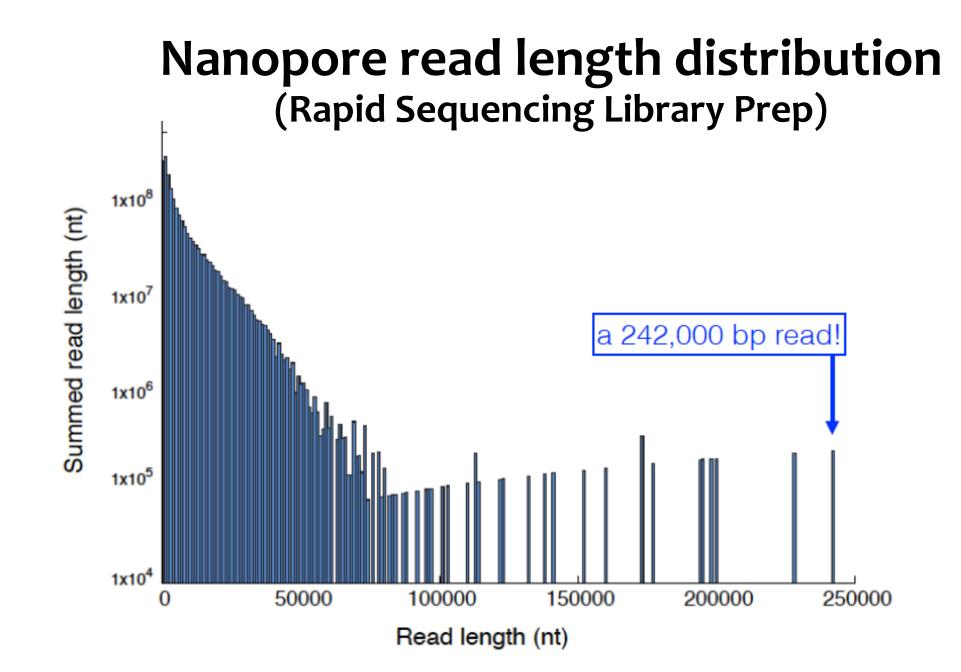
## **Oxford Nanopore Library Preps**

Ligation Prep (longer reads, more prep time)

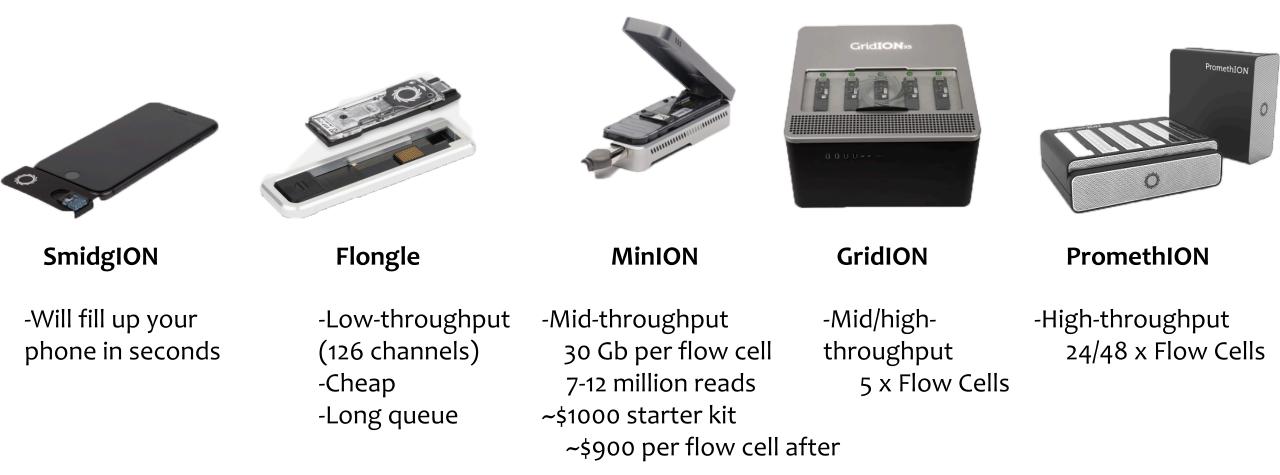


#### Rapid/Field Prep shorter reads, less prep time





## **Oxford Nanopore Sequencing Platforms**



• When is long-read sequencing the right/wrong choice?

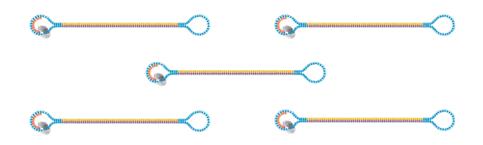
### Long reads have high error rates

	3,180,432 3,180,436 3,180,442		3,180,453 3,180,459	3,180,465 3,180,470	· · · · ·	3,180,486 3,180,492 3,
JVA2 CES	AGGCGCTCGACTTC	GTCGTAGT	CGATCAGCCCG	GTAGCGGTGTC	CAGGCCGTAC	TGCACGGCGTTG
glyA2 gene						
GGCTTATG-TTCACCCC	AGGCGCTCGACTTC	CGTAGT	CGATCAGC	GTAGCGGTGTC	C-GGCCAC	TGCACGGCGTTG
GCTTATGCTCCCCAGCCAGCC	A AGTTTGACTTC	5 - C	CGATCAG-CCG	GT\$AGCGGTGTC	CAGGCCGCAC	C G C A CSS G C G T T G T
GGCTTATGCTCCGGCC	AGGCGCTCGA(CTT-	- T – – – A G –	CG-TCCG	- TAACAGAATC	🗛 – – C 🕰 G T A 😰	T <mark>T</mark> C <mark>G</mark> CGGCGTTG1
GGCTTATGCTCCACGGCC	AGGCGCTCGACTC	STCGTAGT	CGATCAGCCC CGATCAACCCG	GTAGCGGTGTC AT <mark>G</mark> GCAAGTOCTC	CANGGCIGCAA - AGGCCGTAC	TGQ1CGGCGTTG1 TACACGGCGTTG1
	GGGCGCTCGACTTC	STOGTAGT	CGATCAGCCCA	GTAACCGAG-C	CA <b>T</b> TAC	TGCGGCGTT <mark>M</mark> I
:GGCTTATGCTCCACGGCC	AGGCGCTCGACTTC	STCGT-CATI	CAATCAACCAA	- T <mark>G</mark> GC <b>A</b> GT <b>Q</b> TC	AAA-CCA-	TGCA-G-CGTTC1
ICTAATGCTCCACGAC	AGACTCGACT	🤶 T – G 🖸 A G T I	CGATCAGCCCG	GTAGCGGTGTC	CAGGCCGTAC	TGCACGGCGTTGT
CTTCTGG-CCGCCGCC GGCTTATAC-CCACGG-C	🛠 🕻 G C G C T C 🤅 A 🕄 🖪 – C I	G T C\$G 🔂 4\$G T -	-GATCAGCCCG	GTAGCGGTOQTC	CAGGC <b>AA</b> -AC	TGCACGGCGTTG7
GGCTTATGCTCCACGGCC	AGGCGCTCGA	C G T A G T I	C G A T C A 🔂 C 🗹 🗹 G G	GTAGCGGTGTC	CAGGCCGTAC	TGCACGGCGTTG1
GGCTGCTCCACGGC	A - GTGC - CGACCT-I	STCG-AAT STCGT-OT	- GATCAG-CCTC CGATCAACCCCG	GTAGCGGTGTC GTAGCGGTG	CA-GCCGTAC -AG-CCGTAC	TGCACGGC-CAGC
AACTITATGCT COACGG GC	GGGCGCTCGACT	G T AT TA QT I	CGATCAACCCG	GTAGCG <mark>T</mark> TGTC	– A G 🕵 C C G 🗖 A C	TGCGGCGTTAT
GCTT-TGOTOCACOGOC CGGCCCATTTAC	AGGCG-TCGAC-T-		CGATCAGCCC-0 CGATCACCCG0	GGCGGTGTC GTAGCGGTGTC	CAGG <mark>T</mark> CAC CAGGCC <b>CAAA</b>	CIGCAQCGCGTANG- TG <mark>GG</mark> CGCCGTAGC
IGGCTTA-G-TCCACGGTC	AGGC GC - CGACCCC	STO GAAATI	GGATCAGTCCA	GAAGC-GTSGTC	CAGTCAGTAA	T-CACGGCGT <mark>IG</mark> GT
GGCTTATGCTCCACGGCC		STCGTAGT	CGATCAGCCCA CGATCAACCCA	GTAGOGGTGTC G-MSCGGTGTC	CAGGCCATT CAGGCCGTAC	TA-A-G-CGTTG1
SGCTTATGCTCCACGGTG	GOCOCTCOACTTC	TCGGAAT	- GATCAGCCO	GCAGCGGTGTC	CAGGC	TGCACGGCGTTG-
IGGCTTATGCTCCACGGCC	- GGCACTCGACTTC	STCGTAGT	CG-TCAMCCCG	GTAGCGGTGTC	GAG-CCGTAC	TGCACGGCGTTGT
IGGCTTATGC-CCACGGTG IG-CTTGCTCCACATCC	GGCGCTCGACTTC	TCGTAGI	CGATCAGCCCG CGATCAGCCAG	-TAGEGGIGIE	CAGGCEGIAC CAGGCEGTAC	TGCACGGCGIIGI
IGTI-ATA-TACGGCC	■GGCGCTCGACCI	STCGTAGT	CGATCAGCCCG	GTAGCAGTC	GAG-CCGTAC	TGCACGGCGCGTGTGT
GGCTTAGGCTCACGCTG	GGCGCTCGAC-TC	STCGTAGT	CGATCAG-CC-	CGGTGTC	CAGGCCGTAC	TGCACGGCGTTG
SGCTTAT CTCCACSGC-		G-CGGT	CGAGCC-G	- TA-CTG-C	GGCAGT-C	E A C A - G - C - T T G
I-GTTATGCTCCACGG-C	GGCG-TCGACTTC	TCGTAGT	CGATCAG-CCG	GTAGCGGTGTC	CAGGCCGTAC	TGCACGGCGTTGT
IGGCT TATGTTO CAAAGCC	ASGCG-TOSAC-TC	≨TCGTAGT!	CGATCAGCCCG	G <b>C</b> A&GCGGTGTC	CAGGCCGTAC	T G C A C G G C C T T G T
I-GCTTATGC-CCACGG-C						
GGCTTATGCTCCACGGCC						
IGG-TT-T-T-CAC-GCC						
IGGC <mark>GCC</mark> -GCTCCG <mark>A</mark> CC	AAACTCGCGACTTC	ат с <mark>А</mark> т <b>х</b> абат	- GATCAACCCG	G T A\$G C G <mark>AC</mark> G	- A TOCCGTAC	TGCAQ5GCGT-Q1
IGGCTTATGC-CCACGGCC	AGGCGCTCGACTTC		CGATCGGCCCG	GTA-CAGCGTC	CAGGCCGTAC	
GGCACCTCCA-G-CC -GC-CATGC-CCACGG-C						
GGCTTCGGCT-CA <b>RT</b> GCC	-GGCGTTGACTTC	STCG-AAT	GATCAGCCCA	G🗛 A G - 🗖 G T - T <mark>G</mark>	C G∎C G T A C	TGCAC-G-GTTG

### But you can use the consensus sequence for assembly

3,180,414	3,180,4	18 3,18	0,422	3,180,	427	3,180,3	32 3,1	80,436	3	180,44	12	3,180,	447	3,	180,453	3	3,18	1,459		3,180,4	165	3,180	,470		3,180,4	76	3,180	481	3,180	1486	3,18	0,292
GCT		GCT	ĊĊA	CGĠ	CC	A G Ġ	CGC	TC(	SAC	т†с	G	ĊĠ	TA	G T (	GÂ	TC.	AG	cc	GG	ΤÀ	GC	GG	G	СC	AĠ	GCC	GŤ	AC	T G C	AC	GGC	ĠTT
lyA2 CE4																																
lyA2 gen							_						_	_	_					_					_			_			_	
GCT GCT		6 - T		006	CCA	AGG	CGC	ΣĘ S	G A C	TIC	AC	CG	TA(	G T (	GA	TC:	AGO		GG	ΤA	GC	GGI	GI	CC	- G (	GC C		AC.	TGC	AC	GGC	GTT
GCT	t A t d	śč-	CCA	CGG	čči	4	AGT	τŤ	G A C	<del>T</del> ta	G -	- c -			GA	τč	AG-	- čč	GG	TSA .	GC	GGI	GŤ	CC.	ÂG	ŝči	6	AC	GGC	Ad	GGC	GTT
GCT	TATO	GCT	сс-	- G G	CC	A G G	CGC	TCO	G ASC	ΤΤ-	T		- A (	G – (	C G -	TC	(		- G -	TA	ALC 1	AGA	A T	CA	A	- C C	χG Τ	ASC 1	ТТС	GC	GGCI	GТТ
GCT	TATO	GCT	CCA	CGG	CC	<b>4</b> 5 G	CGC	TC	G C	TTC	GT	CA	т-(	<b>≣≹</b> ∏ (	A A	<b>ξ</b> Τ⊂.	AGO		GG	ΤA	GC.	A G I	GT	CA	AA	A-	- T	A C	TGC		GGC	GТ-
GCT	тата	аст	CCA	CGG	cci	A G G	cac	тс	3 A <b>S</b> C	- те		C G	тас	яτα	GA	тс	AGO	. c c	G	ΤA	sc	GGT	бŢ	c.c	150	s c 🗖	GC	A A	тас	<b>S</b>	GGC	атт
ACT	t Ast d	ĞČŤ	ČČA	ČĞĂ	čĞ/	4 G -	čA-	TAT	AAC	тtč	GT	čă	TÃ	ĞΤά	ČĞĂ	τč	AAC	ččč	GA	ΤĜ	ĞČ.	ĂĞĪ	ŏ	·č-	AG	GCC	GT	AC.	TĂČ	AC	GGC	Ğİİ
	-																						-									
GCT GCT																																
-ct																																
ACA:	TAA	-CT	CCG	CGG	CC	AGA	C	TCO	G A C	Τ	्र	- G	CAO	GTC	CGA	Τđ	AGO	CC	GG	TΑ	GC	GGT	G T	CC	AGO	GCC	GT	AC:	TGC	AC	GGC	GTT
-CT	TCT(	GG-	<u>c</u> c <mark>g</mark>	çç	cq	G G	CGC	Τ¢	g A <b>Ş</b> CI	<u>A-</u> S	GŢ	<b>X</b> G	<b>Q {</b>	GΞ:	- GA	TC	AGO	$\subseteq$	GG	ΤA	GC	GGT	হ	C C	AG	GC	A -	٨Q	TGG	AC	GGC	GII
GCT		AC-				GG	GC	TC	S A C	LIC	GI	2-	- A (	GTO	GA	TC	AGO	- 26	GG CC		GC	GGI	GI		AG	A	GI	AC:	C	2A C	GGC	
G	- A - ·		ACA	CGG	$\subset \square$	4 - G	GC	- 0.0	5 ASC	CT-	GT	CG	- A .	AT-	- GA	TC.	AG-	- C C	G	ΤA	GC(	6 G T	G T	CC	A = 0	5 C C	GT	AC:	TGC	: A 🔿	GGC	- CA
GCT	(	GCT	COA	CGG	CCA	AGG	CGC	TO	GAC	TTO	GT	CG	T - 0	ோ	CGA	TC.	AAC	CC	SG G	ΤA	GC	GGT	- G -		AG	- C C	GT	AC.	TGC	GO	G - C I	GTT
ACT:	TAT(	GCT	C CA	CGG	GCC	GGG	CGC	TC	G A C	1212	: G T	A	T <b>S</b> A (	হে (	CGA	TC.	A A (	$C \subset C$	GG	ΤA	GC	G T	ΓGΤ	с-	AGO	SC (	GC	A C	ТGС		GGC	GTT
GCT		gQ	C A	CQG		A G G	cg-	109	G A C	-+-	GC	CG		<u>g + </u> ;	GA	TC	AGO	200	- 6	 	GC	GGI	GI	20	AG	G	-	AC	GC	AQ	GGC	
GCT		б-Т	CCA	CGG	Te /	AGG	dac		5 ASC	ĊĊċ	GT	do	AAA		GA	Τč	AG		AG	AA	GC	- 61	≸i ⊤		AG	i da	GT	AA.	T - C	SA C		G T <b>1</b> 5
GCT:	TAT(	GCT	CCA	CGG	CCA	AGG	CGC	TCO	3 GC	TTC	GT	CG	TAG	G T (	CGA	TO	AGO	CC	AG	ΤA	GC	G 🖌 G 🕇	ГGТ	CC	AGO	GCO	AT	TT	TA-	- A - I	G - C I	GTT
GCC	ССТО	GCT	C	- G -	/	4 – G	CGC	TCO	3 🗛 C	TTO	GT	CG	TAG	GΤC	CGA	TC.	A A C	CCC	AG	- 🐴	GC	GGT	ΓGΤ	CC	AGO	GCC	GT	AC.	TAC	AC	GGC	GTT
GCT	1 A T (	GCT	CCA	CGG	a G	GG	CGC	129	G A C	++2	GI	C G	<b>G</b> A	<u>A</u> + ;	GA	10	AGO	200	<b>S</b>	<b>C</b> A	GC	GGI	GI	CC	AGO	GC.	G	AC.	1 G C	AC	GGC	GII
GCT.	$\pm 2 \pm 2$	SC-		CGG	<b>f</b> G	GG	cac	tea	SAC	-+-	T	- č -	- A (	а ту	GA	τč	A	200	GG	÷Α	GC	GGI	GT	- C C	AG		GT	AC.	TGC	AC	GGC	GTT
- CT	T (	GCT	COA	CAT	CC	GG	c - c	TCO	3 A C	ттс	ΞĠŤ	ĊG	TAC	GТÌ	GA	τc.	AGO	CA	G-	ΤA	GC	AGI	ΓGΤ	Ξč	AG	GCO	्र	AC.	TGC	AC	GGC	GTT
T																																
GCT																																
GCT																																
G T	TAT	5 C T	CCA	CGG	- C	GG	cg-	TEC	3 A C	ттс	GT	CG	TAC	бТŰ	E G A	TC.	AG-	- C C	GG	ΤA	GC	6 G T	GT	CC	AGO	5 C C	GT	A C	TGC	AC	GGCI	GТТ
GCT	TAT	GΤ	🔹 Al	AAG	CCA	4 <b>(</b> G G	CG-	т¢	3 A C	- T 9	্ৰ হা	CG	TAG	GΞ	GA	TC.	AGO	CC	GG		GC	6 G 1	GΤ	CC	AGO	G C C	GT	AC	ТGС	AC	G <b>{</b> G ⊂	्री T
GCT.	TA T(	GC-	CCA	CGG	- 0	GG	CGC	1 S S	GAC.	119	GI	<u>S-</u>	-23	GIS	GA	TC	AGO		GG	ΤA	GC	GGI	GI	60	AG	GCC	GŢ	AA.	15.9	AC	GGC	GTT
GCT G-T																																
G - TI	T - T	T	$- \subset A$	C - G	CC-	G	CGC		3 A C	- T C	: - T	C -	- A 🗸	A T -	- G -	Τ-,	A -	CC	AG	ΤA	GC	- G 🕻	GΤ	CC	AGO	S (	GT	ASCE	T (G C	\$4 <b>T</b>	GGC	G 🖸 T
GCG		GCT	CC-	- GA	CCA	A A A	ССС	GCO	3 A C	TTC	GT	CA	TSAS	GТ·	- GA	TC.	AAO	CC	GG	TA	GC	GAC	G -		AT	ac c	GT	A C	TGC	: A C	GGCI	GT-
GCT	TAT	GC-		CĞG	CCA	AGG	CGC	TC	3 A C	ŢŢĢ	GT	Q		G T (	GA	TC	GGO		GG	TA	- 5	AGC	GŢ	CC	AGO	sçç	GT	ĂC.	TGC	AC	GGC	GTT
GCA GC-	2		CCA	- <u>6</u> -	- 01		CGC	12	A C	++2	2	- 26	+23	2 S	- 24	te	A A G	- 7	A G	+0	-	GGT	-G+	AC	A G	200	6 A	2 C	- 60 T 60	AC.	GGC	8++
GCT.	TOG	GCT	- č A	TTG	cč.	- G G	čĞ	Т	SAC	τtà	ĠŤ	čš	- A)	AT-	GA	τč	AGO	i c'e	AG	AA	G -	G	+	Ğč	220	GTO	GT	ÃČ.	İĞĊ	Ač	- G -	Ğττ

### PacBio sequencing strategies



#### Molecule 1

Molecule n

an na ha a tha a ba a an a an a an a da an a da ann a a tha a tha a tha da a da a tha dan da a ba a an da a' a

Long inserts, few CCS reads De novo assembly



Subread 1	Su	br	ea	d	1
-----------	----	----	----	---	---

<u>\</u> <u>₩₽₩₽₽₩₽₩₽₩₽₽₩₽₽₩₽₽₩₽₩₽₽₩₽₽₩₽₽₩₽₽₩₽₽₩₩₽₽₩₩₽₽₩₽₽₩₽₽₩₽₽₩₩₽</u>₽₽₩₩₽ Subread n 

#### 

Short inserts, many CCS reads Isoform Sequencing (Iso-Seq)

## Which technology would you use?

- Quantifying gene expression among different isoforms in a non-model species
- Linkage analysis between SNPs that are on average 10kb apart
- Assemble a plant mitochondrial genome

### Use long reads

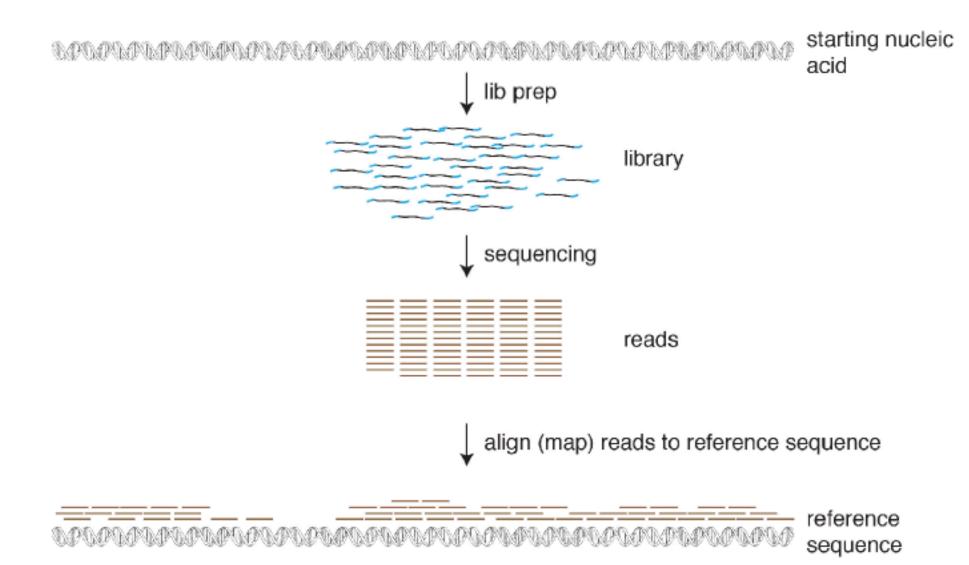
- When linkage is more important than nucleotide identity
- Identify structural variants
- Resolve complex DNA structures
- Sequence though repeats
- Identify distinct splice variants
- Assembling a reference genome

### Don't use long reads

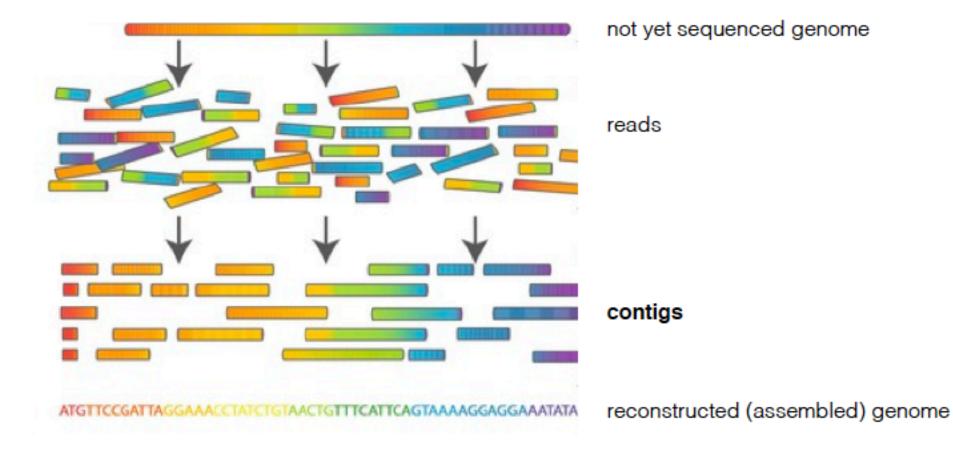
- When nucleotide identity is more important than linkage
- Identify low-frequency SNVs
- Quantify gene expression
- Re-sequencing in populations (for now)

• Genomic and transcriptomic applications

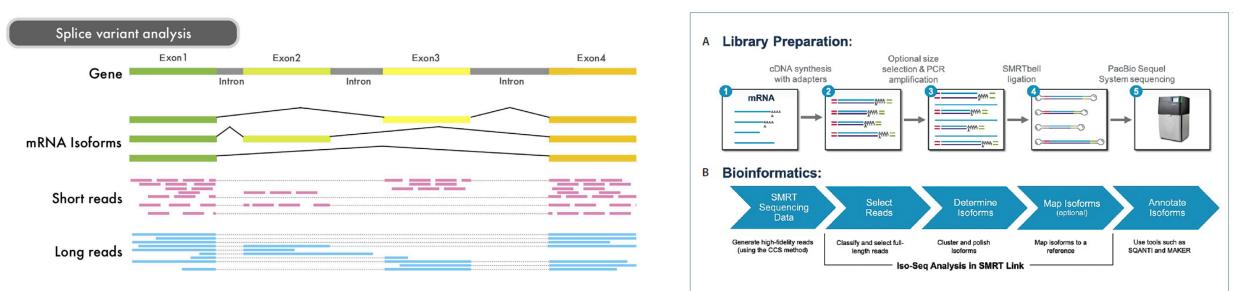
### Long reads can map reads uniquely in a reference



# Using long-read overlaps to perform **de novo assembly**



# Isoform profiling with long reads removes the assembly step



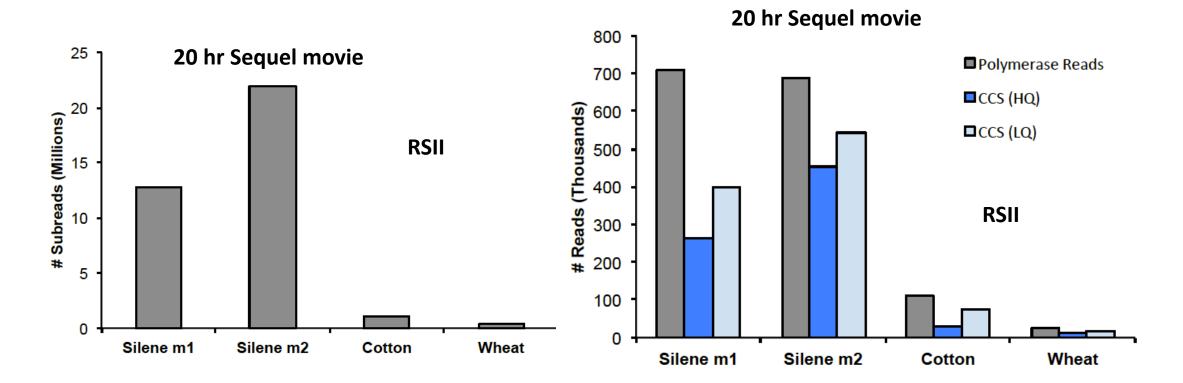
Nanopore RNA sequencing

PacBio IsoSeq

## PacBio Iso-Seq Transcriptomics



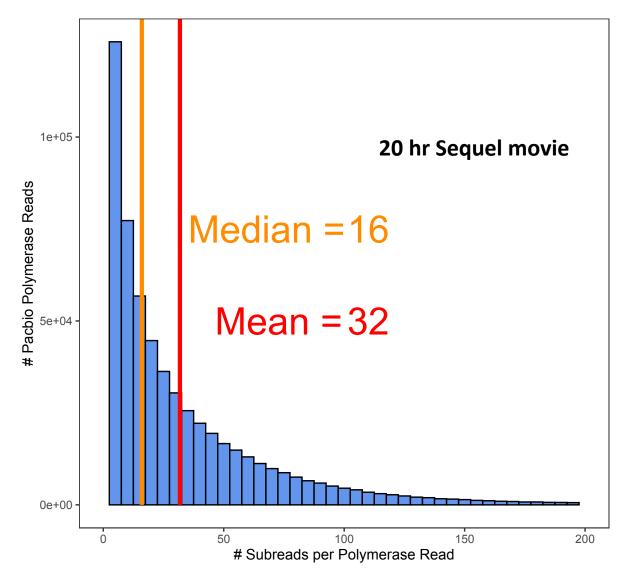
Polymerase reading the (subread + adapter) 3x = 1 CCS read



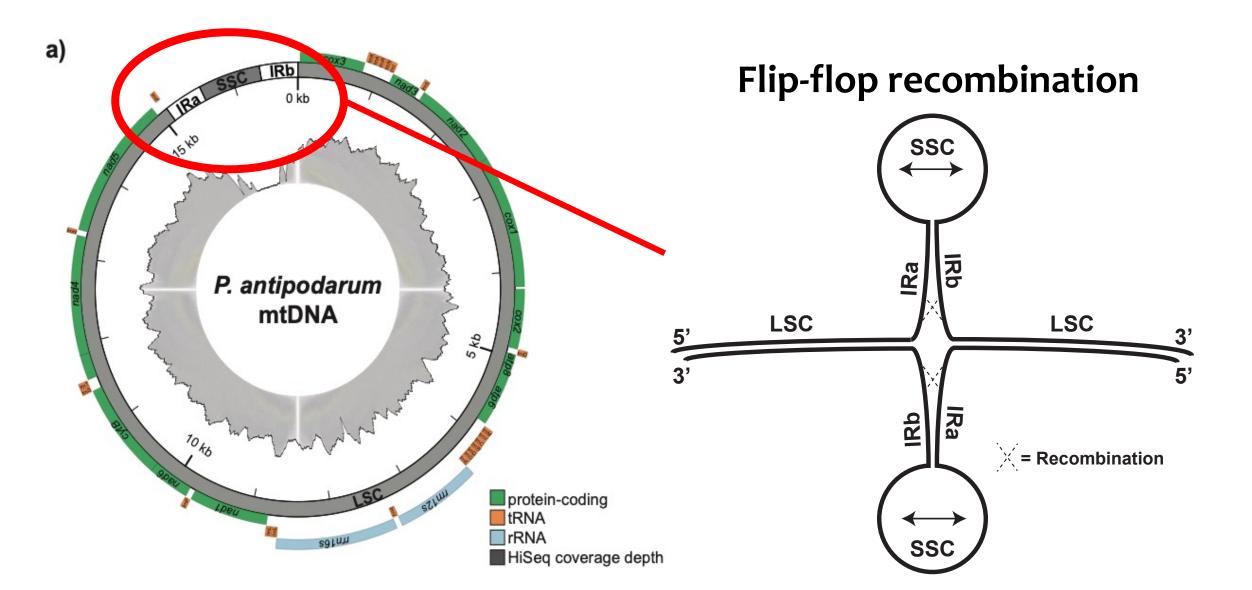
## PacBio Iso-Seq Transcriptomics

Base modifications impede polymerase processivity in a predictable manner

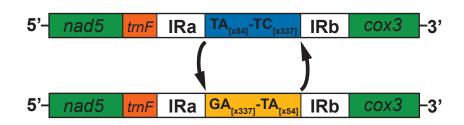
Can be measured with Inter-pulse Distance (IPD)



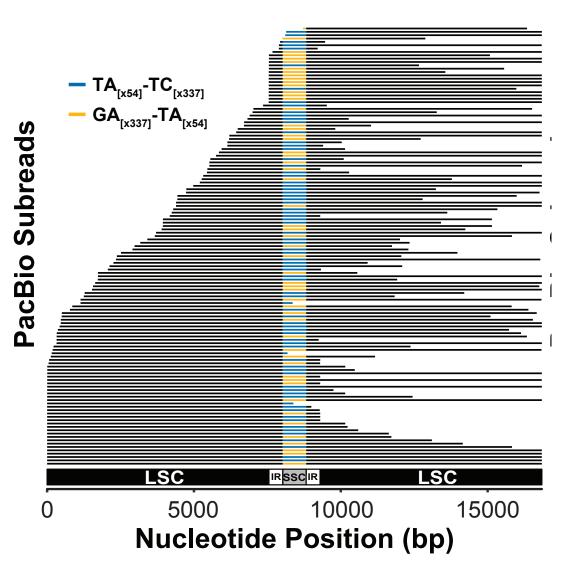
### **Resolving complex genomic features**



### **Resolving complex genomic features**



Long reads can identify structural variants



- How do long-read sequencing technologies work?
- When is long-read sequencing the right/wrong choice?
- Genomic and transcriptomic applications