Long-read: assets and challenges of a (not so) emerging technology

Summary

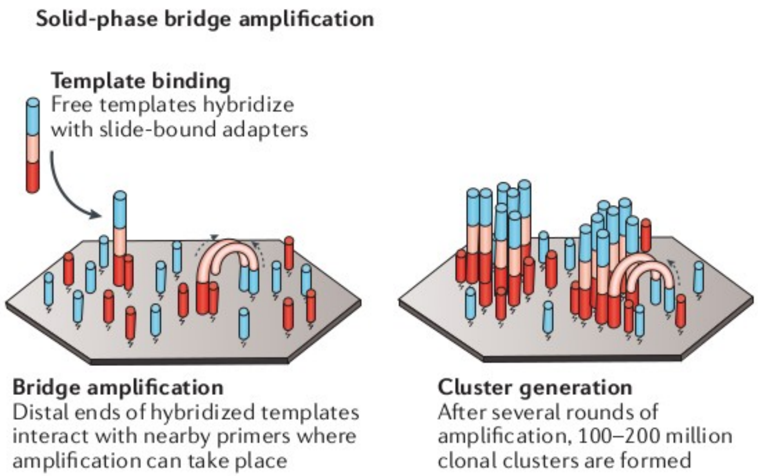
1. Second generation sequencing
2. Long-read technology
3. Error rates & error correction
4. Alternative splicing & isoforms

Second generation sequencing

- Sequencing by synthesis
  - Cyclic reversible termination (Illumina)
    - GeneReader (Qiagen)
  - Single-nucleotide addition
    - 454 pyrosequencing (Roche)
    - IonTorrent (ThermoFisher)
- Sequencing by ligation
  - SOLiD (ThermoFisher)
  - Complete Genomics (Beijing Genomics Institute)

Second generation sequencing
Sequencing by synthesis (Illumina)

1. Amplification of fragments
2. Addition of four types of reversible terminator bases
3. Removal of non-incorporated nucleotides
4. Imaging of the fluorescently labeled nucleotides
5. Removal of dye and terminal 3' blocker
6. New cycle



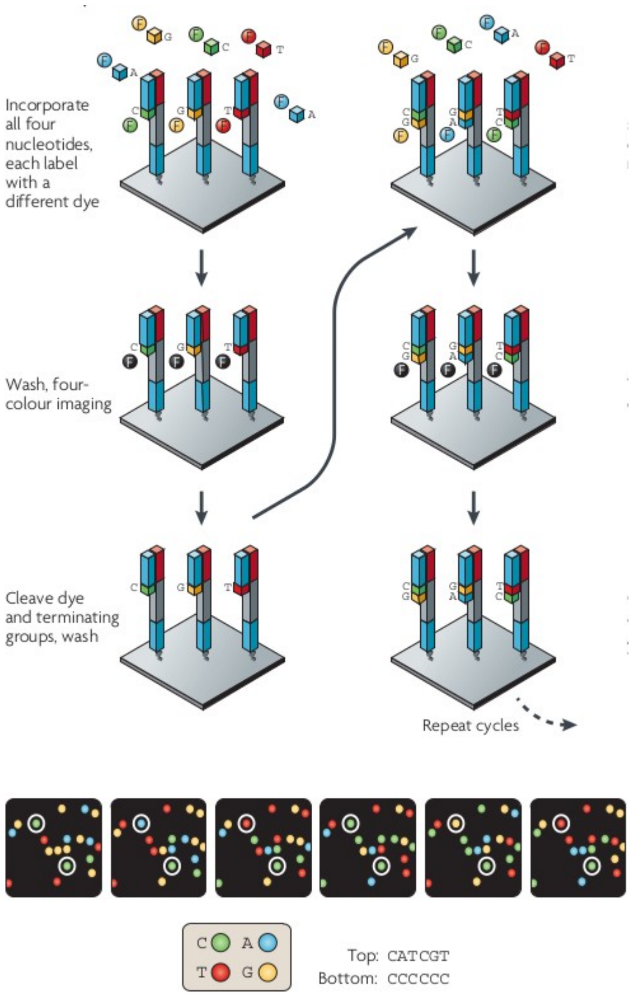
Solid-phase bridge amplification

Template binding
Free templates hybridize with slide-bound adapters

Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place

Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed

(a)



Incorporate all four nucleotides, each label with a different dye

Wash, four-colour imaging

Cleave dye and terminating groups, wash

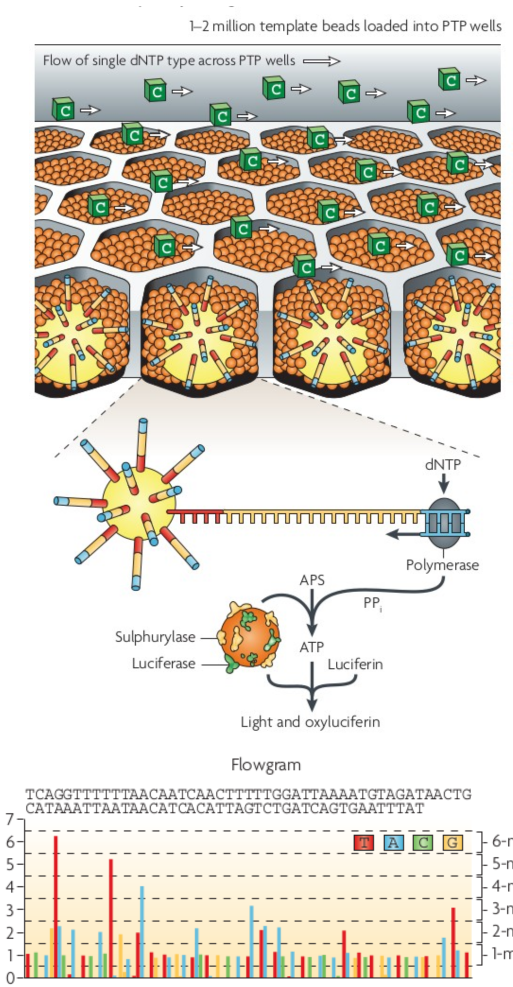
Repeat cycles

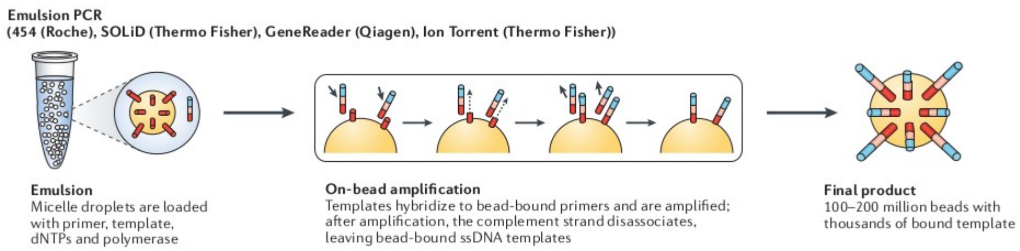
(b)

(a) Goodwin et al., 2016
(b) Metzker, 2010

Second generation sequencing
Sequencing by synthesis (454 pyrosequencing)

1. Amplification PCR emulsion
2. Beads are deposited in wells
3. Addition of a single type of dNTP
4. Emissionof pyrophosphate if dNTP is incorporated
5. Production of luciferase
6. Light is detected and recorded in a flowgram
7. New cycle with another dNTP


1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
C
dNTP
Polymerase
APS
PP$_i$
Sulphurylase
Luciferase
ATP
Luciferin
Light and oxyluciferin
Flowgram
TCAGGTTTTTTAACAATCAACTTTTTGGATTAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT
7
6
5
4
3
2
1
0
T
A
C
G
6-mer
5-mer
4-mer
3-mer
2-mer
1-mer


Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))
Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase
On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates
Final product
100–200 million beads with thousands of bound template

(a)

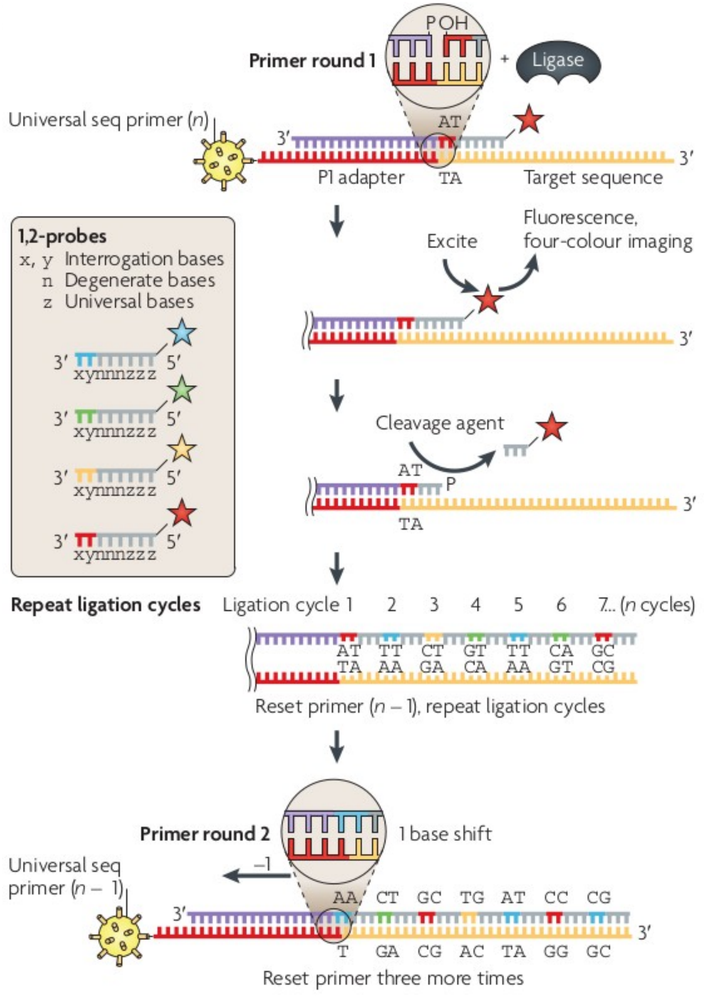
(b)

(a) Goodwin et al., 2016
(b) Metzker, 2010

Second generation sequencing
Sequencing by ligation (SOLiD)

1. Target molecule to be sequenced: single strand of unknown DNA sequence
2. Flanked on at least one end by a known sequence.
3. Addition of short "anchor" strand to bind the known sequence
4. Addition of mixed pool of labeled probe oligonucleotides
5. DNA ligase preferentially joins the molecule to the anchor when bases match the target
6. Cycle repeated
7. Anchor is shifted

...



(a)

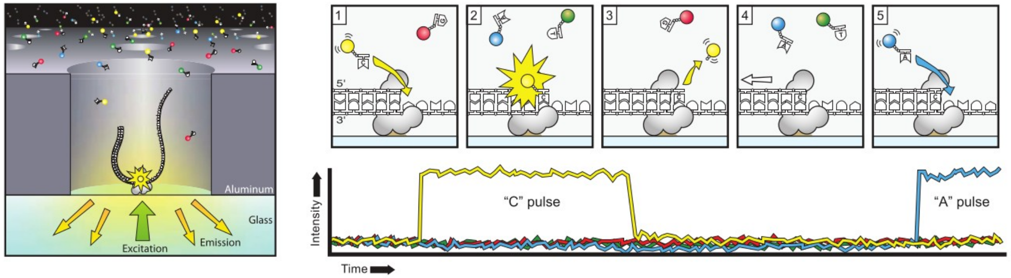
(a) Metzker, 2010

Long-read technology: characteristics

- Single-molecule real-time sequencing (Eid et al., 2009)
  - Main techno:
    - PacBio
    - Oxford Nanopore
  - No library prep, no amplification
  - Requires expensive equipment
- Synthetic long reads (McCoy et al., 2014)
  - Main techno:
    - Illumina
    - 10X Genomics
  - Relies on typical short read sequencing
  - No new equipment required
  - Long reads are constructed in silico using barcodes

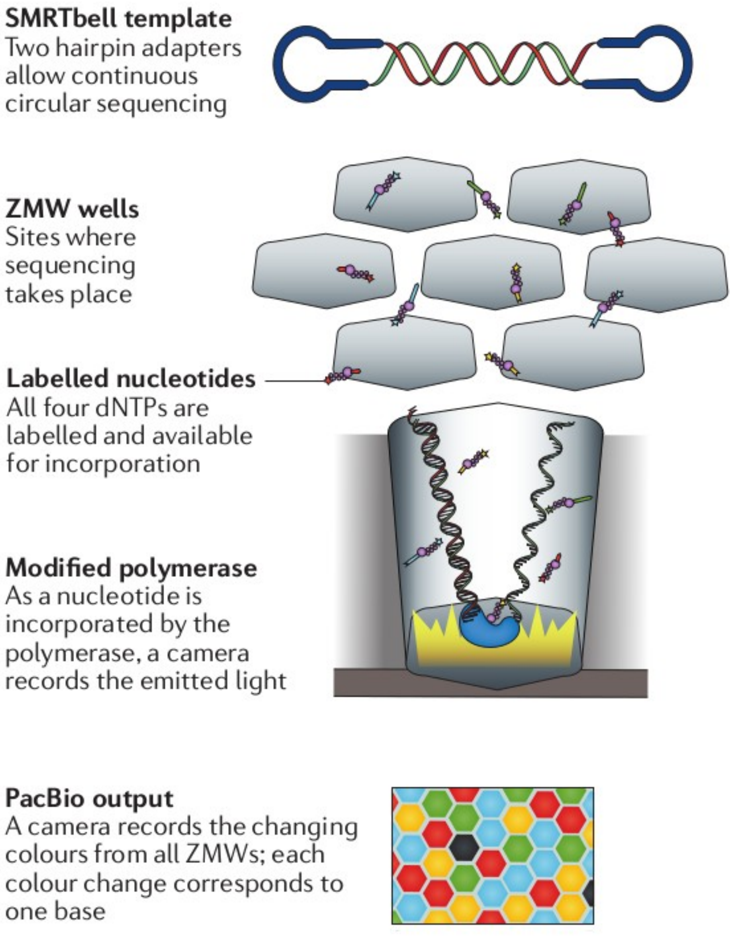
Long-read technology
Single-Molecule Real-Time sequencing

1. Flowcells made of zero-mode waveguides (ZMW) anchored on a glass substrate
2. Polymerase fixed at the bottom of well/waveguide, hence the **single-molecule** focus
3. dNTP incorporation visualized continuously by laser
4. Labelled nucleotide pauses during incorporation, fluorophore is removed
5. Circular templates allow several passes for a single target sequence



(a)

(b)



SMRTbell template
Two hairpin adapters allow continuous circular sequencing

ZMW wells
Sites where sequencing takes place

Labelled nucleotides
All four dNTPs are labelled and available for incorporation

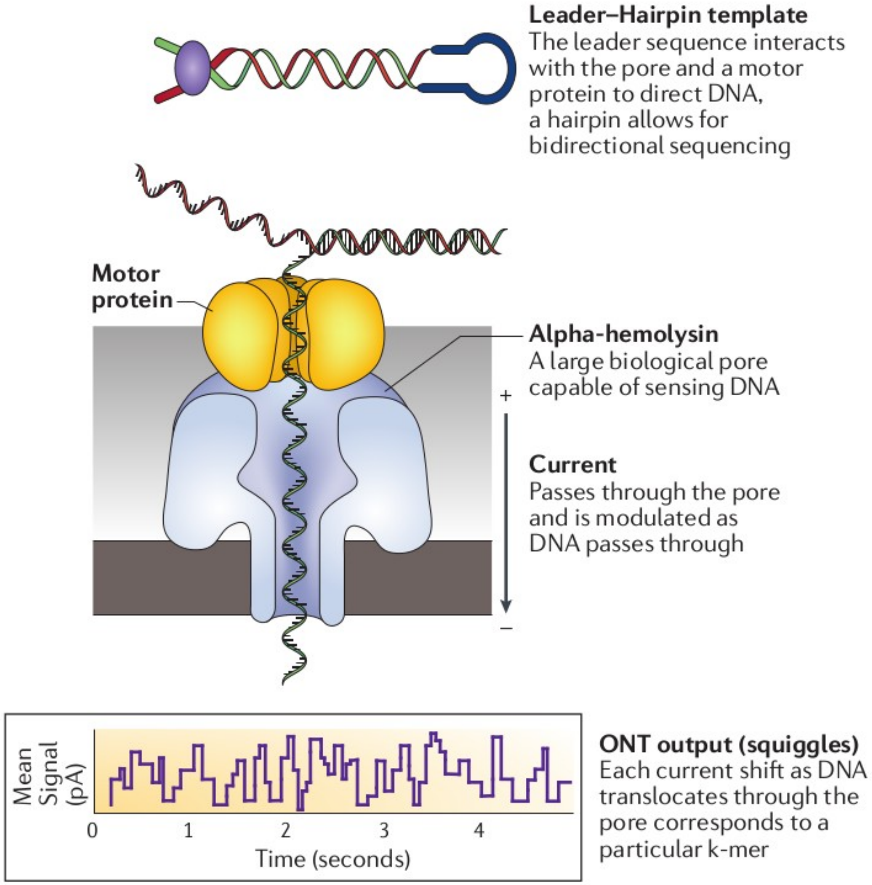
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light

PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base

(a) Eid et al., 2009
(b) Goodwin et al., 2016

Long-read technology
Single-Molecule Real-Time nanopore sequencing

1. Single-stranded DNA is passed through a pore thanks to a secondary motor protein
2. Current passes through the pore
3. The voltage shifts depending on the k-mers passing through
4. System called “squiggle space”
5. More than 1,000 possible levels of signal corresponding to as many different k-mers
6. Hairpin templates allow two passes, forward and reverse
7. Consensus sequence is computed



Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing

Alpha-hemolysin
A large biological pore capable of sensing DNA

Current
Passes through the pore and is modulated as DNA passes through

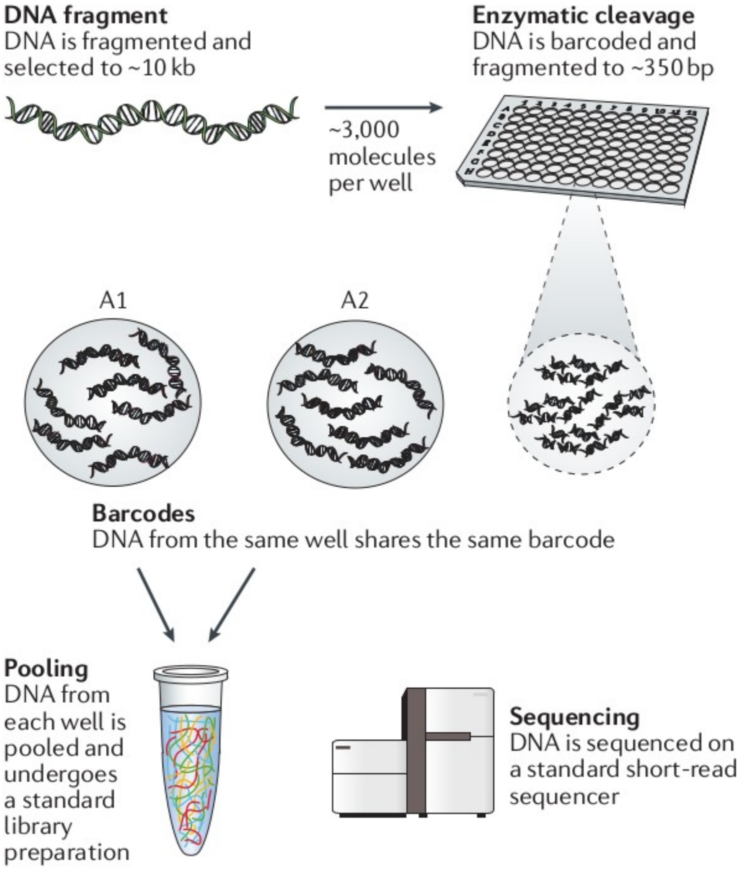
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

(a)

(a) Goodwin et al., 2016

Long-read technology
Synthetic long-reads (Illumina)

1. Large DNA fragments are partitioned into wells
2. Within each well, fragments are sheared into short reads and barcoded
3. DNA from each well is pooled, and short reads are sequenced using standard library preparation and instrumentation
4. Resulting data is split according to barcodes and reassembled

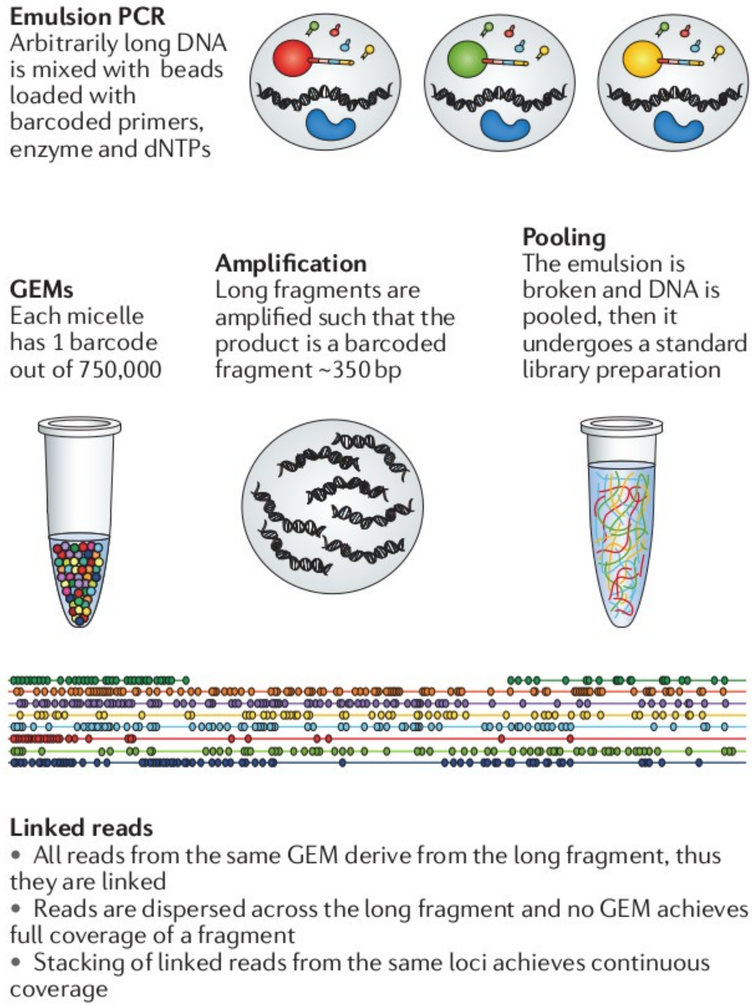

DNA fragment
DNA is fragmented and selected to ~10 kb
~3,000 molecules per well
Enzymatic cleavage
DNA is barcoded and fragmented to ~350 bp
A1
A2
Barcodes
DNA from the same well shares the same barcode
Pooling
DNA from each well is pooled and undergoes a standard library preparation
Sequencing
DNA is sequenced on a standard short-read sequencer

(a)

(a) Goodwin et al., 2016

Long-read technology
Synthetic long-reads (10X Genomics)

1. Large fragments of DNA partitioned into micelles called GEMs using emulsion
2. Each GEM has its own **barcode**
3. Each large fragment is amplified into smaller fragments
4. DNA is pooled and sequenced
5. Reads are aligned and **linked** together
6. Alignment doesn't have to be continuous
7. Coverage is achieved by using a so-called “read cloud”



Emulsion PCR
Arbitrarily long DNA is mixed with beads loaded with barcoded primers, enzyme and dNTPs

GEMs
Each micelle has 1 barcode out of 750,000

Amplification
Long fragments are amplified such that the product is a barcoded fragment ~350 bp

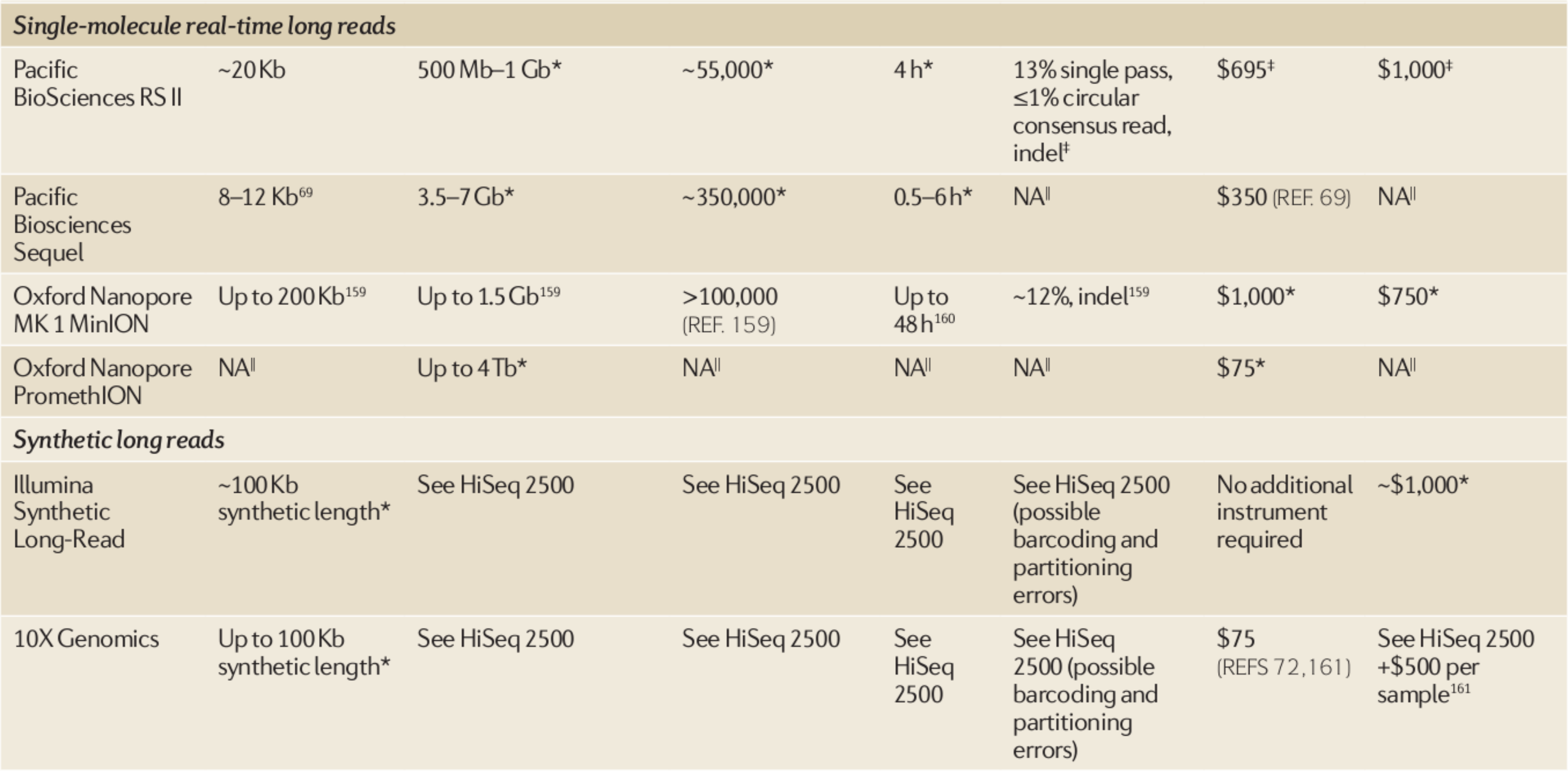
Pooling
The emulsion is broken and DNA is pooled, then it undergoes a standard library preparation

Linked reads
- All reads from the same GEM derive from the long fragment, thus they are linked
- Reads are dispersed across the long fragment and no GEM achieves full coverage of a fragment
- Stacking of linked reads from the same loci achieves continuous coverage

(a)

(a) Goodwin et al., 2016

Long-read technology: summary



Single-molecule real-time long reads							
Pacific BioSciences RS II	~20 Kb	500 Mb–1 Gb*	~55,000*	4 h*	13% single pass, ≤1% circular consensus read, indel‡	$695‡	$1,000‡
Pacific Biosciences Sequel	8–12 Kb[69]	3.5–7 Gb*	~350,000*	0.5–6 h*	NA‖	$350 (REF. 69)	NA‖
Oxford Nanopore MK 1 MinION	Up to 200 Kb[159]	Up to 1.5 Gb[159]	>100,000 (REF. 159)	Up to 48 h[160]	~12%, indel[159]	$1,000*	$750*
Oxford Nanopore PromethION	NA‖	Up to 4 Tb*	NA‖	NA‖	NA‖	$75*	NA‖
Synthetic long reads							
Illumina Synthetic Long-Read	~100 Kb synthetic length*	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500 (possible barcoding and partitioning errors)	No additional instrument required	~$1,000*
10X Genomics	Up to 100 Kb synthetic length*	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500 (possible barcoding and partitioning errors)	$75 (REFS 72,161)	See HiSeq 2500 +$500 per sample[161]

Goodwin et al., 2016

Long-read technology: SMRT

- PacBio RS II (most widely used)
  - Average read length ~ 10-15 kb
  - Single-pass error rate ~ 15%
  - Mostly indel
  - Random distribution → overcome with higher coverage
  - Limited throughput, high cost
- MinION
  - Small, USB-based device (+ library prep equipment)
  - Single-pass error rate up to 30%
  - Base-calling algorithms have improved accuracy recently

Long-read technology: synthetic reads

- Illumina
  - Relies on standard equipment > affordable
  - Throughput and error profiles similar to those of standard techno
  - Requires more coverage due to additional level of partitioning
- 10X Genomics
  - Additional but affordable equipment
  - Works with as little as 1ng of starting material
  - Inefficient DNA partitioning / limited number of barcodes

Long-read technology: applications

- *De novo* assembly
- Sequencing of "challenging" genomes (repetitive regions...)
- Genome finishing
- Genome phasing
  - Analyze compound heterozygotes
  - Measure allele-specific expression
  - Identify variant linkage
  - Phase de novo mutations

  ...

Error rates & correction

- Sequencing errors lead to weaker alignments
  - Mismatches
  - Shorter alignments
- Second generation sequencing > substitutions
- Long-read techno (SMRT) > **insertions**/deletions
  - Median accuracy of 99.3% with 15-fold coverage (Eid et al., 2009)
  - Accuracy 82.1%–84.6% (Koren et al., 2013)
  - Error rate 15% (Salmela et Rivals, 2014)
  - Error rate 13% single pass, <1% circular template (Goodwin et al., 2016)

➢ Errors are considered unbiased and uniformly distributed

Error rates & correction

- Self correction (eg. HGAP)
  - Computing local alignment between long reads
  - Building multiple alignments
  - Calling a consensus sequence
- Hybrid correction (eg. AHA)
  - Aligning short reads on long reads
  - Correcting long reads using short reads' better accuracy
  - Computationally costly

Error rates & correction

- Spectral alignment-based methods (2nd gen.)
  - “With a sufficient coverage, it is possible to compute a minimal threshold such that, with high probability, each error-free k-mer appears at least that number of times in the read set.” (Salmela et Rivals, 2014)
  - A k-mer above/below the threshold is qualified as solid or weak, respectively
  - A de Bruijn graph (DBG) can be constructed using the solid k-mers as nodes

LoRDEC: long-read error correction

- Mix of the hybrid and the spectral approaches
- Construction of a DBG using short-read data
- Correction of long reads by searching an optimal path in the graph

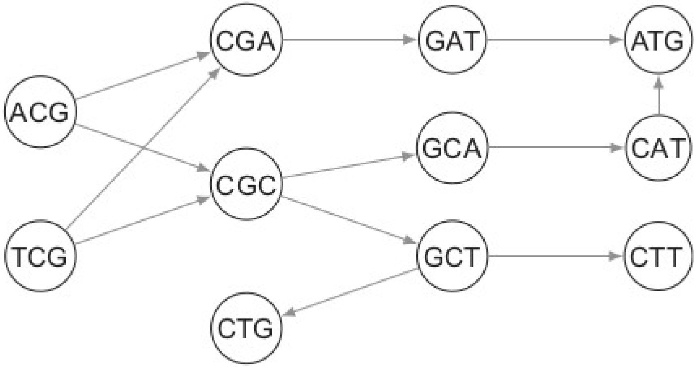


Fig. 1. An example of short read DBG of order $k = 3$. For simplicity reverse complement k-mers are ignored

Salmela et Rivals, 2014

- Any k-mer that occurs less than *s* times within the shorts reads is filtered out
- So-called “solid k-mers” are kept

LoRDEC: long-read error correction

- Long reads are partitioned into weak and solid regions according to the short read DBG
- Several pairs of source/target solid k-mers are investigated to find optimal path over weak region

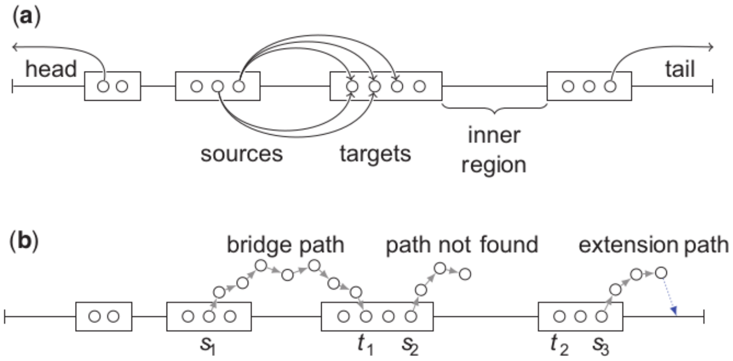


Fig. 2. Long read correction method. **(a)** A long read is partitioned into weak and solid regions (respectively, lines and rectangles) according to the short read DBG. Weak regions starting or ending the long read are called the *head* or the *tail*, respectively, while other weak regions are *inner regions*. Circles in solid regions represent *k*-mers of the DBG. *k*-mers around a weak region serve as source and target nodes to search paths in the DBG. Several source/target pairs are used for each weak inner region. (b) On the second inner region, a *bridging path* between nodes s_1 and t_1 is found in the DBG to correct this region. On the third region, the path search fails to find a path between nodes s_2 and t_2. For the tail, an *extension path* is sought and found from node s_3 toward the end. Once found, the corrective sequence of the path is aligned to the tail to determine the optimal substring (thick dotted arrow)

Salmela et Rivals, 2014

LoRDEC: long-read error correction

- Effects of parameters on accuracy of correction
  - Sensitivity = TP/(TP + FN)
    - How well does the tool recognize erroneous positions?
  - Gain = (TP – FP)/(TP + FN)
    - How well does the tool remove errors without introducing new ones?

LoRDEC: long-read error correction

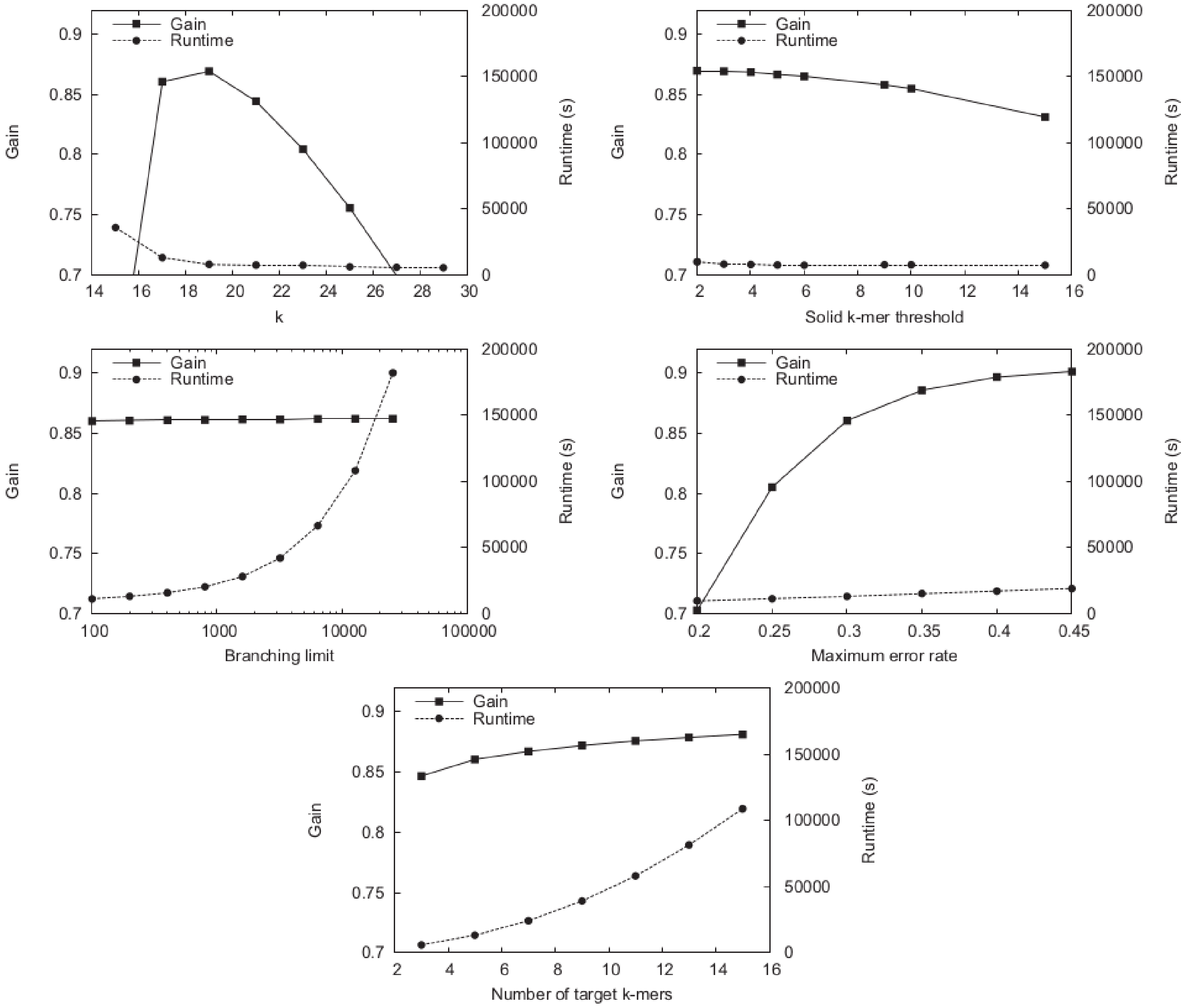
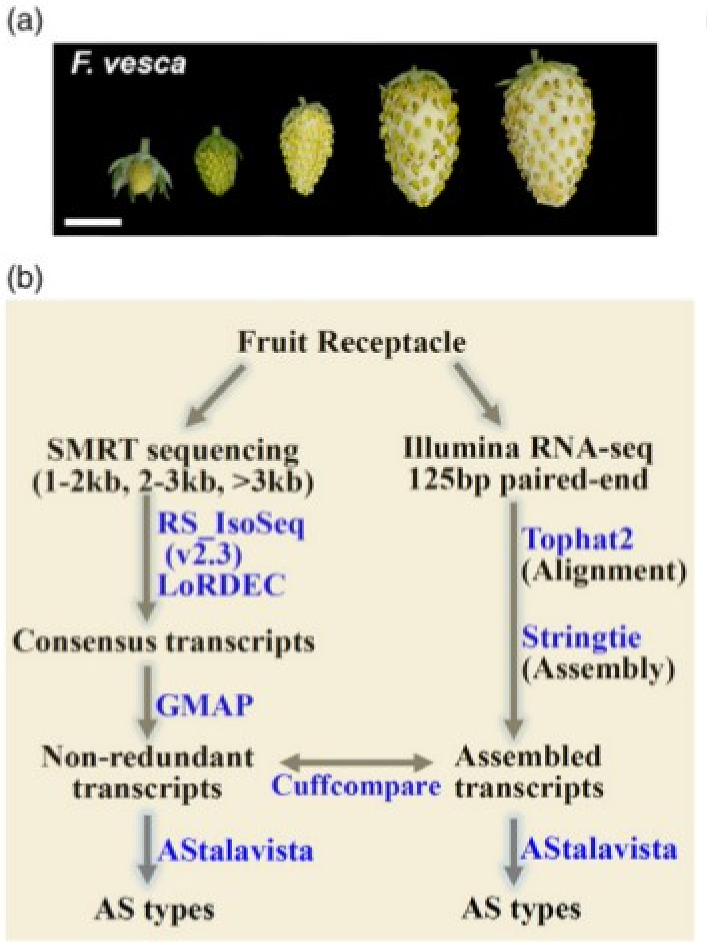


Fig. 3. Effect of parameters on the runtime and gain of our method. We varied *k*, solid *k*-mer threshold, branching limit, maximum error rate and number of target *k*-mers one at a time, while other parameters were kept constant

Salmela et Rivals, 2014

Comparative study of LR and SR

- Comparative study of alternative splicing in strawberry develoment using SMRT sequencing & Illumina short reads
- After filtering and correction (LoRDEC) of SMRT data, 96.4% of consensus transcripts could be aligned to the genome (GMAP)
- Removal of redundant transcripts
- 33,236 full-length isoforms/transcripts
  - 26,737 known transcripts
  - 5,501 novel transcripts



Fruit Receptacle

SMRT sequencing (1-2kb, 2-3kb, >3kb)

RS_IsoSeq (v2.3)

LoRDEC

Consensus transcripts

GMAP

Non-redundant transcripts

AStalavista

AS types

Illumina RNA-seq 125bp paired-end

Tophat2 (Alignment)

Stringtie (Assembly)

Assembled transcripts

Cuffcompare

AStalavista

AS types

Li et al., 2016

Comparative study of LR and SR

- Long-read data (PacBio SMRT)
  - novel transcripts are shorter on average
  - new introns are longer than previously annotated introns
  - more isoforms identified
- Short-read data (Illumina)
  - more annotated genes are found
  - more novel genes are discovered
- Distribution of splicing junctions and splice sites are similar

Quality assessment of LR data

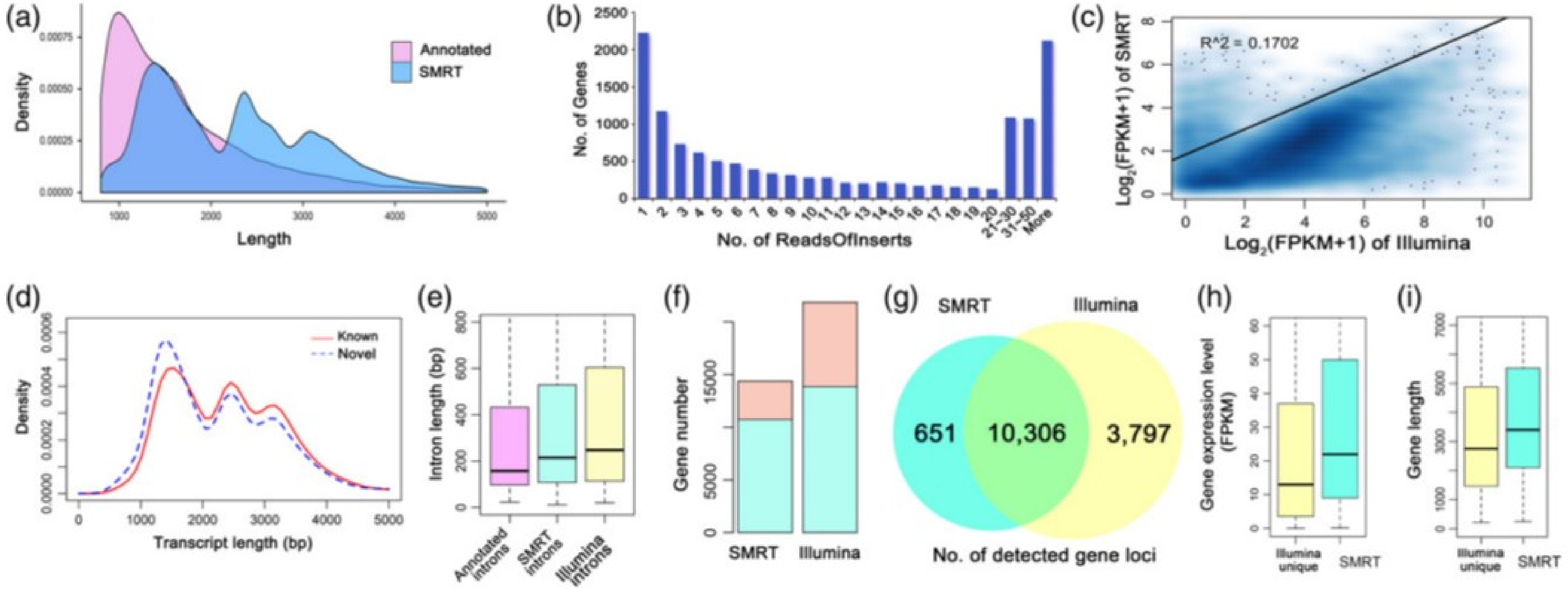


Figure 2. Quality assessment of the SMRT data.

(a) Density plot of the length of all previously annotated genes and all SMRT ReadsOfInserts.

(b) Bar graph showing the number of annotated genes (*y*-axis) with different number of aligned SMRT ReadsOfInserts (*x*-axis).

(c) Scatterplot showing the expression level of each gene estimated by the SMRT and Illumina-based RNA-seq respectively in the fruit receptacle of *F. vesca*. *X*-axis is the log2-transformed FPKM derived from Illumina. *Y*-axis is the log2-transformed FPKM calculated from the aligned SMRT ReadOfInserts. R^2 is calculated by the lm() function in R. From (a) to (c), the 797 718 SMRT ReadOfInserts obtained from relaxed filtering were used (Table 1).

(d) Density distribution of transcript length in both known and novel loci obtained by SMRT sequencing. Known loci are those with class code '=' or 'j', and novel loci are those with the other class codes in the *Cuffcompare* output file.

(e) Boxplot showing the length of annotated introns and new introns uncovered by current SMRT and Illumina sequencing.

(f) The number of genes detected by SMRT and Illumina. 'Novel' (pink) indicates new genes not previously annotated; 'Annotation' (green) indicates previously annotated genes from the *F. vesca* genome version 2.0.a1.

(g) Venn diagram showing the common and unique annotated genes detected by SMRT and Illumina.

(h) Boxplot showing the expression level of the unique genes [yellow region in (g)) in Illumina and all the genes in SMRT (blue circle in (g)].

(i) Boxplot showing the length of unique genes in Illumina and all the genes in SMRT (same as in (h)). From (d) to (i), the SMRT ReadOfInserts obtained from stringent filtering were used (Table 1).

Li et al., 2016

Comparison of SMRT and Illumina in AS events detection

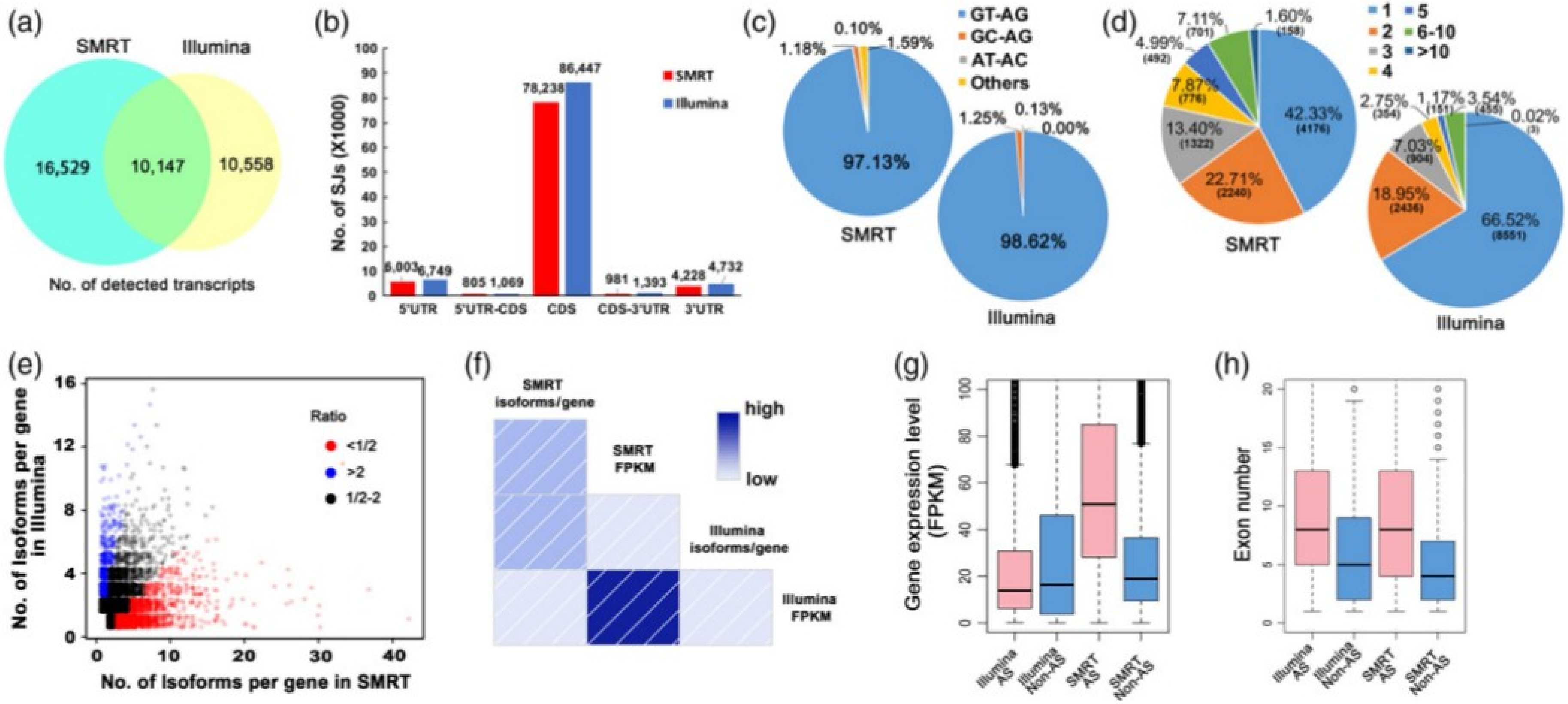


Figure 3. Detailed comparisons of alternative splicing revealed by SMRT- and Illumina-based RNA-seq.
(a) Venn diagram showing the common and unique transcripts from the annotated loci identified by each method.
(b) Distribution of splicing junctions along gene features in the annotated loci. SJ: Splicing Junction.
(c) Pie chart showing the percentage of the splicing donor-acceptor di-nucleotide utilization among all transcripts in the two datasets.
(d) Pie chart showing the number and percentage of multiexon genes with different number of isoforms. Color code indicates the number of isoforms. '1' isoform means no AS.
(e) Scatterplot showing the number of isoforms per annotated gene in Illumina and SMRT.
(f) Corrgram showing the correlation coefficients between the isoform number and FPKM derived from SMRT or Illumina.
(g) Boxplot showing the expression levels of AS and non-AS genes in Illumina and SMRT.
(h) Boxplot showing the exon numbers of AS and non-AS genes in Illumina and SMRT.

Li et al., 2016

Comparison of SMRT and Illumina in AS events detection

- Identification of genes undergoing alternative splicing
  - ➢ Illumina: 33.48%
  - ➢ SMRT: 57.67%
- ➢ Only a few genes have more than 30 isoforms
- ➢ AS events have different profiles depending on tissue & stage of development

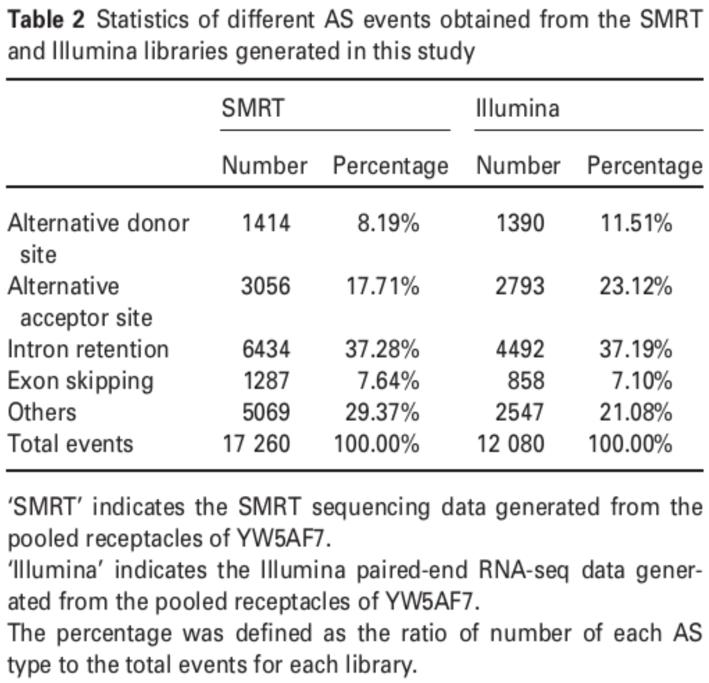
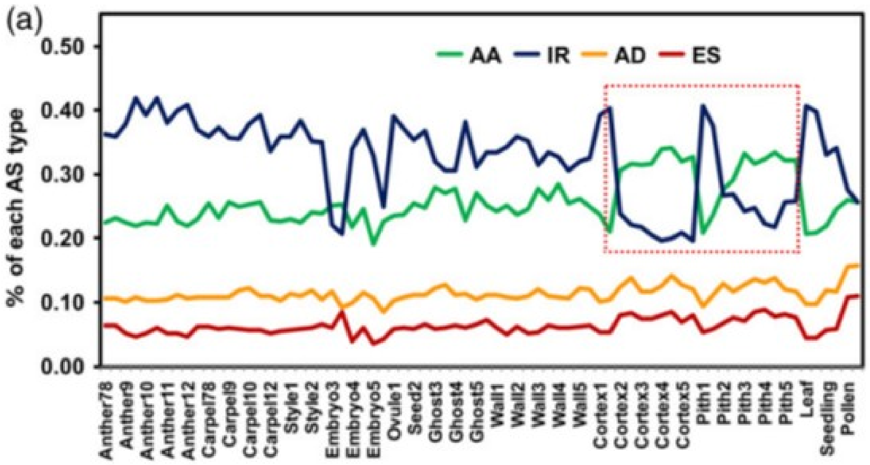


Table 2 Statistics of different AS events obtained from the SMRT and Illumina libraries generated in this study

	SMRT		Illumina	
	Number	Percentage	Number	Percentage
Alternative donor site	1414	8.19%	1390	11.51%
Alternative acceptor site	3056	17.71%	2793	23.12%
Intron retention	6434	37.28%	4492	37.19%
Exon skipping	1287	7.64%	858	7.10%
Others	5069	29.37%	2547	21.08%
Total events	17 260	100.00%	12 080	100.00%

'SMRT' indicates the SMRT sequencing data generated from the pooled receptacles of YW5AF7.
'Illumina' indicates the Illumina paired-end RNA-seq data generated from the pooled receptacles of YW5AF7.
The percentage was defined as the ratio of number of each AS type to the total events for each library.

Li et al., 2016

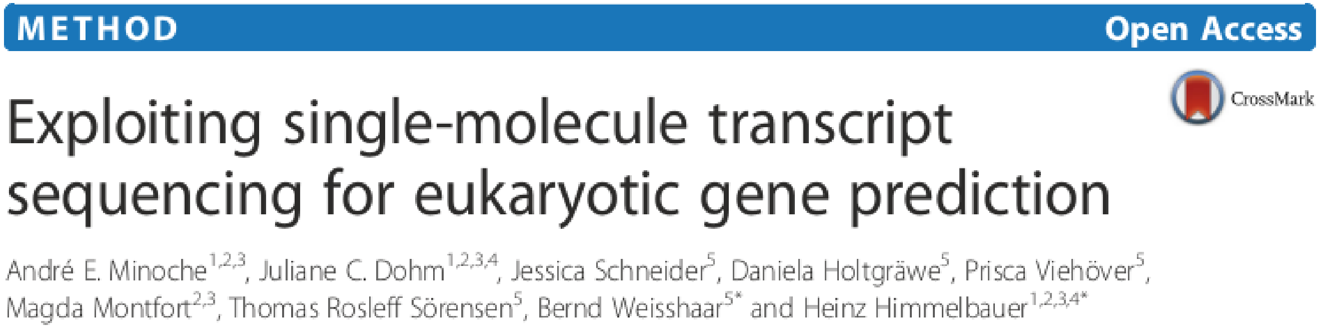


References

- Eid et al., 2009 - Real-Time DNA Sequencing from Single Polymerase Molecules. Science 02 Jan 2009: Vol. 323, Issue 5910, pp. 133-138
- Goodwin et al., 2016 - Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17, 333–351 (2016)
- Koren et al., 2013 - Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012 Jul; 30(7): 693–700
- Li et al., 2016 - Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. Plant J, 90: 164–176. doi:10.1111/tpj.13462
- McCoy et al., 2014 - Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. PLoS ONE 9(9): e106689.
- Metzker, 2010 - Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46
- Salmela et Rivals, 2014 - LoRDEC: accurate and efficient long read error correction. Bioinformatics (2014) 30 (24): 3506-3514
- Stöcker, Köster et Rahmann, 2016 - SimLoRD: Simulation of Long Read Data. Bioinformatics (2016) 32 (17): 2704-2706

Thank you!

Other work related to long-read techno



METHOD Open Access

CrossMark

Exploiting single-molecule transcript sequencing for eukaryotic gene prediction

André E. Minoche[1,2,3], Juliane C. Dohm[1,2,3,4], Jessica Schneider[5], Daniela Holtgräwe[5], Prisca Viehöver[5], Magda Montfort[2,3], Thomas Rosleff Sörensen[5], Bernd Weisshaar[5*] and Heinz Himmelbauer[1,2,3,4*]

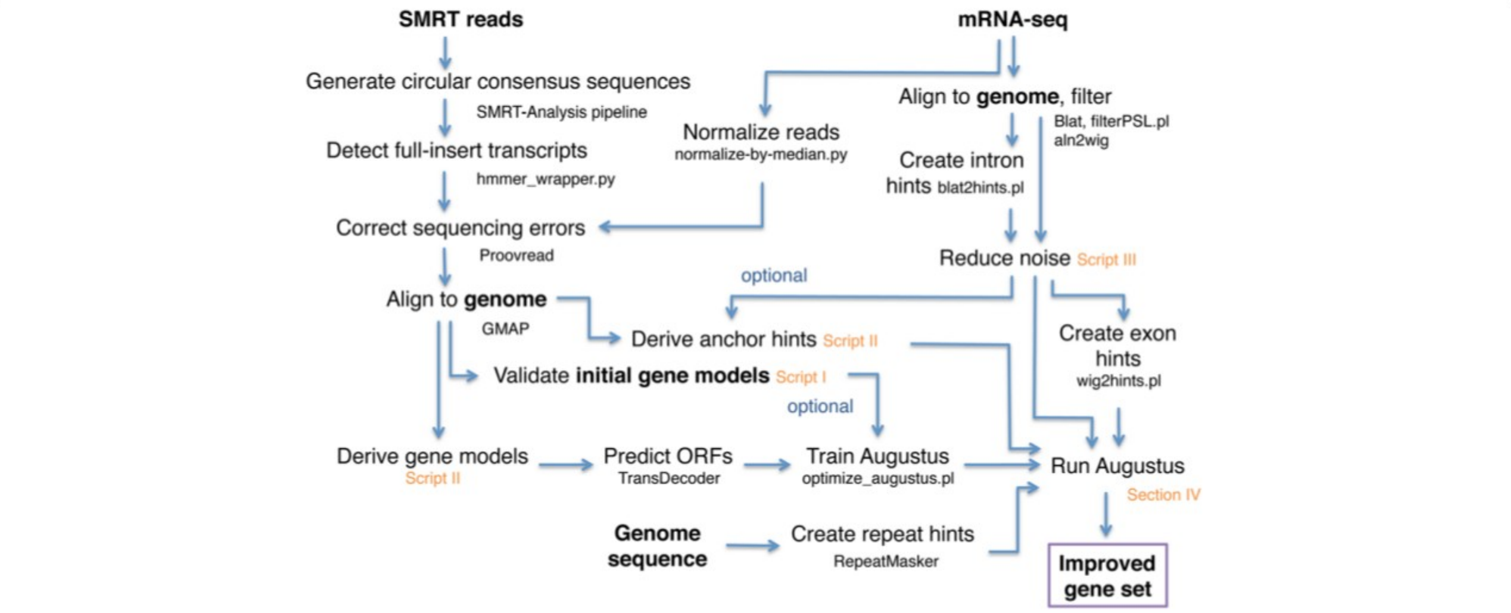
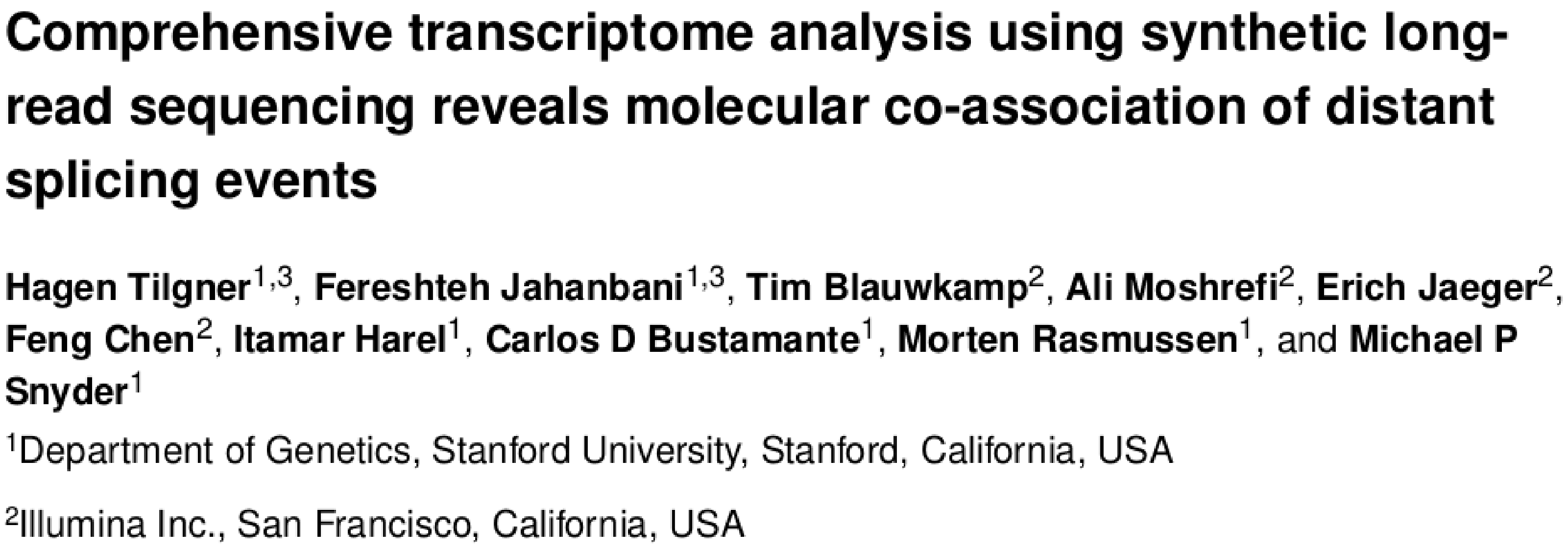


Fig. 6 Workflow of our analyses to improve eukaryotic gene predictions, including the scripts that are part of this publication (highlighted orange). Input and output data are highlighted in bold lettering

Other work related to long-read techno

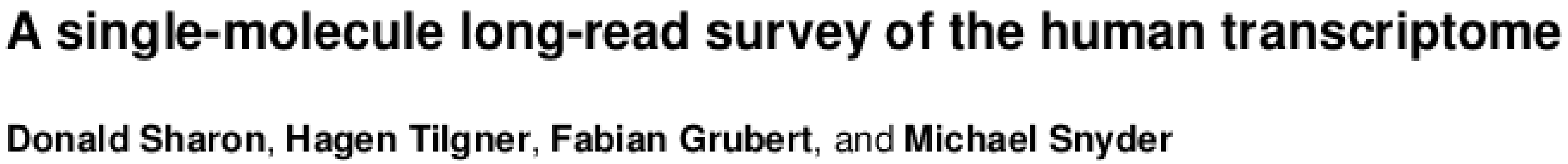


Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events

Hagen Tilgner[1,3], **Fereshteh Jahanbani**[1,3], **Tim Blauwkamp**[2], **Ali Moshrefi**[2], **Erich Jaeger**[2], **Feng Chen**[2], **Itamar Harel**[1], **Carlos D Bustamante**[1], **Morten Rasmussen**[1], and **Michael P Snyder**[1]

[1]Department of Genetics, Stanford University, Stanford, California, USA

[2]Illumina Inc., San Francisco, California, USA



A single-molecule long-read survey of the human transcriptome

Donald Sharon, **Hagen Tilgner**, **Fabian Grubert**, and **Michael Snyder**