Low-input Workflow for PacBio *De Novo* Genome Assemblies

ACKNOWLEDGMENTS

- Sanger Institute, Cambridge, UK:
  - Mara Lawniczak, Juliana Cudini
- PacBio:
  - Christine Lambert, Primo Baybayan, Sarah Kingan

LOW-INPUT INSECT ASSEMBLY PROJECT

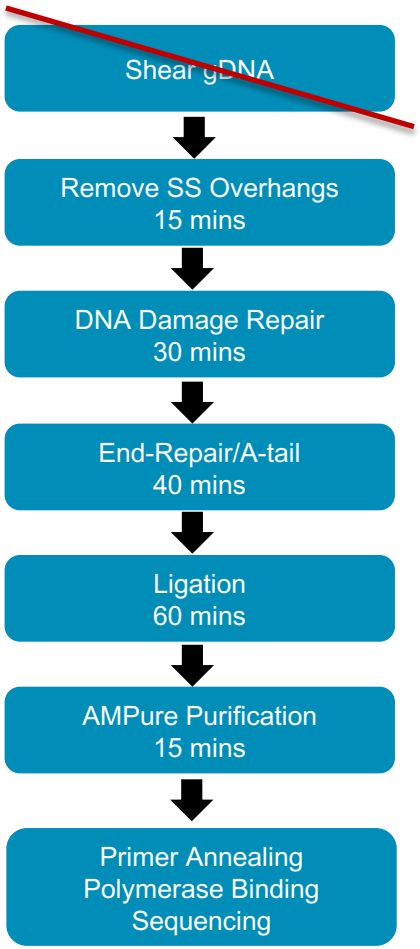
Goal: Obtain high-quality assembly from single insect individual

- Material:
  - Took one half of a single mosquito (*Anopheles spp*) for DNA extraction
  - 100 ng starting DNA amount

- Improved workflow for low-input samples:
  - Eliminate shearing step
  - Using new Express Prep v2, yield >60%
  - Using new Sequel chemistry (v3.0) & software (v6.0)
  - Ran 3 SMRT Cells, yielding total of 72.7 Gb

- Falcon-Unzip *de novo* assembly:
  - Contig N50: 3.5 Mb
  - Completeness: 98% Complete BUSCOs
  - Consensus accuracy: >99.99%
  - ~30% of genome separated into the two haplotypes

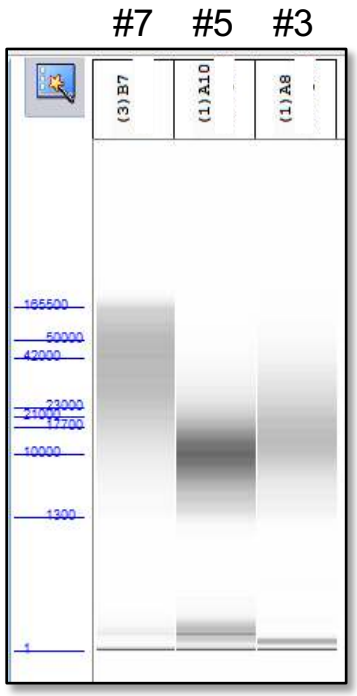
LIBRARY PREP USING EXPRESS PREP V2



- Works well with partially fragmented gDNA

- Single-tube, “addition only” workflow
- Eliminates Buffer exchanges
- No DNA loss

- 1 x 0.45x AMPure purification



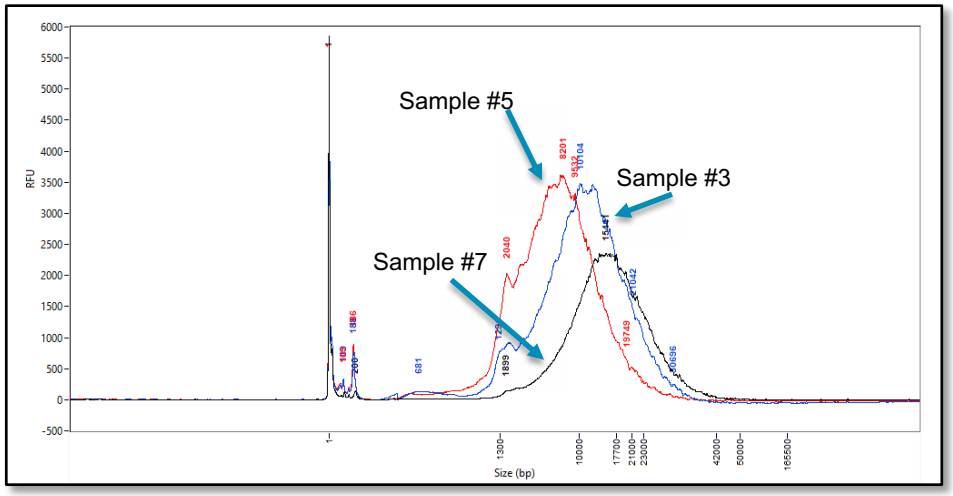
Initial DNA fragment size distribution

Express Prep V2 will be available in the near future

LIBRARY YIELD AND ESTIMATED # SMRT CELLS

Sample	Library Prep Kit	DNA input into Exo	AMPure	Final Yield (ng)	% Library Yield	Estimated SMRT Cells*	Notes
Control	Express v2	112 ng	0.45X +0.8X	60	53	8	Good Loading
Mosquito #7	**Express v2**	**112 ng**	**0.45X +0.8X**	**66**	**59**	**10**	**Good Loading**
Mosquito #5	Express v2	100 ng	0.45X +0.8X	46.6	47	9	Good Loading
Mosquito #3	Express v2	100 ng	0.45X +0.8X	49.7	50	7	Good Loading

- *on-plate concentration
- No-size selection, just 0.45x AMPure



- Library #7: ~17.9 kb
- Library #5: ~9.1 kb
- Library #3: ~12.2 kb

SEQUEL RUNS PRIMARY ANALYSIS METRICS

Sample	Loading concentration	**Gb/Cell**	Pol RL	N50 pol RL	Subread length	**N50 Subread length**	P0	P1	P2
#7	5 pM	**24.1**	40290	116615	8185	**12978**	26.0%	60.1%	13.9%
	5 pM	**23.6**	40077	114807	8254	**13132**	27.1%	59.0%	14.0%
	6 pM	**25.0**	47177	122898	8012	**12751**	35.3%	53.1%	11.7%
#5	4 pM	8.4	38253	109347	4273	7706	55.7%	23.2%	21.1%
#3	4 pM	25.4	50775	121602	3750	6323	33.4%	51.0%	15.6%

- New Chemistry (+120 min pre-extension)
- 20-hour movies
- Diffusion loading

ASSEMBLY RESULTS

	3 Cells	2 Cells	1 Cell
Total bases	72.7 Gb	48.5 Gb	23.6 Gb
Total unique bases	12.8 Gb	8.31 Gb	4.46 Gb
Unique coverage	45 X	31 X	17 X
Assembly size	**271 Mb**	**265 Mb**	150 Mb
Number of contigs	**580**	**815**	3,290
Contig N50	**3.5 Mb**	**1.5 Mb**	0.066 Mb

COMPARISON TO PREVIOUS SHORT-READ ASSEMBLY

	Illumina	PacBio	Improvement
Assembly size	224 Mb	271 Mb	21% longer
No. contigs	27,063	580	47-fold
Contig N50	25 kb	3.5 Mb	140-fold
No. scaffolds	10,521	--	
Scaffold N50	4.4 Mb	--	
Gap length	15 Mb	0	

FALCON-UNZIP RESULTS

- 3 Cells:

	primary	haplotig
Number of contigs	372	665
Assembly size	266 Mb	78.5 Mb
Contig N50	3.5 Mb	0.223 Mb

30% unzipped

- 2 Cells:

	primary	haplotig
Number of contigs	603	888
Assembly size	262 Mb	65.1 Mb
Contig N50	1.6 Mb	0.106 Mb

25% unzipped

INITIAL QC RESULTS

BUSCO ('diptera' set):

- Primary assembly:

C:98.0%[**S:94.1%**,D:3.9%],F:0.9%,M:1.1%,n:2799

2745	Complete BUSCOs (C)
2635	Complete and single-copy BUSCOs (S)
110	Complete and duplicated BUSCOs (D)
25	Fragmented BUSCOs (F)
29	Missing BUSCOs (M)
2799	Total BUSCO groups searched

- Combined assembly:

C:98.1%[S:58.8%,**D:39.3%**],F:0.9%,M:1.0%,n:2799

2747	Complete BUSCOs (C)
1646	Complete and single-copy BUSCOs (S)
1101	Complete and duplicated BUSCOs (D)
24	Fragmented BUSCOs (F)
28	Missing BUSCOs (M)
2799	Total BUSCO groups searched

CEGMA[1]:

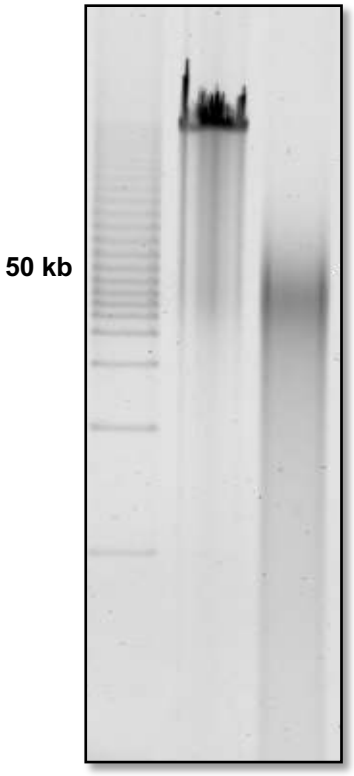
- 228 conserved gene alignments (102 kb sequence)
- 2 single-base frame-shift errors
- >99.99% consensus accuracy

[1] http://korflab.ucdavis.edu/Datasets/genome_completeness/index.html#SCT2 *(human)*

SUMMARY

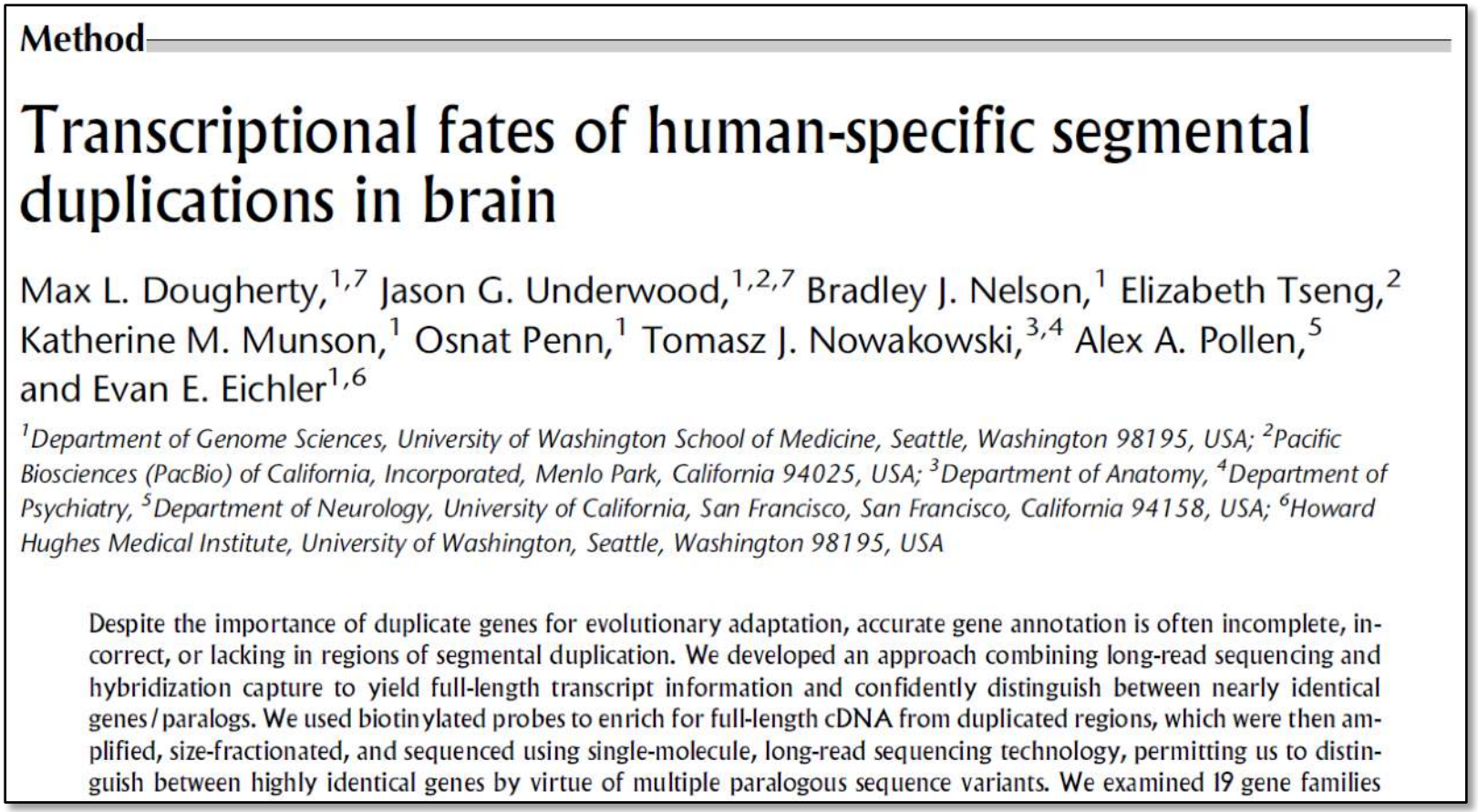
First high-quality *de novo* insect assembly from single individual

- DNA material
  - From a single mosquito
  - 100 ng starting amount
- Improved workflow with high yield for low-input samples
  - Eliminate shearing step
  - Using new Express Prep v2
  - Using new Sequel chemistry (v3.0) & software (v6.0)
  - Ran 3 SMRT Cells, yielding 72.7 Gb
- Falcon-Unzip *de novo* assembly
  - Contig N50: 3.5 Mb
  - BUSCO completeness: 98%
  - Consensus accuracy: >99.99%
  - ~30% of genome separated into the two haplotypes



NEW ISO-SEQ PAPER

Published yesterday in *Genome Research*



Method

Transcriptional fates of human-specific segmental duplications in brain

Max L. Dougherty,[1,7] Jason G. Underwood,[1,2,7] Bradley J. Nelson,[1] Elizabeth Tseng,[2] Katherine M. Munson,[1] Osnat Penn,[1] Tomasz J. Nowakowski,[3,4] Alex A. Pollen,[5] and Evan E. Eichler[1,6]

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA; [2]Pacific Biosciences (PacBio) of California, Incorporated, Menlo Park, California 94025, USA; [3]Department of Anatomy, [4]Department of Psychiatry, [5]Department of Neurology, University of California, San Francisco, San Francisco, California 94158, USA; [6]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA

Despite the importance of duplicate genes for evolutionary adaptation, accurate gene annotation is often incomplete, incorrect, or lacking in regions of segmental duplication. We developed an approach combining long-read sequencing and hybridization capture to yield full-length transcript information and confidently distinguish between nearly identical genes/paralogs. We used biotinylated probes to enrich for full-length cDNA from duplicated regions, which were then amplified, size-fractionated, and sequenced using single-molecule, long-read sequencing technology, permitting us to distinguish between highly identical genes by virtue of multiple paralogous sequence variants. We examined 19 gene families

Dougherty et al. (2018) Genome Research 28: 1-11

MORNING AGENDA

Time	Agenda
8:00 – 8**:55 am**	**Registration and Continental Breakfast**
9:00 – 9:10 am	**Welcome Day 2** Jonas Korlach, Ph.D., Chief Scientific Officer, PacBio
9:10 – 9:35 am	**SMRT-Enabled Clinical Assay for Characterizing the Mutant Transcript in Huntington's Disease** Nenad Svrzikapa, Data Scientist, WAVE Life Sciences
9:35 – 9:45 am	**Catching Large DNA in the SageHLS** Chris Boles, Ph.D., Chief Scientific Officer ***Platinum Sponsor Presentation: Sage Science***
9:50 – 11:05 am	**Breakout Rotation 3***
11:05 – 11:35 am	**Coffee Break**
11:35 – 12:55 pm	**Lightning Talks: Fast-paced Updates from our User Base**
12:55 – 1:55 pm	**Lunch**

*Find your breakout session room assignment on your name badge

AFTERNOON AGENDA

Time	Agenda
1:55 – 2:20 pm	**TrioBinning: Trio-Based Assembly** Sergey Koren, Ph.D., Staff Scientist, Genome Informatics Section, National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH)
2:20 – 2:45 pm	~~**Application of Genome Assembly in Bovinae Species**~~ **"The greatest mammal genome ever!"** Timothy Smith, Ph.D., Molecular Geneticist, USMARC, USDA-ARS
2:45 – 3:00 pm	**Future Developments in SMRT Sequencing** Jonas Korlach, Ph.D., Chief Scientific Officer, PacBio
3:00 pm	**Depart for flights/hotels**