Optimizing High Molecular Weight Conifer gDNA Extraction and Single Molecule Sequencing in the Redwood Genome Project

Rachael Workman

Timp Lab, Johns Hopkins University, Department of Biomedical Engineering

SFAF 2018

Sequoia sempervirens
Coast redwood
California – central/north coast

Sequoiadendron giganteum
Giant sequoia
California – Sierra Nevada

Save The Redwoods LEAGUE®

Two species in our redwood genome project

California endemics

Economic, cultural, and conservation value

”Advanced management strategies”

Huge genomes: 30Gb Coast, 9Gb Giant

Climbing & collection
by
Steve Sillett + team

DNA extraction
Neale lab

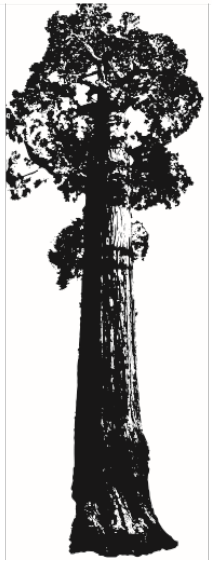
Short read
sequencing @ UCD

DNA extraction
Timp lab

Long read
sequencing @ JHU

Reference
genome

Assembly
@ JHU



Climbing & collection
by
Steve Sillett + team

DNA extraction
Neale lab

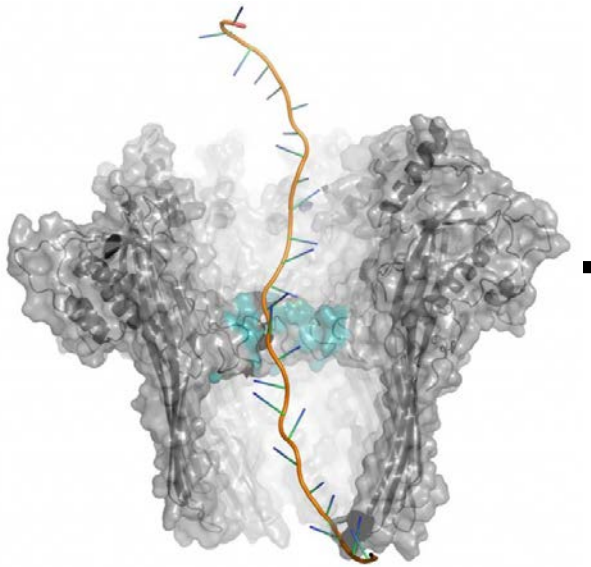
Short read
sequencing @ UCD

DNA extraction
Timp lab

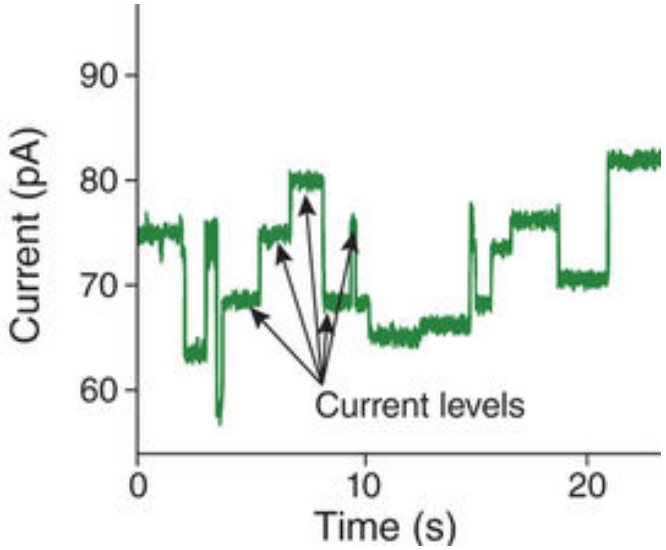
Long read
sequencing @ JHU

Slides courtesy of Neale Lab, UCD

Nanopore Single Molecule Sequencing

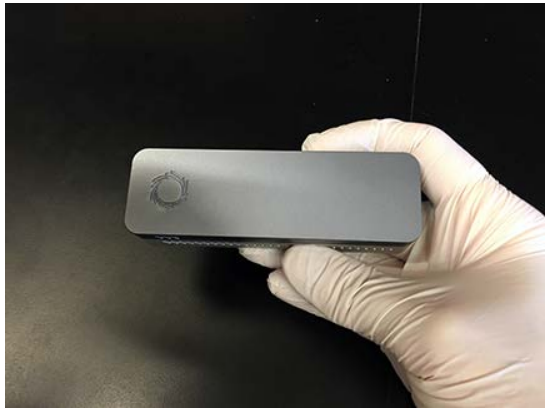


Oxford Nanopore Google Hangout March 2016



Deamer et al 2016, Nature Biotech

ATCGATCGATAG
TATTAGATACGA
CTAGCGATCAG

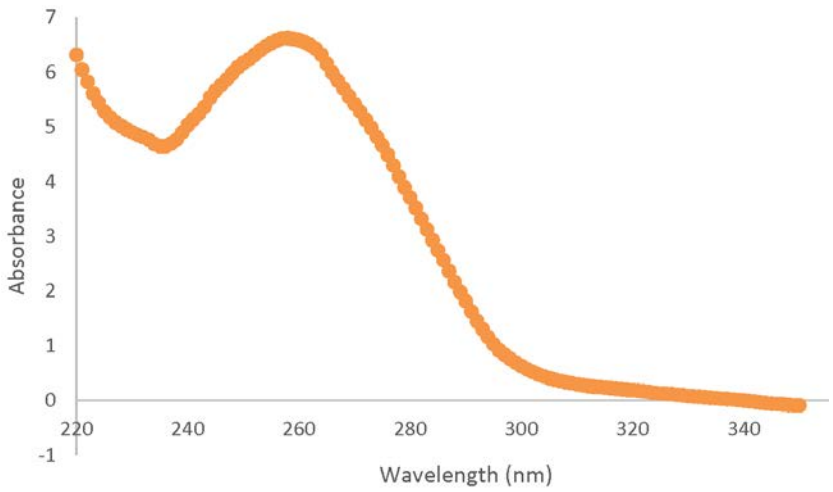


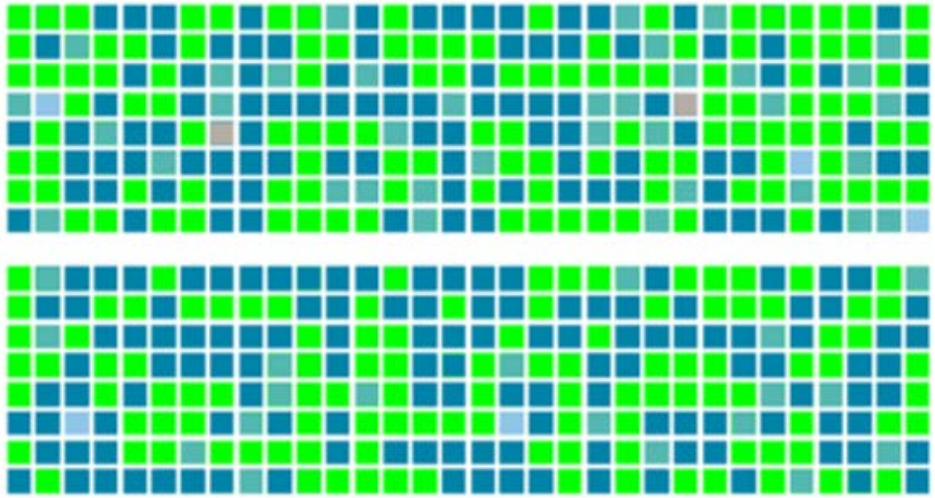
No theoretical upper limit to sequencing read length, practical limit only in preparing long fragment libraries and delivering DNA to the pore intact

Typical user-reported sequencing output 5-15Gb (as of R9.4.1, March 2018)

Sample requirements for sequencing

HMW	Yield	Quality	Reproducibility	Sequencing Yield
100kb+ average	>10ug gDNA From 1g tissue	Nanodrop, gel migration in range	Want no Wizards	>5Gb per run





Sample realities before optimization

LMW

<10kb
average

Low Yield

<1ug gDNA
From 1g tissue

Poor Quality

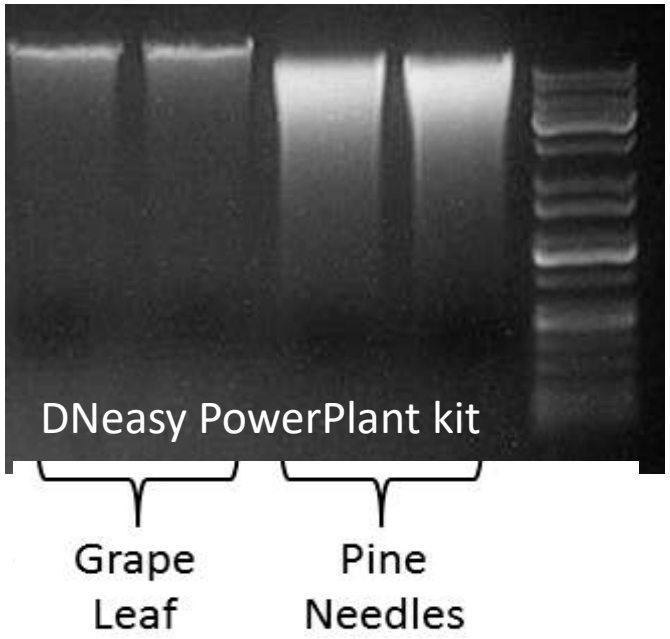
Residual polyphenolics
And polysaccharides

Inconsistent

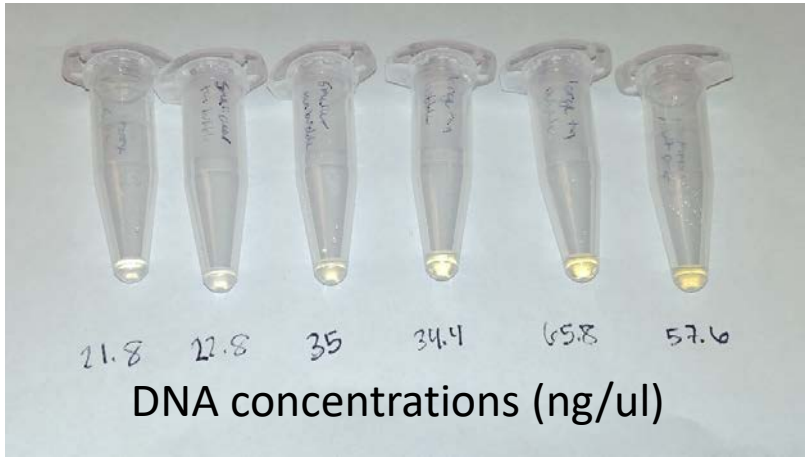
Results varied largely
By sample

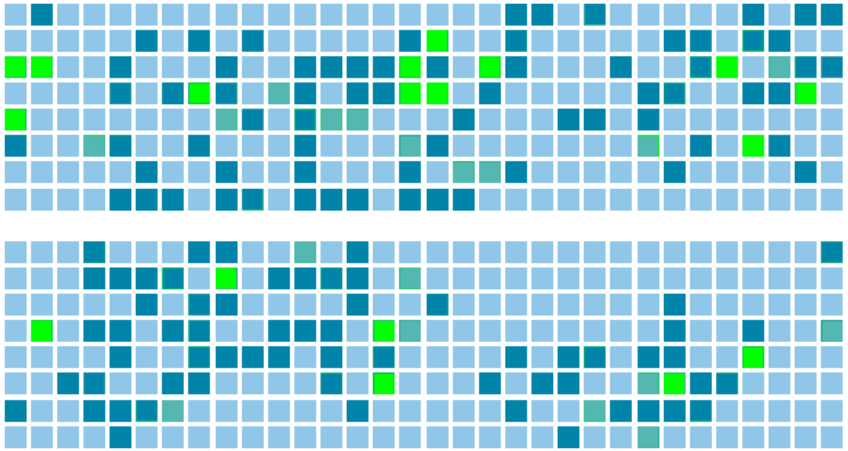
Low Seq Yield

<1Gb per run

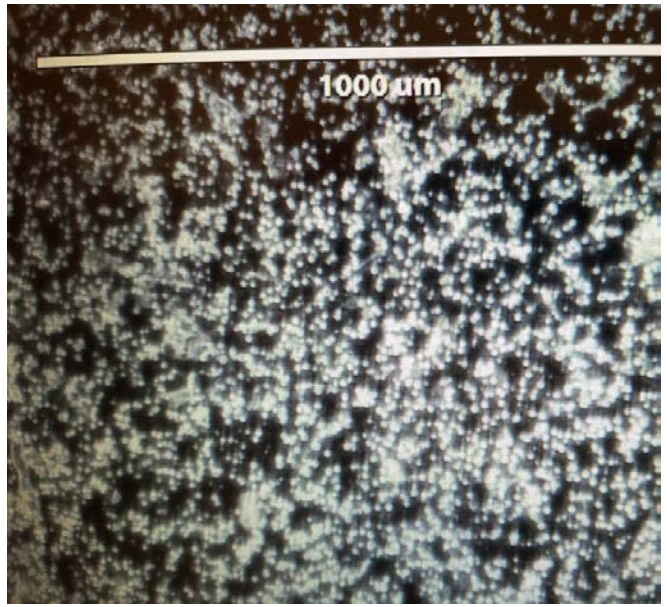


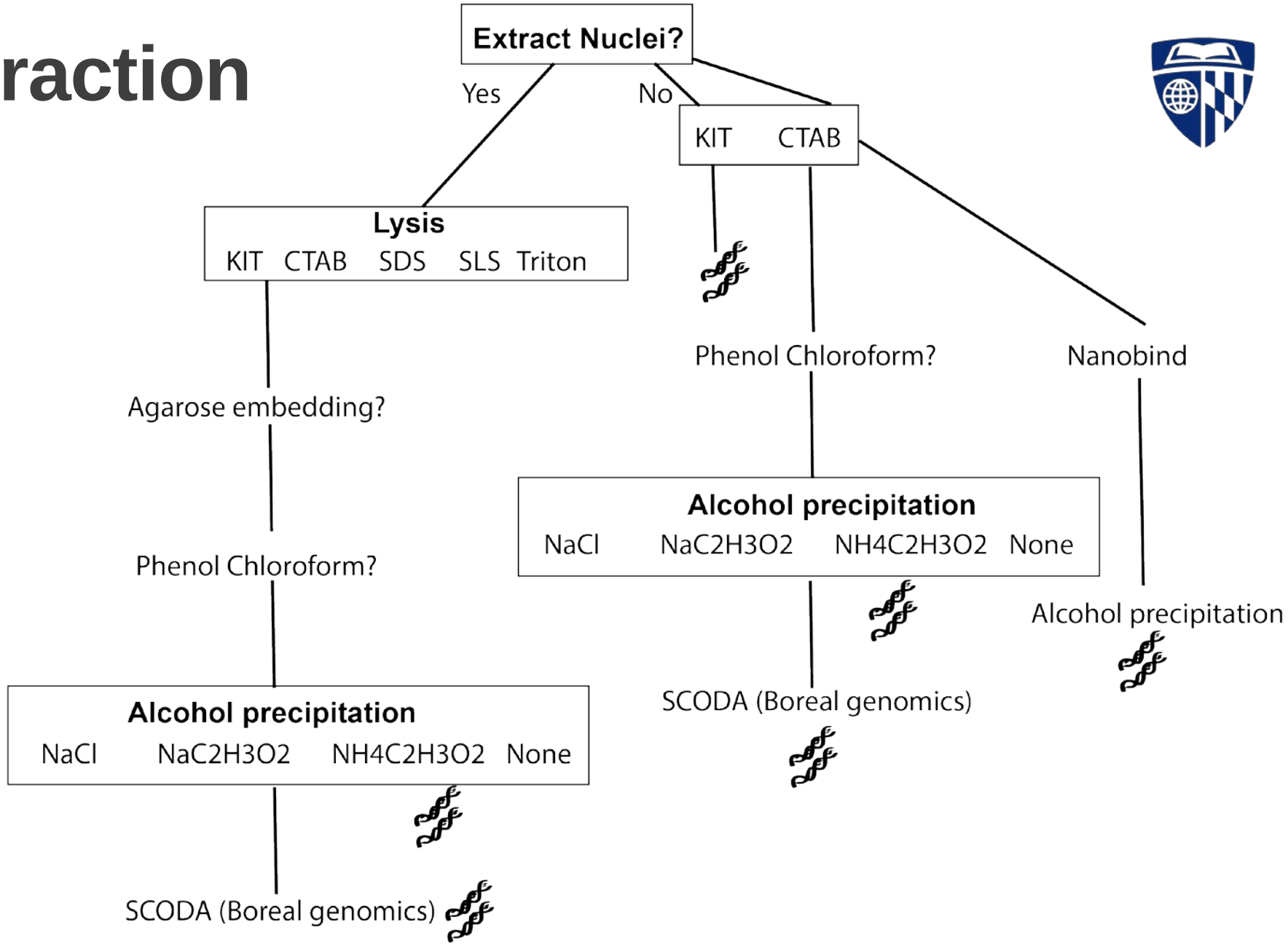
10kb





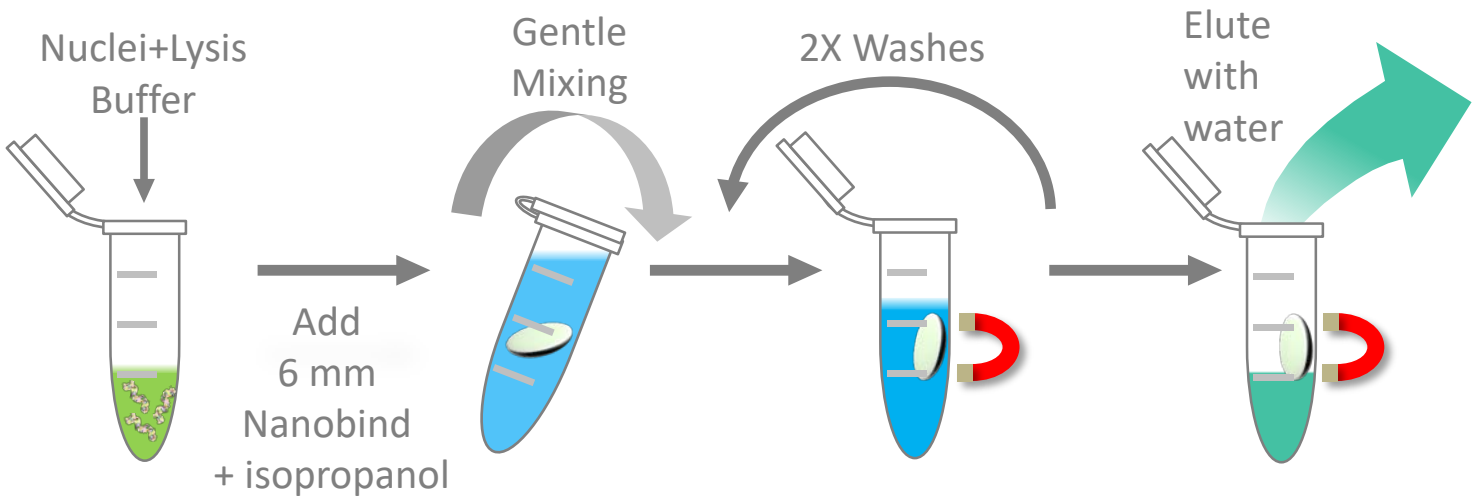
Trials: DNA extraction

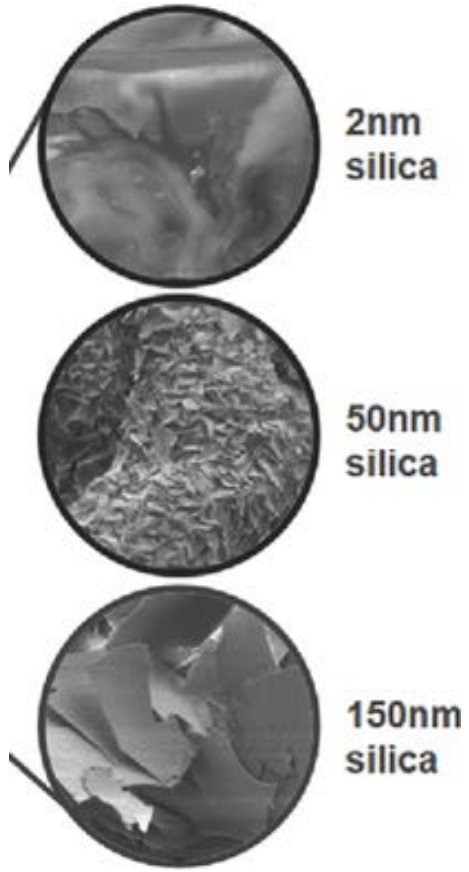




Top extraction protocol: Nanobind

Nuclei →





Nuclei isolation, followed with Nanobind-assisted purification

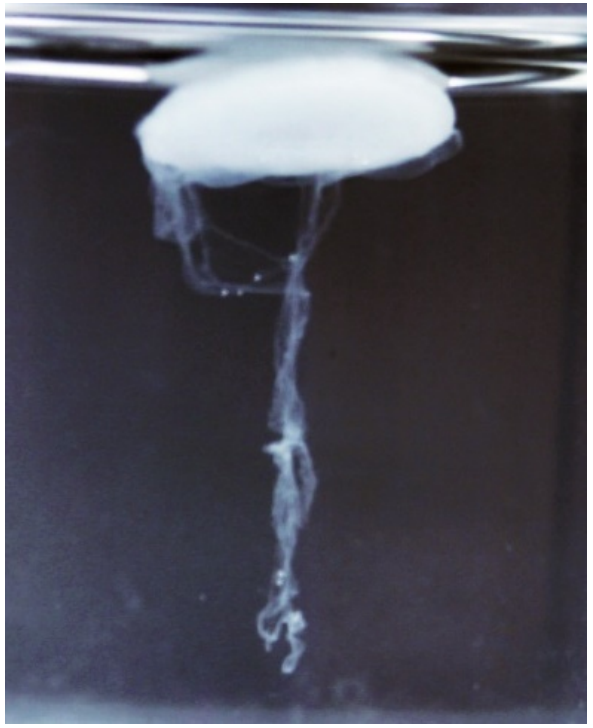
2-3 hours total time (nuclei extract + lysis + purification)

circulomics

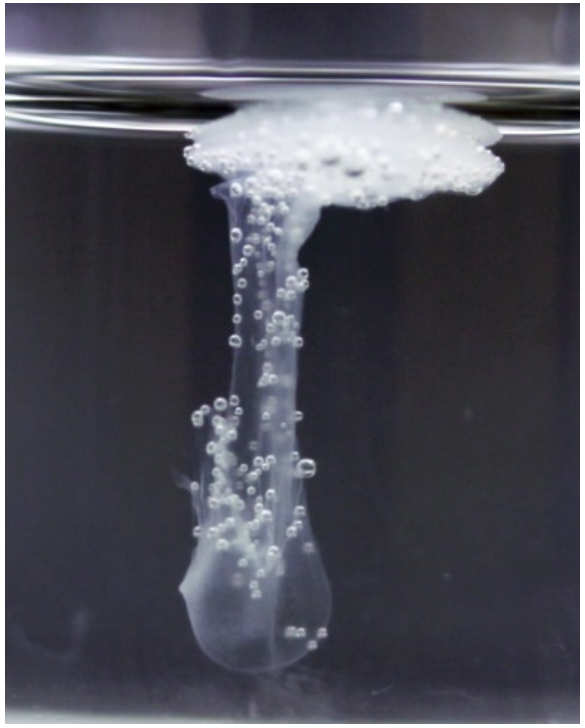
Zhang et al. Adv Mat. (2016)

Tentacle Binding Mechanism

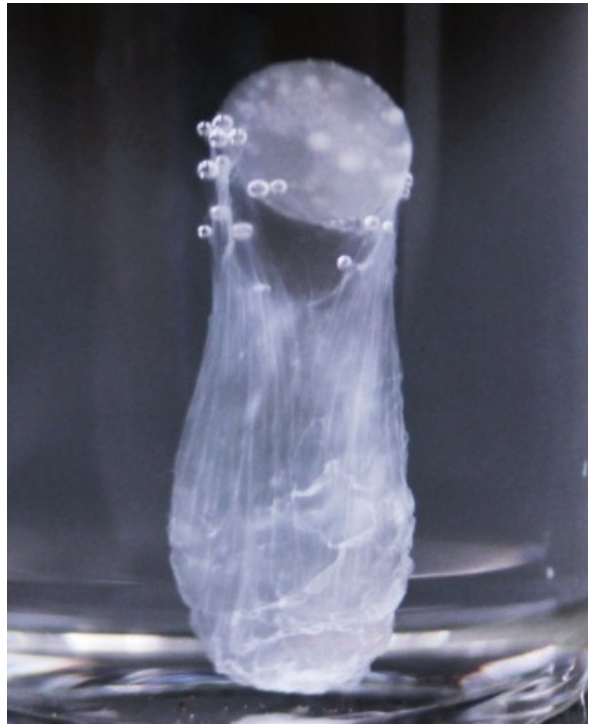
Enhances binding capacity and protects DNA from shear forces



Low Input (10 µg)



Medium Input (50 µg)



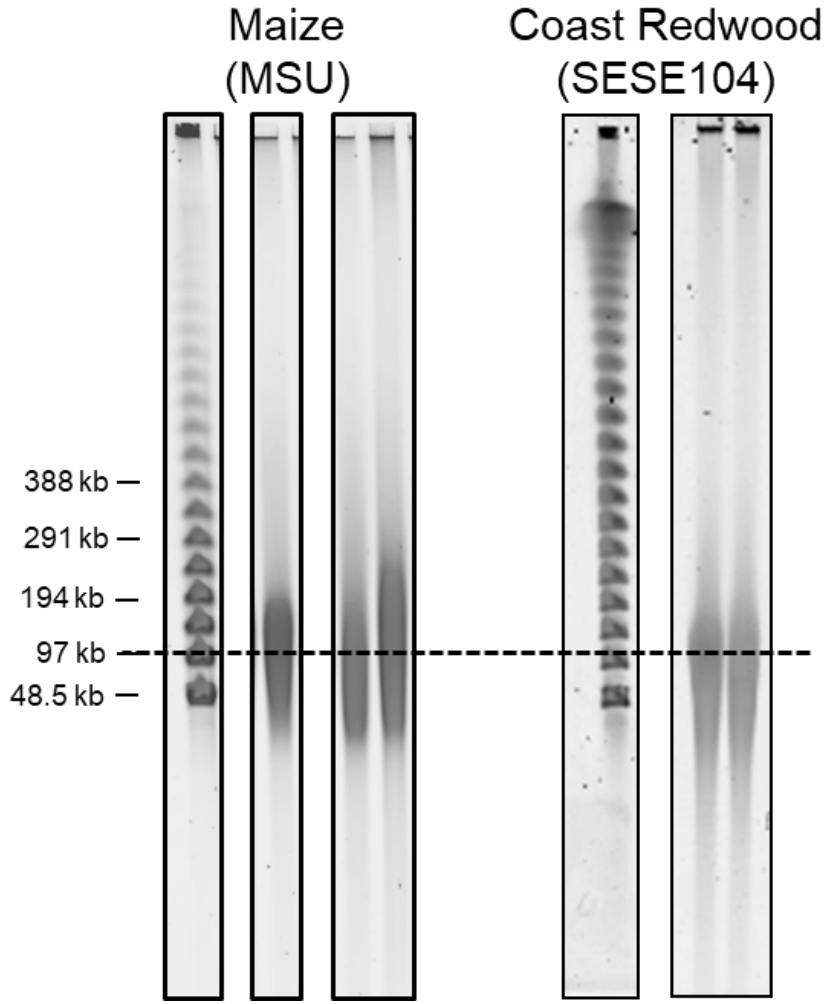
High Input (200 µg)

- **Three material properties needed: low shear, non-porous, high surface area**
- **DNA tentacles form and extend from substrate to get high binding capacity**
- **Low shear unlike beads and columns**

Zhang et al. Adv Mat. (2016)

Current Extraction Yields

	Giant Sequoia	Coast Redwood	Maize (MSU)
Input	1 gram leaf tissue	1 gram leaf tissue	1 gram leaf tissue
Mean gDNA yield (ug)	13.4 ± 1.1 µg (11.5-15.1 ug)	11.5 ± 2.5 µg (7.9-14.8 ug)	5.8 ± 0.9 µg (4.6-6.5 ug)
Mean PFGE sizing[+]	35-150 kb	45-250 kb	45-300 kb
Nanodrop (260/280)	1.77 ± 0.07 (1.70-1.82)	1.77 ± 0.03 (1.73-1.83)	1.85 ± 0.01 (1.83-1.87)
Nanodrop (260/230)	1.41 ± 0.27 (1.12-1.65)	1.40 ± 0.16 (1.20-1.69)	1.87 ± 0.20 (1.48-2.13)



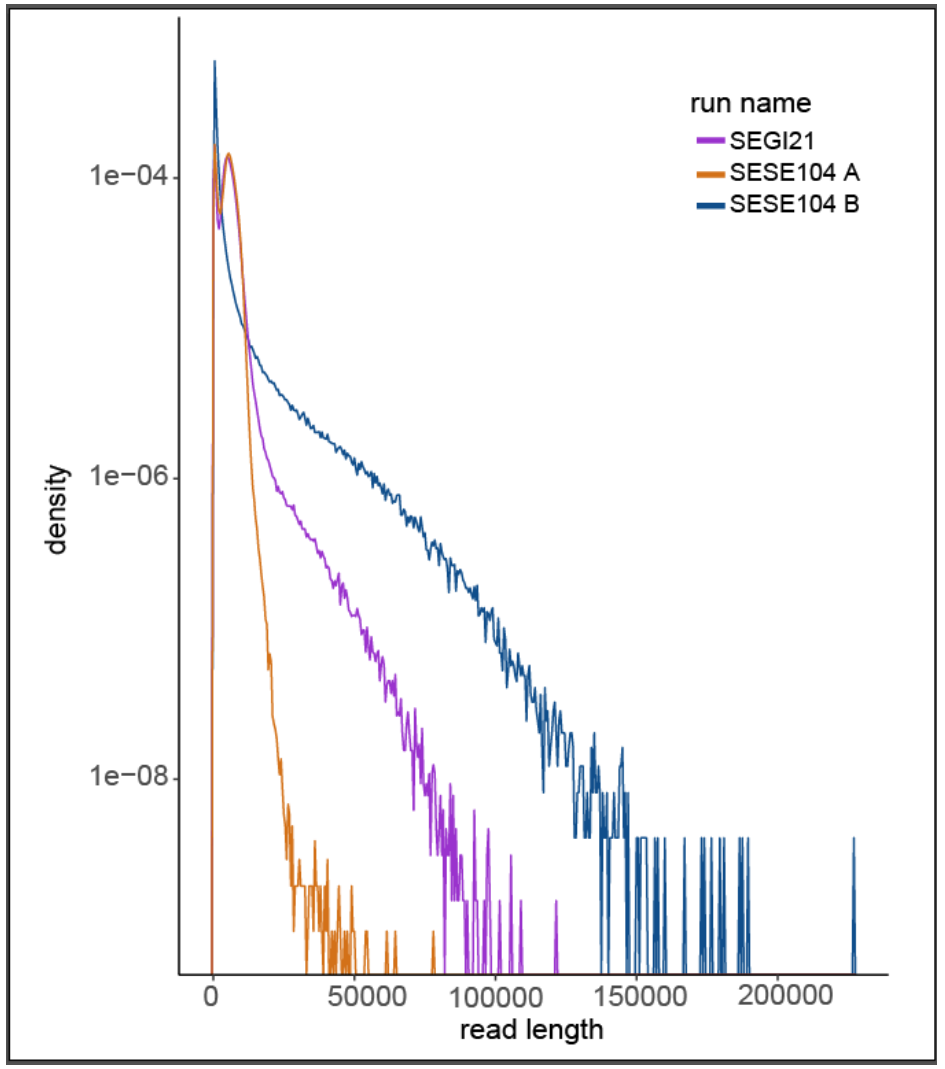
Extraction methods extensible to other plant species as well

Met goals of gDNA yield, length and quality

Maize results courtesy of B. Vaillancourt and Krystle Wiegert-Rininger of the C. Robin Buell Lab at Michigan State University

Current Nanopore Sequencing Yields

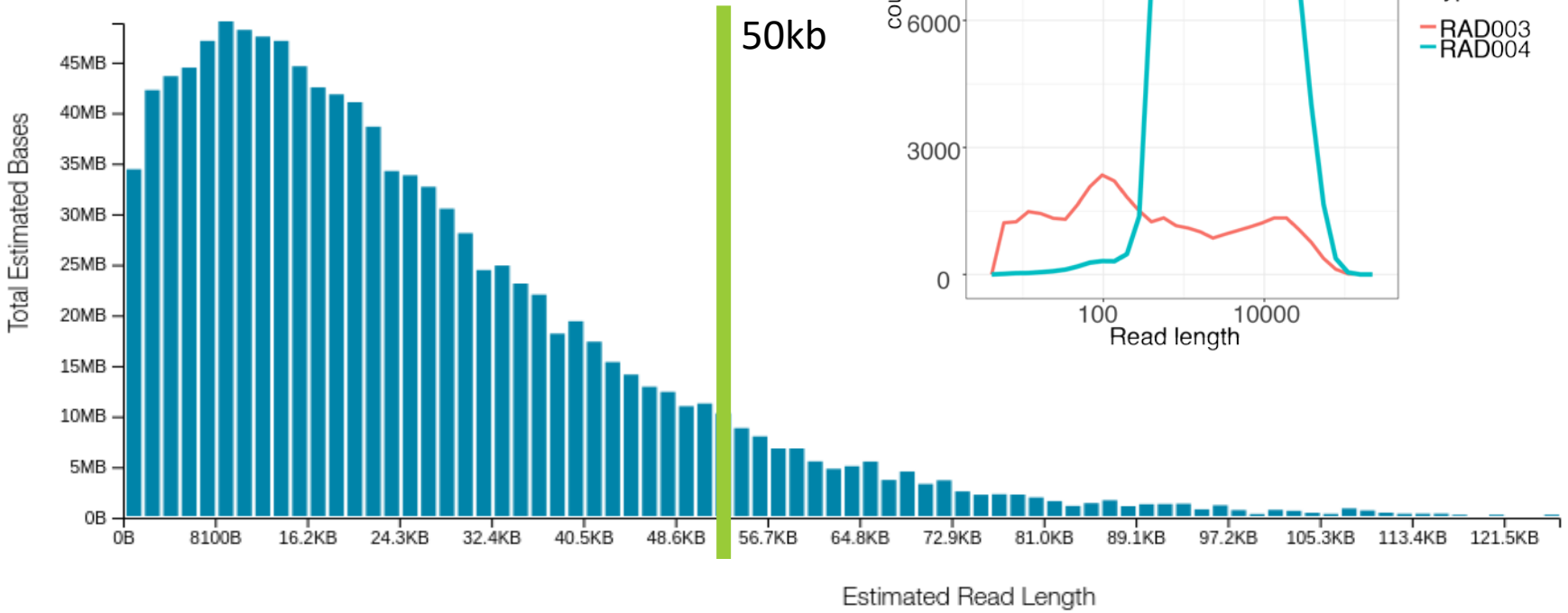
Sample	Giant Sequoia (SEGI21)	Coast Redwood (SESE104A)	Coast Redwood (SESE104B)
Shearing	Megaruptor (10 kb)	Covaris G-tube (8 kb)	26G Needle shear (5X)
Nanopore chemistry	LSK108, R 9.4	LSK108, R 9.4.1	LSK108, R 9.4
Seq yield	6.4 Gb	10.10 Gb	3.3 Gb
Mean read length	5.5 kb	5 kb	6.8 kb
Max read length	121 kb	78 kb	227 kb
Read length N50	6.9 kb	6.6 kb	29 kb

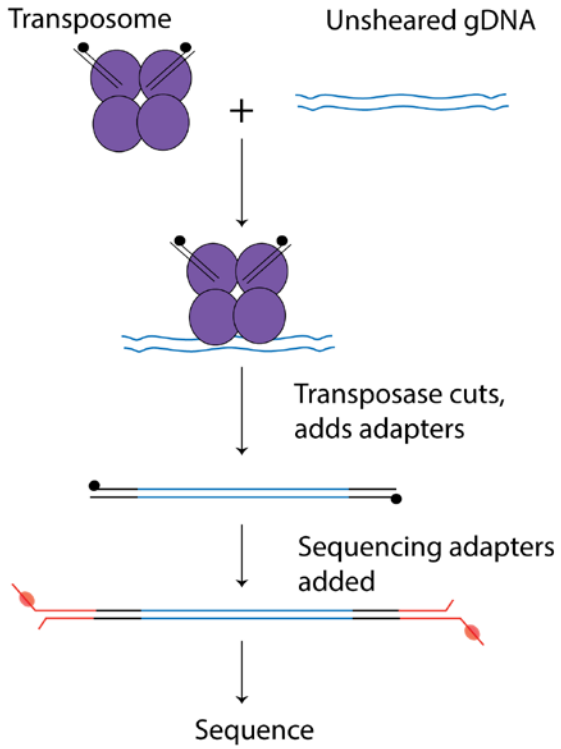


Improving Read Lengths: Rapid kit RAD004

15 minute protocol

Yield:
3Gb from 150K reads
0.89Gb from >50kb reads





Extraction methodology extensible to other sequencing platforms

PacBio	Human (GM12878)	Redwood (SESE104)
Polymerase Read Bases	10,437,674,254	5,559,184,487
Polymerase Reads	475,423	276,987
Polymerase Read Length (mean)	21,954	20,070
Polymerase Read N50	35,750	35,250
Insert Length (mean)	20,925	19,306
Insert N50	33,750	33,250
Yield (Gb)	9.9	5.56

10X analysis ongoing…

Improved extraction and sequencing methods affect assembly contiguity

Conifer species	Year completed	Sequencing methods	Genome size (Gb estimated)	Contig N50 (kb)	Scaffold N50 (kb)	Citation
Giant sequoia	**ongoing	Nanopore + Illumina + …	9	365	495	**ongoing
Loblolly Pine	2017	Illumina PE + MP + DiTag + PacBio	22	26	107.8	Zimin et al, Gigascience 2017
Douglas Fir	2017	Illumina PE + Mate Pair	16	44.1	340.7	Neale et al, G3 2017
Sugar Pine	2016	Illumina + PacBio	31	3.4	246.6	Stevens et al, Genetics 2016

- Improvements in contig and scaffold size over other conifer assemblies afforded by long reads
- MaSuRCA assembler (Zimin et al, Bioinf 2013; http://www.genome.umd.edu/masurca.html)

Conclusions

- Optimization of upstream biochemistry matters for single molecule sequencing and assembly
- Developed an extraction protocol which is extensible to other platforms, and has been successfully used by other labs with other organisms
- Protocol is publically available!

PROTOCOL EXCHANGE | COMMUNITY CONTRIBUTED

High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation Sequencing

Rachael Workman, Renee Fedak, Duncan Kilburn, Stephanie Hao, Kelvin Liu & Winston Timp

Timp Lab, Johns Hopkins

Protocol Exchange (2018) | doi:10.1038/protex.2018.059
Published online 27 April 2018

https://www.nature.com/protocolexchange/protocols/6785




Acknowledgments

Timp Lab
- Stephanie Hao

Salzberg Lab
- Jennifer Lu
- Aleksey Zimin
- Daniela Puiu

Neale Lab
- Alison Scott
- Zane Moore

Circulomics
- Kelvin Liu
- Duncan Kilburn
- Jeffrey Burke
- Renee Fedak

Redwood Genome Project (Neale)

National Human Genome Research Institute
1R01HG009190-01A1 (Timp)
2R44GM109618-02 (Liu)