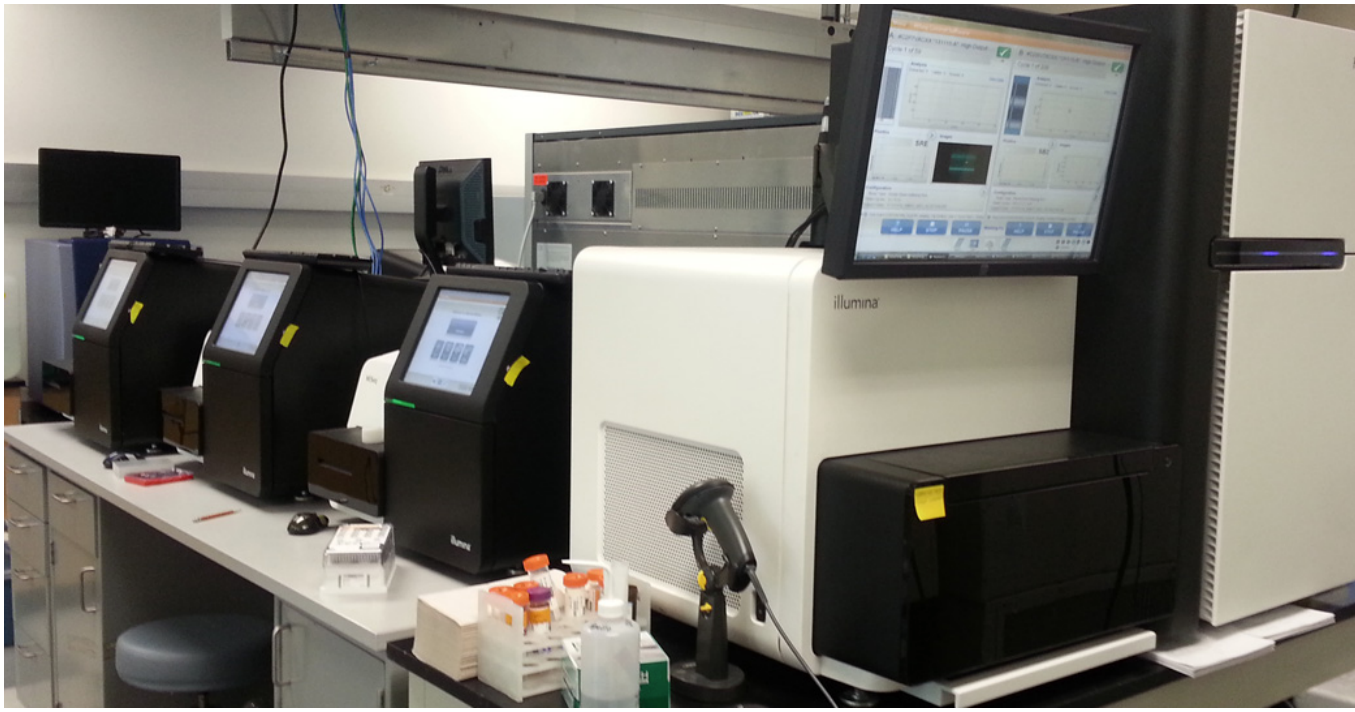
Human Chromosomes. Credit: Jane Ades, NHGRI



The Genome Assembly Workshop

Lutz Froenicke
DNA Technologies & Expression Analysis Cores
UC Davis Genome Center
2020

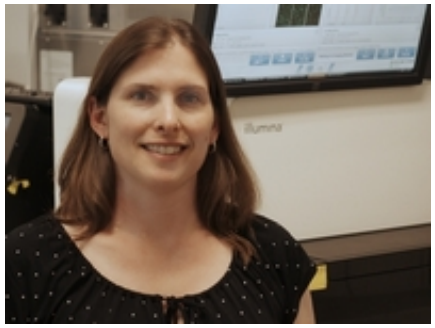
DNA Technologies & Expression Analysis Cores

- HT Sequencing Illumina
- Long-Read & Linked Read Sequencing
  PacBio, Oxoford Nanopore, 10X Genomics
- HMW DNA isolation
- Illumina microarray (genotyping)

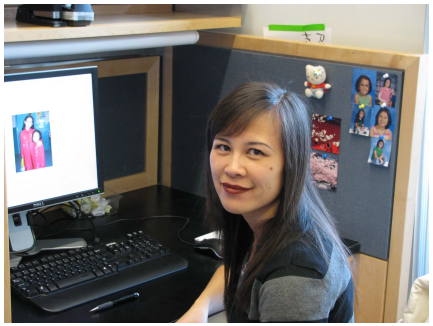
- Consultations → Experimental Design
  (Bioinformatics Core & DNA Tech Core)

- introducing new technologies to the campus
- shared equipment
- teaching (workshops)

The DNA Tech Core Team

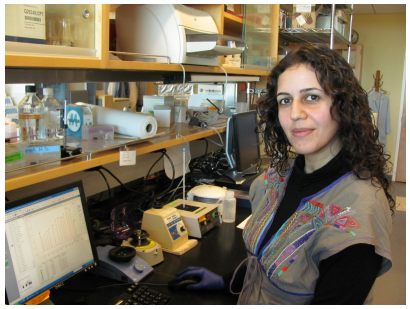


Emily

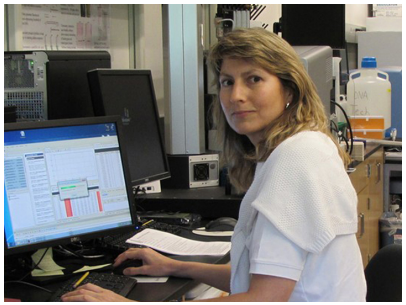


Oanh

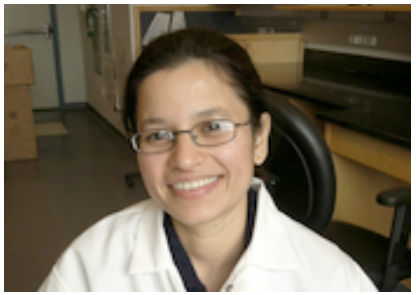
Diana



Siranoosh

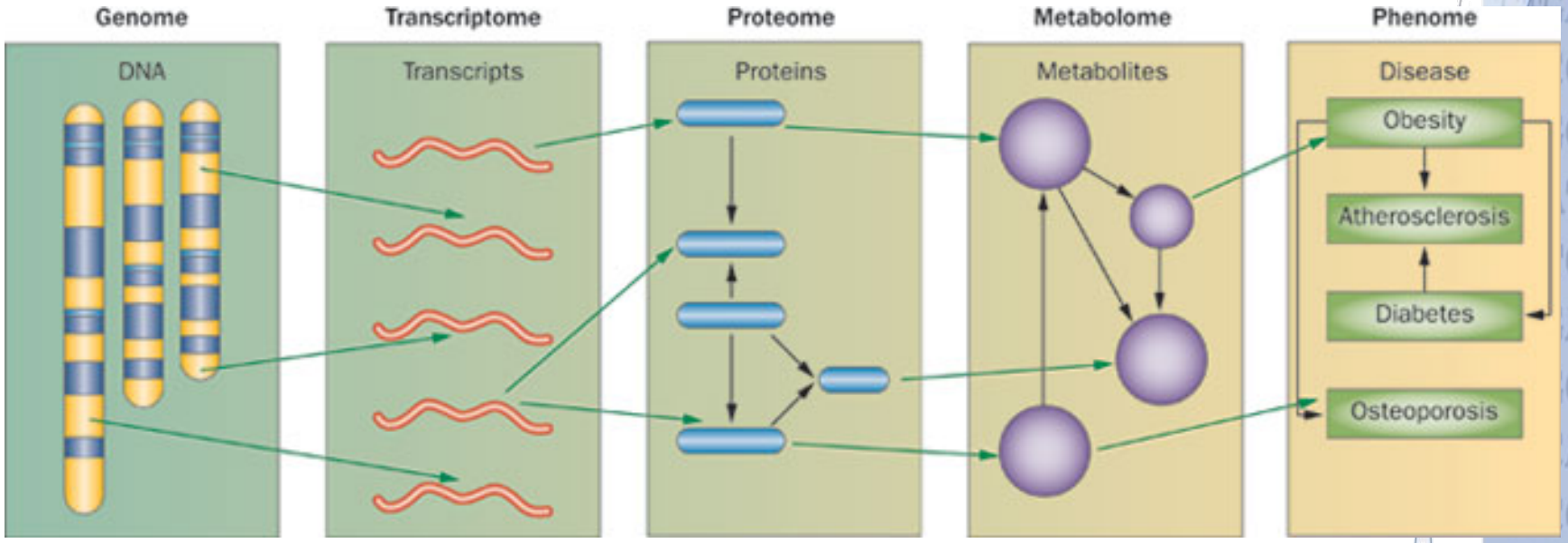


Vanessa



Ruta

The UCD GENOME CENTER



DNA Tech & Expression Analysis Proteomics Core Metabolomics Core

"DNA makes RNA and RNA makes protein"

the Central Dogma of Molecular Biology; simplified from Francis Crick 1958

nature REVIEWS CARDIOLOGY

MacLellan, W. R. *et al.* (2012) Systems-based approaches to cardiovascular disease *Nat. Rev. Cardiol.* doi:10.1038/nrcardio.2011.208

DNA TECH

DNA Technologies & Expression Analysis Core Laboratory

UCDAVIS GENOME CENTER

Welcome to the DNA Technologies & Expression Analysis Core

The DNA Technologies Core has resumed operations. We are working with reduced staffing.

We receive sample shipments via FedEx and UPS.
On campus: Since the Genome Center building doors and lab doors are locked, we can receive samples in front of the building. Please call the lab phone (530-754-9143) between 9 am and noon. We will meet you outside the front door to accept a box with your samples plus submission form.

In order to minimize the likelihood of Covid virus transmissions following changes to our operations remain in place:

- The **lab doors are closed** and only staff members have access to the labs.
- The ***shared equipment*** will not be accessible until further notice. We offer **sample processing** on this equipment **as a service** (please contact Siranoosh Ashtari [sashtari@ucdavis.edu] or call the lab).
- Consultations are carried out via Zoom online calls.

We will update this page as soon anything changes. Other UC Davis Covid-19 updates are available here.

Genomics & Gene Expression Consultations

Please pick a time for a consultation here

The DNA Technologies and Expression Analysis Core at the Genome Center offers high-throughput sequencing, genotyping, and microarray services, as well as training and consultation. Our goal is to enable access to high throughput genome-wide analyses at economical recharge rates, as a functional extension of your laboratory. We operate on the cost-recovery principle. We employ liquid handling robots to minimize sample handling variation and to provide fast turnaround times. We are a designated Campus Research Core Facility.

search here … Go

Recent Posts

- Core operations resume in Phase 2 of the COVID response
- DNA Technologies Core has to ramp down lab work
- Adjusting DNA Tech Core operation to the COVID-19 guidelines
- Join us for the PacBio Day Symposium — February 26th
- PacBio Sequel II Sequencer Up and Running

Latest Tweets

- A potential treatment for the citrus greening disease. https://t.co/ZeixsU1m1f, Jul 18
- Online Bioinformatics Workshops: Our neighbors are running their SECOND set of online workshops in July & August.… https://t.co/hULdc7J2oL, Jul 14
- RNA-Seq shows that circular RNA makes fruit flies live longer - https://t.co/PKcG2oB9qZ, Jul 14
- Automated assembly of CENTROMERES from ultra-long error-prone reads with the "centroFlye" algorithm. Andrey V. Bzik… https://t.co/CtHscxZgLH, Jul 14
- DNA hybridization: Looks like I need to revise some of

DNA Tech Genome Assembly Tools

- 10X Genomics Chromium Genome "linked-reads"

- PacBio Sequel *II* "super-long reads" >20kb
- PacBio Sequel "high-fidelity-long reads" (Q20,Q30, 15kb & 20kb)

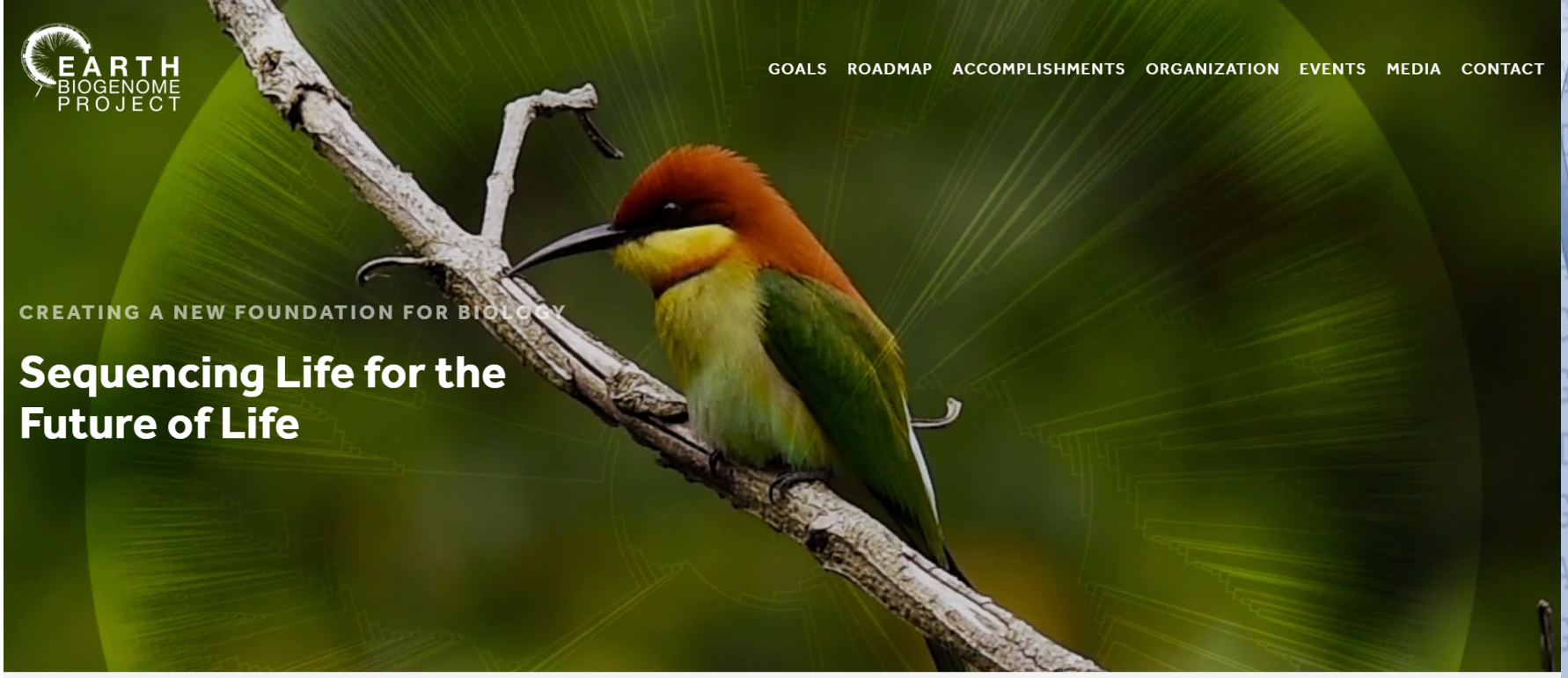
- PromethION Nanopore "super-long reads" >20kb
- MinION Nanopore "ultra-long reads" >100 kb

- Bionano Optical Genome Mapping (scaffolding) > 150 kb
- Hi-C (chromosome scale genome scaffolding)

Bionano Saphyr

Optical Genome Mapping

Saphyr



CREATING A NEW FOUNDATION FOR BIOLOGY

Sequencing Life for the Future of Life

What is the Earth Biogenome Project?

Powerful advances in genome sequencing technology, informatics, automation, and artificial intelligence, have propelled humankind to the threshold of a new beginning in understanding, utilizing, and conserving biodiversity. For the first time in history, it is possible to efficiently sequence the genomes of all known species, and to use genomics to help discover the remaining 80 to 90 percent of species that are currently hidden from science.

A GRAND CHALLENGE

The Earth BioGenome Project (EBP), a *moonshot* for biology, aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years.

A GRAND VISION

Create a new foundation for biology to drive solutions for preserving biodiversity and sustaining human societies.

long-read and linked-read sequencing for high quality genome assemblies

Experience

200+ species/varieties with PacBio

140+ species with 10x Genomics

as of December 2019

DNA Quality !!!

- HMW DNA isolation
  - Physical damage (PFGE image is not fully informative)
  - Chemical damage
  - Chemical contamination

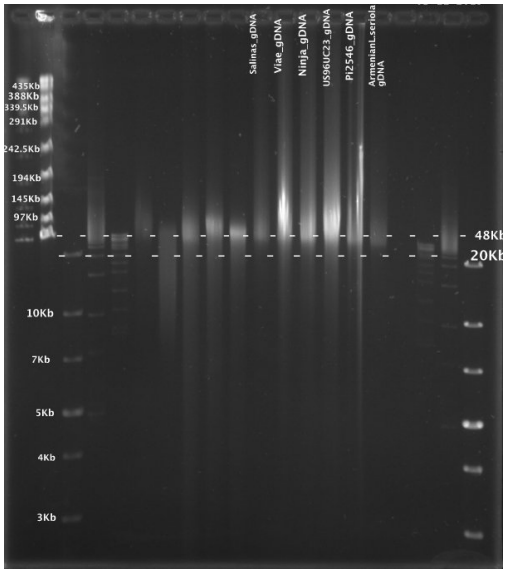
  - Sample specific protocols?
  - Nuclei isolation, agarose plugs
  - Cell culture?
  - Rescue efforts (BluePippin; DNA damage repair) tend to have minimal impact

Input DNA requirements for gDNA sequencing

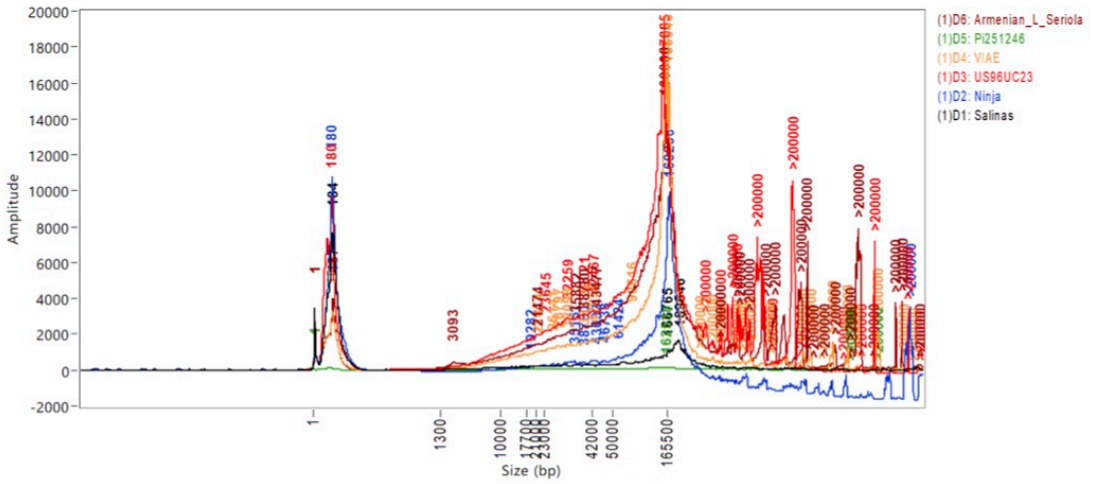
- Good quality, high molecular weigh DNA >50Kb in length
- Free of contaminants such as polysaccharides, proteins, salts, etc
- Nanodrop ratio of 260/280=1.8 260/230=2.0

DNA isolation protocol

Liquid nitrogen grinding + sorbitol pre-wash + CTAB buffer lysis + Sodium acetate and isopropanol precipitation (*Inglis et al*)



Pippin pulse gel image



Femto trace

PromethION run results

ONT library prep protocol: 50Kb shear + SRE, LSK109 library prep

Name	Total yield	Read length N50	Bases >100Kb and Q7 (% of total data)	Bases >200Kb and Q7 (% of total data)
Ninja	81Gb	36Kb	0.4Gb (0.4%)	0.6Mb(0.0007%)
US96UC23	52Gb	36Kb	0.3Gb (0.6%)	0.6Mb(0.001%)
Pi251246	94Gb	37Kb	0.6Gb (0.6%)	4.6Mb (0.04%)
Ar Seriola	98Gb	37Kb	0.8Gb (0.8%)	7.1Mb(0.007%)
Salinas	93Gb	27Kb	0.1Gb (0.19%)	0.4Mb (0.0004%)

Longest read ~200Kb range

Complementary Approaches

Illumina	PacBio CLR	PacBio HiFi	PromethION Nanopore
Still-imaging of clusters (~1000 clonal molecules)	Video recordings fluorescence of single molecules	Video recordings fluorescence of single molecules	Recording of electric current through pores
Short reads - 2x300 bp Miseq	Up to 70 kb, N50 25 kb	Up to 20 kb, N50 18	Up to 100 kb, N50 30 kb
Repeats are mostly not analyzable	spans retro elements	accurate enough to assemble through REs	spans retro elements
High output - up to 2.4 Tb per lane	up to 100 Gb per SMRT-cell	up to 25 Gb HiFi data per cell	Up to 100 Gb per flowcell
High accuracy (< 0.5 %)	Raw data error rate 15 %	CCS data < 0.1%	Raw data error rate 8 %
Considerable base composition bias	No base composition bias	No base composition bias, but still mononucleotide repeat problem	Some systematic errors,
Very affordable	Costs 3 to 5 times higher	Costs 3 to 5 times higher	Costs 2x higher
De novo assemblies of thousands of scaffolds	“Near perfect” genome assemblies	“Near perfect” genome assemblies; lowest error rate	“Near perfect” genome assemblies with suppl. data; highest contiguity

Strengths and weaknesses

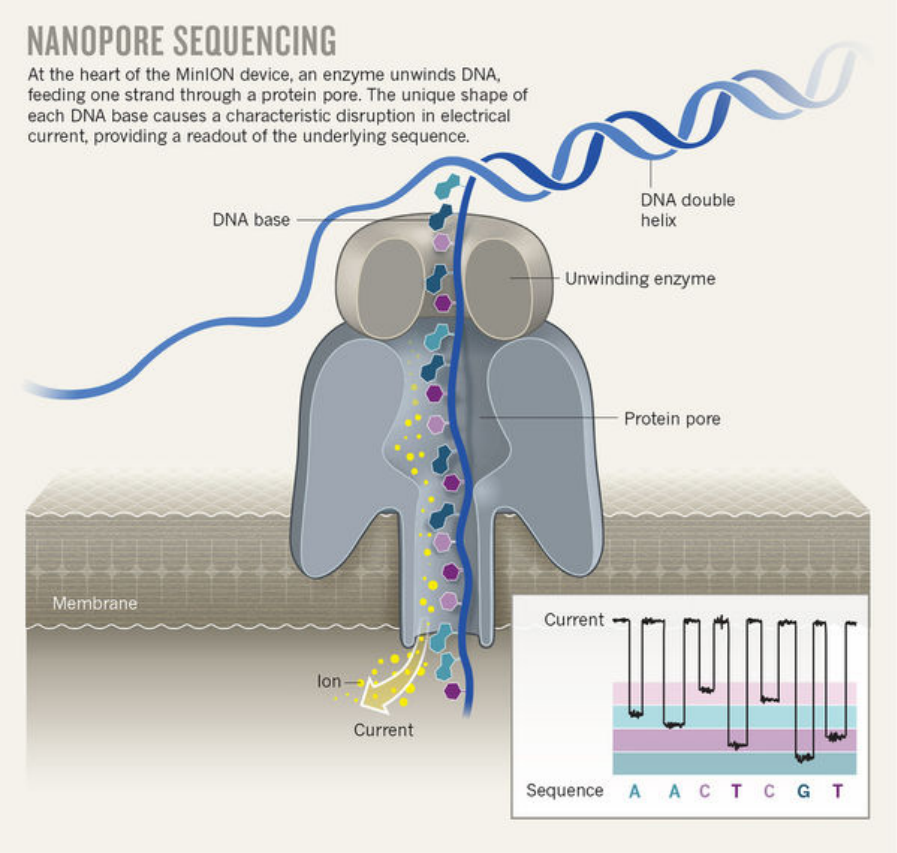
(diploid genomes)

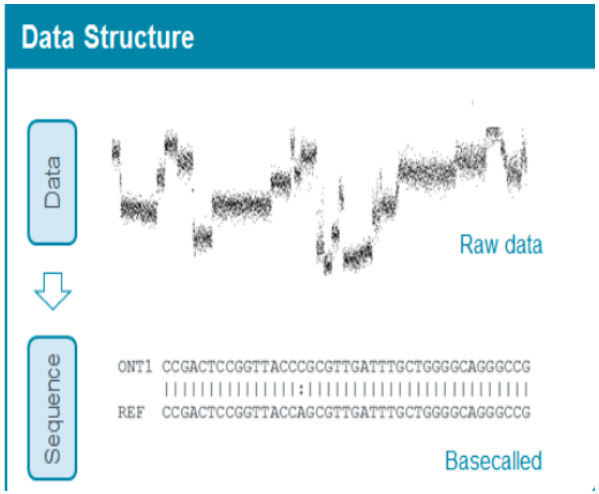
	10X/TELL-Seq	Sequel CLR	Sequel HiFi	PromethION
Accuracy – single read	4	1	4	2
Accuracy - consensus	4	4	4	3
Contiguity	1-2 ?	3-4	3	4
Yield/$	4	4	2	4

- In many case the resulting assembly quality will depend on the DNA sample quality as well genome organization (repeat content/length).

- Pacbio relative even performance (fish ?; jellyfish?)
- Nanopore: significant variance between organisms
  Plants ++; birds -, fish -, jellyfish -

How does nanopore sequencing work ?





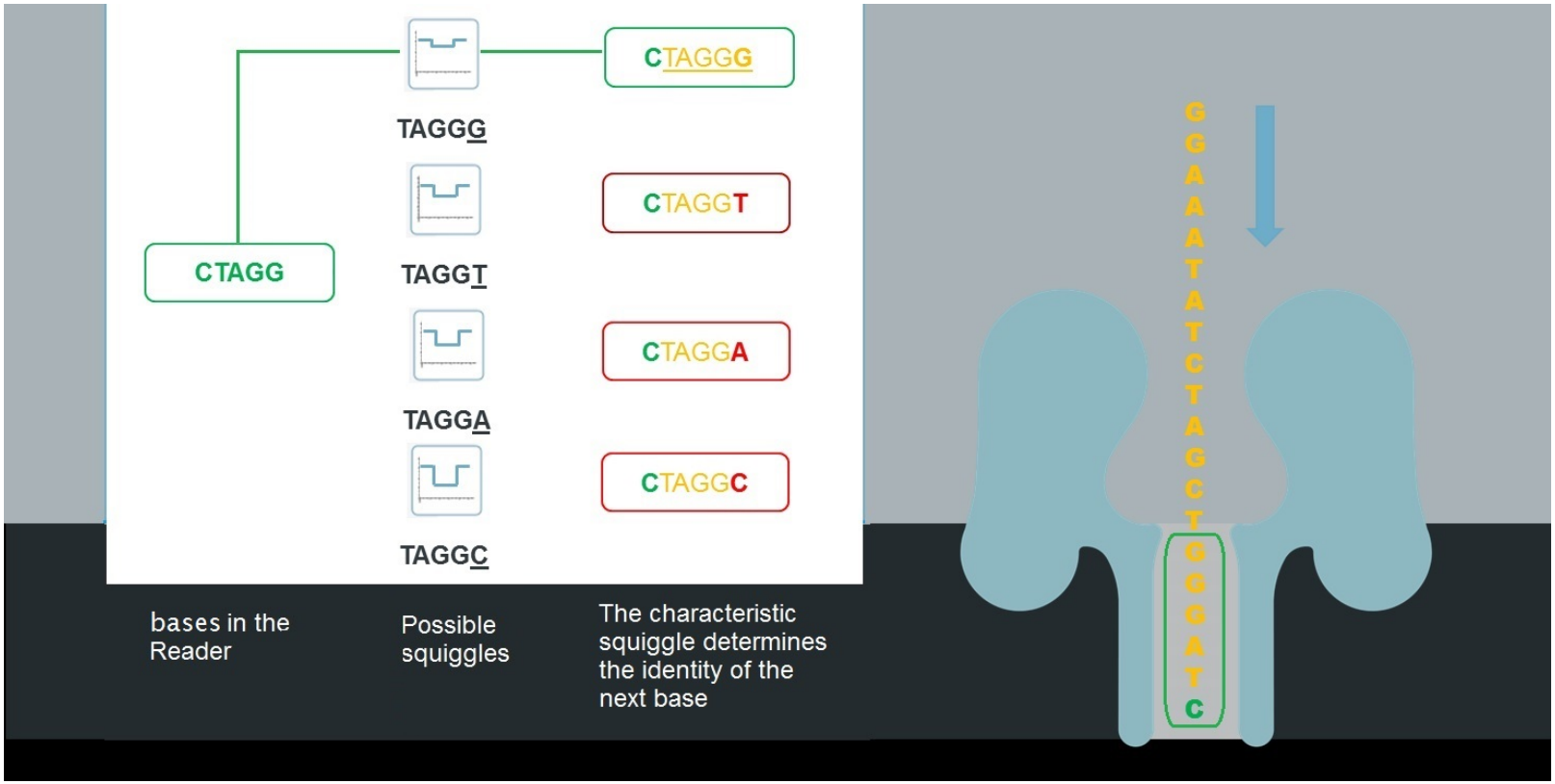
Raw data
fast5

Basecaller
RNN, flip-flop
Guppy 3.2.8

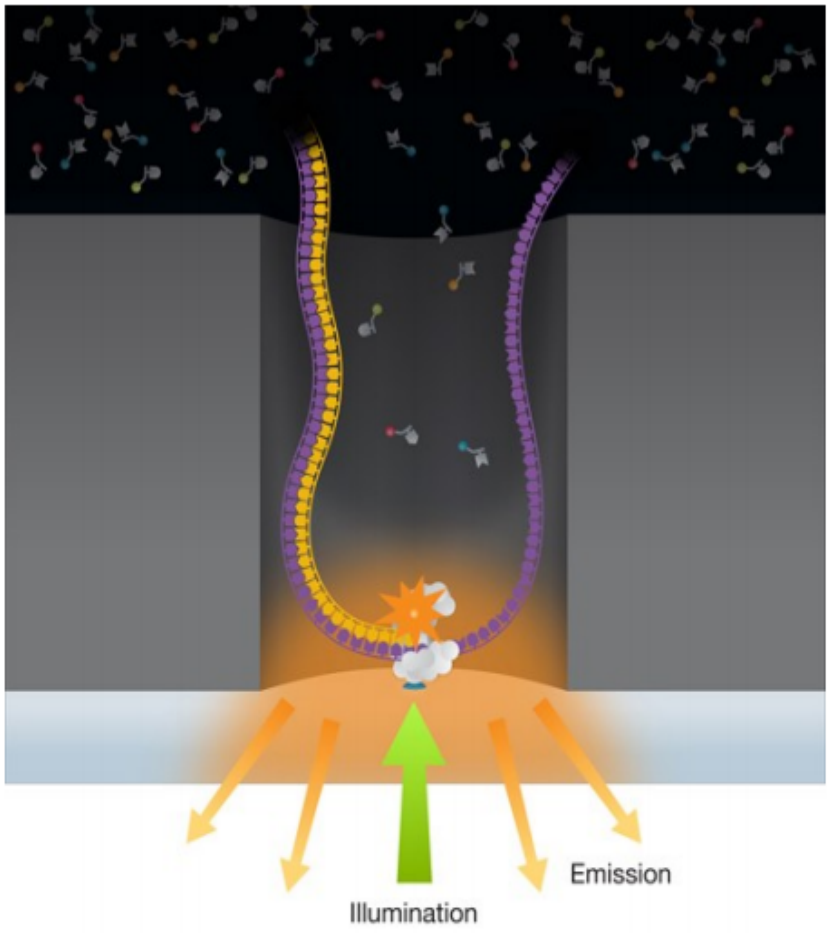
fastq
ATCG

Huge data size

non-random errors



Pacific Biosciences

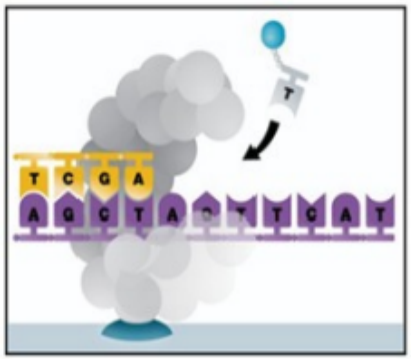


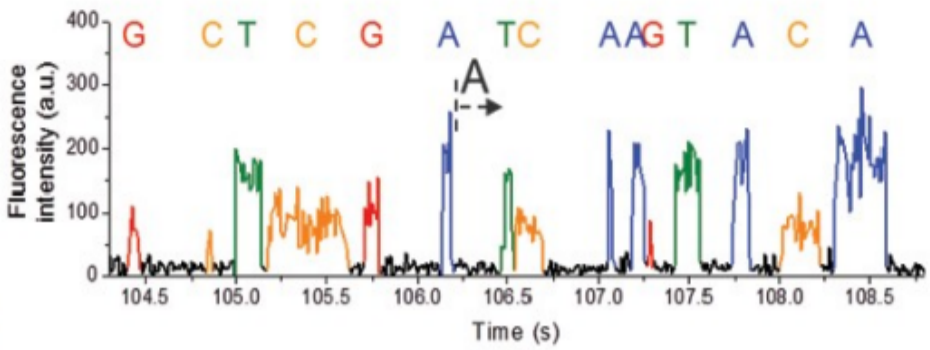
4 nucleotides with different fluorescent dye simultaneous present

2-3 nucleotides/sec
2-3 Kb (up to 50) read length
6 TB data in 30 minutes

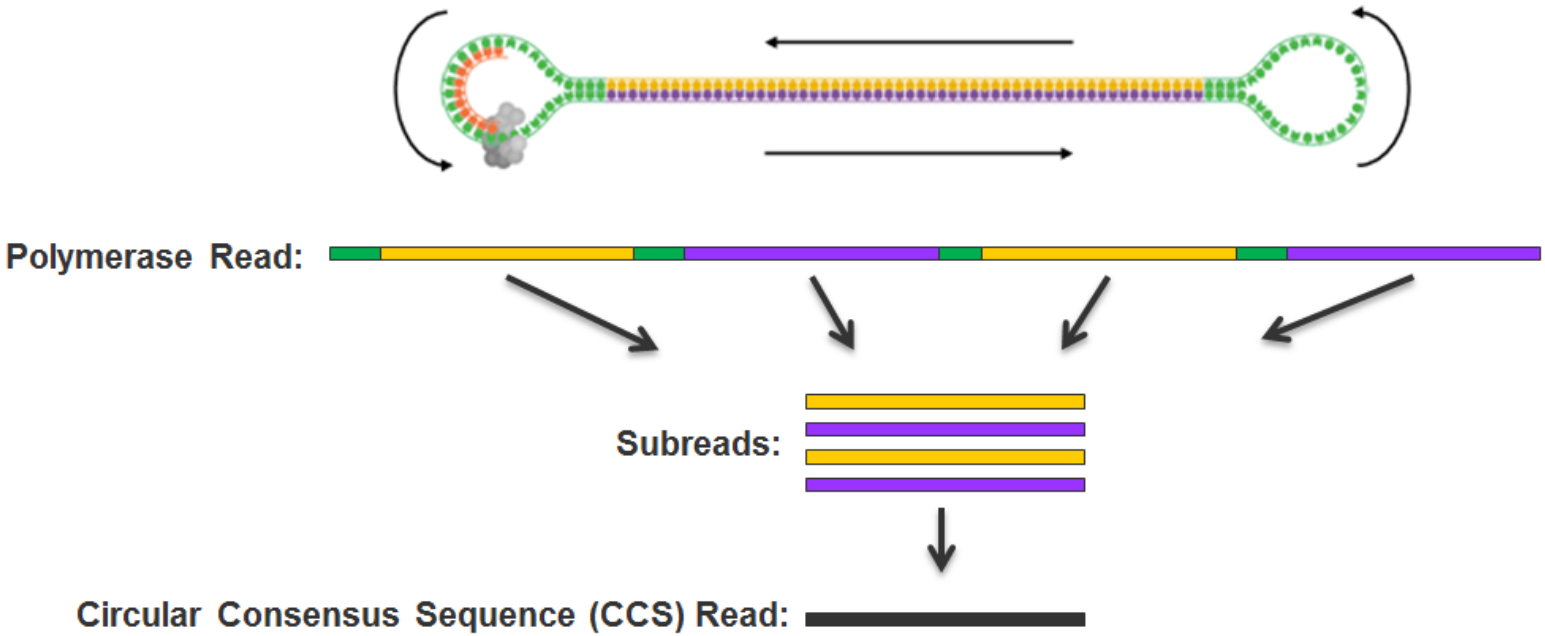
laser damages polymerase

70 nm aperture
"Zero Mode Waveguide"





SMRT-bell adapters circular sequencing

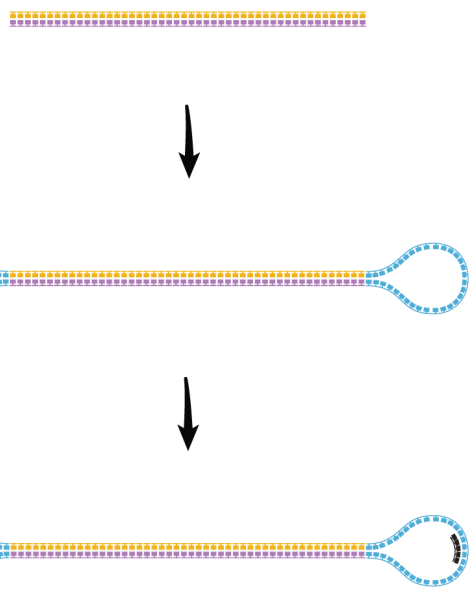


Source: PacBio

Start with high-quality double stranded DNA

Ligate SMRTbell adapters and size select

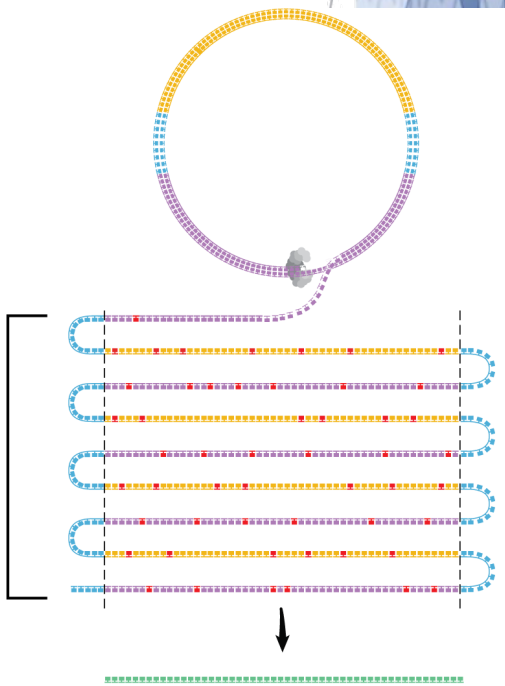
Anneal primers and bind DNA polymerase



Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus is called from subreads

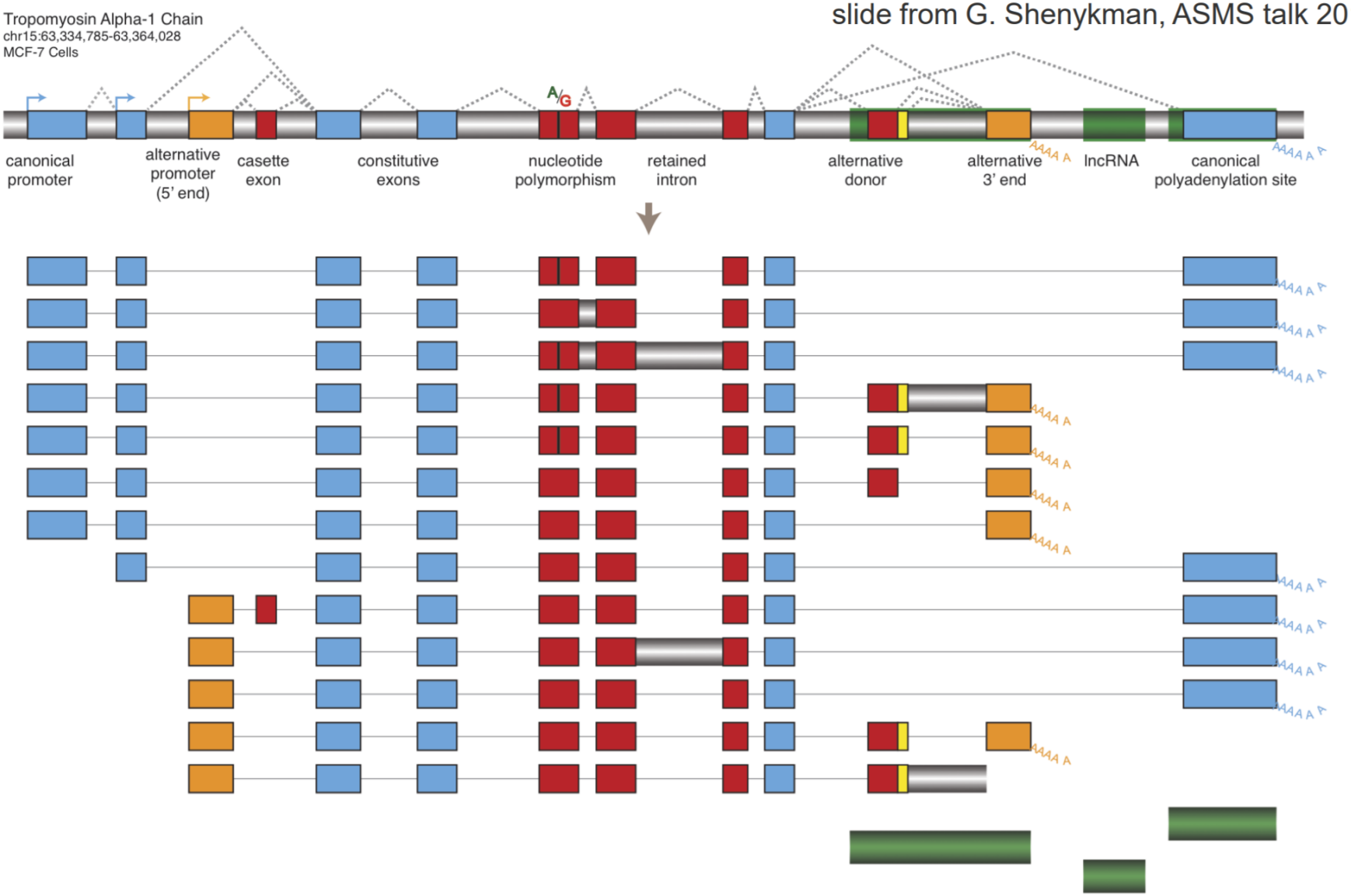


HiFi READ

(>99% accuracy)

A Single Gene Locus ➔ Many Transcripts

slide from G. Shenykman, ASMS talk 2014



Iso-Seq Pacbio

- Sequence full length transcripts
  → no assembly; default for gene annotation

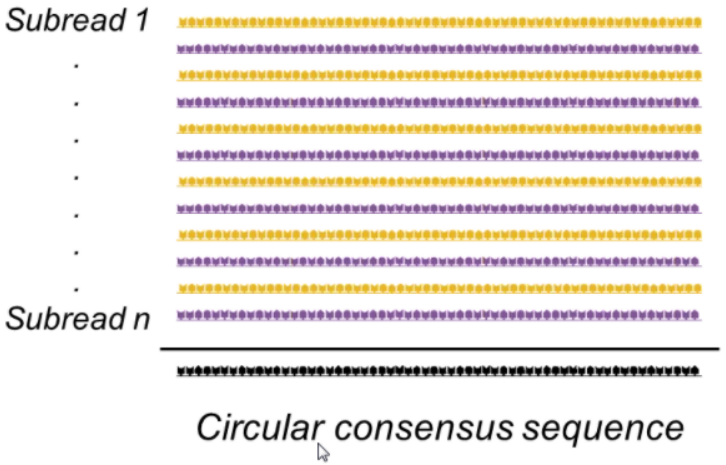
- High accuracy
- More than 95% of genes show alternate splicing
- On average more than 5 isoforms/gene
- Precise delineation of transcript isoforms (PCR artifacts? chimeras?)

NEW CHEMISTRY PERFORMANCE

	v 6.0	v 5.1
Shorter-insert libraries		
- Throughput per SMRT Cell:	**up to 50 Gb**	up to 20 Gb
- Average read length:	**up to 100 kb**	up to 40 kb

Pre-extension:

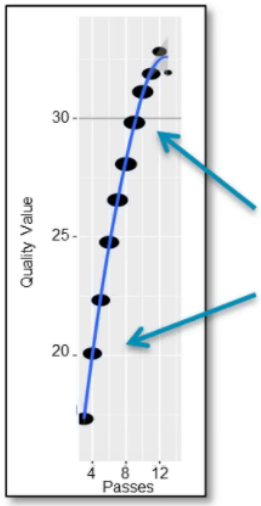
- Expedite polymerase into rolling circle synthesis
- Ensure undamaged template



Source: PacBio

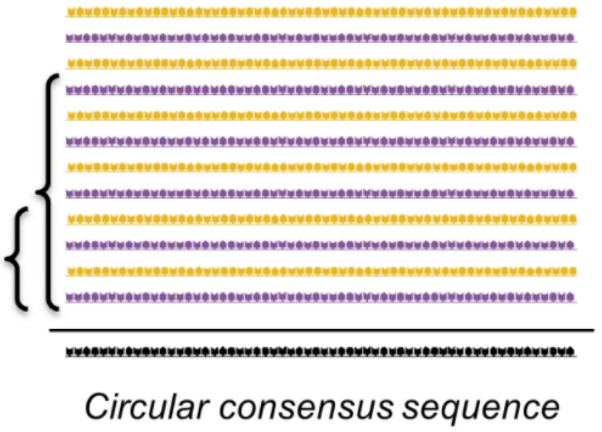
NEW CHEMISTRY PERFORMANCE

Shorter-insert libraries	**v 6.0**	v 5.1
- Throughput per SMRT Cell:	**up to 50 Gb**	up to 20 Gb
- Average read length:	**up to 100 kb**	up to 40 kb



9 passes for Q30 (99.9%)

4 passes for Q20 (99%)

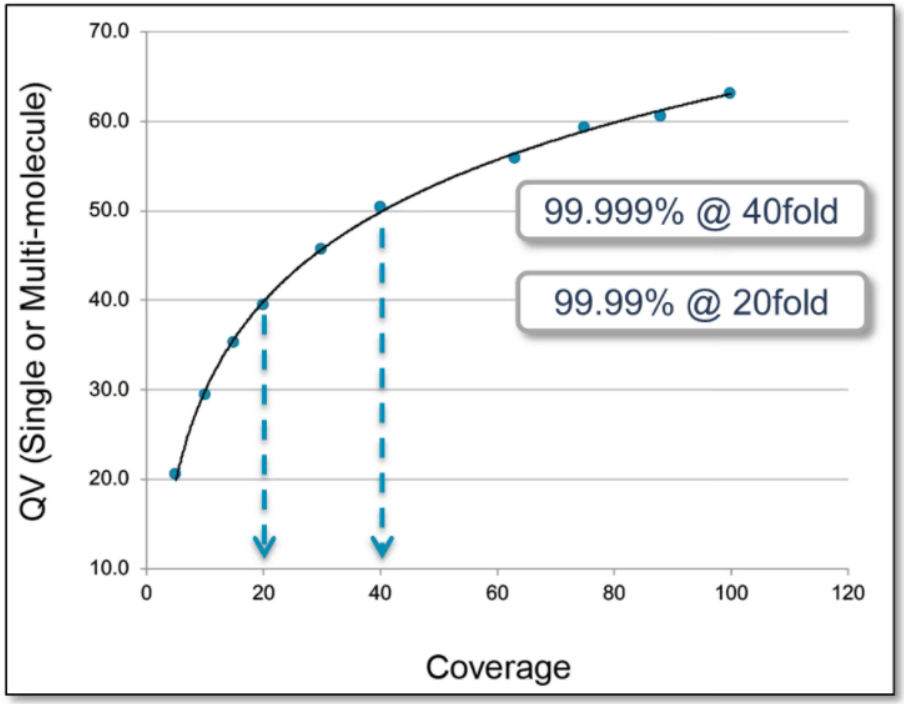


Circular consensus sequence

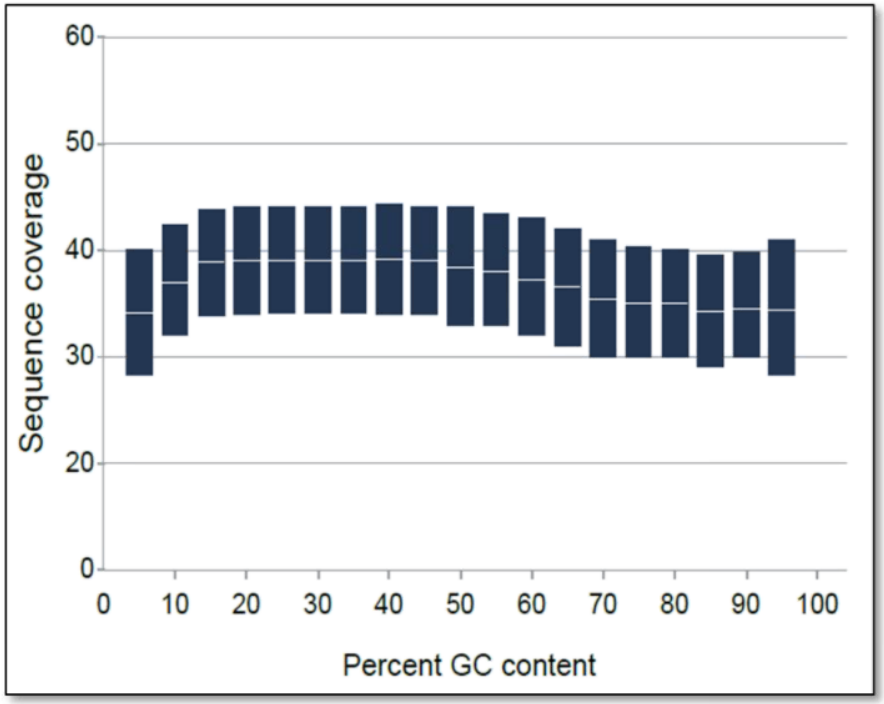
Source: PacBio

NEW CHEMISTRY PERFORMANCE

Consensus accuracy:



Minimal sequence GC% and complexity bias:



Source: PacBio

Coffee

2017:
arabica 1.3 Gb
genome
Medrano et al.

PacBio &
ChICAGO scaffolding
Scaffold N50 = 2.24 Mb
Contig N50= 1.31 Mb

2015:
robusta 0.7 Gb

UCDAVIS

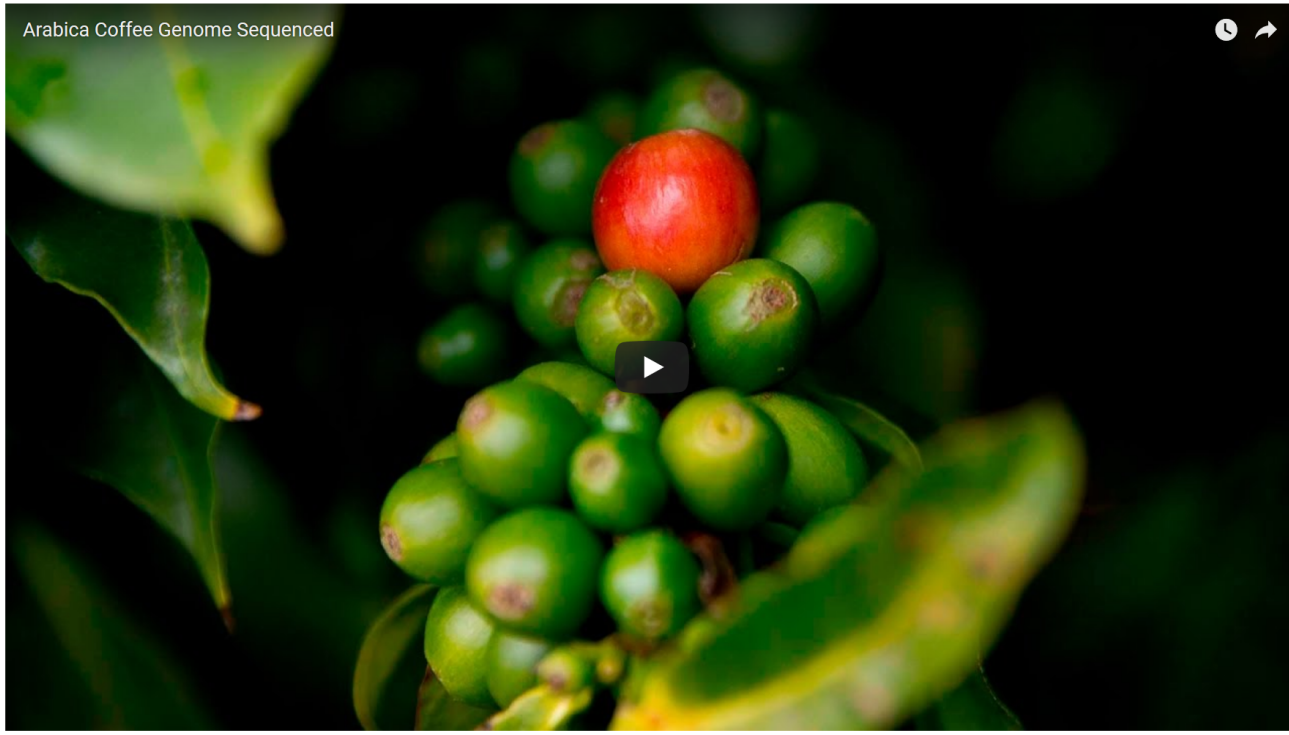
ABOUT US ADMISSIONS ACADEMICS RESEARCH

Home > News > Arabica Coffee Genome Sequenced

Arabica Coffee Genome Sequenced

Coincides With Birth of California-Grown Specialty Coffee Industry

By Pat Bailey on January 13, 2017 in Food & Agriculture



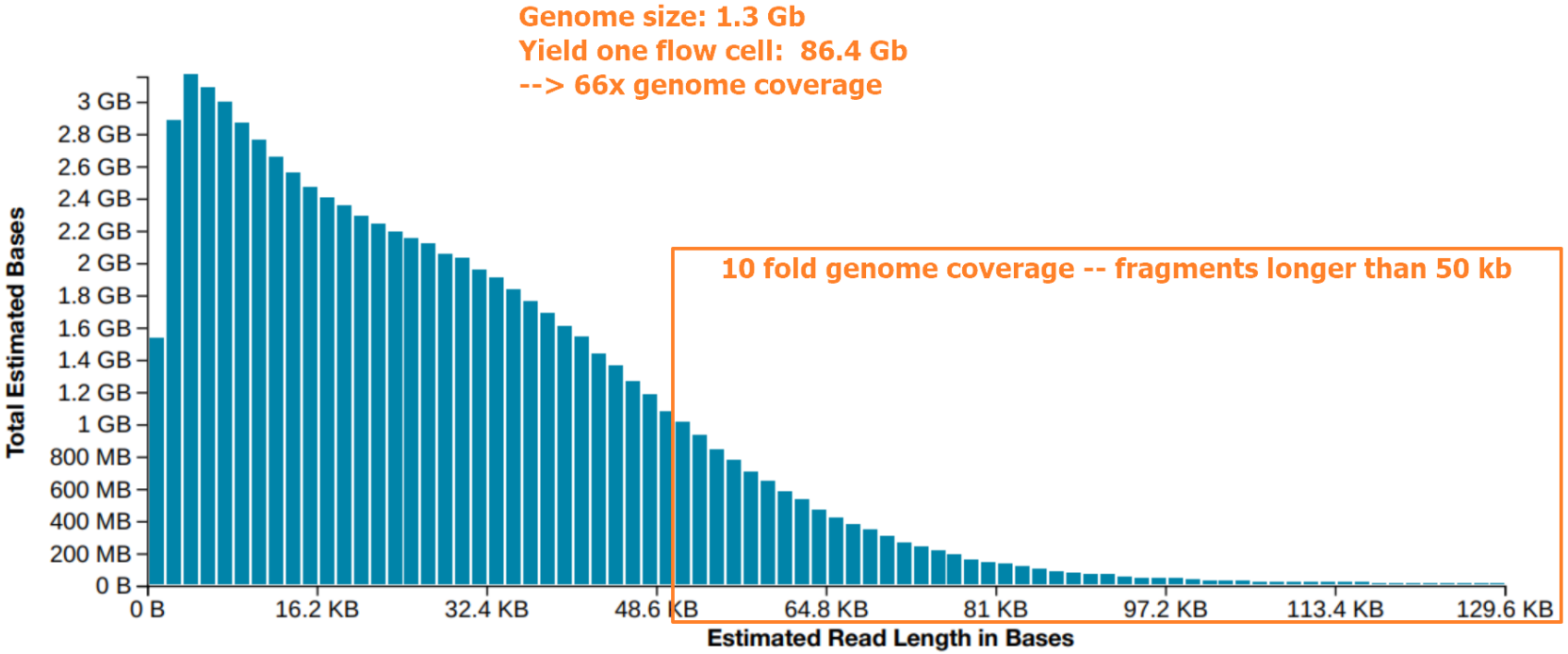
The first public genome sequence for *Coffea arabica*, the species responsible for more than 70 percent of global coffee production, was released today by researchers at the University of California, Davis.

Quick Summary

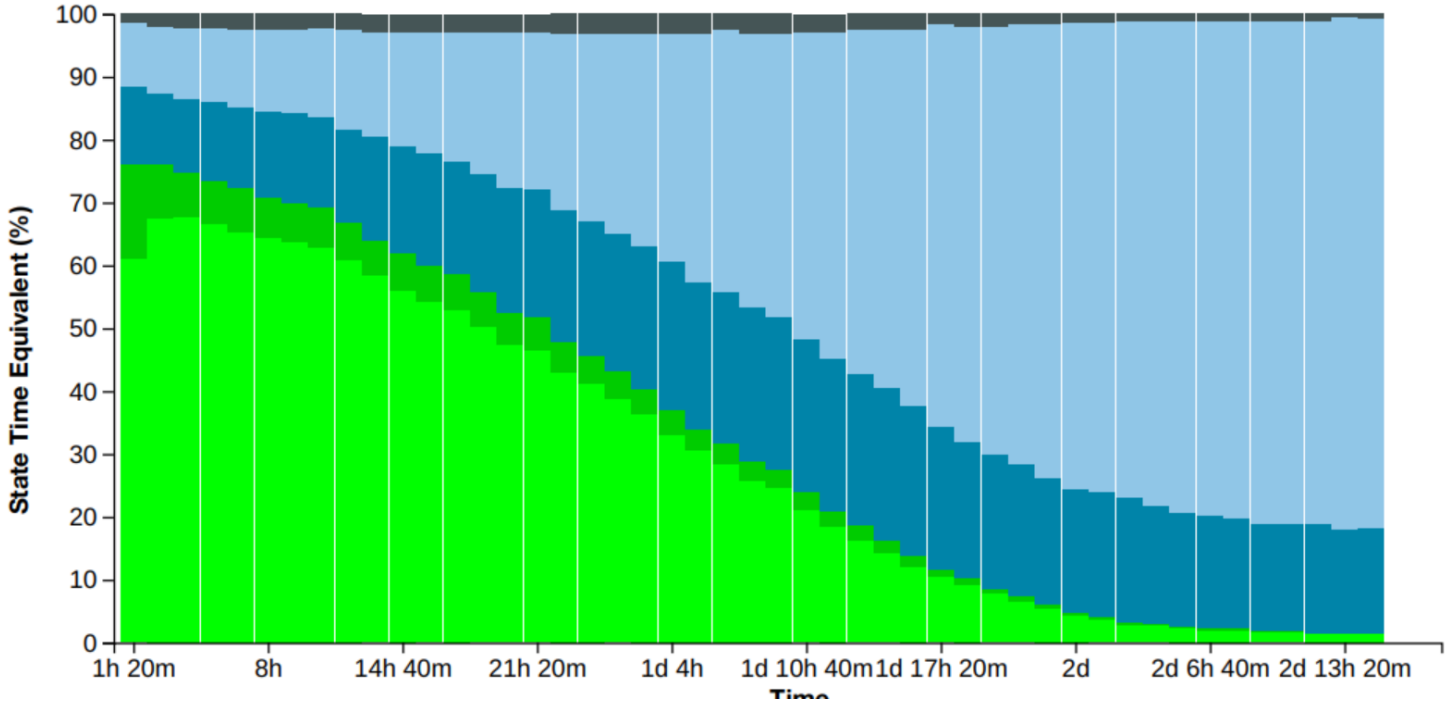
› **Will help develop disease-resistant varieties adaptable to climate**

PromethION read lengths

- RL histogram *arabica*

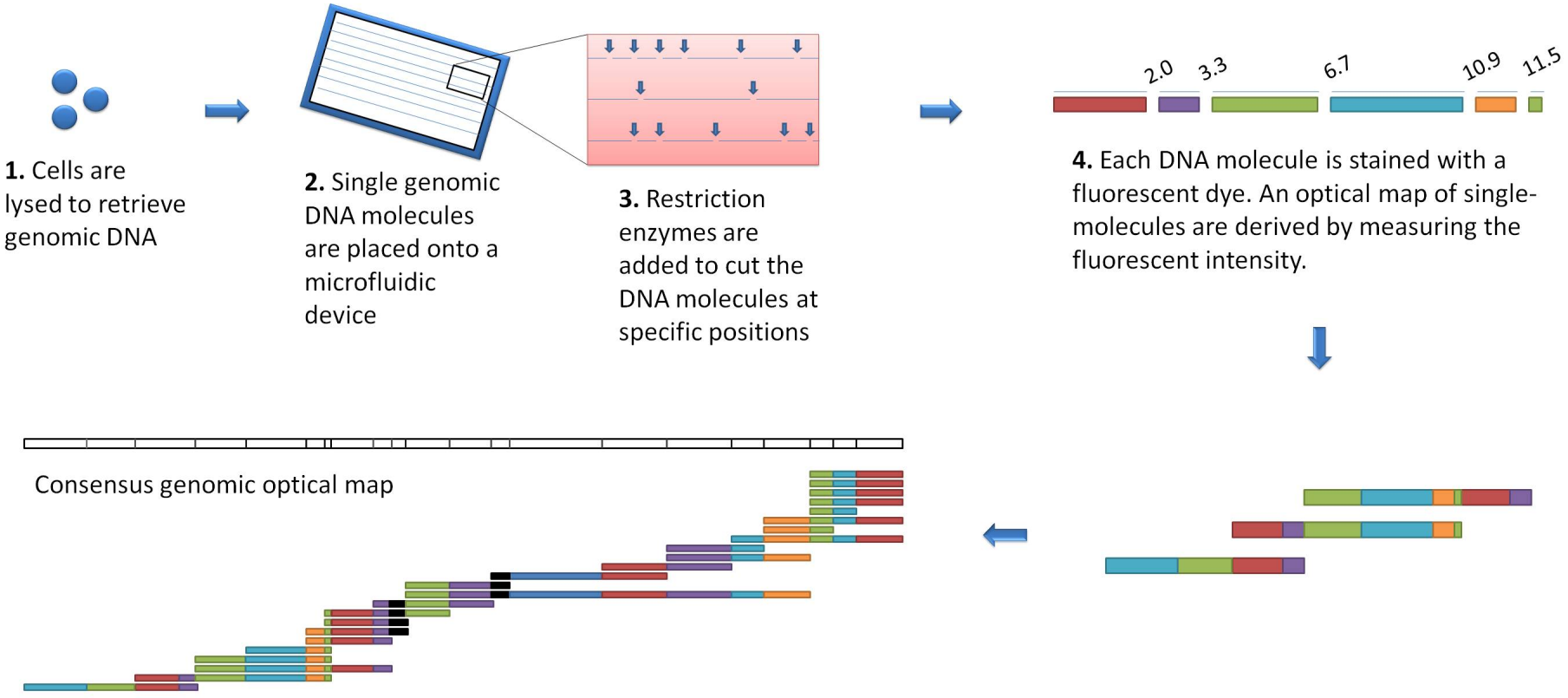


Yield over time



Bionano Optical Genome Mapping

- Contig N50 of assembly > 40 kb
- In silico digests



1. Cells are lysed to retrieve genomic DNA

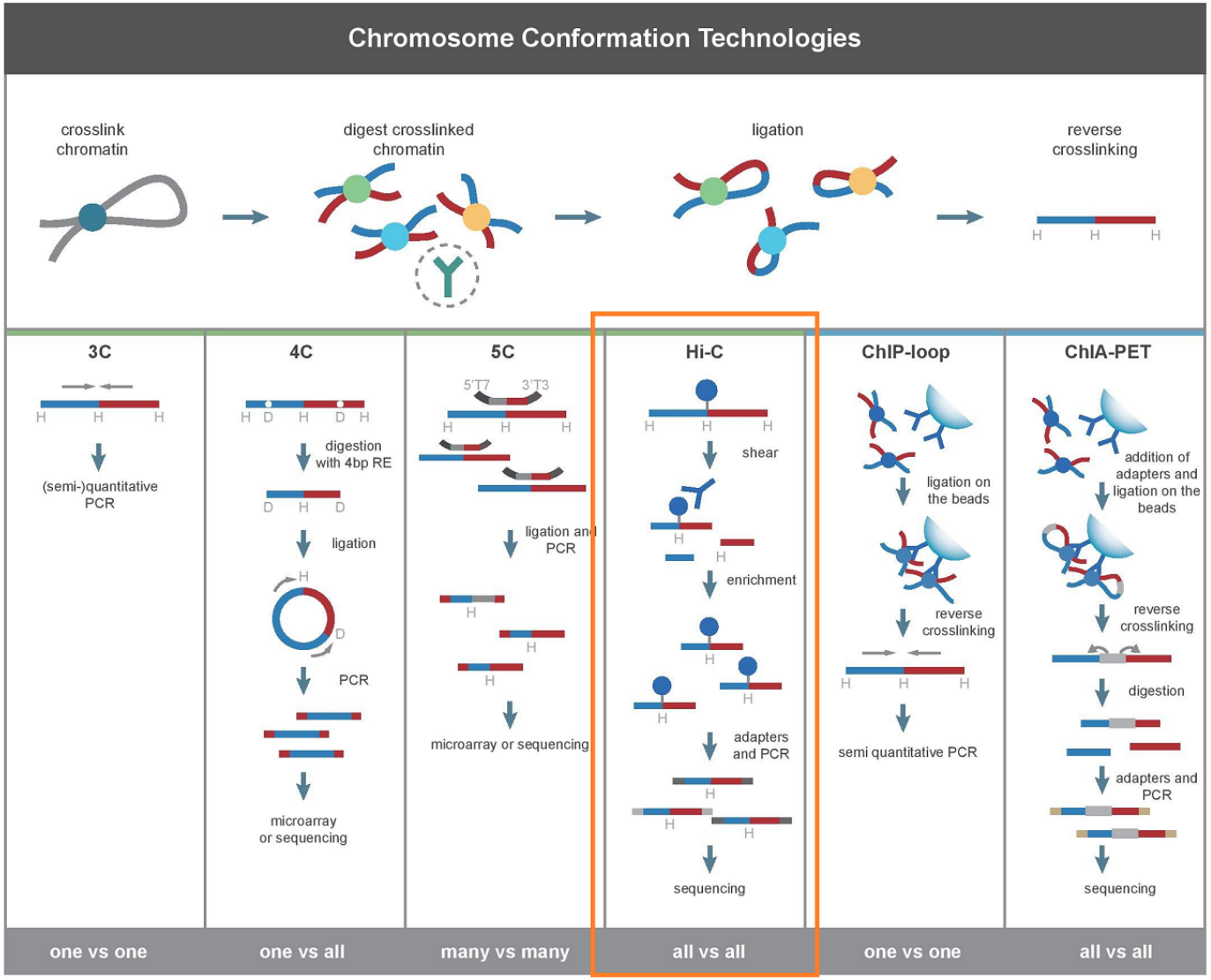
2. Single genomic DNA molecules are placed onto a microfluidic device

3. Restriction enzymes are added to cut the DNA molecules at specific positions

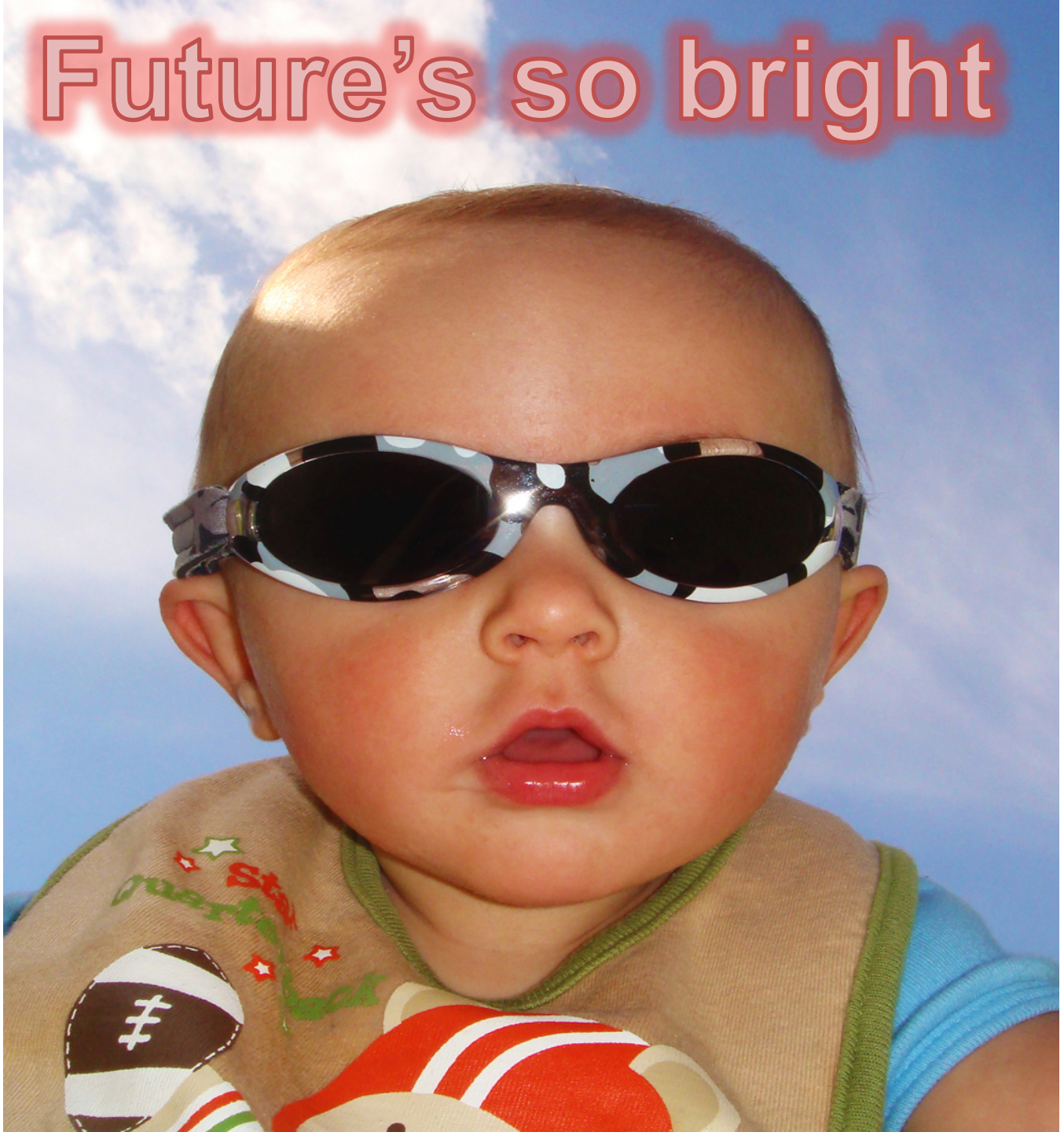
4. Each DNA molecule is stained with a fluorescent dye. An optical map of single-molecules are derived by measuring the fluorescent intensity.

5. Overlapping of the multiple single-molecule maps gives us the consensus genomic optical map

Hi-C chromosome-scale scaffolding



source: Wikipedia

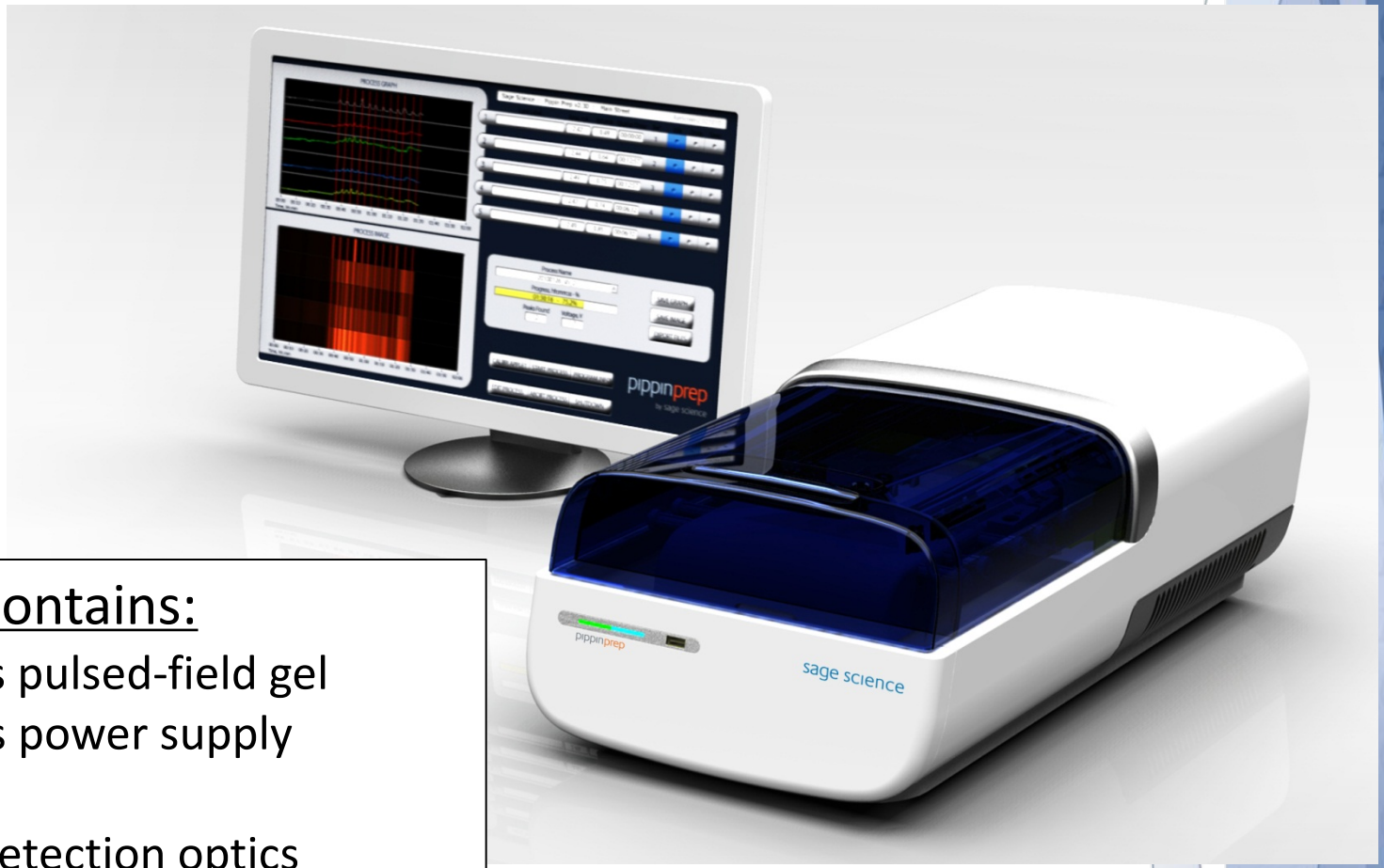

Future's so bright

Thank you!

Let’s get started!

The Blue Pippin Prep System

Automated Preparative Gel Electrophoresis for NGS



Instrument contains:

Electrophoresis pulsed-field gel electrophoresis power supply
Electrode array
Fluorescence detection optics
Single-board PC with control software

The BluePippin Prep System

Automated Preparative Gel Electrophoresis for NGS