SciLifeLab

Democratisation of Genomics and Hidden Figures of Reference Genome Sequencing Projects

Olga Vinnere Pettersson

Scientific Lead of Planetary Biology Capability

ERGA Vice-chair

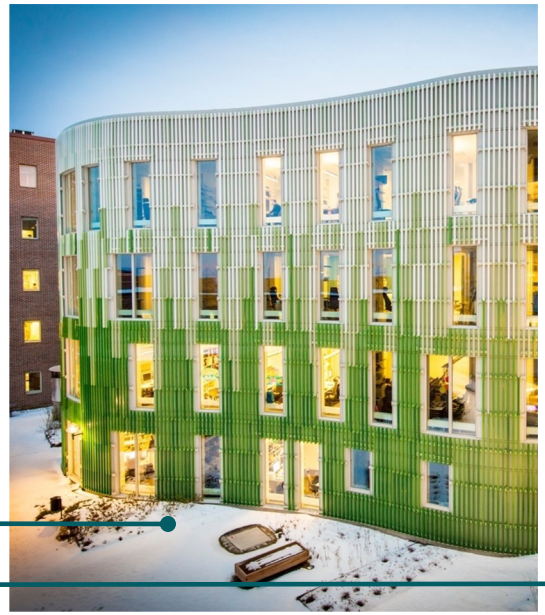
SciLifeLab / Uppsala University

Oslo, 240410

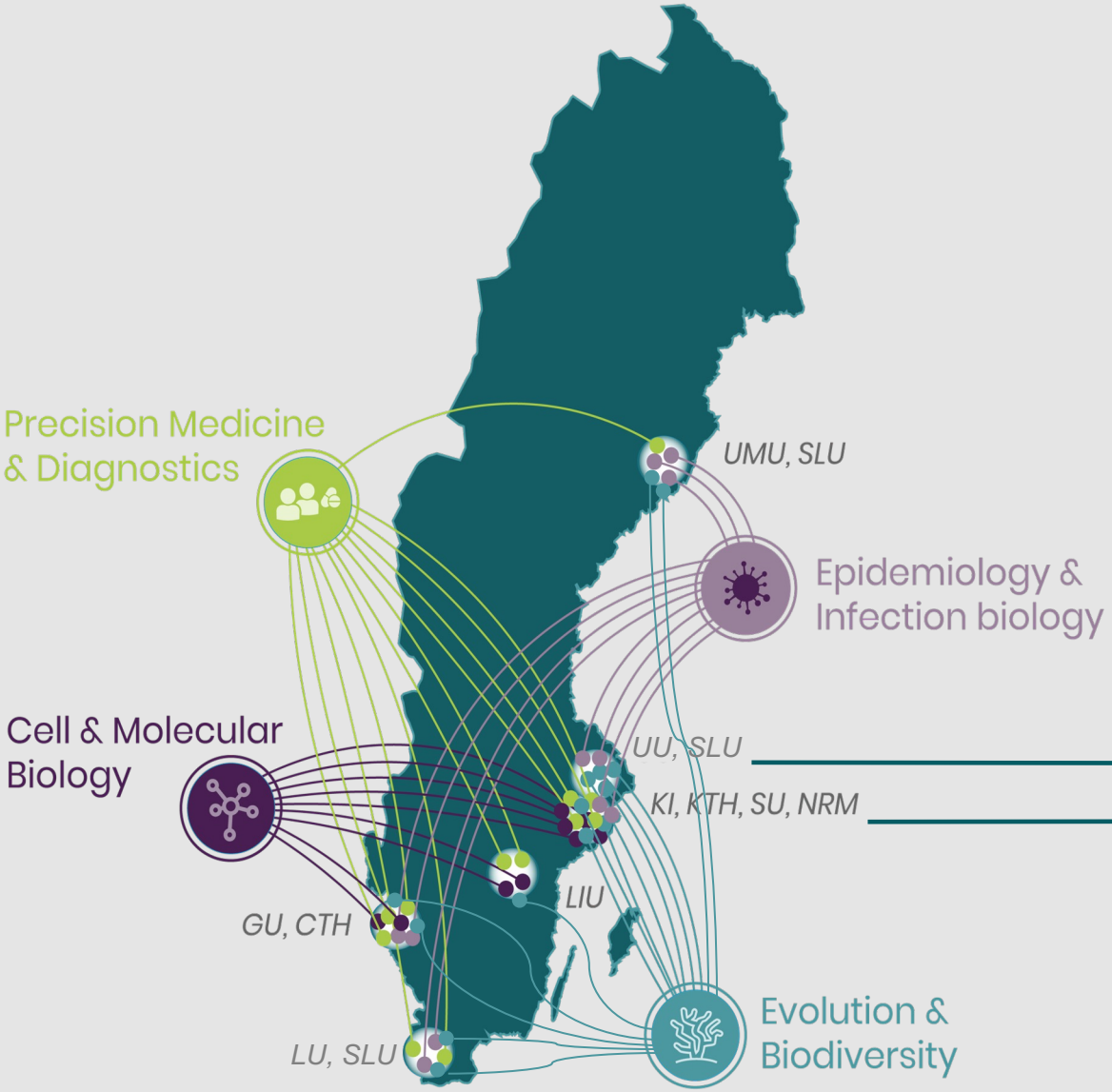
Let’s get acquainted

SciLifeLab

Sweden’s national center for molecular biosciences



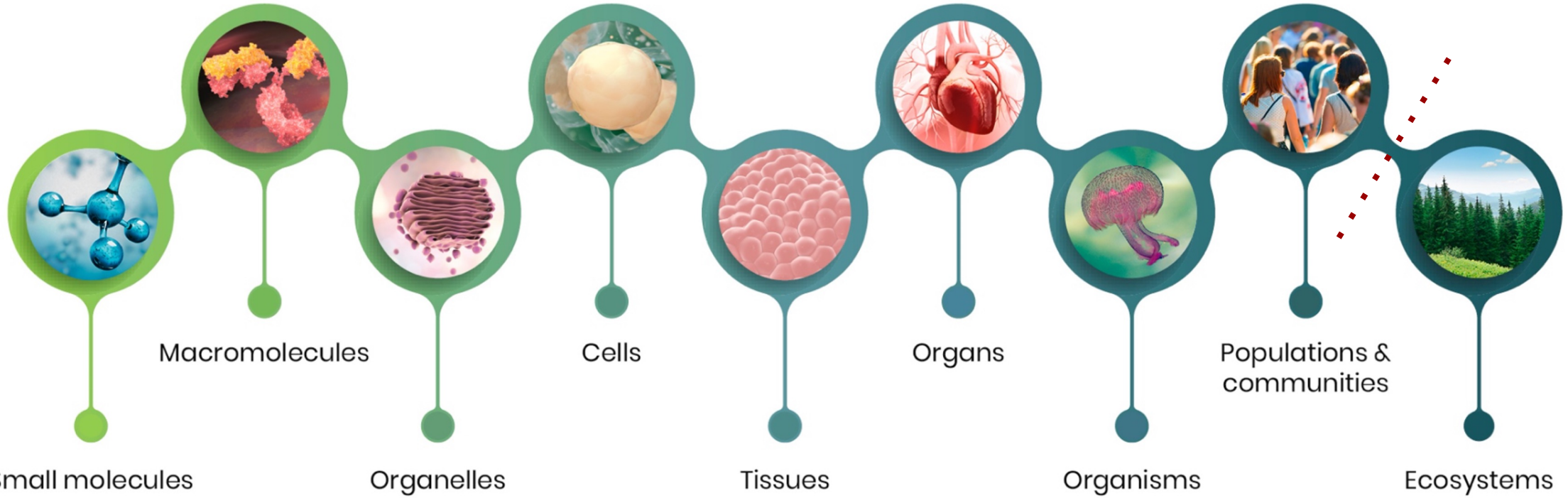
Connecting strong research environments



Enabling research across the full spectrum of life science

SciLifeLab infrastructure technologies:

- Can be used to study the molecular aspects of life **ranging from the atomic scale up to entire ecosystems**
- Are applicable across a **large spectrum of disciplines and research fields** in life science
- Are **available to all academic researchers in Sweden** on equal terms
- Are available to **healthcare and industry** all over the country, as well as international users



And now – to the topic

It all begun in late 19th century…

Johannes Friedrich Miescher
(13 August 1844 – 26 August 1895)

Die Spermatozoen einiger Wirbelthiere.

Ein Beitrag zur Histochemie*)

von

F. Miescher.

Hiezu Tafel I.

Bekanntlich wird in Basel der Fang des Rheinlachses (Salmo Salar) ziemlich lebhaft betrieben. Während der Laichzeit, im November, kann man zuweilen diese stattlichen Fische in grosser Zahl auf dem Markte sehen. Die reifen Geschlechtsprodukte dieser Thiere sind dabei als Abfall in beträchtlicher Menge zu erhalten. Die grosse Anstalt für künstliche Fischzucht in dem benachbarten Hüningen bezieht ihren ganzen Bedarf an Lachseiern, im Betrage von mehreren Millionen jährlich, von Herrn Friedrich Glaser, dem Besitzer der bedeutendsten hiesigen Fischhandlung.

Besonders verlockend ist hier für den Physiologen die Gelegenheit zur Gewinnung von Sperma. Von der rahmigen Flüssigkeit, die man als „Lachsmilch" bezeichnet, habe ich zuweilen mit Erlaubniss der Verkäufer fast einen Schoppen auf einmal
als blendend weisse Crême
lebenden Fisch; bei todten
Galle, Harn oder Blut mit

Der Samen der Fisch
vor andern werthvoll. Keine
ihre Produkte dem Sekret

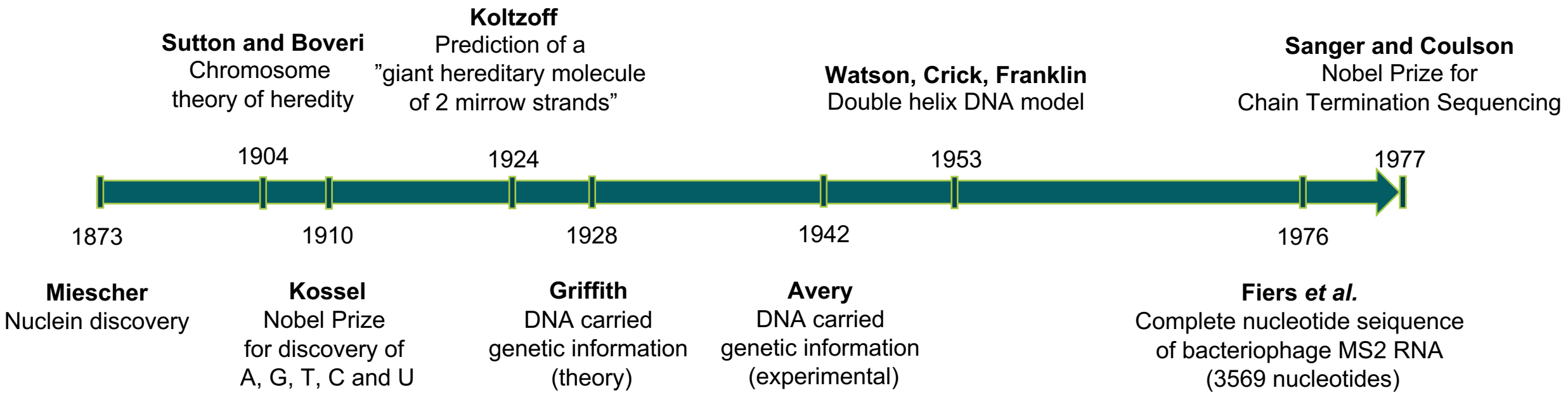
*) Nach Vorträgen, gehalte
ber 1873.

brechung der Spermatozoenköpfe erkennbar sind, ist auch das Protamin nachzuweisen.

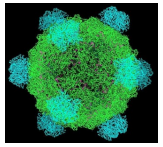
Das Nuclein.

Der Rückstand nach Extraction mit Salzsäure zeigt unter dem Mikroskop noch Hülle und Inhalt und gibt die Millon'sche Reaction. In Kochsalzlösung quillt er nicht mehr, dagegen etwas in destillirtem Wasser.

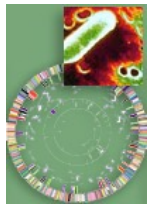
From Miescher to Sanger



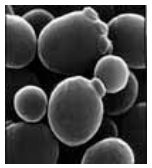
At the very beginning of genome sequencing era…



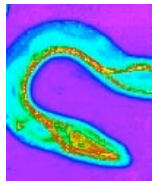
- First DNA genome: virus ϕ X 174 - 5 368 bp (1977)



- First organism: *Haemophilus influenzae* - 1.5 Mb (1995)



- First eukaryote: *Saccharomyces cerevisiae* - 12.4 Mb (1996)



- First multicellular organism: *Cenorhabditis elegans* - 100 MB (1998)

- First plant: *Arabidopsis thaliana* - 157 Mb (2000)

The Yeast Genome project

Life with 6000 Genes

A. GOFFEAU, B. G. BARRELL, H. BUSSEY, R. W. DAVIS, B. DUJON, H. FELDMANN, F. GALIBERT, J. D. HOHEISEL, C. JACQ, [...], AND S. G. OLIVER

& Affiliations

SCIENCE · 25 Oct 1996 · Vol 274, Issue 5287 · pp. 546-567 · DOI: 10.1126/science.274.5287.546

"The genome of the yeast Saccharomyces cerevisiae has been completely sequenced through ***an international effort involving some 600 scientists in Europe, North America, and Japan.*** *It is the largest genome to be completely sequenced so far (a record that we hope will soon be bettered) and is the first complete genome sequence of a eukaryote."*

"New graduate students are already wondering how we all managed in the "dark ages" before the sequence was completed. We must now tackle a much larger challenge, that of elucidating the function of all of the novel genes revealed by that sequence. ***As with the sequencing project itself, functional analysis will require a worldwide effort.*** *In Europe, a new research network called EUROFAN [for European Functional Analysis Network has been established to undertake the systematic analysis of the function of novel yeast genes. Parallel activities are underway in Germany, Canada, and Japan. In the United States, the National Institutes of Health has recently sent out a request for applications for "Large-Scale Functional Analysis of the Yeast Genome." "*

GENETICS

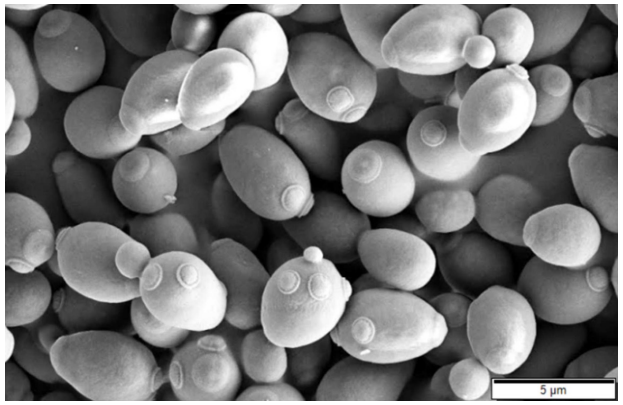
Genetics. 2013 Jun; 194(2): 291–299.
doi: 10.1534/genetics.113.151258

PMCID: PMC3664842
PMID: 23733847

The Modest Beginnings of One Genome Project

David B. Kaback[1]

First genomic references



1/3 of genes related to human by homology

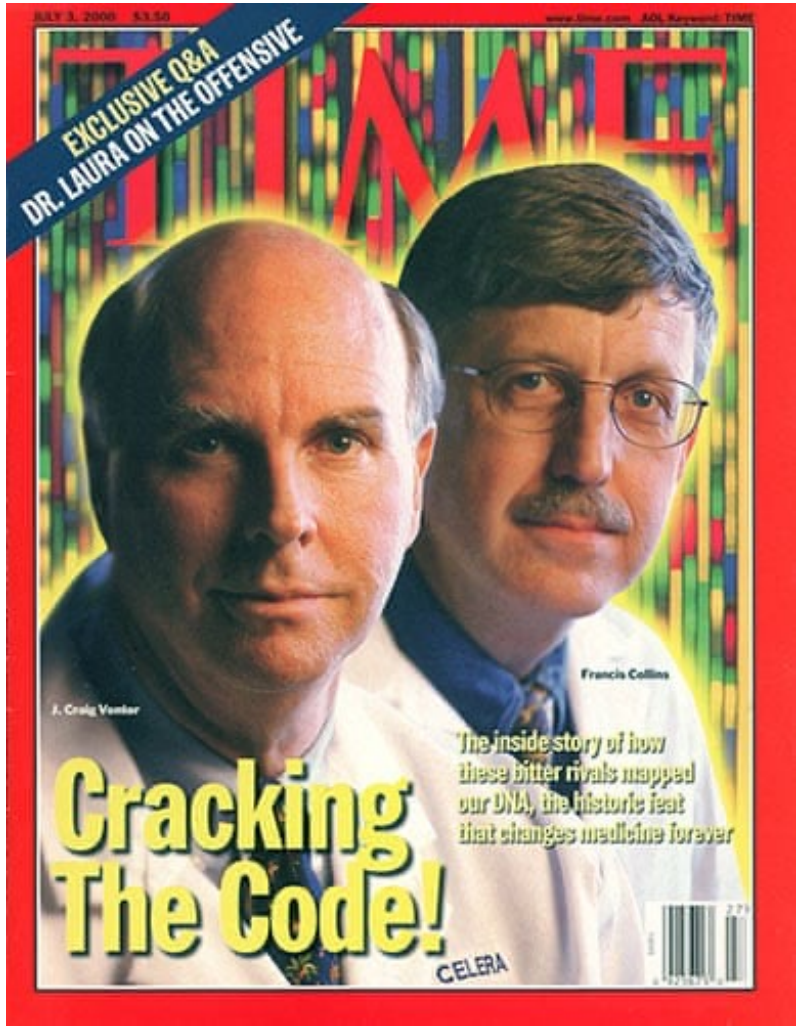
Basic cell functions

Human disease gene discovery

Phenotypical traits

Plant genome structure and function

But it was not enough…



GENOME SEQUENCING WORKSHOP

MARCH 3 & 4, 1986

SANTA FE, NEW MEXICO

SPONSOR

DOE

OFFICE OF HEALTH AND

ENVIRONMENTAL RESEARCH

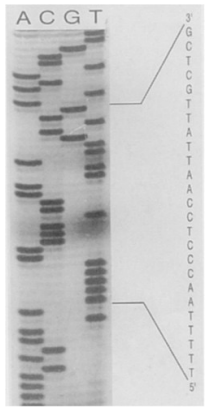
HOST

LIFE SCIENCES DIVISION

LOS ALAMOS NATIONAL LABORATORY

It is thus important that we identify here what real benefits and liabilities might emerge from the contemplated sequencing activity, which would aim at capturing the entire human genome in a period of 10 or 12 years. Do we have the technologies necessary to do this, and do we have the computational power and algorithms needed to integrate and anaylze this data? Will this information provide both clinical and basic benefits of such magnitude to warrant an accelerated effort?

An outcome of HUGO – Genomic Revolution



1977

Sanger

P^{32} labeled gels

1990 HUGO

2000

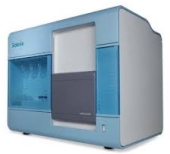
Sanger

ABI 3000 series

NGS

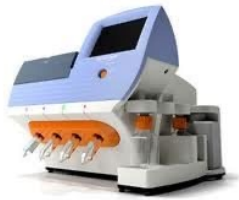
2005

454



Solexa
ABI SOLiD

2006



2010

IonTorrent

PacBio
ONT

2012

2021 HiFi reads

2023

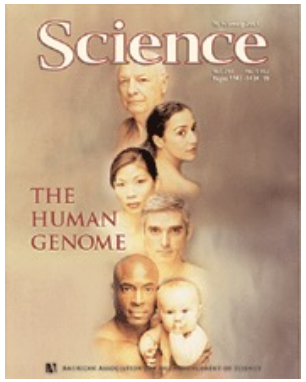
PacBio
REVIO

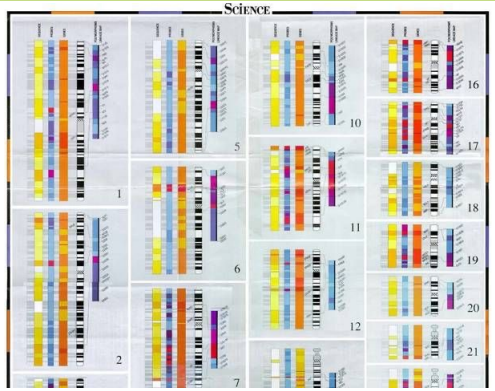
2024 WE ARE HERE

Just a comparison

1990 - 2003
HUGO

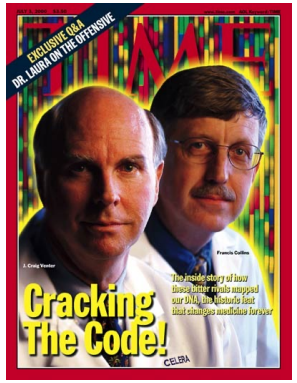
Sanger traditional
$2.7 bln





2007
Craig Venter's genome

Sanger ABI 3730
$300 mln



2008
Jim Watson's genome

454 FLX
$1 mln

TODAY
any human

$800 with Illumina
$1-3k with long reads

Outcome of genomics revolution: paradigm shift

Single genes

Complete genomes

Outcome of genomics revolution: paradigm shift

Single genes

Single transcripts

Complete genomes

Whole transcriptomes

Outcome of genomics revolution: paradigm shift

Single genes	Complete genomes
Single transcripts	Whole transcriptomes
Single organisms	**Metagenomes**

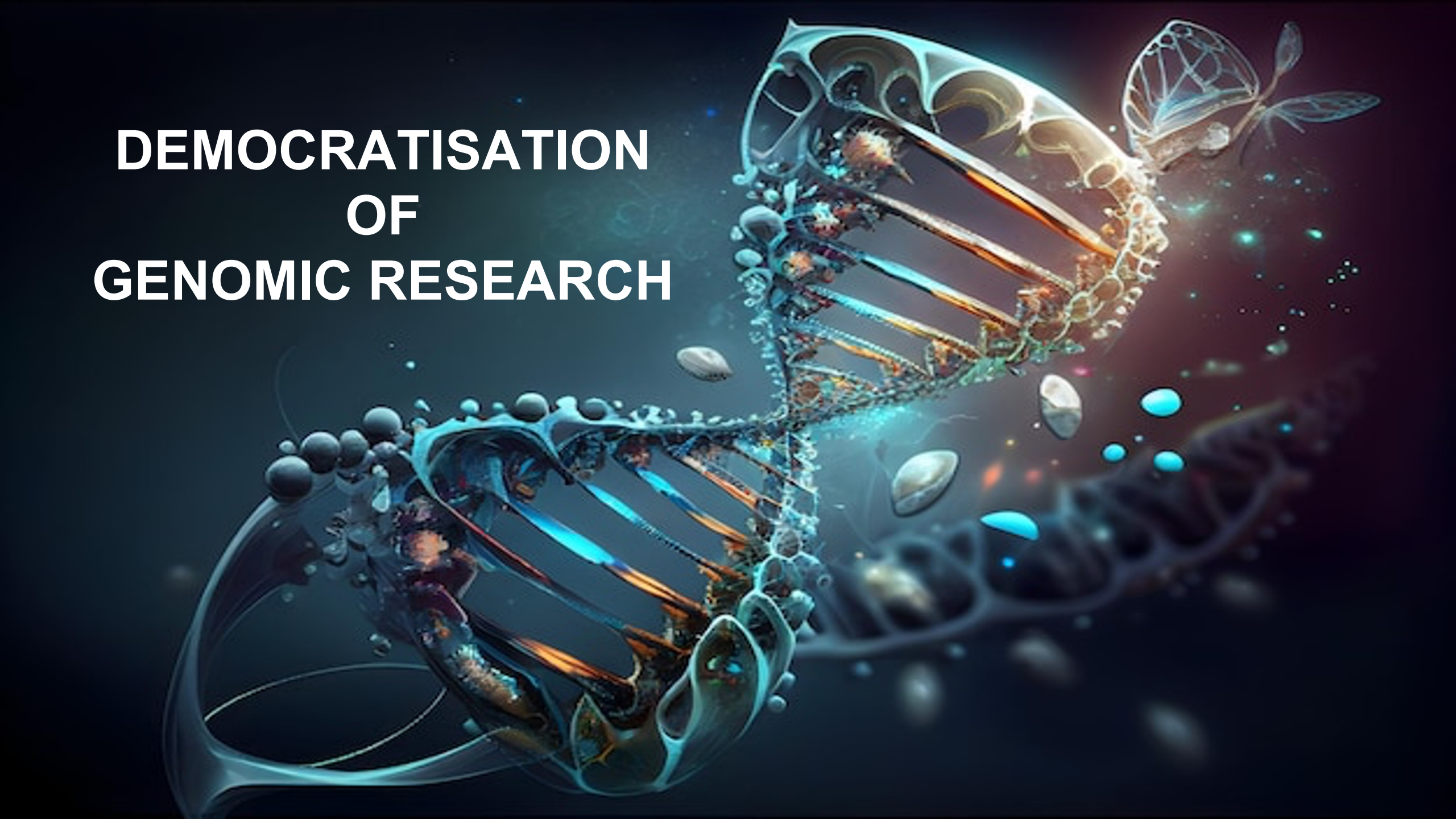
Outcome of genomics revolution: paradigm shift

Single genes	Complete genomes
Single transcripts	Whole transcriptomes
Single organisms	Metagenomes
Model organism	**Any species**

Outcome of genomics revolution: paradigm shift

Single genes	Complete genomes
Single transcripts	Whole transcriptomes
Single organisms	Metagenomes
Model organism	Any species
Available to highly specialized labs	**Available to anyone**

DEMOCRATISATION OF GENOMIC RESEARCH



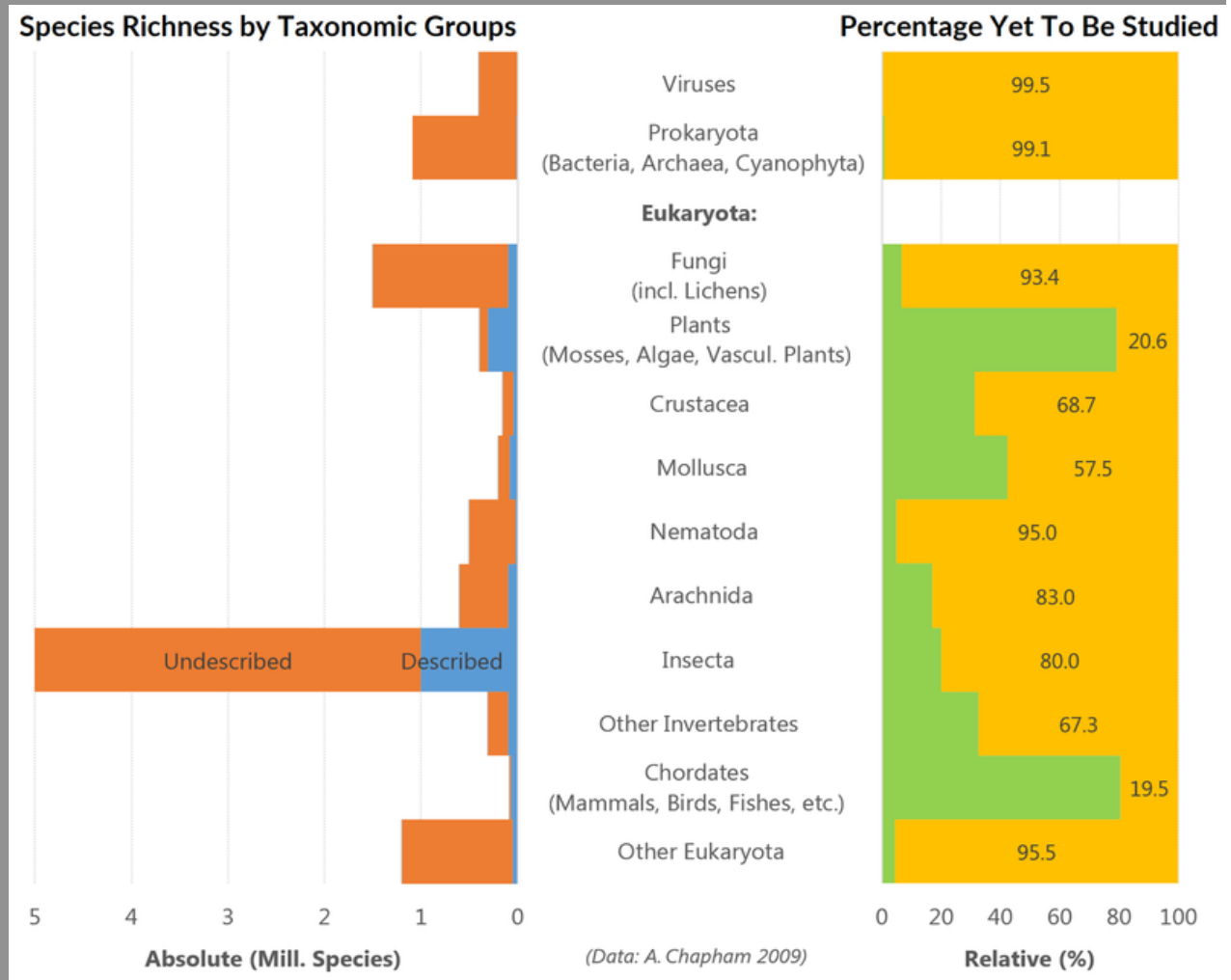
Moving beyond HUGO

Homo sapiens

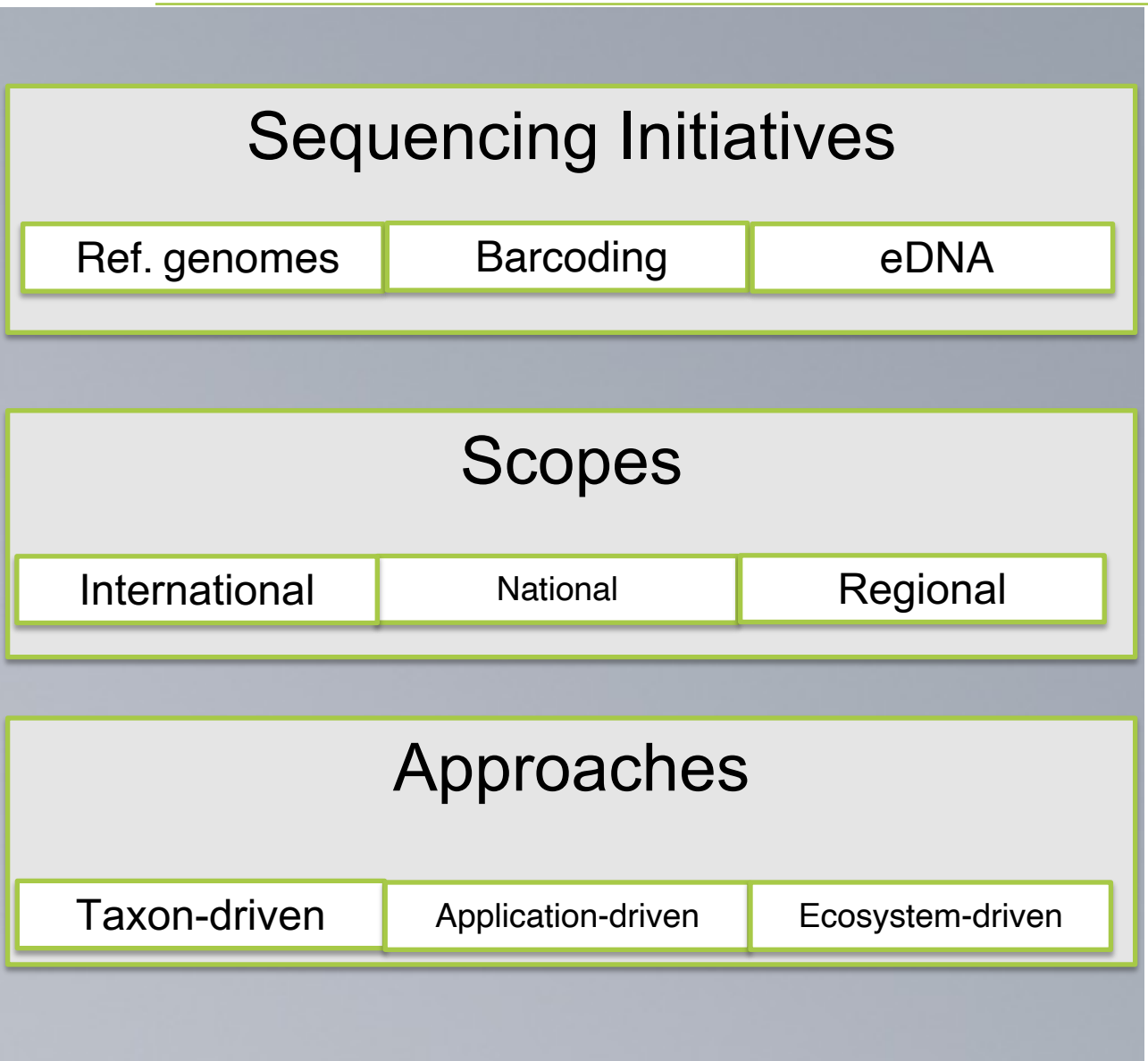
N species = 1

Global Biodiversity

N species = 1 trillion



Why international sequencing efforts?



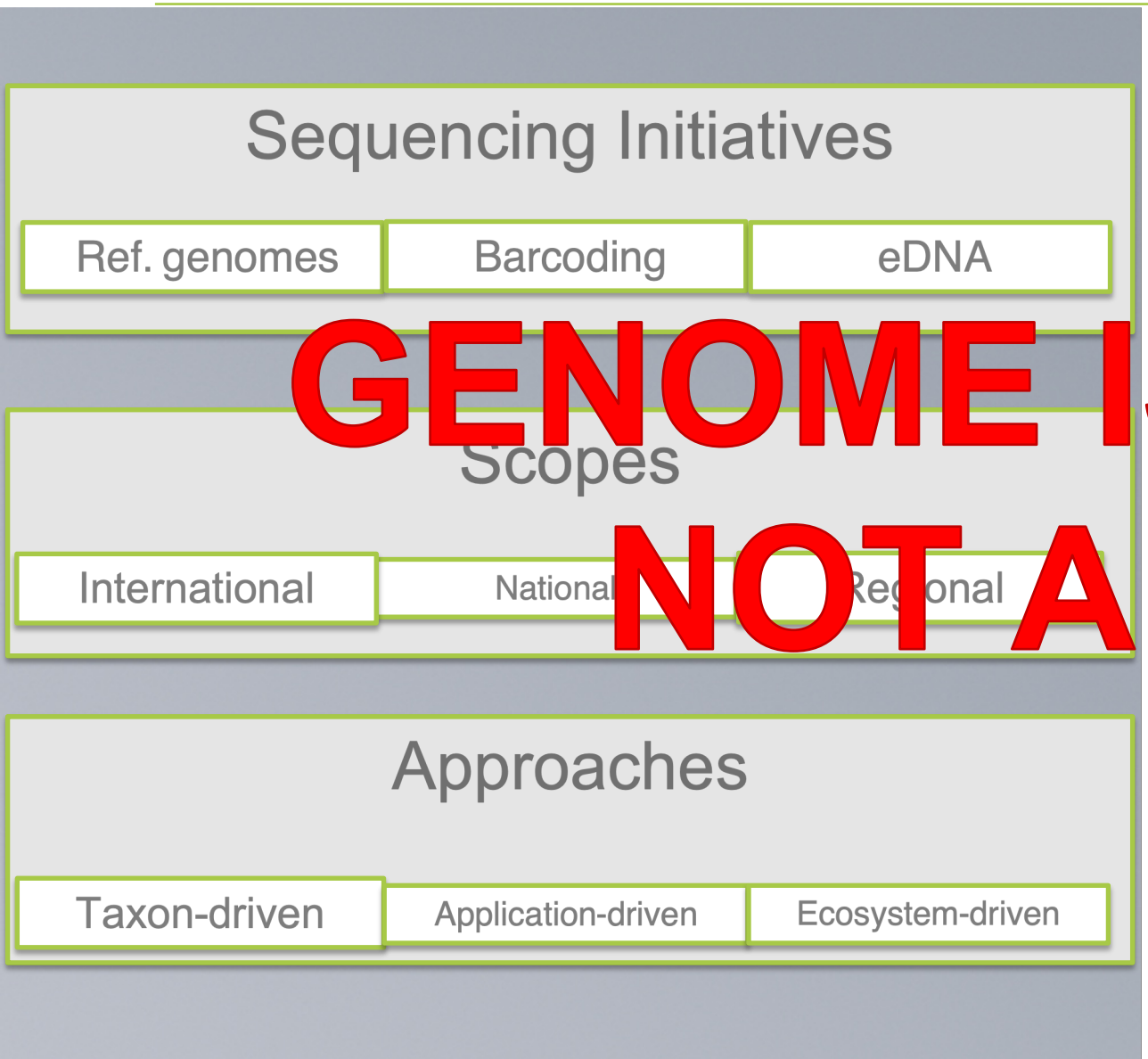
Creating an *open* data repository

- **Biodiversity decline**
- Enabling many disciplines of biology

- Sequencing is still expensive
- Genome is a tool, not a goal
- Standardised methods

- FAIR, ethical and legal

Why international sequencing efforts?



GENOME IS A TOOL

NOT A GOAL

Creating an *open* data repository

- **Biodiversity decline**
- Enabling many disciplines of biology
- Sequencing is still expensive
- Genome is a tool, not a goal
- Standardised methods
- FAIR, ethical and legal

ANYONE CAN SEQUENCE A GENOME!

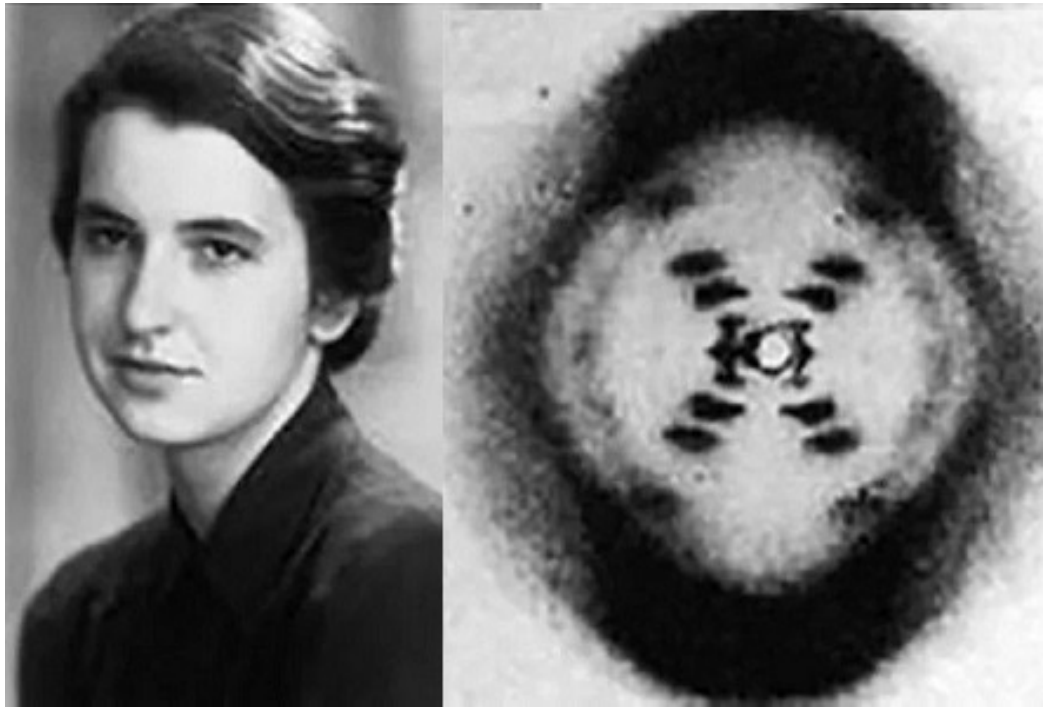
Or can they?!

Hidden Figures

Katherine Johnson
Dorothy Vaughan
Mary Jackson

Hidden Figures

Katherine Johnson
Dorothy Vaughan
Mary Jackson



Rosalind Franklin

Hidden Figures

Alexander Fleming

Hidden Figures

Alexander Fleming

Howard Florey
Ernst Chain

Hidden Figures

Alexander Fleming

Howard Florey
Ernst Chain

Norman Heatley

Hidden Figures in Genomics: reasons

Shift from specialized labs to sequencing facilities

Long-read sequences are VITAL for reference genome generation

Long reads conundrum:

- Heavily reliant on pure, HMW-DNA
- High failure rate both for PacBio and ONT
- Everything is non-model

Hidden Figures: reasons

Shift from specialized labs to sequencing facilities

Long-read sequences are VITAL for reference genome generation

Long reads conundrum:

Heavily reliant on pure, HMW-DNA

Unpredictable yields

Everything is non-model

THE NUCLEIC ACIDS

Chemistry and Biology

Edited by

ERWIN CHARGAFF
Department of Biochemistry
Columbia University
New York, N. Y.

J. N. DAVIDSON
Department of Biochemistry
University of Glasgow
Glasgow, Scotland

Volume I

a. Extraction with Strong Salt Solution. Deproteinization with Chloroform

(1) *Sodium Deoxyribonucleate of Calf Thymus.*[98] Fresh frozen calf thymus glands (54.5 kg.) were minced and suspended in 0.9% sodium chloride (54 l.) and milled to produce a fine suspension. This suspension was centrifuged (6300 r.p.m.) and the solid material resuspended in 0.9% sodium chloride (45.5 l.) and milled and centrifuged as before. The tissues, which were now free of material containing pentose, were suspended in 10% sodium chloride (214 l.) with vigorous mechanical stirring at 0°. At this stage the viscosity of the solution increased considerably. After extraction at 0° for 48 hours, the insoluble material was removed by centrifuging (6300 r.p.m.) and the deoxypentose nucleoprotein precipitated from the resultant solution (pH 6.5) by the addition of an equal volume of industrial methanol. The precipitated solid was washed with 70%, then 100% industrial methanol and dried in a vacuum at room temperature. Yield, 1.69 kg. of a very slightly yellow fibrous solid.

THE PREPARATION OF DEOXYRIBONUCLEIC ACIDS BY THE *p*-AMINOSALICYLATE–PHENOL METHOD

K. S. KIRBY
Chester Beatty Research Institute, Institute of Cancer Research, Royal Cancer Hospital, London (Great Britain)

(Received February 17th, 1959)

A general method for isolation of high molecular weight DNA from eukaryotes

Nikolaus Blin and Darrel W. Stafford

Department of Zoology, University of North Carolina, Chapel Hill, NC 27514, USA

Received 24 June 1976

ABSTRACT

A new method for isolation of high molecular weight DNA from eukaryotes is presented. This procedure allows preparation of DNA from a variety of tissues such as calf thymus or human placenta and from cells which were more difficult to lyse until now (e.g. Crypthecodinium cuhnii, a dinoflagellate). The DNA obtained in such a way has an average molecular weight of about 200 x 10^6 d and contains very few, if any, single strand breaks.

INTRODUCTION

Isolation of large quantities of nick-free, high molecular weight DNA from eukaryotic organisms has heretofore presented considerable technical difficulties. DNA prepared by conventional techniques has been a heterogeneous population of molecules ranging in molecular weight from 10 x 10^6

1983: P C R

Protocol | Published: November 1990

A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue

Thomas H. Tai & Steven D. Tanksley

Plant Molecular Biology Reporter **8**, 297–303(1990) | Cite this article

Journal of Microbiological Methods
Volume 19, Issue 3, March 1994, Pages 167-172

A general method for the extraction of DNA from bacteria

Michael W Lema, Arnold Brown, Jo H Calkins

Show more

https://doi.org/10.1016/0167-7012(94)90066-3

A simple, rapid, inexpensive and widely applicable technique for purifying plant DNA

S Gilmore, PH Weston and JA Thomson

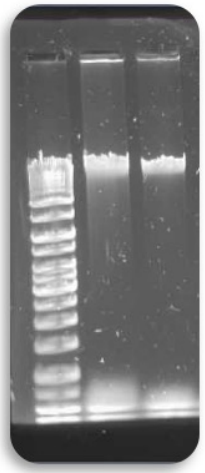
Australian Systematic Botany 6(2) 139 - 148
Published: 1993

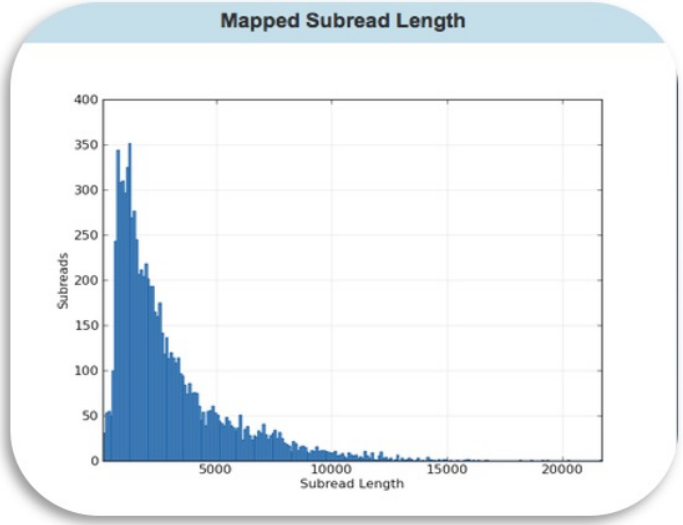
Simple, Efficient, and Nondestructive DNA Extraction Protocol for Arthropods

Aloysius J. Phillips, Chris Simon

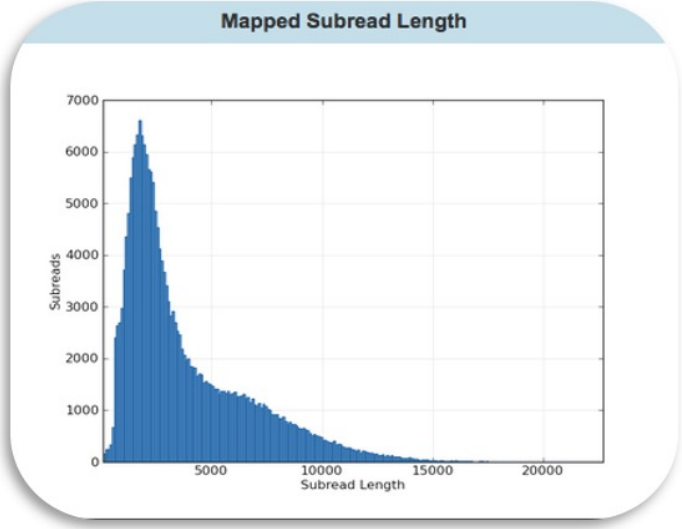
Annals of the Entomological Society of America, Volume 88, Issue 3, 1 May 1995, Pages 281–283, https://doi.org/10.1093/aesa/88.3.281
Published: 01 May 1995 **Article history** ▾

Chemically PURE HMW-DNA





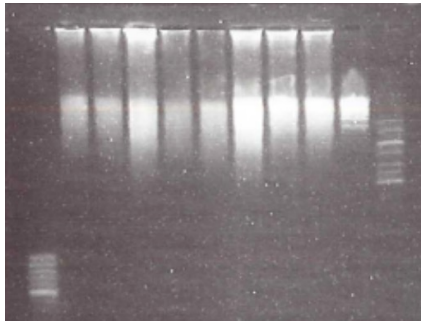
Polished Contigs	223	Max Contig Length	36,298
N50 Contig Length	2,932	Sum of Contig Lengths	480,087



Polished Contigs	9	Max Contig Length	1,508,929
N50 Contig Length	1,353,702	Sum of Contig Lengths	7,813,244

Callosobruchus maculatus

1: High-salt / ethanol protocol, whole body -> black DNA -> **FAIL**
2. MagAttract, entire body ->
260/280 = 2.1 – 2.2
260/230 = 0.32 - 0.45 -> **FAIL**
3. MagAttract, muscle -> A260 within range



-> **FAIL**

The genomic footprint of sexual conflict

Ahmed Sayadi, Alvaro Martinez Barrio, Elina Immonen, Jacques Dainat, David Berger, Christian Tellgren-Roth, Björn Nystedt & Göran Arnqvist

Nature Ecology & Evolution **3**, 1725–1730(2019) | Cite this article

2467 Accesses | **75** Altmetric | Metrics

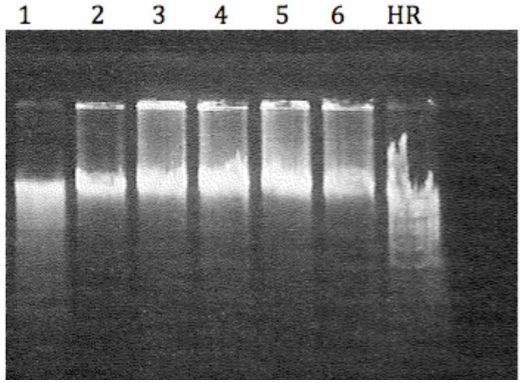
The Evolution of Dark Matter in the Mitogenome of Seed Beetles

Ahmed Sayadi, Elina Immonen, Christian Tellgren-Roth, Göran Arnqvist Author Notes

Genome Biology and Evolution, Volume 9, Issue 10, October 2017, Pages 2697–2706, https://doi.org/10.1093/gbe/evx205

Published: 27 September 2017 **Article history**

4. MagAttract -> Zymo DNA purification -> A260 low -> **FAIL**
5. High-salt / ethanol -> Zymo DNA purification -> A260 low -> **FAIL**
6. GenomicTip 20G stand-alone, muscle -> A260 in range -> **PASS**



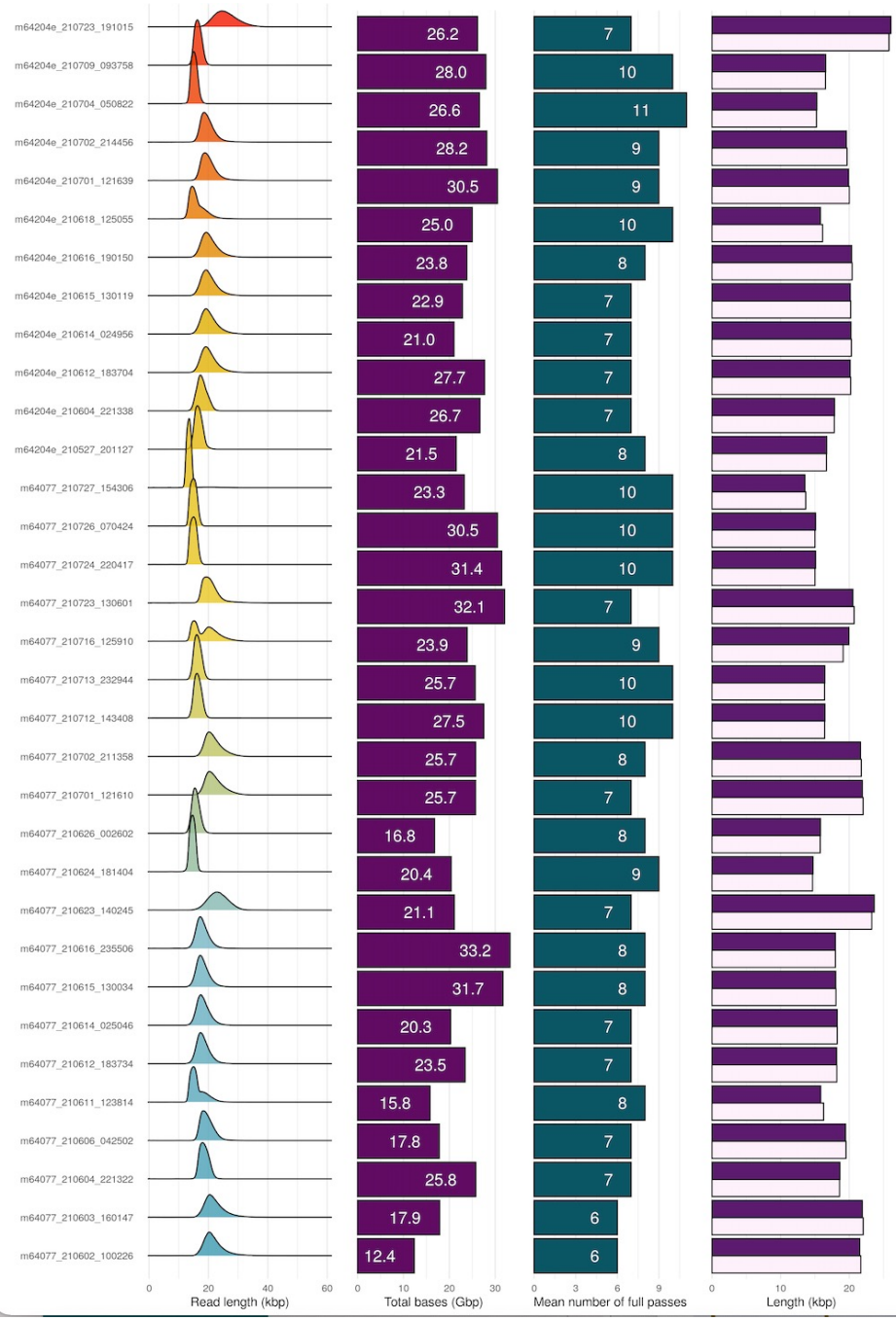
Hidden Figures: reasons

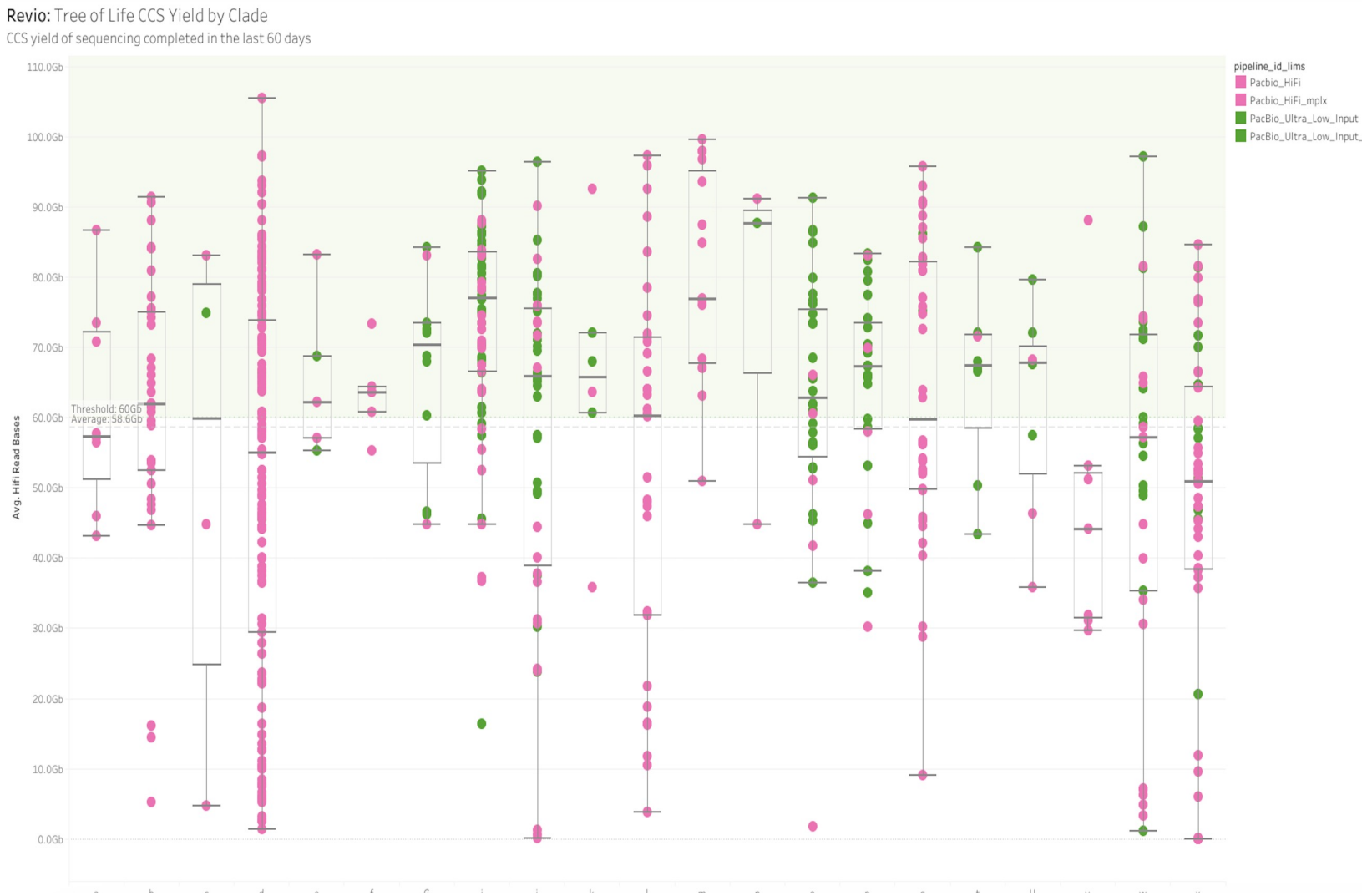
Shift from specialized labs to sequencing facilities

Long-read sequences are VITAL for reference genome generation

Long reads conundrum:

- Heavily reliant on pure, HMW-DNA
- **Unpredictable yield**
- Everything is non-model





Courtesy: James Watt, Sanger Institute (DToL)

Hidden Figures: reasons

Shift from specialized labs to sequencing facilities

Long-read sequences are VITAL for reference genome generation

Long reads conundrum:

- Heavily reliant on pure, HMW-DNA
- Unpredictable yield
- **Everything is non-model**

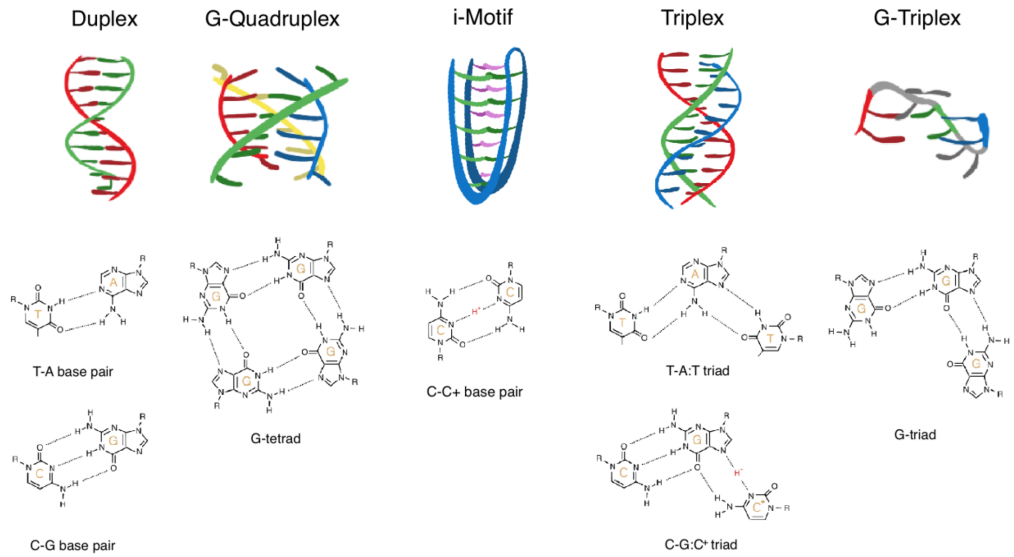
What do we know about non-models?

Biochemistry

Carry-over contaminants

- Proteins
- Polyphenols
- Secondary metabolites (e.g. toxins)
- Pigments
- Polysaccharides

DNA tertiary structures



Open Access Review

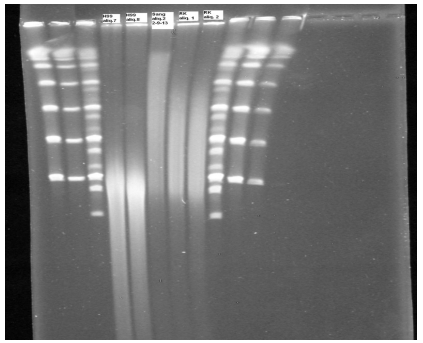
The Dynamic Regulation of G-Quadruplex DNA Structures by Cytosine Methylation

by Aaron John Stevens [1,*], Lucy de Jong [1] and Martin Alexander Kennedy [2]

[1] Department of Pathology and Molecular Medicine, University of Otago, Wellington 6021, New Zealand
[2] Department of Pathology and Biomedical Science, University of Otago, Christchurch 8011, New Zealand
* Author to whom correspondence should be addressed.

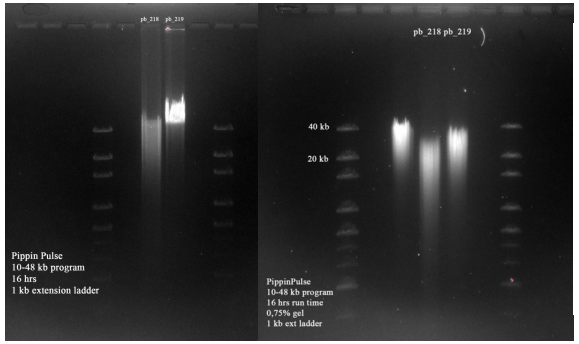
A strain of maize

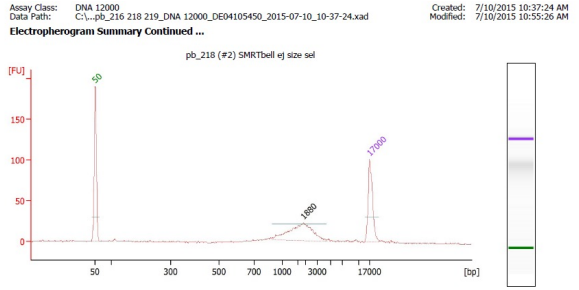
June 26, 2015



260/280 = 1.8
260/230 = 2.0

June 26 – Aug 25 2015





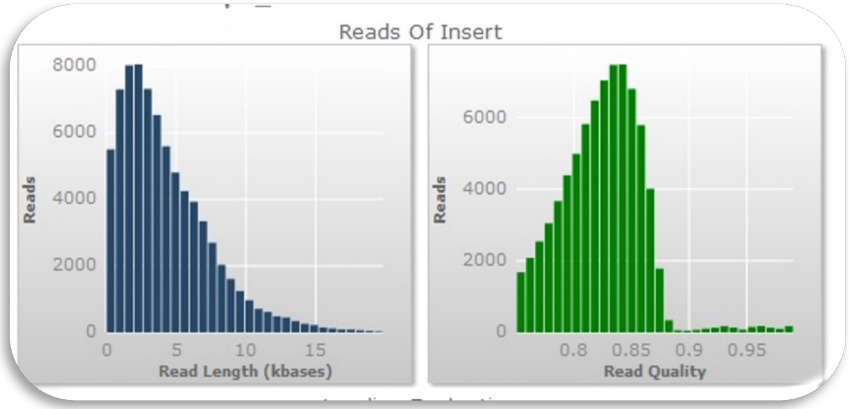
DNA is too fragile; 2 kb insert libraries instead of 20 kb

AUG 25, 2015 – FEB 22, 2016

Troubleshooting

Spectral analysis: DNA chemically pure
Hypothesis: nicks; abundance of transposons / low complexity repeats

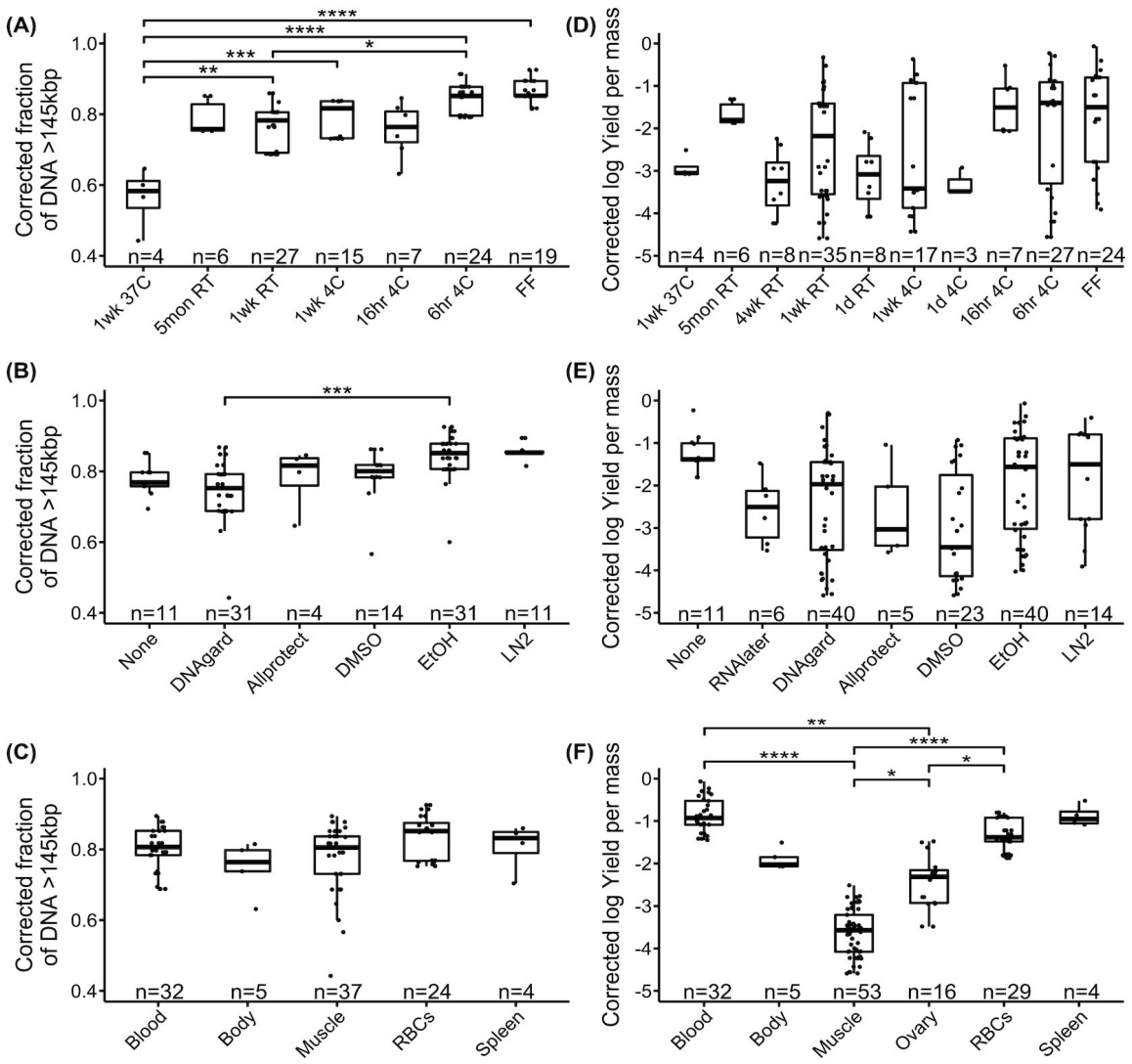
Solution: - DNA repair in all steps of library construction;
- Extra QC after every step
- 10 kb libraries

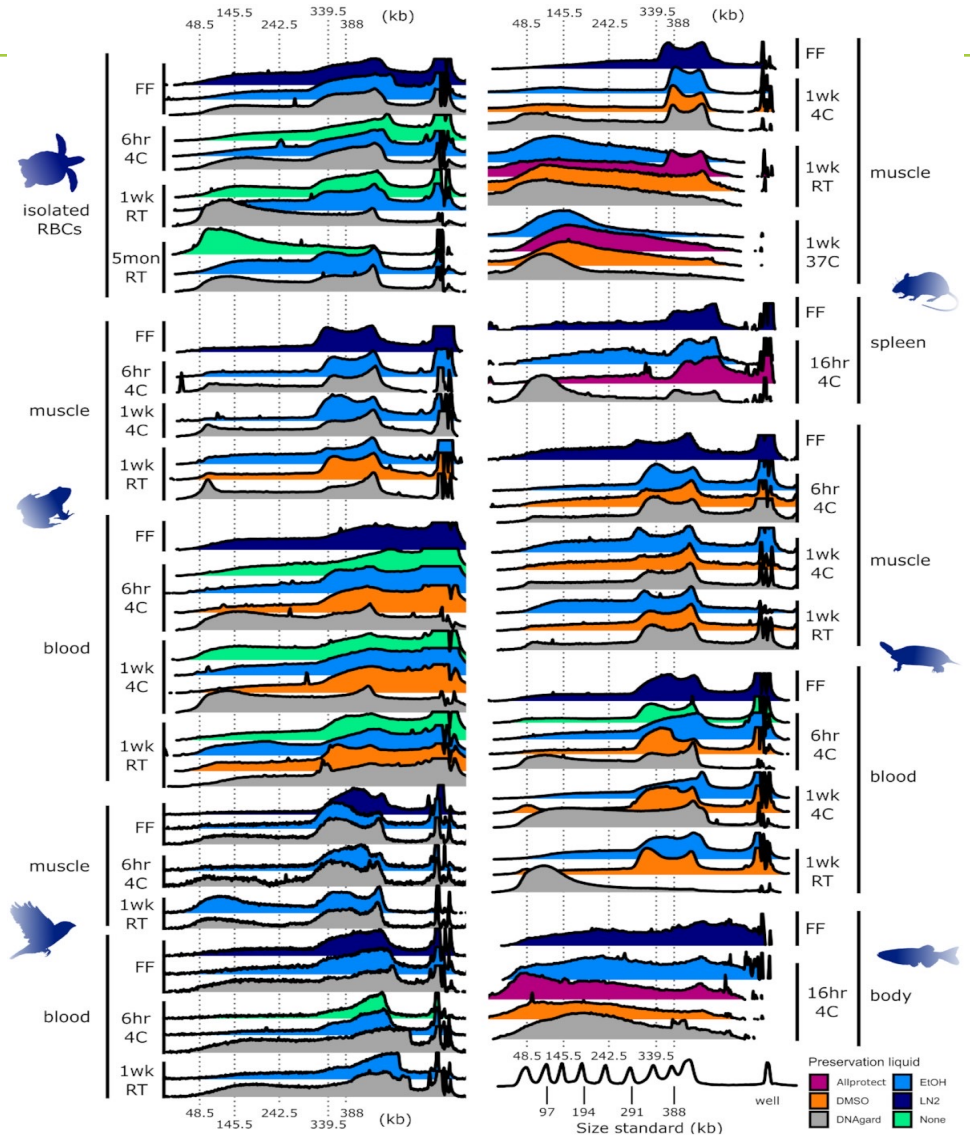


Future of non-model genomics

More R&D should come from sequencing facilities

Sequencing facility R&D: tissue preservation methodology





Dahn et al., Gigascience, Volume 11, , 2022, giac068, https://doi.org/10.1093/gigascience/giac068

Future of non-model genomics

More R&D should come from sequencing facilities

Staff scientists at sequencing facilities need a DISCIPLINE NAME

Future of non-model genomics

More R&D should come from sequencing facilities

Staff scientists at sequencing facilities need a DISCIPLINE NAME

Their commitment must be acknowledged

Future of non-model genomics

More R&D should come from sequencing facilities

Staff scientists at sequencing facilities need a DISCIPLINE NAME

Their commitment must be acknowledged

No more Hidden Figures

Proud to deliver genomic data


Susana Häggqvist


Mai-Britt Mosbech


Tuuli Lundbäck-Larva


Christian Tellgren-Roth


Ignas Bunikis


Adam Ameur


Julia Heintz


Ann-Sofi Strand


Nina Williams


Susanne Hellstedt-Kerje


Lars Feuk


Hannes Yngve


Pernilla Quardford


Ulrika Broström


Linnea Jonsäll

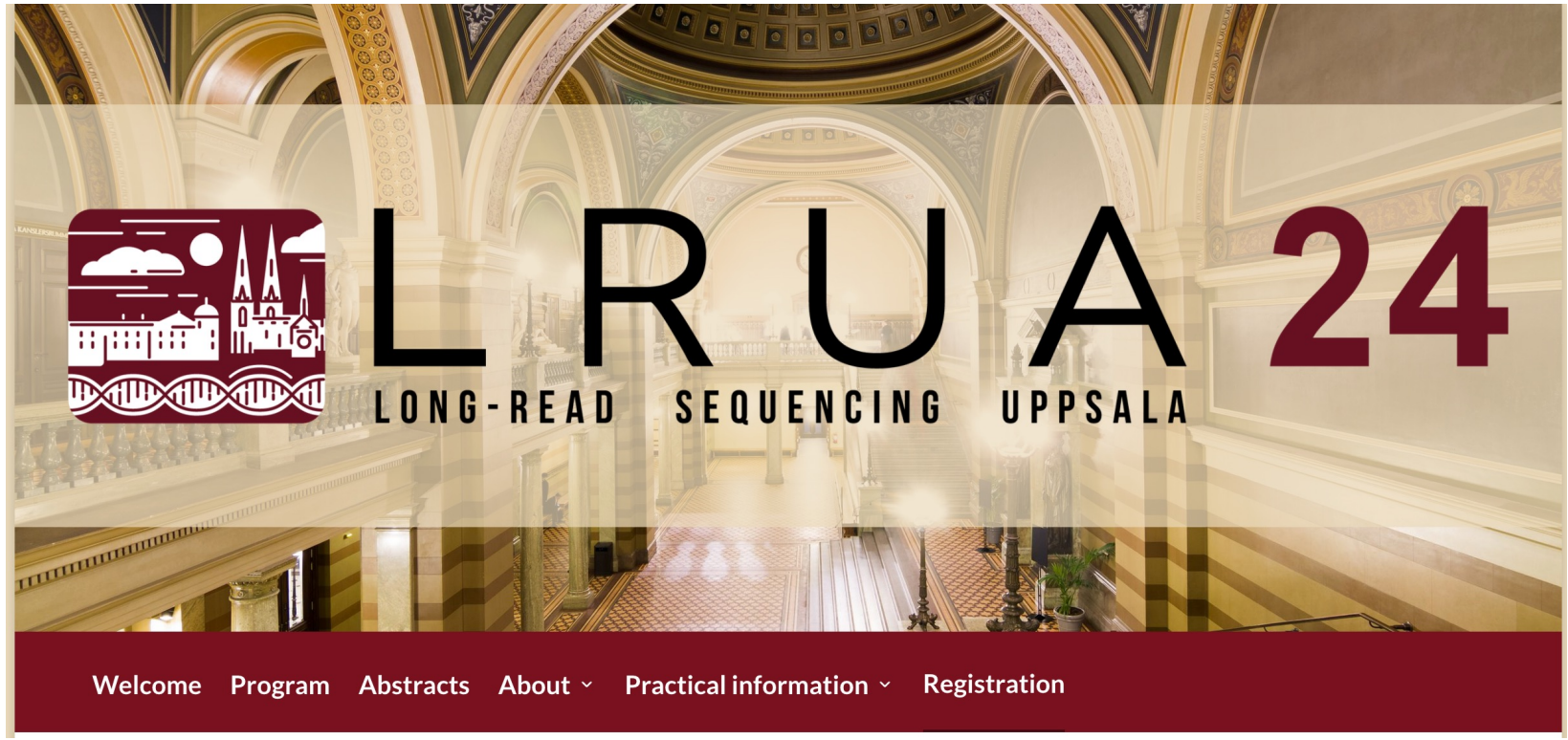
Swedish Research Council

Knut och Alice Wallenbergs Stiftelse

Horizon Europe

SciLifeLab

Karolinska Institutet · KTH Vetenskap och Konst · Stockholms universitet · Uppsala Universitet



October 21-13, 2024 in Uppsala