UPPSALA UNIVERSITET

Experimental design in genomics

Olga Vinnere Pettersson, Uppsala University

Scientific Lead of Planetary Biology Capability, SciLifeLab

ERGA Vice-chair

Český Krumlov
2025-01-10

Origins (green)

Jūrmala

BSc Botany
MSc Plant pathogenic fungi - taxonomy

University of Latvia

Transmutation (genomics)

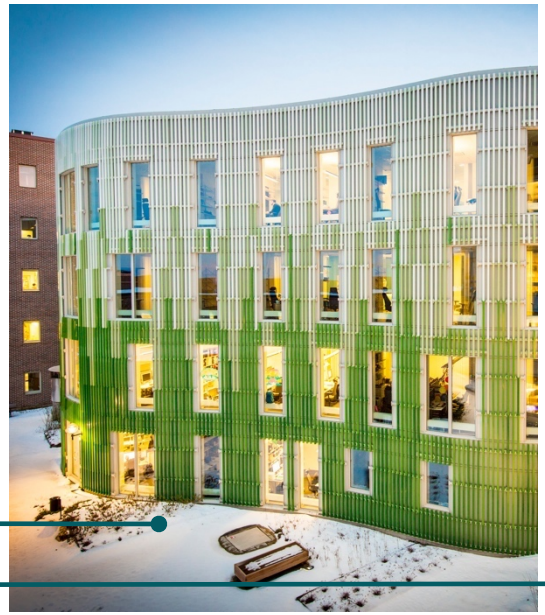
Swedish University of Agricultural Sciences /
Uppsala University

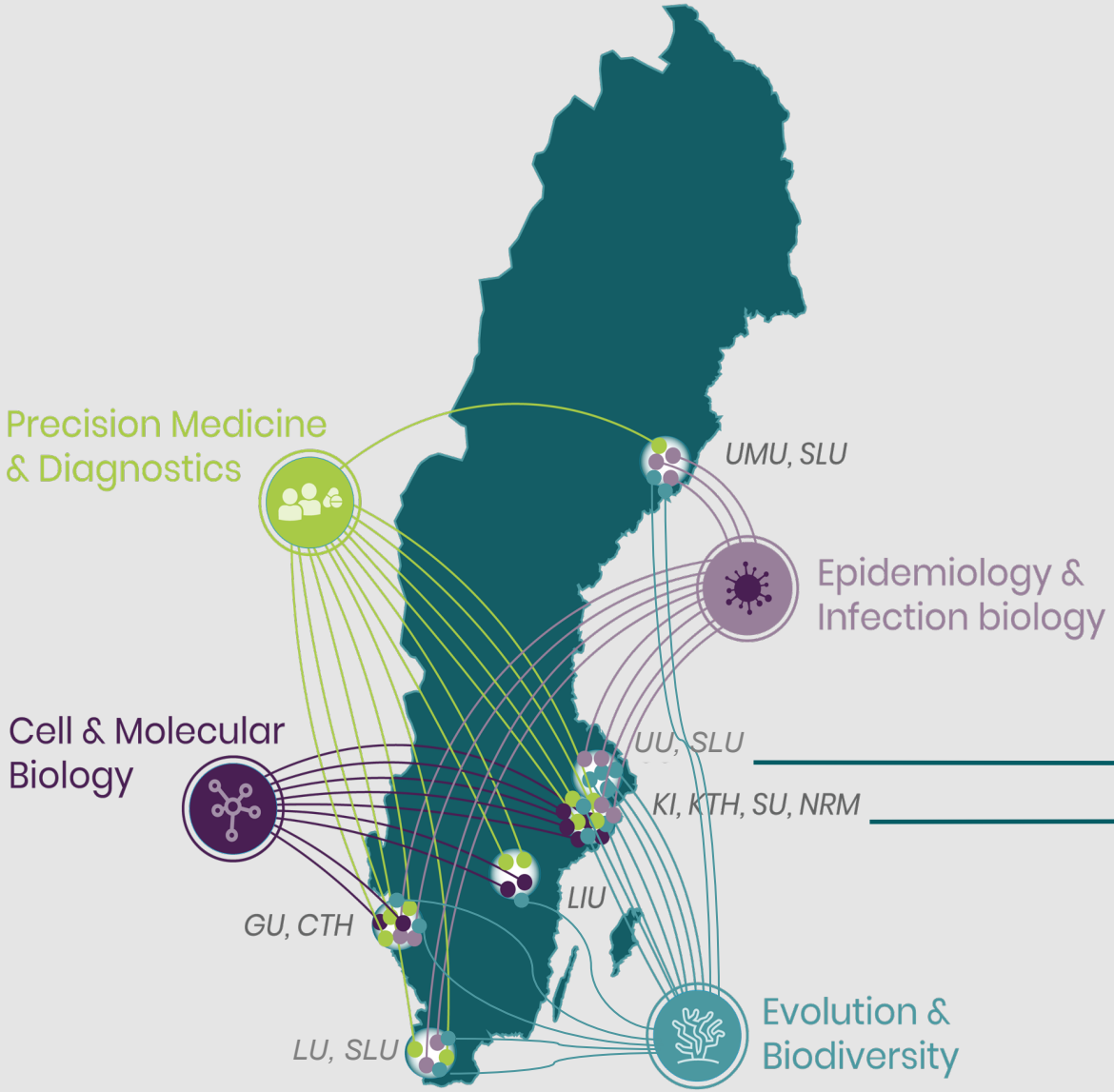
PhD in fungal taxonomy
PostDoc in genome architecture of *Bartonella*
PostDoc in genomics of extremophilic fungi

SciLifeLab Genomics

SciLifeLab

Sweden's national center for molecular biosciences



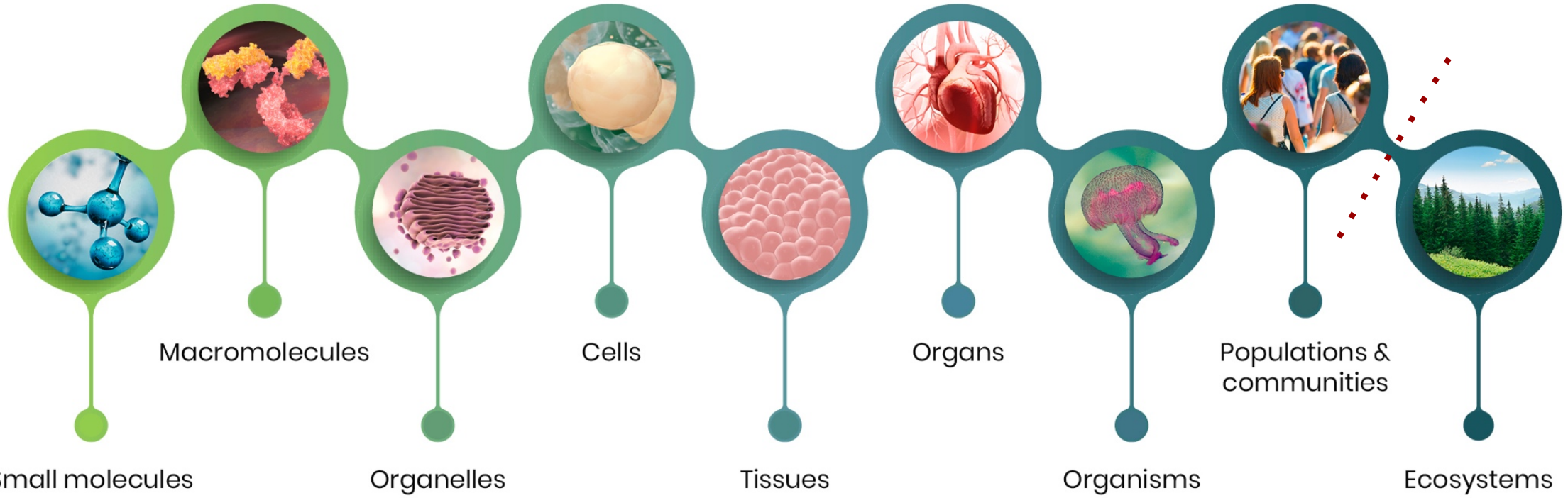


Connecting strong research environments

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



Who am I to tell you things…

- Since 2012 – project manager at SciLifeLab Genomics
- 2013 – long-read sequencing (and why it does not work as it should on many non-models)

- 2018 – VGP SamplePrep committee
- 2019 – EBP Sample Collection and Processing SubCommittee
- 2020 – present at conception of ERGA, since 2023 vice-chair
- 2022 – Biodiversity Genomics Europe
  (building de-centralized European sequencing infrastructure, knowledge transfer)

More than a decade of project management experience, 20 years in genomics of non-models, 12 years in long-read sequencing

Also, I know what you are going through…

Now, to the lecture!

Outline

- Some perspective
- What to think about BEFORE planning a sequencing project (aka Project Design)
- Sequencing applications and experiment design specifics:
    - Whole-genome sequencing
    - Targeted sequencing
    - Transcriptome sequencing
    - Shotgun metagenomics
    - Reference genome sequencing + optimal project workflow example

- Sampling and sample quality requirements
- What every facility wish you knew before sending your samples
- Some perspective

To give you a perspective

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

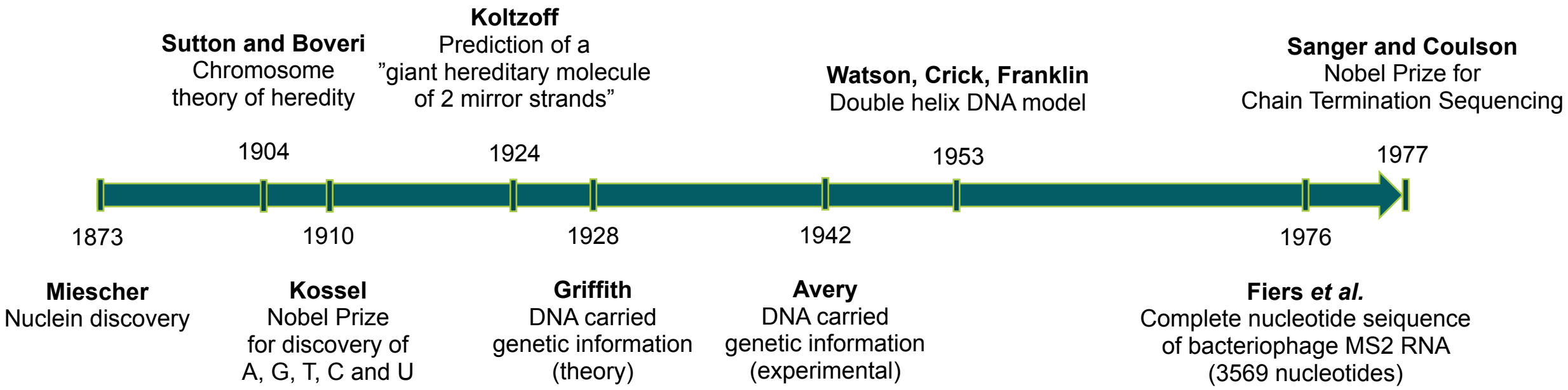
*) Nach Vorträgen, gehalte
ber 1873.

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

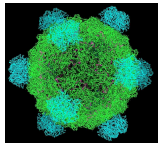
Das Nuclein.

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.

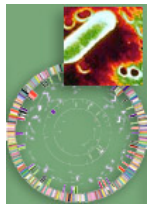
From Miescher to Sanger



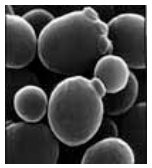
At the very beginning of genome sequencing era…



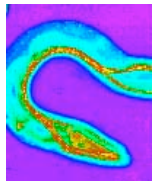
- First DNA genome: virus ϕ X 174 - 5 368 bp (1977)



- 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

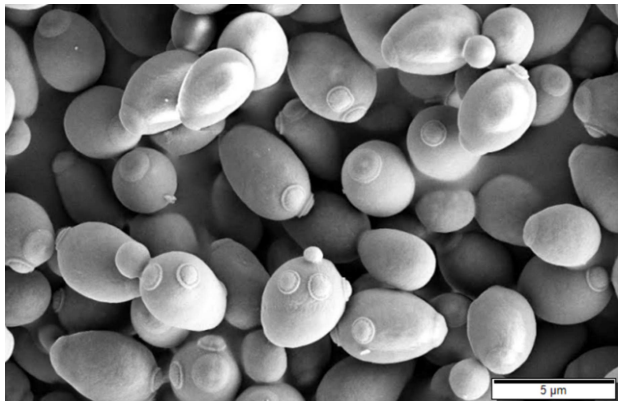
Genetics. 2013 Jun; 194(2): 291–299.
doi: 10.1534/genetics.113.151258

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

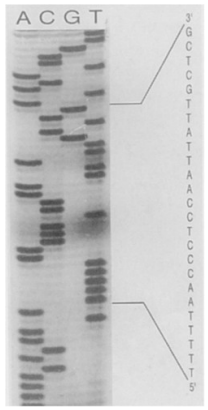
HOST

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

P32 labeled gels

1990 HUGO

2000

Sanger

ABI 3000 series

NGS

2005

454

Solexa
ABI SOLiD

2006

2010

IonTorrent

PacBio
ONT

2012

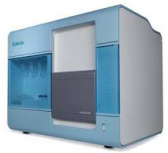
2021 HiFi reads

2023

PacBio
REVIO

Dec 2024 HiFi SPRQ

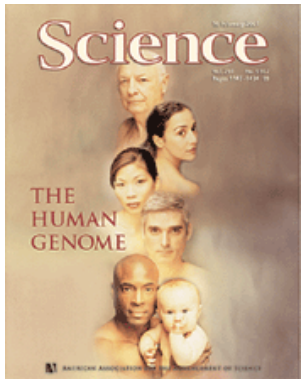
2025 WE ARE HERE

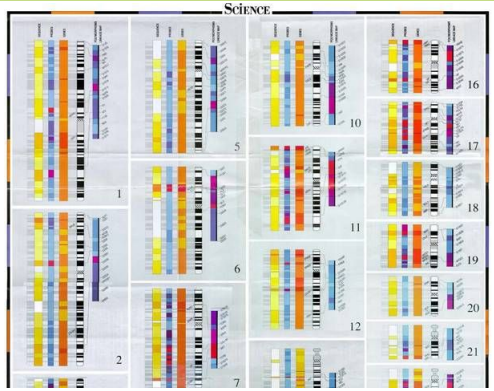


Just a comparison

1990 - 2003
HUGO

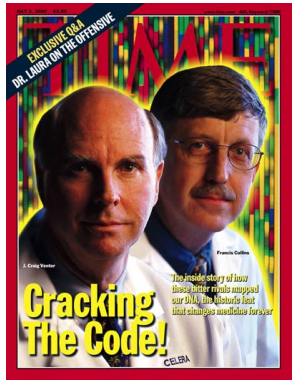
Sanger traditional
$2.7 bln





2007
Craig Venter's genome

Sanger ABI 3730
$300 mln



2008
Jim Watson's genome

454 FLX
$1 mln

TODAY
any human

$500-800 with Illumina
$800-3k 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**

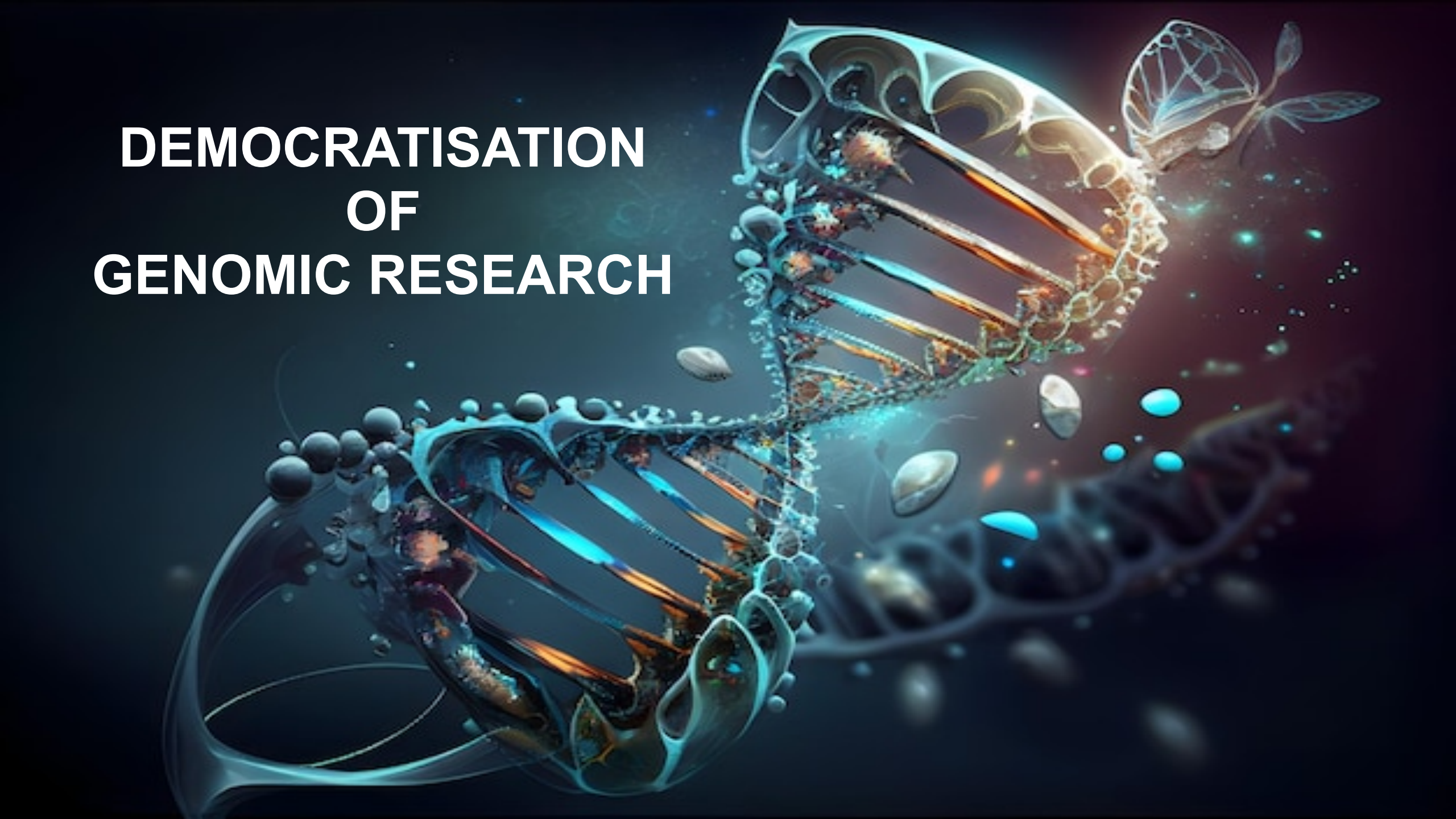
Outcome of genomics revolution: paradigm shift

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



Drawback of genomic revolution: how to stay up to date?

- Immense speed of technological progress

- Inspiration from papers? HA!
  - Design + gathering material + conducting experiments (1 month - years)
  - Sequencing (1-6 months)
  - Analysis (…)
  - Writing paper (month - year)
  - Paper submission (weeks)
  - Reviewer #3 (weeks - months)
    - Design -> published paper 1-3 years

Drawback of genomic revolution: how to stay up to date?

- Immense speed of technological progress

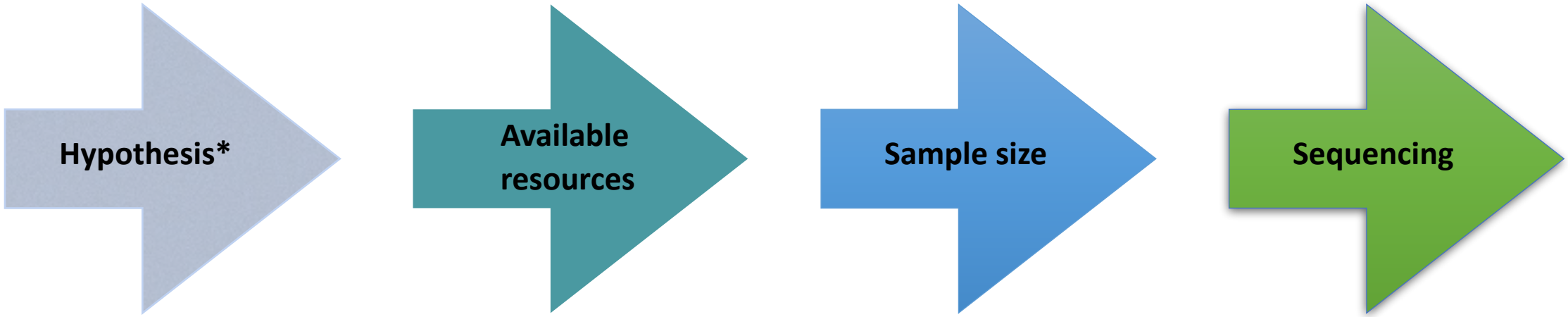
- Inspiration from papers? HA!
  - Design + gathering material + conducting experiments (1 month - years)
  - Sequencing (1-6 months)
  - Analysis (…)
  - Writing paper (month - year)
  - Paper submission (weeks)
  - Reviewer #3 (weeks - months)
    - Design -> published paper 2-3 years

SOLUTION: talk to a sequencing center near you

Planning the sequencing project

WHAT IS MY QUESTION?

- Qualitative or quantitive?
  - Avoid: “let’s sequence and see what happens”
  - Frequent scenario: “here is what I have, do your best/worst”



**or creating a resource (e.g. variant database, reference, etc)*

Available resources

- €€€
- Store and compute
- Who will analyse the data?
- Who will biologically interpret the data?
- Sample availability (number, quality and amount of material)

Sample size

- Statistical sensitivity: #of samples, #of replicates, #of reads (sequencing depth)

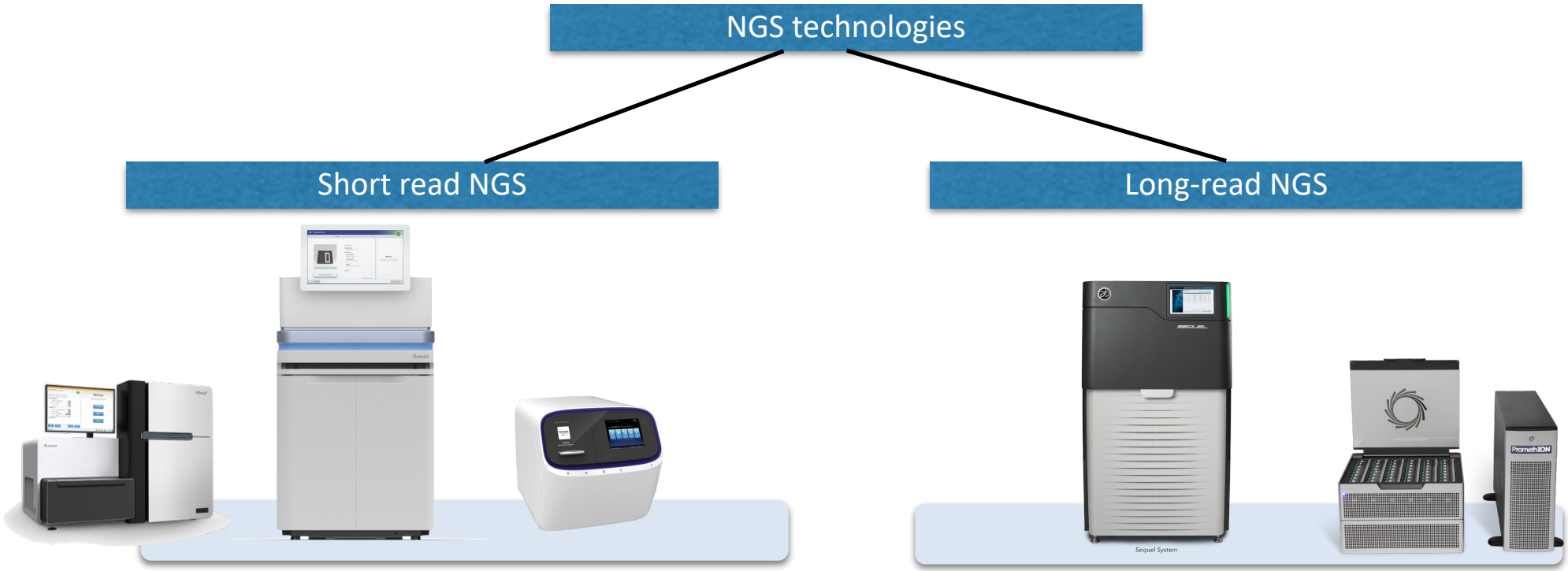
Sequencing

- Choice of application and technology

Early planning **before** writing a grant proposal is highly recommended. Get advise!

Let’s dive into it

NGS Technologies and Applications



Whole genome re-sequencing
RNA-seq
Targeted sequencing
Amplicons up to 600 bp

De novo genome sequencing
RNA-seq
Whole-transcript sequencing
Targeted sequencing
Structural variant resolving
Allele phasing
Amplicons up to 13 kb

Whole genome sequencing (WGS)

Re-sequencing or de novo?

- Re-sequencing (WGS):
  - Pre-requisite: a reference genome to map to.

  Population studies (genotyping, variant discovery, allele frequency, etc)

  SNPs only? Short reads.

  SVs? Long reads.

  SNPs and SVs? PacBio **HiFi** (January 2025).

- *De novo* (Reference genome sequencing):
  - Creating a genomic reference from scratch
  - Long reads (sometimes coupled with short-read skims)

WGS sequencing depth

Population sequencing: individual libraries or pools of many individuals

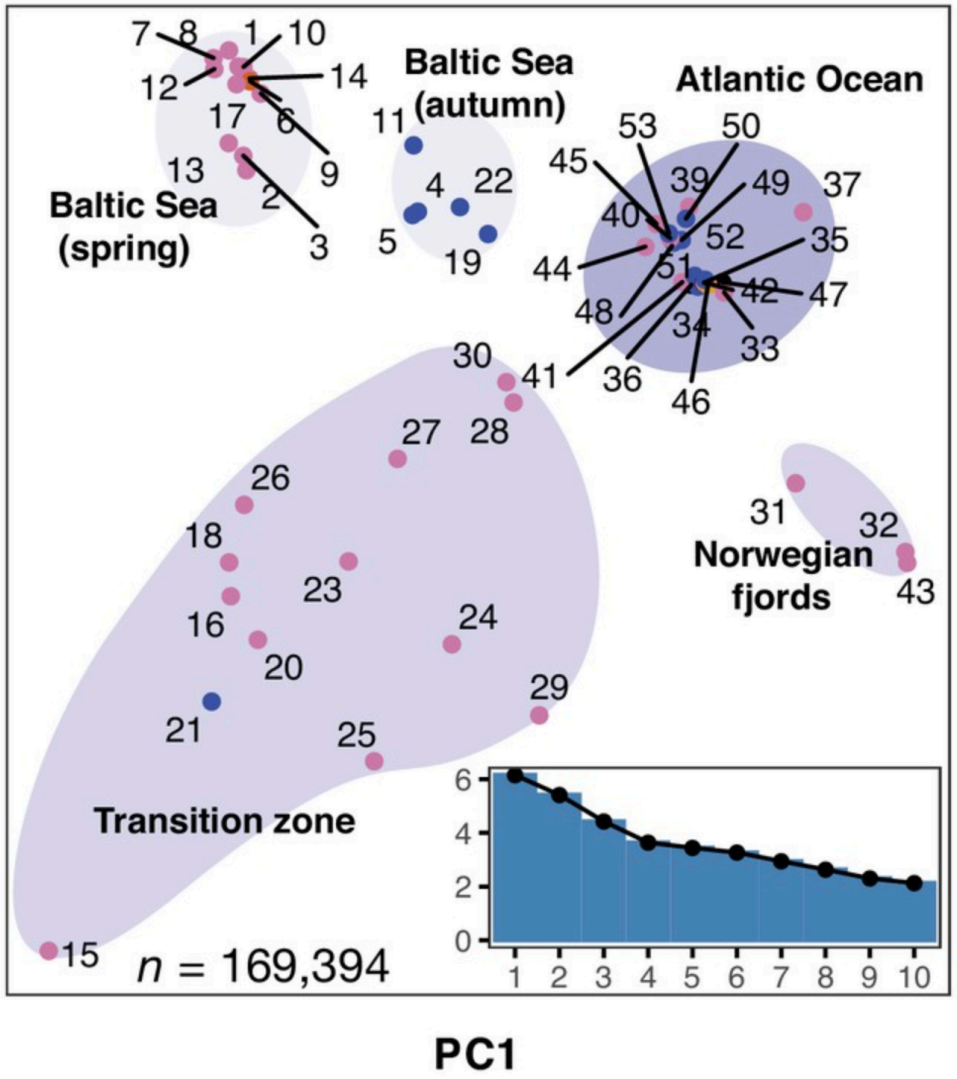
Type of Experiment	Coverage Required
Haploid SNPs/divergence	≥ 10 x
Diploid SNPs/divergence	≥ 30 x
Aneuploid/somatic mutations	≥ 50 x
Population sequencing	≥ 200 x

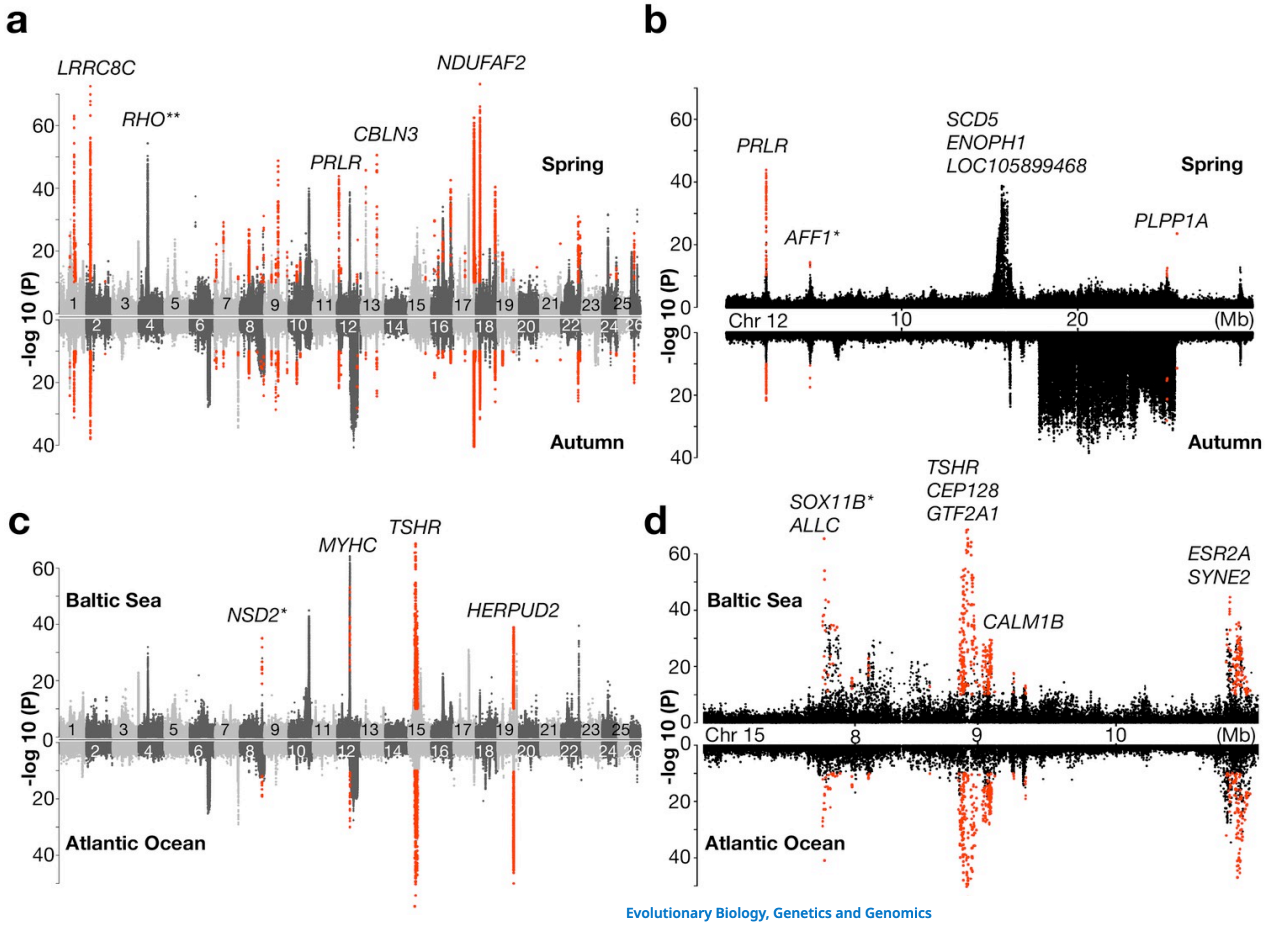
Borrowed from Mike Zody

Individual libraries give better resolution, but are more expensive
Pool sequencing: will pick up main trends (e.g. loci under selection)

Caution not to over sequence: sequencing errors vs true biology

WGS, examples





Evolutionary Biology, Genetics and Genomics

Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci

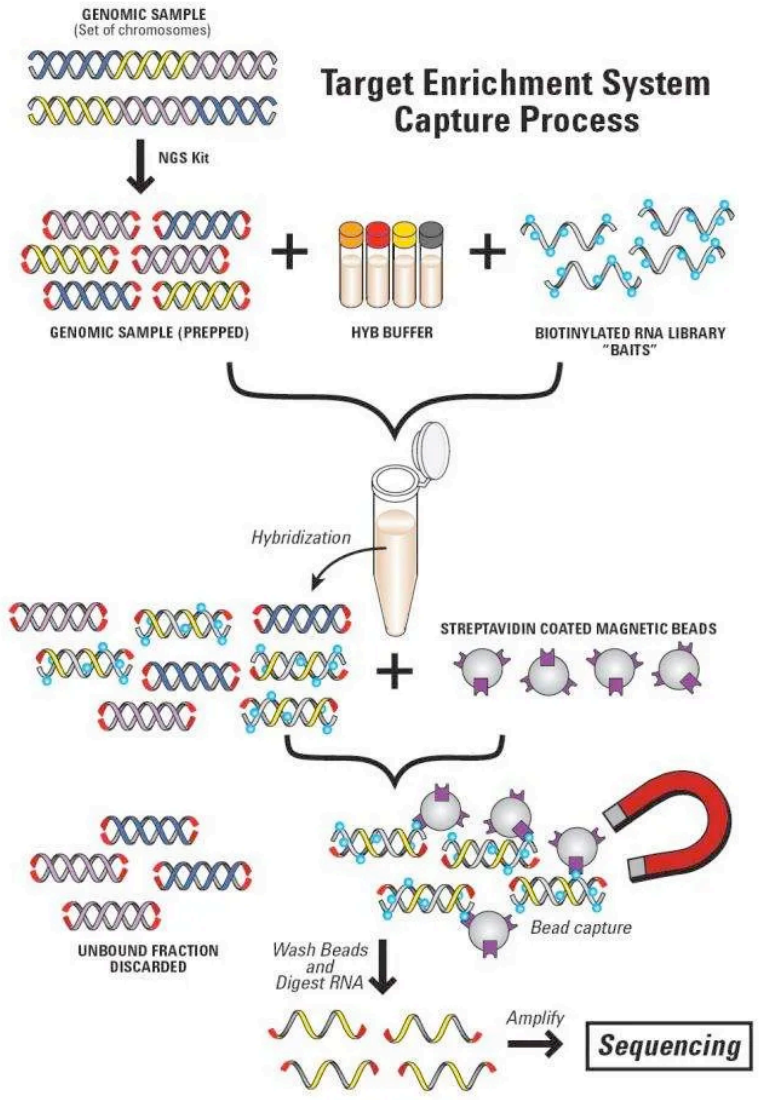
Fan Han, Minal Jamsandekar, Mats E Pettersson, Leyi Su, Angela P Fuentes-Pardo, Brian W Davis, Dorte Bekkevold, Florian Berg, Michele Casini ... Leif Andersson see all »

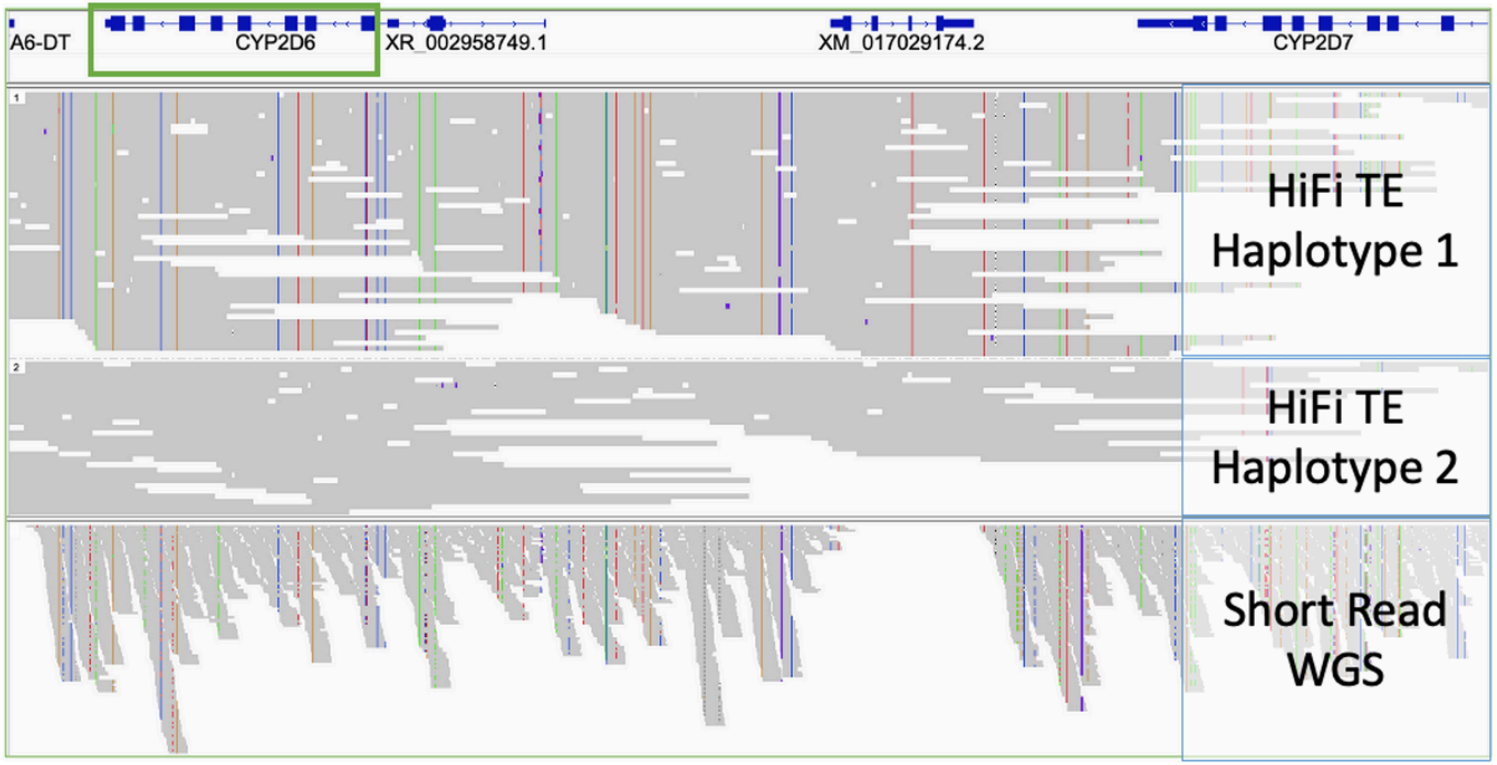
Targeted sequencing

Zooming into portions of a genome - a cost effective screening alternative

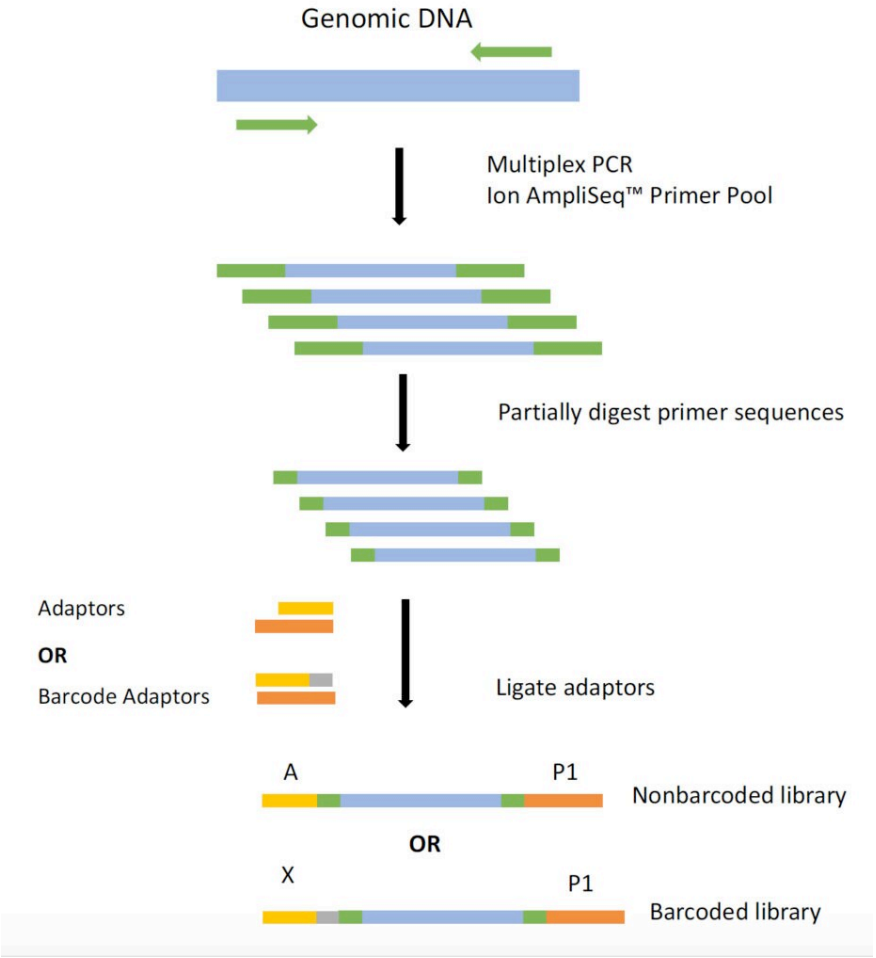
- **Capture with probes**
  - Panels or custom (Agilent, ThermoFisher, Twist, etc)
  - Specificity varies
  - Material requirements (good quality and quantity)
  - Cost
- **PCR**
  - Own primers or custom panels (e.g. Ion AmpliSeq)
  - High sensitivity and specificity
  - Prior knowledge of sequence is needed
  - Bias and product length limitations
- **CRISPR-Cas9**
  - Prior knowledge of sequence is needed
  - Requires a lot of DNA
  - Off-target effects reported

Targeted sequencing: capture with probes

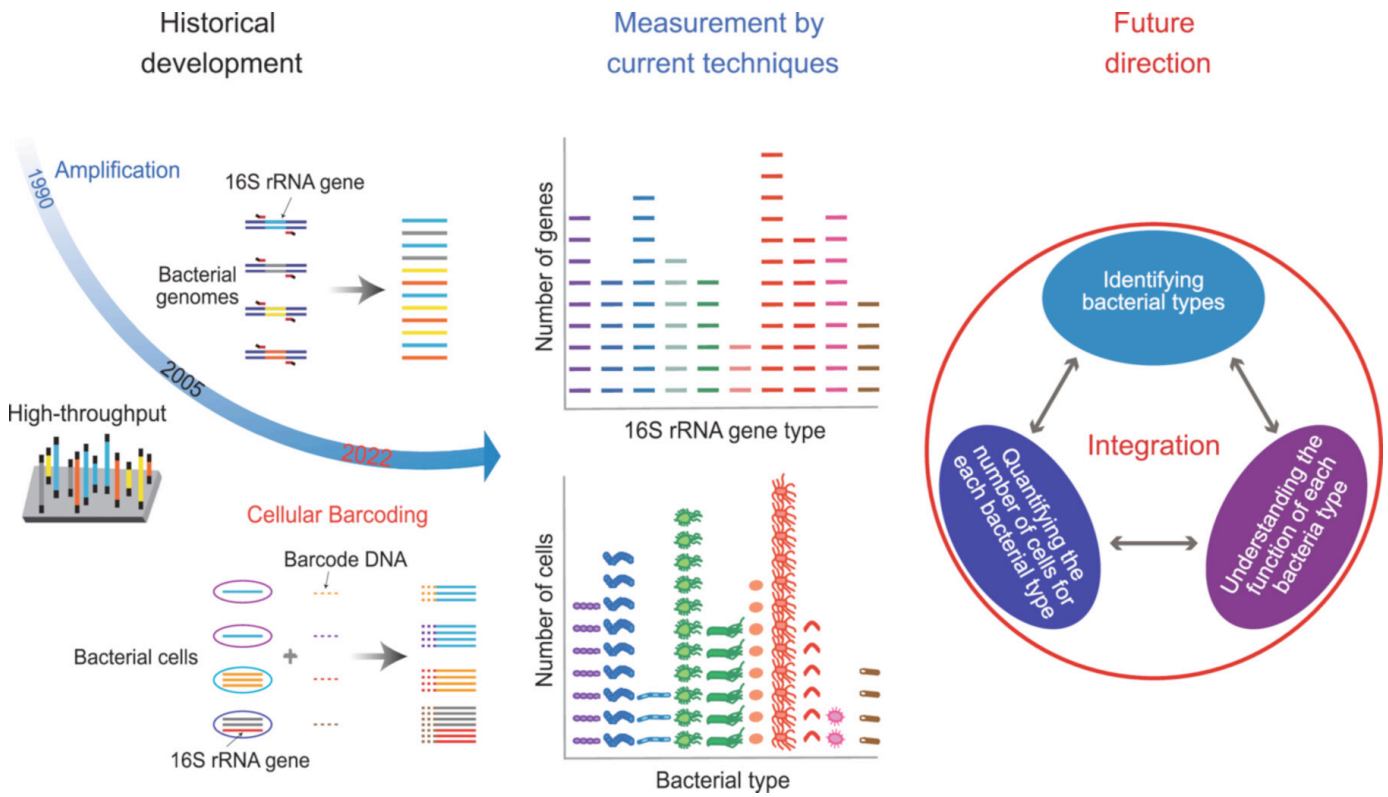




Targeted sequencing: PCR (amplicon seq)



Disease panels with multiple targets (e.g. Ion AmpliSeq)



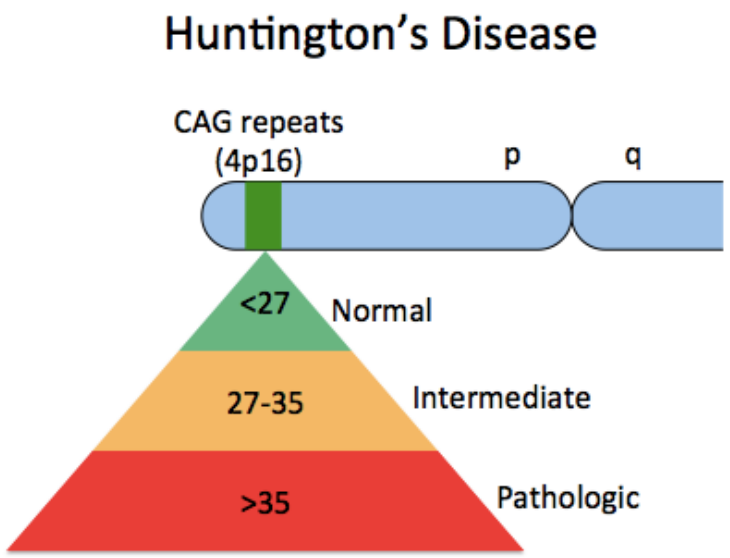
REVIEW ARTICLE | Open Access

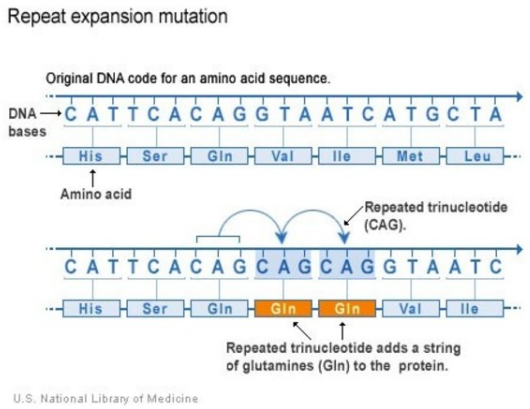
Long journey of 16S rRNA-amplicon sequencing toward cell-based functional bacterial microbiota characterization

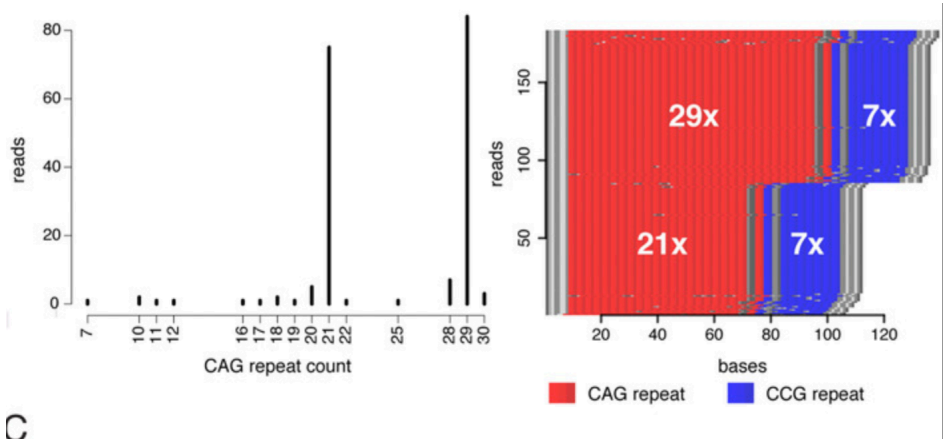
Jianshi Jin, Xiongduo Liu, Katsuyuki Shiroguchi

First published: 11 July 2024 | https://doi.org/10.1002/imo2.9 | Citations: 2

Targeted sequencing: CRISPR-Cas9







Huntington's disease:

- Inherited disorder resulting in brain cell death
- Decline of motor and cognitive functions
- Common onset: 30-50 years of age
- No cure
- Causative genetic variant: CAG-repeat expansion in *HTT* gene

Problem: polymerase slippage – low complexity regions
PCR-based methods do not work

Detailed analysis of *HTT* repeat elements in human blood using targeted amplification-free long-read sequencing

Ida Höijer, Yu-Chih Tsai, Tyson A. Clark, Paul Kotturi, Niklas Dahl, Eva-Lena Stattin, Marie-Louise Bondeson, Lars Feuk, Ulf Gyllensten, Adam Ameur

First published: 22 June 2018 | https://doi.org/10.1002/humu.23580 | Citations: 62

Targeted sequencing: technology & depth?

Talk to different vendors: technology-specific solutions

Run a pilot! (e.g. known truth vs real-life)

Hierarchical data reduction to determine sequencing depth

Ask for advise from your sequencing provider

Transcriptome Sequencing (RNA-seq)

TOTAL RNA

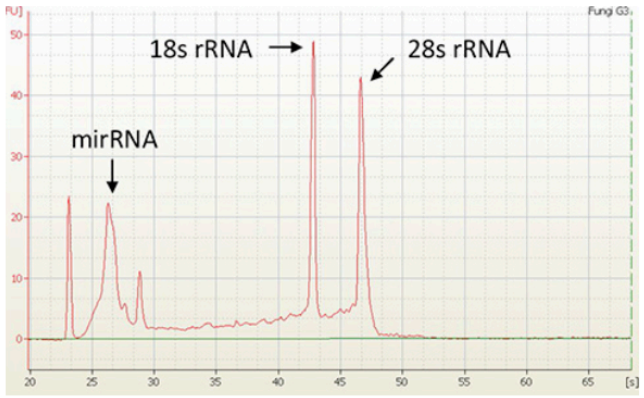
mRNA

- **Differential expression**
- Annotation

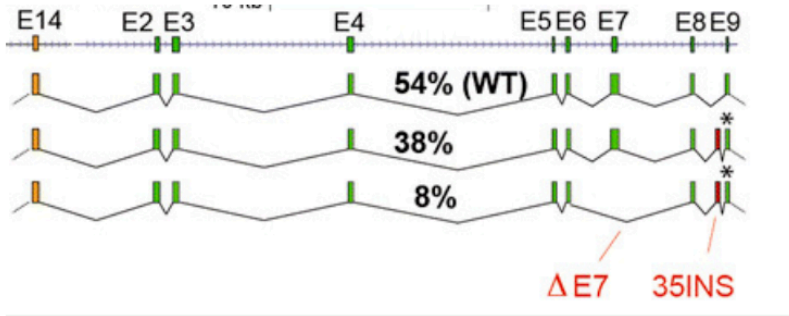
Non-codingRNA

- Transcriptional regulation

miRNA



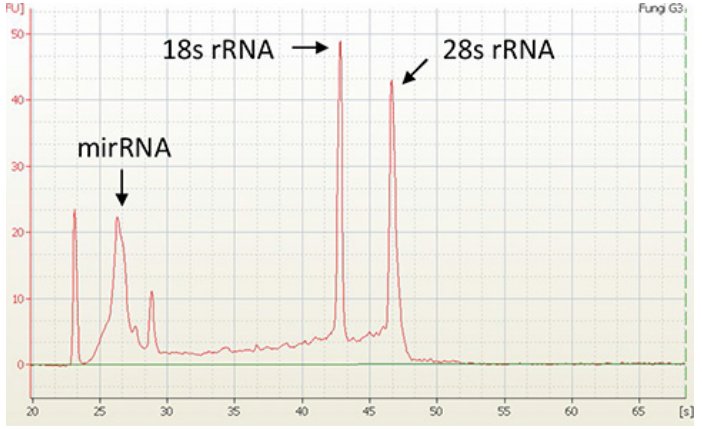
Splice isoforms

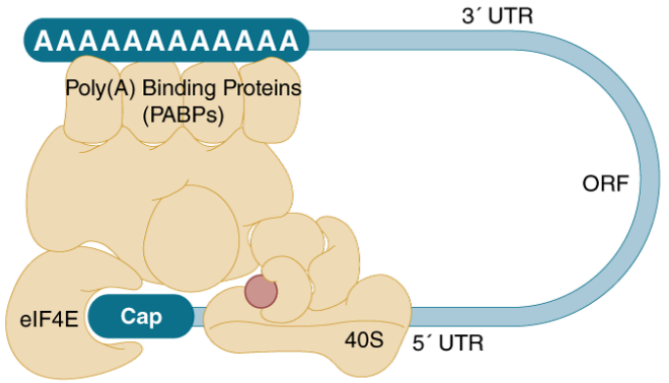


Transcriptome Sequencing (RNA-seq)

One must get rid or rRNA. To PolyA or not to PolyA?

Method	Pros	Cons
rRNA depletion	• Captures on-going transcription • Picks up non-coding RNAs	• Does not get rid of all rRNA • Messy Dif.Ex. profile
polyA selection	• Gives a clean Dif.Ex. Profile • Looses all non-polyA RNAs	• Does not pick many non-coding RNAs



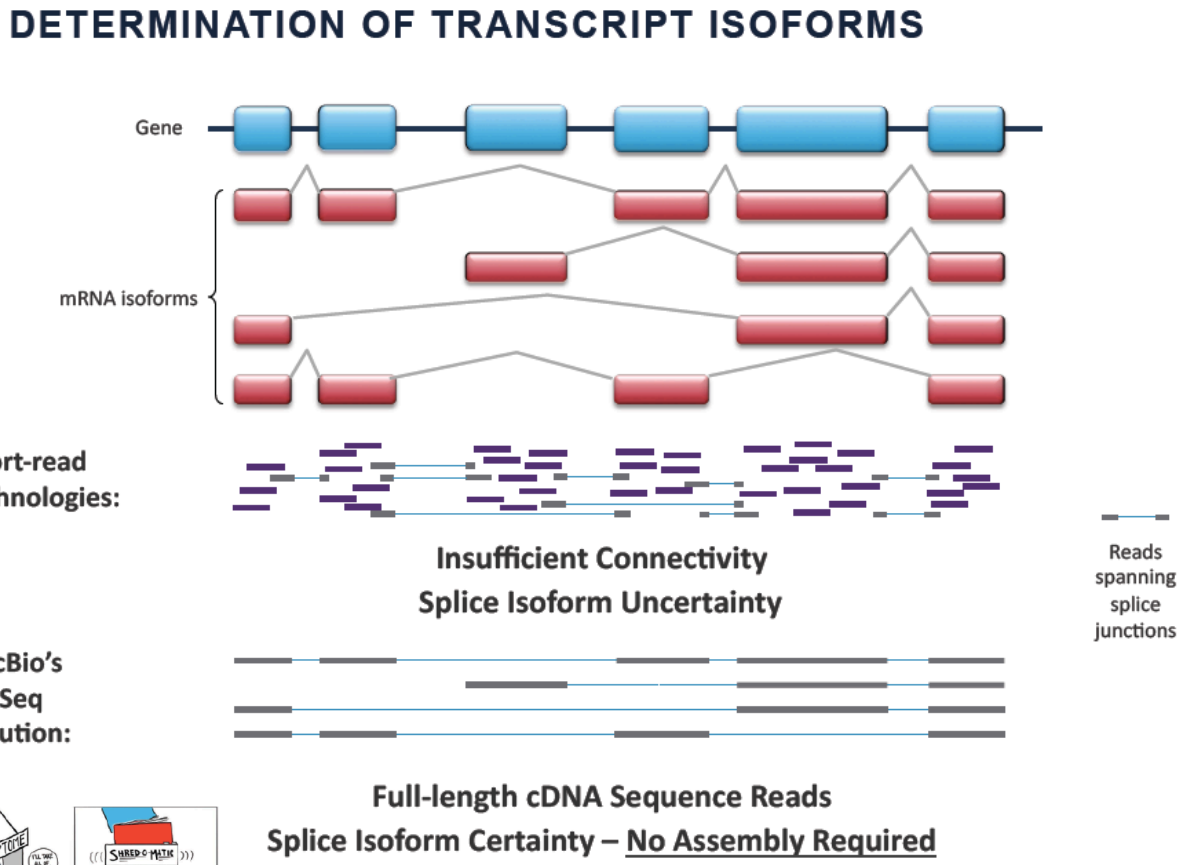


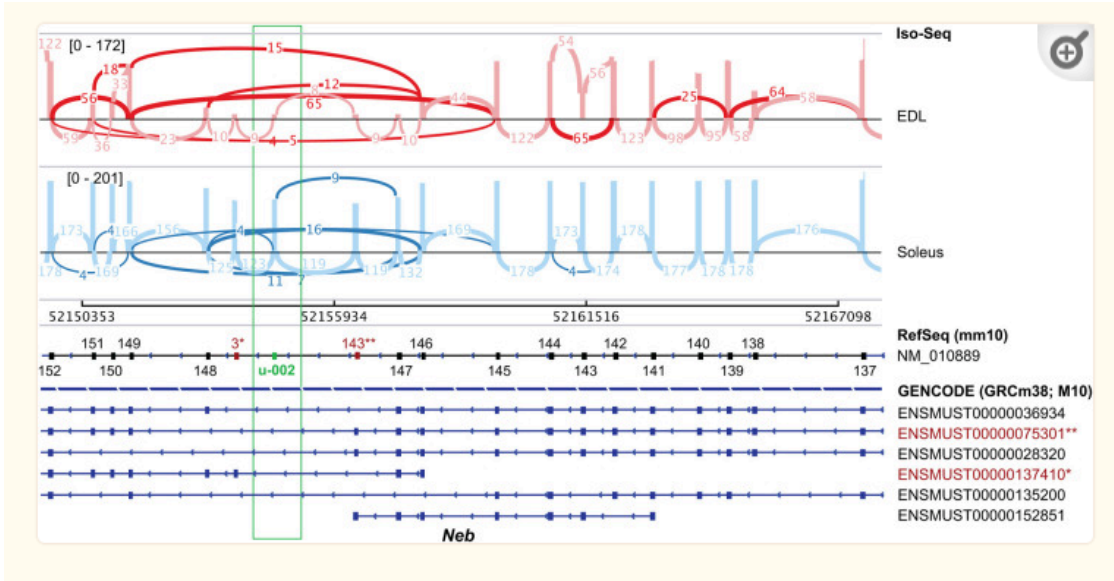
Number of reads

Differential expression with a **good*** reference: 5+M PE reads (up to 100M for rare transcripts)
Annotation: a minimum of 50M PE reads of mixed tissue (rather 100M per tissue)

*well-annotated

RNA-seq with log reads



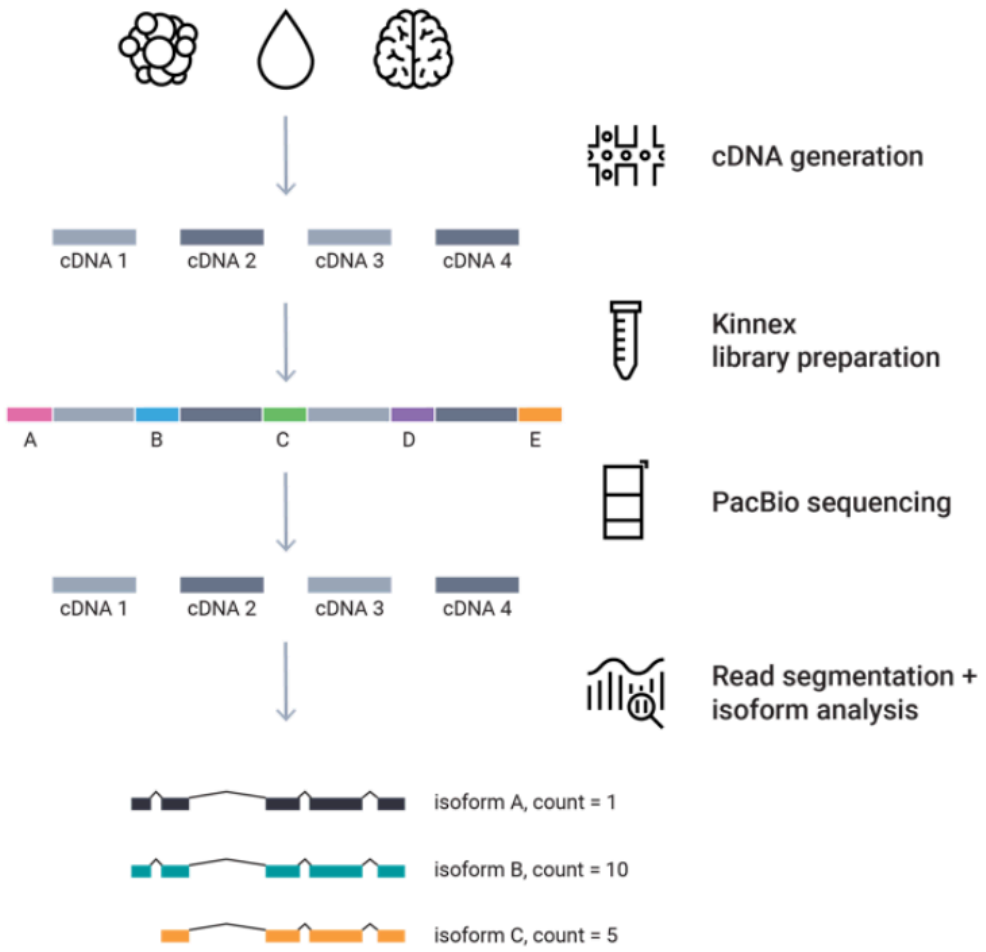


Genome Res. 2020 Jun; 30(6): 885–897.

NEB:
one of the biggest protein coding genes in vertebrates (22kb mRNA, 183 exons)
Codes for nebulin, muscle protein

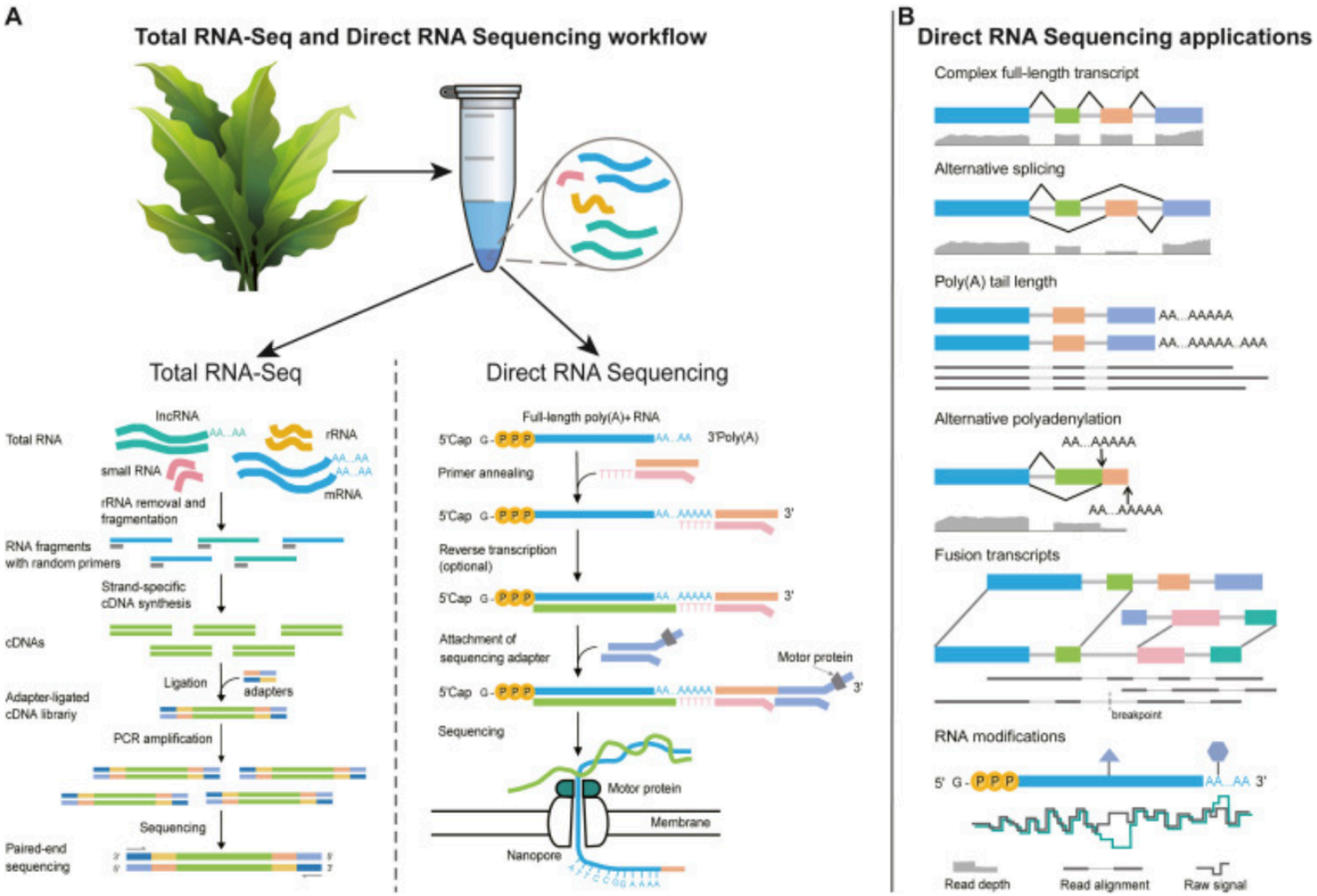
Differential expression with long reads

- Best of both worlds: expression values AND isoform information
- Prior to Q2 2024 - ONT only (polyA or CAP-selected)
- Now: both ONT and PacBio Kinnex



RNA-seq: direct RNA sequencing on ONT

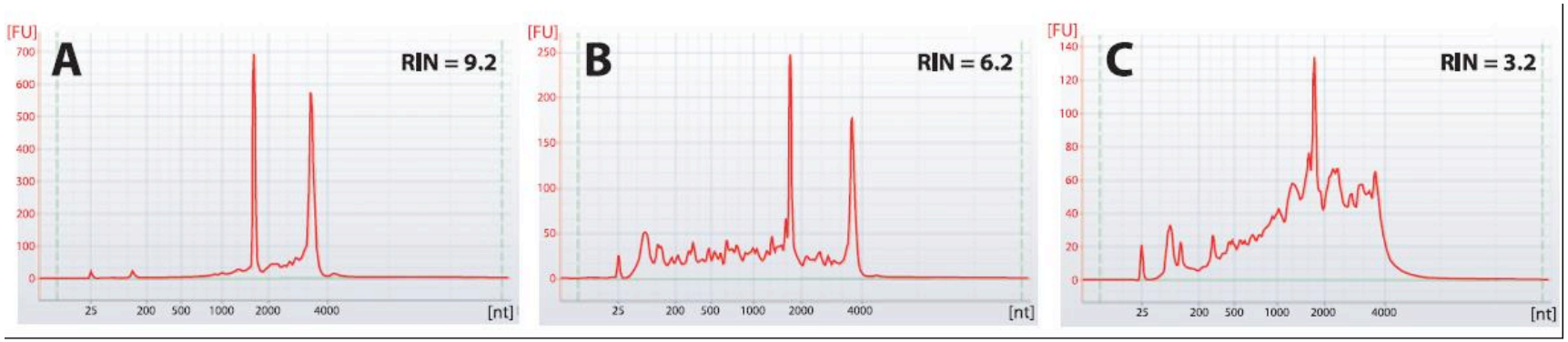
Do not get too excited



One cannot use
direct RNA-sequencing
for differential gene expression
experiments!
(January 2025)

RNA-seq considerations

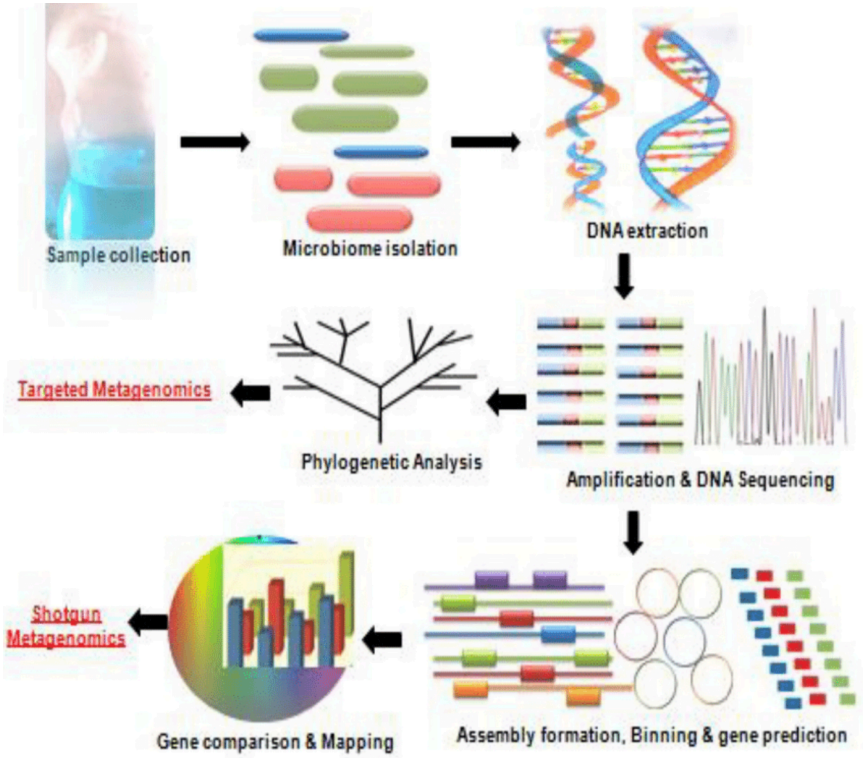
- mRNA only: use any kit (for annotation and long-reads we recommend TRIZOL)
- mRNA **and** miRNA: only specialized kits
- Always use DNase!
- RIN value above 8

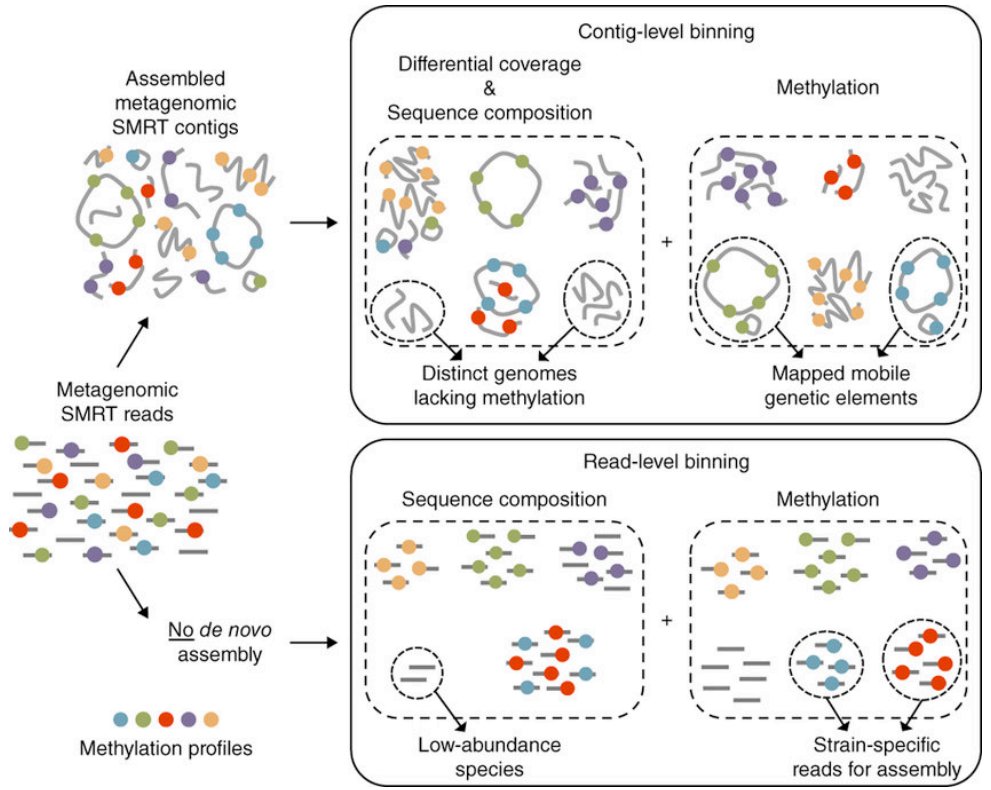


- CONTROL vs experimental conditions
- Biological replicates: a minimum of 4 is strongly recommended

Shotgun metagenomics

- Strongly recommend a pilot + hierarchical data reduction to determine the sequencing depth
- Can be done with both short and long reads
- If with long reads - consider utilizing epigenetic signature for plasmid assignment / OTU binning





Diving into deep: reference genome sequencing of non-models

Coffee break?

De novo (reference genome) sequencing

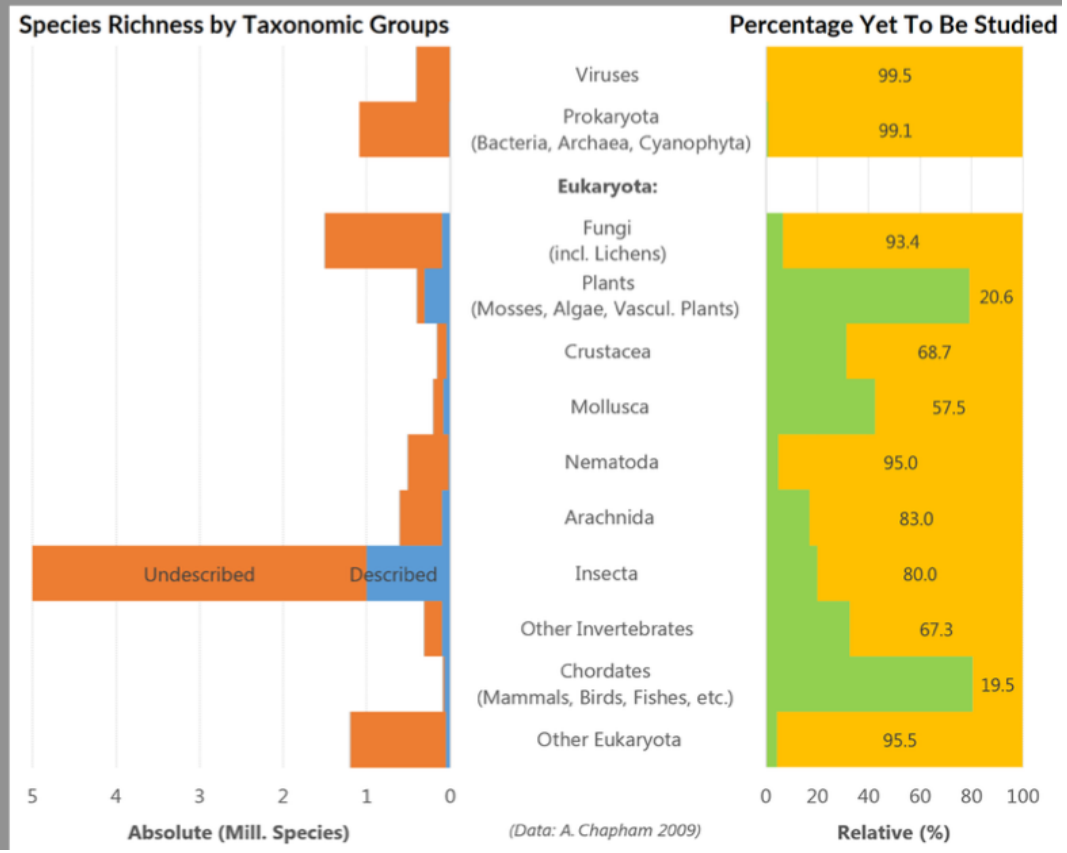
- Sequencing a genome without a prior reference

Homo sapiens

N species = 1

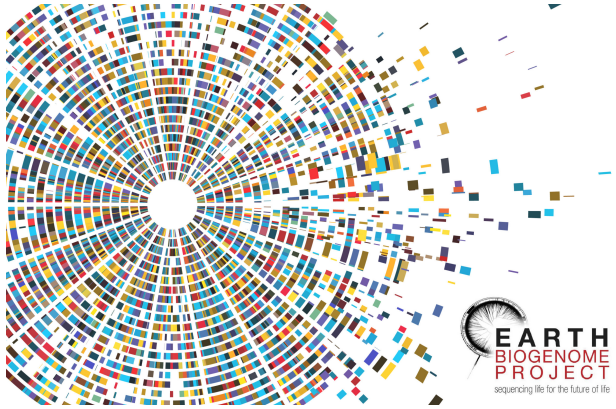
Global Biodiversity

N species = 1 trillion



Reference genome sequencing (RefGen)

- Only closely related taxa can be used for alignment



- Always done with long reads
- Enable any kind of downstream genetic analysis

- Generating a chromosome-scale reference genome is a life-time investment. But it is costly.

ERGA
EUROPEAN REFERENCE GENOME ATLAS

- Sit tight, it is going to be a long one

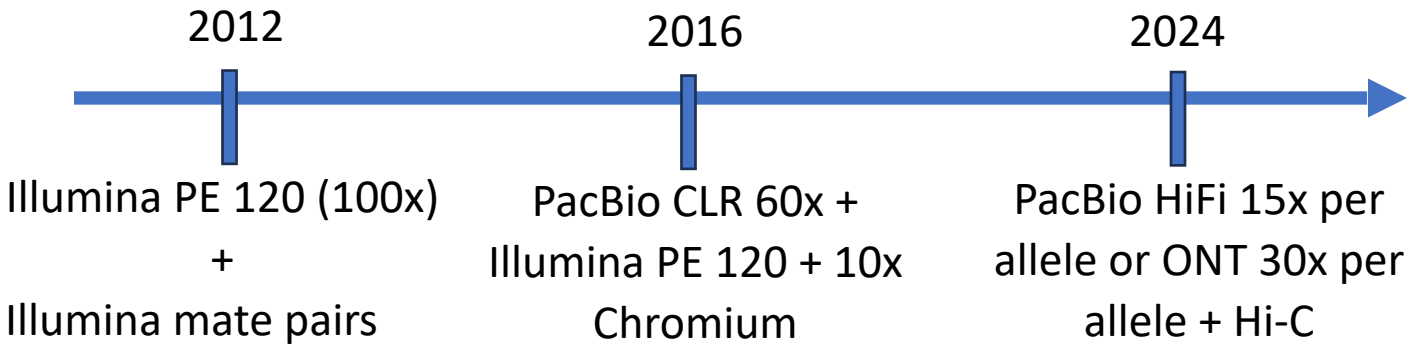
Biodiversity Genomics Europe

RefGen Sequencing recipes

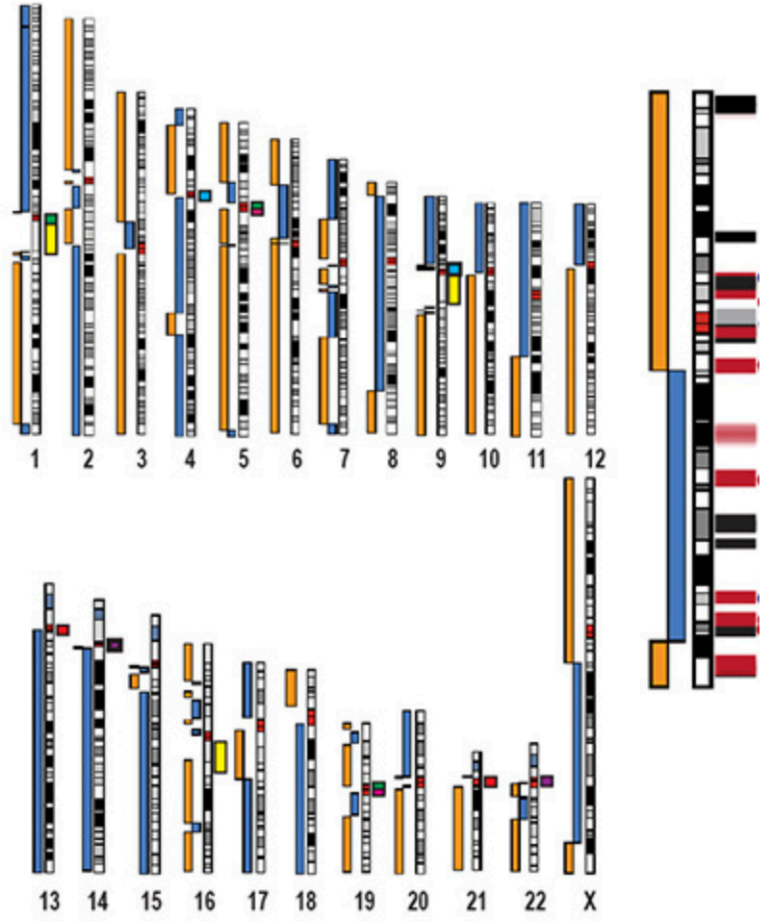
Optimal technology combination for RefGen is a moving target

Technology is constantly developing

So do the methods of analysis



Where to check? ERGA or EBP SOPs

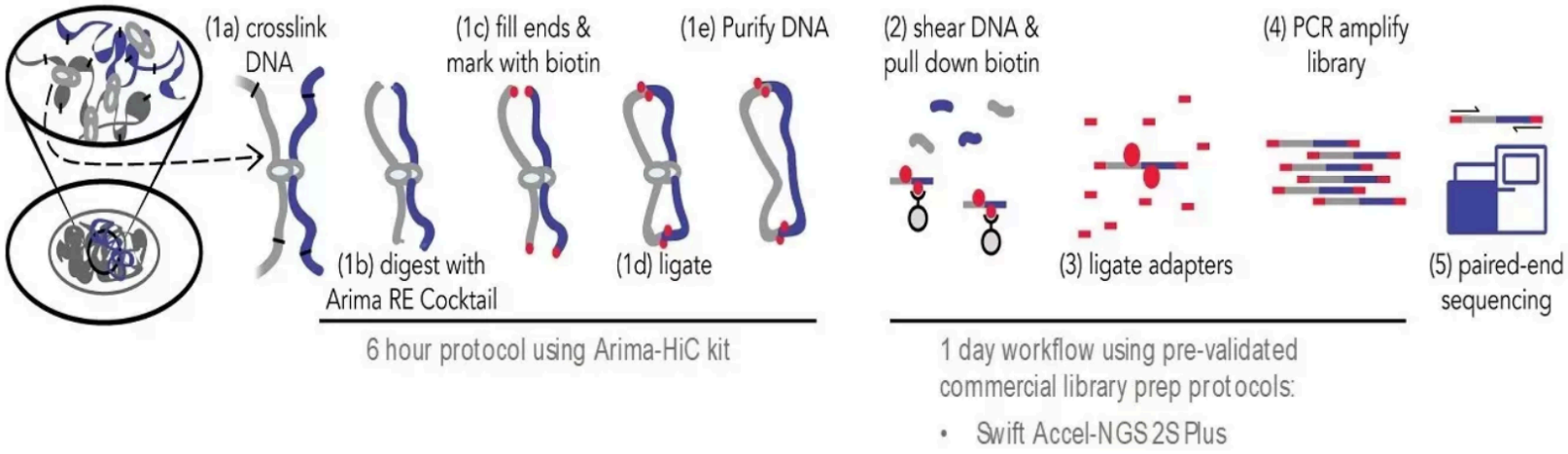


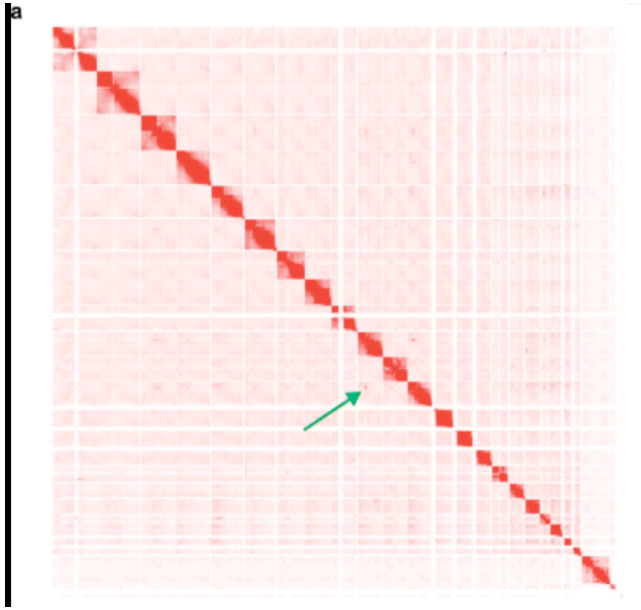
RefGen recipe, details

- 15x PacBio HiFi per allele / ONT 30x per allele + low-coverage Illumina
- HiC 50M reads per Gb of genome (Arima HiC or DoveTail OmniC)
- RNA-seq for annotation
  - 50M reads per tissue mix (EBP standard)
  - Rather 100M reads per tissue, use as many tissues as possible

HiC library principle: DNA arrangement in chromosomes

Invaluable for chromosome reconstruction





FAQ: What data should I add to improve my existing fragmented assembly?

A: Do not waste your time and just do it from scratch.

Time of a bioinformatician is more expensive than sequencing.

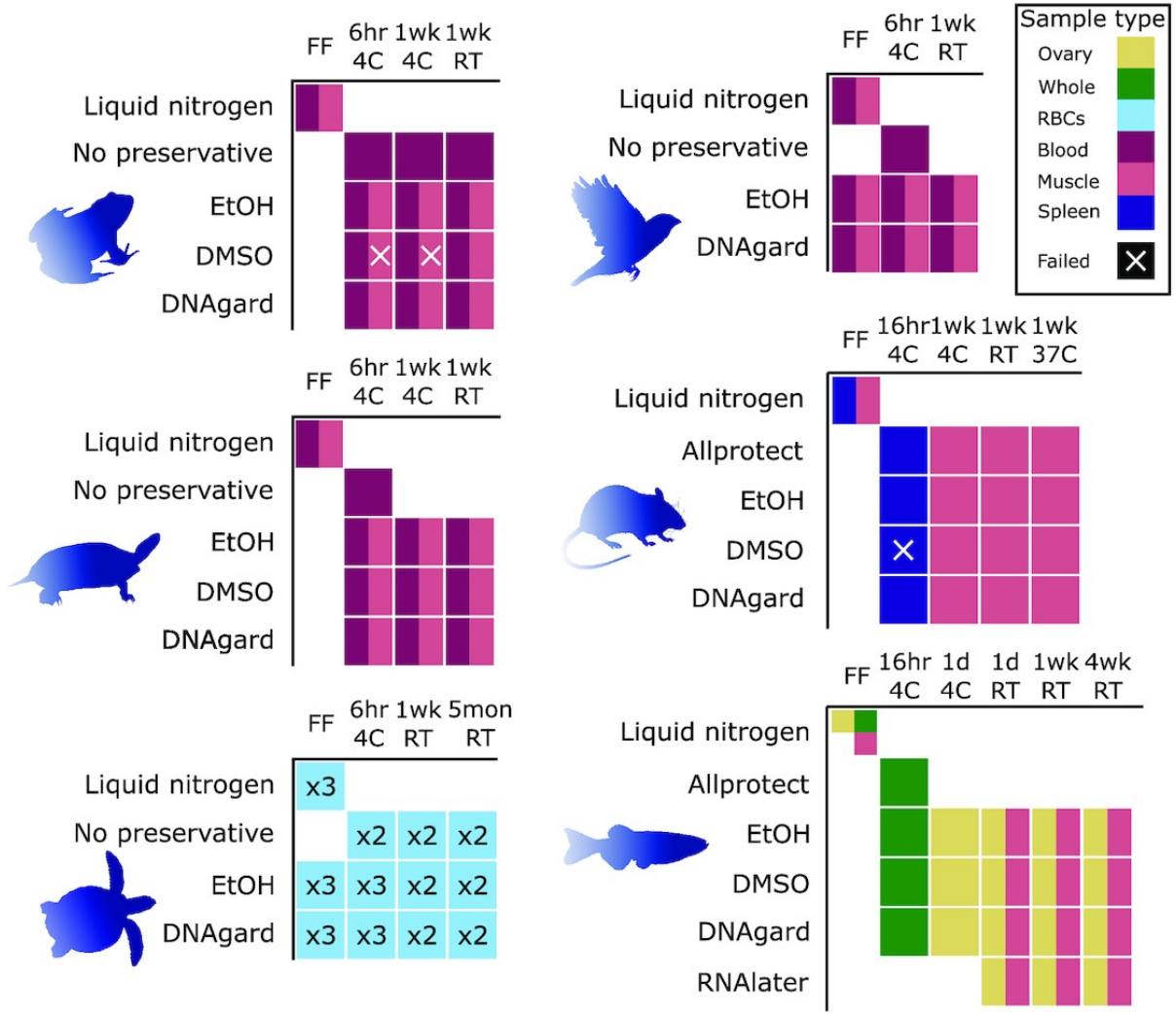
Long-read technologies nowadays is not what they used to be 5 years ago.

Do not “polish” your HiFi reads with Illumina - you will just introduce errors.

Important to keep in mind while planning reference genome sequencing

RefGen: sampling methodology is IMPORTANT

Taxon-specific sampling and preservation



(GIGA)n SCIENCE

Articles Submit Alerts About GigaScience

Volume 11
2022

Article Contents

Abstract

JOURNAL ARTICLE

Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing

Hollis A Dahn, Jacquelyn Mountcastle, Jennifer Balacco, Sylke Winkler, Iliana Bista, Anthony D Schmitt, Olga Vinnere Pettersson, Giulio Formenti, Karen Oliver, Michelle Smith ... Show more

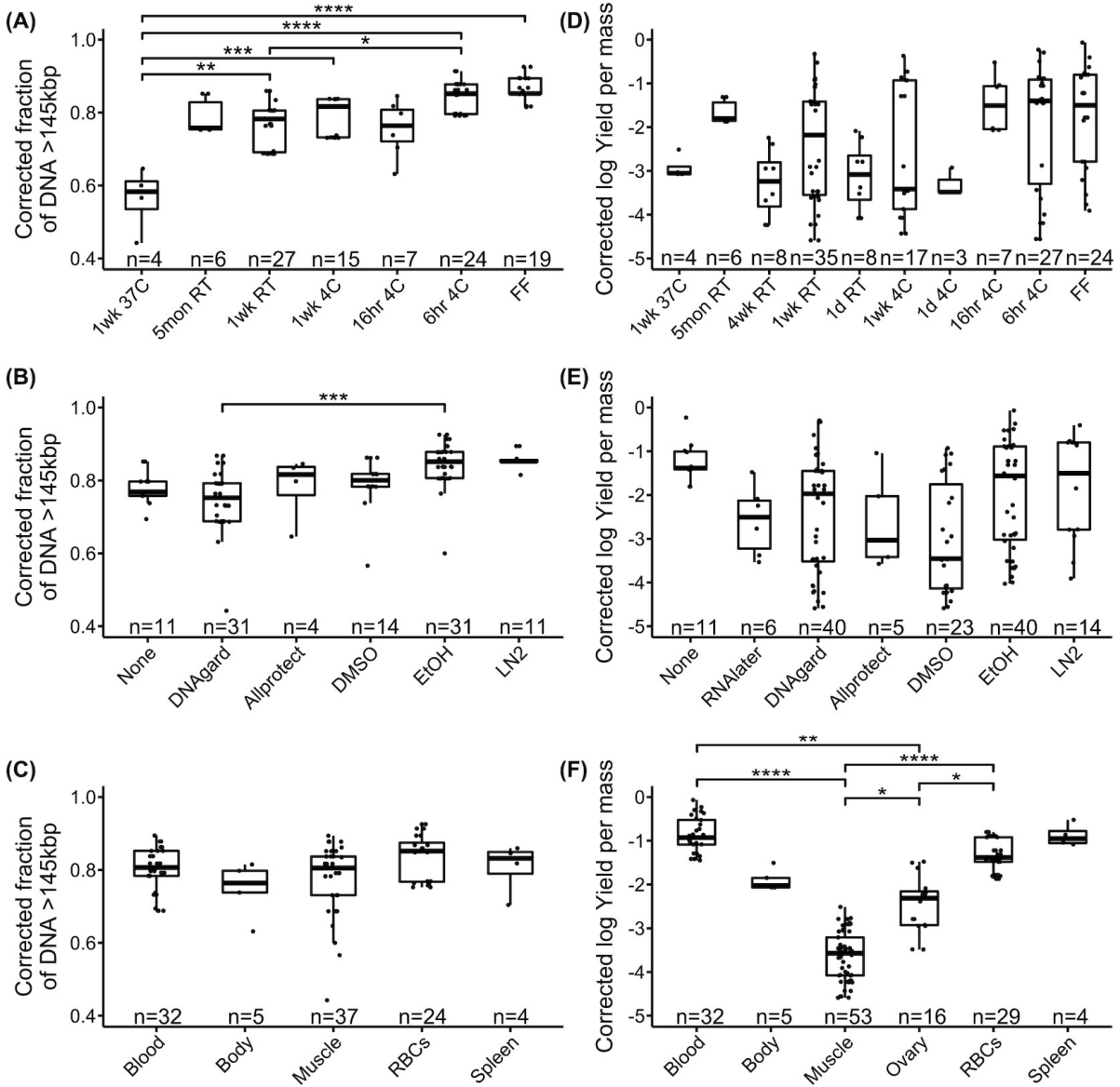
Author Notes

GigaScience, Volume 11, 2022, giac068, https://doi.org/10.1093/gigascience/giac068

Published: 10 August 2022 **Article history**

Dahn et al, 2022

RefGen: Sampling parameters matter



The most important one – TEMPERATURE after collection

Dahn et al, 2022

RefGen: Sample processing in the lab or field

- Optimally – keep the organism alive as long as possible
  - Do not freeze before dissection

- If large organism:
  - Dissect on ice
  - Lentil-size tissue parts are the best

    Weight the sample if possible

Picture by Mara Lawnizak

- Place the parts into **pre-chilled** separate (barcoded) tubes
  - Make sure the tubes are suitable for -70°C

Please, do not send us a whole frozen mammal…..

RefGen: Correct taxonomic identification

- Make sure that the specimen is correctly identified!
- Ask for second opinion if unsure

  Note: the heterogametic sex is always preferred

- Sanger-sequence DNA barcodes
  - Will help ID
  - Will be used as a tracker to safeguard against sample mix-up

- Take a picture including a **measurement instrument** (=digital voucher)

  (if possible – include a colour chart)

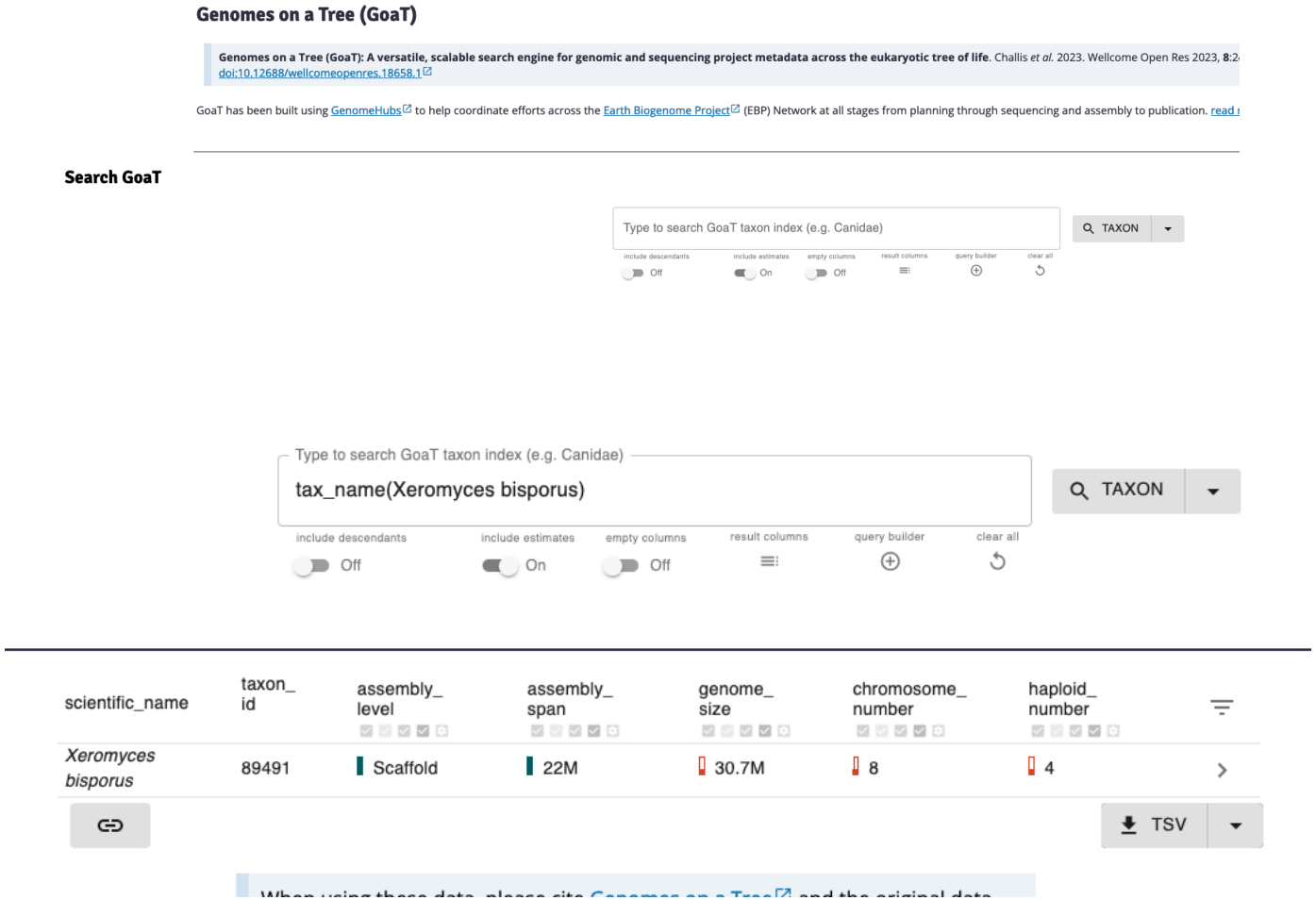
Before going to the field, check list

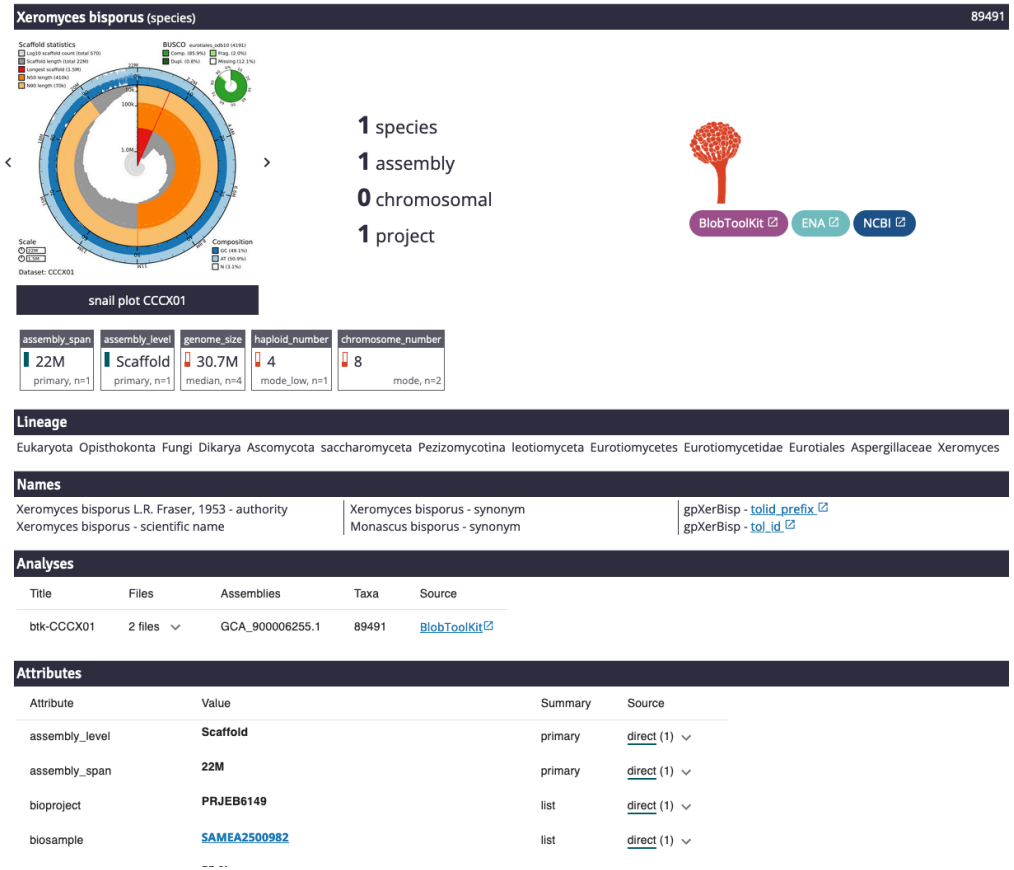
- How much material is needed? How many individuals?
- What should I bring to the field if the sample must be processed there? Can we invest in a dry shipper or a portable fridge?
- Can my sample be preserved in ethanol (check with seq center and literature!)
- If the genome is supposed to be annotated – bring along RNALater or TRIZOL for the dedicated sample

- Record metadata (FAIR):
  - living stage of the organism
  - sex
  - body / organism part
  - time and temperature between sampling and preservation
  - GPS coordinates

How to check if someone is already sequencing the species you are interested in?

- **Talk to the GoaT!** https://goat.genomehubs.org/





Not just RefGen: Legal issues - Nagoya & CITES

TAKE IT SERIOUSLY

Non-compliance: jail sentence, fine, paper retraction, etc

EU ABS regulation (Regulation (EU) Nr. 511/2014) …

→ What KIND OF MATERIAL are you using? (material scope)

„genetic material of actual or potential value“

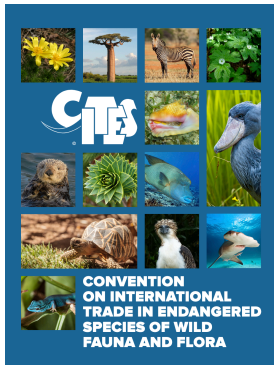
“any material of plant, animal, microbial or other **(non human)** origin containing functional units of heredity i.e. genes.”

Biological material that contains DNA/RNA (dead or alive)

excludes: human DNA ≠ human pathogens & microbiome, plant genetic resources under the ITPGRFA and influenza strains under the PIP framework if they are used under treaty conditions (**plant example**)

Courtesy: Scarlett Sett, Kiel University, Germany

CITES - trade with endangered species (cites.org)



A paperwork nightmare from HELL

Get in touch with your governmental authorities at least 4 months prior to intended shipment

And do not forget all other “normal” import/export permits for shipment of biological material!

Sample quality requirements

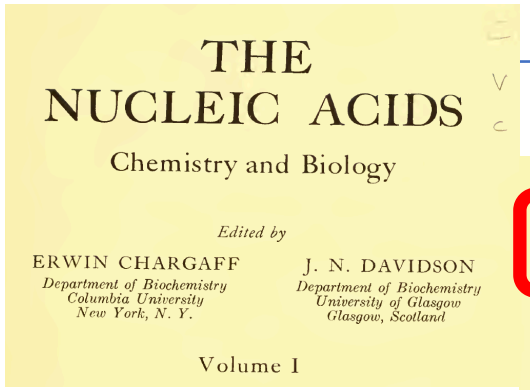
Garbage in – garbage out:

Sequencing success **always** depends on the sample quality

NGS-quality DNA and
PCR-quality DNA
are two completely different things

Especially for long-read sequencing

Considering DNA extractions…



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 to 20 x 10^6 d (1, 2). The single strand molecular weight was often around

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

ELSEVIER

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

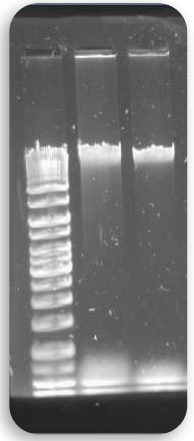
Simple, Efficient, and Nondestructive DNA Extraction Protocol for Arthropods

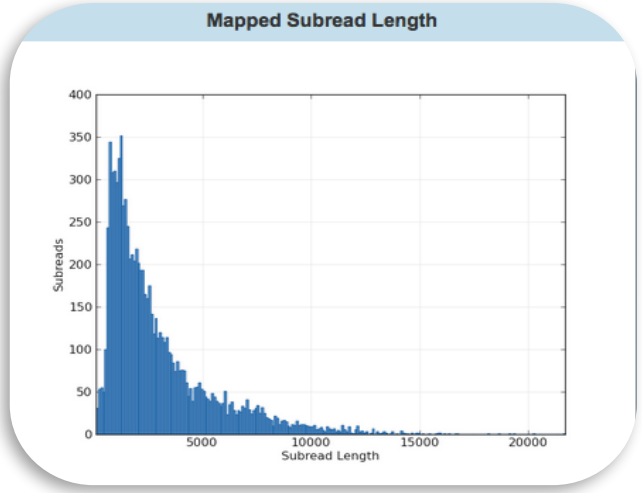
Aloysius J. Phillips, Chris Simon

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

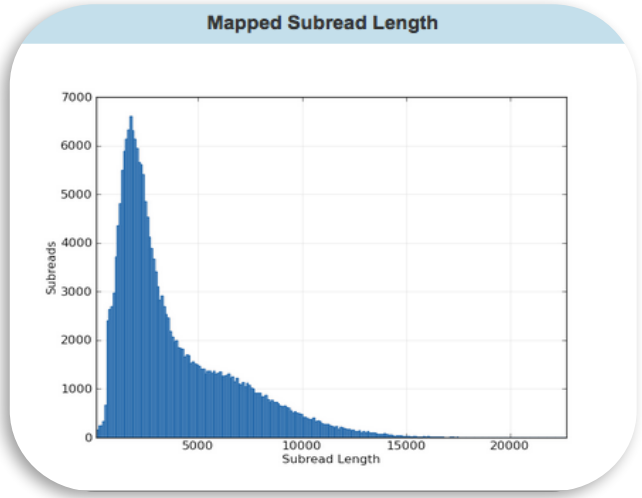
Published: 01 May 1995 **Article history** ▾

2013: a wake-up call





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

For Long Reads one needs to have *long and pure* DNA

SciLifeLab

DNA quality and inhibition of sequencing

Short-read technologies: PCR inhibition
Long-read technologies are PCR-free, but one sequences native DNA “as is”.

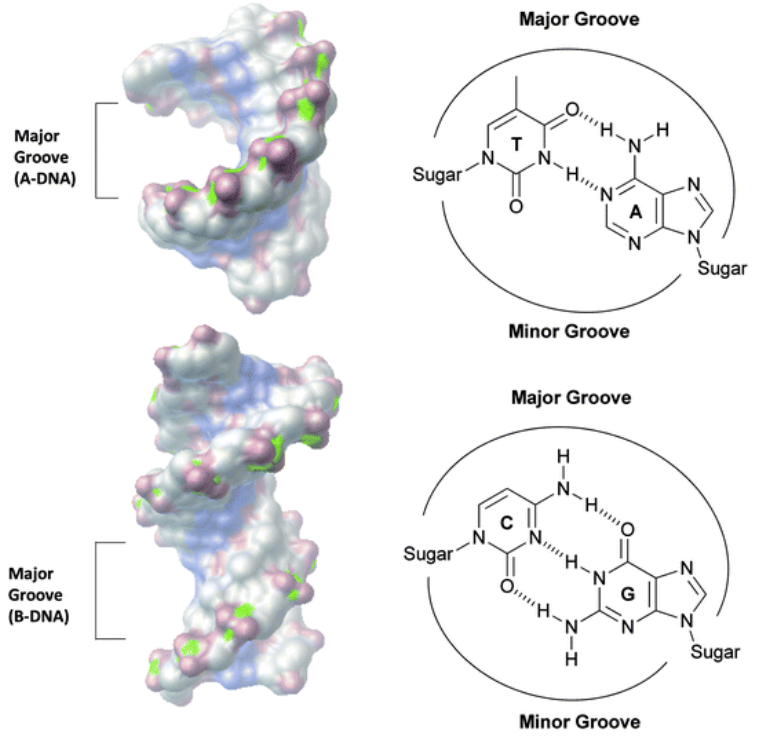
DNA-binders:

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

Polymerase inhibitors:

- Salts
- Phenol
- Alcohols

Physical inhibiting factors – debris



Hamilton & Arya, Nat. Prod. Rep., 2012, **29**, 134-143

What do absorption ratios tell us?

Pure DNA 260/280: 1.8 – 2.0

< 1.8:

Too little DNA compared to other components of the solution; presence of organic contaminants: proteins and phenol; glycogen - **absorb at 280 nm**.

> 2.0:

High share of RNA.

Pure DNA 260/230: 2.0 – 2.2

<2.0:

Salt contamination, humic acids, peptides, aromatic compounds, polyphenols, urea, guanidine, thiocyanates (latter three are common kit components) – **absorb at 230 nm**.

>2.2:

High share of RNA, very high share of phenol, **high turbidity,** dirty instrument, wrong blank.

Photometrically active contaminants:
phenol, polyphenols, EDTA, thiocyanate, protein,
RNA, nucleotides (fragments below 5 bp)

Help! My absorption values are bad!!!

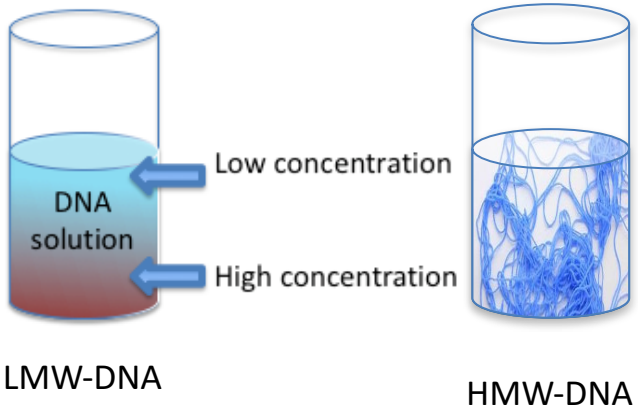
Qiagen DNeasy Power Cleanup Pro

Zymogen gDNA cleanup and concentrator

Besides: AMPure beads, phenol-chloroform-CTAB, etc

Check protocols.io!

How to make a correct DNA measurement



- Thaw DNA completely
- Mix gently (**never vortex!**)
- Put the sample on a thermoblock: 37°C, 15-30 min
- Mix gently
- **Dilute 1:100** (if HMW)
- Mix gently
- Make a measurement with an appropriate blank

- **NANODROP is Bad**. Point.
- Use Qubit, or PicoGreen.
- Nanodrop value : Qubit value ≤ 50%

Causes of DNA degradation/damage

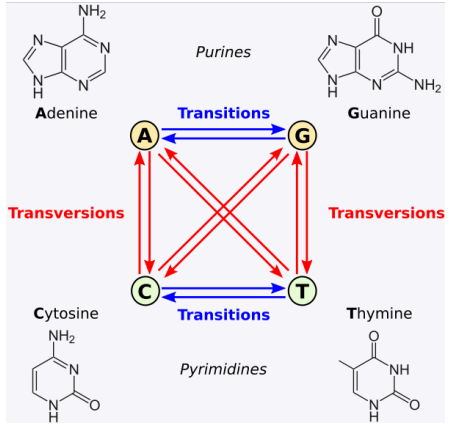
Mechanical damage during tissue homogenization.

Wrong pH and ionic strength of extraction buffer (-> hydrolysis).

Incomplete removal / contamination with **nucleases**.

Phenol: too old, or inappropriately buffered (**pH 7.8 – 8.0**); incomplete removal.

Wrong pH of the **DNA solvent** (acidic water).
Recommended: Low TE for short-term storage, 1xTE for long-term storage.



Vigorous pipetting (wide-bore pipet tips).

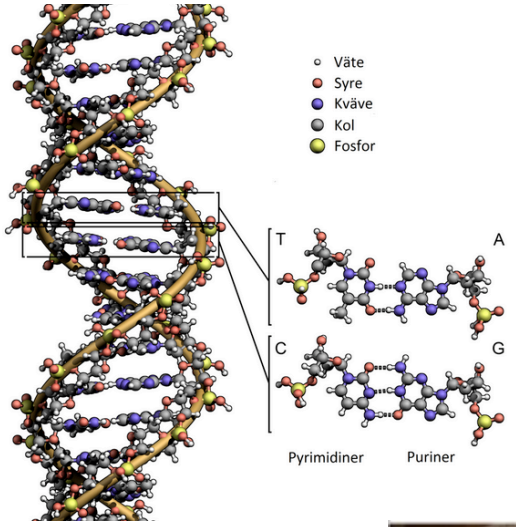
Vortexing of DNA in high concentrations.

Too many **freeze-thaw** cycles (*we tested 5, still Ok*).

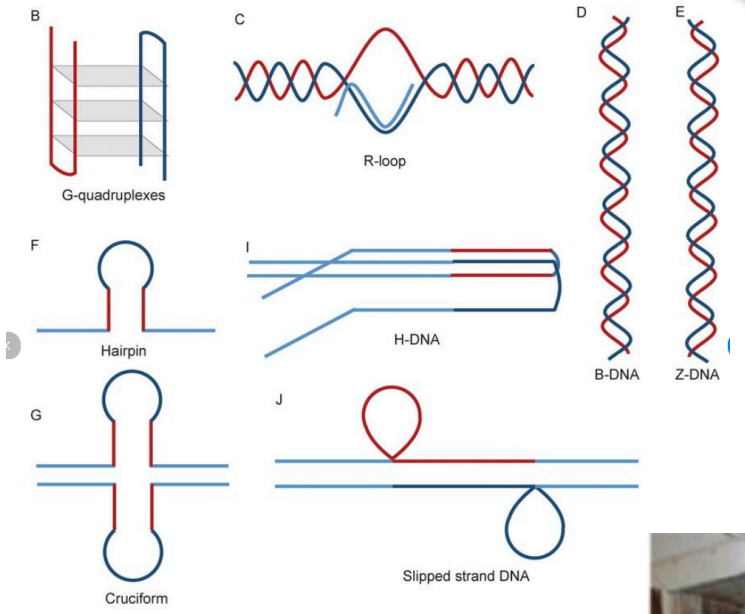
Sequence-dependency: depurination, deamination, T-C transitions… https://www.biorxiv.org/content/10.1101/254276v3

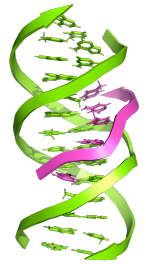
To keep in mind

There are lots of surprises, more so in non-model organisms

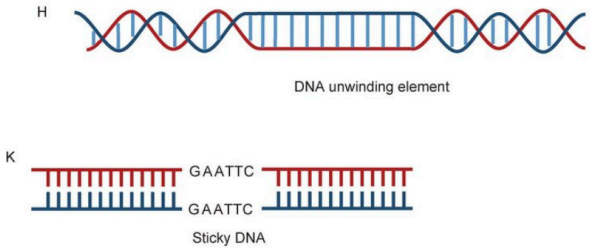


What textbook tells you

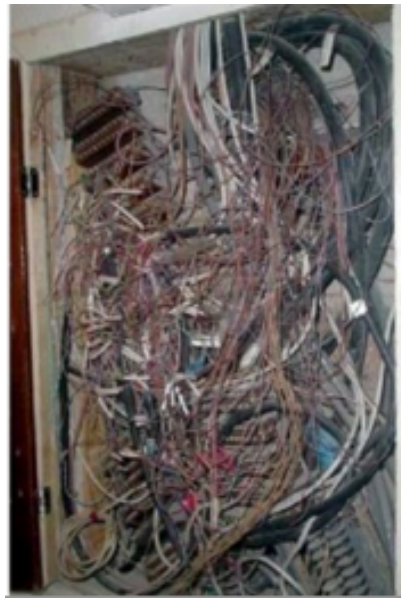




Brutal reality



Do not forget:
DNA in solution behaves differently



What every sequencing facility wished you knew before starting your project

Sequencing facilities and their sample requirements

Two types: commercial and non-profit (university-based)

Some can do DNA / RNA extractions, some do not

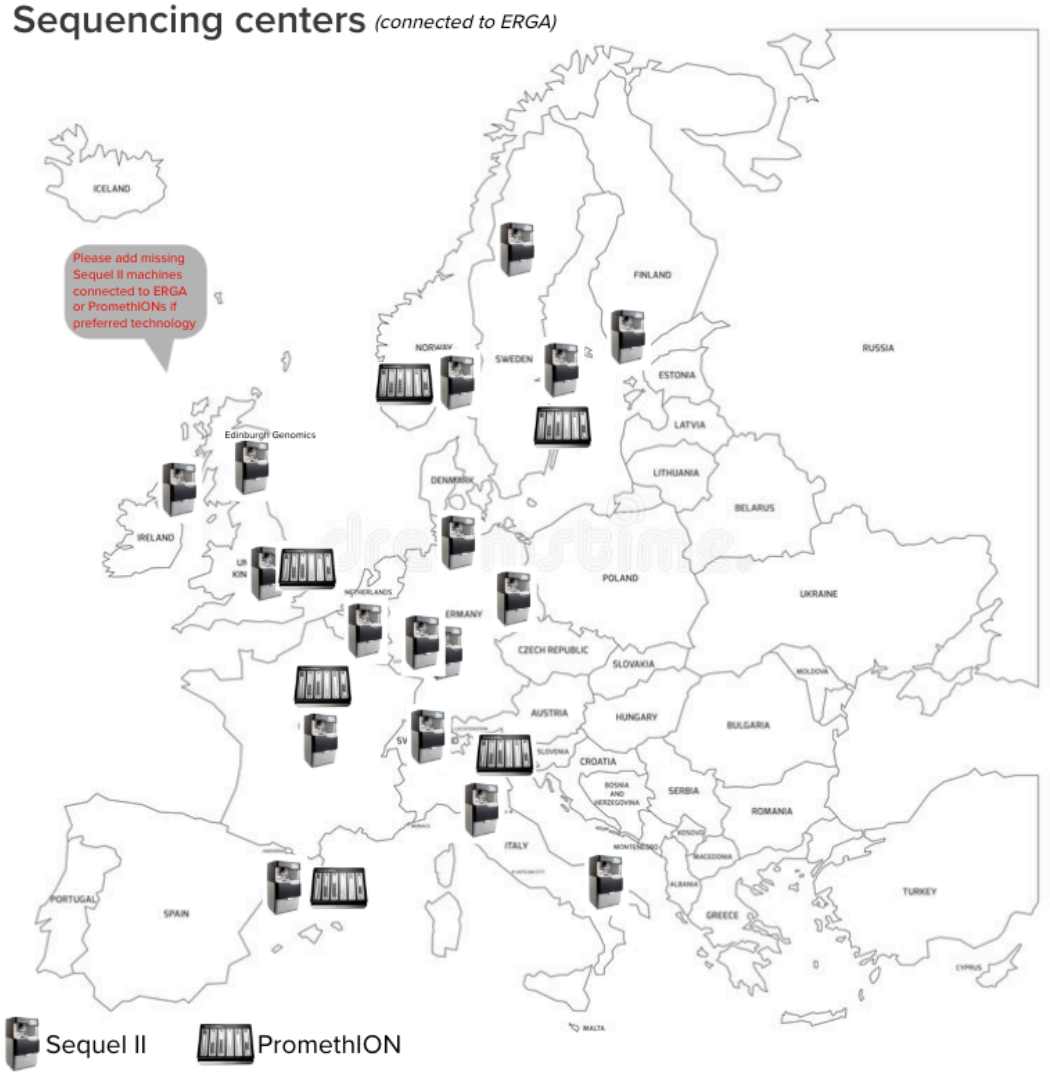
Number of samples and tissues vary (from 1 to 20)

Amount of tissue / DNA vary a lot

Some accept ethanol-preserved samples, some do not

Proximity is important

No-one can answer how long time it will take



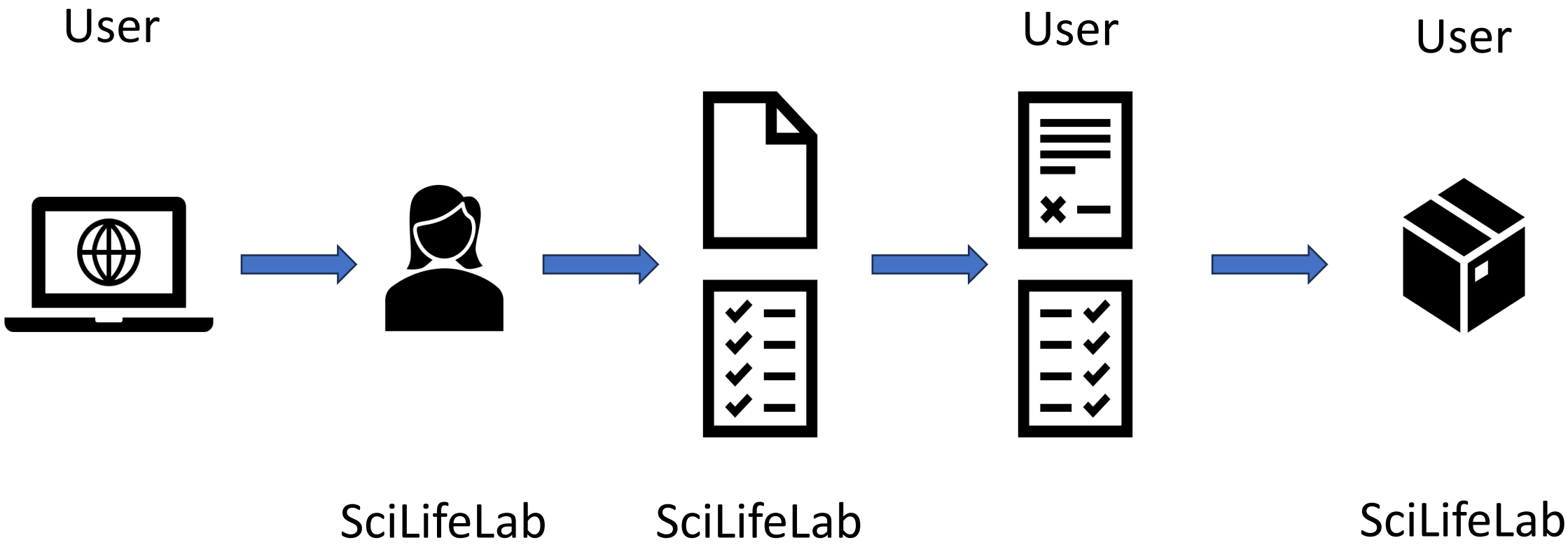
Shipment…

- ALWAYS solid-frozen on dry ice
- Ask for dry-ice top-up
- DHL, FedEx, UPS – all have issues, unfortunately
- BIOCAIR – used to ship human transplants. Expensive, but worth it.
- WorldCourier is very good as well. Not cheap either.

Sequencing facility documentation

Paperwork is a necessary evil.

SciLifeLab example:



Optimal project workflow

1. Check the latest sequencing recipe / application
2. Get in touch with the sequencing center(s), ask for their sample requirements and necessary paperwork
3. Study papers on similar taxonomic groups – check how were the samples collected and preserved
4. Check **ALL THE PERMITS** (ethics, collection, ABS, Nagoya, CITES, import / export)
5. Collect in the field / request from biobank / assess own stock
6. **Record metadata**
7. *For RefGen:* ID the sample (use DNA barcodes for non-models)
8. *For RefGen*: Deposit a voucher / biobank accession
9. Arrange all documentation required by the sequencing facility
10. Get in touch with a courier company
11. Ship to the sequencing center

Considering costs

Sequencing project cost = collection + sample processing + nucleic acid extraction + shipment* + sequencing + data storage + data compute + data analysis + work hours

** Shipment field-lab, lab-sequencing facility, lab-vouchering collection (left-over material?)*

Collection + your & bioinformatican's salaries = **MOST EXPENSIVE** part of the project

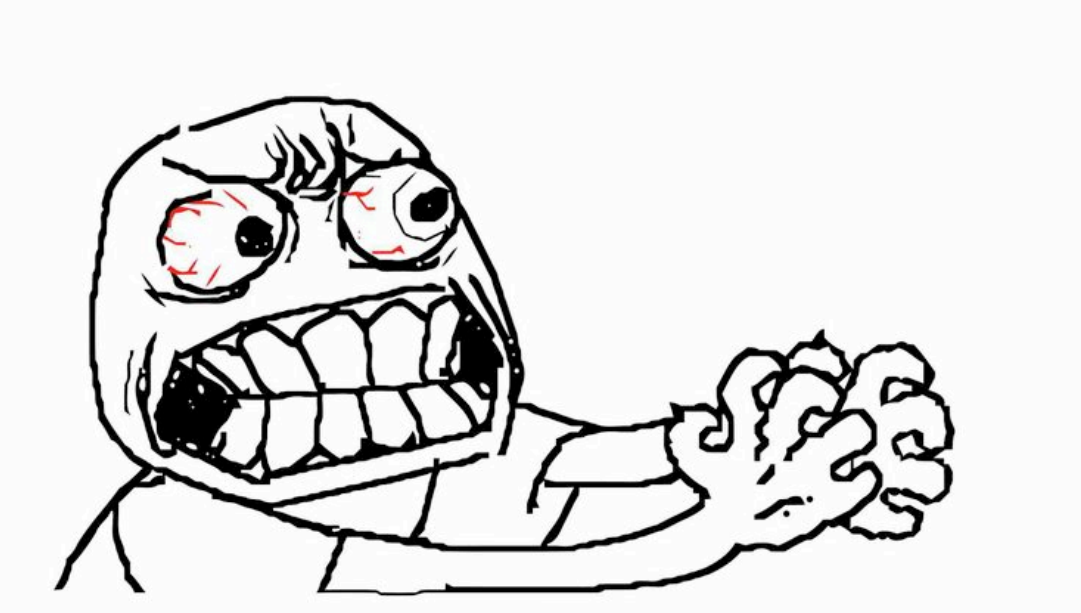
Sequencing is the **CHEAPEST** part

(Never EVER believe the sequencing technology vendor prices!)

Data storage and computing resources cost (a lot of) money

Reagent Cost PacBio, Sequel		Price per kit (SEK)	Units per kit	Price / Unit	Units	Cost (SEK)	Price* / Unit	Units	Cost (SEK)	Cost (SEK)
Sample preparation										
HiFi SMRTbell Express Template Prep Kit 2.0+ Enz Clean Up	Cost per library	16279	9	1809	1	**1809**	2532	1	**2532**	**2532**
AMPure Clean-up	Cost per library	2898	20	145	1	**145**	203	1	**203**	**203**
QC and Size Selection										
Megaruptor DNA shearing 2-20 kb fragments	Per sample up to 10 µg DNA	N/A	N/A	105	2	**210**	155	1	**155**	**155**
SageELF	Cost per library			378	1	**378**	612	1	**612**	**612**
Femto PFGE	Input QC, 1-11 samples	N/A	N/A	960	4	**3840**	1350	4	**5400**	**5400**
Consuambles tubes, tips, Ampure Beads	Cost per library	N/A	N/A	220	1	**220**	220	1	**220**	**220**
Sequencing Reagents										
Sequel™ SMRT® Cell 8M v3 Tray (4 cells)	Per SMRT cell	40801	4	10200	1	**10200**	14280	1	**14280**	**14280**
Sequel Sequencing Kit 2.0 Bundle 4rxn	Per SMRT cell	46292	20	2315	1	**2315**	3240	1	**3240**	**3240**
Sequel Sequencing consumables	Per run (4 SMRT cells)	N/A		850	0,25	**213**	927	0,25	**232**	**232**
Reagent Cost, SEK						**19329**			**26875**	**26875**
Additional cost, SEK										
Instrument related cost, run time (HiFi30, CLR15, IsoSeq24)	Per hour / Per SMRT cell	I		300	0	**Paid**	300	30	**9000**	**9000**
Work hour cost (external users only)	Per hour			400	0	**Paid**	400	32	**0**	**12800**
TOTAL project cost excl OH									**35875**	**48675**
University overhead	Per project 29%								**0**	**29 %**
TOTAL project cost incl OH						**19329**			**35875**	**62790**
***Including costs of re-run, auxilliary equipment, other reagents, etc.**						**1.9 k€**			**3.5 k€**	**6.3 k€**

Expectations vs reality



HUMAN CELL LINES ONLY!!!

HiFi sequencing at scale

With a high-density SMRT Cell, up to 4 SMRT Cells per run, and 24-hour run times[3], the Revio system with SPRQ chemistry delivers up to 480 Gb[2,5] of HiFi reads per day, equivalent to 2,500 human whole genomes[4] per year.

The $500[6] complete, phased genome

HiFi sequencing provides small variants, structural variants, repeat expansions, methylation, and haplotype phasing from a single library and sequencing run. With a comprehensive genome, you can replace multiple assays, saving valuable time and resources while gaining deeper insights in one streamlined process.

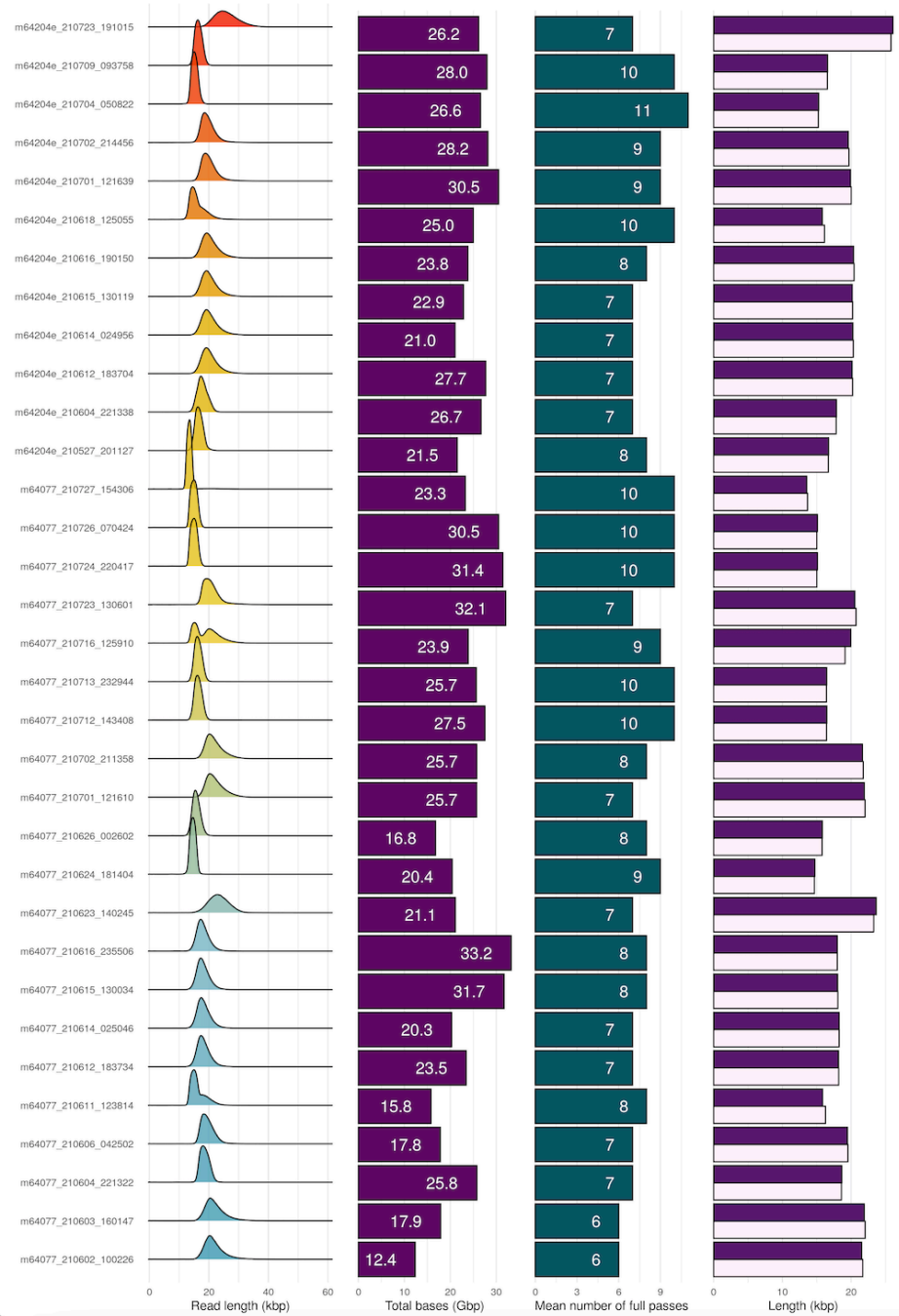
Long-read genomes at scale, with 4x less DNA input required

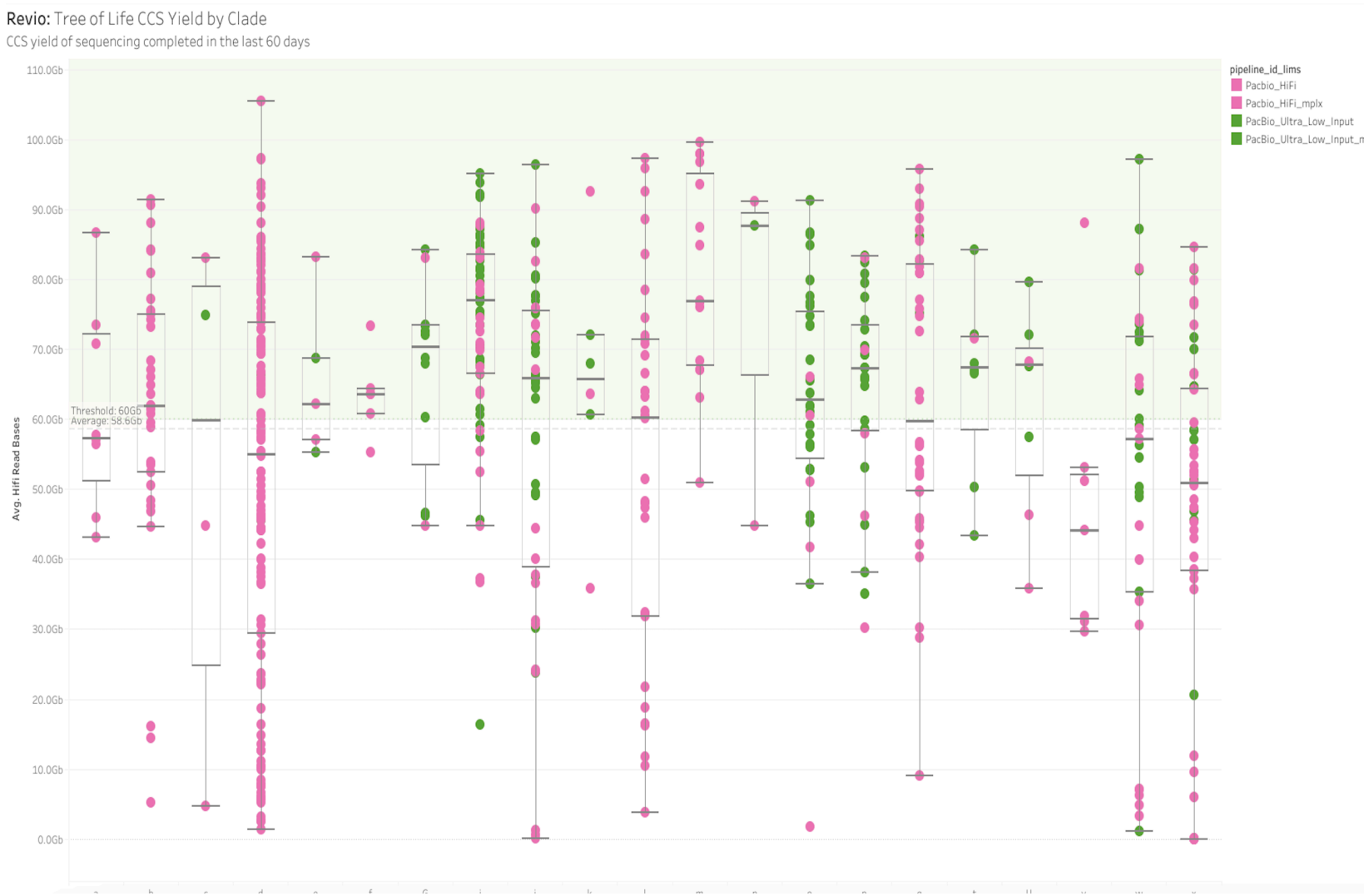
SPRQ chemistry on the Revio system unlocks the ability to sequence more sample types than ever before, using just 500 ng of native DNA – without sacrificing the high throughput or exceptional quality you rely on.

On-instrument 5mC and 6mA caller for multiomic Fiber-seq chromatin assay

SPRQ chemistry, paired with Google Health DeepConsensus algorithms, delivers exceptional read accuracy plus confident 5mC and 6mA methylation detection in every run. Optimized file formats reduce data storage needs through quality value binning and smart read ordering, streamlining data handling and maximizing efficiency.

Applicable to ANY vendor





Courtesy: James Watt, Sanger Institute (DToL)

Finally: Recognize the sequencing facility personnel

Shift from specialized labs to sequencing facilities

Sequencing of non-model organisms, especially for RefGen generation:

Heavily reliant on pure, HMW-DNA

High failure rate both for PacBio and ONT

Everything is non-model
Every project is practically R&D
Some projects require weeks of full-time expert lab engineer

Wrapping it all up:

- Some perspective
- What to think about BEFORE planning a sequencing project (aka Project Design)
- Sequencing applications and experiment design specifics:
    - Whole-genome sequencing
    - Targeted sequencing
    - Transcriptome sequencing
    - Single-cell sequencing
    - Shotgun metagenomics
    - Reference genome sequencing + optimal project workflow example

- Sampling and sample quality requirements
- What every facility wish you knew before sending your samples

Experimental design in genomics – take home

VERY fast development

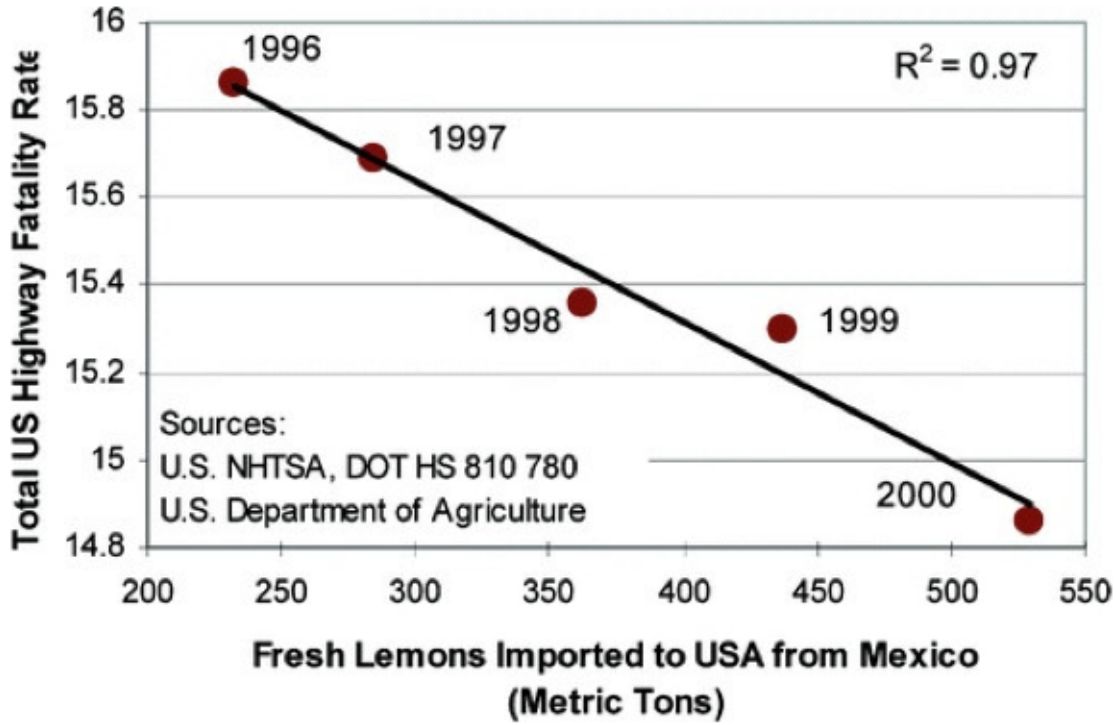
Difficult to keep oneself updated

Ask your sequencing service provider about the latest updates

Sequencing itself is the CHEAPEST part of the project

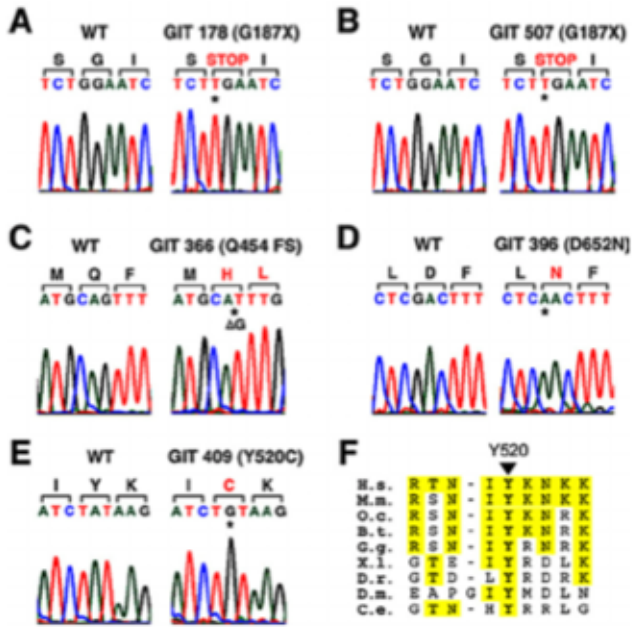
Finally, 2 minutes of philosophy

Never forget: Correlation vs Causation



Reduction in export of fresh lemons from Mexico causes significant reduction of highway traffic fatality rates in the US!

Genome is not a linear string of bases!!



Mutations in coding regions only

↓

Transcriptional & post-transcriptional regulation

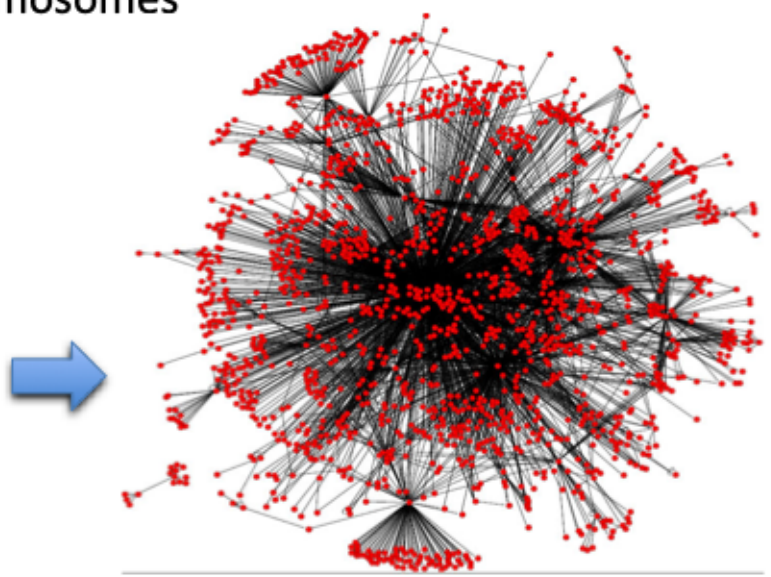
↓

Epigenetics

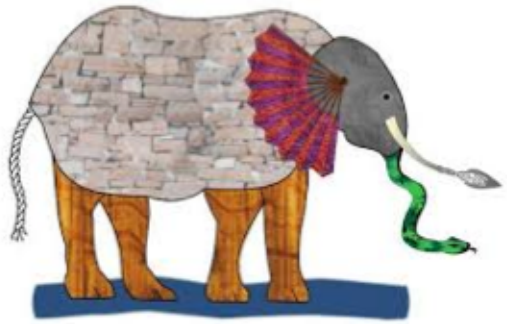
↓

Proximity in chromosomes

```
>gi|30018278|ref|NC_004722.1| Bacillus cereus ATCC 14579, complete genome
TAGCCACTTTTTTTTGATATTATAGTTGTGTTTTCACTTTGAATAAGTTTTCCACATCTTTATCTTATCC
ACAATTTGTGTATAACATGTGGACAGTTTTAATCACATGTGGGTAAATAGTTGTCCACATTTGCTTTTTT
TGTCGAAAACCCTTTCTCAATACAAACGACGTTTTAGGTTTTAAAATCGTTTTCGTATAAATATACATTT
TAATTTATTAGGTTGTACATTTGTTGCACAACCTTTATTCTTTTACCAACTTAGTAAAGGAGGGACACCT
TTGGAAAATATCTCTGATTTATGGAATAGTGCCTTAAAAGAATTAGAAAAAAAGGTAAGCAAGCCTAGTT
ATGAGACATGGTTAAAATCCACAACGGCTCATAACTTGAAGAAAGACGTATTAACGATTACAGCTCCGAA
TGAATTTGCTCGTGACTGGCTAGAATCTCATTACTCCGAACTAATTTCAGAAACACTATACGATTTAACA
GGGGCAAAATTAGCAATTCGCTTTATTATTCCCCAAAGTCAAGCTGAAGAGGACATTGATCTGCCTCCAG
TTAAGCCGAATCCAGCACAAGATGATTCAGCTCATTTACCACAGAGCATGTTAAATCCAAAATATACATT
CGATACATTTGTTATTGGCTCTGGTAACCGTTTTGCCCATGCAGCTTCATTAGCTGTAGCTGAGGCGCCA
GCTAAAGCGTATAATCCACTCTTTATTTACGGGGGAGTTGGACTTGGAAAGACACATTTAATGCACGCAA
TTGGTCATTATGTAATTGAACATAATCCAAATGCAAAAGTTGTATATTTATCATCAGAAAAATTTACAAA
TGAATTTATTAACTCTATTCGTGATAATAAAGCTGTTGATTTTCGTAATAAATATCGTAACGTAGATGTT
TTATTGATAGATGATATTCAATTTCTTGCTGGAAAAGAACAGACTCAAGAAGAGTTTTTCCATACATTTA
ACGCATTACACGAAGAAAGTAAACAAATTGTAATTTCTAGTGACCGACCACCAAAAGAAATTCCAACTTT
```



Blind men & an elephant



Letter

Genome-wide association study identifies five new schizophrenia loci

The Schizophrenia Psychiatric Genom

Nature Genetics **43**, 969–976 (2011)
doi:10.1038/ng.940
Download Citation

Article

Genome-wide association analysis identifies 30 new susceptibility loc schizophrenia

Zhiqiang Li, Jianhua Chen [...] Yongyong Shi

Nature Genetics **49**, 1576–1583 (2017)
doi:10.1038/ng.3973
Download Citation

Received: 17 April 2017
Accepted: 19 September 2017
Published online: 09 October 2017

Current opinion in psychiatry
Author Manuscript HHS Public Access

Genome-wide association studies (GWAS) of schizophrenia: does bigger lead to better results?

Sarah E. Bergen, PhD and Tracey L. Petryshen, PhD

Comment | Open Access

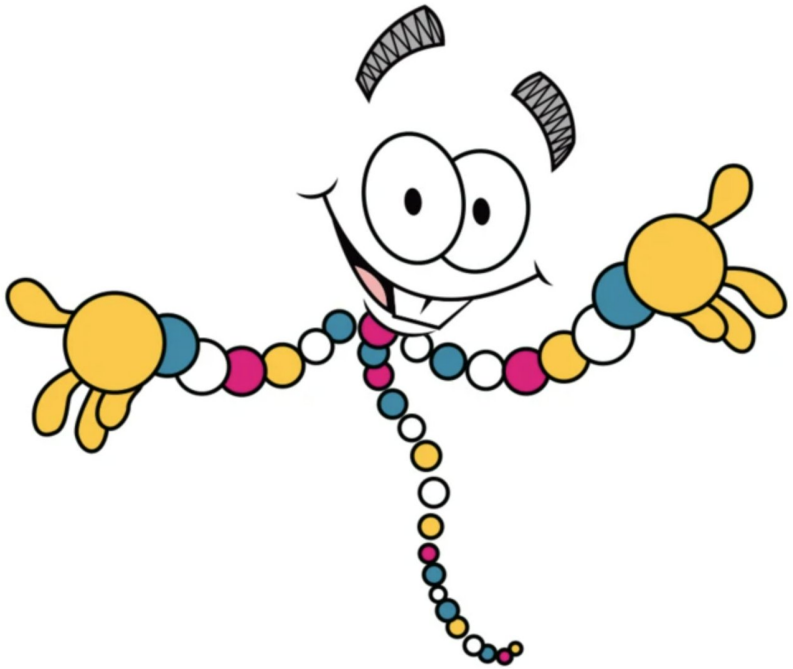
Schizophrenia and the dynamic genome

Patrick F. Sullivan

ımmary

ıriation (CNV) is a widely replicated risk factor for psychiatric disorders
hrenia, although the mechanisms by which CNVs confer risk are
ar. Recent studies have provided robust evidence of CNVs associated with
nd have highlighted a potential role for schizophrenia risk-associated

Thank you!



Swedish Research Council

UPPSALA UNIVERSITET

SciLifeLab

European Commission
Horizon Europe
2021-2027