# Genome-scale technologies 2/ Algorithmic and statistical aspects of DNA sequencing

*What to sequence next? Exciting achievements of the -seq.*

Ewa Szczurek
University of Warsaw, MIMUW

szczurek@mimuw.edu.pl

# The project

- http://students.mimuw.edu.pl/~szczurek/TSG2_Project/project.html
- Report deadline: 20.01.2016
- Presentations: 26.01.2016

# How to do the project?

## STEP 0

- Perform QC
- Report QC before preprocessing
- Remove adapters (e.g. Clip in Galaxy)
- Remove low quality reads only if the quality across bases below 28
- Report QC after preprocessing
- Deduplicate after mapping (?), e.g. with rmdup

# How to do the project?

## STEP 0

- Perform QC
- Report QC before preprocessing
- Remove adapters (e.g. Clip in Galaxy)
- Remove low quality reads only if the quality across bases below 28
- Report QC after preprocessing
- Deduplicate after mapping (?), e.g. with rmdup

## STEPS BETWEEN 0 and 1

- Mapping (e.g. With Bowtie)
- Report the percentage of reads mapped
- MACS (?)
  - not always runs
  - not always gives any peaks with acceptable FDR
  - if any of this occurs, process with the mapped reads only!
    - pysam - An interface for reading and writing SAM files
    - Rsamtools package in R
- Using the gene annotation file, select the first exon start and the last exon end as the gene boundaries.

# How to do the project?

**STEPS 1 – 3: there is no „the way” to answear the questions. The more aproaches to assess the hypotheses, the better.**

## STEP 1:Check whether the histone modifications (A) and the protein (B) bind preferentially in gene regions (5'- or 3'-end or center)?

- Report the % of all binding regions/reads for A and B that land in the genic regions (enywhere in them)

- Compare the number of reads in the genic regions to the number of reads expected as if the reads were uniformly distributed on the genome at random:
    - Compute the expected no of reads per some window in the genome
    - Compare to the avg no of reads in the same window size in the genic regions

- Cut the gene regions into bins (3 --100?). Report how many regions overlap/reads map to each of these bins. Compare the read numbers between the 3’ bins to the 5’ bins to the center bins using Wilcoxon test.

- Plot how many reads map within increasing distances from the ends. Plot the same for randomly shuffled read binding positions.
- ngs.plot

## STEP 2: Is histone modification A and the B protein binding throughout the genome signifficanlty correlated?

- Divide the genome into
  - Equal size bins
  - Regions from the annotation
- Compute the log2( read count A or B/ input ) – normalized read count
- Compute the correlation between the same bins for A and B

- Hypergeometric test: is intersection over genes surprisingly large?
  - Report the number of genes A overlaps with
  - Report the number of genes B overlaps with
  - Report the intersection size
  - Universe: all genes

## STEP 3: Are the genes which are bound by B also differentially expressed between tissues Elav and Repo?

- Again report the numbers and hypergeometric test.

# The last lecture:
# What to sequence next? Exciting achievements of the -seq

## Metagenomics:

- https://www.ted.com/talks/craig_venter_on_dna_and_the_sea?language=en#t-124864

# Metagenomics

## Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman[1], Michelle R Rondon[1], Sean F Brady[2], Jon Clardy[2] and Robert M Goodman[1]

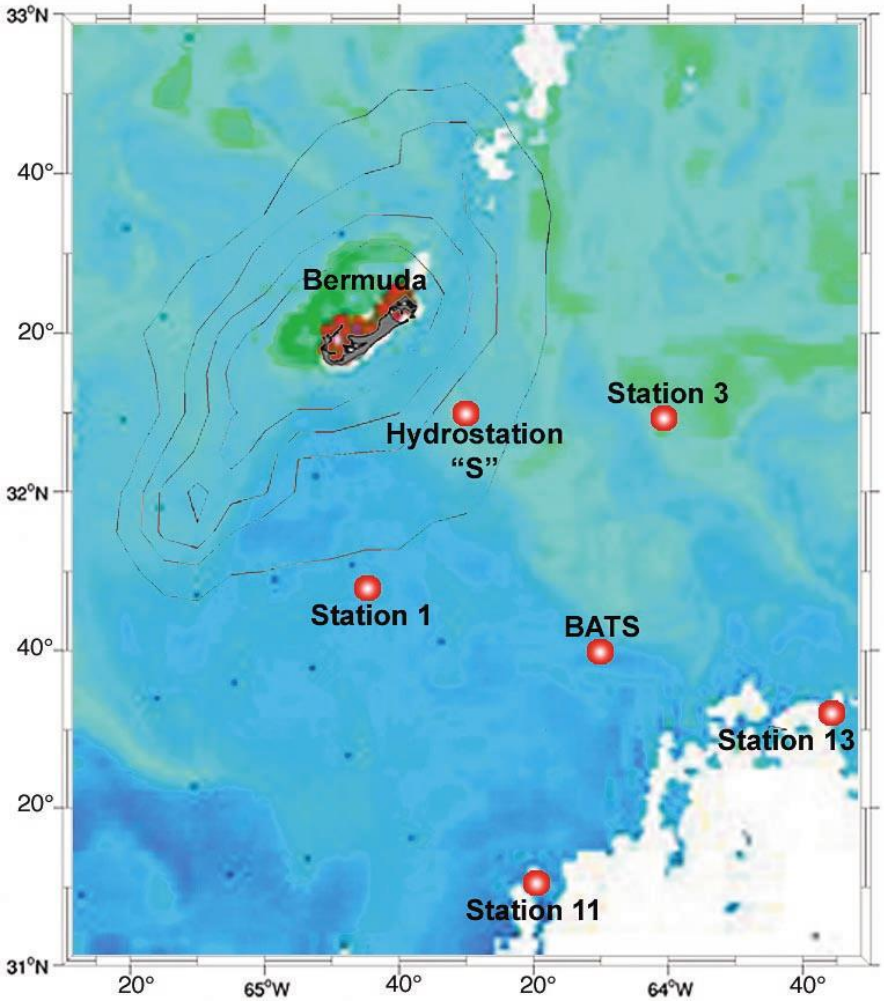
**Cultured soil microorganisms have provided a rich source of natural-product chemistry. Because only a tiny fraction of soil microbes from soil are readily cultured, soil might be the greatest untapped resource for novel chemistry. The concept of cloning the metagenome to access the collective genomes and the biosynthetic machinery of soil microflora is explored here.**

**Figure 1**


Morphological diversity typical of microorganisms cultured from soil on a broad spectrum medium, tryptic soy agar.

# Whole genome shotgun of the microbiome

- Microorganisms:
  - The most abundant and diverse organisms on Earth
  - Not possible to culture
  - Long remained uncharachterised (and vastly still remain)
- Venter et al. Science 2004
  - whole genome shotgun
  - microbial sample
  - from the Sargasso Sea



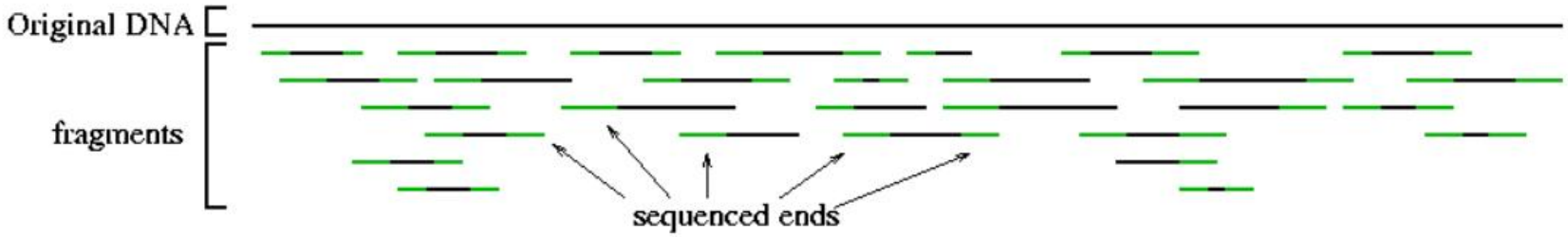

"Venter et al., Science, 2014

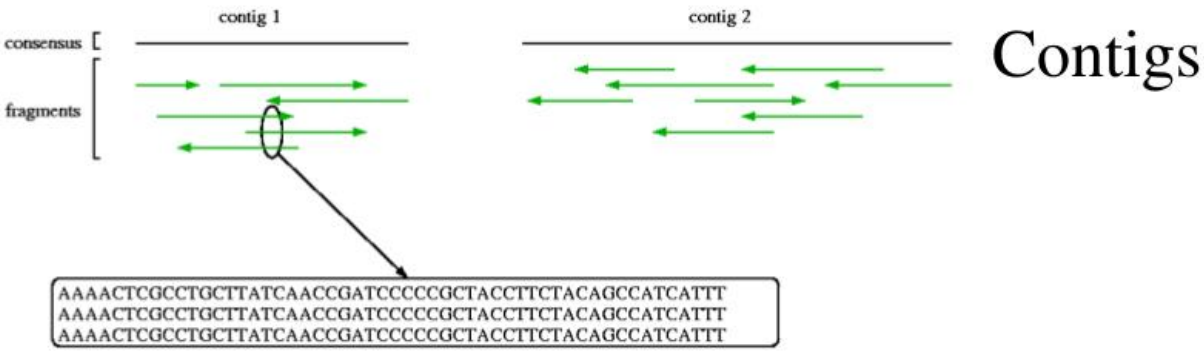
# The metagenome assembly problem

- De novo genome assembly
- Many genomes
- Don't know a priori
  - How many
  - Which proportions (some species dominate the sample, some are represented in just a few copies)

"Venter et al., Science, 2014

# De nove genome assembly




Contigs

Scaffolding

# Determining species

- Venter et al., to sort the assembly pieces into organism „bins“
  - Characterized scaffolds with at least 3x coverage depth
  - Sorted by
    - Coverage
    - Oligonucleotide frequencies
    - Similarity to previously sequenced genomes

- One walk in de Bruijn graph

+ prunning to eliminate sequencing errors and misassemblies
+ information about coverage = one organism

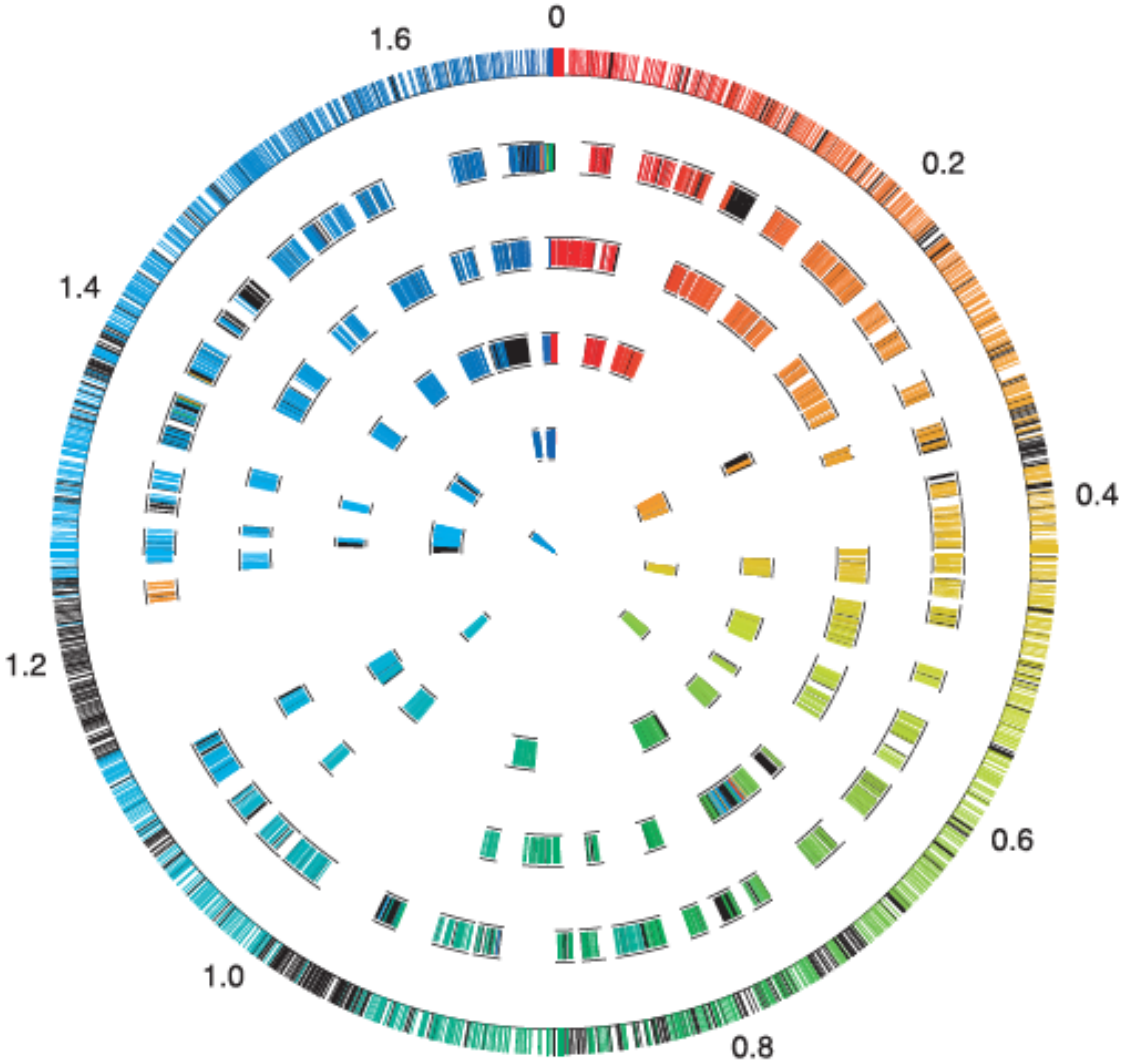


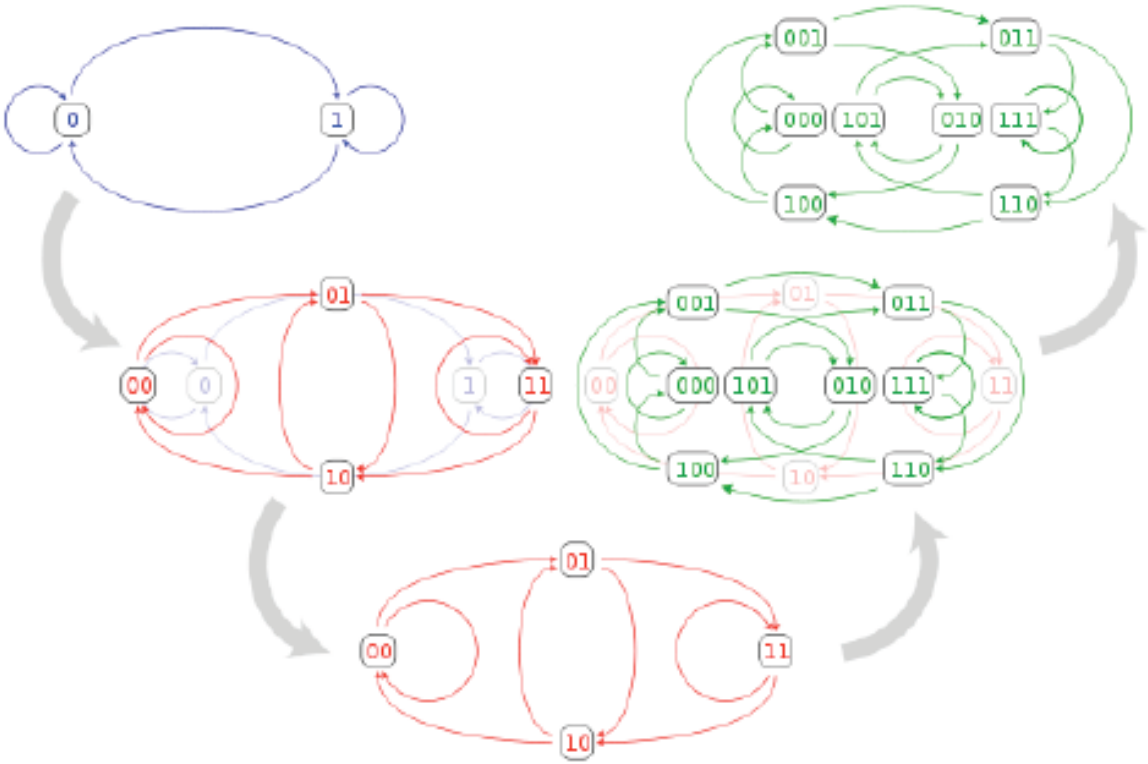

**Fig. 2.** Gene conservation among closely related *Prochlorococcus*. The outermost concentric circle of the diagram depicts the competed genomic sequence of *Prochlorococcus marinus* MED4 (*11*). Fragments from environmental sequencing were compared to this completed *Prochlorococcus* genome and are shown in the inner concentric circles and were given boxed outlines. Genes for the outermost circle have been assigned psuedospectrum colors based on the position of those genes along the chromosome, where genes nearer to the start of the genome are colored in red, and genes nearer to the end of the genome are colored in blue. Fragments from environmental sequencing were subjected to an analysis that identifies conserved gene order between those fragments and the completed *Prochlorococcus* MED4 genome. Genes on the environmental genome segments that exhibited conserved gene order are colored with the same color assignments as the *Prochlorococcus* MED4 chromosome. Colored regions on the environmental segments exhibiting color differences from the adjacent outermost concentric circle are the result of conserved gene order with other MED4 regions and probably represent chromosomal rearrangements. Genes that did not exhibit conserved gene order are colored in black.

# De Bruijn graph

A de Bruijn graph is a directed graph of overlapping sequences, each edge adding one character on the right and removing one character from the left.
A De Briujn graph is called "*n-dimensional*" if the sequences are composed of an alphabet of size n.Graph V

$$V = S^n = \{(s_1, \ldots, s_1, s_1), (s_1, \ldots, s_1, s_2), \ldots, (s_1, \ldots, s_1, s_m), (s_1, \ldots, s_2, s_1), \ldots, (s_m, \ldots, s_m, s_m)\}.$$

where $S := \{s_1, \ldots, s_m\}$ is the alphabet of the sequence

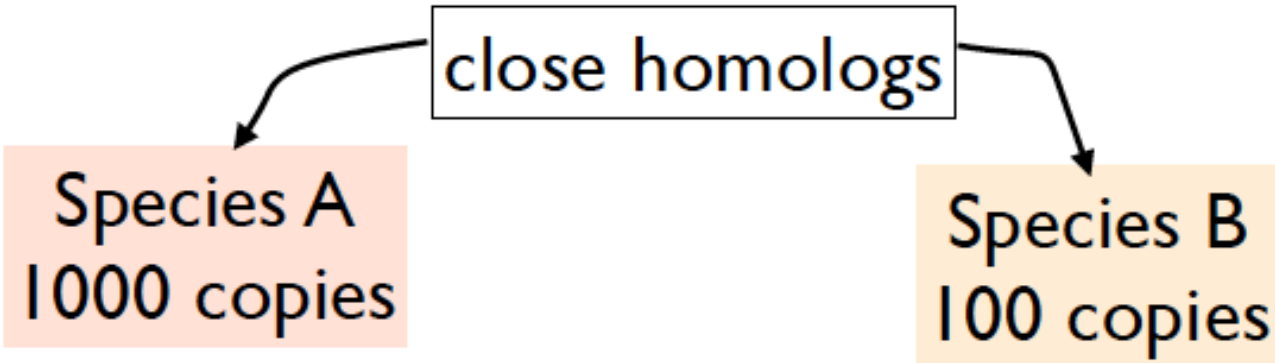


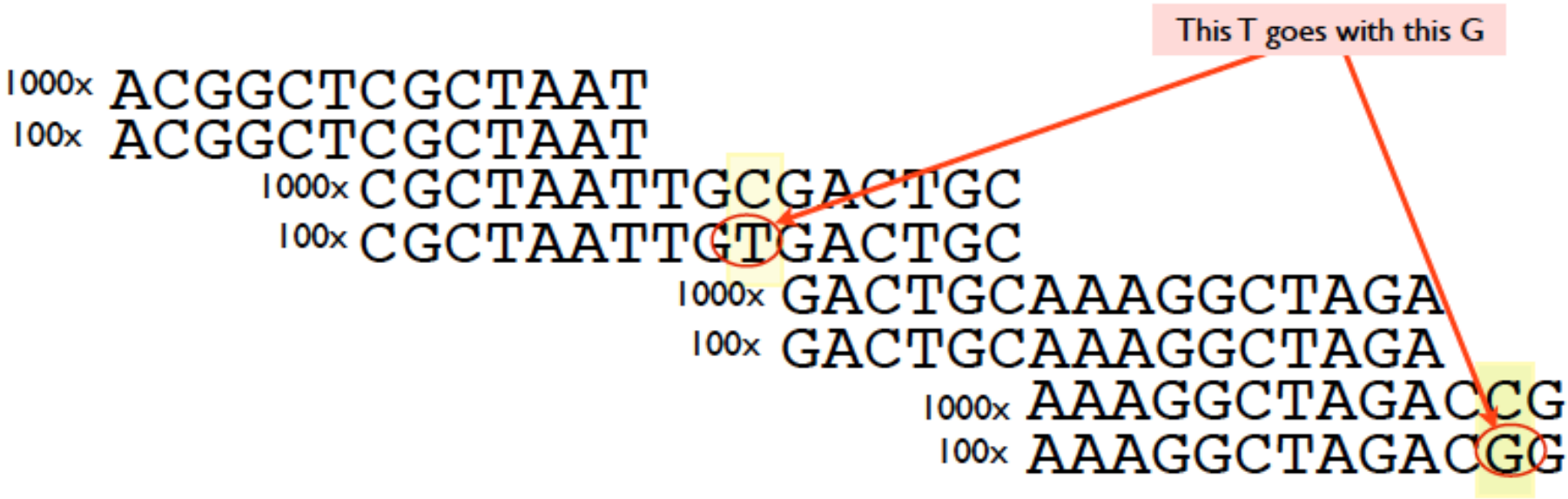

2-dimensional de Bruijn graphs of length 1, 2, 3.

# Path finding in de Bruijn graphs

- 1 path = a sequence of verteces
- 1 path =1 genome
- Edges with only one occurence can be pruned (errors tend to NOT occur in the same place twice).
- *Bubbles* and *tips* may be pruned by the "Tour bus" algorithm. (not included)
- Ambiguous paths may represent multiple strains of a species, or very similar species, and may be separated out using <u>abundances</u>.

# Abundance data

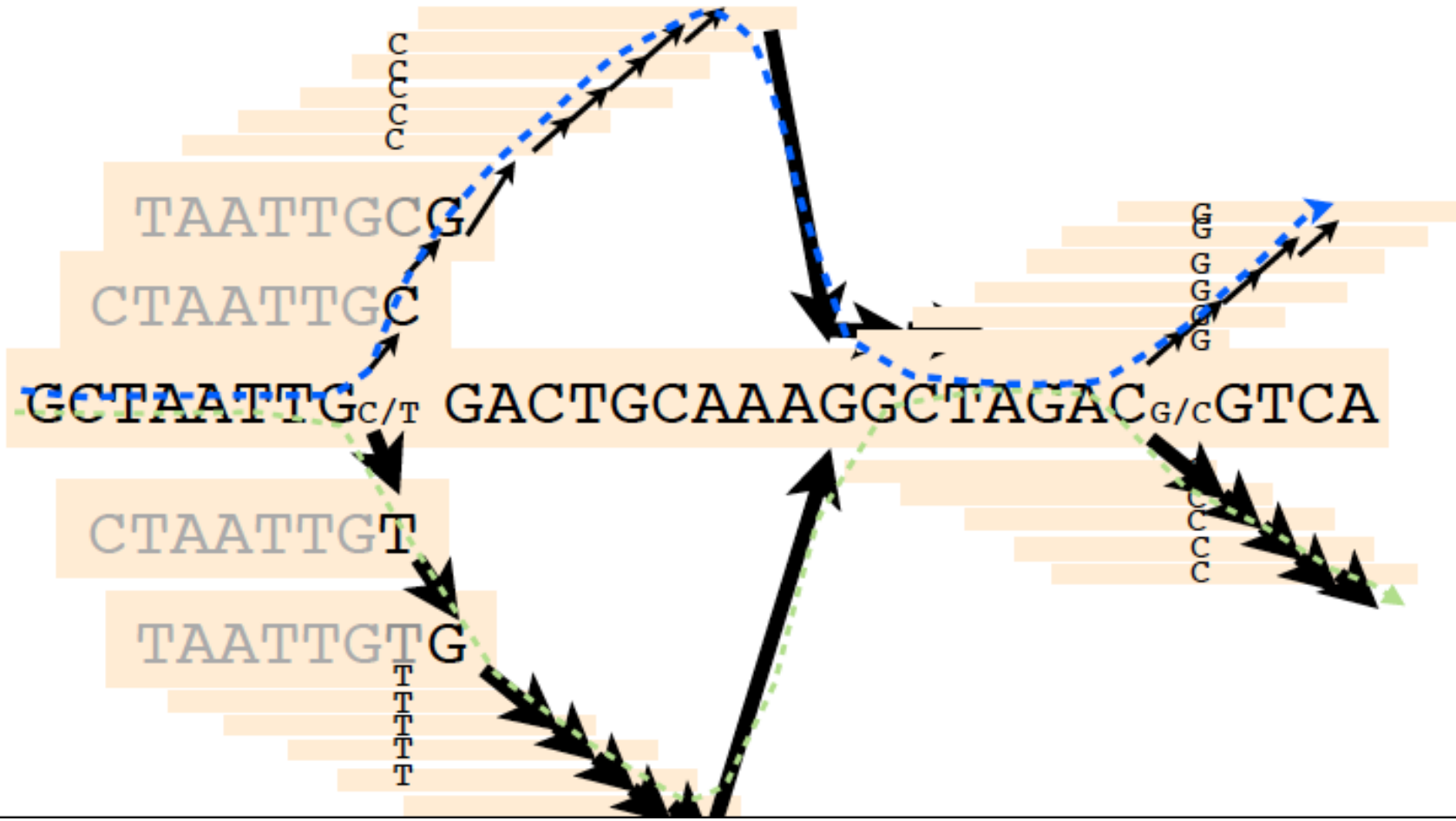



Relative abundance of *mutations* should match relative abundance of *species*, and can be used to resolve ambiguous assemblies.

# Using relative abundance for path finding in De Bruijn graphs

Draw an edge only where overlap is exact.
Initialize all edge weights = number of occurences (arrow thickness)
Identify branched pathways, where there is > 1 way to connect any two verteces.
Classify branches by occurrence weight.
Find a path that stays within an occurrence class.



# Human Microbiome

- Collective of the human microbiome exceeds the number of human cells (somatic and germ cells) by at least an order of magnitude

- The majority of the human microbiome remains unknown

- Many of these microbial interactions endow or enhance human physiology including processes related to development, nutrition, immunity and resistance to pathogens

- Many relationships between the human host and microbiome remain to be determined

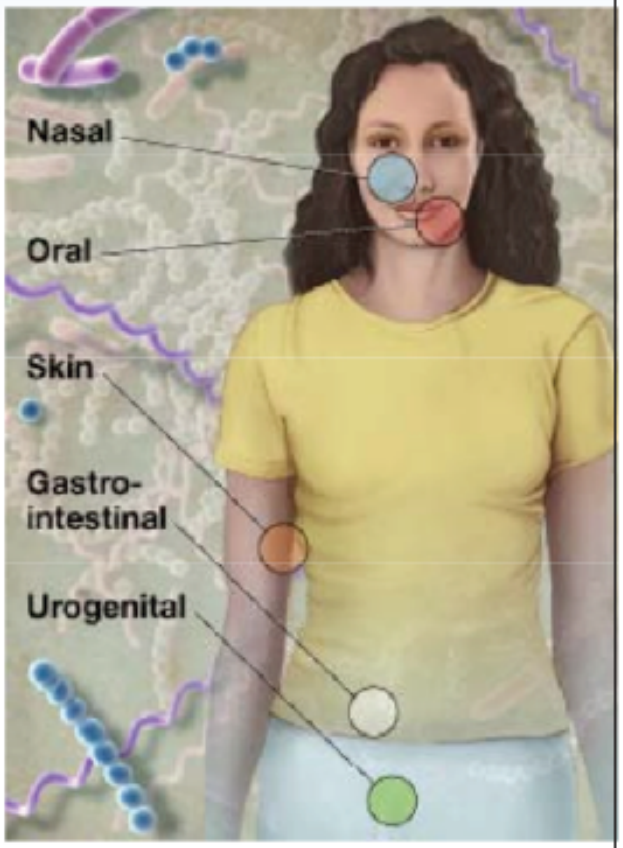



image courtesy of the NIH HMP website
http://nihroadmap.nih.gov/hmp/

Sciencexpress

# Highly Multiplexed Subcellular RNA Sequencing in Situ

Je Hyuk Lee,[1,2]*† Evan R. Daugharthy,[1,2,4]* Jonathan Scheiman,[1,2] Reza Kalhor,[2] Joyce L. Yang,[2] Thomas C. Ferrante,[1] Richard Terry,[1] Sauveur S. F. Jeanty,[1] Chao Li,[1] Ryoji Amamoto,[3] Derek T. Peters,[3] Brian M. Turczyk,[1] Adam H. Marblestone,[1,2] Samuel A. Inverso,[1] Amy Bernard,[5] Prashant Mali,[2] Xavier Rios,[2] John Aach,[2] George M. Church[1,2]†

[1]Wyss Institute, Harvard Medical School, Boston, MA 02115, USA. [2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [3]Department of Stem Cell and Regenerative Biology, Harvard University, Boston, MA 02138, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [5]Allen Institute for Brain Science, Seattle, WA 98103, USA.

* These authors contributed equally to this work.

†Corresponding author. E-mail: jehyuklee@mac.com (J.H.L.); gchurch@genetics.med.harvard.edu (G.M.C.)

**Understanding the spatial organization of gene expression with single-nucleotide resolution requires localizing the sequences of expressed RNA transcripts within a cell in situ. Here, we describe fluorescent in situ RNA sequencing (FISSEQ), in which stably cross-linked cDNA amplicons are sequenced within a biological sample. Using 30-base reads from 8742 genes in situ, we examined RNA expression and localization in human primary fibroblasts with a simulated wound-healing**

# Single-cell sequencing and extensions

- Sequencing of individual cells
  - variation between cells
  - environmental „context“
  - Characterized scaffolds with at least 3x coverage depth
  - Sorted by
    - Coverage
    - Oligonucleotide frequencies
    - Similarity to previously sequenced genomes

- Lee et al. 2016: resolving transcript location within cells
- fluorescent in situ RNA sequencing (FISSEQ)
- amplifies complementary DNA targets by rolling circle amplification,
- in situ cross-linking locks amplicons to produce localized templates for three-dimensional sequencing.
- Tested in fibroblasts to reveal the differences between individual cells during wound repair.

ARTICLES

nature genetics

# Large-scale whole-genome sequencing of the Icelandic population

Daniel F Gudbjartsson[1,2,21], Hannes Helgason[1,2,21], Sigurjon A Gudjonsson[1], Florian Zink[1], Asmundur Oddson[1], Arnaldur Gylfason[1], Soren Besenbacher[3], Gisli Magnusson[1], Bjarni V Halldorsson[1,4], Eirikur Hjartarson[1], Gunnar Th Sigurdsson[1], Simon N Stacey[1], Michael L Frigge[1], Hilma Holm[1,5], Jona Saemundsdottir[1], Hafdis Th Helgadottir[1], Hrefna Johannsdottir[1], Gunnlaugur Sigfusson[6], Gudmundur Thorgeirsson[7,8], Jon Th Sverrisson[9], Solveig Gretarsdottir[1], G Bragi Walters[1], Thorunn Rafnar[1], Bjarni Thjodleifsson[7], Einar S Bjornsson[8,10], Sigurdur Olafsson[8,10], Hildur Thorarinsdottir[10], Thora Steingrimsdottir[8,11], Thora S Gudmundsdottir[11], Asgeir Theodors[10], Jon G Jonasson[8,12,13], Asgeir Sigurdsson[1], Gyda Bjornsdottir[1], Jon J Jonsson[14,15], Olafur Thorarensen[16], Petur Ludvigsson[16], Hakon Gudbjartsson[1,2], Gudmundur I Eyjolfsson[17], Olof Sigurdardottir[18], Isleifur Olafsson[19], David O Arnar[7,8], Olafur Th Magnusson[1], Augustine Kong[1,2], Gisli Masson[1], Unnur Thorsteinsdottir[1,8], Agnar Helgason[1,20], Patrick Sulem[1] & Kari Stefansson[1,8]

**Here we describe the insights gained from sequencing the whole genomes of 2,636 Icelanders to a median depth of 20×. We found 20 million SNPs and 1.5 million insertions-deletions (indels). We describe the density and frequency spectra of sequence variants in relation to their functional annotation, gene position, pathway and conservation score. We demonstrate an excess of homozygosity and rare protein-coding variants in Iceland. We imputed these variants into 104,220 individuals down to a minor allele frequency of 0.1% and found a recessive frameshift mutation in *MYL4* that causes early-onset atrial fibrillation,**

Published in final edited form as:
*Nat Med*. 2014 June ; 20(6): 682–688. doi:10.1038/nm.3559.

formalin-fixed paraffin embedded

# Whole-exome sequencing and clinical interpretation of FFPE tumor samples to guide precision cancer medicine

**Eliezer M. Van Allen**[*,1,2], **Nikhil Wagle**[*,1,2], **Petar Stojanov**[2], **Danielle L. Perrin**[2], **Kristian Cibulskis**[2], **Sara Marlow**[1,2], **Judit Jane-Valbuena**[1,2], **Dennis C. Friedrich**[2], **Gregory Kryukov**[2], **Scott L. Carter**[2], **Aaron McKenna**[2,3], **Andrey Sivachenko**[2], **Mara Rosenberg**[2], **Adam Kiezun**[2], **Douglas Voet**[2], **Michael Lawrence**[2], **Lee T. Lichtenstein**[2], **Jeff G. Gentry**[2], **Franklin W. Huang**[1,2], **Jennifer Fostel**[2], **Deborah Farlow**[2], **David Barbie**[1], **Leena Gandhi**[1], **Eric S. Lander**[2], **Stacy W. Gray**[1], **Steven Joffe**[1,4], **Pasi Janne**[1], **Judy Garber**[1], **Laura MacConaill**[1,5], **Neal Lindeman**[1,5], **Barrett Rollins**[1], **Philip Kantoff**[1], **Sheila A. Fisher**[2], **Stacey Gabriel**[**,2], **Gad Getz**[**,#,2,6], and **Levi A. Garraway**[**,#,1,2]

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, 450 Brookline Avenue, Boston, Massachusetts 02115, USA

[2]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA

[3]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

[4]Children's Hospital Boston, Boston, MA 02115

[5]Brigham and Women's Hospital, Boston, MA 02115

[6]Massachusetts General Hospital Cancer Center and Department of Pathology, Boston, MA 02114

# THANK YOU!

Please fill in the anonymous course evaluation forms in USOS.