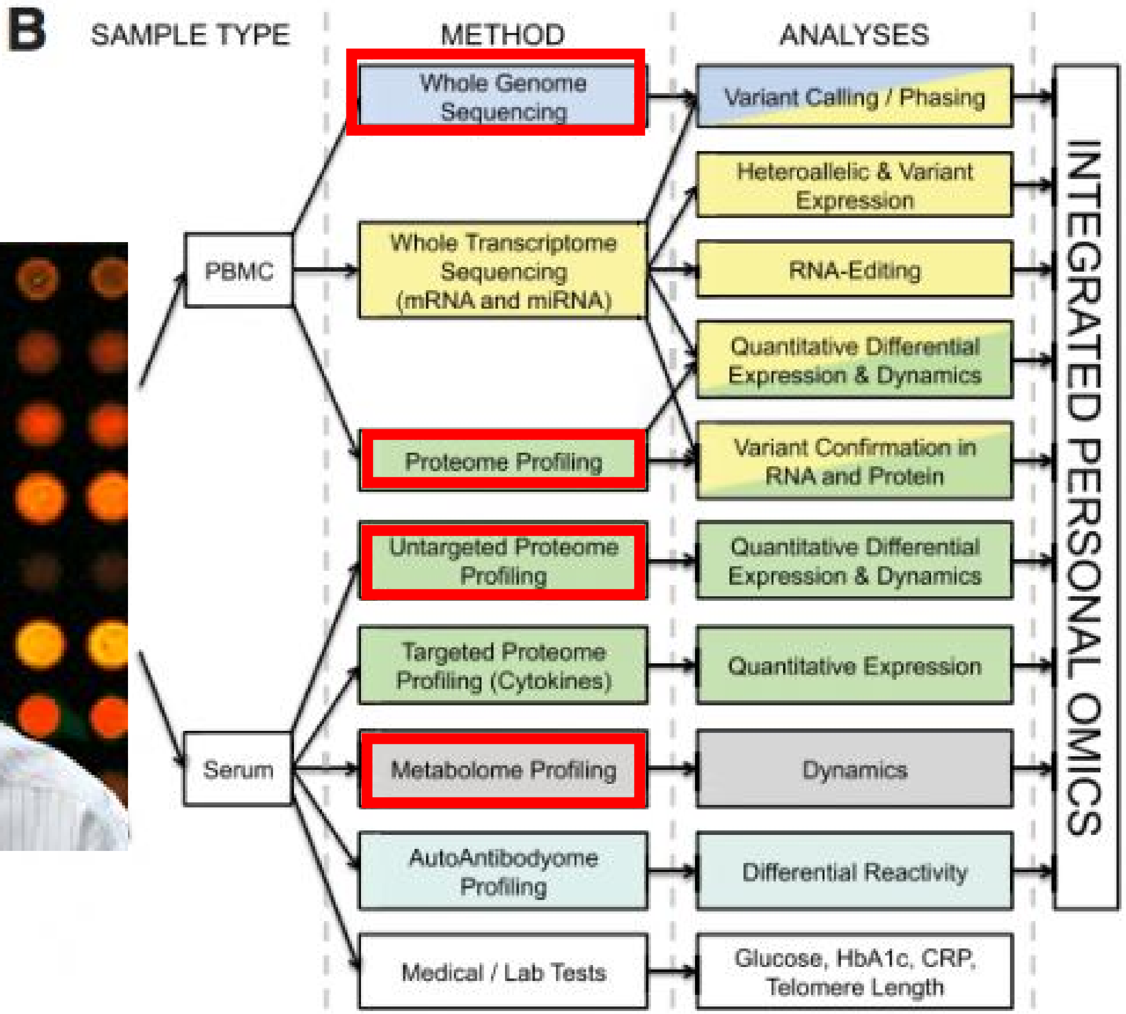
De Novo Genome Assembly from Single Cells

Pavel Pevzner

Department of Computer Science and Engineering
University of California at San Diego

Algorithmic Biology Laboratory
Saint Petersburg Academic University

Michael Snyder Reversed his Own Diabetes by Conducting The Most Extensive Medical Diagnostics Ever (Cell, February 2012)



B

SAMPLE TYPE

METHOD

ANALYSES

PBMC

Serum

Whole Genome Sequencing

Whole Transcriptome Sequencing (mRNA and miRNA)

Proteome Profiling

Untargeted Proteome Profiling

Targeted Proteome Profiling (Cytokines)

Metabolome Profiling

AutoAntibodyome Profiling

Medical / Lab Tests

Variant Calling / Phasing

Heteroallelic & Variant Expression

RNA-Editing

Quantitative Differential Expression & Dynamics

Variant Confirmation in RNA and Protein

Quantitative Differential Expression & Dynamics

Quantitative Expression

Dynamics

Differential Reactivity

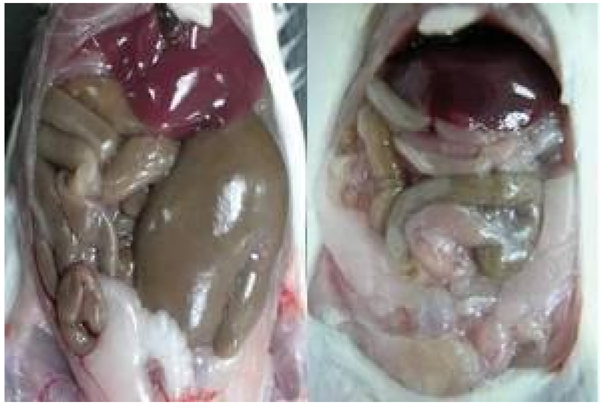
Glucose, HbA1c, CRP, Telomere Length

INTEGRATED PERSONAL OMICS

6000 proteins and 1000 metabolites are measured every month!

What did Michael Snyder Miss?

…Unexpectedly, the cecum in germ-free mice swelled up to several times its normal size and the mice died. **Mice without germs don’t develop normal intestines. ..**



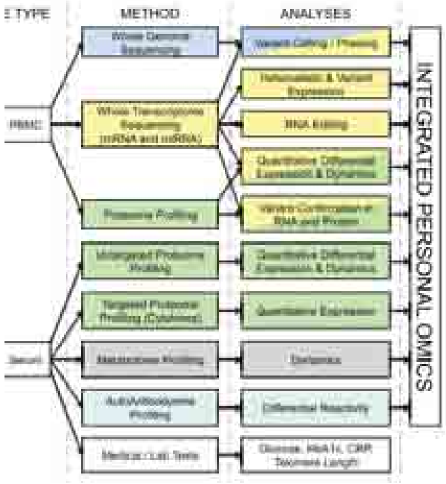
The total size of bacterial genomes from Human Microbiome vastly exceeds the size of human genome.

The number of bacterial cells in our body exceeds the number of human cells by an order of magnitude.

Most human microbes represent **dark matter of life**, ie., their DNA cannot be sequenced with standard DNA sequencing technologies

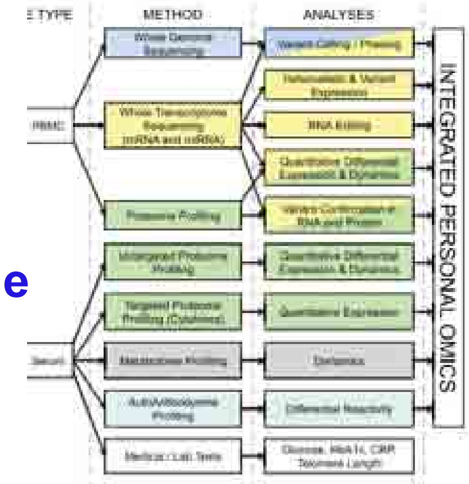
Executive Medical Diagnostics in 2013?

Human genome


E TYPE
METHOD
ANALYSES
PBMC
Serum
Whole Genome Sequencing
Whole Transcriptome Sequencing (mRNA and miRNA)
Proteome Profiling
Untargeted Proteome Profiling
Targeted Proteome Profiling (Cytokines)
Metabolome Profiling
Autoantibodyome Profiling
Medical / Lab Tests
Variant Calling / Phasing
Heteroallelic & Variant Expression
RNA Editing
Quantitative Differential Expression & Dynamics
Variant Confirmation in RNA and Protein
Quantitative Differential Expression & Dynamics
Quantitative Expression
Dynamics
Differential Reactivity
Glucose, HbA1c, CRP, Telomere Length
INTEGRATED PERSONAL OMICS

10^4 human proteins

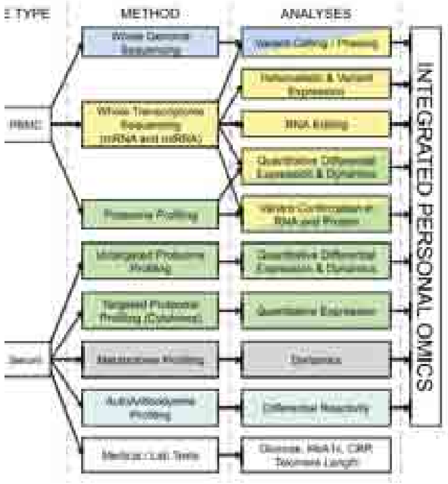
Human microbiome


E TYPE
METHOD
ANALYSES
PBMC
Serum
Whole Genome Sequencing
Whole Transcriptome Sequencing (mRNA and miRNA)
Proteome Profiling
Untargeted Proteome Profiling
Targeted Proteome Profiling (Cytokines)
Metabolome Profiling
Autoantibodyome Profiling
Medical / Lab Tests
Variant Calling / Phasing
Heteroallelic & Variant Expression
RNA Editing
Quantitative Differential Expression & Dynamics
Variant Confirmation in RNA and Protein
Quantitative Differential Expression & Dynamics
Quantitative Expression
Dynamics
Differential Reactivity
Glucose, HbA1c, CRP, Telomere Length
INTEGRATED PERSONAL OMICS

10^6 bacterial proteins

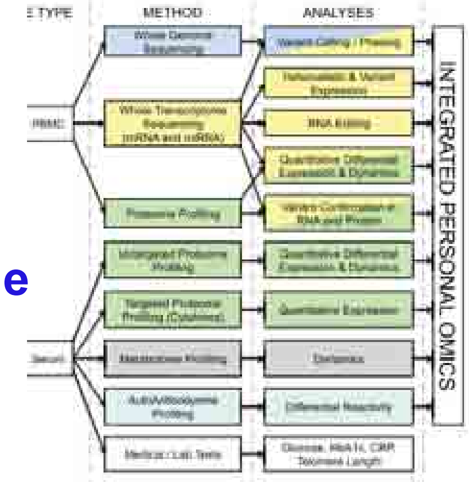
What else did he miss?

Human genome


E TYPE
METHOD
ANALYSES
PBMC
Serum
Whole Genome Sequencing
Whole Transcriptome Sequencing (mRNA and miRNA)
Proteome Profiling
Untargeted Proteome Profiling
Targeted Proteome Profiling (Cytokines)
Metabolome Profiling
AutoAntibodyome Profiling
Medical / Lab Tests
Variant Calling / Phasing
Heteroallelic & Variant Expression
RNA Editing
Quantitative Differential Expression & Dynamics
Variant Confirmation in RNA and Protein
Quantitative Differential Expression & Dynamics
Quantitative Expression
Dietomics
Differential Reactivity
Glucose, HbA1c, CRP, Telomere Length
INTEGRATED PERSONAL OMICS

10^4 human proteins

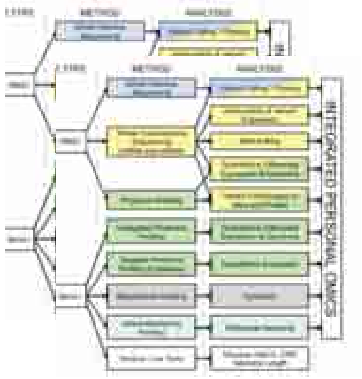
Human microbiome


E TYPE
METHOD
ANALYSES
PBMC
Serum
Whole Genome Sequencing
Whole Transcriptome Sequencing (mRNA and miRNA)
Proteome Profiling
Untargeted Proteome Profiling
Targeted Proteome Profiling (Cytokines)
Metabolome Profiling
AutoAntibodyome Profiling
Medical / Lab Tests
Variant Calling / Phasing
Heteroallelic & Variant Expression
RNA Editing
Quantitative Differential Expression & Dynamics
Variant Confirmation in RNA and Protein
Quantitative Differential Expression & Dynamics
Quantitative Expression
Dietomics
Differential Reactivity
Glucose, HbA1c, CRP, Telomere Length
INTEGRATED PERSONAL OMICS

10^6 bacterial proteins

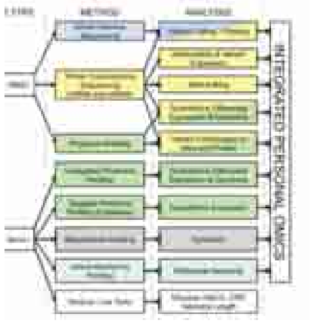
Sequencing of Individual Tumor Cells for Early Cancer Diagnostics/Monitoring

Human genome



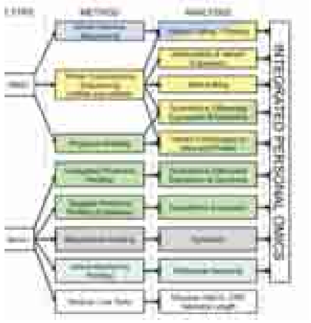
10^4 human proteins

Tumor genome



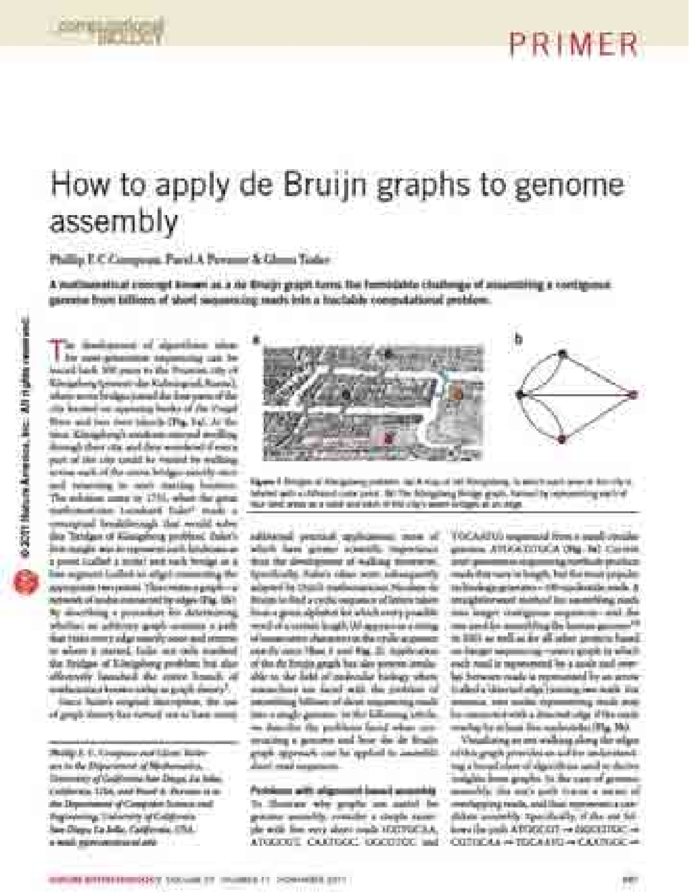
Profiling INDIVIDUAL tumor cells

Human microbiome



10^6 bacterial proteins

Nicolaas de Bruijn



computational BIOLOGY

PRIMER

How to apply de Bruijn graphs to genome assembly

Phillip E C Compeau, Pavel A Pevzner & Glenn Tesler

A mathematical concept known as a de Bruijn graph turns the formidable challenge of assembling a contiguous genome from billions of short sequencing reads into a tractable computational problem.

© 2011 Nature America, Inc. All rights reserved.

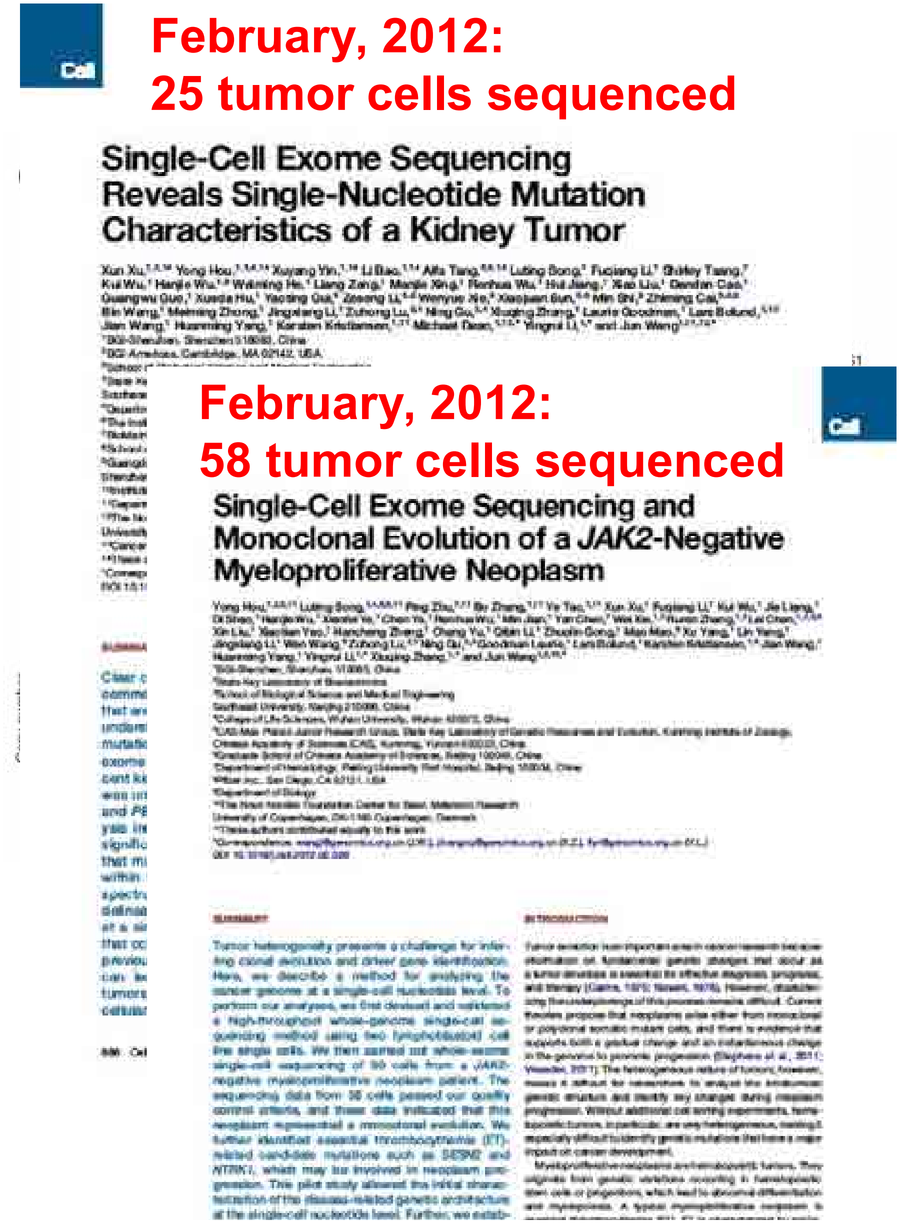
[illegible]

NATURE BIOTECHNOLOGY VOLUME 29 NUMBER 11 NOVEMBER 2011

July 9, 1918 - February 17, 2012

Breakthroughs in Single Cell Genomics in 2011

- **Sequencing phased human chromosomes** (Yang et al., PNAS 2011)
- **Tracing tumor evolution** (Navin et al., Nature 2011)
- **Studying tumor heterogeneity** (Dalerba et al., Nature Biotech. 2011)
- **Characterizing single cell transcriptome** (Islam et al., Genome Res. 2011)
- **Genome-wide haplotyping** (Fan et al., Nature Biotech. 2011)
- **Analyzing uncultivated single cell organisms and revealing the "gray matter of life"** **(**Yoon et al., Science,
Yousseff et al., AIM 2011,
Chitsaz et al., Nature Biotech, 2011)



February, 2012:
25 tumor cells sequenced

Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor

Xun Xu, Yong Hou, Xuyang Yin, Li Bao, Aifa Tang, Luting Song, Fuqiang Li, Shirley Tsang, Kui Wu, Hanjie Wu, Weiming He, Liang Zeng, Manjie Xing, Renhua Wu, Hui Jiang, Xiao Liu, Dandan Cao, Guangwu Guo, Xueda Hu, Yaoting Gui, Zesong Li, Wenyue Xie, Xiaojuan Sun, Min Shi, Zhiming Cai, Bin Wang, Meiming Zhong, Jingxiang Li, Zuhong Lu, Ning Gu, Xiuqing Zhang, Laurie Goodman, Lars Bolund, Jian Wang, Huanming Yang, Karsten Kristiansen, Michael Dean, Yingrui Li, and Jun Wang

BGI-Shenzhen, Shenzhen 518083, China

BGI-Americas, Cambridge, MA 02142, USA

February, 2012:
58 tumor cells sequenced

Single-Cell Exome Sequencing and Monoclonal Evolution of a *JAK2*-Negative Myeloproliferative Neoplasm

SUMMARY

INTRODUCTION

Bacterial Single Cell Genomics

- **Sequencing phased human chromosomes**
  (Yang et al., PNAS 2011)

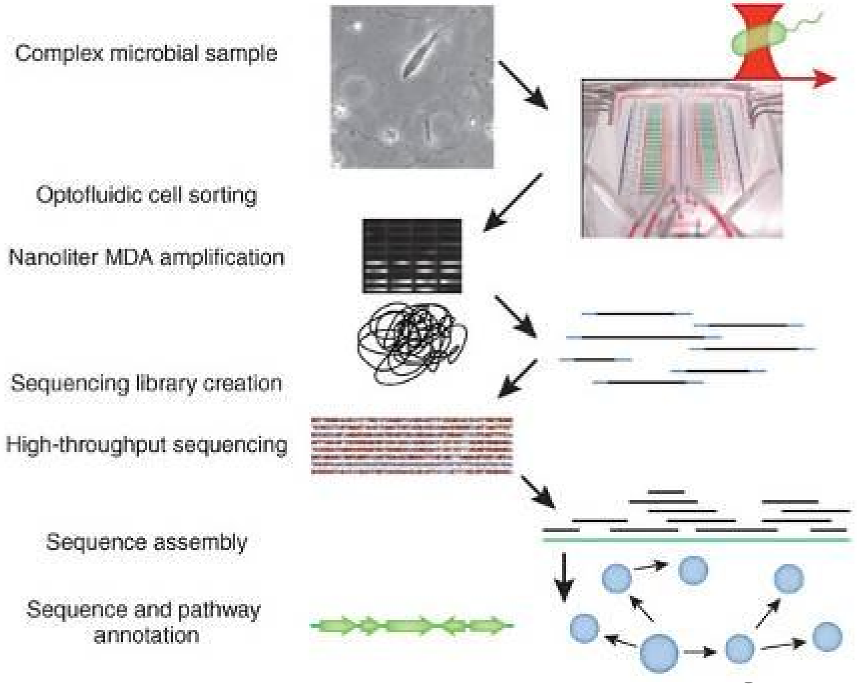
- **Tracing tumor evolution**
  (Navin et al., Nature 2011)

- **Studying tumor heterogeneity**
  (Dalerba et al., Nature Biotech. 2011)

- **Characterizing single cell transcriptome**
  (Islam et al., Genome Res. 2011)

- **Genome-wide haplotyping**
  (Fan et al., Nature Biotech. 2011)

- **Analyzing uncultivated single cell organisms and revealing the "gray matter of life"**
  (Yoon et al., Science,
  Yousseff et al., AIM 2011,
  Chitsaz et al., Nature Biotech, 2011)



Complex microbial sample

Optofluidic cell sorting

Nanoliter MDA amplification

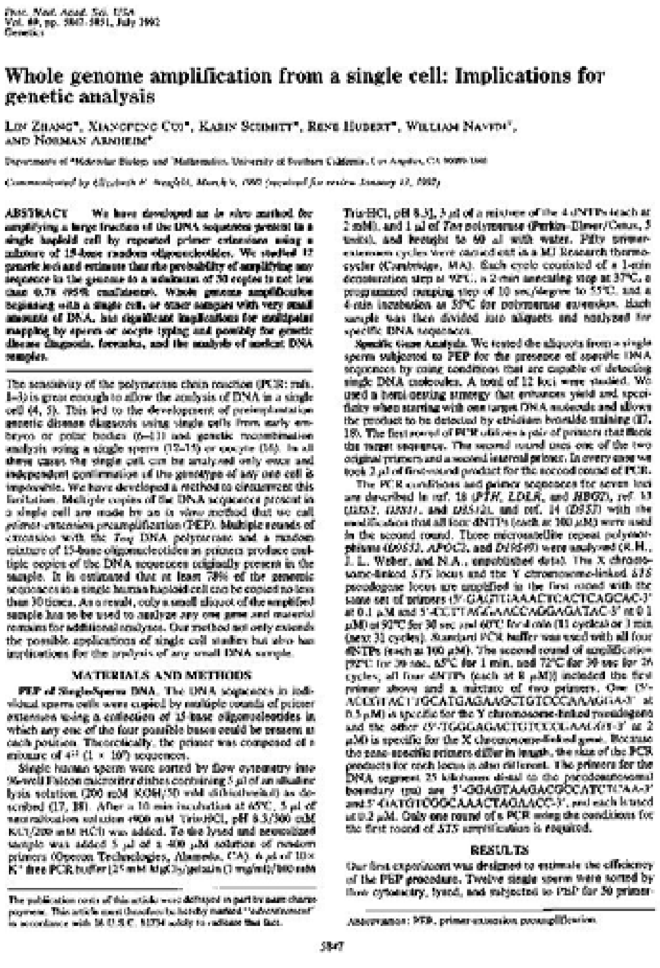
Sequencing library creation

High-throughput sequencing

Sequence assembly

Sequence and pathway annotation

When Did Single Cell Sequencing Started?



Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 5847–5851, July 1992
Genetics

Whole genome amplification from a single cell: Implications for genetic analysis

LIN ZHANG*, XIANGFENG CUI*, KARIN SCHMITT*, RENE HUBERT*, WILLIAM NAVIDI†, AND NORMAN ARNHEIM*

Departments of *Molecular Biology and †Mathematics, University of Southern California, Los Angeles, CA 90089-1340

Communicated by Elizabeth F. Neufeld, March 9, 1992 (received for review January 17, 1992)

ABSTRACT We have developed an *in vitro* method for amplifying a large fraction of the DNA sequences present in a single haploid cell by repeated primer extensions using a mixture of 15-base random oligonucleotides. We studied 12 genetic loci and estimate that the probability of amplifying any sequence in the genome to a minimum of 30 copies is not less than 0.78 (95% confidence). Whole genome amplification beginning with a single cell, or other samples with very small amounts of DNA, has significant implications for multipoint mapping by sperm or oocyte typing and possibly for genetic disease diagnosis, forensics, and the analysis of ancient DNA samples.

The sensitivity of the polymerase chain reaction (PCR; refs. 1–3) is great enough to allow the analysis of DNA in a single cell (4, 5). This led to the development of preimplantation genetic disease diagnosis using single cells from early embryos or polar bodies (6–11) and genetic recombination analysis using a single sperm (12–15) or oocyte (16). In all these cases the single cell can be analyzed only once and independent confirmation of the genotype of any one cell is impossible. We have developed a method to circumvent this limitation. Multiple copies of the DNA sequences present in a single cell are made by an *in vitro* method that we call *primer-extension preamplification* (PEP). Multiple rounds of extension with the *Taq* DNA polymerase and a random mixture of 15-base oligonucleotides as primers produce multiple copies of the DNA sequences originally present in the sample. It is estimated that at least 78% of the genomic sequences in a single human haploid cell can be copied no less than 30 times. As a result, only a small aliquot of the amplified sample has to be used to analyze any one gene and material remains for additional analyses. Our method not only extends the possible applications of single cell studies but also has implications for the analysis of any small DNA sample.

MATERIALS AND METHODS

PEP of Single-Sperm DNA. The DNA sequences in individual sperm cells were copied by multiple rounds of primer extension using a collection of 15-base oligonucleotides in which any one of the four possible bases could be present at each position. Theoretically, the primer was composed of a mixture of 4^{15} (1 × 10^9) sequences.

Single human sperm were sorted by flow cytometry into 96-well Falcon microtiter dishes containing 5 μl of an alkaline lysis solution (200 mM KOH/50 mM dithiothreitol) as described (17, 18). After a 10-min incubation at 65°C, 5 μl of neutralization solution (900 mM Tris·HCl, pH 8.3/300 mM KCl/200 mM HCl) was added. To the lysed and neutralized sample was added 5 μl of a 400 μM solution of random primers (Operon Technologies, Alameda, CA). 6 μl of 10× K⁺ free PCR buffer (25 mM MgCl₂/gelatin (1 mg/ml)/100 mM Tris·HCl, pH 8.3), 3 μl of a mixture of the 4 dNTPs (each at 2 mM), and 1 μl of *Taq* polymerase (Perkin–Elmer/Cetus, 5 units), and brought to 60 μl with water. Fifty primer-extension cycles were carried out in a MJ Research thermocycler (Cambridge, MA). Each cycle consisted of a 1-min denaturation step at 92°C, a 2-min annealing step at 37°C, a programmed ramping step of 10 sec/degree to 55°C, and a 4-min incubation at 55°C for polymerase extension. Each sample was then divided into aliquots and analyzed for specific DNA sequences.

Specific Gene Analysis. We tested the aliquots from a single sperm subjected to PEP for the presence of specific DNA sequences by using conditions that are capable of detecting single DNA molecules. A total of 12 loci were studied. We used a hemi-nesting strategy that enhances yield and specificity when starting with one target DNA molecule and allows the product to be detected by ethidium bromide staining (17, 18). The first round of PCR utilizes a pair of primers that flank the target sequence. The second round uses one of the two original primers and a second internal primer. In every case we took 2 μl of first-round product for the second round of PCR.

The PCR conditions and primer sequences for seven loci are described in ref. 18 (*PTH*, *LDLR*, and *HBG2*), ref. 13 (*D2S2*, *D3S11*, and *D4S10*), and ref. 14 (*D5S3*) with the modification that all four dNTPs (each at 100 μM) were used in the second round. Three microsatellite repeat polymorphisms (*D9S53*, *APOC2*, and *D19S49*) were analyzed (R.H., J. L. Weber, and N.A., unpublished data). The X chromosome-linked *STS* locus and the Y chromosome-linked *STS* pseudogene locus are amplified in the first round with the same set of primers (5′-GAGTTGAAACTCACTCAGCAC-3′ at 0.1 μM and 5′-CCTTAGGAAACCAGGAGATAC-3′ at 0.1 μM) at 92°C for 30 sec and 60°C for 4 min (11 cycles) or 1 min (next 31 cycles). Standard PCR buffer was used with all four dNTPs (each at 100 μM). The second round of amplification (92°C for 30 sec, 65°C for 1 min, and 72°C for 30 sec for 26 cycles, all four dNTPs (each at 8 μM)) included the first primer above and a mixture of two primers. One (5′-ACCCTACTTGCATGAGAAGCTGTCCCAAAGGA-3′ at 0.5 μM) is specific for the Y chromosome-linked pseudogene and the other (5′-TGGGAGACTGTCCCAAAGGA-3′ at 2 μM) is specific for the X chromosome-linked gene. Because the gene-specific primers differ in length, the size of the PCR products for each locus is also different. The primers for the DNA segment 25 kilobases distal to the pseudoautosomal boundary (pa) are 5′-GGAGTAAGACGCCATCTCAA-3′ and 5′-GATGTCGGCAAACTAGAACC-3′, and each is used at 0.2 μM. Only one round of a PCR using the conditions for the first round of *STS* amplification is required.

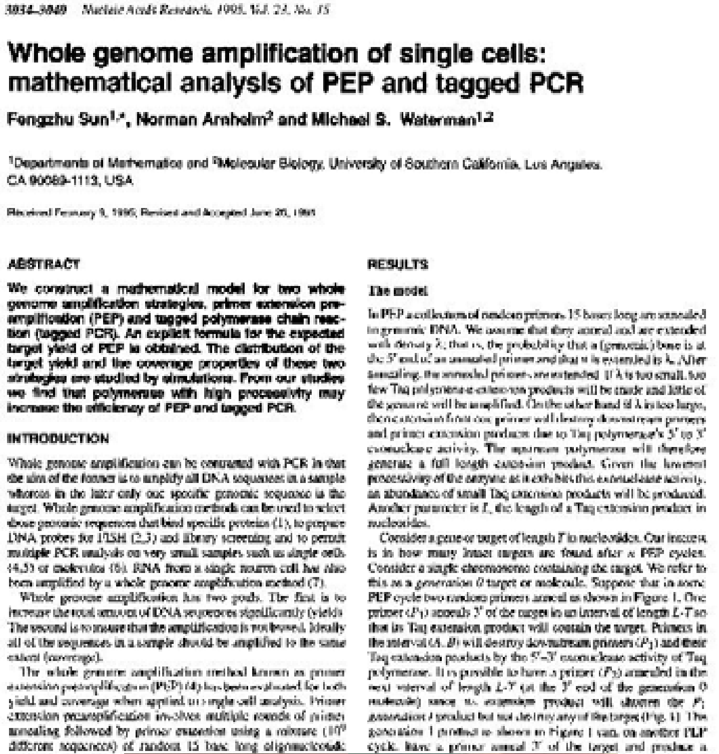
RESULTS

Our first experiment was designed to estimate the efficiency of the PEP procedure. Twelve single sperm were sorted by flow cytometry, lysed, and subjected to PEP for 50 primer-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: PEP, primer-extension preamplification.

5847



3034–3040 *Nucleic Acids Research, 1995, Vol. 23, No. 15*

Whole genome amplification of single cells: mathematical analysis of PEP and tagged PCR

Fengzhu Sun[1,*], Norman Arnheim[2] and Michael S. Waterman[1,2]

[1]Departments of Mathematics and [2]Molecular Biology, University of Southern California, Los Angeles, CA 90089-1113, USA

Received February 9, 1995; Revised and Accepted June 26, 1995

ABSTRACT

We construct a mathematical model for two whole genome amplification strategies, primer extension preamplification (PEP) and tagged polymerase chain reaction (tagged PCR). An explicit formula for the expected target yield of PEP is obtained. The distribution of the target yield and the coverage properties of these two strategies are studied by simulations. From our studies we find that polymerase with high processivity may increase the efficiency of PEP and tagged PCR.

INTRODUCTION

Whole genome amplification can be contrasted with PCR in that the aim of the former is to amplify all DNA sequences in a sample whereas in the latter only one specific genomic sequence is the target. Whole genome amplification methods can be used to select those genomic sequences that bind specific proteins (1), to prepare DNA probes for FISH (2,3) and library screening and to permit multiple PCR analysis on very small samples such as single cells (4,5) or molecules (6). RNA from a single neuron cell has also been amplified by a whole genome amplification method (7).

Whole genome amplification has two goals. The first is to increase the total amount of DNA sequences significantly (yield). The second is to insure that the amplification is not biased. Ideally all of the sequences in a sample should be amplified to the same extent (coverage).

The whole genome amplification method known as primer extension preamplification (PEP) (4) has been evaluated for both yield and coverage when applied to single cell analysis. Primer extension preamplification involves multiple rounds of primer annealing followed by primer extension using a mixture (10^9 different sequences) of random 15 base long oligonucleotide primers. Starting with [illegible]
cycles produce an esti[illegible]
of the genome is repr[illegible]

Given the number [illegible]
inefficient compared [illegible]
ical model of the orig[illegible]
tion (8) to try and de[illegible]
increase yields with[illegible]
applicable to other wh[illegible]
partially degenerate p[illegible]

*To whom correspondence [illegible]
Atlanta, GA 30322, USA [illegible]

RESULTS

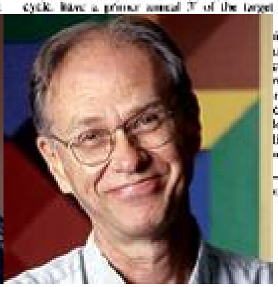
The model

In PEP a collection of random primers 15 bases long are annealed to genomic DNA. We assume that they anneal and are extended with density λ, that is, the probability that a (genomic) base is at the 3′ end of an annealed primer and that it is extended is λ. After annealing, the annealed primers are extended. If λ is too small, too few Taq polymerase extension products will be made and little of the genome will be amplified. On the other hand if λ is too large, the extension from one primer will destroy downstream primers and primer extension products due to Taq polymerase's 5′ to 3′ exonuclease activity. The upstream polymerase will therefore generate a full length extension product. Given the low processivity of the enzyme as it exhibits this exonuclease activity, an abundance of small Taq extension products will be produced. Another parameter is *L*, the length of a Taq extension product in nucleotides.

Consider a gene or target of length *T* in nucleotides. Our interest is in how many intact targets are found after *n* PEP cycles. Consider a single chromosome containing the target. We refer to this as a *generation 0* target or molecule. Suppose that in some PEP cycle two random primers anneal as shown in Figure 1. One primer (P_1) anneals 3′ of the target in an interval of length *L*-*T* so that its Taq extension product will contain the target. Primers in the interval (*A*, *B*) will destroy downstream primers (P_2) and their Taq extension products by the 5′–3′ exonuclease activity of Taq polymerase. It is possible to have a primer (P_3) annealed in the next interval of length *L*-*T* (at the 3′ end of the generation 0 molecule) since its extension product will shorten the P_1 *generation 1* product but not destroy any of the target (Fig. 1). The generation 1 product as shown in Figure 1 can, on another PEP cycle, have a primer anneal 3′ of the target and produce a

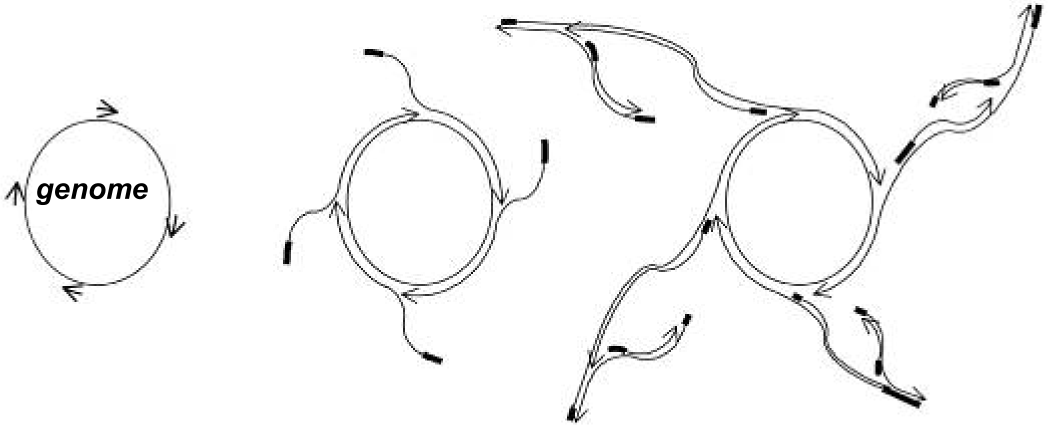
[illegible]ing process. If *p*
[illegible]ut it becomes a
[illegible]cles the expected
[illegible]nential growth of
[illegible]reached. But the
[illegible]c as in PCR. The
[illegible]ld of PEP λ is
[illegible]lity of producing
[illegible]xtension product

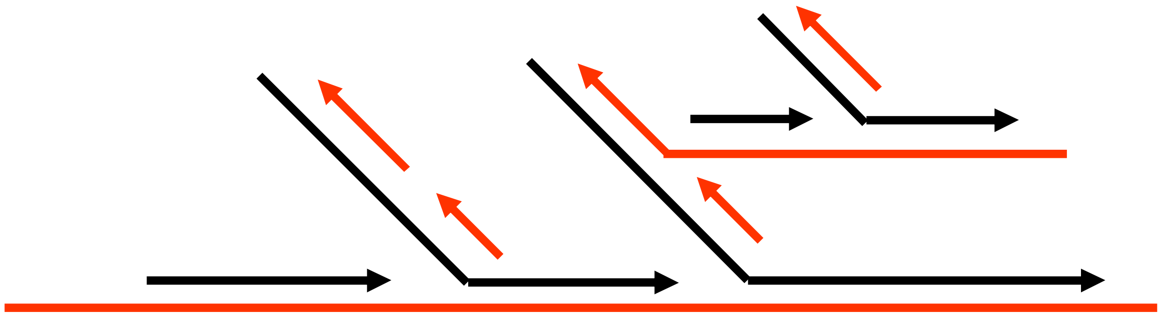
[illegible]mory University,



From Cloning to Single Cell Amplification

Multiple Displacement Amplification (MDA)



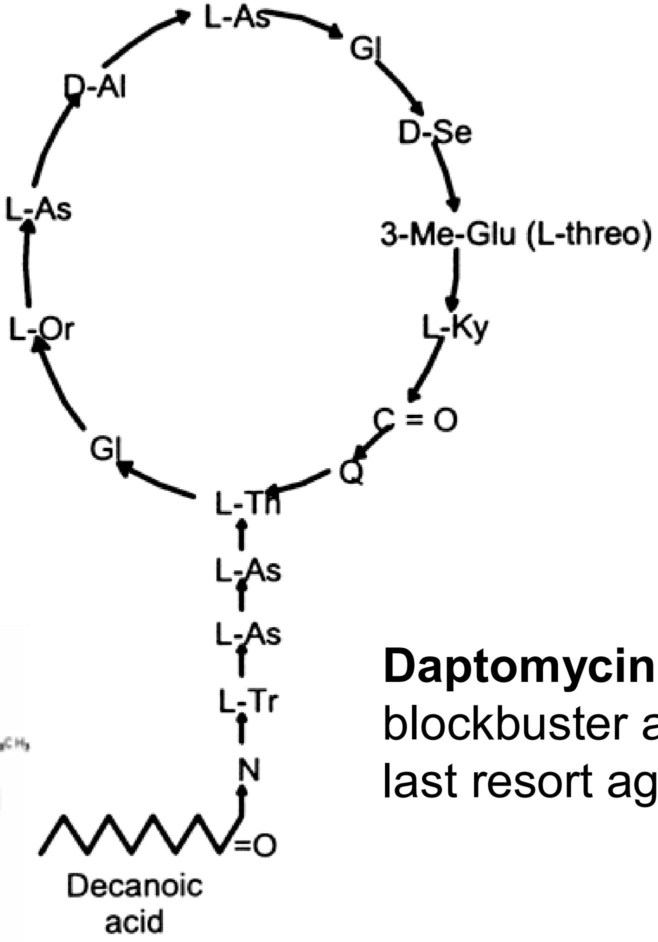


MDA uses random hexamer primers and phi29 DNA polymerase with exceptional ability to displace strands.

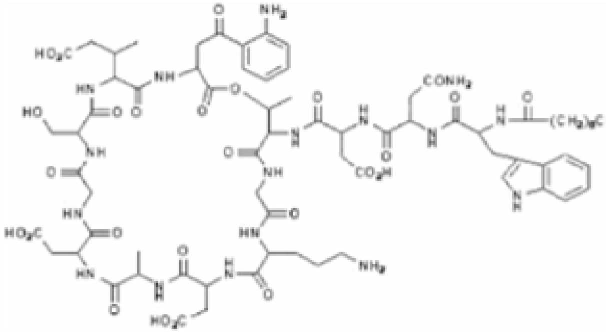
Dean, Nelson, Giesler, Lasken. *Genome Res,* 2001
Dean, Hosono, Fang, ..., Lasken. *PNAS, 2002*

Cycloproteins

Over 50% of antibacterial and anticancer drugs are derived from natural products (many of them are cyclic and branch-cyclic peptides)



This deep tissue infection was positive for MRSA after a partial first ray resection. Daptomycin, a cyclic lipopeptide, has demonstrated activity against MRSA.



Daptomycin:
blockbuster antibiotic of last resort against MRSA

De Novo Sequencing of Cycloproteins is the Only Option Even When Genome is KNOWN

DNA makes RNA makes PROTEIN (central dogma)

transcription

translation

De Novo Sequencing of Cycloproteins is the Only Option Even When Genome is KNOWN

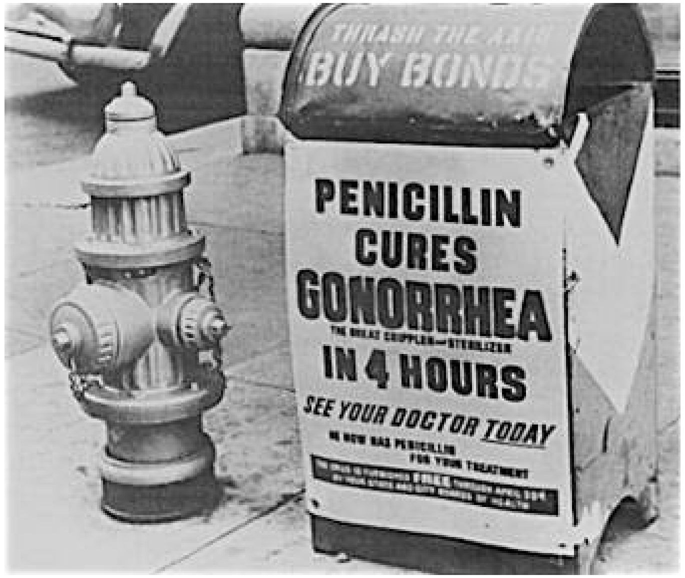
Without any RNA!

DNA makes RNA makes PROTEIN makes … PEPTIDE

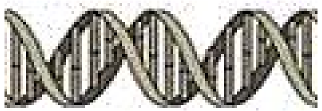
transcription → translation → **non-ribosomal peptide synthesis**

Non-Ribosomal Peptides (NRPs) are excellent compounds for the development of pharmaceutical agents (NRP and other natural products represent 9 out of top 20 bestselling drugs):

• Antibiotics (penicillin, vancomicine, etc.),
• Immunosuppressors (cyclosporin),
• Antiviral agents (luzopeptin A),
• Antitumor agents (bleomycin),
•……………..

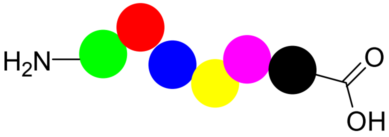


Ribosomal peptide synthesis



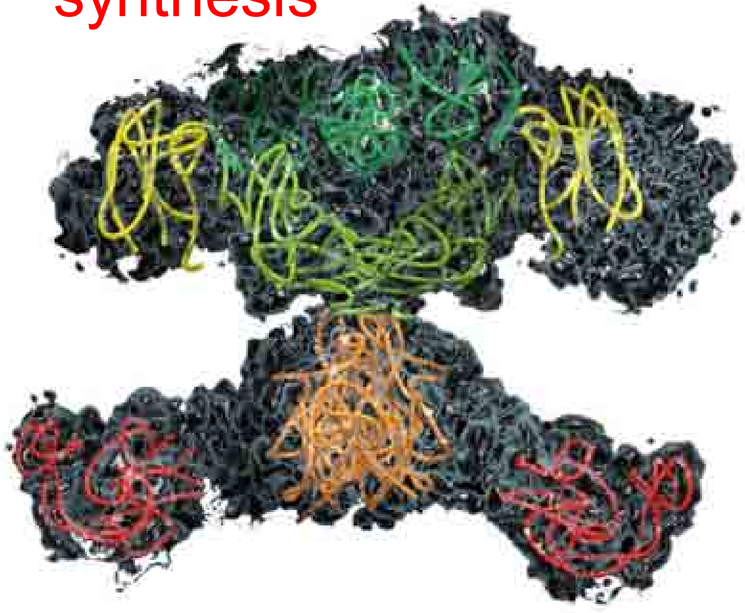
DNA

RNA

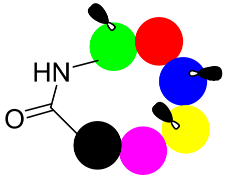


Protein/peptide

Non- Ribosomal peptide synthesis



Modular protein

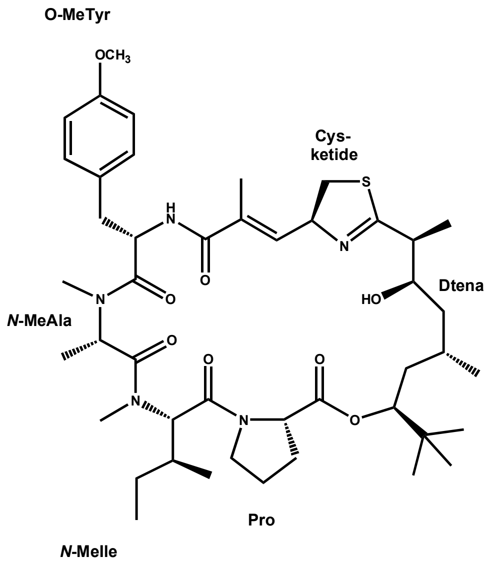


Complex peptides

From Seaside to Bedside

Center for
Marine Biotechnology
and Biomedicine

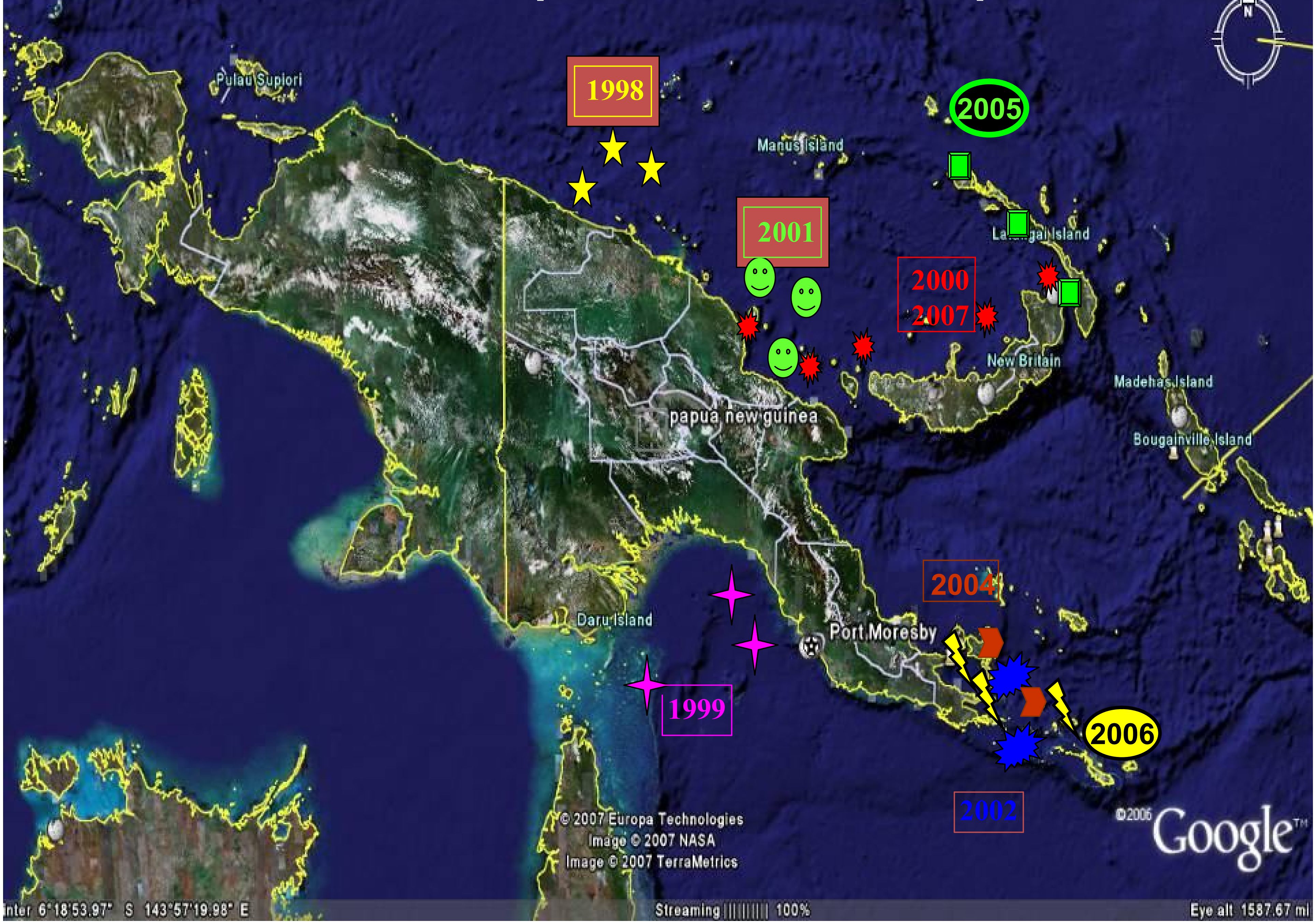
Our colleagues at the Scripps Institute of Oceanography at UCSD found a cyclic peptide apratoxin, a very high priority anticancer toxin. **Novel and still unknown mechanism of action**



They wanted to sequence a 60Kb long aprotoxin gene (that codes for a protein producing apratoxin).

Professor Bill Gerwick at work hunting for new NRPs in New Guinea

Bill Gerwick Papua New Guinea Expeditions



Marine Cycloproteins

Only 1 in 15,000 evaluated compounds becomes an approved drug entity

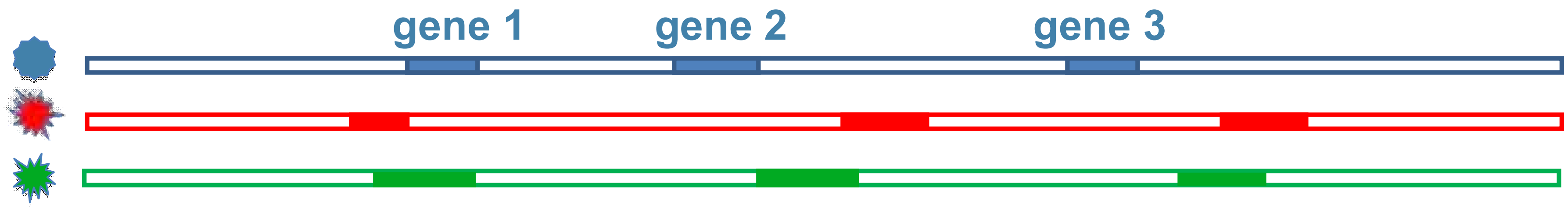
The success record of marine natural products is an order of magnitude better making them one of the most promising drug leads

Single cell sequencing is usually the only way to go for marine bacteria (Grindberg et al., 2011)

From Metagenomics to Single Cell Sequencing

- The lion’s share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies.

- Until recently, **metagenomics** was the only option for studies of microbial communities. However, metagenomics provides information about only a **few genes** (across many species).



From Metagenomics to Single Cell Sequencing

- The lion’s share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies.

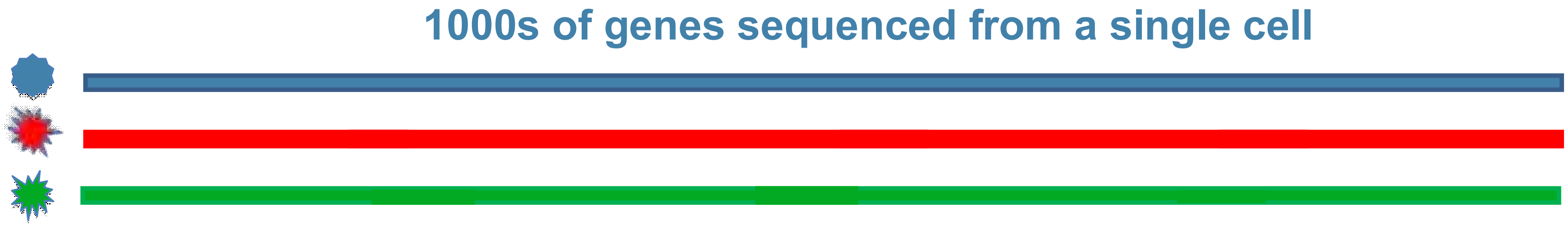
- Until recently, **metagenomics** was the only option for studies of microbial communities. However, metagenomics provides information about only a **few genes** (across many species).

From Metagenomics to Single Cell Sequencing

- The lion’s share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies.

- Until recently, **metagenomics** was the only option for studies of microbial communities. However, metagenomics provides information about only a **few genes** (across many species).

- **Single Cell Bacterial Genomics:** Complementing **gene-centri**c metagenomics data with **whole-genom**e assembly of uncultivated organisms.



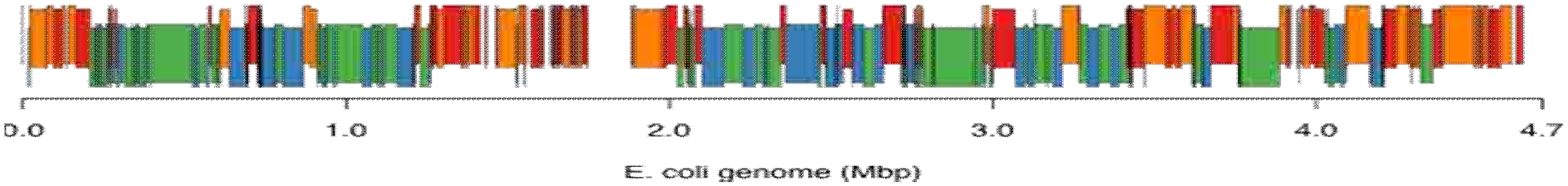
1000s of genes sequenced from a single cell

From Metagenomics to Single Cell Sequencing

- The lion’s share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies.

- Until recently, **metagenomics** was the only option for studies of microbial communities. However, metagenomics provides information about **few genes** (across many species).

- Single Cell Bacterial Genomics: Complementing **gene-centric** metagenomics data with **whole-genome** assembly of uncultivated organisms.



Recently developed single cell assembler SPAdes captures up to 96% of genome and up to 87% of genes from single cell.

In proteomics or antibiotics discovery, capturing a great majority of genes is almost as useful as having a complete assembly.

Introduction to Genome Sequencing
(для школьников и академиков)

What Is Genome Sequencing?

- A genome can be represented as a book written in an alphabet containing only 4 letters, called **nucleotides**: A,T,G, and C.
  - A human genome has roughly 3 billion nucleotides.

```
...CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGA
TCGATCGATCGATTATCTACGATCGATCGATCGATCACTATACGAGCTACTACGTACGTACGATCGCGGGACTATTATCGACTACA
GATAAAACATGCTAGTACAACAGTATACATAGCTGCGGGATACGATTAGCTAATAGCTGACGATATATAGCCGAGCGGCTACGATG
ATGCTAGCTGTACAGCTGATGATCTAGCTATCGATGCGATCGATGCGCGAGTGCGATCGATCACTTCGAGCTAGCTGATCGATCGA
TGCTAGCTAGCTGACTGATCATGGCGTTAGCTAGCTAGCTGATCGTCGATCGTACGTAGCTGATTACGATCGTCCGATCGTGCTAT
GACGTACGAGGCGGCTACGTAGCATGCTAGCTGACTGATGTAGCTAGCTATACGATACTATATATTCGATCGATTTATTACCATGA
CTGACGCGCATCGCTGTACACGTACTAGCTGATCGATGCTAGTCGATCGATCGATCATGTTATATATCGCGGCGCATCGATCGACT
GCTCGATTATCGATACGTCGATCGCTGTATATACGTCTTTATAGCTAGGAGCATAGCGACGCGCTATCGATCGATCGTCTAGTCGA
CTGATCGTACTAGCTGACGCTGACGACTAGCTAGCTATCGACGATCGTAGTGCGATTACTAGCTAGGATCCTACTGTACGTCAGTC
AGTCTGATCGATAGCGAGGAAAGCGAGACTGATCGTTCTCTAGATGTAGCTGATGTGACTACTATACTACTGGCAGCGATCGGGA...
```

- **Genome sequencing** is the process of determining the sequence of nucleotides that make up a genome.

What Is Genome Sequencing?

- Different people have slightly different genomes: all humans share 99.9% of the same genetic code.

- The 0.1% difference accounts for height, eye color, high cholesterol susceptibility, etc.

```
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGA
TCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTAT
CGATCGATCGATCGATTATCTACGATCGATCGATCGATCA
CTATACGAGCTACTACGTACGTACGATCGCGGGACTATTA
TCGACTACAGATAAAACATGCTAGTACAACAGTATACATA
GCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT
```

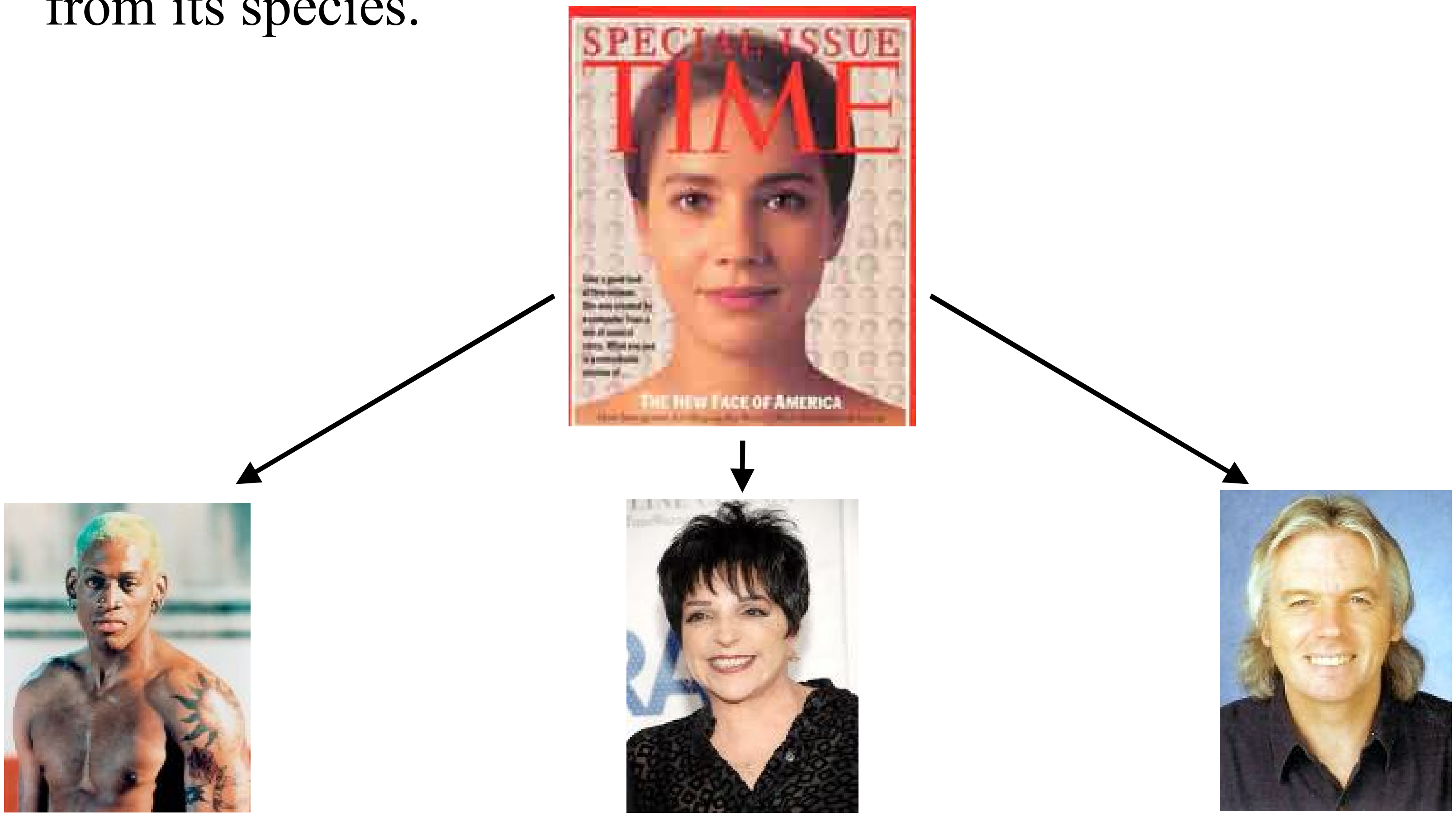
```
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGA
TCAGCTACAACATCGTAGCTACGATGCATTAGCAAGCTAT
CGATCGATCGATCGATTATCTACGATCGATCGATCGATCA
CTATACGAGCTACTACGTACGTACGATCGCGTGACTATTA
TCGACTACAGATGAAACATGCTAGTACAACAGTATACATA
GCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT
```

Species Sequencing vs. Individual Genome Sequencing

- **Species Sequencing**: Determine the “consensus genome” of an entire species.

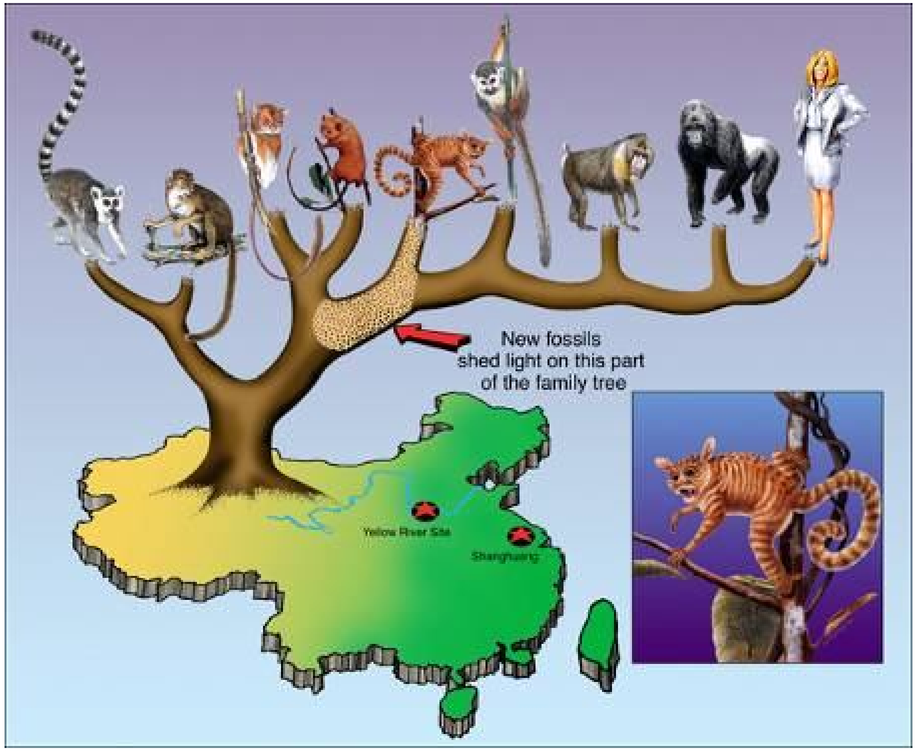
Species Sequencing vs. Individual Genome Sequencing

- **Individual Sequencing**: Determine how an individual differs from its species.



Why Would We Want to Sequence a Genome?

- **Species** genome sequencing:
  - Compare various species (e.g. human and chimpanzee) to understand how their genes function (e.g. which genes are important for brain development).
  - Reveal evolutionary relationships between species.
  - Determine the genetic makeup of our evolutionary ancestors.



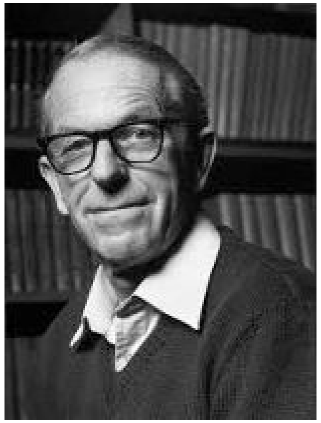
Why Would We Want to Sequence a Genome?

- **Individual** genome sequencing:
  - Unearth the genetic basis of many diseases.
  - Forensics applications.
- **Example**: In 2010, 6-year old Nicholas Volker became the first human being to be saved because of genome sequencing.
  - Doctors could not diagnose his condition, which caused strange infections; he went through nearly 100 surgeries.
  - Genome sequencing revealed a rare mutation in a gene linked to a defect in his immune system.
  - This led doctors to use advanced immunotherapy, which saved the child.

Brief History of Genome Sequencing

- **Late 1970s**: Walter Gilbert and Frederick Sanger develop independent sequencing methods.
- **1980**: They share the Nobel Prize in Chemistry.
- Still, their sequencing methods were too expensive for large genomes: with a $1 per nucleotide cost, it would cost $3 billion to sequence the human genome.

Walter Gilbert

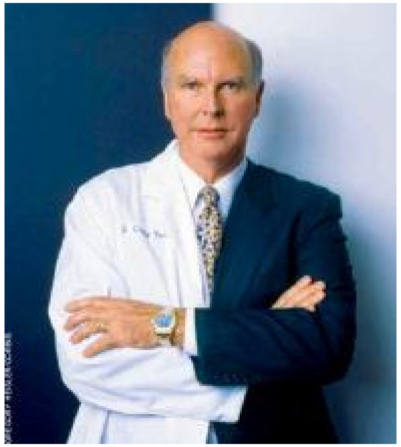


Frederick Sanger

Brief History of Genome Sequencing

- **1990**: The public Human Genome Project, headed by Francis Collins, aims to sequence the human genome.

Francis Collins



Craig Venter

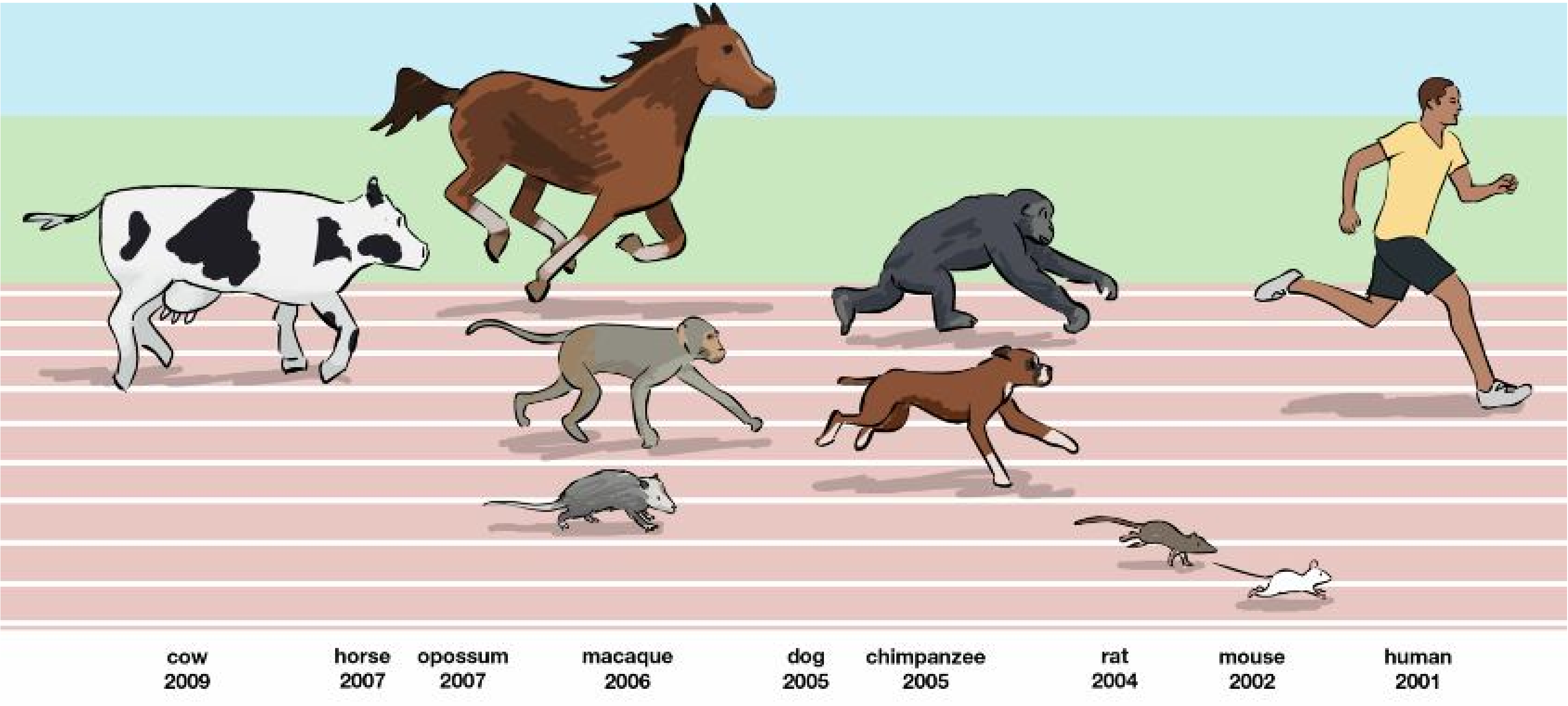
- **1997**: Craig Venter founds Celera Genomics, a private firm, with the same goal.

Brief History of Genome Sequencing

- **2000**: The draft of the human genome is simultaneously completed by the (public) Human Genome Consortium and (private) Celera Genomics.

Brief History of Genome Sequencing

- **2000s**: Many mammalian genomes are sequenced.



The Arrival of Personal Genomics

- **2000s**: Many companies launch projects aimed at reducing sequencing costs by orders of magnitude.
- **2010-2011**: The market for sequencing machines takes off.
  - Illumina reduces the cost of sequencing an individual human genome from $3 billion to $10,000.
  - Complete Genomics builds a genomic factory in Silicon Valley that sequences hundreds of genomes per month.
  - Beijing Genome Institute orders over a hundred of sequencing machines, becoming the world's largest sequencing center.
  - 23andMe offers partial genome sequencing for $499.
  - Many universities introduce new courses in which students study their own genomes.

The Future of Genome Sequencing

- **2012**: Genome sequencing continues to bloom.
  - The $1,000 human genome is expected to arrive later this year.
  - Leading medical centers in the US start the personalized medicine initiatives
  - Hopefully, sequencing an individual genome will soon become as routine as an X-ray.

What Makes Genome Sequencing So Difficult?

- When we read a book, we can read the entire book one letter at a time from the beginning to the end.

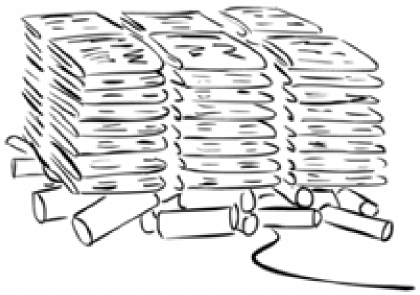
- However, modern sequencing machines cannot read an entire genome one nucleotide at a time from beginning to end. They can only shred the genome and read the short pieces.
  - Thus, we can identify very short fragments of DNA (~100 nucleotides long), called **reads**.
  - But we have no idea which genomic positions these reads come from!
  - **We must figure out how to put the reads back together to assemble a genome.**

The Newspaper Problem

stack of NY Times, June 27, 2000

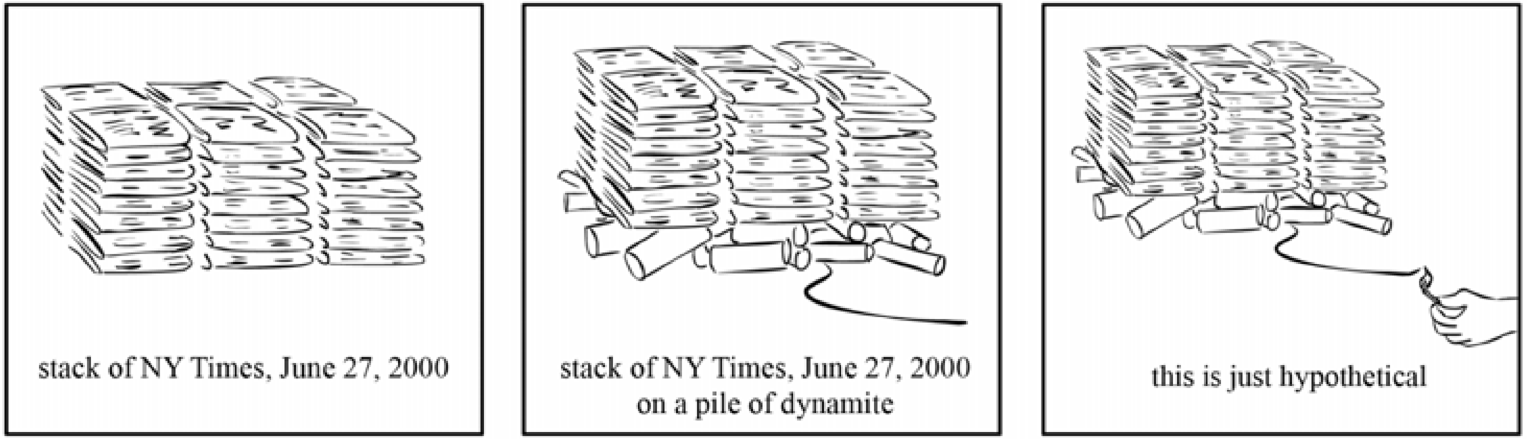
The Newspaper Problem

stack of NY Times, June 27, 2000

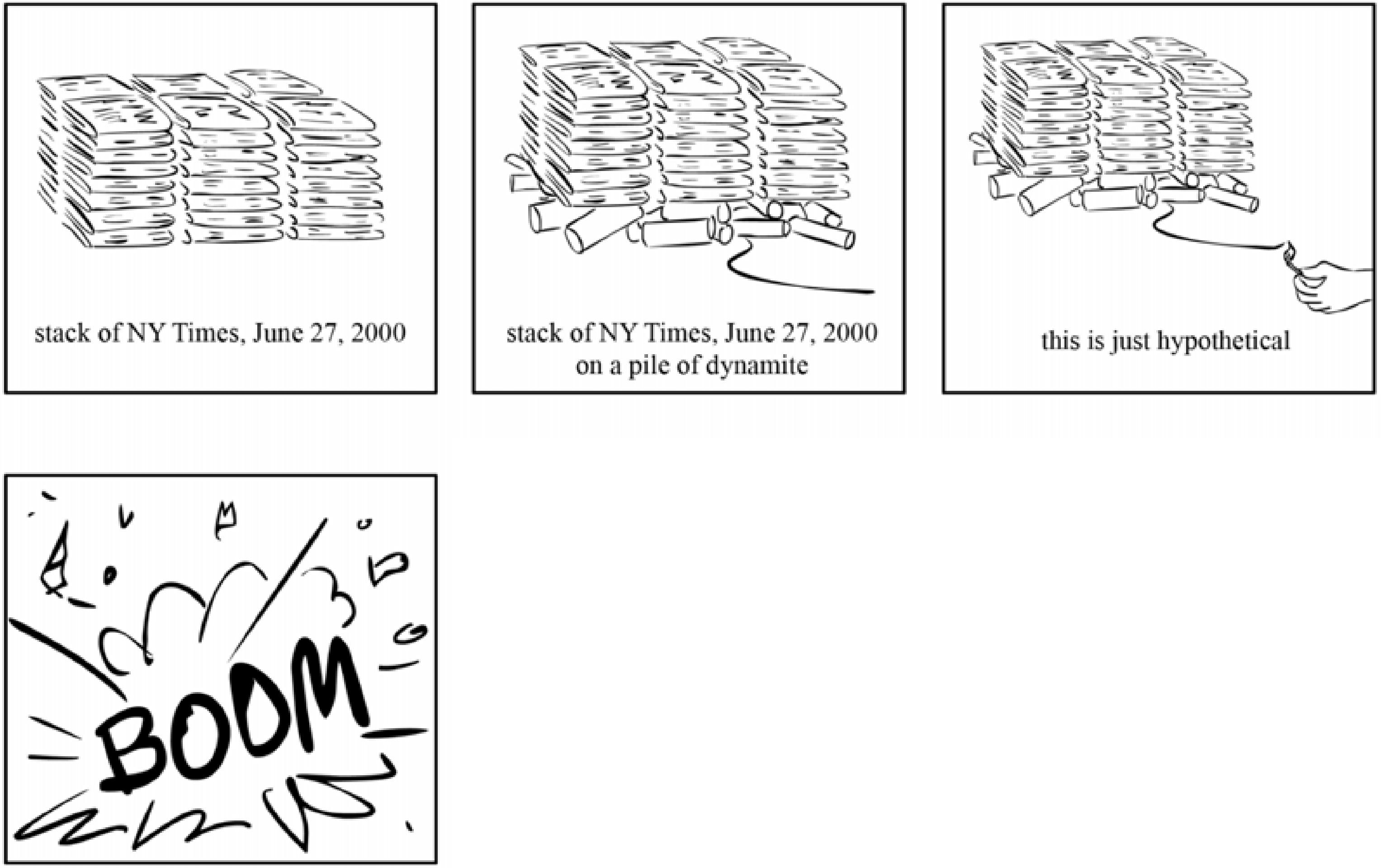


stack of NY Times, June 27, 2000
on a pile of dynamite

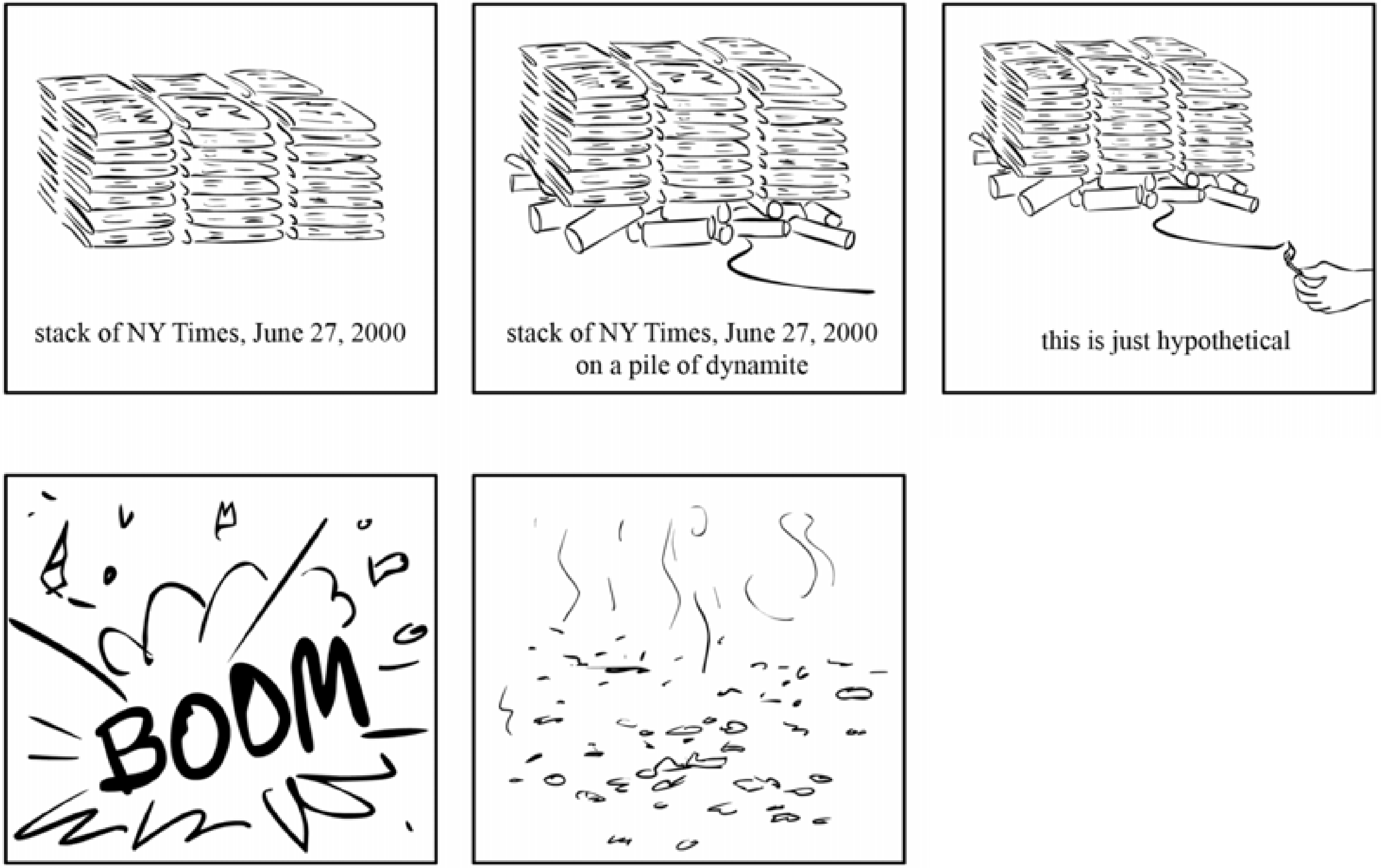
The Newspaper Problem


stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical

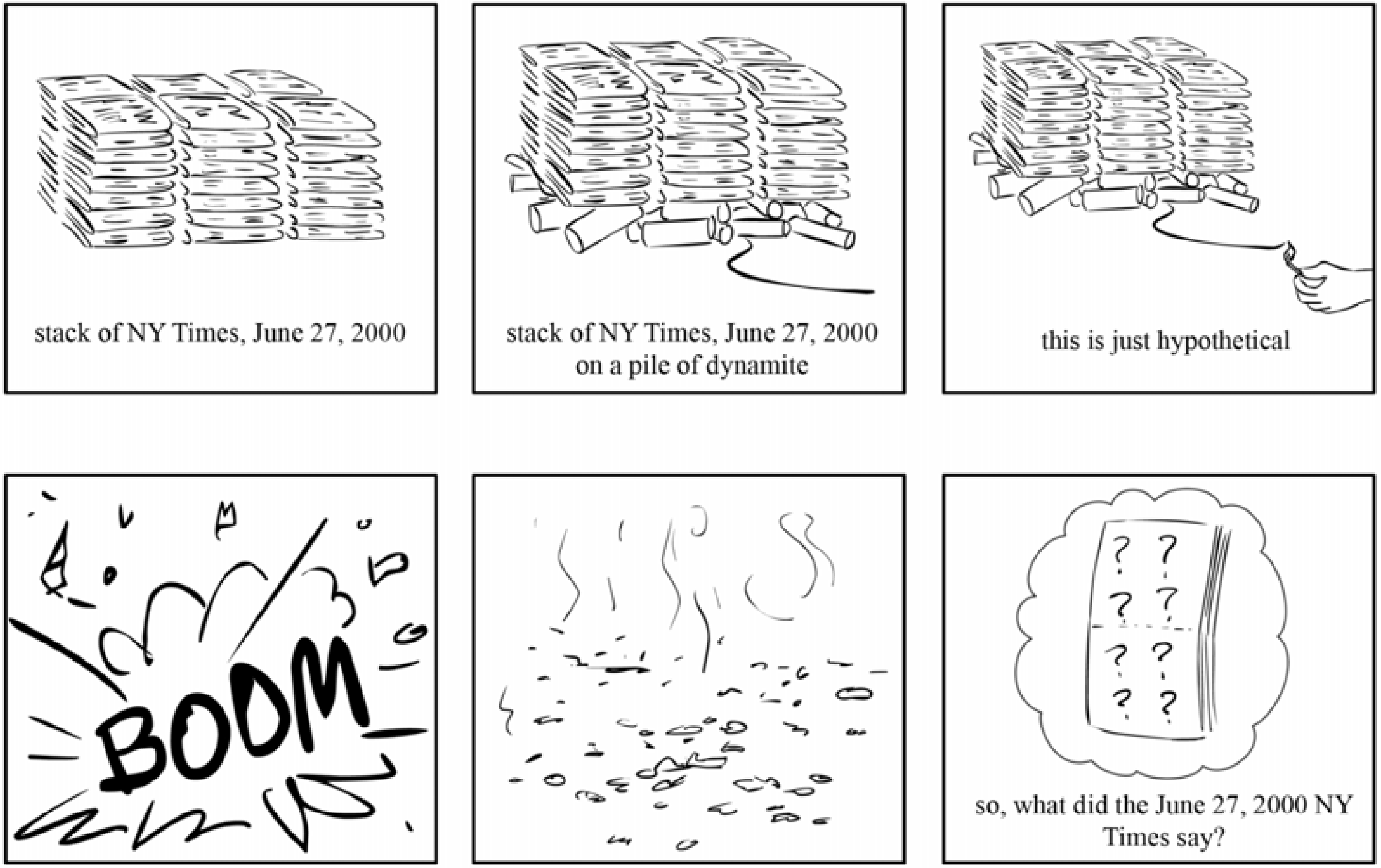
The Newspaper Problem


stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM

The Newspaper Problem

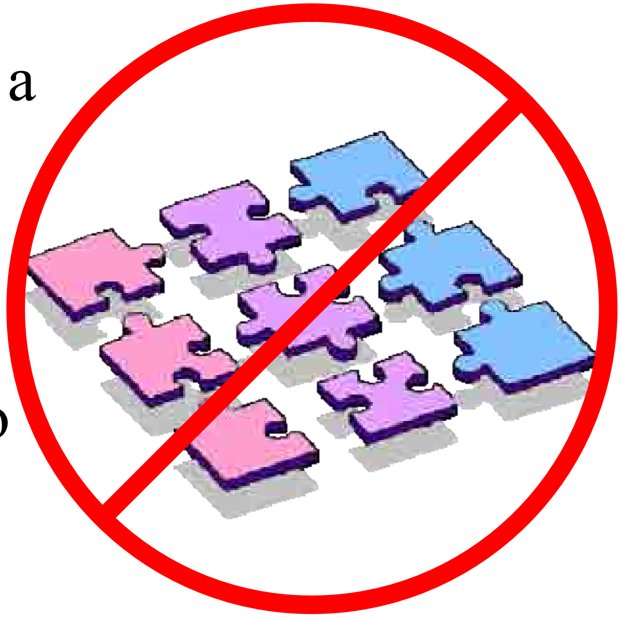

stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM

The Newspaper Problem


stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM
? ? ? ? ? ? ? ?
so, what did the June 27, 2000 NY Times say?

The Newspaper Problem as an "Overlap Puzzle"

- The newspaper problem is not the same as a jigsaw puzzle:
  - We have multiple copies of the *same* edition of a newspaper.
  - Plus, some pieces of paper got blown to bits in the explosion.

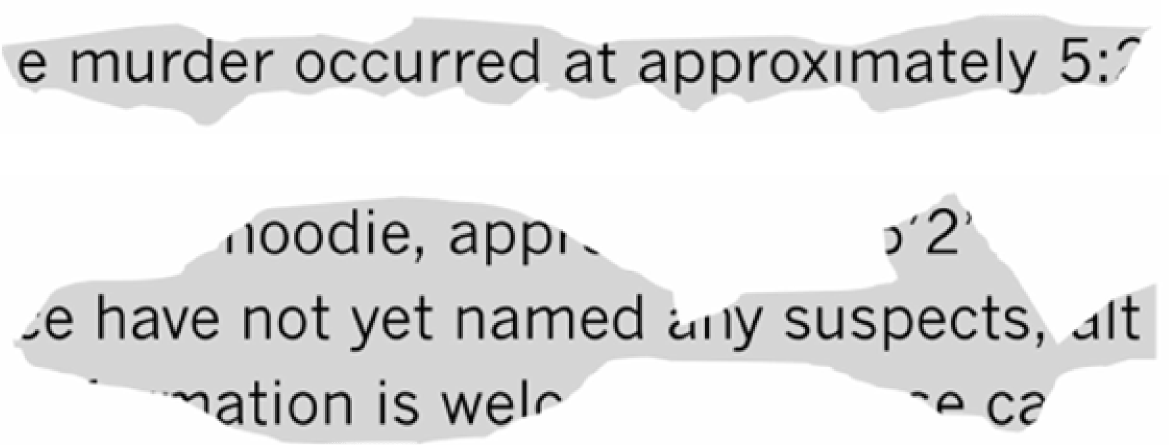


- Instead, we must use *overlapping* shreds of paper to reconstruct what the newspaper said.

- This gives us a giant **overlap puzzle**!

Sequencing is Harder than Newspaper Problem

- In the newspaper problem, we have the rules of language and common sense (e.g. “**murder**” and “**suspect**” would often appear near each other in a newspaper.)



e murder occurred at approximately 5:

hoodie, appr 5’2’
e have not yet named any suspects, alt
mation is welc e ca

- However, the “language” of DNA remains largely unknown.

Sequencing is Harder than Newspaper Problem

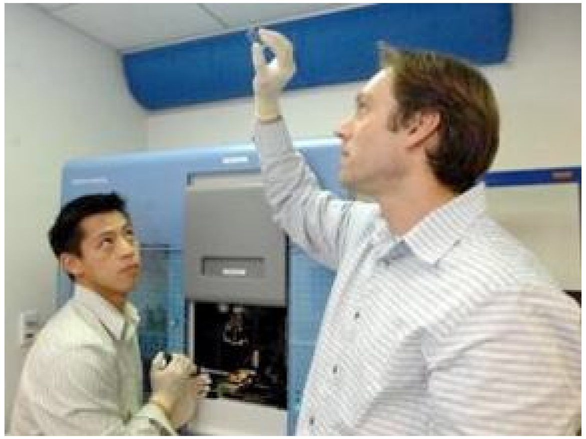
- There are lots of repeated substrings in every genome (50% of human genome is formed by repeats).
  - **Example**: GCTT is repeated 4 times in the following:

    AAGCTTCTATTGCTTAATTGGCTTGCTTCGCTTTG

- **Analogy**: The Triazzle puzzle contains lots of repeated figures. This makes it very difficult to solve (even with just 16 pieces).

Sequencing a Genome: Lab + Computation

- **Read Generation** (Experimental): Generate many reads from multiple copies of the same genome.

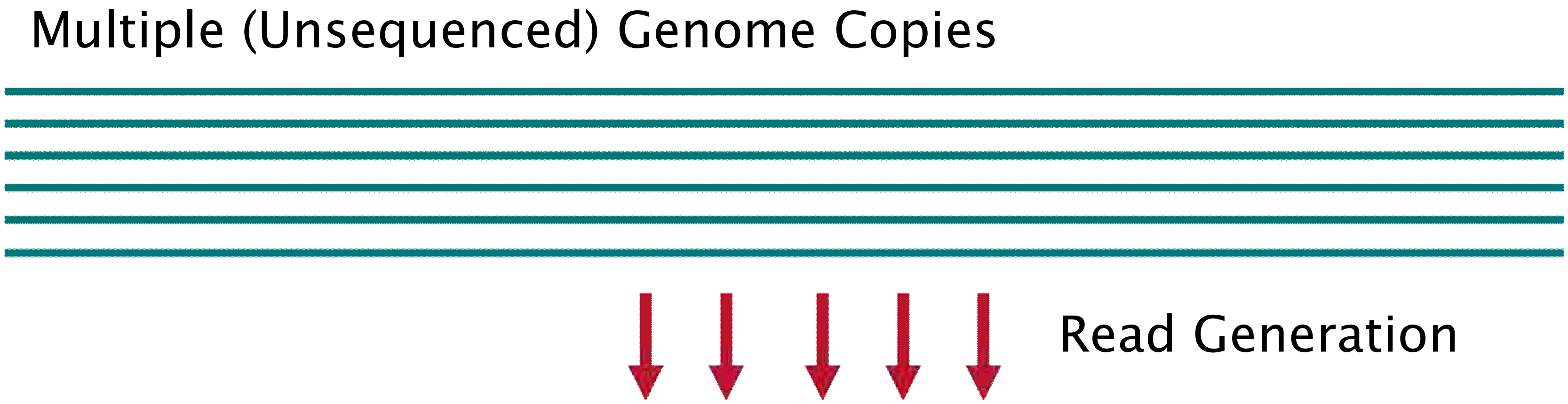


- **Fragment Assembly** (Computational): Use these reads to algorithmically put the genome back together.

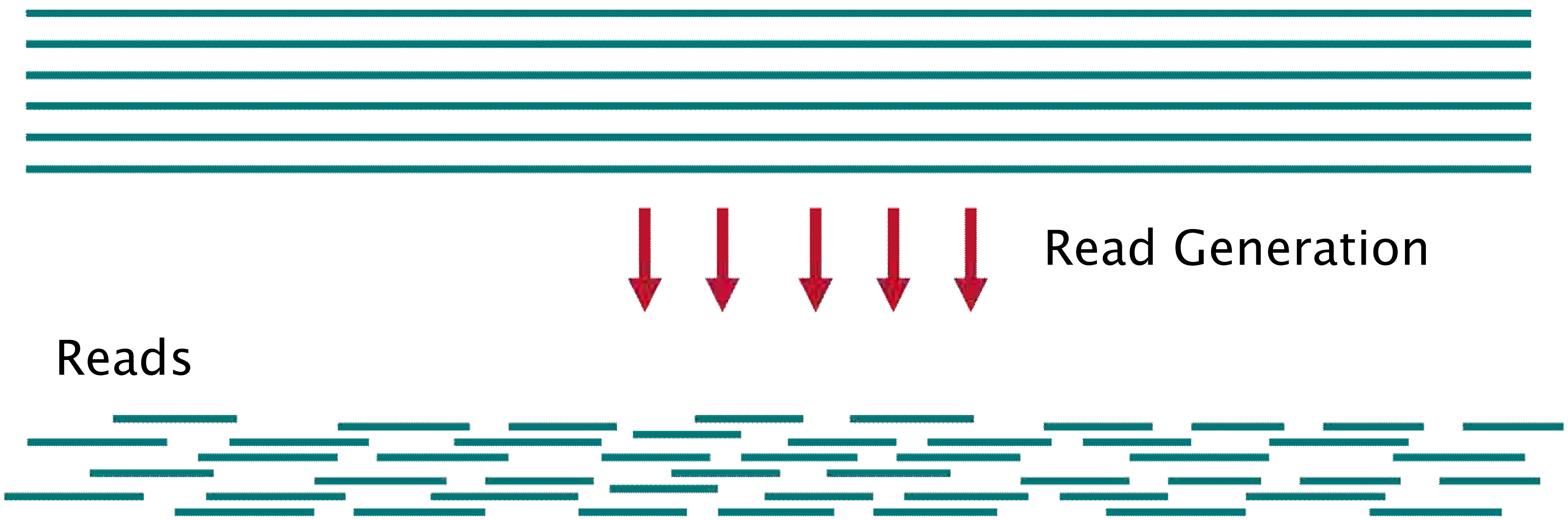
Sequencing a Genome: Illustration

Multiple (Unsequenced) Genome Copies

Sequencing a Genome: Illustration



Sequencing a Genome: Illustration

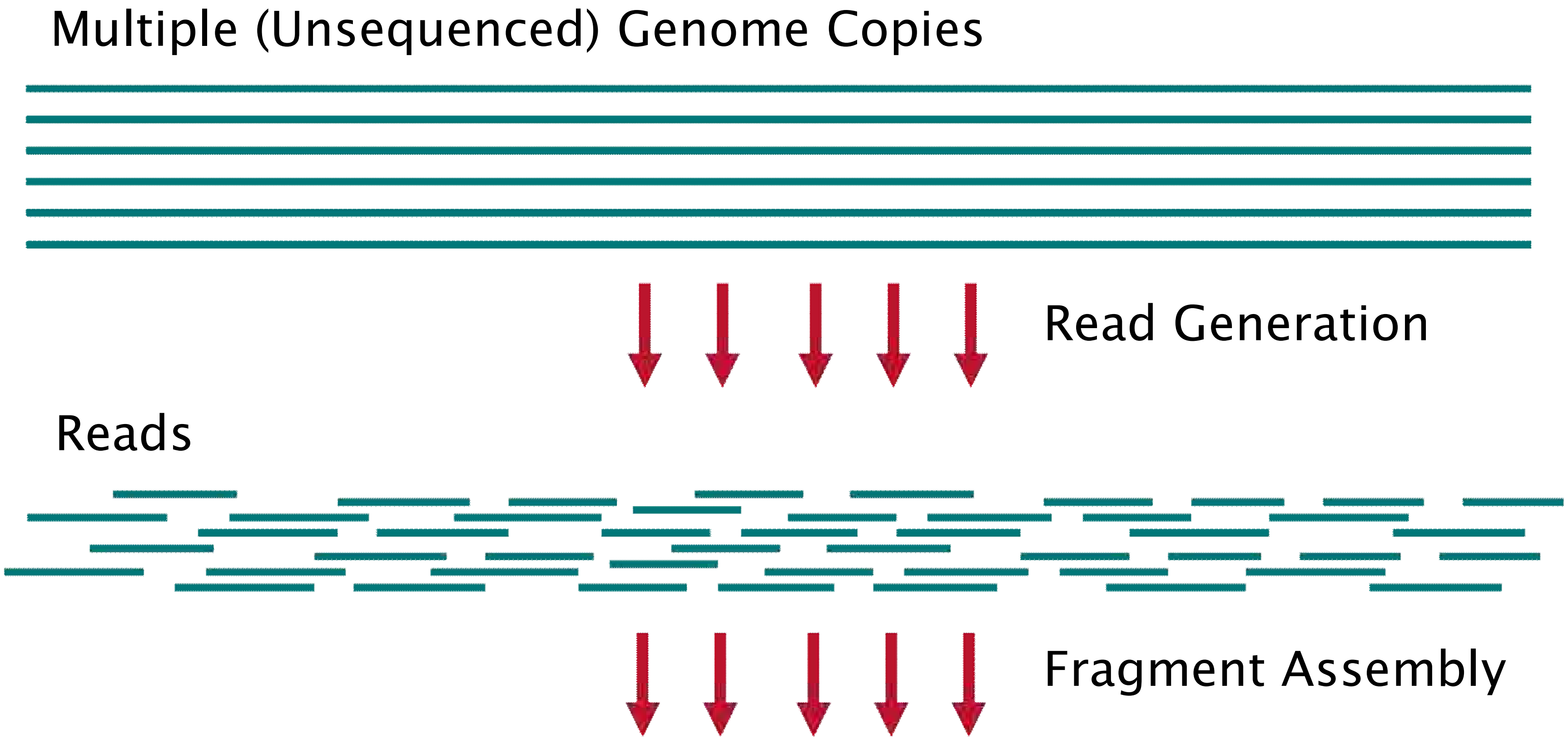


Multiple (Unsequenced) Genome Copies

Read Generation

Reads

Sequencing a Genome: Illustration



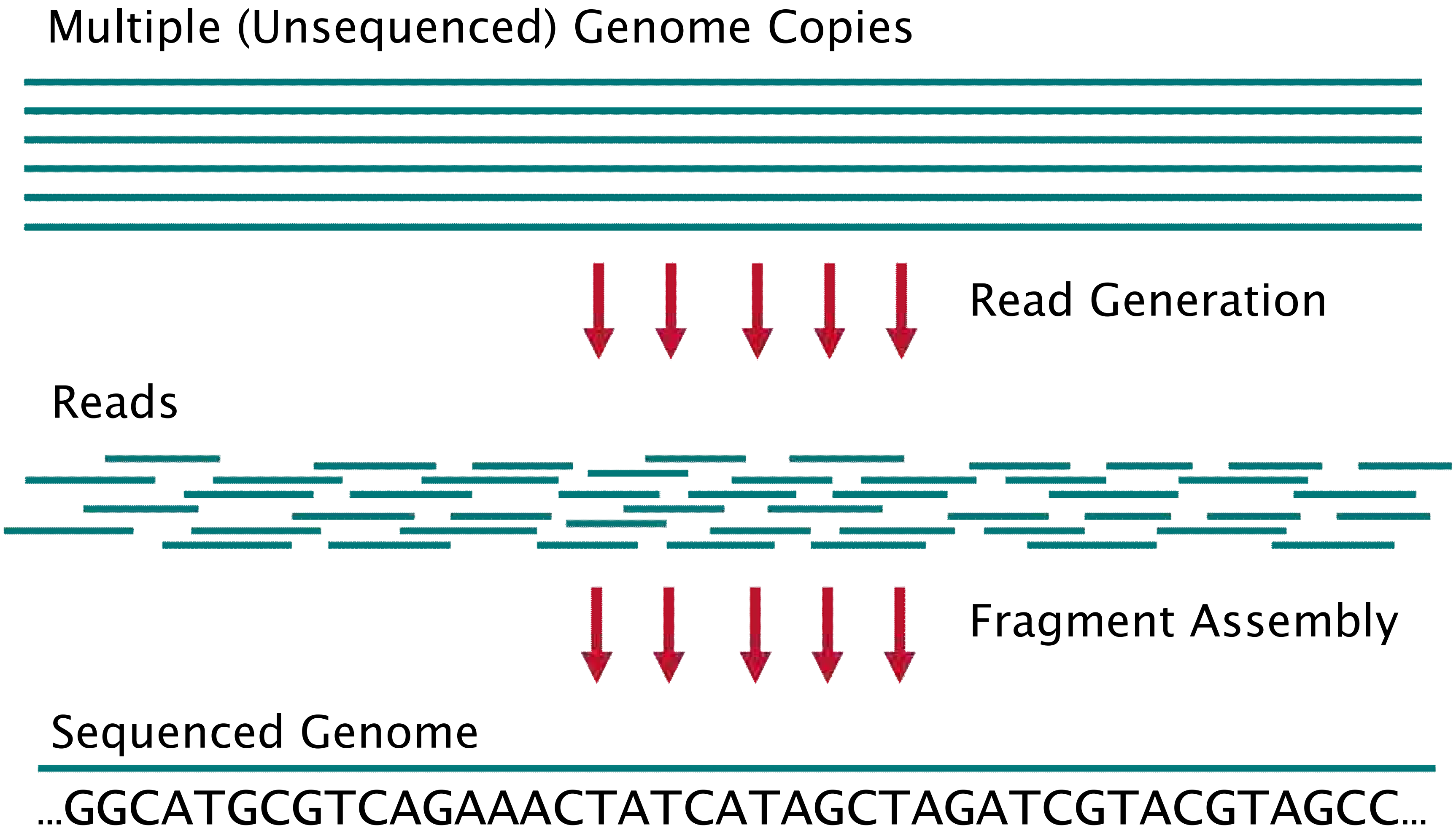
Multiple (Unsequenced) Genome Copies

Read Generation

Reads

Fragment Assembly

Sequencing a Genome: Illustration



Multiple (Unsequenced) Genome Copies

Read Generation

Reads

Fragment Assembly

Sequenced Genome

...GGCATGCGTCAGAAACTATCATAGCTAGATCGTACGTAGCC...

DNA Chips: From an Idea to a New Industry

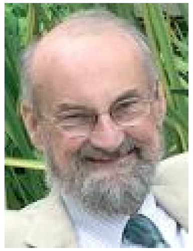
- **1989:** Radoje Drmanac, Andrey Mirzabekov, and Edwin Southern independently invent **DNA chips (arrays)** for read generation.

- **Key Idea**: Generate all ***k*-mers** (see below) from the genome in the hope that they can be assembled to reconstruct the genome.

- **1989**: *Science* magazine writes, "Using DNA arrays for sequencing would simply be substituting one horrendous task for another."

Mirzabekov

Drmanac



Southern

***k*-mer**: A string of length *k* (in an alphabet of 4 nucleotides)

Short Read Sequencing and de Bruijn Graphs

Short read sequencing was first proposed in 1988 under the name of DNA chips or Sequencing by Hybridization (SBH)

- **1988 (Drmanac, Mirzabekov, and Southern's groups)** suggested SBH as an alternative to Sanger sequencing. **Nobody believed it will ever work**

- **1989 (P.P., JBSD 1989)** *de Bruijn* approach for short read SBH assembly

- **2000**: DNA arrays are a multi-billion dollar industry

First SBH array prototype ***(1989)***

First commercial DNA chip by Affymetrix ***(1995)***

Nicolaas de Bruijn

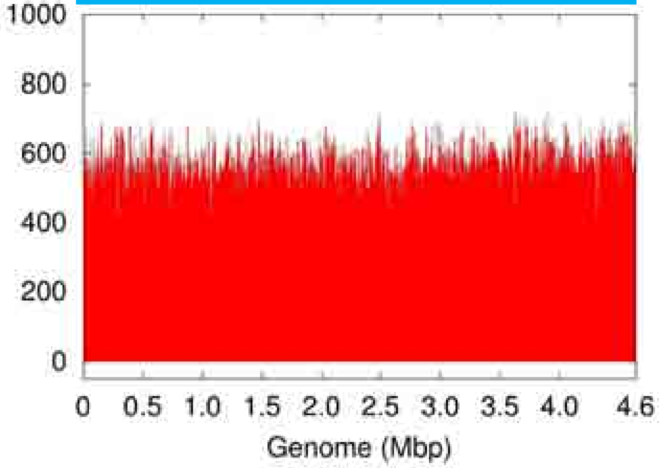
July 9, 1918 - February 17, 2012

Why is Assembly of Single Cell Data Challenging?

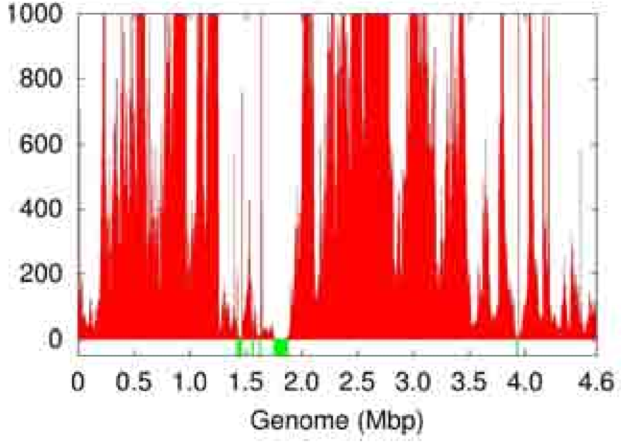
- Orders of magnitude difference in read coverage between different regions
- Elevated number of chimeric reads and chimeric read-pairs
- Elevated number of sequencing errors
- **Existing NGS assemblers were not designed to handle these complications:**

  ``challenges facing the single cell sequencing are increasingly computational rather than experimental" (Rodrigue et al. 2009)

Coverage:
E.Coli **standard** data



Coverage:
E.Coli **single cell** data



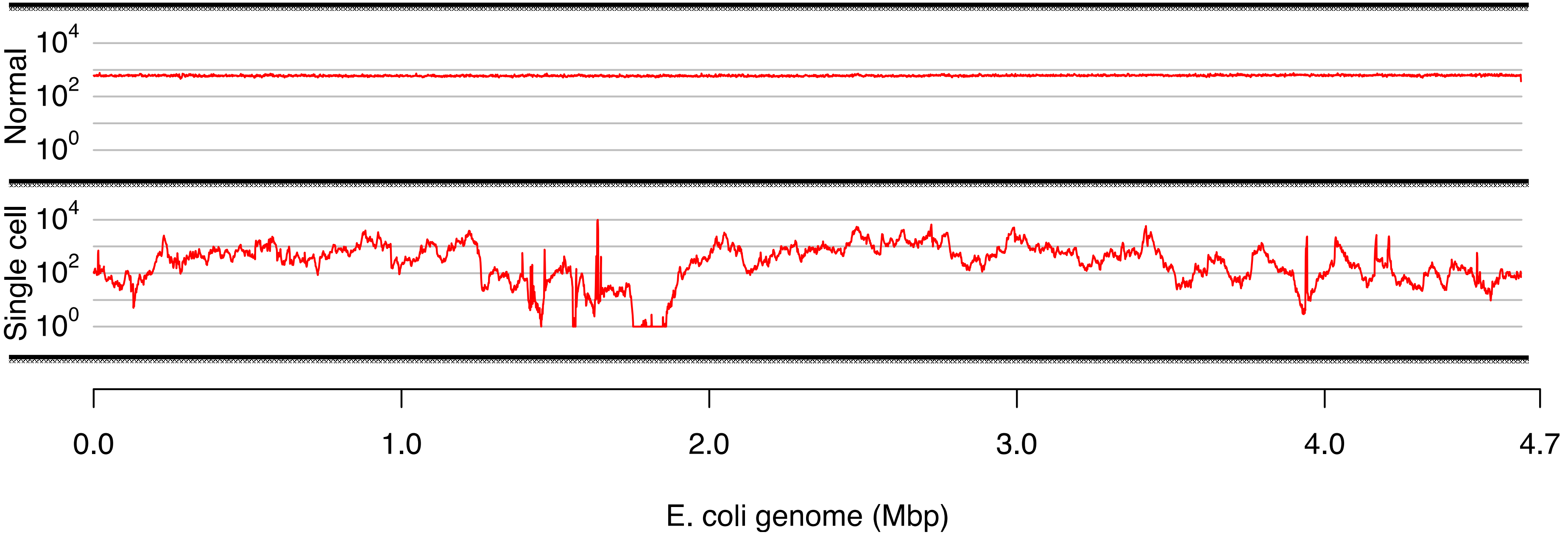
De Bruijn Assemblers

- Idury and Waterman, JCB 1995
- PP, Tang, Waterman, PNAS 2001 (Euler)
- PP, Tang, Tesler, Genome Res, 2004 (A-Bruijn assembly)
- Chaisson and PP, Genome Res. 2008 (Euler-SR)
- Zerbino and Birney, Genome Res. 2008 (Velvet)
- Simpson et al., Genome Res. 2008 (ABySS)
- Butler et al. Genome Res. 2008, Gnuerre et al. Genome Res. 2011 (ALLPATHS)
- Li et al., Genome Res. 2010 (SOAPdenovo)
- and others …

None of them works well with single cell data.
No error correction tool works well with single cell data.

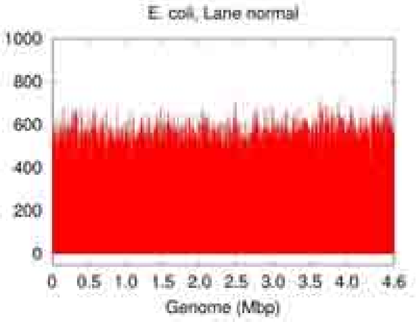
Read Coverage: Multicell vs. Single Cell

E. coli read coverage: logarithmic scale

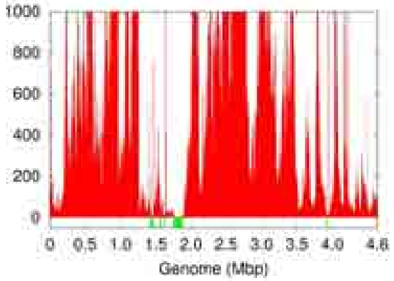


How NGS Assemblers Handle Variations in Coverage?

- **Multicell reads:** read coverage distribution is uniform (average coverage 600X).



- **Single cell reads:** read overage varies widely along the genome (from no coverage to 10000X).

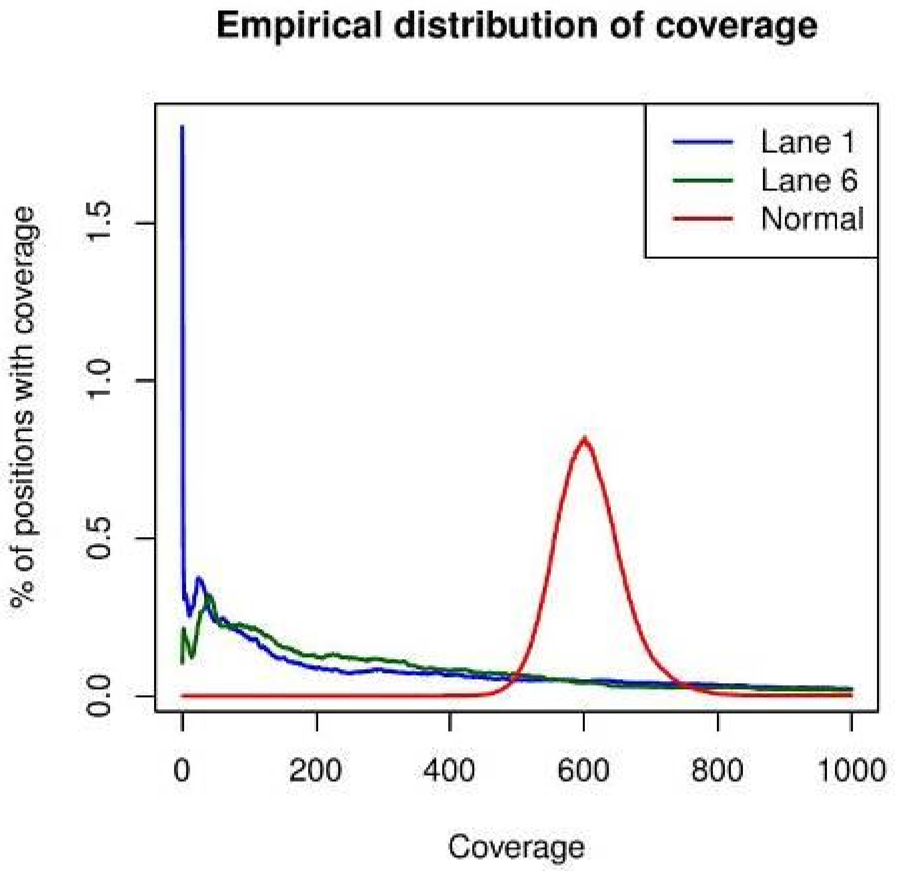


- In single cell projects, correct segments may have 100 times lower coverage then erroneous segments, thus confusing NGS assemblers.

- Existing assemblers (e.g. Velvet) impose a coverage cutoff to avoid assembly errors. **Large cutoff eliminates 25% of valid single cell data. Small cutoff leads to many assembly errors.**

How NGS Assemblers Handle Variations in Coverage?

- **Multicell reads:** coverage distribution is centered sharply at 600X (average coverage).
- **Single cell reads:** coverage varies widely across the entire range (from no coverage to 10000X and higher).
- In single cell projects, correct segments may have coverage 10 and erroneous segments may have coverage 1000, thus confusing NGS assemblers.
- Existing assemblers (e.g. Velvet) impose a coverage cutoff to avoid assembly errors. **A cutoff threshold eliminates 25% of valid data in the single cell case!**



Red: multicell coverage
Blue (or green): single cell coverage

E+V-SC Single Cell Assembler

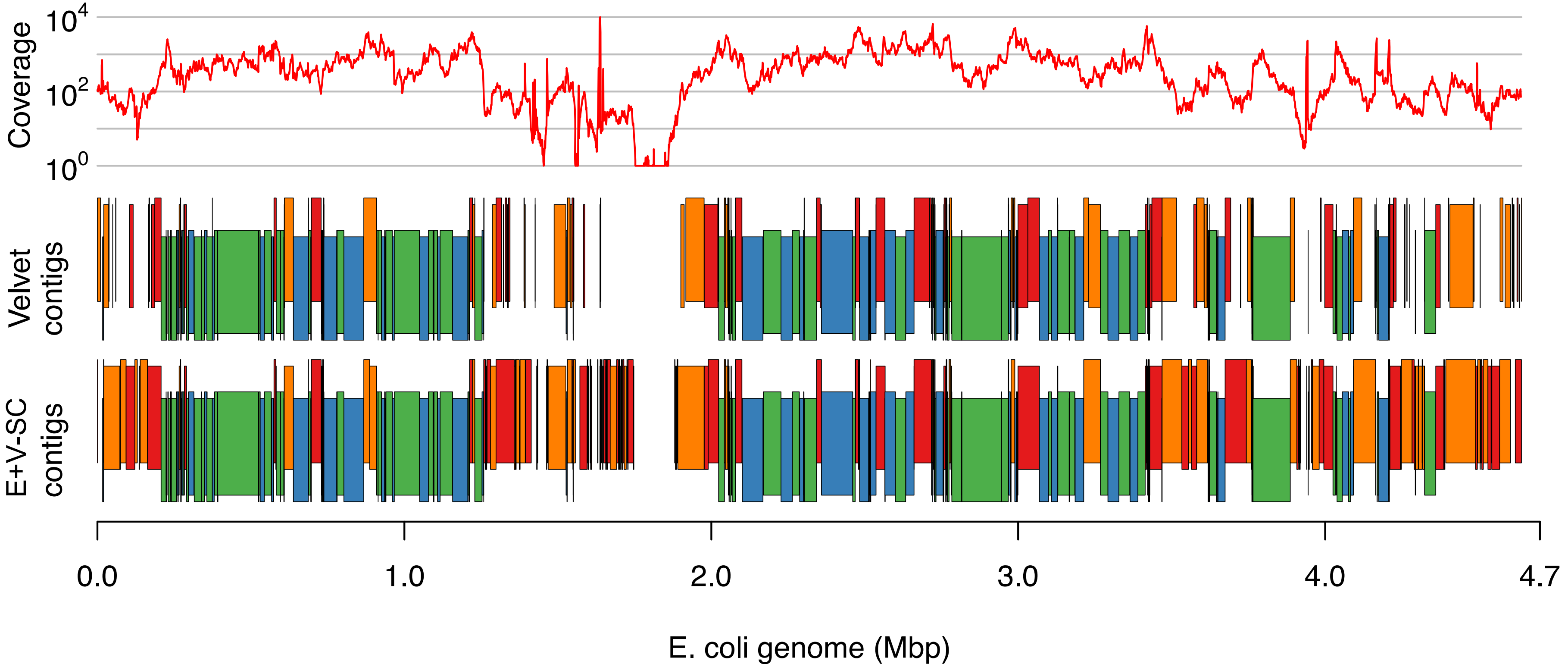
- E+V-SC (**E**uler**+V**elvet**-S**ingle **C**ell assembler) adapted components from EULER and Velvet.

Chaisson & PP, *Genome Res.* (2008)
Zerbino & Birney, *Genome Res.* (2008)

- Error correction in reads from EULER.
- Instead of a global threshold on coverage for the whole de Bruijn graph in Velvet, E+V-SC is adapted to local conditions.
- **E+V-SC has 28% increase in genome coverage and 23% increase in the number of captured genes as compared to Velvet.**

Chitsaz et al., *Nature Biotech.* 2011

E. coli: Single Cell Assemblies



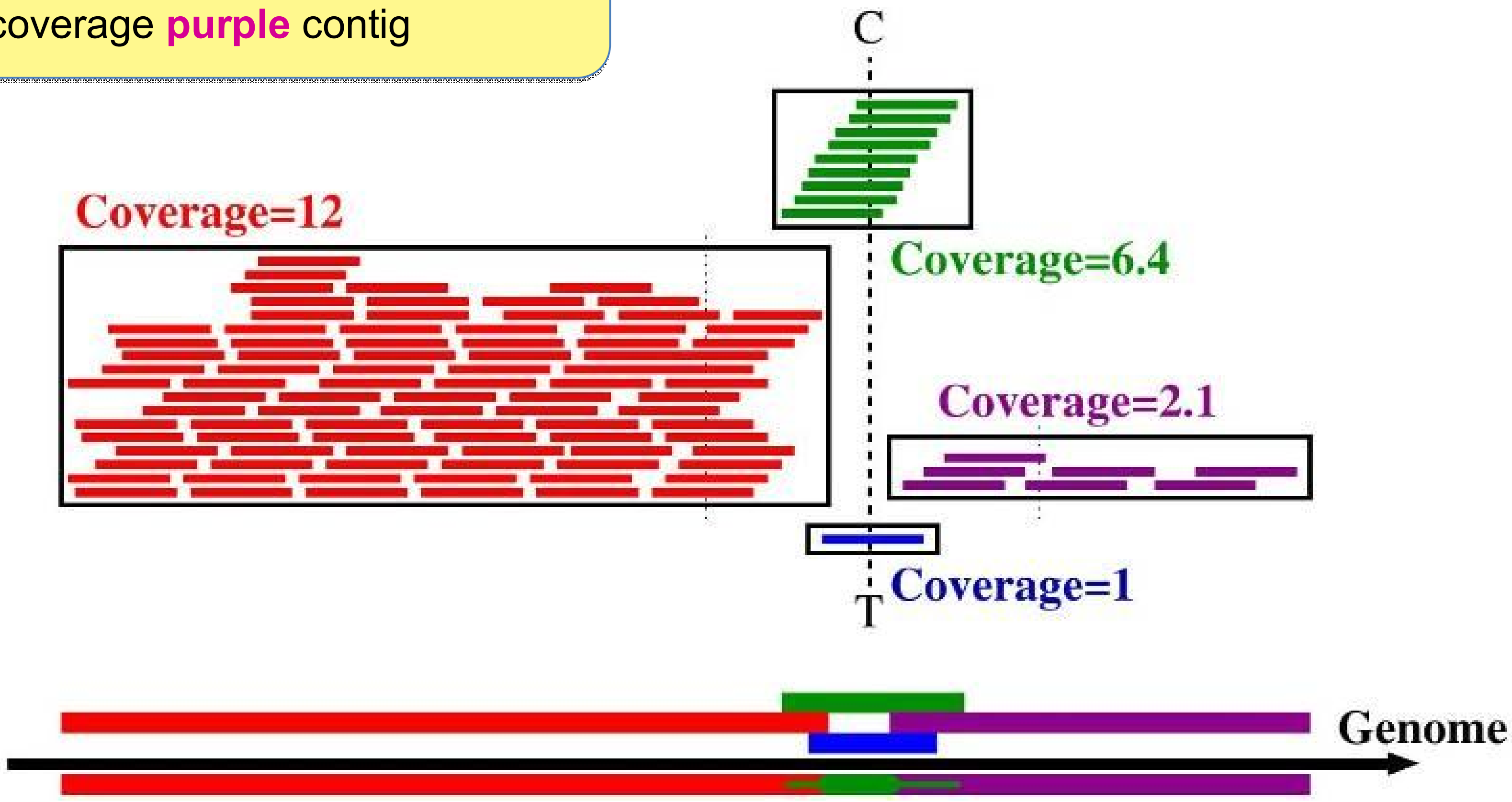
🟩🟦 Matching contigs

🟥🟧 Different lengths, fragmented, missing in one assembly, etc.

Chitsaz, et al., *Nature Biotech.* 2011

Rescuing Low Coverage Contigs

Removing the lowest coverage **blue** contig (edge in de Bruijn graph) rescues the low coverage **purple** contig

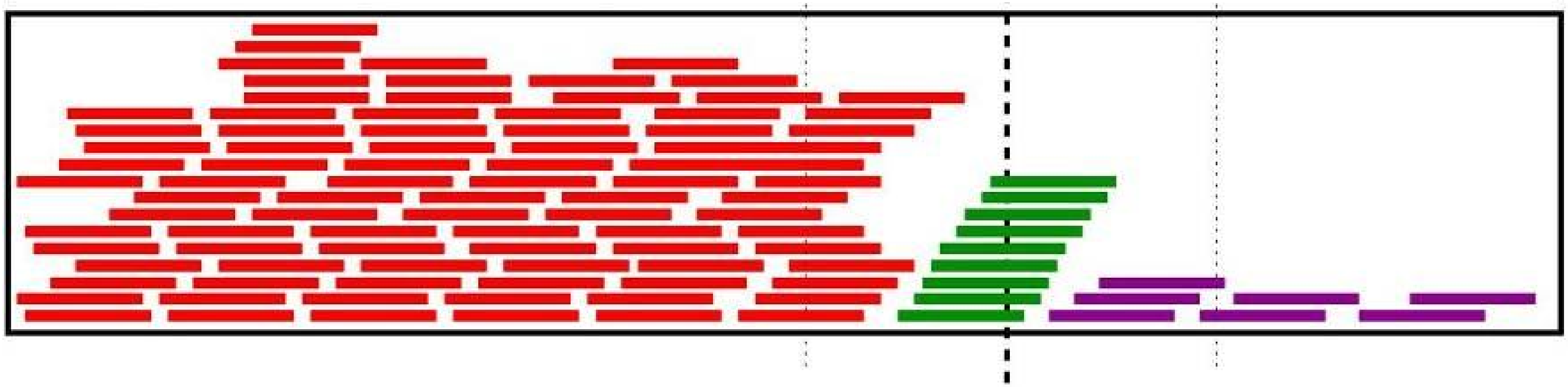


Rescuing Low Coverage Contigs

After removal of erroneous contig

Merged red-green-purple contig:

while purple regions has low coverage, *AVERAGE* coverage across the entire contig is high (preventing the removal of the low coverage purple region)



Velvet vs. Velvet-SC

Velvet is an open source de Bruijn graph based *de novo* assembler from EBI.

Velvet **assembly algorithm**

```
1:  Build a roadmap rdmap from R by indexing all k-mers.
2:  Build a de Bruijn pregraph pg from rdmap.
3:  Clip tips of pg.
4:  Build a graph from pg by threading R.
5:  Condense graph by merging 1-in 1-out vertices.
6:  Clip tips of graph.
7:  Correct graph by the Tour Bus algorithm.
8:  Remove vertices with average coverage < cutoff
9:  Clip tips of graph.
10: Correct graph by the Tour Bus algorithm.
11: Resolve repeats using read pairing.
12: Condense graph by merging 1-in 1-out vertices.
13: Return vertices of graph as contigs.
```

Zerbino & Birney, *Genome Res.* (2008) 18:821-829

Our assembly algorithm ("E+V-SC")

(a) ***EULER-SR*** **error correction**

(b) ***Velvet-SC*** **assembly algorithm**

```
1-7: Same as Velvet assembly algorithm.
8:   for i =2 to cutoff do
9:       Remove vertices with average coverage < i
10:      Clip tips of graph.
11:      Correct graph by the Tour Bus algorithm.
12:      Resolve repeats using read pairing.
13:      Condense graph by merging 1-in1-out vertices.
14:  end for
15:  Return vertices of graph as contigs.
```

Chitsaz et al., *Nat. Biotechnol.* (2011)

Single Cell Assemblies:
Capturing 600 Extra *E. coli* Genes with E+V-SC

Assembler	# contigs	N50 (bp)	Assembly size	Genes
EULER	1344	26662	4369634	3178
Edena	1592	3919	3996911	2425
SOAPdenovo	1240	18468	4237595	3021
Velvet	428	22648	3533351	3055
E+V-SC	501	**32051**	**4570583**	**3753**

N50 = the contig length at which longer contigs represent half of the total genome length.

New Marine Genome: *Deltaproteobacterium*

Assembler	# of contigs	N50 (bp)	Length (bp)	# Conserved single copy genes
Velvet	1,856	11,531	3,921,396	55/111 (46%)
E+V-SC	823	30,293	4,282,110	75/111 (67%)

Over 3800 genes are fully assembled by E+V-SC

New Genome

Deltaproteobacteria (marine bacteria) single cell assembly features

Assembly size	4.3 Mb
Estimated genome size	4.9-6.4 Mb
# assembled genes	3811

Chitsaz, et al., *Nat. Biotechnol.* (2011)

How Complete Are Single Cell Assemblies?

- Jonathan Badger at Venter Institute annotated *Deltaproteobacterium* single cell assembly using metrics from Nelson et al., *Science* (2010)

- Conclusion: **single cell *Deltaproteobacterium* assembly is similar in quality to standard microbial assemblies (before finishing)**

# tRNA genes	20 out of 20 types
# tRNA synthetases	17 of 21 types
# rRNAs	1 each of 5S, 16S, 23S
# conserved single copy genes	75 out of 111 (67%)
# conserved single copy gene clusters	58 out of 66 (87%)

Chitsaz, et al., *Nature Biotech.* 2011

Future Work

- We plan to do sequencing and *de novo* assembly of more unknown single cell genomes, in collaboration with Roger Lasken, JCVI and Pavel Pevzner and Glenn Tesler, UCSD.
- This may revolutionize environmental microbiology and metagenomics.
- Medical application in hospitals to sequence drug resistant pathogens is a future direction.
- As we get more data, we may be able to model MDA biases, potentially using Machine Learning techniques, and design more efficient algorithms to correct such biases.

Agenda

- Fragment Assembly Problem
- De Bruijn Graph
- Paired de Bruijn Graph
- Results
- Questions

Fragments Assembly Problem

- **Previous approaches:**
  - Overlap-layout-consensus
  - De Bruijn Graph

- **New Approach:** Paired de Bruijn graph

From E+V-SC to SPAdes Assembler

La Jolla

+

Saint Petersburg

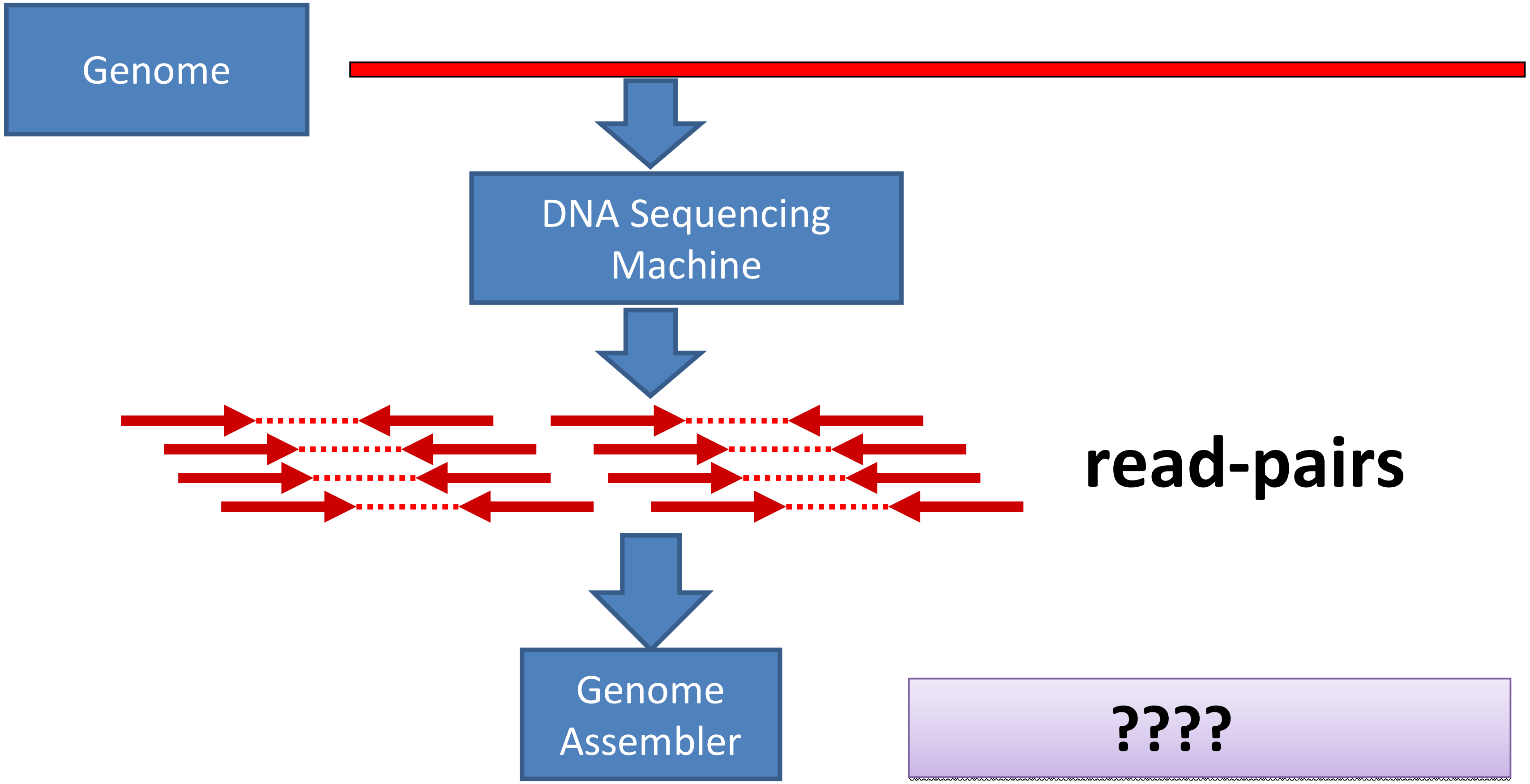
- In single cell projects, correct segments may have coverage 10 and erroneous segments may have coverage 1000, thus confusing NGS assemblers.

- **SPAdes tries not to use coverage in assembly decisions**

Nicolaas de Bruijn

July 9, 1918 - February 17, 2012

Fragment Assembly Problem: from Reads to Read-Pairs



Genome

DNA Sequencing Machine

read-pairs

Genome Assembler

????

Utilization of paired-end reads remains an open problem

From de Bruijn Graphs to Paired de Bruijn Graphs

- Assembling genome from ***k*-mers** (reads): elegant de Bruijn graph algorithm.

- Assembling genome from **paired *k*-mers** (read-pairs): not so elegant post-processing heuristics on de Bruijn graphs that often fail in repeat regions.

- Utilization of paired reads remains arguably the most poorly explored area of assembly.

Saint
Petersburg
Assembler:
SPAdes

Medvedev et al., JCB 2011: assembly of paired *k*-mers using **Paired de Bruijn Graphs (PDBG).**

Finally, an elegant but **IMPRACTICAL** approach to assembling paired *k*-mers ☺

Deja vu from 2001

- **Paired de Bruijn graphs** are impractical since distances between reads within read-pairs are imprecise

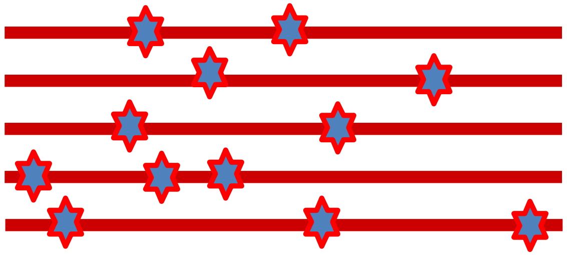
- But in 1995 **de Bruijn graphs** were not very practical either! At least for Sanger reads circa 1995…

Historic Reference

- De Bruijn assembly works when nearly every *k*-mer from genome appears in at least one read without errors
- **Thus, de Bruijn assembly requires either nearly error-free reads or high coverage.**
- **Neither condition held in 1995** when Idury and Waterman proposed de Bruijn assembly for Sanger reads: only ≈13% of 50-mers were correct!
- **Error-correction** (PP, Tang, Waterman, PNAS 2001) made reads nearly error-free (over 90% of 50-mers became correct) and made de Bruijn assembly practical even in low coverage Sanger projects

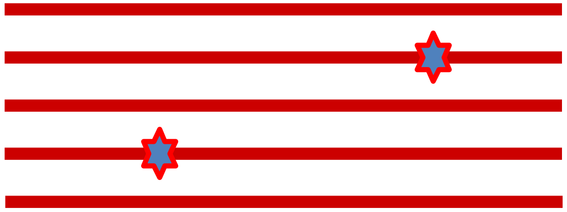
If reads were made nearly error-free in 2001, can we make distances between reads nearly exact in 2012?

Error Correction (2001)

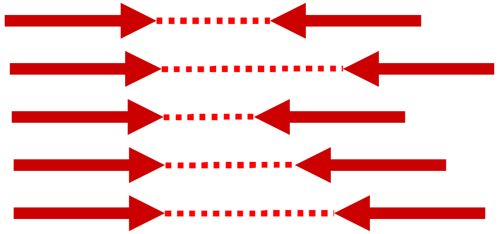


Error-prone reads

PP, Tang, Waterman, PNAS 2001

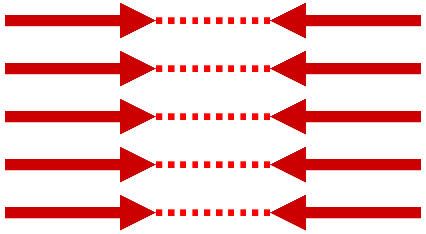


Read-Pair Adjustment (2012)

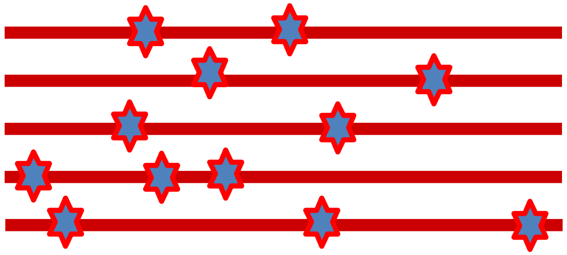


Read-pairs with variable insert sizes

Bankevich et al. JCB 2012

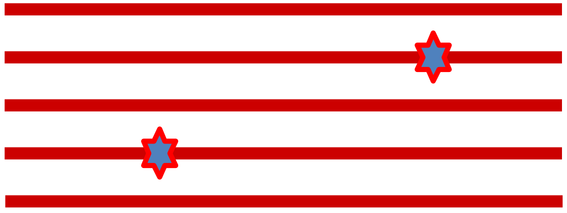


Error Correction



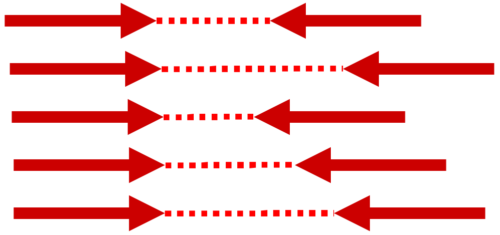
Error-prone reads

PP, Tang, Waterman PNAS 2001



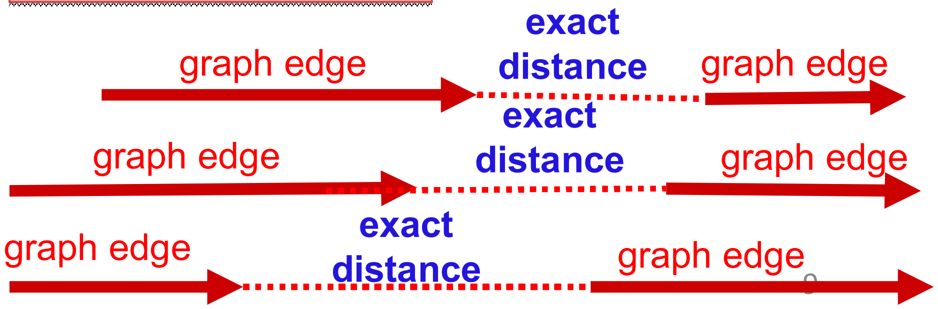
Read-Pair Adjustment

This sequencing machine produces **edge-pairs** instead of **read-pairs**



Read-pairs with variable insert sizes

Bankevich et al. JCB 2012

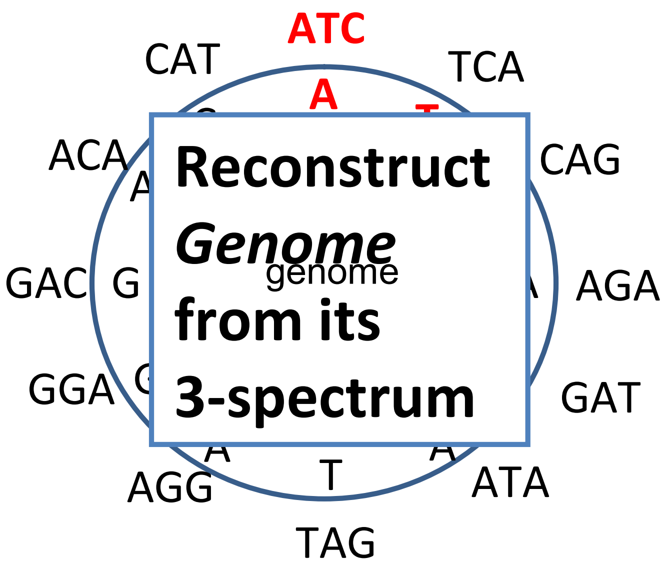


De Bruijn Assemblers

- Idury and Waterman, 1995
- Euler, Pevzner et al., 2001
- Euler-SR: Chaisson and Pevzner 2008
- Velvet: Zerbino and Birney 2008
- ABySS: Simpson et al., 2008
- ALLPATHS: Butler et al., 2008, 2011
- SOAPdenovo: Li et al., 2010
- and others …

None of them works well with single cell data.
No error correction tool works well with single cell data.

Reconstructing Genome from *k*-mers



Reconstruct *Genome* from its 3-spectrum

Generate a 3-mer at each position of a cyclic *Genome*=ATCAGATAGGAC.

The ***k*-spectrum** of *Genome* is the set of all ***k*-mers** of *Genome*.

Reconstructing Genome from *k*-mers

ATC

CAT

TCA

ACA

CAG

GAC

AGA

GGA

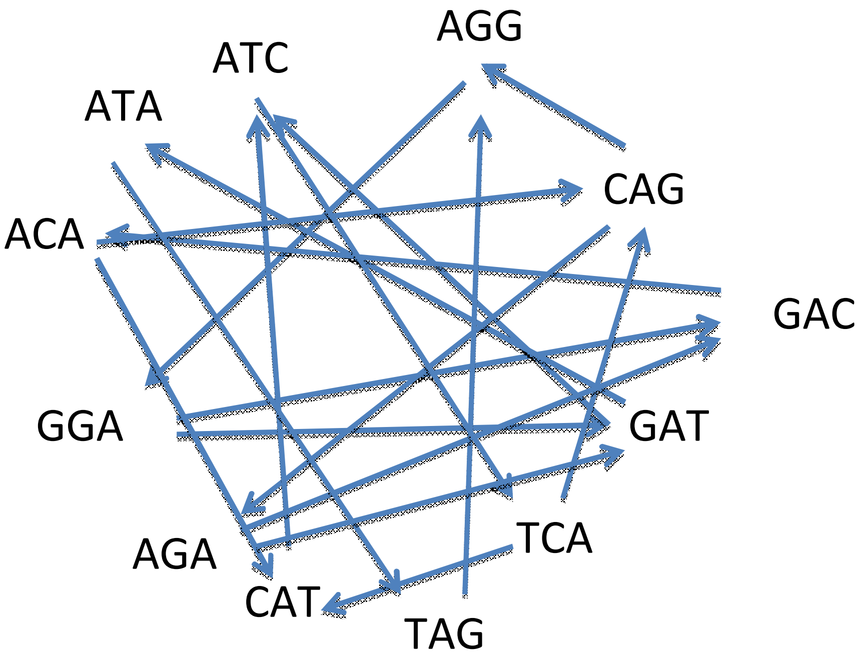
GAT

ATA

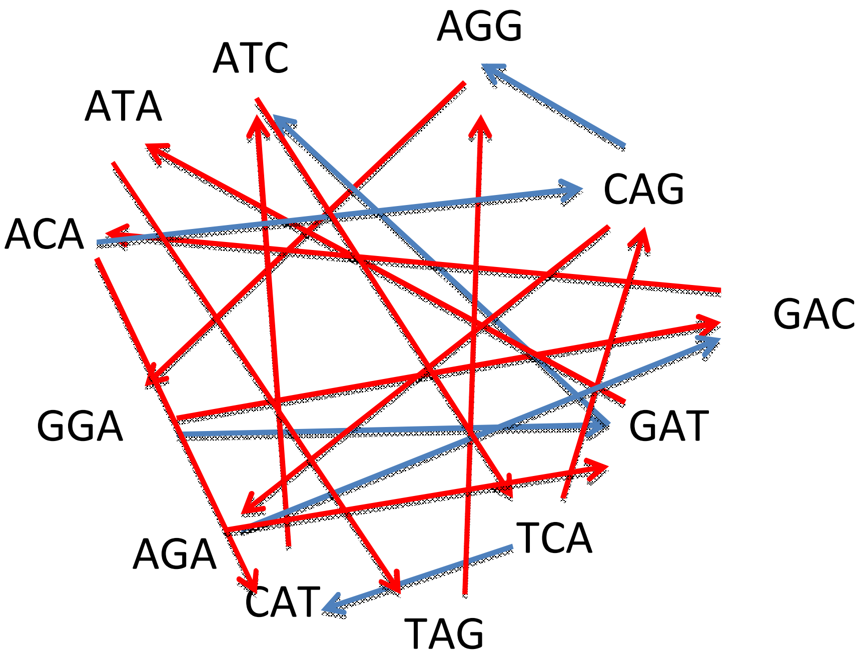
AGG

TAG

Reconstructing Genome from *k*-mers

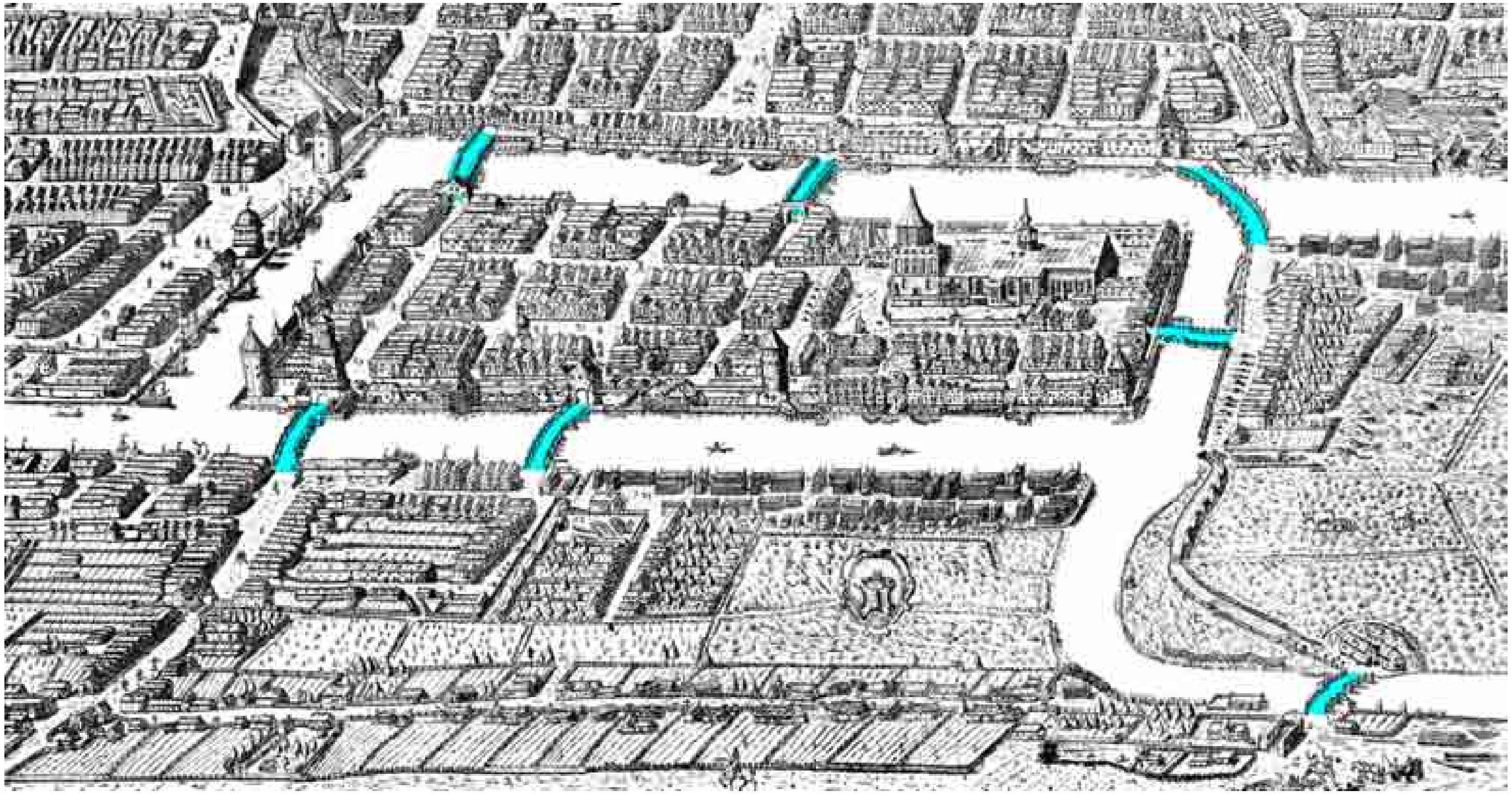


Reconstructing Genome from *k*-mers



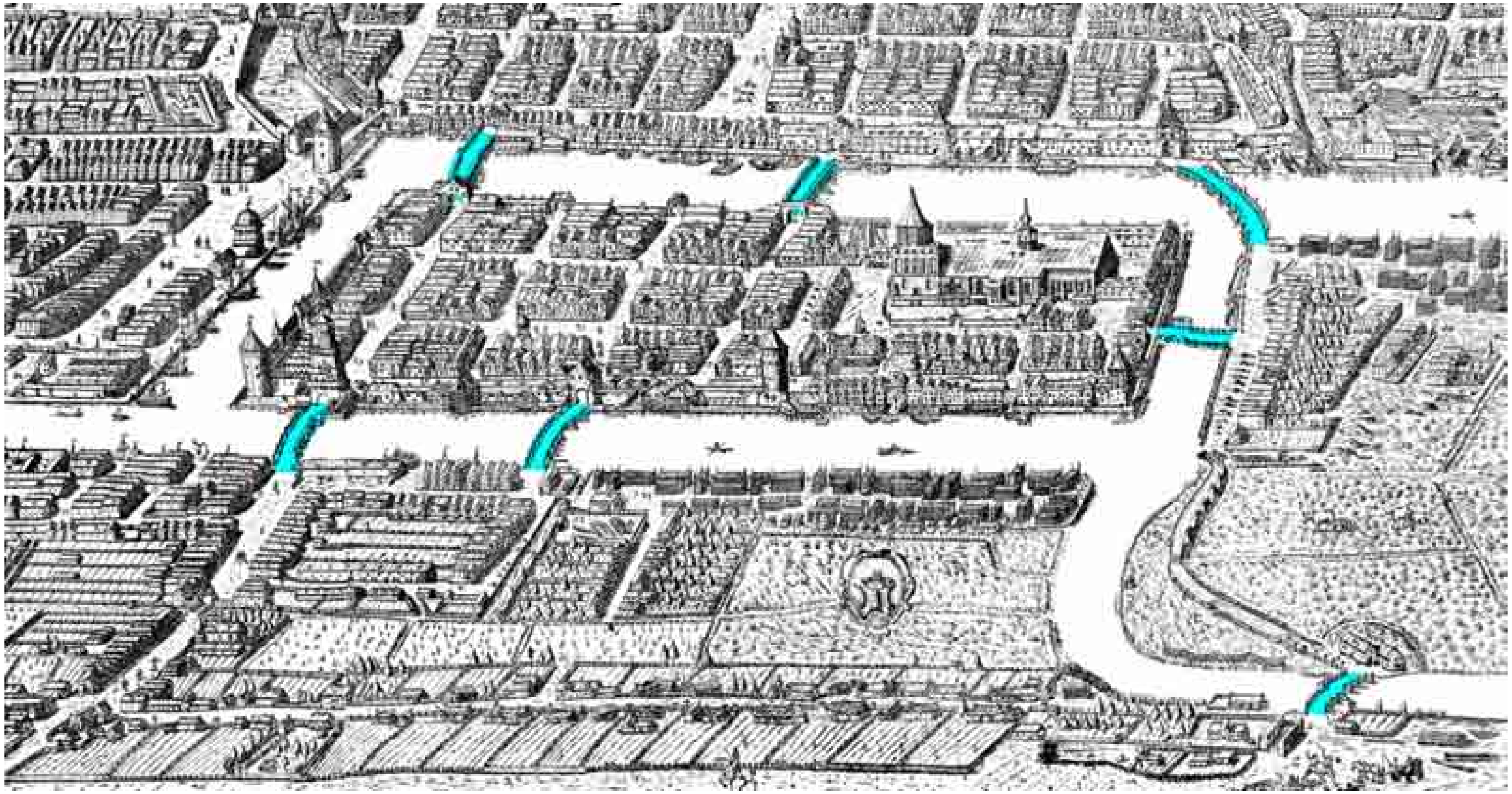
The Bridges of Königsberg

- The people of Königsberg, Prussia (present-day Kaliningrad, Russia) enjoyed taking walks.



The Bridges of Königsberg

- They wondered if they could walk through the city, cross each **bridge** (blue) **exactly once**, and return where they started.



The Bridges of Königsberg

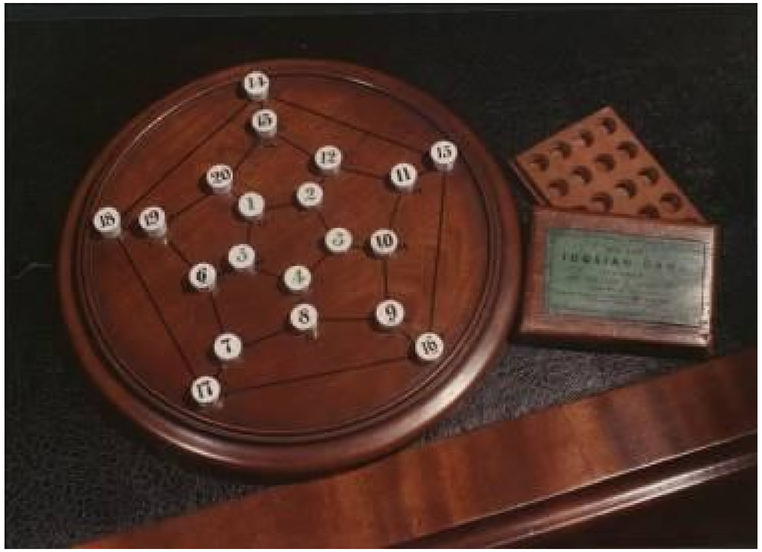
- **1735**: Leonhard Euler develops an approach to answer this question for *any* city, even for a "city" with a billion islands.

Leonhard Euler

The Icosian Game

- Over a century passes…
- **1857**: Irish mathematician William Hamilton designs a game consisting of a board representing 20 "islands" connected by "bridges."
- **Goal**: find a walk that visits every island **exactly once** and returns back where it started.

William Hamilton



Icosian Game

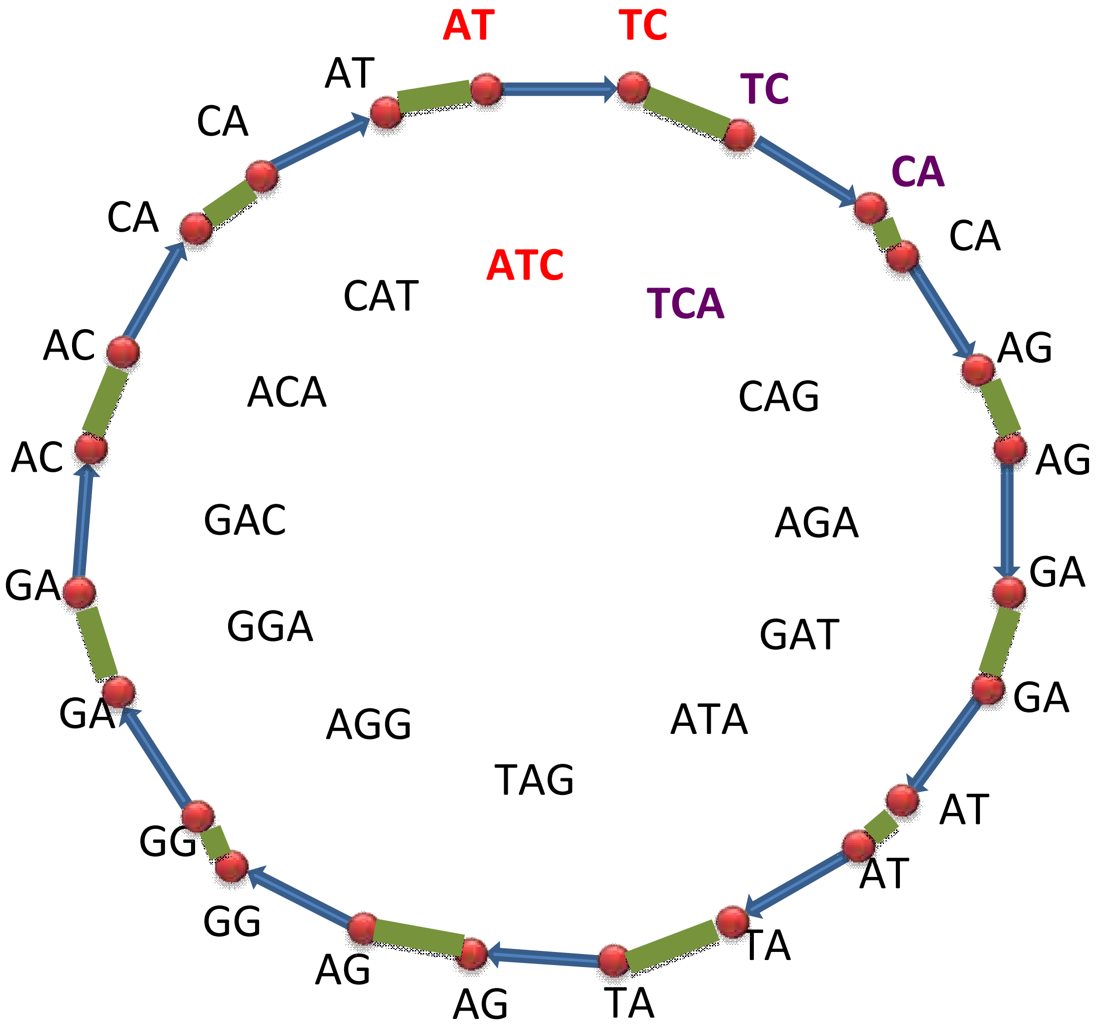
Similar Problems with Very Different Fates

- These two stories have something in common:
  - Find a walk that uses every ***bridge*** once (Konigsberg Bridges Problem)
  - Find a walk that visits every ***island*** once (Hamilton game)

- However, while Euler solved the first problem (even for a city with a million ***bridges***), mathematicians still do not know how to solve the second problem, even for a city with a million ***islands*.**

- **But where are the genomes???**

De Bruijn Graph Approach



Represent each ***k*-mer** as an edge from its **prefix** to its **suffix**. E.g., **ATC** is represented as **AT→TC**

Edges corresponding to consecutive *k*-mers share a node with the same label: **prefix(ATC)=suffix(TCA)=TC**.

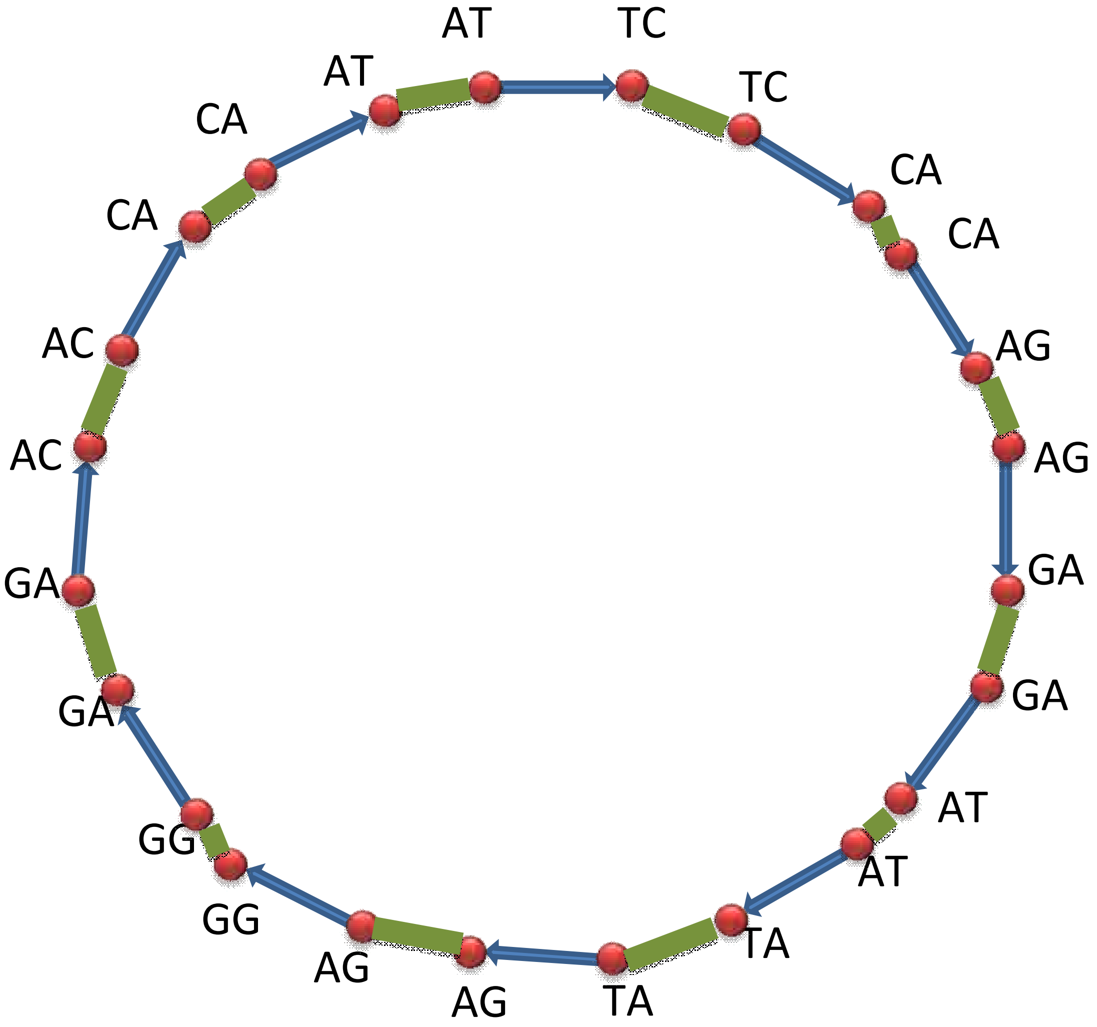
Glue identically labeled nodes

De Bruijn Graph

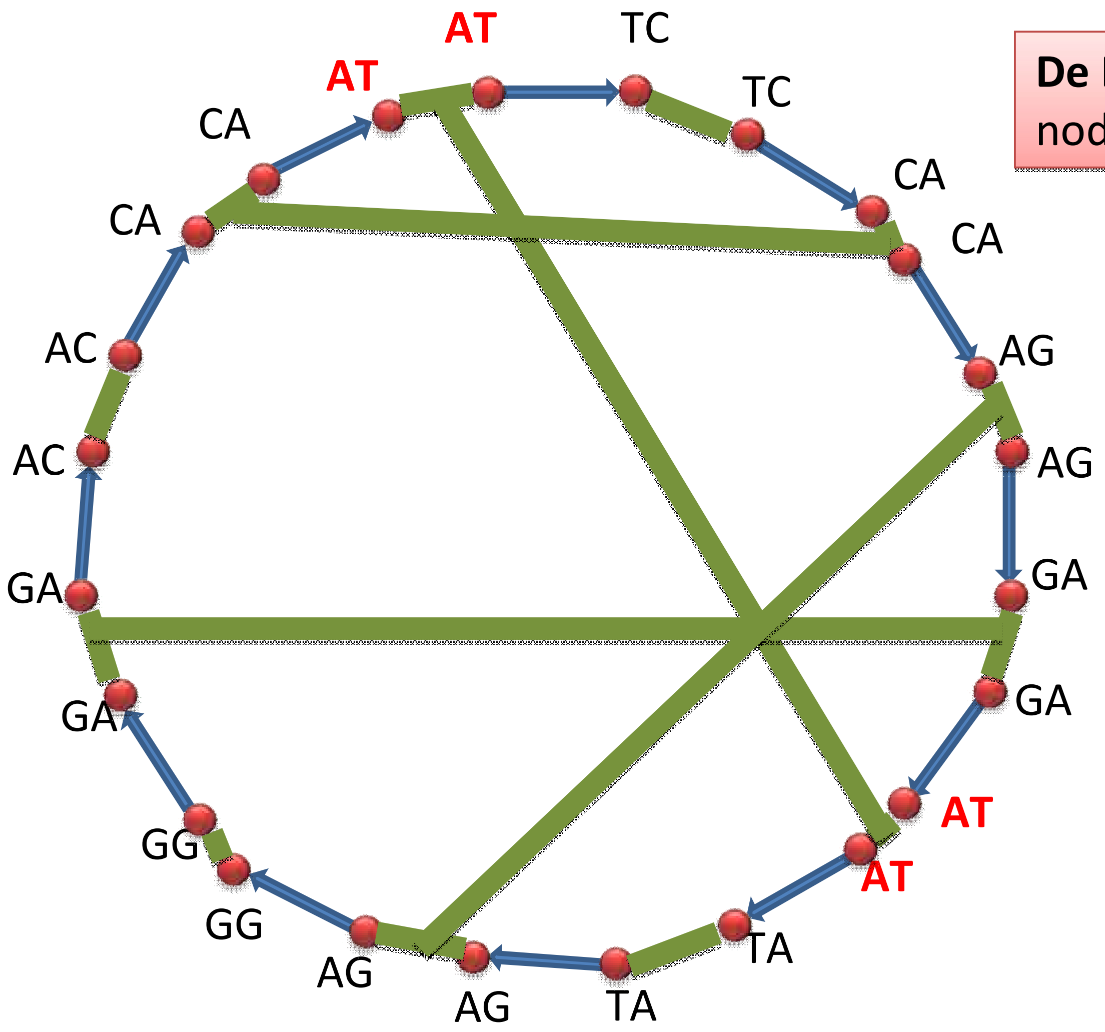
(presented as *A-Bruijn graph*, PP, Tang, Tesler, Genome Res. 2004)

- **De Bruijn graph of a *k*-spectrum:**
  - Represent every *k*-mer as an edge between its prefix and suffix
  - Glue **ALL** nodes with identical labels.

De Bruijn Graphs and Node Gluing

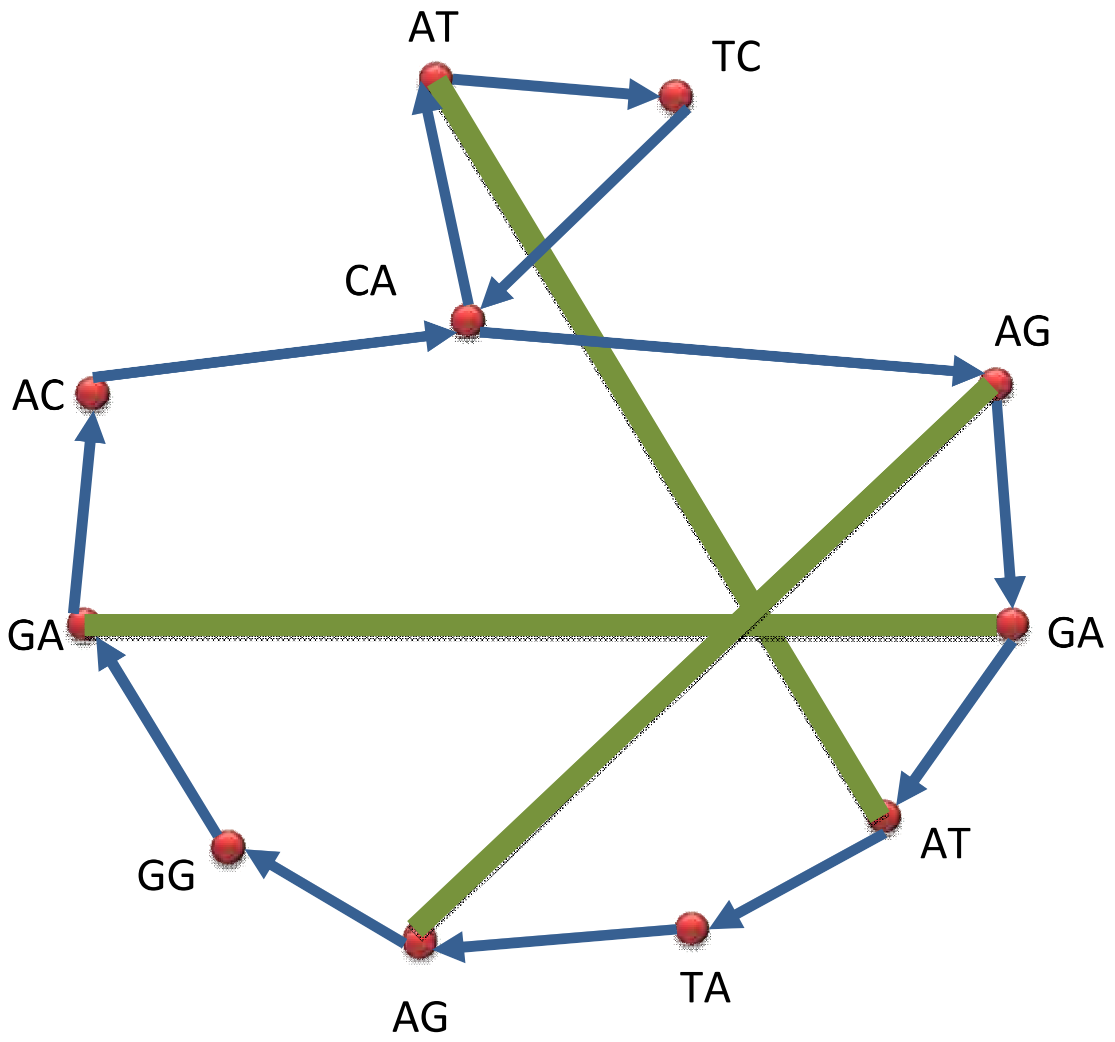


De Bruijn Graphs and Node Gluing

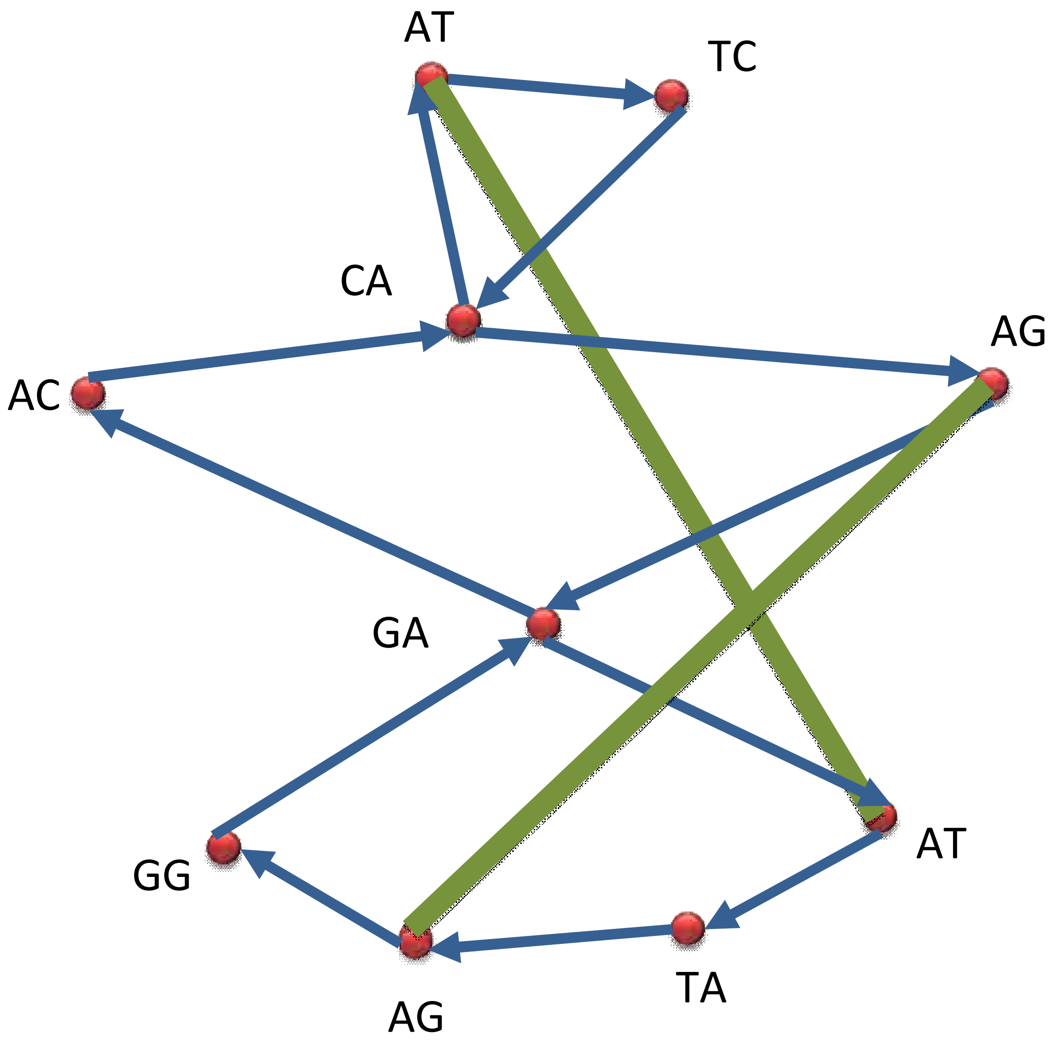


De Bruijn graph = gluing **ALL** nodes with same labels.

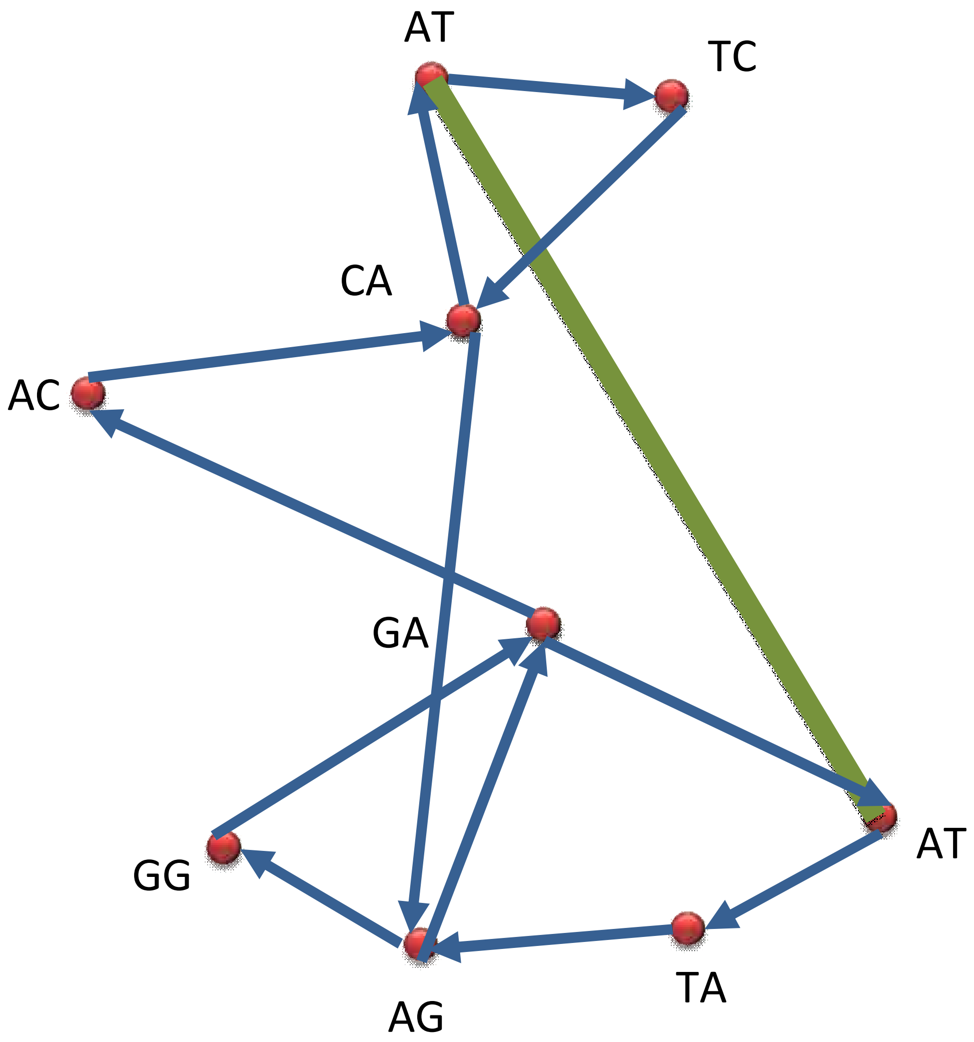
De Bruijn Graph: Gluing in Progress



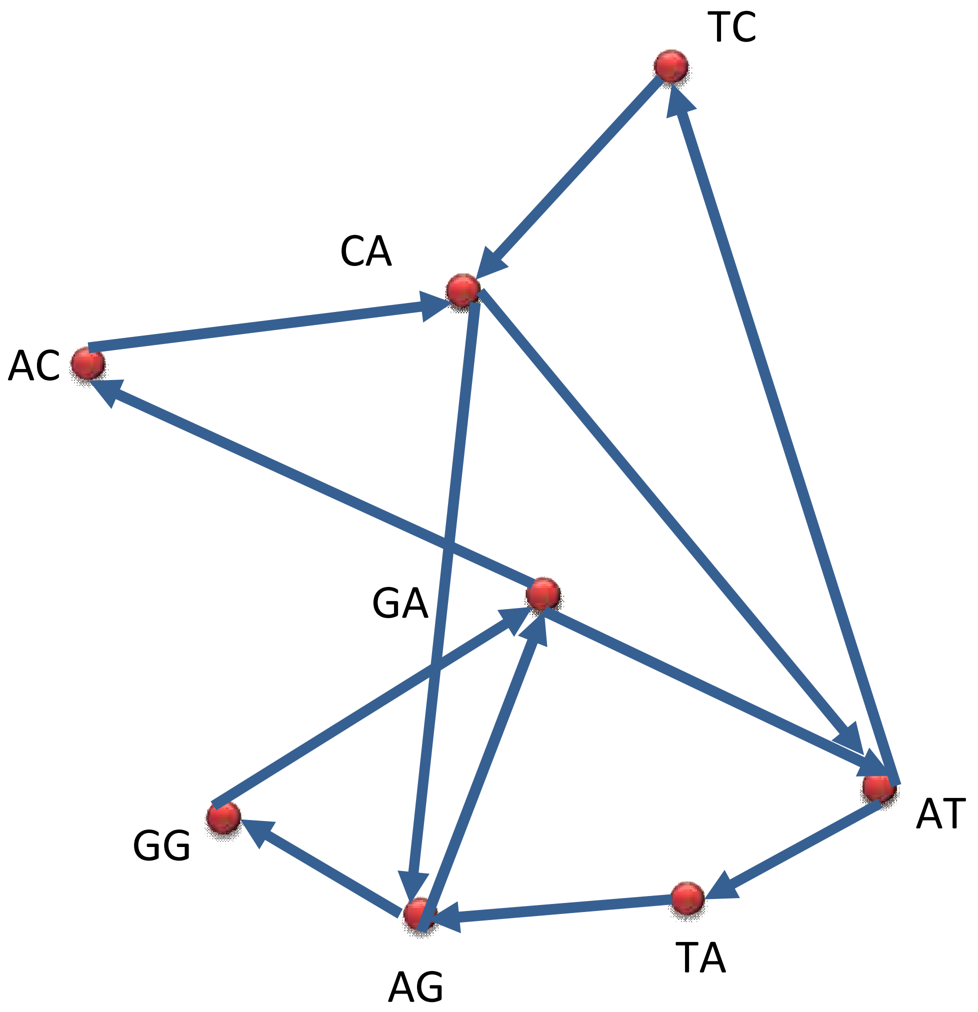
De Bruijn Graph: Gluing in Progress

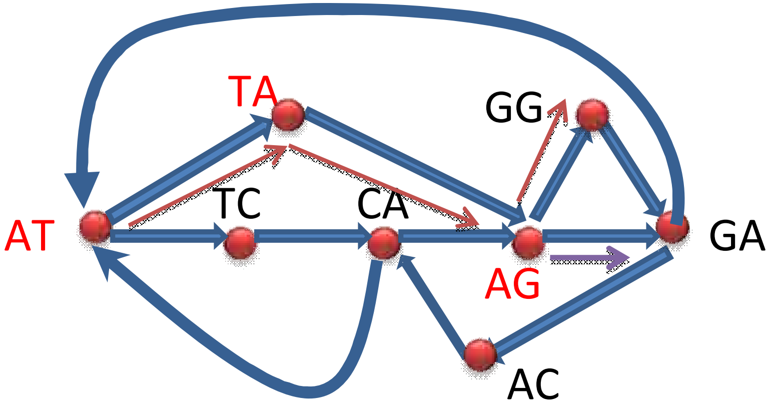


De Bruijn Graph: Gluing in Progress

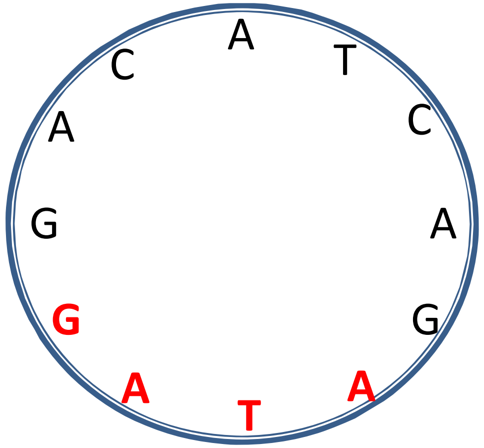


De Bruijn Graph: Gluing in Progress





Genome is an Eulerian cycle in the de Bruijn graph **but we don't know how *Genome* traverses the graph beyond branching vertices.**

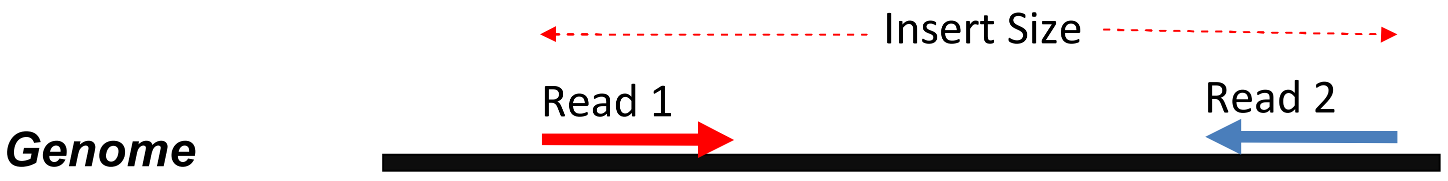


Repeats – A major problem in genome assembly

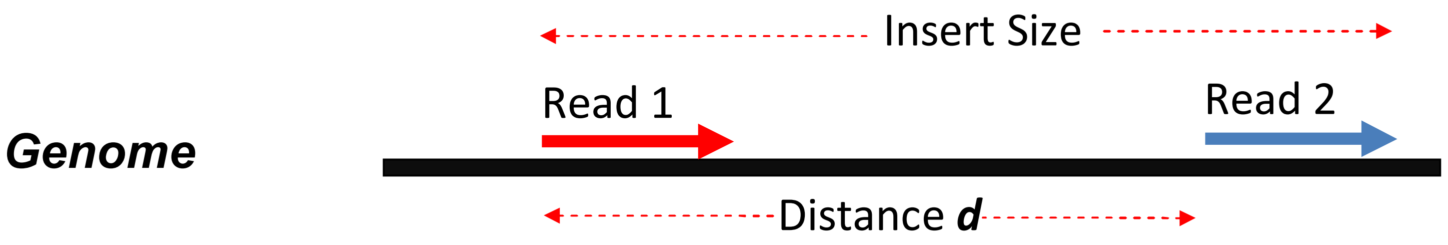
Repeats: A **major** problem for fragment assembly

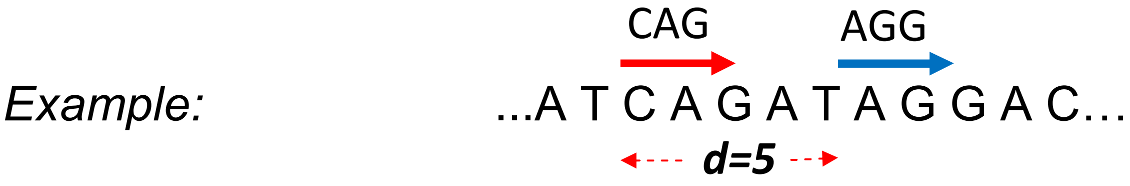
- More than 50% of human genome is repeats
  - over 1 million *Alu* repeats (about 300 bp)
  - about 200,000 LINE repeats (1000 bp and longer)

From Reads to Read-Pairs



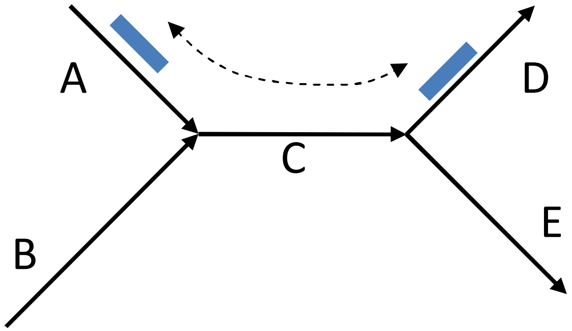
From *k*-mers to paired *k*-mers





A **paired *k*-mer** is a pair of *k*-mers at a **fixed** distance *d* apart in *Genome.*
E.g. **CAG** and **AGG** are at distance *d=5* apart.

Utilizations of Read-Pairs in de Bruijn Assemblers

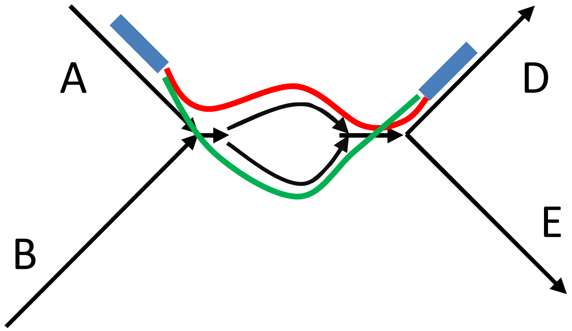


Read-pair transformation (PP and Tang, ISMB 2001)

- Map the read-pairs to the edges of the de Bruijn graph
- Find a unique path between these mapped reads
- The length of this path equal to the insert size.
- Transform the pair of **SHORT** reads into a **LONG** virtual read
- Assemble long virtual reads

VELVET and ALLPATHS describe related approaches to utilize read-pairs.

Utilizations of Read-Pairs in de Bruijn Assemblers



Read-pair transformation (PP and Tang, ISMB 2001)

- Map the read-pair to the edges of the de Bruijn graph
- Find a **unique** path between these mapped reads
- The length of this path equal to the insert length.
- Transform the pair of **SHORT** reads into a **LONG** virtual read
- Assemble long virtual reads

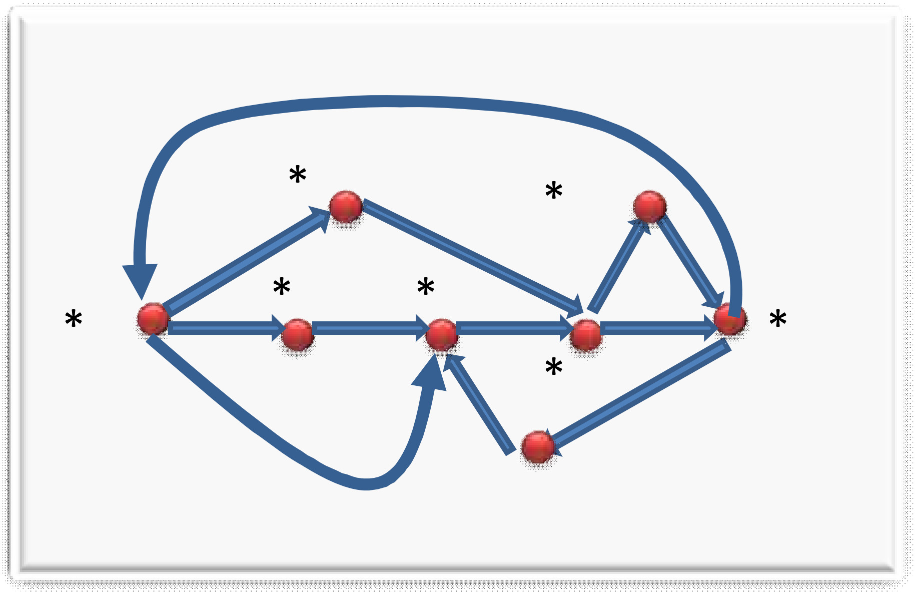
- **Read-pair transformation fails when there exist multiple paths between reads**

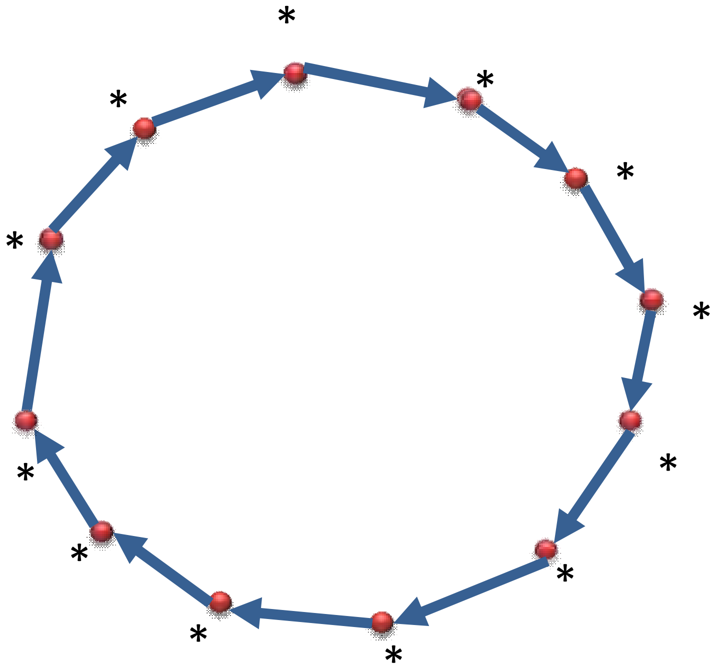
What Would de Bruijn Do?

Read-pair transformation fails when there exist multiple paths between reads

Paired de Bruijn graph

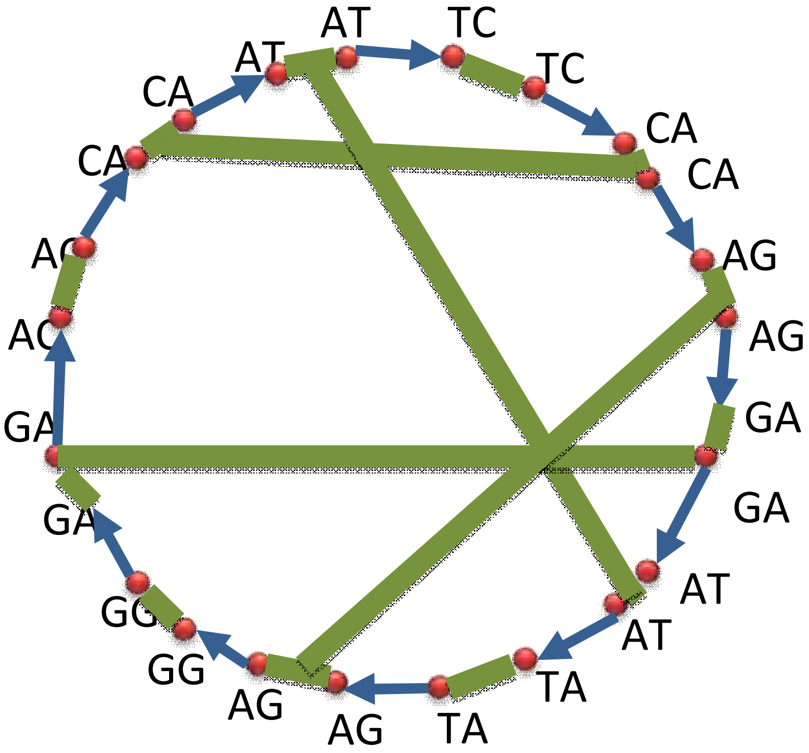
Medvedev et al., J. Comp. Biol. 2011

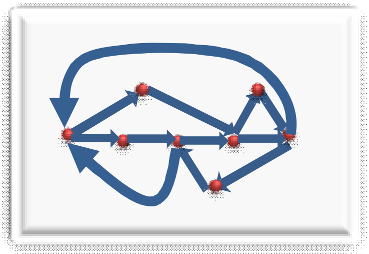


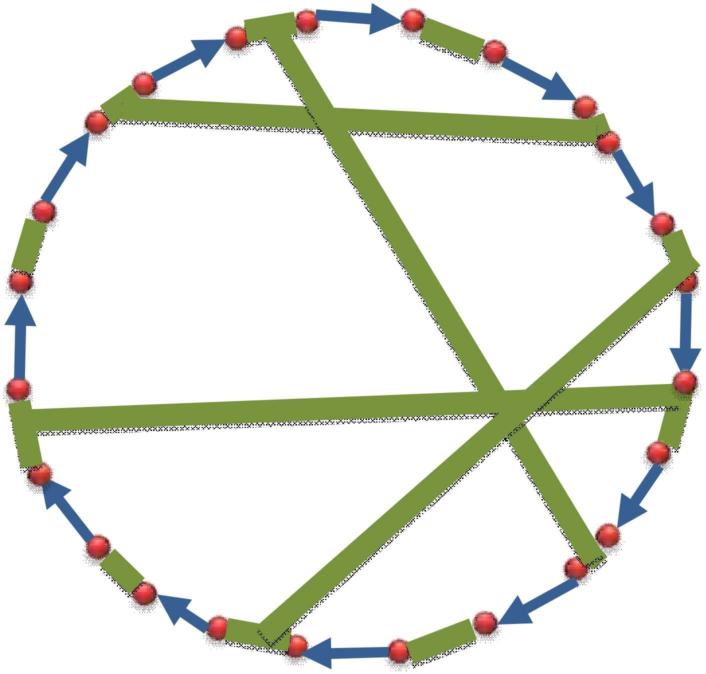


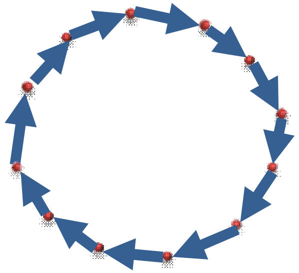
To assemble the original sequence, which graph do we want?

How to get rid of these excessive glues?



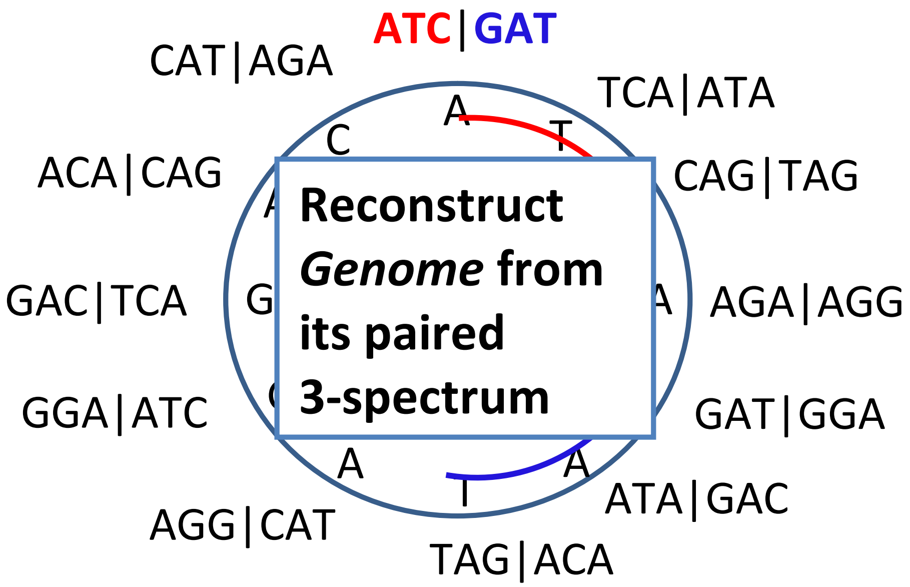






Reconstructing Genome from Paired Spectrum

Generate all paired 3-mers of *Genome (read starts separated by distance 4)*



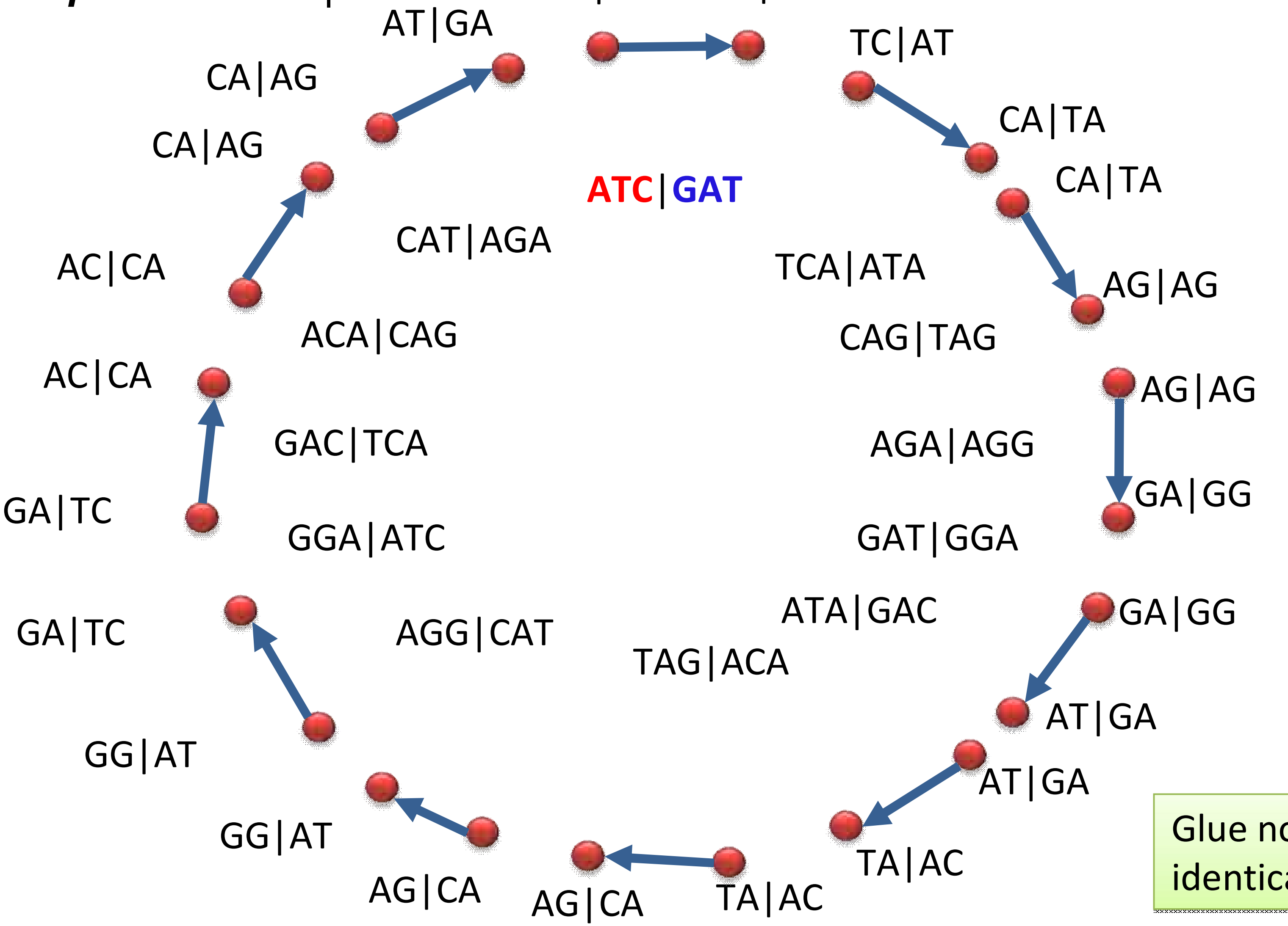
Reconstruct *Genome* from its paired 3-spectrum

A **paired *k-mer*** is a pair of *k*-mers at a fixed distance ***d*** apart in *Genome.*

The **paired *k-spectrum*** of *Genome*: all paired *k-mers* of *Genome (for a fixed distance **d**)*.

Paired de Bruijn Graph

***paired prefix of* ATC|CAT** ⟶ **AT|GA** **TC|AT** ⟵ ***paired suffix of* ATC|CAT**

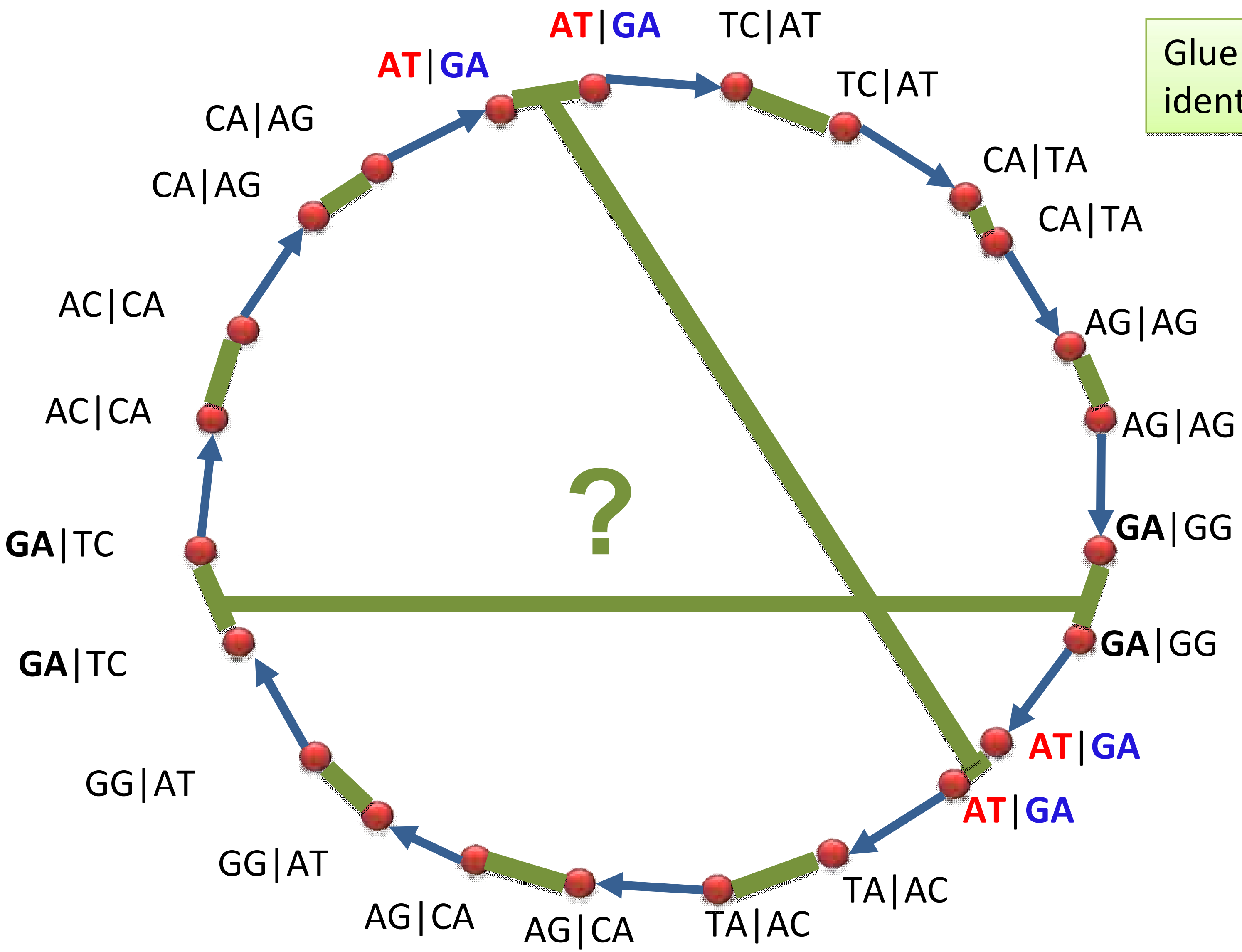


Glue nodes with identical labels!

Paired de Bruijn Graph

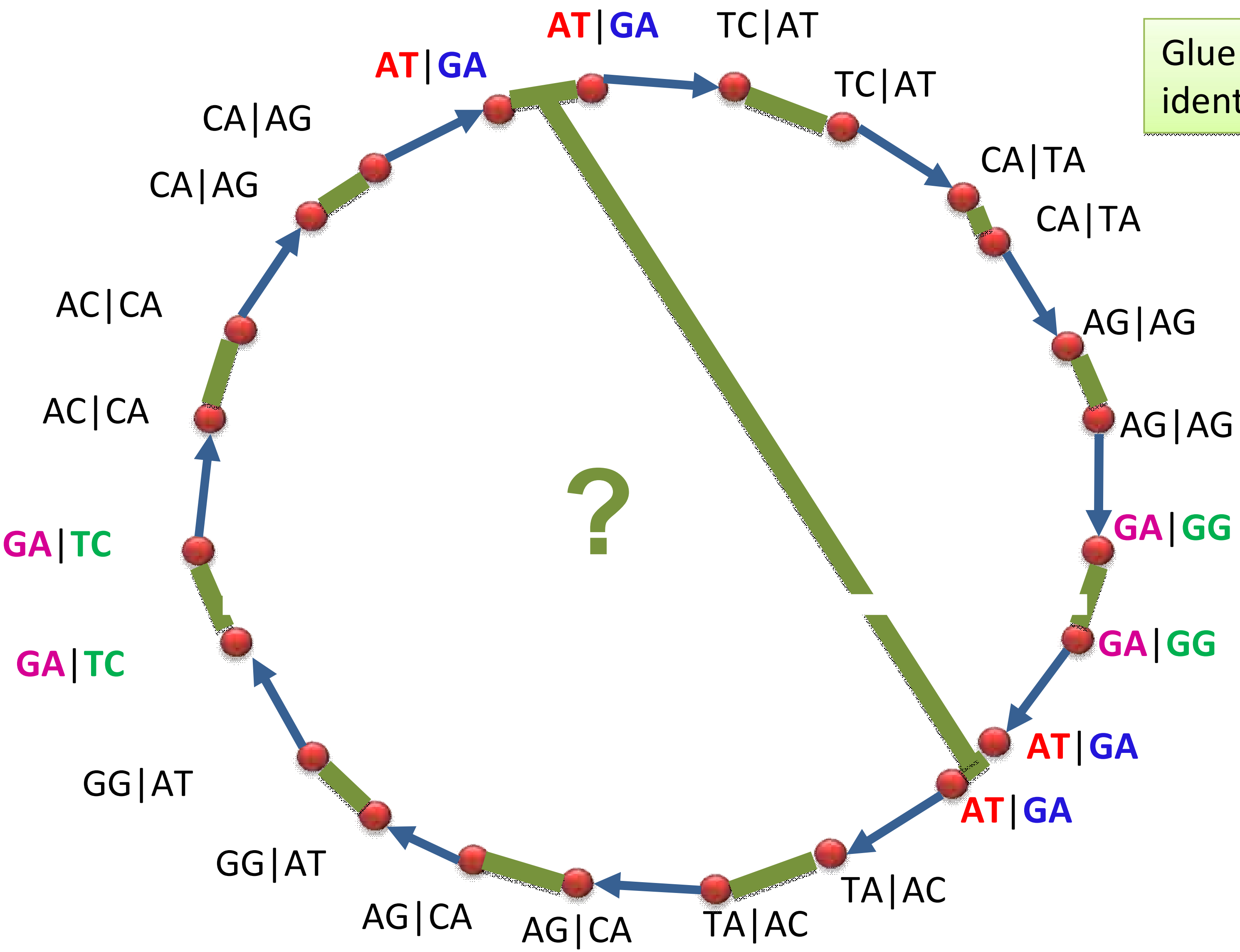
- **Paired de Bruijn graph of a paired *k*-spectrum:**
  - Represent every paired *k*-mer as an edge between its paired prefix and paired suffix.
  - Glue **ALL** nodes with identical labels.

Paired de Bruijn Graph: Gluing in Progress



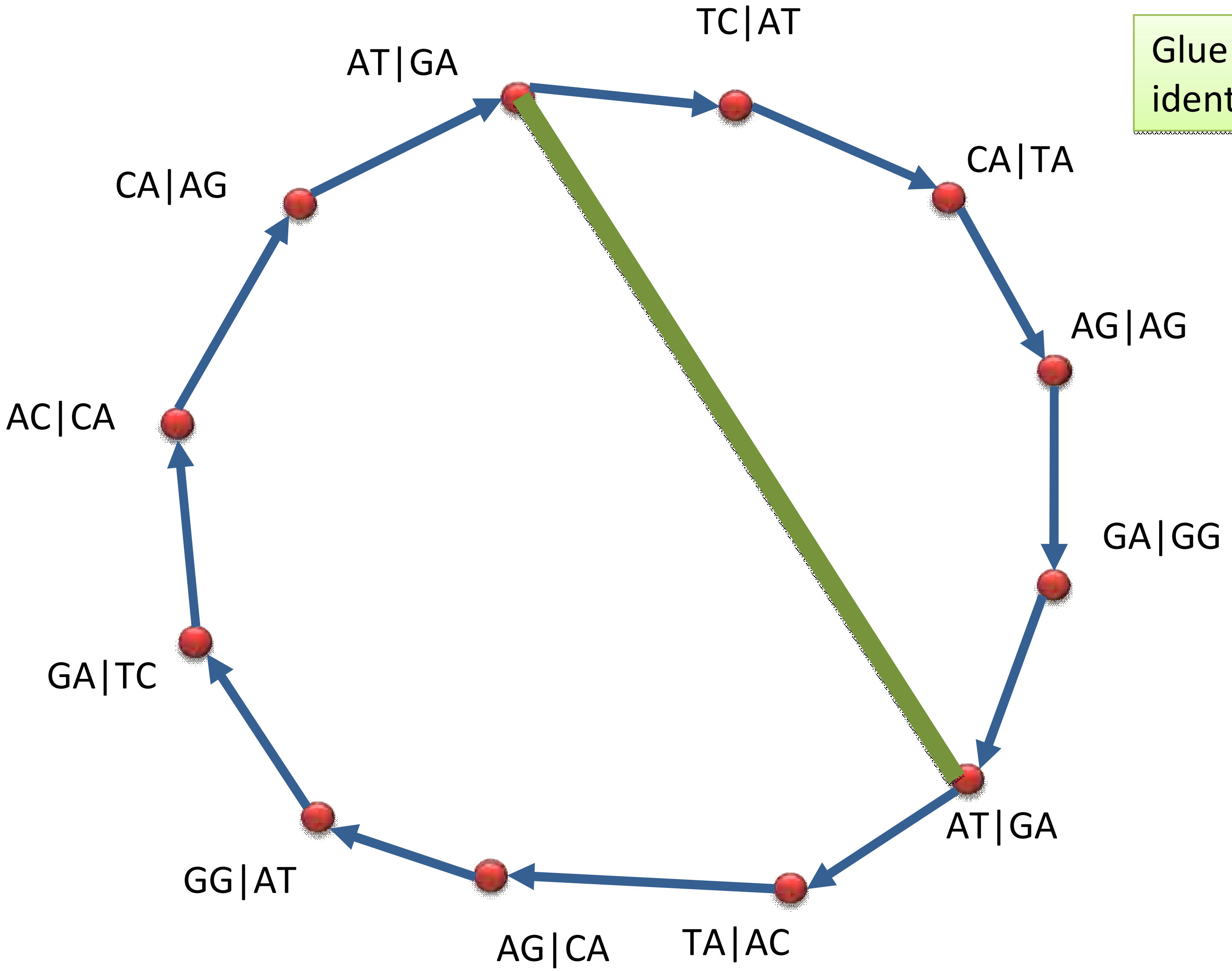
Glue nodes with identical labels!

Paired de Bruijn Graph: Gluing in Progress



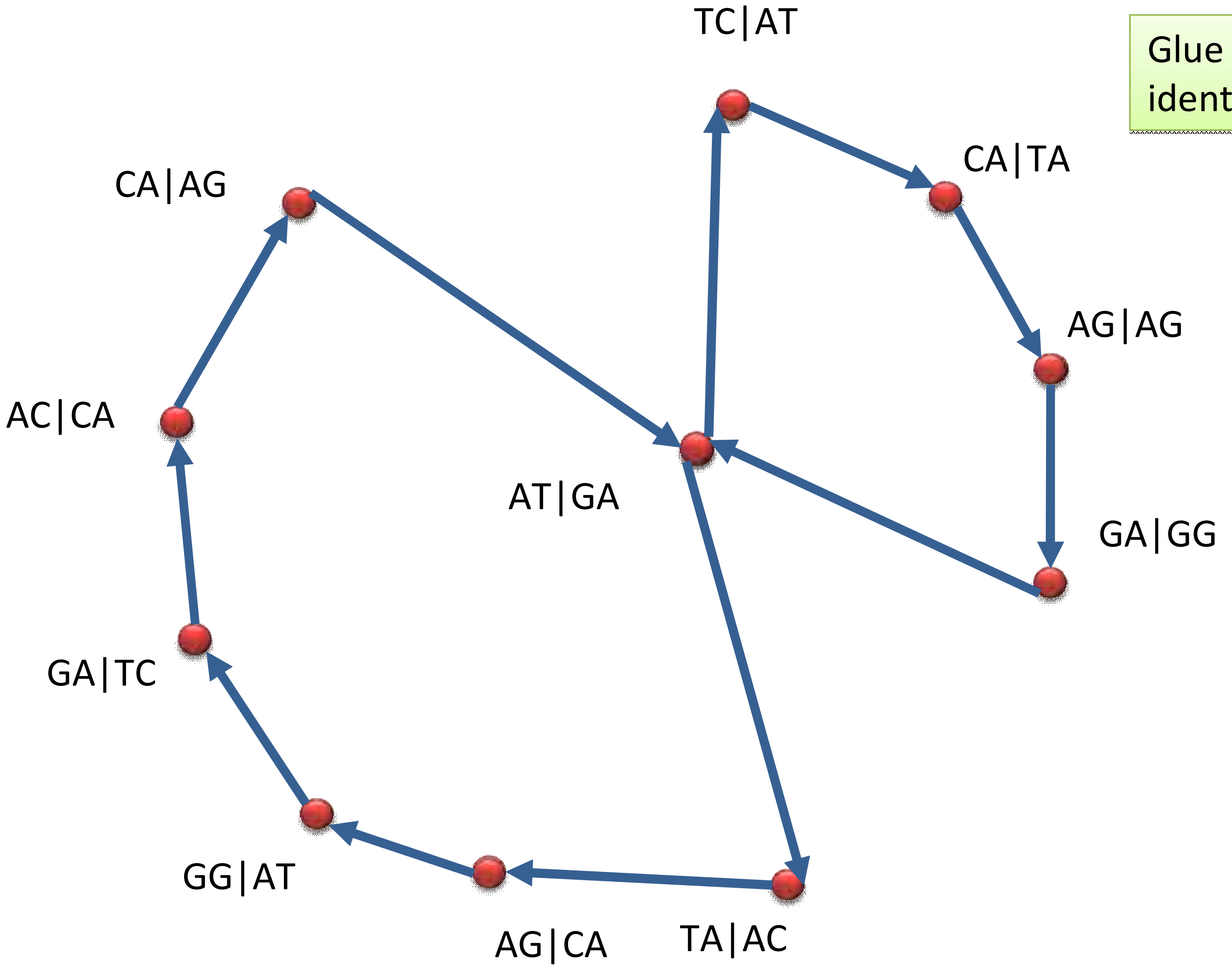
Glue nodes with identical labels!

Paired de Bruijn Graph



Glue nodes with identical labels!

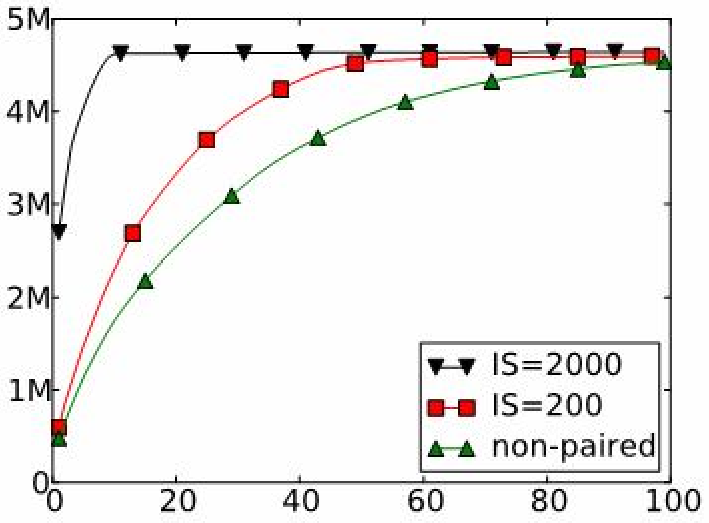
Paired de Bruijn Graph



Glue nodes with identical labels!

Cumulative Contig Length: EXACT insert size

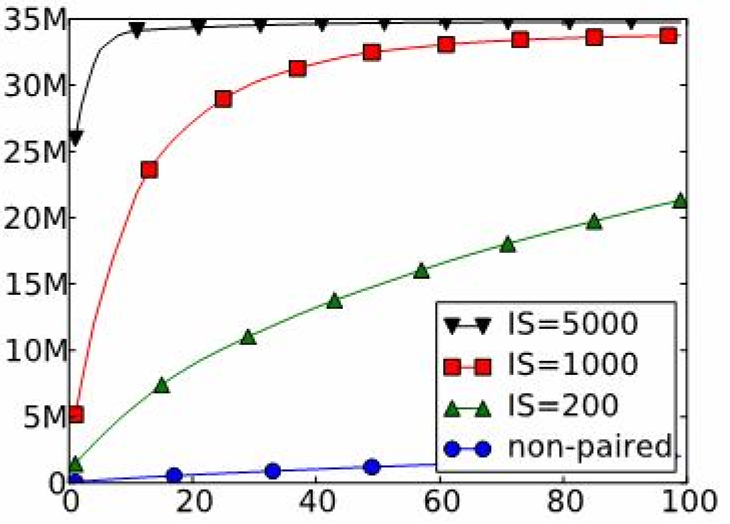
E. Coli



Cumulative length of m longest contigs

m contigs

Human (Chr. 22)



m contigs

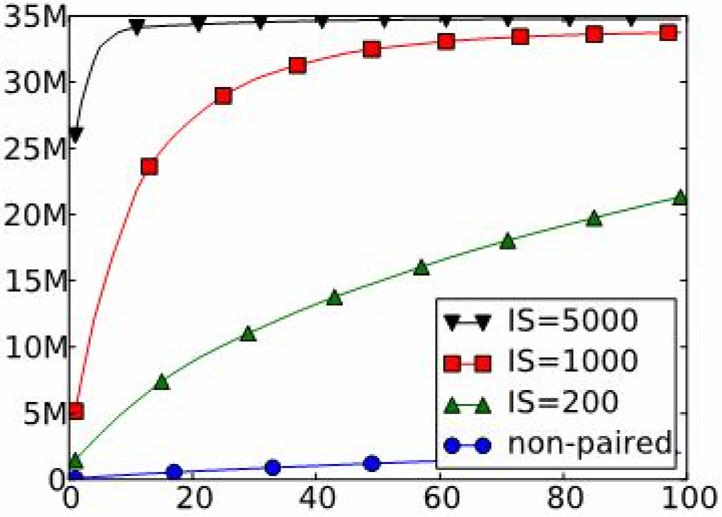
k-mer size = 50

For EXACT distance *d=1000* (let alone *5000)*, the PDBG approach generates an excellent assembly of human genome even with very short reads (*k=50*).

Cumulative Contig Length:
EXACT Distance between Reads

Human

Cumulative length of m longest contigs

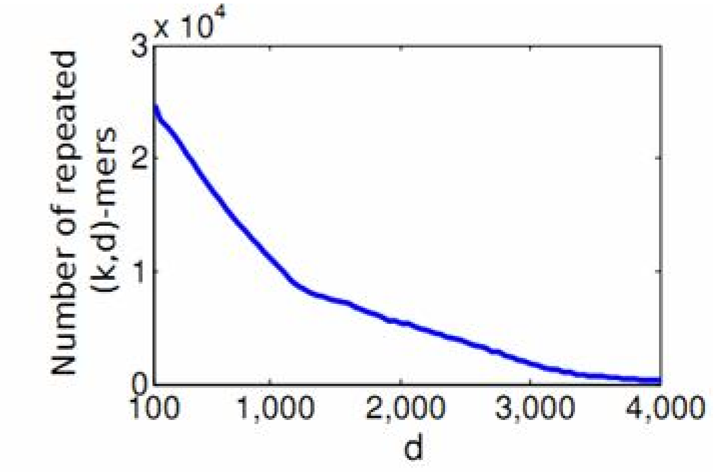


m contigs

k-mer size = 50

For EXACT distance *d=1000* (let alone *5000*), the PDBG approach generates an excellent assembly even with very short reads (*k=50*).

Number of repeated paired *k*-mers drops as distance *d* increases

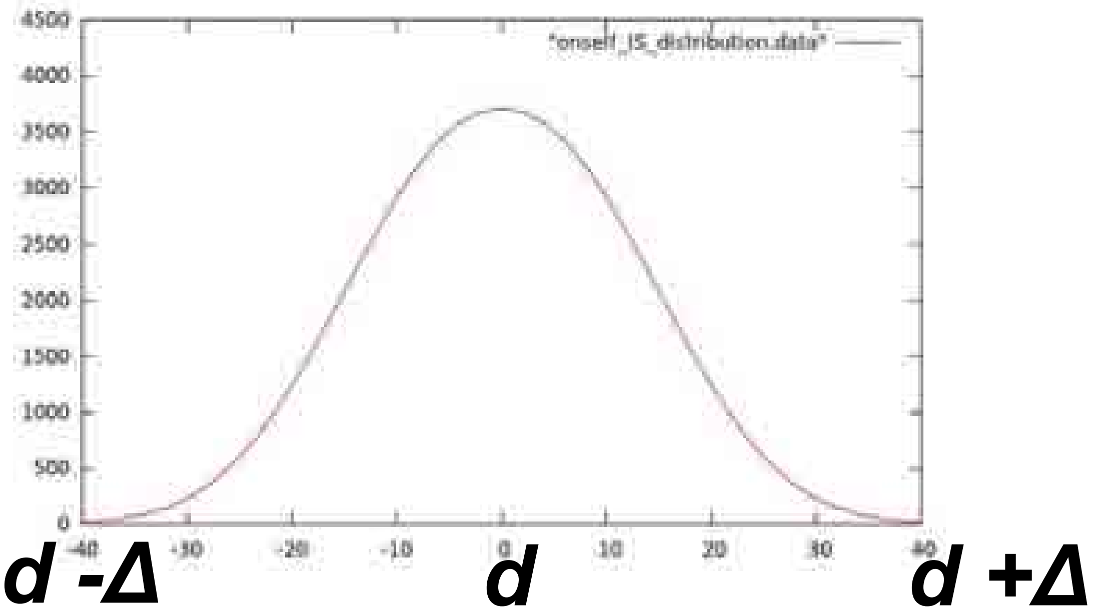


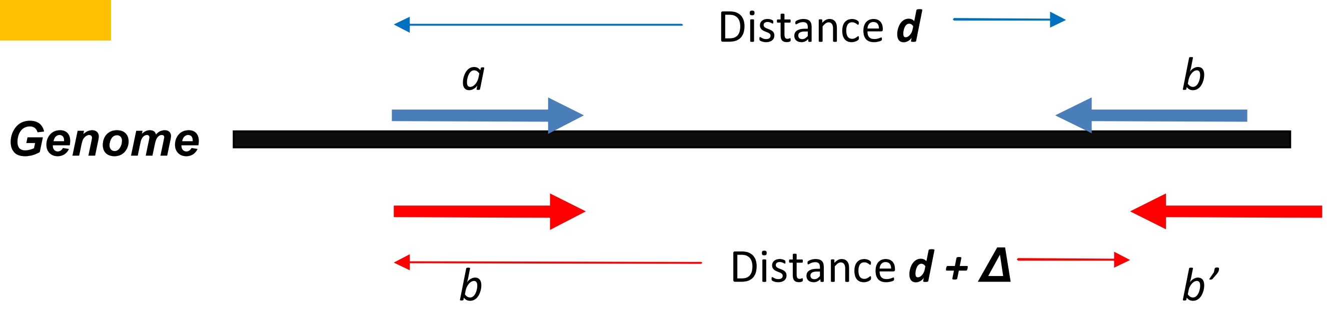
Number of repeated paired *k*-mers for *k=50* and varying distance *d*

For distance *d=4000,* from the perspective of paired 50-mers, the *E. coli* genome has no repeats. Assembly becomes trivial!

Back to Reality:
Distances between Reads are INEXACT

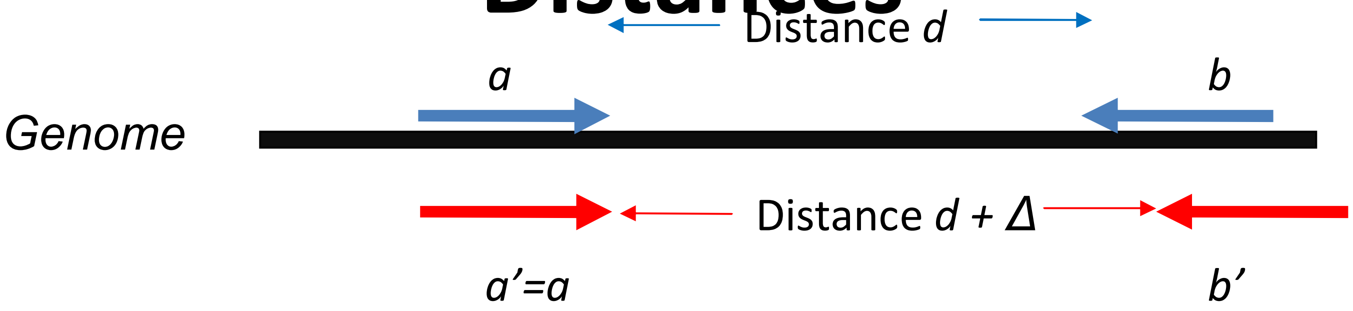
The distances between reads lie within some range ***d ±Δ.***





Constructing Paired de Bruijn Graph from Paired *k*-mers with Approximate Distances

Note: "Distance" in these slides is different from the one described in the paper



If paired *k*-mers *(a|b)* and *(a'|b')* are generated at the same position of the genome:

- *a = a'*
- *b* and *b'* should be at distance at most *2Δ* apart in *Genome*

But the *Genome* is unknown!

If *b* and *b'* are at distance at most *2Δ* in *Genome*, they should be at distance at most *2Δ* in the de Bruijn graph!

- Paired *k*-mers *(a|b), (a'|b'*) are **SIMILAR** if:
  1. *a = a'*
  2. *b* and *b'* are at distance at most *2Δ* in the de Bruijn graph.

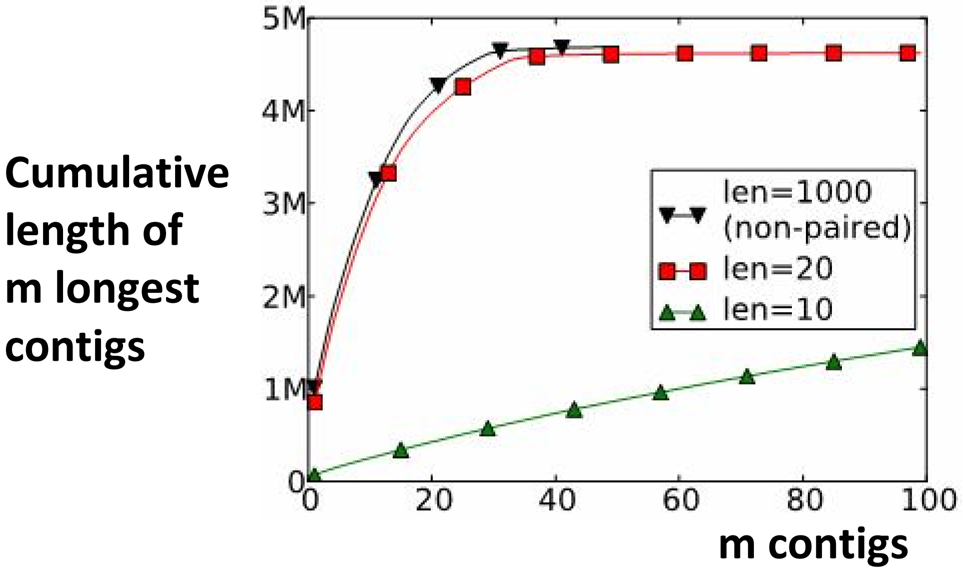
Approximate Paired de Bruijn Graph

- **Approximate Paired de Bruijn graph of a paired *k*-spectrum:**
  - Represent every paired *k*-mer as an edge between its paired prefix and paired suffix:
  - Glue **ALL** nodes with ***SIMILAR*** labels.

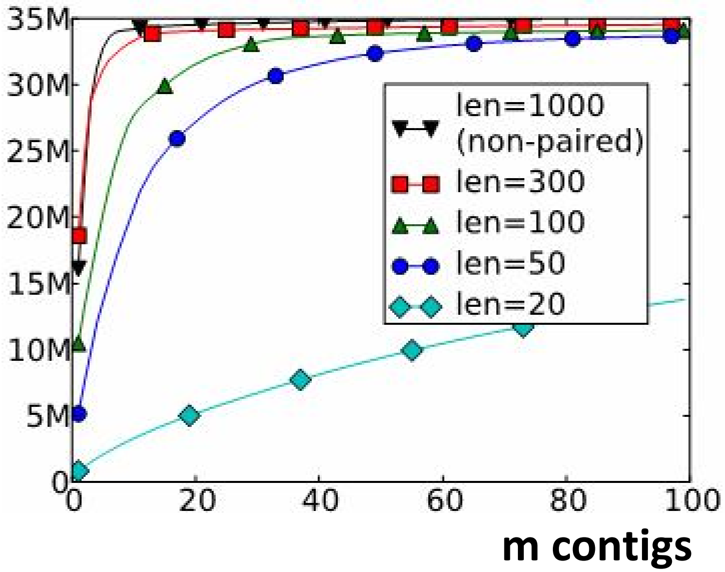
The notion of "SIMILAR" is defined in Medvedev et al., 2011

Cumulative Contig Length (fixed insert size, varying *k*-mer size)

E. Coli



Human (Chr. 22)

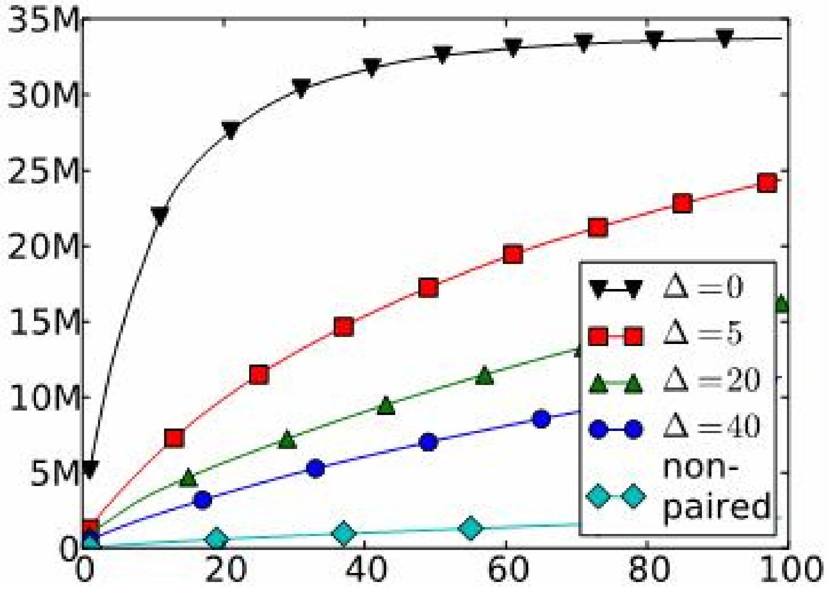


Insert size = 1000

Cumulative Contig Length: INEXACT Distance (with error *Δ*)

Human

Cumulative length of m longest contigs



Insert size = 1000, *k* = 50

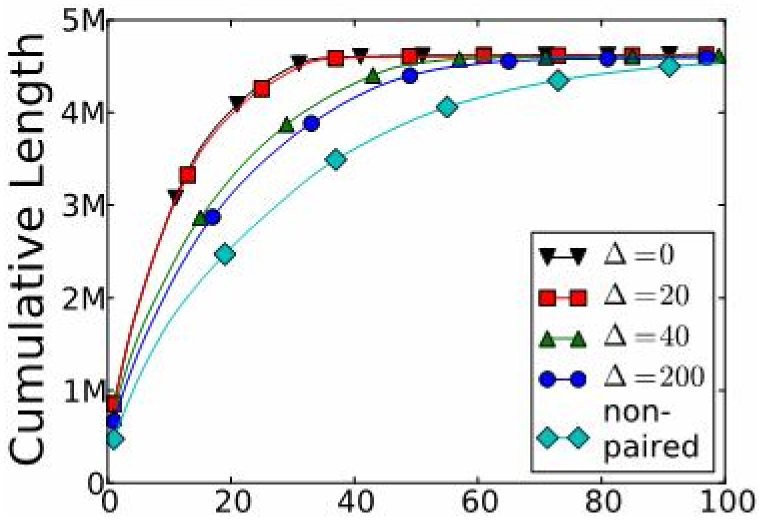
For INEXACT distance, the assembly quickly deteriorates even for small distance error, e.g., *Δ=20*

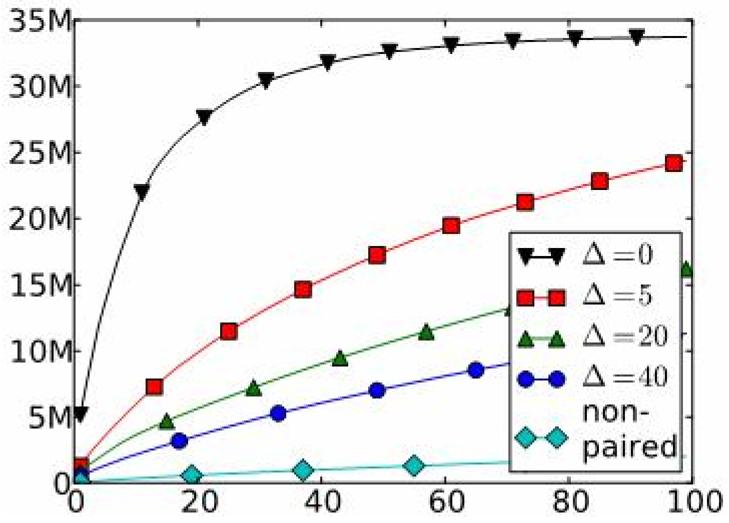
Cumulative Contig Length: INEXACT insert size (with error *Δ*)

E. Coli

Human (Chr. 22)

Cumulative length of m longest contigs





Insert size = 1000, k = 50

For INEXACT distance *d*, the assembly deteriorates even for small distance error, e.g., *Δ=20*

The Key Deficiency of Paired de Bruijn Graphs

Medvedev et al., 2011: assembly of paired *k*-mers using **Paired de Bruijn graphs (PDBG).** Finally, an elegant approach to assembling paired *k*-mers BUT …

PDBGs only work when the EXACT (or nearly exact) distances between reads within read-pairs are known.

Deja vu from 2001

- **Paired de Bruijn graphs** are impractical since distances are imprecise

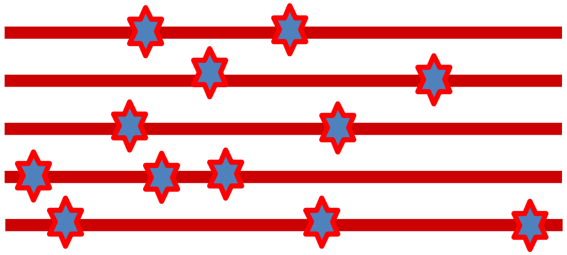
- But in 1995 **de Bruijn graphs** were not very practical either! At least for Sanger reads circa 1995…

Historic Reference

- De Bruijn assembly works when nearly every *k*-mer from genome appears in at least one read without errors
- **Thus, de Bruijn assembly requires either nearly error-free reads or high coverage.**
- **Neither condition held in 1995** when Idury and Waterman proposed de Bruijn assembly for Sanger reads: only ≈13% of 50-mers were correct!
- **Error-correction** (PP, Tang, Waterman, PNAS 2001) made reads nearly error-free (over 90% of 50-mers became correct) and made de Bruijn assembly practical even in low coverage Sanger projects

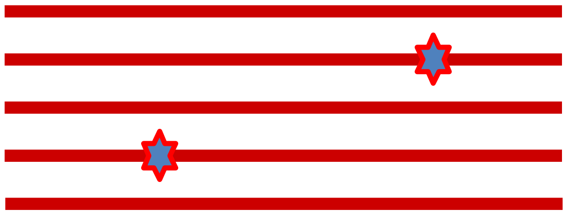
If reads were made nearly error-free in 2001, can we make distances between reads nearly exact in 2012?

Error Correction (2001)

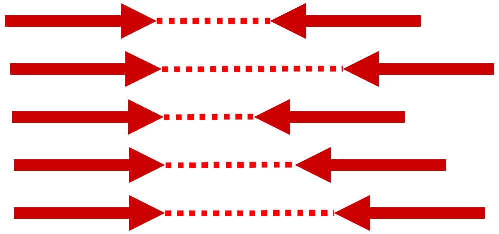


Error-prone reads

PP, Tang, Waterman PNAS 2001

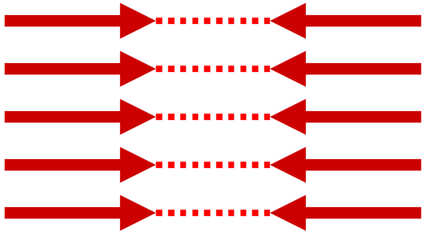


Read-Pair Adjustment (2012)

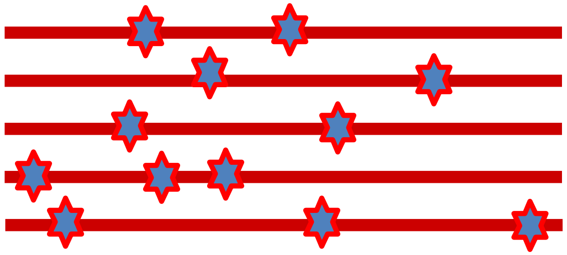


Read-pairs with variable insert sizes

Bankevich et al. JCB 2012 (in press)

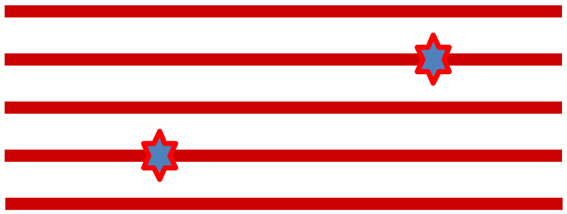


Error Correction



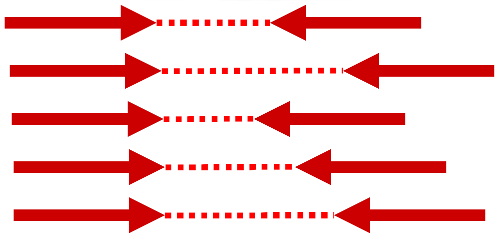
Error-prone reads

PP, Tang, Waterman PNAS 2001



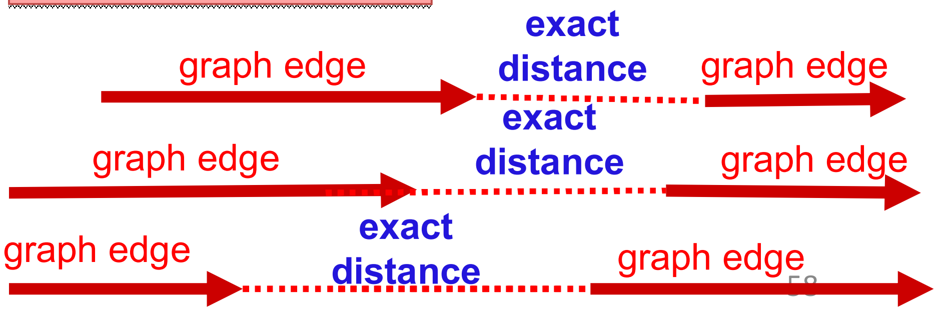
Read-Pair Adjustment

This sequencing machine produces **edge-pairs** instead of **read-pairs**

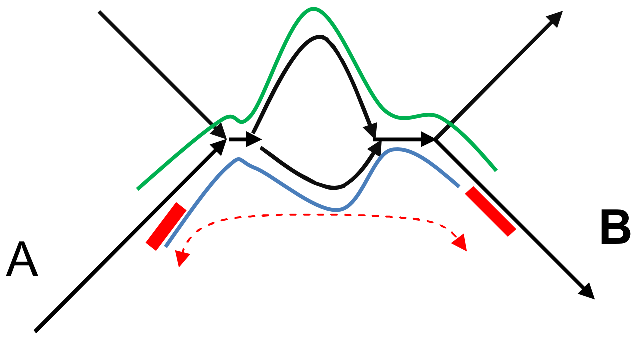


Read-pairs with variable insert sizes

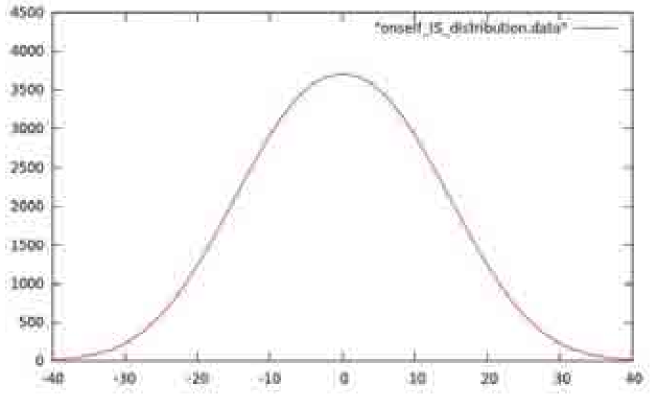
Bankevich et al. JCB 2012 (in press)



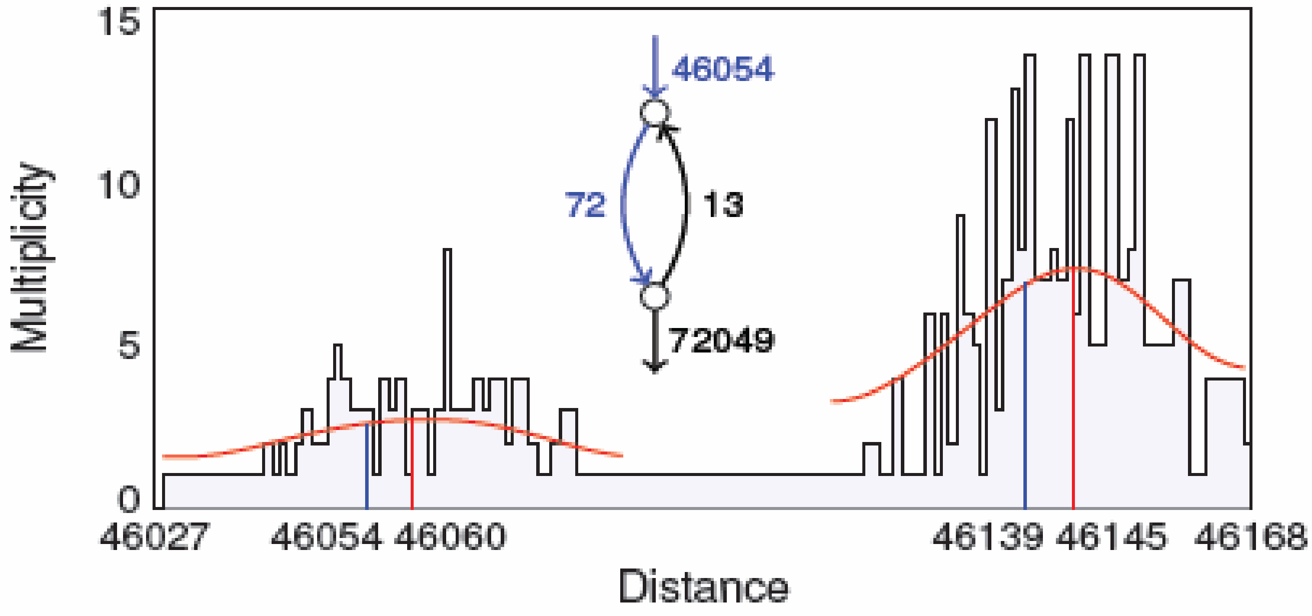
What is the Correct Genomic Path between Edges A and B?



Variation in insert size



Is the correct path between red reads
short (passing through lower edge)
or **long** (passing through upper edge)?



The genomic distance between edges A and B can be estimated when they are linked by a read-pair

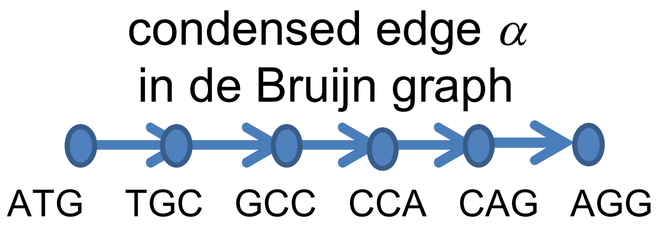
A single read-pair provides an unreliable distance estimate

But many read-pairs accurately estimate the distance and vote for the blue path

While original read-pairs have large errors in distance estimates (e.g. 210 ± 40 bp), nearly 100% of edge-pairs feature exact distances after distance adjustment by SPAdes

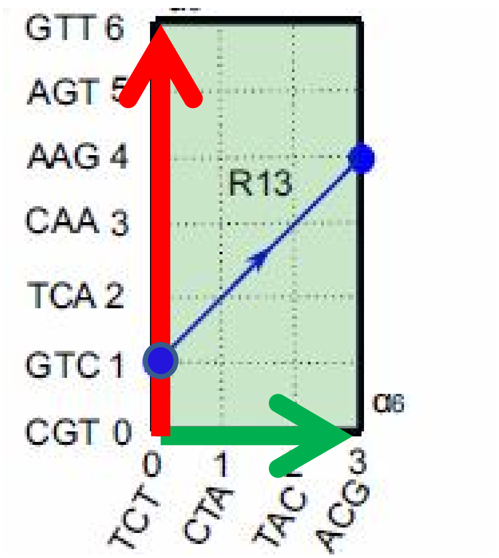
Representing Edge-Pairs as Rectangles

- An edge-pair formed by (condensed) edges α and β at the estimated distance ***D*** in the de Bruijn graph forms a **rectangle** *(α/β,**D**)* of size ***|** α **|· |** β **|***


condensed edge α
in de Bruijn graph

- Every integer point within rectangle projects into *k*-mers on **green** and **red** sides

- The *k*-mers separated by distance *d* (fixed average distance between reads) form a **45 degree blue line** in the rectangle



Blue line starts in (TCT|GTC)
and ends in (ACG|AAG)

Generating Rectangles

- A ***Genome*** (with repeats **P1** and **P2**) is spelled as:

  P3, **P1**, P6**, P2**, P4, **P1**, P5, **P2**

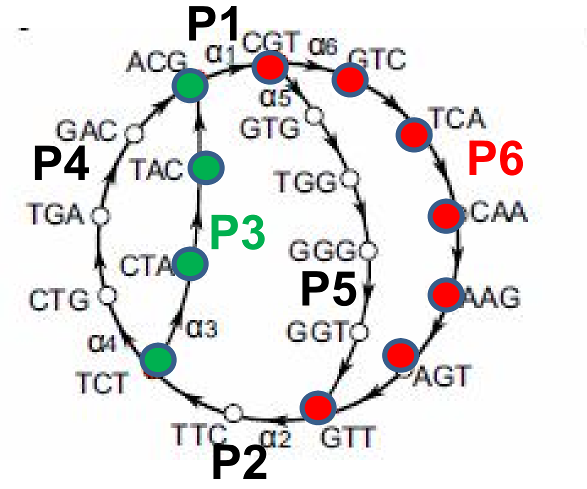
- P3=TCTACG
  P6=CGTCAAGTT

- Green and red edges P3 and P6 are distance ***D=4*** apart resulting in a **rectangle** (P3|P6,4).

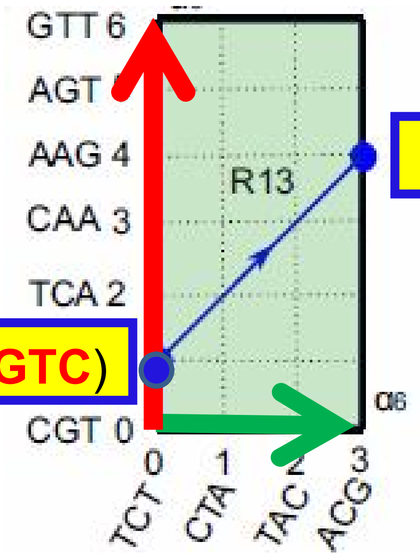
- The **blue 45 degree line** within rectangle reveals all *k*-mers separated by the default distance ***d=5***, e.g., TCT and GTC.

- The **blue line** connects **starting** (TCT|GTC) **with ending** (ACG|AAG) blue points of the rectangle

De Bruijn graph with 6 condensed edges P1,…, P6



P1: ACGT
P2: GTTCT
P3: TCTACG
P4: TCTGACG
P5: CGTGGGTT
P6: CGTCAAGTT



(ACG|AAG
)

ending blue point

(TCT|GTC)

starting blue point

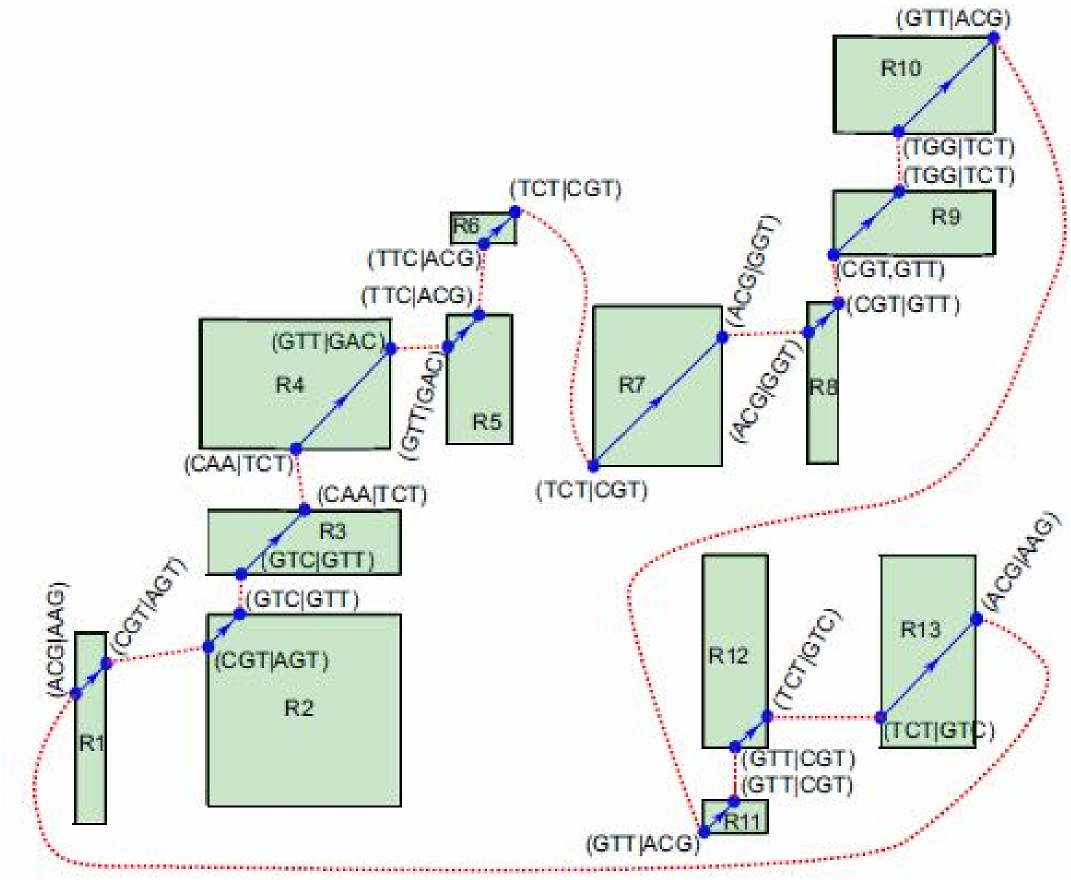
Rectangle Graph
(yet another A-Bruijn graph)

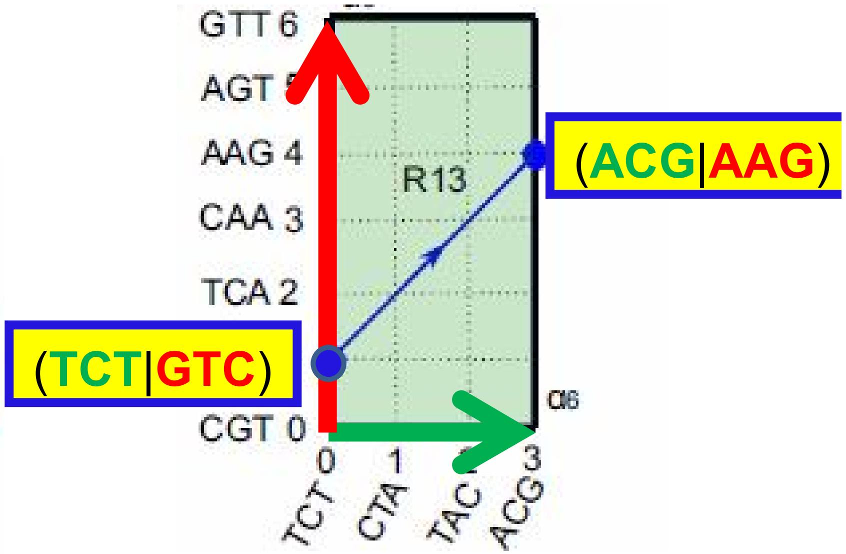
- **Rectangle graph on edge-pairs:**
  - Represent every edge-pair formed by edges α and β at distance ***D*** as a **blue edge** within a rectangle *($\alpha|\beta$,**D**).* The blue edge connects its **starting** and **ending** nodes labeled as:

$$\mathbf{start_d(\alpha|\beta,D) \rightarrow end_d(\alpha|\beta,D)}$$

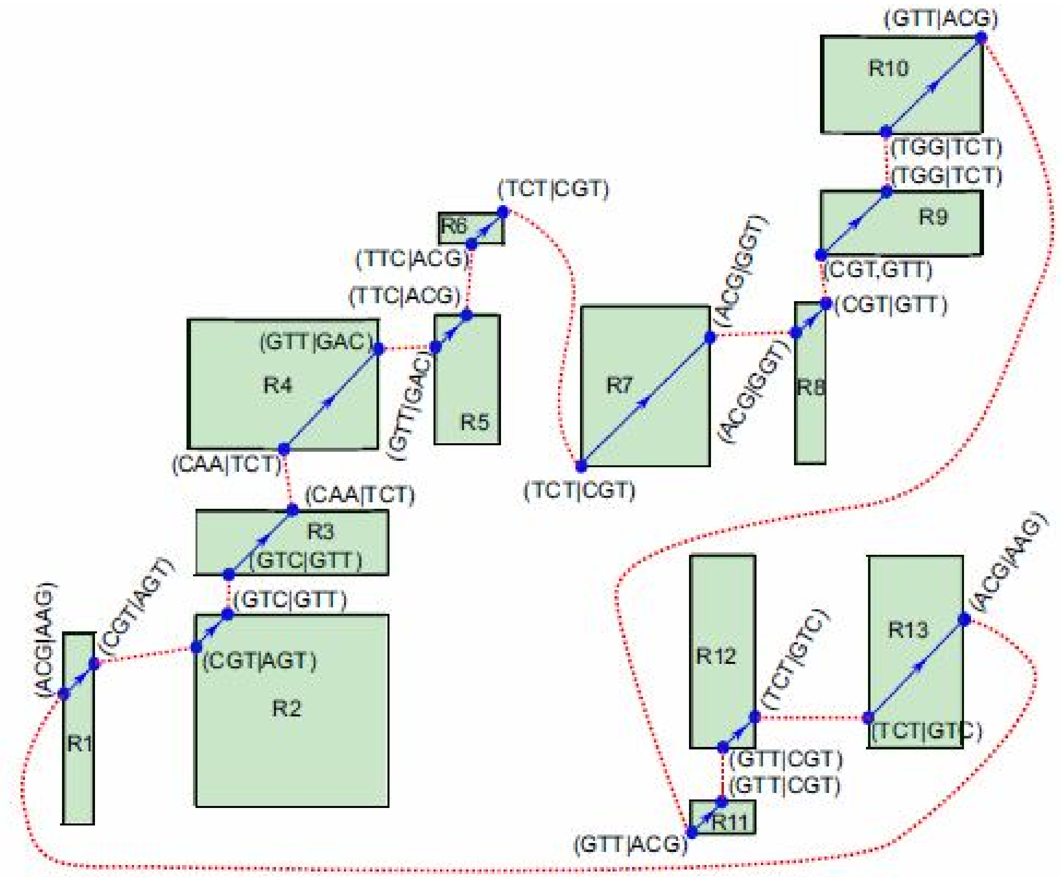
  - Glue **ALL** nodes with identical labels

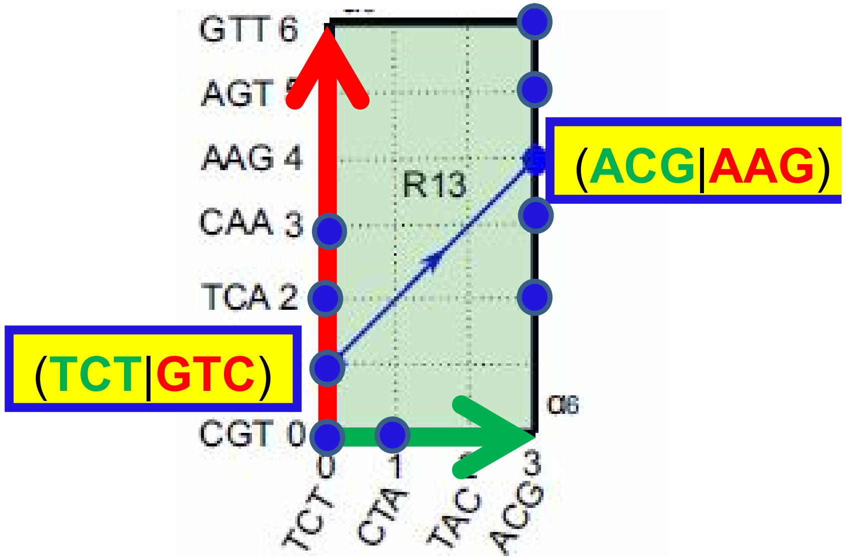
Rectangle Graph Assembly





What if Distance Estimates Are (Slightly) Imprecise?





Benchmarking SPAdes:
87% of *E. coli* genes fully captured from single cell data

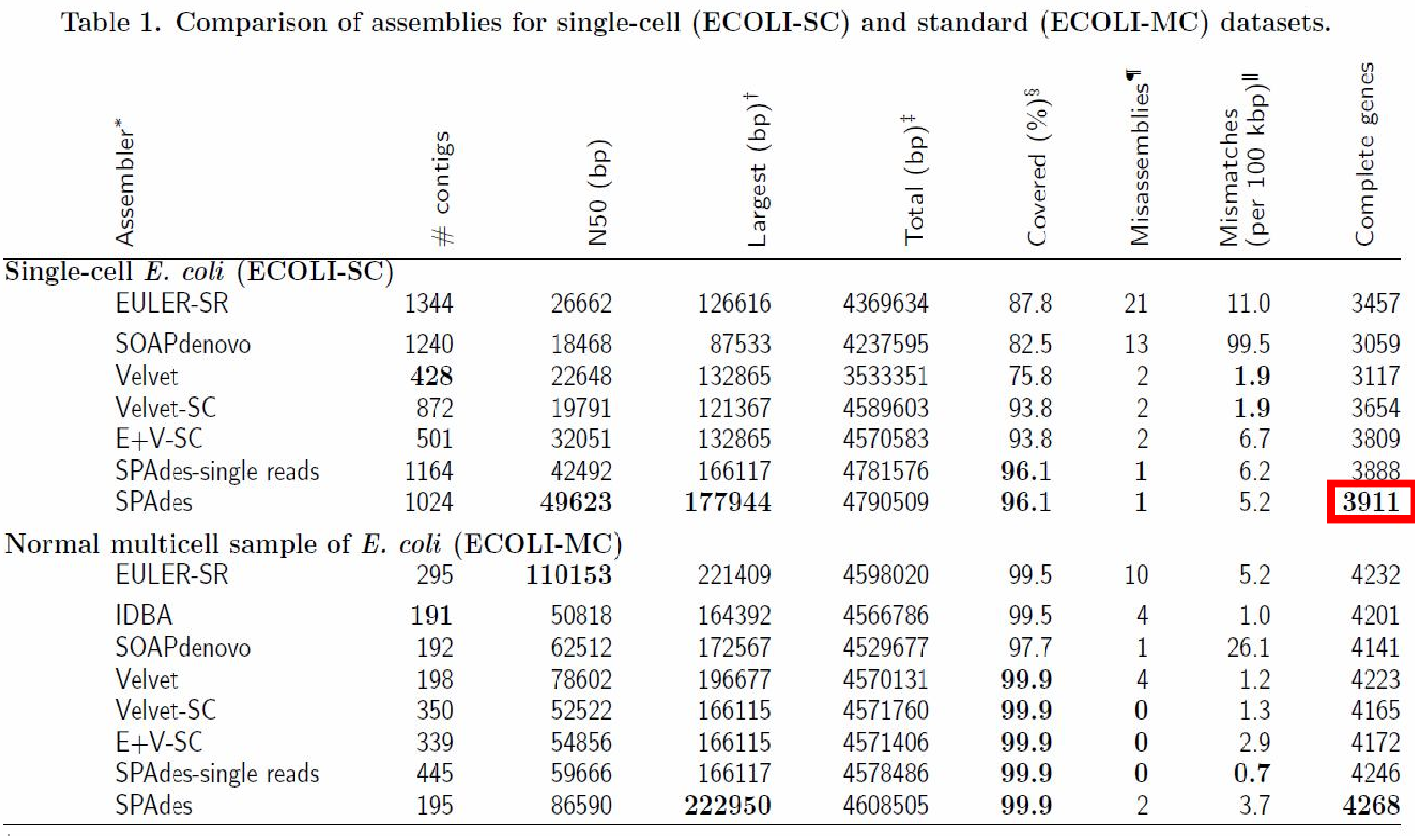


Table 1. Comparison of assemblies for single-cell (ECOLI-SC) and standard (ECOLI-MC) datasets.

Assembler*	# contigs	N50 (bp)	Largest (bp)†	Total (bp)‡	Covered (%)§	Misassemblies¶	Mismatches (per 100 kbp)‖	Complete genes
Single-cell *E. coli* (ECOLI-SC)								
EULER-SR	1344	26662	126616	4369634	87.8	21	11.0	3457
SOAPdenovo	1240	18468	87533	4237595	82.5	13	99.5	3059
Velvet	**428**	22648	132865	3533351	75.8	2	**1.9**	3117
Velvet-SC	872	19791	121367	4589603	93.8	2	**1.9**	3654
E+V-SC	501	32051	132865	4570583	93.8	2	6.7	3809
SPAdes-single reads	1164	42492	166117	4781576	**96.1**	**1**	6.2	3888
SPAdes	1024	**49623**	**177944**	4790509	**96.1**	**1**	5.2	**3911**
Normal multicell sample of *E. coli* (ECOLI-MC)								
EULER-SR	295	**110153**	221409	4598020	99.5	10	5.2	4232
IDBA	**191**	50818	164392	4566786	99.5	4	1.0	4201
SOAPdenovo	192	62512	172567	4529677	97.7	1	26.1	4141
Velvet	198	78602	196677	4570131	**99.9**	4	1.2	4223
Velvet-SC	350	52522	166115	4571760	**99.9**	**0**	1.3	4165
E+V-SC	339	54856	166115	4571406	**99.9**	**0**	2.9	4172
SPAdes-single reads	445	59666	166117	4578486	**99.9**	**0**	**0.7**	4246
SPAdes	195	86590	**222950**	4608505	**99.9**	2	3.7	**4268**

Bankevich et al., *J. Comp. Biol.*, 2012

Ongoing SPAdes Collaborations

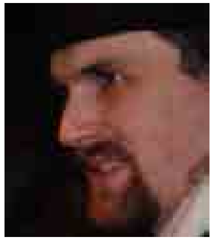
- Sequencing uncultivated bacteria representing gray matter of life (**Roger Lasken, Venter Institute**)

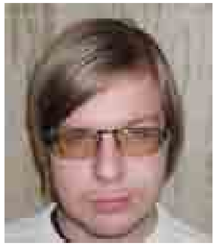
- Sequencing pathogens isolated from hospital environment (**Jeff McLane, Venter Institute**)

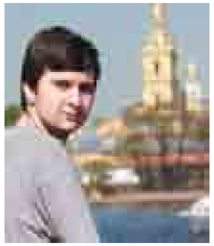
- Sequencing antibiotics producing bacteria (**Bill Gerwick, Scripps Institute of Oceanography**)

- Sequencing drug-resistant pathogens (**Nik Schork, Scripps Translational Medicine**)

Acknowledgments: SPAdes Assembler


Dmitry
Antipov


Anton
Bankevich


Mikhail
Dvorkin

Valery
Lesin

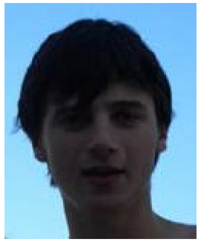
Alexander
Kulikov

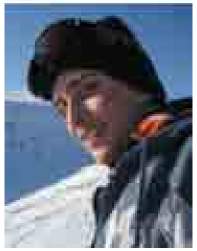
Sergey
Nurk

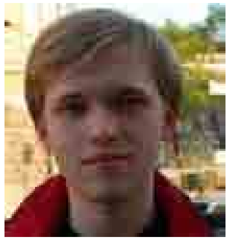
Nikolay
Vyahhi

Alexander
Sirotkin

Alexey
Gurevich

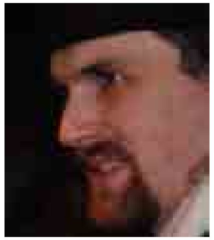

Alexey
Pyshkin

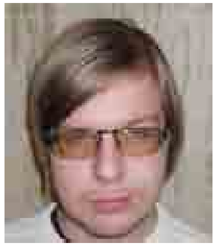

Andrey
Przhibelsky

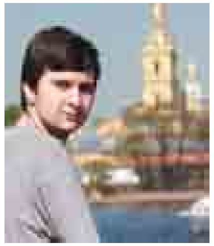

Sergey
Nikolenko

Saint Petersburg Academic University, Russian Academy of Sciences

Acknowledgments: SPAdes Assembler


Dmitry Antipov


Anton Bankevich


Mikhail Dvorkin

Valery Lesin

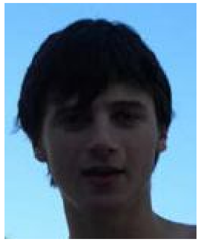
Alexander Kulikov

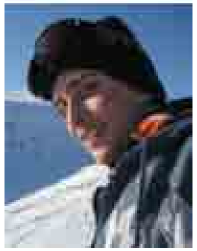
Sergey Nurk

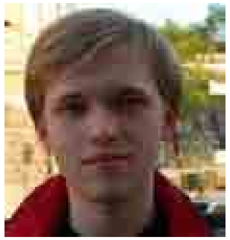
Nikolay Vyahhi

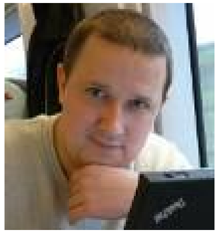
Alexander Sirotkin

Alexey Gurevich


Alexey Pyshkin


Andrey Przhibelsky

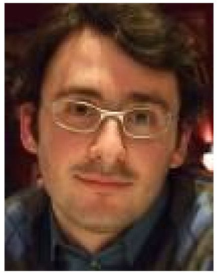

Sergey Nikolenko


Max Alekseyev
University of South Carolina

Glenn Tesler
UCSD

Son Pham
UCSD

Acknowledgments: PDBGs



Paul Medvedev
UCSD

Son Pham
UCSD

Glenn Tesler
UCSD

Mark Chaisson
Pacific Biosciences

Acknowledgements

E+V-SC assembler

Hamid Chitsaz
Wayne State

Glenn Tesler
UCSD

Roger Lasken
Venter Institute

Mary-Jane Lombardo
Venter Institute

J. Craig Venter Institute

Joyclyn L. Yee-Greenbaum
Christopher L. Dupont
Johnathan H. Badger
Mark Novotny
Douglas B. Rusch

Illumina

Louise J. Fraiser
Niall A. Gormley
Ole Schulz-Trieglaff
Geoffrey P. Smith
Dick J. Evers

Funding

NHGRI (NIH-2 R01 HG003647)
Alfred P. Sloan Foundation
NIH 3P41RR024851-02S1
Megagrant from Russian Ministry of Science

RECOMB 2012 Satellite Conferences in Saint Petersburg, Russia

Open Problems in Algorithmic Biology (1st)

August 27-29, 2012

http://bioinf.spbau.ru/ab2012

RECOMB-AB brings together leading researchers in the mathematical, computational, and life sciences to discuss current challenges in computational biology, with an emphasis on open algorithmic problems. The program will consist of invited speakers, contributed speakers, posters, and discussion panels.

Submission Deadline: **April 27, 2012**

Due to the close deadlines, contact us right away if you are interested but would need a short extension.

Bioinformatics Education (4th)

August 26, 2012

http://bioinf.spbau.ru/be2012

RECOMB-BE will consist of invited presentations, oral presentations selected from submitted educational problems, and discussion panels, all of which focus on improving bioinformatics education.

Submission Deadline: **May 7, 2012**