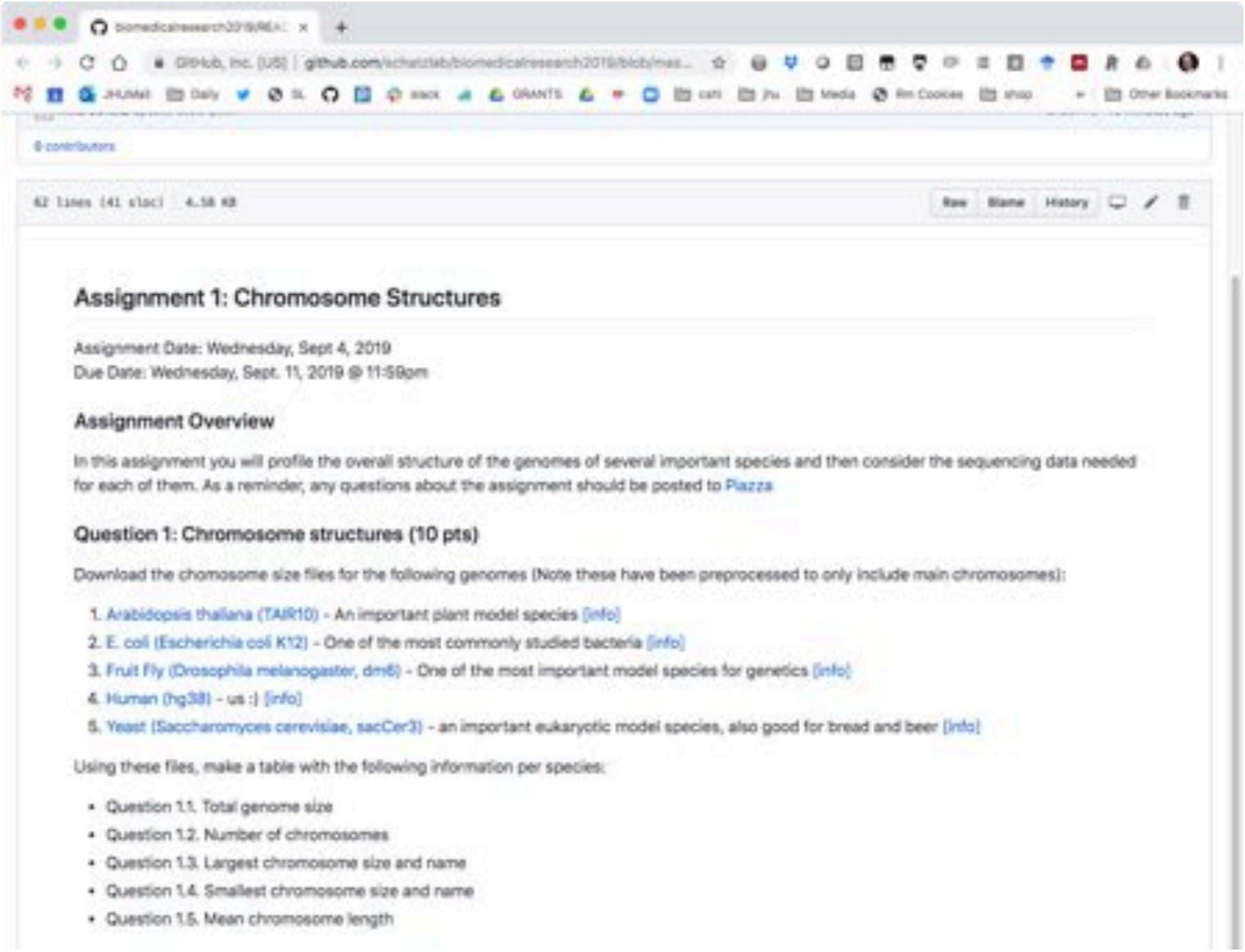
The human genome

Michael Schatz

Sept 11, 2019

Lecture 4: Computational Biomedical Research

Assignment 1: Chromosome Structures
Due Wed Sept 11 @ 11:59pm



Assignment 1: Chromosome Structures

Assignment Date: Wednesday, Sept 4, 2019
Due Date: Wednesday, Sept. 11, 2019 @ 11:59pm

Assignment Overview

In this assignment you will profile the overall structure of the genomes of several important species and then consider the sequencing data needed for each of them. As a reminder, any questions about the assignment should be posted to Piazza

Question 1: Chromosome structures (10 pts)

Download the chomosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):

1. Arabidopsis thaliana (TAIR10) - An important plant model species [info]
2. E. coli (Escherichia coli K12) - One of the most commonly studied bacteria [info]
3. Fruit Fly (Drosophila melanogaster, dm6) - One of the most important model species for genetics [info]
4. Human (hg38) - us :) [info]
5. Yeast (Saccharomyces cerevisiae, sacCer3) - an important eukaryotic model species, also good for bread and beer [info]

Using these files, make a table with the following information per species:

- Question 1.1. Total genome size
- Question 1.2. Number of chromosomes
- Question 1.3. Largest chromosome size and name
- Question 1.4. Smallest chromosome size and name
- Question 1.5. Mean chromosome length

https://github.com/schatzlab/biomedicalresearch2019

Assignment 2: De novo Assembly
Due Wed Sept 18 @ 11:59pm

Assignment 2: De novo assembly

Assignment Date: Wednesday, Sept 11, 2019
Due Date: Wednesday, Sept. 18, 2019 @ 11:59pm

Assignment Overview

In this assignment you will assemble a few simple datasets by hand and then by using an automated assembler called Spades. As a reminder, any questions about the assignment should be posted to Piazza

Question 1: de Bruijn graph construction (10 pts)

1a. Draw (by hand or in code) the de Bruijn graph for the following reads using k=3 (assume all reads are from the forward strand, no sequencing errors, complete coverage of the genome)

```
ACTA
ATCA
ATCG
ATCT
CATC
CTAC
CTAT
GATC
TACT
TATC
TCAT
TCTA
```

https://github.com/schatzlab/biomedicalresearch2019

Part 1: Recap

de Bruijn Graph Construction

- $G_k = (V,E)$
  - V = Length-*k* sub-fragments
  - E = Directed edges between consecutive sub-fragments
    - Sub-fragments overlap by *k*-1 words

Fragments |f|=5

It was the best of

was the best of times

Sub-fragment *k*=4

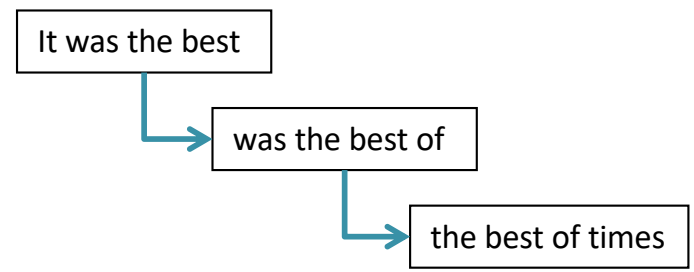
It was the best

was the best of

was the best of

the best of times

Directed edges (overlap by *k*-1)



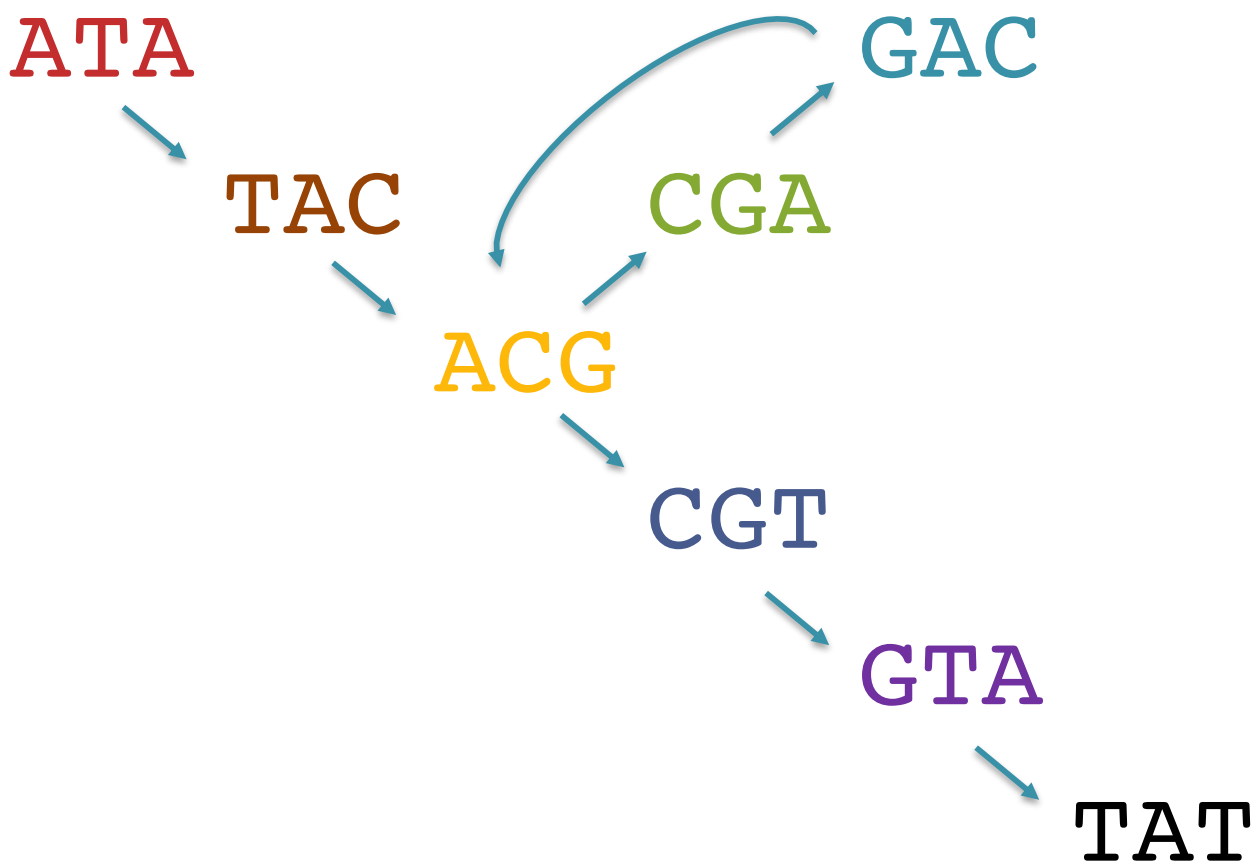
– Overlaps between fragments are implicitly computed

de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~

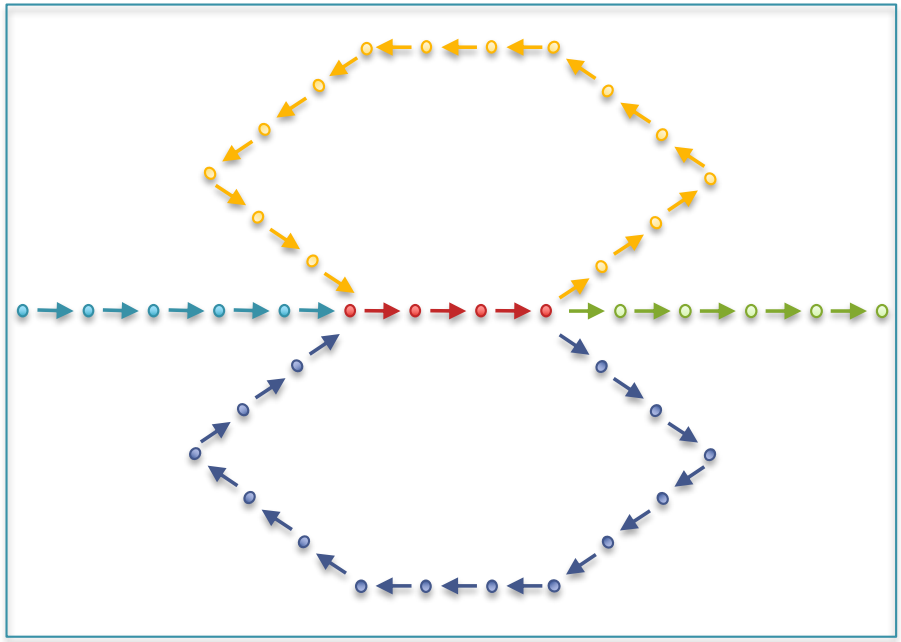


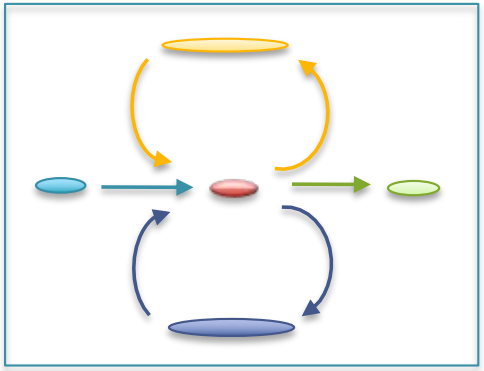
Note: there is no edge from ATA to TAT

ATACGACGTAT

Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"





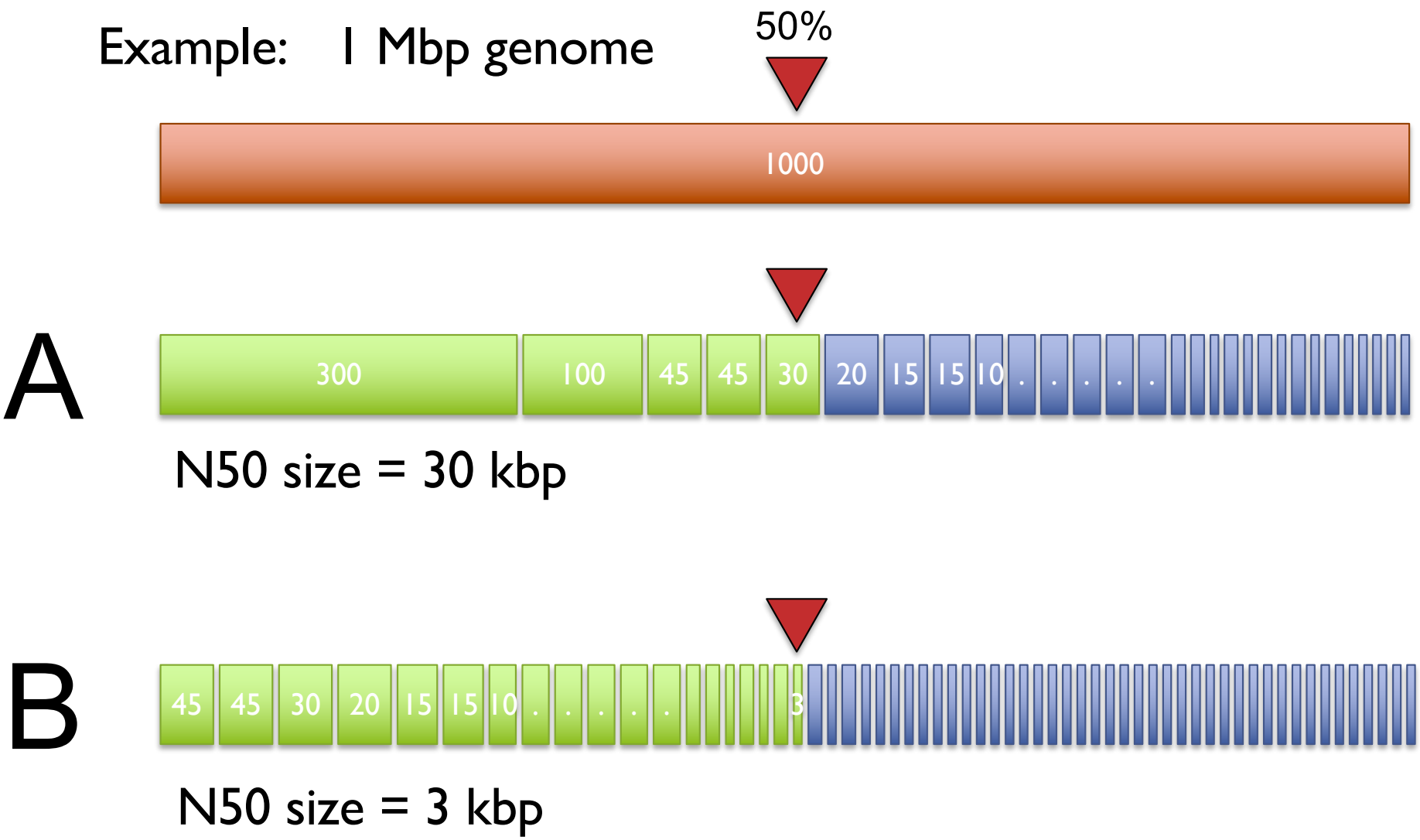
Why do contigs end?

(1) End of chromosome! ☺, (2) lack of coverage, (3) errors, (4) heterozygosity and (5) repeats

Contig N50

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



Part 2: The human genome

The scale of DNA in our body is staggering.

- **A typical human is comprised of roughly 40 trillion human cells**
  (excluding trillions of bacterial cells in our gut)

- If stretched out, each haploid genome would be **roughly 2 meters**.
- So, each cell has 4 meters of DNA.
- 40 trillion * 4 meters = 160 trillion meters.
- 160 trillion meters / 1609.34 = 99,750,623,441 miles
- 99,750,623,441 / 92,960,000 = 1,073.05 trips to the sun.

A typical cell replicates about 100 times

160 trillion meters x 100 =

1.69123746 light years

More info

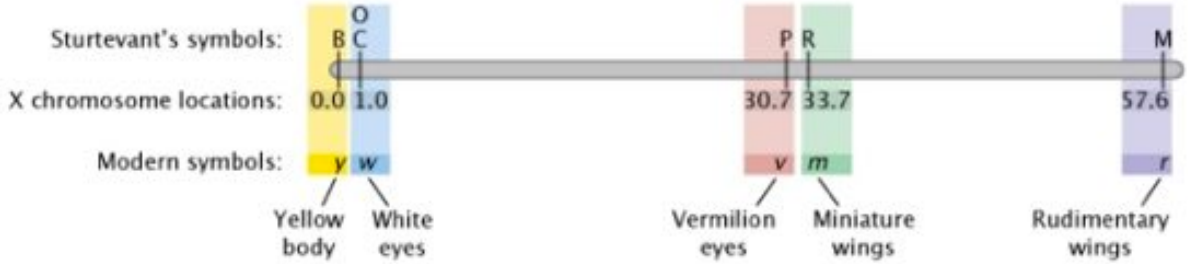
The first genetic map

http://www.caltech.edu/news/first-genetic-linkage-map-38798

Mendel's Second Law (The Law of Independent Assortment) states alleles of one gene sort into gametes independently of the alleles of another gene: ***Pr(smooth/wrinkle) is independent of Pr(yellow/green)***

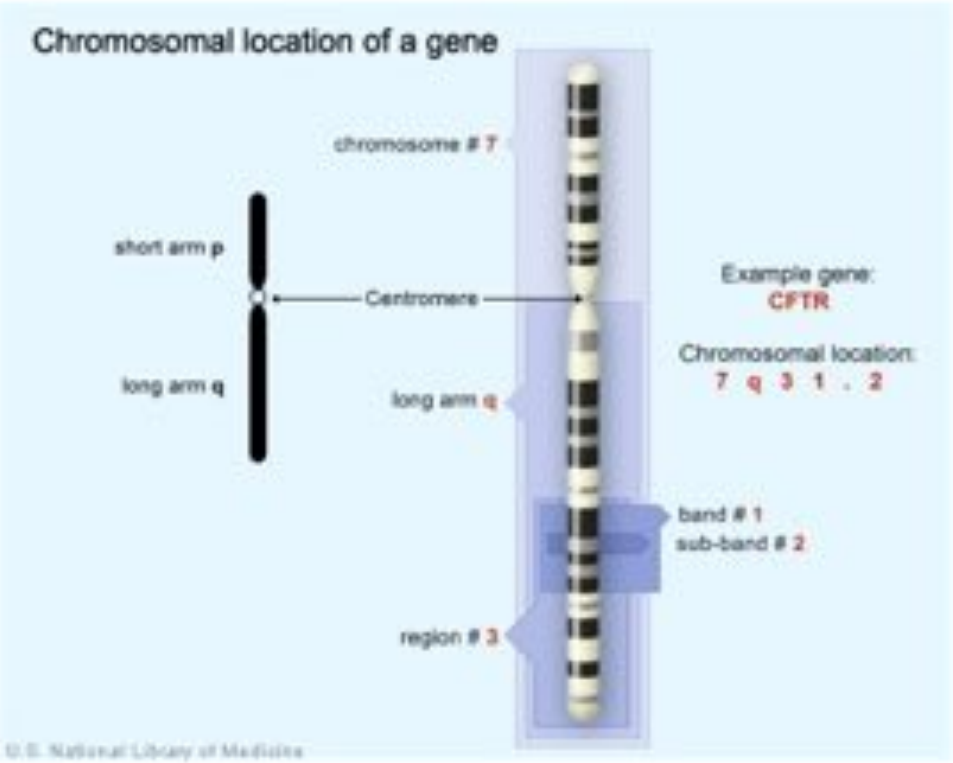
Morgan and Sturtevant noticed that the probability of having one trait given another was **not** always 50/50– those traits are ***genetically linked***

Sturtevant realized the probabilities of co-occurrences could be explained if those alleles were arranged on a linear fashion: traits that are most commonly observed together must be locates closest together



The Linear Arrangement of Six Sex-Linked Factors in Drosophila as shown by their mode of Association
Sturtevant, A. H. (1913) *Journal of Experimental Zoology*, 14: 43-59

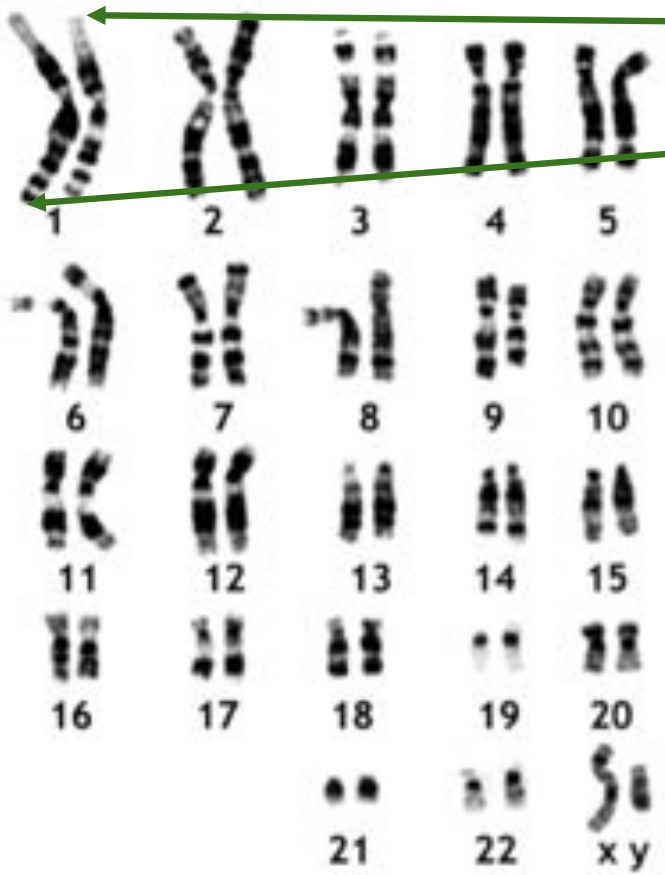
Chromosome Giemsa banding (G-banding)



- Heterochromatic regions, which tend to be rich with adenine and thymine (AT-rich) DNA and relatively gene-poor, **stain more darkly** with Giemsa and result in G-banding
- Less condensed ("open") chromatin, which tends to be (GC-rich) and more transcriptionally active, incorporates less Giemsa stain, resulting in **light bands in G-banding**.
- Cytogenetic bands are labeled p1, p2, p3, q1, q2, q3, etc., **counting from the centromere out toward the telomeres**. At higher resolutions, sub-bands can be seen within the bands.
- For example, the locus for the CFTR (cystic fibrosis) gene is **7q31.2**, which indicates it is on **chromosome 7**, **q arm**, **region 3, band 1, and sub-band 2. (Say 7,q,3,1 dot 2)**

https://en.wikipedia.org/wiki/G_banding, https://ghr.nlm.nih.gov/chromosome/1#ideogram

The human karyotype

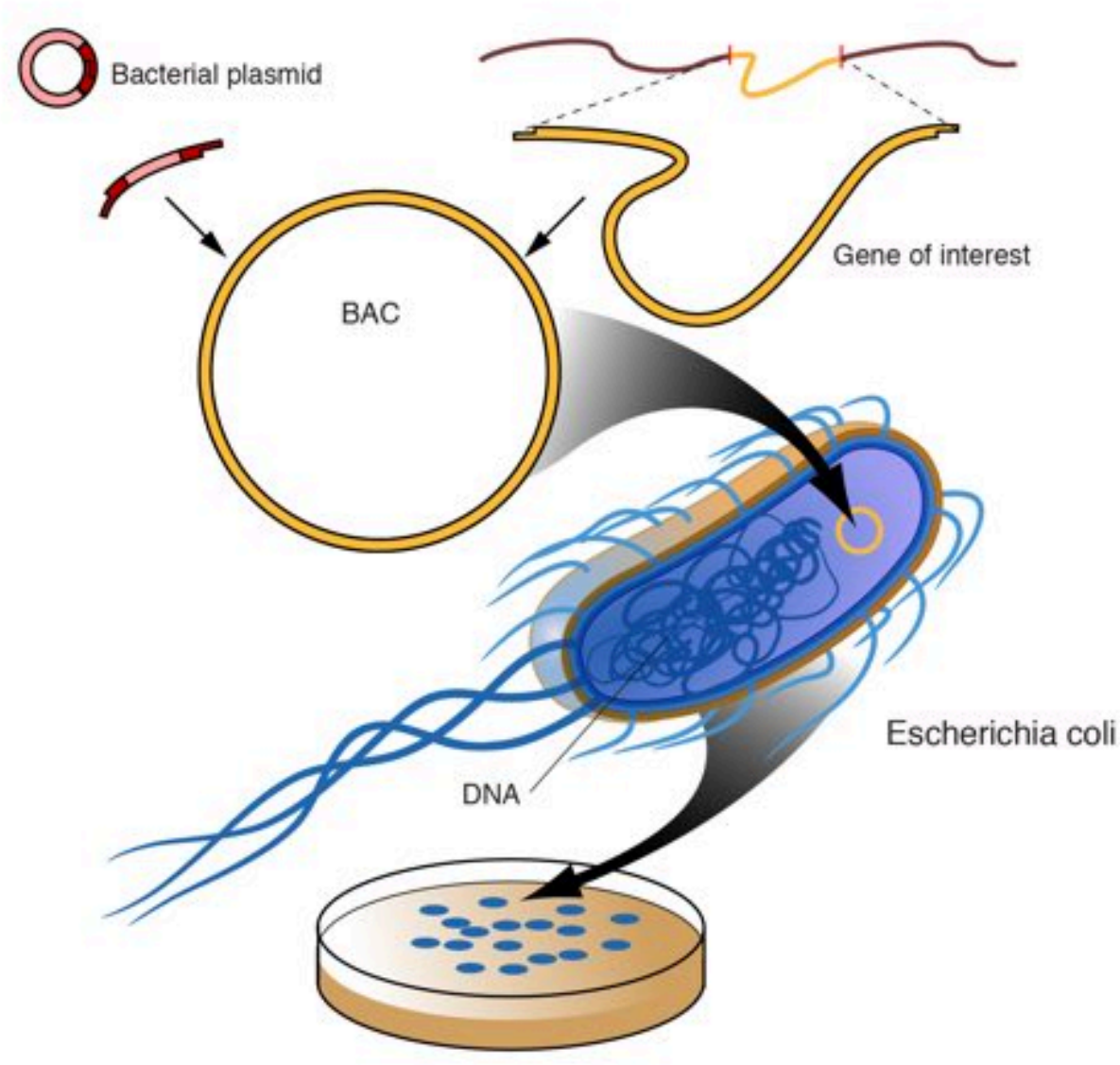


Parental haploid copy 1

Parental haploid copy 2

Male

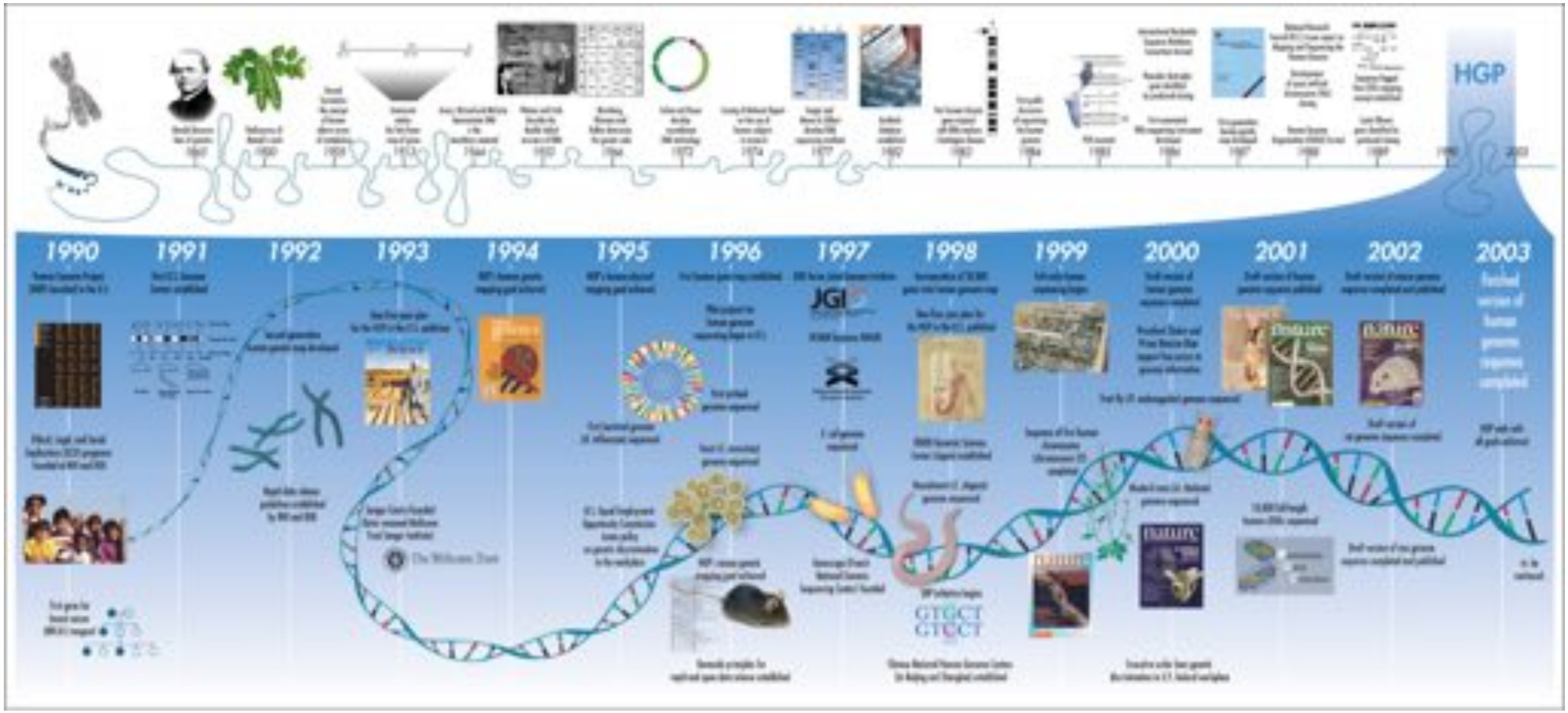
Bacterial Artificial Chromosomes (BACs)



- A BAC is an engineered DNA molecule used to clone DNA sequences in bacterial cells (for example, E. coli).
- BACs are often used in connection with DNA sequencing.
- Segments of an sample's DNA, ranging from 100,000 to about 300,000 base pairs, can be inserted into BACs.
- The BACs, with their inserted DNA, are then taken up by bacterial cells.
- As the bacterial cells grow and divide, they amplify the BAC DNA, which can then be isolated and used in sequencing DNA.

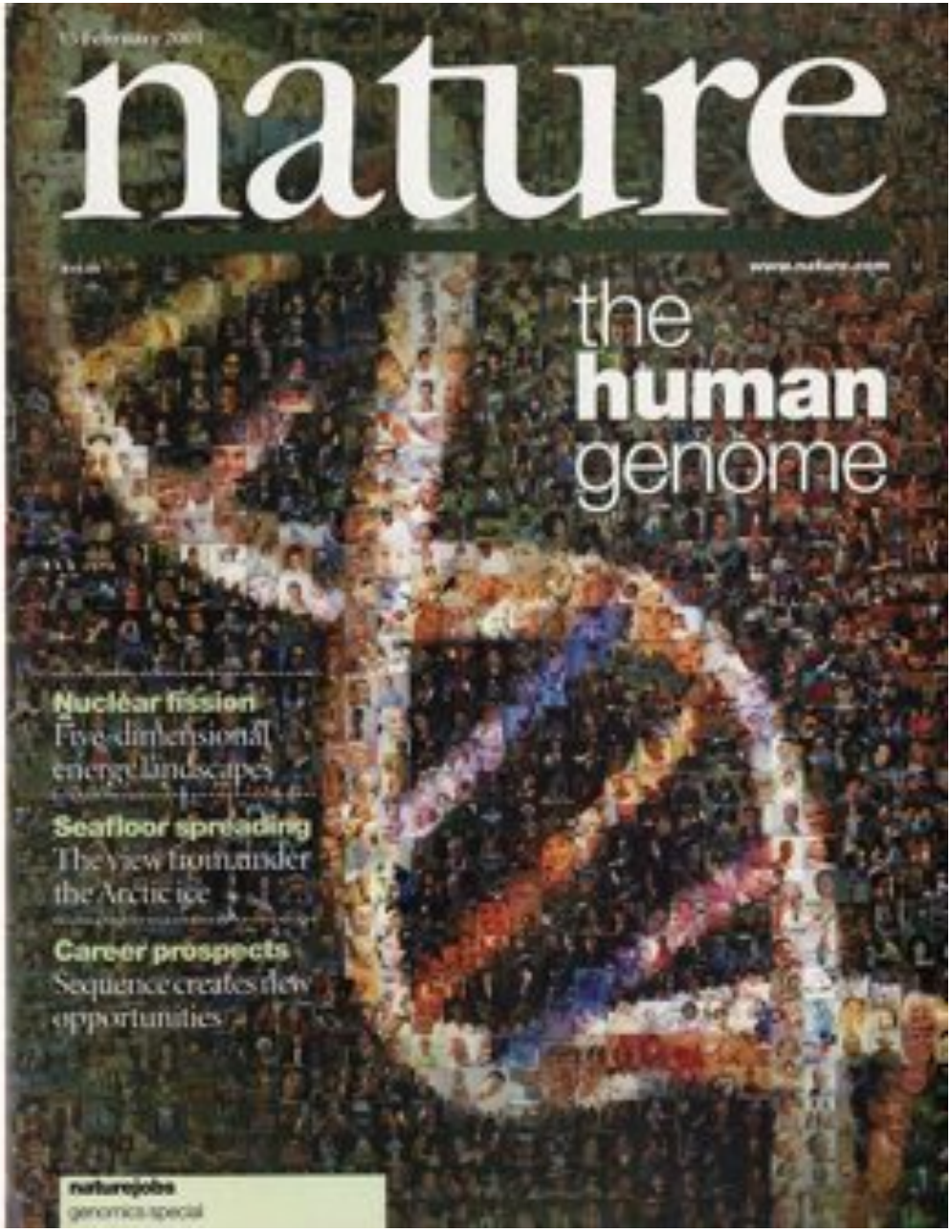
https://www.genome.gov/genetics-glossary/Bacterial-Artificial-Chromosome

History of the Human Genome Project



The reference human genome

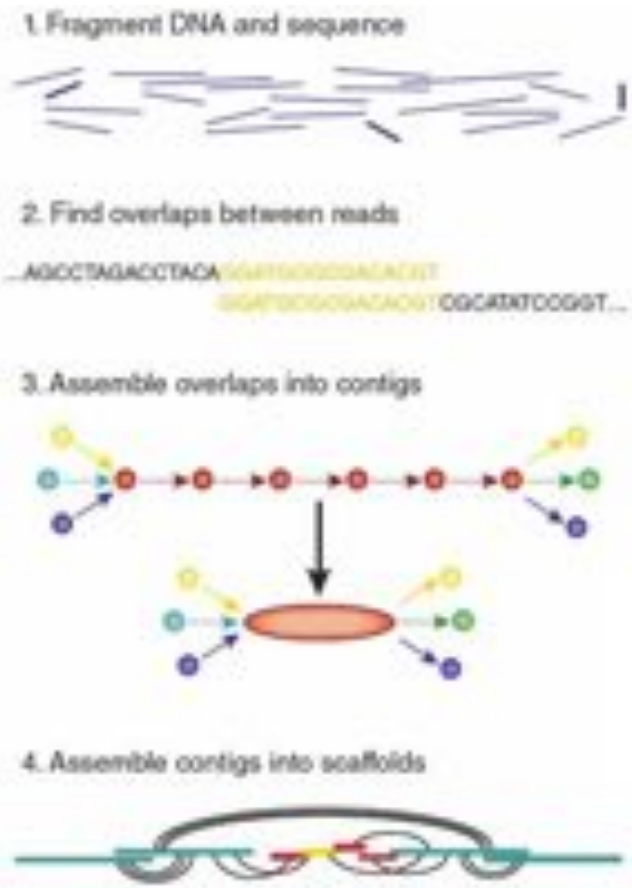
The reference human genome

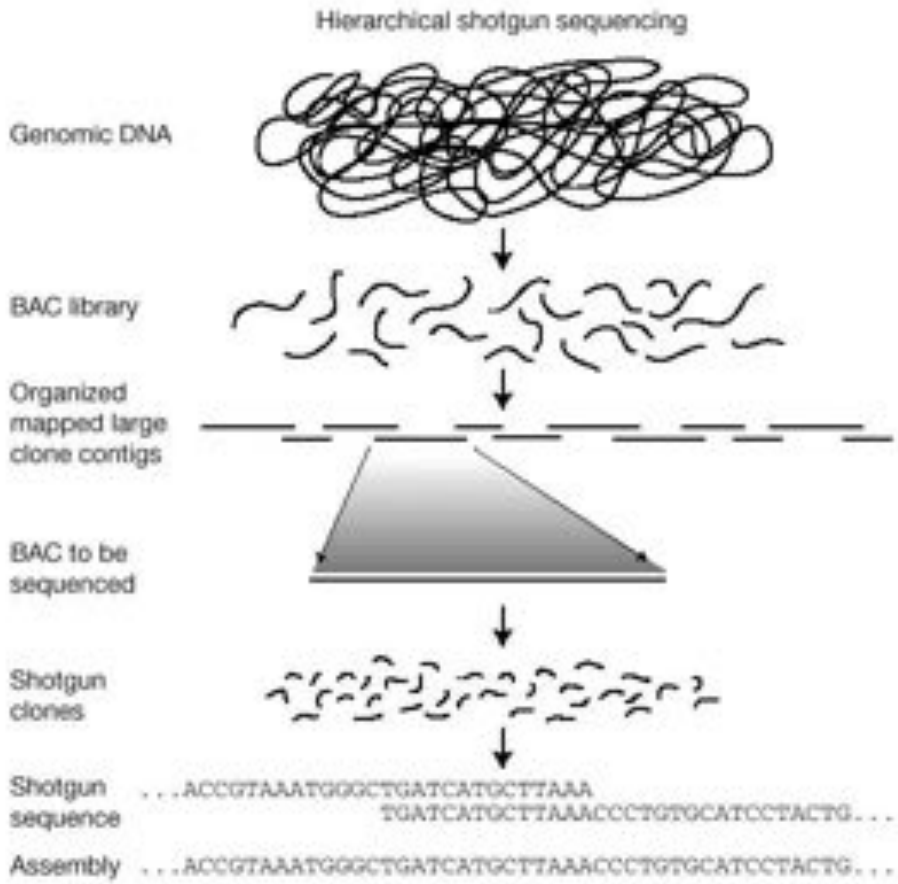


The Sequence of the Human Genome
Venter et al.
Science 291. pp 1304-1351 (2001)

Initial sequencing and analysis of the human genome
International Human Genome Sequencing Consortium
Nature 409, pp 860–921 (2001)

Two Human Genomes?





The Sequence of the Human Genome
Venter et al.
Science 291. pp 1304-1351 (2001)

(Figure from Baker (2012) Nature Methods)

Initial sequencing and analysis of the human genome
International Human Genome Sequencing Consortium
Nature 409, pp 860–921 (2001)

Who is the reference human?

The Buffalo News/Sunday, March 23, 1997

ment abuse, civil disobedience

opic. But the very nature of gov- creates a mind set that inspires increase their authority, always at cric of the people."

y, government has forgotten that rvant of the people," Parlato add- acting more like it's the master."

o and the Lapps share an abiding non-violent civil disobedience.

insist on being respectful in our of resistance," Barbara Lyn Lapp ut if we claim to care about our rs, we must protest government in-

violence has to be the watchword, said, calling civil disobedience the st of the violent militia movement. on-violence can serve as an anti- government oppression, he added. law is unjust or you're given an ithout moral or legal authority, you should refuse it," Parlato said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls the aggressors in society. Instead, it too often comes down on the side of the aggressors, who enforce child-protection laws, compulsory education, disclosure rules on tax forms and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapps and Parlato will be joined by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Pieter de Jong, RPCI

Who is the reference human?

The Buffalo News/Sunday, March 23, 1997

ment abuse, civil disobedience

opic. But the very nature of gov- creates a mind set that inspires increase their authority, always at cise of the people."

y, government has forgotten that rvant of the people," Parlato add- acting more like it's the master."

o and the Lapps share an abiding non-violent civil disobedience.

insist on being respectful in our of resistance," Barbara Lyn Lapp ut if we claim to care about our rs, we must protest government in-

violence has to be the watchword, said, calling civil disobedience the st of the violent militia movement. on-violence can serve as an anti- government oppression, he added. law is unjust or you're given an ithout moral or legal authority, you should refuse it," Parlato said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls the aggressors in society. Instead, it too often comes down on the side of the aggressors, who enforce child-protection laws, compulsory education, disclosure rules on tax forms and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapps and Parlato will be joined by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Pieter de Jong, RPCI

Who is the reference human?

The Buffalo News/Sunday, March 23, 1997

ment abuse, civil disobedience

ople. But the very nature of gov- creates a mind set that inspires increase their authority, always at cise of the people."

y, government has forgotten that rvant of the people," Parlato add- acting more like it's the master."

o and the Lapps share an abiding non-violent civil disobedience.

insist on being respectful in our of resistance," Barbara Lyn Lapp ut if we claim to care about our rs, we must protest government in-

violence has to be the watchword, said, calling civil disobedience the st of the violent militia movement. on-violence can serve as an anti- government oppression, he added. law is unjust or you're given an ithout moral or legal authority, you should refuse it," Parlato said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls the aggressors in society. Instead, it too often comes down on the side of the aggressors, who enforce child-protection laws, compulsory education, disclosure rules on tax forms and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapps and Parlato will be joined by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Pieter de Jong, RPCI

Appendix: Identifying the ancestry of segments of the human genome reference sequence

To compare Neandertal to present-day human haplotypes for the purpose of population genetic analysis, we needed to have long haploid sequences from present-day humans that were of known ancestry. To identify such segments, we took advantage of the fact that the human reference sequence is haploid over scales of tens of kilobases, because it is comprised of a tiling-path of Bacterial Artificial Chromosomes (BACs) or other clone types that are of typical size 50-150 kb (*S92*). We do not know of any other substantial source of high quality human haploid sequences of the requisite size.

Determining the ancestries of the libraries in the human genome reference sequence using HAPMIX

It is crucial to know the 'ancestry' of a clone to use it in a meaningful population genetic analysis. In what follows, we define 'ancestry' as the geographic region in which a clone's ancestor lived 1,000 years ago, inferred based on its genetic proximity to other individuals from that region today. This definition allows us to classify clones from Chinese Americans as "East Asian," from European Americans as "European", and from African Americans as either "West African" or "European".

To identify the ancestries of the libraries comprising most of the human genome reference sequence, we used a list of 26,558 clones tiling the great majority of the genome, most of which we were able to assign to a library of origin. Restricting to the autosomes, we identified 21,156 clones that seemed to fall into 9 libraries based on the naming scheme: CTA (n=199), CTB (n=356), CTC (n=452), CTD (n=1,426), RPCI-1 (n=740), RPCI-3 (n=456), RPCI-4 (n=716), RPCI-5 (n=802) and RPCI-11 (n=16,009). (In a subsequent re-examination, we identified additional clones that we likely could have classified into libraries, including 953 from RPCI-11, 632 from RPCI-1, and 490 from another library RPCI-13.) The median span of the 21,156 clones we analyzed was 112 kb, and 80% are >50kb in size. About 2/3 came from a single library, RPCI-11.

1. RPCI-11 is an African American: RPCI-11, the individual who contributed most of the human genome reference sequence, is consistent with having African American ancestry, with 42% of the clones of confident West African ancestry and 42% of the clones of confident European ancestry, and the ancestry of the remaining clones less confidently inferred. The finding of likely African American ancestry for RPCI-11 was previously reported in a study of the ancestry of RPCI-11 clones spanning the Duffy blood group locus (*S93*), and here we confirm this finding, and also expand the inference to the whole genome.
2. CTD is an East Asian: The majority of clones from CTD, the second largest library in its contribution to the human genome sequence, is likely an East Asian. In a HAPMIX analysis with CEU (European) – CHB+JPT (East Asian) as the proposed ancestral populations, the majority of clones are of confident East Asian origin, and there is no secondary mode of confident European ancestry, as might be expected from a Latino or South Asian individual.
3. The remaining 7 libraries are European: The remaining libraries (CTA, CTB, CTC, RPCI-1, RPCI-3, RPCI-4 and RPCI-5) are inferred to be of European ancestry, since they all have consistent distributions of inferred clone ancestries, with the majority of clones of confident European ancestry in both our HAPMIX analyses and no secondary modes.

A Draft Sequence of the Neandertal Genome

Green et al (2010) Science. DOI: 10.1126/science.1188021
Supplemental Note 16 (pg 145-146)

Who is the reference human?

nature methods
Techniques for life scientists and chemists

EDITORIAL

Nature Methods **7**, 331 (2010)
doi:10.1038/nmeth0510-331

E pluribus unum

If the human reference genome is to reflect more of the actual genomic diversity in humans, community participation is needed.

Please visit methagora to view and post comments on this article.

The human genome is ten years old. We acknowledge its reference assembly as an invaluable resource essential for many purposes such as the assembly of short reads from high-throughput sequencing platforms into chromosome context during resequencing projects. At the same time, we think necessary improvement of the reference genome depends on the willingness of the research community to provide data for the genome's less accessible regions.

First published in 2001, the human reference genome has, since 2007, been in the hands of the Genome Reference Consortium (GRC) a small group of fewer than 20 scientists from the European Bioinformatics Institute, the US National Center for Biotechnology Information, The Sanger Institute and The Genome Center at Washington University in St. Louis, who have committed to the improvement and completion of this reference, with very little financial support.

The reference genome is now in its 19th rendition, and probably the best measure of its improvement over the last ten years is the number of fragments it consists of. The very first version had ~150,000 gaps; the most recent build, GRCh37, has only around 250 gaps.

The only other publicly accessible *de novo* assembly of a human genome that contains chromosome sequences is HuRef. Obtained by traditional capillary sequencing, HuRef is the diploid genome of Craig Venter. It comes in 4,500 pieces and, like any individual genome, it contains many rare alleles.

GRCh37, in contrast, is a mosaic haploid genome derived from about 13 people. It still contains rare alleles, but the GRC recently decided to convert these to common haplotypes. Deciding which alleles are common and which are rare is proving challenging, and the GRC members are collaborating with members of the 1000 Genomes project to collect enough data to make these decisions.

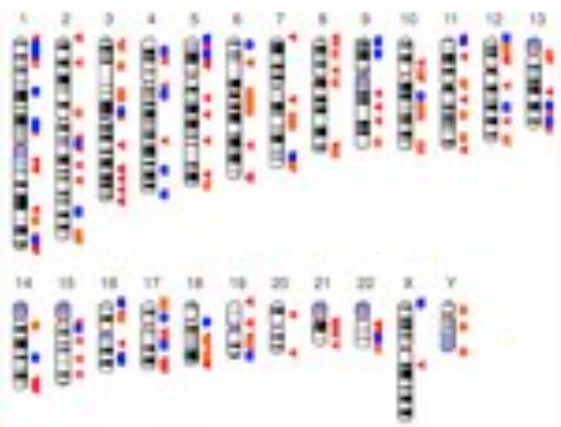
Genome Reference Consortium

GRC Home | Data | Help | Report an issue | Contact Us | Credits | Curators Only

Human Overview | Human Genome Issues | Human Assembly Data

Human Genome Overview

Information about the continuing improvement of the human genome



Region containing alternate loci
Region containing fix patches
Region containing novel patches

Ideogram of the latest human assembly, GRCh38.p11

The GRC is working hard to provide the best possib
by both generating multiple representations (alterna
represented by a single path. Additionally, we are re
allows users who are interested in a specific locus to
affecting users who need chromosome coordinate s

Download data:

- GRCh38.p11 (latest minor release) FTP
- GRCh38 (latest major release) FTP
- Genomic regions under review FTP
- Current Tiling Path Files (TPFs)

Transitioning to GRCh38? Try the NCBI Remapp
assembly alignments used by the GRC.

Next assembly update
The next assembly update (GRCh38.p12) will be

GRCh38.p11 | GRCh37.p13 | GRCh37

GRCh38.p11

Release date: June 14, 2017
Release type: minor
Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordin
of patch scaffolds is now: 64 FIX and 59 NOVEL.
Assembly accessions: GenBank: GCA_000001405.26 , RefSeq: GCF_000001405.37

Pseudoautosomal regions

Name	Chr	Start	Stop
PAR#1	X	10,001	2,781,479
PAR#2	X	155,701,383	156,030,895
PAR#1	Y	10,001	2,781,479
PAR#2	Y	56,887,903	57,217,415

Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Register on Piazza & GradeScope
4. Submit HW 1
5. Work on HW2