

Applied Comparative Genomics

Michael Schatz

August 28, 2023

Lecture I: Course Overview



Welcome!

The primary goal of the course is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses.

- We will study the leading computational and quantitative approaches for comparing and analyzing genomes starting from raw sequencing data.
- The course will focus on human genomics and human medical applications, but the techniques will be broadly applicable across the tree of life.
- The topics will include (pan)-genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics.

Course Webpage:

<https://github.com/schatzlab/appliedgenomics2023>

Course Discussions:

<http://piazza.com>

Class Hours:

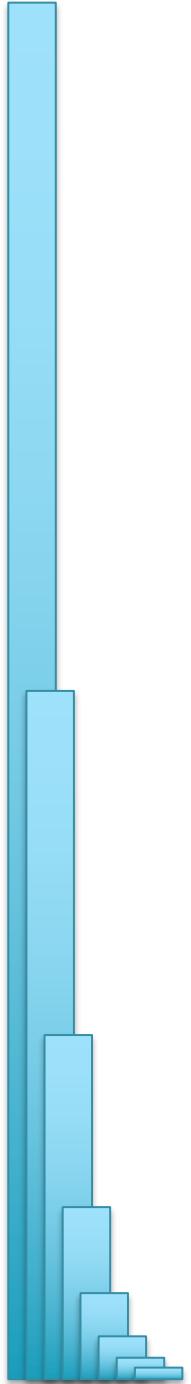
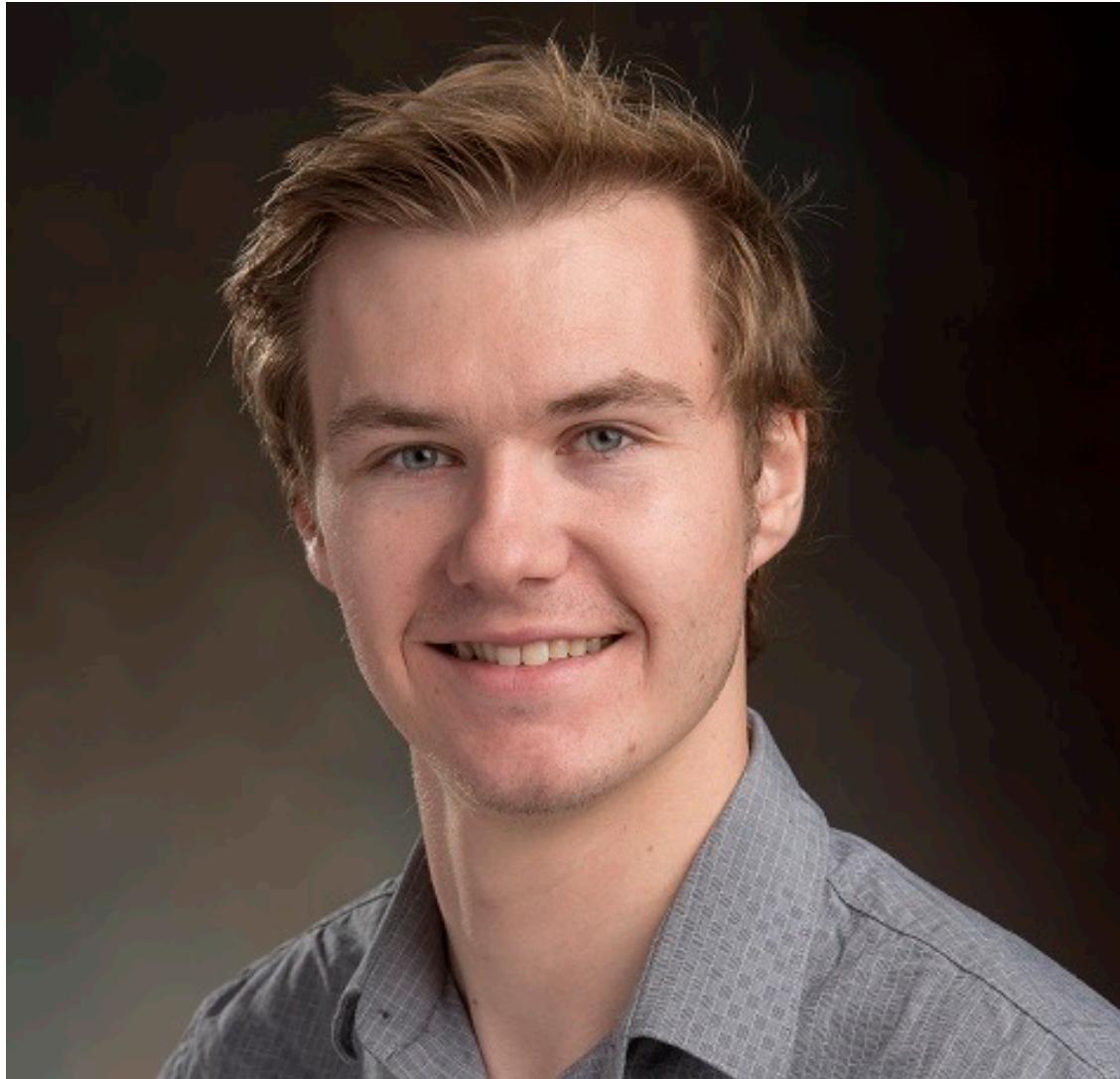
Mon + Wed @ 3:00p – 4:15p, Hodson 311

Schatz Office Hours: TBD and by appointment

Sweeten Office Hours: TBD and by appointment

Please try Piazza first!

TA:Alex Sweeten



Prerequisites and Resources

Prerequisites

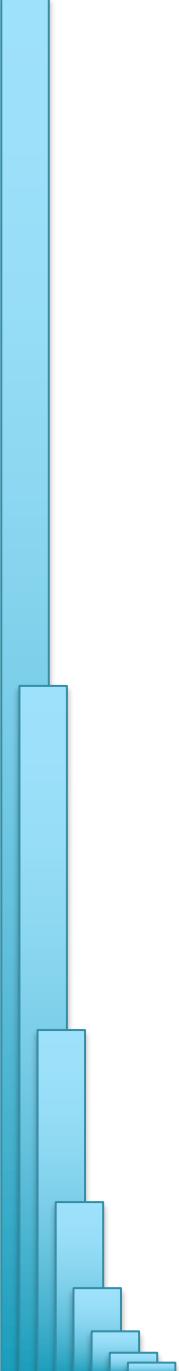
- No formal course requirements
- Access to an Apple or Linux Machine, or Install VirtualBox
- Familiarity with the Unix command line for exercises
 - bash, ls, grep, sed, + install published genomics tools
- Familiarity with a major programming language for project
 - C/C++, Java, R, Perl, Python

Primary Texts

- None! We will be studying primary research papers

Other Resources:

- Google, SEQanswers, Biostars, StackOverflow
- Applied Computational Genomics Course at UU
- <https://github.com/quinlan-lab/applied-computational-genomics>
- Ben Langmead's teaching materials:
- <http://www.langmead-lab.org/teaching-materials/>



Grading Policies

Assessments:

- 5 Assignments: 30% Due at 11:59pm a week later
Practice using the tools and methods we are discussing
- 1 Exam: 30% In Class (Tentatively Nov 6)
Assess your performance, focusing on the concepts
- 1 Class Project: 40% Presented last week of class
Significant project developing a novel analysis/method
- In-class Participation: Not graded, but there to help you!

Policies:

- Scores assigned relative to the highest points awarded
- Automated testing and grading of assignments
- ***Late Days:***
 - A total of 96 hours (24×4) can be used to extend the deadline for assignments, but not the class project, without any penalty; after that time assignments will not be accepted

On ChatGPT

The screenshot shows a web browser window for ChatGPT. On the left, there's a sidebar with a list of previous chats categorized by date: Today, Yesterday, Previous 7 Days, Previous 30 Days, and July. Under Today, there are two entries: "Using ChatGPT Effectively" and "Balanced Approach: ChatGPT & Learn". Under Yesterday, there are several entries related to genomics and data science. Under Previous 7 Days, there are entries like "Retirement Savings Calculation" and "Ankle Popping and Relief". Under Previous 30 Days, there are entries like "SOLscape Genomics Project" and "Late-Night Party Jokes". Under July, there are entries like "Churned User Analysis" and "Mark Ebert Inquiry".

The main area shows a conversation between a user (Michael Schatz) and GPT-4. Michael asks: "what's a good policy on the use of chatgpt for a university class on genomics and data science. the instructor is generally favorable but also wants students to learn and understand the content presented in the class". GPT-4 responds with a detailed policy:

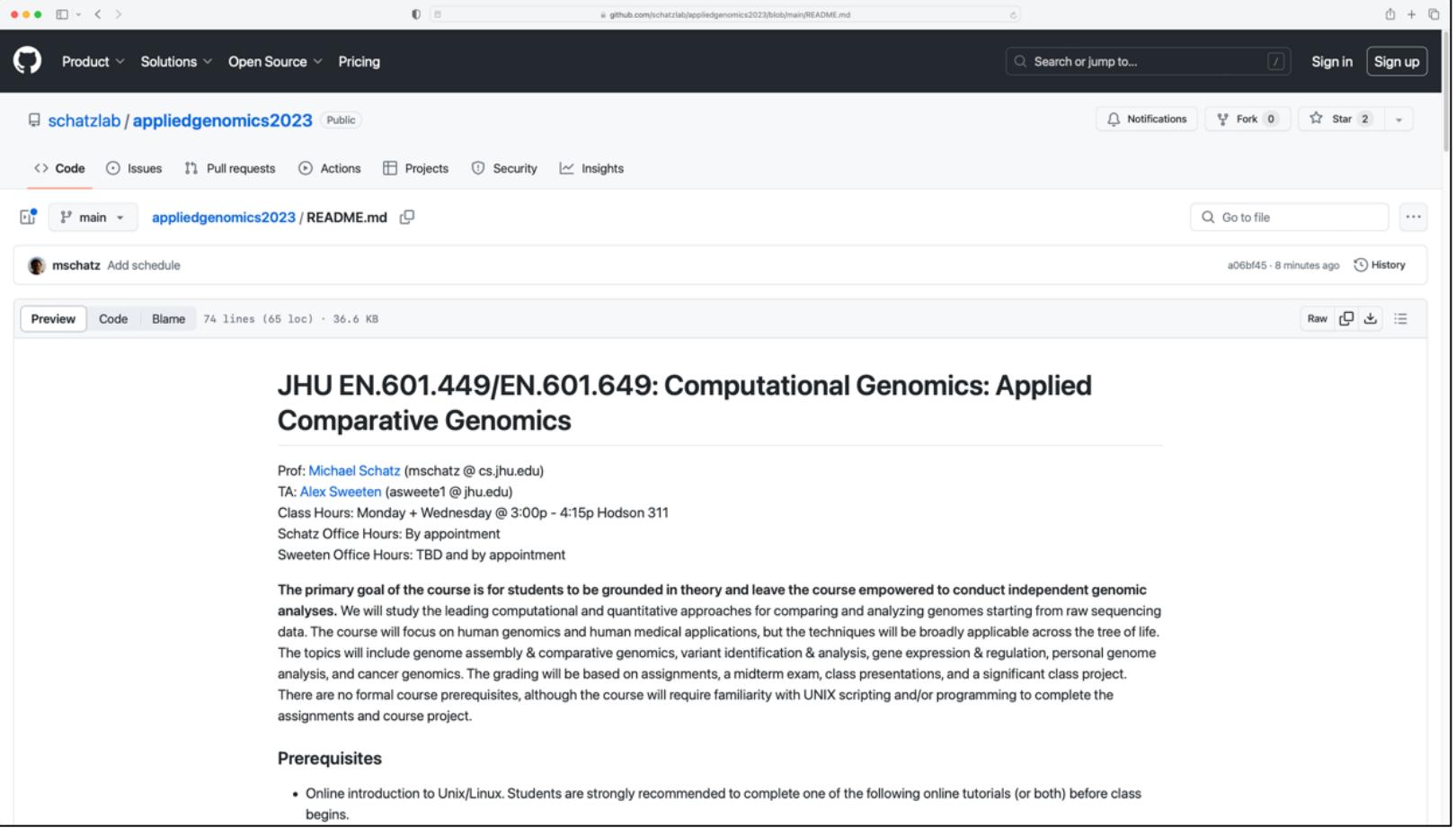
Using ChatGPT in a university class on genomics and data science can be very beneficial, as it can assist students in understanding complex topics, provide immediate feedback, and offer personalized learning experiences. However, it's important to strike a balance to ensure that students are still actively engaging with the material and developing critical thinking skills. Here's a good policy that balances these goals:

- Supplemental Use:** ChatGPT should be used as a supplemental tool, not as a replacement for the instructor or the primary course materials. It can be used for clarification of concepts, additional practice, or answering questions outside of class hours.
- Encourage Independent Thinking:** Students should be encouraged to try to answer questions and solve problems on their own before consulting ChatGPT. This will help them develop critical thinking and problem-solving skills.
- Limit Dependence:** Set some limitations on the use of ChatGPT. For example, restrict its use during certain assignments or assessments to encourage independent thinking and to ensure that students are not overly reliant on it.
- Monitor Usage:** If possible, monitor the usage of ChatGPT to ensure that it is being used appropriately and effectively. This can help identify if students are becoming too dependent on it or if they are using it to enhance their understanding.
- Provide Guidance:** Provide clear guidelines on how and when to use ChatGPT. Make sure students understand its purpose and its limitations.
- Feedback Loop:** Encourage students to provide feedback on the utility of ChatGPT, and adjust its usage accordingly. If students find it particularly helpful for certain topics or tasks, consider incorporating it more for those areas. Conversely, if students find that it is hindering their learning or understanding, consider limiting its use.
- Ethical Use:** Emphasize the importance of ethical use of AI tools like ChatGPT. Make sure students understand the implications of AI and its potential impact on society.

Remember, the goal is to use ChatGPT to enhance the learning experience, not to replace traditional methods of instruction or to enable students to bypass critical thinking and problem-solving.

At the bottom, there's a message input field with "Send a message" and "Regenerate" buttons, and a note: "ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version".

Course Webpage



The screenshot shows a GitHub repository page for 'appliedgenomics2023'. The repository is public and contains a single file, 'README.md'. The content of the README is as follows:

JHU EN.601.449/EN.601.649: Computational Genomics: Applied Comparative Genomics

Prof: [Michael Schatz](#) (mschatz @ cs.jhu.edu)
TA: [Alex Sweeten](#) (asweete1 @ jhu.edu)
Class Hours: Monday + Wednesday @ 3:00p - 4:15p Hodson 311
Schatz Office Hours: By appointment
Sweeten Office Hours: TBD and by appointment

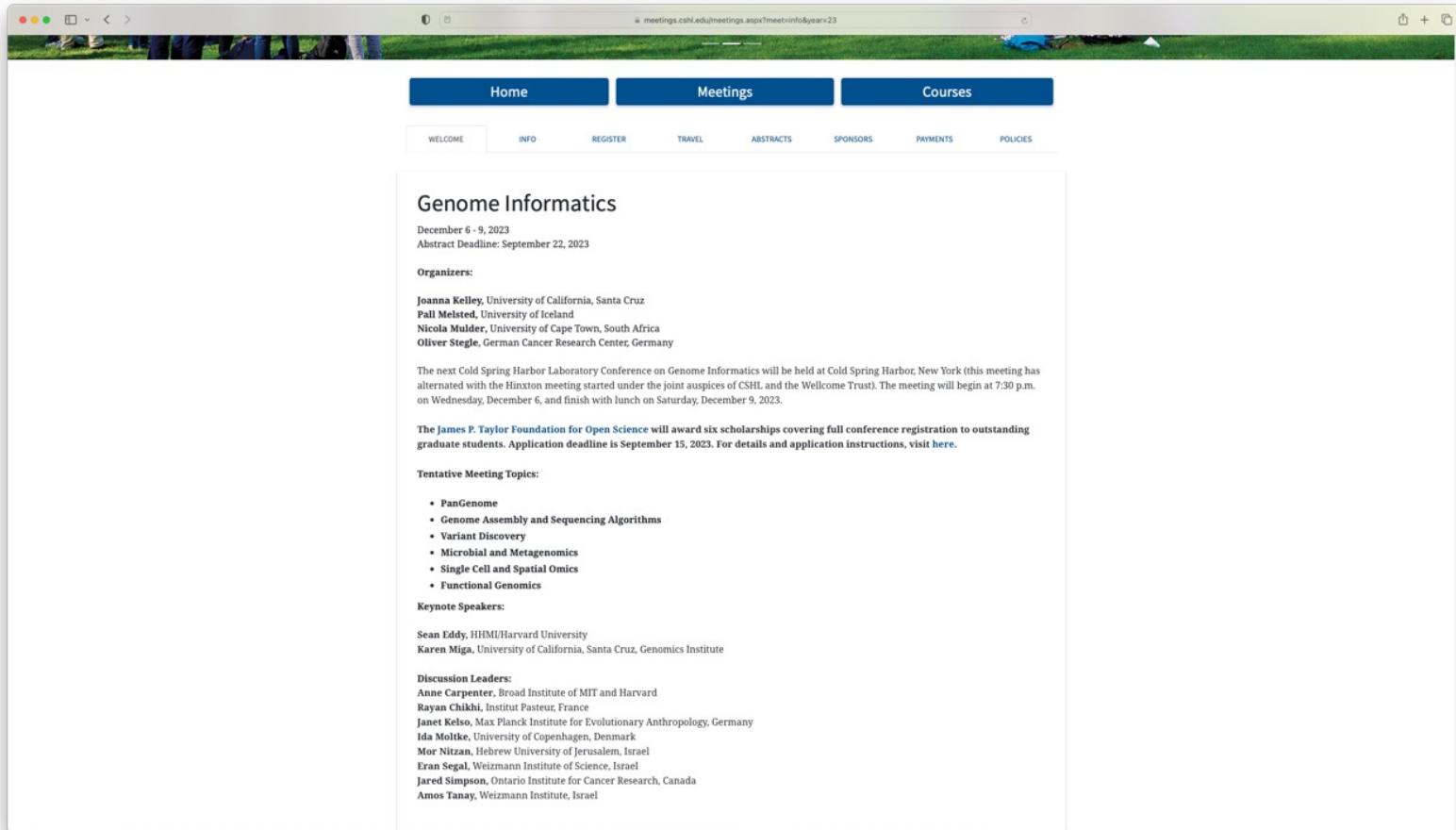
The primary goal of the course is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses. We will study the leading computational and quantitative approaches for comparing and analyzing genomes starting from raw sequencing data. The course will focus on human genomics and human medical applications, but the techniques will be broadly applicable across the tree of life. The topics will include genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics. The grading will be based on assignments, a midterm exam, class presentations, and a significant class project. There are no formal course prerequisites, although the course will require familiarity with UNIX scripting and/or programming to complete the assignments and course project.

Prerequisites

- Online introduction to Unix/Linux. Students are strongly recommended to complete one of the following online tutorials (or both) before class begins.

<https://github.com/schatzlab/appliedgenomics2023>

Genome Informatics



The screenshot shows a web browser displaying the "Genome Informatics" conference page. The page has a header with a navigation menu including "Home", "Meetings" (which is the active tab), and "Courses". Below the menu are links for "WELCOME", "INFO", "REGISTER", "TRAVEL", "ABSTRACTS", "SPONSORS", "PAYMENTS", and "POLICIES". The main content area features a banner image of people at a conference, followed by the title "Genome Informatics" and the dates "December 6 - 9, 2023". It also mentions the "Abstract Deadline: September 22, 2023". A section for "Organizers" lists Joanna Kelley, University of California, Santa Cruz; Pall Melsted, University of Iceland; Nicola Mulder, University of Cape Town, South Africa; and Oliver Stegle, German Cancer Research Center, Germany. A note states that the meeting will begin at 7:30 p.m. on Wednesday, December 6, and finish with lunch on Saturday, December 9, 2023. A scholarship information block mentions the James P. Taylor Foundation for Open Science awarding six scholarships. A "Tentative Meeting Topics" section includes a bulleted list: PanGenome, Genome Assembly and Sequencing Algorithms, Variant Discovery, Microbial and Metagenomics, Single Cell and Spatial Omics, and Functional Genomics. A "Keynote Speakers" section lists Sean Eddy, HHMI/Harvard University, and Karen Miga, University of California, Santa Cruz, Genomics Institute. A "Discussion Leaders" section lists Anne Carpenter, Broad Institute of MIT and Harvard; Rayan Chikhi, Institut Pasteur, France; Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Germany; Ida Moltske, University of Copenhagen, Denmark; Mor Nitzan, Hebrew University of Jerusalem, Israel; Eran Segal, Weizmann Institute of Science, Israel; Jared Simpson, Ontario Institute for Cancer Research, Canada; and Amos Tanay, Weizmann Institute, Israel.

No class on Dec 6

<https://meetings.cshl.edu/meetings.aspx?meet=info&year=23>

Piazza

The screenshot shows the Piazza platform interface. At the top, there's a navigation bar with tabs for Q & A, Resources, Statistics, and Manage Class. Below the navigation bar is a search bar and a 'New Post' button. The main content area displays a note titled 'Welcome to Piazza!' by Michael Schatz. The note content is as follows:

Welcome to Piazza! We'll be conducting all class-related discussion here this term. The quicker you begin asking questions on Piazza (rather than via emails), the quicker you'll benefit from the collective knowledge of your classmates and instructors. We encourage you to ask questions when you're struggling to understand a concept—you can even do so anonymously.

-Michael Schatz

Below the note, there's a section for 'followup discussions, for lingering questions and comments'. It includes a button to 'Start a new followup discussion' and a text input field for 'Compose a new followup discussion'. At the bottom of the note, it says 'Updated 2 days ago by Michael Schatz'.

On the left side, there's a sidebar with sections for 'PINNED' (containing a note about 'Search for Teammates!'), 'LAST WEEK' (containing notes like 'Welcome to Piazza!', 'Introduce Piazza to your stu...', 'Get familiar with Piazza', and 'Tips & Tricks for a successf...'), and a general 'Welcome to Piazza!' message. The sidebar also includes links for 'LIVE Q&A', 'Drafts', and various assignment categories like hw1, hw2, hw3, hw4, hw5, project, exam, logistics, and other.

<https://piazza.com/jhu/fall2023/600449600649>

GradeScope

gradescope by Turnitin

EN.601.449/EN.601.649 Fall 2023
Course ID: 587880

Description: Applied Comparative Genomics

Things To Do: Create your first assignment from the [Assignments](#) page.

Active Assignments	Released	Due (EDT)	Submissions	% Graded	Published	Regrades
--------------------	----------	-----------	-------------	----------	-----------	----------

You currently have no assignments.
Create an assignment to get started.

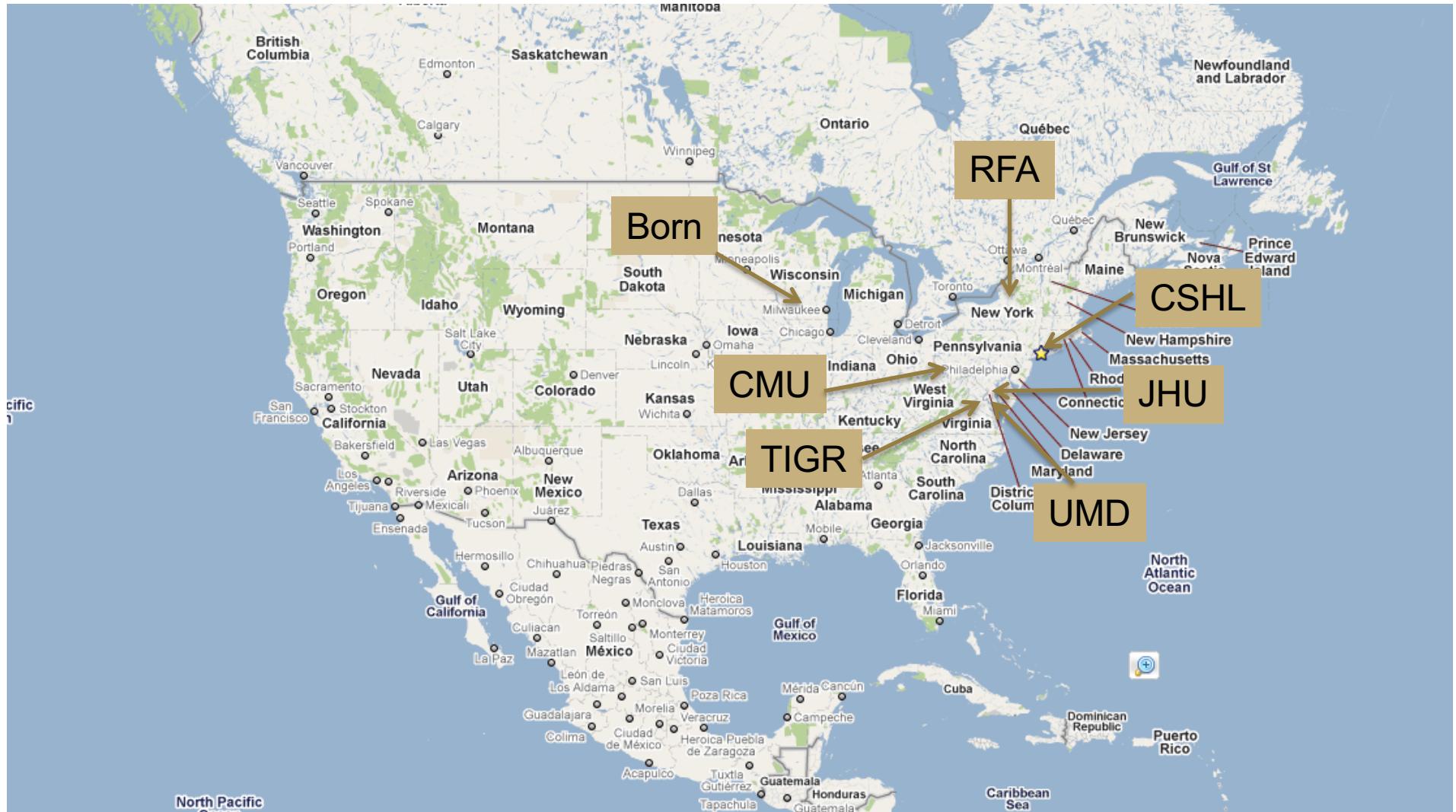
Create Assignment

Account

?

<https://www.gradescope.com/>
Entry Code: **JK5VB4**

A Little About Me



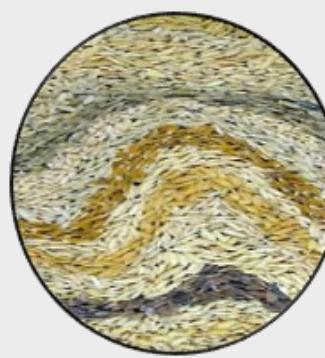
Schatzlab Overview



Human Genetics

Role of mutations
in disease

Nurk *et al.* (2022)
Aganezov *et al.* (2020)



Agricultural Genomics

Genomes &
Transcriptomes

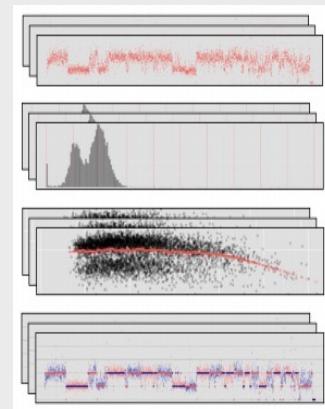
Naish *et al.* (2021)
Alonge *et al.* (2020)



Algorithmics & Systems Research

Ultra-large scale
biocomputing

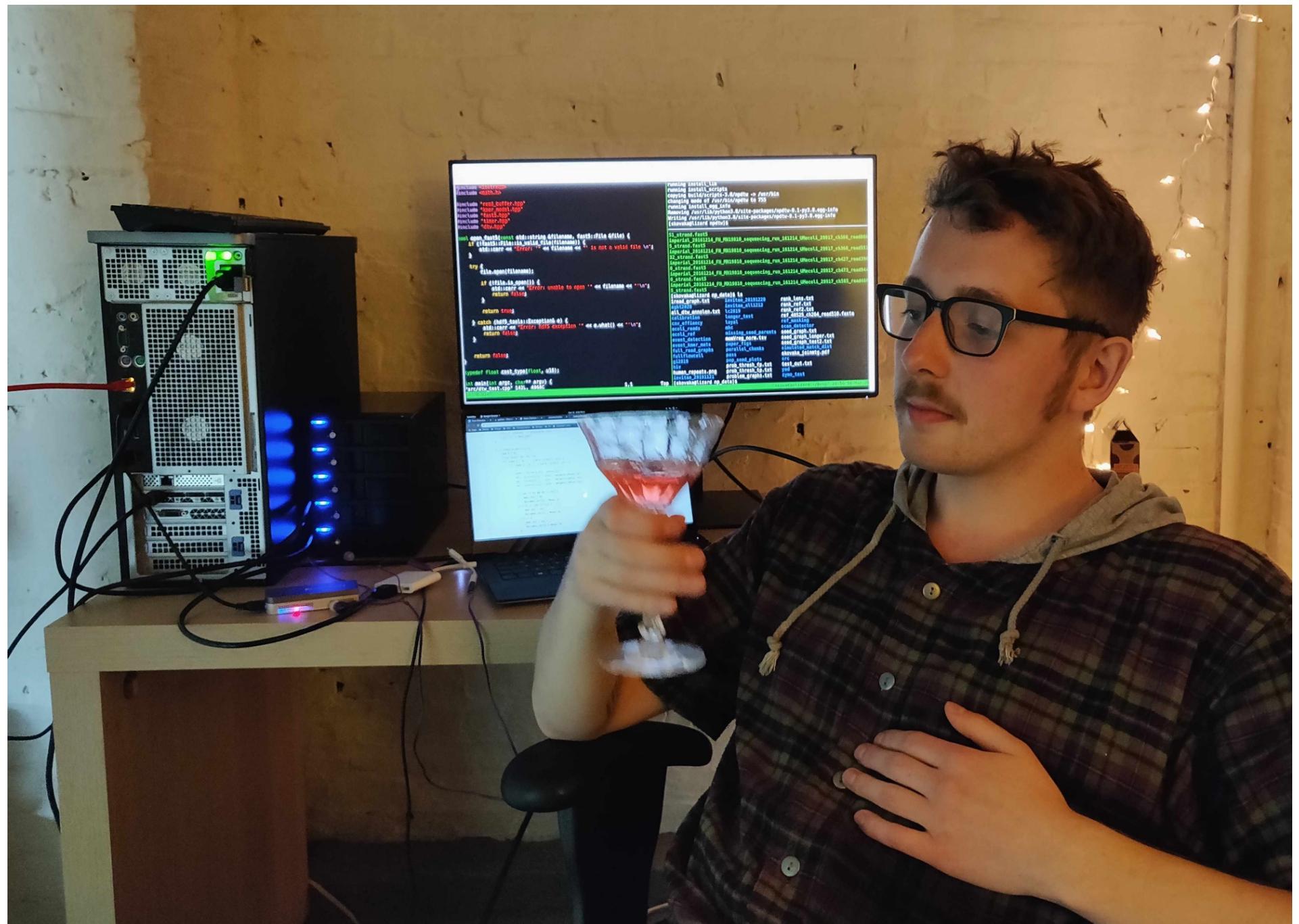
Kirsche *et al.* (2023)
Schatz *et al.* (2022)



Biotechnology Development

Single Cell + Single
Molecule Sequencing

Rozowsky *et al.* (2023)
Kovaka *et al.* (2020)



Targeted nanopore sequencing | nature.com/articles/s41587-020-0731-9

nature biotechnology

Explore content ▾ Journal information ▾ Publish with us ▾ Subscribe View all Nature Research journals Search My Account

nature > nature biotechnology > articles > article

Article | Published: 30 November 2020

Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED

Sam Kovaka✉, Yunfan Fan, Bohan Ni, Winston Timp & Michael C. Schatz

Nature Biotechnology (2020) | Cite this article

5715 Accesses | 2 Citations | 261 Altmetric | Metrics

Abstract

Conventional targeted sequencing methods eliminate many of the benefits of nanopore sequencing, such as the ability to accurately detect structural variants or epigenetic modifications. The ReadUntil method allows nanopore devices to selectively eject reads from pores in real time, which could enable purely computational targeted sequencing. However, this requires rapid identification of on-target reads while most mapping methods require computationally intensive basecalling. We present UNCALLED (<https://github.com/skovaka/UNCALLED>), an open source mapper that rapidly matches streaming of nanopore current signals to a reference sequence. UNCALLED probabilistically identifies reads that fall within a user-specified threshold of the total signal amplitude.

You have full access to this article via Johns Hopkins Libraries

Download PDF

Sections Figures References

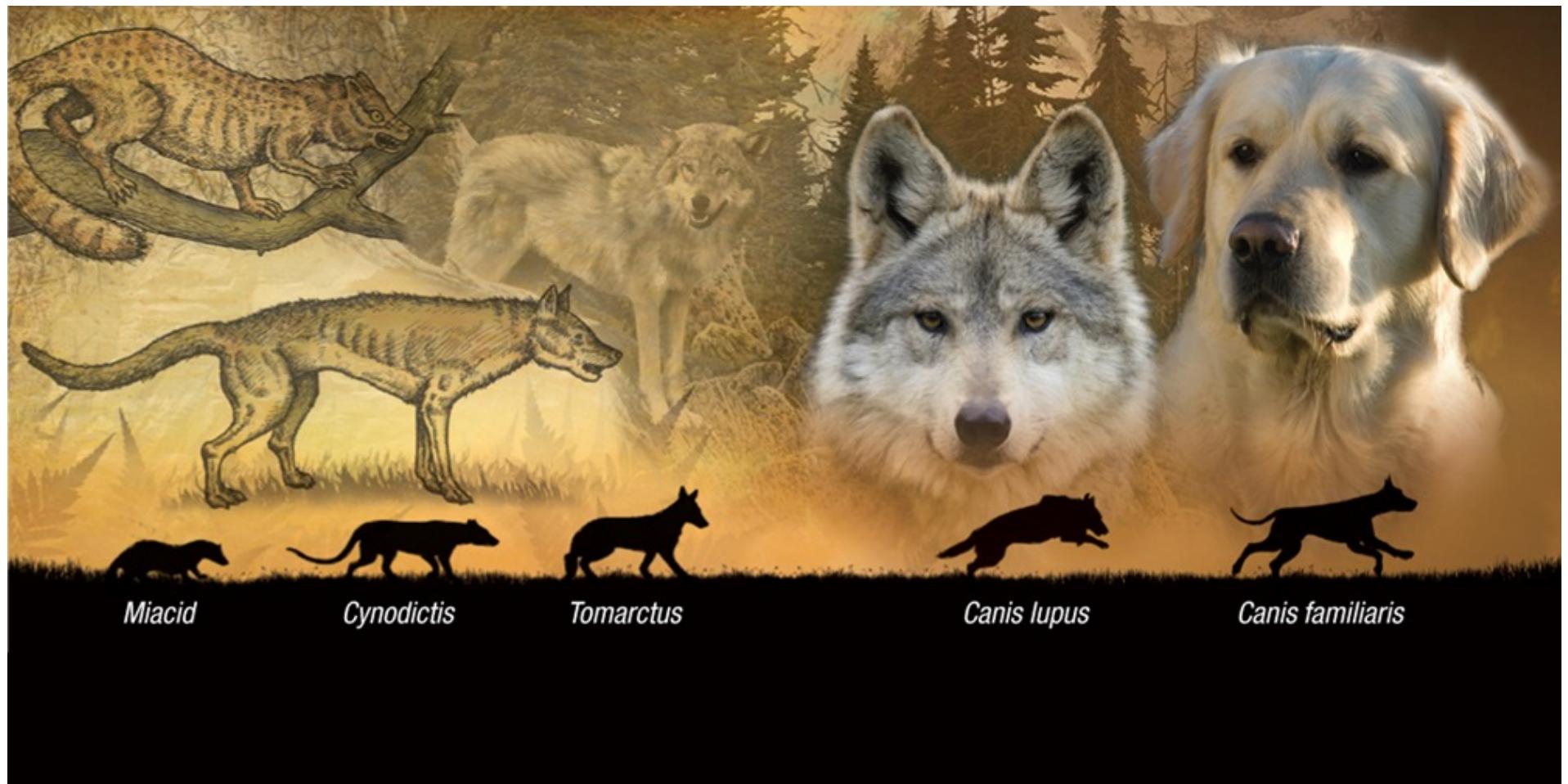
Abstract Main Results Discussion Methods Data availability Code availability References Acknowledgements Author information Ethics declarations

Why Genomics?

Earliest Genomics

Any Guesses?

Earliest Genomics



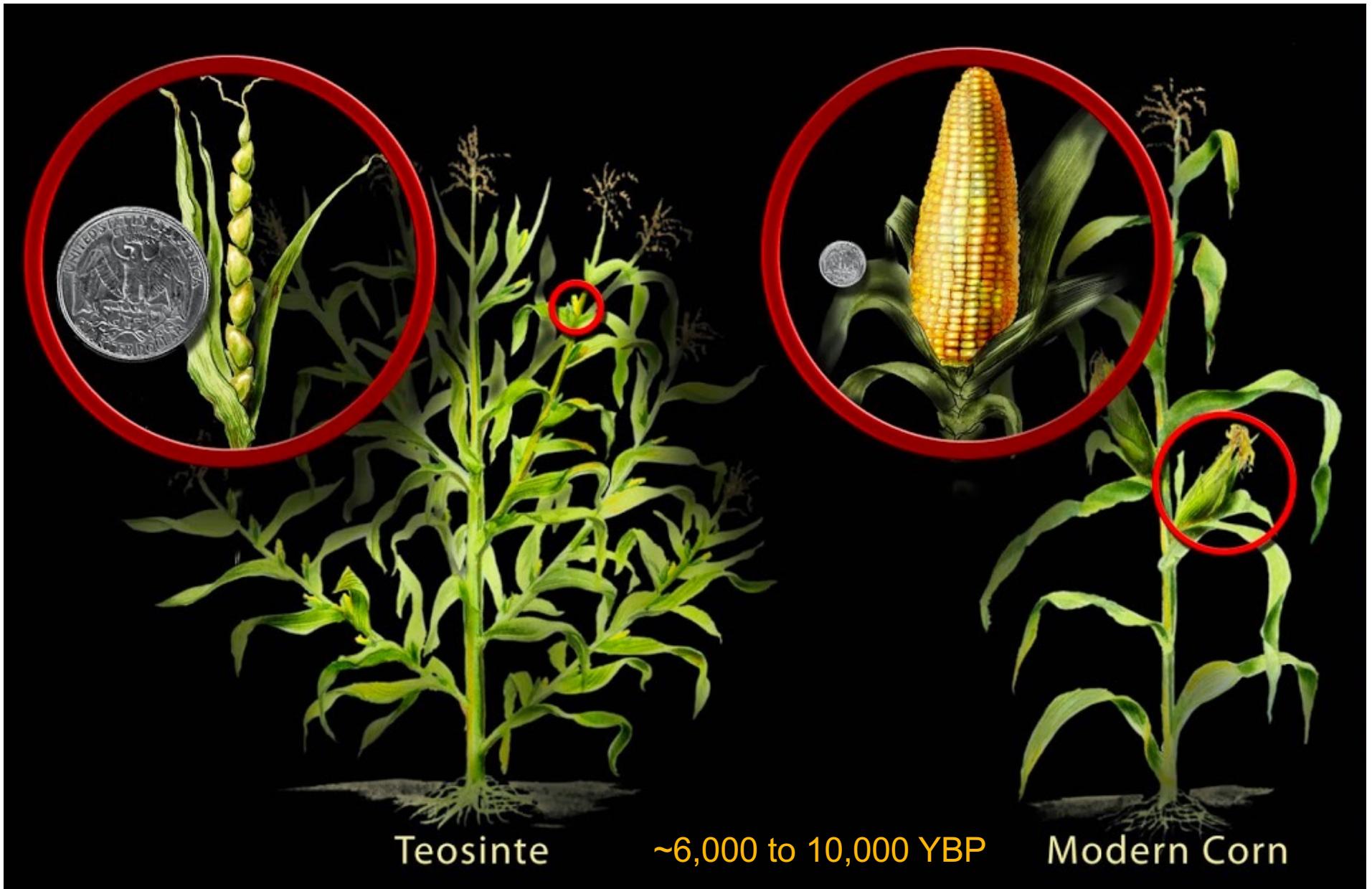
15,000 to 35,000 YBP

Earliest Genomics



~1,000 to 10,000 YBP

Earliest Genomics



Discovery of Chromosomes

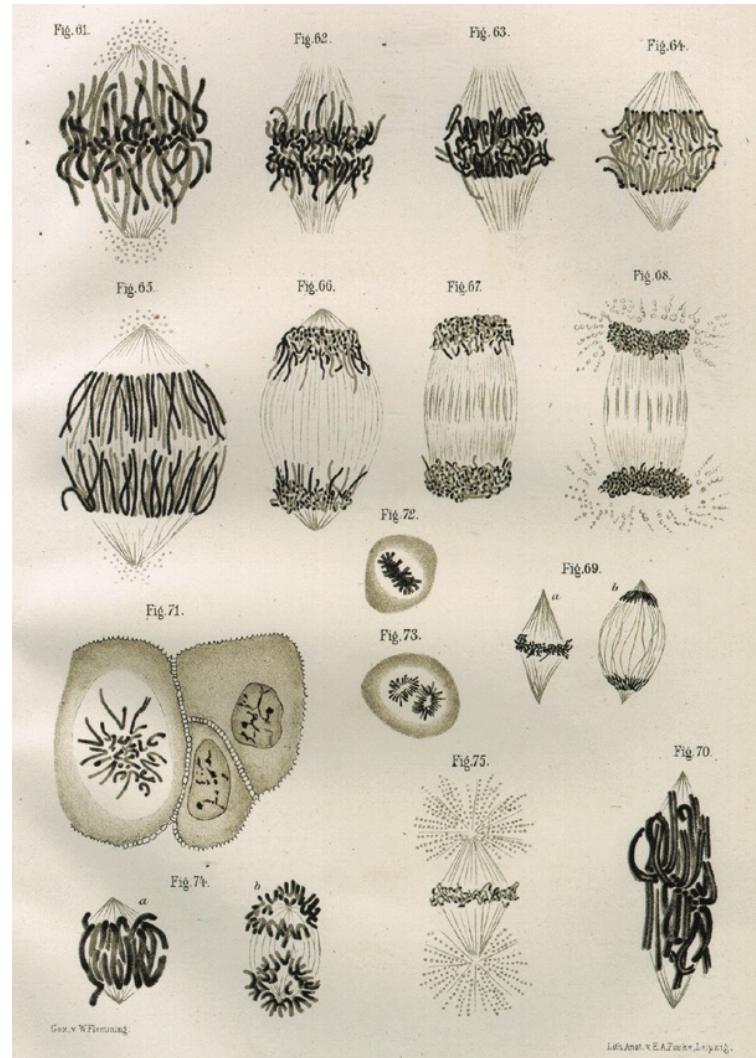
By the mid-1800s, microscopes were powerful enough to observe the presence of unusual structures called “chromosomes” that seemed to play an important role during cell division.

It was only possible to see the chromosomes unless appropriate stains were used

“Chromosome” comes from the Greek words meaning “color body”

Today, we have much higher resolution microscopes, and a much richer varieties of dies and dying techniques so that we can visualize particular sequence elements.

When you see something unexpected that you think might be interesting, give it a name



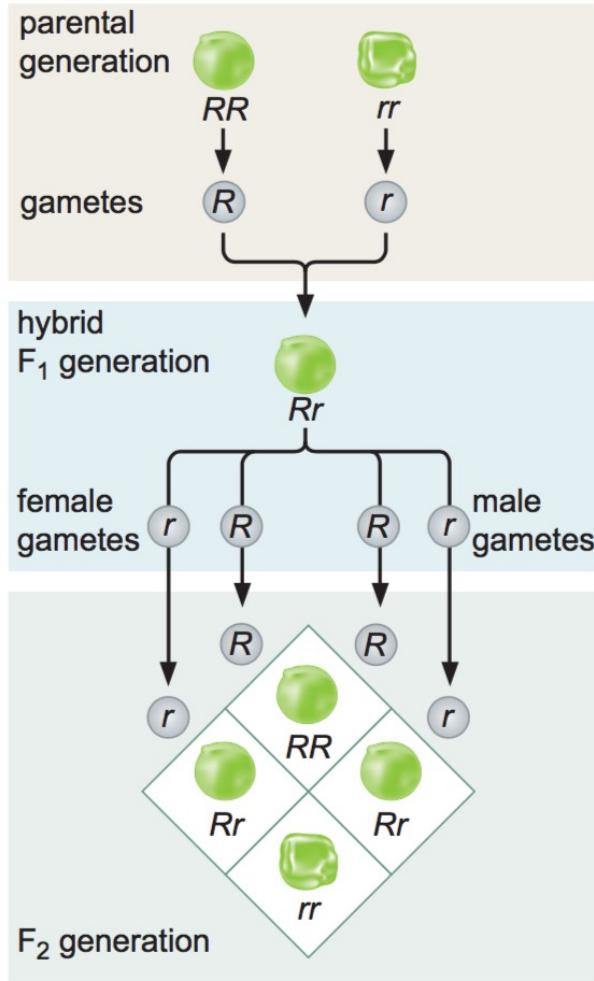
Drawing of mitosis by Walther Flemming.

Flemming, W. Zellsubstanz, Kern und Zelltheilung (F. C. W. Vogel, Leipzig, 1882).

The “first” quantitative biologist

Any Guesses?

Laws of Inheritance



Seed		Flower		Pod		Stem	
Form	Cotyledons	Color		Form	Color	Place	Size
Grey & Round	Yellow	White		Full	Yellow	Axial pods, Flowers along	Long (6-7ft)
White & Wrinkled	Green	Violet		Constricted	Green	Terminal pods, Flowers top	Short (<1ft)
1	2	3		4	5	6	7

http://en.wikipedia.org/wiki/Experiments_on_Plant_Hybridization

Observations of 29,000 pea plants and 7 traits

Generation	in Verhältniss gestellt:			
	A	Aa	a	$A : Aa : a$
1	1	2	1	1 : 2 : 1
2	6	4	6	3 : 2 : 3
3	28	8	28	7 : 2 : 7
4	120	16	120	15 : 2 : 15
5	496	32	496	31 : 2 : 31
n				$2^n - 1 : 2 : 2^n - 1$

Versuche über Pflanzen-Hybriden. Verh. Naturforsch (Experiments in Plant Hybridization)
Mendel, G. (1866). Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32).

The first genetic map

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***



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

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 located closest together



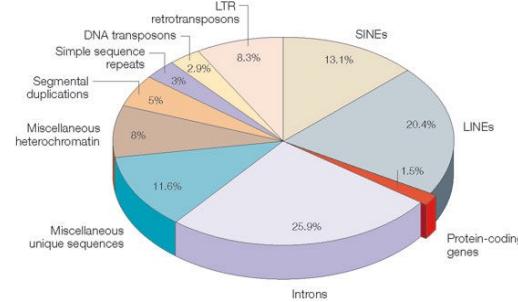
The Linear Arrangement of Six Sex-Linked Factors in Drosophila as shown by their mode of Association
Sturtevant, A. H. (1913) *Journal of Experimental Zoology*, 14: 43-59

Jumping Genes



Previously, genes were considered to be stable entities arranged in an orderly linear pattern on chromosomes, like beads on a string

Careful breeding and cytogenetics revealed that some elements can move (cut-and-paste, DNA transposons) or copy itself (copy-and-paste, retrotransposons)



(Gregory, 2005, Nature Reviews Genetics)

(Much) later analysis revealed that nearly 50% of the human genome is composed of transposable elements, including LINE and SINE elements (long/short interspersed nuclear elements) which can occur in 100k to 1M copies

“The genome is a graveyard of ancient transposons”

The origin and behavior of mutable loci in maize.

McClintock, B. (1950) PNAS. 36(6):344–355.

Nobel Prize in Physiology or Medicine in 1983

Discovery of the Double Helix

No. 4356 April 25, 1953

NATURE

737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

¹ Young, F. B., Gerard, H., and Jeavons, W. *Phil. Mag.*, **40**, 149 (1920).

² Lonsdale, Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Suppl.* **5**, 285 (1949).

³ Von Arx, W. S., Woods Hole Papers in Phys. Oceanogr. Meteor., **11**, 131 (1960).

⁴ Ekman, V. W., *Arkiv. Mat. Astron. Fysik.* (Stockholm), **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey¹. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diffraction pattern of the fibre must be the acidic hydroxyl atoms; it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains, each linked end-to-end along the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β -D-deoxyribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There

is a residue on each chain every 3.4 Å. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 Å. The distance of a phosphorus atom from the fibre axis is 10 Å. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The outstanding feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-coordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

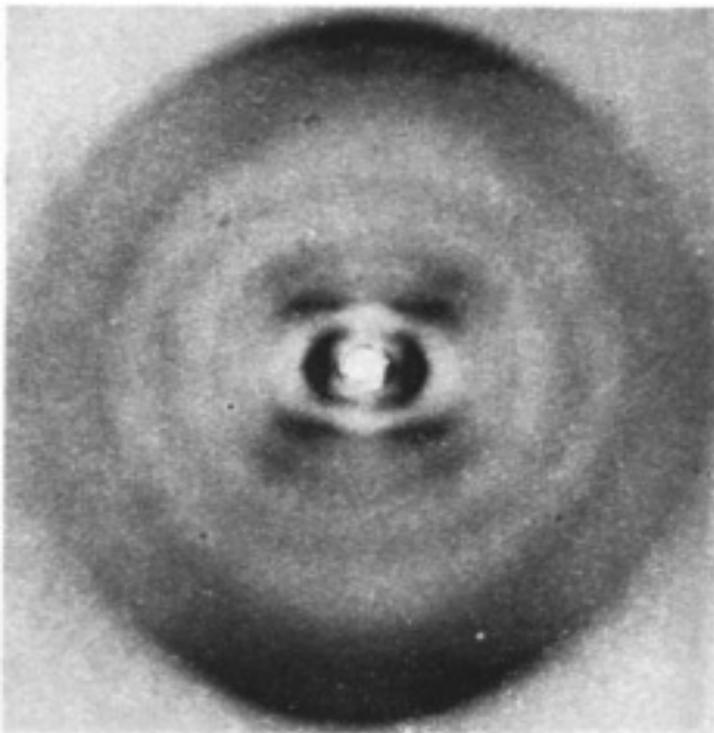
It has been found experimentally^{2,3} that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data^{4,5} on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the con-



It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the con-

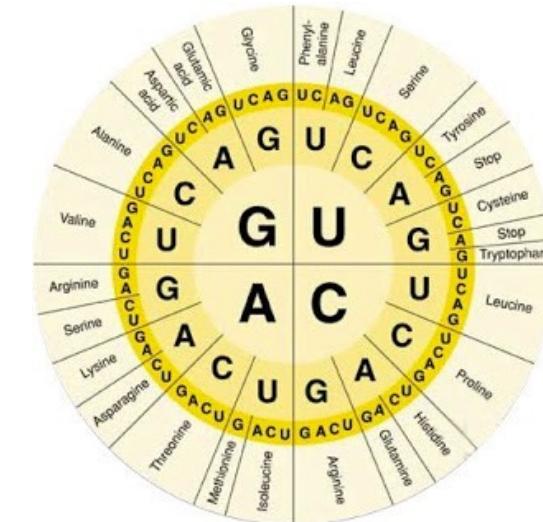
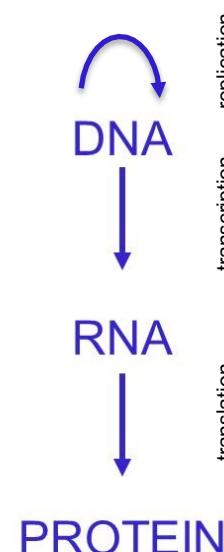
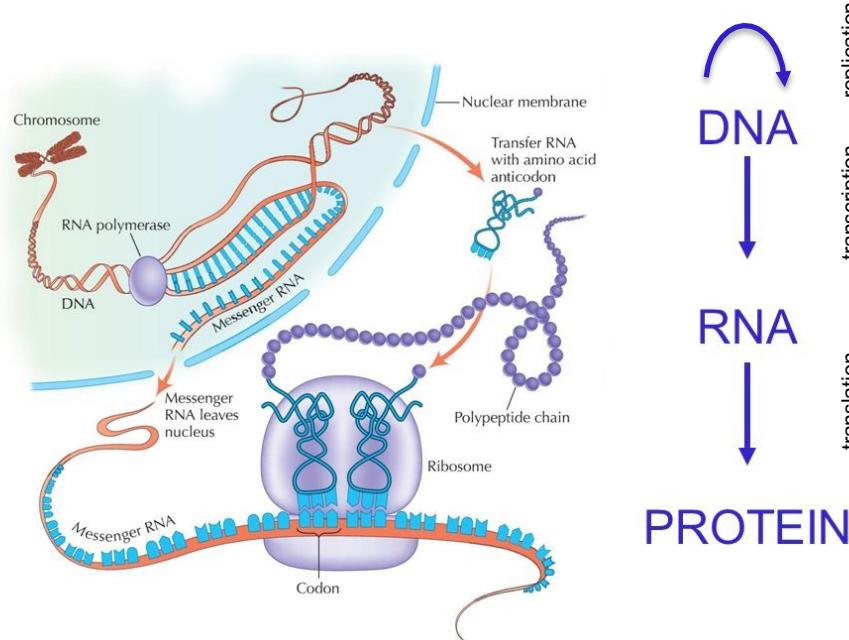
Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid

Watson JD, Crick FH (1953). Nature 171: 737–738.

Nobel Prize in Physiology or Medicine in 1962

Central Dogma of Molecular Biology

“Once ‘information’ has passed into protein it cannot get out again. In more detail, the transfer of information **from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible**, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein”

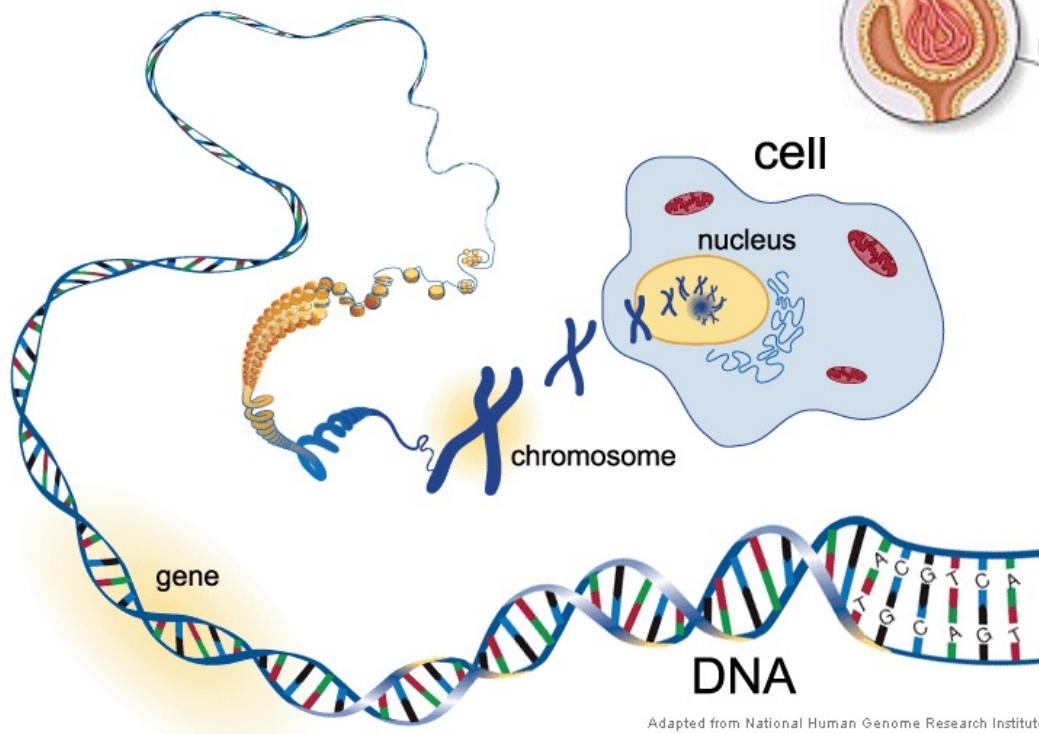
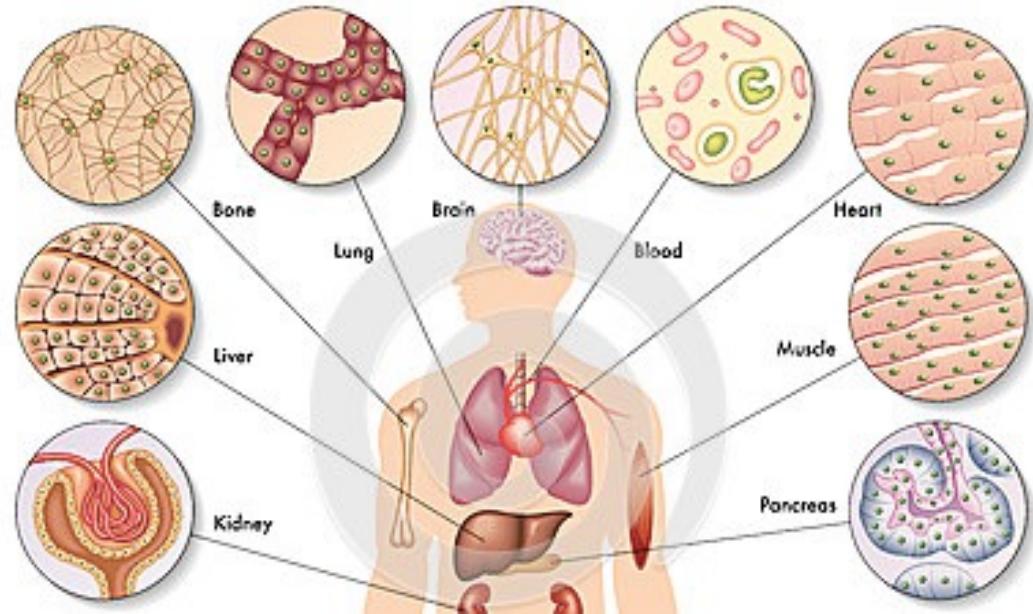


On Protein Synthesis

Crick, F.H.C. (1958). *Symposia of the Society for Experimental Biology* pp. 138–163.

One Genome, Many Cell Types

Each cell of your body contains an exact copy of your 3 billion base pair genome.



Adapted from National Human Genome Research Institute

Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

Unsolved Questions in Biology

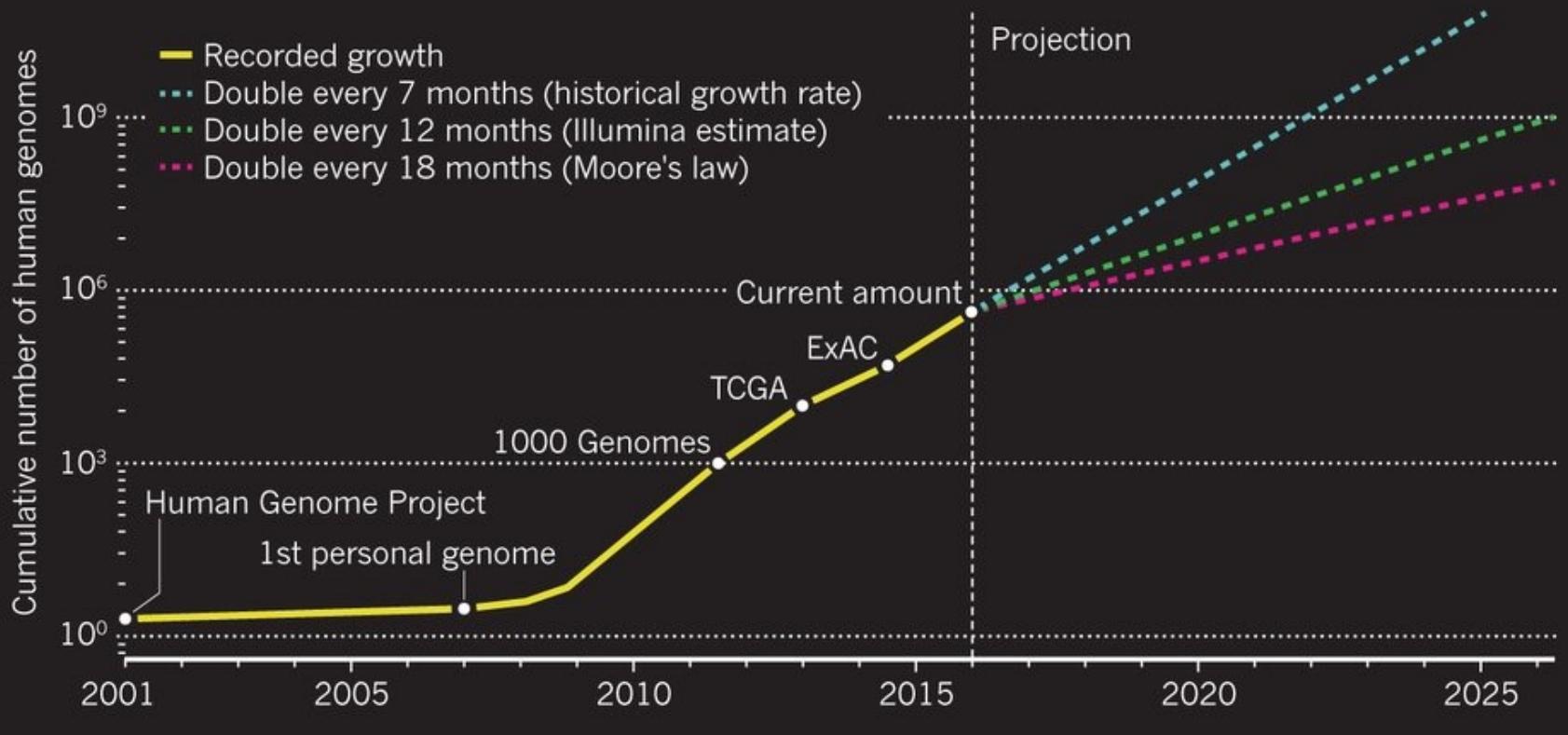
- What is your genome sequence?
- How does your genome compare to my genome?
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?
- How does methylation change during development?
- How does chromatin change during development?
- How does your genome folded in the cell?
- Where do proteins bind and regulate genes?
- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- What drugs and treatments should we give you?
- ***Plus thousands and thousands more***



Sequencing Capacity

DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TCGA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.



Big Data: Astronomical or Genomical?

Stephens, Z, et al. (2015) PLOS Biology DOI: [10.1371/journal.pbio.1002195](https://doi.org/10.1371/journal.pbio.1002195)

Sequencing Capacity

DNA SEQUENCING SOARS

Human
aggreg
the Ex
three p

Cumulative number of human genomes

10^9

10^6

10^3

10^0

2001

2003

2005

2007

2009

2011

The instruments provide the data, but
none of the answers to any of these
questions.

What software and systems will?

And who will create them?

Big Data: Astronomical or Genomical?

Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195