Johns Hopkins University

Johns Hopkins University
Center for Computational Biology
CCB

Diversifying Genomics

Identifying large variations in genomes of African ancestry individuals

Rachel Sherman

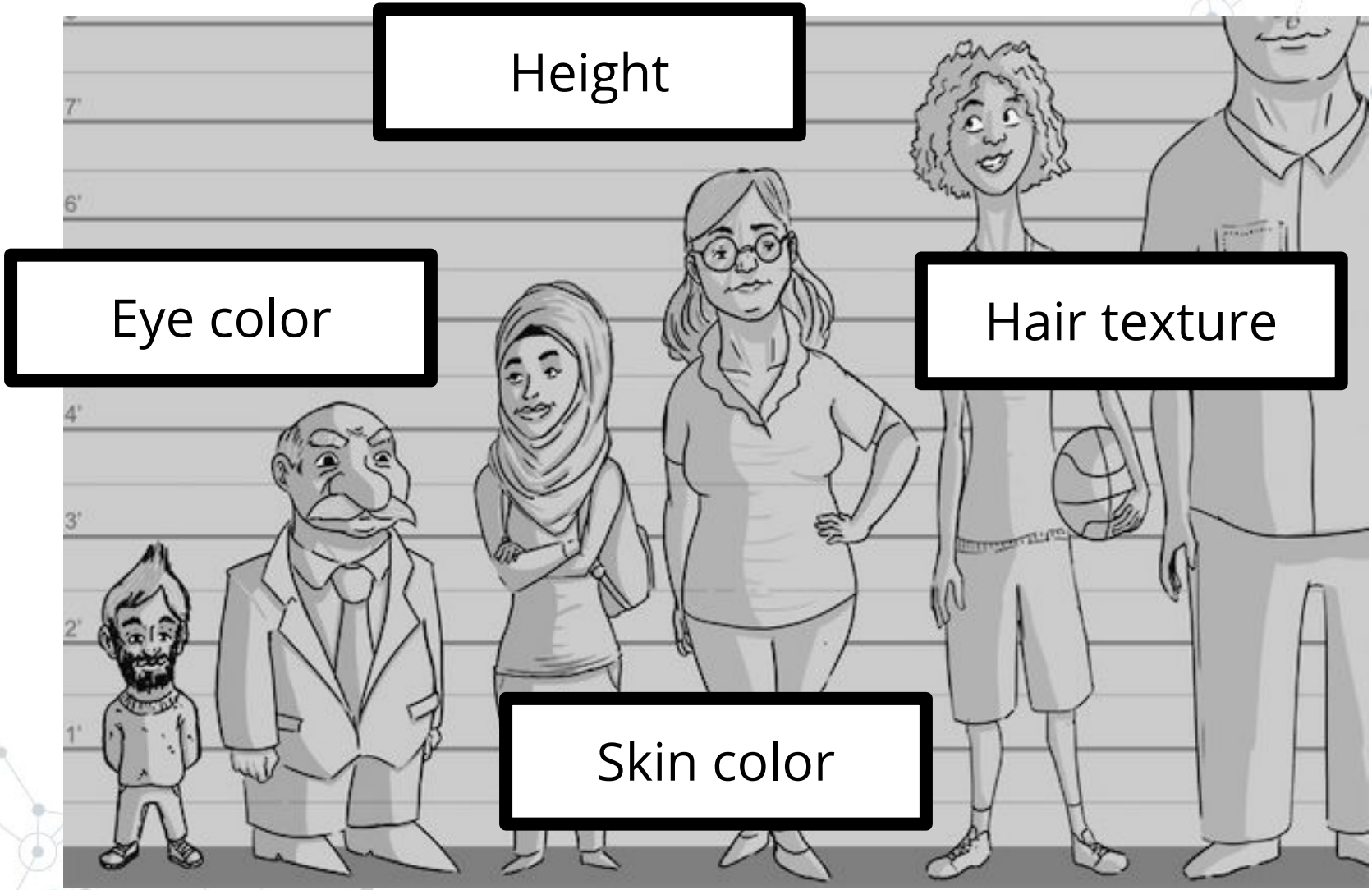
Johns Hopkins University

March 1, 2019

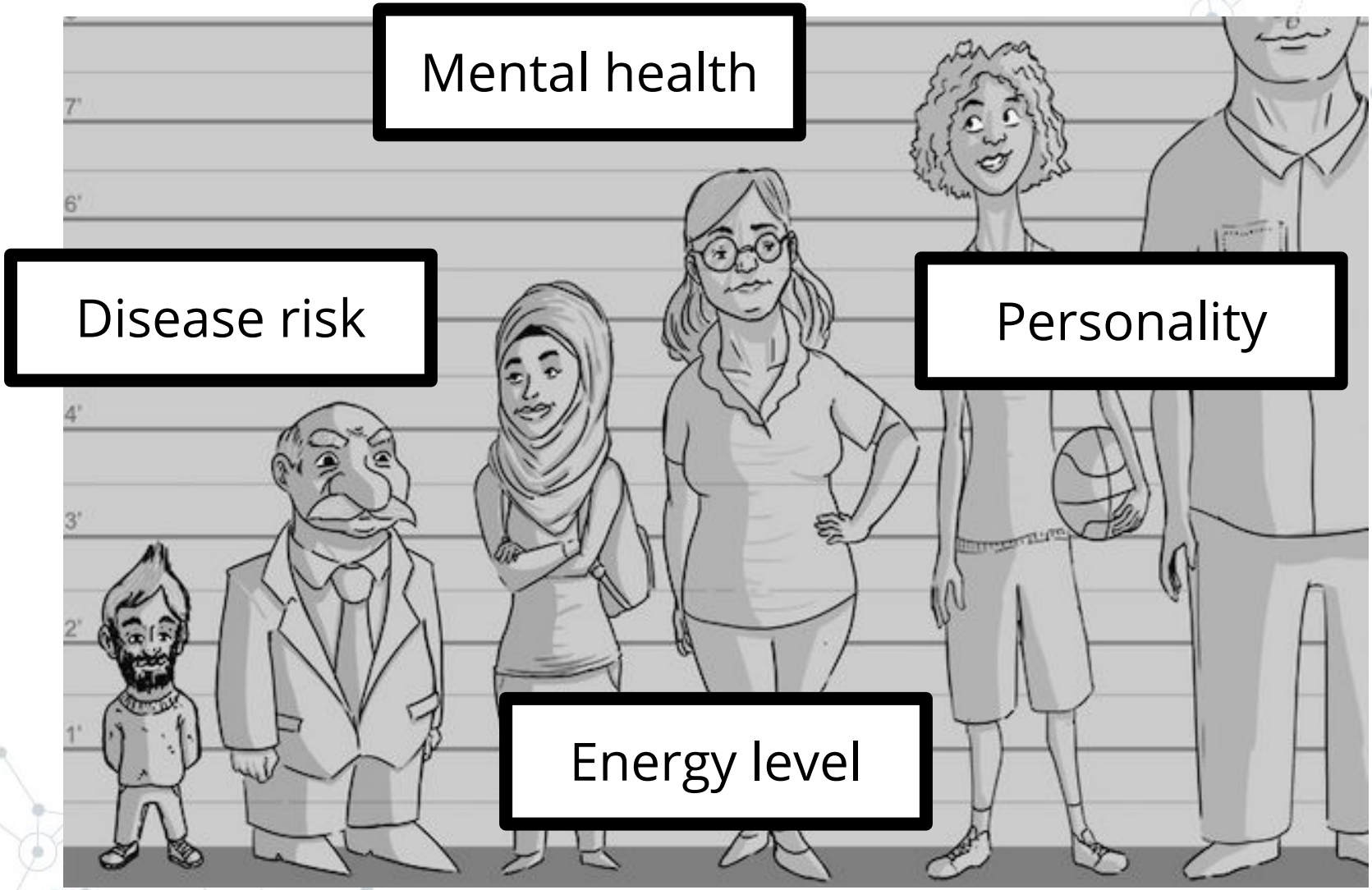
Sources and types of variation

Image source: https://steemit.com/science/@sallyquin/variations-in-population-2

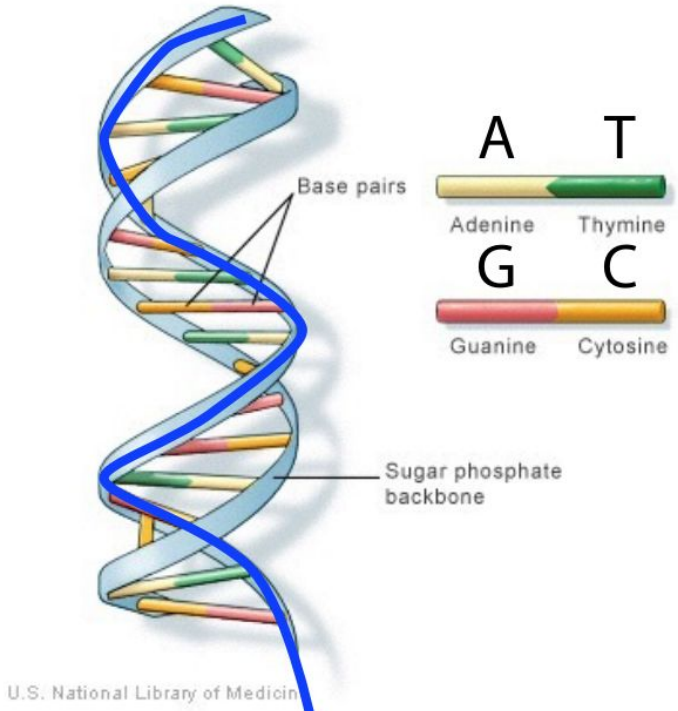
Sources and types of variation



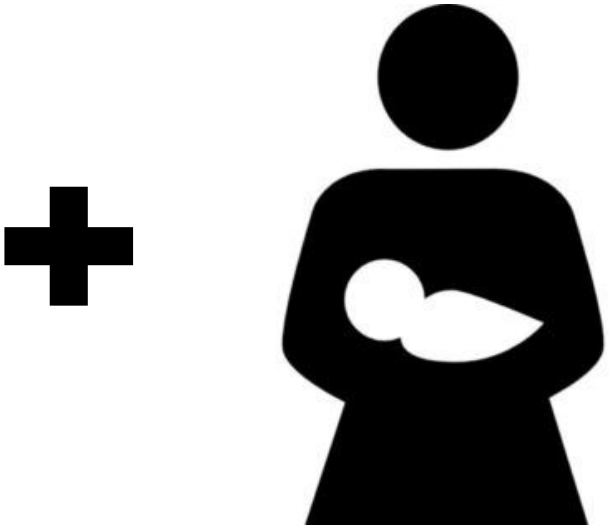
Sources and types of variation



Sources and types of variation



TCACACTGAGCGTGCTG



The human genome project

The Buffalo News/Sunday, March 23, 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 CANCER INSTITUTE

For more information please contact the
Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

The human genome project


The Buffalo News/Sunday, March 23, 1997

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

~3 billion bases

23 chromosome pairs

Can only “read” 500 bp at once

$$$ 3 billion dollars $$$

The human genome project


The Buffalo News/Sunday, March 23, 1997

WANTED

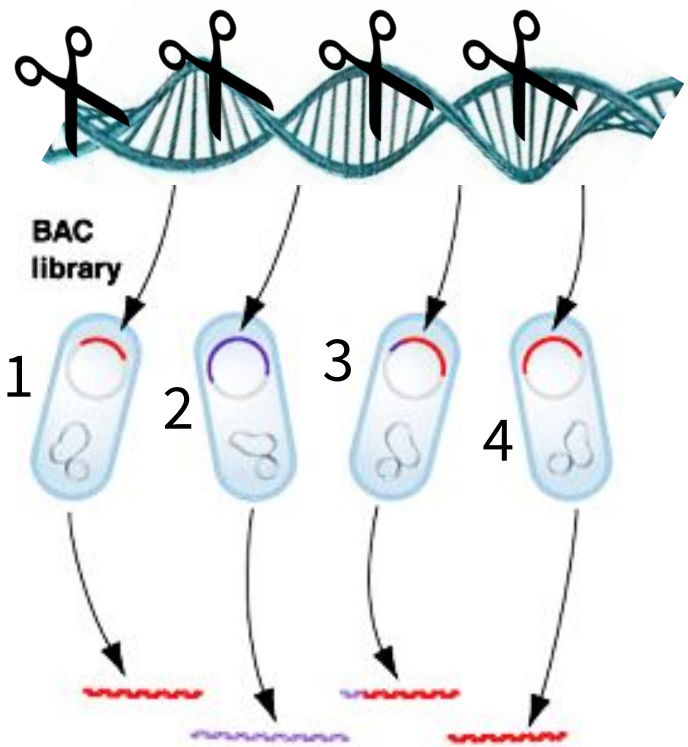
20 Volunteers

to participate in the

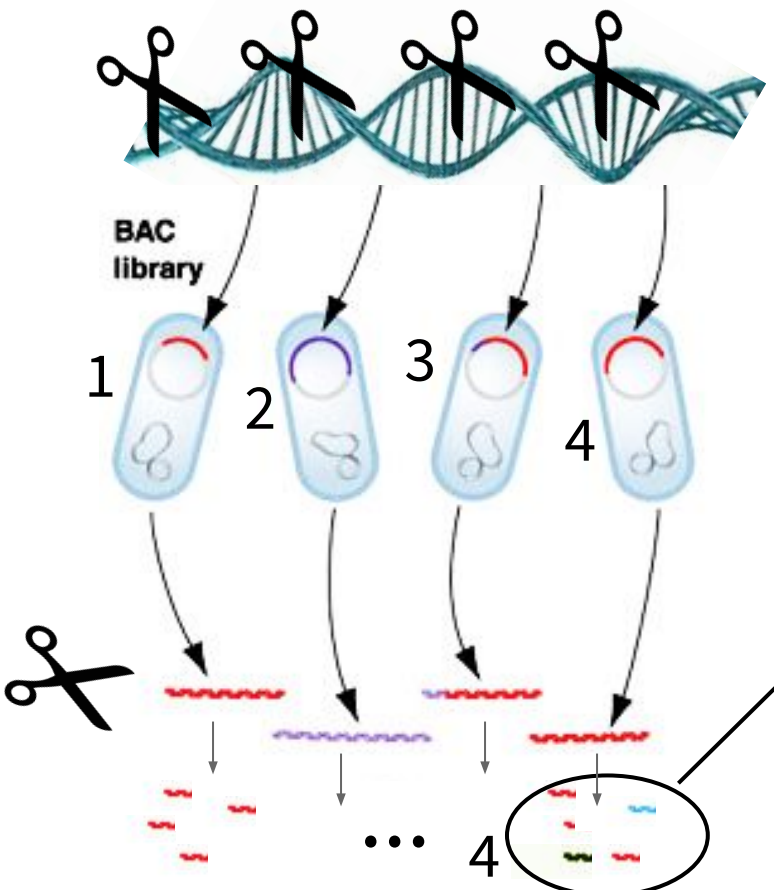
Human Genome Project

a very large international scientific research effort.

The human genome project



The human genome project

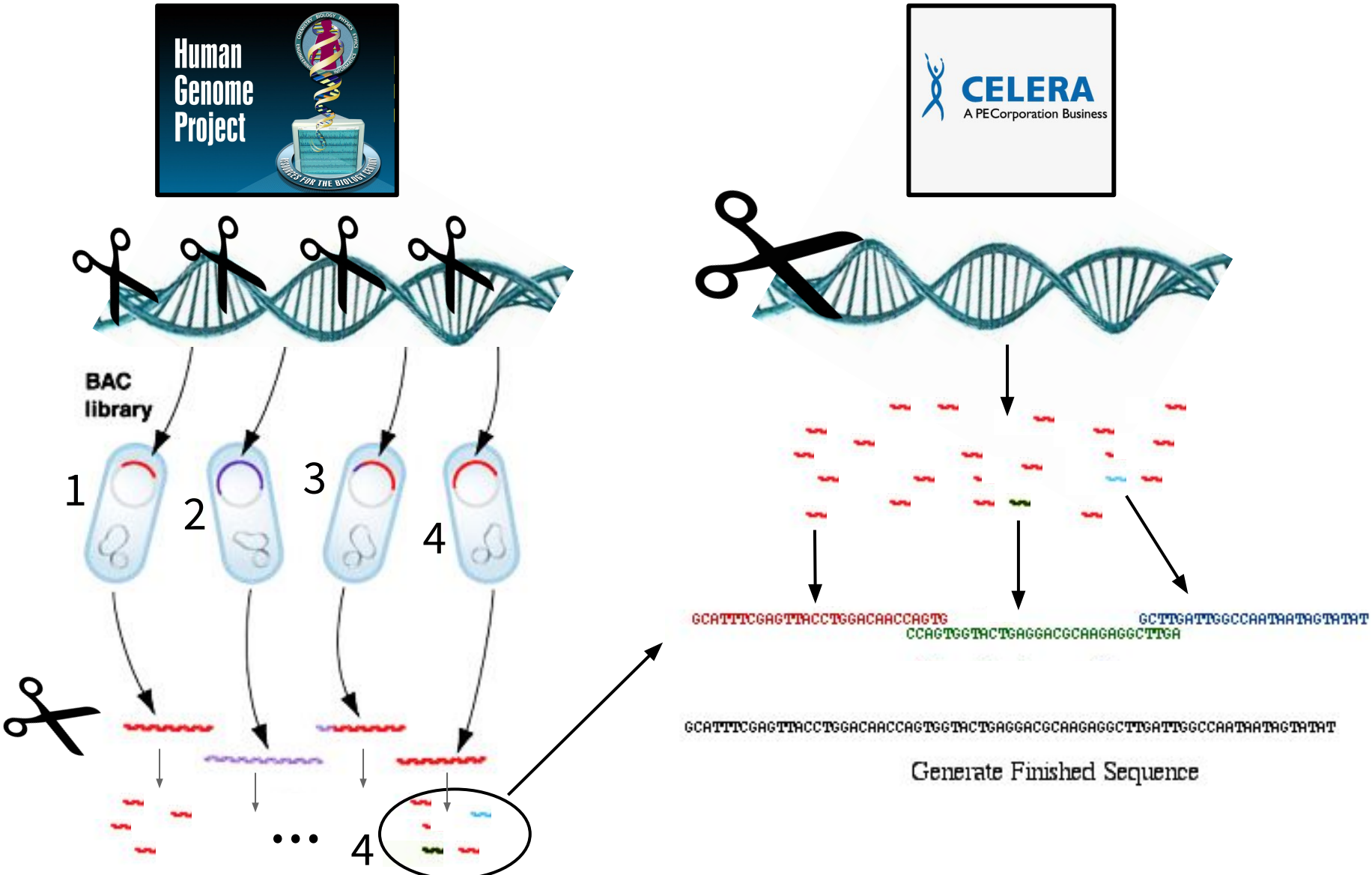


GCATTTCGAGTTACCTGGACAACCAGTG GCTTGATTGGCCAATAATAGTATAT
CCAGTGGTACTGAGGACGCAAGAGGCTTGA

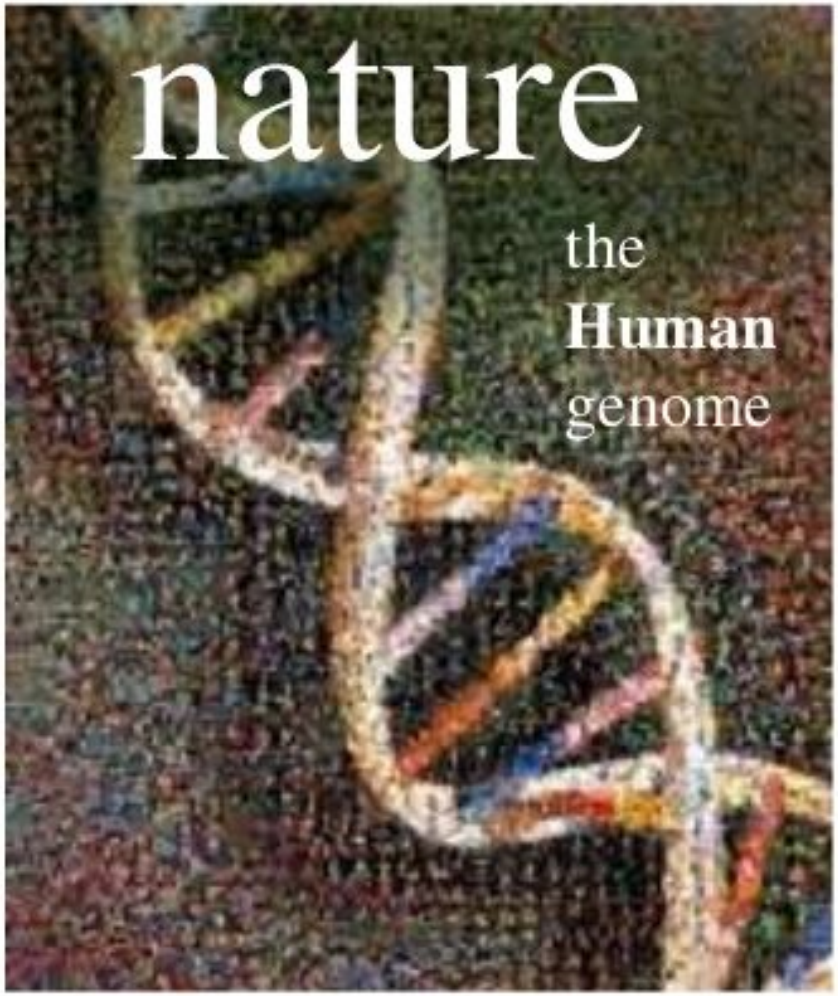
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT

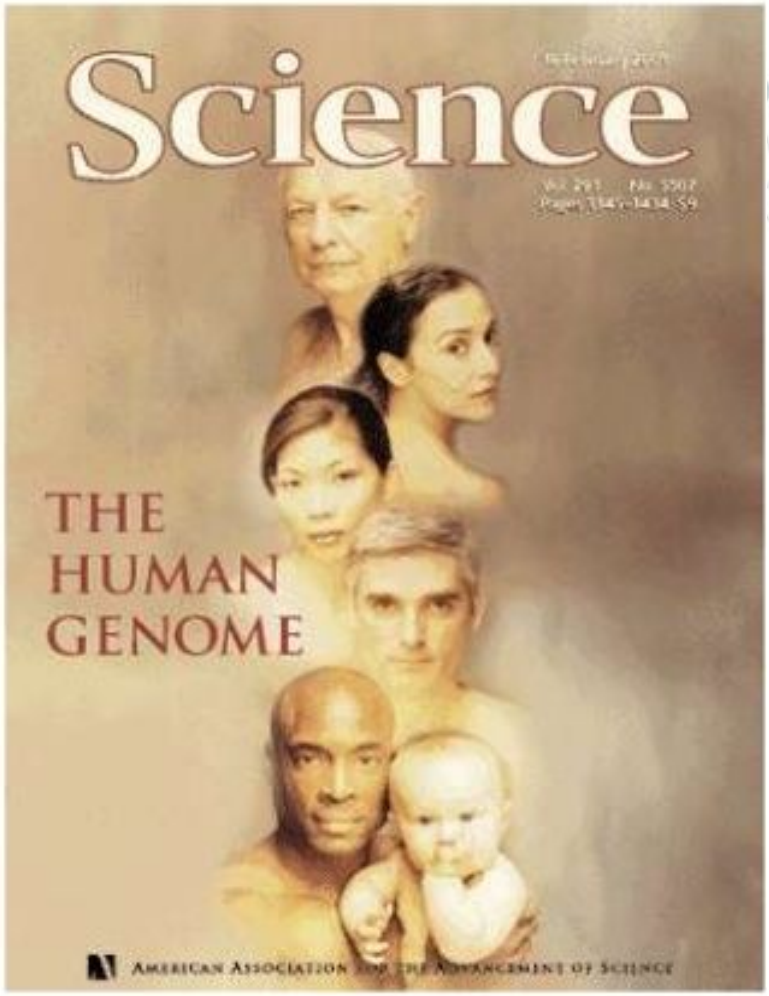
Generate Finished Sequence

The human genome project



The human genome project





2001

The human genome project

SECTIONS HOME SEARCH The New York Times

ARCHIVES | 2003

Once Again, Scientists Say Human Genome Is Complete

By NICHOLAS WADE APRIL 15, 2003

The human genome is complete and the Human Genome Project is over, leaders of a public consortium of academic centers said today.

2003

The human genome project


The Buffalo News/Sunday, March 23, 1997

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The human genome project


The Buffalo News/Sunday, March 23, 1997

WANTED

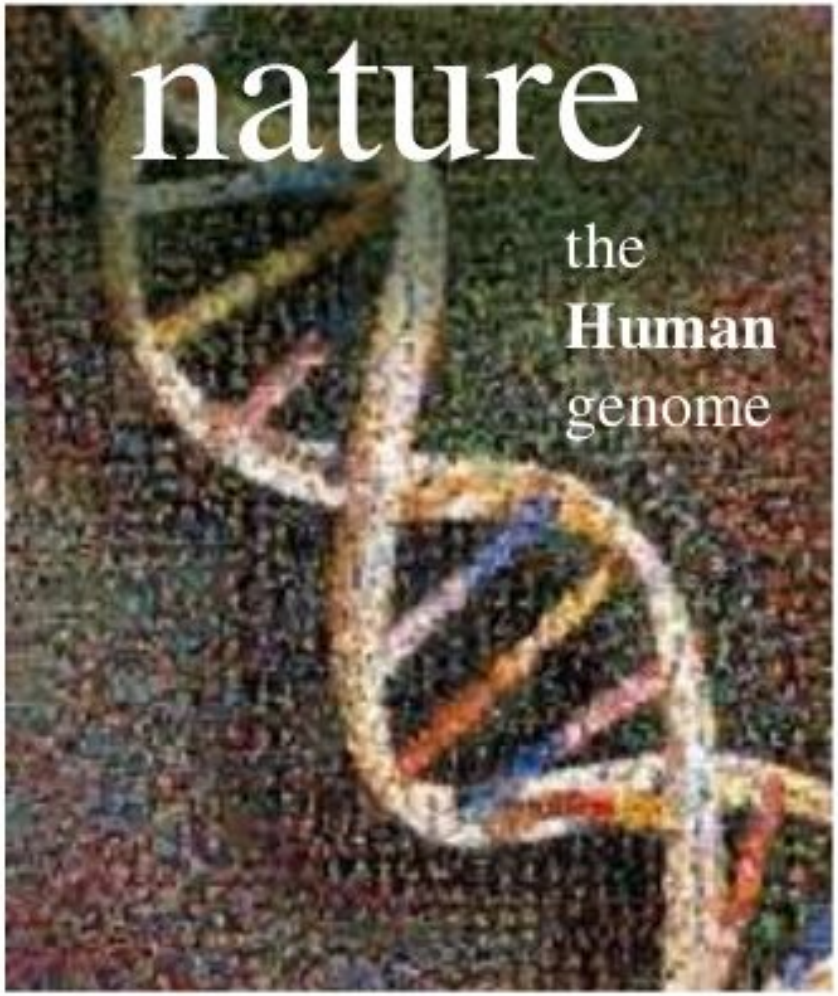
20 Volunteers

Human Genome Project

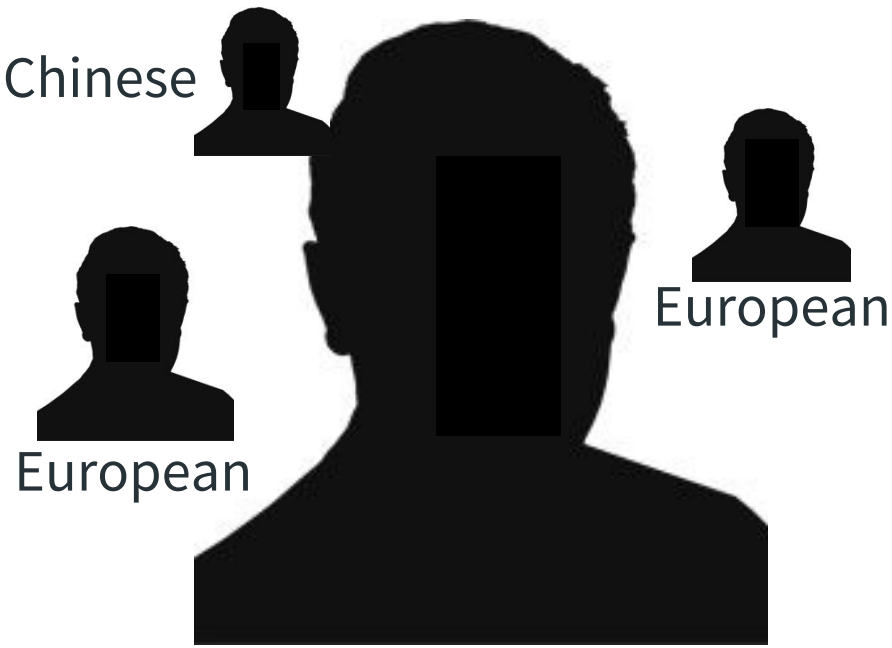
a very large international scientific research effort.

The “reference” genome

“RP-11”



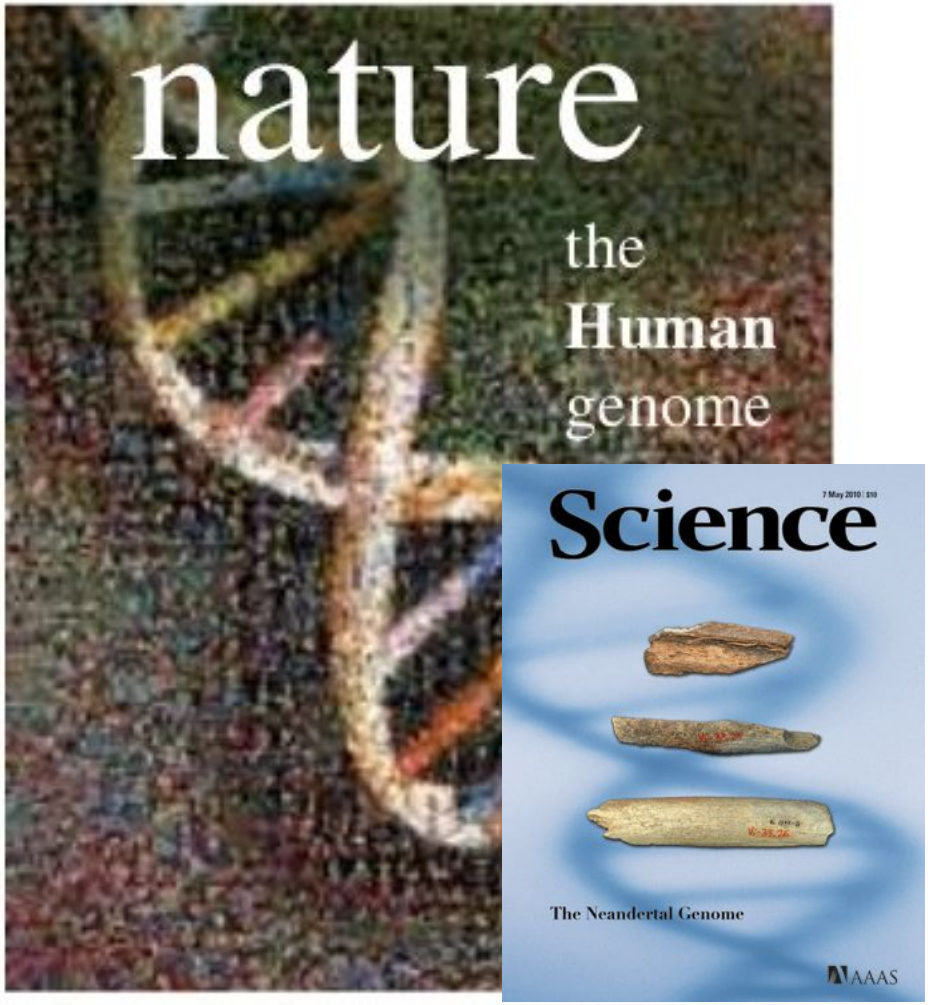
The “reference” genome



RP-11

half African, half European

70% of the reference



2010

The utility of the reference genome

All humans are ~99.9% similar on a DNA level

CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTG
GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAA**G**TGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACGCTTGATGTTTT
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCTTTTTTTTTCTTC**T**CCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAATCATTATACATAT
TTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAATTTTAAAA
AATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT
CTGCATATAAATATTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTA
CCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCA
TACCTCTTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC
CCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTT
TCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTT
GAGCATCTGGATTCTGCCTAATAAAAAACAT**T**TATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTA
AAAAGGGAATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATA

On average unrelated people have a 1 base difference every 1,000 bases.

What does genomic variation look like?

Reference ...ATCGGAATAGCGAGTA...

Person of interest ATGGGAATAGCTAGTA

Reference ...ATCGGAATAGCGAGTA...

Person of interest ATCGGAATAGTA

Reference ...ATCGGAATAGCGAGTA...

Person of interest ATCTCAGGAATAGCGAGTA

What does genomic variation look like?

SNP

```
.ATCGGAATAGCGAGTA...
ATGGGAATAGCTAGTA
```

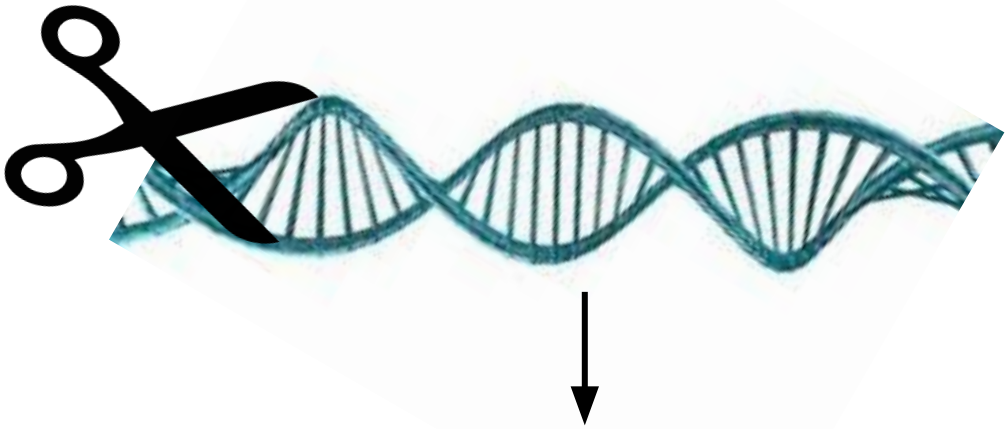
Deletion

```
.ATCGGAATAGCGAGTA...
ATCGGAATAG     TA
```

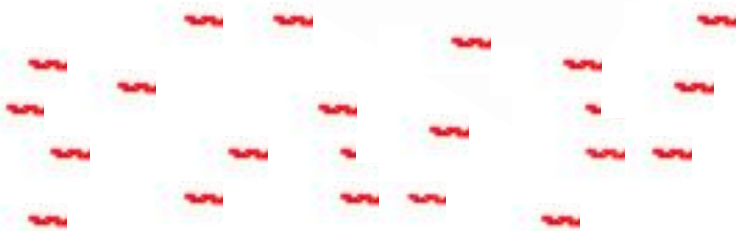
Insertion

```
.ATC    GGAATAGCGAGTA...
ATCTCAGGAATAGCGAGTA
```

Analyzing genomes: alignment



Chop up genome



Line pieces up to reference genome to find where they go

Human reference genome

Analyzing genomes: alignment

Analyzing genomes: alignment

```
                ATGCGCCC                GGTATAC...
...CCATAG    TATGCGCC    CGGAAATTT
...CCAT   CTATATGCG     TCGGAAATT  CGGTATAC
...CCAT GGCTATATG    CTATCGGAAA   GCGGTATA
...CCA AGGCTATAT    CCTATCGGA   TTGCGGTA  C...
...CCA AGGCTATAT  GCCCTATCG    TTTGCGGT
...CC  AGGCTATAT   CCCTATCG AAATTTGC   ATAC...
...CC TAGGCTATA GCGCCCTA   GAAATTTG  GTATAC...
```

```
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

Human reference genome

Analyzing genomes: alignment

```
                ATGCGCCC                GGTATAC...
...CCATAG    TATGCGCC    CGGAAATTT
...                     TCGGAAATT  CGGTATAC
...                     TCGGAAA   GCGGTATA
...                     TCGGA   TTGCGGTA  C...
...                     TCG    TTTGCGGT
...                     TCG AAATTTGC   ATAC...
...CC TAGGCTATA GCGCCCTA   GAAATTTG  GTATAC...
```

We found a SNP!

```
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

The utility of the reference genome

From the National Institutes of Health website:

- The Human Genome Project has already fueled the **discovery of more than 1,800 disease genes**.
- There are **now more than 2,000 genetic tests for human conditions**. These tests enable patients to learn their genetic risks for disease and also help healthcare professionals to diagnose disease.
- In 2010, the third phase of the HapMap project was published, with data from 11 global populations. **HapMap data have accelerated the search for genes involved in common human diseases, and have already yielded impressive results in finding genetic factors involved in conditions ranging from age-related blindness to obesity.**

The utility of the reference genome

SECTIONS HOME SEARCH

The New York Times

RESEARCH

A Decade Later, Genetic Map Yields Few New Cures

By NICHOLAS WADE JUNE 12, 2010

See how this article appeared when it was originally published on NYTimes.com

Ten years after President Bill Clinton announced that the first draft of the human genome was complete, medicine has yet to see any large part of the promised benefits.

Subscribe

SCIENTIFIC AMERICAN

Cart 0 Sign In | Stay Informed

THE SCIENCES MIND HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS BLOGS PUBLICATIONS

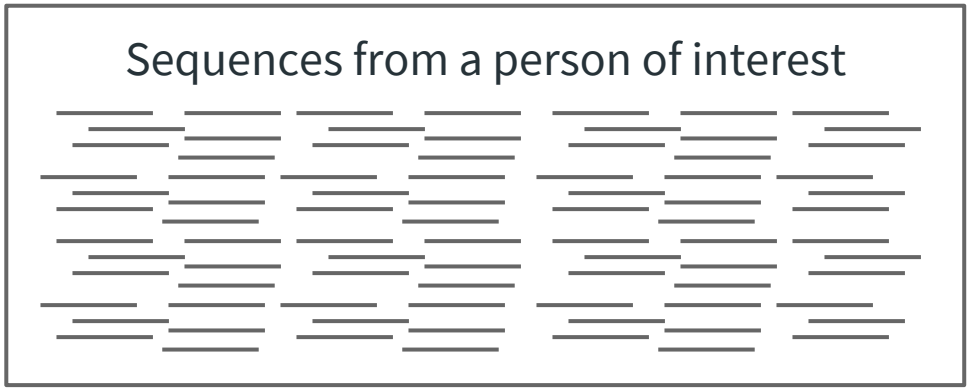
SCIENTIFIC AMERICAN OCTOBER 2010

Revolution Postponed: Why the Human Genome Project Has Been Disappointing

The Human Genome Project has failed so far to produce the medical miracles that scientists promised. Biologists are now divided over what, if anything, went wrong—and what needs to happen next

By Stephen S. Hall

Thought experiment: extra chromosome



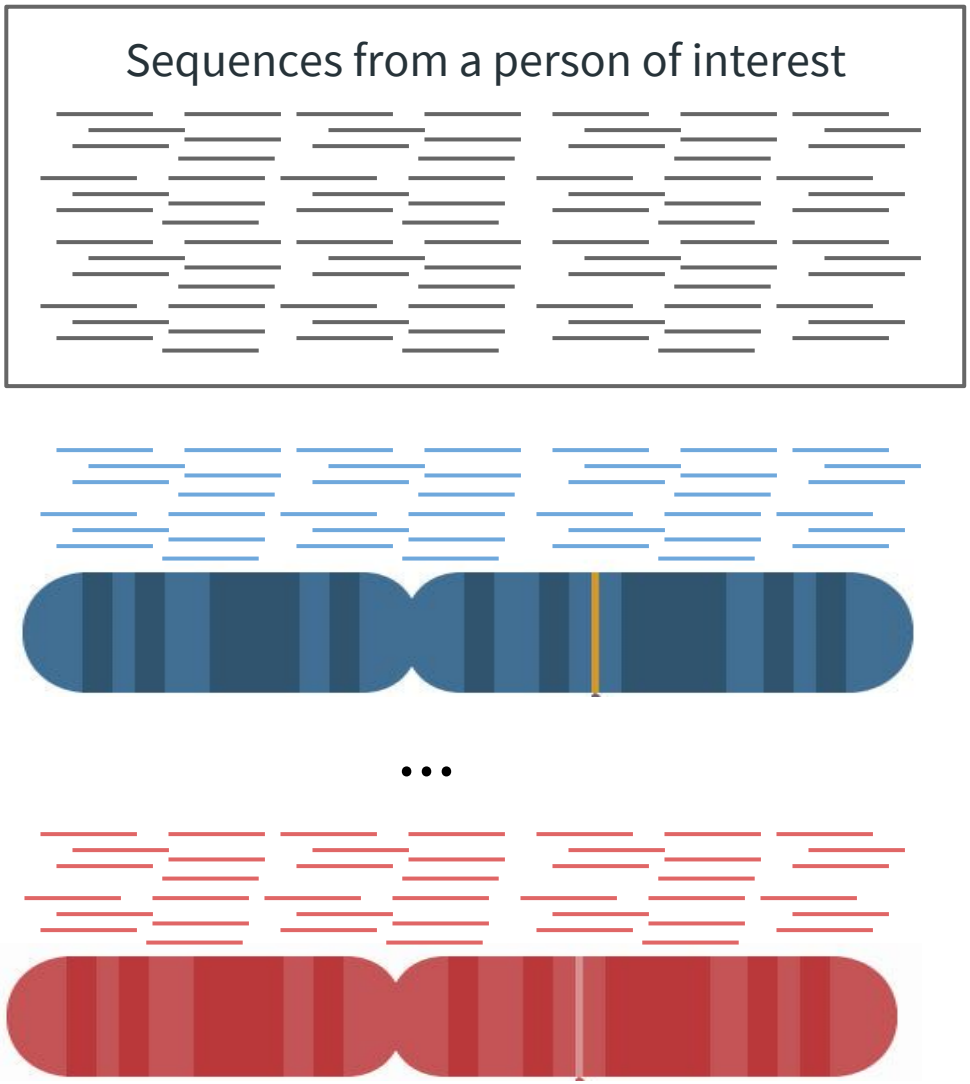
What if the individual these sequences came from has an extra, completely new chromosome?

Reference chromosome

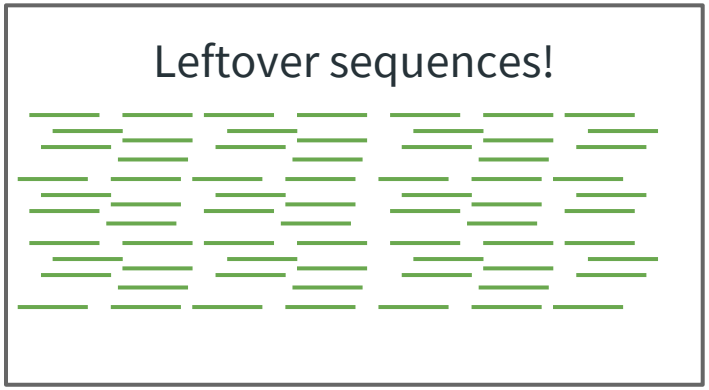
...

Reference chromosome

Thought experiment: extra chromosome



What if the individual these sequences came from has an extra, completely new chromosome?



African-ancestry populations

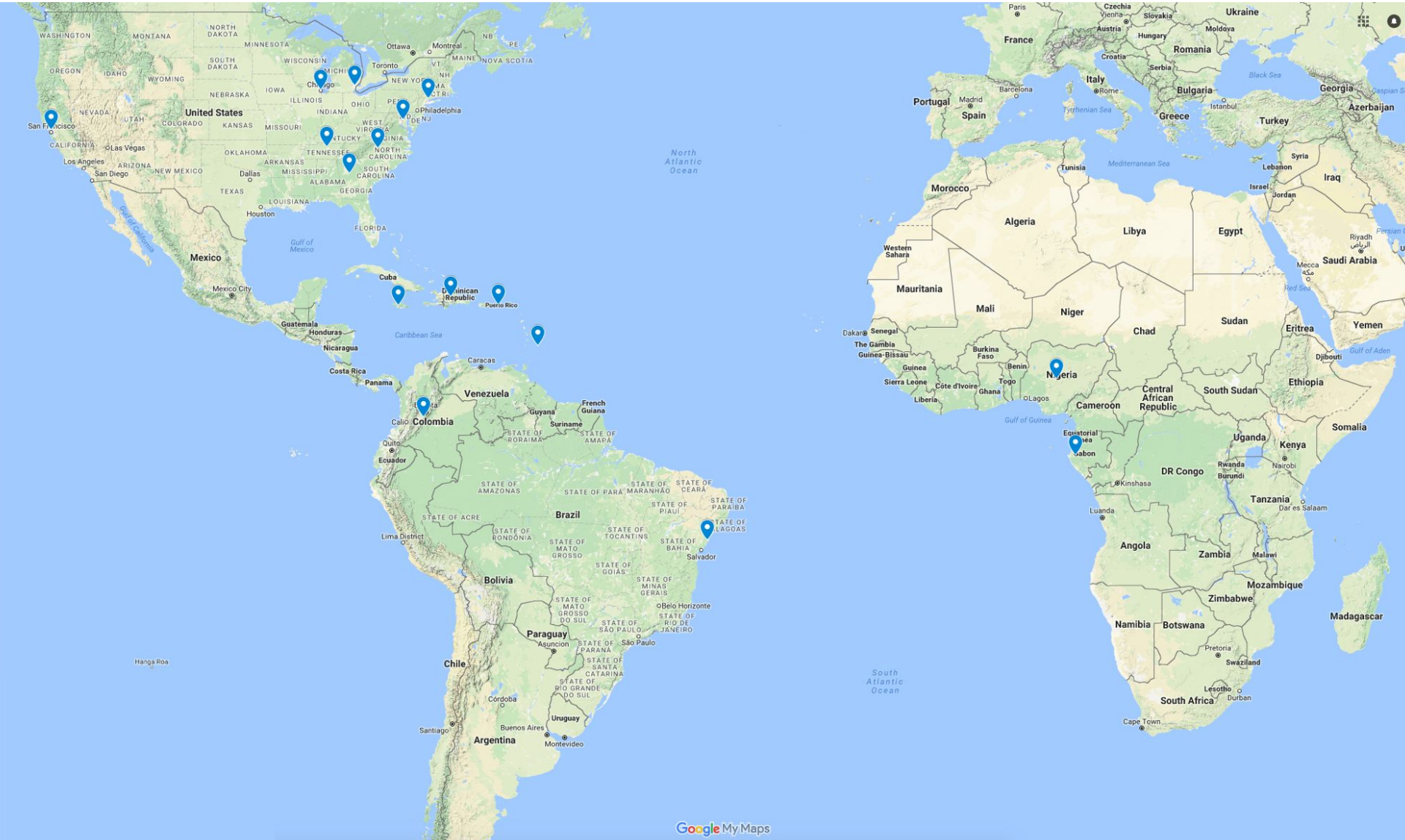
Known to be more genetically diverse than other populations

Higher prevalence of asthma than other populations (controlling for environment)

Understudied relative to European populations

Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) collected sequencing data from ~900 individuals, half with asthma, half without

CAAPA data overview



CAAPA data overview

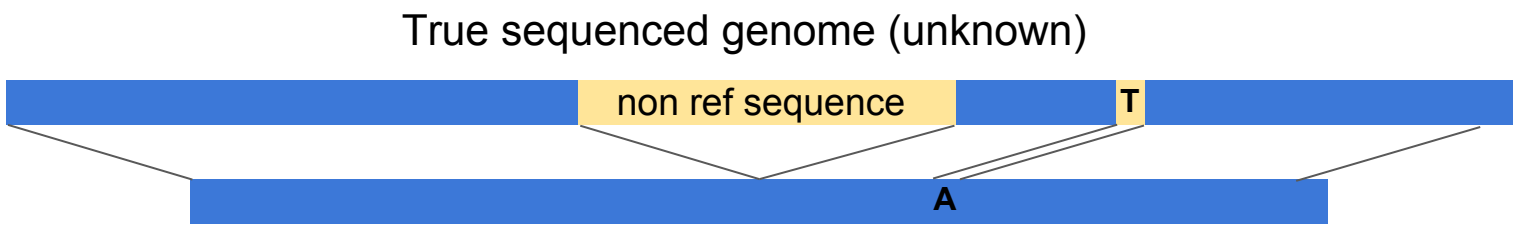
Supplementary Table 6 | Cohorts of CAAPA samples.

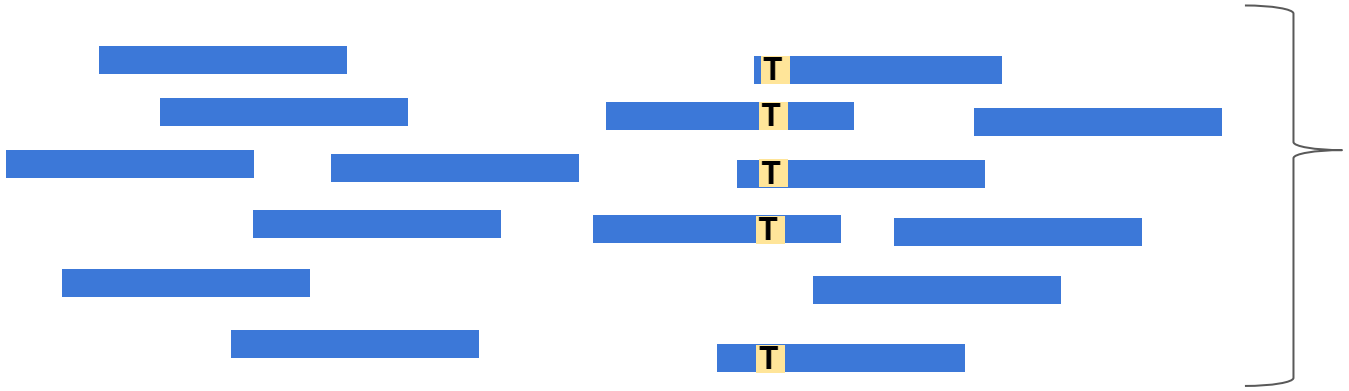
Cohort	Number of Samples
African American (Atlanta)	50
African American (Baltimore-DC)	50
African American (Chicago)	50
African American (Detroit)	50
African American (Jackson, MS)	50
African American (Nashville)	48
African American (NYC)	48
African American (San Francisco)	50
African American (Winston-Salem)	50
Barbados	49
Brazil	47
Colombia	50
Dominican Republic	47
Gabon	34
Honduras	50
Jamaica	50
Palenque	34
Nigeria	50
Puerto Rico	53

Data was collected from 19 distinct cohorts across the Americas, the Caribbean, and Africa resulting in 910 analyzed samples.

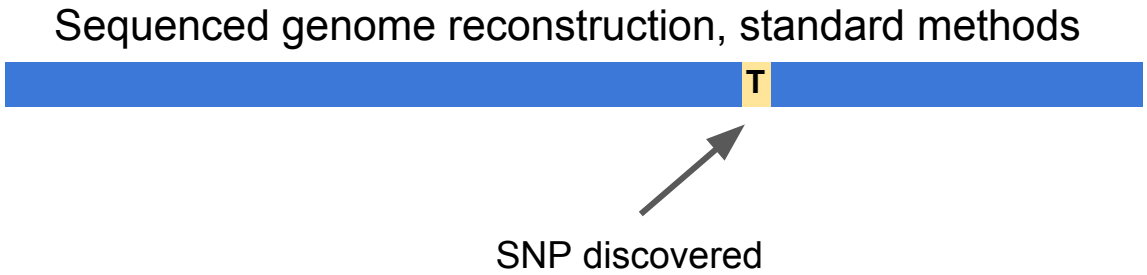
Sherman *et al* 2019

Sequences missed by alignment



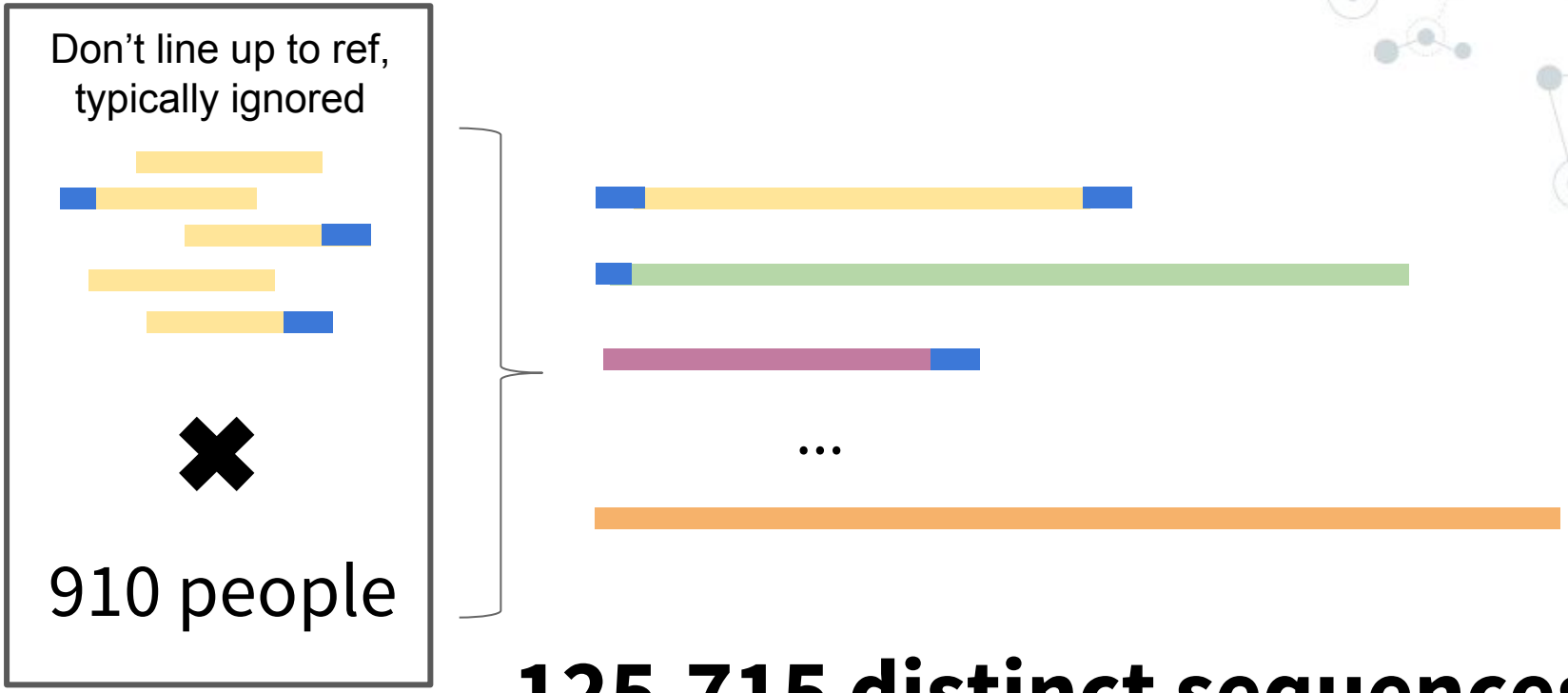


Sequencing reads get aligned to reference genome



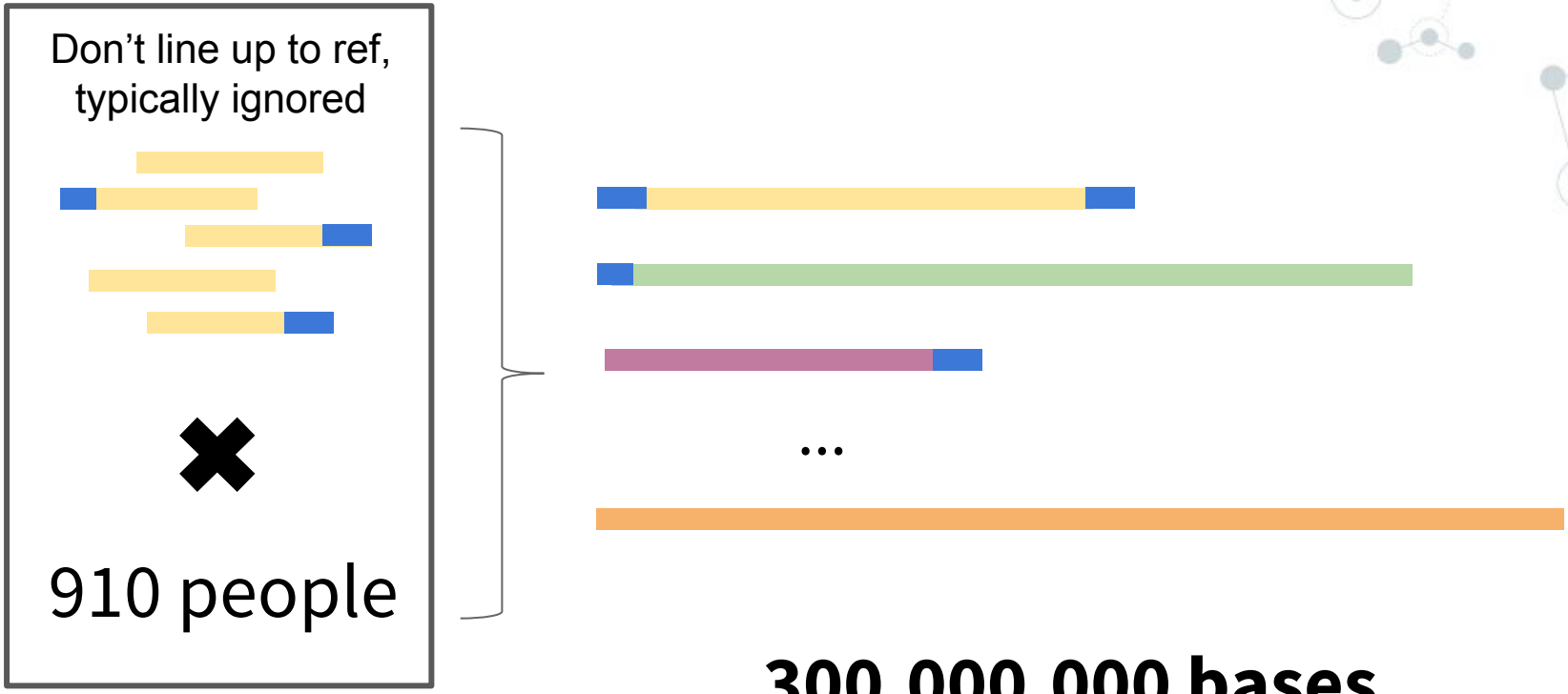
Don't line up to ref, typically ignored

Discovering insertions



125,715 distinct sequences
totaling nearly
300,000,000 bases

Discovering insertions

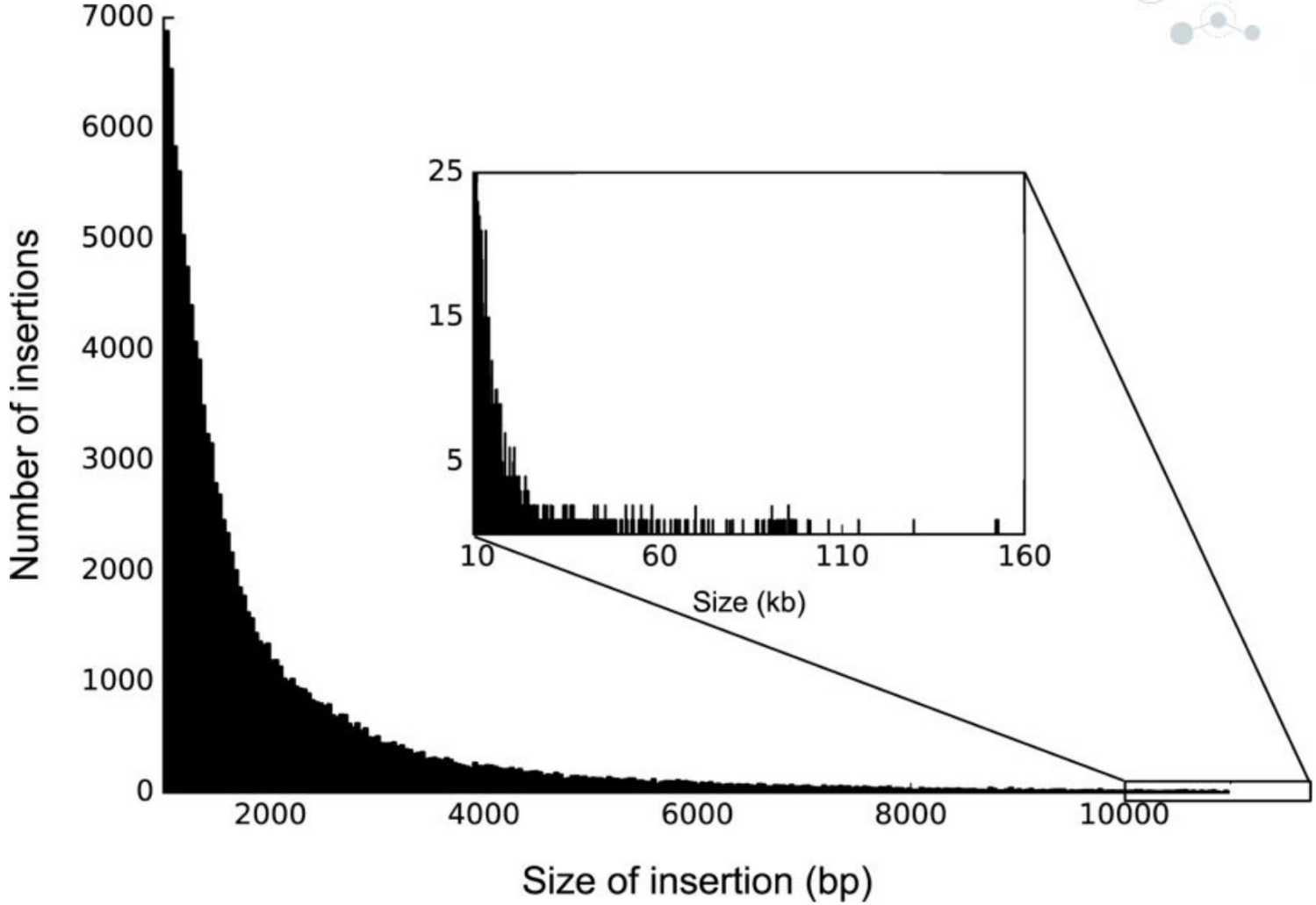


300,000,000 bases

=

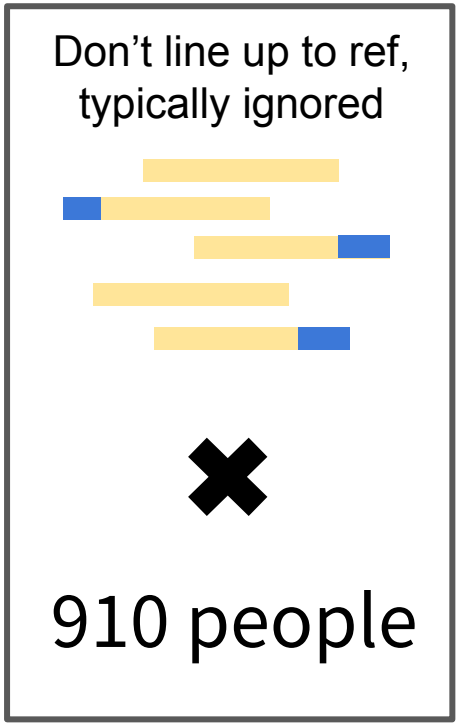
10% of the size of the human genome!

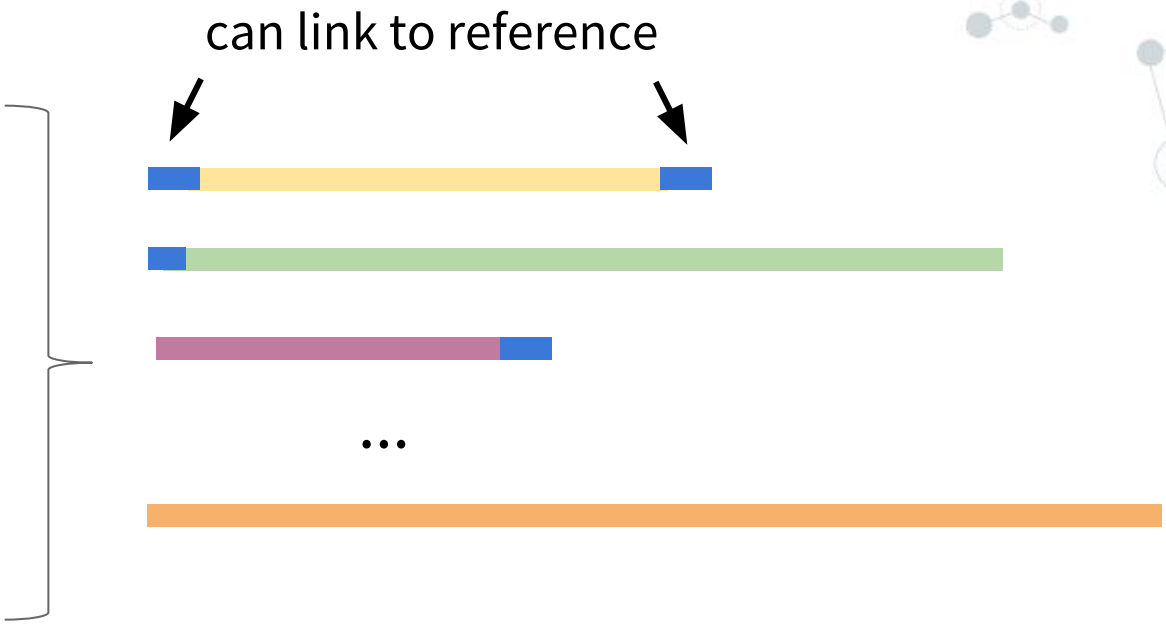
Insertion size distribution



Sherman *et al* 2019

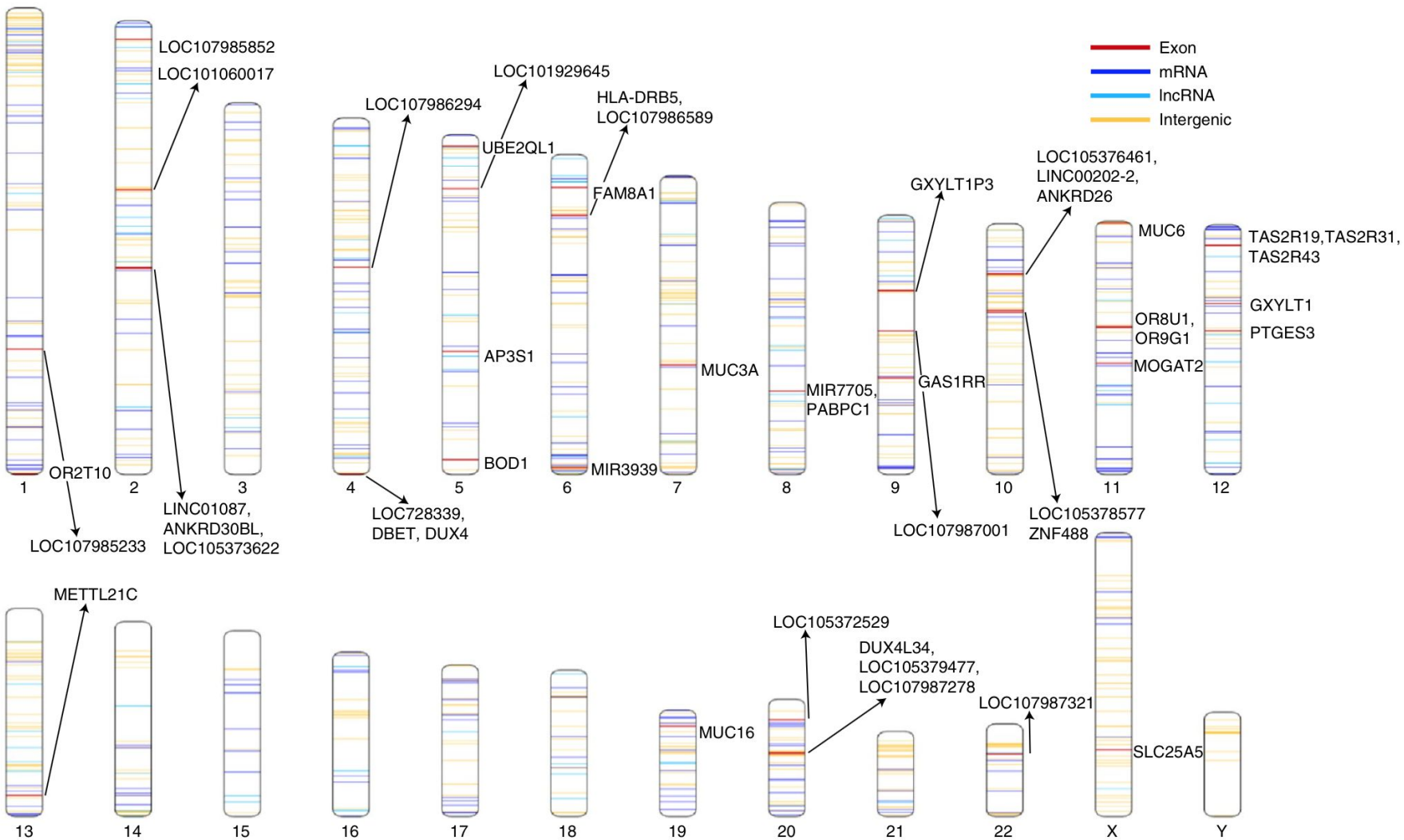
Discovering insertions





125,715 distinct sequences
totaling nearly
300,000,000 bases

Known insertion locations in reference



Sherman *et al* 2019

Prevalence in the 910 individuals

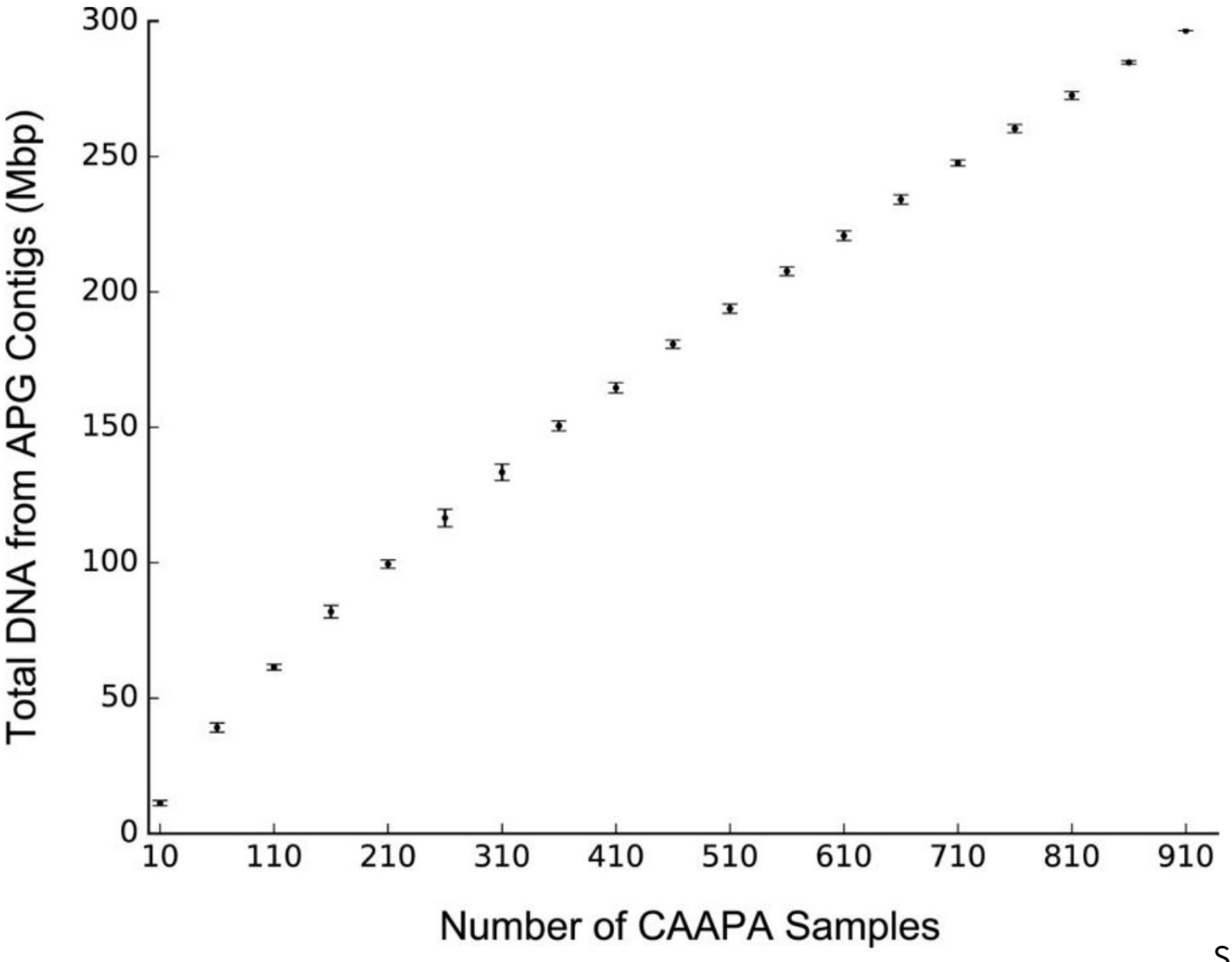
- 26.7% (33,599/125,715) found in multiple people
  - In total, these make up 80,098,092 bases
- On average, each person had **859 of these insertions**
- 16,068,045 bases match well to a Chinese genome or a Korean genome assembly, with another 105,098,989 matching somewhat well

Are we really all 99.9% identical?

It depends how we measure.

But probably not.

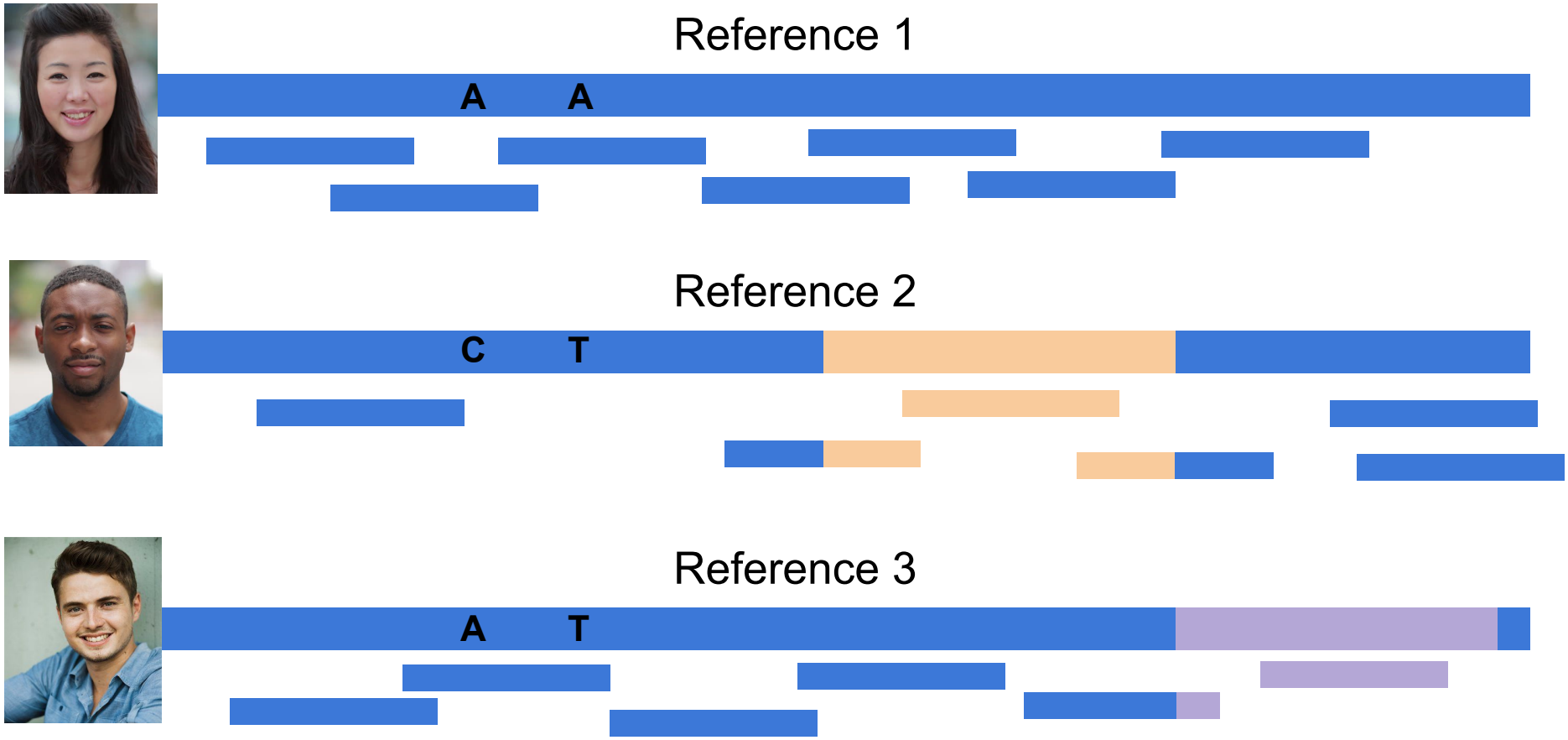
There's still more to be found



Sherman *et al* 2019

What next?

We're diverse! We need more references to represent us all.



Acknowledgments

Steven Salzberg

Daniela Puiu

Geo Pertea

Juliet Forman

Kathleen Barnes and the rest of the CAAPA team

...and many others who gave advice along the way

JOHNS HOPKINS UNIVERSITY

Consortium on Asthma among African-ancestry Populations in the Americas

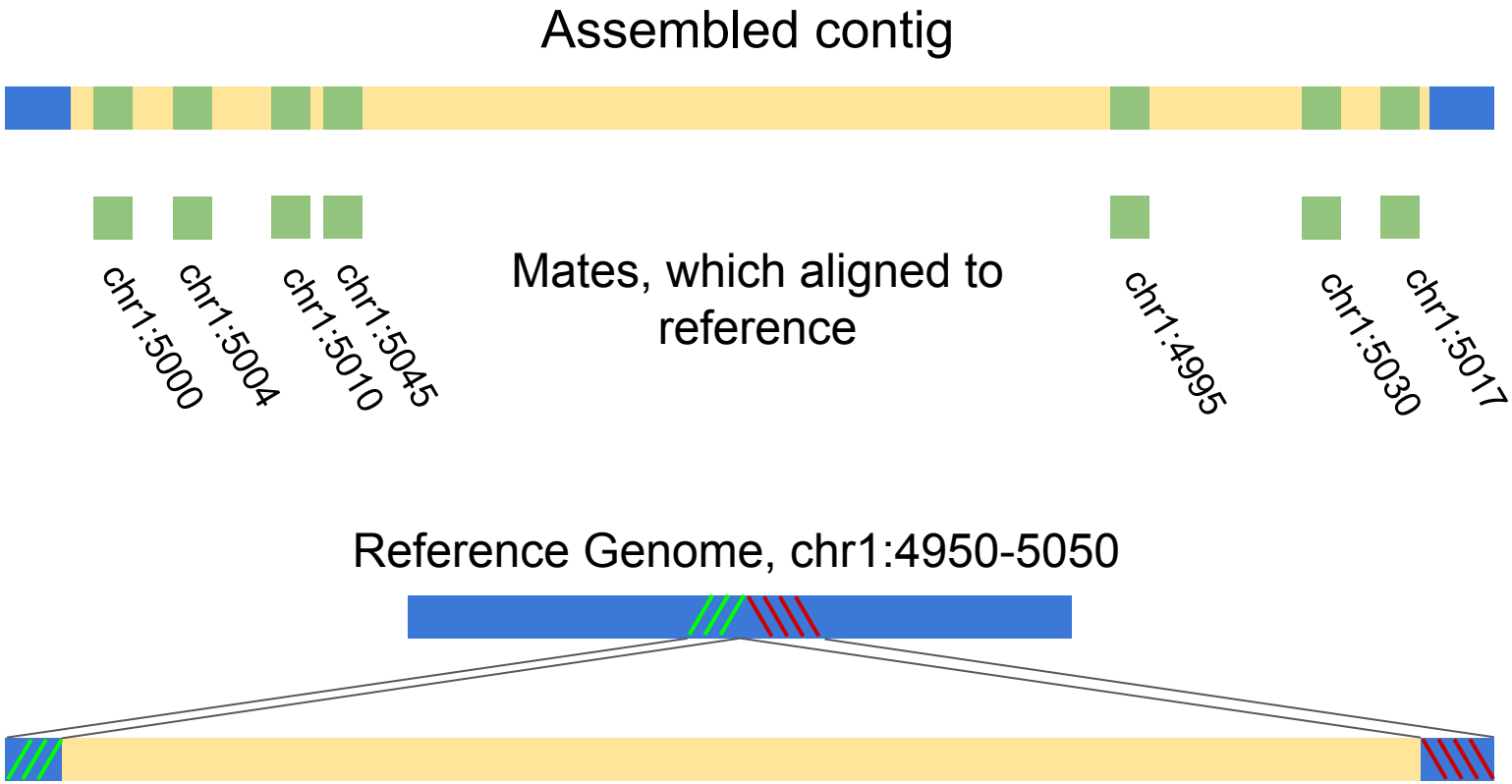
CAAPA

NIH National Heart, Lung, and Blood Institute

Questions?

Additional Slides

Placing assembled contigs



Pan-genome contigs in SGDP individuals

Supplementary Table 5 | APG contig presence in Simons Genome Diversity Project individuals

Sample ID	Population	Country	Sex	Number of APG Contigs Present
LP6005442-DNA_E10	English	England	M	796
LP6005442-DNA_F10	English	England	F	680
LP6005441-DNA_A05	French	France	M	963
LP6005441-DNA_B05	French	France	F	810
LP6005441-DNA_C11	Sardinian	Italy	M	943
LP6005441-DNA_D11	Sardinian	Italy	F	905
LP6005442-DNA_A11	Spanish	Spain	M	817
LP6005442-DNA_B11	Spanish	Spain	F	1011
LP6005442-DNA_C10	Finnish	Finland	M	893
LP6005442-DNA_D10	Finnish	Finland	F	892
LP6005442-DNA_A08	Hungarian	Hungary	M	1041
LP6005442-DNA_B08	Hungarian	Hungary	F	1007
LP6005441-DNA_G08	Mozabite	Algeria	M	1034
LP6005441-DNA_H08	Mozabite	Algeria	F	980
LP6005443-DNA_A01	Bantu	Kenya	M	791
LP6005441-DNA_B02	Bantu	Kenya	F	991
LP6005442-DNA_G10	Gambian	Gambia	M	710
LP6005442-DNA_H10	Gambian	Gambia	F	690
LP6005442-DNA_G11	Mende	Sierra Leone	M	720
LP6005442-DNA_H11	Mende	Sierra Leone	F	711
LP6005592-DNA_C03	Mbuti	Congo	M	690
LP6005441-DNA_B08	Mbuti	Congo	F	914
LP6005442-DNA_A02	Yoruba	Nigeria	M	925
LP6005442-DNA_B02	Yoruba	Nigeria	F	980

Twenty-four individuals from the Simons Genome Diversity Project from 12 populations, 6 African and 6 European, were examined to determine presence/absence of the APG contigs. Each individual's assembled contigs were aligned to the APG contigs to determine the number of APG contigs present in the individual.

296.5 Mb, 16.6 Mb in Korean and Chinese assemblies

	Number of sequence contigs	Total length (bp)	Bases with no alignment to GRCh38 (<80% identity)	Longest contig (bp)
Two ends placed	302	667,668	431,656	20,732
One end placed	1,246	3,687,028	1,866,699	79,938
Unplaced	124,167	292,130,588	202,629,979	152,806
Total	125,715	296,485,284	204,928,334	152,806
Non-private only	33,599	80,098,092	50,044,650	152,806

	Best GRCh38 alignment is 80–90% identical with 50–80% coverage		Best GRCh38 alignment is <80% identical or <50% coverage		Total	
	Contigs	Length (bp)	Contigs	Length (bp)	Contigs	Length (bp)
Matches Chinese only	1,625	2,898,106	7,607	25,475,277	9,232	28,373,383
Matches Korean only	2,242	3,989,277	15,635	48,642,664	17,877	52,631,941
Matches both	5,385	9,720,662	9,713	29,981,048	15,098	39,701,710
Total	9,252	16,608,045	32,955	104,098,989	42,207	120,707,034

Sherman *et al* (2018). Nature Genetics.

Pan-genome contig presence/absence

	Number of contigs	Mean number of insertions per individual	Mean number individuals per insertion
Two ends placed	302	120 (39.7%)	363 (of 910)
One end placed	1,246	212 (17.0%)	155 (of 910)
Unplaced	124,167	527 (0.4%)	4 (of 910)
Total	125,715	859 (0.7%)	6 (of 910)
Non-private only	33,599	758 (2.2%)	21 (of 910)