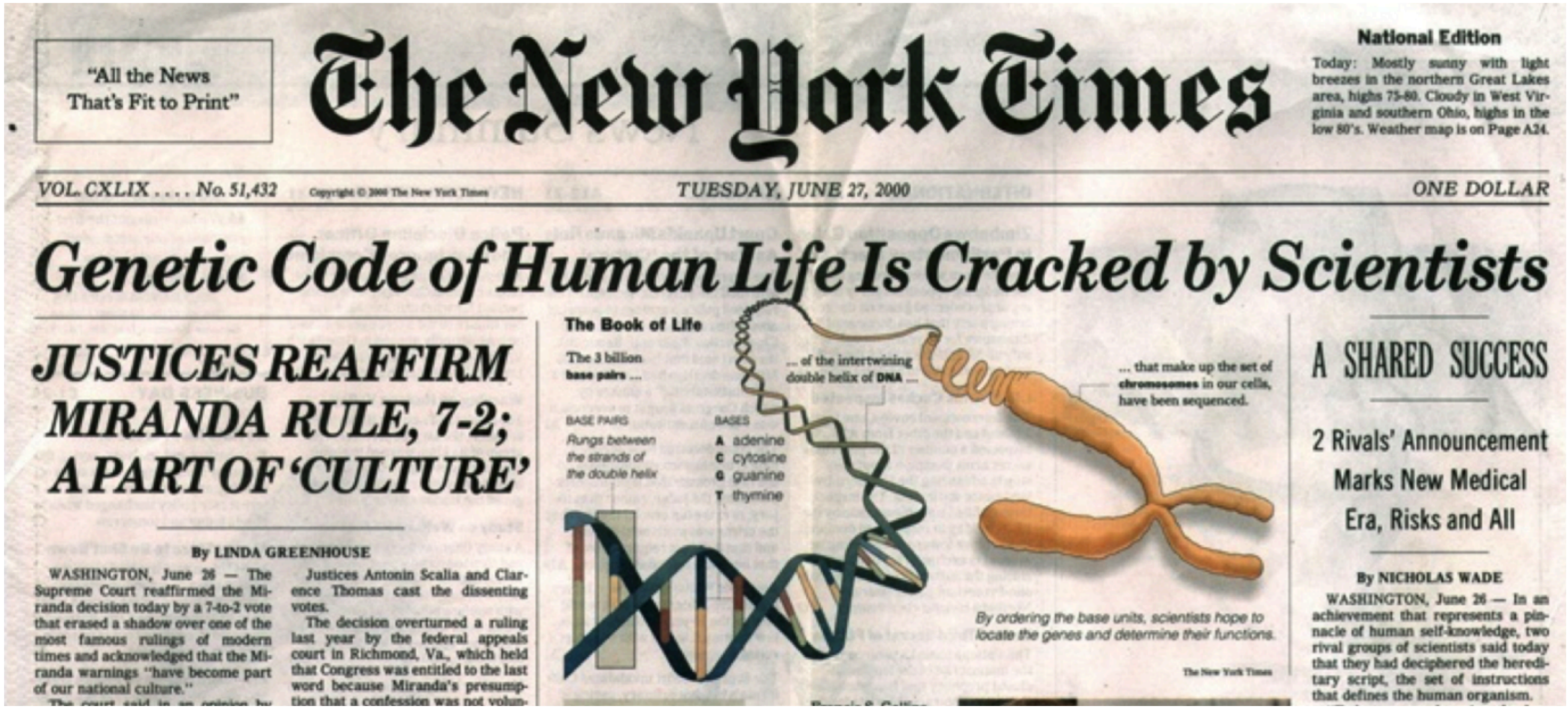
Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History

Human genome project&reference genome

Human Genome Project



"All the News That's Fit to Print"

The New York Times

National Edition

Today: Mostly sunny with light breezes in the northern Great Lakes area, highs 75-80. Cloudy in West Virginia and southern Ohio, highs in the low 80's. Weather map is on Page A24.

VOL. CXLIX . . . . No. 51,432 Copyright © 2000 The New York Times TUESDAY, JUNE 27, 2000 ONE DOLLAR

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part of our national culture."

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption that a confession was not volun-

The Book of Life

The 3 billion **base pairs** ...

... of the intertwining double helix of **DNA** ...

... that make up the set of **chromosomes** in our cells, have been sequenced.

BASE PAIRS

Rungs between the strands of the double helix

BASES

A adenine
C cytosine
G guanine
T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

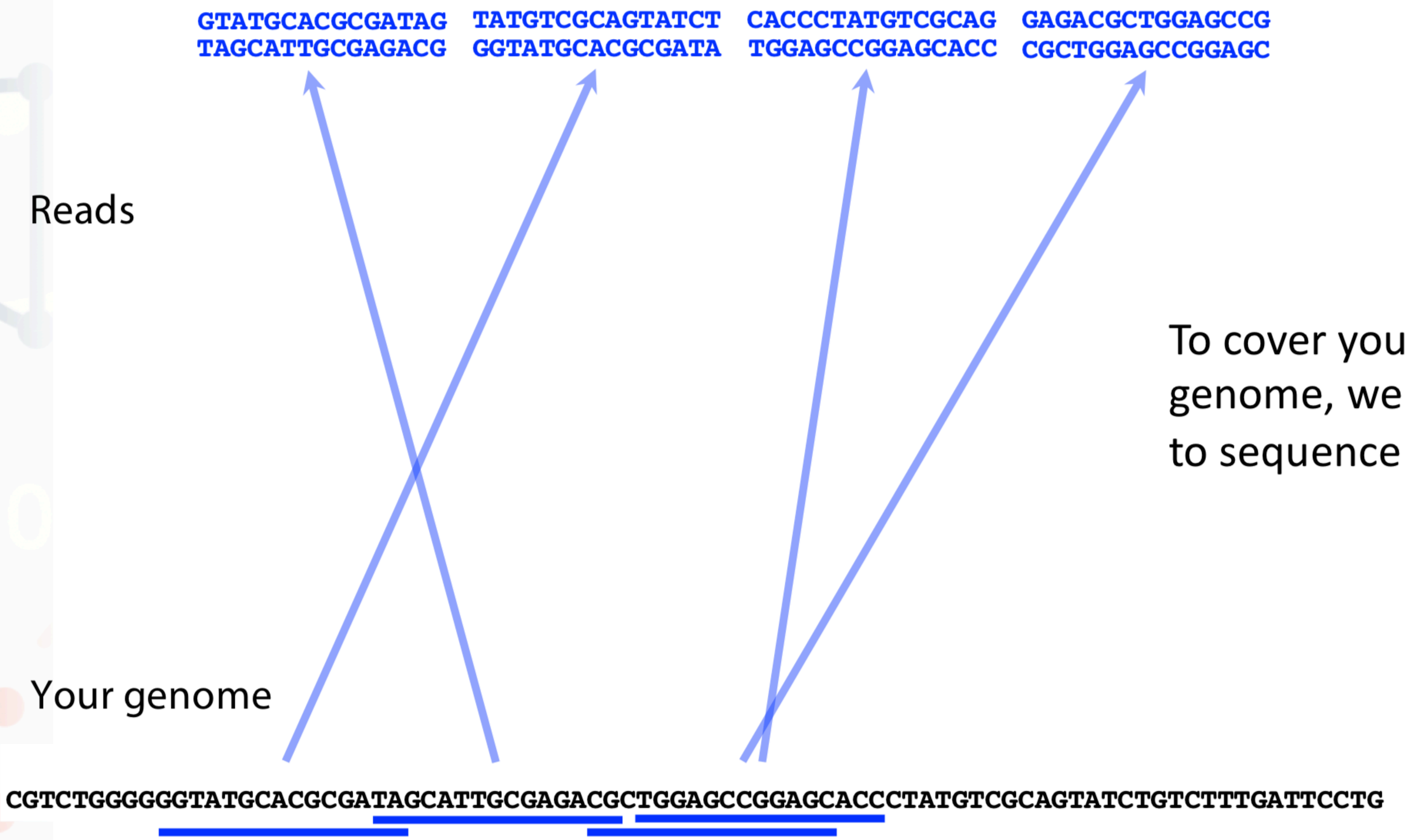
WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

- 70% come from a male donor. The remainders are from one American male donor and two American female donor.

- 3 billion US dollars & 15 years.

- Chinese scientists sequenced 1% of the reference genome.

How to get our genome data?



GTATGCACGCGATAG
TAGCATTGCGAGACG

TATGTCGCAGTATCT
GGTATGCACGCGATA

CACCCTATGTCGCAG
TGGAGCCGGAGCACC

GAGACGCTGGAGCCG
CGCTGGAGCCGGAGC

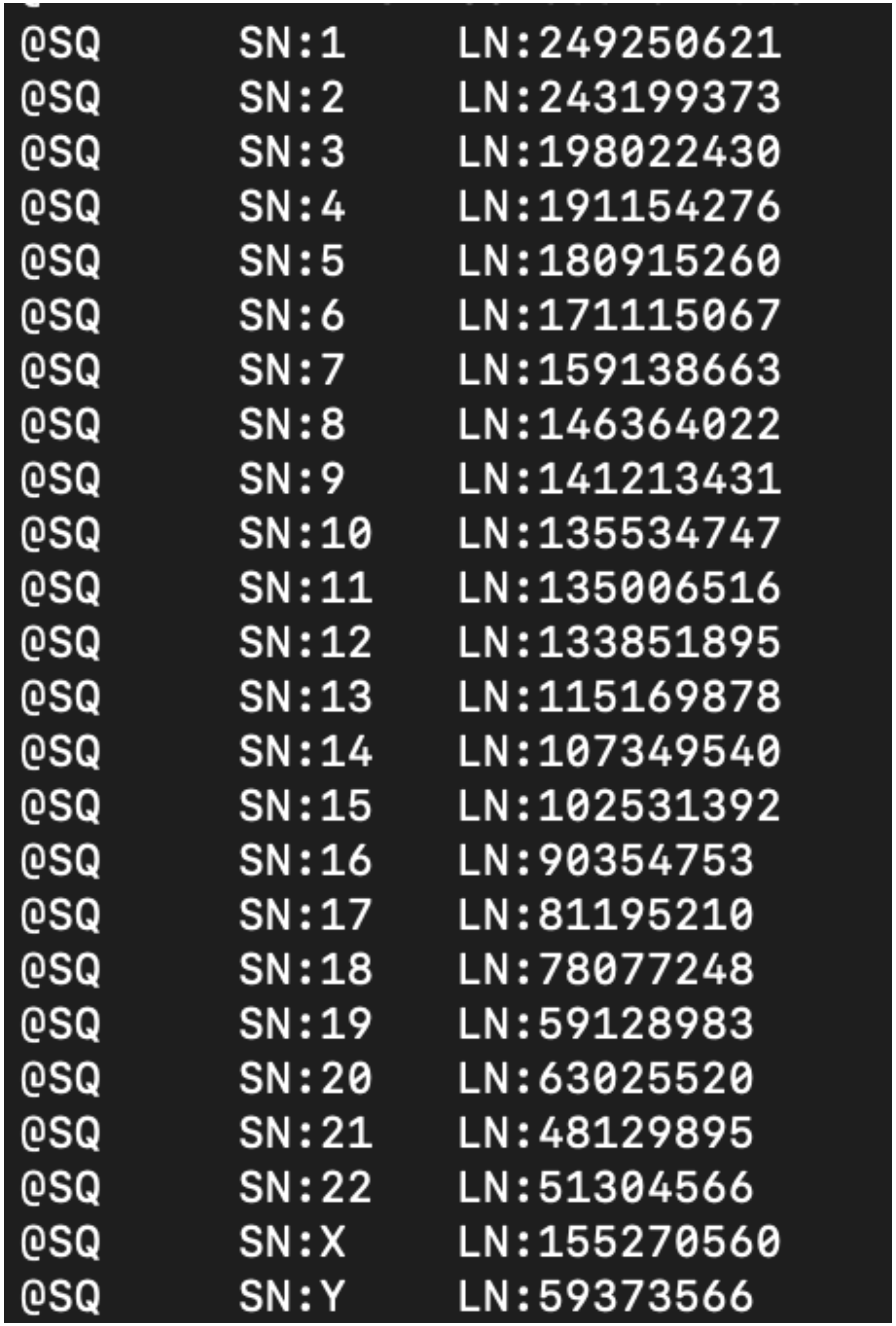
Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

To cover your genome, we need to sequence more

“Genomic number”



```
@SQ     SN:1    LN:249250621
@SQ     SN:2    LN:243199373
@SQ     SN:3    LN:198022430
@SQ     SN:4    LN:191154276
@SQ     SN:5    LN:180915260
@SQ     SN:6    LN:171115067
@SQ     SN:7    LN:159138663
@SQ     SN:8    LN:146364022
@SQ     SN:9    LN:141213431
@SQ     SN:10   LN:135534747
@SQ     SN:11   LN:135006516
@SQ     SN:12   LN:133851895
@SQ     SN:13   LN:115169878
@SQ     SN:14   LN:107349540
@SQ     SN:15   LN:102531392
@SQ     SN:16   LN:90354753
@SQ     SN:17   LN:81195210
@SQ     SN:18   LN:78077248
@SQ     SN:19   LN:59128983
@SQ     SN:20   LN:63025520
@SQ     SN:21   LN:48129895
@SQ     SN:22   LN:51304566
@SQ     SN:X    LN:155270560
@SQ     SN:Y    LN:59373566
```

Allele frequency calculation for a population

Allele Frequency Calculation

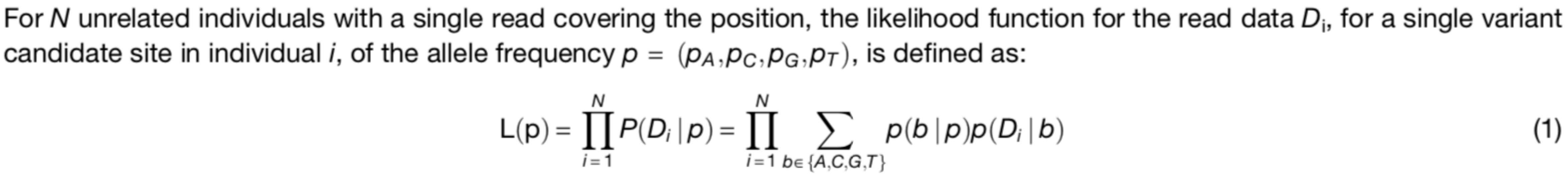
- Why not simply count the number of “A,T,C,G”?

Because of sequencing errors, systemic bias, etc, the probability should be incorporated.

Allele frequency calculation

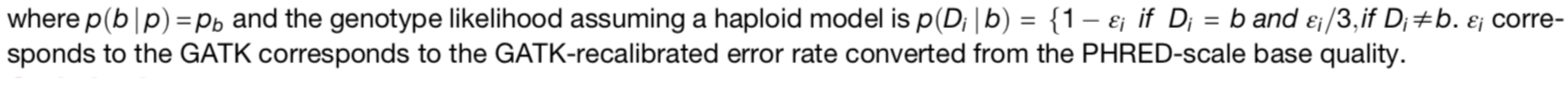
- **Maximum Likelihood estimation**

idea: given observations (sequenced reads from different individuals), MLE attempts to find the parameters of probabilistic model to maximize the likelihood L(p)



For N unrelated individuals with a single read covering the position, the likelihood function for the read data D_i, for a single variant candidate site in individual i, of the allele frequency $p = (p_A, p_C, p_G, p_T)$, is defined as:

$$\mathrm{L(p)} = \prod_{i=1}^{N} P(D_i \mid p) = \prod_{i=1}^{N} \sum_{b \in \{A,C,G,T\}} p(b \mid p) p(D_i \mid b) \tag{1}$$



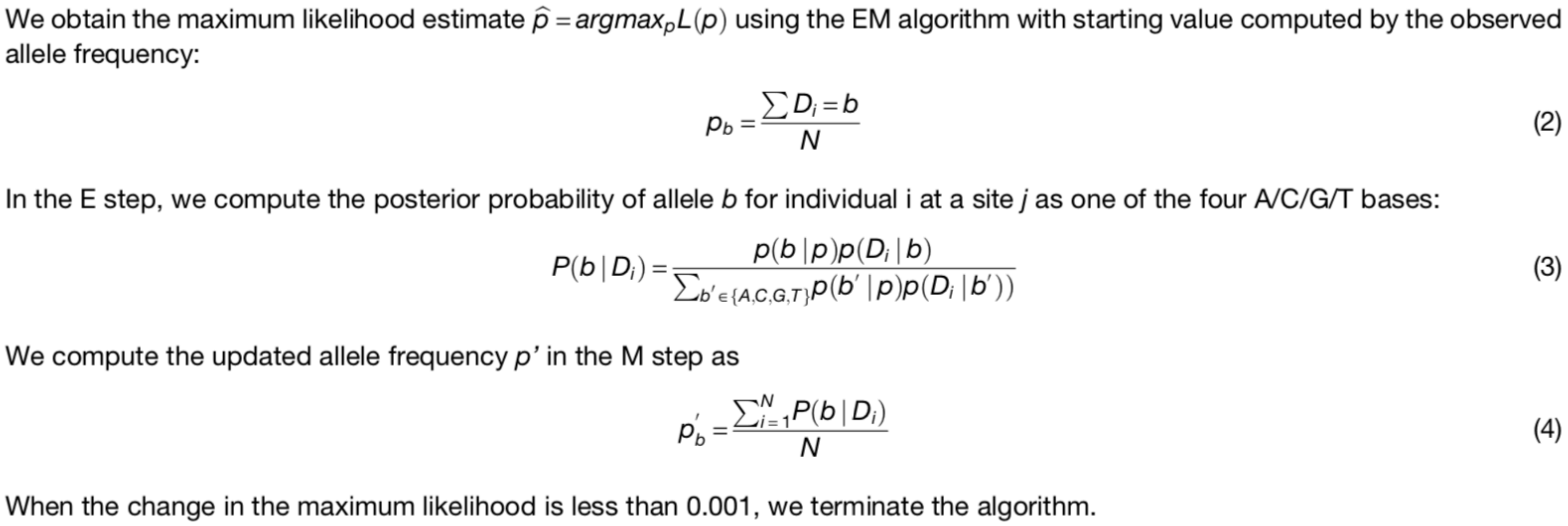
where $p(b \mid p) = p_b$ and the genotype likelihood assuming a haploid model is $p(D_i \mid b) = \{1 - \varepsilon_i$ *if* $D_i = b$ *and* $\varepsilon_i/3$,*if* $D_i \neq b$. ε_i corresponds to the GATK corresponds to the GATK-recalibrated error rate converted from the PHRED-scale base quality.

Ref: Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History

Allele frequency calculation

- **EM algorithm**

EM is an iterative method to find maximum likelihood estimates of parameters in statistical model, where the model depends on unobserved latent variables.



We obtain the maximum likelihood estimate $\widehat{p} = argmax_p L(p)$ using the EM algorithm with starting value computed by the observed allele frequency:

$$p_b = \frac{\sum D_i = b}{N} \qquad (2)$$

In the E step, we compute the posterior probability of allele *b* for individual i at a site *j* as one of the four A/C/G/T bases:

$$P(b \mid D_i) = \frac{p(b \mid p) p(D_i \mid b)}{\sum_{b' \in \{A,C,G,T\}} p(b' \mid p) p(D_i \mid b'))} \qquad (3)$$

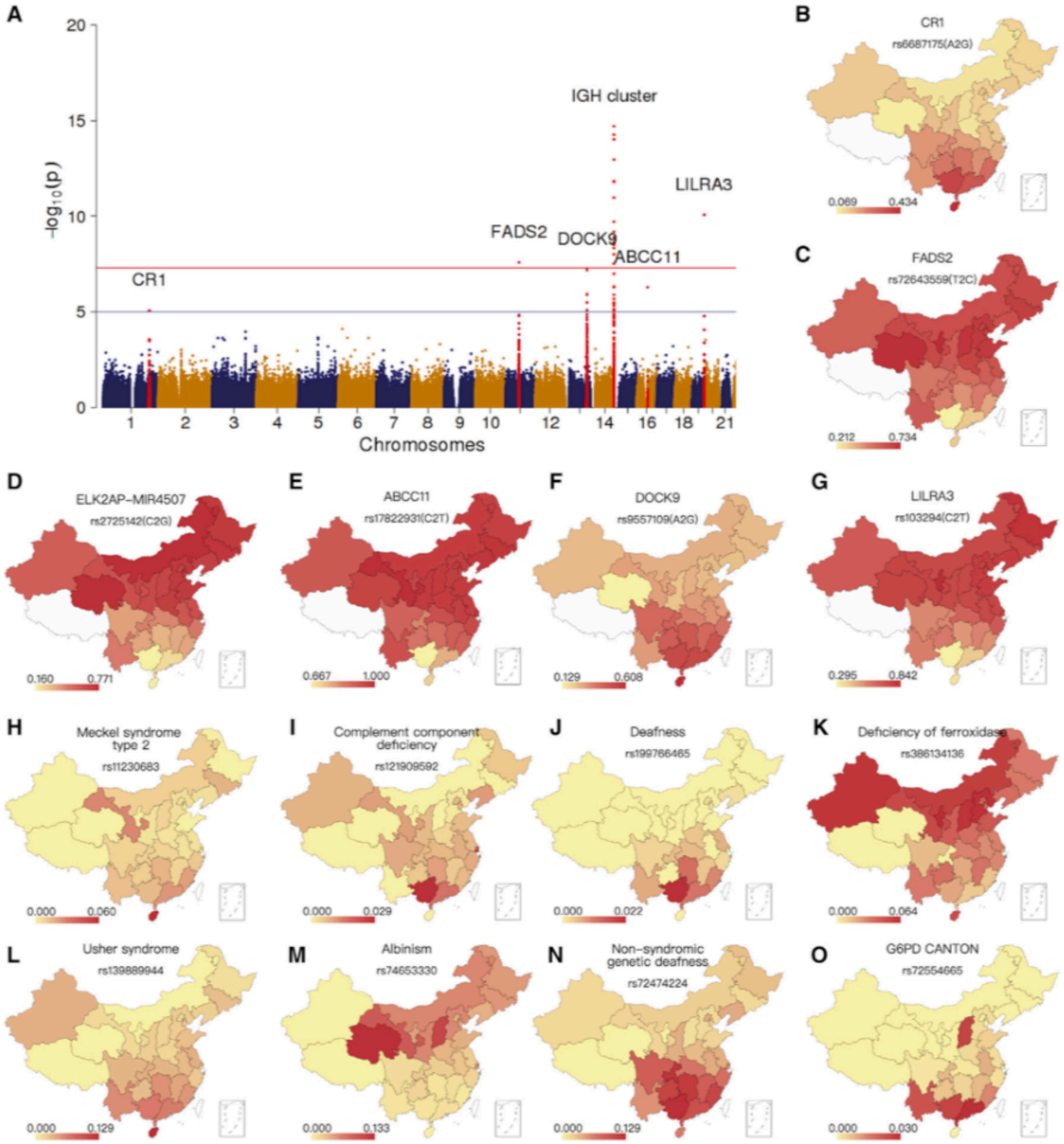
We compute the updated allele frequency *p'* in the M step as

$$p_b^{'} = \frac{\sum_{i=1}^{N} P(b \mid D_i)}{N} \qquad (4)$$

When the change in the maximum likelihood is less than 0.001, we terminate the algorithm.

Ref: Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History

Allele frequency calculation



Fitness test

Decision of allelic type

- Method: log-likelihood ratio test

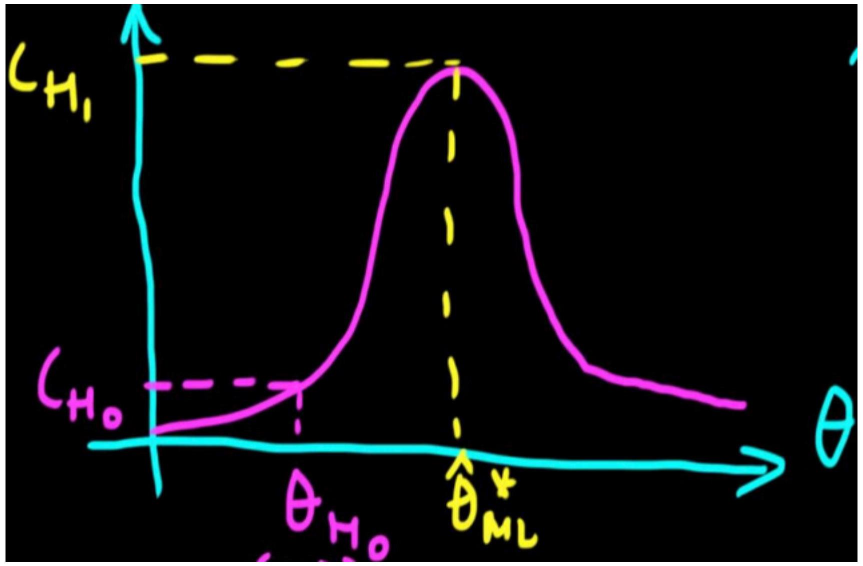
Fitness test

$$\mathrm{H}_0 : \theta = \theta_0$$

$$\mathrm{H}_1 : \theta = \theta_{ML}$$

$$LR = 2(\log L(O \mid \theta_{ML}) - \log L(O \mid \theta))$$

$$LR \sim \chi^2(1)$$



Ref: https://www.youtube.com/watch?v=Tn5y2i_MqQ8

Decision of allelic type

1. iteratively set the allele frequency of one of the four nucleotides to zero to obtain models of tri-allelic loci .

$$LRT_{4vs3} = -2log\left(\frac{\widehat{f_3}(p_x=0)}{\widehat{f_4}}\right)$$

where x is one of the 4 bases, f is likelihood functior $L(p) = \prod_{i=1}^{N} P(D_i \mid p) = \prod_{i=1}^{N} \sum_{b \in \{A,C,G,T\}} p(b \mid p) p(D_i \mid b)$

2. If the p values of LRT_{4vs3} test are significant, the variant will be classified as a tetra-allelic loci (H_0 is rejected). If not, we further to testify:

1.

$$LRT_{3vs2} = -2log\left(\frac{\widehat{f_2}(p_x=0, p_y=0)}{\widehat{f_3}(p_x=0)}\right)$$

where x is the base which makes the p value of LRT_{4vs3} maximum, y is one of the rest 3 bases.

Ref: Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History

Decision of allelic type

$$LRT_{3vs2} = -2log\left(\frac{\widehat{f_2}\left(p_x = 0, p_y = 0\right)}{\widehat{f_3}\left(p_x = 0\right)}\right)$$

3. Similarly, if p value of LRT_{3vs2} is significant, this loci is classified as tri-allelic loci. Otherwise, we choose the base y which makes p value of LRT_{3vs2} maximum, to continue test the bi-allelic versus mono-allelic assumption:

$$LRT_{2vs1} = -2log\left(\frac{\widehat{f_1}\left(p_x = 0, p_y = 0, p_z = 0\right)}{\widehat{f_2}\left(p_x = 0, p_y = 0\right)}\right)$$

Ref: Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History