Opinionated Lessons in Statistics

by Bill Press

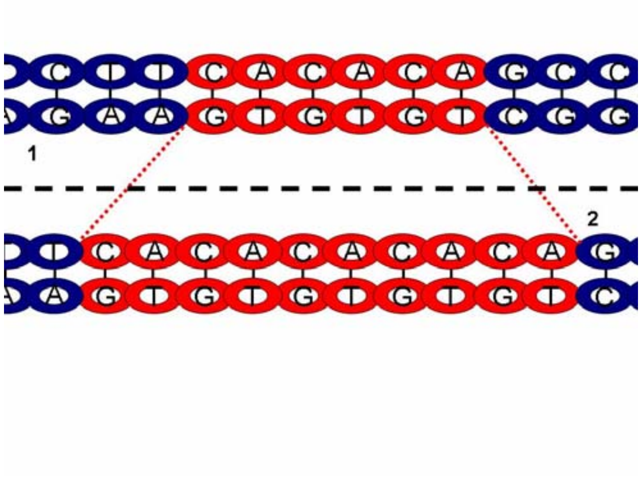
#6 The Towne Family Tree

Neat example (with some biology):

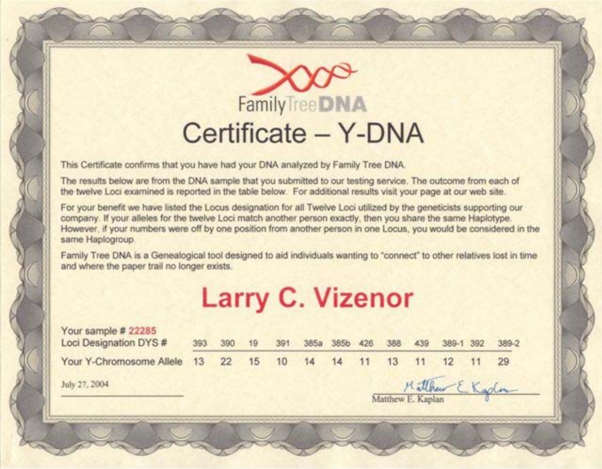
Individual identity, or ancestry, can be determined by “variable length short tandem repeats” (STRs) in the genome.

~0.5% mutation prob per STR per generation (though highly variable)

if use Y chromosome only, get paternal ancestry



There are companies that sell “certificates” with your genotype. A bit opportunistic, since in a few years your whole genome will be sequenced by your health plan.



FamilyTreeDNA

Certificate – Y-DNA

This Certificate confirms that you have had your DNA analyzed by Family Tree DNA.

The results below are from the DNA sample that you submitted to our testing service. The outcome from each of the twelve Loci examined is reported in the table below. For additional results visit your page at our web site.

For your benefit we have listed the Locus designation for all Twelve Loci utilized by the geneticists supporting our company. If your alleles for the twelve Loci match another person exactly, then you share the same Haplotype. However, if your numbers were off by one position from another person in one Locus, you would be considered in the same Haplogroup.

Family Tree DNA is a Genealogical tool designed to aid individuals wanting to “connect” to other relatives lost in time and where the paper trail no longer exists.

Larry C. Vizenor

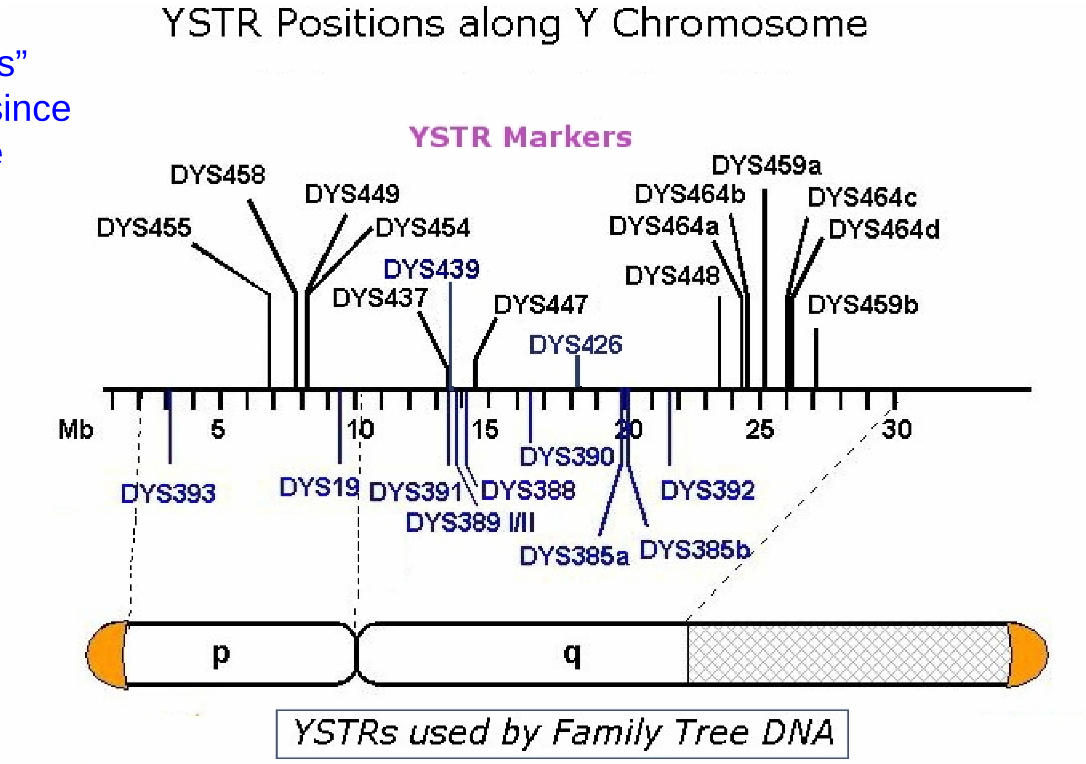
Your sample # 22285

Loci Designation DYS #	393	390	19	391	385a	385b	426	388	439	389-1	392	389-2
Your Y-Chromosome Allele	13	22	15	10	14	14	11	13	11	12	11	29

July 27, 2004

Matthew E. Kaplan

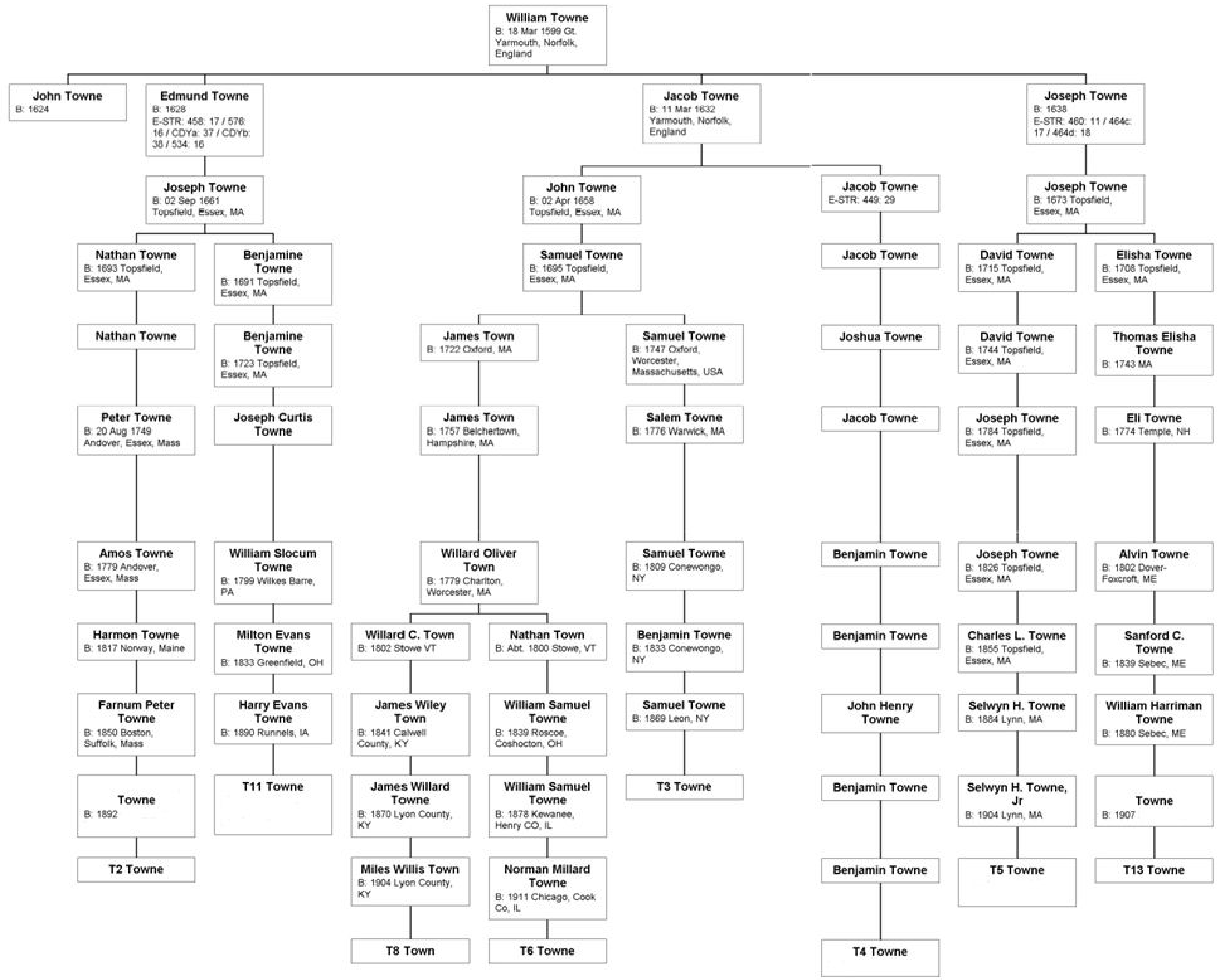
Matthew E. Kaplan



YSTRs used by Family Tree DNA

Margaret, my ex-wife, is really into the Towne family. (And, she’s neither a biologist nor a Towne.)

Descendant Chart for William Towne



William Towne B: 18 Mar 1599 Gt. Yarmouth, Norfolk, England

John Towne B: 1624

Edmund Towne B: 1628 E-STR: 458: 17 / 576: 16 / CDYa: 37 / CDYb: 38 / 534: 16

Joseph Towne B: 02 Sep 1661 Topsfield, Essex, MA

Nathan Towne B: 1693 Topsfield, Essex, MA

Nathan Towne

Peter Towne B: 20 Aug 1749 Andover, Essex, Mass

Amos Towne B: 1779 Andover, Essex, Mass

Harmon Towne B: 1817 Norway, Maine

Farnum Peter Towne B: 1850 Boston, Suffolk, Mass

Towne B: 1892

T2 Towne

Benjamine Towne B: 1691 Topsfield, Essex, MA

Benjamine Towne B: 1723 Topsfield, Essex, MA

Joseph Curtis Towne

William Slocum Towne B: 1799 Wilkes Barre, PA

Milton Evans Towne B: 1833 Greenfield, OH

Harry Evans Towne B: 1890 Runnels, IA

T11 Towne

Jacob Towne B: 11 Mar 1632 Yarmouth, Norfolk, England

John Towne B: 02 Apr 1658 Topsfield, Essex, MA

Samuel Towne B: 1695 Topsfield, Essex, MA

James Town B: 1722 Oxford, MA

James Town B: 1757 Belchertown, Hampshire, MA

Willard Oliver Town B: 1779 Charlton, Worcester, MA

Willard C. Town B: 1802 Stowe VT

James Wiley Town B: 1841 Calwell County, KY

James Willard Towne B: 1870 Lyon County, KY

Miles Willis Town B: 1904 Lyon County, KY

T8 Town

Nathan Town B: Abt. 1800 Stowe, VT

William Samuel Towne B: 1839 Roscoe, Coshocton, OH

William Samuel Towne B: 1878 Kewanee, Henry CO, IL

Norman Millard Towne B: 1911 Chicago, Cook Co, IL

T6 Towne

Samuel Towne B: 1747 Oxford, Worcester, Massachusetts, USA

Salem Towne B: 1776 Warwick, MA

Samuel Towne B: 1809 Conewongo, NY

Benjamin Towne B: 1833 Conewongo, NY

Samuel Towne B: 1869 Leon, NY

T3 Towne

Jacob Towne E-STR: 449: 29

Jacob Towne

Joshua Towne

Jacob Towne

Benjamin Towne

Benjamin Towne

John Henry Towne

Benjamin Towne

Benjamin Towne

T4 Towne

Joseph Towne B: 1638 E-STR: 460: 11 / 464c: 17 / 464d: 18

Joseph Towne B: 1673 Topsfield, Essex, MA

David Towne B: 1715 Topsfield, Essex, MA

David Towne B: 1744 Topsfield, Essex, MA

Joseph Towne B: 1784 Topsfield, Essex, MA

Joseph Towne B: 1826 Topsfield, Essex, MA

Charles L. Towne B: 1856 Topsfield, Essex, MA

Selwyn H. Towne B: 1884 Lynn, MA

Selwyn H. Towne, Jr B: 1904 Lynn, MA

T5 Towne

Elisha Towne B: 1708 Topsfield, Essex, MA

Thomas Elisha Towne B: 1743 MA

Eli Towne B: 1774 Temple, NH

Alvin Towne B: 1802 Dover-Foxcroft, ME

Sanford C. Towne B: 1839 Sebec, ME

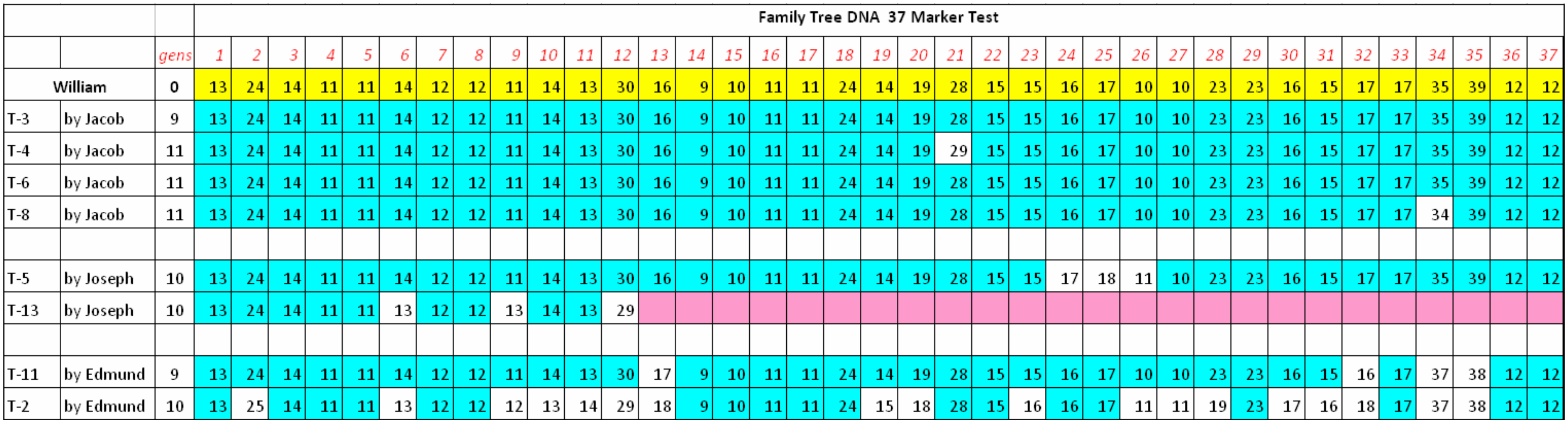
William Harriman Towne B: 1880 Sebec, ME

Towne B: 1907

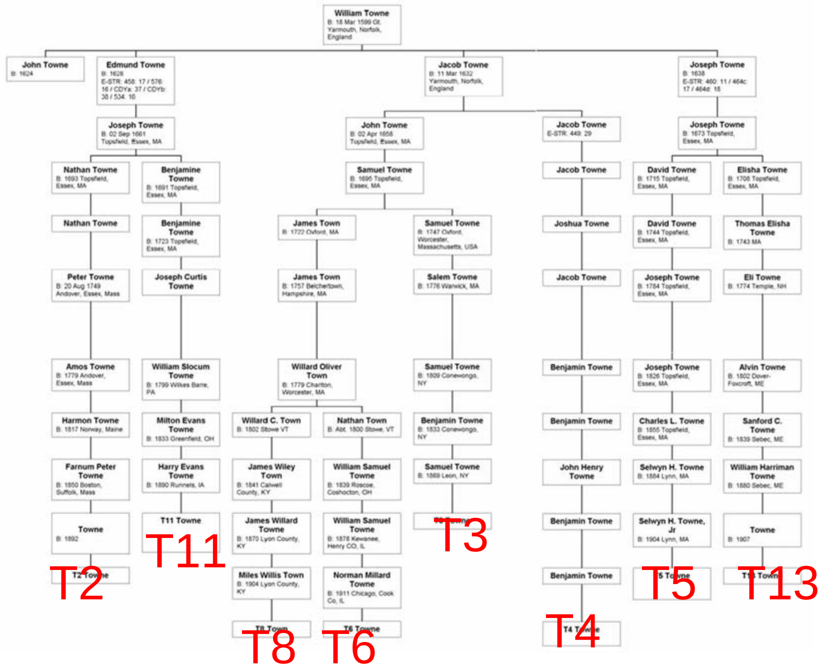
T13 Towne

as of March 6, 2008

Here’s data from Margaret on 8 recent Townes (identified only by T code). (We’ll use this data several times in the next few of lectures.)

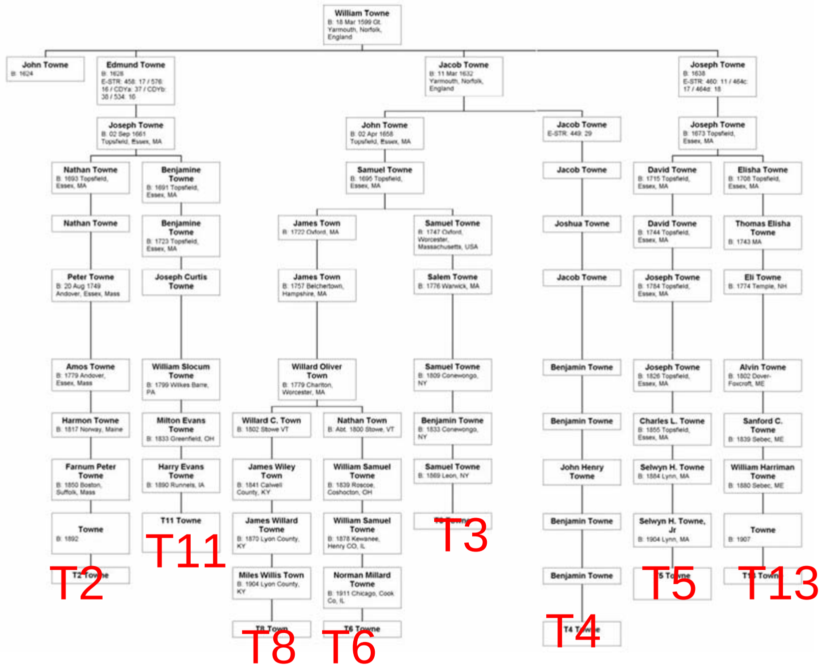


			Family Tree DNA 37 Marker Test																																		
		gens	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
William		0	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-3	by Jacob	9	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-4	by Jacob	11	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	29	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-6	by Jacob	11	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-8	by Jacob	11	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	34	39	12	12
T-5	by Joseph	10	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	17	18	11	10	23	23	16	15	17	17	35	39	12	12
T-13	by Joseph	10	13	24	14	11	11	13	12	12	13	14	13	29																									
T-11	by Edmund	9	13	24	14	11	11	14	12	12	11	14	13	30	17	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	16	17	37	38	12	12
T-2	by Edmund	10	13	25	14	11	11	13	12	12	12	13	14	29	18	9	10	11	11	24	15	18	28	15	16	16	17	11	11	19	23	17	16	18	17	37	38	12	12

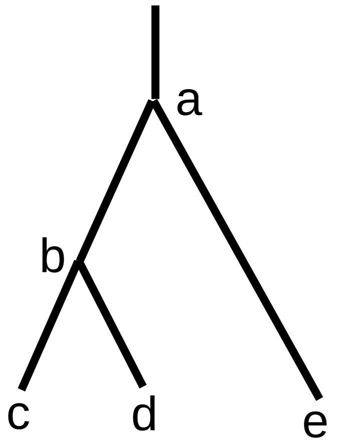


Or, just showing the changes from what we impute for William:

			Family Tree DNA 37 Marker Test																																			
		gens	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
William		0	13	24	14	11	11	14	12	12	11	14	13	30	16	9	10	11	11	24	14	19	28	15	15	16	17	10	10	23	23	16	15	17	17	35	39	12	12
T-3	by Jacob	9																																					
T-4	by Jacob	11																					1																
T-6	by Jacob	11																																					
T-8	by Jacob	11																																		-1			
T-5	by Joseph	10																								1	1	1											
T-13	by Joseph	10						-1			2			-1																									
T-11	by Edmund	9													1																			-1		2	-1		
T-2	by Edmund	10		1				-1			1	-1	1	-1	2						1	-1			1			1	1	-4		1	1	1		2	-1		



Unraveling dependencies



$$
\begin{aligned}
P(abcde) &= P(e|a\cancel{bcd})P(abcd) \\
&= P(e|a)P(c|\cancel{a}b\cancel{d})P(abd) \\
&= P(e|a)P(c|b)P(d|\cancel{a}b)P(ab) \\
&= P(e|a)P(c|b)P(d|b)P(b|a)P(a)
\end{aligned}
$$

Another important idea is “conditional independence”

Example: b and e are “conditionally independent given a”

$$
\begin{aligned}
P(be|a) &= P(b|\cancel{e}a)P(e|a) \\
&= P(b|a)P(e|a)
\end{aligned}
$$

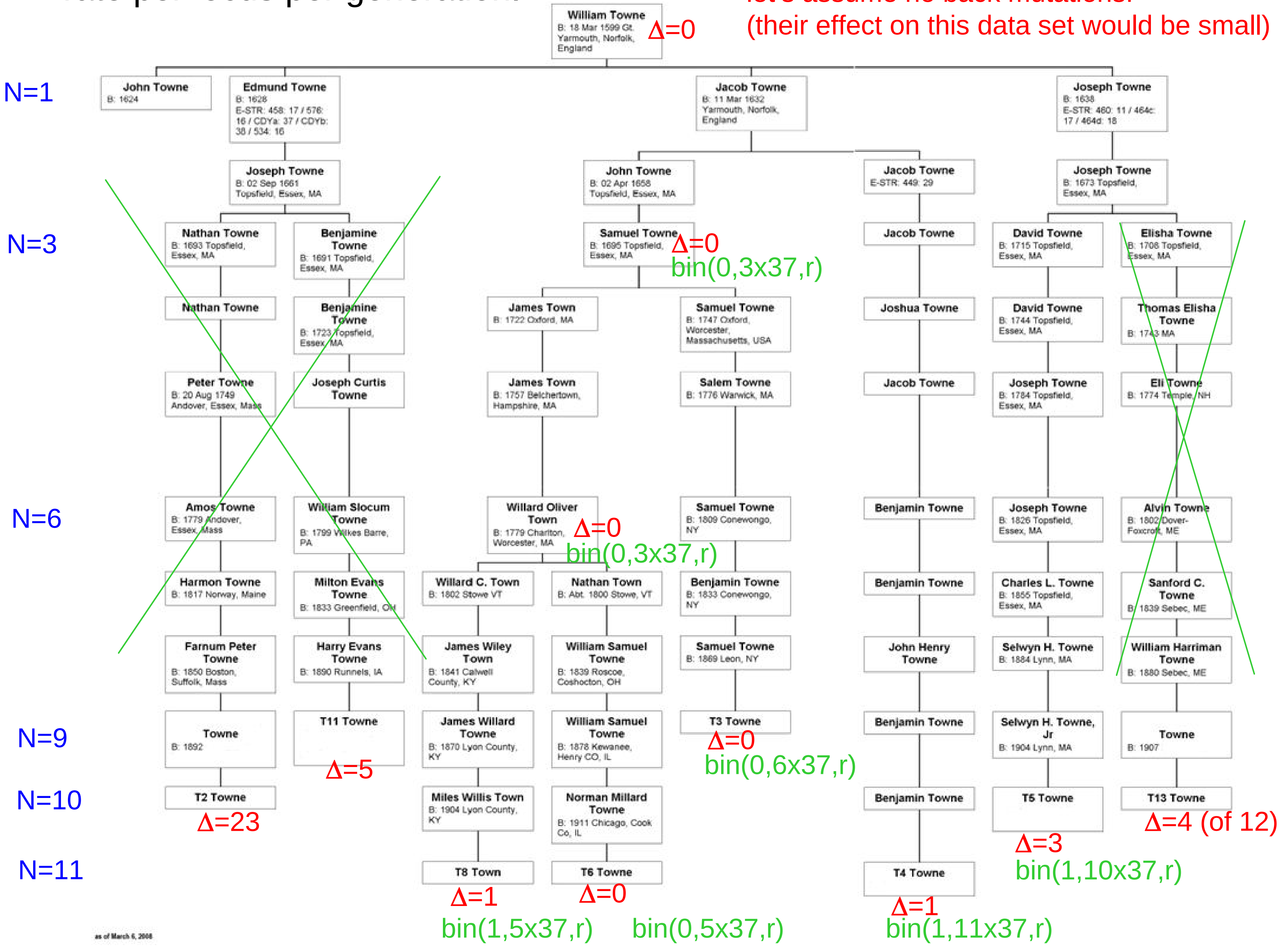
while b and d are <u>not</u> conditionally independent given a:

$$P(bd|a) = P(b|da)P(d|a)$$

?

Let’s do a Bayesian estimation of the parameter r, the mutation rate per locus per generation.

let’s assume no back mutations!
(their effect on this data set would be small)



William Towne
B: 18 Mar 1599 Gt. Yarmouth, Norfolk, England
Δ=0

N=1
John Towne
B: 1624
Edmund Towne
B: 1628
E-STR: 458: 17 / 576: 16 / CDYa: 37 / CDYb: 38 / 534: 16
Jacob Towne
B: 11 Mar 1632 Yarmouth, Norfolk, England
Joseph Towne
B: 1638
E-STR: 460: 11 / 464c: 17 / 464d: 18

Joseph Towne
B: 02 Sep 1661 Topsfield, Essex, MA
John Towne
B: 02 Apr 1658 Topsfield, Essex, MA
Jacob Towne
E-STR: 449: 29
Joseph Towne
B: 1673 Topsfield, Essex, MA

N=3
Nathan Towne
B: 1693 Topsfield, Essex, MA
Benjamine Towne
B: 1691 Topsfield, Essex, MA
Samuel Towne
B: 1695 Topsfield, Essex, MA
Δ=0
bin(0,3x37,r)
Jacob Towne
David Towne
B: 1715 Topsfield, Essex, MA
Elisha Towne
B: 1708 Topsfield, Essex, MA

Nathan Towne
Benjamine Towne
B: 1723 Topsfield, Essex, MA
James Town
B: 1722 Oxford, MA
Samuel Towne
B: 1747 Oxford, Worcester, Massachusetts, USA
Joshua Towne
David Towne
B: 1744 Topsfield, Essex, MA
Thomas Elisha Towne
B: 1743 MA

Peter Towne
B: 20 Aug 1749 Andover, Essex, Mass
Joseph Curtis Towne
James Town
B: 1757 Belchertown, Hampshire, MA
Salem Towne
B: 1776 Warwick, MA
Jacob Towne
Joseph Towne
B: 1784 Topsfield, Essex, MA
Eli Towne
B: 1774 Temple, NH

N=6
Amos Towne
B: 1779 Andover, Essex, Mass
William Slocum Towne
B: 1799 Wilkes Barre, PA
Willard Oliver Town
B: 1779 Charlton, Worcester, MA
Δ=0
bin(0,3x37,r)
Samuel Towne
B: 1809 Conewongo, NY
Benjamin Towne
Joseph Towne
B: 1826 Topsfield, Essex, MA
Alvin Towne
B: 1802 Dover-Foxcroft, ME

Harmon Towne
B: 1817 Norway, Maine
Milton Evans Towne
B: 1833 Greenfield, OH
Willard C. Town
B: 1802 Stowe VT
Nathan Town
B: Abt. 1800 Stowe, VT
Benjamin Towne
B: 1833 Conewongo, NY
Benjamin Towne
Charles L. Towne
B: 1855 Topsfield, Essex, MA
Sanford C. Towne
B: 1839 Sebec, ME

Farnum Peter Towne
B: 1850 Boston, Suffolk, Mass
Harry Evans Towne
B: 1890 Runnels, IA
James Wiley Town
B: 1841 Calwell County, KY
William Samuel Towne
B: 1839 Roscoe, Coshocton, OH
Samuel Towne
B: 1869 Leon, NY
John Henry Towne
Selwyn H. Towne
B: 1884 Lynn, MA
William Harriman Towne
B: 1880 Sebec, ME

N=9
Towne
B: 1892
T11 Towne
Δ=5
James Willard Towne
B: 1870 Lyon County, KY
William Samuel Towne
B: 1878 Kewanee, Henry CO, IL
T3 Towne
Δ=0
bin(0,6x37,r)
Benjamin Towne
Selwyn H. Towne, Jr
B: 1904 Lynn, MA
Towne
B: 1907

N=10
T2 Towne
Δ=23
Miles Willis Town
B: 1904 Lyon County, KY
Norman Millard Towne
B: 1911 Chicago, Cook Co, IL
Benjamin Towne
T5 Towne
Δ=3
bin(1,10x37,r)
T13 Towne
Δ=4 (of 12)

N=11
T8 Town
Δ=1
bin(1,5x37,r)
T6 Towne
Δ=0
bin(0,5x37,r)
T4 Towne
Δ=1
bin(1,11x37,r)

as of March 6, 2008

So we have a statistical model for the data!
Any statistical model is just a way to compute $P(\text{data}|\text{parameters})$

Our model is not “exact”, but statistical models rarely (never?) are.

neglects backmutations
assumes single probability for all loci
etc.

The model is:

$$P(\text{data}|r) = \text{bin}(0,3\times 37,r)\,\text{bin}(0,3\times 37,r)\,\text{bin}(1,5\times 37,r)\,\text{bin}(0,5\times 37,r)$$
$$\times\,\text{bin}(0,6\times 37,r)\,\text{bin}(1,11\times 37,r)\,\text{bin}(3,10\times 37,r)$$

Bayes estimation of the parameter:

$$P(r|\text{data}) \propto P(\text{data}|r)\times P(r) \propto P(\text{data}|r)\times \frac{1}{r}$$

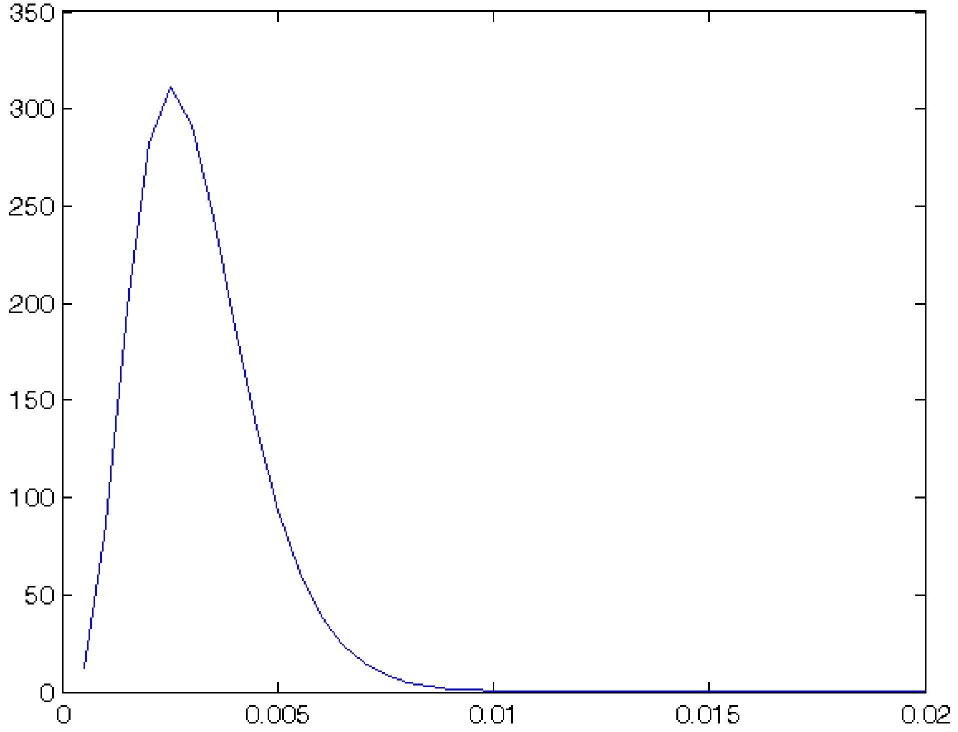
What kind of prior is this???
It is called “log-uniform”

The log-uniform prior has equal probability in each order of magnitude.

$$\int_r^{10r} P(r)dr = \int_r^{10r} \tfrac{1}{r}dr = \log 10$$

It is often taken as the non-informative prior when you don’t even know the order of magnitude of the (positive) quantity.
It is an “improper prior” since its integral is infinite.
This is almost always ok, but it is possible to construct paradoxes with improper priors (e.g., the “marginalization paradox”)

Here is the plot of the (normalized) $P(r|\mathrm{data})$

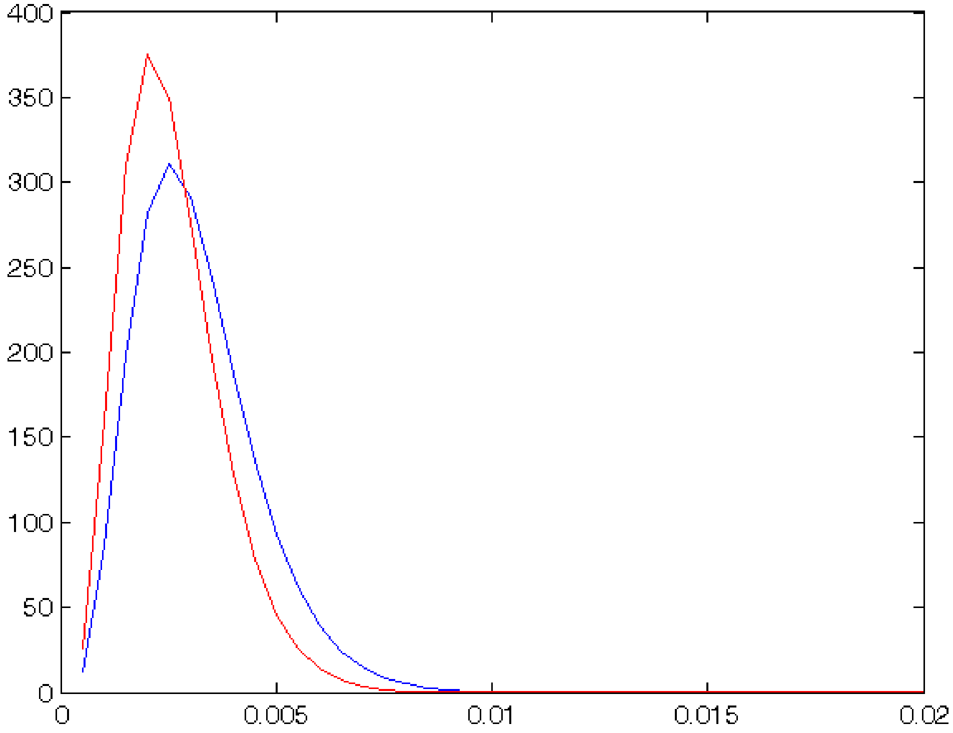


This is (almost) real biology. We've measured the mutation probability, per locus per generation of Y chromosome STRs. This tells us something about the actual DNA replication machinery!

It really did matter (a bit) that we sorted out the conditional dependencies correctly.
Here's a comparison to doing it wrong by assuming all data independent:

The true dependencies allow somewhat larger values of r, because we don't wrongly count the Δ=0 branches multiple times

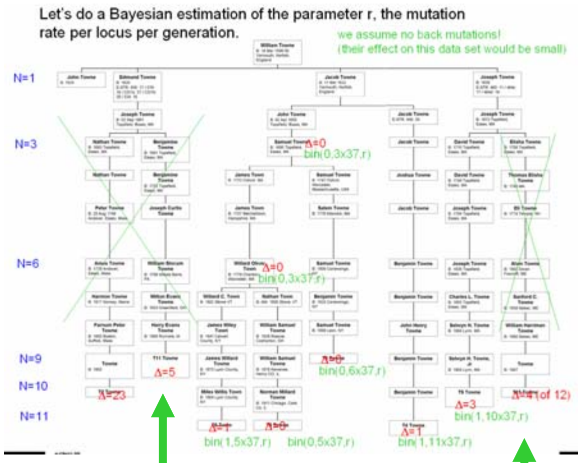
We'll come back to the Towne family for some fancier stuff later!



Ignoring conditional dependencies and just multiplying the probabilities of the data as if they were independent is called naïve Bayes. People often do this. It is mathematically incorrect, but sometimes it is all you can do!

A small cloud: The way we "trimmed" the data mattered. (And should trouble us a bit!) Here's the effect of including T11 and T13, both of which seemed to be outliers:

$$P(\text{data}|r) = [\text{old model}] \times \text{bin}(5, 9 \times 37, r)\,\text{bin}(4, 10 \times 12, r)$$

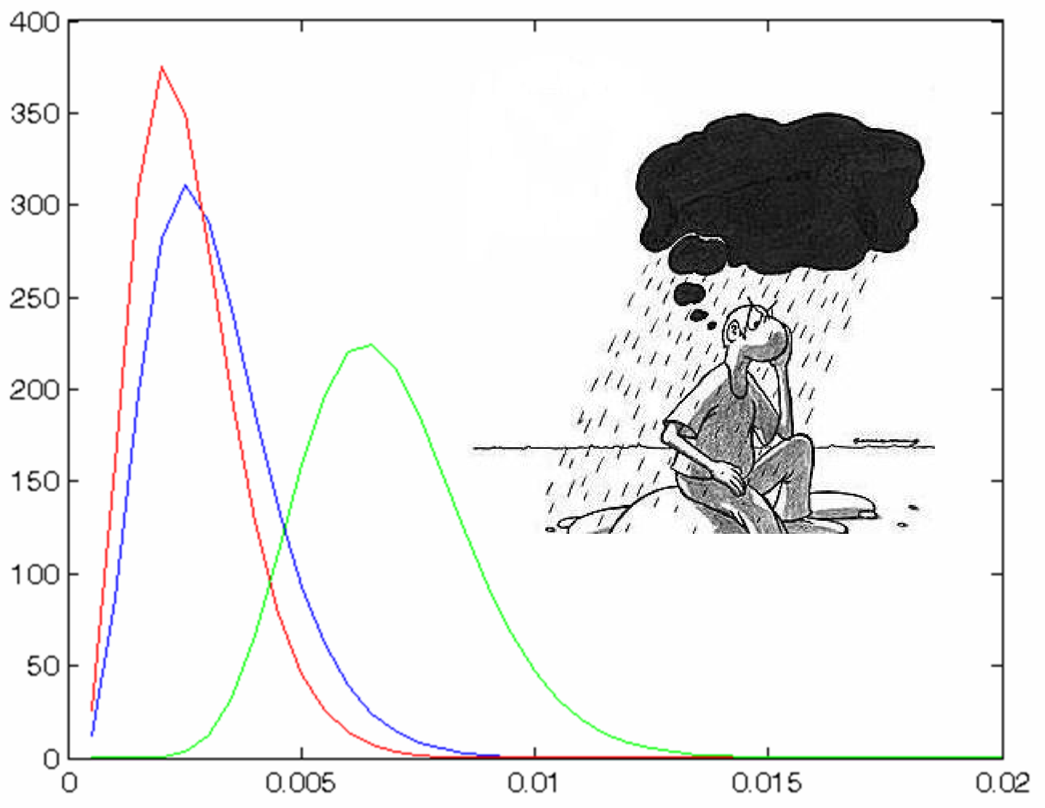


Let's do a Bayesian estimation of the parameter r, the mutation rate per locus per generation.

we assume no back mutations! (their effect on this data set would be small)

now include these

Editing outliers is a tricky issue that we will return to when we learn about mixture models.



400
350
300
250
200
150
100
50
0
0
0.005
0.01
0.015
0.02