The history of U106

(1) INTRODUCTION

This deep phylogenic tree of the human population represents our current understanding of the way the human family tree has divided along its male lines. This is a rapidly-evolving field, thus the information is subject to considerable change over time.

This tree summarises the extensive tree that lies above U106, which contains the vast majority of the world's population. Below U106, we show every known branch (whether or not it has a known associated SNP mutation) and include the results of a comprehensive dating analysis of nearly 1500 tests that was undertaken in autumn 2013 by the U106 group.

(2) OUT OF AFRICA

Ultimately, we all descend from the first life-forms, which lived approximately three billion years ago. Through a long and convoluted process, they evolved into *homo sapiens*. While *H. sapiens* has only been around for about half a million years, this is still older than the common ancestor of the male lines of every person alive today. We call this person Y-chromosomal Adam, because we all descend from him via our father's father's father's… etc. Recent estimates of his age vary widely from 120,000 to 581,000 years ago.

The vast majority of people descend through Haplogroup A. In fact, it's only recently that researchers discovered our most-distant relations hiding among remote Africa tribes. Haplogroup A arose in Africa about 60,000 to 120,000 years ago, when the most of the human population consisted of a small number of tribes living in the Horn of Africa.

The human genetic tree continued to diversify and flourish as mankind expanded throughout Africa. Around 50,000 to 60,000 years ago, a small group of migrants is thought to have crossed the Red Sea into Arabia, starting the most important in a series of Out of Africa migrations.

Our base haplogroup, R, arose from this migration between 20,000 and 34,000 years ago. By this time, humans had probably reached the Caucasus or Russian Steppe, where they existed as hunter gatherers. Before this point, about 40,000 to 45,000 years ago, we split from haplogroup I, which forms about a fifth of Europe's population.

(3) INTO EUROPE

Within haplogroup R, most people are part of R1, descended from an individual living 12,500 to 25,700 years ago. The majority of western Europe is descended from the R1 founder. Within R1, there is a bifurcation into two groups: R1a, or M420, and R1b, or M343. R1a is strongest in eastern populations, where it can exceed 60% of individuals in Poland and the south-west Russian states. Many British R1a may have Viking roots.

R1b (M343) is thought to have arisen less than 18,500 years ago. It is very much dominated by R1b1a2, or M269. This group alone makes up over half the population in Western Europe, and makes up over 90% of some populations. Despite this, its origins are still thought to have been in western Asian populations, and it came to dominate Europe as it expanded throughout the continent at the end of the last Ice Age.

R1b1a2 is estimated to be some 4,000 to 10,000 years old, and is in turn dominated by a sub-clade, L11 (equivalent to P310). Again, the hypothesis is that this population arose in western Asia, and migrated through eastern Europe (possibly up the Danube valley) to dominate western European lines.

L11 splits between the larger P312, which is more frequent south-west of the Rhine, and the slightly smaller U106, dominating regions north and east of the Rhine. Generally, P312 occurs more frequently within Celtic populations, while U106 overlaps with Germanic populations, though there is a broad Celtic-Germanic mix in both. Our age analysis supports a rapid expansion in population at this time.

U106 is only slightly younger than L11 and represents about 1/8th of Europe. Our current hypothesis is that it originated somewhere near the upper Danube valley. It is found throughout Europe, but is concentrated in areas later settled by Germanic groups, north of the boundaries of the Roman Empire.

Structure of R-U106
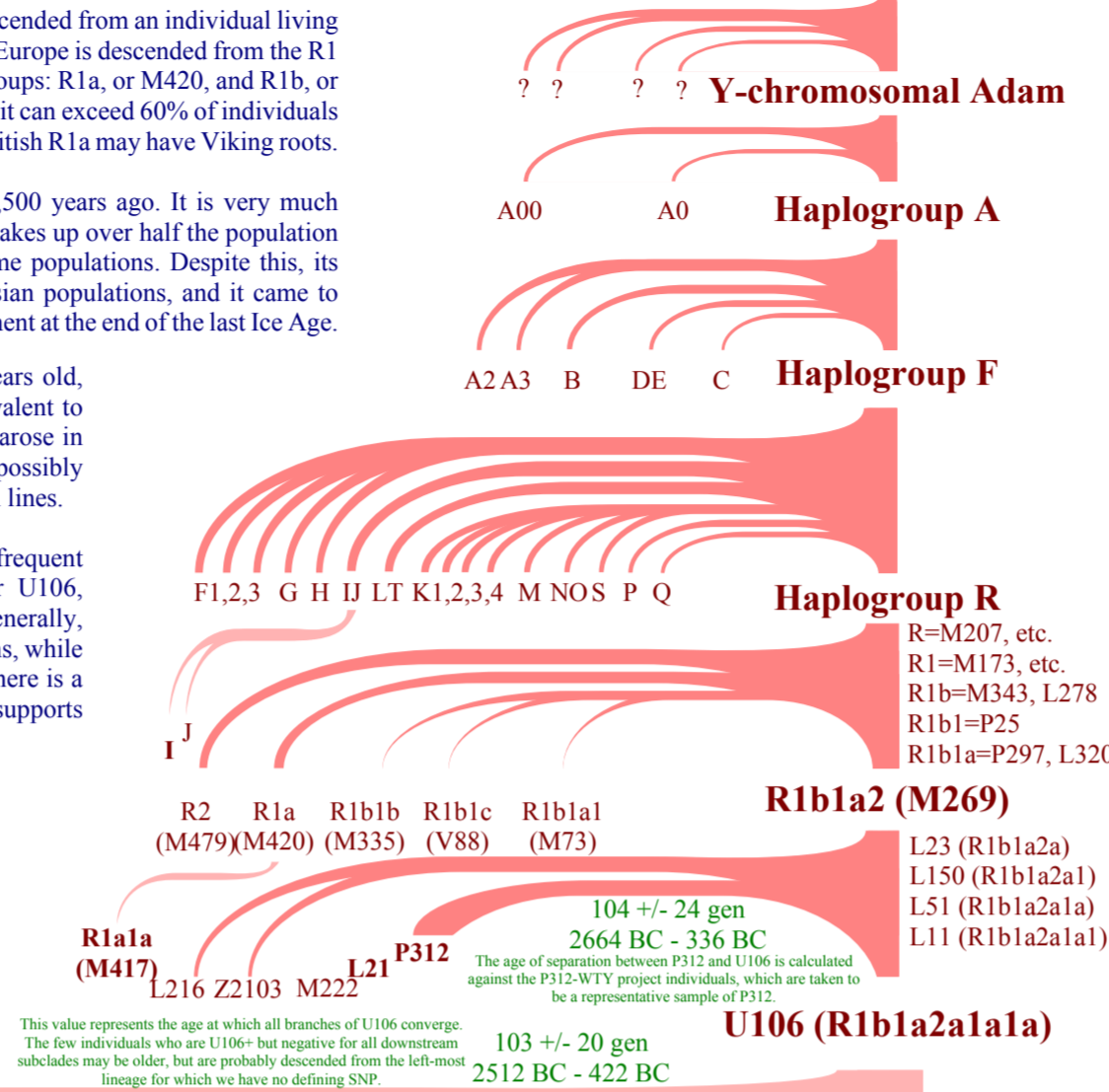
How to read this chart

What is shown

This chart shows how the male-line genetic (phylogenic) tree splits from its foundation down to the present. The entire tree structure below U106 is shown. Different ages and distances are shown on the chart, which should be interpreted carefully. In particular, we have:

GREEN AGES: These represent *inter-clade* ages: the date given is where the branch below where the date is placed diverged. These are the most-accurate dates we have.

Where these dates are given in *italics*, they are less accurate, as they represent comparisons to people who are negative for the only main downstream SNP (e.g. Z305+ and Z305- within Z156).

MAGENTA AGES: These represent *intra-clade* ages. These are used where inter-clade ages are not possible to calculate, or where an inter-clade age only describes a fraction of the population as most people have not tested positive for sub-clades, we have also provided intra-clade ages. These compare the typical age of relationships within a group and are less accurate, normally they will under-estimate the age of the whole group. Intra-clade ages have been calculated by comparing two different parts of the group to each other to minimise this effect.

Ages are given as 95.5% confidence intervals, what we call "2-sigma". We are 95.5% sure that the real dates lie between these two boundaries. By dividing the uncertainty in half, we can recover the 68% confidence interval, or "1-sigma" range. For example, we are 95.5% sure that the U106 founder lived between 83 and 123 generations ago, between 2512 BC and 422 BC. We are 68% sure that he lived between 93 and 113 generations ago, between 1990 BC and 945 BC.

Note that this variance-based method has several strengths and weaknesses compared to the Bayesian-based method used by, e.g., the McGee tool. It better accounts for back mutations (reversions) in the DNA, but it cannot account for multi-step mutations (e.g. markers moving from 12 to 10) and does not take into account markers that do not mutate at all. These differences mean it tends to give an older age than the the Bayesian-based method. Variance-based methods are better for older clusters, Bayesian methods for younger clusters. The cross-over where the uncertainties are the same for both methods lies somewhere close to 2000 years old.
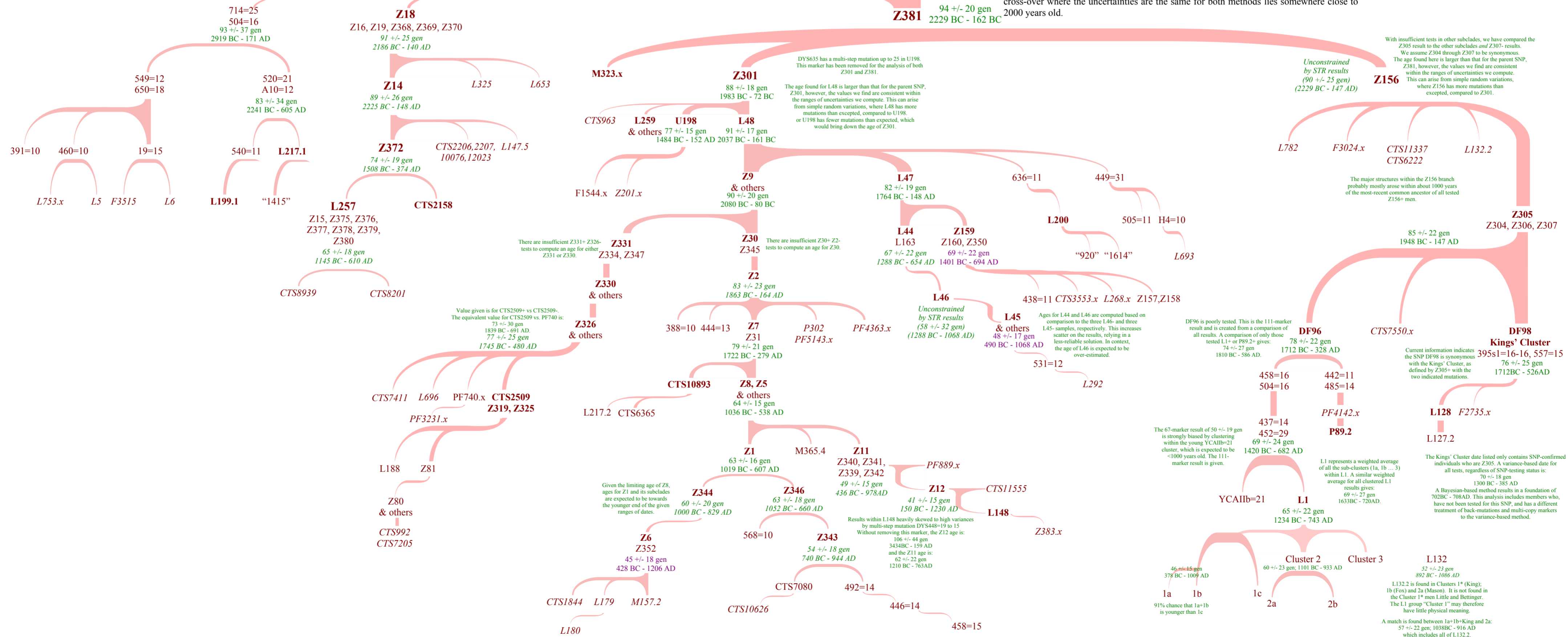
Caveats

It is important to remember the limitations of this approach, and the assumptions we made in producing it. Full details of this can be found in the attached documentation. We have tried to account for all sources of uncertainty, but there are always "unknown unknowns" outside our control. This chart is based on:
* An average of four sets of mutation rates.
* A mixture of 67 and 111 marker data.
* A sliding scale of 35 years/generation at present to 25 years/generation before 2000 BC.

Tree labels (chart)

Homo sapiens

? ? ? **Y-chromosomal Adam**

A00 A0 **Haplogroup A**

A2 A3 B DE C **Haplogroup F**

F1,2,3 G H IJ LT K1,2,3,4 M NOS P Q **Haplogroup R**

R=M207, etc.
R1=M173, etc.
R1b=M343, L278
R1b1=P25
R1b1a=P297, L320

I J

R2 (M479) R1a (M420) R1b1b (M335) R1b1c (V88) R1b1a1 (M73) **R1b1a2 (M269)**

L23 (R1b1a2)
L150 (R1b1a2a1)
L51 (R1b1a2a1a)
L11 (R1b1a2a1a1)

R1a1a (M417)

L216 Z2103 M222 **L21** **P312**
104 +/- 24 gen
2664 BC - 336 BC
The age of separation between P312 and U106 is calculated against the P312-WTY project individuals, which are taken to be a representative sample of P312.

U106 (R1b1a2a1a1a)

103 +/- 20 gen
2512 BC - 422 BC

This value represents the age at which all branches of U106 converge. The few individuals who are U106+ but negative for all downstream subclades may be older, but are probably descended from the left-most lineage for which we have no defining SNP.

Z381 94 +/- 20 gen
2229 BC - 162 BC

714=25
504=16
93 +/- 37 gen
2919 BC - 171 AD

Z18
Z16, Z19, Z368, Z369, Z370
91 +/- 25 gen
2186 BC - 140 AD

520=21
A10=12
83 +/- 34 gen
2241 BC - 605 AD

549=12
650=18

M323.x

Z301
88 +/- 18 gen
1983 BC - 72 BC

DYS635 has a multi-step mutation up to 25 in U198. This marker has been removed for the analysis of both Z301 and Z381.

Unconstrained by STR results (90 +/- 25 gen) (2229 BC - 147 AD)

Z156
With insufficient tests in other subclades, we have compared the Z305 result to the other subclades and Z307- results. We assume Z304 through Z307 to be synonymous. The age found here is larger than that for the parent SNP, Z381, however, the values we find are consistent within the ranges of uncertainties we compute. This can arise from simple random variations, where Z156 has more mutations than excepted, compared to Z301.

Z14
Z16, Z19 ...
89 +/- 26 gen
2225 BC - 148 AD

L325 *L653*

CTS963
L259
& others 77 +/- 15 gen
1484 BC - 152 AD

U198
91 +/- 17 gen
2037 BC - 161 BC

L48
The age found for L48 is larger than that for the parent SNP, Z301, however, the values we find are consistent within the ranges of uncertainties we compute. This can arise from simple random variations, where L48 has more mutations than expected, compared to U198. or U198 has fewer mutations than expected, which would bring down the age of Z301.

L782 *F3024.x* *CTS11337* *L132.2*
S36222

391=10 460=10 19=15

540=11 **L217.1**

Z372
74 +/- 19 gen
1508 BC - 374 AD

CTS2206,2207, L147.5 10076,12023

Z9
& others 90 +/- 20 gen
2080 BC - 80 BC

L47
82 +/- 19 gen
1764 BC - 148 AD

636=11 449=31

The major structures within the Z156 branch probably mostly arose within about 1000 years of the most-recent common ancestor of all tested Z156+ men.

L753.x *L5* *F3515* *L6* **L199.1** "1415"

L257
Z15, Z375, Z376, Z377, Z378, Z379, Z380
65 +/- 18 gen
1145 BC - 610 AD

CTS2158

F1544.x *Z201.x*

Z30
Z345

L44
L163
67 +/- 22 gen
1288 BC - 654 AD

Z159
Z160, Z350
69 +/- 22 gen
1401 BC - 694 AD

L200 505=11 H4=10

"920" "1614" *L693*

Z305
Z304, Z306, Z307

85 +/- 22 gen
1948 BC - 147 AD

Z331
Z334, Z347
There are insufficient Z331+ Z326- tests to compute an age for either Z331 or Z330.

Z30
Z345
There are insufficient Z30+ Z2- tests to compute an age for Z30.

Z2
83 +/- 23 gen
1863 BC - 164 AD

L46
Unconstrained by STR results (58 +/- 32 gen) (1288 BC - 1068 AD)

438=11 *CTS3553.x* *L268.x* Z157,Z158

Ages for L44 and L46 are computed based on comparison to the three L46- and three L45- samples, respectively. This increases scatter on the results, relying in a less-reliable solution. In context, the age of L46 is expected to be over-estimated.

DF98
Kings' Cluster
395s1=16-16, 557=15
76 +/- 25 gen
1712BC - 526AD

Z326
& others
Value given is for CTS2509+ vs CTS2509-. The equivalent value for CTS2509 vs. PF740 is: 73 +/- 30 gen
1839 BC - 691 AD.
77 +/- 25 gen
1745 BC - 480 AD

388=10 444=13

Z7
Z31

P302 *PF5143.x*

PF4363.x

L45
& others
48 +/- 17 gen
490 BC - 1068 AD

531=12

DF96 is poorly tested. This is the 111-marker result and is created from a comparison of all results. A comparison of only those tested L1+ or P89.2+ gives:
74 +/- 27 gen
1810 BC - 586 AD.

DF96
78 +/- 22 gen
1712 BC - 328 AD

CTS7550.x

DF98
Current information indicates the SNP DF98 is synonymous with the Kings' Cluster, as defined by Z305+ with the two indicated mutations.

L128 *F2735.x*

L127.2

Z330
& others

CTS8939 **CTS8201**

Z8, Z5
& others
64 +/- 15 gen
1036 BC - 538 AD

L292

458=16
504=16

442=11
485=14

PF4142.x

L132.2 is found in Clusters 1* (King); 1b (Fox) and 2a (Mason). It is not found in the Cluster 1* men Little and Bettinger. The L1 group "Cluster 1" may therefore have little physical meaning.

A match is found between 1a+1b+King and 2a: 57 +/- 23 gen, 1038BC – 916 AD which includes all of L132.2.

CTS7411 *L696* **PF740.x** **CTS2509**
Z319, Z325

CTS10893

L217.2 CTS6365

Z1
63 +/- 16 gen
1019 BC - 607 AD

M365.4

Z11
Z340, Z341, Z339, Z342
49 +/- 15 gen
436 BC - 978AD

PF889.x

437=14
452=29

P89.2

The 67-marker result of 50 +/- 19 gen is strongly biased by clustering within the young YCAIIb=21 cluster, which is expected to be <1000 years old. The 111-marker result is given.

PF3231.x

L188 Z81

Z8, Z5
& others

Z344
60 +/- 20 gen
1000 BC - 829 AD
Given the limiting age of Z8, ages for Z1 and its subclades are expected to be towards the younger end of the given ranges of dates.

Z346
63 +/- 18 gen
1052 BC - 660 AD

Z12
150 BC - 1230 AD

CTS11555

L148
Results within L148 heavily skewed to high variances by multi-step mutation DYS448=19 to 15. Without removing this marker, the Z12 age is: 106 +/- 44 gen
3434BC – 159 AD and the Z11 age is: 62 +/- 22 gen
1210 BC – 763AD

YCAIIb=21

L1
65 +/- 22 gen
1234 BC - 743 AD

L1 represents a weighted average of all the sub-clusters (1a, 1b ...) within L1. A similar weighted average for all clustered L1 results gives: 69 +/- 27 gen
1633BC – 720AD.

Z80
& others

CTS992
CTS7205

Z6
Z352
45 +/- 18 gen
428 BC - 1206 AD

568=10

Z343
54 +/- 18 gen
740 BC - 944 AD

Z383.x

Cluster 2 **Cluster 3**

L132
52 +/- 23 gen
892 BC - 1086 AD

L132.2 is found in Clusters 1* (King); ...

46 +/- 15 gen
378 BC - 1009 AD

CTS1844 *L179* *M157.2*

CTS7080 492=14

CTS10626 446=14

1a 1b 1c 2a 2b

91% chance that 1a=1b is younger than 1c

60 +/- 23 gen; 1101 BC - 933 AD

L180

458=15