Works of the Eurasian Society for Genetic Genealogy

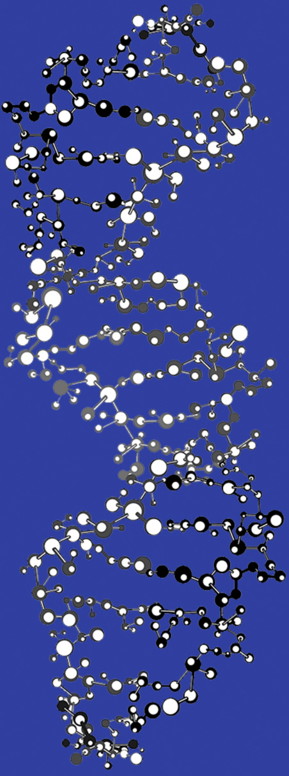
Genetic History of Eurasian Populations

Collection of articles, 2015

Труды Евразийского общества генетической генеалогии

Генетическая история народов Евразии

Сборник статей, 2015



Труды Евразийского общества генетической генеалогии

Генетическая история народов Евразии

Сборник статей, 2015

Works of the Eurasian Society for Genetic Genealogy

Genetic History of Eurasian Populations

Collection of articles, 2015

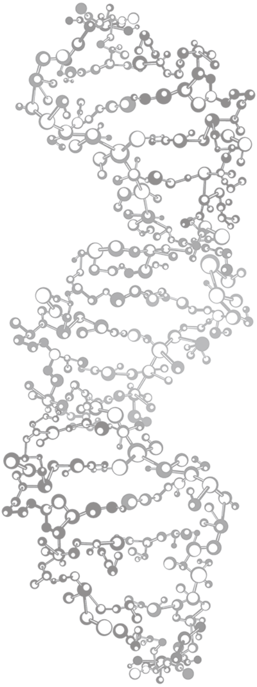
Издательские решения
По лицензии Ridero
2016

УДК 57
ББК 28

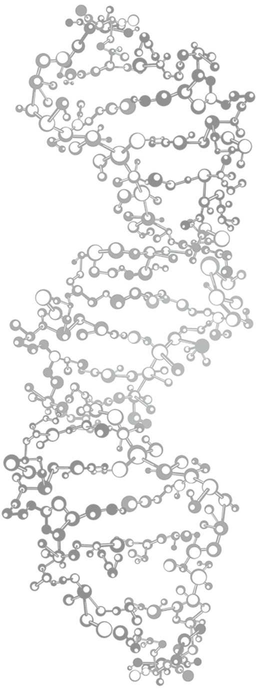
Труды Евразийского общества генетической генеалогии : Генетическая история народов Евразии. Сборник статей, 2015. — [б. м.] : Издательские решения, 2016. — 318 с. — ISBN 978-5-4474-8041-7

УДК 57
ББК 28

ISBN 978-5-4474-8041-7



Оглавление



The Russian Journal of Genetic Genealogy (Русская версия): Том 7, №1, 2015 год, ISSN: 1920-2997. Принята 9 февраля 2015 года; опубликована 24 февраля 2015 года.

Defining a New Rate Constant for Y-Chromosome SNPs based on Full Sequencing Data

Dmitry Adamov
Vladimir Guryanov
Sergey Karzhavin
Vladimir Tagankin
Vadim Urasin

Abstract

Two important advances: 1) the accumulation of Big Y and FGC test data, and 2) the publication of Y-chromosome sequences for three ancient samples (Anzick-1, Ust-Ishim, and K14), have made it possible to estimate the average rate of base-substitutions (SNPs). The authors of this study have developed a new method of selecting true mutations in modern and ancient samples, and have defined with high accuracy the rate constant of SNP mutations: $0.82 \cdot 10^{-9}$ per year per bp (95% CI: $(0.70 - 0.94) \cdot 10^{-9}$).

Introduction

A single nucleotide polymorphism (SNP) is a DNA sequence variation in which a single nucleotide (A, T, G or C) in a genome (or another shared sequence) differs among members of a species or between paired chromosomes. The authors use “SNP mutation”, “base-substitution”, and “mutation” interchangeably.

Substitution of one nucleotide with another during meiosis occurs at random. The probability of SNP mutations is very small: few replacements per one hundred million base pairs (nucleotide sites) occur in the course of a single meiosis. The mutation flow is rare (standard probability) and subsequent mutations do not depend on previous mutations. These characteristics determine the mutations flow as a Poisson process. The probability of mutations in the meiosis sequence within T generations is estimated using a Poisson distribution:

$$P_k = \frac{(\mu T)^k}{k!} e^{-\mu T},$$

where P_k is the probability of k mutations occurring across T generations within the same nucleotide site of the chromosome, and μ is the rate constant of base-substitutions.

In practice, many nucleotide sites are measured simultaneously. Let a total number of measured base pairs be denoted as B. The average number of SNP mutations (N_{SNP}) is determined by the ratio

$$N_{SNP} = \mu_{SNP} TB, \qquad (1)$$

where $\mu_{SNP} \equiv \mu$ is the rate constant of SNP mutations. For brevity's sake, we call this the «mutation rate».

Measuring the number of mutations N_{SNP} for genealogical purposes became possible a few years ago when high-performance Next Generation Sequencing (NGS) technologies capable of large-scale parallel reading of genomes became available at a reasonable cost.

The commercial NGS testing of Y-chromosome samples began in 2013. Table 1 summarizes the data on coverage of Y-chromosome sequences by commercial and research laboratories based on NGS technology.

The mapped sequences of the Y-chromosome supplied by NGS technology average about 23 Mbp in length.

Skaletsky et al. (2003) noted that the structure of the Y-chromosome is heterogeneous. Y-chromosome euchromatin consists of the nucleotide sequences of the following types:

1) X-transposed, which have 99% identical analogues in the X-chromosome (total length is approximately 3.7 Mbp),

2) X-degenerated, which are unique and easily mapped sequences (total length is approximately 8.6 Mbp),

3) Ampliconic, which are segments 99.9% similar to the sequences located in other parts of the Y-chromosome (total length is approximately 10.2 Mbp, including 8 palindromic segments of 5.7 Mbp).

Table 1.

The overall size and reading quality of Y-chromosome segments.

Research (Test)	**Mapping area, Mbp**	**Coverage, X times**
FTDNA Big Y	up to 11.38	60
FGC Elite	~ 23	60
1000 Genomes Project	~ 23	2-4
Personal Genome Project	~ 23	No data
Poznik et al. (2013)	9.99	3.1
Wei et al. (2013)	8.97	28.4
Francalacci et al. (2013)	8.97	2.16
Yan et al. (2014)	3.90	10
Hallast et al. (2015)	3.72	51

In scientific works, research is limited, as a rule, to the X-degenerated area. Researchers try to avoid X-transposed and ampliconic sequences. Wei et al. (2013) wrote:

> *We identified unique regions within the male-specific part of the Y chromosome reference sequence . . . where we expected read mapping and variant detection to escape complications introduced by repeated sequences. That was achieved by excluding the pseudoautosomal, heterochromatic, X-transposed, and ampliconic segments.*

In Poznik D. et al. (2013) 9.99 Mbp segments of euchromatin which are the most suitable for mapping short reads (about 100 bp) were defined. These segments are basically X-degenerated and, to a lesser extent, ampliconic. Palindromic segments were excluded.

To estimate the age of male genealogical lines using the number of detected derived variants the ratio (1) is transformed as follows:

$$T = \frac{N_{SNP}}{\mu B} \qquad (2)$$

The size B of the measured and mapped area is evaluated using the BED file.

The mean mutation rate is estimated by using different calibration methods (see Wang, Gilbert, Jin, Li, 2014).

1. Kuroki et al. (2006) compared the Y-chromosome of humans and chimpanzees. Assuming that the species division occurred 6 million years ago, the average mutation rate was estimated at $1.5 \cdot 10^{-9}$ per year per bp, 95% CI: $(0.767 - 2.10) \cdot 10^{-9}$.

2. In their well known work, Xue et al. (2009) used samples with modern genealogies to calibrate mutation rates. Four mutations were detected in 13 generations in area with a total length of 10.15 Mbp. The mutation rate estimate was $1.0 \cdot 10^{-9}$ per year per bp, 95% CI: $(0.3 - 2.5) \cdot 10^{-9}$.

3. The calculations of Mendez et al. (2013), which were based on the mutation rate in autosomal chromosomes led to a slower rate at $0.617 \cdot 10^{-9}$ per year per bp, 95% CI: $(0.439 - 0.707) \cdot 10^{-9}$.

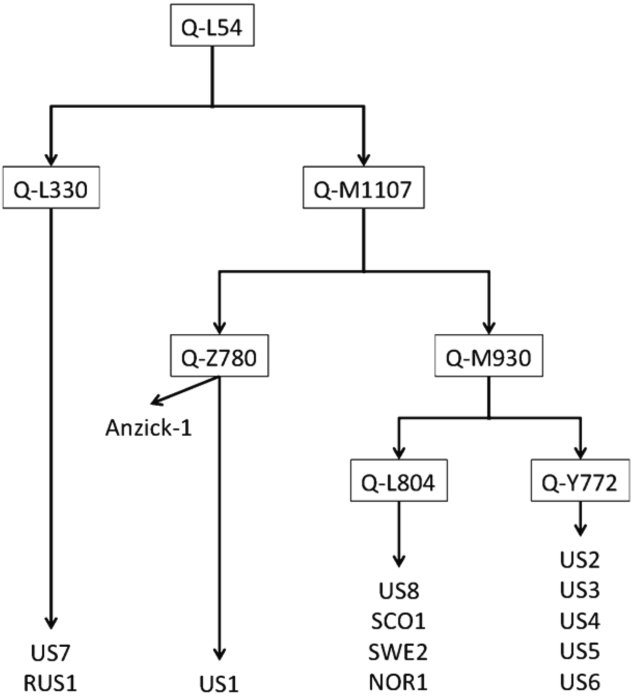
4. Working from the premise that America was settled by humans 15,000 years ago, Poznik et al. (2013) estimated the average mutation rate at $0.82 \cdot 10^{-9}$ per year per bp, 95% CI: $(0.72 - 0.92) \cdot 10^{-9}$.

5. Recent progress in the sequencing of ancient human skeletal remains has made it possible to perform a direct calibration of the average mutation rate using the number of derived alleles accumulated from the time of ancient humans. Fu et al. (2014) published the data on the complete genome of an ancient Ust-Ishim Man who lived about 45 thousand years ago in Western Siberia. On analyzing a Y-chromosomal area 1.86 Mbp long, Fu et al. estimated the rate for SNP mutations at $0.76 \cdot 10^{-9}$ per year per bp, 95% CI: $(0.67 - 0.86) \cdot 10^{-9}$.

Results

Selection of derived alleles

The variants of derived alleles differing from the reference sequence are contained in VCF files. Not all of them reflect actual mutations, however. Because of the peculiarities of NGS technology, the imperfections of mapping methods, the presence of repetitive chains of nucleotide sequences in the Y-chromosome, and the coincidence of sequences with analogues in other chromosomes, some of the derived variants are not real mutations. The share of erroneous choices for X-degenerated segments is not high (between 15% and 25% of the total). In ampliconic areas, however, erroneous options can be greater than the number of real mutations.



US1	Native American descendant
US7	European origin
RUS1	Russian
US2	Native American descendant
US3	Native American descendant
US4	Native American descendant
US5	Native American descendant
US6	Native American descendant
US8	European origin
SCO1	Scotland origin
SWE2	Sweden origin
NOR1	Norwegian origin

Figure 1. The family tree of haplogroup Q-L54 constructed on base-substitutions of 12 private samples and the ancient Anzick-1 sample.

The method of selection of real mutations developed in the present study allows for elimination of erroneous alternatives from the analysis. The description of our method is contained under "Materials and Methods" below. Our method is based on the selection of X-degenerated sequences. Our SNP mutation rate calibration was carried out in what we call the "combBED" area (combined BED), which contains start and end coordinates in the hg19 system of the Y-chromosome segments, in which we expect our samples to have SNP variants. Table 1 of the Supplement to this article shows the location of 857 "good" regions of the Y-chromosome (total length of 8,473,821 bp). SNP mutation rate calibration was carried out for these areas, which will be further referred to as "combBED area".

Calibration using the ancient Anzick-1 sample

Rasmussen et al. (2014) wrote about a sample (Anzick-1) which was obtained from the bone remains of a boy who lived about 12.6 thousand years ago in the territory of what is now the State of Montana, USA. The sample was perfectly preserved, which allowed Rasmussen et al. (2014) to derive a high-quality genome sequencing for the boy. The mutations in the boy's Y-chromosome placed him in haplogroup Q-L54. The age of the remains was determined with very high precision by radiocarbon dating. According to Rasmussen et al. (2014), the boy lived between 12,707 and 12,556 years before the present. On average, 12632 years BP.

The YFull database has accumulated unique samples of Q-L54 which we used to construct a detailed family tree for haplogroup Q-L54. The tree is shown in Figure 1. Anzick-1 is located on the same branch (Z780) as a contemporary sample taken from a present-day Native American descendant (US1). Twelve thousand six hundred years ago, US1's ancestor was a close relative of the Anzick-1 boy.

The structure of the Q-L54 tree required that the mutation rate be calibrated in two stages. In the preliminary stage, we computed the relatively short period between the time the Anzick-1 boy lived (Z780) and the separation of L330 and M1107. Our calibrations revealed that the separation of these branches occurred very close to the generally accepted time of the settlement of Anatomically Modern Humans (AMH) in America (about 15 thousand years ago).

For our pre-calibration, we used the data of Anzick-1 and the data of US1 who had common ancestors at level Z780. After Z780, their lines were distinguished by distinct, differing mutations. To date, the male line of US1 has accumulated 78 mutations. In the Anzick-1 sample only one mutation is revealed later Z780 level. Based on the radiocarbon dating of the Anzick-1 sample the average rate of base-substitutions is estimated as $0.831 \cdot 10^{-9}$ per year per bp, 95% CI: $(0.66 - 1.04) \cdot 10^{-9}$.

We used this evaluation only for calculating the time interval between the split of the L330 and the M1107 branches and the time that Anzick-1 lived. According to the number of mutations detected in the Anzick-1 and US1 samples, the interval is approximately 2,989 years (95% CI: 1,814-4,751 years). The age of the L330 and the M1107 branches is 15,621 years before the present (95% CI: 14,446-17,383 years). The dating does not depend on any model, since the data are based on radiocarbon analysis of the Anzick-1 sample and the molecular clock of the US1 and Anzick-1 samples.

We secured eleven haplotypes from Big Y and one from FGC which were provided by contemporary individuals who belong to the branches under investigation. The samples set up four branches: Z780, L330, Y772, L804 (see Figure 1 above). With a high degree of accuracy, the branches can be assumed to be independent. This allows us to reduce the relative variance in the calibration. The results of the calibration process are summarized in Table 2.

Based on the average number of mutations (392 out of 30.65 million b.p.), and the age of the L330 and the M1107 branches (15,621 years), we arrive at a final calibration rate of SNP mutations

$$\frac{392}{30.65 \cdot 10^{6} \cdot 15621} = 0.819 \cdot 10^{-9} \text{ per year per bp.}$$

$$95\% \text{ CI: } (0.704 - 0.935) \cdot 10^{-9}.$$

To recalculate an average rate per one generation, we take an average interval of one generation of men equal to 31.5 years (Fenner, 2005):

$$0.819 \cdot 10^{-9} \cdot 31.5 = 2.58 \cdot 10^{-8} \text{ per generation per bp.}$$

Table 2.

Number of mutations in 12 contemporary samples selected to calibrate the average rate of SNP mutations at the time of separation of the L330 and the M1107 branches.

Branch	Sample	Number of mutations later than L54 level	combBED area size, Mbp	Average combBED area size, Mbp	Average umber of mutations per branch	Lower 95% confidence interval	Upper 95% confidence interval
Z780	US1	92	7.335	7.335	92.00	74.19	112.83
L330	US7	102	7.527	7.639	104.50	85.46	126.56
L330	RUS1	107	7.751				
Y772	US2	93	8.129	7.741	88.00	70.60	108.42
Y772	US3	90	7.856				
Y772	US4	85	7.365				
Y772	US5	88	7.693				
Y772	US6	84	7.660				
L804	US8	105	7.895	7.937	107.50	88.17	129.85
L804	SCO1	100	7.900				
L804	SWE2	115	8.072				
L804	NOR1	110	7.882				
	Total:			30.65	392.00	354.26	432.83

Calibration with modern genealogies

Some of the samples in the YFull database belong to relatives with paternal lines which can be proved by documentation. We chose 41 samples representing 14 genealogies. The number of generations to the most recent common ancestor in the samples varied from 1 to 23. We used the property that the sum of random numbers distributed according to the Poisson is a random number itself distributed according to the Poisson. In our case, it is the sum of the accumulated mutations.

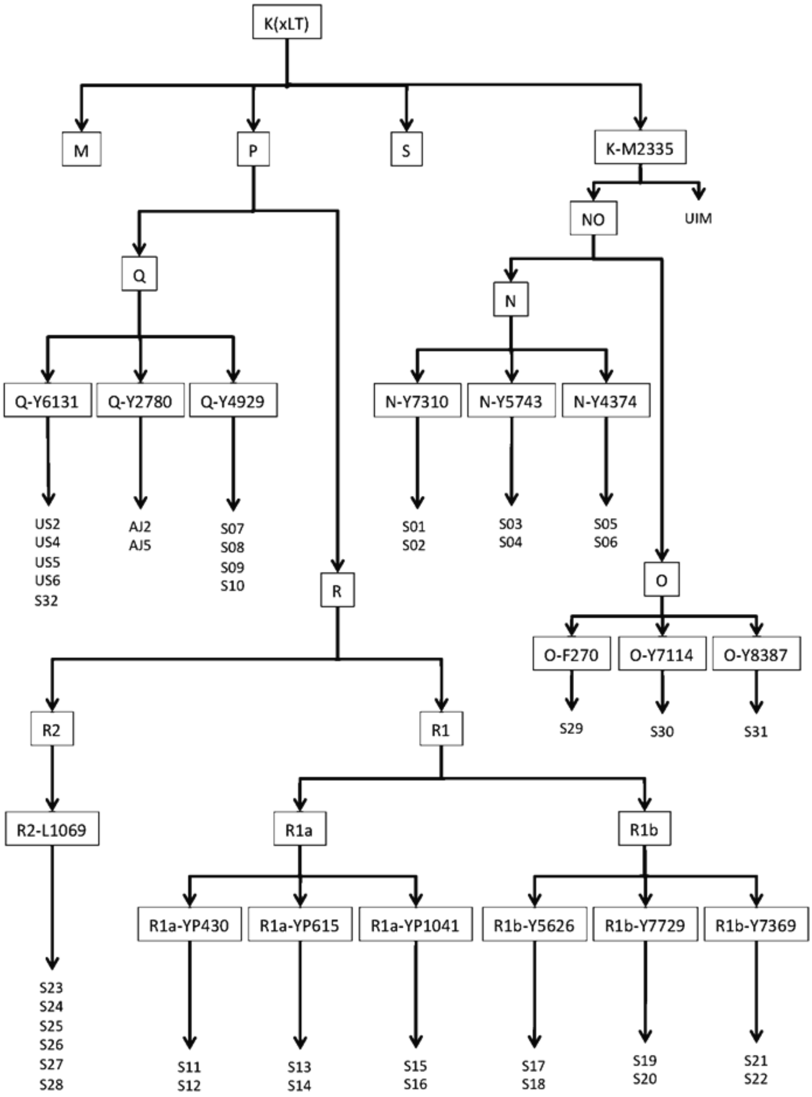


Figure 2. Haplogroup K(xLT) family tree constructed on base-substitutions of 38 private samples and the ancient sample of Ust-Ishim Man (UIM).

The advantage of this calibration is that the exact number of generations in which private mutations occurred is known. The result of the calibration based on modern genealogies is:

$$2.56 \cdot 10^{-8} \text{ per generation per bp, 95\% CI: } (2.03 - 3.16) \cdot 10^{-8}.$$

This is in agreement with the calibration on the dating of Anzick-1 calculated for the average interval of one generation of 31.5 years.

The average interval between generations in the chosen genealogies is 32.1 years. Therefore, the absolute calibration for one year corresponds to $0.798 \cdot 10^{-9}$ per year per bp.

Calibration using the Ust-Ishim sample

According to radiocarbon dating reported by Fu et al. (2014), Ust-Ishim Man (UIM) lived about 45 thousand years ago. On the family tree below, he is at the beginning of branch K-M2335 (see Fig. 2). Fu et al. (2014) studied the full genome of the ancient sample. The presence of a mutation at M526 placed UIM at the beginning of haplogroup K(xLT). Detailed examination of derived options revealed a later mutation (M2308/Z4842). Studying the BAM file, we identified 11 mutations (including M2308), which emerged after M526 in the combBED area (total length of 8.463 Mbp). This number was used to calculate the average mutation rate.

For the calibration, haplogroups N, O, Q, and R are appropriate.

In order to reduce possible errors to a minimum, we selected from the YFull database 35 samples with at least one fairly close relative and three samples (of haplogroup O) who were not close relatives (see Fig. 2). Table 3 shows the number of identified mutations in these samples downstream of K(xLT).

The average number of mutations occurred later K(xLT) level is 36.36 mut./Mbp.

The correction for UIM mutations is:

$$\frac{11}{8.463} = 1.30 \text{ mut./Mbp.}$$

The estimate of the mutation rate is

$$\frac{36.36 - 1.30}{10^6 \cdot 44890} = 0.781 \cdot 10^{-9} \text{ per year per bp.}$$

$$95\% \text{ CI: } (0.709 - 0.853) \cdot 10^{-9}.$$

Table 3.

The number of actual mutations in 38 samples collected to calibrate an average rate of SNP mutations at Ust-Ishim Man's dating.

№	ID	Branch	combBED area size, bp	Number of mutations later than K(xLT) level	Average number of mutations per 1 Mbp, within branch	Average number of mutations per 1 Mbp, within haplogroup	Mutation rate estimate, $x10^9$ within haplogroup
1	S01	N-Y7310	7452832	260	34.44	34.14	0.732
2	S02	N-Y7310	7415734	252			
3	S03	N-Y5743	7480204	253	33.74		
4	S04	N-Y5743	7517710	253			
5	S05	N-Y4374	7500387	258	34.25		
6	S06	N-Y4374	7769783	265			
7	US2	Q-Y6131	8129326	293	35.09	36.58	0.786
8	US4	Q-Y6131	7364970	258			
9	US5	Q-Y6131	7693497	269			
10	US6	Q-Y6131	7659810	264			
11	S32	Q-Y6131	7650626	267			
12	AJ2	Q-Y2780	8455137	324	38.30		
13	AJ5	Q-Y2780	8230718	315			
14	S07	Q-Y4929	8419169	311	36.34		
15	S08	Q-Y4929	7649280	279			
16	S09	Q-Y4929	7461979	270			
17	S10	Q-Y4929	7510483	268			

18	S11	R1a-YP430	7232725	276	38.15	37.95	0.816
19	S12	R1a-YP430	7418179	283			
20	S13	R1a-YP615	7181725	282	39.22		
21	S14	R1a-YP615	7632739	299			
22	S15	R1a-YP1041	6795545	242	36.47		
23	S16	R1a-YP1041	7902430	294			
24	S17	R1b-Y5626	7650748	270	34.46	35.19	0.755
25	S18	R1b-Y5626	7235989	243			
26	S19	R1b-Y7729	7515672	265	34.89		
27	S20	R1b-Y7729	7446027	257			
28	S21	R1b-Y7369	7855639	287	36.23		
29	S22	R1b-Y7369	7462871	268			
30	S23	R-L1069	7412514	290	39.00	39.00	0.840
31	S24	R-L1069	7640179	298			
32	S25	R-L1069	7777760	302			
33	S26	R-L1069	7826576	304			
34	S27	R-L1069	7476862	287			
35	S28	R-L1069	8070902	321			
36	S29	O-F270	7527844	271	36.00	35.31	0.758
37	S30	O-Y7114	7157490	237	33.11		
38	S31	O-Y8387	7658568	282	36.82		
	average		7611596		36.03	36.36	0.781

Calibration using the K14 sample

Seguin-Orlando et al. (2014) studied a genetic sample (K14) secured from male bones found in Kostenki, Russia. According to radiocarbon dating, the sample is 37 thousand years old. The quality of the K14 genetic sample is not as good as either the Ust-Ishim sample or the Anzick-1 sample.

Published accounts of the sequencing reveal that it is only possible to estimate the upper limit of the mutation rate. Still, in the K14 sample we can confidently read two derived alleles at K29 and CTS6773 (combBED area length of 1.381 Mbp). According to YFull, this is the beginning of haplogroup C1. Unfortunately, the other 15 candidates identified in the sample have poor coverage, mainly 3X-5X and cannot be used for calibration. Still, with the confirmed SNPs, we were able to calculate an upper limit for the mutation rate.

In the YFull database there are four samples from branch C1-K29; details are given in Table 4.

The upper limit of SNP mutation rate is estimated as:

$$\frac{1297}{32.174 \cdot 10^{6} \cdot 37470} = 1.08 \cdot 10^{-9} \text{ per year per bp.}$$

This result agrees with other calibrations, but is not of practical importance.

Table 4.

The number of actual mutations in 4 samples collected to calibrate an average rate of SNP mutations at K14 dating.

Sample	**Number of mutations later than C1-K29 level**	**combBED area size, Mbp**	**Average number of mutations per 1 Mbp**
S33	326	8.042	40.5
S34	303	7.866	38.5
S35	329	8.022	41.0
S36	339	8.244	41.1
Total:	1297	32.174	average 40.3

Coefficients for age estimate

It is convenient to estimate genealogical age directly from the actual number of mutations. The formula for calculation of the coefficient is derived from equation (2). It is:

$$k = \frac{1}{\mu B} \qquad (3)$$

It is essential for researchers to know the value B of the measured area length. For Big Y, the confidence regions length averaged over individual bed files is 10.31 Mbp (Big Y White Paper, 2014). For FGC Elite, the length is 23 Mbp. At the average mutation rate $0.82 \cdot 10^{-9}$ the coefficient (formula 3) is 118 years for Big Y, and 53 years for FGC.

If the size of the measured area changes, it is necessary to recalculate the coefficient. For example, if the measured area of Big Y were not 10.31 Mbp but 11.0 Mbp, then the coefficient would be

$$k = 118 \cdot \frac{10.31}{11} = 111 \text{ years per base-substitution.}$$

For a more effective selection of actual mutations, we recommend that any research area be within the boundaries of the combBED area. The size of the combBED area in individual Big Y samples varies and the average is about 7.6 Mbp. The appropriate conversion factor, therefore, is 160 years per base-substitution.

Discussion

We used four methods of calibrating the rate of SNP mutations. The four methods are within measurement accuracy, and are consistent.

The values of the rate constant obtained by the dating of the ancient Anzick-1 sample and of contemporary genealogies are in agreement. These are the most accurate calibrations.

The difference between the results of the calibrations between Anzick-1 and UIM is only 5%. However, we should note two things about the sequencing UIM calibrations:

1) Radiocarbon dating estimates the age of UIM as approximately 45,000, which is quite close to the reliability limit of radiocarbon dating (50,000 years).

Calibrating UIM with the Anzick-1 sample (using the rate $0.819 \cdot 10^{-9}$) we estimate Ust-Ishim Man's age to be 42,800 years (95% CI: 49,800-37,500). Fu et al. (2014) estimate the age of UIM at between 46,880-43,210 years, with a 95.4% confidence level. Although our value (42,800 years) falls slightly outside the radiocarbon calibration, the lower estimate falls within measurement accuracy.

2) We cannot be sure of the average interval of generations for men who lived thousands of years ago.

The rate constant of base-substitutions per calendar year (μ_a) is associated with a constant per one generation (μ_g) using the ratio

$$\mu_a = \frac{\mu_g}{\bar{t}},$$

where $\bar{t}$ is average interval of one generation (father's average age).

Kong et al. (2012) made an analysis of the complete genomes of 78 parent-offspring trios (i.e. father-mother-offspring). The study strictly confirmed the fact that the number of mutations per meiosis increases with the age of the father. For the purposes of our study, we represented the increase of base-substitutions in a father's gametes versus his age t as a power-law

$$\mu_g \sim t^{\gamma}.$$

According to genomes of five parent-offspring trios (Kong et al., 2012) $\gamma \approx 1.02$.

If we assume that the number of mutations in the Y-chromosome of the son increases in proportion to the age of the father (γ = 1), the average rate of SNP mutations based on a calendar year will not depend on the average interval of one male generation.

In general, the mutation rate per year will be described by a ratio:

$$\mu_a \sim \bar{t}^{\gamma-1}. \tag{4}$$

This ratio is useful for understanding the relationship between the change in the average age of a father and the corresponding change in the mutation rate for one year. Suppose we observe a decrease in the rate constant of mutations in one calendar year during the transition from one calibration to another. When $\gamma > 1$, this reduction corresponds to the decrease in the father’s average age. If $\gamma < 1$, the decrease in the rate of mutations indicates, on the contrary, an increase in the father's generation interval.

Our understanding of the lives of humans during the Paleolithic (12,600-45,000 years BP) makes us think that the average interval of one male generation was less than 31 or 32 years. The average mutation rate calibrations of Anzick-1 and Ust-Ishim do not show any significant difference. This suggests a linear relation of the mutation rate per generation μ_g with the age of the father. In any event, the mutation rate per one calendar year slightly depends on the male generation interval.

In general, our results are consistent with recent data available, including Poznik et al. (2013) $0.82 \cdot 10^{-9}$, Fu et al. (2014) $0.76 \cdot 10^{-9}$. However, our rates are more accurate.

According to Poznik et al (2013), their method of mutation rate evaluation depends on the age of separation between the lines represented by Maya samples HGDP00856 and HGDP00877. HGDP00856 enters the Q-Y772 branch and HGDP00877 enters the Q-Z780 branch. Poznik et al. supposed this time to be 15,000 years BP (the first settlement of America). In our work, we don't use an arbitrary parameter to estimate the date at which these two genetic lines split, instead, we base our calculations using the Anzick-sample. Replacement of the age parameter (15,621 instead of 15,000) reduces the mutation rate in Poznik et al. (2013) to $0.79 \cdot 10^{-9}$.

The base of mutation rate evaluation in Fu et al. (2014) paper is $45000 \cdot 1.86 = 83{,}700$ Mbp·year. Our evaluation base, using the Anzick-1 sample, is much larger: $15621 \cdot 7.7 = 120{,}000$ Mbp·year. A 7.7 Mbp length is calculated by averaging over 12 modern samples bed files (see Table 2). The quality of the data in the Anzick-1 sample is higher than the quality of the data collected from the Ust-Ishim sample.

Kuroki et al. (2006) estimates the mutation rate $(1.5 \cdot 10^{-9})$, since the ancestors of humans and chimpanzees separated. This estimate is 80% higher than our.

The pioneering work of Xue et al. (2009) established the rate $1.0 \cdot 10^{-9}$ per year per bp which contributed greatly to the development of genealogical research on the Y-chromosome. Our research, however, has shown that the calibration Xue et al. obtained in four derived alleles is outdated.

The rate $0.617 \cdot 10^{-9}$ from the work of Mendez et al. (2013) seems to us to be underestimated. The criticism of Mendez et al.'s mutation rate conversion method from data on autosomal chromosomes is contained in Sayres (2013).

Overall, the results of our research on the rate constant of base-substitutions are consistent with the conclusions of Wang, Gilbert, Jin, Li (2014) in their paper, "Evaluating the Y chromosomal timescale in human demographic and lineage dating".

The probable time America was first settled is revealed by the structure of the branches of Q-M1107 (see Figure 1, above). Currently, two branches are known to extend from M1107: branch Z780 and branch M930. Z780 occurs exclusively in the male lines of Native Americans (Indians). The L804 sub-branch of M930 is classed as European (the members of L804 seem to have remained in Eurasia) while those with the Y772 sister-mutation appear to have migrated to America. There are only four mutations at the same level as M930 in combBED area. There are five samples of sub-branch Y772 and four samples of sub-branch L804. The age separation of M930 is estimated at about 14,800 years. We believe this is closest to the time of the settlement of America. Our estimate agrees with the results of the those who study the settlement of America specifically: they assert that humans appeared in the America no later than 14 thousand years ago (Barnosky et al., 2014). We note that the first male immigrants to America do not belong to only one haplogroup: they belong not only to M930 but also to Z780 like the ancestors of Anzick-1.

Hallast et al. (2015) published research on two samples of the aboriginal population of Australia. Mutations F3393 and K35 (measuring 3.7 Mbp) indicated that the samples belong to haplogroup C1 (see YFull tree at http://www.yfull.com/tree/C/). We found that 129 SNPs (of Hallast et al., 2015) in the Australian samples are distinct from the C1a and C1b branches in the Y Full data base. To determine the date of the first settlement of Australian Aborigines, we sought to calibrate the time of the separation of haplogroup C1 from its subclades C1a and C1b. Using four modern samples C-V20 (see Table 4 above) we arrived at 49,200 years (95% CI: 43,900-54,600). This result is in agreement with the estimates archaeologists O'Connell and Allen (2004) and Hiscock (2013) give for the peopling of Sahul: approximately 50 thousand years ago.

On the basis of probability theory we obtained from formula (2) an estimate of the root-mean-square (r.m.s.) error:

$$\frac{\sigma(T)}{T} = \sqrt{\left(\frac{\sigma(N)}{N}\right)^2 + \left(\frac{\sigma(\mu)}{\mu}\right)^2} \qquad (5)$$

With the large number of independent measurements, the average mutation rate can be evaluated with high relative accuracy:

$$\frac{\sigma(\mu)}{\mu} \ll \frac{\sigma(N)}{N} \qquad (6)$$

In this case, the age estimation accuracy would be determined only by the number of mutations:

$$\frac{\sigma(T)}{T} = \frac{\sigma(N)}{N} = \frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}} \qquad (7)$$

Equation (7) allows us to calculate the theoretical limit of accuracy of the SNP branch age estimation. Table 5 shows the relative confidence intervals for the 95% probability (i.e. 1.96σ/T) for the next parameters: $\mu = 0.82 \cdot 10^{-9} (year \cdot bp)^{-1}$, $B = 8 \cdot 10^6$ bp.

We could double the accuracy of the limits specified in Table 5 if we selected actual mutations among ampliconic segments (+10 Mbp added to the combBED area), and take into account second SNP branches (+N substitutions).

Table 5.

The theoretical limit of accuracy of the age estimation versus the SNP branch age.

Branch age, years	**1.96σ/T**
1000	76%
2000	54%
5000	34%
10000	24%

Materials and methods

In order to calibrate the average rate of mutations, we used private Y-chromosome NGS data from two commercial laboratories, FTDNA and Full Genome Corporation. The private samples were provided by the YFull team in compliance with the confidentiality requirements for personal data. All the persons had given individual permits for the use of their data for research.

The data on the ancient Y-chromosome samples were taken from scientific articles: Anzick-1 from Rasmussen et al. (2014), Ust-Ishim Man from Fu et al. (2014), and K14 from Seguin-Orlando et al. (2014).

We developed a selection method which effectively excluded from consideration false options with derived alleles ("false positives"). We used the following filtration criteria:

1. "Reg" criterion. There are derived variants (i.e. alleles different from the reference sequence) revealed in the BAM files. The nucleotide sequences under investigation had a total length between 13-15 Mbp for Big Y, and about 23 Mbp for FGC. Single base read coverage varied from 1X to 8000X. The average coverage of commercial samples is about 60X. From this set of variants, we selected only those coordinates that fell into the combBED regions. As it was mentioned above, the combBED area was designed by the authors to select X-degenerate segments. The combBED area borders were formed by mutual overlapping BED file taken from the work of Poznik et al. (2013) (total length of 10.45 Mbp) and by the generalized Big Y BED file (11.38 Mbp long), published in the Big Y White Paper (2014). The result was 857 continuous segments of the Y-chromosome with a total length of 8,473,821 bp. The coordinates of the beginning and the end of these regions are contained in Table 1 of Supplement.

2. "Indel" criterion. We excluded insertions and deletions (indels), as well as multiple nucleotide polymorphism (more than one base position in derived alleles, MNP) variants.

3. "Locs" criterion. We excluded variants which were detected in more than five different localizations. (Note: "localization" is defined as a group of samples from the YFull database [2,900 samples at February, 2015] belonging to the same subclade and having derived allele nomination that have been studied). In some cases, the same derived variants were revealed in samples from different subclades or haplogroups.

One of the reasons consists of the fact that standard reference sequence is based on haplogroup R1b data and also to a lesser extent on haplogroup G data. Thus, some variants in some haplogroups are ancestral allele, not derived. Another reason is mapping errors. We found limit of five localization empirically. This criterion is soft but effective.

4. "Reads" criterion. We excluded from consideration any one or two read variants.

5. "Qual" criterion. We excluded variants with a read quality less than 90%. Quality is defined as weighted average of the quality index where correct values are taken with the positive and error values, with the negative.

6. "Post mortem" criterion. It's applied only to the ancient samples. Postmortem damages of DNA, lead to the replacement of these base pairs: C→T and G→A (Briggs et al., 2007) were excluded.

7. "Single SNP" criterion. We excluded variants with Double Nucleotide Polymorphisms (DNP). Our program interpreted DNP as a base-substitution in two adjacent positions and therefore were not excluded by our Indel criterion. This secondary criterion allowed us to reject both options.

8. "Trash" criterion. We excluded suspicious variants which have alignment error or reading error. In general, these are variants in palindromic segments and segments with repetitive copies at other Y-chromosome segments.

Variants that pass our criteria are actually base-substitutions. Criteria 2-8 collectively eliminated up to one-third of entered variants, an average of 20%.

The individual combBED area includes all nucleotide positions with three or more reads. The modern samples selected for study ranged from 7.2 to 8.45 Mbp in length (depending on the quality of the sequencing and reflected in the BAM file [see the tables 2 and 3]). The length of the combBED area in the ancient samples varied: Anzick-1 - 6.355 Mbp, Ust-Ishim Man - 8.463 Mbp, K14 - 1.381 Mbp.

There is no single formula for calculating the mutation rate because of the difference in the calibration methods. In all calculations ratio (1) $N_{SNP} = \mu_{SNP}TB$ was used.

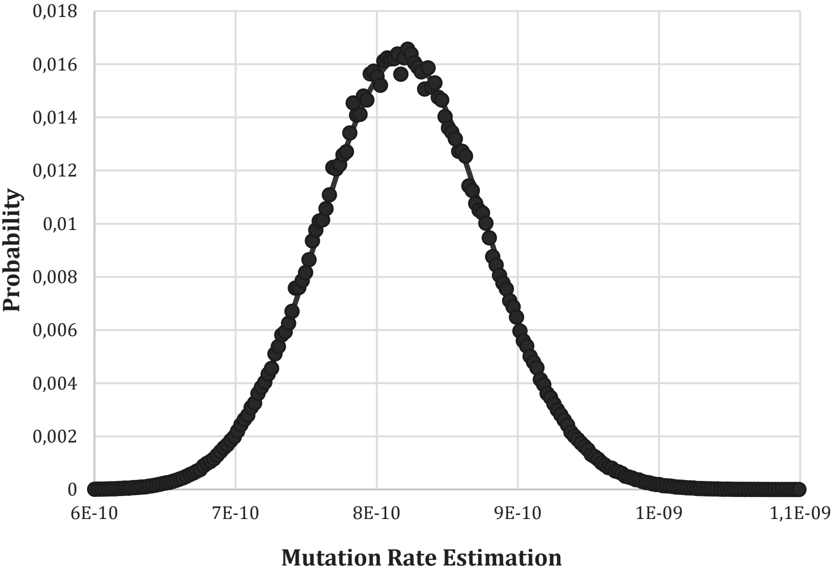


Figure 3. Distribution of the mutation rate estimate based on the ancient Anzick-1 sample dating. The full brown line represents a normal distribution.

An evaluation of 95% confidence intervals was performed in each case individually based on the properties of the Poisson distribution. Figure 3 shows the distribution of the mutation rate estimate at a two-step calibration based on Anzick-1 dating. The distribution was obtained by computer simulation of $5 \cdot 10^6$ random events. It is clear that the curve is very close to a normal distribution.

The variance calculation method for simple one-step calibration is given in Poznik et al. (2013).

Conclusion

Our method of base-substitution variants filtration allowed us effectively select actual mutations and exclude false positives in individual samples.

Using four independent calibrations and ranking them in order of validity and reliability yielded independent but similar rates constant for SNP mutations ($0.82 \cdot 10^{-9}$ per year per bp, 95% CI: $(0.70 - 0.94) \cdot 10^{-9}$).

Our analysis of the Big Y and FGC data collected in the YFull database allowed us to fine tune the probable date of the arrival of humans in the Western Hemisphere (14,800 b.p.), and the arrival of humans in Australia (49,200 b.p.).

We express gratitude to Roman Sychev for the help and discussion.

We thank Judith Remy Leder for help with translation and discussion.

We thank Olga Vassilieva for help with translation.

Bibliography

1. Barnosky A. et al. (2014) Prelude to the Anthropocene: Two new North American Land Mammal Ages (NALMAs). The Anthropocene Review, 1–18. DOI: 10.1177/2053019614547433.

2. Big Y White Paper (2014) ttps://www.familytreedna.com/learn/y-dna-testing/big-y/white-paper/ https://www.familytreedna.com/documents/bigy_targets.txt

3. Briggs A. et al. (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA, 104: 14616–14621.

4. Fenner J.N. (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am.J.Phys.Anthropol. 128(2): 415-423.

5. Francalacci P. et al. (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341: 565-569.

6. Fu Q. et al. (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. Nature, 514: 445-449.

7. Hallast et al (2015) The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol, 32, no. 3: 661-673.

8. Hiscock P. (2013) Occupying New Lands: Global Migrations and Cultural Diversification with Particular Reference to Australia. In Kelly E Graf, Caroline V Ketron, Michael R Waters (Eds.), Paleoamerican Odyssey, (pp. 3-11). Texas: Center for the Study of the First Americans.

9. Kong A. et al. (2012) Rate of de novo mutations and the importance of father’s age to disease risk. Nature, 488 (7412): 471-475.

10. Kuroki Y. et al. (2006) Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. Nature Genetics 38: 158 – 167.

11. Mendez F. et al. (2013) An african american paternal lineage adds an extremely ancient root to the human y chromosome phylogenetic tree. Am. J. Hum. Genet. 92: 454-459.

12. O'Connell J. and Allen J. (2004) Dating the colonization of Sahul (Pleistocene Australia–New Guinea): a review of recent research. Journal of Archaeological Science 31: 835–853.

13. Poznik D. et al. (2013) Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females. Science, 341: 562-565.

14. Rasmussen M. et al. (2014) The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature, 506: 225-229.

15. Sayres M. (2013) Timing of ancient human Y lineage depends on the mutation rate: A comment on Mendez et al. arXiv preprint: arXiv:1304.6098.

16. Seguin-Orlando A. et al. (2014) Genomic structure in Europeans dating back at least 36,200 years.
http://www.sciencemag.org/content/early/recent/ 6 November 2014/ Page 1/ 10.1126/science.aaa0114.

17. Skaletsky H. et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature, 423(6942): 825-837.

18. Wang C.-C., Gilbert T., Jin L., Li H. (2014) Evaluating the Y chromosomal timescale in human demographic and lineage dating. Investigative Genetics, 5:12.
http://www.investigativegenetics.com/content/5/1/12.

19. Wei W. et al. (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res., 23(2): 388-395.

20. Xue Y. et al. (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr. Biol. 19: 1453-1457.

21. Yan S. et al. (2014) Y Chromosomes of 40% Chinese Are Descendants of Three Neolithic Super-grandfathers. PLoS ONE 9(8): e105691. doi:10.1371/journal.pone.0105691.

Supplement.

Table 1.

List of 857 Y-chromosome start and end positions (hg 19 coordinates) of target sequences (“combBED area”).

The start position is not included in the target sequence, the end position is included in the target sequence.

	Start position	End position
1	2655000	2669950
2	2680900	2683020
3	2683240	2684220
4	2684440	2688930
5	2689390	2697630
6	2699410	2708370
7	2708470	2744910
8	2745730	2768038
9	2773800	2775090
10	2775160	2786260
11	2787070	2788503
12	2792920	2856720
13	2858890	2865900
14	2866444	2895330
15	2900860	2913000
16	6619420	6621660
17	6626650	6637410
18	6642670	6645650
19	6646720	6648240
20	6648460	6654290
21	6654780	6663699
22	6666100	6682350
23	6688450	6690090
24	6694660	6709320
301	14550565	14552625
302	14553005	14554575
303	14554965	14556545
304	14558425	14559405
305	14560525	14561505
306	14561635	14563015
307	14566345	14567325
308	14569565	14574215
309	14575685	14584745
310	14585995	14589065
311	14589631	14605385
312	14607965	14612985
313	14620755	14623105
314	14623338	14624786
315	14625805	14630585
316	14635455	14648875
317	14649415	14660649
318	14661328	14668038
319	14691439	14695028
320	14695728	14700888
321	14701388	14703908
322	14704168	14712048
323	14712768	14714948
324	14720208	14724708
601	18925326	18928386
602	18928656	18930246
603	18930486	18932006
604	18932496	18967146
605	18971806	18974736
606	18976366	18987126
607	18990886	19000276
608	19000986	19012836
609	19020046	19025106
610	19031226	19040646
611	19044366	19055712
612	19056046	19062366
613	19065886	19071226
614	19073190	19075336
615	19075446	19081416
616	19085906	19097296
617	19102526	19137766
618	19138606	19145846
619	19146376	19155236
620	19155886	19163999
621	19188000	19210016
622	19211430	19213536
623	19213666	19235416
624	19235446	19244866

25	6714900	6720570
26	6731140	6747310
27	6747820	6748950
28	6752380	6756090
29	6756940	6765700
30	6765780	6770070
31	6773890	6790340
32	6791590	6798940
33	6799060	6807020
34	6807070	6819200
35	6822371	6825510
36	6825940	6826920
37	6830520	6832150
38	6833100	6881370
39	6882280	6892160
40	6892325	6901280
41	6903490	6916230
42	6917610	6922680
43	6922900	6925890
44	6930914	6976140
45	6977380	6983940
46	6984550	6998220
47	7008550	7015440
48	7034000	7039350
49	7045470	7048310
50	7048410	7053960
51	7054090	7058460
52	7063690	7080930
53	7081600	7085500
54	7086630	7088090
55	7088270	7090830

325	14726988	14736348
326	14743708	14755791
327	14761211	14764571
328	14764971	14785111
329	14785371	14788111
330	14788631	14792961
331	14798441	14815451
332	14815751	14823811
333	14823911	14908514
334	14909394	14959764
335	14965116	14990530
336	14991036	14992636
337	14992740	15011876
338	15012546	15044416
339	15045356	15066106
340	15066236	15073366
341	15073506	15080836
342	15084476	15098256
343	15103476	15109436
344	15109616	15113756
345	15113886	15115016
346	15115446	15116825
347	15122106	15126466
348	15126996	15141206
349	15143436	15146636
350	15151086	15168206
351	15171349	15177686
352	15180126	15183916
353	15189426	15201666
354	15202316	15210976
355	15210996	15212756

625	19245916	19248366
626	19249996	19251626
627	19254286	19297066
628	19304486	19310346
629	19310816	19324396
630	19331486	19340266
631	19340276	19351706
632	19356176	19375935
633	19375986	19381999
634	19391000	19415546
635	19416576	19417556
636	19418076	19423986
637	19429306	19433456
638	19435536	19437136
639	19438236	19459406
640	19460556	19461896
641	19464756	19479336
642	19479686	19483000
643	19493000	19510016
644	19510036	19517496
645	19517667	19539936
646	19542106	19544266
647	19544346	19551000
648	21049052	21054562
649	21061442	21072462
650	21073142	21077202
651	21078782	21095972
652	21097042	21107022
653	21109782	21118722
654	21125482	21139562
655	21139802	21150000