Robert L. Dorit, Hiroshi Akashi und Walter Gilbert berichten in

> Absence of Polymorphism at the ZFY Locus on the Human Y Chromosome, *Science* 268, 1183–1185 (1995)

Ergebnisse einer genetischen Studie:

- Weltweite[1] Stichprobe von 38 Männern (*homo sapiens*)
- Ein 729 Basenpaare langes, nicht-kodierendes Stück des Y-Chromosoms (das 3. Intron des ZFY-Gens) wurde für jede Stichprobe sequenziert

[1] Loc. cit., S. 1184: "Human DNA samples were obtained from male volunteers who donated hair follicle samples or from cell lines provided by L. L. Cavalli-Sforza and K. K. Kidd. Geographic origins were determined by interview. Whenever possible, geographic origins of parents and grandparents were also ascertained. The samples are grouped by continent of origin, and the number of individuals is given in parentheses. Africa: Nigeria* (1), Ivory Coast (1), Tanzania (1), Southern Africa (2), Algeria (1), Central African Republic* (2), African American (2); Americas: Mexico (2), Guatemala (1), Peru* (1), Argentina (1), Native American (2); Asia: China* (2), Korea (1), Japan* (2), Taiwan (2), Indonesia (1), India (1); Europe/Middle East: Ireland* (1), Belgium (1), Italy* (1), Spain (1), Russia* (2), Poland* (1), Saudi Arabia* (1), Turkey (1); South Pacific: Melanesia (1), New Guinea* (1), Australia* (1). (*) Indicates samples where the 3'-most zinc-finger exon was also sequenced."

Robert L. Dorit, Hiroshi Akashi und Walter Gilbert berichten in

> Absence of Polymorphism at the ZFY Locus on the Human Y Chromosome, *Science* 268, 1183–1185 (1995)

Ergebnisse einer genetischen Studie:

- Weltweite[1] Stichprobe von 38 Männern (*homo sapiens*)
- Ein 729 Basenpaare langes, nicht-kodierendes Stück des Y-Chromosoms (das 3. Intron des ZFY-Gens) wurde für jede Stichprobe sequenziert
- Es wurden keinerlei Mutationen gefunden: Alle 38 Stichproben identisch

[1] Loc. cit., S. 1184: "Human DNA samples were obtained from male volunteers who donated hair follicle samples or from cell lines provided by L. L. Cavalli-Sforza and K. K. Kidd. Geographic origins were determined by interview. Whenever possible, geographic origins of parents and grandparents were also ascertained. The samples are grouped by continent of origin, and the number of individuals is given in parentheses. Africa: Nigeria* (1), Ivory Coast (1), Tanzania (1), Southern Africa (2), Algeria (1), Central African Republic* (2), African American (2); Americas: Mexico (2), Guatemala (1), Peru* (1), Argentina (1), Native American (2); Asia: China* (2), Korea (1), Japan* (2), Taiwan (2), Indonesia (1), India (1); Europe/Middle East: Ireland* (1), Belgium (1), Italy* (1), Spain (1), Russia* (2), Poland* (1), Saudi Arabia* (1), Turkey (1); South Pacific: Melanesia (1), New Guinea* (1), Australia* (1). (*) Indicates samples where the 3'-most zinc-finger exon was also sequenced."

R.L. Dorit, H. Akashi und W. Gilbert, *Science* 268:

- 38 Y-Chromosomen aus weltweiter Stichprobe,
  jeweils 729 Bp langes Stück des ZFY-Gens sequenziert,
  keine Mutationen in der Stichprobe sichtbar

R.L. Dorit, H. Akashi und W. Gilbert, *Science* 268:

- 38 Y-Chromosomen aus weltweiter Stichprobe,
  jeweils 729 Bp langes Stück des ZFY-Gens sequenziert,
  keine Mutationen in der Stichprobe sichtbar
- Inter-spezies-Vergleich mit Schimpanse, Gorilla, Orang-Utan (und Pavian als "outgroup") zeigt, dass am betrachteten Lokus Mutationen vorkommen können
- Molekulare Uhr-Annahme und auf Fossilien beruhende Annahmen über die Zeit seit der Aufspaltung von der Vorfahren von Mensch und Schimpanse bzw. Orang-Utan ergeben geschätzte Rate von (fixierten) Mutationen

$$1{,}35 \times 10^{-3}\ \mathit{Mutationen\ pro\ Basenpaar\ pro\ Million\ Jahre}$$

R.L. Dorit, H. Akashi und W. Gilbert, *Science* 268:

- 38 Y-Chromosomen aus weltweiter Stichprobe,
  jeweils 729 Bp langes Stück des ZFY-Gens sequenziert,
  keine Mutationen in der Stichprobe sichtbar
- Inter-spezies-Vergleich mit Schimpanse, Gorilla, Orang-Utan (und Pavian als "outgroup") zeigt, dass am betrachteten Lokus Mutationen vorkommen können
- Molekulare Uhr-Annahme und auf Fossilien beruhende Annahmen über die Zeit seit der Aufspaltung von der Vorfahren von Mensch und Schimpanse bzw. Orang-Utan ergeben geschätzte Rate von (fixierten) Mutationen

$$1{,}35 \times 10^{-3}\ \textit{Mutationen pro Basenpaar pro Million Jahre}$$

Was können wir angesichts dieser Beobachtungen über die Zeit bis zum jüngsten gemeinsamen Vorfahren der gezogenen 38 Y-Chromosomen (und damit implizit auch über den jgV aller heute lebenden Männer) sagen?

Sei t_{MRCA} die Zeit (in Jahren) bis zum jüngsten gemeinsamen Vorfahren der 38 gezogenen Männer.

Wir verwenden den Koaleszenten als Modell der Genealogie.

1. A-priori-Verteilung Ohne Berücksichtigung der Beobachtungen würden wir annehmen, dass

$$T_{\text{MRCA}} \stackrel{d}{=} T_{38} + T_{37} + \cdots + T_2$$

wo T_{MRCA} die Zeit (in Koaleszenten-Zeiteinheiten) bis zum jüngsten gemeinsamen Vorfahren, die T_k unabhängig mit $T_k \sim \text{Exp}\left(\binom{k}{2}\right)$

Sei t_{MRCA} die Zeit (in Jahren) bis zum jüngsten gemeinsamen Vorfahren der 38 gezogenen Männer.

Wir verwenden den Koaleszenten als Modell der Genealogie.

1. A-priori-Verteilung Ohne Berücksichtigung der Beobachtungen würden wir annehmen, dass

$$T_{\text{MRCA}} \stackrel{d}{=} T_{38} + T_{37} + \cdots + T_2$$

wo T_{MRCA} die Zeit (in Koaleszenten-Zeiteinheiten) bis zum jüngsten gemeinsamen Vorfahren, die T_k unabhängig mit $T_k \sim \text{Exp}\left(\binom{k}{2}\right)$

$$1 \text{ Koaleszenten-Zeiteinheit } \widehat{=} \; N_{\text{eff}} \times g \text{ Jahre}$$

mit N_{eff} ... effektive Populationsgröße (für Männer),

g ... Generationslänge (in Jahren)

1. A-priori-Verteilung: $\mathscr{L}(T_{\mathrm{MRCA}}) = \ast_{k=2}^{38} \mathrm{Exp}\left(\binom{k}{2}\right)$

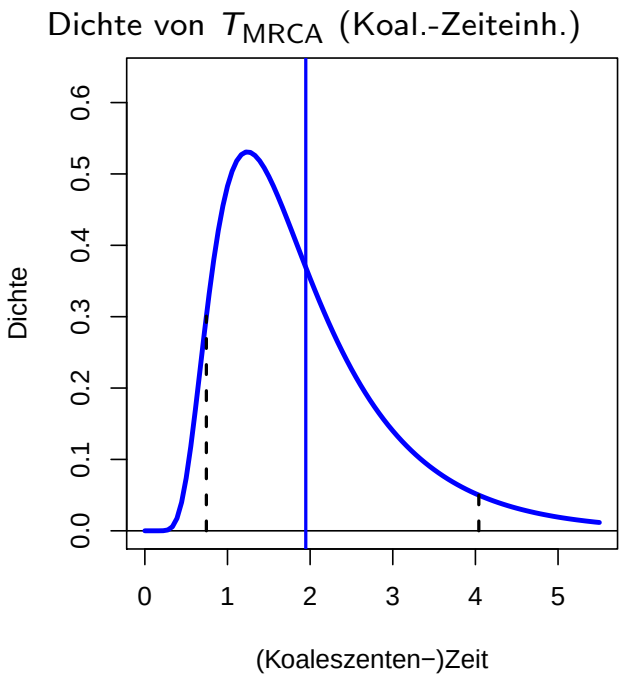
$$\mathbb{E}[T_{\mathrm{MRCA}}] = \sum_{k=2}^{38} \frac{2}{k(k-1)} = 2\left(1 - \tfrac{1}{38}\right) = \frac{37}{19} \doteq 1{,}947,$$

5%-Quantil von T_{MRCA}: $q_{0,05} \doteq 0{,}744$, 95%-Quantil: $q_{0,95} \doteq 4{,}041$.

1. A-priori-Verteilung: $\mathscr{L}(T_{\mathrm{MRCA}}) = \ast_{k=2}^{38} \mathrm{Exp}\left(\binom{k}{2}\right)$

$$\mathbb{E}[T_{\mathrm{MRCA}}] = \sum_{k=2}^{38} \frac{2}{k(k-1)} = 2\left(1 - \tfrac{1}{38}\right) = \frac{37}{19} \doteq 1{,}947,$$

5%-Quantil von T_{MRCA}: $q_{0{,}05} \doteq 0{,}744$, 95%-Quantil: $q_{0{,}95} \doteq 4{,}041$.

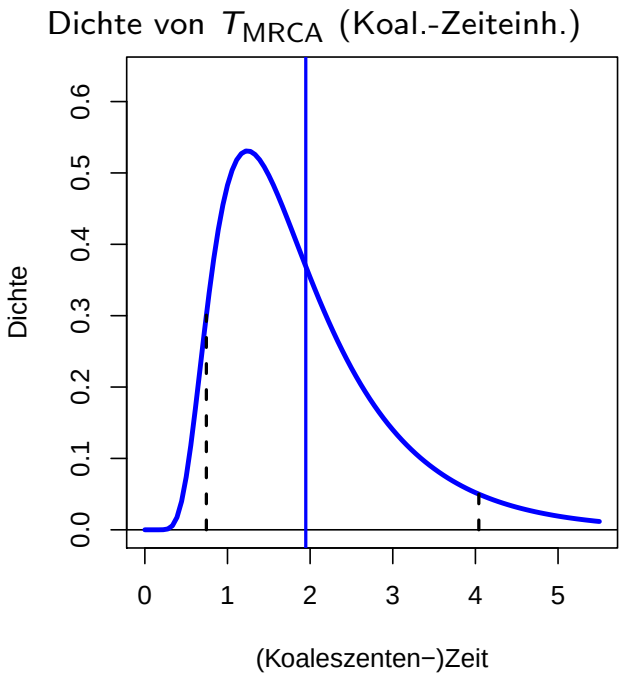


1. A-priori-Verteilung: $\mathscr{L}(T_{\mathrm{MRCA}}) = \ast_{k=2}^{38} \mathrm{Exp}\left(\binom{k}{2}\right)$

$$\mathbb{E}[T_{\mathrm{MRCA}}] = \sum_{k=2}^{38} \frac{2}{k(k-1)} = 2\left(1 - \tfrac{1}{38}\right) = \frac{37}{19} \doteq 1{,}947,$$

5%-Quantil von T_{MRCA}: $q_{0,05} \doteq 0{,}744$, 95%-Quantil: $q_{0,95} \doteq 4{,}041$.

Mit Annahmen $N_{\mathrm{eff}} = 5.000$, $g = 20\mathrm{a}$ übersetzt sich dies zu
MW $\doteq$ 195.000a, $q_{0,05} \doteq$ 74.000a, $q_{0,95} \doteq$ 404.000a

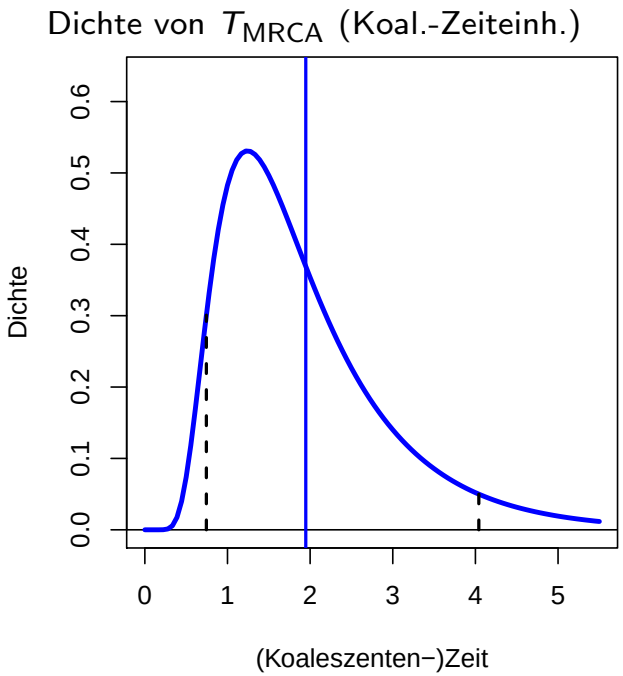


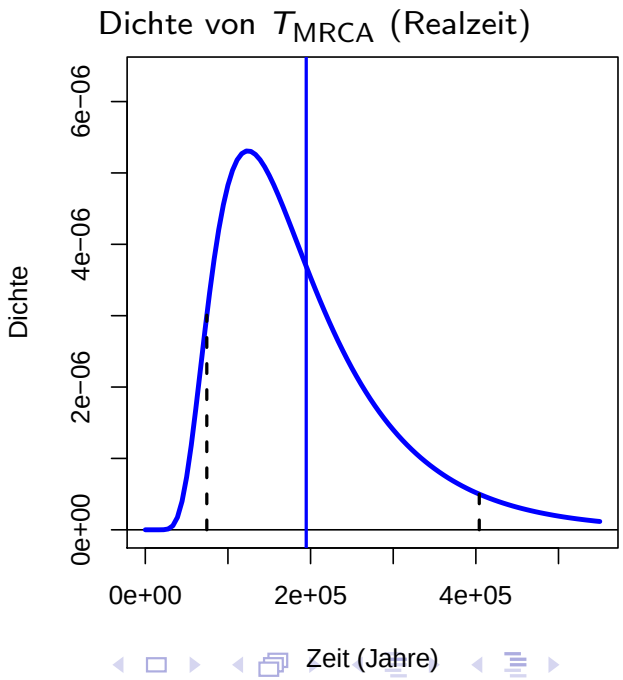
1. A-priori-Verteilung: $\mathscr{L}(T_{\mathrm{MRCA}}) = \ast_{k=2}^{38} \mathrm{Exp}\left(\binom{k}{2}\right)$

$$\mathbb{E}[T_{\mathrm{MRCA}}] = \sum_{k=2}^{38} \frac{2}{k(k-1)} = 2\left(1 - \tfrac{1}{38}\right) = \frac{37}{19} \doteq 1{,}947,$$

5%-Quantil von T_{MRCA}: $q_{0,05} \doteq 0{,}744$, 95%-Quantil: $q_{0,95} \doteq 4{,}041$.

Mit Annahmen $N_{\mathrm{eff}} = 5.000$, $g = 20\mathrm{a}$ übersetzt sich dies zu
MW $\doteq$ 195.000a, $q_{0,05} \doteq$ 74.000a, $q_{0,95} \doteq$ 404.000a





Wir verwenden den Koaleszenten als Modell der Genealogie.

2. A-posteriori-Verteilung $T_k \ldots$ Länge des Zeitintervalls (in Koaleszenten-Zeiteinheiten) währenddessen k Linien in der Genealogie,
$M_k \ldots$ Anzahl Mutationen, die während dieses Intervalls in der Genealogie auftreten

Gegeben $T_k = t$ ist M_k Poisson-verteilt mit Parameter $tk\frac{\theta}{2}$, d.h.

$$\mathbb{P}(M_k = m | T_k = t) = \exp\left(-tk\tfrac{\theta}{2}\right)\frac{\left(tk\frac{\theta}{2}\right)^m}{m!}$$

wobei

$$\theta = 2N_{\text{eff}} \times g \times \mu$$

mit N_{eff} effektive Populationsgröße, g Generationslänge (in Jahren),
μ Mutationsrate der betrachteten Region im Genom (pro Jahr)

(und gegeben $T_2, \ldots, T_n$ sind $M_2, \ldots, M_n$ unabhängig)

Frage: Wie ist $T_{38} + \cdots + T_2$ verteilt, gegeben dass $M_{38} + \cdots + M_2 = 0$?

2. A-posteriori-Verteilung: $T_k \sim \mathrm{Exp}\left(\binom{k}{2}\right)$, $\mathscr{L}(M_k | T_k = t) = \mathrm{Poi}(tk\theta/2)$, dann ist

$$\mathbb{P}(M_k = m) = \frac{k-1}{k-1+\theta}\left(\frac{\theta}{k-1+\theta}\right)^m,$$

d.h. $\mathscr{L}(M_k) = \mathrm{Geom}\left(\frac{k-1}{k-1+\theta}\right)$ und

$$\mathscr{L}(T_k | M_k = 0) = \mathrm{Exp}\left(\frac{k(k-1+\theta)}{2}\right)$$

($\theta = 2N_{\text{eff}} \times g \times \mu$).

Bedingt auf $M_2 = \cdots = M_{38} = 0$ sind $T_2, \ldots, T_{38}$ (weiterhin) unabhängig.

2. A-posteriori-Verteilung: $T_k \sim \mathrm{Exp}\left(\binom{k}{2}\right)$, $\mathscr{L}(M_k|T_k = t) = \mathrm{Poi}(tk\theta/2)$, dann ist

$$\mathbb{P}(M_k = m) = \frac{k-1}{k-1+\theta}\left(\frac{\theta}{k-1+\theta}\right)^m,$$

d.h. $\mathscr{L}(M_k) = \mathrm{Geom}\left(\frac{k-1}{k-1+\theta}\right)$ und

$$\mathscr{L}(T_k|M_k = 0) = \mathrm{Exp}\left(\frac{k(k-1+\theta)}{2}\right)$$

($\theta = 2N_{\text{eff}} \times g \times \mu$).

Bedingt auf $M_2 = \cdots = M_{38} = 0$ sind $T_2, \ldots, T_{38}$ (weiterhin) unabhängig.

Demnach: Verteilung der Zeit bis zum jüngsten gemeinsamen Vorfahren (in Koaleszenten-Zeiteinheiten), bedingt auf $M := M_2 + \cdots + M_{38} = 0$ ist

$$T_{\text{MRCA}}|_{\{M=0\}} \stackrel{d}{=} T'_{38} + T'_{37} + \cdots + T'_2$$

mit T'_k u.a., $T'_k \sim \mathrm{Exp}\left(\frac{k(k-1+\theta)}{2}\right)$.

2. A-posteriori-Verteilung: $\mathscr{L}(T_{\text{MRCA}}|M=0) = \overset{38}{\underset{k=2}{*}} \text{Exp}\left(\frac{k(k-1+\theta)}{2}\right)$ mit $\theta = 2N_{\text{eff}} \times g \times \mu$.

Wir fixieren $g = 20\text{a}$, $\mu = 729 \times 1{,}35 \cdot 10^{-9}\text{a}^{-1} \doteq 0{,}98 \cdot 10^{-6}\text{a}^{-1}$ (aus Dorit et al (1995), diese Werte waren auch in der Literatur unstrittig), so hängt die bedingte Verteilung von T_{MRCA} (und nicht nur ihre „Übersetzung in Realzeit“) vom Parameter N_{eff} ab.

2. A-posteriori-Verteilung: $\mathscr{L}(T_{\text{MRCA}}|M=0) = \overset{38}{\underset{k=2}{*}} \text{Exp}\left(\frac{k(k-1+\theta)}{2}\right)$ mit $\theta = 2N_{\text{eff}} \times g \times \mu$.

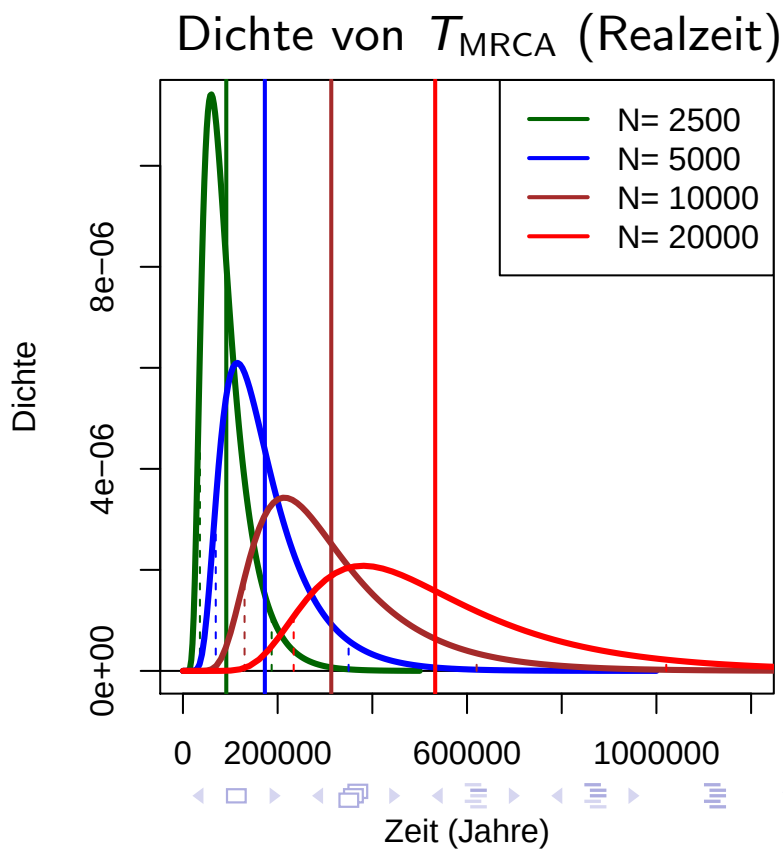
Wir fixieren $g = 20\text{a}$, $\mu = 729 \times 1{,}35 \cdot 10^{-9}\text{a}^{-1} \doteq 0{,}98 \cdot 10^{-6}\text{a}^{-1}$ (aus Dorit et al (1995), diese Werte waren auch in der Literatur unstrittig), so hängt die bedingte Verteilung von T_{MRCA} (und nicht nur ihre „Übersetzung in Realzeit“) vom Parameter N_{eff} ab.

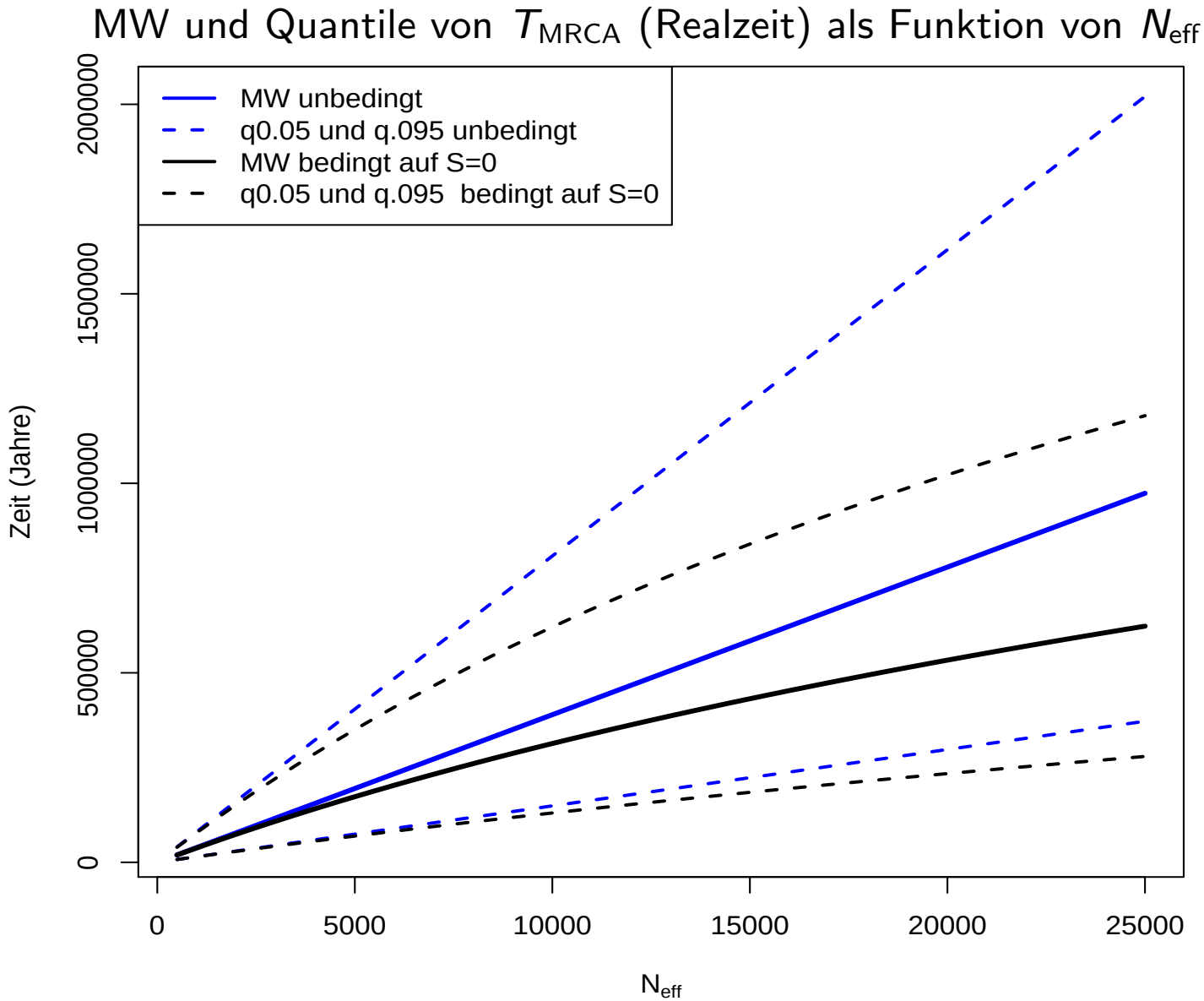
N_{eff}	EW	$q_{0,05}$	$q_{0,95}$
2.500	91.519	35.851	187.369
5.000	173.007	69.263	349.909
10.000	313.234	130.095	620.279
20.000	532.785	233.853	1.020.819

2. A-posteriori-Verteilung: $\mathscr{L}(T_{\text{MRCA}}|M=0) = \overset{38}{\underset{k=2}{*}} \text{Exp}\left(\frac{k(k-1+\theta)}{2}\right)$ mit $\theta = 2N_{\text{eff}} \times g \times \mu$.

Wir fixieren $g = 20\text{a}$, $\mu = 729 \times 1{,}35 \cdot 10^{-9}\text{a}^{-1} \doteq 0{,}98 \cdot 10^{-6}\text{a}^{-1}$ (aus Dorit et al (1995), diese Werte waren auch in der Literatur unstrittig), so hängt die bedingte Verteilung von T_{MRCA} (und nicht nur ihre „Übersetzung in Realzeit“) vom Parameter N_{eff} ab.

N_{eff}	EW	$q_{0,05}$	$q_{0,95}$
2.500	91.519	35.851	187.369
5.000	173.007	69.263	349.909
10.000	313.234	130.095	620.279
20.000	532.785	233.853	1.020.819





Insbesondere: Die Verteilung von T_{MRCA} hängt in nicht-linearer Weise von N_{eff} ab.

Literatur

- Robert L. Dorit, Hiroshi Akashi und Walter Gilbert, Absence of Polymorphism at the ZFY Locus on the Human Y Chromosome, *Science* 268, 1183–1185 (1995)
- Diskussionsbeiträge ("technical comments") dazu in *Science* 272, 1356–1362 (1996) von
  - Y.-X. Fu, W.-H. Li
  - P. Donnelly, S. Tavaré, D.J. Balding, R.C. Griffiths
  - G. Weiss, A. von Haeseler
  - J. Rogers, P.B. Samollow, A.G. Comuzzie
- Kapitel 8.1 in J. Wakeley, *Coalescent Theory: An Introduction*, Roberts & Company, 2008