Supplementary Note for

Repping *et al*.,

Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection

Supplementary Note 1. Stratification is unlikely to explain association of the gr/gr deletion with spermatogenic failure.

For stratification to account for the observed association, there would have to be two sub-populations marked by:

1. similar Y haplotype distributions,
2. different risks of spermatogenic failure (for reasons unrelated to the gr/gr deletion), and
3. an excess of gr/gr deletions in the at-risk population (in at least one Y haplotype group) for reasons unrelated to spermatogenic failure.

The following analysis indicates that stratification is extremely unlikely to account for the observed association. Consider the raw data gathered both in the preliminary study and in the formal association study. Inspection of the data reveals that:

- In the preliminary study, the gr/gr deletion was found in five Y haplotypes present in both cases and controls (**Supplementary Table 2**).
- In the association study, the gr/gr deletion was found in three Y haplotypes present in both cases and controls (**Supplementary Table 4**).

In all eight groups, the frequency of the gr/gr deletion allele is higher among cases than among controls. If the gr/gr deletion were randomly distributed among cases and controls, the chance of this consistent bias would be $(0.5)^8 = 0.4\%$. Even if the association in one haplotype were spurious—due to stratification—and even if that haplotype were present in both studies, then the likelihood that the observed bias in the remaining haplotypes occurred by chance would only rise to $(0.5)^6 = 1.6\%$. In sum, it is very unlikely that chance would account for such a consistent bias—always in the same direction—in multiple haplotypes in both the preliminary and formal association studies.

This argument is all the more telling when one considers that, as demonstrated in the main text, the gr/gr deletion allele had >13 independent origins in the eight informative haplotypes in **Supplementary Tables 2** and **4**. Finally, one should consider the biological reasonableness of an association between the gr/gr deletion and spermatogenic failure the outset (the prior odds), for it is well established that a 3.5-Mb deletion (*AZFc*) that encompasses the 1.6-Mb gr/gr deletion severely impairs spermatogenesis[1].

Supplementary Note 2. Branch D2b Y chromosomes, the YAP insertion in Japan, and the gr/gr deletion.

Work by Kuroki and colleagues suggested that Japanese Y chromosomes with the polymorphic YAP Alu insertion[2] confer increased risk of spermatogenic failure[3]. Since most Japanese Y chromosomes with the YAP insertion are in branch D2b (ref. 4 and our data), Kuroki and colleague's at-risk haplotype in Japan is roughly equivalent to branch D2b.

Two phenomena could explain the high proportion of gr/gr deletions (12/12) that we observed in branch D2b: (1) all gr/gr-deleted chromosomes in this branch having descended from a single gr/gr-deleted founder, or (2) a very high rate of deletion events. Support for the former possibility comes from a recent report that Yfm1, a microsatellite marker, has reduced copy number in 82 of 82 YAP-inserted Japanese Y chromosomes examined[5]. Our analysis shows that Yfm has copies in *AZFc* amplicons r1, r2, r3, and r4 (**Fig. 1a**, main text). This suggests that gr/gr deletions are universal in branch D2b (82/82, 95% CI 0.96 to 1.0) and that the D2b founder was probably gr/gr-deleted.

Supplementary Note 3. What proportion of Y chromosomes might be gr/gr-deleted in the absence of selection?

As discussed in the text, it seems clear that the last common ancestor of extant human Y chromosomes was not gr/gr deleted. Recent estimates of the number of generations since that last common ancestor range from 2388 (95% CI 1624-5796, ref. 6) to 2758 (95% CI 1818 – 4333) (from ref. 7 and a 33-year generation[8,9]). If the gr/gr deletion had no effect on fitness and new gr/gr deletions arose at a rate of 1 per 4000 male births, after 2388 generations we would expect only $(1 - (1 / 4000))^{2388} = 55\%$ of Y chromosomes to have escaped gr/gr deletion. After more generations, that proportion would be lower. The mutation rate assumption is based on the mutation rate that has been estimated for *AZFc* (b2/b4) deletions[1]. Even if the rate of *de novo* gr/gr deletions were one tenth this estimate, there would still be more gr/gr deletions than we observed.

We use the term "gr/gr deletion" to refer to the particular organization of amplicons shown in the bottom half of **Figure 1c** (main text). The term is derived from the amplicons that demarcate the deletion, and is not intended to refer to any particular homologous recombination event or series of events that might generate this deletion. Thus, in the calculations above, the mutation rate is that at which the gr/gr-deleted organization is generated *de novo* from all other *AZFc* organizations. We base the mutation rate used in the calculation above on the rate at which the b2/b4-deleted organization is generated *de novo* from all other *AZFc* organizations (1/4000 newborn males; ref. 1).

Supplementary Note 4. Exploratory estimate of the effect of the gr/gr deletion on fitness.

To estimate *s*, the reduction in fitness due to the gr/gr deletion, we have $s = \mu/q$, where μ is the rate at which new gr/gr deletions arise by mutation, and q is the proportion of gr/gr deletions in the general population (see derivation below). If $\mu = 1/4000$ new gr/gr deletions per male birth, and $q = 4/178$, then s would be 1%. We expect that genetic drift also significantly influenced the observed q.

Derivation

The male-specific region of the Y chromosome is hemizygous and transmitted clonally from generation to generation. In this case, the derivation for the relationship between the frequency of a deleterious allele, the rate at which the allele is created by mutation, and the coefficient of selection, is similar to but simpler than the well-known deterministic approximations for deleterious autosomal dominant or X-linked recessive alleles (see, for example, refs. 10,11), but to our knowledge has not been published. We have

$$q_{n+1} = (q_n w_2 + (1-q_n)\mu) / ((1-q_n)w_1 + q_n w_2),$$

where q_n is the prevalence of the deleterious allele (for example the gr/gr deletion) in the population in the current generation, q_{n+1} is the prevalence of the deleterious allele in the subsequent generation, $w_1 = 1$ is the coefficient of fitness of the fittest ("wild-type") allele, $w_2 = w_1 - s$, $s > 0$ is the coefficient of fitness of the deleterious allele, and μ is the rate per generation at which new deleterious alleles arise by mutation. At equilibrium we have $q_{n+1} = q_n$. Substituting q for q_{n+1} and q_n, 1 for w_1, and $1 - s$ for w_2 we have

$$q = (q(1 - s) + \mu - q\mu) / (1 - sq).$$

Solving for s yields

$$s = (\mu - q\mu) / (q - q^2) = \mu/q.$$

This equation provides the relationship between fitness, mutation rate, and allele frequency in a deterministic model. For small populations, genetic drift will also influence q.

Supplementary Note 5. Alternative analysis, including gr/gr deletions with b2/b4 duplication or 47,XYY karyotype.

The gr/gr deletion removes 9 of the 21 genes and transcription units located in the *AZFc* region, but leaves at least one copy of each gene family intact (Table 1, main text). In our studies, the hypothesis investigated is that the gr/gr deletion increases risk of spermatogenic failure because it reduces the dosage of *AZFc* genes. Therefore, in our association analysis, we excluded one man (AMC0105) with both a gr/gr deletion and a b2/b4 duplication (**Fig. 3**, main text).

Alternative hypotheses for the phenotypic effect of gr/gr deletions include functional specialization among *AZFc* genes or regulatory effects due to the proximity of remaining *AZFc* genes to Y-chromosome long-arm heterochromatin. If we define the genotype of interest to include men with the gr/gr deletion and duplication (either by the b2/b4 duplication or because of a 47,XYY karyotype) we still observe a significant association with spermatogenic failure. The resulting table, analogous to Table 2 in the main text is:

	gr/gr deletion (incl. b2/b4 dupl. and 47,XYY)	No gr/gr deletion	Total
Spermatogenic failure	10	237	247
Normal spermatogenesis	0	148	148
Total	10	385	395

$P < 0.009$ [a]

[a] By Fisher's exact test, one-sided.

References for Supplementary Note

1. Kuroda-Kawaguchi, T. *et al*. The *AZFc* region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279-286 (2001).
2. Hammer, M.F. A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* **11**, 749-761 (1994).
3. Kuroki, Y. *et al*. Spermatogenic ability is different among males in different Y chromosome lineage. *J. Hum. Genet.* **44**, 289-292 (1999).
4. Underhill, P.A. *et al*. Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358-361 (2000).
5. Ewis, A.A., Lee, J., Shinka, T. & Nakahori, Y. Microdeletions of a Y-specific marker, Yfm1, and implications for a role in spermatogenesis. *J. Hum. Genet.* **47**, 257-261 (2002).
6. Thomson, R., Pritchard, J., Shen, P., Oefner, P. & Feldman, M. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6927-6929 (2001).
7. Tang, H., Siegmund, D.O., Shen, P., Oefner, P.J. & Feldman, M.W. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* **161**, 447-159 (2002).
8. Tremblay, M. & Vezina, H. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.* **66**, 651-658 (2000).
9. Helgason, A., Hrafnkelsson, B., Gulcher, J., Ward, R. & Stefansson, K. A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am. J. Hum. Genet.* **72**, 1370-1388 (2003).
10. Crow, J.F. & Kimura, M. *An Introduction to Population Genetics Theory*, 591 (Harper & Row, New York, Evanston, and London, 1970).
11. Cavalli-Sforza, L.L. & Bodmer, W.F. *The Genetics of Human Populations*, 965 (W H Freeman & Co, 1971).