Supporting Information

Neumann et al. 10.1073/pnas.1003634107

Fig. S1. Gel-electrophoretic enrichment of γ-globin duplication and deletion molecules. Typical distributions are shown for the fractionation of EcoRV–NdeI-digested sperm DNA from man 1, with the size range of DNA fragments present in each fraction shown by blue bars and the predicted sizes of mutant and progenitor molecules indicated by horizontal dashed lines. Size fractions potentially containing deletions were identified using a control 6.86-kb EcoRV-NdeI genomic fragment matched to the size of deletions (6.94 kb). This fragment is located on chromosome 11 5,576,578 to 5,583,437 (assembly GRCh37, release 56) and was amplified using primers CL2.5F (TCC AGA TCA GAG ATG CTC CC) and CL6.3R (TCA AGA GAA TGC CTG ACT CC) to generate a 3.84-kb amplicon. Likewise, fractions containing 16.78-kb duplications were located using a size-matched 17.06-kb genomic EcoRV-NdeI DNA fragment located on chromosome 1 9,359,406 to 9,376,466 and amplified with primers CG12.0F (CAG ATC CTT TCT TCT CCC GC) and CG14.8R (AGC AGG AGA TTC GAG ATC CC) to produce a 2.81-kb amplicon. Deletion and duplication molecules were distributed across fractions as expected from these genomic control restriction fragments and were not, therefore, PCR artifacts generated from progenitor molecules. The biphasic distributions, seen in particular for deletions and deletion control molecules, were because of the difficulty of achieving complete digestion with NdeI; in each case the larger component corresponded to the DNA fragment length predicted from incomplete NdeI digestion. This partial digestion did not compromise the purification of mutants, given the presence of a nearby EcoRV restriction site that was fully cleaved in these experiments (Fig. 1*A*).

Fig. S1. (EPS)

Fig. S2. Structures of γ-globin gene rearrangements detected in men 1 to 3, shown as in Fig. 2. The haplotype of origin of each rearrangement is indicated, except for man 3, who is homozygous for haplotype D. The length of each exchange interval and the Poisson-corrected number of rearrangements of each type are shown at right. Recurrent base-changes are shown in blue, with the number of recurrences indicated. The two duplications in man 2 marked with an asterisk were detected in the same PCR and might instead represent a single unrepaired heteroduplex molecule (Fig. 4*B*).

Fig. S2. (PDF)

Fig. S3. DNA sequence traces of recurrent base mutations detected in γ-globin gene rearrangements. C, sequence showing a complete base switch; M, sequence showing this switch as a mixed site; P, nonmutant sequence; Details of the mutations are given in Fig. 4*B* and Table S2. Mutation 5 was detected in a PCR containing either two different duplications or an unrepaired heteroduplex.

Fig. S3. (PDF)

Fig. S4. Spectrum of base changes seen in γ-globin DNA molecules. (*A*) Comparison of the types of base change seen at mixed sites, showing both the correct and an incorrect base in sequence traces, with those showing completely switched bases (excluding recurrences). (*B*) Frequency of base changes at different admixture levels in γ-globin DNA molecules, binned into 10% intervals. Switched bases are shown in black and do not appear to belong to the mixed site distribution. Mixed sites were identified by aligning batches of sequence traces using AutoAssembler and checking for trace positions showing a mixed base, as evidenced by reduction in peak height of the correct base and the compensating appearance of an alternative base. The degree of admixture was estimated from peak heights and admixtures of <20% were rejected, a threshold chosen to ensure that all mixed sites would be captured at any trace position. Admixture estimates obtained separately from both DNA strands agreed on average within 5%. Mixed sites were surveyed in 478 mutant and progenitor sequences (see below) totalling 2.0 Mb in length; their incidence in mutants and progenitors was indistinguishable. Mixed sites showed a very different substitution spectrum from that seen at switched sites ($P < 0.0001$). The spectrum of changes seen at 19 different switched sites in blood was not significantly different from that seen at 64 different switches detected in sperm rearrangements and progenitor molecules ($P = 0.33$). Control progenitor molecules remaining in sperm DNA fractions from man 2 that had been enriched for deletions or duplications were amplified using primers G18.3F (AAT AGT AAG CCT GAG CCC TTC TG) and G23.3R (TAC AAG GCT AGA GTA AAG CAT G), then reamplified with the nested primers G18.4F (CGT ACT TTA GGC TTG TAA TGT G) and G23.2R (GTT AGA GAG AAG GGC GCA G) to produce a 4.76 kb amplicon spanning the $^{A}\gamma$ -globin gene. A total of 138 progenitor PCRs were fully sequenced from this haplotype C/D heterozygote, yielding 69 PCRs containing only haplotype C, 62 containing D and 7 containing a mixture of C plus D. Approximately 124 of these PCRs were therefore derived from single input molecules, with haplotypes C and D being equally efficiently amplified. None of these progenitor molecules showed any rearrangement, and six complete base switches were detected in 0.53 Mb DNA sequence derived from single molecules.

Fig. S4. (EPS)

Table S1. Common γ-globin haplotypes

SNP	Position	Haplotype A	B	C	D
$^{G}\gamma$ Homology block					
rs2855121	199	A	G	G	G
rs2855122	254	G	A	G	A
rs2855123	412	T	A	T	A
rs2011051	672	C	A	C	A
rs7482144	1321	T	C	C	C
rs1894398	1688	C	T	C	T
rs2070973	2083	A	G	A	G
rs11036476	2147	G	A	G	A
–	2201	G/A	A	A	A
rs11036475	2250	C	T	C	T
rs11036474	2312	G	A	A	A
–	2607	G	A/G	G	G
rs2070972	2773	T	G	T	G
rs34879481	3037	+	–	–	–
rs2236794	3223	G	A	G	A
rs10768707	3568	G	G	A	G
–	3857	T/C	C	C	C
rs2255519	3949	C	T	C	T
rs2855126	4343	G	C	G	C
rs74049345	4695	C	T/C	C	C
rs2855036	4808	A	G	G	G
rs34306743	4932	+	–	–	–
$^{A}\gamma$ Homology block					
rs2855038	5336	A	G	A	G
–	5565	G	A/G	G	G
rs2855039	5819	A	G	G	G
rs11575900	6181	+	+	–	+
rs1894397	6427	A	G	G	G
rs1894398	6612	C	T	C	T
rs1061234	6804	T	T	C	T
rs33993529	6916	A	C	C	C
rs33988501	6951	T	T	G	T
rs2070973	7007	A	G	G	G
–	7050	C	T	T	T
rs11036455	7072	T	C	T	C
–	7079	G	A	A	A
rs59374378	7097	C	T	T	T
–	7144*	+	–	+	–
rs2855041	7155	T	C	C	C
rs2856807	7157	G	A	A	A
rs11036475	7174	C	T	T	T
rs2855042	7182	C	T	T	T
rs2855043	7183	T	C	C	C
–	7192†	+	–	–	–
rs2187608	7559	G	C	C	C
rs28379094	7684	G	A	G	A
rs28440105	7691	G	G	T	G
–	7905‡	C	C	T	C
–	7935	A	A	C/A	A
rs3841756	7956	+	+/–	+	+
rs916111	8147	A	T	A	T
rs6578592	8350	T	T	G	T
rs10488676	8693	C	T	C	T
rs7483789	8868	A	C	A	C
rs7480197	9084	G	A	G	A
	Frequency	0.30	0.17	0.17	0.29

Haplotypes were experimentally determined from 90 individuals of northern European origin, of which 93% fell into four haplotype classes, A to D. Within each class, haplotypes were identical or differed by the occasional derived-state variant restricted to that class, as indicated. SNP coordinates are given relative to the beginning of the $^{G}\gamma$-globin homology block at chromosome 11, position 5,277,490 (assembly GRCh37, release 56). SNPs shared by both homology blocks are indicated in matching colors, and alleles at complex SNPs and indels are shown by footnotes. The $^{G}\gamma$ and $^{A}\gamma$ homology blocks also share a complex variable $(GT)_n/(GC)_n$ microsatellite, at positions 2540 to 2585 and 7460 to 7504, respectively, but these were not surveyed over all individuals.

*At position 7144: – represents TA; + represents TCTTTA.

†At position 7192: – represents TG; + represents TATTCG.

‡At position 7905: T = CTCACT, C = CCTCTT.

Table S2. Recurrent de novo base mutations identified in γ-globin gene rearrangements

Mutant	Man	Tissue	Rearrangement type	Mutation	Location*	No. molecules with complete switch	No. additional molecules with mixed sites†	Bonferroni-corrected P‡
1	2	Sperm	Deletion	C→G	A	3	0	—
2	3	Sperm	Deletion	T→C	B	3	2	0.044
3	3	Blood	Deletion	T→A	C	2	1	0.0069
4	3	Sperm	Duplication	C→A	D	1	1	0.0034
–	3	Blood	Duplication	A→G	B	1	1	0.91
5	2	Sperm	Duplication	T→C	B	1–2§	0	—

*All mutations were in single-copy DNA and none affected coding sequences. Locations were: A, upstream of the $^{G}\gamma$-globin gene; B, between $^{G}\gamma$ and $^{A}\gamma$; C, in intron 2 of $^{A}\gamma$; D, upstream of $^{G}\gamma$ or between $^{G}\gamma$ and $^{A}\gamma$.

†Additional recurrences of base switches detected as mixed sites in PCRs that presumably contain more than one mutant. These were found by systematically surveying each base switch site over all sequence traces. In every case, these mixed sites were only seen in the man showing the switched base, and only in PCRs containing the same rearrangement as seen for the molecule with the switch.

‡Probability of obtaining such mixed sites by chance through PCR misincorporation, estimated from the frequency and spectrum of mixed bases seen in progenitor and rearranged sequences (Fig. S4). By way of example, consider the three sperm deletions from man 3 carrying a T→C switch, also seen twice as mixed sites. Mixed bases occurred at a frequency of 1.7×10^{-4}/bp of sequenced DNA and appeared to be randomly distributed along sequences. From base composition over the γ-globin gene region and from the mixed base spectrum, the probability of a T→C misincorporation at a given T base is 2.2×10^{-4}. The chance that two or more of the 578 additional deletion sequences analyzed over all men in sperm and blood would show this T→C misincorporation is 0.0073. The switched base was seen on a common class of deletion in man 3 and was present as a mixed site in two of his 155 remaining sequenced PCRs containing this deletion. These account for 155 of 578 (27%) of all deletion sequences from all men examined for mixed sites, and the chance that both mixed sites would be in PCR reactions containing this specific rearrangement in this man is 0.072. The overall probability of obtaining two mixed sites showing the correct switch in appropriate PCRs through PCR misincorporation is therefore $0.0073 \times 0.072 = 0.00053$, or 0.044 after Bonferroni-correction for the 83 different base switches analyzed. Neither of the instances of recurrence at mixed sites in sperm is likely to be the result of PCR misincorporation, and we therefore included these additional molecules in our final tally of candidate mutations. One of the blood recurrences could be explained by chance misincorporation and was rejected. The final tally showed that 4 of the 14 to 15 recurrent switches (27%) were present in mixed PCRs containing more than one mutant; this was not unexpected, given that 26% of all positive PCR reactions contained more than one mutant.

§Either two different duplications sharing the same switch or a heteroduplex molecule with the same (complementary) switch on both DNA strands (Fig. 4*B*). Each of the other recurrent switches was seen on rearranged molecules with identical structures and the same haplotype of origin.