Human fetal globin DNA sequences suggest novel conversion event

C.J.Stoeckert, Jr., F.S.Collins and S.M.Weissman

Department of Human Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA

Received 14 March 1984; Accepted 7 May 1984

ABSTRACT

DNA sequencing studies of two recently cloned human A_γ globin alleles has revealed a number of base differences which are clustered in the large intron (IVS-2). One allele has a previously undescribed IVS-2 sequence. Most of the allelic differences can be explained as resulting from a gene conversion event involving G_γ as a donor. A novel feature of this event is that three G_γ -like regions occur interspersed among unconverted areas of the A_γ gene. We propose that an alternating purine-pyrimidine run which is located between two of the converted sites is the initiation site of the conversion event. Consistent with models of gene conversion, this poly (purine-pyrimidine) tract has single-stranded characteristics in supercoiled plasmids as assayed by S1-nuclease.

INTRODUCTION

Gene conversion, or non-reciprocal recombination, is a well established phenomenon in fungi such as yeast. Based on genetic and molecular analysis of progeny, conversion of part or all of a gene by another allele or a neighboring related gene has been characterized (1-6). Furthermore, models of the conversion event have been proposed (7-10) (see Fig. 1) to explain these events on a molecular basis.

Gene conversion appears not to be restricted to fungi. Compelling evidence for conversion has been shown in cultured mouse cells (11). Recent DNA sequencing studies in gene families such as the globins (12,13,14), the immunoglobulins (15,16) and the histocompatibility antigens (17,18) have provided evidence for gene conversion-like events in mammals. One such study by Smithies and colleagues (12) contained evidence for the conversion of the A_γ fetal human globin gene by the neighboring G_γ fetal globin gene. The two fetal genes, G_γ and A_γ , are the result of gene duplication estimated to have occurred about 30 million years ago (19). The genes are close in position (coding regions are separated by 3.3 kb) and close in coding sequence [0.35% divergence (20)]. A number of differences are found in the large intron (IVS-2) between the two genes. However, in comparing the homology between

two A_Y alleles and that between a G_Y (G_Y-A) gene and an A_Y (A_Y-A) gene on the same chromosome, (12) closer homology is seen between the G_Y-A and A_Y-A than between the A_Y alleles (Fig. 2, 3). A boundary for homology between G_Y-A and A_Y-A genes occurs in the 3' part of IVS-2 at a run of poly purine-pyrimidine (poly R·Y). For 1.5 kb 5' to this boundary, G_Y-A matches A_Y-A very closely (8 differences). In contrast, there are 11 differences in 150 bases 3' to the poly (R·Y) run between G_Y-A and A_Y-A where the A_Y alleles were identical. Slightom *et al.* proposed (12) that a G_Y converted A_Y over that 1.5 kb with a putative border of conversion occurring at the poly (R·Y) run. Poly (R·Y) sequences have been shown to form single stranded regions in supercoiled DNA (21,22) as well as having the potential to form Z-DNA (23,24,25). Thus such a site would be a good candidate for DNA strand invasion to occur (Fig. 1).

Figure 1. Pathways of gene conversion. A) Meselson-Radding model adapted from Ref. 10. Initiation occurs when single stranded DNA from donor gene (thick lines) invades the acceptor gene (thin lines) displacing a "D" loop. The "D" loop is degraded leaving the acceptor gene containing a heteroduplex region. The bubbles are region of mismatch. The heteroduplex mismatches can be corrected by either strand independent (random) or concerted (co-conversion) mismatch repair. A third pathway (patchy) is suggested by the A_Y-S gene sequence. B) Examples of D-loop strand invasion. The invasion site is marked by vertical arrows. Heteroduplexes are formed flanking the invasion site which grow by displacement of the acceptor gene strand by the invading loop. The bottom example is adapted from the double strand break model (9).

MATERIALS AND METHODS

Cloning of fetal genes

Leukocyte DNA (26) was digested with Kpn I, size selected to 30-45 kb over a 10-40% sucrose gradient, and ligated to the cosmid vector pFC1 (a modification of pJB8 (27) wherein a Kpn I cloning site is introduced). DNA was packaged, transfected into HB101 and plated using standard methods (28), and nitrocellulose filters were screened with a purified 1.9-Kb Bam fragment containing the 5' end of the β gene.

DNA sequence analysis

Fetal genes were subcloned into M13mp8 (29) by digestion of total cosmid DNA with either Pst I or Pvu II and also by sonication of total cosmid DNA (30). Clones were identified by screening with the purified 2.6 kb EcoRI fragment containing the 5' end of the A_Y gene. Phages were prepared for dideoxy sequencing procedure of Sanger (31) by standard methods (32). Independent cosmids of both alleles were digested by Hha I or Fnu 4HI and analyzed on 1% agarose gels to check consistency of sequences determined. Fnu 4HI digests were transferred to nitrocellulose by the method of Southern (33) and probed with the Pst I fragment containing the A_Y IVS-2.

S1 nuclease analysis of fetal genes

A cosmid containing the sickle mutation was cut with Hind III and subcloned into the Hind III site of pBR322. Plasmids were supercoiled as isolated from bacteria. 10 μ g of plasmid were treated with 500U of S1 nuclease (BRL) for 90 seconds at 37 $^{\circ}$ in 100 μ l of 0.3 M NaCl, 30mM NaAcetate pH4.5, 3mM ZnCl $_2$. Reactions were stopped on ice and extracted with phenol/chloroform immediately. Samples were cut with either EcoRI or BstEII, the 5' overhanging ends filled in using α^{32} PdATP or α^{32} PdGTP respectively, and run on a 4% polyacrylamide gel. Sizes were estimated using HinfI cut pBR322 as a marker.

RESULTS

Fetal gene sequences

We have cloned into cosmids (26) the 39kb Kpn I fragment which contain the γ , δ , and β genes from a patient with sickle cell trait. The sickle mutation, which changes a Mst II restriction enzyme site, allows unambiguous identification of which allelic chromosome 11 a cosmid is derived from. As the cosmids contain both fetal genes as well as the β gene, it is also unambiguous to determine which fetal genes are syntenic.

The A_Y sequence gene from the sickle mutation containing cosmid (A_{Y-S})

described (12) in that the base changes result in A_{γ} -S resembling G_{γ} -S more closely.

The base differences observed between the G_{γ} and A_{γ} alleles are presented in Fig. 2. Gaps are treated as one difference. Sixteen changes scattered over the approximately 900bp IVS-2 (1.8% divergence) are found between A_{γ} -S and A_{γ} -B. Comparison of A_{γ} -S to the other published A_{γ} sequence [A_{γ} -A: clone 165.24 of Slightom (12)] reveals 20 differences (2.2%) of which 11 are at sites at the same positions as to the A_{γ} -S/ A_{γ} -B differences. Of all the changes noted between A_{γ} -S and the two published A_{γ} sequences, all but five are at sites noted as G_{γ} - A_{γ} differences. If positions where A_{γ} -S resembles G_{γ} are excluded, then A_{γ} -S differs from A_{γ} -B at only 4 bases (0.45% divergence). For comparison, G_{γ} -S differs from G_{γ} -A by 7 differences (0.8% divergence), 3 of which are clustered at the highly polymorphic poly (R·Y) tract. Only one of the G_{γ} - G_{γ} differences is a base found in A_{γ} and that is the common Hind III polymorphism (34) at position 807 (Fig. 2). The β globin gene has been sequenced for numerous alleles in the course of identifying thalassemic mutations. The 850 bp IVS-2 of Mediterraneans and Asians varies at most by 0.35% (35). Thus the A_{γ} gene has accumulated mutations at a rate higher than G_{γ} and β .

Precautions were taken to insure that differences observed are real and not artifacts of sequencing or cloning. Independent overlapping clones were sequenced to rule out sequencing errors and initial cloning was done in a $RecA^{-}$ host bacterial strain. In addition, restriction enzyme digests were done on cosmids independent of the ones sequenced to examine the fidelity of the base differences observed in A_{γ} -S. One difference at position 180 alters a HhaI site. Another, Fnu4HI, is gained by two base changes at positions 718-719 in A_{γ} -S (See Fig. 2). Both enzymes show consistent cutting patterns in A_{γ} -S of independent cosmids (data not shown). Thus, recombination did not occur during the cloning.

S1 analysis of fetal genes

The fetal genes, G_{γ} -S and A_{γ} -S, were subcloned as 7.8 kb and 3.4 kb Hind III fragments, respectively, into the Hind III sites of pBR322. The plasmids were supercoiled as derived in isolation from the bacteria. These plasmids were treated with the single strand specific nuclease S1 in order to detect any single stranded regions. Under the conditions used, linearization of the plasmids was obtained in both G_{γ} -S and A_{γ} -S containing plasmids. To map the positions of the S1 sites, treated plasmids were cut with either EcoRI or BstEII and run against size markers (Fig. 4). Analysis of the bands produced by

Human fetal globin IVS-2

Figure 3. Schematic depiction of two conversion events in A_γ IVS-2. G_γ and A_γ -B are the sequences reported in ref. 1 as clones 165.24 and 51.1 respectively. Sites of differences between the two are marked at the top of the figure. Also, marked is the poly(R·Y) tract as (GT)_n. A_γ -A and A_γ -S are the converted alleles from ref. 1 and this report. The filled in regions are the areas where G_γ has converted A_γ . The hatched areas indicate the uncertainty in the boundaries of conversion due to lack of differences between A_γ and G_γ in those areas. Also, given below A_γ -S are the S1 sensitive sites (S) as determined in Fig. 4.

the digestions positioned S1 sites within the poly (R·Y) tract. Two prominent sites are seen (Fig. 3 and 4). About a five-fold excess of G_γ -S plasmid over A_γ -S plasmid was needed to produce comparable intensities of S1-Bst EII bands.

DISCUSSION

The A_γ alleles appear to have diverged to a greater extent than β and G_γ alleles. An explanation for an increased mutation rate in A_γ alleles is higher mutability of the A_γ large intron. Increased random mutations however, do not account for the observed differences between A_γ alleles which result in an A_γ resembling G_γ . This point is illustrated by examining some of the allelic differences closely. The sequence of A_γ -S is the same as A_γ -B at most of the positions where A_γ -A and A_γ -B vary (nucleotides 24, 115, 158, etc.). Since 2 alleles, A_γ -S and A_γ -B have a C at position 24, we suggest it to be the unconverted A_γ sequence. Thus A_γ -A is the variant with an A at position 24 which was converted by G_γ (also an A at position 24). Using the same reasoning, positions 716, 722 and 749 are identical in A_γ -A and A_γ -B and

Figure 4. S1 nuclease sensitive sites in A_γ and G_γ . The $Hind$ III fragment containing A_γ or G_γ from the HbS cosmid was cloned into pBR 322 and treated with S1 as supercoiled plasmid. Top) After S1 treatment, samples were digested with Bst EII and run on a 4% acrylamide gel. Lane one (1) $Hinf$ I digested pBR 322 (2) A_γ (3) G_γ . Bottom) Schematic depiction of the fetal gene indicating Bst EII sites (b) and prominent S1 sites (S). Positions of the S1 sites were confirmed using Eco RI in place of Bst EII.

thus represent the unconverted A_γ sequence. Then A_γ -S is the variant whose changes are identical to G_γ at those sites and presumably converted. Positions 180, 205 and 251 are probably also converted in A_γ -S by G_γ , however, to be consistent, we must consider the possibility that these sites are variant in A_γ -B as A_γ -A, A_γ -S and G_γ are all identical (except for T to G at 251 in A_γ -S). As more fetal genes are sequenced, the unconverted or common sequence for A_γ and G_γ will be determined. If as Slightom *et al.* (1) proposed these positions were converted in A_γ -A, then these sites would be converted in A_γ -S.

A double crossover between G_γ and A_γ is an alternative recombinational event which would result in A_γ partially resembling G_γ . This would require three such events to occur in a region of 700 base pairs. One cross-over would be in a 20bp region which seems statistically unlikely. Therefore, recombination through gene conversion is a more likely explanation for the observed data.

A_γ -S is clearly the result of a different conversion event than A_γ -A. However, the conversion event may have initiated at the same poly (R·Y) tract

described earlier. Slightom *et al.* (12) proposed that since the tract was at the 3' end of the converted region and consists of a simple repeat, it was a hotspot for recombination. As mentioned earlier, we recognized that the poly (R·Y) tract might be single stranded under torsional stress due to its similarity to sequences known to be sensitive to single strand specific nuclease S1. Single stranded regions play a key role in models concerning gene conversions both as a site for strand invasion to occur and as a site for breakage and formation of the invading DNA strand (see Fig. 1). Thus, we treated supercoiled plasmids containing either A_Y-S or G_Y-S with S1 nuclease. Indeed, as shown in Fig. 4, both genes are sensitive to S1 with A_Y more sensitive than G_Y. A direct link between this type of sequence [poly (R·Y)] and increased recombination has been shown in prokaryotes (24). Evidence is accumulating for such a link in eukaryotes. For example, a *Hinf* I restriction polymorphism just 5' to the β globin gene and near a poly (R·Y) tract found to be S1-sensitive (22) shows an unexpected lack of linkage disequilibrium to flanking polymorphic sites (35). In addition, the α globin gene duplication is bordered by a (CA)₄ block (36) and in one case insertion of SV40 into the rat genome involved recombination between (CA)_n sequences. (37)

The presence of the poly (R·Y) sequence in the fetal IVS-2 may explain why two examples of conversion have been found in only four genes analyzed. It is curious that both cases of conversion in the fetal genes were of A_Y by G_Y and not the reverse. This may simply be due to the small number of genes analyzed. Alternatively, the greater propensity of A_Y to form single stranded regions in IVS-2 than G_Y (as implicated by the increased S1 sensitivity) may enhance the role of A_Y as the acceptor in conversion (see Fig. 1).

A novel feature of the conversion event in A_Y-S is that there are 3 sites of conversion interspersed among unconverted regions. This situation has not been reported in fungal studies or any mammalian system except the γ 2a- γ 2b conversion, which shows a much higher degree of interspersion of converted and unconverted regions to the point that correction of mismatches appears to be random (10, 16). The A_Y conversion is, on the other hand, "patchy", i.e., consists of blocks of converted and unconverted regions up to a hundred or so bases in length (see Fig. 2). The patchy conversions could represent the result of three separate conversions. Alternatively, the base change at position 180 might be a point mutation. If so, then two conversions could have produced the A_Y-S sequence. These scenarios cannot be ruled out and would place the conversions into the same class as most others have

reported. For such to be the case, however, the conversions must be frequent and non-overlapping, or require some back conversions from an allelic A_Y gene. A more economical explanation would involve only one event. While this work was in progress, Michelson and Orkin (39) reported on a patchwork of conversion-like events in the α -globin gene system. They propose that the two contiguous patches of highly homologous sequence between the nonallelic α_2 and α_1 genes result from independent conversions based on differing levels of homology in the two patches. In the fetal gene IVS-2, no such clear cut distinction between homologies preserved in converted regions can be made.

Assuming that the A_Y -S conversion was a single event, it is interesting that the proposed initiation site [the poly (R·Y) tract) is in the middle of the converted region rather than bordering it. This finding suggests an alternative to a nicked single strand (Fig. 1) such as an invading D-loop (9,38) leading to a pair of heteroduplexes flanking the site of invasion. Another curious feature of A_Y -S is the additional number of unique allelic mutations found in and near the converted sites. These might reflect errors in the correction of the heteroduplex formed during conversion. The T to G change at position 251 (Fig. 2) is particularly suggestive. A similar finding was reported in the $\gamma 2A$ - $\gamma 2B$ (16) conversion.

The novel features of the A_Y -S conversion may reflect an alternate pathway of mismatch repair. Mismatch repair has been proposed (7) as a mechanism by which base differences in the heteroduplex (Fig. 1) formed during conversion are corrected. Coconversion such as seen in A_Y -A (12) could occur by differences being excised and repaired using only one strand (the donor) as a template. Independent or random conversion could occur, as in $\gamma 2a$ - $\gamma 2b$ (16), by no strand preference for correction. In this case, the probability of any two neighboring differences being corrected off of the same strand is 0.50. The A_Y -S may be a third pathway intermediate between these two cases in which the repair enzymes correct a difference in the heteroduplex and continue correction over a range of 20-100 bases. The corrected region is replaced using one strand as template resulting in a patch of coconversion. No preference of strand for initiation of the excision would result in interspersion of converted (donor strand used as template) and unconverted (acceptor strand used as template) as seen in A_Y -S. Mismatch repair has been shown to occur in cultured mouse and monkey cells (40,41). In those experiments, SV40 molecules were constructed to contain heteroduplex regions and analyzed after infection with regard to segregation of distinguishing restriction enzyme sites. Both independent and cocorrec-

tion was observed. Precedent for different pathways of correction comes from the fungi where *Ascomobolus* and *Saccharomyces* show different strand preference in correcting heteroduplex differences particularly insertions and deletions (2,5,9)

To summarize, we report a unique sequence in the IVS-2 of an A_Y allele which is most likely the result of G_Y converting A_Y . In addition, the converted region contains single strand forming sites. The data presented give evidence that: 1) conversion is a common event; 2) the heteroduplex formed can be resolved by patchy mismatch repair and 3) may occur in a bidirectional manner from the poly (R $\cdot$ Y) initiation site. Experiments looking at conversion in cultured cells (11) will be needed to confirm and extend these properties. Based on the example presented here, gene conversion in man may use a pathway of mismatch repair in addition to that than reported in fungi.

ACKNOWLEDGEMENT

The authors wish to thank Joe Chamberlain for excellent technical assistance and Ann Mulvey for the preparation of this manuscript. Subclones of the fetal genes in pBR322 were kindly provided by J. Metherall. C. S. was supported by N.I.H. Grant 5T32CA0959-08. F. C. was supported by fellowships from the Cooley's Anemia Foundation, Inc., and the Charles E. Culpeper Foundation. S. M. W. was supported by NIH Grant GM-32156.

REFERENCES

1. Petes, T. and Fink, G.R. (1982) *Nature* 300, 216-217.
2. Fogel, S., Mortimer R.C. and Lusnak, K. (1982) in the *Molecular Biology of the Yeast Saccharomyces*, Y.N. Strathern, E.W. Jones, J.R. Broach, Eds., Cold Spring Harbor Laboratory, New York.
3. Egel, R. (1981) *Nature* 290, 191-192.
4. Jackson, J.A. and Fink, G.R. (1981) *Nature* 292, 306-311.
5. Rossignol, J.L. Paquette, N. and Nicholas, A. (1979) *Cold Spring Harbor Symp. Quant. Biol.* 43, 1343-1356.
6. Klein, H.L. and Petes, T.D. (1981) *Nature* 289, 144-148.
7. Holliday, R. (1964) *Genet. Res* 5, 282-304.
8. Meselson, M.S. and Radding, C. M. (1975) *Proc. Natl. Acad. Sci. USA* 72, 358-361.
9. Szostak, J.W., Orr-Weaver, T.L., Rothstein R.J. and Stahl, F.W. (1983) *Cell* 33, 25-35.
10. Kourilsky, P. (1983) *Biochimie* 65, 85-93.
11. Liskay, R.M. and Stachelek, J.L. (1983) *Cell* 35, 157-165.
12. Slightom, J.L., Blechl, A.E. and Smithies, O. (1980) *Cell* 21, 627-638.
13. Liebhaver, S.A., Goossens, M., and Kan, Y.W. (1981) *Nature* 290, 26-29.
14. Martin, S.L., Vincent K.A., and Wilson, A.C. (1983) *J. Mol. Biol.* 164, 513-528.
15. Bentley, D.L. and Rabbitts, T.H. (1983) *Cell* 32, 181-189.
16. Ollio, R., and Rougeon, F. (1983) *Cell* 32, 515-523.

17. Weiss, E.H., Mellor, A., Golden, L., Fahrner, K., Simpson, E., Hurst J. and Flavell, R.A. (1983) *Nature* 301, 671-674.
18. Pease, L.R., Schulze, D.H., Pfaffenbach, G.M. and Nathenson, S.G. (1983) *Proc. Natl. Acad. Sci. USA* 80, 242-246.
19. Shen, S., Slightom J.L. and Smithies, O. (1981) *Cell* 26, 191-203.
20. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C.O., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O.S., Baralle, F.E., Shoulders, C. C. (1980) *Cell* 21, 653-667.
21. Singleton, E.K., Klysik, J., Stirdivant S.M. and Wells, R.D. (1982) *Nature* 299, 312-316.
22. Cockerill, P.N. and Goodwin, G.H. (1983) *BBRC* 112, 547-554.
23. Arnott S., Chandrasekaran, R., Birdsall, D.L. Leslie, A.G.W., and Ratliff, R.L. (1980) *Nature* 283, 743-745.
24. Klysik, J., Stirdivant S.M. and Wells, R.D. (1983) *Journ. Biol. Chem.* 257, 10152-10158.
25. Nordheim, A. and Rich, A. (1983) *Nature* 303, 674-679.
26. Collins, F.S., Stoeckert, C.J., Serjeant, G., Forget, B., Weissman, S.M. (Manuscript in preparation).
27. Ish-Horowicz, D. and Burke, J. (1981) *Nucleic Acids Research* 9, 2987-2998.
28. Grosveld, F.G., Dahl, H.H., DeBoer, E., and Flavell, R.A. (1981) *Gene* 13, 227-237.
29. Messing, J. and Vieira, J. (1982), *Gene* 19, 269-276.
30. Deininger, P.S. (1983) *Anal. Biochem.* 129, 216-223.
31. Sanger, F., Nicklen, S., and Coulson, A.E. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5468.
32. Godson, G.N. in *Methods of DNA and RNA Sequencing*, S.M. Weissman, Ed. (Praeger, New York, 1983) 69-111.
33. Southern, E.M. (1982) *J. Mol. Biol.* 97, 503-517.
34. Jeffreys, A.J. (1979) *Cell* 18, 1-10.
35. Orkin, S.H., Antonarakis, S.E. and Kazazian Jr., H.H. (1983) *Progress in Hematology* 12, 49-73.
36. Proudfoot N.J., and Maniatis, T. (1980) *Cell* 21, 537-544.
37. Stringer, J.R. (1982) *Nature* 296, 363-366.
38. Bianchi, M., DasGupta, C. and Radding, C.M. (1983) *Cell* 34, 931-939.
39. Michelson, A.M. and Orkin, S.H. (1983) *Journ. Biol. Chem.* 258, 15245-15254.
40. Miller, L.K., Cooke, B.E. and Fried, M. (1976) *Proc. Natl. Acad. Sci. USA* 73, 3073-3077.
41. Wilson, J.H. (1977) *Proc. Natl Acad. Sci. USA* 74, 3503-3507.