Cell, Vol. 21, 627–638, October 1980.

Human Fetal $^{G}\gamma$– and $^{A}\gamma$–Globin Genes: Complete Nucleotide Sequences Suggest That DNA Can Be Exchanged between These Duplicated Genes

Jerry L. Slightom, Ann E. Blechl and Oliver Smithies
Laboratory of Genetics
University of Wisconsin
Madison, Wisconsin 53706

Summary

We present the nucleotide sequences of the $^{G}\gamma$– and $^{A}\gamma$–globin genes from one chromosome (A) and of most of the $^{A}\gamma$ gene from the other chromosome (B) of the same individual. All three genes have a small, highly conserved intervening sequence (IVS1) of 122 bp located between codons 30 and 31 and a large intervening sequence (IVS2) of variable length (866–904 bp) between codons 104 and 105. A stretch of simple sequence DNA occurs in IVS2 which appears to be a hot spot for recombination. On the 5′ side of this simple sequence, the allelic $^{A}\gamma$ genes differ considerably in IVS2 whereas the nonallelic $^{G}\gamma$ and $^{A}\gamma$ genes from chromosome A differ only slightly. Yet on the 3′ side of the simple sequence, the allelic genes differ only slightly whereas the nonallelic genes differ considerably. We hypothesize that the 5′ two thirds of the $^{A}\gamma$ gene on chromosome A has been "converted" by an intergenic exchange to become more like the $^{G}\gamma$ gene on its own chromosome A than it is like the allelic $^{A}\gamma$ gene on the other chromosome B. Our sequence data suggest that intergenic conversions occur in the germ line. The DNA sequence differences between two chromosomes from a single individual strongly suggest that DNA sequence polymorphisms for localized deletions, additions and base substitutions are very common in human populations.

Introduction

The loci specifying the amino acid sequences of the globin chains of mammalian hemoglobins usually occur in duplicated nonallelic pairs which frequently differ in their relative expressions at different stages of development (Bunn, Forget and Ranney, 1977). For example, most humans have two α genes (Orkin, 1978); two adult β-type genes, δ and β (Lawn et al., 1978); and two fetal β-type genes, $^{G}\gamma$ and $^{A}\gamma$ (Fritsch, Lawn and Maniatis, 1979; Bernards et al., 1979; Little et al., 1979a; Ramirez et al., 1979; Tuan et al., 1979). In mice a similar situation exists; some strains have two adult β genes which code for different although closely related globins, β^{major} and β^{minor} (Tiemeier et al., 1978), while in other strains either the two adult β genes code for identical products, β^{s}, or only one of

duplicated globin genes varies, but in many cases the products of a pair of duplicated genes within a given species are more like each other than either is like the comparable pair of globins in other species. For example, adult human δ– and β–globins, which differ in 10 amino acid residues out of 146, are more different than are the fetal human globins $^{G}\gamma$ and $^{A}\gamma$, which differ in only 1 residue, but the adult human δ– and β–globins are still more similar to each other than either is to the adult mouse β–globins (the minimum human/mouse difference is 26 residues). The conventional explanation of these findings is that during evolution the globin genes have increased and decreased their number by unequal crossovers and/or duplication events. A high degree of similarity between the members of a duplicated pair within a species is usually taken to indicate a short evolutionary time since the last duplication occurred. A lack of similarity between the comparable members of duplicated pairs in two different species is taken to mean that the last duplications occurred after the species had diverged. (See Little et al., 1979b, for a discussion of the problems of this argument in relation to the fetal globin genes.)

Duplications probably first arise by rare breakage and reunion events at nonhomologous points on two chromosomes, resulting in an unequal exchange between them (Muller, 1936; Smithies, 1964). Duplications may span part of a gene with no intergenic DNA (such as the haptoglobin Hp^{2} allele; Smithies, Connell and Dixon, 1962) or may involve several genes and intergenic DNA (such as the Bar locus in Drosophila; Bridges, 1936). The DNA between duplicated genes which have recently arisen by a single nonhomologous breakage and reunion event must also be duplicated, either completely on one side or partly on both sides of the duplicated genes.

These considerations make it difficult to account, by events involving duplication followed by unequal but homologous crossing over, for the occurrence of closely related genes adjacent to DNA which does not itself appear to be duplicated. Such cases are known. For example, the two adult mouse β–globin genes appear to be the result of a relatively recent duplication, since their structural genes readily hybridize with each other to form a DNA heteroduplex (Tiemeier et al., 1978); however, the DNA flanking these two genes does not show sufficient homology to form a heteroduplex. The problem is that of understanding how duplicated genes can continue to share many species-specific features while their flanking DNA shows little or no evidence of once having had a common origin.

In this paper we present the complete nucleotide sequence of the two human fetal globin genes, $^{G}\gamma$ and $^{A}\gamma$, from one chromosome (A), and of most of the $^{A}\gamma$

example at the molecular level of a mechanism permitting the co-evolution of related linked genes without the need to involve the DNA between these genes. Thus the data indicate that the $^{G}\gamma$– and $^{A}\gamma$–globin genes on a given chromosome can exchange DNA sequences by a recombinational event like a gene conversion, and that a stretch of simple sequence DNA in IVS2 is a hotspot for initiating these exchanges. The sequence data suggest that conversions take place in the germ line and can occur between chromosomes as well as within a single chromosome.

In an accompanying paper (Efstratiadis et al., 1980) we present a detailed analysis of the nucleotide sequences of the human β-type globin genes [embryonic ϵ (Baralle, Shoulders and Proudfoot, 1980), fetal $^{G}\gamma$ and $^{A}\gamma$ (this paper), adult δ (Spritz et al., 1980) and adult β (Lawn et al., 1980)], and we compare these sequences with published globin gene sequences from other species.

Results and Discussion

Chromosomal Maps and Clones Studied

We have isolated clones covering the fetal globin region of both chromosomes of a diploid female donor and have found that the chromosomes differ in a number of places. Restriction enzyme sites used in defining the two chromosomes, arbitrarily labeled A and B, are presented in Figure 1 together with the code names and extents of the clones. Asterisks emphasize restriction sites which differ between the two chromosomes. The coding regions for the $^{G}\gamma$ and $^{A}\gamma$ genes are shown by heavy raised bars. About 14 kb to the 5′ side of the $^{G}\gamma$ gene of chromosome B we have identified another globin gene by hybridization. We presume that it is an ϵ–globin gene on the basis of restriction maps and sequence data presented by Proudfoot and Baralle (1979) and Fritsch, Lawn and Maniatis (1980), and have labeled it accordingly.

Clone 165.24 is the key clone from chromosome A. The restriction map of 165.24 (and the sequence data presented below) establish that it contains 14.3 kb of DNA which include the coding sequences for the expected two fetal globins, with $^{G}\gamma$ on the 5′ side of $^{A}\gamma$. The restriction map of 165.24 agrees within experimental error with the fetal portion of more extensive maps of the human fetal and adult globin region published by Bernards et al. (1979), Fritsch et al. (1979), Little et al. (1979a), Ramirez et al. (1979) and Tuan et al. (1979), using DNA from different donors. These previous studies showed one copy each of the $^{G}\gamma$ and $^{A}\gamma$ genes per chromosome. Since clone 165.24 contains one copy of each fetal globin gene in the same DNA environment found in the earlier studies, as judged by the restriction maps, we conclude that mosomes (Deisseroth et al., 1978) of our diploid donor.

Clone 164.6 is the key clone for chromosome B. It was an independent isolate from the same unamplified collection of in vitro packaged phages as 165.24. We identify it as being from chromosome B because its restriction map (Figure 1) and several critical sequenced regions (see below) differ from those of 165.24.

Clone 51.1 was isolated and initially characterized in earlier studies from this laboratory (Blattner et al., 1978; Smithies et al., 1978, 1979) using DNA from the same donor as for 165.24. DNA sequence data presented below show that the sequence of the $^{A}\gamma$ gene in 51.1 is substantially different from that of the $^{A}\gamma$ gene in 165.24, but is identical to the sequenced regions of the $^{A}\gamma$ gene in clone 164.6. Consequently we identify the $^{A}\gamma$ gene in clone 51.1 as being from chromosome B.

Organization of the Human Fetal Globin Genes

The restriction sites and strategy used in the DNA sequencing are shown in Figure 2. The same basic strategy was applicable to all three γ–globin genes. Figure 3 presents and compares the complete DNA sequences of the $^{G}\gamma$– and $^{A}\gamma$–globin genes from chromosome A and most of the DNA sequence of the $^{A}\gamma$–globin gene from chromosome B of our donor.

The data obtained from this sequence analysis support our earlier finding (Smithies et al., 1978) that the coding region of the human fetal $^{A}\gamma$–globin gene is divided into three segments by two noncoding intervening sequences, and extend this finding to the human fetal $^{G}\gamma$–globin gene. The smaller intervening sequence (IVS1) interrupts the coding region between codons 30 and 31, and the larger intervening sequence (IVS2) interrupts the coding region between codons 104 and 105. The arrows at the 5′ and 3′ boundaries of both IVS1 and IVS2 in Figure 3 point out a possible splicing frame for the removal of these intervening sequences which would conform to the GT/AG rule observed at the boundaries of the intervening sequences of many other genes (Breathnach et al., 1978). The human fetal globin gene intervening sequences occur in exactly the same positions as they occur in adult globin genes from mouse (Konkel, Tilghman and Leder, 1978; Konkel, Maizel and Leder, 1979), rabbit (van den Berg et al., 1978; van Ooyen et al., 1979) and human (Lawn et al., 1980; Spritz et al., 1980). Clearly the difference in expression between adult and fetal globins cannot be the result either of the fetal genes lacking these intervening sequences or of their presence in different places in the coding region.

IVS1 is 122 bases in length in all three γ–globin genes. The length of IVS2 is different in each of the

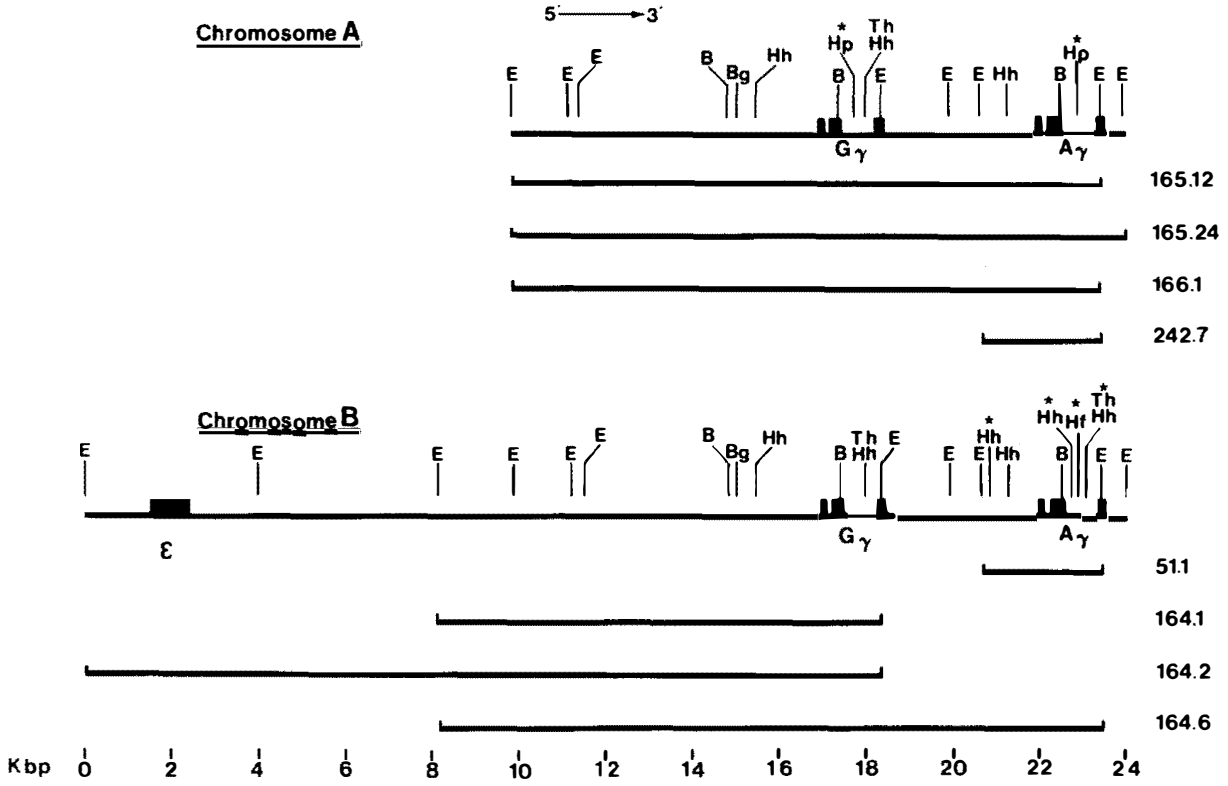

Figure 1. Maps Outlining Restriction Enzyme Sites Used to Define Chromosomes A and B of the DNA Donor

The coding regions of the $^{G}\gamma$– and $^{A}\gamma$–globin genes and the general location of a gene presumed to code for ϵ–globin are shown by heavy raised bars. The direction of transcription of the fetal globin genes is shown. Asterisks emphasize restriction sites which differ in the two chromosomes. The brackets under the two chromosome maps show the extents of the clones considered in this paper; their respective code numbers are listed alongside. The scale is in kilobase pairs. The restriction enzyme sites are (B) Bam HI; (Bg) Bgl II; (E) Eco RI; (Hf) Hinf I; (Hh) Hha I; (Hp) Hph I; (Th) Tha I.

of chromosome A and 876 bases in the $^{A}\gamma$ gene of chromosome B. Preliminary data (not presented here) indicate that the IVS2 in the $^{G}\gamma$ gene from chromosome B is the largest of the four, having about 904 bases. A difference in the size of IVS2 in the nonallelic $^{G}\gamma$– and $^{A}\gamma$–globin genes is not surprising, but we did not expect that the lengths of the allelic globin genes would differ.

Comparison and Analysis of Sequence Data

The general similarities and differences in the three γ–globin genes are illustrated in Figure 4 by a bar diagram showing the distribution of the differences. Two striking features are revealed by this comparison. First, substantial portions of the three genes have virtually identical sequences: the 5′ flanking and 5′ untranslated regions, the complete coding sequences, all of IVS1, and three regions at the ends and middle of IVS2. Second, in the 3′ third of the genes there are more differences between the nonallelic $^{G}\gamma$ and $^{A}\gamma$ genes on the same chromosome (hatched areas) than between the allelic $^{A}\gamma$ genes (unhatched areas); how-

5′ Flanking and 5′ Untranslated Region

Only one difference occurs in the DNA sequences of the 5′ flanking untranscribed region and 5′ transcribed but untranslated region of the three genes; this is at position 25, where the $^{A}\gamma$ gene from chromosome A has an adenine residue whereas the other two genes have a guanine. Chang et al. (1978) have published a sequence for (presumably mixed G and A) γ–globin mRNA; they only report a guanine at this position. At position 19 they report both guanine and cytosine, where we find only guanine. It is not known whether these differences are due to genetic polymorphisms or to problems in sequencing. In the remainder of the 5′ untranslated region of mRNA, the two sets of sequence data are in complete agreement.

Included in the 5′ flanking region are some sequences with features common to many globin genes and to other eucaryotic genes. These sequences are considered in detail in the accompanying comparison paper (Efstratiadis et al., 1980). They include a hexanucleotide sequence (AATAAA) starting 31 bases before the first nucleotide of the mRNA (overlined

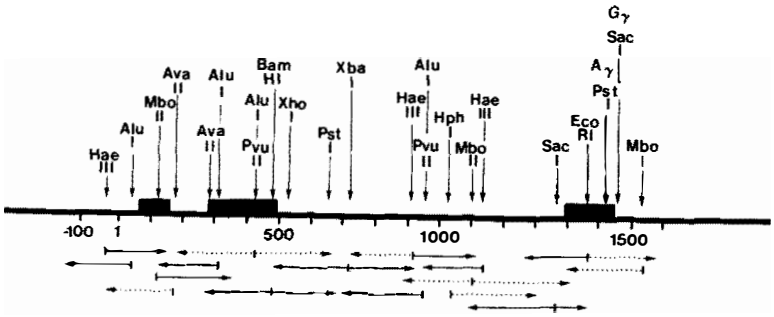

Figure 2. Detailed Map of the Restriction Enzyme Sites, Shown by Vertical Arrows, Used in Sequencing the $^{G}\gamma$ and $^{A}\gamma$ Genes of Clone 165.24 and the $^{A}\gamma$ Gene of Clone 51.1

The $^{A}\gamma$ above the Pst I site and the $^{G}\gamma$ above the Sac I site indicate that each site occurs only in the specified gene. The directions of sequencing shown by the solid horizontal arrows apply to all three genes; the dotted arrows apply only to the genes in 165.24. The positive and negative numbers on the scale refer to nucleotide positions; position 1 is the first adenine of the fetal globin mRNAs. The coding regions (common to all three genes) are shown by heavy bars.

(1979) in the Drosophila histone genes. [The hexanucleotide sequence (AATAAA) from the 5′ side of the γ–globin genes is also the same as the poly(A) addition signal found about 20 bases before the 3′ end of many eucaryotic mRNAs (Proudfoot and Brownlee, 1976), including our γ–globin genes. This identity is probably coincidental, but we cannot exclude the possibility that these sequences on the 5′ side of the γ–globin genes are part of the 3′ end of mRNAs transcribed from DNA on the 5′ side of the γ–globin genes.] A second region of possible functional importance in the 5′ untranslated region of many eucaryotic genes is the "capping box," first pointed out by Ziff and Evans (1978). The capping box consists of 12 nucleotides, 10 of which are on the 5′ side of the first nucleotide of the mRNA. The capping boxes in the three γ–globin genes all have the same sequence (GCAGTTCCACAC), which is underlined in Figure 3.

$^{G}\gamma$– and $^{A}\gamma$–Globin Coding Sequences

It is most remarkable that only one nucleotide difference occurs in the 438 bases comprising the coding region of the three sequenced γ–globin genes. This difference is in codon 136, where the $^{G}\gamma$–globin codon is GGA (coding for glycine) and the $^{A}\gamma$–globin codon is GCA (coding for alanine). The fact that the DNA sequences in the coding regions of the $^{G}\gamma$ and $^{A}\gamma$ genes are otherwise identical clearly indicates either that strong selection pressures are being exerted at the nucleotide level, and/or that some type of molecular mechanism exists whereby these duplicated genes have avoided evolutionary divergence.

The nucleotide sequence for $^{G}\gamma$ mRNA has previously been determined by Forget et al. (1979) from cDNA clones. A comparison of our $^{G}\gamma$–globin coding sequence with their recently revised sequence (B. Forget, personal communication) reveals no differences.

IVS1 Is Very Conserved

Comparison of the 122 bp IVS1 from the three γ–globin genes shows their DNA sequences to be almost

chromosome A have identical IVS1 DNA sequences, and there is only one difference between the IVS1 of the allelic $^{A}\gamma$ genes; this difference is at position 210, about in the middle of the IVS. The strict conservation of these nucleotide sequences again suggests that the sequences, like the coding region, must either be under strong selection pressure or else are maintained in identical form by some other mechanism. A result similar to ours has been reported by Konkel et al. (1979) for the IVS1 sequences from the mouse β^{major} and β^{minor} globin genes; three nucleotide differences were found in the 116 nucleotides of the IVS1 sequences of their mouse β–globin genes.

The 3′ Untranslated Region Contains a Considerable Number of Differences

The nucleotide sequences of the 3′ untranslated regions of both $^{G}\gamma$ mRNA (Forget et al., 1979) and $^{A}\gamma$ mRNA (Poon, Kan and Boyer, 1978) have already been determined. Both are 90 bp in length, but their sequences differ at six positions. Our $^{A}\gamma$ sequence agrees completely with that of Poon et al. (1978). We also find six differences between the two nonallelic genes at the same positions as in these previous reports. Four of the differences are in a block just after the terminator codon (positions 1508–1511 in Figure 3); the other two occur at positions 1522 and 1583. Whether these differences are related to the varied synthesis of the $^{G}\gamma$– versus $^{A}\gamma$–globins during development (Bunn et al., 1977) cannot be determined from the data currently available. Our $^{G}\gamma$ sequence differs in two positions from the $^{G}\gamma$ mRNA sequence reported by Forget et al. (1979). The first difference is at position 1510 (where a G was found in the mRNA and we find an A), and the second is at position 1583 (where a T was found in the mRNA and we find an A). These two positions also differ between the nonallelic genes.

Proudfoot and Brownlee (1976) have suggested, as mentioned above, that the hexanucleotide (AATAAA) forms a signal for poly(A) addition about 20 bp 5′ to the first nucleotide of poly(A) in many mRNAs. We

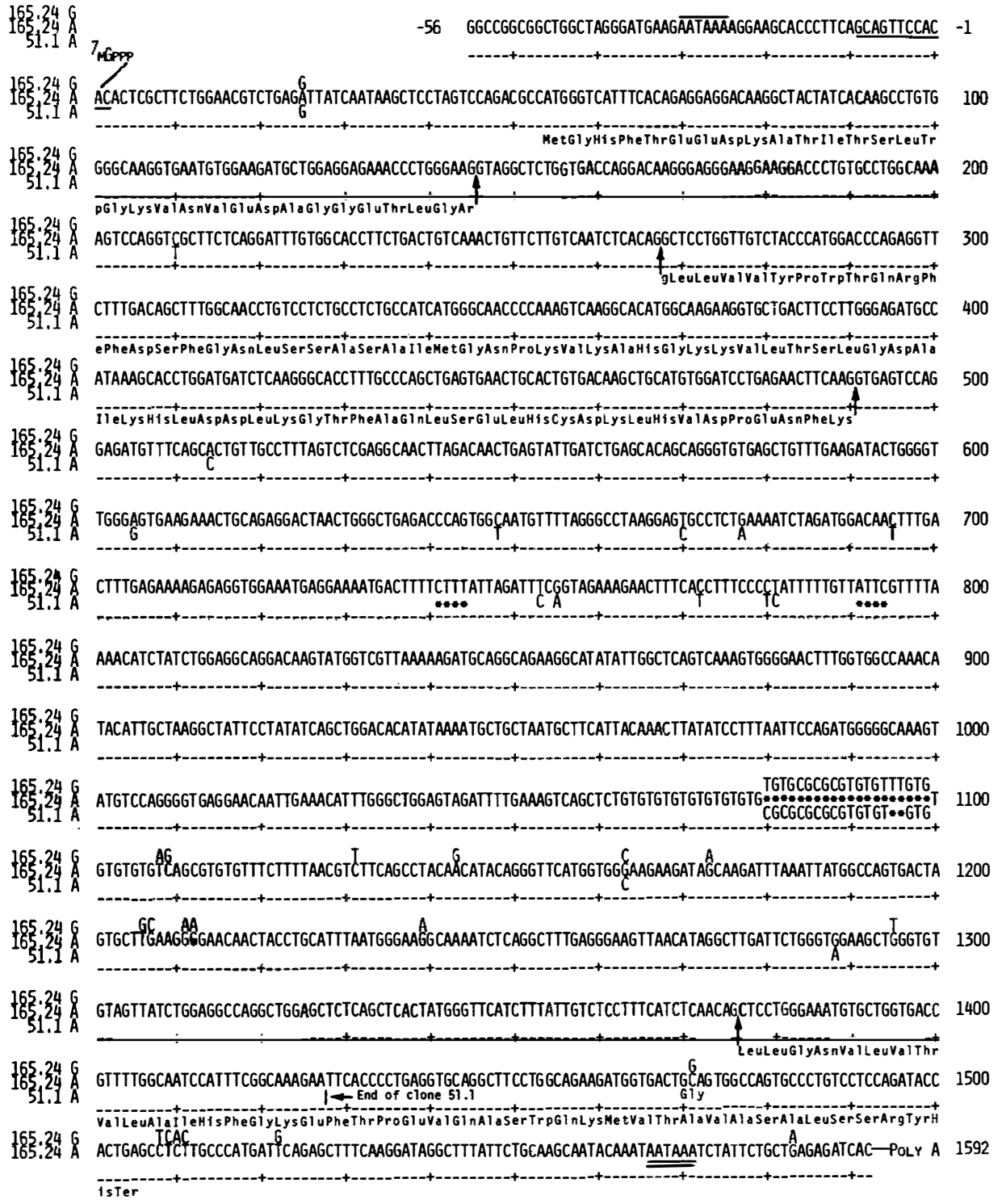

Figure 3. Nucleotide Sequences of the $^{G}\gamma$ and $^{A}\gamma$ Genes from Clone 165.24 and of the $^{A}\gamma$ Gene from Clone 51.1

The numbering system is taken from the $^{G}\gamma$ gene of 165.24, the largest of the three sequenced genes, with position 1 corresponding to the first adenine of globin mRNA to which the cap, 7mGppp, is added (Chang et al., 1978). The fully listed sequence is that of the $^{A}\gamma$ gene of 165.24; nucleotides which are different in the $^{G}\gamma$ gene of 165.24 or in the $^{A}\gamma$ gene of 51.1 are shown respectively above or below the sequence of the $^{A}\gamma$ gene of 165.24. Asterisks indicate gaps. Underlined and overlined nucleotides denote regions of possible biological importance (see text). The

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.