BLOOD

The Journal of
The American Society of Hematology

VOL 73, NO 5 APRIL 1989

REVIEW

A Review of the Molecular Genetics of the Human α-Globin Gene Cluster

By D.R. Higgs, M.A. Vickers, A.O.M. Wilkie, I.-M. Pretorius, A.P. Jarman, and D.J. Weatherall

DURING THE relatively brief period in which molecular biology has been applied to human disease, the structure, function, and synthesis of hemoglobin have been studied most intensively. Over the past 10 years we have acquired a detailed knowledge of the control of individual globin genes and gained some insight into the mechanisms underlying the tissue and developmental stage specific control of the coordinately regulated α- and β-globin gene clusters.

Much of our understanding of globin gene expression has come from observations on experimental systems and the analysis of naturally occurring mutants of α-like (α-thalassemia) and β-like (β-thalassemia) globin chain synthesis. This summarizes our current knowledge of the structure and function of the normal α-globin gene cluster and the molecular basis for α-thalassemia.

HISTORICAL PERSPECTIVE

The major clinical syndromes resulting from α-thalassemia (Hb H disease and the Hb Bart's hydrops fetalis syndrome) were first recognized in the mid 1950s and early 1960s through the association of the abnormal hemoglobins (Hb H and Hb Bart's) with hypochromic microcytic anemia in the absence of iron deficiency.[1-4] Identification of tetramers of excess β-like globin chains in these two syndromes (Hb Bart's γ4: Hb H, β4: and δ4 tetramers) first provided evidence that these conditions result specifically from defects in the production of α-globin chains[5-9] as proposed by Ingram and Stretton.[10]

The genetics of these syndromes were initially more confusing than in other hemoglobinopathies described at that time. This was because the adult carriers of α-thalassemia do not produce large amounts of either Hb H or Hb Bart's. Although the relatives of the affected individuals do not have a readily defined phenotype, it was eventually shown that the offspring of individuals with Hb H disease have raised levels of Hb Bart's (γ_4) in the neonatal period,[11] and the parents of individuals with Hb H disease and the Hb Bart's hydrops fetalis syndrome have mildly hypochromic, microcytic red cell indices[12]; sometimes Hb H inclusions could be demonstrated in occasional red cells.[13] Furthermore, such obligate carriers could be divided into mild (α-thalassemia 2) and severe (α-thalassemia 1) types depending on the level of Hb Bart's present at birth or the degree of abnormality in the hematologic indices. By 1969 it had been shown that Hb H disease results from the inheritance of α-thalassemia 1 × α-thalassemia 2 and the Hb Bart's hydrops fetalis syndrome results from α-thalassemia 1 × α-thalassemia 1.[11,14]

When in 1965 it became possible to measure directly the relative rates of α- and β-globin chain synthesis it was found that there is a progressive decrease in α-chain synthesis in carriers of α-thalassemia, Hb H disease, and the Hb Bart's hydrops fetalis syndrome, respectively.[15-18] This is reflected in the relative levels of α-mRNA activity and quantity in these syndromes.[19-21] However, even using these techniques in combination it still often proved difficult to distinguish unambiguously carriers of α-thalassemia from non-thalassemic individuals.

The molecular genetics of α-thalassemia were further complicated by uncertainty over the number of α genes present in normal individuals (reviewed in reference 22). Based on observations of α-globin structural mutants it was clear that some individuals have four α-chain loci[23,24] but other data suggested that there could be polymorphism for the number of α genes even in apparently normal individuals (reviewed in reference 25). This important question was finally resolved when specific α-, β-, and γ-cDNAs were isolated; solution hybridization studies showed that normal individuals have four α genes, and that the α-thalassemia syndromes usually resulted from the inheritance of 3, 2, 1, or 0 α-globin genes.[26-29]

These findings explained the genetics of α-thalassemia in Southeast Asia but did not explain the findings of a high frequency of α-thalassemia, as judged from cord blood surveys for Hb Bart's in many regions such as Africa and the Middle East, where Hb H disease is relatively uncommon and the Hb Bart's hydrops fetalis syndrome does not occur (reviewed in reference 30). This was ultimately explained by understanding the structural organization of the α-globin genes revealed by blot hybridization analysis.[31] Normal

From the M.R.C. Molecular Haematology Unit, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Headington, Oxford, England.

Submitted September 29, 1988; accepted November 30, 1988.

Address reprint requests to D.R. Higgs, MB, BS, M.R.C. Molecular Haematology Unit, Nuffield Department of Clinical Medicine, Level 7, John Radcliffe Hospital, Headington, Oxford, England OX3 9DU.

0006-4971/89/7305-0014$3.00/0

Table 1. Summary of Hematologic Findings in Individuals With α-Thalassemia

Phenotype	Equivalent Number of Functional α-Genes	Level of Hb Bart's at Birth (%)	Hb H% (Inclusions)	MCV*	MCH*	α/β Globin Chain Synthesis Ratio†
Normal	4	0	0 (none)	85-100	~30	~1.0
α-Thalassemia‡ trait (mild)	3	0-2	0 (rare)	75-85	~26	~0.9
α-Thalassemia‡ trait (severe)	2	2-8	0 (occasional)	65-75	~22	~0.7
Hb H disease	1	10-40	1-40 (many)	60-70	~20	~0.4
Hb Bart's Hydrops Fetalis	0	~80	present (present)	110-120	reduced	0.0

*These values vary considerably depending on the age of the patient and the figures given are a guide to the indices seen in adults. More detail is given in Table 8.

†These values overlap to a considerable degree (see reference 30 and Table 8).

‡The mild and severe forms of α-thalassemia trait have often been referred to as α-thalassemia 2 and α-thalassemia 1, respectively.

individuals have two α genes on each chromosome 16 ($\alpha\alpha/\alpha\alpha$) and carriers for α-thalassaemia have either three ($-\alpha/\alpha\alpha$) or two ($--/\alpha\alpha$) α genes. Thus, the most frequently encountered genotype of Hb H disease is $--/-\alpha$ and Hb Bart's hydrops fetalis is $--/--$.[31-34] In some regions only the $-\alpha$ determinant is frequent, therefore although α-thalassemia is common ($-\alpha/\alpha\alpha$ and $-\alpha/-\alpha$), neither Hb H disease nor the Bart's hydrops fetalis syndrome occur.[35,36]

Thus by 1980 the molecular genetics of α-thalassemia was understood in outline (Table 1). However, subsequent analysis has shown that there are many different molecular defects that underly this simple model. These mutants are now classified as α^+ and α° thalassemia to indicate chromosomes in which there is reduced (α^+) or absent (α°) output from the affected complex. Furthermore, many non-thalassemic variants of the complex are now known that not only serve as useful markers for genetic analysis and anthropological studies but also provide information on which segments of the α-globin complex are required for its correct expression.

LOCALIZATION OF THE HUMAN α-GLOBIN GENE CLUSTER

Genetic studies in a family with α- and β-globin variants, performed in the late 1950s, showed that these globins are encoded by genetically distinct loci.[37] The α- and β-globin gene clusters were initially localized to chromosomes 16 and 11, respectively, by observing the specific pattern of cDNA-DNA hybridization to human × mouse somatic cell hybrids.[38,39] Subsequently, the localization of the α-locus has been refined to the distal segment (p13.1-pter) of the short arm of chromosome 16 by a variety of methods.[40-43] Recently, two independent studies of individuals with unbalanced karyotypes have unequivocally located the α-locus to the Giemsa negative band 16p13.3 at the very tip of chromosome 16,[44,45] although its orientation in the chromosome is not yet known (Fig 1).

The genetic linkage map of 16p is also consistent with this localization. Over 30 genetic markers, localized to 16p13.11-pter, including the locus for adult polycystic kidney disease (PKD1), lie centromeric to α-globin, none lies telomeric.[41,46-48]

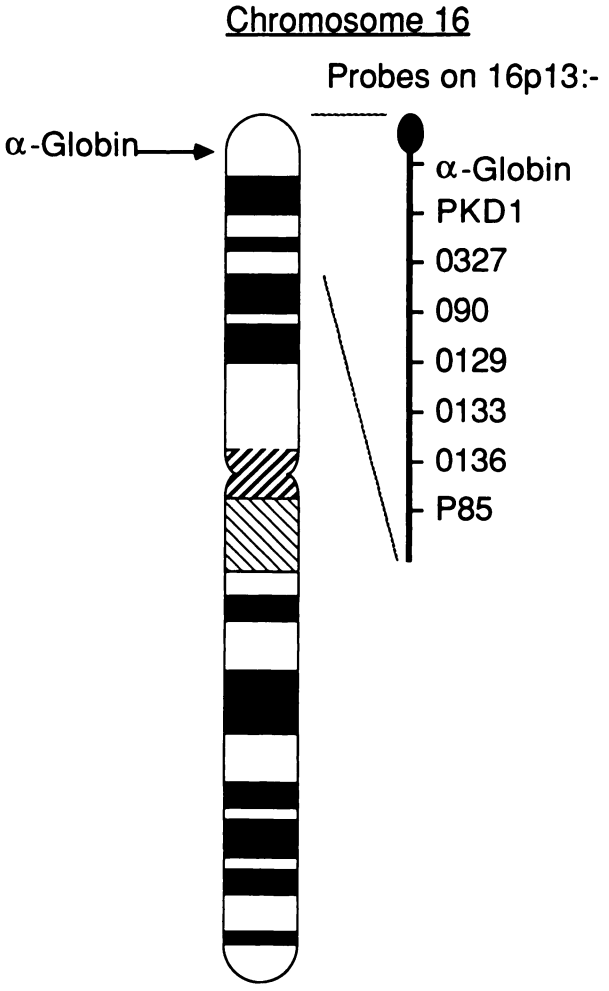

Fig 1. The α-globin genes at 16p13.3[44,45] are shown on the left. The order of chromosome 16 specific probes from this region, as described in reference 46, is shown on the right of the diagram. The cosmids CR1-090, -0136, -0129, -0327, 0133 used to provide the markers indicated, were isolated from a chromosome 16 only mouse/human hybrid. P85 was isolated from a Los Alamos flow-sorted chromosome 16 library. PKD1 denotes the disease locus for adult polycystic kidney disease.

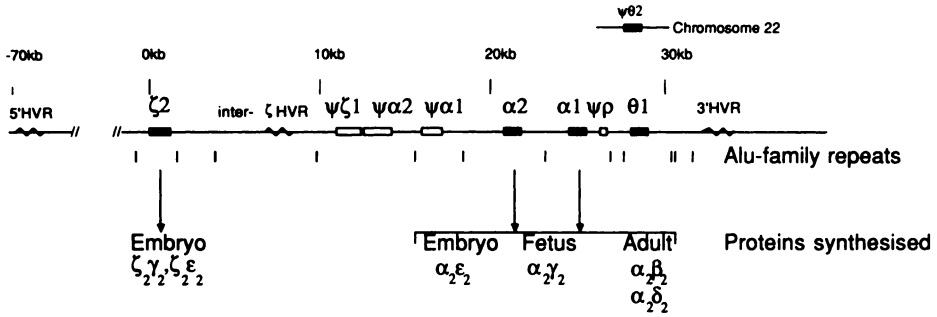

Fig 2. The organization of the α-globin complex. Filled boxes indicate functional genes and open boxes pseudogenes. Position 0 represents the ζ-globin mRNA CAP site. The nonlinked θ2 gene is shown on chromosome 22. Hypervariable regions are denoted by zig-zag lines. The positions of *Alu* family repeats are shown below the complex. The hemoglobins synthesized at each stage of development are indicated below the α complex. Each molecule of hemoglobin comprises a tetramer of two α-like (α- or ζ-) and two β-like (β-, γ-, δ- or ε-) globin chains.

THE STRUCTURAL ORGANIZATION OF THE α-GLOBIN CLUSTER

Evidence from buoyant density fractionation,[49,50] pulsed field gel electrophoresis studies,[51] and DNA sequence analysis (summarized in reference 51) suggests that the 26 kb segment of DNA containing the α-globin genes and the extended cloned segment (150 kb) flanking the cluster are part of a very long (>200 kb) G+C rich DNA segment (isochore). The α-globin cluster and the surrounding DNA have many of the characteristics associated with such regions; they are G+C rich (60%),[51] early replicating (reviewed in reference 52) present within a Giemsa negative band[44,45] (see previous section), are associated with nonmethylated CpG-rich islands[53,54] and contain many Alu-family repeats.[55] In general such regions of the genome are thought to contain a high proportion of "housekeeping genes"; the tissue specific α-globin genes thus constitute an interesting exception. These findings are in contrast to the human β-globin cluster, which has a more normal nucleotide composition (G+C = 39.5% [reviewed in reference 54]), contains no CpG-rich islands, and replicates late in nonhematopoietic cells (reviewed in reference 52).

The α-globin cluster includes the duplicated α genes (α2 and α1), an embryonic α-like gene (ζ2), three pseudogenes (ψζ1, ψα2, ψα1), and a gene of undetermined function (θ1) arranged in the order 5′ – ζ2 – ψζ1 – ψα2 – ψα1 – α2 – α1 – θ1 – 3′ (Fig 2).[56-58] Recently, a pseudo gene for the Ro family of small cytoplasmic RNAs was identified downstream of α1[59] and a truncated, processed copy of the θ gene family (ψθ2) was found on chromosome 22.[60]

Several regions of the cluster contain tandemly repeated segments of DNA (minisatellites). They were first identified as hypervariable regions (HVRs) located at the 3′ end of the complex (α-globin 3′HVR), between the ζ2 and ψζ1 genes (interzeta-HVR) and within the introns (IVS 1 and IVS 2) of the ζ-like genes (ζ-intron HVRs),[61,62] (summarized in Fig 2 and Table 2). We have recently identified a further hypervariable region approximately 70 kb upstream of ζ2 (called the α-globin 5′HVR.[63] The structure of these regions and their interrelationships are shown in Table 2. It is of interest that such regions may be particularly frequent at the telomeres of human chromosomes.

EVOLUTION OF THE α-GLOBIN FAMILY

The ancestral α- and β-globin genes started to diverge from each other approximately 500 million years (MY) ago.[67] Each gene family has subsequently evolved through a series of gene duplications modified by a variety of genetic processes (divergence, deletion, gene conversion, and retroposition). Comparisons of the α-globin cluster from several species suggest that the ancestral α-globin family in mammals (85 MY ago) included a ζ-gene, duplicated α-genes, and a θ gene (proto ζ – 5′α – 3′α – θ).[68] It has been calculated (Fig 3) that the ζ/α divergence took place approximately 400 MY ago[62] and the θ/α divergence occurred 260 to 280 MY ago,[58] although, because θ1 may not serve as a globin gene

Table 2. Summary of the Repeat Elements Within the α-Globin Hypervariable Regions (HVRs)

Name	Sequence	Unit Size (bp)	Number of Repeats (Range)	Enzymes Commonly Used to Detect Alleles	Reference
Core	GNGGGGNACAG				64
5′HVR	GGGGAGCATTCAGGAGGCCTTCCCGGAGGTAGGGTGGTGGGAAGAAGGGG-TCAGCGT	57	5-55	*Rsa*I, *Dde*I	63
ζHVR (IV↑S1)	RGGAGGGG-ACAGTG	14	12-30	*Sac*I, *Pvu*II	62
IZ-HVR	TGTGAGGGTGCCCGGGACGGCTTGTGGGGCACAG-GT	36	12-65	*Sac*I, *Sau*96I	61, 65, 66
3′HVR	ACACGGGGGGAACAGCG	17	70-450	*Pvu*II, *Rsa*I, *Hinf*I	61, 64
ζHVR (IV↑S2)	CGGGG	5	35-52	*Sac*I, *Pvu*II	62

Each hypervariable region comprises a segment of DNA containing a variable number of repeats of unit sizes ranging from 5 to 57 bp as indicated.

GENES COMPARED	SUBSTITUTIONS PER SITE		% MISMATCH OF AMINO ACIDS			
	Replacement	Silent	Exon 1	Exon 2	Exon 3	Total
α2 vs α1	0.00	0.00	0	0	0	0
α2 vs ζ2	0.38	0.60	66	37	33	42
ζ2 vs ψζ1	0.01	0.00	6	0	2	2
θ1 vs α2	0.28	0.67	55	34	35	40
θ1 vs ζ2	0.45	0.72	71	41	40	47
ψα1 vs α2	0.22	0.48	38	47	50	46
ψα2 vs α2	0.54	0.66	69	60	48	58
ψα2 vs ζ2	0.61	0.94	91	57	48	62

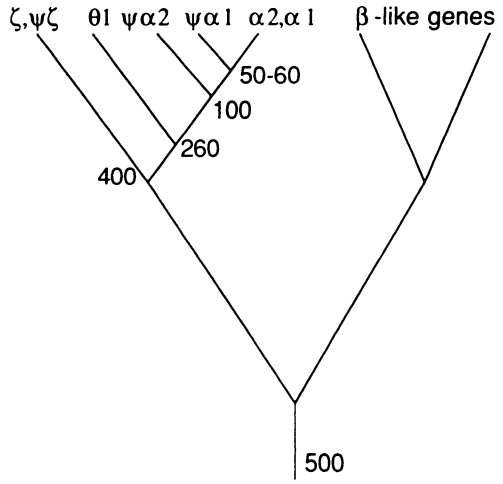

Fig 3. Comparison of the coding regions of the α-like genes (above). Substitutions per site and percent mismatch of amino acids in the coding regions were calculated as previously described. The data for most of this table are derived from references 57 and 58. Below, the proposed timing of the evolutionary events in the α-globin gene family are summarized. The dates for these divergence times are from references 57, 58, 62, 67, and 73.

and may have evolved rapidly to fulfil another function, the latter may be an overestimate.[69]

At least three further duplications are required to account for the present day human α-gene family. Mapping of the α-globin cluster in apes shows that the ζ-gene duplication predates the divergence of human and chimpanzee (estimated to be ~5 MY ago.[70] Provisional data based on comparisons of the horse and human ζ-globin regions suggests that it may even have predated the separation of the equid and primate lineages.[71] However, more precise estimates of the timing of the duplication event cannot yet be made since there are insufficient data from other species and direct comparison of the human ζ2 and ψζ1 genes (Fig 3 and reference 62) shows that they have recently undergone a gene conversion event (see below) that has eradicated any DNA sequence differences by which to calculate the timing of the duplication.

The ψα1-α2-α1 cluster is derived from the proto 5′α-3′α genes. Sequence comparisons of orthologous α-genes in humans and galago show that the nonfunctional pseudogene ψα1 is derived from the proto 5′α gene and α1 from proto 3′α gene.[72] Thus, unlike the duplicated α genes of many mammals that are derived from the proto 5′ α and proto 3′ α genes, the human α2/α1 pair results from a further α gene duplication of the proto 3′ α gene that occured after divergence from the prosimian lineage (52 to 72 MY ago). The ψα1 (proto 5′α) started to diverge from the functional α1 and α2 genes 60 MY ago and was inactivated 45 MY ago[73] after the divergence of prosimians and higher primates, but before the divergence of monkeys and great apes.

The evolutionary history of the ψα2 gene has not yet been elucidated. However, the extensive differences between this nonfunctional pseudogene and the functional α gene (Fig 3) suggest that they have been diverging for a considerable time,[57] possibly since before the mammalian radiation.

DNA sequence analysis of the human α1 and α2 genes shows that they have remained virtually identical to each other during evolution[74,75] despite the fact that they have both diverged as expected from the α-globin sequence of other species. Sequence homology between the ζ2 and ψζ1 gene has been similarly maintained.[62] This type of evolution-in-tandem is called concerted evolution, which occurs through two related genetic mechanisms, crossover fixation and gene conversion (reviewed with respect to the α-genes in reference 76). The intermediates of crossover fixation (chromosomes with single [−α, −ζ] or triplicated [ααα, ζζζ] genes) are frequently observed in extant human populations (see below).

The structure of the human α-globin cluster has been further modified during evolution through insertion by retroposition of many copies of the *Alu* family of repetitive elements[55,72,75] (Fig 2). The origin of the tandemly repeated segments of DNA within the minisatellite regions is not clear. However, it is interesting that a sequence similar to the 14 bp repeat present in the ζ intron HVR is also present in the ζ intron of the goat[77] suggesting that such structures may have preceeded the mammalian radiation.

THE STRUCTURE AND EXPRESSION OF INDIVIDUAL α-LIKE GLOBIN GENES

The α-globin gene family has evolved through a series of gene duplications and sequence divergence such that now, although there are similarities in polypeptide structure (Fig 3), the functional α and ζ genes show only 58% homology in their 141 amino acids. In contrast the α1 and α2 genes are highly homologous (Fig 3), encode identical proteins, and only differ within IVS2 and their 3′ noncoding regions (17% divergence).[74] Curiously, all of the α-like genes share a surprisingly high G+C content, which is associated with a bias in codon usage.[78] Similar extremes in base composition and codon usage have been noted for other genes that, like the α-gene family, lie within GC-rich isochores.[49]

Comparisons of many globin genes within and between species demonstrate that some structural features have been conserved.[79,80] In general, the globin genes are compact (1 to 2 kb, Table 3) and are divided into three exons (protein coding regions) by two noncoding intervening sequences (IVS or introns). The positions of the introns (Fig 4) are conserved in all globin genes and may reflect the early (>500 MY ago) evolutionary events that brought together the functional domains of the protein.[81] For example, exon 2

Table 3. Consensus Sequences in the α-Like Globin Genes

Gene (Approximate size bp)	Upstream Elements* −40 to −200 SP1†	"CAAT" Box (position with respect to CAP site)	−30 "ATA" Box (position with respect to CAP site)	Exon 1 Length of 5′ UT‡ (N)	Initiator	IVS I 5′ Donor	Length (N)	3′ Acceptor	IVS II 5′ Donor	Length (N)	3′ Acceptor	Exon 3 Terminator	Length of 3′ UT (N)	Poly(A) add Signal	Poly(A) add Site (Distance from signal)	References
					CCRCCATGG#	↓ C/AAGGTRAGT#		↓ (Y)NYAGG#	↓ C/AAGGTRAGT#		↓ (Y)NYAGG#	TAA TAG TGA		AATAAA	(10-30)CA	
α1 (850)	+	AGCCAATGA(−69)	CATAAAC(−28)	40	CCACCATGG	GAGGTGAGG	117	GCAGG	AAGGTGAGC	149	ACAGC	TAA	113	AATAAA	(18) CA	74
α2 (850)	+	AGCCAATGA(−69)	CATAAAC(−28)	40	CCACCATGG	GAGGTGAGG	117	GCAGG	AAGGTGAGC	142	ACAGC	TAA	113	AATAAA	(18) CA	74
ψα1 (850)		AGCCAATGA(−ND)	CATAAGA(−ND)	ND	CTGCCGTGG	GAGGCAAGA	127	GAAGC	AGGGAGAGC	134	ACAGT	TAA	ND	AATGAA	(18) CA	73
ψα2 (1,500)	−	−	−	ND	GTCCGGGGG	GGTCGGAGG	138§	GCAGG	CCGGTGAGG	107	GCAGC	TGA	ND	AGCAAA	(18) GT	57
ψζ1 (ζ1) (2,200)	0	GACCAATGG(−61)	TATATAA(−29)	58	CCGCCATGT	GAGGTGAGT	1,261‖	CCAGG	AAGGTGCGC	340‖	CCAGC	TGA	102	AATAAA	(17) AG	62
ζ2 (2,200)	0	GACCAATGG(−61)	TATATAA(−29)	58	CCGCCATGT	GAGGTGAGT	886‖	CCAGG	AAGGTGCGC	239‖	CCAGC	TGA	108	AATAAA	(17) CA	62
θ1 (700)	+	GGCCAATTT (ND)	CATATAG (ND)	ND	CGGGGATGG	AAGGTGCGG	84	GCAGG	CAGGTGAGC	109	GCAGC	TGA	72	AGTAAA	(18) CA¶	58

Abbreviations: ND, not determined; UT, untranslated; bp, base pairs; Poly A add, poly (A) addition.

* Includes both constitutive and regulating elements (see reference 96 for review). Some regulatory elements may also exist 3′ and 5′ of the genes they interact with. The function of these upstream elements may be further modified by enhancer elements.

† Potential SP1 binding sites.

‡ Includes ATG initiation codon.

§ Two additional inserts present.

‖ The site of each of these introns may range as set out in references 66 and 82.

¶ Putative poly(A) addition site.

In the consensus sequences Y indicates any pyrimidine and R denotes any purine.

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.