Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 3381–3385, May 1986
Genetics

The *qa* repressor gene of *Neurospora crassa*: Wild-type and mutant nucleotide sequences

(repressor protein/repressor mutants/eukaryotic gene regulation/gene cluster)

LAYNE HUIET* AND NORMAN H. GILES

Department of Genetics, University of Georgia, Athens, GA 30602

Contributed by Norman H. Giles, December 31, 1985

ABSTRACT The *qa-1S* gene, one of two regulatory genes in the *qa* gene cluster of *Neurospora crassa*, encodes the *qa* repressor. The *qa-1S* gene together with the *qa-1F* gene, which encodes the qa activator protein, control the expression of all seven *qa* genes, including those encoding the inducible enzymes responsible for the utilization of quinic acid as a carbon source. The nucleotide sequence of the *qa-1S* gene and its flanking regions has been determined. The deduced coding sequence for the qa-1S protein encodes 918 amino acids with a calculated molecular weight of 100,650 and is interrupted by a single 66-base-pair intervening sequence. Both constitutive and noninducible mutants occur in the *qa-1S* gene and two different mutations of each type have been cloned and sequenced. All four mutations occur within the predicted coding region of the *qa-1S* gene. This result strongly supports the hypothesis that the *qa-1S* gene encodes a repressor. All four mutations are located within codons for the last 300 amino acids of the qa-1S protein. The mutations in three of the mutants involve amino acid substitutions, while the fourth mutant, which has a constitutive phenotype, contains a frameshift mutation. The two constitutive mutations occur in the most distal region of the gene, possibly implicating the COOH-terminal region of the *qa* repressor in binding to its target. The two noninducible mutations occur in a region proximal to the constitutive mutations, possibly implicating this region of the *qa* repressor in binding the inducer.

The utilization of quinic acid as a carbon source by *Neurospora crassa* is controlled by a cluster of five structural genes and two regulatory genes (1). Both genetic and molecular analyses indicate that regulation of the *qa* gene cluster occurs primarily at the level of transcription. One of the regulatory genes, *qa-1F*, is inferred to encode an activator protein that is required for the induced transcription of all the *qa* genes, including itself and the other regulatory gene, *qa-1S*. The *qa-1S* gene is proposed to encode a negative regulatory protein, which interacts with quinic acid, and in the absence of quinic acid, blocks the action of *qa-1F*. The negative role of the *qa-1S* gene in quinic acid regulation is based on the phenotypes of two types of *qa-1S* mutants. One group contains mutants that are noninducible and dominant ($qa\text{-}1S^-$), while the other group contains mutants that are constitutive and recessive ($qa\text{-}1S^c$).

The two regulatory genes have been previously localized to one end of the 17.3-kilobase (kb) *qa* cluster adjacent to one another. This paper reports the complete nucleotide sequence of the region containing the wild-type *qa-1S* gene and its flanking regions from strain 74A. From these data, it is deduced that the qa-1S protein contains 918 amino acid residues with an approximate molecular weight of 100,650.

The *qa-1S* genes from two constitutive mutants have also been sequenced, revealing that the mutations are located in the distal region of the gene and suggesting that the COOH-terminal region of the repressor protein may be involved in binding to its target, which may be the activator protein (1). The *qa-1S* genes from two noninducible mutants had amino acid substitutions just upstream of the constitutive mutations, possibly implicating this region in binding the inducer. Sequence analysis of these mutations also revealed a number of polymorphisms, primarily in the noncoding region of this gene.

MATERIALS AND METHODS

Strains Used. The nucleotide sequence determination of the *qa-1S* gene was carried out on cloned wild-type (74-OR23-1A) DNA (2). The two noninducible mutations that have been cloned and sequenced are M141 and M105^{ts}. Both mutants were isolated in the same filtration–concentration experiment after ultraviolet irradiation (mutant 105^{ts}) or ethyl methanesulfonate treatment (mutant 141) from an *arom-9* mutant in wild-type 74-OR8-1a background (3). The 105^{ts} mutant is unusual in that it behaves as a noninducible mutant at 25°C but as a constitutive mutant at 35°C (4). The homokaryotic isolates of the two constitutive mutations sequenced were 105-R12-1.5 (mutant 105^c) and A1-1.4 (5, 6). Mutant 105^c was derived as a secondary constitutive "revertant" of mutant 105^{ts} after ultraviolet irradiation and selection of rapidly growing colonies from conidia plated on quinic acid as the sole carbon source at 25°C (5). Mutant A1 was induced by ultraviolet irradiation of wild-type 74-OR23-1A conidia and recovered by a special selection technique (6). Isolates 105-R12-1.5 (105^c) and A1-1.4 were obtained as homokaryotic strains from backcrosses to the closely linked marker gene *met-7* (5, 6). Both homokaryotic isolates M105^c and A1-1.4 are constitutive at both 25°C and 35°C.

Genomic Cloning. The cloning of the *qa* cluster from 74-OR23-1A wild type has been reported (2). DNA from the four mutant strains was isolated as described (7). The DNA was digested with *Bgl* II and then fractionated on a glycerol gradient. DNA·DNA hybridization indicated that the *qa-1S* region was contained within a 13-kb *Bgl* II fragment in mutants 105^c, 105^{ts}, and 141, and within a 24-kb *Bgl* II fragment in A1-1.4. Fractions containing these 105^c, 105^{ts}, and 141 *Bgl* II fragments were cloned into the λ vector EMBL4 by standard procedures (8). λ arms and packaging extracts were a generous gift of James Baum and Robert Geever. λ clones containing the *qa-1S* region were isolated by hybridization to the previously isolated wild-type *qa-1S* gene. The 24-kb *Bgl* II fragment from mutant A1-1.4 was further digested with *Bcl* I to yield a smaller fragment, which could

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase(s); bp, base pair(s).

*Present address: Commonwealth Scientific and Industrial Research Organization Division of Plant Industry, P. O. Box 1600, Canberra City, Australian Capital Territory 2601, Australia.

be cloned into the λ vector. Fragments containing the *qa-1S* coding region were subcloned into pUC plasmids for DNA sequence analysis.

DNA Sequencing. The DNA sequence determination was carried out by using the method of Maxam and Gilbert (9). The only alteration was for the G+A reaction. We used 25 μl of 88% formic acid (Baker) rather than 1.0 M piperidine formate. All the fragments, except two, were labeled at the 5′ end using T4 polynucleotide kinase (New England Nuclear). Fragments with blunt ends or 5′ recessed ends were labeled by first digesting the fragments for 10 min with 5 units of T4 DNA polymerase (Bethesda Research Labs) with no nucleoside triphosphates present. The fragments were then treated with calf intestinal alkaline phosphatase (a gift of Frank Buxton) and subsequently phosphorylated with polynucleotide kinase. In all cases, fragments labeled at a single end were generated by digestion with a second restriction enzyme and were purified by fractionation on a 5% acrylamide/10% glycerol gel.

S1 Nuclease Mapping. S1 nuclease mapping was carried out using a modification of the method of Berk and Sharp (10, 11). Uniquely ^{32}P-end-labeled DNA fragments were hybridized to 5–10 μg of *Neurospora* poly(A)$^+$ mRNA in 10 μl of 80% formamide/80 mM Pipes, pH 6.8/0.4 M NaCl/10 mM EDTA, for 6–24 hr at 45°C. After hybridization, the DNA·RNA hybrids were digested with 50–150 units of S1 endonuclease (Sigma) in 250 mM NaCl/60 mM Na acetate, pH 4.5/1 mM $ZnSO_4$/alkali denatured calf thymus DNA (5 μg/ml) for 30 min at 37°C. The reaction was stopped by adding to a final concentration 10 mM EDTA/50 mM Tris·HCl, pH 9.0/1 μg of tRNA. The digestion products were then precipitated with ethanol and fractionated on an 8% acrylamide/7 M urea sequencing gel.

RNA Sequencing. A *Sal* I/*Aha* II DNA fragment 5′-end-labeled only at the *Sal* I site was hybridized to 10 μg of *Neurospora* poly(A)$^+$ mRNA isolated from wild-type induced mycelia under conditions described above, except hybridization was at 40°C. The RNA·DNA hybrids were recovered by ethanol precipitation, redissolved in 12.5 mM Tris·HCl, pH 8.3/1.25 mM EDTA/125 mM NaCl, then split into five 25-μl reaction mixtures of 66 mM Tris·HCl, pH 8.3/55 mM NaCl/10 mM $MgCl_2$/10 mM dithiothreitol (final concentrations). One reaction (N) also contained 200 μM each of dNTPs. The remaining four reactions (G, A, T, C) contained 1.6 μM of one ddNTP, 10 μM of the corresponding dNTP, and 20 μM of the remaining dNTPs. Avian myeloblastosis virus reverse transcriptase (10 units) (Seikagaku America, St. Petersburg, FL) was added to each reaction for 20 min at 40°C. The reaction products were recovered by ethanol precipitation and were fractionated on a sequencing gel (12).

RESULTS

DNA Sequence Analysis of the Wild-Type *qa-1S* Gene. The approximate location of the *qa-1S* gene and its direction of transcription had previously been determined by *Neurospora* transformation and DNA·RNA hybridization analyses (7). The results of these experiments indicated that the *qa-1S* gene was contained within a 5.0-kb *Eco*RI restriction fragment in plasmid pMSK366. A detailed restriction map was generated for the region, and the DNA sequence was determined by the method of Maxam and Gilbert (9). Over 65% of the 4.0-kb sequence was determined on both strands. All labeling sites were sequenced through from adjoining sites and overlaps were at least 20 base pairs (bp). A 3522-bp sequence of this region from *Hpa* I to *Bgl* II is shown in Fig. 1. This region includes a potential coding region of 918 codons interrupted by a 66-bp intervening sequence (see below) as well as ≈600 bp of 5′ flanking sequences and 100 bp of the 3′ flanking sequences. The coding region would encode a protein of M_r 100,650.

RNA Mapping and Sequencing. S1 nuclease digestion of RNA·DNA hybrids was used to map the 5′ end of the gene and to determine the location(s) of any intervening sequences. Previous S1 nuclease experiments used to determine the direction of transcription indicated that the 5′ end of the mRNA or an intervening sequence boundary was ≈175 bp from a *Sal* I site (7). Further S1 nuclease experiments indicated that this was a 3′ exon/intron boundary. To determine the 5′ end of the intron, a 570-bp *Sph* I/*Sal* II fragment was 3′-end-labeled at the *Sph* I cleavage site and hybridized to poly(A)$^+$ RNA. A single band ≈340 bp long was protected from S1 nuclease digestion, placing the 5′ end of the intron 340 bp downstream of the *Sph* I site (data not shown). The rest of the *qa-1S* gene appeared to have no other intervening sequences.

To determine the precise exon/intron boundaries, the sequence of the mRNA was determined by primer extension with reverse transcriptase in the presence of dideoxynucleotides (described in *Materials and Methods*). The results (Fig. 2) revealed that the intervening sequence was 66 bp long with a donor site of GC rather than GT. Splicing of this intron joins a 75-residue open reading frame with an 843-residue open reading frame.

RNA mapping of the 5′ end(s) of the *qa-1S* transcript(s) was determined by S1 nuclease digestion of DNA·RNA hybrids. A 490-bp *Hpa* II/*Eco*RI DNA fragment 5′-end-labeled at the *Hpa* II site was used as the DNA probe. The results are shown in Fig. 3. In lanes d and e, respectively, are G+A, and T+C sequencing products, while lane c shows the protected fragments. A range of fragments from 98 to 145 nucleotides were protected. The most intense bands of 98, 144, and 145 nucleotides indicate that there are two major sites of initiation. Results using a *Taq* I-end-labeled probe, which was 12-bp shorter than the *Hpa* II probe, were identical (data not shown). The 3′-end-mapping experiments will be published elsewhere (unpublished data).

Nucleotide Sequences of *qa-1S* Mutants. The *qa-1S* genes from two noninducible mutants, 141 and 105^{ts}, and from two constitutive mutants, 105^c and A1-1.4, were also cloned and completely sequenced. The results, summarized in Fig. 4, indicate that mutant 141 contains a transition mutation changing glycine to aspartic acid at amino acid residue 627. (This change creates a *Sal* I restriction site.) The second mutant, 105^{ts}, which is noninducible at 25°C but constitutive at 35°C, results from a transversion mutation changing asparagine to tyrosine at amino acid 743. One of the two constitutive mutants that were cloned and sequenced was derived as an ultraviolet-induced constitutive revertant at 25°C of the noninducible mutant 105^{ts}. This mutant (105c) is constitutive at both 25°C and 35°C (4). The nucleotide sequence data confirm that mutant 105^c is a double mutant. It contains both the original mutation present in the parental strain 105^{ts} (at amino acid 743) and an additional transition mutation, which changes a proline to a leucine at amino acid 791. The second constitutive mutant analyzed, A1-1.4, derived from wild-type 74-OR23-1A, is constitutive at both 25°C and 35°C (6). The mutation in A1-1.4 consists of two single-base-pair deletions in the adjacent codons, 789 and 790. A G·C base pair has been deleted from the third position of codon 789, while a T·A base pair has been deleted from the middle position of codon 790. The resulting frameshift mutation generates the nonsense codon UGA at residue 813. The relative positions in the *qa-1S* gene and the nucleotide and amino acid changes occurring in the four mutations sequenced are indicated in Fig. 4. Approximately 60% of the *qa-1S* gene was sequenced for mutants 105^{ts}, 141, and A1-1.4, while 90% of the gene was sequenced in mutant 105^c.

```
Hpa I
GTTAACGTGA AAACACCCGC CATGGTTACA GATGACTGTA CCTAGGTAAA TACTGGAAGG TTTGACATGG
-535
AGGCTTGGAC GTATATCACC TCATATGAGA GAGGCAGGCT CCGTCCTAGG ACAGTACCTA CCTCTCTACT
-465          . .**                                   .. ..  ...  ...*.
TAGGTAATTA AATGCCCGCC CTCCTTCCAT TCCGTGCTTG CCGCATGTAC ATACCTACAG ACCTCGCTGT
-395
ACCTTATCCA AGTGCAAGCA AGGTGTGAAT ATAAGGGGGT TTCAGTTTCT TTGCCGCAAG GCATTTATTA
  Taq I       HpaII                          Sph I
ACTCGAAGGC ACACCGGGCG GGCTGGCATC TCTCAGCTTG CATGCTAGAG GACGCTGTTT CCTTCCCGAC
-255
CTCAAGCAAC CGACCGCTAC TGTACTGCGA CAAGCACAGA ACAAAACTCA TGAAGCCTCC CCCTATTTTC
-185
TGATCGGATG AGTGATTCTC CAGACATACC TCTAGTTTCC AATTTGCAGA ATTGACATTT GCCATGGCGC
-115
ATGCGAGAAT GCCGATGCAC AACTTAAGTA TGCGCACAAC TAGGTAGGCA CCCGGCTTTT CACTCCCCGC
- 45
TGCGATAGAG ACACCACTAC CACCACCCCC GCCGCCGCCG CCATC   ATG AAC ACC ATC CCG
                                                    MET Asn Thr Ile Pro     5

GCA CGC CAT GTC GGG GAT GTC GCC GCC CGC GAT CCT CTA CCT CTA CCG CAC ATA
Ala Arg His Val Gly Asp Val Ala Ala Arg Asp Pro Leu Pro Leu Pro His Ile    23

TCA TCC TCC GTC GCC AGC GGC ATG AAG CGT TCC TTC GCA ACC ATG GCC ATG CTC
Ser Ser Ser Val Ala Ser Gly Met Lys Arg Ser Phe Ala Thr Met Ala Met Leu    41

TAC AAC GAC ACT GGC AAC AGC AAC GAT GTC GGT GCC CAT GCC AGG CGA CCA CCA
Tyr Asn Asp Thr Gly Asn Ser Asn Asp Val Gly Ala His Ala Arg Arg Pro Pro    59

CGA ACC CTC TCC AAT AGT CGG AGC ACT TCC GCC CAC AGA GTA CCT CTA G GCACGT
Arg Thr Leu Ser Asn Ser Arg Ser Thr Ser Ala His Arg Val Pro Leu G          75

CTGTATTGCTTCGCGATTTTTACCTCAAGTCTTGATCGTACTAACAGTCAGTCTCACCAG GC TCT TGG
                                                                ly Ser Trp   78
  AhaII
TCG GCG CCC AAT TCC CCG CCC CGC CGT GCG CTG CCG CAT CAT CCC ATC ACC GCC
Ser Ala Pro Asn Ser Pro Pro Arg Arg Ala Leu Pro His His Pro Ile Thr Ala    96

AGC TTC GAT CCC GAT GCT TCC ATT GTT ATC GCC GGC ATT CGT GGC GCT GGC AAG
Ser Phe Asp Pro Asp Ala Ser Ile Val Ile Ala Gly Ile Arg Gly Ala Gly Lys   114
                                                         SalI
TCT ACG CTG GCC ATC ATG GCA TCT ACC GCC ATG AAG CGC AAG ATA GTC GAC CTG
Ser Thr Leu Ala Ile Met Ala Ser Thr Ala Met Lys Arg Lys Ile Val Asp Leu   132

GAA TCC GAG TTC CAT CAT CTT ACT GGC TTG TCT AGT TCC AGC TAC AAG AAG ACA
Glu Ser Glu Phe His His Leu Thr Gly Leu Ser Ser Ser Ser Tyr Lys Lys Thr   150

CAC GGC CCG GTC GAC TAT GGG AGG CGC CAG ATC GCC ATC TTG CAG AAC ATC TTG
His Gly Pro Val Asp Tyr Gly Arg Arg Gln Ile Ala Ile Leu Gln Asn Ile Leu   168

AAT CTG CAC AGG ACC CGC GCC ATT CTC GTC TGC TCT TGG CTG GAG CGG GAT GTG
Asn Leu His Arg Thr Arg Ala Ile Leu Val Cys Ser Trp Leu Glu Arg Asp Val   186

CAG GCC ATG TTG CAG GAT TTC AGC GTG TCT AAT CCT GTC ATT TAC GTT CTG CGC
Gln Ala Met Leu Gln Asp Phe Ser Val Ser Asn Pro Val Ile Tyr Val Leu Arg   204

GAT GCC AAA GCT ATC GAG GCC CAT CTG AAG GGA TAC GAC AAG TCC AAA GTT GGC
Asp Ala Lys Ala Ile Glu Ala His Leu Lys Gly Tyr Asp Lys Ser Lys Val Gly   222

ACC CTC CTT GAT GCC ACC AGT ACT GTC CTT CGC CGC TGC ACC CGT TTC GAG TTC
Thr Leu Leu Asp Ala Thr Ser Thr Val Leu Arg Arg Cys Thr Arg Phe Glu Phe   240

TTC AAC GTC TCC GAA GAG AAC CTG GAC ACC CAC TCC GCT TCA ACA TCA CCA CCT
Phe Asn Val Ser Glu Glu Asn Leu Asp Thr His Ser Ala Ser Thr Ser Pro Pro   258

GCT GTT CCG GAC CAG CGG CAT ACC GCG CCG TAT CTA ACG CTT AAA CGA GCC GAG
Ala Val Pro Asp Gln Arg His Thr Ala Pro Tyr Leu Thr Leu Lys Arg Ala Glu   276

CGC CAC TTC CTC AAA TTT CTC TCC TTG ATT CTA CCC AAG GGG ACC ATA CCT TTT
Arg His Phe Leu Lys Phe Leu Ser Leu Ile Leu Pro Lys Gly Thr Ile Pro Phe   294

GTC GAG TCC GCC TTT CCC CTG GCT TCC GTC CCC GTT GAA CAG CGC CGC TTT ACC
Val Glu Ser Ala Phe Pro Leu Ala Ser Val Pro Val Glu Gln Arg Arg Phe Thr   312

TAC GCC CTC GCC TTG CCT GTA TCT GCC TTG CTC GAC AAA GGC GTC GAT ATC CAA
Tyr Ala Leu Ala Leu Pro Val Ser Ala Leu Leu Asp Lys Gly Val Asp Ile Gln   330

GAG CTT GAT GTC GGT GTA GAC GCA ATC GAG ATC ATT GTA GAC GAT CTT GCA ACG
Glu Leu Asp Val Gly Val Asp Ala Ile Glu Ile Ile Val Asp Asp Leu Ala Thr   348

AGC GAA TCC GGC CCA ACG AGC CCC TTG GGT CTT GCG CCC CAC CGA GCG AGC GAG
Ser Glu Ser Gly Pro Thr Ser Pro Leu Gly Leu Ala Pro His Arg Ala Ser Glu   366

ATC AGT CGT GTT GTA GGC GAA ATC AGG AGG GAC ACA GTG ATC CCC ATC ATT CTG
Ile Ser Arg Val Val Gly Glu Ile Arg Arg Asp Thr Val Ile Pro Ile Ile Leu   384

CAC GTG GTC TTT CCA GAA AGA GCG CTA TAT GAA GAA GCT CTG CTC GCG CTT TAC
His Val Val Phe Pro Glu Arg Ala Leu Tyr Glu Glu Ala Leu Leu Ala Leu Tyr   402

ATG ACT TAC CTG AAC CAT GCC TTA AGG CTT GCG CCA GAT TAT CTT ACG GTC GAT
Met Thr Tyr Leu Asn His Ala Leu Arg Leu Ala Pro Asp Tyr Leu Thr Val Asp   420

CTG GGG CTC GAT TCT GGC TTG CTT GGG CAA CTA ACC ACC GTT CAA GGA ACC ACC
Leu Gly Leu Asp Ser Gly Leu Leu Gly Gln Leu Thr Thr Val Gln Gly Thr Thr   438

AAG GTC ATC GGC AAT AAA CAA CTT GCA GAG GTC AAT TCG CCG CGC TGG GGG GAT
Lys Val Ile Gly Asn Lys Gln Leu Ala Glu Val Asn Ser Pro Arg Trp Gly Asp   456

CCA TCT TGG TTA CAA GCC TAT GAA AAG GCC CAG AAT ACA GGA TGC GAC TTG GTA
Pro Ser Trp Leu Gln Ala Tyr Glu Lys Ala Gln Asn Thr Gly Cys Asp Leu Val   474

AGG TTG ACC AGA CCG GCT TCA AAT CCC CGG GAC AAC ACA GAC ATT CGG CAG TTC
Arg Leu Thr Arg Pro Ala Ser Asn Pro Arg Asp Asn Thr Asp Ile Arg Gln Phe   492

CAC GTT GCT GTA GAG GCC GTC GGG GGT CCA AGG CTC CCA TTT ATT GCT TAC AAC
His Val Ala Val Glu Ala Val Gly Gly Pro Arg Leu Pro Phe Ile Ala Tyr Asn   510

ACA GGA CGC CTA GGT CGG ACA TCG ATG TGT TTT AAC GAG ATC CTG ACT CCA GTT
Thr Gly Arg Leu Gly Arg Thr Ser Met Cys Phe Asn Glu Ile Leu Thr Pro Val   528

ACA CCA GTG CCT ACC AAG GAG GAT GCA ATC GGG CTC CGC AAT CCA GCC CAT CGC
Thr Pro Val Pro Thr Lys Glu Asp Ala Ile Gly Leu Arg Asn Pro Ala His Arg   546

TAT CTC CAG CCT CCG CTC ACG GCT CTG GAA GCA ACA CAG GCT CTC TAC TCG GCA
Tyr Leu Gln Pro Pro Leu Thr Ala Leu Glu Ala Thr Gln Ala Leu Tyr Ser Ala   564

TTT GTC CAC GAC CCA ATG AAG CTG TAT GTC TTT GGC GCA AAT GTG GGA TAT AGC
Phe Val His Asp Pro Met Lys Leu Tyr Val Phe Gly Ala Asn Val Gly Tyr Ser   582

TTG TCC CCA GCC ATG CAC AAT GCC GCA CTC AAG GCC TGT GGC ATT CCA CAC CAT
Leu Ser Pro Ala Met His Asn Ala Ala Leu Lys Ala Cys Gly Ile Pro His His   600

TAC AAG CCC CTT TCC ACA GCA AAC ATC GGG ACT TTG CGC GAG GTT ATC AGC GAT
Tyr Lys Pro Leu Ser Thr Ala Asn Ile Gly Thr Leu Arg Glu Val Ile Ser Asp   618

CCG CAG TTT GCT GGA GCC TCG GTC GGC CTG CCG TTT AAG GTG GAA ATC ATC AGC
Pro Gln Phe Ala Gly Ala Ser Val Gly Leu Pro Phe Lys Val Glu Ile Ile Ser   636

CTC ACA CAC TCG CTG AGC CGG CAC GCG AAA GCC ATC GGA GCC GTC AAC ACC TTG
Leu Thr His Ser Leu Ser Arg His Ala Lys Ala Ile Gly Ala Val Asn Thr Leu   654

ATT CCG GTA CGA CAC CTT ACC GCG GAC GGT GGA ATA CCG GAC GAG GTG TCC ATG
Ile Pro Val Arg His Leu Thr Ala Asp Gly Gly Ile Pro Asp Glu Val Ser Met   672

TTC AAC AAT ATC AGC CAA GCC GGC GCT GTC AGA GCT CTC TAC GGC GAG AAC ACG
Phe Asn Asn Ile Ser Gln Ala Gly Ala Val Arg Ala Leu Tyr Gly Glu Asn Thr   690

GAT TGG ATT GGT ATC CGA GCC TGC CTT CGC CGC GGT TTA TCG CCC GCC AAT GCC
Asp Trp Ile Gly Ile Arg Ala Cys Leu Arg Arg Gly Leu Ser Pro Ala Asn Ala   708

GTG AGA TCA ACA AGC ACT GGT CTT GTC ATC GGC GCT GGC GGA ATG GCT AGG GCA
Val Arg Ser Thr Ser Thr Gly Leu Val Ile Gly Ala Gly Gly Met Ala Arg Ala   726

GCT GTC TAT GCC ATG CTT CAA CTG GGA GTC AAG AAG ATT TTG ATC TTT AAC CGA
Ala Val Tyr Ala Met Leu Gln Leu Gly Val Lys Lys Ile Leu Ile Phe Asn Arg   744

ACA TTT GCT AAT GCC GAG AAG CTG GTT CTA CAC TTC GAG AAC CTG TTG GTC AGA
Thr Phe Ala Asn Ala Glu Lys Leu Val Leu His Phe Glu Asn Leu Leu Val Arg   762

GAC GCA TTG CCT CTG TTG AGC ACA GGG CCA AGA TCC CAC GAC AAC ACC TGT TTT
Asp Ala Leu Pro Leu Leu Ser Thr Gly Pro Arg Ser His Asp Asn Thr Cys Phe   780

CAC ATC ATT CGA TCT CGA GAC GAT CCG CTC CCA GAA AAC TTC AAA AAC CCG ACC
His Ile Ile Arg Ser Arg Asp Asp Pro Leu Pro Glu Asn Phe Lys Asn Pro Thr   798

ATG ATC GTT TCC TGC ATA CCG ACA CAC ACA GTG GAC AAC ACC CCT GAC CCT GAA
Met Ile Val Ser Cys Ile Pro Thr His Thr Val Asp Asn Thr Pro Asp Pro Glu   816

TTT ACT GTG CCT TTG CAC TGG CTC GAC AAC CCC ACT GGC GGC ATT GTA CTA GAA
Phe Thr Val Pro Leu His Trp Leu Asp Asn Pro Thr Gly Gly Ile Val Leu Glu   834

CTC GAC TAC AAA TGT CTC ACA TCA CCC TTG CTC GAA CAA ACA CGA CGC GAG GCT
Leu Asp Tyr Lys Cys Leu Thr Ser Pro Leu Leu Glu Gln Thr Arg Arg Glu Ala   852

CAC AGA GGC TGG GTC GCA ATG GAT GGA CTT GAC CTC TTG CCA GAA CAA GGG TTT
His Arg Gly Trp Val Ala Met Asp Gly Leu Asp Leu Leu Pro Glu Gln Gly Phe   870

GCC CAA TTT GAA CTG TTC ACC GGG CGG CGA GCA CCT CGT CGC TTG ATG AGG CGC
Ala Gln Phe Glu Leu Phe Thr Gly Arg Arg Ala Pro Arg Arg Leu Met Arg Arg   888

GAG GTT TTG CGA GCA TAC CCA GAT GAT CAA GCA AAA TCT CAT ACC GCG CGA TTA
Glu Val Leu Arg Ala Tyr Pro Asp Asp Gln Ala Lys Ser His Thr Ala Arg Leu   906

CGA CCT CGC CTC AAC GGA ATT GCA ACG CAA ATA TCT TGA AACAAAGAGACATGCGGATC
Arg Pro Arg Leu Asn Gly Ile Ala Thr Gln Ile Ser                           918

TCACGGTCTT GATTTCGGGG CGGGCTGTCG GCATTAGGGT ACCACTCGGA GCGATTTGGC CGGACAGCAA
 Bgl II
CAAGATCT
```

FIG. 1. Nucleotide and deduced amino acid sequence of the *N. crassa qa-1S* gene. The untranslated region is in negative numbers and the amino acids are numbered from ATG as amino acid 1. There is a 66-bp intervening sequence between amino acids 75 and 76. An asterisk indicates a major transcription start site, and a dot indicates a minor site.

Nucleotide Polymorphisms in the Coding Region of the *qa-1S* Gene. The DNA sequence analysis of the four *qa-1S* mutations also revealed several polymorphisms within the deduced coding region and especially within the 5′ flanking region of the gene. Within the first 857 amino acids of the coding region, there are three positions at which the wild-type 74-OR23-1A sequence differs by a single base pair from that of all three mutants (141, 105^{ts}, and 105^{c}) that were derived from 74-OR23-1a. Two of the changes result in no amino acid change, while the third causes a change of valine (*G*TA) to isoleucine (*A*TA) at amino acid 319. The region comprising amino acids 858–918 was not sequenced in the mutants. The 5′ flanking region of the gene in these three mutants also showed differences from the wild-type 74-OR23-1A. There are a number of base substitutions, deletions, and additions in the region between the *Hpa* I site and the major 5′ ends of the gene, as shown in Fig. 5. However, none of these differences observed in mutants 141, 105^{ts}, and 105^{c} were found in mutant A1-1.4, which was derived from 74-OR23-1A. Therefore, these differences presumably constitute polymorphisms in chromosome VII in the different wild-types 74-OR23-1A and 74-OR8-1a.

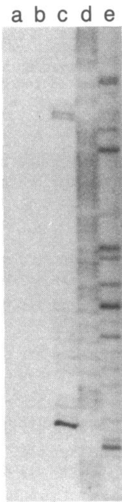

FIG. 2. S1 nuclease mapping of the 5′ ends of the qa-1S mRNA. Lanes: a, *Escherichia coli* RNA (10 μg); b, uninduced *N. crassa* RNA (10 μg); c, quinic acid-induced RNA (10 μg); d, G+A chemical sequencing reactions of the probe fragment; e, T+C chemical sequencing reactions.

DISCUSSION

The *qa-1S* Gene Encodes a Very Large Protein. The *qa-1S* gene had previously been shown by *Neurospora* transformation to be located on a 5.0-kb *Eco*RI fragment and DNA·RNA hybridization analysis had indicated that related 3.4-kb and 4.1-kb mRNAs were transcribed from this region (7). The nucleotide sequence of this *Eco*RI fragment has been determined. It contains a long open reading frame of 918 codons interrupted by an intervening sequence of 66 bp, which contains the only two in-frame stop codons. The approximate boundaries of the intervening sequence were determined by comparing the S1 nuclease digestion products with DNA size standards (data not shown). To determine the exact splice sites, the sequence of the mRNA was directly determined by primer extension with reverse transcriptase in the presence of dideoxynucleotides. The sequences at the intron boundaries match the normal consensus sequences, except that the 5′ donor site is GC rather than GT. However, a GC donor site occurs in chicken and duck α-globin genes and has been directly shown to be efficiently and precisely removed (13). Another interesting feature of the intervening sequence is that it contains an exact copy of a yeast consensus splice sequence TACTAACA at the same position found in yeast introns. Although *Neurospora* introns do not usually contain this exact sequence, a similar more variable consensus sequence has been observed (14). Splicing at the GC donor position allows readthrough from an upstream open reading frame into the 2500-bp downstream open reading frame. The initiator methionine of the deduced coding region is the furthest upstream methionine in the reading frame 5′ to the intron. This is most likely the initiation codon, since the sequences immediately upstream of the ATG have many similarities to those upstream of the initiation codons of other *Neurospora* genes. The nucleotides at positions −4 and −3 are C and A, respectively, as in five of the seven *qa* genes and the six other sequenced *Neurospora* genes (14–20). It also has the second most common consensus site observed by Kozak, AXXATGA (21). In addition, the region around −25 is 80% A/C, which is also common before *Neurospora* genes (19, 20). The sequence CCACC, which occurs immediately before the ATGs of several other *Neurospora* genes (14–20), occurs 25 bp before the putative ATG of the qa-1S protein. It cannot be ruled out, however, that one of the methionines at positions 31, 38, or 40 is the true initiator. This awaits the determination of the NH_2-terminal sequence of the protein.

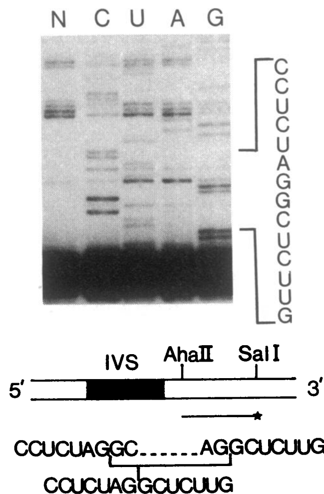

FIG. 3. Sequencing of the qa-1S mRNA across the splice site. The 5′-end-labeled *Sal* I/*Aha* II fragment indicated was used to prime DNA synthesis by reverse transcriptase in the presence of no ddNTPs (N), ddGTP (C), ddATP (U), ddTTP (A), or ddCTP (G). C, U, A, and G indicate the nucleotides of the inferred RNA sequence. At the bottom of the diagram is shown the inferred splicing reaction. IVS, intervening sequence.

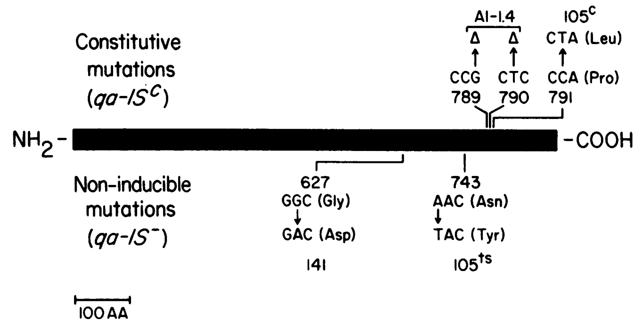

FIG. 4. Diagram of the relative locations, nucleotide changes, and inferred amino acid changes in two constitutive (*qa-1S*^{c}) and two noninducible (*qa-1S*$^-$) repressor mutations. Numbers refer to *qa-1S* codons. (The total number of codons in the *qa-1S* gene is 918.) Mutant 105^c is derived from mutant 105^{ts} as a constitutive revertant at 25°C. In mutant A1-1.4, the single base pair deletions in codons 789 and 790 generate the nonsense codon UGA at position 813.

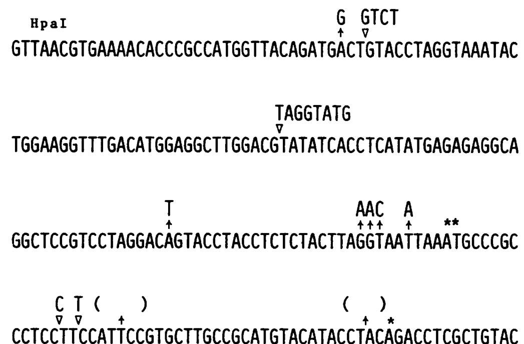

FIG. 5. Locations of nucleotide polymorphisms in the *qa-1S* 5′ flanking sequences. The sequence shown is that of wild-type 74-OR23-1A as in Fig. 1. The bases above the wild-type sequence are substitutions, insertions, and deletions found in all three mutants derived from wild-type 74-OR8-1a. ↑, Substitution; ▽, insertion; ( ), deletion; *, the major transcription start sites of the qa-1S mRNAs.

The 5′-end-mapping experiments indicate that the qa-1S mRNA has two major transcription initiation sites and several minor sites. Most of the sites, including the major sites at -407, -453, and -454 are on an adenine residue. Many of the minor sites frequently occur at the sequence TAC. Similar 5′ end heterogeneity is displayed by many of the qa mRNAs (18, 22).

As in many other *Neurospora* genes, the sequence TATA does not occur 30 bp before any of the putative 5′ ends (18–20). The closest TATA is 70 bp from the major start site at -453, -454.

The codon usage of the *qa-1S* gene exhibits little codon bias, in contrast to highly expressed *Neurospora* genes such as *am* and histones (14, 17). *qa-1S* utilizes all 61 codons at least four times, and this result is therefore consistent with the observation made for yeast genes that codon bias is related to the level of expression of a specific gene (23). For example, the yeast galactose regulatory genes also show little bias (24, 25).

Nucleotide Sequences of *qa-1S* Regulatory Mutations. The nucleotide sequences of four *qa-1S* mutations have been determined. Two (141 and 105^{ts}) were semi-dominant noninducible types that revert to constitutivity (super-repressor types). The other two were constitutive mutants, one derived directly from wild type (A1-1.4) and one (105^{c}) derived as a revertant of one of the noninducible mutants (105^{ts}). All the mutations occurred within the presumptive *qa-1S* coding region, strongly supporting the hypothesis that *qa-1S* encodes a negative regulatory protein.

The constitutive mutation in mutant 105^{c} is a missense mutation at amino acid 791, and that in mutant A1-1.4 is a frameshift mutation in amino acids 789 and 790, producing the nonsense codon UGA at position 813 (Fig. 4). Therefore the COOH-terminal region of the repressor protein appears to contain the domain involved in binding to the target of the repressor, which may be the activator protein (1). Alternatively, it is possible that the region defined by the constitutive mutations is involved in forming active dimers or tetramers (if the repressor is a multimeric protein), assuming that the inability to form an active multimeric repressor would give rise to a constitutive phenotype.

The localization of the noninducible mutation in mutant 141 at amino acid 627, in a more proximal region of the *qa-1S* gene relative to the positions of the two constitutive mutations (Fig. 4), suggests that this region of the protein may be involved in inducer binding. The other noninducible mutant sequenced, mutant 105^{ts}, is noninducible at 25°C, but is constitutive at 35°C. It contains a missense mutation, changing an asparagine to a tyrosine at amino acid 743. Therefore, it is likely that this region may also be involved in inducer binding. Evidence from transcriptional studies supports the hypothesis that this mutant produces a thermolabile repressor that is almost fully active at 25°C but is largely inactive at 35°C (26). Mutant 105^{c}, which was derived as a second site constitutive revertant of mutant 105^{ts}, is constitutive at both temperatures. The second site mutation in 105^{c} changes a proline to a leucine at amino acid 791 and, presumably, makes the protein inactive at the low temperature also. The fact that the second-site mutation changes a noninducible mutant into a constitutive mutant at normal temperature is expected if noninducible $qa\text{-}1S^{-}$ mutants make active "super-repressors" that bind to the target of the repressor, but not to the inducer, whereas constitutive $qa\text{-}1S^{c}$ mutants make inactive repressors that do not bind to the target of the repressor, irrespective of their ability to bind inducer. It is interesting that all four mutations sequenced map well beyond the midpoint of the *qa-1S* gene, although the sample size is small. Studies of the *lac* repressor indicate that the inducer-binding domain occupies a considerable region of the repressor molecule and has several rather widely spaced positions where i^{s} mutations affect inducer binding (27). In the yeast galactose system, the three *gal80* super-repressor mutants sequenced are clustered between amino acids 301 and 325 in the COOH terminus of the 435-amino acid repressor protein (25). To our knowledge, this is the only other eukaryotic negative regulatory gene in which mutations have been sequenced. Obviously, sequence analysis will be required from many more *qa-1S* mutations in order to define more precisely the critical functional regions of this protein that are involved in binding inducer and in binding the repressor to its target.

The authors would like to acknowledge with thanks the assistance of a number of individuals: Mary Case for growing the mutant strains, Paulette Geever for DNA preparation, Carmel Todd for excellent technical assistance, and Colleen McElfresh for preparation of the manuscript. L.H. would especially like to thank Robert Geever for help with some of the computer analysis and printouts and also with the mRNA sequencing. Finally, thanks go to Brett Tyler for editorial and scientific advice. This research was supported by National Institutes of Health Grant GM28777 to N.H.G. L.H. was supported in part by National Institutes of Health Predoctoral Training Grant GM07103.

1. Giles, N. H., Case, M. E., Baum, J., Geever, R., Huiet, L., Patel, V. & Tyler, B. (1985) *Microbiol. Rev.* **49**, 338–358.
2. Schweizer, M., Case, M. E., Dykstra, C. C., Giles, N. H. & Kushner, S. R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5086–5090.
3. Rines, H. W., Case, M. E. & Giles, N. H. (1969) *Genetics* **61**, 789–800.
4. Case, M. E. & Giles, N. H. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 553–557.
5. Valone, J. A., Jr., Case, M. E. & Giles, N. H. (1971) *Proc. Natl. Acad. Sci. USA* **68**, 1555–1559.
6. Partridge, C. W. H., Case, M. E. & Giles, N. H. (1972) *Genetics* **72**, 411–417.
7. Huiet, L. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1174–1178.
8. Frischauf, A., Lehrach, H., Poustka, A. & Murray, N. (1983) *J. Mol. Biol.* **170**, 827–842.
9. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
10. Berk, A. S. & Sharp, P. A. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1274–1278.
11. Favaloro, J., Treisman, R. & Kamen, R. (1980) *Methods Enzymol.* **65**, 718–749.
12. Treisman, R., Proudfoot, N. J., Shander, M. & Maniatis, T. (1982) *Cell* **29**, 903–911.
13. Fischer, H. D., Dodgson, J. B., Hughes, S. & Engel, J. D. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2733–2737.
14. Kinnaird, J. H. & Fincham, J. R. S. (1983) *Gene* **26**, 253–260.
15. Arends, H. & Sebald, W. (1984) *EMBO J.* **3**, 377–382.
16. Viebrock, A., Perz, A. & Sebald, W. (1982) *EMBO J.* **1**, 565–571.
17. Woudt, L. P., Pastink, A., Kempers-Veenstra, A. E., Jansen, A. E. M., Mager, W. H. & Planta, R. J. (1983) *Nucleic Acids Res.* **11**, 5347–5360.
18. Rutledge, B. J. (1984) *Gene* **32**, 275–287.
19. Alton, N. K., Buxton, F., Patel, V., Giles, N. H. & Vapnek, D. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1955–1959.
20. Schechtman, M. & Yanofsky, C. (1983) *J. Mol. Appl. Genet.* **2**, 83–99.
21. Kozak, M. (1984) *Nucleic Acids Res.* **12**, 857–872.
22. Tyler, B. M., Geever, R. F., Case, M. E. & Giles, N. H. (1984) *Cell* **36**, 493–502.
23. Bennetzen, J. L. & Hall, B. D. (1982) *J. Biol. Chem.* **257**, 3026–3031.
24. Laughon, A. & Gesteland, R. F. (1984) *Mol. Cell. Biol.* **4**, 260–267.
25. Nogi, Y. & Fukasawa, T. (1984) *Nucleic Acids Res.* **12**, 9287–9298.
26. Patel, V. B. & Giles, N. H. (1985) *Mol. Cell. Biol.* **5**, 3593–3599.
27. Miller, J. H. (1979) *J. Mol. Biol.* **131**, 249–258.