Blood Cells, Molecules, and Diseases (2000) **26**(3) February 15: 37–46
doi:10.1006/bcmd.2000.0274, available online at http://www.idealibrary.com on IDEAL®

Genomic Structure of the Human p47-phox (*NCF1*) Gene

Submitted 01/24/00; revised 02/01/00
(communicated by Bernard M. Babior, M.D., 02/02/00)

Stephen J. Chanock,[1] Joachim Roesler,[2,3] Shixing Zhan,[1,4] Penelope Hopkins,[5,6] Pauline Lee,[5] David T. Barrett,[1] Barbara L. Christensen,[1] John T. Curnutte,[2,5] and Agnes Görlach[5,7]

ABSTRACT: The cytosolic factor p47-phox, encoded by the *NCF1* gene, is an essential component of the phagocyte NADPH-oxidase system. Upon activation of this multicomponent system, p47-phox translocates to the membrane and participates in the electron transfer from NADPH to molecular oxygen. A deficiency or absence of p47-phox is the most common autosomal form of chronic granulomatous disease (CGD). We have cloned and characterized the *NCF1* gene from four bacteriophage clones, a P1 clone and genomic DNA from normal individuals. The gene is 15,236 base pairs long and includes 11 exons. It is 98.6% homologous in sequence to at least one pseudogene that maps to the same region of chromosome 7q11.23. Slightly more than half (50.37%) of the wild-type *NCF1* gene consists of repetitive elements. In particular, the density of Alu sequences is high (1.4 Alu/kb); there are 21 Alu repeats interspersed through 10 introns. These findings are consistent with the observation that recombination events between the wild-type gene and its highly homologous pseudogenes account for the majority of potentially lethal mutations in p47-phox-deficient chronic granulomatous disease. Analysis of 1.96 kb of sequence 5′ of the start of translation revealed a high homology (99.6%) between wild-type and pseudogene clones. Characterization of *NCF1* establishes a foundation for detailed molecular analysis of p47-phox-deficient CGD patients as well as for the study of the regulation of the *NCF1* gene and pseudogenes, both of which are present as full-length transcripts in normal individuals.

Keywords: gene; NADPH-oxidase; chronic granulomatous disease; respiratory burst.

INTRODUCTION

In phagocytic cells, the NADPH-oxidase is responsible for the oxidative burst in which electrons are transported from NADPH to molecular oxygen according to the reaction

$$NADPH + 2O_2 \rightarrow NADP^+ + 2O_2^- + H^+ \quad (1).$$

This multicomponent enzyme complex consists of at least five components, two membrane bound subunits (gp91-phox and p22-phox) and three cytosolic factors (p40-phox, p47-phox and p67-phox) (2). The respiratory burst oxidase is dormant in resting cells but upon activation is assembled at the membrane. The cytosolic proteins, p40-phox, p47-phox and p67-phox translocate to the membrane where they associate with the gp91-phox and p22-phox components (3–6). p47-

Reprint requests to: Dr. Stephen J. Chanock, Pediatric Oncology Branch, National Cancer Institute, Advanced Technology Center, 8717 Grovemont Circle, Gaithersburg, MD 20877. Fax: 301-402-3134. E-mail: sc83a@nih.gov.

Abbreviations used: CGD, chronic granulomatous disease; NADPH-oxidase, nicotinamide adenine dinucleotide phosphate-oxidase; *NCF1*, gene encoding p47-phox, cytosolic protein of the NADPH-oxidase; *NCF2*, gene encoding p67-phox, cytosolic protein of the NADPH-oxidase; *NCF4*, gene encoding p40-phox, cytosolic protein of the NADPH-oxidase; *CYBB*, gene encoding gp91-phox, membrane protein of the NADPH-oxidase; *CYBA*, gene encoding p22-phox, membrane protein of the NADPH-oxidase; LINE, Long Interspersed Elements; MIR, Mammalian Interspersed Repetitive; SINE, Short Interspersed Elements; SH3, *src*-homology region #3.

[1] Pediatric Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland.
[2] Research Immunology, Genentech, Inc., South San Francisco, California.
[3] Current address: Technische Universitaet Dresden, Dresden, Germany.
[4] Current address: Scripps Research Institute, La Jolla, California.
[5] Department of Molecular and Experimental Medicine, Scripps Research Institute, La Jolla, California.
[6] Current address: Sequana Therapeutics, Inc., La Jolla, California.
[7] Current address: Institut für Kardiovaskuläre Physiologie, Klinikum der JWG-Universität, Frankfurt am Main, Germany.

1079-9796/00

phox, a basic protein, is phosphorylated upon activation of the enzyme complex (7–9). In addition, one or more cytosolic GTP-binding proteins appear to be required for oxidase activity (10–14).

Chronic granulomatous disease, CGD, is an inherited disorder of phagocytic cells characterized by the inability of phagocytes and B-lymphocytes to generate superoxide (15,16). Superoxide and its toxic oxygen derivatives are essential for microbicidal activity. CGD patients, who have deficient or absent production of superoxide, are susceptible to recurrent and life-threatening bacterial and fungal infections. This heterogeneous group of patients is lacking one or more of four components of the respiratory burst oxidase, gp91-phox, p22-phox, p47-phox and p67-phox. In rare cases, the functional activity of one of these components is significantly decreased, resulting in the CGD phenotype (15,17).

The gene structures of the NADPH-oxidase components *CYBB, CYBA,* and *NCF2* have been determined (18–20). The genes of the NADPH-oxidase have been mapped to Xp21.1 (*CYBB*), 16q24 (*CYBA*), 1q25 (*NCF2*), 22q13.1 (*NCF4*) and 7q11.23 (*NCF1*) (19,21–23). Recently, we have discovered the existence of a highly homologous pseudogene, which localizes to the same region of 7q11.23 as the wild-type gene (24).

The identification of the structure of these genes has permitted detailed molecular analyses of CGD patients deficient in gp91-phox, p22-phox, and p67-phox. Characterization of the mutations in gp91-phox-deficient CGD patients has revealed a wide spectrum including deletion of the entire gene, point insertions, substitutions or deletions within exons, splice junctions and the 5′ upstream regulatory region of the gene (17,25,26). Similarly, analysis of two of the autosomal recessive forms of CGD due to p22-phox or p67-phox deficiency is notable for a heterogenous group of mutations, including single base deletions, amino acid substitutions and splice junction errors (17,19).

Approximately 30% of patients with CGD do not have detectable p47-phox protein by Western blot analysis and an inheritance pattern that is autosomal recessive (3,27). Interestingly, the first four patients (who were unrelated) were reported to be homozygous for an identical mutation, a GT deletion (ΔGT) at the start of exon two, resulting in a frameshift (28). In two subsequent reports, the same GT deletion was detected exclusively in five of six unrelated patients (29,30). One patient was a compound heterozygote with the ΔGT and a single base pair deletion at bp 502 (29). Subsequently, analysis of additional p47-phox-deficient CGD patients has confirmed the importance of the GT deletion in over 85% of reported patients (31). In addition, it has been suggested that recombination events between one or more of the highly homologous pseudogenes and the p47-phox gene accounts for this observation (24,31).

Here we report the structure of the wild-type *NCF1* gene which had been previously mapped to chromosome 7q11.23 (23,24). 1960 bp of the 5′ upstream region has been sequenced and compared to the sequence corresponding to the pseudogenes. The determination of the map of the gene as well as the 5′ upstream region establishes an important foundation for further molecular analysis of patients with p47-phox-deficient CGD and for the investigation of the regulation of the *NCF1* gene and pseudogenes.

METHODS

Screening of Genomic Library for NCF1 Genomic Clones

Two separate Sau3A partial digest lambda EMBL 3A bacteriophage libraries prepared from human DNA were screened by plaque hybridization using a radiolabeled cDNA probe of the full length *NCF1*. The cDNA used to screen the library was generated by polymerase chain reaction with primers, 47F-1, ATGGGGGACA CCTTCATCCGT and 47R-1, CACTCCAAGCAACATTTATTG. The first library screened was prepared from human leukocyte DNA and the second was prepared from human sperm DNA (gift of Dr. Jeremy Nathans). Clones corresponding to the wild-type gene (L4A, L-HDLA, L27II, and L24) and the pseudogene (L14, L 25A) from the bacteriophage libraries were plaque purified, and the phage DNA analyzed by restriction digests and

Blood Cells, Molecules, and Diseases (2000) **26**(3) February 15: 37–46
doi:10.1006/bcmd.2000.0274, available online at http://www.idealibrary.com on IDEAL®

Southern blot hybridizations using established methods.

P1 clones were obtained from a human library by PCR screening with two sets of oligonucleotides (GenomeSystems, Inc., St. Louis, MO) corresponding to the region in the immediate 5′ untranslated region (using the oligonucleotide primers TTTTCCTTGTCCCTGCAGGT and GACTGGGTGGCCTCCAGTGCTCCCT) and the region between exon 11 and the 3′ untranslated region (using the oligonucleotide primers AGACGCAGCGCTCTAAACCGCA and CTATAGAGCCTGGCGTCTGGA). One of ten contained the wild-type gene (P42) and nine of ten contained the ΔGT signature sequence for the pseudogene (P38, 39, 40, 41,43, 47, 48, 49, and 50) previously reported elsewhere (24).

Sequence Analysis of Clones and Control Genomic DNA

Fragments of bacteriophage and P1 clones were subcloned into pUC and Bluescript KS+ and directly sequenced for confirmation of intron/exon borders using the dideoxynucleotide chain termination method. Each intronic sequence length was confirmed by both Southern hybridization of restriction digests of the phage, P1 and plasmid subclones and polymerase chain reaction-generated fragments derived from control genomic DNA. The full-length gene sequence was obtained from bacteriophage clone L24 and portions of the P1 clone, P42; the latter was used to complete or verify intronic regions. The sequence 5′ of the start of translation was determined for two wild-type clones, L24 and P42, and two pseudogene clones, L14 and P43. The coding region and immediate flanking regions of the introns were confirmed by direct sequencing of fragments amplified from genomic DNA derived from healthy controls.

DNA Amplification and Characterization of Sequences

PCRs contained 0.5 μM of each primer, 50 mM KCl, 10 mM Tris–HCl, 2.5 mM $MgCl_2$, 200 μM dATP, dCTP, and dTTP, 50 μM dGTP, and 150 μM 7-deaza-2′-GTP. The amplification conditions were as follows: 94°C for 1 min, 60°C for 1 min and 72°C for 2 min, for a total of 30 cycles. In selected circumstances, PCR-generated fragments from genomic DNA were directly sequenced by cycle sequencing (fmol Sequencing Kit, Promega, Madison, WI) using the following conditions: 7 min denaturation at 98°C, 30 cycles of 94°C for 30 s, 58°C for 30 s and 72°C for 45 s and a 7-min extension at 72°C.

Sequences were analyzed with MacVector 6.5 (Oxford Molecular, Madison, WI). Further sequence analysis utilized Repeat Masker2 (version 6/16/98) and Mat Inspector v2.2 (http://transfac.gbf-braunschweig.de/cgi-bin/matSearch/matsearch2.pl) (32).

RESULTS

Isolation of Genomic Clones and Mapping of the NCF1 Gene

Four separate recombinant λEMBL 3A bacteriophage clones corresponding to the wild-type *NCF1* gene were isolated from two different human genomic libraries by screening with a full-length cDNA probe containing the coding region of *NCF1* mRNA (Fig. 1). Clones L4A, L-HDLA and L27AII were obtained from a human genomic DNA leukocyte library and clone L24 was derived from a human genomic DNA sperm library. The P1 clone P42 was isolated from a human genomic P1 library as described previously (24) and, in addition, used to confirm known sequence.

Southern blot analysis of the genomic clones digested with four different enzymes (*Bam*HI, *Bgl*II, *Eco*RI, and *Hin*dIII) confirmed that no large rearrangements had occurred during the cloning procedure (data not shown). Based upon the restriction digestion map and direct sequence analysis of these clones, a composite map of the p47-phox gene was constructed (Fig. 1). Exon positions identified by Southern blot analysis matched exactly with the sequence analysis.

The most 5′ bacteriophage clone, L4A, extended at least 6 kb upstream of the initiation codon. L24 extended more than 4 kb upstream of exon 1. One clone, L-HDLA, overlapped the three clones containing the 5′ end of the gene and extended 6 kb beyond the 3′ untranslated region. The restriction map of the P1 clone, P42, and

doi:10.1006/bcmd.2000.0274, available online at http://www.idealibrary.com on IDEAL®

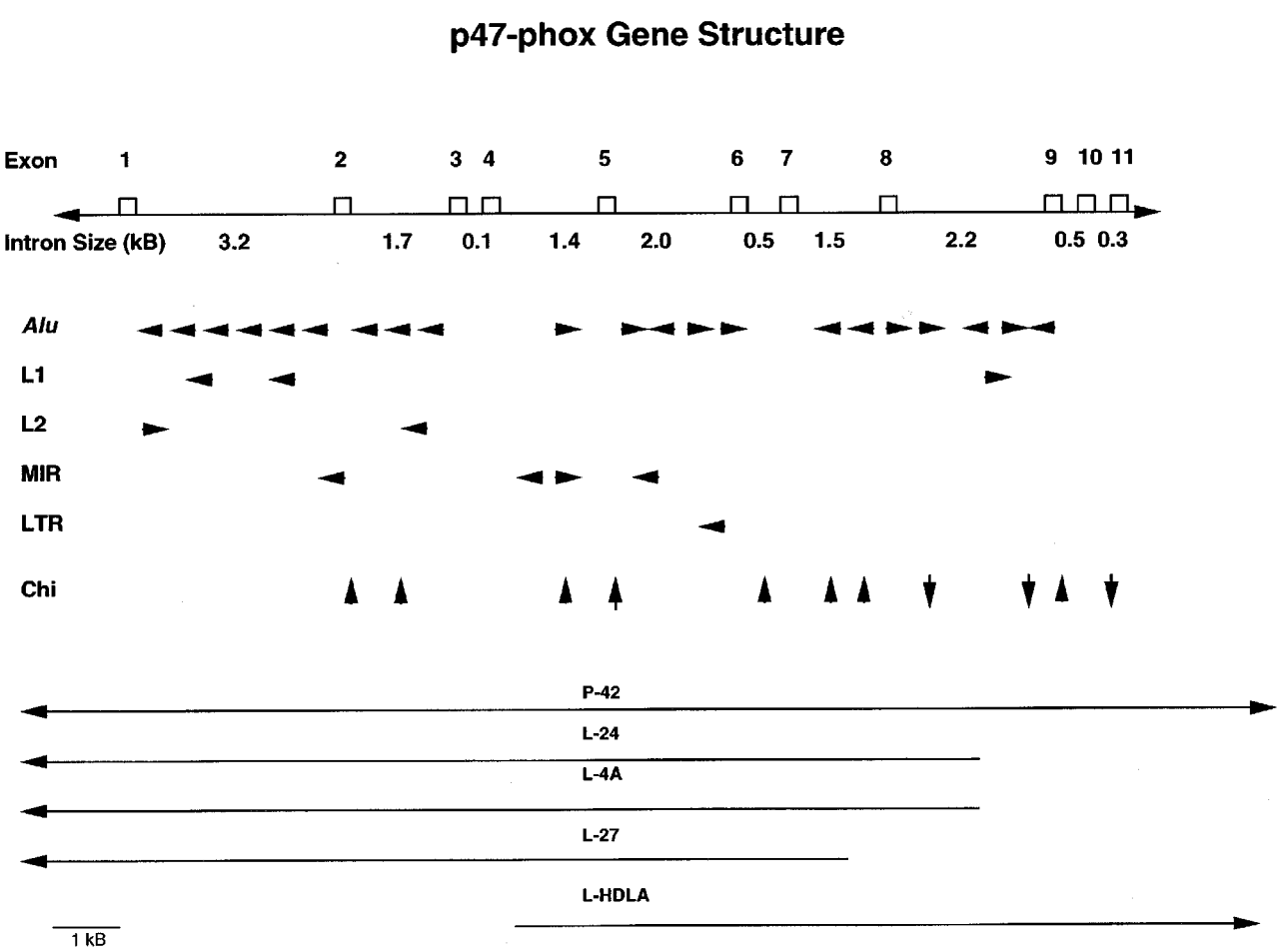

Figure 1. Gene structure for the wild-type *NCF1* gene based upon sequence analysis of clones L24 and P42. Bacteriophage clones (L24, L4A, L27, L-HDLA) were isolated from two separate human genomic libraries using a full-length *NCF1* probe. The P1 clone P42 was obtained as described in the text. Intron sizes are indicated and the approximate positions of repetitive elements are shown underneath. These include 21 Alu sequences, three L1 type LINEs, two L2 type LINEs, four MIR, and one LTR. Also shown are the Chi sequences; exact matches with the consensus sequence are indicated by down arrows whereas the up arrows indicate sites with a one-base mismatch.

sequence analysis confirmed the organizational structure and sequence of the phage clones. The P1 clone extended approximately 45 kb upstream and 20 kb downstream of the p47-phox gene.

Sequence Analysis of the NCF1 Wild-Type Gene

Sequence analysis of the bacteriophage clone, L24, and the P1 clone, P42, revealed that the *NCF1* wild-type gene contains 15,236 base pairs from the initiation ATG to the start of the poly(A) tail. The GC content is 54.12%. A surprisingly high portion of the gene (50.37%) consists of repetitive elements. 41.09% of the gene consists of SINEs (Short Interspersed Elements) and 6.33% of sequence consists of LINEs. (Long Interspersed Elements). Notably, there are 21 Alu elements and 3 MIR elements (Mammalian Interspersed Repetitive) elements throughout the gene. The density of Alu sequences is very high, one every 0.69 kb of sequence and 66% of the Alu sequences are biased in the 3′ to 5′ orientation. Interestingly, all 9 Alu sequences in intron 1 and 2, which flank exon 2 (where the ΔGT is present in the p47-phox pseudogenes) are in the 3′ to 5′ orientation.

Sequence analysis of 1,960 bases upstream of the start of translation is presented in Fig. 2. Interestingly, there is 99.6% homology between the gene and pseudogene sequence, derived from two clones, L14 and P43 (24). There are two

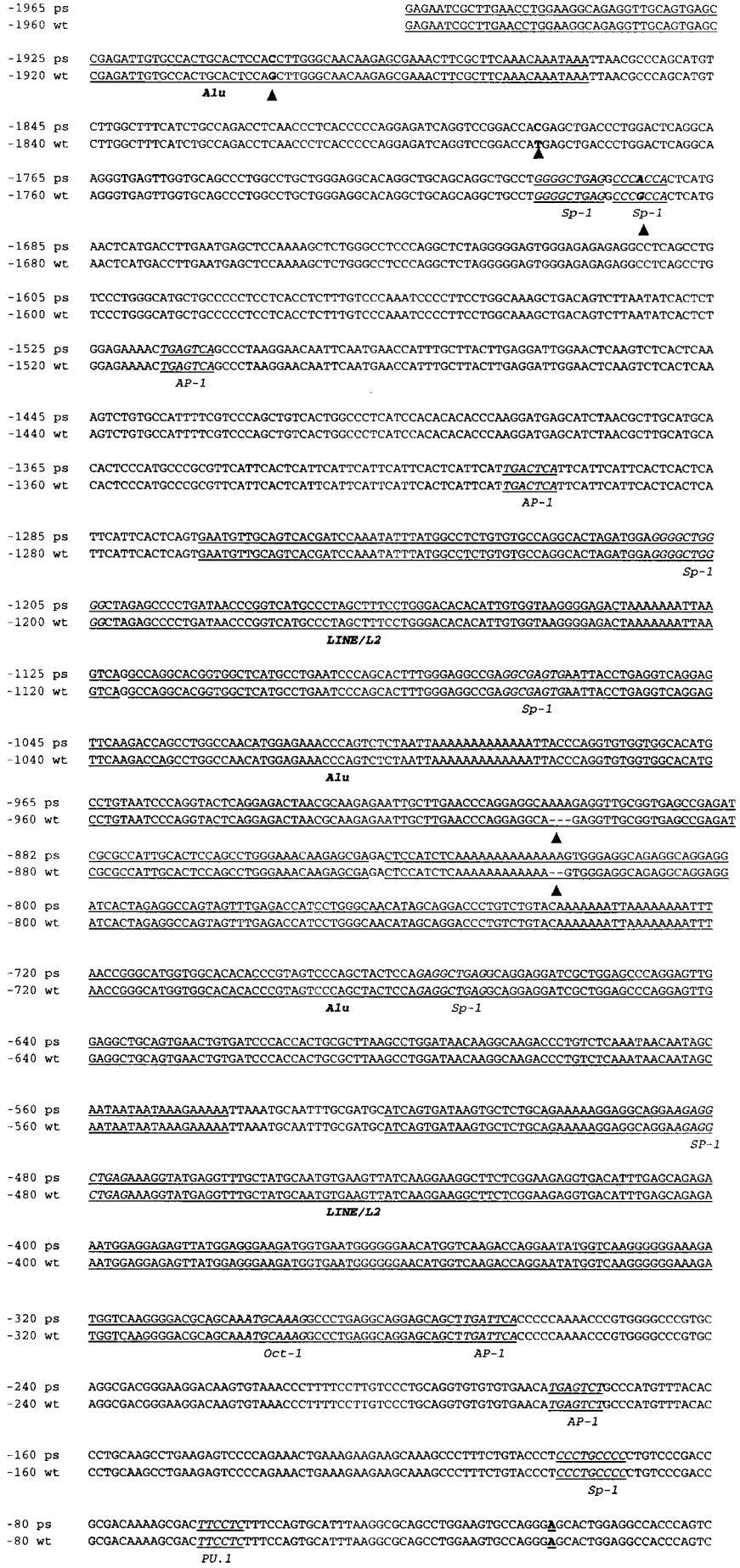
-1965 ps GAGAATCGCTTGAACCTGGAAGGCAGAGGTTGCAGTGAGC
-1960 wt GAGAATCGCTTGAACCTGGAAGGCAGAGGTTGCAGTGAGC
-1925 ps CGAGATTGTGCCACTGCACTCCACCTTGGGCAACAAGAGCGAAACTTCGCTTCAAACAAATAAATTAACGCCCAGCATGT
-1920 wt CGAGATTGTGCCACTGCACTCCAGCTTGGGCAACAAGAGCGAAACTTCGCTTCAAACAAATAAATTAACGCCCAGCATGT
Alu
-1845 ps CTTGGCTTTCATCTGCCAGACCTCAACCCTCACCCCCAGGAGATCAGGTCCGGACCACGAGCTGACCCTGGACTCAGGCA
-1840 wt CTTGGCTTTCATCTGCCAGACCTCAACCCTCACCCCCAGGAGATCAGGTCCGGACCATGAGCTGACCCTGGACTCAGGCA
-1765 ps AGGGTGAGTTGGTGCAGCCCTGGCCTGCTGGGAGGCACAGGCTGCAGCAGGCTGCCTGGGGCTGAGGCCCACCACTCATG
-1760 wt AGGGTGAGTTGGTGCAGCCCTGGCCTGCTGGGAGGCACAGGCTGCAGCAGGCTGCCTGGGGCTGAGGCCCGCCACTCATG
Sp-1 Sp-1
-1685 ps AACTCATGACCTTGAATGAGCTCCAAAAGCTCTGGGCCTCCCAGGCTCTAGGGGGAGTGGGAGAGAGAGGCCTCAGCCTG
-1680 wt AACTCATGACCTTGAATGAGCTCCAAAAGCTCTGGGCCTCCCAGGCTCTAGGGGGAGTGGGAGAGAGAGGCCTCAGCCTG
-1605 ps TCCCTGGGCATGCTGCCCCCTCCTCACCTCTTTGTCCCAAATCCCCTTCCTGGCAAAGCTGACAGTCTTAATATCACTCT
-1600 wt TCCCTGGGCATGCTGCCCCCTCCTCACCTCTTTGTCCCAAATCCCCTTCCTGGCAAAGCTGACAGTCTTAATATCACTCT
-1525 ps GGAGAAAACTGAGTCAGCCCTAAGGAACAATTCAATGAACCATTTGCTTACTTGAGGATTGGAACTCAAGTCTCACTCAA
-1520 wt GGAGAAAACTGAGTCAGCCCTAAGGAACAATTCAATGAACCATTTGCTTACTTGAGGATTGGAACTCAAGTCTCACTCAA
AP-1
-1445 ps AGTCTGTGCCATTTTCGTCCCAGCTGTCACTGGCCCTCATCCACACACACCCAAGGATGAGCATCTAACGCTTGCATGCA
-1440 wt AGTCTGTGCCATTTTCGTCCCAGCTGTCACTGGCCCTCATCCACACACACCCAAGGATGAGCATCTAACGCTTGCATGCA
-1365 ps CACTCCCATGCCCGCGTTCATTCACTCATTCATTCATTCATTCACTCATTCATTGACTCATTCATTCATTCACTCACTCA
-1360 wt CACTCCCATGCCCGCGTTCATTCACTCATTCATTCATTCATTCACTCATTCATTGACTCATTCATTCATTCACTCACTCA
AP-1
-1285 ps TTCATTCACTCAGTGAATGTTGCAGTCACGATCCAAATATTTATGGCCTCTGTGTGCCAGGCACTAGATGGAGGGGCTGG
-1280 wt TTCATTCACTCAGTGAATGTTGCAGTCACGATCCAAATATTTATGGCCTCTGTGTGCCAGGCACTAGATGGAGGGGCTGG
Sp-1
-1205 ps GGCTAGAGCCCCTGATAACCCGGTCATGCCCTAGCTTTCCTGGGACACACATTGTGGTAAGGGGAGACTAAAAAAATTAA
-1200 wt GGCTAGAGCCCCTGATAACCCGGTCATGCCCTAGCTTTCCTGGGACACACATTGTGGTAAGGGGAGACTAAAAAAATTAA
LINE/L2
-1125 ps GTCAGGCCAGGCACGGTGGCTCATGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGAGTGAATTACCTGAGGTCAGGAG
-1120 wt GTCAGGCCAGGCACGGTGGCTCATGCCTGAATCCCAGCACTTTGGGAGGCCGAGGCGAGTGAATTACCTGAGGTCAGGAG
Sp-1
-1045 ps TTCAAGACCAGCCTGGCCAACATGGAGAAACCCAGTCTCTAATTAAAAAAAAAAAAAATTACCCAGGTGTGGTGGCACATG
-1040 wt TTCAAGACCAGCCTGGCCAACATGGAGAAACCCAGTCTCTAATTAAAAAAAAAAAAATTACCCAGGTGTGGTGGCACATG
Alu
-965 ps CCTGTAATCCCAGGTACTCAGGAGACTAACGCAAGAGAATTGCTTGAACCCAGGAGGCAAAAGAGGTTGCGGTGAGCCGAGAT
-960 wt CCTGTAATCCCAGGTACTCAGGAGACTAACGCAAGAGAATTGCTTGAACCCAGGAGGCA---GAGGTTGCGGTGAGCCGAGAT
-882 ps CGCGCCATTGCACTCCAGCCTGGGAAACAAGAGCGAGACTCCATCTCAAAAAAAAAAAAAAAGTGGGAGGCAGAGGCAGGAGG
-880 wt CGCGCCATTGCACTCCAGCCTGGGAAACAAGAGCGAGACTCCATCTCAAAAAAAAAAAA--GTGGGAGGCAGAGGCAGGAGG
-800 ps ATCACTAGAGGCCAGTAGTTTGAGACCATCCTGGGCAACATAGCAGGACCCTGTCTGTACAAAAAAATTAAAAAAAATTT
-800 wt ATCACTAGAGGCCAGTAGTTTGAGACCATCCTGGGCAACATAGCAGGACCCTGTCTGTACAAAAAAATTAAAAAAAATTT
-720 ps AACCGGGCATGGTGGCACACACCCGTAGTCCCAGCTACTCCAGAGGCTGAGGCAGGAGGATCGCTGGAGCCCAGGAGTTG
-720 wt AACCGGGCATGGTGGCACACACCCGTAGTCCCAGCTACTCCAGAGGCTGAGGCAGGAGGATCGCTGGAGCCCAGGAGTTG
Alu Sp-1
-640 ps GAGGCTGCAGTGAACTGTGATCCCACCACTGCGCTTAAGCCTGGATAACAAGGCAAGACCCTGTCTCAAATAACAATAGC
-640 wt GAGGCTGCAGTGAACTGTGATCCCACCACTGCGCTTAAGCCTGGATAACAAGGCAAGACCCTGTCTCAAATAACAATAGC
-560 ps AATAATAATAAAGAAAAATTAAATGCAATTTGCGATGCATCAGTGATAAGTGCTCTGCAGAAAAAGGAGGCAGGAAGAGG
-560 wt AATAATAATAAAGAAAAATTAAATGCAATTTGCGATGCATCAGTGATAAGTGCTCTGCAGAAAAAGGAGGCAGGAAGAGG
SP-1
-480 ps CTGAGAAAGGTATGAGGTTTGCTATGCAATGTGAAGTTATCAAGGAAGGCTTCTCGGAAGAGGTGACATTTGAGCAGAGA
-480 wt CTGAGAAAGGTATGAGGTTTGCTATGCAATGTGAAGTTATCAAGGAAGGCTTCTCGGAAGAGGTGACATTTGAGCAGAGA
LINE/L2
-400 ps AATGGAGGAGAGTTATGGAGGGAAGATGGTGAATGGGGGGAACATGGTCAAGACCAGGAATATGGTCAAGGGGGGAAAGA
-400 wt AATGGAGGAGAGTTATGGAGGGAAGATGGTGAATGGGGGGAACATGGTCAAGACCAGGAATATGGTCAAGGGGGGAAAGA
-320 ps TGGTCAAGGGGACGCAGCAAATGCAAAGGCCCTGAGGCAGGAGCAGCTTGATTCACCCCCAAAACCCGTGGGGCCCGTGC
-320 wt TGGTCAAGGGGACGCAGCAAATGCAAAGGCCCTGAGGCAGGAGCAGCTTGATTCACCCCCAAAACCCGTGGGGCCCGTGC
Oct-1 AP-1
-240 ps AGGCGACGGGAAGGACAAGTGTAAACCCTTTTCCTTGTCCCTGCAGGTGTGTGTGAACATGAGTCTGCCCATGTTTACAC
-240 wt AGGCGACGGGAAGGACAAGTGTAAACCCTTTTCCTTGTCCCTGCAGGTGTGTGTGAACATGAGTCTGCCCATGTTTACAC
AP-1
-160 ps CCTGCAAGCCTGAAGAGTCCCCAGAAACTGAAAGAAGAAGCAAAGCCCTTTCTGTACCCTCCCTGCCCCCTGTCCCGACC
-160 wt CCTGCAAGCCTGAAGAGTCCCCAGAAACTGAAAGAAGAAGCAAAGCCCTTTCTGTACCCTCCCTGCCCCCTGTCCCGACC
Sp-1
-80 ps GCGACAAAAGCGACTTCCTCTTTCCAGTGCATTTAAGGCGCAGCCTGGAAGTGCCAGGGAGCACTGGAGGCCACCCAGTC
-80 wt GCGACAAAAGCGACTTCCTCTTTCCAGTGCATTTAAGGCGCAGCCTGGAAGTGCCAGGGAGCACTGGAGGCCACCCAGTC
PU.1

Figure 2

TABLE 1

Intron/Exon Borders of the Wild-Type *NCF1* Gene

Intron	5′ boundary	Intron length (Kb)	3′ boundary	
	72			73
I.	AGC CAG CAC TAT	gtgagtagctgctggagggcatccccgtg	3.2	ggtcccccgatgtggctttcccccag	GTG TAC ATG
	153			154
II.	TAC GAG TTC CAT	gtgagtgtggggacggaggagggacagg	1.7	cacgtcttcttgtctttttaatgtttag	AAA ACC TTA
	229			230
III.	CC CAC CTC CCA G	gtgagcacggggctgagccgcctgtcagg	0.099	cttgacctcatgttctctggtgccag	CT CCC AAG
	395			396
IV.	ACG GAC AAC CA	gtgagtgaacttttcaccctgccaggtggg	1.4	tctcacccagactggttctctcctcag	G ACA AAA
	451			452
V.	GT ACC GCG ACA G	gtgagaggacgggggggcagccggcggg	2.0	ctcaccctgccctccctcttgccccag	AC ATC ACC
	574			575
VI.	G AGC GAG AGC G	gtcagacctcccaccttacggggctccttc	0.5	gtcacattcccgcacctctggcacag	GT TGG TGG
	682			683
VII.	CC AAC TAT GCA G	gtgccccctgccctccgaggctgtaggggt	1.5	tgctctgtgcccctgccgtggaccag	GT GAG CCA
	800			801
VIII.	TGG GTC ATC AG	gtaggagggcccctctccatccagagcacc	2.2	gcctcaagggctgcctctgttgcag	G AAA GAC
	905			906
IX.	G CCG CCC CGC AG	gtaagcgggggtccccggggctgggcgg	0.5	cgacgccccgtcccgctgggccag	G TCG TCC
	1051			1052
X.	GG AGC CCG CTC G	gtgagtgcagctgagagagggcaggaag	0.3	actcggccccgctctctgcccgcag	AG GAG GAG

insertions in the pseudogene 5′ upstream sequence, an insertion of two adenosines at −830 bp and an insertion of three adenosines at bp −903. Three single base pair changes were found at bp −1895, bp −1783, and bp −1690.

The region is also notable for a high degree of repetitive sequences, including three Alu sequences, two of which, (bp −1115 to −821 and bp −819 to −543) are back to back and flanked on both sides by LINE sequences. There is no consensus site for a TATA or CCAAT box in close proximity to the start of transcription (see below) at −21 bp. A PU.1 binding site which has been shown to be myeloid specific, is located between bp −61 and −66 (33). Further upstream in the sequence are a number of potential binding sites for transcription factors, seven SP-1 sites (bp −92 to −100, bp −476 to −484 and bp −670 to −678, bp −1061 to −1069, bp −1200 to 1208, bp −1687 to −1693 and bp −1695 to −1703), four AP-1 sites (bp −175 to −181 (one mismatch) bp −266 to −272, bp −1301 to −1307, bp −1503 to −1510) and an Oct1 site (bp −291 to −299).

Exon–Intron Structure of the Coding Region of NCF1

Exon positions were identified based on Southern blot analysis of restriction-digested

Figure 2. Comparison of the 5′ upstream sequence of 1960 nucleotides of the wild-type *NCF1* gene and 1965 bp of the pseudogene but not including the initiation ATG codon. The pseudogene (*ps*) clones (L14 and P43) differ at only five different locations from the wild-type (*wt*) clones (L24 and P42), two insertions of AA at −830 and AAA at −903 and three single base pair changes (at −1895, 1783 and −1690 of the *wt* sequence). Differences in the pseudogene sequence are shown in bold (for single base changes) and with bars for insertions. A triangle symbol is indicted underneath the noted differences between the wild-type and pseudogene sequences. The major start of transcription is indicated in boldface type, the A at bp −21 (33, 40). Repetitive sequence elements are underlined, three Alu elements (bp −1960 to −1856, bp −1116 to −821 and bp −819 to −543) and two LINE/L2 elements (bp −1246 to −1116 and bp −542 to −270). Consensus sequences for binding of transcription factors are indicated by underlined and italicized bases. These include seven Sp-1 sites, four AP-1 sites and one Oct-1 site. The PU.1 binding site, shown by Li *et al.*, 1997 is at bp −61 to −66.

Blood Cells, Molecules, and Diseases (2000) **26**(3) February 15: 37–46
doi:10.1006/bcmd.2000.0274, available online at http://www.idealibrary.com on IDEAL®

cloned genomic DNA. Consecutive sequence analysis of the restriction fragments revealed the exon/intron boundaries (Table 1). Polymerase chain reaction amplification of genomic DNA extracted from normal individuals followed by direct sequence analysis of the amplicons confirmed the junctions. The gene is divided into 11 exons (Fig. 1). The predicted product of translation is 390 amino acids, which corresponds to the reported size of the isolated protein (4,6). All splice junction sequences conformed to the GT/AG rule (see Table 1). The sizes of individual exons ranged from 55 to 165 bp corresponding to 18 to 55 amino acids.

Several features are notable in the organization of the gene. The intron between exon 3 and exon 4 is only 99 bases long and lacks a termination codon. If translated in frame with exon 3, an alternatively spliced form of *NCF1* could exist. However, this sequence has not been observed by direct sequence analysis of cDNA derived from healthy controls or p47-phox-deficient CGD patients nor from cDNA clones from a range of libraries of hematopoietic origin (neutrophil, monocyte or lymphocyte) (data not shown). Furthermore, using this 99 bp fragment as a probe, it has not been possible to identify an expressed mRNA species containing this sequence by Northern blot analysis or RT-PCR (data not shown). The two functionally important SH3 domains, corresponding to amino acid residues 161–211 and 231–281, are distributed between contiguous exons; SH3-#1 lies in exons 6 and 7 and SH3-#2 lies in exon 8 and 9.

DISCUSSION

The complete wild-type *NCF1* gene has been sequenced and analyzed. The *NCF1* gene is 15,327 bp long, contains 11 exons and has an intron/exon structure identical to the recently identified highly homologous pseudogenes (24). Sequence comparison of the *NCF1* gene and pseudogenes revealed several remarkable features. First, there is a greater than 98% homology in regions sequenced. Secondly, a set of signature differences distinguishes wild-type clones from pseudogene clones, including the presence or absence of the ΔGT sequence at the start of exon 2. Thirdly, the *NCF1* gene (and pseudogenes) contains a high density of repetitive sequences. 50.37% of the sequence between exon 1 and the poly(A) tail and 59.03% of the first 1960 bp of the 5′ region contain repetitive units.

Determination of the structure and complete sequence of the wild-type *NCF1* gene provides an important foundation for further investigation of the molecular basis of p47-phox-deficient CGD. This autosomal recessive form of CGD is four to five times more common than the other autosomal forms of CGD (due to a deficiency of either p22-phox or p67-phox) and bears the unusual feature of an overrepresentation of a single mutation, the GT deletion at the beginning of exon 2. Recently, we have discovered the existence of highly homologous pseudogenes, which co-localize to the same region of chromosome 7q11.23 as the wild-type, *NCF1* gene (23,24). Localization of the gene and pseudogene to the same region of chromosome 7q11.23 is particularly important for analyses of patients with p47-phox-deficient CGD because it has been proposed that recombination events between the gene and pseudogenes occur in p47-phox-deficient CGD patients who possess only the ΔGT sequence at the start of exon 2 (31).

Analysis of the exonic regions is consistent with the reported cDNA sequence at the genomic level (4,6). The size and distribution of exons is unremarkable. Analysis of the complete sequences for each of the ten introns is remarkable for several reasons. The shortest intron, number 3, is only 99 bases long and its predicted translation product is in frame with the flanking exons. However, we have been unable to identify an alternatively spliced form of *NCF1* that contains this 99 base fragment by Northern blot analysis nor by RT-PCR. Overall, the intronic sequence is remarkable for a high density of repetitive sequences, including 21 *Alu* sequences resulting in a density of 1.4 *Alu* repeats every kilobase of intron sequence. Moyzis *et al.* have estimated the average density for *Alu* sequences to be at least one every four kilobases (34). Furthermore, it is remarkable that all nine *Alu* sequences in introns 1 and 2 are in the 3′ to 5′ orientation. In this region, there is one *Alu* sequence every five hundred

bases. A high density of *Alu* sequences, particularly oriented in one direction, has been associated with recombination mutational events in the LDL receptor (35) and the complement C1q inhibitor gene (36). Therefore, it is interesting to note that the most common form of p47-phox-deficient CGD results from recombination events that exchange the ΔGT of the pseudogene for the wild-type GTGT signature sequence in exon 2 (31).

The identification of the 5′ upstream region is notable because the primary sequence has several GC-rich regions in close proximity to the start of transcription; in particular, there is a G + C content of 58.5% in the region between bp +200 to −400 and no consensus sequence for either a TATA or CAAT box. The absence of these motifs differs from the gp91-phox upstream region which has a well defined CCAAT region and a region where the CAT displacement factor binds (18,37,38). However, the PU.1 site has been shown to be critical for myeloid-specific expression of *NCF1* (33) and the site is conserved in both the wild-type and the pseudogene 5′ upstream sequence. A PU.1 binding site has been identified as an important myeloid-specific transcription factor for the expression of the *CYBB* in neutrophils, monocytes and B lymphoblastoid cells (39).

Comparison of the sequence between the wild-type and pseudogene sequence in the vicinity of the start of transcription, reveals no differences, particularly in the region of the start of transcription, mapped by others to within 21–22 bases of the start methionine (33,40). It is notable that the wild-type and pseudogene transcripts are easily detected in all normal healthy controls (24,31). This implies that expression of both genes is regulated by elements common to both the wild type and pseudogene, perhaps present in this highly homologous region. Although one base pair substitution in the pseudogene 5′ upstream sequence is situated within an SP-1 consensus site, the high homology between the gene and pseudogene 5′ upstream sequence over the first 1,960 bp suggests that any major difference in expression between gene and pseudogene transcripts, if it exists, lies elsewhere. Ongoing studies may further delineate additional cis-acting elements in the transcriptional regulation of the *NCF1* gene.

In conclusion, a complete map of the wild-type *NCF1* gene is now presented and establishes a foundation for further characterization of the sites of recombination responsible for p47-phox CGD. In addition, this sequence will also be useful in mapping the 7q11.23 region, a complex region that has eluded linear mapping because of the high density of large blocks of sequence. This region has also been of interest to groups studying the molecular basis of Williams syndrome, which maps to the same region (41,42). A preliminary model for this region suggests that the wild-type gene is telomeric and at least two pseudogenes lie closer to the centromere. The Williams locus is interposed between the wild-type and pseudogene loci (41–44).

Lastly, knowledge of the wild-type genomic structure will also permit further investigation of the regulation of *NCF1* expression in hematopoietic cells, such as myeloid and B lymphoid cells. The sequence of the promoter region might also be important for lineage specific expression of *NCF1* in future gene therapy (45,46). In addition, this data might be potentially useful for the study of the regulation and function of the p47-phox gene in non hematopoietic cells, including fibroblasts, mesangial cells, HepG2 cells, neuroepithelial cells and type I cells of the carotid body (47–51).

ACKNOWLEDGMENTS

We thank Drs. Mary Dinauer and Stuart Orkin for their advice and comments. This study was supported in part by NIH Grant AI 24838, a Faculty Development Award, an Immune Deficiency Society, Cutter Biologic Award, and the Stein Endowment Fund. A.G. was supported by an Otto-Hahn Fellowship (Max-Planck Society, Germany) and a DAAD Fellowship (DAAD, Germany). Sequence data have been submitted to GenBank and are accessible with the following numbers: AF184614 for the complete wild-type p47-phox and AF184613 for the 5′ upstream region of the pseudogene clone L14.

doi:10.1006/bcmd.2000.0274, available online at http://www.idealibrary.com on IDEAL®

REFERENCES

1. Halliwell B. Oxidant and human disease: Some concepts. *FASEB J* 1:358–364, 1987.
2. Babior BM. NADPH oxidase: An update. *Blood* 93: 1464–1476, 1999.
3. Clark R, Volpp B, Leidal K, Nauseef W. Two cytosolic components of the human neutrophil respiratory burst oxidase translocate to the plasma membrane during cell activation. *J Clin Invest* 85:714–721, 1990.
4. Lomax K, Leto T, Nunoi H, Gallin JI, Malech HL. Recombinant 47-kilodalton cytosol factor restores NADPH oxidase in chronic granulomatous disease. *Science* 245:409–412, 1989.
5. Leto T, Lomax K, Volpp B, *et al.* Cloning of a 67-kDa neutrophil oxidase factor with similarity to a noncatalytic region of p60 c-src. *Science* 248:727–730, 1990.
6. Volpp B, Nauseef W, Donelson J, Moser D, Clark RA. Cloning of the cDNA and functional expression of the 47-kilodalton cytosolic component of human neutrophil respiratory burst oxidase. *Proc Natl Acad Sci USA* 86:7195–7199, 1989. [Correction: Volpp B. *Proc Natl Acad Sci USA* 86:9563, 1989]
7. Segal A, Heyworth P, Cockcroft S, Barrowman M. Stimulated neutrophils from patients with autosomal recessive chronic granulomatous disease fail to phosphorylate a Mr-44,000 protein. *Nature* 316:547–549, 1985.
8. Hayakawa T, Suzuki K, Suzuki S, Andrews P, Babior B. A possible role for protein phosphorylation in the activation of the respiratory burst in human neutrophils. *J Biol Chem* 261:9109–9115, 1986.
9. Faust L, El Benna J, Babior B, Chanock S. The phosphorylation targets of p47-phox, a subunit of the respiratory burst oxidase: Functions of the individual target serines as evaluated by site-directed mutagenesis. *J Clin Invest* 96:1499–1505, 1995.
10. Abo A, Pick E, Hall A, Totty N, Teahan C, Segal A. Activation of the NADPH oxidase involves the small GTP-binding protein p21rac. *Science* 353:668–670, 1991.
11. El Benna JE, Ruedi J, Babior B. Cytosolic guanine nucleotide-binding protein rac2 operates *in vivo* as a component of the neutrophil respiratory burst oxidase. *J Biol Chem* 269:6729–6734, 1994.
12. Dorseuil O, Vasquez A, Lang P, Bertoglio J, Gacon G, Leca G. Inhibition of superoxide production in B lymphocytes by rac antisense oligonucleotides. *J Biol Chem* 267:20540–20547, 1992.
13. Knaus U, Heyworth P, Evans P, Kinsella R, Curnutte JT, Bokoch G. rac 2 is a regulator of phagocyte oxygen radical production. *Science* 254:1512–1515, 1991
14. Kwong C, Malech HL, Rotrosen D, Leto T. Regulation of the human neutrophil NADPH oxidase by rho-related G-proteins. *Biochemistry* 32:5711–5717, 1993.
15. Curnutte JT. Chronic granulomatous disease: The solving of a clinical riddle at the molecular level. *Clin Immunol Immunopathol* 67:S2–S15, 1993.
16. Thrasher A, Keep N, Wientjes F, Segal A. Chronic granulomatous disease. *Biochim Biophys Acta* 1227: 1–24, 1994.
17. Roos D, de Boer M, Kuribayashi F, *et al.* Mutations in the X-linked and autosomal recessive forms of chronic granulomatous disease. *Blood* 87:1663–1681, 1996.
18. Skalnik D, Dorfman D, Perkins A, Jenkins N, Copeland N, Orkin SH. Targeting of transgene expression to monocyte/macrophages by the gp91-phox promoter and consequent histiocytic malignancy. *Proc Natl Acad Sci USA* 88:8505–8509, 1991.
19. Dinauer M, Pierce E, Bruns G, Curnutte JT, Orkin SH. Human neutrophil cytochrome b light chain (p22-phox): Gene structure, chromosomal location, and mutation in cytochrome-negative autosomal recessive chronic granulomatous disease. *J Clin Invest* 86:1729–1737, 1990.
20. Kenney R, Malech HL, Epstein N, Roberts R, Leto T. Characterization of the p67-phox gene: Genomic organization and restriction fragment length polymorphism analysis for prenatal diagnosis in chronic granulomatous. *Blood* 82:3739–3744, 1993.
21. Royer-Pokora B, Kunkel L, Monaco A, *et al.* Cloning of the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location. *Nature* 322:32–38, 1986.
22. Zhan S, Vazquez N, Zhan S, *et al.* The genomic structure, chromosomal localization, tissue specificity and regulation of p40-phox gene. *Blood* 88:2714–2721, 1996.
23. Francke U, Hsieh C-L, Foellmer B, Lomax K, Malech HL, Leto T. Genes for two autosomal recessive forms of chronic granulomatous disease assigned to 1q25 (NCF2) and 7q11.23 (NCF1). *Am J Hum Genet* 47: 483–492, 1990.
24. Görlach A, Lee P, Roesler J, Hopkins P, Christensen BL, Green ED, Chanock SJ, Curnutte JT. The p47-phox pseudogene carries the most common mutation causing p47-phox deficient chronic granulomatous disease. *J Clin Invest* 100:1907–1918, 1997.
25. Rae J, Newburger P, Dinauer M, Noack D, Hopkins P, Kuruto R, Curnutte J. X-linked chronic granulomatous disease: Mutations in the CTBB gene encoding the gp91-phox component of respiratory-burst oxidase. *Am J Hum Genet* 62:1320–1331, 1998.
26. Roesler J, Heyden S, Burdelski M, *et al.* Uncommon missense and splice mutations and resulting biochemical phenotypes in German patients with X-linked

chronic granulomatous disease. *Exp Hematol* 27:505–511, 1999.

27. Volpp B. Nauseef WM, Clark RA. Two cytosolic neutrophil oxidase components absent in autosomal chronic granulomatous disease. *Science* 242:1295–1297, 1988.
28. Casimir C, Bu-Ghanim H, Rodaway A, Bentley D, Rowe P, Segal A. Autosomal recessive chronic granulomatous disease caused by a deletion at a dinucleotide repeat. *Proc Natl Acad Sci USA* 88:2753–2756, 1991.
29. Volpp B, Lin Y. *In vitro* molecular reconstitution of the respiratory burst in B lymphoblasts from p47-phox-deficient chronic granulomatous disease. *J Clin Invest* 91:201–207, 1993.
30. Iwata M, Nunoi H, Yamazaki H, *et al.* Homologous dinucleotide (GT or TG) deletion in Japanese patients with chronic granulomatous disease with p47-phox deficiency. *Biochem Biophys Res Commun* 199:1372–1377, 1994.
31. Roesler J, Curnutte J, Rae J, Barrett D, Patino P, Chanock S, Görlach A. Recombination events between the p47-phox gene and its highly homologous pseudogenes are the main cause of autosomal recessive chronic granulomatous disease (CGD). *Blood* 95: 2150–2156, 2000.
32. Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* 23:4878–4884, 1995.
33. Li S-L, Valente A, Zhao S-J, Clark R. PU.1 is essential for p47-phox promoter activity in myeloid cells. *J Biol Chem* 272:17802–17809, 1997.
34. Moyzis RK, Torney DC, Meyne J. The distribution of interspersed repetitive DNA sequences in the human genome. *Genomics* 4:273–289, 1989.
35. Lehrman MA, Russel DW, Goldstein JL, and Brown MS. Alu–Alu recombination deletes splice acceptor cites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. *J Biol Chem* 262:3354–3361, 1997.
36. Stoppa-Lyonnet D, Carter PE, Meo T, and Tosi M. Clusters of intragenic Alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements. *Proc Natl Acad Sci USA* 87:1551–1555, 1990.
37. Skalnik D, Strauss E, Orkin SH. CCAAT displacement protein as a repressor of the myelomonocytic specific gp91-phox gene promotor. *J Biol Chem* 266: 16736–16744, 1991.
38. Neufeld EJ, Skalnik D, Lievens P, Orkin SH. Human CCAAT displacement protein is homologous to the Drosophila homeoprotein, cut. *Nat Genet* 1:50–55, 1992.
39. Suzuki S, Kumatori A, Haagen I-A, *et al.* PU.1 as an essential activator for the expression of gp91-phox gene in human peripheral neutrophils, monocytes and B lymphocytes. *Proc Natl Acad Sci USA* 95:6085–6090, 1998.
40. Rodaway A, Teahan C, Casimir C, Segal A, Bentley D. Characterization of the 47-kilodalton autosomal chronic granulomatous disease protein: Tissue specific expression and transcriptional control by retinoic acid. *Mol Cell Biol* 10:5388–5396, 1990.
41. DeSilva U, Massa H, Trask B, Green E. Comparative mapping of the region of human chromosome 7 deleted in Williams syndrome. *Genome Res* 9:428–436, 1999.
42. Hockenhull E, Carette M, Metcalfe K, Donnai D, Read A, Tassabehji M. A complete physical contig and partial transcript map of the Williams syndrome critical region. *Genomics* 58:138–145, 1999.
43. Bouffard G, Idol J, Braden V, *et al.* A physical map of human chromosome 7: An integrated YAC contig map with average STS spacing of 79 kb. *Genome Res* 7:673–692, 1997.
44. Perez L, Wang Y-K, Peoples R, Coloma A, Cruces J, Francke U. A duplicated gene in the breakpoint regions of the 7q11.23 Williams–Beuren syndrome deletion encodes the initiator binding protein TFII-I and BAP-135, a phosphorylation target of BTK. *Hum Mol Genet* 7:325–334, 1998.
45. Malech H, Maples P, Whiting-Theobald N, *et al.* Prolonged production of NADPH oxidase-corrected granulocytes after gene therapy of chronic granulomatous disease. *Proc Natl Acad Sci USA* 94:12133–12138, 1997.
46. Malech H, Bauer T, Hickstein D. Prospects for gene therapy of neutrophil defects. *Semin Hematol* 34:355–361, 1997.
47. Acker H, Dufau E, Huber J, Sylvester D. Indications to an NADPH oxidase as a possible pO_2 sensor in rat carotid body. *FEBS Lett* 256:75–78, 1989.
48. Görlach A, Holtermann G, Jelkmann W, Hancock J, Jones O, Acker H. Photometric characteristics of haem proteins in erythropoietin-producing hepatoma cells (HepG2). *Biochem J* 290:771–776, 1993.
49. Meier B., Cross A, Hancock J, Jones FKO. Identification of a superoxide-generating NADPH oxidase system in human fibroblasts. *Biochem J* 275:241–245, 1991.
50. Radeke H, Cross A, Hancock J. Functional expression of NADPH oxidase components (alpha and beta subunits of cytochrome b558 and 45 kDa flavoprotein) by intrinsic human glomerular mesangial cells. *J Biol Chem* 266:21025–21029, 1991.
51. Youngson C, Nurse C, Yeger H, Cutz E. Oxygen sensing in airway chemoreceptors. *Nature* 365:153–155, 1993.