In Silico comparative analysis of DNA and Amino Acid sequence for Alport syndrom gene family

Abdulla A. AL-Harthi[1], Ayman M. Sabry[1&2] and Manal M. Said[3]

[1]Biotechnology and Genetic Engineering Unit, Scientific Research Deanship, Taif University, TAIF, Zip Code 21974, Post Box 888 KSA.
[2]Cell Biology Department, National Research Center, Dokki, Giza, Egypt
[3]Dept. of Biotechnology, Faculty of Science, Taif University
amsabry@gmail.com

Abstract: Alport syndrome is a hereditary renal disorder caused by mutation in COL4A genes. In the present study, *in silico* comparative analysis of nucleotides sequence was preformed to shade some insight into characteristics of the COL4A gene family nucleotide structure. A total of 14 CDS variants were collected from Genbank for Alport syndrome genes, namely COL4A3, COL4A4, COL4A5, COL4A6 of COL4A gene family. The sequence length of these genes varied greatly where the longest sequence belongs to COL4A3 (10152 bp), where the shortest was for COL4A5 (2088 bp). Stop codon did not vary within the variants of the same gene except for the COL4A5 where it's stop codon varied. CG content of COL4A genes ranged from 55 to 59%, with COL4A4 has the highest GC content 59%. The results of the DnaSp analysis indicated that the selected region (15244) of the14 sequences from different genes (*and their alleles*) have 1484 sites excluding sites with gaps Sites with alignment gaps or missing data: 3760. There are 339 invariable sites and 1145 variable sites include 96 singletone variable site and 1049 parsimony informative sites. The nucleotide diversity (π= 0.38) and the average number of nucleotide differences (K = 563). A total of 25 conserved regions were detected in 14 aligned sequences with the smallest being 65 nucleotides long and the largest 310 nucleotides long. These conserved regions are unevenly distributed throughout the COL4A genes. The logo analysis of the amino acid sequence of COL4A gene family showed that the conserved regions are not evenly distributed. To our knowledge, this is the first attempt to carry out an *in silico* comparison for the nucleotides sequence and amino acid for all five genes of Alport syndrome. This work shades some insight into the assessment and characteristics of the nucleotides and amino acid sequences of COL4A gene family. That is, the advances in the next generation sequencing provided a wider insight about the nature of nucleotide structure and amino acids of COL4A genes.
[Abdulla A. AL-Harthi, Ayman M. Sabry and Manal M. Said. ***In Silico* comparative analysis of DNA and Amino Acid sequence for Alport syndrom gene family.** Life Sci J 2013; 10(3): 1373-1379] (ISSN: 1097-8135).http://www.lifesciencesite.com 207

Keywords: Alport Syndrome, Sequence analysis, Logo analysis

1. Introduction

Alport syndrome (MIM #301050, AS) is a hereditary disorder of the basement membrane, resulting in progressive renal failure due to glomerulonephropathy, variable sensorineural hearing loss, and ocular anomalies (Gross *et al.* , 2002). The disease is characterised by family history of hematuria, progressive renal failure, sensorinural deafness and typical ocular changes anomalies (Alport, 1927 and Flinter *et al.,* 1988). The pathogenesis of AS is related to the defect in the COL4A3, COL4A4 , COL4A5 and COL4A6 genes encoding type IV collagen α-chain isoforms (α3, α4 and α5). These four genes are necessary for proper development of the glomerular basement membranes, which plays a crucial role in the purification of blood plasma in the kidney (Kalluri *et al.,* 1977). AS is a genetically heterogeneous disease. About 85% of AS patients have X-linked AS (XLAS), caused by mutations in the COL4A5 gene, which is located on chromosome X (Kalluri *et al.,* 1977 and Barker, 1990). COL4A3 and COL4A4 genes, located on chromosome 2, are involved in approximately 15% of autosomal recessive AS cases and the rarer autosomal dominant forms of AS ((Jefferson, 1997), Mochizuki, 1994 and van der Loop, 2000). XLAS males have severer phenotypes and usually progress to end-stage renal disease (ESRD), whereas the affected females, heterozygous for the COL4A5 mutant gene, have more variable phenotypes, from microscopic hematuria to ESRD (Kashtan, 2000). COL4A6 is also located on X chromosome. This gene encodes one of the six subunits of type IV collagen, the major structural component of basement membranes. Like the other members of the type IV collagen gene family, this gene is organized in a head-to-

head conformation with another type IV collagen gene, α5 type IV collagen, so that the gene pair shares a common promoter. Deletions in the α5 gene that extend into the α6 gene result in diffuse leiomyomatosis accompanying the X-linked Alport syndrome caused by the deletion in the α5 gene. Two splice variants have been identified for this gene (Uliana *et al.*, 2011)

Mutations in COL4A4 and COL4A3 genes have been reported in both autosomal recessive and autosomal dominant ATS. The conventional mutation screening, performed by DHPLC and/or Sanger sequencing, is time- consuming and has relatively high costs because of the absence of hot spots and to the high number of exons per gene: 51 (COL4A5), 48 (COL4A4) and 52 (COL4A3) (Artuso *et al.* , 2012). To date, 688 COL4A5 mutations have been identified according to the Human Gene Mutation Database (HGMD Professional 2012.1. Release date 30th March 2012) (http://www.hgmd.org/). certain correlations between genotypes and phenotypes have been established in males (Jais, 2000, Gross *et al.,* 2002 and Bekheirnia, 2010,). Patients with large deletions, nonsense or frameshift mutations demonstrated severer symptoms as compared to those with missense or splice site mutations. However, such genotype–phenotype correlations have not been found in females with variable phenotypes, even among family members (Jais, 2003).

Diagnosis of Alport syndrome is complex and requires urinalysis, renal function studies, audiometry, ophthalmic evaluation, and skin and/or kidney biopsy. Molecular testing for mutations in the COL4A5 gene is useful for diagnosis of XLAS as other diagnostic methods may be inconclusive in the early stages of renal disease. As Alport Syndrome gene family are targeting the same organ (*kidney),* one can therefore hypothesize that are a certain amount of molecular similarities among the sequences of these genes. So, *if yes*, what is the amount of this similarity? and how this similarity or dissimilarity affecting the expression of these genes. In the present study, *in silico* comparative analysis of nucleotides sequence was preformed to shade some insight into characteristics of the COL4A gene family nucleotide structure. Particularly, this study compared: (i) the nucleotides structure to characterize and assess, polymorphic sites, GC content and conserved regions (ii) amino acid sequence using logo analysis of COL4A gene family.

2. Material and Methods

Fourteen complete coding sequence (CDS) of the 4 genes affecting Alport syndrome 1 were collected from GenBank. These CDS's were obtained through seqinr R package (Charif & Lobry, 2007) . All sequences were aligned using the Clustal W program (Larkin *et al.* , 2007). DnaSP (version 5.10.01) software was used to analyze the haplotype diversity (H_d), the average number of nucleotide differences, the average number of nucleotide differences (Tajima, 1983), the nucleotide diversity (π), synonymous nucleotide diversity (π_S), non- synonymous nucleotide diversity (π_a) with the Jukes and Cantor correction, the polymorphic site (S), the singleton variable sites (SP), and the parsimony informative sites (PIP) (Lynch & Crease, 1990). Nucleic acid logo (Schneider & Stephens, 1990) analysis was carried out using SeqLogo package (Bembom, 2013)

3.Results and Discussion

A total of 14 CDS variants were collected from Genbank for Alport syndrome genes, namely COL4A3, COL4A4, COL4A5, COL4A6 (Table 1). All members of COL4A gene family were found to have only 2 variants on the Genbank where COL4A5 has 8 variants. The sequence length of these genes varied greatly where the longest sequence belongs to COL4A3 (10152 bp), where the shortest was for COL4A5 (2088 bp).

Stop codon did not vary within the variants of the same gene except for the COL4A5 where it's stop codon varied. This variation may be ascribed to the large number of variants for COL4A5 compared to the other members of Alport syndrome gene family. That is, five variants have stop codon TAA, three variants have stop codon TAG and one variant has stop codon TGG.

In general the analysis of mammalian chromosome sequences revealed complex genomic landscapes: some regions of the genome are very gene-rich, whereas some other large regions are devoid of genes (Lander *et al.,* 2001). The analysis of nucleotides structure of COL4A5 genes revealed its high CG content (Zhou *et al.,* 1994). In this study, the *in silico* analysis showed that CG content of COL4A genes ranged from 55 to 59 % (table 1), with COL4A4 has the highest GC content 59%. The variation in GC content of COL4A5 was higher than that for the rest of COL4A gene family members, that is CG content for COL4A3, COL4A4, and COL4A5

was 56, 59, and 57%. where for COL4A5 CG content ranged between 55 to 57%.

The alignment of 14 sequences with region of 5244 bp and containing gaps was carried out using Clustal W program (Larkin *et al.* , 2007). The results of the DnaSp analysis indicated that the selected region (15244) of the14 sequences from different genes (*and their alleles*) have 1484 sites excluding sites with gaps Sites with alignment gaps or missing data: 3760. There are 339 invariable sites and 1145 variable sites include 96 singletone variable site and 1049 parsimony informative sites. The nucleotide diversity (π= 0.38) and the average number of nucleotide differences (K = 563). Polymorphic information and haplotype diversity of COL4A gene for informative species are presented in table (2)

Table 1: Accession numbers, sequence length and stop codons of COL4A gene family

Gene	Accession Number	Length	GC Content%	Stop Codon
COL4A3	NM000091	5013	0.56	TGA
COL4A3	X80031	5013	0.56	TGA
COL4A4	NM000092	10152	0.59	TAG
COL4A4	X81053.1	5073	0.59	TAG
COL4A5	NM033380	5076	0.55	TAA
COL4A5	NM000495	5058	0.55	TAA
COL4A5	BC035387	2088	0.56	TAG
COL4A5	M90464	2731	0.56	TGG
COL4A5	BC151846	5076	0.55	TAA
COL4A5	M58526	4816	0.55	TAA
COL4A5	M31115	2319	0.55	TAA
COL4A5	D21337	5037	0.57	TAG
COL4A6	NM033641	5073	0.57	TAG
COL4A6	NM001847	5076	0.57	TAG

Table 2: Genetic diversity of the COL4A gene Diversity parameter [a]

	h	H_d	K	π
COL4A	9	0.9	1970	0.38

[a] h, Number of haplotypes; H_d, haplotype diversity; K, average number of nucleotide differences; π, nucleotide diversity

Conserved DNA regions were detected among the 14 sequences of the COL4A gene family (Table 3). A total of 25 conserved regions were detected in 14 aligned sequences with the smallest being 65 nucleotides long and the largest 310 nucleotides long. These conserved regions are unevenly distributed throughout the COL4A genes. The vast majority of the conserved regions are less than 100 nucleotides long (Figure 1). Table 3 shows that conservation Index (proportion of conserved columns). In other words, conservation (C) is measured as the proportion of conserved sites in the alignment region. The conservation ranged from 0.31 to 0.35, all conserved regions were statistically significant where α ranged between less than 0.01 than 0.03 (un- der the hypergeometric distribution) Table 3 also shows, H, Homozigosity (1-Heterozigosity) where H ranged between 0.60 to 0.67.

Table 3: Conserved Regions , length , conservation, homozigosity and *P*- values of COL4A gene family

Region	Start-End	Length	Conservation	Homozigosity	*P*-value
1	225-534	310	0.35	0.66	< 0.01
2	568-786	219	0.33	0.67	< 0.01
3	825-897	73	0.31	0.63	0.03
4	828-899	72	0.32	0.63	0.03
5	864-1088	225	0.31	0.64	0.01
6	1213-1289	77	0.31	0.66	0.03
7	1453-1522	70	0.31	0.66	0.03

8	1459-1528	70	0.31	0.67	0.03
9	1465-1530	66	0.32	0.67	0.03
10	1558-1696	139	0.31	0.64	0.01
11	1636-1735	100	0.31	0.66	0.02
12	1665-1736	72	0.32	0.68	0.02
13	1757-1822	66	0.32	0.68	0.02
14	2169-2261	93	0.31	0.69	0.02
15	2195-2271	77	0.31	0.68	0.03
16	2205-2279	75	0.32	0.69	0.02
17	3456-3711	256	0.32	0.60	< 0.01
18	3716-3782	67	0.31	0.63	0.03
19	3719-3788	70	0.31	0.63	0.03
20	3722-3801	80	0.31	0.64	0.03
21	3735-3802	68	0.32	0.65	0.02
22	3742-3903	162	0.31	0.63	< 0.01
23	4294-4360	67	0.31	0.60	0.03
24	4297-4401	105	0.31	0.60	0.01
25	4345-4409	65	0.32	0.60	0.02

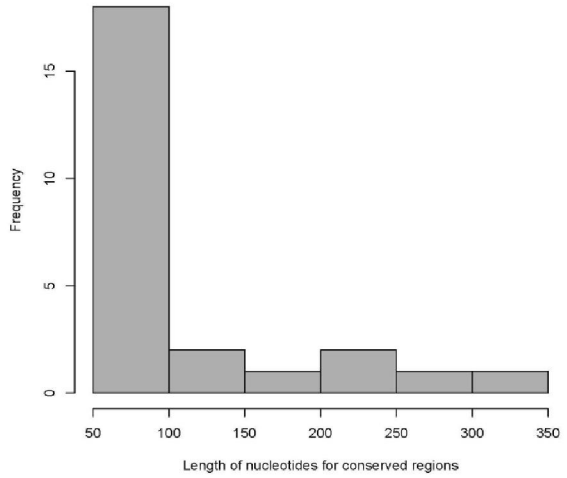
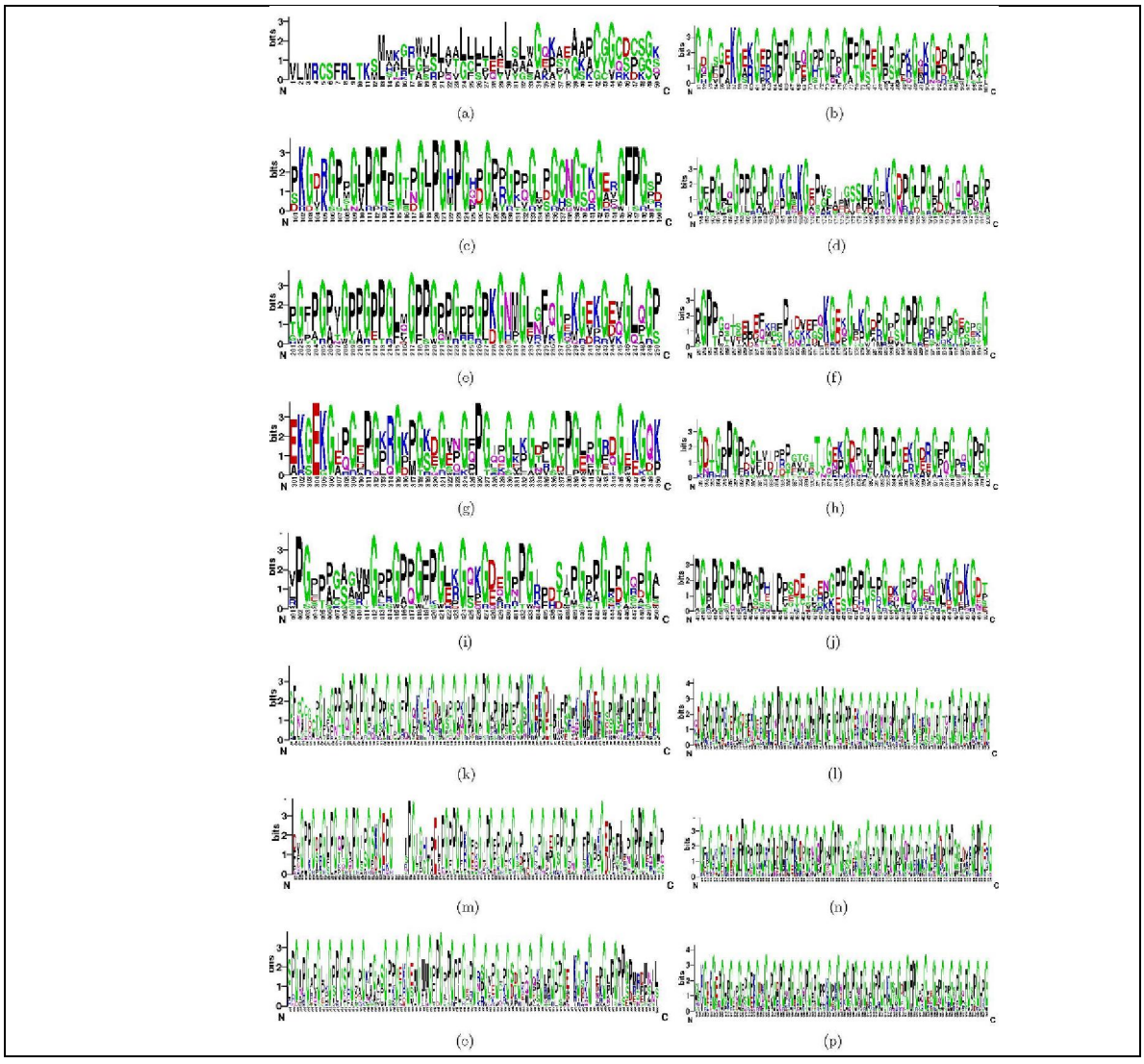


Figure 1: Distribution of conserved regions in COL4A gene family

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Schneider & Stephens (1990). Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence. The logo analysis of the amino acid sequence of COL4A gene family showed that the conserved regions are not evenly distributed. This result is supporting the earlier findings of nucleotide sequence analysis. Apart from the first 50 positions, it can be seen from figure (2) that Glycine representing the most of conserved amino acid where Proline is also conserved but to a lesser extent.

To our knowledge, this is the first attempt to carry out an *in silico* comparison for the nucleotides sequence and amino acid for all five genes of Alport syndrome. The large size and rich GC content of COL4A genes have

presented diagnostic laboratories with problems in identifying mutations with greater than a 50% success rate. For example (King *et al.*, 2006) identified twenty-one mutations in twenty-five patients with clear Alport syndrome. This work shades some insight into the assessment and characteristics of the nucleotides and amino acid sequences of COL4A gene family. That is, the advances in the next generation sequencing provided a wider insight about the nature of nucleotide stutter and amino acids of COL4A genes.



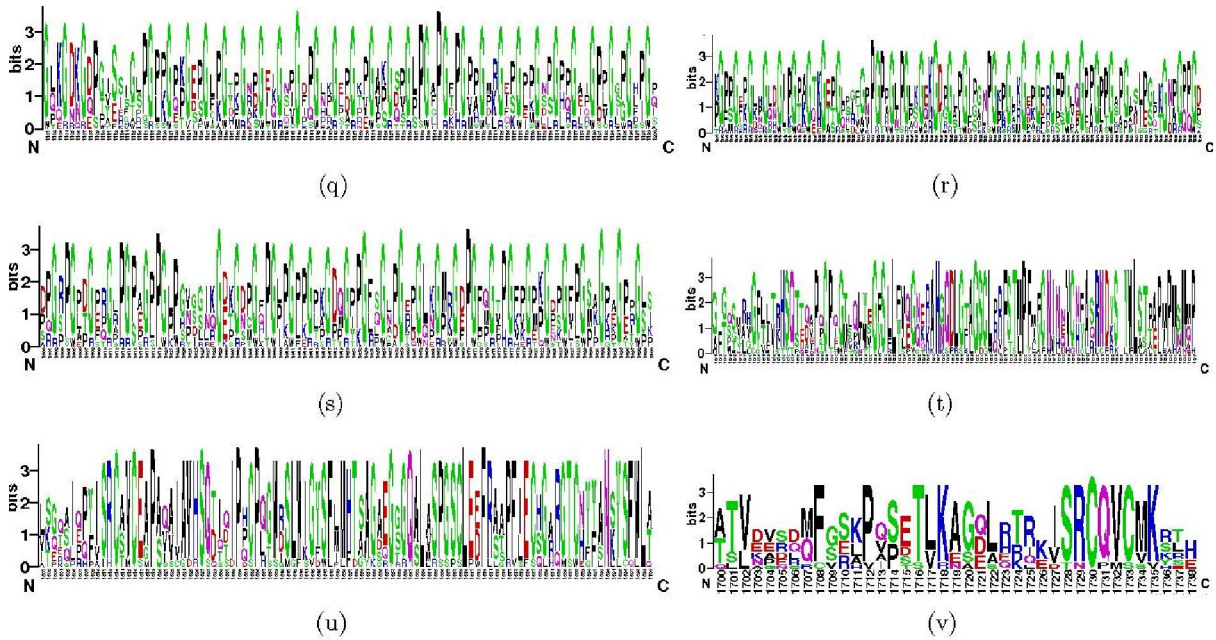


Figure 2: Logo representation of the degree of amino acid conservation along the 14 protein sequences. The height of each letter is proportional to the degree of conservation of each amino acid (in one letter code). Numbers below each amino acid represent their position in the protein sequence.

References

1. Alport, AC. 1927. Hereditary familial congenital haemorrhgic Nephritis. *Br. Med. J*, **1**(504–507).
2. Artuso, R., Fallerini, C., Dosa, L., Scionti, F., Clementi, M., Garosi, G., Massella, L., Epistolato, M.C., Mancini, R., Mari, F., Longo, I, Ariani, F., Renieri, A., & Bruttin, M. 2012. Advances in Alport syndrome diagnosis using next-generation sequencing. *European Journal of Human Genetics*, **20**, 50–57.
3. Barker, D.F., SL Hostikka, J Zhou, *et al.* 1990. Identification of mutations in the COL4A5 collagen gene in Alport syndrome. *Science*, **284**, 1224–1227.
4. Bekheirnia, M.R., Reed B, Gregory MC *et al.,* 2010. Genotypephenotype correlation in X- linked Alport syndrome. *J. Am. Soc. Nephrol.*, **21**, 876–883.
5. Bembom, Oliver. 2013. *seqLogo: Sequence logos for DNA sequence align- ments*. R package version 1.26.0.
6. Charif, D., & Lobry, J.R. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *Pages 207–232 of:* Bastolla, U., Porto, M., Roman, H.E., & Vendruscolo, M. (eds), *Structural approaches to se- quence evolution: Molecules, networks, population s*. Biological and Medi- cal Physics, Biomedical Engineering. New York: Springer Verlag. ISBN : 978-3-540-35305-8.
7. Flinter, FA., Cameron, JS., Chantler, C., Housten, I., & Bo- brow, M. 1988. Gnetics of classic Aport's syndrome. *Lancet*, **ii**, 1005– 1007.
8. Gross, O., Netzer, K.O., Lambrecht, R., Seibold, S., & Weber, M. 2002. Meta-analysis of genotypephenotype correlation in X-linked Al- port syndrome: impact on clinical counseling. *Nephrol. Dial. Transplant*, **17**, 1218–1227.
9. Jais, J.P., Knebelmann B, Giatras I, *et al.,* 2000. X-linked Alport syndrome: natural history in 195 families and genotype phenotype correlations in males. *J. Am. Soc. Nephrol.*, **11**, 649-657.
10. Jais, J.P., ., Knebelmann B, Giatras I, *et al.,* 2003. X-linked Alport syndrome: natural history and genotype phenotype correlations in girls and women belonging to 195 families: a European Community Alport Syndrome Concerted Action study. *J. Am. Soc. Nephrol.*, **14**, 2603–2610.
11. Jefferson, J.A., Lemmink HH, Hughes AE *et al.,* 1997. Autosomal dominant Alport syndrome linked to the type IV collage alpha

3 and alpha 4 genes (COL4A3 and COL4A4). *Nephrol. Dial. Transplant*, **12**, 1595–1599.
12. Kalluri, R., Shield, C.F., Todd, P., Hudson, B.G., & Neilson, E.G. 1977. Isoform switching of type IV collagen is developmentally ar- rested in X-linked Alport syndrome leading to increased susceptibility of renal basement membranes to endoproteolysis. *J. Clin. Invest*, **99**, 2470–2478.
13. Kashtan, C.E. 2000. Alport syndromes: phenotypic heterogeneity of pro- gressive hereditary nephritis. *Pediatr. Nephrol.*, **14**, 502–512.
14. King, K., Flinter, F. A., & Green, P. M. 2006. A Tow-Tier approach to mutation detected in the COL4A5 gene for Alport Syndrome. *Human Mutation in Brief*, 1–8.
15. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., & et al. 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
16. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., & Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics Applications Note*, **23**(21), 2947–2948.
17. Lynch, M., & Crease, T.J. 1990. The analysis of population survey data on DNA sequence variation. *Mol Biol Evol*, **7**, 377–394.
18. Mochizuki, T., et al. 1994. Identification of mutations in the alpha 3(IV) and alpha 4(IV) collagen genes in autosomal recessive Alport syndrome. *Nat. Genet.*, **8**, 77–81.
19. Schneider, TD., & Stephens, RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res*, **18**, 6097–6100.
20. Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.
21. Uliana, V., Marcocci, E., Mucciolo, M., Meloni, I., Izzi, C., Manno, C., Bruttini, M., Mari, F., Scolari, F., Renieri, A., & Salviati, L. 2011. Alport syndrome and leiomyomatosis: the first deletion extending beyond COL4A6 intron 2. *Pediatr. Nephrol*, **26**(5), 717–724.
22. van der Loop, F.T., et al. 2000. Autosomal dominant Alport syndrome caused by a COL4A3 splice site mutation. *Kidney Int.*, **58**, 1870–1875.
23. Zhou, J., Leinonen, A., & Tryggvason, K. 1994. Structue of the human type IV collagen COL4A5. *J. Biol. Chem.*, **269**, 6608–6614.

7/21/2013