COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY **18(2)**, 95-101 (2012)

Linkage Analysis as an Approach for Disease-related Loci Identification

Dorota M. Nowak[1], Jose A. Pitarque[2], Andrea Molinari[2], Bassem A. Bejjani[3], Marzena Gajecka[1]*

[1]*Institute of Human Genetics, Polish Academy of Sciences, Poznan, Poland*
[2]*Department of Ophthalmology, Hospital Metropolitano, Quito, Ecuador*
[3]*Revemic Systems, Spokane, Washington, USA*
**e-mail: gamar@man.poznan.pl*

(Received: 10 February 2012; accepted: 6 June 2012; published online: 23 August 2012)

Abstract: Extensive progress in human genetics and clinical diagnostics allowed identification of the majority of genetically-related diseases. Still, the genes responsible for numerous diseases have not been recognized and frequently the disease etiology remains unknown. This is true for keratoconus, the study subject of this article. Therefore, before mutation analysis or other sequence variant assessment, it is essential to identify the chromosomal region, where a gene or genes causing the disease phenotype are located. To achieve that, advanced bioinformatics methods are applied to data obtained from molecular research to narrow the chromosomal region containing the disease locus.

Key words: linkage analysis, multifactorial disease, keratoconus

I. INTRODUCTION

Despite the development of many molecular biology tools, finding the particular gene or genes responsible for any given human genetic disease is still a difficult task. Linkage analysis is one of the possible methods used to find causal genes. Genetic linkage is the tendency of loci or alleles to be inherited together. Genetic loci that are physically close to one another on the same chromosome tend to stay together during meiosis, and are therefore transmitted together to the offspring. The aim of linkage analysis is to find the approximate location of the targeted gene (usually a gene causing an abnormal phenotype) relative to a genetic marker. DNA sequences of known location in the genome that show polymorphisms (variations in size or sequence) in the population are used as genetic markers. The most popular markers are microsatellite and single nucleotide polymorphisms (SNPs).

Genetic Markers

A microsatellite marker is a piece of DNA sequence in which the repeated motif consisting of di-, tri-, or tetranucleotide sequences (e.g., CA, TGA, or CAAG) is present. These markers are usually located in non-coding regions of the genome. Heterozygosity, the frequency of different alleles at the same location on a chromosome (locus), is a measure of their polymorphism. Average heterozygosity for microsatellite markers is approximately 0.70. Initially, analyses were carried out under conditions of low density distribution of markers, i.e. every 10 cM, which gives about 400 markers to be used throughout the human genome. Under conditions of high-density analysis, markers are at intervals of every 5 cM on average which translates to about 800 loci in the genome.

Today, analyses using SNPs markers are more frequent. Single nucleotide polymorphism is the phenomenon of a DNA sequence variability, which consists of changing a single nucleotide (A, T, C or G) at a certain position in DNA sequences between individuals from the same species or between corresponding chromosomes of a given individual. Single nucleotide polymorphism (e.g., GAC**G**TA, GAC**C**TA, GAC**A**TA, and GAC**T**TA) can occur in coding or non-coding sequences of genes, as well as in regions between genes. SNPs in the coding sequence do not neces-

sarily lead to changes in the amino acid sequence of protein due to the degeneracy of the genetic code. The advantage of these markers is its frequent occurrence in the human genome (on average once every 100-300 nucleotides), thus allowing for a detailed analysis of a small DNA fragment.

Family Pedigree

In linkage analysis multigenerational families are studied, which include both affected and unaffected individuals. Figure 1A shows an example of the structure of such examined families. Typically, a pedigree diagram shows the relationship between family members (Fig. 1A). The term "founders" describes individuals whose parents are not specified in the pedigree. Additionally, parents and their offspring are called the nuclear family.

Linkage analysis

Human autosomal chromosomes come in pairs, in which one chromosome is inherited from the mother, and the other one from the father. Each pair of normal/typical chromosomes (without chromosomal rearrangement e.g., duplication, deletion, insertion and inversion) contains the same genes localized in the same order, but the gene sequences may not be identical. These different forms of a gene or a genetic locus are called alleles. Recombination by the crossing-over of chromosomes may occur during meiosis, in which DNA is exchanged between chromosomes of the same pair. Thus two genes that were previously unlinked, being on separate chromosomes, can become linked because of recombination; and conversely, linked genes may become unlinked (Fig. 1B). A recombination occurs more or less at random. If there is a large distance between two DNA sequences on a chromosome, there is a great chance that recombination will occur between them, and the maternal and paternal alleles will be mixed up. In contrast, if two DNA sequences are localized very close on the chromosome, they will recombine infrequently. Disease genes are mapped by measuring recombination against a set of different markers spread over the entire genome or a part of the genome. When recombination occurs frequently, it indicates that the disease gene and marker are far apart. However, other markers are said to be linked to a disease-causing gene because they are localized in close proximity to that disease gene and will not tend to recombine during crossing-over. Markers that flank the disease gene and define a candidate region of the genome usually are localized in a range between 1 and 5 million base pairs (bp). The gene responsible for the studied disease is positioned somewhere in this interval.

Linkage analysis is based on the assumption that loci close to each other are linked, inherited together and do not undergo recombination during crossing-over. To determine whether two loci are linked (Fig. 1A), the frequency of recombination (θ) between the loci is established, and the logarithm of odds (LOD) score is calculated [1]. An LOD score is the decimal logarithm of the ratio of the probability that the marker and disease are closely linked to the likelihood of lack of linkage, which is equal with $\theta = 0.5$. An LOD score greater than 3.3 is considered as an evidence for linkage in a genome-wide scale, while LOD less than -2.0 indicates no linkage. Additionally, an LOD score of 1.9 indicates suggestive linkage [2].

Considering the number of analyzed loci, linkage analysis can be divided into two-point and multipoint. In two-point analysis a marker-disease co-segregation with one locus only at a time is analyzed. However, very often the analyzed markers do not provide sufficient information to determine whether recombination occurs or not. For example, this happens when one or both parents of the genotyped individual are homozygous in certain chromosomal region. In that case identical alleles on both homologous chromosomes are present, and it is impossible to determine whether recombination takes place in that locus. In such case of low informativeness of the markers, the power of the analysis can be increased by the simultaneous analysis of multiple markers located near each other, instead of each separately.

In linkage analysis, parametric (LOD score) and non-parametric (NPL) analyses may be performed. Parametric linkage analysis, as a model-based method, requires a few parameters, including frequency of disease allele, penetrance, and phenocopy rate. Penetrance is the proportion of individuals with a predisposing genotype that developed the studied disease. The term phenocopy describes an individual exhibiting certain phenotype as a result of e.g. environmental factors, although he/she does not have a predisposing genotype. These parameters are easy to obtain for monogenic diseases. However, in the case of multifactorial disorder, in which these parameters are unknown, non-parametric linkage analysis would be more appropriate. The general principle of this approach is that pairs of affected relatives share significantly more alleles at markers in locus linked to the disease than expected by chance. Allele sharing is commonly described by (i) identical by state (IBS) and (ii) identical by descent (IBD). Two alleles in the same genetic locus are IBS if their DNA sequences are identical (Fig 1C), while they are IBD if additionally both alleles are inherited from a recent common ancestor. In the non-parametric approach, methods based on IBD are used more often, because IBD alleles come from a common ancestor shared between pairs of affected

relatives, thus IBD carries more information on linkage. In this type of analysis, IBD alleles probability is used to determine whether the marker is linked to the trait. Linkage would be suggested if the pairs of affected relatives share significantly more alleles IBD than expected by chance. For affected sib-pairs, for sharing 0, 1, and 2 IBD alleles, the expected probabilities are 0.25, 0.5, and 0.25, respectively.

For multipoint analysis of extended families, Kruglyak et al. [3] proposed non-parametric linkage test (NPL), which is based on IBD. However, an NPL test implemented in GENEHUNTER [3] assumes complete IBD information, and in case of incompleteness, the test results in a loss of power. That method has been modified by Kong and Cox [4], to provide accurate likelihood-based tests, and implemented in GNEHUNTER-PLUS [4] and ALLEGRO [5]. Scoring functions S_{pairs} and S_{all} belong to the most commonly used NPL statistics. S_{pairs} is based on numbers of alleles shared IBD by pairs of affected relatives, whereas S_{all} is based on IBD patterns among all affected individuals in a pedigree.

Keratoconus

Keratoconus (KTCN, OMIM 148300) is an eye disorder in which the cornea becomes cone-shaped because of weakening and thinning of its central part, and results in altered refractive powers, and loss of visual acuity. Its prevalence is estimated to be 1:2,000 in the general population [6]. Most diagnosed cases are sporadic, but also familial cases are observed [7]. The reasons for development and progression of this disease remain elusive, despite numerous studies. KTCN is thought to be a multifactorial disorder, whose development involves the participation of both environmental and genetic factors. Among the environmental factors, frequent eye rubbing [8] and contact lenses wearing [9] are mentioned. Also, in some reports, coexistence of KTCN with atopy is presented [10-12]. Genetic studies have led to the identification of several loci on many chromosomes, linked to KTCN, including 1p36.23-36.21, 2p24, 2q13, 3p14-q13, 5q14.3-q21.1, 5q21.2, 5q32-q33, 9q34, 13q32, 14q11.2, 14q24.3, 15q15.1, 15q22.33-24.2, 16q22.3-q23.1, 17p13, 20q12 [13-25]. To date, most of the identified loci have not been replicated in other populations.

II. METHODS

Subjects

Twenty individuals from an Ecuadorian family (KTCN-015) participated in the study. All individuals were examined in the Hospital Metropolitano in Quito, Ecuador, undergoing a complete ophthalmic evaluation. The detailed examination process has been previously described [20]. The possible consequences of the study were explained and informed consent was obtained from all family members, according to the Declaration of Helsinki. The research protocol was approved by the Institutional Review Board at Poznan University of Medical Sciences in Poland.

Linkage analysis

A genome-wide screen was performed by genotyping the KTCN-015 family with fluorescently labeled microsatellite markers, as previously described [20]. PEDSTATS v.0.6.10 [26] was used to identify potential Mendelian inconsistencies. Because the model of inheritance for keratoconus is unknown, whole-genome multipoint non-parametric linkage analyses were performed with the SimWalk2 v.2.91 program [27, 28]. SimWalk2 contains five NPL statistics: BLOCKS, MAX-TREE, ENTROPY, NPL PAIR and NPL ALL (http://www.genetics.ucla.edu/software/simwalk_doc/) for each marker. BLOCKS is most powerful at detecting linkage to a recessive trait, MAX-TREE was designed for traits best modeled by dominant inheritance, ENTROPY is a measure of the entropy of the alleles among the affected cases, and NPL PAIR and NPL ALL are most powerful at detecting linkage to an additive trait. Mega2 v.4.5.1 [29] was used to construct the input files for the SimWalk2 program. The Rutgers Combined Linkage-Physical Map was used to estimate loci genetic map distances [30].

Haplotypes were reconstructed using the SimWalk2 v.2.91 [27, 31] and visualized with HaploPainter v.1.043 program [32].

III. RESULTS

A genome-wide screen of family KTCN-015 indicated one suggestive locus. Table 1 displays NPL multipoint scores for five statistics: BLOCKS, MAX-TREE, ENTROPY, NPL PAIR and NPL ALL in family KTCN-015 for chromosome 11. Maximum NPL PAIR score was obtained for chromosome 11 at marker D11S1760 (NPL PAIR score = 1.667; NPL ALL score = 1.718). Distance between the markers D11S4046 and D11S1760 is 8.73 cM. Such interval might cause the inaccuracies in NPL calculation. Because of that, additional analyses are needed. Haplotype reconstruction allowed to narrow the chromosomal region of interest to a fragment located in locus 11p15.5-p15.4, and flanked by markers D11S1363 and

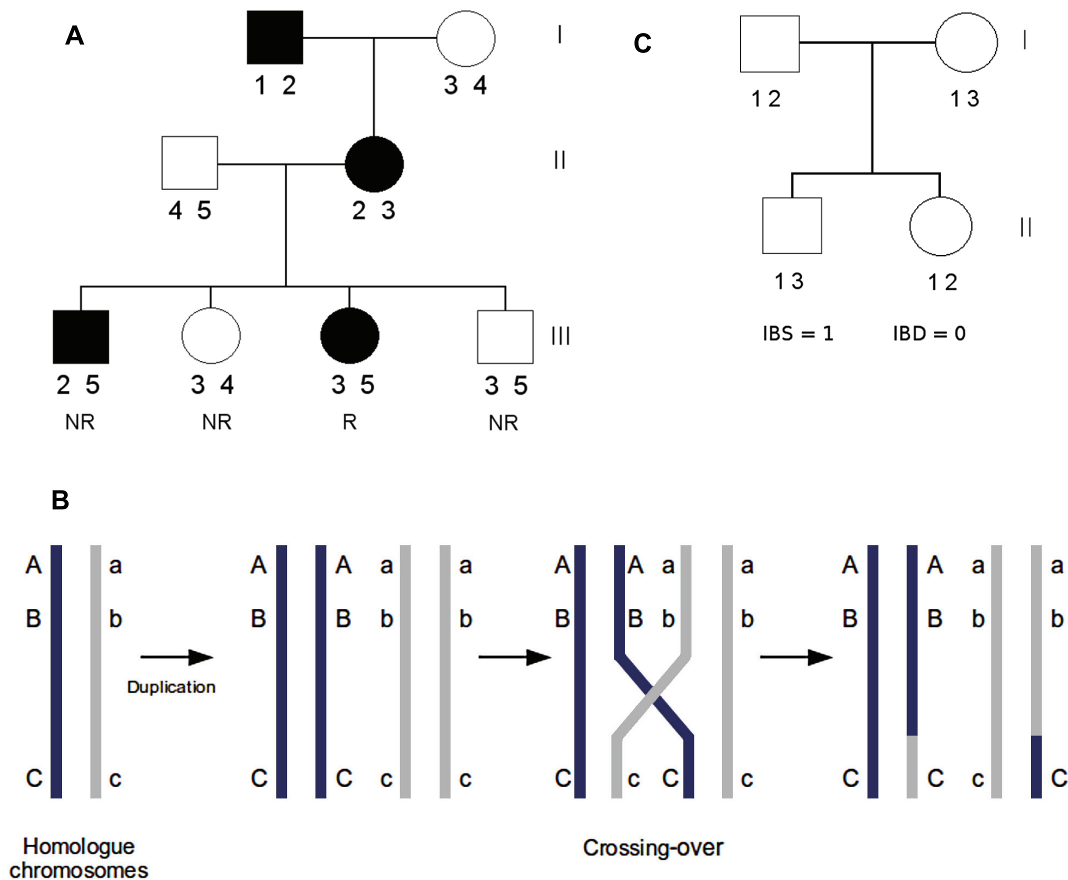

Fig. 1. A. An example of a family pedigree; a symbol of circle points to women and square to men. Black filled symbol indicates an affected status, while the open one – an unaffected individual. Genotypes of grandparents show that allele 2 segregates with the disease phenotype, which allows to determine whether there was a recombination (R) or not (NR) in children genotypes. That information allows for calculation of LOD. C. The diagram presents identity by state (IBS) and identity by descent (IBD) alleles in an exemplary pedigree. In this case alleles IBS = 1, because sibling share allele 1. However, IBS = 0, because allele 1 is transmitted from mother to daughter, and from father to son. B. Schema of a recombination between homologous chromosomes, during crossing-over. In these chromosomes alleles of genes A, B, and C are mapped. Genes in close proximity tend to stay together due to genetic linkage (A and B). Recombination more frequently takes place between genes located far apart (B and C)

Table 1. Multipoint Nonparametric Analysis Results of Selected STR Marker Loci on Chromosome 11 in Family KTCN-015

Marker names	Genetic Position (cM)*	Physical Position (start, bp)**	Bloks	Max-Tree	Entropy	NPL pair	NPL all
D11S1363	4.27	961,991	0.063	0.756	0.361	0.702	0.820
D11S4046	4.96	1,863,635	0.232	1.021	0.867	1.089	1.266
D11S1760	13.69	5,284,337	0.615	1.269	1.404	1.667	1.718
D11S1338	14.84	5,887,912	0.616	1.259	1.401	1.661	1.714
D11S4149	20.45	9,029,910	0.608	1.261	1.391	1.631	1.691
D11S4116	26.91	12,850,599	0.602	0.589	0.943	0.905	0.982
D11S902	30.69	17,388,442	0.605	0.587	0.938	0.891	0.962
D11S4190	38.99	20,331,158	0.228	0.589	0.465	0.551	0.646
D11S915	44.32	23,502,567	0.227	0.593	0.466	0.554	0.651
D11S914	51.53	31,263,277	0.222	0.497	0.460	0.534	0.623
D11S935	58.22	35,923,164	0.060	0.019	0.064	0.047	0.082

*The Rutgers Combined Linkage-Physical Map of The Human Genome
**NCBI build 37.1 genome assembly

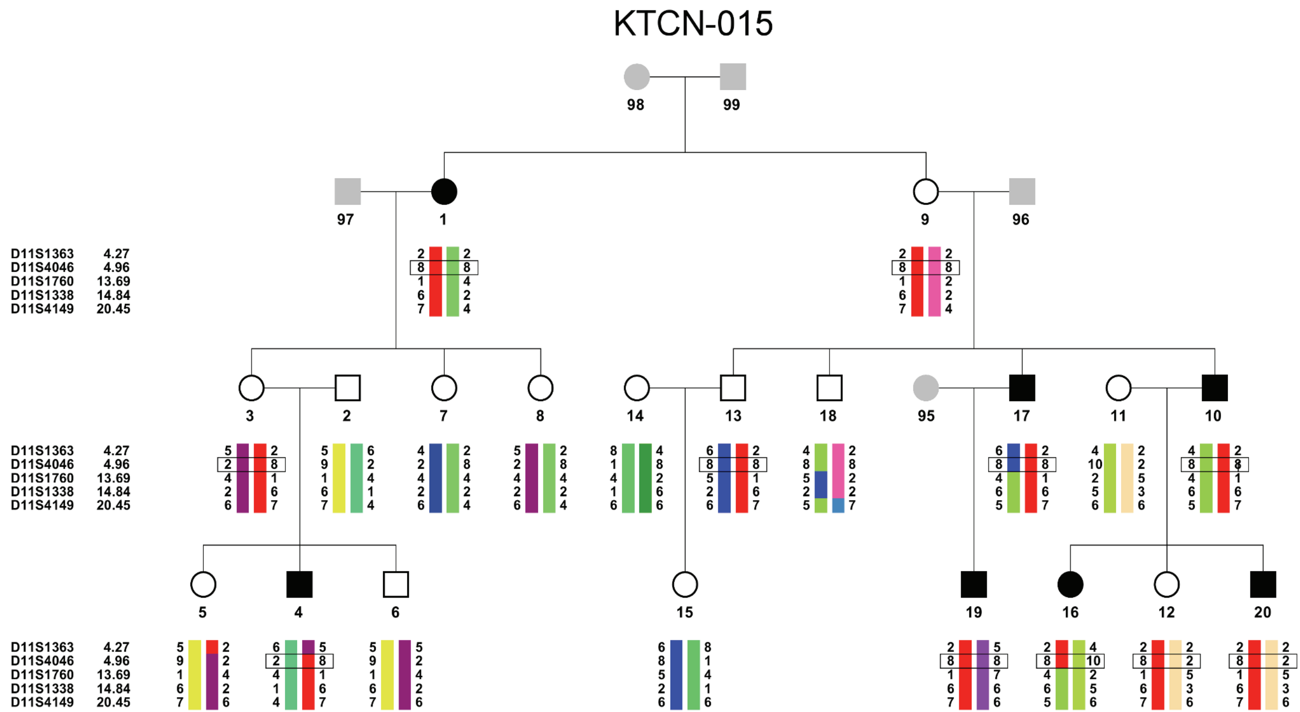

Fig. 2. Pedigree of family KTCN-019 and haplotype analysis. Black filled symbols indicate individuals with keratoconus phenotype, gray filled symbols – individuals of unknown phenotype, and the open symbols unaffected family members. Haplotype regions in different colors indicate the pattern of inheritance

D11S1760 (Fig. 2). However, that haplotype inherited by 7 affected individuals was also observed in 4 unaffected family members.

DISSCUSION

During that study a whole-genome, non-parametric linkage analysis of a multigenerational Ecuadorian family (KTCN-015) was performed. Based on previous studies, keratoconus is considered a multifactorial disease. Therefore, we decided to use non-parametric linkage analysis in our study. Additionally, we used program SimWalk2, which allows to analyze a very large pedigree, without a need to remove any family members or split the pedigree into smaller nuclear families. For the KTCN-015 family, the highest NPL scores, indicating suggestive linkage, were obtained in locus 11p15.5-p15.4. However, haplotype reconstruction at that locus did not reveal a specific chromosomal region completely segregating with the keratoconus phenotype. Incomplete penetrance in this family could cause that result. On the other hand, such low linkage signal could also be a false positive result.

The monogenic diseases characterized by Mendelian inheritance are a group of the diseases in which the identification of genetic causes of disorder development is easiest. To determine the locus related to the presence of monogenic disorder, the right tool is a parametric linkage analysis. However, numerous loci identified for keratoconus suggest that it is a polygenic disorder. What is more, many studies indicate the participation of environmental factors in keratoconus causation. It is difficult to obtain the full picture of a multifactorial disease, which includes both genetics and environmental factors. The genetics of diseases caused by many different genes are more problematic to elucidate. Each gene may have a different type of inheritance (autosomal dominant, autosomal recessive, sex-linked). In addition, parameters necessary for parametric analysis are unknown for most polygenic diseases. While the wrong choice of model can weaken the power of analysis, non-parametric methods or association analysis are more suitable. Still, even the analysis based on non-parametric methods can be unsuccessful.

In summary, our study did not allow us to confirm the connection between keratoconus and locus 11p15.5-p15.4. Despite the fact that many linkage studies of keratoconus

have been performed, there is not much knowledge about the genetic factors of this disorder. This may suggest the need to find another approach to further investigate the genetics of keratoconus.

Acknowledgements

Computational calculations were performed at Poznan Supercomputing and Networking Center. This work was supported by the Polish Ministry of Science and Higher Education, Grant N N402 097 837.

References

[1] N.E. Morton, *Sequential tests for the detection of linkage*. Am. J. Hum. Genet. 7, 277-318 (1955).

[2] E. Lander, L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results*. Nat. Genet. 11, 241-247 (1995).

[3] L. Kruglyak, M.J. Daly, M.P. Reeve-Daly, E.S. Lander, *Parametric and nonparametric linkage analysis: a unified multipoint approach*. Am. J. Hum. Genet. 58, 1347-1363 (1996).

[4] A. Kong, N.J. Cox, *Allele-sharing models: LOD scores and accurate linkage tests*. Am. J. Hum. Genet. 61, 1179-1188 (1997).

[5] D.F. Gudbjartsson, T. Thorvaldsson, A. Kong, G. Gunnarsson, A. Ingolfsdottir, *Allegro version 2*. Nat. Genet. 37, 1015-1016 (2005).

[6] Y.S. Rabinowitz, *Keratoconus*. Surv Ophthalmol 42, 297-319 (1998).

[7] D.M. Nowak, M. Gajecka, *The genetics of keratoconus*. Middle East Afr J Ophthalmol 18, 2-6 (2011).

[8] P.L. Jacq, Y. Sale, B. Cochener, P. Lozach, J. Colin, *Keratoconus, changes in corneal topography and allergy. Study of 3 groups of patients*. J Fr Ophtalmol 20, 97-102 (1997).

[9] A.R. Gasset, W.L. Houde, M. Garcia-Bengochea, *Hard contact lens wear as an environmental risk in keratoconus*. Am. J. Ophthalmol. 85, 339-341 (1978).

[10] R.J. Harrison, P.T. Klouda, D.L. Easty, M. Manku, J. Charles, C.M. Stewart, *Association between keratoconus and atopy*. Br J Ophthalmol 73, 816-822 (1989).

[11] A. Rahi, P. Davies, M. Ruben, D. Lobascher, J. Menon, *Keratoconus and coexisting atopic disease*. Br J Ophthalmol 61, 761-764 (1977).

[12] L. Wachtmeister, S.O. Ingemansson, E. Möller, *Atopy and HLA antigens in patients with keratoconus*. Acta Ophthalmol (Copenh) 60, 113-122 (1982).

[13] K.P. Burdon, D.J. Coster, J.C. Charlesworth, R.A. Mills, K.J. Laurie, C. Giunta, A.W. Hewitt, P. Latimer, J.E. Craig, *Apparent autosomal dominant keratoconus in a large Australian pedigree accounted for by digenic inheritance of two novel loci*. Hum. Genet. 124, 379-386 (2008).

[14] H. Hutchings, H. Ginisty, M. Le Gallo, D. Levy, F. Stoësser, J.F. Rouland, J.L. Arné, M.H. Lalaux, P. Calvas, M.P. Roth, A. Hovnanian, F. Malecaze, *Identification of a new locus for isolated familial keratoconus at 2p24*. J. Med. Genet. 42, 88-94 (2005).

[15] S. Kim, J. Mok, H. Kim, C.K. Joo, *Association of -31T>C and -511 C>T polymorphisms in the interleukin 1 beta (IL1B) promoter in Korean keratoconus patients*. Mol. Vis. 14, 2109-2116 (2008).

[16] F. Brancati, E.M. Valente, A. Sarkozy, J. Fehèr, M. Castori, P. Del Duca, R. Mingarelli, A. Pizzuti, B. Dallapiccola, *A locus for autosomal dominant keratoconus maps to human chromosome 3p14-q13*. J. Med. Genet. 41, 188-192 (2004).

[17] Y.G. Tang, Y.S. Rabinowitz, K.D. Taylor, X. Li, M. Hu, Y. Picornell, H. Yang, *Genomewide linkage scan in a multigeneration Caucasian pedigree identifies a novel locus for keratoconus on chromosome 5q14.3-q21.1*. Genet. Med. 7, 397-405 (2005).

[18] L. Bisceglia, P. De Bonis, C. Pizzicoli, L. Fischetti, A. Laborante, M. Di Perna, F. Giuliani, N. Delle Noci, L. Buzzonetti, L. Zelante, *Linkage analysis in keratoconus: replication of locus 5q21.2 and identification of other suggestive Loci*. Invest. Ophthalmol. Vis. Sci. 50, 1081-1086 (2009).

[19] X. Li, Y.S. Rabinowitz, Y.G. Tang, Y. Picornell, K.D. Taylor, M. Hu, H. Yang, *Two-stage genome-wide linkage scan in keratoconus sib pair families*. Invest. Ophthalmol. Vis. Sci. 47, 3791-3795 (2006).

[20] M. Gajecka, U. Radhakrishna, D. Winters, S.K. Nath, M. Rydzanicz, U. Ratnamala, K. Ewing, A. Molinari, J.A. Pitarque, K. Lee, S.M. Leal, B.A. Bejjani, *Localization of a gene for keratoconus to a 5.6-Mb interval on 13q32*, Invest. Ophthalmol. Vis. Sci. 50, 1531-1539 (2009).

[21] P. Liskova, P.G. Hysi, N. Waseem, N.D. Ebenezer, S.S. Bhattacharya, S.J. Tuft, *Evidence for keratoconus susceptibility locus on chromosome 14: a genome-wide linkage screen using single-nucleotide polymorphism markers*. Arch. Ophthalmol. 128 1191-1195 (2010).

[22] A.E. Hughes, D.P. Dash, A.J. Jackson, D.G. Frazer, G. Silvestri, *Familial keratoconus with cataract: linkage to the long arm of chromosome 15 and exclusion of candidate genes*. Invest. Ophthalmol. Vis. Sci. 44, 5063-5066 (2003).

[23] H. Tyynismaa, P. Sistonen, S. Tuupanen, T. Tervo, A. Dammert, T. Latvala, T. Alitalo, *A locus for autosomal dominant keratoconus: linkage to 16q22.3-q23.1 in Finnish families*. Invest. Ophthalmol. Vis. Sci. 43, 3160-3164 (2002).

[24] A. Hameed, S. Khaliq, M. Ismail, K. Anwar, N.D. Ebenezer, T. Jordan, S.Q. Mehdi, A.M. Payne, S.S. Bhattacharya, *A novel locus for Leber congenital amaurosis (LCA4) with anterior keratoconus mapping to chromosome 17p13*. Invest. Ophthalmol. Vis. Sci. 41, 629-633 (2000).

[25] J. Fullerton, P. Paprocki, S. Foote, D.A. Mackey, R. Williamson, S. Forrest, *Identity-by-descent approach to gene localisation in eight individuals affected by keratoconus from north-west Tasmania, Australia*. Hum. Genet. 110, 462-470 (2002).

[26] J.E. Wigginton, G.R. Abecasis, *PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data*. Bioinformatics 21, 3445-3447 (2005).

[27] E. Sobel, K. Lange, *Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics*. Am. J. Hum. Genet. 58, 1323-1337 (1996).

[28] E.M. Lange, K. Lange, *Powerful allele sharing statistics for nonparametric linkage analysis*. Hum. Hered. 57 49-58 (2004).

[29] N. Mukhopadhyay, L. Almasy, M. Schroeder, W.P. Mulvihill, D.E. Weeks, *Mega2: data-handling for facilitating genetic linkage and association analyses*. Bioinormatics 21, 2556-2557 (2005).

[30] T.C. Matise, F. Chen, W. Chen, F.M. De La Vega, M. Hansen, C. He, F.C.L. Hyland, G.C. Kennedy, X. Kong, S.S. Murray, J.S. Ziegle, W.C.L. Stewart, S. Buyske, *A second-generation combined linkage physical map of the human genome*. Genome Res. 17, 1783-1786 (2007).

[31] D.E. Weeks, E. Sobel, J.R. O'Connell, K. Lange, *Computer programs for multilocus haplotyping of general pedigrees*. Am. J. Hum. Genet. 56, 1506-1507 (1995).

[32] H. Thiele, P. Nürnberg, *HaploPainter: a tool for drawing pedigrees with complex haplotypes*. Bioinformatics 21, 1730-1732 (2005).

DOROTA M. NOWAK studied bioinformatics at the Adam Mickiewicz University in Poznań, where received her M.Sc. in Biology. Since 2008 she has been working in the Institute of Human Genetics PAS in Poznań. She is a co-author of several scientific publications in Polish and international journals. Her main scientific interests are linkage analysis and genetics of ocular diseases.