Gene finding in genetically isolated populations

Peter Heutink* **and Ben A. Oostra**

Institute of Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands

Received July 3, 2002; Accepted July 17, 2002

The struggle to identify susceptibility genes for complex disorders has stimulated geneticists to develop new approaches. One approach that has gained considerable interest is to focus on genetically isolated populations rather than on the general population. There remains much controversy and theoretical debate over the feasibility and advantages of such populations, but recent results speak in favor of the feasibility of this approach, and will be reviewed here.

In July 2003, the complete sequence of the human genome is expected to be finished, but already over the past few years the progress of the Human Genome Project has dramatically changed the daily live of human geneticists. Gene finding for disorders with Mendelian inheritance is rapidly shifting from cumbersome positional cloning projects requiring many years to complete (e.g. with Huntington's disease) (1) to positional candidate gene approaches using *in silico* candidate gene selection (e.g. hemochromatosis) (2). Although for the large majority of our estimated 32 000 genes the biological function is still largely unknown, we have now at least acquired knowledge of their existence and localization within our genome.

The expectations on the tradeoffs of the Human Genome Project have also resulted in a shift of attention of scientists towards gene finding for more complex disorders. This is often accompanied by statements that all interesting Mendelian disorders have been, or soon will be, mapped. This statement neglects the current status of human disease mapping. In the Online Mendelian Inheritance of Man (OMIM) (http://www3.ncbi.nlm.nih.gov/omim/), >10 000 disease entries are listed, but for only ~10% have the responsible genes been identified (http://www.ncbi.nlm.nih.gov/locuslink/).

The importance of the continuation of mapping Mendelian disorders is demonstrated by our increased understanding of many biological processes through studying the pathogenesis of rare disorders such as mental retardation (3). Especially instrumental have been those diseases in which genes responsible for rare familial forms have been identified and have helped our understanding of the pathways involved in the more common and complex forms of diseases. An example is neurodegenerative disorders, where common pathways are being unraveled using Mendelian forms of the disease (4–6).

COMPLEX DISEASE

In Mendelian disorders, mutations are in general directly causal, and segregation of phenotypes can be followed in families, allowing family-based linkage approaches. For complex disorders, genetic influences are less clear cut, and so far the progress in gene identification for these disorders has been disappointing. Very few genetic risk factors with convincing evidence have been identified. The most widely used examples for successful identification of risk factors for complex disorders are apolipoprotein E for Alzheimer's disease (7,8) and factor V Leiden clotting factor for thrombophilia (9,10).

The main reason for this very limited success is that it is simply not known how the genetics of complex disorders should be described or modeled (11). Approaches that have been so successfully used for Mendelian disorders were originally designed for use with well-defined genetic parameters. However, complex disorders do not clearly follow Mendelian laws of inheritance because of complicating factors such as phenocopies, genetic heterogeneity, variable clinical expression, age at onset, new mutations, incomplete penetrance, polygenic inheritance and environmental risks. Only a combination of these genetic and (often) environmental risk factors leads to the presentation of the phenotype; for genetic risk factors with small effects, classical linkage analysis using independent pedigrees therefore has very limited power to detect a locus.

The extent to which genetic risk factors contribute to the onset of common diseases is largely unknown. Widely used measures for heritability and relative attributable risk not only measure the effect of all genetic risk factors combined, but also include the effects of shared environmental risk factors and the interactions between these risk factors. Since these measures pool together the effects of all genetic risk factors, it is not clear how many and how strong the individual risk factors are (12).

There are two main hypotheses for explaining the genetics of complex disorders. Firstly, in the multi-equivalent risk model, many rare variants exist in the population with varying effect on risk (each variant being a strong risk factor). Secondly, in the restricted polymorphism model, a relatively small pool of common disease alleles exists (each factor being a weak risk factor).

*To whom correspondence should be addressed at: Department of Clinical Genetics, Erasmus MC Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands. Tel: +31 104088136; Fax: +31 104089384; Email: heutink@kgen.fgg.eur.nl

These risk factors can be both genetic and environmental, and can increase or decrease the risk for disease. It is likely that both these and intermediate (oligogenic) situations exist—as for diabetes, where interactions between environmental and genetic risk factors are well documented. For example, the effect of changes towards a more Westernized life style by Pima Indians and Japanese is associated with genetic susceptibility to the disease (13–15). And susceptibility in the NOD mouse model can be explained by several factors, with only a few having a strong effect (16). Other disorders with oligogenic inheritance are Hirschprung's disease (17,18) and late-onset Alzheimer's disease (19–21).

ALTERNATIVE METHODS FOR GENE FINDING

The use of alternative linkage approaches such as affected sib-pair analysis circumvents some of the problems associated with linkage analysis on larger families by avoiding extensive modeling of the genetic parameters. Other researchers have looked for more direct approaches to circumvent genetic modeling by directly testing the effect of sequence variations in candidate genes using cases versus controls in association studies on genes that are predicted to harbor risk factors because of their known biological function. In its basic design, this approach is very appealing if the function of the candidate gene is fully known. Because of the progress in our understanding of the function of many genes, this approach is becoming more and more feasible. However, at this point for most genes we still have very limited knowledge, and therefore candidate genes are often put forward with very few arguments supporting their biological role in a disease process. This is reflected in the very small number of genetic association studies that have actually been replicated and accepted in the scientific community. The association of the E4 allele of apolipoprotein E is often used as an example of the successful identification of a genetic risk factor for Alzheimer's disease by association studies, but this risk factor was originally localized by linkage analysis (7,8). Another approach that holds great promise is to look at quantitative traits associated with disease that might segregate as more Mendelian traits—thus to subdivide patient populations or to use congenic animal models (22,23).

LINKAGE VERSUS ASSOCIATION

A comparison of the statistical power of linkage-based approaches versus association studies strongly favored the latter over linkage-based approaches such as sib-pair studies, especially for mild risk factors with a genotype risk ratio $\gamma < 4$ (24,25). For risk factors with $\gamma \leq 2$, unrealistic sample sizes for linkage studies (~2500 families) would be required. It should be realized that the measure for describing a risk factor γ is different from the more commonly used sibling recurrence risk ratio λ (26); $\gamma = 2$ correspondents to $\lambda = 1.3$. Linkage studies for risk factors of $\lambda \geq 2$ are certainly feasible, although we should be aware that we consider the λ for the specific risk factor under investigation and not for the disease as a whole (27). Another important observation is that it was assumed in the calculations that the tested markers were in complete linkage disequilibrium with the biologically relevant risk factor within the whole population (28). Linkage disequilibrium (LD) refers to the non-random association of marker alleles within a genomic region in which the presence of an allele of one marker predicts the presence of an allele of a nearby marker. LD is eroded by recombination and gene conversion (Fig. 1). Thus the predictions above are only true if there has been a single mutation event in the gene under study. Otherwise, no LD will be detected. Given the experience with allelic heterogeneity in many Mendelian disorders, this is an unlikely situation. In contrast, linkage can be detected even if multiple mutation events have occurred, and therefore genome-wide LD mapping in population samples potentially has high power to detect such genes, but it is restricted by the amount of allelic heterogeneity.

Risch and Merikangas (24) recognized that as part of the Human Genome Project large numbers of single-nucleotide polymorphisms (SNPs) would become available and that these SNPs could serve as an excellent tool for genome-wide association studies. This recognition caused considerable debate on how many SNPs would then be needed for genome-wide association studies and how genome-wide significance could be achieved if hundreds of thousands of SNPs had to be tested.

SIGNAL-TO-NOISE RATIO

In general, the statistical power to detect a real association or linkage is limited by the background noise in the population under study. This noise consists of all possible combinations of environmental and genetic factors present in the population. Therefore, in heterogeneous populations, large sample sizes would be needed to obtain sufficient statistical power to detect genetic risk factors. More homogeneous populations such as genetically isolated populations have been proposed as a possible alternative for these large sample sizes, because environmental variation might be lower and the genetic make-up of these populations is expected to be less complex owing to founder effects, thus improving the signal-to-noise ratio. The use of genetically isolated populations is not new; for example, in Finland, there are numerous examples of Mendelian disorders with increased prevalence (29). This has been especially valuable for mapping rare recessive disorders, but many researchers believe this could be a solution for more complex disorders as well because of the relatively uniform genetic background of the population. Some culturally and genetically isolated populations have a more similar way of living, eating habits and natural environment that reduces environmental variation. Often these populations have been founded by a small number of individuals, followed by a period of genetic isolation, during which genetic drift might have been seen and population expansion mainly occurred by population growth and not by immigration. In addition, if genealogical records are available, the kinship coefficient of patients can be determined, which is often higher than in heterogeneous populations. In countries from Scandinavia, for example, state healthcare registries have been maintained over centuries (30). Increased kinship should also be suitable for detecting recessive effects of common risk factors.

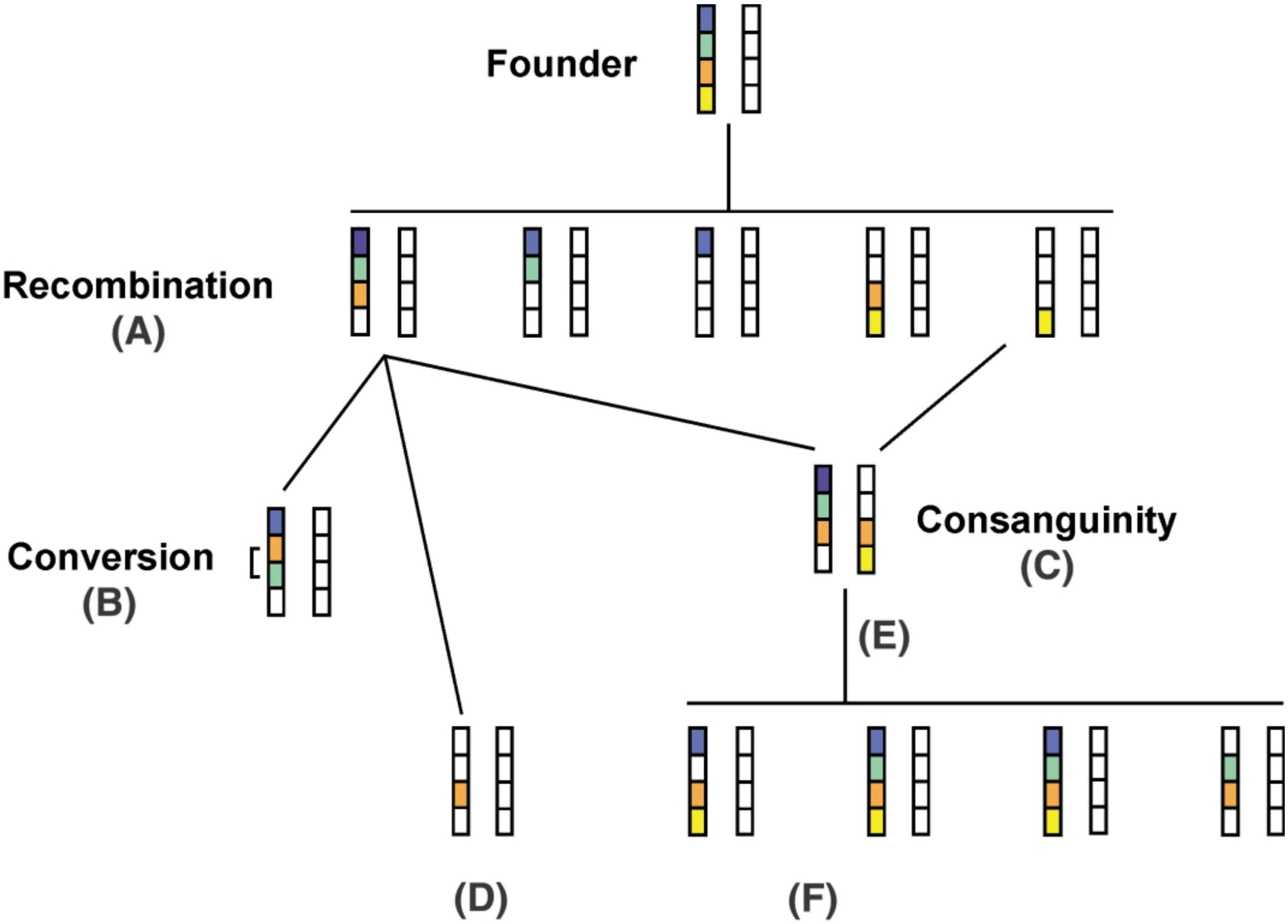

Figure 1. History of present-day linkage disequilibrium is shaped by recombination, gene conversion and consanguinity. If a mutation is present in the orange segment, recombination events (**A** and **D**) degrade the region of LD around the mutation. Gene conversion degrades short-range LD (**B**). Increased consanguinity (**C**) can result in enrichment of disease-associated alleles (**E**) but patterns of LD (**F**) can be complex as a result of converging haplotypes.

Each genetically isolated population has its own demographic history, and each might have its own advantages and disadvantages. Very old isolates (>100 generations) with limited population growth such as the Saami in northern Scandinavia are carrying very old (general) mutations or new (population-specific) mutations. Because of the prolonged period of isolation, considerable genetic drift has occurred. For mapping disease genes in these populations, in general, short regions of LD are expected because many meiotic events have occurred in founder chromosomes. However, if considerable consanguinity in such populations exists and only a limited number of founder chromosomes were present, it might be difficult to distinguish disease-associated haplotypes from the background haplotypes.

In younger isolates (≤100 generations) such as exist in Finland (early and late settlement), Iceland, Sardinia and Japan with low immigration and expansion, allelic heterogeneity might still be reduced by genetic drift before a more recent expansion started, but the reduction is less strong than in the older isolates. Furthermore, LD will exist over slightly longer genomic regions. These populations have been useful for rare Mendelian disease (29), but for common disorders the number of founders might have been relatively high for some of these populations, resulting in many different alleles present in the population. This problem can be circumvented by using extensive genealogical studies so that relatively shallow pedigrees can be identified (31). This is especially feasible in the population of Iceland, where the genealogy since the original settlement is available. In this way, disease loci for tremor, stroke and arteriosclerosis have been identified (32–35).

On the other hand, very young isolates (<20 generations), such as the population of the Central Valley of Costa Rica (CVCR) (36), Tristan da Cunha (37), Hutterites (12,38), small population sub-isolates in the Netherlands (39) and French Canada, have been identified with exponential population growth in the last generations. In these populations, a very tight founding bottleneck probably limits the number of mutations in the current population. Long regions of LD have been detected in these populations for bipolar illness (36,40) and type 1 diabetes (39). Although disease gene localization can benefit from the use of isolated populations, there might be disadvantages as well. Findings in isolated populations might not be valid in the general population, especially if in very old isolated populations new mutations arise or if old mutations remain while they became extinct in the general population. For example, linkage findings for multiple sclerosis in Finland (41) have not been replicated in other populations, but this is not generally true for other diseases or in younger isolates, which are still in many ways very similar to the general population because of their recent separation. In addition, these findings are very important for our understanding of the biological process leading to disease.

THE DISTRIBUTION OF LD IN THE HUMAN GENOME

In the recent literature, two central questions have been discussed; how many SNPs would be needed for genome-wide association studies and is there really a benefit using genetically isolated populations? Central to this debate is the question of how much linkage disequilibrium exists in different populations. LD is eroded by recombination and gene conversion (Fig. 1) and influenced by the age and history of the mutation and population size and structure. A number of groups have studied the distribution of LD on the human genome. Kruglyak (42) performed extensive computer simulation studies, which provided a theoretical basis for our understanding of the distribution of LD. Assuming an out-of-Africa human founding population of ~100 000 years ago, simulations were performed taking genetic drift and different models of population history into account, using bi-allelic markers equivalent to SNPs. Under this scenario, a useful level of LD is not likely to extend >3 kb in the general population. In addition, LD in isolated populations would not be expected to differ much, except if the founding bottleneck were very narrow (10–100 unrelated individuals) or if the frequency of the variant were low (<5%). In this scenario, ~500 000 SNPs would be needed for genome-wide association studies.

However, a series of 'wet-lab' genotyping studies did not fully support these simulation studies (43–52). Although a large variation in the extent of LD has been described, depending on the chromosomal region studied, in general LD can be detected over much larger genomic distances than expected in populations of northern European descent. There are large differences in LD for different genomic regions, and the distribution of LD can be described by blocks of extensive LD that are interspersed by blocks showing little LD (49) (Fig. 2). The borders of these blocks appear to be quite sharp and related to recombination frequencies. There are some indications that specific sequence content or motifs are important, but it is not known whether this is a general mechanism (52–55). Thus, in the general population, LD is more extensive than originally predicted, but the choice of the density of a SNP map to catch a risk factor would have to take the regional variation in LD into account. In order to be able to make this choice, a genome-wide LD map should be constructed, and the first attempts are becoming available (56–58).

Preliminary comparisons of LD between populations with different demographic histories have been made, and, indeed (as could be expected) regions of LD in older populations tend to be shorter in length than in younger populations. Most regions of LD are shared between populations such as between European and African populations, but also population-specific LD exists for some regions, which has probably arisen after the separation of both populations in history (59). In order to obtain a useful LD map, LD should be measured across a dense set of loci covering the entire genome in a range of populations.

LD IN ISOLATED POPULATIONS

In several of the studies mentioned above, samples from populations considered as genetic isolates were included, and not much difference in the length of LD with admixed populations could be detected. This has been interpreted by some that genetic isolates might not be more useful for gene mapping than the general population. However, much of this debate is clouded by the fact that some of these statements are imprecise. For example, in several studies on LD, samples from the population of Sardinia have been used. Sardinia is regarded as a genetically isolated population by several authors, but this is not generally true for the whole population of Sardinia. On many coastal areas, there have been influences from outside populations (60,61). Unless authors clearly give evidence that the samples are from real genetically isolated populations, these data cannot be properly interpreted. In line with this, one should realize that Iceland is not a simple homogeneous genetic isolate, but instead has been founded by a mix of populations (62–64). The Norse founders were mostly male and the founding females were mostly of Gaelic origin. In addition, in Finland, two founding populations exist, one of which is much older (~2000 years) than the other (~500 years) (29).

In contrast to the findings above, several studies on very young isolates (<20 generations) show very promising results by detecting LD over long distances (>1 cM), such as in the populations of the Central Valley of Costa Rica and Palau (56,57) and Dutch sub-isolates (manuscript in preparation). It should further be noted that the extent of background LD in the population under study is not the most important measure to consider, but the difference of LD around risk factors in patients compared to this background LD in these populations (65).

FROM GENE MAPPING TO GENE FINDING

Developing methodology to successfully map genetic risk factors for complex disorders is only a first step towards the identification of these risk factors. The use of very young isolates, with extended LD, should make it relatively easy to find a locus with a limited number of markers. The ease of finding LD, however, has a downside as well, because in order to reduce the region of interest large population samples are needed. The second step towards the identification of the biologically relevant mutation will therefore be more difficult in younger isolates compared with older isolates, in which on average shorter LD exists. In older isolates, more markers would be needed for the initial screen, but the identified region can be expected to be shorter than for younger isolates. An attractive strategy would be to carry out the initial screen in a young isolate followed by fine mapping in an older isolate or even in the general population. Therefore it would be useful to

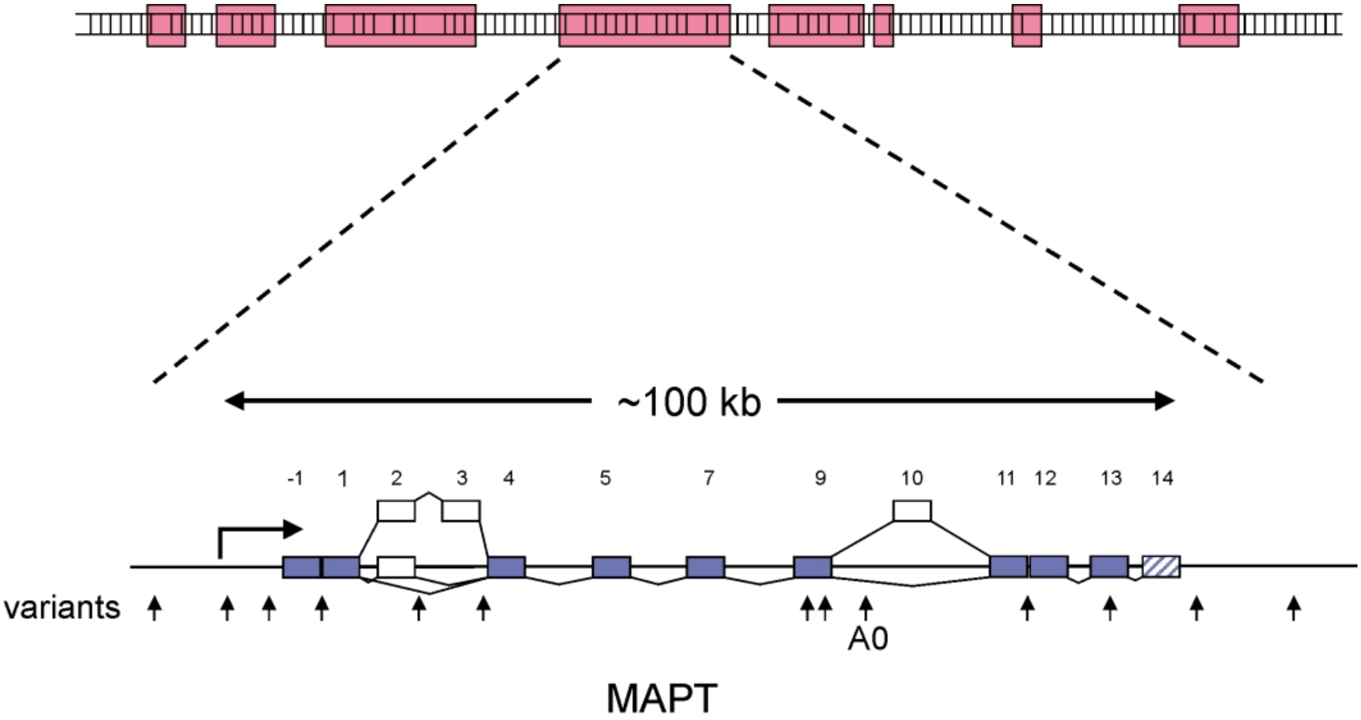

Figure 2. Top: Linkage disequilibrium (LD) in the human genome (double line) is organized in a block-like structure. SNPs are indicated by vertical lines. Regions of high LD (boxes) are interspersed with regions of low LD (49). The strength of long-range LD is inversely correlated with recombination frequencies (82). Short-range LD is mainly degraded by gene conversion (83). Bottom: Genomic organization of the microtubule-associated protein *tau* (*MAPT*) gene. Exons are indicated by colored boxes. A large number of SNP variants (indicated by arrows) show strong LD over >100 kb. Early association studies demonstrated a strong association of a polymorphic short tandem repeat allele within intron 9 (A0) and several neurodegenerative disorders. This association now extends over the whole gene, including the promoter region (5).

collect additional samples from other (isolated) populations or to organize repositories of population samples.

Even though, in this way, the number of candidate genes can be reduced considerably, the problem of identifying the biologically relevant mutation remains. There is a need to screen for variants in the whole region that is in LD, since not all SNPs are known. It is often estimated that an SNP exists in 1 in 1000 bp, but some studies reached a much higher number of variants (1 in 200 bp) by re-sequencing larger number of individuals, which means that many SNPs remain to be detected (66). An alternative is to only test SNPs within functional candidate genes. However, the function of most genes is not known, and their selection is therefore not (yet) very specific. Once all SNPs have been identified and tested, several will demonstrate evidence for association. The problem will be how to separate the biological relevant mutations from harmless SNPs that are in complete LD. Testing the biological effect of each of these SNPs is a possible solution. However, this approach is not feasible if extended regions of LD exist: an example is the region around the microtubule-associated protein *tau* (*MAPT*) gene on chromosome 17q21 (Fig. 2). Mutations in *MAPT* are associated with frontotemporal dementia with parkinsonism linked to chromosome 17 (FTDP-17) (67–69). In addition there is strong evidence that *MAPT* harbors a risk factor for other neurodegenerative disorders such as progressive supranuclear palsy, corticobasal degeneration and even late-onset Parkinson's disease (5). However, the detection of the biologically relevant mutation has so far puzzled scientists, because the region shows extensive LD for a number of SNPs spanning the whole *MAPT* gene (>100 kb). In populations of European descent, only two major haplotypes exist for the gene. A possible solution might come from testing LD in additional populations with a different ethnicity. It is currently not known how much difference in LD exists between populations, but there are indications that population-specific LD exists (59), again demonstrating the importance of population-specific LD maps.

MUTATIONS IN DISTANT REGULATORY ELEMENTS

Since mutations involved in complex disorders are often not directly causal, they can be expected to have mild biochemical effects. Mutations can be located within coding regions of genes and affect the function of the protein, or can have an effect on the function of the protein by changing alternative splicing or secondary structure, stability, translation or localization of the mRNA. To test such effects, functional assays can be designed. Although this is already a major task, in addition mutations can be located outside the coding region of the gene in elements that control timing, location and/or level of gene expression. Several studies have demonstrated that by using comparative genomics and expression data, conserved sequences can be identified in the genome that might act as regulatory elements (70–72). However, we do not know

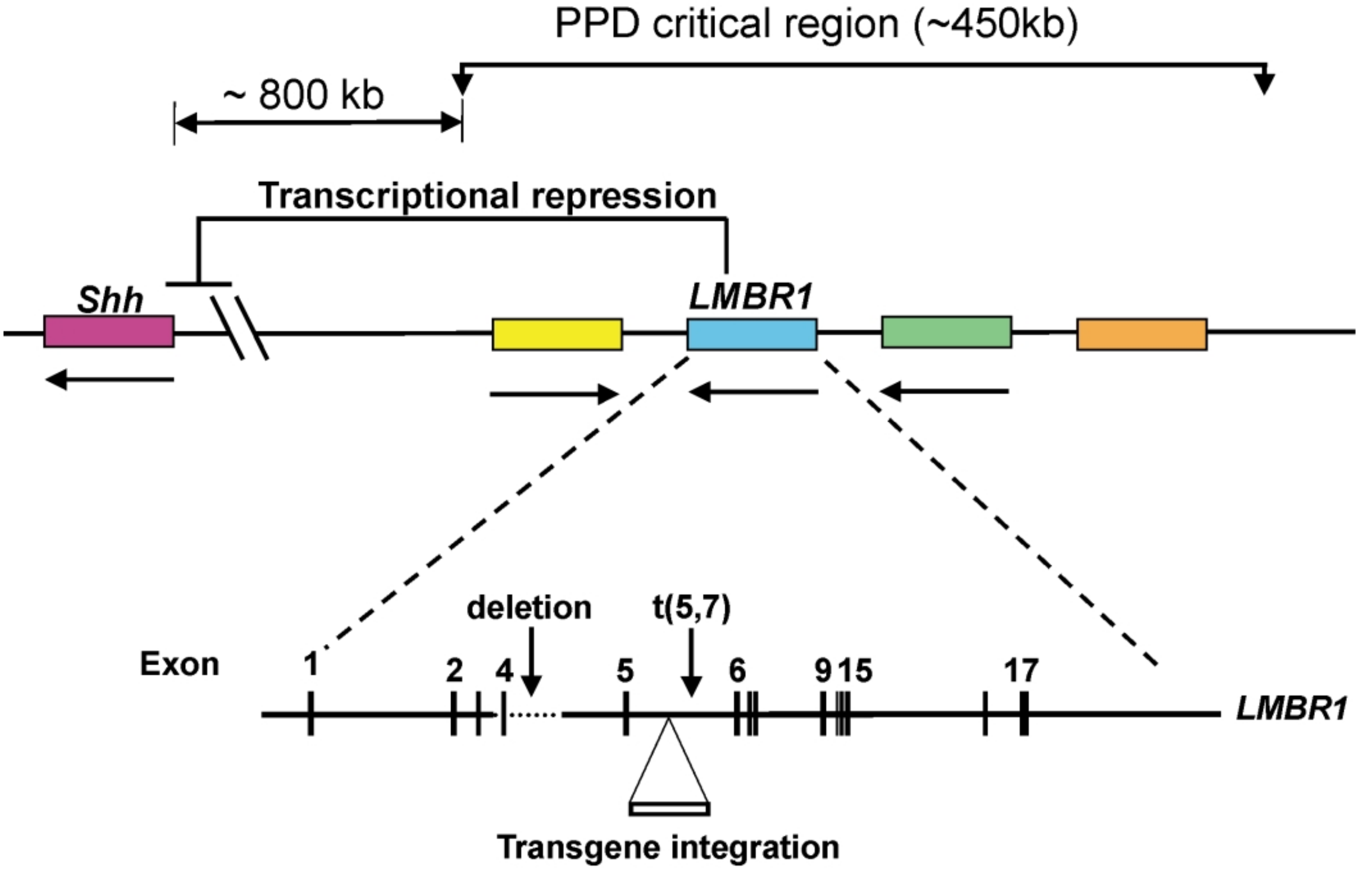

Figure 3. Top: Genomic organization around the pre-axial polydactyly (PPD) critical region (79). Colored boxes indicate transcripts. Arrows indicate direction of transcription. *LMBR1* contains a *cis*-acting limb-specific transcriptional repressor for the *Sonic hedgehog* (*Shh*) gene (75). Bottom: Intron/exon structure of LMBR1. The deletion reported for acheiropodia (84), the translocation t(5,7)(q11,q36) and the Sasquatsh transgene integration leading to abolishment of limb-specific *Shh* repression are indicated (75).

how far such regulatory sequences extend. Most studies investigate a limited amount of sequence around a gene, but is this really sufficient? For most cases, it probably will be, but there is increasing evidence for the existence of long-ranging *cis*-acting elements in the genome. One example is adult-type hypolactasia, or lactose intolerance, an autosomal recessive disease with a physiological decline in the activity of the lactase–phlorizin hydrolase (LPH) (73). Families linked to the gene encoding LPH (the *LCT* gene) showed no mutations within the gene, but the associated haplotype spanning the *LCT* gene showed differences in transcript levels of non-persistence, suggesting a *cis*-acting allele. Subsequent LD and haplotype analysis in Finnish families demonstrated that the mutation had to be located within a 47 kb interval around the gene. Two point mutations, a C/T variant 14 kb and a G/A change 22 kb upstream of the *LCT* gene, were identified. These variants were also associated in distantly related populations, indicating that they are very old. This example is certainly not unique or a worst-case scenario. Long-range *cis*-acting regulatory elements have been described for the *Sonic hedgehog* (*Shh*) gene, which is associated with polydactyly and holoprosencephaly. Chromosomal rearrangements >250 kb upstream of the *Shh* gene have been identified in patients with holoprosencephaly (74). In addition, in families with polydactyly, *cis*-acting limb-specific regulatory sequences for the *Shh* gene are located at least 800 kb upstream of the gene and embedded in an intron of another gene (75) (Fig. 3). Additional examples such as SOX9 (76), PAX6 (77) and β-globin (78) demonstrate that this is not an exceptional phenomenon. The examples above have only been found because there was a strong functional candidate gene and because a relatively small region of LD could be determined owing to the genetic isolation of the population under study and the availability of genealogical data. It might be more difficult for less obvious factors. In the case of polydactyly, recombination events excluded the responsible gene by 800 kb (79). This demonstrates that the selection of candidate genes should not only be based on function and location within the borders of the critical region, but also include extensive *in silico* genome analysis of a more broadly defined region.

A PLEA TO ASCERTAIN GENETICALLY ISOLATED POPULATIONS

Some of the promises of the use of isolated populations are beginning to be apparent, but many questions remain. Since each disease has its own characteristics, it has been suggested to improve study design by adapting the study population choice to the specific research question under study (11,80). This would imply that there is a need to investigate a large number of different genetically isolated populations, each with its own demographic history and health characteristics. Where are these populations to be found? This seems not to be an easy task, but it should be realized that even within very densely populated countries such as the Netherlands, genetically isolated subpopulations have been identified based on religious, behavioral or socio-economic isolation. If isolated populations hold the key to finding genes for complex disorders that affect maybe 60% of our population (81), the timely collection of samples from genetically isolated populations is crucial, because the increasing mobility of people from all over the world will lead to the disappearance of many of these populations in the nearby future.

REFERENCES

1. The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.
2. Njajou, O.T., Vaessen, N., Joosse, M., Berghuis, B., van Dongen, J.W., Breuning, M.H., Snijders, P.J., Rutten, W.P., Sandkuijl, L.A., Oostra, B.A. *et al.* (2001) A mutation in SLC11A3 is associated with autosomal dominant hemochromatosis. *Nat. Genet.*, **28**, 213–214.
3. Chiurazzi, P. and Oostra, B.A. (2000) Genetics of mental retardation. *Curr. Opin. Pediatr.*, **12**, 529–535.
4. Lovestone, S. and McLoughlin, D.M. (2002) Protein aggregates and dementia: Is there a common toxicity? *J. Neurol. Neurosurg. Psychiatry*, **72**, 152–161.
5. Heutink, P. (2000) Untangling tau-related dementia. *Hum. Mol. Genet.*, **9**, 979–986.
6. Kruger, R., Eberhardt, O., Riess, O. and Schulz, J.B. (2002) Parkinson's disease: one biochemical pathway to fit all genes? *Trends Mol. Med.*, **8**, 236–240.
7. Pericak-Vance, M.A., Bebout, J.L., Gaskell, P.C., Jr, Yamaoka, L.H., Hung, W.Y., Alberts, M.J., Walker, A.P., Bartlett, R.J., Haynes, C.A., Welsh, K.A. *et al.* (1991) Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am. J. Hum. Genet.*, **48**, 1034–1050.
8. Strittmatter, W.J., Saunders, A.M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G.S. and Roses, A.D. (1993) Apolipoprotein E: high-avidity binding to β-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl Acad. Sci. USA*, **90**, 1977–1981.
9. Bertina, R.M., Koeleman, B.P., Koster, T., Rosendaal, F.R., Dirven, R.J., de Ronde, H., van der Velden, P.A. and Reitsma, P.H. (1994) Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature*, **369**, 64–67.
10. Majerus, P.W. (1994) Human genetics. Bad blood by mutation. *Nature*, **369**, 14–15.
11. Terwilliger, J.D. and Weiss, K.M. (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.*, **9**, 578–594.
12. Ober, C., Tsalenko, A., Parry, R. and Cox, N.J. (2000) A second-generation genomewide screen for asthma-susceptibility alleles in a founder population. *Am. J. Hum. Genet.*, **67**, 1154–1162.
13. Price, R.A., Charles, M.A., Pettitt, D.J. and Knowler, W.C. (1993) Obesity in Pima Indians: large increases among post-World War II birth cohorts. *Am. J. Phys. Anthropol.*, **92**, 473–479.
14. Knowler, W.C., Saad, M.F., Pettitt, D.J., Nelson, R.G. and Bennett, P.H. (1993) Determinants of diabetes mellitus in the Pima Indians. *Diabetes Care*, **16**, 216–227.
15. Nemoto, M., Sasaki, T., Deeb, S.S., Fujimoto, W.Y. and Tajima, N. (2002) Differential effect of PPARγ2 variants in the development of type 2 diabetes between native Japanese and Japanese Americans. *Diabetes Res. Clin. Pract.*, **57**, 131–137.
16. Risch, N., Ghosh, S. and Todd, J.A. (1993) Statistical evaluation of multiple-locus linkage data in experimental species and its relevance to human studies: application to nonobese diabetic (NOD) mouse and human insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **53**, 702–714.
17. Gabriel, S.B., Salomon, R., Pelet, A., Angrist, M., Amiel, J., Fornage, M., Attie-Bitach, T., Olson, J.M., Hofstra, R., Buys, C. *et al.* (2002) Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat. Genet.*, **31**, 89–93.
18. Passarge, E. (2002) Dissecting Hirschsprung disease. *Nat. Genet.*, **31**, 11–12.
19. Olson, J.M., Goddard, K.A. and Dudek, D.M. (2002) A second locus for very-late-onset alzheimer disease: a genome scan reveals linkage to 20p and epistasis between 20p and the amyloid precursor protein region. *Am. J. Hum. Genet.*, **71**, 154–161.
20. Li, Y.J., Scott, W.K., Hedges, D.J., Zhang, F., Gaskell, P.C., Nance, M.A., Watts, R.L., Hubble, J.P., Koller, W.C., Pahwa, R. *et al.* (2002) Age at onset in two common neurodegenerative diseases is genetically controlled. *Am. J. Hum. Genet.*, **70**, 985–993.
21. Bertram, L. and Tanzi, R.E. (2001) Dancing in the dark? The status of late-onset Alzheimer's disease genetics. *J. Mol. Neurosci.*, **17**, 127–136.
22. Doerge, R.W. (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.*, **3**, 43–52.
23. Brockman, G.A. and Bevova, M.R. (2002) Using mouse models to dissect the genetics of obesity. *Trends Genet.*, **18**, 367–376.
24. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
25. Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
26. Scott, W.K., Pericak-Vance, M.A. and Haines, J.L. (1997) Genetic analysis of complex diseases. *Science*, **275**, 1327; discussion 1329–1330.
27. Risch, N. and Merikangas, K. (1997) Genetic analysis of complex disease. *Science*, **275**, 1327–1328; discussion 1329–1330.
28. Muller-Myhsok, B. and Abel, L. (1997) Genetic analysis of complex diseases. *Science*, **275**, 1328–1329; discussion 1329–1330.
29. Peltonen, L., Jalanko, A. and Varilo, T. (1999) Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.*, **8**, 1913–1923.
30. Peltonen, L., Palotie, A. and Lange, K. (2000) Use of population isolates for mapping complex traits. *Nat. Rev. Genet.*, **1**, 182–190.
31. Ombra, M.N., Forabosco, P., Casula, S., Angius, A., Maestrale, G., Petretto, E., Casu, G., Colussi, G., Usai, E., Melis and P., Pirastu, M. (2001) Identification of a new candidate locus for uric acid nephrolithiasis. *Am. J. Hum. Genet.*, **68**, 1119–1129.
32. Lindqvist, A.K., Steinsson, K., Johanneson, B., Kristjansdottir, H., Arnasson, A., Grondal, G., Jonasson, I., Magnusson, V., Sturfelt, G., Truedsson, L. *et al.* (2000) A susceptibility locus for human systemic lupus erythematosus (hSLE1) on chromosome 2q. *J Autoimmun.* **14**, 169–178.
33. Gulcher, J.R., Jonsson, P., Kong, A., Kristjansson, K., Frigge, M.L., Karason, A., Einarsdottir, I.E., Stefansson, H., Einarsdottir, A.S., Sigurthoardottir, S., Baldursson, S., Bjornsdottir, S., Hrafnkelsdottir, S.M., Jakobsson, F., Benedickz, J. and Stefansson, K. (1997) Mapping of a familial essential tremor gene, FET1, to chromosome 3q13. *Nat. Genet.*, **17**, 84–87.
34. Gretarsdottir, S., Sveinbjornsdottir, S., Jonsson, H.H., Jakobsson, F., Einarsdottir, E., Agnarsson, U., Shkolny, D., Einarsson, G., Gudjonsdottir, H.M., Valdimarsson, E.M. *et al.* (2002) Localization of a susceptibility gene for common forms of stroke to 5q12. *Am. J. Hum. Genet.*, **70**, 593–603.
35. Gudmundsson, G., Matthiasson, S.E., Arason, H., Johannsson, H., Runarsson, F., Bjarnason, H., Helgadottir, K. , Thorisdottir, S., Ingadottir, G., Lindpaintner, K. *et al.* (2002) Localization of a gene for peripheral arterial occlusive disease to chromosome 1p31. *Am. J. Hum. Genet.*, **70**, 586–592.
36. Freimer, N.B., Reus, V.I., Escamilla, M.A., McInnes, L.A., Spesny, M., Leon, P., Service, S.K., Smith, L.B., Silva, S., Rojas, E. *et al.* (1996) Genetic mapping using haplotype, association and linkage methods suggests a locus for severe bipolar disorder (BPI) at 18q22–q23. *Nat. Genet.*, **12**, 436–441.
37. Zamel, N., McClean, P.A., Sandell, P.R., Siminovitch, K.A. and Slutsky, A.S. (1996) Asthma on Tristan da Cunha: looking for the genetic link. The University of Toronto Genetics of Asthma Research Group. *Am. J. Respir. Crit. Care Med.*, **153**, 1902–1906.

38. Abney, M., McPeek, M.S. and Ober, C. (2000) Estimation of variance components of quantitative traits in inbred populations. *Am. J. Hum. Genet.*, **66**, 629–650.
39. Vaessen, N., Heutink, P., Houwing-Duistermaat, J.J., Snijders, P.J., Rademaker, T., Testers, L., Batstra, M.R., Sandkuijl, L.A., van Duijn, C.M. and Oostra, B.A. (2002) A genome-wide search for linkage-disequilibrium with type 1 diabetes in a recent genetically isolated population from The Netherlands. *Diabetes*, **51**, 856–859.
40. McInnes, L.A., Service, S.K., Reus, V.I., Barnes, G., Charlat, O., Jawahar, S., Lewitzky, S., Yang, Q., Duong, Q., Spesny, M. *et al.* (2001) Fine-scale mapping of a locus for severe bipolar mood disorder on chromosome 18p11.3 in the Costa Rican population. *Proc. Natl Acad. Sci. USA*, **98**, 11485–11490.
41. Tienari, P.J., Wikstrom, J., Sajantila, A., Palo, J. and Peltonen, L. (1992) Genetic susceptibility to multiple sclerosis linked to myelin basic protein gene. *Lancet*, **340**, 987–991.
42. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
43. Laan, M. and Paabo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat. Genet.*, **17**, 435–438.
44. Varilo, T., Laan, M., Hovatta, I., Wiebe, V., Terwilliger, J.D. and Peltonen, L. (2000) Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur. J. Hum. Genet.*, **8**, 604–612.
45. Taillon-Miller, P., Bauer-Sardina, I., Saccone, N.L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J.P. and Kwok, P.Y. (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.*, **25**, 324–328.
46. Dunning, A.M., Durocher, F., Healey, C.S., Teare, M.D., McBride, S.E., Carlomagno, F., Xu, C.F., Dawson, E., Rhodes, S., Ueda, S. *et al.* (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.*, **67**, 1544–1554.
47. Eaves, I.A., Merriman, T.R., Barber, R.A., Nutland, S., Tuomilehto-Wolf, E., Tuomilehto, J., Cucca, F. and Todd, J.A. (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **25**, 320–323.
48. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
49. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
50. Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, **68**, 191–197.
51. Mohlke, K.L., Lange, E.M., Valle, T.T., Ghosh, S., Magnuson, V.L., Silander, K., Watanabe, R.M., Chines, P.S., Bergman, R.N., Tuomilehto, J. *et al.* (2001) Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. *Genome Res.*, **11**, 1221–1226.
52. Schulze, T.G., Chen, Y.S., Akula, N., Hennessy, K., Badner, J.A., McInnis, M.G., DePaulo, J.R., Schumacher, J., Cichon, S., Propping, P. *et al.* (2002) Can long-range microsatellite data be used to predict short-range linkage disequilibrium? *Hum. Mol. Genet.*, **11**, 1363–1372.
53. Badge, R.M., Yardley, J., Jeffreys, A.J. and Armour, J.A. (2000) Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum. Mol. Genet.*, **9**, 1239–1244.
54. Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, **29**, 217–222.
55. Eisenbarth, I., Striebel, A.M., Moschgath, E., Vogel, W. and Assum, G. (2001) Long-range sequence composition mirrors linkage disequilibrium pattern in a 1.13 Mb region of human chromosome 22. *Hum. Mol. Genet.*, **10**, 2833–2839.
56. Service, S.K., Ophoff, R.A. and Freimer, N.B. (2001) The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum. Mol. Genet.*, **10**, 545–551.
57. Devlin, B., Roeder, K., Otto, C., Tiobech, S. and Byerley, W. (2001) Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania. *Hum. Genet.*, **108**, 521–528.
58. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **10**, 10.
59. Menashe, I., Man, O., Lancet, D. and Gilad, Y. (2002) Population differences in haplotype structure within a human olfactory receptor gene cluster. *Hum. Mol. Genet.*, **11**, 1381–1390.
60. Zavattari, P., Deidda, E., Whalen, M., Lampis, R., Mulargia, A., Loddo, M., Eaves, I., Mastio, G., Todd, J.A. and Cucca, F. (2000) Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Hum. Mol. Genet.*, **9**, 2947–2957.
61. Angius, A., Melis, P.M., Morelli, L., Petretto, E., Casu, G., Maestrale, G.B., Fraumene, C., Bebbere, D., Forabosco, P. and Pirastu, M. (2001) Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. *Hum. Genet.*, **109**, 198–209.
62. Arnason, E., Sigurgislason, H. and Benedikz, E. (2000) Genetic homogeneity of Icelanders: fact or fiction? *Nat. Genet.*, **25**, 373–374.
63. Gulcher, J., Helgason, A. and Stefansson, K. (2000) Genetic homogeneity of Icelanders. *Nat. Genet.*, **26**, 395.
64. Helgason, A., Hickey, E., Goodacre, S., Bosnes, V., Stefansson, K., Ward, R. and Sykes, B. (2001) mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *Am. J. Hum. Genet.*, **68**, 723–737.
65. Freimer, N.B., Service, S.K. and Slatkin, M. (1997) Expanding on population studies. *Nat. Genet.*, **17**, 371–373.
66. Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T., Stanley, S.E., Jiang, R., Messer, C.J., Chew, A., Han, J.H. *et al.* (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**, 489–493.
67. Hutton, M., Lendon, C.L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A. *et al.* (1998) Association of missense and 5′-splice-site mutations in *tau* with the inherited dementia FTDP-17. *Nature*, **393**, 702–705.
68. Spillantini, M.G., Murrell, J.R., Goedert, M., Farlow, M.R., Klug, A. and Ghetti, B. (1998) Mutation in the *tau* gene in familial multiple system tauopathy with presenile dementia. *Proc. Natl Acad. Sci. USA*, **95**, 7737–7741.
69. Poorkaj, P., Bird, T.D., Wijsman, E., Nemens, E., Garruto, R.M., Anderson, L., Andreadis, A., Wiederholt, W.C., Raskind, M. and Schellenberg, G.D. (1998) *Tau* is a candidate gene for chromosome 17 frontotemporal dementia. *Ann. Neurol.*, **43**, 815–825.
70. Yaspo, M.L. (2001) Taking a functional genomics approach in molecular medicine. *Trends Mol. Med.*, **7**, 494–501.
71. Pilpel, Y., Sudarsanam, P. and Church, G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.
72. Sreekumar, K.R., Aravind, L. and Koonin, E.V. (2001) Computational analysis of human disease-associated genes and their protein products. *Curr. Opin. Genet. Dev.*, **11**, 247–257.
73. Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L. and Jarvela, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.*, **30**, 233–237.
74. Belloni, E., Muenke, M., Roessler, E., Traverso, G., Siegel-Bartelt, J., Frumkin, A., Mitchell, H.F., Donis-Keller, H., Helms, C., Hing, A.V. *et al.* (1996) Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly. *Nat. Genet.*, **14**, 353–356.
75. Lettice, L.A., Horikoshi, T., Heaney, S.J., Van Baren, M.J., Van Der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N. *et al.* (2002) Disruption of a long-range *cis*-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA*, **99**, 7548–7553.
76. Pfeifer, D., Kist, R., Dewar, K., Devon, K., Lander, E.S., Birren, B., Korniszewski, L., Back, E. and Scherer, G. (1999) Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to *SOX9*: evidence for an extended control region. *Am. J. Hum. Genet.*, **65**, 111–124.
77. Lauderdale, J.D., Wilensky, J.S., Oliver, E.R., Walton, D.S. and Glaser, T. (2000) 3′ deletions cause aniridia by preventing PAX6 gene expression. *Proc. Natl Acad. Sci. USA*, **97**, 13755–13759.
78. Grosveld, F., van Assendelft, G.B., Greaves, D.R. and Kollias, G. (1987) Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell*, **51**, 975–985.

79. Heus, H.C., Hing, A., van Baren, M.J., Joosse, M., Breedveld, G.J., Wang, J.C., Burgess, A., Donnis-Keller, H., Berglund, C., Zguricas, J. *et al.* (1999) A physical and transcriptional map of the preaxial polydactyly locus on chromosome 7q36. *Genomics*, **57**, 342–351.
80. Terwilliger, J.D. and Goring, H.H. (2000) Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum. Biol.*, **72**, 63–132.
81. Baird, P.A., Anderson, T.W., Newcombe, H.B. and Lowry, R.B. (1988) Genetic disorders in children and young adults: a population study. *Am. J. Hum. Genet.*, **42**, 677–693.
82. Stumpf, M.P. (2002) Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet.*, **18**, 226–228.
83. Ardlie, K., Liu-Cordero, S.N., Eberle, M.A., Daly, M., Barrett, J., Winchester, E., Lander, E.S. and Kruglyak, L. (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.*, **69**, 582–589.
84. Ianakiev, P., van Baren, M.J., Daly, M.J., Toledo, S.P., Cavalcanti, M.G., Neto, J.C., Silveira, E.L., Freire-Maia, A., Heutink, P., Kilpatrick, M.W. and Tsipouras, P. (2001) Acheiropodia is caused by a genomic deletion in *C7orf2*, the human orthologue of the *Lmbr1* gene. *Am. J. Hum. Genet.*, **68**, 38–45.