Chapter 16

Short Tandem Repeats and Genetic Variation

Bo Eskerod Madsen, Palle Villesen, and Carsten Wiuf

Abstract

Single nucleotide polymorphisms (SNPs) are widely distributed in the human genome and although most SNPs are the result of independent point-mutations, there are exceptions. When studying distances between SNPs, a periodic pattern in the distance between pairs of identical SNPs has been found to be heavily correlated with periodicity in short tandem repeats (STRs). STRs are short DNA segments, widely distributed in the human genome and mainly found outside known tandem repeats. Because of the biased occurrence of SNPs, special care has to be taken when analyzing SNP-variation in STRs.

We present a review of STRs in the human genome and discuss molecular mechanisms related to the biased occurrence of SNPs in STRs, and its implications for genome comparisons and genetic association studies.

Key words: SNPs, Short tandem repeat, Pattern, Variation, Mechanism, Mutation, Polymorphism

Abbreviations

SNP	single nucleotide polymorphism
bp	base pair
STR	short tandem repeat

1. Introduction

Single nucleotide polymorphisms (SNPs) are widely distributed in the human genome, and are not restricted to any type of genetic elements such as exons, transcripts, transposons or tandem repeats. There are 11.9 million reported SNPs in the human genome (dbSNP (1, 2), build 128) and panels of up to 650 k SNPs have been used as markers for genetic disease susceptibility variants in genome wide association studies (3–7). SNPs are generally thought to be the result of independent mutational events which subsequently have spread in the human population

Michael R. Barnes and Gerome Breen (eds.), *Genetic Variation: Methods and Protocols*, Methods in Molecular Biology, vol. 628, DOI 10.1007/978-1-60327-367-1_16, © Springer Science+Business Media, LLC 2010

and thereby lead to nucleotide diversity in the genome (8). Much effort has gone into identifying new SNPs in the human genome and studying the frequencies of SNPs in different human populations. For example in the HapMap project, 4.0 million non-redundant SNPs (release #23, January 2008) have been genotyped in 270 individuals from four different human populations from Africa, Asia and Central Europe (9). SNP information from the HapMap project has been used to select the SNP panels for genome wide association studies, and has contributed to the validation of some of the SNPs that have been reported to dbSNP.

By comparing genomes from different species or individuals, single nucleotide variation has been used to estimate how genomes evolve over time. Simplified models such as the Jukes-Cantor (10), Felsenstein (11) and HKY (12) models are typically applied to compare the evolution of different segments of the genome. Such comparisons can identify DNA segments that are highly conserved and/or under selection, and thereby identify functionally important elements in the genome. Knowledge about functional elements is then again used in studies of how genetic variation influences the resulting phenotype (e.g. disease).

In this review, we focus on how SNP occurrence may depend on periodicity in the nucleotide composition. Nucleotides occurring in a periodic manner are known as tandem repeats, microsatellites, simple repeats or simple sequence repeats (SSR). We especially focus on short (imperfect) tandem repeats (STRs) in the human genome, relate the findings to possible molecular mechanisms for generating STRs and discuss what implications the findings may have on genetic association studies and genome comparison studies.

2. Identification of Short Tandem Repeats

In this review, we use the definition of STRs given by Madsen et al. (13) (originally called periodic DNA). In brief, a DNA segment is defined as an STR if (1) it is at least 9 bp long, (2) the repeat-unit (e.g. AT in ATATATATAT) is repeated at least three times, (3) only a few base pairs in the segment do not match the repeat-unit. To allow for sequence ambiguity, all possible SNP alleles are used in the identification of an STR (see Fig. 1).

Several algorithms, such as Tandem Repeat Finder (14), mreps (15) and TROLL (16), have been developed for the identification of tandem repeats. These algorithms are designed for general identification of tandem repeats, but care should be taken since the algorithms differ significantly in what they detect as tandem repeats (17). None of the above mentioned algorithms incorporate

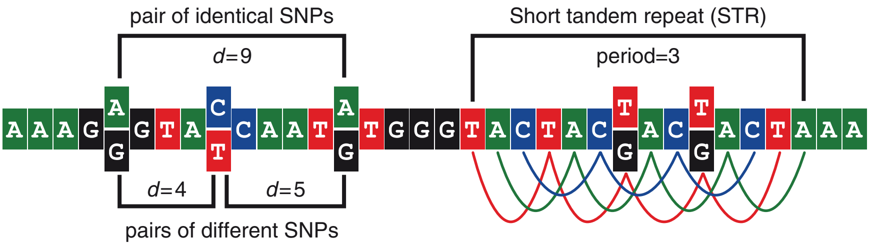
pair of identical SNPs
d=9
Short tandem repeat (STR)
period=3
A A A G A/G G T A C/T C A A T A/G T G G G T A C T A C T/G A C T/G A C T A A A
d=4
d=5
pairs of different SNPs

Fig. 1. Definitions of distances in pairs of SNP and an example of an STR. Distances are calculated between all pairs of SNPs, thus the figure shows three pairs with three distances. The distance (*d*) between any two SNPs is defined as the positive difference between the two genomic SNP positions, for example, *d*=1 corresponds to contiguous SNPs. A pair of identical SNPs is defined as two SNPs with identical alleles (here SNP1: A/G, SNP2: A/G, *d*=9). Pairs of different SNPs are defined as two SNPs with different alleles (here SNP1: A/G, SNP2: C/T, *d*=4; SNP1: C/T, SNP2: A/G, *d*=5). To the right, an example of an STR is shown. The period (*p*) is 3, and it is shown that SNPs are allowed in the pattern. Adapted in part from Madsen et al. (13), with permission from Genome Research

information about known SNP variations in the human genome, and we previously implemented a specialized algorithm for the identification of STRs (13).

STRs are widely distributed in the human genome; i.e. STRs make up 4.3% of the entire human genome, whereas 2.87% of exons and 4.3% of the entire transcribed regions are tagged as STRs (13) Furthermore, STRs are generally different from the "Simple Repeats" track from the UCSC Table Browser (18) (found using Tandem Repeat Finder (14)), as 97.17% of all STRs are found outside the track (13). The genomic content of tandem repeats in general has been investigated in several studies, and is described elsewhere (19–26).

3. A Periodic Pattern in SNP Distances

One feature of STRs is a periodic pattern in the distance between pairs of "identical SNPs" (SNPs with identical alleles). In contrast to non-STR, pairs of identical SNPs are common and clearly non-uniformly distributed in STRs (Fig. 2). If SNPs occur with the same probability independently at all sites in the genome, then the distance between two random SNPs is uniformly distributed. This does not hold true for immediately adjacent SNPs because of the high CpG mutation rate (27). Inside STR regions, pairs of identical SNPs positioned 2, 4, 6 or 8 bp apart are much more frequent than pairs of identical SNPs positioned 3, 5, 7 or 9 bp apart, whereas this pattern is completely absent for pairs of different SNPs (13). This 2, 4, 6, 8 pattern is most likely explained by biased introduction of SNPs in STRs (see Molecular mechanisms) and in concordance, there are found 1.8 times more SNPs in STRs than would have been expected by chance (13).

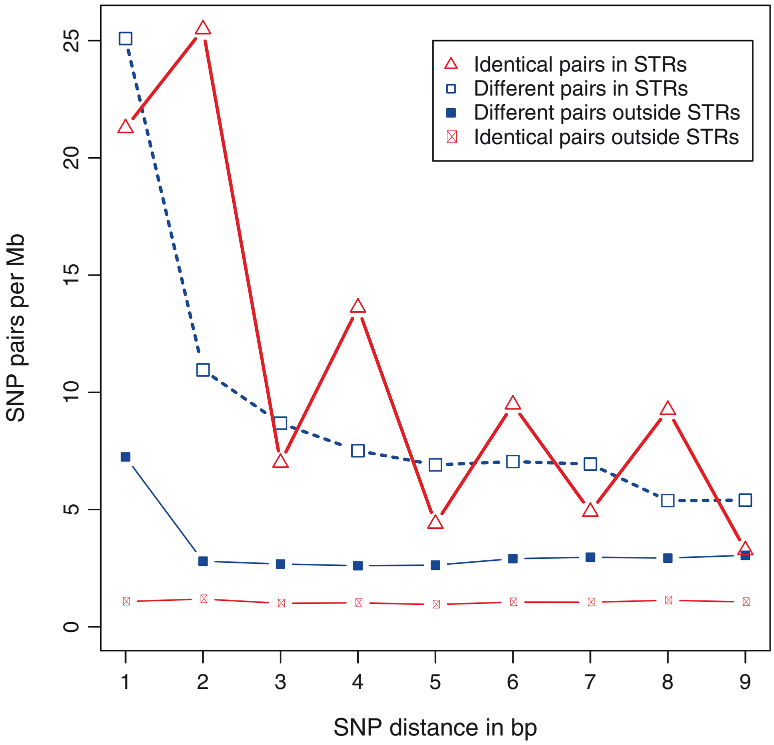

Fig. 2. Pairs of SNPs inside and outside STRs. Shown is the distance between pairs of SNPs inside and outside STRs. Both pairs of identical SNPs and pairs of different SNPs are overrepresented inside STRs when compared to outside STRs. Pairs of identical SNPs show the highest overrepresentation in STRs and identical SNPs 2, 4, 6 or 8 bp apart are much more common than identical SNPs 3, 5, 7 or 9 bp apart. Adapted in part from Madsen et al. (13), with permission from Genome Research

As for tandem repeats in general, the majority of STRs have periods of 1 or 2 bp (13, 28, 29). The 2, 4, 6, 8 pattern in SNP distances are therefore likely to be due to SNPs emerging according to the periods of STRs; i.e. if an A/G SNP is present in an STR segment ATATATATAT, then another A/G SNP in the same segment occurs more often than is expected by chance, generating pairs of identical SNPs 2, 4, 6 or 8 bp apart. If this biased emergence of SNPs is equally probable for all periods of STRs, then the 2, 4, 6, 8 pattern would be generated simply because STRs with period $p=2$ are common.

4. Molecular Mechanisms

Length variations in tandem repeats are generally thought to be generated by polymerase slippage and uneven cross over (30–35). Polymerase slippage is a mechanism, whereby the DNA polymerase skips one (or more) repeat-unit(s) in a tandem repeat, or copies a repeat-unit more than once from the template strand (36, 37). Uneven cross over is a mechanism whereby the two homologous DNA strands do not break in the same position before recombination, leading to a strand with a deletion of a segment and a strand with an insertion of the same segment (38). If these irregularities are not caught by the repair mechanisms, they lead to length variations in tandem repeats.

The observed 2, 4, 6, 8 pattern in STRs cannot be explained by misalignments of sequences due to length variations in STR segments, since only SNPs which are mapped to an exact location in the reference genome are used (13). However, this does not rule out that length variation mediates the bias towards an excess of pairs of identical SNPs in STRs. E.g. if a repeat-unit is inserted at the left side of the C in ATCTATATAT, generating the "temporary" sequence ATATCTATATAT, and a repeat-unit subsequently is removed on the right side of C, we get the two sequences ATCTATATAT and ATATCTATAT in the population, which will be interpreted as two A/C SNPs in distance $d = 2$ bp (see Fig. 3). Repair mechanisms may tend to correct for insertions in the same meiotic cycle as they are introduced and thereby generate pairs of identical SNPs in STRs, as just explained. Alternatively, an inversion of 3 bp (e.g CTA) yields a pair of identical SNPs too. A second independent length-mutation in a STR can result in the

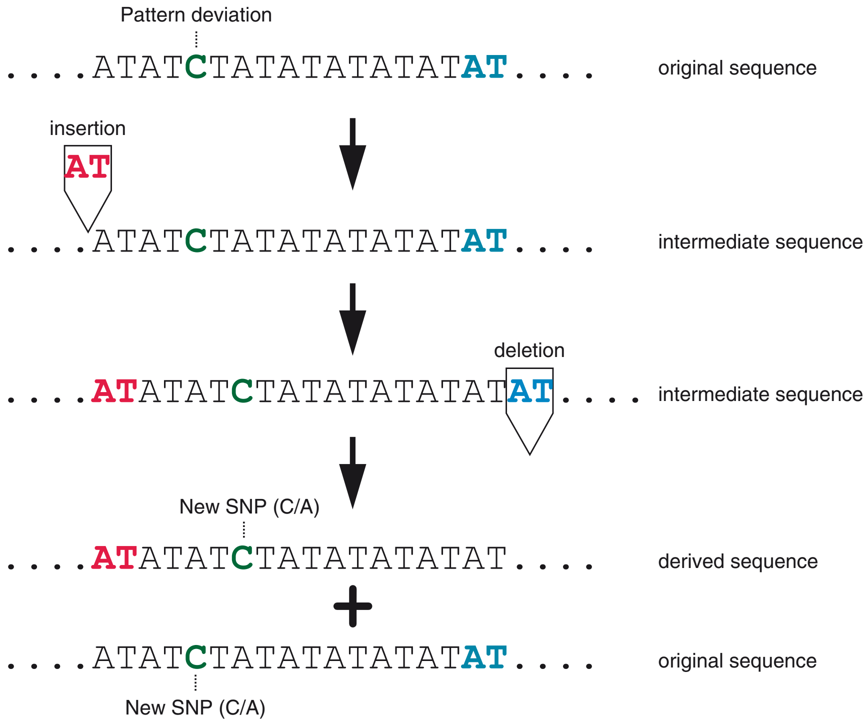

Fig. 3. A molecular mechanism for generating a pair of identical SNPs

same, but this scenario is less probable since two independent mutations are needed. Another possibility is gene conversion, where a DNA segment is copied to a new position without creating a length polymorphism (39, 40). Complex mechanisms of context dependent generation of point mutations could explain the observed pattern as well, but no such mechanism are known. It is worth noting, however, that the elevated mutation rate in CpG islands (27) is context dependent, and the importance of such a mechanism can not be ruled out per se.

5. Genetic Association Studies

Like other forms of genetic variation, insertion deletion polymorphisms (indels) are of great interest because they may influence gene function and cause disease. An example is Fragile X Syndrome that is caused by expansion of a three-nucleotide tandem repeat in the FMR-1 gene (41–44). Likewise, cystic fibrosis is frequently caused by a three bp deletion that eliminates a single amino acid from the protein encoded by the CFTR gene (45–48). Next-generation sequencing technologies may enable identification of new disease susceptibility variants by resequencing a large number of disease cases and controls. However, sequencing the entire genome of a large group of affected individuals may still be prohibitively expensive for years to come and identification of probable targets for disease causing variants may be useful. Hypermutable segments of functional genomic elements (exons) are probable targets for disease related mutations and may therefore be good candidates for resequencing studies. Tandem repeats are well known to be hypermutable and to have an excess of indels compared to the rest of the genome, but tandem repeats are rare in functional elements such as exons (20, 28, 35, 49–51). In contrast, STRs are widely distributed in the human genome (13, 28) and since they share the hypermutability of longer tandem repeats (unpublished results), they may be targets for disease causing mutations. If hypermutable segments are located in “junk” (uninformative) DNA, mutations are not disease causing. Tandem repeats are mainly thought to be “junk” DNA, but several studies have shown that tandem repeats can have a functional role. Examples of tandem repeat related functions are differentiated transcription activity of human genes (52), and the ability of pathogens to adapt to their host (26). Additional examples of functional tandem repeats are reviewed elsewhere (24, 35, 53–56).

The call-rate for genotyping SNPs in the HapMap (9) study has been shown to be significantly lower for SNPs located inside STRs (13). This supports that STRs are hypermutable and emphasizes that care should be taken when SNP studies are designed and analyzed. Besides affecting the call-rate, structural variants may lead to genotyping errors, if the DNA sequence is altered close to a SNP position and a wrong genomic position is read for the SNP. Such a bias may be difficult to identify and precautionary steps should be taken in the study design. A strategy to minimize the impact of STRs in genotyping studies is simply to avoid SNPs inside or near STRs. The downside of this strategy is that variants in some parts of the genome are poorly covered in the study and hence disease associated variants may be missed. Resequencing STR segments would solve the problem, but that approach may be too expensive in many studies.

As it has been debated for tandem repeats (24, 52, 54), STRs may serve functional roles in the genome. One possibility is that DNA and/or RNA fold according to the repeated sequence of STRs and thereby influence gene function (35). A mutation in such an STR may alter the folding and thus the function. Furthermore, hypermutable regions (e.g. in exons) may introduce a high level of phenotypic variation and thereby allow for fast adaptation to a changing environment. Although hypermutability in functional elements may have been beneficial throughout evolution, disease related variants may also be introduced in an elevated rate in such regions. Hypermutable segments with functional roles may be obvious candidates for resequencing studies, since a high density of rare disease susceptibility variants are expected.

6. Genome Comparison Studies

Models for genome comparison usually assume that mutations occur independently and a violation on this assumption may bias findings. The excess of pairs of identical SNPs in STRs clearly show that the assumption of independent mutations is not always valid, and hence care must be taken. Since it is not known whether the underlying molecular mechanism(s) is (are) restricted to STRs or just visible in these segments, excluding STRs from genome comparison studies may not guarantee that the analyzed variation have occurred independently.

7. Concluding Remarks

The presence of a periodic pattern of SNPs in STRs emphasizes that care should be taken when using SNPs in disease association studies and genome comparisons. Further studies are needed to clarify what mechanisms underlie the excess of pairs of identical

SNPs in STRs. Investigations of how common insertion or deletion of repeat-units is in STR regions may help to distinguish between some of the possible mechanisms, whereas identifying the exact mechanism(s) may be difficult.

Whether STRs are associated with gene function, or are a probable target for disease-causing mutations, remains an open question, but it is worth giving a second thought.

References

1. Sherry, S.T., Ward, M. and Sirotkin, K. (1999) dbSNP – database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **9**, 677–679.
2. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
3. Eberle, M.A., Ng, P.C., Kuhn, K., Zhou, L., Peiffer, D.A., Galver, L., *et al.* (2007) Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet.*, **3**, e170.
4. Fan, J.-B., Chee, M.S. and Gunderson, K.L. (2006) Highly parallel genomic assays. *Nat. Rev. Genet.*, **7**, 632–644.
5. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D.P., Thompson, D., Ballinger, D.G., *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.
6. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**, 881–885.
7. The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
8. Stoneking, M. (2001) Single nucleotide polymorphisms. From the evolutionary past. *Nature*, **409**, 821–822.
9. The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
10. Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In Munro, H.N. (ed.), *Mammalian Protein Metabolism.* Academic Press, New York.
11. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
12. Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.
13. Madsen, B.E., Villesen, P. and Wiuf, C. (2007) A periodic pattern of SNPs in the human genome. *Genome Res.*, **17**, 1414–1419.
14. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
15. Kolpakov, R., Bana, G. and Kucherov, G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
16. Castelo, A.T., Martins, W. and Gao, G.R. (2002) TROLL – tandem repeat occurrence locator. *Bioinformatics*, **18**, 634–636.
17. Leclercq, S., Rivals, E. and Jarne, P. (2007) Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics*, **8**, 125.
18. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
19. Boby, T., Patch, A.M. and Aves, S.J. (2005) TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics*, **21**, 811–816.
20. Borstnik, B. and Pumpernik, D. (2002) Tandem repeats in protein coding regions of primate genes. *Genome Res.*, **12**, 909–915.
21. O'Dushlaine, C., Edwards, R., Park, S. and Shields, D. (2005) Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biol.*, **6**, R69.
22. Hancock, J.M. and Simon, M. (2005) Simple sequence repeats in proteins and their significance for network evolution. *Gene*, **345**, 113–118.
23. Alba, M.M. and Guigo, R. (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.*, **14**, 549–554.

24. Kashi, Y. and King, D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.*, **22**, 253–259.
25. Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F. and Makova, K.D. (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.*, 18, 30–38.
26. Mrazek, J., Guo, X. and Shah, A. (2007) Simple sequence repeats in prokaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 8472–8477.
27. Hwang, D.G. and Green, P. (2004) Inaugural article: Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 13994–14001.
28. Lai, Y. and Sun, F. (2003) The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. *Mol. Biol. Evol.*, **20**, 2123–2131.
29. Almeida, P. and Penha-Goncalves, C. (2004) Long perfect dinucleotide repeats are typical of vertebrates, show motif preferences and size convergence. *Mol. Biol. Evol.*, **21**, 1226–1233.
30. Levinson, G. and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.
31. Pearson, C.E., Edamura, K.N. and Cleary, J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.*, **6**, 729–742.
32. Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
33. Chambers, G.K. and MacAvoy, E.S. (2000) Microsatellites: consensus and controversy. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.*, **126**, 455–476.
34. Kruglyak, S., Durrett, R.T., Schug, M.D. and Aquadro, C.F. (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 10774–10778.
35. Mirkin, S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.
36. Weber, J.L. and Wong, C. (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.*, **2**, 1123–1128.
37. Walsh, P.S., Fildes, N.J. and Reynolds, R. (1996) Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Res.*, **24**, 2807–2812.
38. Jeffreys, A.J., Barber, R., Bois, P., Buard, J., Dubrova, Y.E., Grant, G., *et al.* (1999) Human minisatellites, repeat DNA instability and meiotic recombination. *Electrophoresis*, **20**, 1665–1675.
39. Holliday, R. (1964) A mechanism for gene conversion in fungi. *Genet. Res.*, **5**, 282–304.
40. Lewin, B. (2004) *Genes VIII*. Prentice Hall, New Jersey.
41. Warren, S.T., Zhang, F., Licameli, G.R. and Peters, J.F. (1987) The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. *Science*, **237**, 420–423.
42. Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., *et al.* (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science*, **252**, 1711–1714.
43. Verkerk, A.J.M.H., Pieretti, M., Sutcliffe, J.S., Fu, Y.-H., Kuhl, D.P.A., Pizzuti, A., *et al.* (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905–914.
44. Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., *et al.* (1991) Fragile X genotype characterized by an unstable region of DNA. *Science*, **252**, 1179–1181.
45. Collins, F.S., Drumm, M.L., Cole, J.L., Lockwood, W.K., Vande Woude, G.F. and Iannuzzi, M.C. (1987) Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science*, **235**, 1046–1049.
46. Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245(4922)**, 1073–1080.
47. Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., *et al.* (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, **245(4922)**, 1066–1073.
48. Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., *et al.* (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, **245(4922)**, 1059–1065.
49. Ellegren, H. (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.*, **16**, 551–558.
50. Toth, G., Gaspari, Z. and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.

51. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
52. Lawson, M.J. and Zhang, L. Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5′-UTR region. *Gene*, 407, 54–62.
53. Thomas, E.E. (2005) Short, local duplications in eukaryotic genomes. *Curr. Opin. Genet. Dev.*, **15**, 640–644.
54. Li, Y.-C., Korol, A.B., Fahima, T. and Nevo, E. (2004) Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.*, **21**, 991–1007.
55. Sutherland, G.R. and Richards, R.I. (1995) Simple tandem DNA repeats and human genetic disease. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 3636–3641.
56. Zuckerkandl, E. (2002) Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine. *Genetica*, **115**, 105–129.