SSRs as genetic markers in the human genome and their observable relationship to hereditary diseases

“...short sequence repeats have been popularly used as genetic markers for artificial trait selection and forensic applications.”

First draft submitted: 1 April 2016; Accepted for publication: 4 April 2016; Published online: 27 May 2016

Keywords: genetic disorder • microsatellite • next-generation sequencing (NGS)
• SSR instability • SSR somatic mutations

Short sequence repeats (SSRs) are abundantly distributed throughout the genome of all species. Owing to high levels of polymorphism in the repeat lengths between different species and individuals, and functional genetic characteristics, SSRs have been popularly used as genetic markers for artificial trait selection and forensic applications. In addition, numerous SSRs have been considered for their association with genetic diseases and gene regulatory functions, since expansion of unstable SSRs could induce neurodegenerative and neuromuscular disorders. Only a limited number of SSR markers were discovered so far. Thorough scanning and prediction of potential SSR markers from the genome of a species and identification of significant features of SSRs, either for functional conservation or polymorphic properties between different species/individuals, are challenges that cannot be achieved through empirical research alone. At present, high-throughput sequencing technologies such as next-generation sequencing (NGS), incorporated with powerful computing techniques, provide efficient and low-cost mechanisms of genome-scale analyses for biomarker discovery [1]. Groups of candidate SSR biomarkers could be detected by cross-genome comparisons with respect to a specific functional gene group, and successful identification would reveal polymorphic SSRs in functional domains and thus facili-

tate experimental designs for various biological applications. However, sometimes the NGS sequenced reads have relatively short read lengths as compared with the original long SSR motifs or the sequenced reads contain only repeated segments, which may cause uncertainty and errors during genome assembly procedures. This is especially true in the case of *de novo* species assembly where reference genomes are lacking, and the misassembled contigs and fragmented assemblies are formulated mainly due to the nature of de Bruijn graph-based algorithms [2]. This situation can be improved by selecting long-read sequencing technologies for *de novo* species and paired-end sequencing technologies for model species [3]. Here, we mainly focus on discussing SSR markers for the human genome. All SSR motifs from the Ensembl updated version of the human genome (GRCh38.p5) were identified by using the CGSSR tool, and three different tolerance types including insertion, deletion and substitution were considered simultaneously [4]. The obtained SSRs were set with two tolerance rates of 80% (imperfect SSRs) and 100% (perfect SSRs), and a minimum total repetition length of 18 bp was considered for different fundamental repeat unit lengths. Using the updated version of the human genome, we found that the proportion of identified perfect SSRs with repeat unit lengths ranging from 1 to 6 bp are 36.19, 20.31, 12.59, 15.21, 10.29 and

Tun-Wen Pai

Author for correspondence:
Department of Computer Science
& Engineering, National Taiwan Ocean
University, No. 2, Pei-Ning Road,
Keelung 20224, Taiwan, R.O.C.
Tel.: +886 224 622 192 ext. 6618
twp@mail.ntou.edu.tw

Chien-Ming Chen

Department of Computer Science
& Engineering, National Taiwan Ocean
University, No. 2, Pei-Ning Road,
Keelung 20224, Taiwan, R.O.C.

5.42%, respectively. The percentage of SSRs located within different genetic regions are, coding region of 0.42%, intron of 52.55%, 5' UTR of 0.38%, 3' UTR of 0.99%, upstream of 13.30% (5' flank segment of 5 kbp), downstream of 12.84% (3' flank segment of 5 kbp), noncoding exon of 1.38% and intergenic regions of 36.84% (SSRs in intergenic regions may overlap with defined upstream, downstream and noncoding exon regions). It was observed that only 0.42% SSRs are located within coding regions, of which more than 54.27% are trinucleotide SSRs.

Conserved & polymorphic features of SSR markers for various applications

In our previous report, we compared seven human genomes and created a virtual template SSR profile for detecting polymorphic SSRs among different human family genomes [5]. All conserved and polymorphic SSRs across different individual sequence reads and genes were successfully identified. In addition, we verified the SSR markers of the family genomes based on Combined DNA Index System (CODIS) and concluded that distinguishability for the original 13 SSRs could be further expanded in the post-NGS era, especially for forensic identification at the individual level. Meanwhile, the FBI Laboratory announced in 2015 that seven additional SSR loci were selected to enhance greater discrimination potential for human identification applications [6]. To investigate the extent of SSR polymorphisms between different gene groups, we created two more contrast gene sets including a homologous gene set (225 orthologous genes with high-sequence similarities among ten species from fish to mammalian species) and a human-unique gene set (492 human-unique genes compared with five other mammalian species: gorilla, chimpanzee, macaque, orangutan and mouse). This observation suggests that characteristics of conserved SSRs in the homologous gene set are more than fivefold greater among various species compared with CODIS and human-unique gene sets. In other words, the highest average occurrence rate of polymorphic SSRs is for human-unique genes (1.70 polymorphic SSRs/Mbp) compared with that of CODIS genes (0.35 polymorphic SSRs/Mbp) and homologous genes (0.32 polymorphic SSRs/Mbp), suggesting that SSRs that are well conserved across species might convey important messages for gene regulation in the cell and polymorphic SSR markers found in human-unique genes can be applied as distinctive features for identifying different individuals. This, in turn may provide a strategy for selecting SSR markers specific for gene function-oriented and strain or species identification-oriented applications.

Unstable SSR expansion associated with human diseases

By performing cross-genome comparisons for various species, it was shown that the distribution of SSRs is nonrandom and interspersed within the whole genome. In the human genome, the most common SSRs are mononucleotide repeats of the (A/T) SSRs (~34.25% among all 501 SSR patterns), and they are highly polymorphic among individuals [7]. Three mutation mechanisms have been proposed for mononucleotide repeats, including slipped-strand misalignment (insertion/deletion), dNTP-mediated misalignment (deletion) and misincorporation-misalignment (insertion or deletion) [8]. These indels of mononucleotide repeats are expected to be dependent on the DNA sequence and DNA polymerase. Nevertheless, the mutation mechanisms of SSRs including mononucleotides and non-monomonucleotides are recognized by replication slippage, repair, recombination or sometimes in combination [9]. When mutations evade DNA mismatch repair mechanisms, they may lead to SSR instability (or called as microsatellite instability) and are responsible for human hereditary neurodegenerative/neuromuscular disorders. Presently, more than 40 reported monogenic disorders are caused by SSR instability [10–12]. Additionally, unstable SSRs exhibit genetic anticipation mechanisms whereby the symptoms of the genetic disorder become apparent and increase in severity at an earlier age with each generation. Larger repeat copy number mutations could induce more severe phenotypes, and generally, they tend to increase the repeat numbers for the next generation. A few comprehensive reviews of human diseases associated with SSR instability were reported recently [13,14]. The pathogenic mechanisms of SSR expansions can be categorized into three major types:

- Located within noncoding regions resulting in loss of protein function;
- Located within coding regions resulting in altered protein function;
- Located within noncoding regions resulting in altered RNA function.

The first type of human diseases caused by SSR expansions of noncoding repeats include fragile X (FRAXA) with (CGG)_n expansion in 5' UTR, fragile XE (FRAXE) with (CGG)_n expansion in 5' UTR, Friedreich ataxia with (GAA)_n expansion in the first intron and progressive myoclonic epilepsy (EPM1) with (C₄GC₄GCG)_n expansion in the promoter. These repeat expansions mediate pathogenesis by transcriptional silencing, with larger repeats

(>200 repeats) resulting in greater gene silencing. The second type of expansion caused by SSR mutations in coding regions describes disorders caused by gain of function. A total of 14,255 SSRs (0.42%) are located within coding regions of the human genome and more than half of them are trinucleotide SSR motifs. The most abundant triplet patterns in coding regions include CCG, AGC, AGG and AAG (including frame-shift and complementary patterns), resulting in translation of these SSR motifs into corresponding repeat amino acid peptides. It was observed that the most frequently occurring repeat amino acid peptides in the human genome are poly-E (glutamic acid), A (alanine), P (proline) and L (leucine) representing 11.16, 9.53, 9.49 and 7.14%, respectively, of all amino acid repeats in human proteins. The most well-known expanded SSR repeats causing human diseases are polyglutamine (poly-Q, 3.63% appearance rate) disorders. There is a group of genetically distinct disorders caused by expansions of translated CAG repeats including dentatorubral-pallidoluysian atrophy, Huntington's disease, spinobulbar muscular atrophy and spinocerebellar ataxia 1, 2, 3, 6, 7 and 17 (SCA1/2/3/6/7/17). The other type of SSR unstable in the coding region is the triplet (GCN)_n, which causes polyalanine-mediated disorders including blepharophimosis-ptosis-epicanthus inversus, cleidocranial dysplasia, congenital central hypoventilation syndrome, hand-foot-genital syndrome, holoprosencephaly 5, X-linked infantile spasm syndrome, mental retardation, x-linked, with isolated growth hormone deficiency (MRGH), oculo-pharyngeal myotonic dystrophy syndrome and synpolydactyly syndrome disorders. Mutant proteins of polyglutamine and polyalanine form aggregated forms and lead to cell death and tissue degeneration through a series of inappropriate interactions such as conformational changes in mutant proteins that interact with numerous transcription regulators and lead to transcriptional dysregulation and mitochondrial dysfunction [15]. Furthermore, both expansion mutations were shown to be bidirectionally transcribed and located either within or outside the ATG-initiated open reading frame. The latter type of no ATG initiation codon, called the repeat associated non-ATG translation and the translated proteins in three open reading frames containing no N-terminal methionine initiation codon are toxic in general, such as the (GGGGCC)_n expansion in *C9orf72* open reading frame translated into aggregating dipeptide-repeat proteins in frontotemporal dementia and amyotrophic lateral sclerosis [16]. The last type of SSR unstable disorders are caused by mutations in non-coding regions and result in altered RNA function [17]. These disorders include dystro-

phia myotonia 1 with triplet (CTG)_n in the 3' UTR, dystrophia myotonia 2 (DM2) with tetranucleotide (CCTG)_n expansion in intron 1, mutants of FXTAS with expansion of (CGG)_n in the 5' UTR, spinocerebellar ataxia 8 (SCA8) with expansion of (CTG)_n in the 3' UTR, spinocerebellar ataxia 10 with expansion of (ATTCT)_n in the first intron, spinocerebellar ataxia 12 (SCA12) with expansion of (CAG)_n in the 5' UTR, Huntington's-disease-like 2 with expansion of (CTG)_n in the 3' UTR region. The mechanisms for disease-associated RNA gain-of-function were mainly categorized into protein sequestration, RNA degradation and chromatin silencing. Although the above-mentioned disorders might be caused by different SSR unstable expansions, all mutations involve aberrant binding of expanded RNA to a specific set of RNA-binding proteins, leading to dysregulation of protein function. However, further studies are required to elucidate the spectrum of toxic RNA gain-of-function mutations involved in disease development [17].

Genome-scale analysis for SSR instability

Recently, SSR instability has become a popular research issue for detecting its association with somatic mutations, which can be comprehensively verified by the emergence of NGS approaches. Through cross-genome and/or cross-exome comparison, perhaps the SSR instability and associated genetic markers can be used to distinguish between DNA from tumors and matched normal samples [18,19]. These identified markers may play a significant role in medical diagnostic applications and personalized cancer therapy. Several cancers such as colorectal, prostate, gastric, endometrial, ovarian and breast cancers were reported for their association with SSR instability [20]. However, for accurate assessment of SSR instability status for a specific cancer type, a few challenges need to be overcome, such as limited published clinical information about tumor sequences for reference or comparison, the inability to allocate nonexonic SSR mutations obtained by NGS differential expression analysis and the inability to accurately measure low somatic length changes from long SSR markers. These challenges perhaps can be resolved by incorporating whole-genome and high-coverage rate sequencing by longer read sequencers. A more intensive cooperation between biologists, bioinformaticians and statisticians is required to identify effective SSR markers comprehensively. We believe that the multipurpose applications of SSR genetic markers in the human genome can play an important role and further gain more popularity in the near future, especially in the fields of medical diagnosis, personalized therapy and DNA profiling in forensic kinship identification.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employ-

ment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

- 1 Yu JN, Won C, Jun J, Lim Y, Kwak M. Fast and cost-effective mining of microsatellite markers using NGS technology: an example of a Korean water deer *Hydropotes inermis argyropus*. *PLoS ONE* 6(11), e26933 (2011).
- 2 Paszkiewicz K, Studholme DJ. *De novo* assembly of short sequence reads. *Brief Bioinform.* 11(5), 457–472 (2010).
- 3 Zalapa JE, Cuevas H, Zhu H *et al.* Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. Bot.* 99(2), 193–208 (2012).
- 4 Pai TW, Chen CM, Hsiao MC, Cheng R, Tzou WS, Hu CH. An online conserved SSR discovery through cross-species comparison. *Adv. Appl. Bioinform. Chem.* 2, 23–35 (2009).
- 5 Chen CM, Sio CP, Lu YL, Chang HT, Hu CH, Pai TW. Identification of conserved and polymorphic STRs for personal genomes. *BMC Genomics* 15(Suppl. 10), S3 (2014).
- 6 Hares DR. Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17, 33–34 (2015).
- 7 Montgomery SB, Goode DL, Kvikstad E *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* 23(5), 749–761 (2013).
- 8 Baptiste BA, Jacob KD, Eckert KA. Genetic evidence that both dNTP-stabilized and strand slippage mechanisms may dictate DNA polymerase errors within mononucleotide microsatellites. *DNA Repair (Amst.)* 29, 91–100 (2015).
- 9 Bowater RP, Wells RD. The intrinsically unstable life of DNA triplet repeats associated with human hereditary disorders. *Prog. Nucleic Acid. Res. Mol. Biol.* 66, 159–202 (2001).
- 10 Lopez Castel A, Cleary JD, Pearson CE. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat. Rev. Mol. Cell. Biol.* 11(3), 165–170 (2010).
- 11 Gemayel R, Vences MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* 44, 445–477 (2010).
- 12 Pearson CE, Nichol Edamura K, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6(10), 729–742 (2005).
- 13 Mirkin SM. Expandable DNA repeats and human disease. *Nature* 447(7147), 932–940 (2007).
- 14 Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.* 6(10), 743–755 (2005).
- 15 Panov AV, Gutekunst CA, Leavitt BR *et al.* Early mitochondrial calcium defects in Huntington's disease are a direct effect of polyglutamines. *Nat. Neurosci.* 5(8), 731–736 (2002).
- 16 Cleary JD, Ranum LP. Repeat associated non-ATG (RAN) translation: new starts in microsatellite expansion disorders. *Curr. Opin. Genet. Dev.* 26, 6–15 (2014).
- 17 Ranum LP, Cooper TA. RNA-mediated neuromuscular disorders. *Annu. Rev. Neurosci.* 29, 259–277 (2006).
- 18 Huang MN, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG. MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. *Sci. Rep.* 5, 13321 (2015).
- 19 Sonay TB, Koletou M, Wagner A. A survey of tandem repeat instabilities and associated gene expression changes in 35 colorectal cancers. *BMC Genomics* 16(1), 702 (2015).
- 20 Woerner SM, Benner A, Sutter C *et al.* Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene* 22(15), 2226–2235 (2003).