5728–5741 *Nucleic Acids Research, 2014, Vol. 42, No. 9* *Published online 20 March 2014*
doi: 10.1093/nar/gku212

Large-scale analysis of tandem repeat variability in the human genome

Jorge Duitama[1,2,†], Alena Zablotskaya[3,4], Rita Gemayel[1], An Jansen[1,3,4], Stefanie Belet[3,4], Joris R. Vermeesch[5], Kevin J. Verstrepen[1] and Guy Froyen[3,4]

[1]VIB lab for Systems Biology & CMPG Lab for Genetics and Genomics, KU Leuven, B-3001 Leuven, Belgium, [2]Agrobiodiversity Research Area, International Center for Tropical Agriculture (CIAT), Cali, Colombia, [3]Human Genome Laboratory, VIB Center for the Biology of Disease, Leuven, Belgium, [4]Human Genome Laboratory, Department of Human Genetics, KU Leuven, B-3000 Leuven, Belgium and [5]Center for Human Genetics, University Hospitals Leuven, KU Leuven, B-3000 Leuven, Belgium

Received November 1, 2013; Revised February 27, 2014; Accepted February 28, 2014

ABSTRACT

Tandem repeats are short DNA sequences that are repeated head-to-tail with a propensity to be variable. They constitute a significant proportion of the human genome, also occurring within coding and regulatory regions. Variation in these repeats can alter the function and/or expression of genes allowing organisms to swiftly adapt to novel environments. Importantly, some repeat expansions have also been linked to certain neurodegenerative diseases. Therefore, accurate sequencing of tandem repeats could contribute to our understanding of common phenotypic variability and might uncover missing genetic factors in idiopathic clinical conditions. However, despite long-standing evidence for the functional role of repeats, they are largely ignored because of technical limitations in sequencing, mapping and typing. Here, we report on a novel capture technique and data filtering protocol that allowed simultaneous sequencing of thousands of tandem repeats in the human genomes of a three generation family using GS-FLX-plus Titanium technology. Our results demonstrated that up to 7.6% of tandem repeats in this family (4% in coding sequences) differ from the reference sequence, and identified a *de novo* variation in the family tree. The method opens new routes to look at this underappreciated type of genetic variability, including the identification of novel disease-related repeats.

INTRODUCTION

Repetitive DNA sequences make up a significant portion of all genomes. Almost half of the human genome is comprised of repeats (1). A subset of repeated DNA is composed of tandem repeats, which are stretches of DNA that consist of tandemly repeated short sequence units (e.g. CAG) next to each other. The terms microsatellites and minisatellites are also frequently used to denote tandem repeats of short ($\leq$ 9bp) or long ($>$ 9bp) repeated units, respectively (for a complete list of terms used in this manuscript, see Supplementary Text S1). Tandem repeats can be mutational hotspots due to their repetitive nature; slippage during DNA replication or recombination events generate alleles that differ in the number of repeated units (called 'copy numbers'). Compared to other genomic loci, the mutation rates of tandem repeats are 10 to 10 000 fold higher (2). Because of this instability and apparent lack of genetic information, most tandem repeats were thought to be devoid of direct biological function and termed 'junk' DNA (3). Tandem repeats did prove extremely useful as genetic markers in fine-scale genotyping and forensics. They also provide an added advantage to genome-wide linkage studies because of their higher diversity compared to single nucleotide polymorphisms (SNPs) (4). In certain cancers, the mutational spectrum of microsatellites appears to be tumor-type specific, thus opening new avenues for the use of microsatellites as genetic markers for disease diagnosis (5).

While many tandem repeats (also called 'repeats' further on) are present in gene deserts, the accumulation of whole genome sequencing data showed that repeats are also present in functional (coding and regulatory) regions of the genomes.

Past research demonstrated that tandem repeats located within regulatory or coding regions can act as variable "tuning knobs" that can tune the function or expression of a

†To whom correspondence should be addressed. Tel: +32 16 330130; Fax: +32 16 300084; Email: Guy.Froyen@med.kuleuven.be
Correspondence may also be addressed to Kevin Verstrepen. Tel: +32 16 751390; Fax: +32 16 751391; Email: Kevin.verstrepen@BIW.VIB-kuleuven.be
*The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

gene (6). Most of this research was focused on simple model organisms such as *Saccharomyces cerevisiae* (7). However, recent findings suggest that repeats are equally important sources of phenotypic variability and disease in higher eukaryotes, including plants, metazoans, mammals and humans (8,9). For example, an intriguing study by Fondon and Garner (10) uncovered a strong correlation between variation in repeats located in two key regulatory genes (*Alx4* and *Runx2*) and skeletal morphology in dogs, suggesting that repeat variation in these genes may affect skull shape. Moreover, instability in such coding or regulatory repeats can have devastating consequences for human health. There are several examples where expansion of a repeat close or even within a human gene leads to severe diseases: Huntington disease, fragile X syndrome, and spinal and bulbar muscular atrophy are among them (11–13). In all these progressive disorders the severity and the age of onset of symptoms are directly correlated with repeat copy number in a particular gene, since allelic differences in tandem repeat copy numbers can influence allelic expression, e.g. in case of the *ATRX* gene (14). In addition, tandem repeat variability in certain genes (e.g. Thymidylate Synthase gene) are associated with a poor prognosis in a number of cancers (15,16). However, despite the high number of genes potentially affected by tandem repeats, and despite the growing evidence of the functional role of variable repeats, most studies that report genetic variation do not consider repeat variability and only focus on SNPs and copy number polymorphisms of larger segments (>1 kb up to few Mb).

The ubiquitous presence of repeats in functional parts of genomes in spite of the potentially devastating consequences of their variability suggests that repeats might also serve a beneficial functional role. Moreover, tandem repeats are not randomly present in coding sequences. Genes that encode regulatory, cell-wall and stress-induced proteins are particularly enriched in repeats, whereas genes encoding metabolic enzymes are depleted. Strikingly, this functional enrichment is evolutionary conserved from yeasts to humans (2,8,17). As several reports documented, variable tandem repeats can provide functional diversity allowing rapid evolution of phenotypes. In *S. cerevisiae*, gradual changes in intragenic tandem repeats in the *FLO1* gene (coding for a cell-surface protein) lead to gradual changes in the adhesion properties of the cells, allowing tuning of biofilm formation (2). Similarly, variable tandem repeats in promoters allow fine-tuning and rapid divergence of downstream gene expression (7,18,19). In the human genome, evidence suggests that some tandem repeat polymorphisms are under positive selection in certain parts of the world (20).

Despite their prevalence in functional parts of genomes and their association with almost 20 human neurological diseases, and despite their usefulness as genetic markers, tandem repeats remain understudied. A precise mapping of all tandem repeats in the human genome is not fully achieved, and the relevant literature often contains conflicting results. For example, the lobSTR software (21) called 45 461 microsatellite loci from a trio of sequenced genomes, far less than the 376 685 microsatellites detected by another study using the same genome (22). This is mostly due to the technical difficulties associated with sequencing repeats. Although improved methods are being introduced, the current standard next generation sequencing (NGS) techniques are not adequate for tandem repeat genotyping because reads with short lengths cannot be confidently aligned to genomic regions with tandem repeats. This technical shortcoming implies that biologically important variations are being missed in many of today's genomics studies. As a result, even basic aspects of tandem repeats, such as the degree of variability between individuals and whether this (non-pathological) variability has functional consequences, remain open questions. Such research questions require long reads (>300 bp) that are currently restricted to the relatively low-throughput NGS methods GS-FLX and SMRT (Roche).

Although new analysis software packages specifically designed for genotyping tandem repeats from short reads have been recently published (21,23), they are only able to genotype repeats with short unit length and low copy numbers from Illumina whole-genome sequencing libraries. Here, we set out to develop a strategy that permits a more accurate mapping of tandem repeats and also allows better assessment of repeat variability between individuals. We report a method based on targeted enrichment for tandem repeats in the human genome, followed by sequencing using the Roche 454 platform. Using this method, we were able to reliably genotype over 1600 tandem repeats in seven members of a three-generation family. We performed extensive polymerase chain reaction (PCR) validation experiments to confirm the accuracy of our method and we investigated the properties of tandem repeats that can be targeted using this technology. Our genotype calls reveal a surprising degree of variability within tandem repeats, even within repeats located in coding regions and between direct relatives.

MATERIALS AND METHODS

Selection of DNA samples

The protocol was approved by the Medical Ethical Commission of the University Hospitals Leuven (Belgium) with number B322201111336. The family was chosen because of the availability of sufficient DNA from seven family members in three generations. Informed consent was obtained from each family member or their parents. Genomic DNA was extracted from peripheral blood according to standard procedures. The pedigree of the selected family is shown in Figure 1 and includes the grandparents (grandfather: GF, and grandmother: GM), their daughter (mother: Mo) and her husband (father: Fa), and the three children (Ch1, Ch2, Ch3). This pedigree thus has four trios, which allowed us to look at *de novo* mutations in tandem repeats and at general variation compared to the reference genome and within the family.

Selection of tandem repeats

We downloaded the last human reference genome (hg19) from the UCSC server (24). To retrieve regions with tandem repeats in this reference, we ran the tool ETANDEM available in the EMBOSS package (18) with default parameters. ETANDEM calculates a consensus sequence for a putative repeat region and scores potential repeats based on

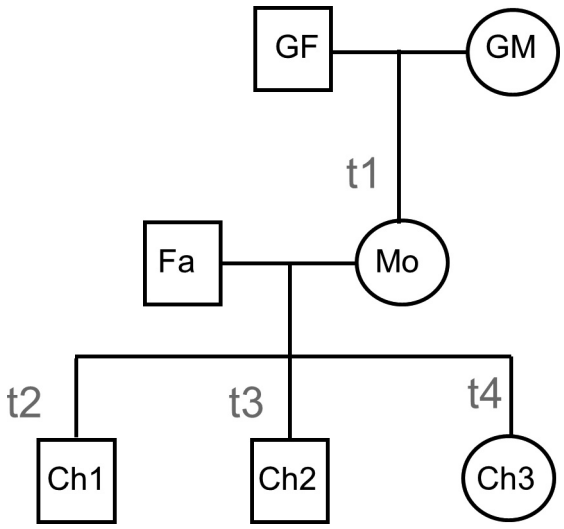

Figure 1. Pedigree of the three-generation family consisting of seven family members from whom the tandem repeats were sequenced. The four trios are indicated with t1, t2, t3 and t4. In the first 454 runs, samples GF, GM, Fa and Mo were sequenced. In the second run, samples Mo, Ch1, Ch2 and Ch3 were analyzed. GF: grandfather; GM: grandmother; Fa: father; Mo: mother; Ch1–3: child 1–3.

the number of matches and mismatches there are to the consensus: the score is incremented (+1) for a match and decremented (−1) for a mismatch (http://emboss.sourceforge.net/apps/release/6.0/emboss/apps/etandem.html). Because ETANDEM does not predict mononucleotide repeats (repeats with unit length equal to one), we downloaded from the UCSC server annotated repeats predicted by the Tandem Repeats Finder (TRF) tool (25), and we selected all the mononucleotide repeats from this dataset. We included the mononucleotide repeats in order to obtain the most comprehensive set of repeats irrespective of the NGS platform and associated software. However, the use of 454 GS-FLX for sequencing, which allowed us to obtain long reads, did not allow us to perform reliable genotyping of mononucleotide repeats in the downstream analysis. We also downloaded the track of microsatellites, which are mostly repeats with unit length two and three with high variability (25). Finally, we considered repeats known to be variable from three additional sources: repeats used for genotyping and forensic applications (see: http://www.cstl.nist.gov/biotech/strbase/), repeats related to diseases (8) and repeats studied by Matsumoto et al. (26). We classified all these repeats based on their location relative to the catalog of annotated genes in the following categories: coding, intron, 5′ UTR, 3′ UTR, upstream (1 kb before the 5′ UTR) and downstream (300 bp after the 3′ UTR). We also created separate categories for genes spanning annotated microRNAs, regulatory elements (transcription factor binding sites) annotated in the ORegAnno database (27) and CpG Islands annotated also in the UCSC database (28). Repeats that did not span any of these functional elements were classified as intergenic.

To select the set of targeted repeats for sequencing, we first set the maximal repeat length (total length of a repeat region, as observed in the reference genome) arbitrarily to 250 bp, which retained 96% of the repeats gathered (792 394) as explained above. To predict genotype variability, we calculated a 'variability score' for each repeat using the SERV algorithm (17). This algorithm uses a support vector machine to provide for each repeat a score, which is based on unit length, total repeat length and purity. The phenotypic variability is predicted as follows: we classified the repeats in two groups, one having only repeats in deep intronic sequences or intergenic regions (called the RI group), and the other having repeats located within functional regions (termed the RF group), including coding and potential regulatory regions (5′ UTR, 3′ UTR, upstream, downstream, microRNA, transcription factor binding sites and CpG Islands). It is generally assumed that repeats in the RF group are more likely to produce phenotypic variation when compared to those in the RI group.

From the RF group (33 807), we selected repeats for which we could design at least two unique probes, and which belong to at least one of the following subgroups: (i) mononucleotide repeats (211); (ii) repeats with SERV score equal or greater than 1 (2299); (iii) repeats in transcription factor binding sites (1090); (iv) repeats in coding regions with ETANDEM score equal or larger than 21 (3889) and (v) repeats with ETANDEM score larger than 45 (2093). The scores for the latter two RF subgroups were arbitrarily selected to yield a total number of selected repeats that would be feasible to sequence, when added up with those selected for the RI group. The final number for the RF group (as a union) here was 7724.

Because the total size of the RI group is much larger than the size of the group RF and repeats in this group are less likely to produce phenotypic variation, we applied more stringent rules to select repeats from the RI group. First, we selected repeats for which we can design three or more probes, with SERV score equal to or larger than 1, and with ETANDEM score larger than 100 (75 for mononucleotides). We selected a total of 673 repeats from RI following these conditions. Second, to compare variability between the RF and the RI group for different kind of repeats, we built a histogram with the selected repeats in the RF group based on their unit length, copy number and GC content. Then, from all repeats in RI for which we can design at least three probes, we selected 2338 random repeats following the probability distribution defined by the histogram and added them to the list of 673 RI repeats (details can be provided upon request).

We completed the set of selected tandem repeats in the RF (7724) and RI (3011) groups by adding the repeats from the forensic, disease related datasets and the repeats analyzed by Matsumoto and colleagues (26) for which we could design at least one probe: 23, 16 and 6 repeats, respectively.

Probes design for selected tandem repeats

For each candidate probe, we used the following two strategies to predict specific hybridization: first, we called a probe not unique if a stretch of more than 25 consecutive bases, not including the repeat, spans a region that is masked in the reference genome by RepeatMasker. Second, for each probe passing the previous filter, we simulated three consecutive fragments corresponding to the first, middle and

last 40 bp of the probe and we mapped those to the reference genome using Bowtie (29). We used a base quality score of 20 for each base of each simulated read and asked Bowtie to retain alignments with a sum of mismatch qualities less than or equal to 200, which retains alignments with up to 10 mismatches. We also set a seed length of 15 bp and allowed each alignment to have up to three mismatches in the seed. We finally set the 'try hard' option to find as many good alignments as possible. Given the Bowtie alignments, we calculated the number of mismatches of the second best hit for each fragment of each probe. We finally called a probe unique if the number of mismatches of each fragment with its second best alignment is greater than 3 and the total sum of mismatches over the whole probe is greater than 19. This procedure ensures that there will be at least 20 well-distributed mismatches with the second best hit of each probe that we call unique. For spanning probes, we also asked that the unit length of the corresponding repeat is at least 3, and for the repeats with unit length equal to 3, the total repeat length is at most 80 to avoid hybridization to undesired di- or trinucleotides with the same motif of a targeted repeat.

Capture and sequencing of tandem repeats

Library preparation was done following the manufacturer's instructions by the Genomics Core of the University Hospitals Leuven (http://gc.uzleuven.be/). Briefly, genomic DNA of each family member was sonicated in a Bioruptor (Diagenode) to yield fragments of 200–1500 bp, which was size-selected on agarose gels to yield DNA bands of 500–800 bp. Selected fragments were blunt-ended, adapters were ligated and then purified again. SureSelect (Agilent) sequence capture was performed by the Genomics Core according to the manufacturer's protocols. Four samples were loaded in a four-lane gasket of the GS FLX+ instrument (Roche) and run at one lane per sample with the GS-FLX Titanium XL+ kit (Roche), which produces single-end reads up to 600 bp.

Bioinformatic analysis of sequence data

We aligned reads to the reference genome using bwa-sw with default parameters (30). We implemented a custom script to perform local alignment and identify the 45-bp long flanking regions in the reads mapping next to targeted repeats, and extract the segment corresponding to the repeat region. We also implemented a script that gathers all reads genotyping the same repeat in different samples, performs multiple sequence alignment, and corrects for homopolymer errors. After corrections, the script again splits the reads over the different samples and performs independent genotype calling by choosing the two alleles with the highest read counts. We set a minimum threshold of two reads for the second most frequent allele to call a genotype heterozygous. Software information for this type of analysis is available as Supplementary Methods S1. For each family member, we thus generated five data points for each repeat: (i) the copy number of the most frequent allele (C1), (ii) the number of reads of that allele (R1), (iii) the copy number of the second (most frequent) allele (C2), (iv) the number of reads of that second allele (R2) and (v) the total number of reads (RT) for that repeat. In most cases, RT = R1 + R2, although RT could be >R1 + R2 because of the presence of single reads, which are not called but included in the RT count, or because of the presence of alleles of an apparent third allele with a lower read number than R2, or slippage associated with the sequencing protocol. Polymorphism in coding repeats was analyzed with the PROVEAN Protein tool (31). PROVEAN calculates a delta alignment score based on the reference and variant versions of a protein query sequence with respect to sequence homologs collected from the non-redundant (NR) protein database at the National Center for Biotechnology Information (NCBI) through Basic Local Alignment Search Tool. If the PROVEAN score is equal to or below a predefined threshold, the protein variant is predicted to have a 'deleterious' effect, otherwise it is predicted as 'neutral'.

Validation of sequence data

The genotypes of the family members obtained by 454 sequencing were validated by the method of fragment analysis. For that, two rounds of PCR were performed followed by capillary electrophoresis. The first PCR consisted of 15 cycles performed on 50 ng genomic DNA in a 25 μl mixture using GoTaq Flexi DNA Polymerase (Promega) and 0.2 μM unlabeled specific primers designed in CLC Main Workbench (CLC bio, Denmark). A 21 bp extension of the M13 primer sequence was added to the 5′-end of each forward primer. All primer sequences can be obtained upon request. The second PCR consisted of 20 cycles and was performed on 2 μl of the first PCR product using a FAM-labeled M13 primer in combination with each respective reverse primer. Final products were mixed with the GeneScan 500 ROX Size Standard (Lifetechnologies) and subjected to capillary electrophoresis on an ABI3500xL Genetic Analyzer (Lifetechnologies). Fragment lengths were determined with the GeneMapper v4.1 software (Lifetechnologies). Conclusions on allele calls were made only by comparing the three samples of a trio in the same run. To verify potential *de novo* mutations, we performed Sanger sequencing as follows: 35 cycles were performed with the respective unlabeled primers and the BigDye v3.1 cycle sequencing kit, analyzed on the ABI3500xL instrument. In case of heterozygosity, however, we first ligated the purified PCR product in the pGEM-T Easy vector (Promega) and sequenced clones with the different alleles. Sequences were analyzed and aligned in the BioEdit v. 7.1 software (Ibis Biosciences).

RESULTS

A novel strategy for targeted sequencing of tandem repeats

We developed a novel strategy to sequence thousands of repeats in the human genome, which can be summarized in three main steps. First, we aimed to capture specific repeats from sheared total genomic DNA. These fragments are subsequently sequenced using the 454 GS-FLX-Plus Titanium system, a technology that yields the long read lengths (>500 bp) required to span complete tandem repeats plus some of the flanking regions. In a final step, the reads are mapped onto the reference genome to identify the specific repeat,

and subsequently the reads are filtered and analyzed to yield an accurate estimate of the total repeat length.

We started gathering 792 394 predicted and validated tandem repeats in the current reference genome (hg19) from a wide variety of sources: ETANDEM predictions, UCSC tracks, databases for genotyping and forensic applications, and previous reports describing disease-related or highly variable repeats (see Materials and Methods for details). For each tandem repeat we predicted its most likely genetic annotation based on the UCSC genome browser (24). Table 1 summarizes the number of repeats gathered from the different sources and classified in the different genetic annotation classes. We found tandem repeats of lengths <250 bp in the coding region of 4180 genes and in the promoter of 5859 additional genes, which represents, respectively, 9% and 12.61% of the catalog of 46 454 protein coding genes available in the UCSC database. Supplementary Figure S1 shows the distribution of repeats for different unit lengths. In contrast with other categories, repeats in coding regions often have units that contain multiples of three nucleotides. This finding is expected given the variability of repeats and selection against frameshift mutations. We estimated the degree of variability for each repeat using the SERV model (17). Predictions performed with this model indicate that repeats in coding regions are on average less variable than repeats in other regions, even though as many as 873 repeats located within coding regions are predicted to be extremely variable (VARScore > 1) and more than 3000 coding repeats are predicted to be highly variable (VARScore between 0 and 1) (Supplementary Figure S2).

Because we can genotype only a few thousand repeats with the reads produced by a single run of the GS FLX+ system, we designed a set of rules to select tandem repeats for which (i) we can design at least two different unique probes, (ii) a total repeat length can be covered by a GS FLX+ read including the flanking sequences, (iii) there is a high probability of being variable among different individuals and (iv) there is a potential to induce phenotypic variability (see Materials and Methods for details). We selected 7728 repeats in presumed functional domains of which 3891 are in coding genes, 245 are in noncoding RNA and 1836 are in promoter or terminator regions. The remaining 1756 repeats are in intergenic regions with a predicted regulatory function according to UCSC annotations. Together, this set is referred to as 'repeats in functional regions (RF group)'. A second set of 3018 repeats located in gene deserts and deep intronic sequences (RI group) and with the same distribution as the presumed functional repeats was also included, yielding a total of 10 746 repeats targeted for sequencing. This total set covers a wide range of unit lengths, copy numbers, GC content and functional characteristics (see Supplementary Table S1 for detailed information of each repeat).

Because tandem repeat sequences are by nature not unique, they cannot be specifically captured using traditional approaches. Hence, we designed a novel sequence strategy to capture the (unique) sequences that flank these target repeats and purify genomic DNA fragments that contain one of the targeted repeats. For the enrichment of the selected tandem repeats, we used custom-designed 120 nt RNA capture probes. For each repeat in our initial database, we defined three classes of probes: (i) flanking probes ('left' and 'right'), which bind the region immediately flanking the tandem repeat including 20 nt inside the repeat; (ii) 'special' probes, carrying 60 nt from both flanking sides of the repeat and (iii) 'spanning' probes, which include the repeat itself as well as part of the flanking regions (Figure 2). To maximize the success rate of our capture strategy, we ensured that all designed probes will uniquely hybridize in the genome (see Materials and Methods for details). To balance the number of probes per repeat, we included unique flanking probes twice if only two different probes could be designed, or if no spanning probe could be designed. All spanning probes were also added two times. Finally, because it is well known that probes with a too low (<40%) or too high (>70%) GC content are less effective for DNA capture, we also included every such probe twice. As a result, a minimum of four and a maximum of seven probes (of two to four different types) were present for each repeat, resulting in a total of 54 752 probes for the 10 746 selected tandem repeats.

To assess the effectiveness of the different types of probes that we designed, we selected 257 random repeats from the dataset of repeats in RF for which only one type of probe could be defined (114 with a left probe, 52 with a right probe, 48 with a spanning probe and 43 with a special probe). Each probe was included four times. We also selected 172 random repeats from the dataset of repeats in RF for which both flanking probes could be designed and we included each flanking probe twice.

Genotyping accuracy and efficiency

Sequence capture was done with the SureSelect kit from Agilent. For NGS, the GS FLX+ system is expected to yield about 700 million sequenced nucleotides per single run with an accuracy of 99.997%. We simultaneously sequenced four samples per run using the four-lane gasket, so we did not need to tag the samples. In the first run, we sequenced the samples GF, GM, Mo and Fa. In the second run, the samples were Mo, Ch1, Ch2 and Ch3. The sample of the mother was again included in the second run because the number of reads was lower compared to the other three samples. For each sample we obtained on average 200 000 reads (161 340–228 591) with an average read length of 405 bp (Table 2). The distribution of read lengths is given in Supplementary Figure S3. Over 95% of the reads aligned to a unique location in the reference genome. About 60% of the reads aligned close to or in a targeted repeat region. For about 40% of these reads (25% of the total) we could reliably identify the two 45 bp long sequences flanking the targeted repeat and genotype the repeat region using the sequence between the flanks. We thus obtained 338 046 reads useful for genotyping of our selected tandem repeats, with an average read length of 501 bp (Table 2), which represents an average of 48 292 reads per sample (40 929–63 279). Sequencing reads are available in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRP033260.

We obtained at least one useful read that allowed to make allele calls for 6161–6632 tandem repeats (57.3–61.7%) per sample, >4 reads for 3233–3957 repeats (30.1–36.8%) and >10 reads for 1171–2285 tandem repeats (10.9–21.3%). Of

Table 1. Number of repeats gathered from different sources and classified using UCSC annotations

Categories	ETANDEM	TRF Unit Length = 1	UCSC Microsatellites	Forensic	Matsumoto	Disease	Combined Datasets
Intergenic	414 188	34 543	22 218	31	77	0	438 833
Intron	294 430	33 593	17 830	17	59	4	319 754
TFBS	1495	60	89	0	0	0	1540
CpG Islands	8939	23	57	0	1	0	8945
Coding	6681	4	5	0	0	10	6682
MicroRNA	2593	82	96	0	1	1	2649
5′ UTR	3627	48	43	0	0	2	3651
3′ UTR	4341	326	257	0	2	1	4572
Upstream	4312	427	250	1	1	0	4646
Downstream	981	169	74	0	0	0	1122
Total	741 587	69 275	40 919	49	141	18	792 394

Upstream repeats are located at most 1 kb before the 5′ UTR of a gene. Downstream repeats are located at most 300 bp after the 3′ UTR of a gene.

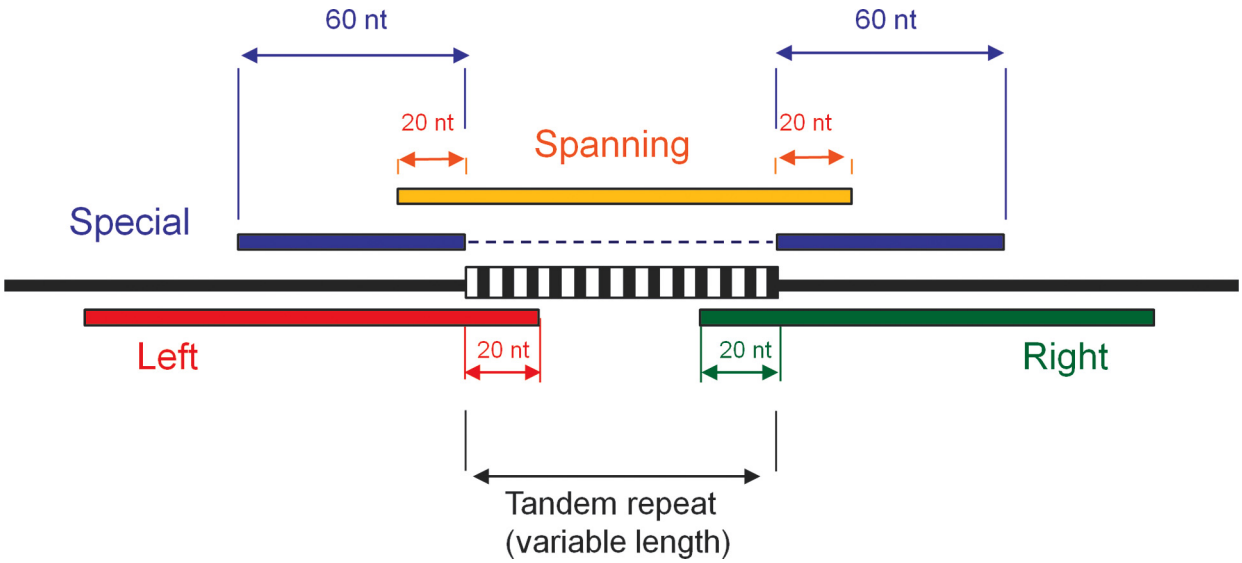

Figure 2. Schematic of probes design for an example repeat of total length 80 bp (striped box). The length of all probes is 120 nt. The flanking probes on the left and right (red and green bars, respectively) are composed of 20 nt inside the repeat and 100 nt of unique flanking sequence. The special probe (split blue bar) mixes together the 60 nt of unique sequences flanking the repeat at the left and right. The spanning probe (orange bar) covers the whole repeat plus a few unique bases in the flanks. nt: nucleotides.

those with at least one read, a mean coverage which we calculated as an average number of useful reads obtained per locus was between 6.3 (for Ch1) and 9.8 (for GM). Analysis of the percentage of targeted repeats that were sequenced demonstrated only a modest effect with increasing of repeat lengths (Supplementary Figure S4). The likelihood that longer repeats are fully sequenced is smaller than those of shorter ones but still is about 50% for repeat lengths between 200 and 250 bp.

The distribution of the 'number of repetitive loci' for different coverage levels (Supplementary Figure S5) indicates that the reads are not evenly distributed among the targeted repeat regions and that we did not obtain any reads for about 2500 repeats. Because this behavior is consistent among the different samples (Supplementary Figure S6), we investigated the main factors that determine the coverage of a targeted repeat. First, despite the doubling of probes with extreme GC content, repeats with GC content <20% or >60% (Figure 3A) and repeats with total lengths >200 (Figure 3B) in general show lower coverage (*P*-value of a Wilcoxon rank test $<10^{-16}$). Further analysis revealed that the capture efficiency increases with the number of different probe types (Figure 3C). It is of interest to note that the inclusion of 'special' probes seems to greatly increase the capturing efficiency. However, the coverage obtained for the 429 control repeats with only one type of probe present (in four-fold) was not different between the various probe types (data not shown). Finally, the coverage was also correlated with the functional classification, with repeats in CpG islands and 5′ UTRs showing less coverage than repeats in other regions (Figure 3D). It is important to note, however, that we cannot assess whether these biases are due to differences in capture efficiency, or in sequencing.

For each tandem repeat, we calculated the copy number in both alleles of each individual (eg. CAGCAGCAG are three copies of a CAG repeat) (see Materials and Methods for details). Because the seven family members make four different trios, we used Mendelian inheritance as an initial cross-validation of the genotype calls. For each trio,

Table 2. Mapping statistics of the reads of all seven family members of the family

	Grandmother (GM)		Grandfather (GF)		Mother (Mo)		Father (Fa)		Child 1 (Ch1)		Child 2 (Ch2)		Child 3 (Ch3)		Total	
	# reads	%	# reads	%	# reads	%	# reads	%	# reads	%	# reads	%	# reads	%	# reads	%
Total reads	228 591	100	207 982	100	227 761	100	161 340	100	196 990	100	194 937	100	182 939	100	1 400 540	100
Reads useful for repeats	63 279	27.68	54 843	26.37	49 805	21.87	40 929	25.37	41 198	20.91	43 919	22.53	44 073	24.09	338 046	24.14
Reads in the boundary of repeats	84 977	37.17	75 345	36.23	85 236	37.42	57 725	35.78	76 089	38.63	75 879	38.92	71 051	38.84	526 302	37.58
Reads outside repeats	70 808	30.98	68 119	32.75	82 670	36.30	53 944	33.43	72 854	36.98	68 719	35.25	61 868	33.82	478 982	34.20
Reads mapped uniquely	219 064	95.83	198 307	95.35	217 711	95.59	152 598	94.58	190 141	96.52	188 517	96.71	176 992	96.75	1 343 330	95.92
Reads with multiple mappings	7377	3.23	7771	3.74	7873	3.46	6764	4.19	5737	2.91	5370	2.75	4911	2.68	45 803	3.27
Unmapped reads	2137	0.93	1898	0.91	2165	0.95	1972	1.22	1109	0.56	1047	0.54	1031	0.56	11 359	0.81
Read length total (mean)	429		418		385		411		387		398		407		mean	405
Read length useful (mean)	549		539		497		544		448		460		472		mean	501

Reads useful for repeats: reads in which the tandem repeat as well as 45 bp at the left and the right are present.
Reads in the boundary of repeats: reads mapping at the tandem repeat region (+/− 100 bp) not including reads spanning the repeat +/− 45 bp.
Reads outside repeats: reads that do not map within +/− 100 bp of the repeat.

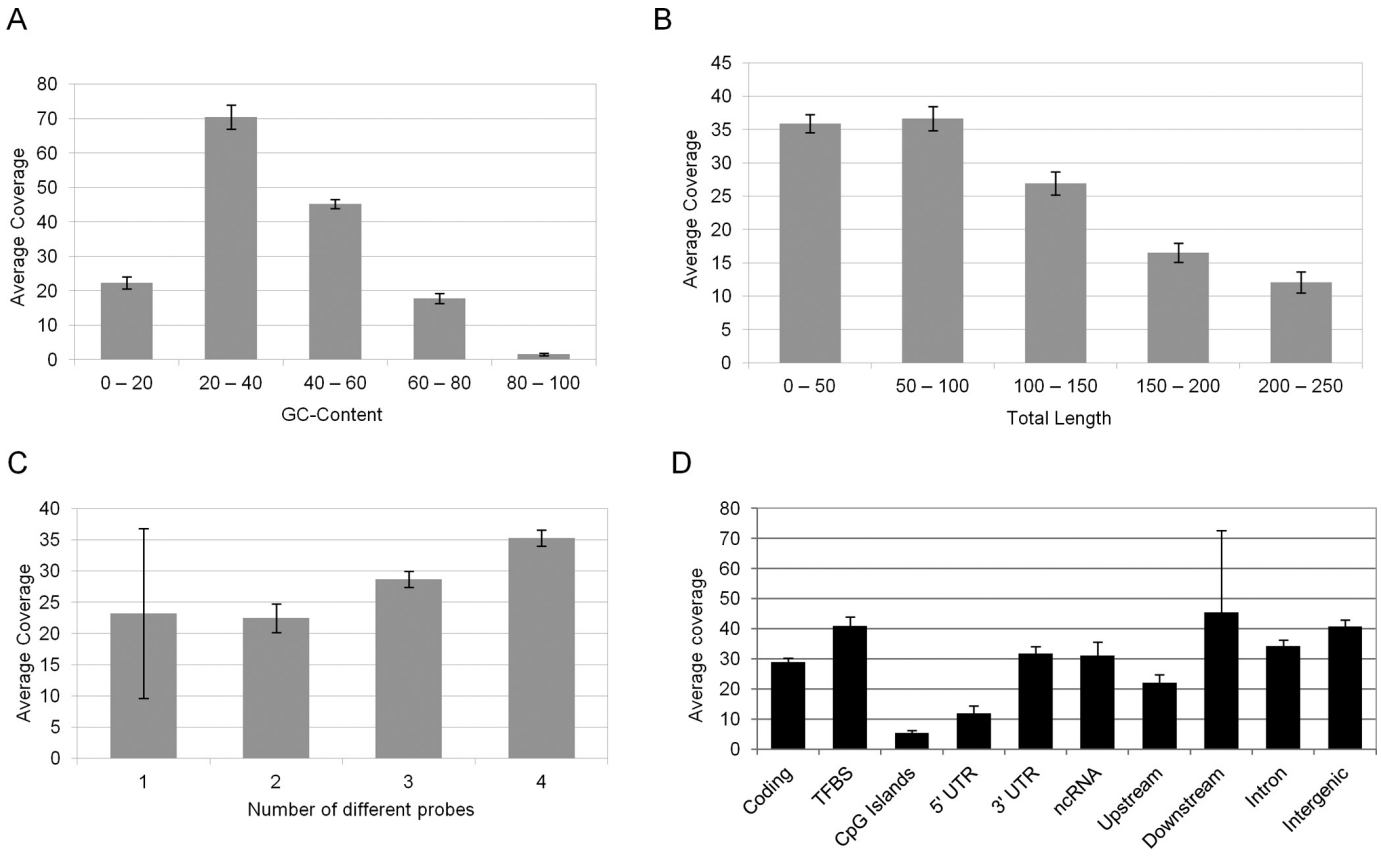

Figure 3. Coverage for repeats with (A) different GC content, (B) different total repeat lengths, (C) different numbers of distinct probes, and (D) different functional roles. Data are provided as mean ± SE for the coverage between different repeats within each discriminating feature represented in the different panels.

we could test Mendelian inheritance on about 5200 repeats and observed Mendelian consistency in about 88% of those.

Validation of sequence data and sequencing output filtering

To verify the apparent Mendelian inconsistencies, we performed validation of 66 repeats by fragment analysis. In the same way, we analyzed an additional 42 repeats for which no Mendelian inconsistencies were detected in any of the trios. Together, this also allowed us to assess the quality of our genotyping algorithm. The fragment lengths for each repeat were compared with the copy number obtained by GS-FLX sequencing. We found that most of the genotyping errors in our sequencing strategy were caused by 'PCR stutters' produced at the PCR amplification step during the library preparation. PCR stutters due to slippage are expected to be most frequent for those repeats having a short repetitive unit and a high copy number (32). Moreover, PCR stuttering usually results in a PCR product that is one or two copies shorter or longer than the actual repeat, and the PCR yield of the stutter product is typically much lower than that of the PCR product with the correct length (33,34). Bearing this in mind, we corrected for those apparent heterozygous genotype calls in which the copy numbers differed by one or two copies, e.g. 11 and 10 copies, and for which the number of reads was significantly different between both alleles, e.g. 15 and 2 reads, respectively. In this example, the 10-copy allele is likely a stutter product of a homozygous genotype of 11 copies. As a consequence, the 10-copy allele should be corrected to an allele with 11 copies. We confirmed such stutter artifacts for 78 out of 108 (72%) apparently heterozygous genotypes for which two alleles had copy number differences of one or two (see Supplementary Table S2). We also noticed that in most cases of these erroneously called 'stutter alleles' there was at least a two-fold difference in the read numbers of the true allele and its stutter, with the stutter always having less reads. Based on these results we deduced a 'stutter correction rule' for each repeat with a difference of one or two copies between alleles; both alleles are considered as true alleles only if the percentage of the lowest read number relative to the highest one, $(R2/R1)\times 100\%$, is $\geq 50\%$, which we called the read number imbalance (I). For the given example $I = (2/15)\times 100\% = 13.33\%$, which thus is below the 50% threshold, meaning that the genotype should be corrected for a stutter and thus becomes a homozygous genotype with 11 copies. Following this rule we *in silico* corrected 64 out of 108 tested genotypes (59.3%) with $I < 50\%$, and from these, 55 (86%) were confirmed and only 9 (14%) were truly heterozygous (Supplementary Table S2), demonstrating that stutters could reliably be detected based on the read number imbalance rule (I). However, from the remaining 44 with $I \geq 50\%$ that were not corrected, 23 were also confirmed as stutter cases, which thus are false negatives (21%) that escape our correction rule. Setting the threshold

higher reduced the number of false negatives but also increased the number of false positives. In order to maximize the number of heterozygous calls, we opted to keep the $I \geq 50\%$.

The main reason why we could not correct some stutter products is that the total coverage for some repeats was simply too low to be able to differentiate true alleles from stutter products by read depth imbalance. Moreover, the low number of reads can also prevent the detection of both alleles in a true heterozygous position. We thus implemented an additional filter step stating that the total read number (RT) should be above a certain threshold, which we arbitrarily set to RT > 4 in all seven individuals.

We also identified a few genotype calls for which the total read number was higher than the sum of the reads of two alleles (RT>R1+R2). For example, a heterozygous repeat being called with C1 = 49 and C2 = 59, and read numbers R1 = 5 and R2 = 2 (RT = 7) was found by fragment analysis to consist of alleles with 49 and 60 copies. Analysis of the aligned reads revealed that a wrongly heterozygous genotype was obtained because the longest allele (60 copies) had only one read while its stutter (59 copies) had two (Supplementary Figure S7). Out of 184 genotypes (368 alleles) that we validated by fragment analysis, we detected 24 cases (6.5%) where one of the true alleles was missing, being replaced by a stutter in the final genotype call. Following a conservative approach, we filtered out every genotype call for which RT≠R1+R2, because this data inconsistency is most likely caused by stutter products or misaligned reads.

After applying the stutter correction and read number (RT) filtering steps to our initial sequencing data, we corrected 23 804 genotypes in the seven samples (3000–3600 genotypes per sample) and filtered out 9092 repeats. Then, from the 1654 repeats passing all filters as explained above (for each family member), we selected 10 random repeats to perform validation of the genotype calls by fragment analysis. We did not find any erroneous calls in any of the seven family members (70 genotypes) (Supplementary Table S3). However, all 10 randomly selected repeats happened to be monomorphic (the same homozygous copy number in all seven family members). To also test our filtering criteria on the polymorphic repeats, we randomly selected 10 additional repeats from the polymorphic dataset and we validated those in all family members. Fragment analysis revealed an error rate of 7.9% (11/140 alleles). Taken together, despite the difficulties and technical limitations, the overall success rate in allele calling on our final dataset after filtering was estimated at 99.4% (Supplementary Table S3).

Patterns of variability in tandem repeats

In order to estimate the variability of tandem repeats in the human genome, we selected only those repeats that passed our filtering criteria and were genotyped in every individual of the family (1654 repeats). For each repeat, we measured its 'variability' in three different ways: first, we calculated the standard deviation of the alleles (i.e. their copy numbers) observed across the seven individuals (Figure 4); second, we counted the number of heterozygous individuals (Figure 5); finally, we counted the number of repeats with at least one allele that is different from the reference sequence observed over the seven individuals. We found variants different from the reference genome in 125 repeats (7.6%), hence regarded as polymorphic repeats. This variation was only 3.85% for minisatellites, but 9.95% for microsatellites showing that the variability of microsatellites is higher, as expected. Variation revealed positive correlation with the total length of tandem repeats, with their copy number and sequence purity (Supplementary Figure S8). We also found that based on the level of heterozygosity, as expected, repeats in coding regions are significantly less variable ($P < 0.01$) than repeats in introns, intergenic regions, 3′ UTRs, and noncoding RNAs (Figures 4 and 5).

From the 640 repeats in coding sequences genotyped in the seven individuals, 26 (4.06%) were polymorphic and thus were predicted to result in changes of a polypeptide length. For 17 repeats we examined the source of variation and its effect on the amino acid sequence of a corresponding protein (Supplementary Table S4). As one repeat is a pentanucleotide, deletion of one repetitive unit causes a frameshift in *GPR126*, which generates a stop codon, at least in some transcripts, soon after this variation. The unit lengths of the remaining 16 repeats were multiples of 3 bp thus causing insertion or deletion of one or more amino acids, mostly within a poly amino acid tract. For four of the latter variations, the PROVEAN protein software predicted a deleterious effect on the function of the proteins TAF7L, HEG1, ODFP1 and HYDIN (Supplementary Table S4). Analysis of these 17 variations by fragment analysis confirmed all except one, which validated our stringent filtering criteria. As a result, one frameshift and three potential harmful in-frame variants were detected in this family.

Interestingly, repeats in 5′ UTRs seem highly conserved (Figure 4), which, however, could be due to the fact that we only obtain full reads for 11 repeats in all seven family members.

We also performed a correlation analysis of the variability predicted by the SERV score and the observed variability in our data but this correlation was only 0.13. However, we verified that the average SERV score for repeats with at least one heterozygous individual was significantly higher than that for conserved (monomorphic) repeats (Wilcoxon rank test, $P = 2.925 \times 10^{-8}$), thereby confirming the accuracy of the SERV algorithm to predict the variability of a tandem repeat (Supplementary Figure S9).

Our selection of repeats for sequencing was biased in two ways, because we specifically targeted potentially polymorphic repeats as well as repeats located within functional regions. To estimate how this bias affects the observed variation rate, we analyzed the variability of four groups of repeats that represent combinations of the above mentioned two factors: (i) repeats predicted as non-polymorphic and located within regions without a clear biological function (gene deserts); (ii) repeats that are predicted to be polymorphic and that are located in gene deserts; (iii) repeats that are predicted to be non-polymorphic and are located within functional regions and (iv) repeats that are predicted to be polymorphic and are located within functional regions. When we extrapolate the variability for each of the four repeat categories in our dataset to the whole genome (taking into account the proportion of each repeat class in

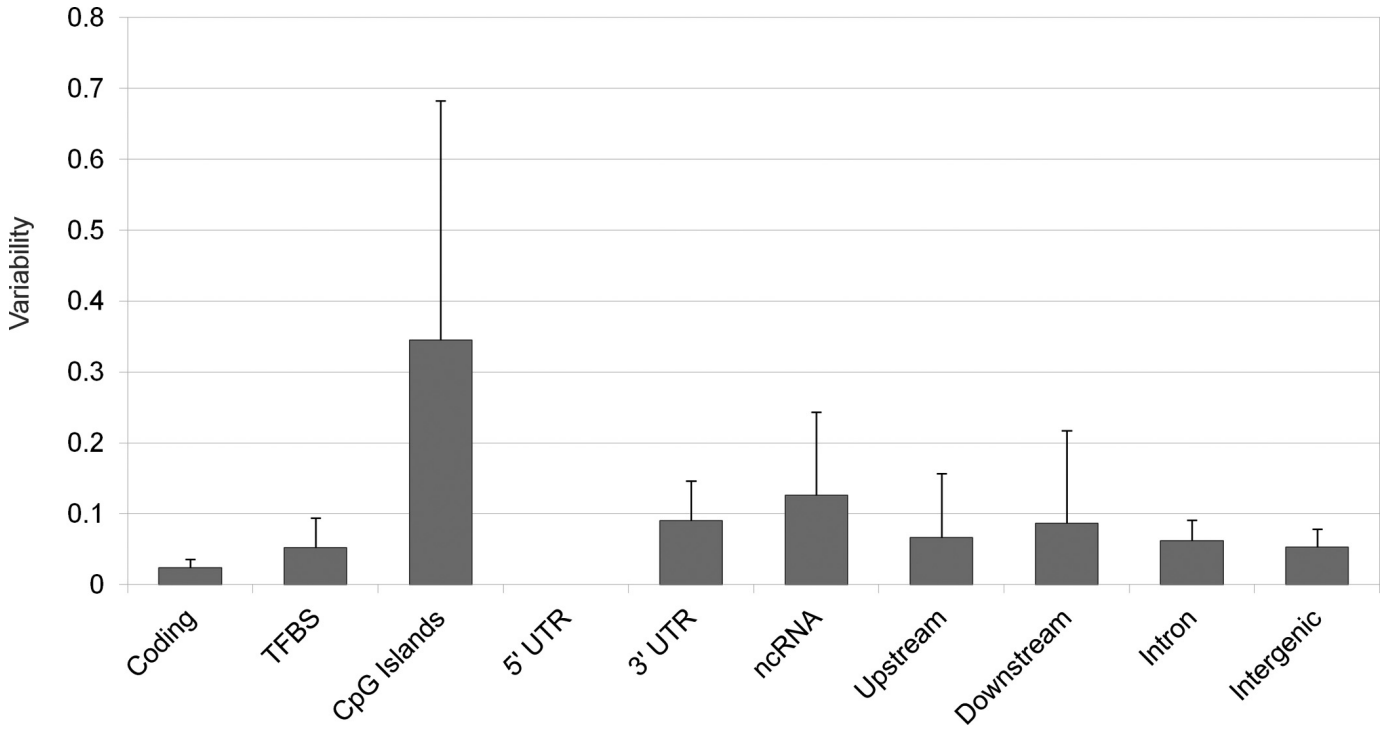

Figure 4. Variability, measured as standard deviation of called alleles, for tandem repeats plotted according to their different functional roles. The number of repeats obtained for each class are Coding, 640; TFBS, 230; CpG Islands, 3; 5′ UTR, 11; 3′ UTR, 96; ncRNA, 32; Upstream, 25; Downstream, 16; Intron, 1566; and Intergenic, 320. Data are provided as mean ± SE.

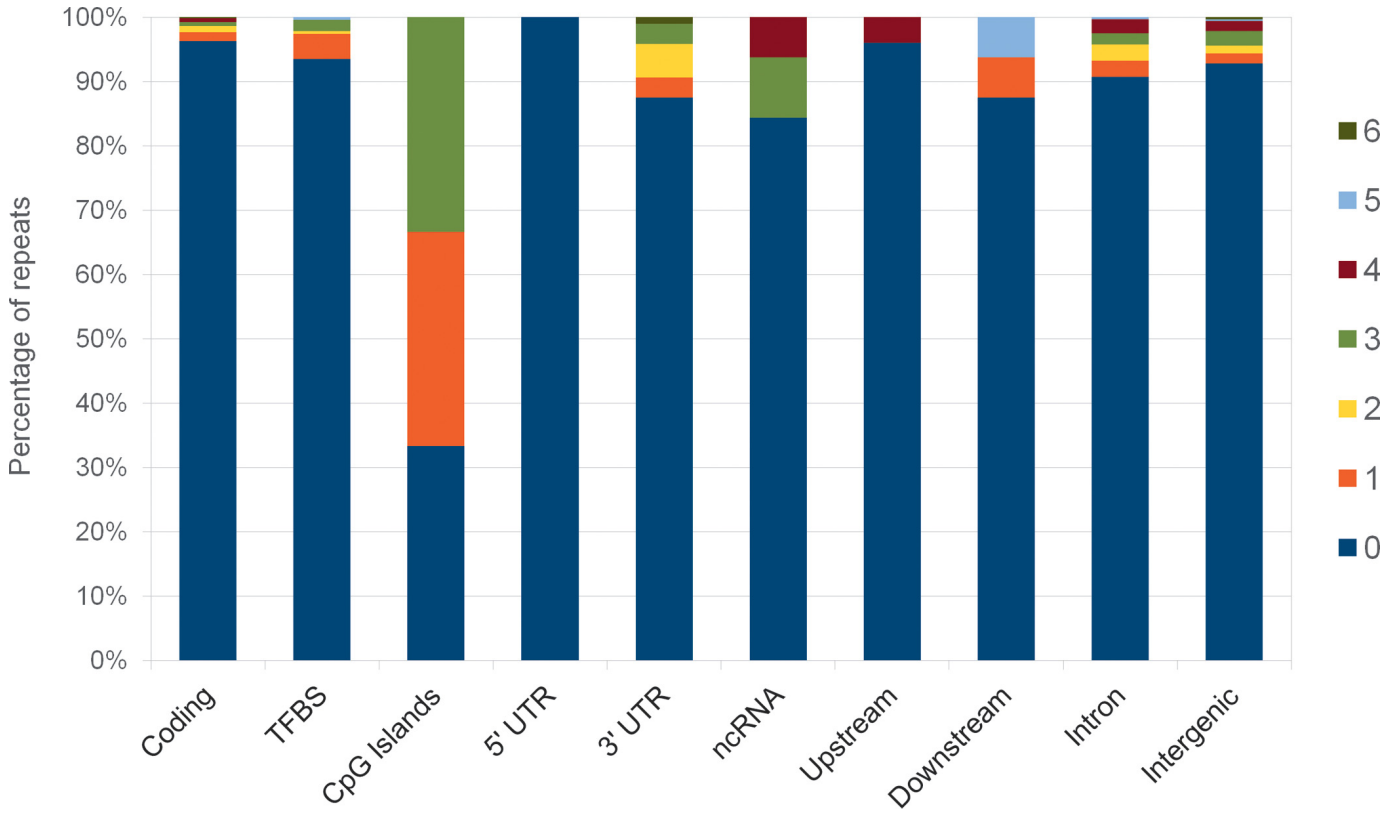

Figure 5. Distribution of heterozygous repeats in none to all seven individuals in relation to the location of tandem repeats and their functional roles. 0: percentage of repeats that are heterozygous in none of the individuals; 1: percentage of repeats that are heterozygous in 1 individual; etc.

the full genome), we estimate about 9.3% of all repeats to be polymorphic in this family (Supplementary Table S5).

De novo tandem repeat variation

Based on the final filtered dataset of 1654 tandem repeats we looked for Mendelian inconsistencies within the four trios we analyzed. After applying our set filtering criteria ($I \geq 50\%$ and ≥ 5 reads in all individuals) we were left with only 55 repeats that showed an apparent Mendelian inconsistency in at least one trio. Fifteen of those repeats were tested by fragment analysis, but none of these proved to be a true *de novo* mutation. From the 15 candidate repeats tested, in seven cases a stutter was causing an erroneous heterozygous call even with $I \geq 50\%$. The remaining eight true heterozygous calls, with $I < 50\%$, were erroneously corrected into homozygotes marking them as false positives after the correction step. Based on these validation data, the remaining 40 repeats with apparent inconsistencies were manually examined, and revealed that they fit one of both types of errors described above.

Though no *de novo* events were detected in the final corrected and filtered dataset, one *de novo* mutation was identified and confirmed by fragment analysis before these correction steps. The AAGA tetranucleotide repeat at Chr1:84 267 437–84 267 512 showed an inconsistency of inheritance in child1 (Ch1). The called alleles (C1 and C2) in the mother were 17 and 20 while Ch1 carried alleles with copy numbers 18 and 19. The 18-copy allele was inherited from his father indicating that the 19-copy allele should have been derived from a maternal allele i.e. by contraction of the 20-copy allele with one copy or by expansion of the 17-copy allele with two copies (Figure 6A). Validation of this repeat by fragment analysis confirmed the allelic inconsistency in the mother and her son. To prove that the fragment length difference was indeed caused by a copy number change, we subcloned the alleles from the mother and child1 and confirmed the *de novo* variation by regular Sanger sequencing (Figure 6B).

DISCUSSION

Like other genomes, the human genome is scattered with tandem repeats. Using a combination of repeat databases and *de novo* searches with a commonly used algorithm for repeat detection, we identified almost 800 000 tandem repeats in the human genome including mononucleotide repeats, microsatellites and minisatellites. More than 6000 of these repeats are located within coding regions, and almost half of these loci in the human genome are located near or within genes where variation in the repeats might have functional consequences.

In this study, we describe a method to simultaneously characterize thousands of selected tandem repeats in humans. More specifically, we first developed a strategy for the selection of tandem repeats that can be captured using unique flanking sequences, then captured about 10 000 selected repeats from the genomes of seven family members followed by massive parallel sequencing of those by 454 GS-FLX+ technology. Finally, we mapped the millions of reads with subsequent determination of the repeat copy numbers using our own software. Lastly, we developed data filtering algorithms based on validation by fragment analysis and Sanger sequencing. Our study of this three-generation family demonstrated that overall, 7.6% of tandem repeats are polymorphic. Importantly, 4.06% of repeats located within coding regions have an allele that differs from the reference sequence. Finally, of the 4447 repeats called in all seven family members, we found one example of a *de novo* variation that illustrates the instability of these regions.

A few recent studies have contributed to a better view of repeat variability in the human genome, but they were significantly limited in the number of targets to study because, due to technical limitations, they could focus only on short repeats. A genome-wide population-scale microsatellite analysis of >500 individuals of the 1000 Genomes Project exome sequencing pilot study, performed on 454 and/or Illumina instruments, was analyzed for 8342 repeats with the majority (94.5%) <20 bp in total length (35). Similarly, whole genome paired-end Illumina sequencing in 200 Drosophila inbred strains (36) or human gastric cancer cell lines and primary tissues (37), as well as the targeted capture MiSeq sequencing of thousands of selected short tandem repeats (38) had to be restricted to short microsatellites. However, sequencing longer reads is especially crucial to study tandem repeat variability because this variability increases with repeat length (39,40). Our success in genotyping a larger number of coding and potential regulatory tandems thus is mainly due to the longer reads that were obtained by the GS-FLX+ Roche technology (up to 700 bp) by which we could increase the actual repeat size up to 250 bp. In previous reports, the GS-FLX system with maximal read lengths of 450 bp had been validated for efficient genotyping of a selection of five microsatellites in 10 individuals (41), or a much larger pool of tandem repeats in several different microbial species (42). In our study, we enlarged the selection of tandem repeats by number (>10 000) and range of characteristics (both micro- and minisatellites, total repeat size up to 250 bp, low and high copy numbers included) for a more comprehensive analysis in seven individuals of the same family.

Our method for selective capture of tandem repeats by flanking, spanning and special probes is relatively efficient. It was demonstrated that spanning probes, which we designed to capture short-length repeats, do not compromise accuracy. In addition, we showed that a novel capture probe design including both flanks of the repeat (called special probes) can also capture a repeat with an efficiency comparable to that of flanking probes. The use of a similar capture method with spanning probes for eight repeat motifs (42), or with flanking probes for 7851 target loci (38) have recently been described. In the latter case, the capture efficiency was 38.7%, while in our study the combined usage of up to four types of probes allowed us to reach a capture efficiency >60%. In the latter reported study, however, only 2.2% of all HiSeq reads completely spanned the repeat, thereby severely limiting the use for large-scale genotyping (38). Our study showed that longer reads yielded a very significant increase in sequencing efficiency: specifically, about 25% of all sequencing reads covered a complete repeat sequence together with 3′ and 5′ flanking regions that allowed mapping the repeat. We also found a bell-shaped relation-

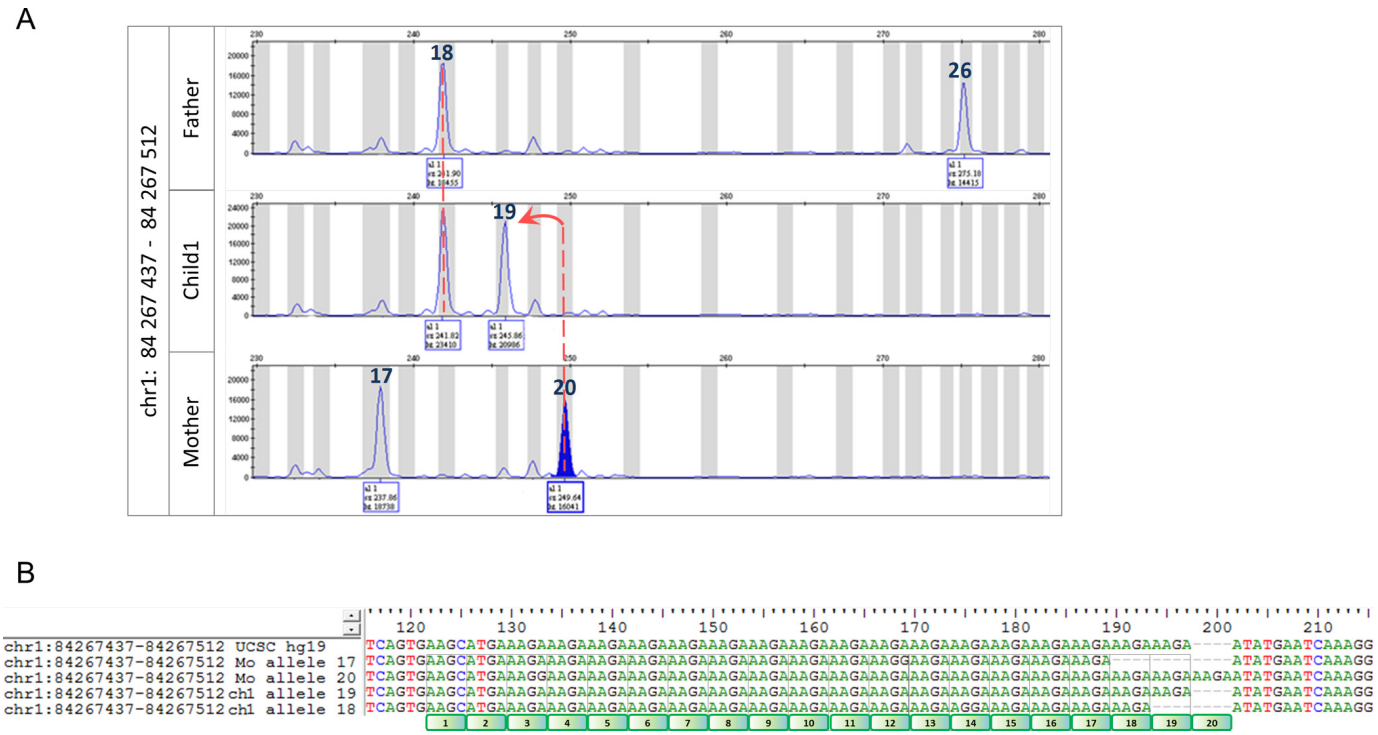

Figure 6. *De novo* generation of a tandem repeat allele in child1. (**A**) Fragment analysis view of the alleles for the repeat chr1:84 267 437–84 267 512 in the father, mother and child1 demonstrating that the 19-copy allele in child1 occurred *de novo*, most likely by a contraction event. (B) Alignment of the Sanger sequenced cloned alleles of the mother (Mo) and child1 (Ch1). Copy numbers of the (imperfect) AAGA repeat are indicated below.

ship between the sequencing coverage and GC content with a preference to moderate GC content as has been reported by others (38). Despite our attempt to solve this issue by doubling the number of probes for the loci with both low and high GC content (<40% and >70%), the data showed that this procedure is not sufficient to capture repeats in GC-rich genomic regions.

The biggest technical challenge encountered with tandem repeat genotyping is PCR stuttering, which can lead to erroneous calls of the copy number for a given tandem repeat. Recently, Highnam and colleagues developed new software to deal with this problem by estimating probabilities for errors to appear in repeats with a specific set of characteristics. Applying these probabilities to the genotype calls of the respective repeats should allow making better decisions whether a variant is more likely to be a stutter artifact or a real allele (23). This predictive approach requires prior knowledge of the likelihood to generate stutters for each type of repeat, which unfortunately was lacking. Therefore, we deduced the filtering rules for stutter correction based on fragment analysis validation data. The only way to distinguish a real allele from a stutter artifact was to compare the read frequency with that of the main allele. The reliability of this proportional approach in general is very good but obviously increases with a higher total number of reads per locus.

After these filtering steps, 92.4% of the sequenced repeats turned out to be monomorphic, leaving 7.6% classified as polymorphic in the studied family (having at least one allele different from the reference sequence). This variation in copy numbers was unevenly distributed between minisatellites (3.85%) and microsatellites (9.95%). The latter rate is higher than the 5.9% estimated for microsatellites by McIver et al. in the CEU population from the 1000 Genome project (35). We anticipate, however, that the true percentage of variability is somewhat lower due to stutters that still escape our correction step. On the other hand, our study only comprised seven individuals from one family consisting of only three unrelated individuals (GF, GM and Fa). Sequencing more unrelated individuals would therefore increase the observed repeat variability. Even though our initial selection of targeted repeats was biased toward those with SERV scores >1.0, i.e. predicted to be more likely variable, the selection was also based on their location in the genome as an indicator for their potential to induce phenotypic variation. Therefore, natural selection against unfavorable variants might also have influenced the observed variability in our set of tandem repeats. The very high occurrence of monomorphic repeats and thus high stability of the copy number could be explained by a (yet unknown) phenotypically important localization of the repeats, especially taking into account that 70% of the 10 746 selected repeats belong to the presumably functional (RF) group.

The genome-wide variation of tandem repeats was estimated by extrapolation of the observed percentage of polymorphisms among repeats with SERV scores ≥ 1 versus SERV scores < 1, and in repeats from functional (RF) versus non-functional (RI) groups. Such approach yields an estimation of the overall portion of polymorphic repeats at about 9%. This variability rate is 7.4 times higher than that observed for SNPs (1.25%) (43), though it was expected to be up to 1000-fold higher (2). A potential explanation

might relate to the unclear definition of tandem repetitive sequences. Thereby, a large number of sequences recognized as repeats in the genome might be still fairly resistant to DNA slippage during replication due to specific features such as repeat imperfectness, long repeat motifs or low copy number. In any case, our findings show a much higher level of variability compared to the previous estimate in tandem repeats of 1% (35). Although larger scale genotyping on unrelated individuals would be needed to confirm our data, the percentage predicted in our study is encouraging to further investigate the variability of tandem repeats and its consequences on different traits.

Interestingly, we also detected one *de novo* mutation in this small family. This sequence change is most likely due to a contraction of a maternal allele with one repetitive unit although we cannot exclude a two-unit expansion of the other maternal allele. However, since a one-unit change is more likely than a gain of two units at once (44,45), the former option is preferred. When focusing on repeats located within the coding part that were sequenced in all seven individuals, we found 17 repeats for which at least one variation was detected when compared to the reference sequence. Four of these variants are potentially deleterious, which demonstrates that our strategy is capable of detecting protein altering variants with possible functional consequences. However, it is clear that subsequent functional studies are required to determine the impact of each of these repeat variants.

Because repeats are mostly ignored in today's comparative genomics and GWAS studies, they could be partly responsible for the so-called "missing heritability", i.e. instances where a phenotype has a clear genetic basis but where no genetic aberration could be found. The capture and sequencing strategy described in this study may provide a stepping stone for routine genome-scale repeat characterization. It is worth mentioning that our method is not restricted to humans but can be applied for a comprehensive analysis of repeats in any species with a reference genome. With the constant improvements in read length and sequencing cost reductions, we expect that this method can be applied for larger sets of repeats and with deeper coverage, increasing its accuracy and cost effectiveness. Emerging long-read NGS technologies such as PacBio, Ion Proton or Nanopore can provide the throughput needed to make this happen within the next year.

In summary, we report on a family-based analysis of 10 000 selected tandem repeats using long sequence reads that yielded the complete repeat sequences in 25% of reads. Using this method, we provide a better estimate of their variability and demonstrate the occurrence of a *de novo* mutation event. This novel method provides major opportunities for genome-wide population-based genotyping for the association of repeat variability with common traits as well as disease.

ACCESSION NUMBER

Sequencing reads are available in the NCBI Sequence Read Archive as accession number SRP033260 (http://www.ncbi.nlm.nih.gov/sra).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT

We would like to thank Jeroen Van Houdt of the Genomics Core (UZ Leuven, Belgium) for help with capturing and sequencing the repeats.

FUNDING

European Research Council (ERC) Starting Grant [241426 to K.J.V.]; Human Frontier Science Program (HFSP) [RGP0050/2013 to K.J.V.]; EMBO YIP program and Agentschap voor Innovatie door Wetenschap en Technologie (IWT) [K.J.V.]; VIB [K.J.V., G.F.]; Fonds voor Wetenschappelijk Onderzoek (FWO)-Vlaanderen [G.0795.11 to K.J.V., J.R.V., G.F.]; University of Leuven (KU Leuven) [GOA/12/015 to J.R.V., G.F.]. Marguérite-Marie Delacroix [GV/B-155 to A.Z., G.F]. Funding for open access charge: FWO-Vlaanderen [G.0795.11].
Conflict of interest statement. None declared.

REFERENCES

1. Jelinek,W.R., Toomey,T.P., Leinwand,L., Duncan,C.H., Biro,P.A., Choudary,P.V., Weissman,S.M., Rubin,C.M., Houck,C.M., Deininger,P.L. *et al.* (1980) Ubiquitous, interspersed repeated sequences in mammalian genomes. *Proc. Natl. Acad. Sci. U.S.A*, **77**, 1398–1402.
2. Verstrepen,K.J., Jansen,A., Lewitter,F. and Fink,G.R. (2005) Intragenic tandem repeats generate functional variability. *Nat. Genet.*, **37**, 986–990.
3. Ohno,S. (1972) So much "junk" DNA in our genome. *Brookhaven. Symp. Biol.*, **23**, 366–370.
4. Gulcher,J. (2012) Microsatellite markers for linkage and association studies. *Cold Spring Harb. Protoc.*, **2012**, 425–432.
5. Kim,T.M., Laird,P.W. and Park,P.J. (2013) The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell*, **155**, 858–868.
6. Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.*, **22**, 253–259.
7. Vinces,M.D., Legendre,M., Caldara,M., Hagihara,M. and Verstrepen,K.J. (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, **324**, 1213–1216.
8. Gemayel,R., Vinces,M.D., Legendre,M. and Verstrepen,K.J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.*, **44**, 445–477.
9. Jansen,A., Gemayel,R. and Verstrepen,K.J. (2012) Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. *Genome Dyn.*, **7**, 108–125.
10. Fondon,J.W. III and Garner,H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. U.S.A*, **101**, 18058–18063.
11. La Spada,A.R. and Taylor,J.P. (2010) Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.*, **11**, 247–258.
12. Gatchel,J.R. and Zoghbi,H.Y. (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.*, **6**, 743–755.
13. Brouwer,J.R., Willemsen,R. and Oostra,B.A. (2009) Microsatellite repeat instability and neurological disease. *Bioessays*, **31**, 71–83.
14. Law,M.J., Lower,K.M., Voon,H.P., Hughes,J.R., Garrick,D., Viprakasit,V., Mitson,M., De Gobbi,M., Marra,M., Morris,A. *et al.* (2010) ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. *Cell*, **143**, 367–378.

15. Biason,P., Visentin,M., Talamini,R., Stopar,A., Giorda,G., Lucia,E., Campagnutta,E. and Toffoli,G. (2012) Polymorphic thymidylate synthase gene impacts on overall survival of patients with epithelial ovarian cancer after platinum-based chemotherapy. *Pharmacogenomics.*, **13**, 1609–1619.
16. Lecomte,T., Ferraz,J.M., Zinzindohoue,F., Loriot,M.A., Tregouet,D.A., Landi,B., Berger,A., Cugnenc,P.H., Jian,R., Beaune,P. *et al.* (2004) Thymidylate synthase gene polymorphism predicts toxicity in colorectal cancer patients receiving 5-fluorouracil-based chemotherapy. *Clin. Cancer Res.*, **10**, 5880–5888.
17. Legendre,M., Pochet,N., Pak,T. and Verstrepen,K.J. (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.*, **17**, 1787–1796.
18. Rockman,M.V. and Wray,G.A. (2002) Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.*, **19**, 1991–2004.
19. Hammock,E.A. and Young,L.J. (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science*, **308**, 1630–1634.
20. Rockman,M.V., Hahn,M.W., Soranzo,N., Loisel,D.A., Goldstein,D.B. and Wray,G.A. (2004) Positive selection on MMP3 regulation has shaped heart disease risk. *Curr. Biol.*, **14**, 1531–1539.
21. Gymrek,M., Golan,D., Rosset,S. and Erlich,Y. (2012) lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.*, **22**, 1154–1162.
22. McIver,L.J., Fondon,J.W. III, Skinner,M.A. and Garner,H.R. (2011) Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics*, **97**, 193–199.
23. Highnam,G., Franck,C., Martin,A., Stephens,C., Puthige,A. and Mittelman,D. (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.*, **41**, e32.
24. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
25. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
26. Matsumoto,C., Shinkai,T., Hori,H., Ohmori,O. and Nakamura,J. (2004) Polymorphisms of dopamine degradation enzyme (COMT and MAO) genes and tardive dyskinesia in patients with schizophrenia. *Psychiatry Res.*, **127**, 1–7.
27. Griffith,O.L., Montgomery,S.B., Bernier,B., Chu,B., Kasaian,K., Aerts,S., Mahony,S., Sleumer,M.C., Bilenky,M., Haeussler,M. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.
28. Meyer,L.R., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M., Sloan,C.A., Rosenbloom,K.R., Roe,G., Rhead,B. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
29. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
30. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.*, **26**, 589–595.
31. Choi,Y., Sims,G.E., Murphy,S., Miller,J.R. and Chan,A.P. (2012) Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE.*, **7**, e46688.
32. Shinde,D., Lai,Y., Sun,F. and Arnheim,N. (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.*, **31**, 974–980.
33. 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.
34. Hauge,X.Y. and Litt,M. (1993) A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Hum. Mol. Genet.*, **2**, 411–415.
35. McIver,L.J., McCormick,J.F., Martin,A., Fondon,J.W. III and Garner,H.R. (2013) Population-scale analysis of human microsatellites reveals novel sources of exonic variation. *Gene*, **516**, 328–334.
36. Fondon,J.W. III, Martin,A., Richards,S., Gibbs,R.A. and Mittelman,D. (2012) Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. *PLoS ONE.*, **7**, e33036.
37. Yoon,K., Lee,S., Han,T.S., Moon,S.Y., Yun,S.M., Kong,S.H., Jho,S., Choe,J., Yu,J., Lee,H.J. *et al.* (2013) Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. *Genome Res.*, **23**, 1109–1117.
38. Guilmatre,A., Highnam,G., Borel,C., Mittelman,D. and Sharp,A.J. (2013) Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Hum. Mutat.*, **34**, 1304–1311.
39. Legendre,M. and Verstrepen,K.J. (2008) Using the SERV applet to detect tandem repeats in DNA sequences and to predict their variability. *CSH. Protoc.*, **2**, pdb.ip50.
40. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
41. Fordyce,S.L., Avila-Arcos,M.C., Rockenbauer,E., Borsting,C., Frank-Hansen,R., Petersen,F.T., Willerslev,E., Hansen,A.J., Morling,N. and Gilbert,M.T. (2011) High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *BioTechniques*, **51**, 127–133.
42. Malausa,T., Gilles,A., Meglecz,E., Blanquart,H., Duthoy,S., Costedoat,C., Dubut,V., Pech,N., Castagnone-Sereno,P.,, Delye,C. *et al.* (2011) High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Mol. Ecol. Resour.*, **11**, 638–644.
43. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1092 human genomes. *Nature*, **491**, 56–65.
44. O'Dushlaine,C.T., Edwards,R.J., Park,S.D. and Shields,D.C. (2005) Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biol.*, **6**, R69.
45. Vogler,A.J., Keys,C., Nemoto,Y., Colman,R.E., Jay,Z. and Keim,P. (2006) Effect of repeat copy number on variable-number tandem repeat mutations in Escherichia coli O157:H7. *J. Bacteriol.*, **188**, 4253–4263.