Published in final edited form as:
Gene. 2013 March 10; 516(2): . doi:10.1016/j.gene.2012.12.068.

Population-scale analysis of human microsatellites reveals novel sources of exonic variation

L. J. McIver[a], **J. F. McCormick**[a], **A. Martin**[a], **J. W. Fondon III**[b], and **H.R. Garner**[a,*]
[a]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

[b]Department of Biology, University of Texas at Arlington, TX, USA

Abstract

Using our microsatellite specific genotyping method, we analyzed tandem repeats, which are known to be highly variable with some recognized as biomarkers causative of disease, in over 500 individuals who were exon sequenced in a 1000 Genomes Project pilot study. We were able to genotype over 97% of the microsatellite loci in the targeted regions. A total of 25,115 variations were observed, including repeat length and single nucleotide polymorphisms, corresponding to an average of 45.6 variations per individual and a density of 1.1 variations per kilobase. Standard variant detection did not report 94.2% of the exonic repeat length variations in part because the alignment techniques are not ideal for repetitive regions. Additionally some standard variation detection tools rely on a database of known variations, making them less likely to call repeat length variations as only a small percent of these loci (~6,000) have been accurately characterized. A subset of the hundreds of non-synonymous variations we identified was experimentally validated, indicating an accuracy of 96.5% for our microsatellite-based genotyping method, with some novel variants identified in genes associated with cancer. We propose that microsatellite-based genotyping be used as a part of large scale sequencing studies to identify novel variants.

Keywords

tandem repeats; 1000 Genomes Project; genomics; cancer

1. Introduction

Microsatellites are tandemly repeated units of 1–6 base pairs in length that comprise approximately 3% of the human genome (Toth et al., 2000; Lander et al., 2001). They are often highly variable with mutation rates dependent on several factors, including the length of the microsatellite and its location in the genome (Fondon et al., 1998; Ellegren, 2004). Microsatellite mutations within genes have been shown to frequently affect gene expression and function (Ritz et al., 2001; Fondon et al., 2008). They are responsible for more than 20 neurological disorders and have been implicated in several others including autism, Parkinson's disease, and Huntington's disease (Duyao et al., 1993; Eerola et al., 2010;

[*]**Corresponding author:** Harold R. Garner, Ph.D., Director of Medical Information Systems, Virginia Bioinformatics Institute, Virginia Tech, Washington Street, MC0477, Blacksburg, VA 24061-0477, USA, garner@vbi.vt.edu, Phone: (540) 231-2582, Fax: (540) 231-1388.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Vedrine et al., 2011). Microsatellite mutations are also involved in several cancers. For example, a microsatellite locus 4,400 bases from the transcription start site of *ERBB2* was recently associated with breast cancer risk (Breyer et al., 2009). Also breast cancer patients' germ line DNA was found to have a global microsatellite signature involving a dozen AT-rich motif families (Galindo et al., 2010).

Microsatellites are highly polymorphic yet difficult to analyze en masse from next-generation sequencing data, and thus variations at all loci have likely not been identified. The vast difference in the reporting of microsatellites polymorphisms when compared to other variations, such as single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels), is apparent considering that the latest release of dbSNP contains over 40 million entries for the human genome of which only 5,198 (less than 0.02%) are labeled as microsatellite polymorphisms (Sherry et al., 2001). This is as expected as the largest genetic map of microsatellite variation, the deCODE map, genotyped 5,136 microsatellite markers in 146 families (Kong et al., 2002), with the NCBI sequence-tagged sites database currently containing only 581 microsatellite sequences (Olson et al., 1989). The deCODE map was created when microsatellites were the main loci used for linkage studies. However, now that SNPs can be discovered faster with less genotyping cost (Syvanen, 2005) they are now the variant most commonly studied and most thoroughly characterized.

Recent advances in methods for obtaining reliable microsatellite genotypes from next-generation sequencing data have potential to provide a more complete view of variations at repetitive loci (McIver et al., 2011; Fondon et al., 2012; Gymrek et al., 2012). However, some methods are limited with lobSTR not able to call monomers and RepeatSeq not written to accept reads from the LS454 platform (Highnam et al., 2012). lobSTR is also limited in the number of microsatellite loci it can call and thus the number of variations it will detect (Highnam et al., 2012). For example, using a whole genome sequenced 1000 Genomes Project trio (father, mother, daughter), lobSTR called ~57% loci (25,885 of the 45,461 callable microsatellites) variable (Gymrek et al., 2012) while our original method called ~46% loci (49,316 of the 108,154 callable microsatellites) variable (McIver et al., 2011). RepeatSeq was tested on the trio but the global variation data was not reported (Highnam et al., 2012). We would expect the global number of polymorphic microsatellites to be in the range of 75,000 to 500,000 if the sequencing coverage of this study was high enough that all ~2 microsatellites in the human genome could be called (Wren et al., 2000; Payseur et al., 2011). RepeatSeq, like our method, is able to call significantly more microsatellites than lobSTR. However, in validating a subset of 40 discordant RepeatSeq and lobSTR calls, only 62.5% of the RepeatSeq calls were confirmed by Sanger sequencing (Highnam et al., 2012). In this study we modify our original methods to include local alignment and introduce the ability to accurately identify novel SNPs in repeat sequences. Additionally, with Sanger sequencing and data from HapMap, we are able to validate 96.5% of a subset of 85 non-synonymous variants called with our revised methods.

We applied our microsatellite genotyping method to high coverage sequencing data obtained from a 1000 Genomes Project pilot study (Durbin et al., 2010). This pilot study used targeted next-generation sequencing to obtain the sequences of the exons of 906 randomly selected genes in 697 individuals (Durbin et al., 2010). We identified and experimentally validated novel variants that were not detected by common variant calling methods. Some of these variants are located in genes associated with cancer.

2. Results and Discussion

2.1. Microsatellite genotypes gathered from 551 individuals representing seven populations

The 697 genomes included in the 1000 Genomes Project Pilot study were sequenced on a variety of second generation sequencing platforms (Durbin et al., 2010) with the samples representing seven populations from six countries (Supplementary Table 1). Of the 697 individuals, 570 were sequenced at the minimum read length required by this study (45 bases), as this read length could span at least 98.6% of all microsatellites identified in the targeted region, based on the allele in the reference genome. However, 19 of these 570 individuals lacked adequate read coverage to call at least one reliable microsatellite genotype and were not considered further, leaving 551 individuals.

The average depth of coverage in targeted regions (comprising 1,423,559 bases) for the 551 individuals analyzed was 45. 5× (Supplementary Table 1). The average number of reads which completely spanned a microsatellite locus plus at least 10 flanking bases on both ends was 26.7 in exonic regions and 14 overall. Therefore the effective number of quality reads for calling microsatellite genotypes was at least half the average coverage in targeted regions. Standard variant calling methods, indel/SNP-based methods, use all reads mapped to a microsatellite locus which artificially increases the coverage and assumed confidence of the call. Using these reads can also lead to inaccurate microsatellite genotyping.

Microsatellite genotypes were obtained from at least one of the 551 individuals for 8,124 of the 8,342 (97.4%) microsatellite loci, of which 2,304 exhibited some form of polymorphism, including 335 repeat length variations and 2,086 single nucleotide polymorphisms (SNPs) (Table 1 and 2). Variation, of either a SNP or repeat length, in 935 of the 2,304 polymorphic microsatellites were observed in more than one individual (Table 1 and 2), with a total of 25,115 variations identified. The frequency of variations was 45.6 per genome or 1.1 variations per repetitive sequence kilobase considering 4,106.9 (standard deviation (SD) ± 1,467.9) ~10 base microsatellites, on average, could be reliably genotyped for each of the 551 genomes studied. We were only able to call on average 4,106.9 microsatellites per genome because short (≤50 base) read lengths comprised the bulk of this early-second-generation sequence data resulting in lower coverage in microsatellites within targeted regions. The variation density of 1.1 variations per kilobase attributable to the repetitive regions falls within the widely accepted density of polymorphic loci in the human genome, ranging from one to three per kilobase (Pumpernik et al., 2008). Considering only variations in coding regions, we identified 0.6 variations per kilobase, which is lower than the predicated density of human exonic SNPs at 0.9 per kilobase (Sachidanandam et al., 2001), possibly due to the small region we are analyzing.

The average individual variation for all genomes was approximately 1.0% with the lowest amount of variation detected in exon regions. In a prior study, microsatellite-based genotyping reported the average repeat length variation of the two trios (mother, father, and daughter) sequenced by a 1000 Genomes Project pilot study was approximately 1% globally (McIver et al., 2011). Significantly less variation was seen in microsatellites in exons in the trio study, with rates varying from 0.0% to 0.2% (McIver et al., 2011). The average individual divergence of microsatellite loci for this study (0.5%) is higher than that found in the trio perhaps attributable to the larger statistical sampling (over 500 genomes). However, it is in general a possibly low estimation of microsatellite variation in exons because the majority of the microsatellites in this study (94.5%) were less than 20 bases in length, as microsatellite mutability increases with microsatellite length (Ellegren, 2004).

2.2. Variation at microsatellite loci shows differences between populations

The total number of variations at microsatellite loci was computed for each population sample. The Kenyan (LWK) sample contained the highest number of variations from the reference genome (Supplementary Table 2). This held true as well for SNPs and indels found using standard variant calling, indel/SNP-based genotyping (Durbin et al., 2010). Indel/SNP-based genotyping, like microsatellite-based genotyping, also identified the LWK sample as containing the most variations from the reference genome (a total of 4,255) (Durbin et al., 2010).

Average divergence at microsatellite loci per individual, computed by comparing each individual in a population to the reference genome, was the highest in the Kenyan (LWK) samples when considering only those individuals for which we could accurately genotype at least 300 microsatellite loci (Supplementary Table 3). However, the population with the highest number of variations also had the most individuals for which a large number of microsatellite loci could be genotyped so these differences are most likely artifacts of sampling.

2.3. Validation of non-synonymous exonic variations in repeat regions

We used Sanger capillary sequencing and data from HapMap (2003) to evaluate allele calls for a subset of 85 non-synonymous exonic variations identified by microsatellite-based genotyping. Indel/SNP-based genotyping consisted of 4.9% false negatives while microsatellite-based genotyping resulted in 16 (18.8%) false positives. Increasing the stringency in calling novel SNPs by requiring at least 3 reads covering a variation at 99.9% accuracy, results in microsatellite-based genotyping calling 96.5% of the loci correctly and reduces the rate of false positives to 3.5% while maintaining a 0% false negative rate. Indel/SNP-based genotyping is able to call 91.9% of the loci correctly, in part because, due to low coverage, it did not report genotypes for three loci. However, Indel/SNP-based genotyping was slightly more successful at distinguishing between homozygous and heterozygous variations, reporting 98.3% of the variations called accurately, with microsatellite-based genotyping reporting 92.0% of the variations called accurately. This difference could be due in part to the fact that microsatellite-based genotyping called all loci in the validation set while indel/SNP-based genotyping only called 96.5% of the loci. Microsatellite-based genotyping was able to accurately call variations at the three loci that were not genotyped by indel/SNP-based methods. Although these loci were called homozygous, Sanger sequencing confirmed them to be heterozygous.

The repeat length polymorphisms identified by microsatellite-based genotyping in three of the four microsatellite loci were confirmed, including a novel repeat length variation in cadherin-related family member 2 (*CDHR2*) a gene associated with colon cancer tumor suppression (Okazaki et al., 2002). None of these confirmed variations were identified using indel/SNP-based genotyping methods.

One of the confirmed repeat alleles differed in the Sanger sequencing data from the microsatellite-based genotype; a repeat length contraction identified in *CLSPN*, claspin, as homozygous was determined to be heterozygous by Sanger sequencing, with the other allele matching the sequence found in the reference genome. Microsatellite-based genotyping failed to report the allele matching the reference as there were only six spanning reads, all of which supported the shorter allele. We believe this is due to low coverage combined with allelic bias, which is common in target enrichment (Sherlock et al., 1998; Wells et al., 1999). The other variation captured at low coverage (6×) was in *KANK1*, KN motif and ankyrin repeat domains 1; Sanger sequencing indicated this microsatellite as identical to the reference. The differences between the microsatellite-based genotyping calls and the Sanger sequencing for these two loci is due to very low coverage, as the average indel calling software requires 30× coverage for maximum accuracy (Neuman et al., 2012).

All novel variants confirmed with Sanger sequencing have been submitted to dbSNP under the handle SGARNER. Additionally all microsatellite variations identified in the 551 individuals, along with those found in the two trios analyzed previously (McIver et al., 2011), are publically available on-line at MicrosatDB (http://discovery.vbi.vt.edu/MicrosatDB/).

2.4. Microsatellite-based genotyping improves the detection of variations

Indel/SNP-based genotyping identified over 86% of the exonic SNPs that were identified using microsatellite-based genotyping though only reported 5.8% of the exonic repeat length variations. The small number of microsatellites characterized in the database commonly used by standard variation detection, dbSNP, is in part why indel/SNP-based genotyping only reported a small percent of the microsatellite variations. Overall ~63% of the exonic repeat length variations not identified by indel/SNP-genotyping were also not found in dbSNP.

Over 96.3% of the exonic variations called using indel/SNP-based genotyping were located at indels/SNPs that were previously recorded in dbSNP, indicating the database training set has an influence on the variants reported. Indel/SNP-based genotyping using the Genome Analysis Toolkit (GATK) (McKenna et al., 2010) along with GigaBayes (Marth, 2012) and Atlas-Indel2 (Danny Challis, 2012) is considered to be highly reliable in identifying previously recognized variations but requires a library, like dbSNP. Though dbSNP has over 40 million entries, only approximately 5,000 are labeled as microsatellite variations, with these likely used in the past for linkage studies (Kong et al., 2002). Adding indels located in repetitive regions to our microsatellite count of variants in dbSNP increases it by 27,682 loci. However, the new total of possibly ~33,000 microsatellite, repeat length, variants in dbSNP is still considerably smaller than expected; with estimates that 25% of microsatellites are highly variable (Wren et al., 2000), we would expect to see at least 500,000 entries. Additionally the accuracy of the 27,682 putative microsatellite variations is difficult to determine as they were likely called with indel/SNP-based genotyping which can result in calling errors in repetitive regions (Treangen and Salzberg, 2012). To increase the number of microsatellite variants accurately characterized, we have created a publically available on-line resource which contains all of the repeat length variations we have identified in genomes sequenced from the 1000 Genomes Project (http://discovery.vbi.vt.edu/MicrosatDB/).

The alignment techniques used with indel/SNP-based genotyping can negatively impact their ability to identify microsatellite variations in a variety of ways (Treangen and Salzberg, 2012). For example, the polymorphic microsatellite locus in *COL3A6* (collagen (type VI) alpha 3), which is recorded in dbSNP, was identified in this study using microsatellite based genotyping but was not called with indel/SNP-genotyping. Considering a traditional alignment at this locus, the majority of these reads (81.5%) do not indicate any difference from the human reference sequence (Figure 1A). This causes standard variant detection software to call the locus as the same as the reference genome. However, by altering the alignment to only contain the reads that completely span the repeat and include flanking non-repetitive bases on both ends, microsatellite-based genotyping mitigates the reference-biased effects of these non-informative reads, so that 41.6% of the remaining reads support a shorter CAG repeat allele (Figure 1B), resulting in a heterozygous call confirmed by Sanger sequencing (Figure 1C). Our microsatellite-based genotyping, unlike traditional alignment techniques, also performs local realignment allowing for approximately an additional million reads in this study to be used in calling variants which were not able to be aligned using BWA.

2.5. Exonic variations identified in genes associated with cancer

From genotyping microsatellites in over 500 samples, we were able to identify many novel polymorphic sites, some of which were located in exons. Of the 19 exonic repeat length variations identified, only seven appear in dbSNP or have been previously published (Table 3). Both novel non-synonymous SNPs and exonic repeat length variations are important to discover, as repeat length variations can affect the gene function in a variety of ways (Fondon et al., 2008). For example, repeat length variations can affect transcription rates, protein-protein interactions, and transcript stability with some variations turning genes on or off (Fondon et al., 2008).

Two repeat length variations are in genes associated with cancer. We identified a variant in *COL6A3*, a gene significantly up regulated in pancreatic cancer (Arafat et al., 2011). Another variation was identified in *CLSPN*; though increased expression of *CLSPN* has been seen in cancer cell lines and this gene is associated with fragile site expression and genome instability, both of which are associated with cancer, this repeat was previously studied and not found to be significantly associated with breast cancer (Erkko et al., 2008; Focarelli et al., 2009). A novel variation of a single repeat unit was identified in *CDHR2*, a candidate for tumor suppression since elevated expression of *CDHR2* in colon cancer cells has been shown to prevent tumor formation in vivo (Okazaki et al., 2002); the colon sample used in the *CDHR2* study was HCT116 which is known to exhibit high-frequency microsatellite instability (Huang et al., 2011).

A novel non-synonymous SNP was identified in exon 8 of *TEX14*, testis-expressed protein 14 (also named cancer/testis antigen 113). A SNP in the 5’UTR region of *TEX14*, rs302864, is associated with a significantly increased risk of pancreatic cancer for individuals with high body mass index (BMI>=30) (Couch et al., 2010). The rs302864 SNP, which is 64,936 bps upstream of the novel exonic variant we identified, has also been associated with breast cancer (Kelemen et al., 2009). Another novel non-synonymous SNP, was identified in exon 13 of *HEATR6*, HEAT repeat-containing protein 6; this gene is highly expressed in breast cancer (Sinclair et al., 2003).

3. Conclusion

Our microsatellite-based genotyping method identified novel repeat length variations and SNPs with the majority of repeat length variations not identified by standard variant calling methods. We found standard variation detection software did not identify these variations due to alignment techniques not tailored towards non-unique regions and the use of a database which only characterizes a limited number of repeat length variations. With the accuracy of microsatellite-based genotyping estimated at over 96%, we propose this method be used in addition to standard variant calling methods for large scale genome sequencing studies.

4. Materials and methods

4.1. Microsatellite identification

Microsatellites at least 10 base pairs long, with no more than one interruption to the canonical repeat sequence for each ten bases in length (≥ 90% “pure”), and within 500 base pairs of the 1000 Genomes Project Pilot targeted exon regions, were identified in human reference genome (NCBI36/hg18) using Tandem Repeat Finder (Benson, 1999). The parameters provided to TRF for our initial microsatellite set were: matching weight=2, mismatching penalty=5, indel penalty=5, match probability=80, indel probability=10, minimum alignment score to report=14, maximum period size to report=4. Running TRF again with the maximum period size to report set to 6 allowed us to supplement our initial data set with 5-mer and 6-mers. Microsatellites within or immediately adjacent to other microsatellites or larger repetitive elements identified using RepeatMasker were removed (Smit AFA, 1996–2012). After removing all monomers, the resulting set of microsatellite loci totaled 8,342: 2-mer (n=353), 3-mer (n=857), 4-mer (n=2,066), 5-mer (n=1,483), and 6-mer (n=3,603). Using genomic locations, these microsatellites were associated with all genes they were in or near using the Refseq data provided by the UCSC Genome Browser (Rhead et al., 2010). Microsatellites that were located in two gene regions were labeled as belonging to the region in which most of their sequence was contained. The points 1,000 bases from the transcription start and end sites of each gene were defined as the upstream and downstream boundaries.

4.2. Identifying variations at microsatellite loci using microsatellite-based genotyping

The quality filtered reads from the 1000 Genomes Project (Durbin et al., 2010), at least 45 base pairs in length, were aligned to the human reference genome (NCBI36/hg18) using BWA, with BWA-SW for 454 reads (Li and Durbin, 2009; Li et al., 2009; Durbin et al., 2010). Next we applied microsatellite-based genotyping, which uses non-repetitive flanking sequences to ensure reliable mapping and alignment at microsatellite loci. This approach incorporated heuristics that were optimized to obey Mendelian inheritance of informative loci using deep sequencing data of two trios produced in the first phase of the 1000 Genomes Project (McIver et al., 2011). For this study we have modified our methods to allow us to call SNPs in tandem repeats.

We start by identifying reads that completely span the repeat region plus some unique flanking sequence on both ends. We further filtered these results using a 10 base flanking sequence to enable comparison to the common SNP filtering window used for MAQ and GATK (McKenna et al., 2010). Increasing this minimum flank length from 5 bases to 10 bases reduced the number of callable loci by less than 5% but increases confidence in our alignments by relying on additional unique sequences. Those reads that were not aligned by BWA to the reference along with the reads that were aligned to a microsatellite locus by BWA but did not meet our unique flanking sequence criteria were run through our custom code to determine if they should be aligned to another microsatellite locus based on flanking sequences and a short portion of the repeat. This allows us to maximize our use of reads with repetitive sequences, adding almost a million useable reads, and it also removes the restriction associated with the length of BWA indel calling on our method.

Using a small portion of the repeat is essential as over half of the microsatellites in our set have multiple alignments in the human genome if we allow the flanking sequences to be separated by at most 200 bases. Two hundred bases was chosen as it is slightly less than the average sequencing read length for this study considering half are from the Illumina with a maximum of 100 bases with the remaining from the 454 with an average of 400 bases.

Next reads were grouped for each microsatellite locus based on the repeat length variations or SNPs they contained. Then we applied the same heuristics from our first study, modified slightly to account for the increase in coverage and the larger quantity of 454 reads. More specifically the microsatellite loci were not allowed to have more than two allelotypes to call a variation, after filtering those alleles supported by fewer than five reads or more than 50 reads, the average depth of coverage for all populations studied (Hudson, 2008). Average coverage was determined for each population by selecting 10 random points in the targeted sequencing regions for each sample, finding the coverage using SAMTOOLS, and calculating an average for the group of individuals (Li et al., 2009).

In the case of microsatellites which could possibly be heterozygous, they were only considered to be heterozygous if the reads for each allele were no more than two times the reads of the second allele. This allowed for unequal amplification, which is an issue with whole genome sequencing, with only 17–40% of microsatellite alleles sequencing equally, and even more of an issue with targeted sequencing (Sherlock et al., 1998; Wells et al., 1999). Since the 454 platform is known to have errors in sequencing homopolymer regions (Margulies et al., 2005; Wicker et al., 2006), reads from the 454 platform with homopolymer indels were thrown out prior to performing microsatellite-based genotyping. For population comparisons, Illumina reads were processed to remove homopolymer indels so that the methods applied to all samples would be identical.

4.3 Validation of non-synonymous variations

A majority of the SNPs in exons that were found at known SNP locations were validated using the alleles provided by HapMap (2003). The variations identified at known SNP locations for which alleles were not provided by HapMap were validated by Sanger sequencing.

DNA (cell-culture) was obtained from the Coriell Cell Repositories (Camden, New Jersey) for the four individuals (NA12717, NA19220, NA19321, and NA20763) whose genomes contained possible repeat length variations in exons along with 22 additional individuals for which we identified putative SNPs. Forward and reverse primers were designed for each of the non-synonymous variations that could not be verified using HapMap to amplify these specific repeat regions (Supplementary Table 4). Per the manufacturer’s instructions, PCR was performed using the Promega 2X PCR Master Mix. Next the PCR products were cleaned using ExoSAP-IT (Affymetrix, Cleveland, Ohio) and then sequenced by the Virginia Bioinformatics Institute Core Facility at Virginia Tech.

4.4 Comparing variations found at microsatellite loci using microsatellite-based genotyping to those found using indel/SNP-based genotyping

The genomic locations of all variant calls, found with indel/SNP-based genotyping using GATK (McKenna et al., 2010) along with Gigabayes (Marth, 2012) and Atlas-Indel2 (Danny Challis, 2012), were downloaded from the 1000 Genomes Project website (Durbin et al., 2010). We searched the regions of each of the microsatellite loci analyzed in this study for variants, including three bases on either side of the repeat.

4.5 Determining which of the variations identified were novel

The genomic locations of all known variations from dbSNP (v130) corresponding to the hg18 release of the human reference genome were downloaded from the UCSC Genome Tables (Rhead et al., 2010). We checked all of the variants found in all of 551 individuals studied to determine if any variant in dbSNP was located within this region. Any variation in this region was recorded as a possible match. Exon variations were also manually verified using the latest release of dbSNP (v132) corresponding to hg19 to determine if there were any additional variations recorded (Sherry et al., 2001).

4.6 Counting the total indels in dbSNP in repetitive regions

Using the same rules as those to create the targeted set of microsatellites analyzed in this study, we created a set of global microsatellites. This microsatellite set, composed in part of monomers, totaled 830,153. This set includes less than half of the ~2 million which are present in the human genome because we limited them by length, at least 12 bases on average, and we also limited them by a purity of 90%. Those microsatellites not included are highly unlikely to be variable as polymorphism rates increase with length and purity

(Ellegren, 2004; Fondon et al., 2008). The dbSNP database for hg18 (v130) was searched for any indels within these repeats, including two bases of flanking sequence on both sides of the repetitive regions. A possible repeat length variation was recorded if any indel was included in the dbSNP database in these regions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the Virginia Bioinformatics Institute Director’s Funds and the National Institute of Health, National Human Genome Research Institute, 1000 Genomes Project Dataset Analysis Grant (1U01HG005719-01). The high performance computing infrastructure on which this analysis was conducted was supported by a grant from the National Science Foundation (OCI-1124123). We thank Michael Snow, Dominik Borkowski, David Bynum, Tiffany Trent, and Laurie Spotswood for technical support.

All authors have agreed to all the content in the manuscript, including the data as presented.

References

The International HapMap Project. Nature. 2003; 426:789–796. [PubMed: 14685227]

Arafat H, Lazar M, Salem K, Chipitsyna G, Gong Q, Pan TC, Zhang RZ, Yeo CJ, Chu ML. Tumor-specific expression and alternative splicing of the COL6A3 gene in pancreatic cancer. Surgery. 2011; 150:306–315. [PubMed: 21719059]

Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27:573–580. [PubMed: 9862982]

Breyer JP, Sanders ME, Airey DC, Cai Q, Yaspan BL, Schuyler PA, Dai Q, Boulos F, Olivares MG, Bradley KM, Gao YT, Page DL, Dupont WD, Zheng W, Smith JR. Heritable variation of ERBB2 and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2009; 18:1252–1258. [PubMed: 19336545]

Couch FJ, Wang X, Bamlet WR, de Andrade M, Petersen GM, McWilliams RR. Association of mitotic regulation pathway polymorphisms with pancreatic cancer risk and outcome. Cancer Epidemiol Biomarkers Prev. 2010; 19:251–257. [PubMed: 20056645]

Danny Challis, UEaFY. Atlas-Indel2. 2012

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

Duyao M, Ambrose C, Myers R, Novelletto A, Persichetti F, Frontali M, Folstein S, Ross C, Franz M, Abbott M, et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet. 1993; 4:387–392. [PubMed: 8401587]

Eerola J, Luoma PT, Peuralinna T, Scholz S, Paisan-Ruiz C, Suomalainen A, Singleton AB, Tienari PJ. POLG1 polyglutamine tract variants associated with Parkinson's disease. Neurosci Lett. 2010; 477:1–5. [PubMed: 20399836]

Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004; 5:435–445. [PubMed: 15153996]

Erkko H, Pylkas K, Karppinen SM, Winqvist R. Germline alterations in the CLSPN gene in breast cancer families. Cancer Lett. 2008; 261:93–97. [PubMed: 18077083]

Focarelli ML, Soza S, Mannini L, Paulis M, Montecucco A, Musio A. Claspin inhibition leads to fragile site expression. Genes Chromosomes Cancer. 2009; 48:1083–1090. [PubMed: 19760606]

Fondon JW 3rd, Hammock EA, Hannan AJ, King DG. Simple sequence repeats: genetic modulators of brain function and behavior. Trends Neurosci. 2008; 31:328–334. [PubMed: 18550185]

Fondon JW 3rd, Martin A, Richards S, Gibbs RA, Mittelman D. Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. PLoS One. 2012; 7:e33036. [PubMed: 22427938]

Fondon JW 3rd, Mele GM, Brezinschek RI, Cummings D, Pande A, Wren J, O'Brien KM, Kupfer KC, Wei MH, Lerman M, Minna JD, Garner HR. Computerized polymorphic marker identification: experimental validation and a predicted human polymorphism catalog. Proc Natl Acad Sci U S A. 1998; 95:7514–7519. [PubMed: 9636181]

Galindo CL, McCormick JF, Bubb VJ, Abid Alkadem DH, Li LS, McIver LJ, George AC, Boothman DA, Quinn JP, Skinner MA, Garner HR. A long AAAG repeat allele in the 5' UTR of the ERR-gamma gene is correlated with breast cancer predisposition and drives promoter activity in MCF-7 breast cancer cells. Breast Cancer Res Treat. 2010; 130:41–48. [PubMed: 21153485]

Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 2012

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res. 2012

Huang SC, Lee JK, Smith EJ, Doctolero RT, Tajima A, Beck SE, Weidner N, Carethers JM. Evidence for an hMSH3 defect in familial hamartomatous polyps. Cancer. 2011; 117:492–500. [PubMed: 20845481]

Hudson ME. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour. 2008; 8:3–17. [PubMed: 21585713]

Kelemen LE, Wang X, Fredericksen ZS, Pankratz VS, Pharoah PD, Ahmed S, Dunning AM, Easton DF, Vierkant RA, Cerhan JR, Goode EL, Olson JE, Couch FJ. Genetic variation in the chromosome 17q23 amplicon and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2009; 18:1864–1868. [PubMed: 19454617]

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. A high-resolution recombination map of the human genome. Nat Genet. 2002; 31:241–247. [PubMed: 12053178]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

Marth GT. GigaBayes. 2012

McIver LJ, Fondon JW 3rd, Skinner MA, Garner HR. 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. 2011; 97:193–199. [PubMed: 21223998]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]

Neuman JA, Isakov O, Shomron N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. Brief Bioinform. 2012

Okazaki N, Takahashi N, Kojima S, Masuho Y, Koga H. Protocadherin LKC, a new candidate for a tumor suppressor of colon and liver cancers, its association with contact inhibition of cell proliferation. Carcinogenesis. 2002; 23:1139–1148. [PubMed: 12117771]

Olson M, Hood L, Cantor C, Botstein D. A common language for physical mapping of the human genome. Science. 1989; 245:1434–1435. [PubMed: 2781285]

Payseur BA, Jing P, Haasl RJ. A genomic portrait of human microsatellite variation. Mol Biol Evol. 2011; 28:303–312. [PubMed: 20675409]

Pumpernik D, Oblak B, Borstnik B. Replication slippage versus point mutation rates in short tandem repeats of the human genome. Mol Genet Genomics. 2008; 279:53–61. [PubMed: 17926066]

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 2010; 38:D613–D619. [PubMed: 19906737]

Ritz D, Lim J, Reynolds CM, Poole LB, Beckwith J. Conversion of a peroxiredoxin into a disulfide reductase by a triplet repeat expansion. Science. 2001; 294:158–160. [PubMed: 11588261]

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409:928–933. [PubMed: 11237013]

Sherlock J, Cirigliano V, Petrou M, Tutschek B, Adinolfi M. Assessment of diagnostic quantitative fluorescent multiplex polymerase chain reaction assays performed on single cells. Ann Hum Genet. 1998; 62:9–23. [PubMed: 9659974]

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]

Sinclair CS, Rowley M, Naderi A, Couch FJ. The 17q23 amplicon and breast cancer. Breast Cancer Res Treat. 2003; 78:313–322. [PubMed: 12755490]

Smit AFA, H R, Green P. RepeatMasker Open-3.0. 1996–2012

Syvanen AC. Toward genome-wide SNP genotyping. Nat Genet. 2005; 37(Suppl):S5–S10. [PubMed: 15920530]

Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 2000; 10:967–981. [PubMed: 10899146]

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012; 13:36–46. [PubMed: 22124482]

Vedrine SM, Vourc'h P, Tabagh R, Mignon L, Hofflin S, Cherpi-Antar C, Mbarek O, Paubel A, Moraine C, Raynaud M, Andres CR. A functional tetranucleotide (AAAT) polymorphism in an Alu element in the NF1 gene is associated with mental retardation. Neurosci Lett. 2011; 491:118–121. [PubMed: 21236316]

Wells D, Sherlock JK, Handyside AH, Delhanty JD. Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation. Nucleic Acids Res. 1999; 27:1214–1218. [PubMed: 9927758]

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N. 454 sequencing put to the test using the complex genome of barley. BMC Genomics. 2006; 7:275. [PubMed: 17067373]

Wren JD, Forgacs E, Fondon JW 3rd, Pertsemlidis A, Cheng SY, Gallardo T, Williams RS, Shohet RV, Minna JD, Garner HR. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet. 2000; 67:345–356. [PubMed: 10889045]

Highlights

- We use microsatellite-based genotyping to indentify variations in 551 individuals.
- Over 68% of the exonic repeat length variations we identify are novel.
- Indel-based genotyping only reports 5.8% of the exonic repeat length variations.
- Microsatellite-based genotyping accuracy, from experimental validation, is 96.5%.
- Novel non-synonymous variations we identify are in cancer genes.

(A)

```
Ref:                CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_74592:       CGCCACTGGTTTTGCAGCAG
SEQ_ID_2869537:     CGCCACTGGTTTTGCAGCAG
SEQ_ID_7679106:     CGCCACTGGTTTTGCAGCAG
SEQ_ID_8807551:     CGCCACTGGTTTTGCAGCAG
SEQ_ID_12070758:    CGCCACTGGTTTTGCAGCAG
SEQ_ID_12206959:    CGCCACTGGTTTTGCAGCAG
SEQ_ID_1531358:     CGCCACTGGTTTTGCAGCAGCAGC
SEQ_ID_2923266:     CGCCACTGGTTTTGCAGCAGCAGC
SEQ_ID_3587544:     CGCCACTGGTTTTGCAGCAGCAGC
SEQ_ID_3587643:     CGCCACTGGTTTTGCAGCAGCAGC
SEQ_ID_3587715:     CGCCACTGGTTTTGCAGCAGCAGC
SEQ_ID_10263160:    CGCCACTGGTTTTGCAGCAGCAGCAGC
SEQ_ID_1931311:     CGCCACTGGTTTTGCAGCAGCAGCAGCG
SEQ_ID_264343:      CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTA
SEQ_ID_7962571:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTC
SEQ_ID_2979623:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTA
SEQ_ID_6180285:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTA
SEQ_ID_3395352:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_9120508:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_9138431:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_2012466:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_3220589:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_6308493:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_7472536:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_4125481:      GCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_7374254:      GCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_5446439:                               CGGGGGGTCTTACA
```

(B)

```
Ref:                CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_264343:      CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTA
SEQ_ID_2979623:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTA
SEQ_ID_6180285:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTA
SEQ_ID_3395352:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_9120508:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_9138431:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_2012466:     CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_3220589:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_6308493:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_7472536:     CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
SEQ_ID_4125481:      GCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
SEQ_ID_7374254:      GCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
```

(C)

```
Ref:                CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
PCR Sequence1:      CGCCACTGGTTTTGCAGCAGCAGC-------GGGGGGTCTTACA
PCR Sequence2:      CGCCACTGGTTTTGCAGCAGCAGCAGCGGGGGGTCTTACA
```

Figure 1.

The raw reads obtained from sequencing individual NA12717 are aligned to the human reference sequence at a polymorphic microsatellite locus in exon 39 of *COL6A3*. The repeat region is shown in bold. All aligned reads are shown for indel/SNP-based genotyping, which display a majority of reads that do not indicate a variation from the reference sequence, (A) and microsatellite-based genotyping, which indicates a heterozygous locus with a single codon deletion, (B) along with the consensus sequences obtained from Sanger sequencing, which, like microsatellite-based genotyping, also indicate a heterozygous locus with a single codon deletion in one allele with the other allele matching the reference (C).

Table 1

Repeat length variations identified using microsatellite-based genotyping

Region	Microsatellite Loci	% Reliable Alignments	Repeat Length Variations	% Homozygous	dbSNP
Upstream	194	81.4%	6	68.9%	66.7%
5’UTR	296	88.9%	7	44.4%	-
Exon	1,311	98.9%	19	32.8%	31.6%
Intron	6,024	98.0%	286	69.4%	37.8%
3’UTR	389	96.9%	12	79.3%	33.3%
Downstream	128	96.9%	5	91.3%	60.0%
Total	8,342	97.4%	335	69.1%	37.3%

The total microsatellites examined were within 500 base pairs of the targeted exon regions. Microsatellite counts are shown by region with upstream and downstream defined as the regions 1,000 base pairs from the start and end of transcription. A reliable alignment indicates a microsatellite genotype was captured in at least one of the 551 individuals. The total number of microsatellite loci with repeat length variations differing from the reference genome (hg18) using microsatellite-based genotyping is shown. Also shown is the percent of all repeat length variations in all samples which were homozygous. The dbSNP column displays the number of variations which were previously recorded in the dbSNP database.

Table 2

Single nucleotide polymorphisms identified using microsatellite-based genotyping

Region	Microsatellite Loci	SNPs	% Homozygous
Upstream	194	38	51.6%
5’UTR	296	63	63.6%
Exon	1,311	317	64.3%
Intron	6,024	1,534	65.1%
3’UTR	389	103	70.4%
Downstream	128	31	72.4%
Total	8,342	2,086	65.1%

The total number of microsatellite loci with SNPs identified using microsatellite-based genotyping is shown. Also displayed is the percent of all SNPs in all samples which were homozygous.

Table 3

Repeat length variations at exonic microsatellite loci

Gene	Exon	Motif	Variation	Total Genomes (homo/het)	Sequencing Platforms	dbSNP
ACCN3	4	$(CCCCAG)_3$	−/−CAGCCC	9 (6/3)	454	rs3217353
ACCN4	1	$(CCAGCA)_2$	−/+CCAGCA	1 (1/0)	454	-
APCDD1	4	$(ACA)_4$	−/−ACA	1 (0/1)	454	-
		−/+CGGGAC	1 (0/1)	454		
C1orf63	2	$(CGGGAC)_2$	−/−CGGGAC	1 (1/0)	454	-
CDHR2	8	$(ACA)_3$	−/−CAA	1 (0/1)	454	-
CLSPN	22	$(TTC)_5$	−/−GAA	3 (2/1)	Both	-
COL6A3	39	$(GCA)_4$	−/−GCT	4 (1/3)	Both	rs71704006
FBXO2	2	$(GCG)_3$	−/+GCG	1 (1/0)	454	rs148874459
			$−/−(TCC)_5$	1 (1/0)	454	-
FTSJ3	12	$(TCC)_9$	$−/−(TCC)_2$	1 (0/1)	454	-
			−/+GGA	1 (0/1)	454	
KANK1	5	$(GAG)_5$	−/−GGA	1 (1/0)	Illumina	rs113586916
OXA1L	10	$(AGC)_4$	−/+AGC	67 (17/50)	Both	rs148216086
PBK	7	$(TCA)_5$	−/−TCA	1 (0/1)	454	-
POLE	44	$(CA)_6$	−/−CACA	1 (0/1)	454	-
RNF25	7	$(GGCAGT)_3$	−/+G	1 (0/1)	454	-
RNF145	10	$(GAAT)_2$	−/−AA	1 (0/1)	Illumina	-
USP31	16	$(CAGCCC)_2$	−/−CAGCCC	2 (0/2)	454	-
UTRN	48	$(CAG)_3$	−/−C	1 (0/1)	454	-
ZKSCAN5	2	$(GAA)_3$	−/−GAA	1 (0/1)	454	-
ZNF474	2	$(GAATTT)_2$	−/−T	36 (15/21)	Both	rs140716087

Exonic repeat variations at microsatellite loci, with average coverage ranging from 5× to 24×, are shown. The total number of individuals in which the variation is found in, including the number which are homozygous and heterozygous for this variation, is also included along with the sequencing platform.