VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data

Yevgeniy Gelfand, Yozen Hernandez, Joshua Loving and Gary Benson

Jclub in Bioinformatics
Tarmo Puurand
05.11.2014

VNTR

- DNA tandem repeats (TRs) are typically divided into two classes,
- *1. microsatellites which have short pattern sizes, generally* 1–6 nucleotides (nt), and
- *2. minisatellites which have* longer patterns.
- Tandem repeat variants, or VNTRs (variable
- number of tandem repeats), are loci in which the number of internal copies in the repeat varies in the population.

1000g and esv-s

1 8976892 esv2670913
ATGAAACCCTGTCTCTACTAAAAATACAAAAATTAGCCGGGCATGGTGGCACGCGCCTGTAGTCCCAGCAACTTGGGAGGGTGAGGCAGGAGAATCACTTGAACCTGGGTGGTGGAGGCTGTGGTGAGCTGAGA
TTGCGCCACTGCACTCCAGCCTGGGTTACAGAAGGAGACTATGTATCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGTCCGGGCGCGGTGGCTCACCTTGTAATCCCAGCGCTTTGGGAGGCCGAGGCGG
GCAGATCACGAGGTCAGGAGTTTGAGATTATCCTGGCCAACATGGTGAAACCCTATCTGTACTAAAAACACAAAAATTAGCTGGGTGTGGTGGCTCGTGCCTGTAATCCCAGCTACTCGGGAGGCCAAGGCAGGA
GAATTGCTTGAAACAGGGAGTTGGAGGTTGCAGTGAGCCGAGATCACACCACTGCACTCCAGCCTGGCGACAGACTGAGACTCTGTTCAAGAAAAAAAATATATAAAAAAAGTTGCTCGTGCCTGTAATCACAGCA
CTTTAGGAGGCCGAGACGGGAGGATTGTTTGAGCCCAGTAGTTCAAGATCAGCCTGGGCAACATAGCAAGACCCTGTCTCTATAAAATAGGAAAAAAAAAGAGGTTAATGTTGTATTATATTACTTTAATTTTACCC
CAATATGAGTTTGTAAACAAATCAAGTTCTCTGTAAACAAGTATTTGATGGTCTGGAAAATGGGATTGCCCAATATCCAAACAGCCGTTTCTGGGCAGCTTCCTGACCATCCGGTAGGTTTCTTGGTCCAGCCCTGA
GTCTCCTTGGCAATGCATCAGGGAGACACCAGGCTCATTCCTCTTTCCTCGTGGCTTCCATATCTTTCCATCCCTCAGCCCTCCGCAGAACCAGCTGTTTGCGGTCACACCGTGGATCTTGTATGAAGCAAAAACAAA
AAAAACAGGGCCAGGCGCGGTGGCTCACGCCTGTAATCCCAGCACGTTGGAGGCTGAGGTGGGCAGATCACTTGAGCTCAGGAATTCAAGACCAGCCTGACCAACATGGTGAAACCCCATCTCTACTAAAAATAC

AAACATTAGCCAGGCACGATGGCTCATGCCTGTAATTCCAGCACTTTGAGAGGCCAAGGTGGGTGGATCAAACTCCTGAGGTCAGGAGTTTGAGACCAGCCTGACCAACATGG **A** . PASS
AC=15;AF=0.01;AFR_AF=0.01;AMR_AF=0.01;AN=2184; AVGPOST=0.9787;CIEND=-39,141;CIPOS=-
18,150;END=8978092;ERATE=0.0032;EUR_AF=0.01;HOMLEN=48;LDAF=0.0147;RSQ=0.4204;SVLEN=-
1200;SVTYPE=DEL;THETA=0.0002;VT=SV;HOMSEQ=TGAAACCCTGTCTCTACTAAAAATACAAAAATTAGCCGGGCATGGTGG

1 9134123 esv2674776
CGCAAAACTCCATCTCCAAAAATTAAAAAAAAAAAAATACGTCAGTAGAATGCAGAAATTGTGTTTGTGTTCTTAGCAGCATGTTCATATAATAGTTTGAGGTGGTGGTGGAGCGGAGGGGGGCAAGCATTCTCTC
AATGAAAGAGCAAGGCTTAGCAACTGTAAACCAAAAAGTATCTGAGACAGGTCTCAATCAATTTAGAAGTCTATTTTGCCAAGGTTAAGGACATGCCCGAAGAAAAAAGCAGAATCTCAGAAATAGTCTGTGGTC
TGTGCCTTTCTCCAAAGATGATTTTGAGGGCGTCGATATTTAAAGAAAAGCTGGCTGGAGGGGAAAGAGGGAGAGTATGGTAACCTACATGTTACAAGAGAAAAGGTACAAGTAGAGGAATCAGCAATTACATG
TCTGTCTCTTGCTCAGTAAATCAGCACTTTACATAAAGTGAGGTGAACACAGAGTAGCTACTTGTGGGGATACTTAACCTTTTACTGTCGCTGTCTGCTTAGGAACATAAGGAAAGACAGCTCCTTGCATGACTCAG
CTTTCAGCTTAATTTTGTTTTTCCTTTTTGGCAGAGTGAATTGGGGTCCTGAGTTTTTTATTTTCCTTTACACAACATATGAAGACCTTAAAGAGAAAACGCTGAAAATTCTGCAGGGCTCTGGGCTGGTTTTGTGATT
TCAAGAAGCGCTTTCTCCTGTAATCCCAGCATTTTGGGAGGCTGAGGCAGGAGGATCGCTTGACCCAGGAGTTCGAGACCAGCCTGGGCAACACAGTGAGACCTTATCTCTACAAATAATCAAAAATTATCTGGGT
GTGGTGGTGTGCACCTATAGTCCCAGCTACTTGAGAGGCTGAGGCAGAAGAATTGCTTGAGCCAGGAACTTCAGGCTGCAGTGAGATGTGACTGAGCCACAGAACTCGGCTTGAGCAACAGAGTGAGATCCCAT
CTCAAAAAACAAAACAAAAAAAAAAAGAAGTGCTTTCCAAGCTTTCTGGGAGTCCTAATTAGGGAAAAGGAGTCAGGTTGGCAGCACCAGGGGAAAGCAAAAGAGAAAGCAAATAAGCTATAAATCTGCCTCTC
TTCATGGTCCACAACAGATAAACAAAGAGGAAGCAGATGAGTTATAGGTCTGTCTGTGTTTATGTCCCAGGAAGTGTAGCCCTCCTGAGCAAATAACACACATAACTCACAGACTTCCCGTTTCACATCAAACACCT
CAATTTATCAAACATCCCGGTTGACAGAAGAAGCAGGTTAGCTCCTGCAGCCGTGGCGTAATCCAACATCCCAAGAGCCATCCTATAAAATCTCCAGCAAGCCTGTTTCCTTGCAGTCCGCTCCTCTTCTGCTGATAC
CGCCCGTTGCCTCCTTGCAACATATTTTCCTACTTTCTCTAATAAATCTATCTTTCTCTACCTACAACTGTCTTGGTAAATTCTTTTACTCCCGTGCCACTGGCCCAGACAGTTGTCACTCCCCTGTGACCCTTTCAATAA
TCTTAAACAAATGGTGGAGTTAGCAGGACAGTTGAAGATATTTTCATATGTAATTTGTTTGTTTTAATTGATTATAGCCTGCATTGGATTTTCAATTTTTTTTTTTTGAGACGAAGTCTCACTCTTGTCCCCCAGGCTG
GAGTGCAATGGCATGATCTCAACTCACTGTAACCTCGGCCTCCCAAGTTCAAGTGATTCTCCTGCCTCAACCTCCCAAGTAGTTGGGATTACAGGCACCTGCCACCATGCCCAGCTAATTTTTGATTTTTAGTAGAAA
CGGGGTTTCACCATGTTGGTCAAGCTGGTCTCGAACTCCTGACCTGAGGTGATCCACCCACCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCGCGCCCGGCATTTTTTTTTTTTTGAGACGGAGTCT
CACAGTGTTGCCCAGGCTGGGGTGCAGTGGCACAATCTCAGCTCACTGCAACCTCCACCTCCCAGATTCAAGCGATTCTCTTGCCTCAGCCTCCCGAGTAACTGAATTACAGACGCACGCCACTACGCCTGGCTAAT
TTTTGTATTTTTAGTAGAGGCGGGGTTTCGCCATGTTGGCCAGGCTGTTCTAGAACTCCTGACCTCAAATGATCTGACCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACTGCACCTGGCCGGATT
TTCAATTTTGATGAAACAGCTCTACTAGGAGTGAAAGGCCCAAAGACCCTAATCTGAGAAGAGAAAGCAGAAGTTCCCCCAGGCGTGGAGATCTGCCCTAGCTGCTGTGCTTATCAAGACTGCATTTATACTTGTT
TGTGTTCAGATTGTAAAACCGCATAGGATTCAAGTGGTCCTTGCCAGTCATTTCTCCCCTTAAGCCCTGTTATTTTGATGTGCCACTTGGCAGAATGTCATGTTCTTTAGGAACAGGTATATCACATGATAGCAGAAA
TGCCTATATGTGCCTTTGATGCACAAAACGGGGTTGCTATCAAGGTGGTGTCTATATTTGGCTTGCTGATTGATGGGGAGAGGAAATGGCTCTTGGCCACACCCATTTATTAATCTAAAGGAAAAAACTGAGGTGA
AATCCATACAAGTAGAGAGTTTATTTGGGCCAAGTTTGAGGACTGCAACCCAGGAGCATAGATTCAAGTTGCCCTGAATTTATGTTCCTATCAGCAATAGTTACAAGTGTGGTGTTTTTTTTGTTTTTTTTTTGAGA
TGCAGTCTCACTCTTGTTGCCCAGACTGGAATGCAGTGGCGTGATCTGGCTCACTGCAACCTCCGTCTCCC **CTCCTTGCATGACTCAGTGCAGTGGCATGATCTCCGACTCCT**
. PASS AC=6;AF=0.0027;AMR_AF=0.0028;AN=2184;AVGPOST=0.9987;CIEND=-17,33;CIPOS=-
18,19;END=9136941;ERATE=0.0006;EUR_AF=0.01;HOMLEN=2;LDAF=0.0034;RSQ=0.8314;SVLEN=-2777;SVTYPE=DEL;THET A=0.0056;VT=SV;HOMSEQ=GC

1 9171842 esv2675581 A **<DEL>** . PASS AC=1;AF=0.0005;AN=2184;ASN_AF=0.0017;AVGPOST=0.9989;CIEND=-17,33;CIPOS=-
18,19;END=9176284;ERATE=0.0006;LDAF=0.0010;RSQ=0.4932;SVTYPE=DEL;THETA=0.0002;VT=SV

VNTRseek program

chr1 9085006 td175348920
GGCCCACCTTGTGCCCACCTTGTGCCCACCTTGTGC
GGCCCACCTTGTGCCCACCTTGTGCCCACCTTGTGCCCACCTTGTGC

. PASS RC=3.18;RPL=11;RAL=35;RCP=**GCCCACCTTGT**;ALGNURL=http://orca.bu.edu/vntrview/index.php?db=VNTRPIPE_NA12891&ref=-175348920&isref=1&istab=1&ispng=1&rank=3 GT:SP:CGL 0/1:7,8:0,1

chr1 9131935 td175348951
CCCTCTGCCTCAGCCCGGATCCCCCTCGGCCTCTGCCTCAGCCCGGGTCCCCCTCGGCCTCTGCCTCAGCCCGGGTCCCCCTCAGCCTC
CCCTCTGCCTCAGCCCGGATCCCCCTCGGCCTCTGCCTCAGCCCGGATCCCCCTCGGCCTCTGCCTCAGCCCGGGTCCCCCTCGGCCTCTGCCTCAGCCCGGG CCCCCTCAGCCTC

. PASS RC=3.14;RPL=28;RAL=88;RCP=**CCTCTGCCTCAGCCCGGGTCCCCCTCGG**;ALGNURL=http://orca.bu.edu/vntrview/index.php?db=VNTRPIPE_NA12891&ref=-175348951&isref=1&istab=1&ispng=1&rank=3 GT:SP:CGL 0/1:3,2:0,1

Other programs

- Most methods use the Tandem Repeats Finder (TRF) program to identify TRs in the reads and to establish reference sets.
- The Garner lab's method
- lobSTR
- RepeatSeq

Workflow of the program

- Our program, *VNTRseek,* is, to the best of our knowledge, the first software for genome-wide detection of minisatellite VNTRs. In outline, it works as follows:
- (i) TRF is used to identify a reference set of TR loci and to identify TRs in the reads;
- (ii) the read TRs are mapped to the reference TRs based on similarity in the repeat consensus patterns, and the TR array profiles;
- (iii) mappings are confirmed based on comparison of the read and reference flanking sequences, adjacent upstream and downstream to the TR arrays; and
- (iv) TR genotypes are called based on the number of pattern copies in the mapped reads. In particular, a locus is called a VNTR if it has at least two mapped reads which exhibit a common copy number different from that in the reference.

Creating reference genome TR list

- TRF program and hg19 (only files chrXXX.fa.gz)
- TRF command line parameters used were 2 5 7 80 10 50 2000 (match weight, mismatch penalty, indel penalty, match probability, indel probability, minimum score, maximum period size).
- The results were filtered in the Tandem Repeats Database to remove:
- (i) low-quality TRs with many indels and mismatches (average per column alignment score ≤ 1.3);
- (ii) TRs having significant overlap (*>20% of total length* overlapping) with common interspersed repeat elements including SINEs, LINEs, LTRs and DNA transposons identified by RepeatMasker;
- (iii) redundant TRs reported for the same locus using the TRDB Redundancy Elimination tool, and
- (iv) microsatellite TRs (pattern size ≤ 6).
- The final reference set contained 230,306 TRs (ref-TRs).

Subject data

Individual	Seq. tech	2nd gen reads	Coverage	Reads count	Read-TRs count
Watson	454	74,2 milj.	6,3x	2,9 milj	4,8 milj.
Khoisan	454	83,3 milj.	15,7x	15 milj.	59,6 milj.
CEPH and YRI trios	Illumina		68-81x	33,3-39,6 milj.	60,6-68,4 milj.

Alignments

- Spaced-seed indexing of the TR consensus patterns was used to determine candidate pairings of read-TRs to ref-TRs. Pairings were confirmed with three types of alignment:
- (i) longest common subsequence (LCS) comparison of consensus patterns;
- (ii) profile alignment of TR arrays;
- (iii) edit-distance alignment of flanking sequences.

Mappings

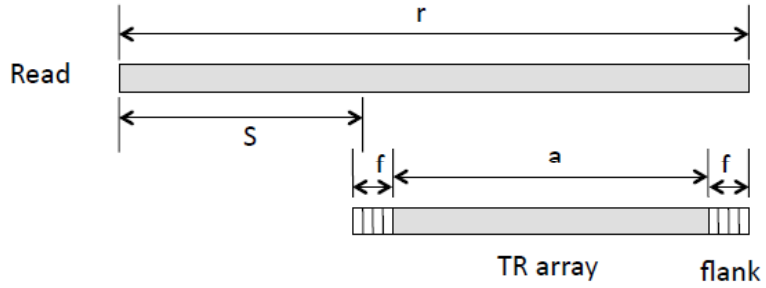
- Read-TR*/ref-TR pairs passing the alignment* filters were retained for mapping and were sorted into two lists, one by profile WS score and one by flank score. Within each list,
- (i) each read-TR picked its best scoring ref-TR (or all best scoring if ties occurred) and then
- (ii) each ref-TR picked its best read-TR if more than one occurred in the same read (or in the case of ties, the highest numbered read-TR).

Allele support and VNTR calling

- An allele containing *n* tandem copies was considered to have *support if the ref-TR* had at least two mapped reads with *n copies. Ref-TRs with* supported alleles were categorized as follows:
- 1. Single allele, same as reference—Not a VNTR.
- 2. Single allele, different from reference—An *inferred* VNTR assuming that the reference is correct, i.e. not an artifact.
- 3. Two alleles—An *observed VNTR.*

Performance

- We evaluated the performance of VNTRseek in five ways:
- (i) simulation studies to determine nominal accuracy for read mapping and VNTR calling,
- (ii) comparison of the expected number of mapped ref-TRs to those actually mapped with the Watson data,
- (iii) a BLAST search of unmapped ref-TRs from the Watson analysis against the Watson reads to determine how many mappings were missed,
- (iv) an analysis of indels deposited in dbSNP by the authors of the Watson sequencing paper in order to determine which are coincident with the VNTRs we report, and
- (v) analysis to determine the degree of observed Mendelian inheritance in VNTRs in two family trios from the 1000 Genomes Project.



Read Length (nt)	Minimum Flank Length					
	20			10		
	Coverage			Coverage		
	T	30	5	T	30	5
50	0	0	0	10	8 / 5	3 / 0
75	25	20 / 14	7 / 1	69	67 / 64	45 / 22
100	73	71 / 67	45 / 21	82	81 / 80	67 / 45
150	87	87 / 86	75 / 54	89	89 / 88	82 / 66
250	94	94 / 93	88 / 75	95	94/ 94	90 / 80

S is the length of the genomic region where a read of fixed length can start and span the TR array.

Expected fraction, in percent, of human reference TRs that will be spanned by at least one read and at least two reads for various read lengths in terms of genome coverage and minimum flanking sequence length. ”T” is the theoretical fraction spanned based only on reference array length. With Illumina 100 nt reads, a coverage of 30, and a minimum flank length of 20 nt, 71% of the references are expected to be scanned by at least one read and 67% are expected to be spanned by at least two reads. With 250 nt reads, under the same conditions, 94% of the references are expected to be spanned by at least one read and 93% by at least two reads.

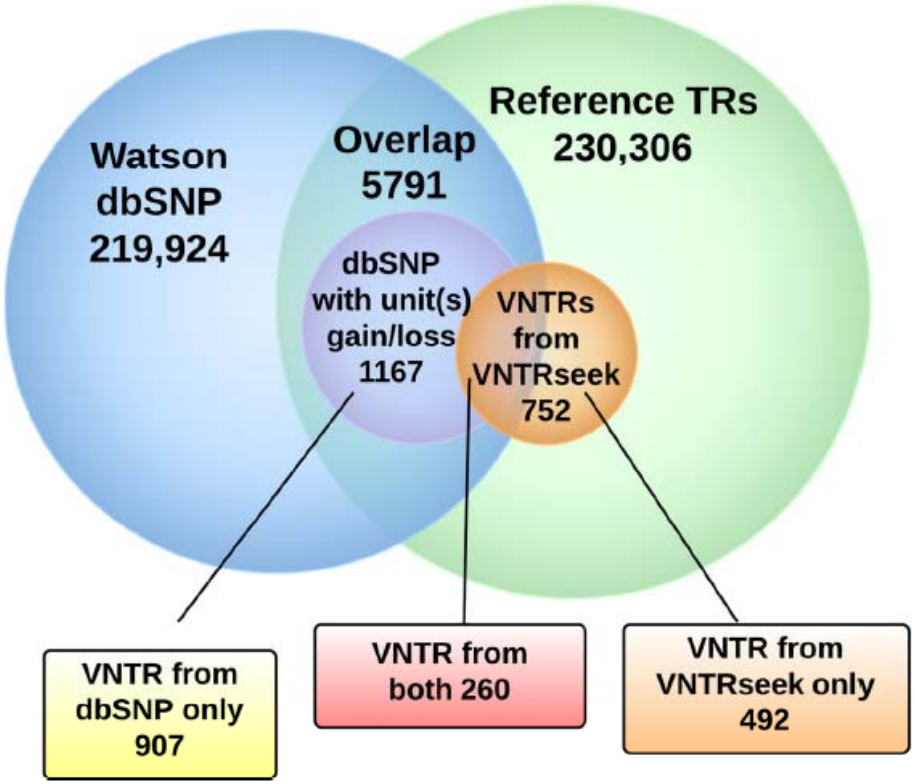
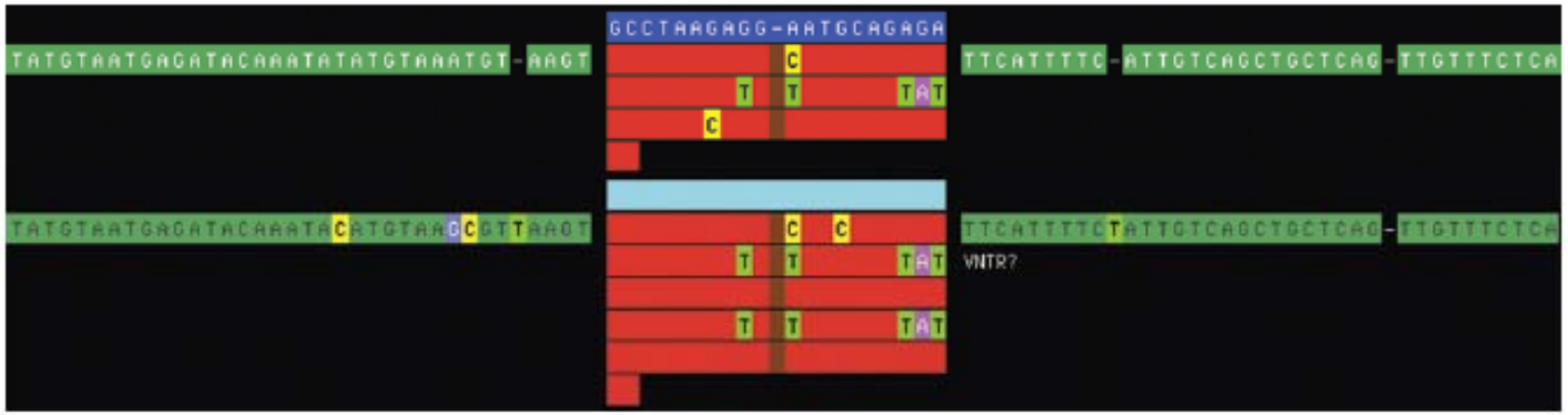


Figure 3. Comparison of VNTRseek detected VNTRs to Watson indels in dbSNP.

Program usage

- VNTRseek is a combination of C source code and Perl scripts which interact with a MySQL database created as part of the processing. Input is a set of FASTA or FASTQ files holding the subject reads and TR reference set data.
- Output consists of web pages which summarize the results of each program step;
- (i) sortable tables which list the characteristics of each mapped ref-TR, and each called VNTR;
- (ii) visualizations of read-TR to ref-TR alignments (e.g. Figure 1);
- (iii) Latex output of mapping statistics (Table 3 and Supplementary Tables 14 and 15; and
- (iv) two VCF format files, one for VNTRs detected and the other for all genotyped ref-TRs whether they are found to be variable or not. VCF files contain URL links to the alignment visualizations.



Watson read-TR (bottom) mapped to a ref-TR (top) from Chr 15:23,215,373. The number of copies, pattern motifs and motif order differ between the two, conditions which favor profile-based alignment. Blue—consensus pattern of ref-TR; red—multiple alignment of individual copies within a repeat, red matches the consensus, differences are shown explicitly; order of copies vertically matches order in the tandem array; light blue—consensus pattern of read-TR, here with no differences from the ref-TR consensus; green—flanking sequence with differences; not all available flanking sequence is shown due to page width limitation. Vertical gap due to insertion in another repeat (not shown).

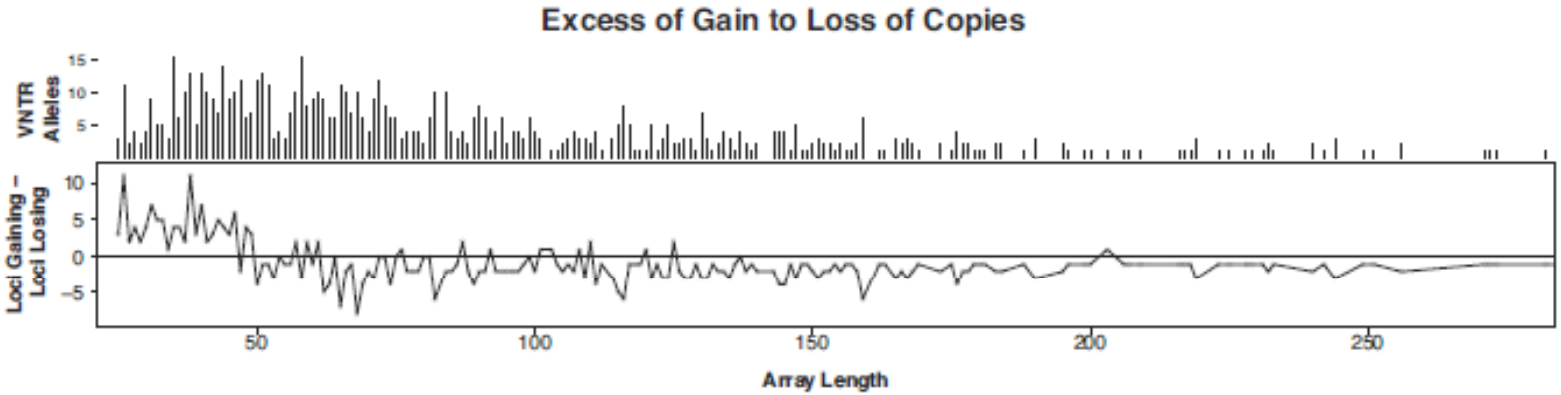
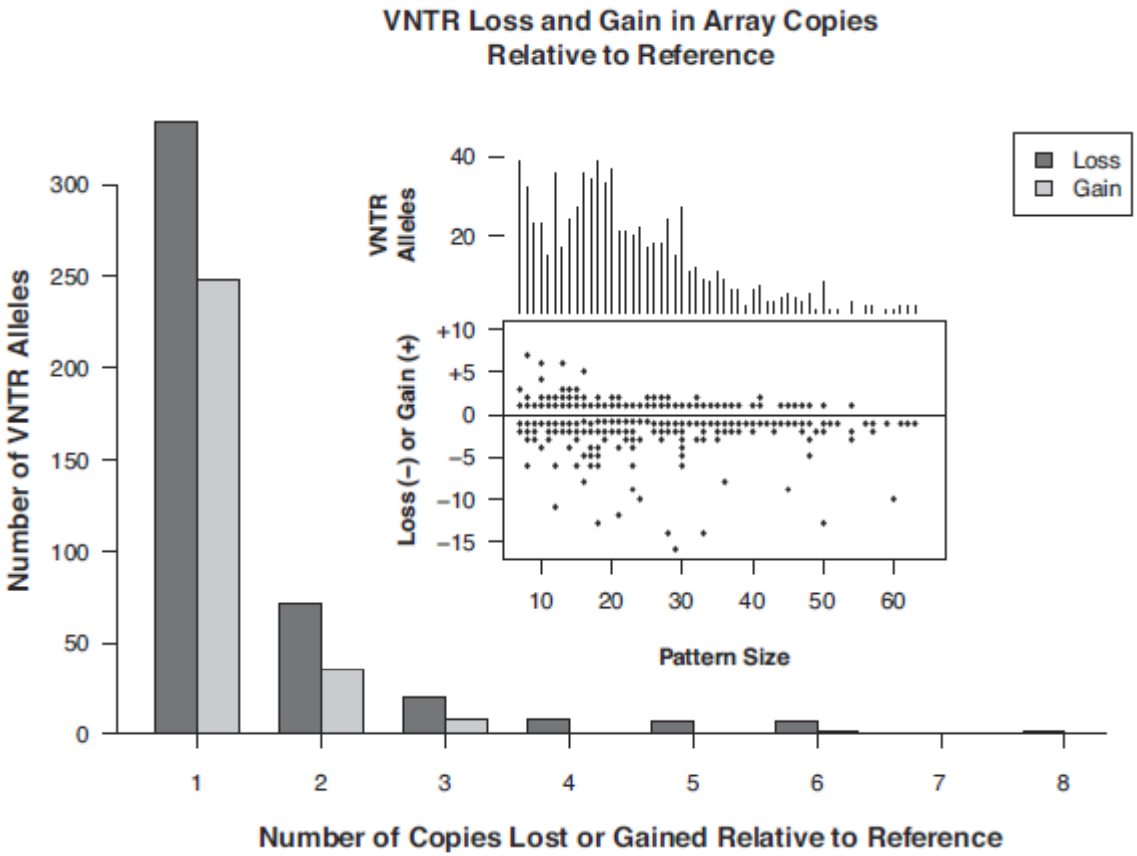


Figure 4. Watson VNTRs. Top: distribution of VNTRs by array length in the reference. Bottom: difference between number of loci that show copy gain and number that show copy loss, relative to the reference. More VNTRs show gain than loss at array lengths under 50 nt (trace above line). Abruptly, loss of copies becomes more common at longer array lengths. (Omitted from the graph are 15 VNTRs with reference array length longer than 282 nt.)



Watson VNTRs. Large graph: distribution of copy loss or gain relative to the reference. Singe copy change is by far the most commonly detected VNTR allele. Loss of copies overall is more frequent than gain.

Inset top: distribution of VNTRs by pattern size.

Inset bottom: number of copies gained or lost, by pattern size.

Note that frequency for each gain or loss is not shown, only occurrence.Data are for 759 variant alleles from752 reference TRs called as VNTRs. (Omitted from large graph are 13 VNTRs with loss*/gain greater than 8. Omitted from offset bottom are two VNTRs* with loss greater than 16 copies. Omitted from both insets is one VNTR with pattern size = 84 nt.)

Results

Individuals	Pattern size nt	Count
Watson	7-84	752
Khoisan	7-105	2572
Trios		2660-3822

Table 3. Watson VNTRseek results

A. Mapping					B. Mapped reference results				
	Total	After filters	Mapped	%	Number of reads mapped		At least one allele supported	By reference category	
Ref-TRs	1,188,939	230,306	169,463	74	≥ One	≥ Two		Singleton	Indist.
Read-TRs	13,080,867	4,826,849	532,960	11	169,463	131,855	131,707	164,080	5,383
Reads	74,198,831	2,925,732	525,748	18	100%	78%	78%	97%	3%

C. VNTR results						
		Alleles supported				
		One	Two			
		⋆	•	•	By reference category	
	Total	Diff	Same/Diff	Diff/Diff	Singleton	Indist.
	752	627	118	7	720	32
	100%	83%	16%	1%	96%	4%

A. Input data and data after filtering the reference set (for quality, common interspersed repeats, redundancy and pattern size) and the read set (for pattern size and sufficient flanking sequence); B. Counts and percentages of mapped references that were assigned at least one read, at least two reads, had at least one allele supported, and were either singleton or indistinguishable. An allele was *supported* if at least two reads were assigned to the ref-TR and they agreed on the pattern copy number. A ref-TR is indistinguishable if it is highly similar in both profile and flank alignments with another reference. All others are singletons. C. Counts and percentages of total VNTRS, number of alleles supported and reference category. For one allele supported, 'Diff' means the number of copies is different from the reference. These are inferred VNTRs because the reference is assumed to be correct, i.e. not an artifact. For two alleles supported, 'Same/Diff' means one allele has the same number of copies as the reference; 'Diff/Diff' means neither does; these are observed VNTRs because two alleles are observed. ⋆ Inferred VNTR • Observed VNTR

Table 2. Mendelian inheritance of VNTRs in 1000 Genomes trios

Utah family						Nigerian family					
Daughter	Mother	Father	Loci		Incon-	Daughter	Mother	Father	Loci		Inconsistent
NA12878	NA12892	NA12891	All	Diff	sistent	NA19240	NA19238	NA19239	All	Diff	
1241	1327	1402	274	20	1	1963	1979	1956	437	55	0

Shown are the number of VNTR loci for which two alleles were supported in each individual (sum of Same/Diff and Diff/Diff as in Table 3 C), number of loci in common for the trio (All), the subset of loci in common for which all three have different genotypes (Diff), and the number of loci inconsistent with Mendelian inheritance. The inconsistency count applies to all the loci in common, although the subset of loci for which all the family members are heterozygous AND have different genotypes provides the strongest test that VNTRseek is not systematically mis-assigning alleles from different loci to the same locus. Note that the VNTR loci with two alleles supported ranged from 46% to 52% of the total VNTRs in these individuals (data not shown). This contrasts with the lower coverage Watson genome in which 17% of loci exhibited two alleles.