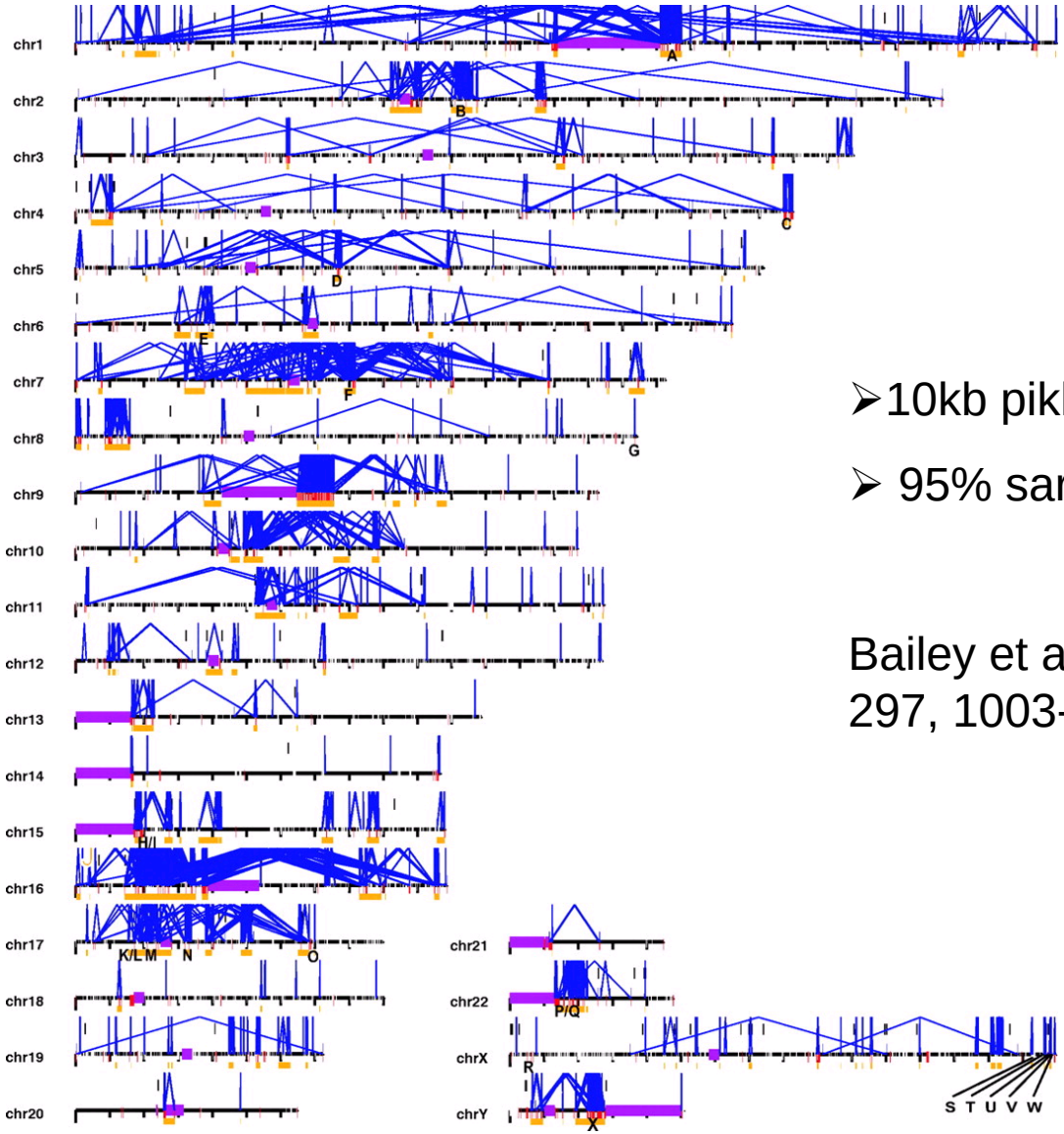
Complex SNP-related sequence variation in segmental genome duplications

Fredman et al. Nature Genetics 36, 861-866 põhjal

Tarmo Puurand 30.09.2004

Duplitseerunud piirkonnad inimesel



- 10kb pikkust
- 95% sarnasust

Bailey et al. Science 297, 1003-1007 (2002)

Uurimisobjekt

* uuriti 157 SNP-d, millest 107 olid polümorfsed

* 16 rootslannat ja 8 CHM'i (complete hydatidiform mole). CHM on günekoloogiline kõrvalekalle sagedusega 1:500-1:2000 rasedustest. CHMil on homosügootne genoom, mis on pärit mehelt.

DASH signaalide variandid

Table 1 Possible genotype patterns and allele ratios for several types of polymorphisms

	Polymorphism type	Position	Allele 1	Allele 2	Ordinary DNA samples		CHM samples	
					Genotype pattern	Allele ratios	Genotype pattern	Allele ratios
Example 1	Ordinary SNP	Site 1	A	G	A	2:0	A	2:0
		Site 2	none	none	AG	1:1	G	0:2
					G	0:2		
Example 2	PSV	Site 1	A	A	AG	2:2	AG	2:2
		Site 2	G	G				
Example 3	Duplicon SNP	Site 1	A	G	A	4:0	A	4:0
		Site 2	A	A	AG	3:1 or 2:2	AG	2:2
Example 4	MSV	Site 1	A	G	A	4:0	A	4:0
		Site 2	A	G	AG	3:1 or 2:2 or 1:3	AG	2:2
					G	0:4	G	0:4

Example 1: Ordinary SNP. Three genotype patterns (AA, AG, GG) are observed in ordinary DNA whereas only the two homozygous patterns (A, G) are observed in the CHM samples. Example 2: PSV. Only heterozygous genotype patterns are observed in ordinary and CHM samples. Example 3: Duplicon SNP. Only one of the homozygous patterns is observed. Example 4: MSV. Paralogous sites are variant, producing three genotype patterns in CHM samples.

DASH signaalid

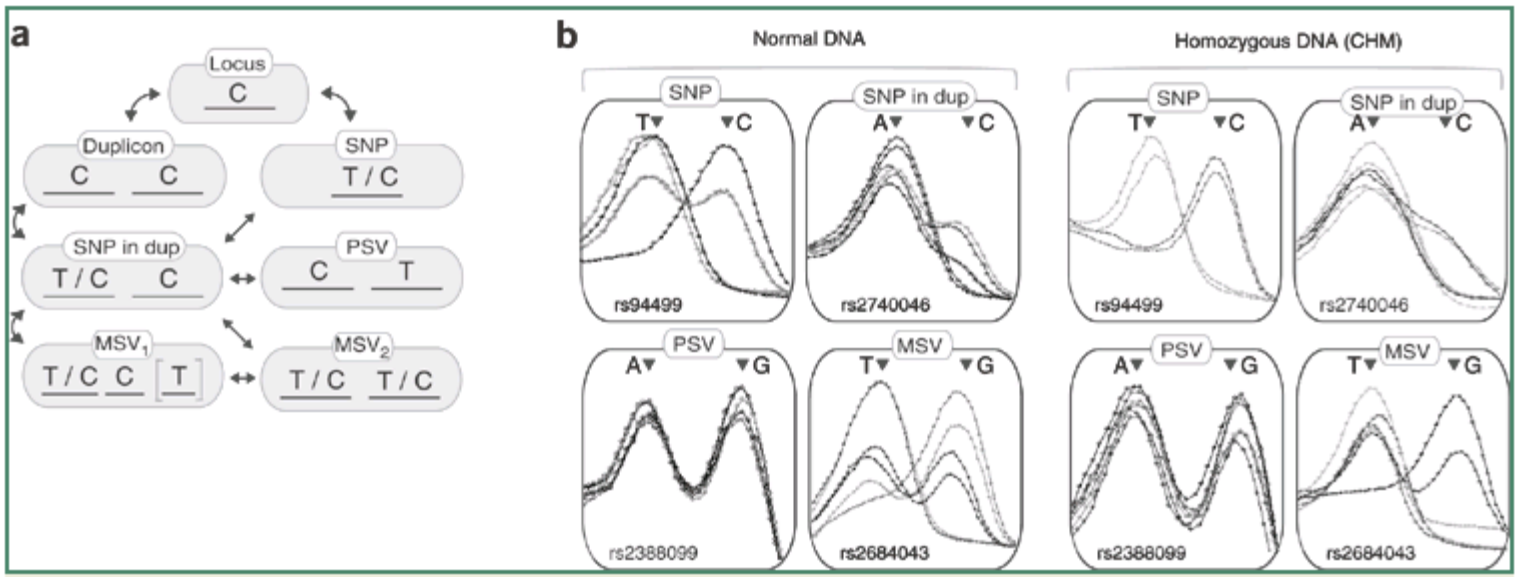


Figure 1. Genotyping patterns identifying evolutionary sequence states.
(**a**) Evolutionary sequence changes from a monomorphic base to a polymorphic MSV. Arrows depict processes such as mutation, fixation, duplication, deletion and gene conversion. Most events are reversible. (**b**) Representative DASH genotyping patterns observed in normal and CHM samples for the corresponding structures in **a**. Each line shows the negative derivative of the melting curve of a probe–target duplex for one DNA sample. The temperature on the *x* axis ranges from 45 to 75 °C. Peaks marked by arrowheads indicate the presence of each particular allele as marked, with peak heights indicating the relative amount of each allele present in the tested DNA. Dup, duplicon.

Uuritud regioonid

Table 1 Target regions

Region	WSSD	NCBI	Chrom	ChromStart (bp)	ChromEnd (bp)	Size (bp)	Name	Dispersal
A	Dup	Unique	1	85,402,915	85,427,399	24,485	–	Unknown
B	Dup	Unique	2	89,796,158	89,812,623	16,466	–	Unknown
C	Dup	Unique	16	18,167,513	18,191,332	23,820	–	Unknown
D	Dup	Unique	16	69,832,810	69,854,823	22,013	–	Unknown
E	Dup	Dup <98%	7	75,865,780	75,891,118	25,339	–	Intra
F	Dup	Dup <98%	9	85,988,721	86,012,093	23,373	–	Inter
G	Dup	Dup <98%	10	46,657,428	46,672,624	15,197	–	Intra
H	Dup	Dup <98%	11	88,972,901	88,996,892	23,992	–	Intra
I	Dup	Dup <98%	16	32,022,851	32,039,556	16,706	–	Inter
J	Dup	Dup >98%	8	7,161,589	7,293,710	132,121	8p23	Intra
K	Dup	Dup >98%	15	20,852,650	20,890,966	38,316	HERC2	Intra
L	Dup	Dup >98%	15	30,161,462	30,293,362	131,900	CHRNA7	Intra
M	Dup	Dup >98%	16	16,603,367	16,682,029	78,662	LCR16a	Intra
N	Dup	Dup >98%	17	44,072,366	44,126,506	54,140	MS	Intra
O	Unique	Dup >98%	1	57,845,958	57,856,075	10,117	–	Intra
P	Unique	Dup >98%	11	133,555,034	133,578,684	23,650	–	Intra
Q	Unique	Dup >98%	12	51,307,117	51,382,529	75,412	–	Intra
R	Unique	Unique	16	21,560,883	21,636,826	75,943	–	Unique
S	Unique	Unique	22	20,825,861	20,875,861	50,000	–	Unique
T	Unique	Unique	Various	Random validated SNPs	–	Unique		

Coordinates are from the July 2003 NCBI assembly. These comprise 17 duplicons and additional controls, covering a total of 1 Mb, taken from 12 different chromosomes. The target regions were grouped into four broad classes: A–D, domains that are present uniquely in the NCBI assembly but that are indicated to be duplicons by WSSD; E–I, duplicated domains in the NCBI assembly having 90–98% sequence similarity and WSSD support; J–N, duplicated domains in the assembly with >98% similarity and WSSD support; O–Q, duplicated domains in the assembly with >98% similarity but no WSSD support. Regions R–T are unique control sequences.

Genotüübid CHM ja tavalise DNA juhul

Table 2 Identification of genomic structures by analysis of DASH genotypes for CHMs and normal DNA

Genetic structure	Material	Number of alleles	Genotypes	Het. allele ratios	Constraints
SNP	DNA	1 or 2	M, H, m	Fixed ratio	–
	CHM	1 or 2	M, m	–	–
SNP in duplication	DNA	1 or 2	M, H	2 different ratios	One DNA H ratio must match CHM ratio
	CHM	1 or 2	M, H	Fixed ratio	
PSV	DNA	2	H	Fixed ratio	Same H ratio in DNA and CHM
	CHM	2	H	Fixed ratio	
MSV	DNA	1 or 2	M, H, m	Variable ratio	–
	CHM	1 or 2	M, H, m	Variable ratio	–

Samples are either homozygous with respect to one allele (M or m) or apparently heterozygous (H). Single-locus SNPs produce consistent homozygous and heterozygous signals in normal individuals, and no heterozygotes in CHMs. For a true SNP present in one copy of a duplicon (SNP in duplicon), one of the alleles is additionally represented at the other duplicon version(s), generating a heterozygote signal in one or more CHM. In normal DNA, these completely lack one homozygote pattern and generate two distinctive heterozygote patterns with different allele ratios. PSVs render heterozygote signals of identical allele ratios in all tested samples. MSVs produce two or more heterozygote types in CHMs, three or more heterozygote types in normal DNA, or both homozygotes combined with at least one type of heterozygote in CHMs.

Genotüübid

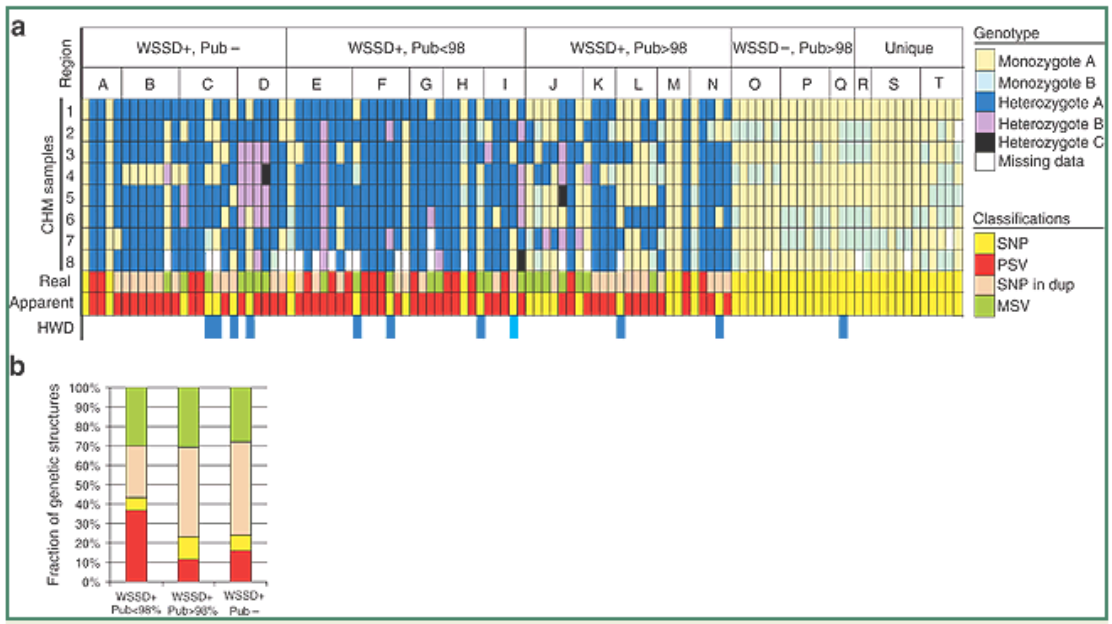


Figure 2. Summarized genotyping results.
(**a**) Marker results. Individual CHM data, along with a single line summary (Real) of marker classification based on data from the CHMs and the normal individuals. Purely qualitative genotyping methods used on normal DNA could misinterpret SNPs in duplicons as PSVs and MSVs as SNPs (Apparent), and only sometimes will HWE considerations resolve the latter (HWD). Dup, duplicon. Regions A–T are as described in Table 1. (**b**) Duplicon results. Whereas SNPs in duplicons are the largest category in the >98% similar (presumably recent) duplicons, PSVs are the biggest group in the <98% similar (presumably older) duplicons.MSVs have a similar representation in these two duplicon classes. PSVs can thus be viewed as a genetic remnant of duplicon sequence variation, representing the path duplicons follow towards sequence divergence and uniqueness.

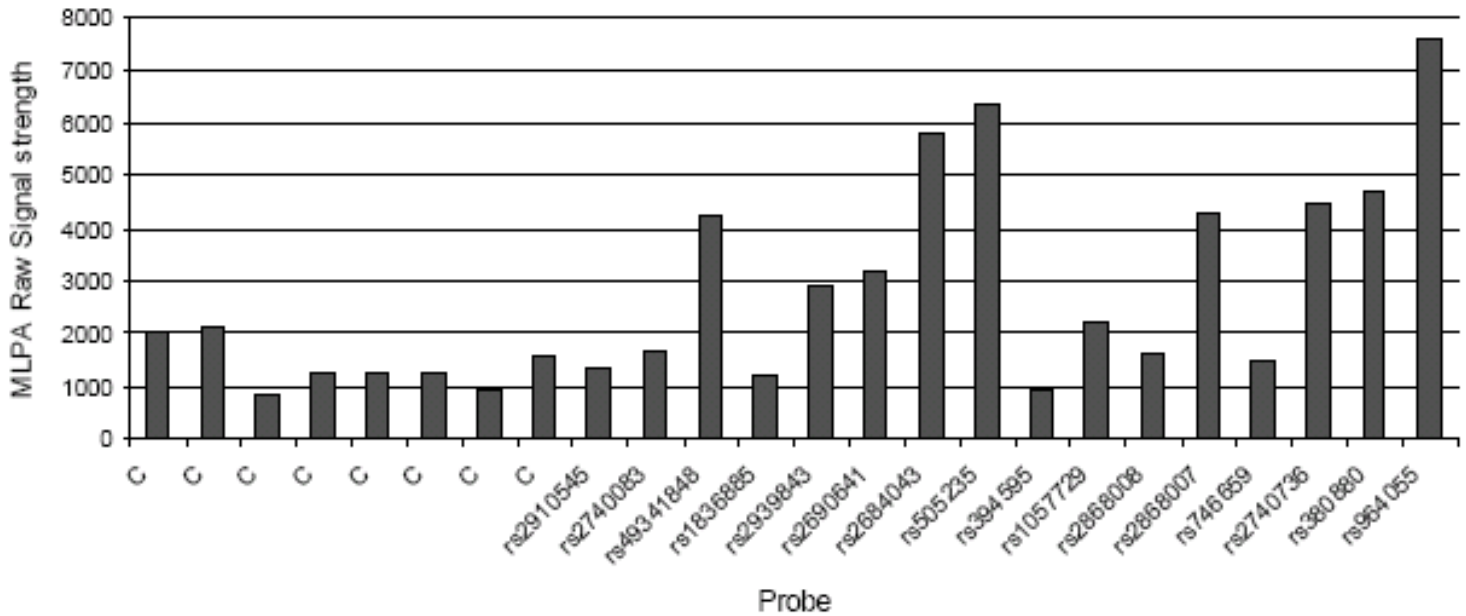
Geenikonversioon ja koopiaarvu varieeruvus

Table 3 MLPA analysis of 16 MSVs and two single-copy reference sequences

Nearest rs ID	Dup. region	Normalized MLPA ratios (triplicate means)								s.d.	Copy-number variation
		CHM1	CHM2	CHM3	CHM4	CHM5	CHM6	CHM7	CHM8		
–	Unique	–	0.87	1.12	1.11	0.85	0.93	1.03	0.92	0.11	No
–	Unique	0.93	0.89	1.1	1.09	0.93	0.98	1.03	1.06	0.08	No
394595	B	1.16	1.05	0.97	0.63	0.91	1.01	–	1.04	0.18	Yes
2910545	C	1.13	1.01	1.01	1.00	0.94	0.93	0.93	1.04	0.07	No
1057729	D	1.28	1.22	0.85	0.85	0.77	0.86	1.17	1.02	0.2	Yes
2868008	D	1.28	1.17	0.83	0.92	0.73	0.89	–	0.96	0.19	Yes
2868007	D	1.35	1.18	0.89	0.78	0.74	0.93	1.00	1.14	0.21	Yes
2690641	E	1.04	0.94	1.09	1.16	0.88	0.91	–	0.82	0.12	No
505235	F	1.03	1.02	1.04	0.98	0.96	0.96	0.94	1.06	0.04	No
1836885	H	1.01	0.98	0.94	0.96	1.11	0.93	0.92	1.16	0.09	No
964055	I	1.05	1.18	0.95	1.01	1.01	1.18	0.72	–	0.16	Yes
2939843	I	1.04	1.05	0.92	1.11	1.07	0.94	0.85	1.03	0.09	No
2684043	J	1.15	1.1	1.02	1.16	0.79	0.97	0.92	0.89	0.13	No
2740736	J	1.17	1.1	1.21	1.24	0.7	0.82	1.03	0.74	0.22	Yes
2740083	J	1.03	1.1	1.01	1.11	0.91	0.89	0.97	0.98	0.08	No
746659	J	1.37	1.3	1.00	–	0.73	0.83	0.95	0.78	0.25	Yes
296349	K	0.99	1.00	1.12	1.06	0.89	1.02	0.81	1.1	0.1	No
380880	K	0.75	1.26	1.05	0.86	1.00	1.08	0.93	1.08	0.15	Yes

Half of the MSV sequences show substantial evidence of copy-number variation. The remainder, including the two reference sequences, either have a fixed number of sequence copies or have a relative difference below the threshold of detection (s.d. < 0.15 across the eight CHMs).

Koopia arvud



Supplementary Figure 1. Average raw MLPA signal strength correlates with target sequence copy number[3]. For tested sequences in duplicons (labeled with the corresponding closest rsID) strong signals prevail, but stay well below 10-fold the signal strength observed for unique control sequences (labeled 'C'), indicating that there are typically 10 or fewer (haploid) copies of the sequences we are assaying.

Kokkuvõte

- Duplitseerunud alades on tõelisi SNP’i ~50%, PSV’i 22% ja MSV’i 23%.

Lingid

- http://www.dynametrix-ltd.com
- http://humanparalogy.gene.cwru.edu/