Data: 2958 1Mb non-overlapping windows over the whole genome (198 eliminated due to undefined variables).

The 7 variables non repetitive aln, interspersed repeat density, GC content, delta GC, recombination rate, exon density, and snp density (from TSC) were standardized (minus mean, divided by st. dev.) to eliminate overall location and variation scale differences. The linear regression of aln (nr) on the remaining 6 variables provides the **direction of strongest linear explanation** of the response in the predictor space. To it, we can associate a **share of explained (standardized) variability of the response** -- R2 of the regression. Below are the results, overall and by chromosome:

Least square coefficients from linear regression; response = aln (nr)

	intercept	intersp rep	GC	delta GC	recomb	exon	snp (tsc)	share of var (R2)
Overall	**0.00000**	**-0.320666**	**-0.171324**	**-0.115630**	**0.011233**	**0.25750**	**-0.129328**	**0.144**
chr1	0.18491	-0.393928	-0.088810	-0.021954	-0.013012	0.18351	-0.133822	0.124
chr2	-0.07559	-0.542904	-0.161485	-0.536636	-0.052700	0.26743	-0.052980	0.417
chr3	0.00038	-0.419904	0.177110	-0.112827	0.093195	0.21119	0.146186	0.219
chr4	-0.18065	-0.397484	0.181114	-0.341840	0.011845	0.09516	-0.043833	0.259
chr5	0.10994	-0.414866	0.147021	-0.476363	-0.131950	0.10413	-0.261286	0.339
chr6	-0.27811	-0.373218	-0.230734	-0.236830	-0.029376	0.12787	-0.126814	0.117
chr7	0.08270	-0.384894	-0.408089	-0.395759	0.037767	0.50500	-0.058995	0.550
chr8	-0.17661	-0.230123	-0.221628	-0.301019	0.061636	0.16972	-0.136349	0.135
chr9	0.16377	-0.451124	-0.360056	-0.052870	0.014257	0.42290	-0.063366	0.174
chr10	0.08443	-0.242507	-0.138261	-0.035262	0.003205	0.34695	-0.065072	0.123
chr11	0.44518	-0.336092	0.324117	-0.441767	0.271057	0.05652	-0.216063	0.370
chr12	-0.15456	-0.574566	-0.422803	-0.024646	-0.015177	0.39957	-0.231253	0.368
chr13	-0.84024	-0.600242	-0.132399	-0.387117	-0.093475	0.09200	-0.026458	0.248
chr14	0.24038	-0.494908	-0.185609	-0.350425	0.038005	0.28379	-0.016953	0.341
chr15	0.36798	-0.596453	-0.216380	-0.030469	-0.039083	0.14490	-0.180727	0.488
chr16	0.36501	-0.205788	-0.603910	-0.187575	-0.011370	0.45333	0.015692	0.371
chr17	0.22762	-0.146081	-0.067281	-0.274155	-0.004111	0.32023	0.004145	0.264
chr18	-0.50385	-0.399859	-0.130809	-0.373458	-0.078899	-0.16506	-0.154097	0.200
chr19	-0.21057	-0.366291	-0.597422	0.166754	-0.094962	0.33467	-0.388348	0.342
chr20	0.57474	-0.279829	-0.972352	0.177166	0.062638	0.50828	-0.117798	0.451
chr21	-1.03256	-0.510089	-0.279723	-0.650333	-0.072513	0.51646	0.020117	0.588
chr22	-0.29449	0.044596	-0.350202	0.166660	0.008008	0.44078	-0.137411	0.622
chrX	-0.43225	-0.246934	0.118540	0.026844	-0.234549	0.15377	-0.617391	0.164
chrY	0.63676	-0.261864	-0.910238	0.152440	0.000000	1.28507	0.915282	0.416

These results can be better expressed transforming the coefficients (neglecting the intercept) into norm one vectors in a 6D space, as reported below (so these are objects of the same nature of the first PC's in the other analysis):

Norm one least square vectors from linear regression; response = aln (nr)

	intersp rep	GC	delta GC	recomb	exon	snp (tsc)	share of var (R2)
Overall	**-0.670520**	**-0.358242**	**-0.241785**	**0.023488**	**0.538445**	**-0.270427**	**0.144**
chr1	-0.848972	-0.191399	-0.047315	-0.028042	0.395487	-0.288406	0.124
chr2	-0.655528	-0.194984	-0.647960	-0.063632	0.322910	-0.063971	0.417
chr3	-0.773009	0.326045	-0.207705	0.171565	0.388783	0.269117	0.219
chr4	-0.704015	0.320785	-0.605461	0.020979	0.168551	-0.077636	0.259
chr5	-0.576886	0.204437	-0.662398	-0.183480	0.144793	-0.363327	0.339
chr6	-0.702918	-0.434563	-0.446044	-0.055326	0.240835	-0.238840	0.117
chr7	-0.450103	-0.477228	-0.462809	0.044165	0.590558	-0.068990	0.550
chr8	-0.465980	-0.448780	-0.609540	0.124807	0.343677	-0.276095	0.135
chr9	-0.626189	-0.499781	-0.073387	0.019790	0.587020	-0.087956	0.174
chr10	-0.537200	-0.306275	-0.078111	0.007100	0.768558	-0.144148	0.123
chr11	-0.458845	0.442497	-0.603117	0.370057	0.077163	-0.294978	0.370
chr12	-0.675793	-0.497292	-0.028988	-0.017851	0.469961	-0.271995	0.368
chr13	-0.812637	-0.179248	-0.524098	-0.126551	0.124555	-0.035820	0.248
chr14	-0.711050	-0.266671	-0.503467	0.054603	0.407726	-0.024357	0.341
chr15	-0.880677	-0.319490	-0.044988	-0.057706	0.213945	-0.266847	0.488
chr16	-0.255617	-0.750142	-0.232995	-0.014123	0.563099	0.019491	0.371
chr17	-0.323738	-0.149104	-0.607569	-0.009111	0.709686	0.009186	0.264
chr18	-0.654115	-0.213986	-0.610928	-0.129068	-0.270014	-0.252081	0.200
chr19	-0.411920	-0.671843	0.187527	-0.106792	0.376363	-0.436725	0.342
chr20	-0.242524	-0.842721	0.153547	0.054287	0.440518	-0.102093	0.451
chr21	-0.501690	-0.275117	-0.639625	-0.071319	0.507956	0.019786	0.588
chr22	0.073752	-0.579150	0.275614	0.013244	0.728950	-0.227245	0.622
chrX	-0.337419	0.161977	0.036681	-0.320496	0.210117	-0.843623	0.164
chrY	-0.141818	-0.492959	0.082557	0.000000	0.695956	0.495690	0.416

Overall, the linear explanation of aln (nr) in terms of these 6 variables is not very strong (overall R2 ~ 14%). Also, the coefficients for interp rep, snp (neg) and exon (pos) are as expected, that for recomb is pos but very small, but those for GC ad deltaGC are more ambiguous (sizeable and neg?). Note that these coefficients express sign and size of effects on the response, “adjusted for” the presence of the other predictors in the pool (this is true also for the coefficients computed within each of the chromosomes).

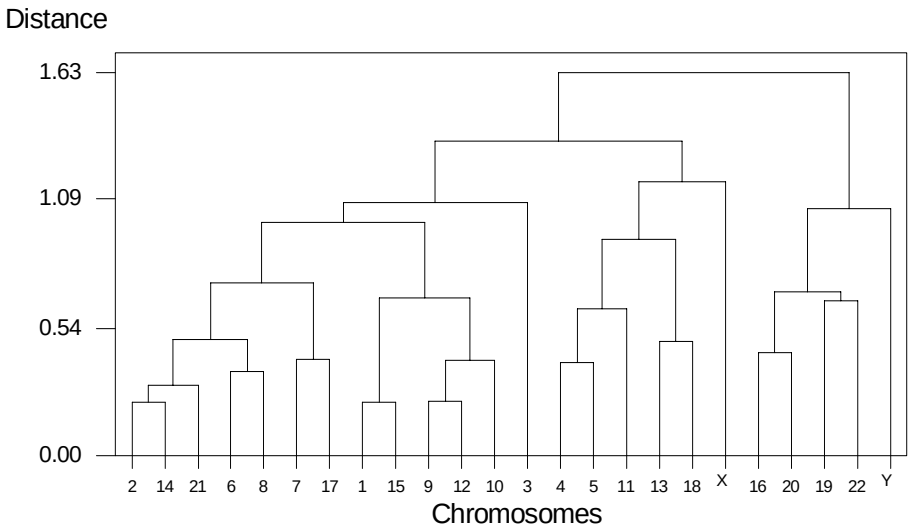
Again, both the share of explained variability (R2) and the coefficients size and sign patterns vary dramatically when the analysis is repeated within chromosomes. For example, chromosomes 7, 21 and 22 have a much larger R2 (above .5). Chromosome 22 presents almost no effect of intersp rep, chromosome 18 presents a negative effect of exon, etc.

Next, I attempted clustering of chromosomes on the basis of

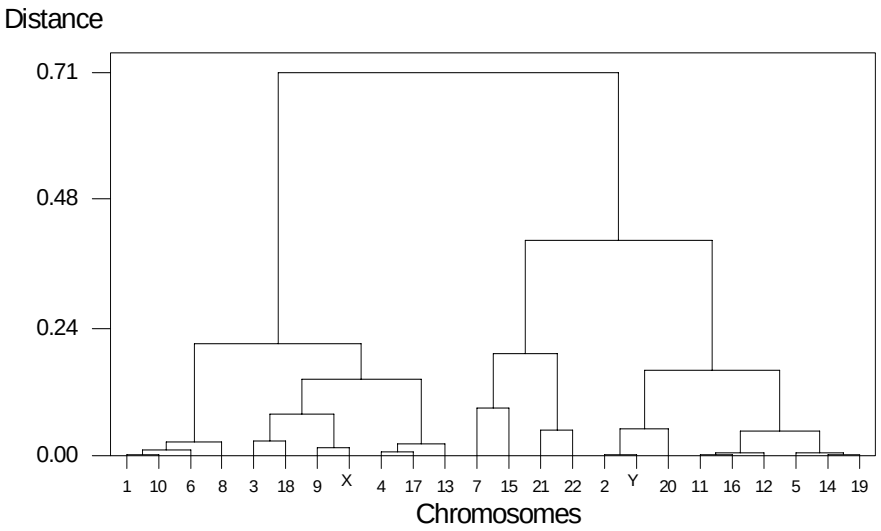
1. The distances among their norm one least square vectors (hierarchical agglomeration)

2. The differences among their R2 (hierarchical agglomeration)

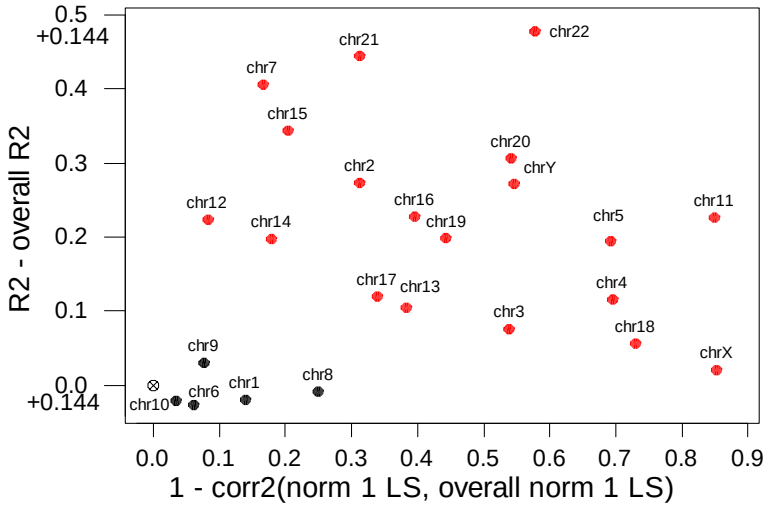
3. Their resemblance of the overall behavior, in terms of both norm one least square vectors, and R2 (visual inspection of a 2D plot).



Agglomeration (complete linkage) of chromosomes based on euclidean distance between their norm 1 LS's – in a 6D space



Agglomeration (complete linkage) of chromosomes based on difference between their R2.



Chromosomes located in terms of (Horiz) discrepancy between their norm 1 LS and the overall one, as measured by 1 minus the squared correlation; and (Vert) difference between their R2 and the overall one, which is 0.144. Here (0,0) (circled black cross) is the position of the overall norm 1 LS. The grouping isn't as clear as in the case of the PC analysis, but there is an obvious separation between chromosomes that are/ are not similar to the overall behavior (color-coding above)