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 leading principal component computed on these standardized variables represents the **direction of strongest linear association** among them. To it, we can associate a **share of explained (standardized) variability**. Below are the results, overall and by chromosome:

Leading (first) Principal Component

	aln(nr)	intersp rep	GC	delta GC	recomb	exon	snp (tsc)	share of var
Overall	**-0.054457**	**0.005378**	**-0.610328**	**-0.445098**	**-0.123844**	**-0.577668**	**0.278126**	**0.294**
chr1	-0.147574	0.071831	-0.624086	-0.478953	-0.093615	-0.520673	0.272616	0.334
chr2	-0.821878	0.396684	0.098507	0.361281	0.087430	-0.138753	0.005892	0.295
chr3	-0.061819	0.262956	0.499856	0.290577	-0.219587	0.547822	-0.494383	0.290
chr4	-0.497121	0.180134	-0.322456	0.333500	-0.389765	-0.225375	0.550011	0.296
chr5	-0.223617	0.075552	-0.345634	-0.152342	-0.057720	-0.864217	0.226745	0.323
chr6	0.008047	0.156750	-0.566541	-0.076166	-0.277537	-0.755625	-0.024472	0.297
chr7	-0.515598	0.311729	0.608382	0.427979	0.031298	0.242755	-0.154209	0.344
chr8	-0.076815	-0.073595	-0.422375	-0.086312	0.046291	-0.371680	0.813968	0.271
chr9	-0.026791	0.167607	-0.641997	-0.432936	-0.225487	-0.393365	0.407451	0.330
chr10	-0.109079	0.481171	-0.459518	-0.599292	-0.404202	-0.129776	-0.077766	0.336
chr11	-0.350917	0.152451	-0.582481	-0.296917	-0.131492	-0.523237	0.367567	0.408
chr12	-0.154699	0.043737	-0.512100	-0.429913	-0.109901	-0.691622	0.191479	0.374
chr13	-0.642614	0.363986	-0.321736	0.378451	-0.403872	-0.185792	-0.100952	0.324
chr14	-0.307376	0.373847	-0.420422	-0.013799	-0.689877	-0.326180	-0.080560	0.316
chr15	0.074816	0.184342	0.174249	0.047896	-0.483512	0.461861	-0.693300	0.333
chr16	0.040915	0.099669	-0.437256	-0.478837	0.181172	-0.660026	0.315367	0.457
chr17	-0.129235	0.099233	-0.465202	-0.397433	-0.111711	-0.737163	0.207839	0.392
chr18	0.023343	-0.024181	0.099565	-0.281712	0.951383	0.047831	0.046668	0.292
chr19	-0.088307	0.156974	-0.446713	-0.420522	-0.291890	-0.711061	0.019006	0.489
chr20	0.036025	0.196577	-0.435819	-0.563084	0.433686	-0.482916	0.178232	0.376
chr21	0.124548	0.126531	-0.724779	-0.325699	0.140249	-0.459197	0.326439	0.388
chr22	-0.319491	-0.070488	-0.181284	-0.244585	0.483759	-0.579779	0.479692	0.374
chrX	-0.420770	0.813276	-0.222536	-0.061285	-0.110021	-0.305047	-0.055661	0.409
chrY	-0.133410	0.972739	-0.115785	-0.141735	0.000000	0.048895	-0.009679	0.675

Overall, the linear association among these 7 variables is not very strong (overall share of explained variability ~ 30%). Also, coefficients for aln(nr), GC, delta GC and exon share a sign (negative), and those for intersp rep and snp (tsc) share the opposite (positive). The coefficient for recomb, which has the same sign as those for alignment, GC and exons, is more ambiguous.

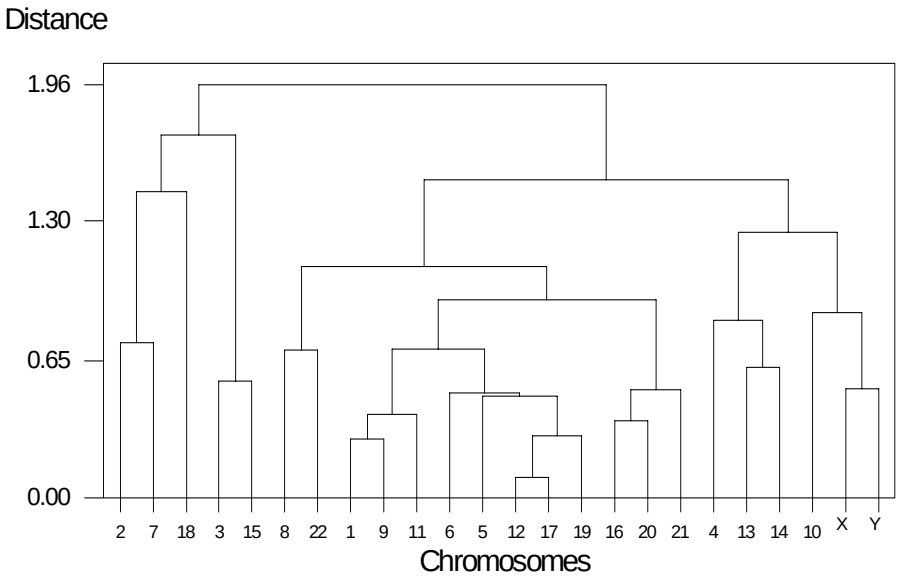
However, both the share of explained variability and the coefficients size and sign patterns vary dramatically when the analysis is repeated within chromosomes. For example, chromosomes Y, 19 and 16 have much stronger linear associations among the 7 variables. As another extreme example, chromosomes 3, 7 and 15 have same sign and sizeable coefficients for intersp rep and exon.

Next, I attempted clustering of chromosomes on the basis of

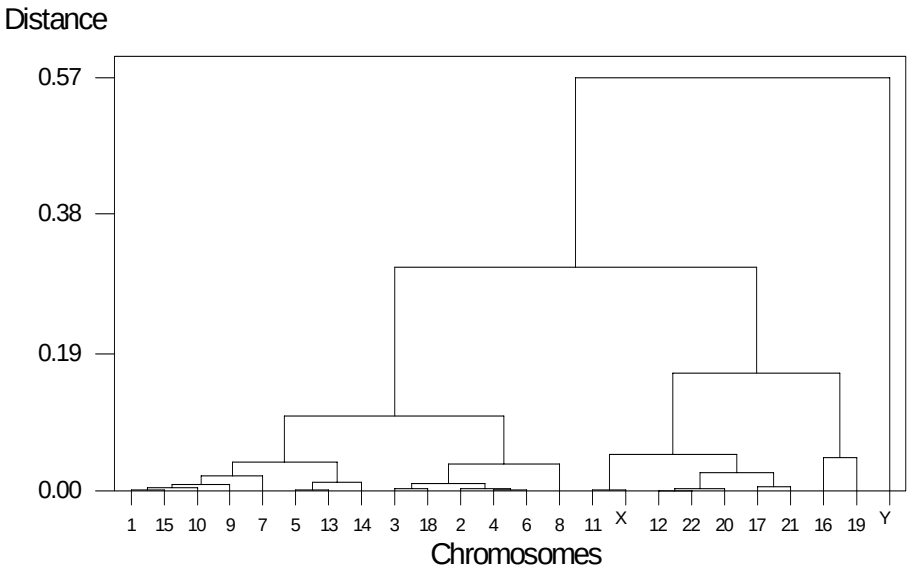
1. The distances among their first PC's (hierarchical agglomeration)

2. The differences among their first PC shares of explained variability (hierarchical agglomeration)

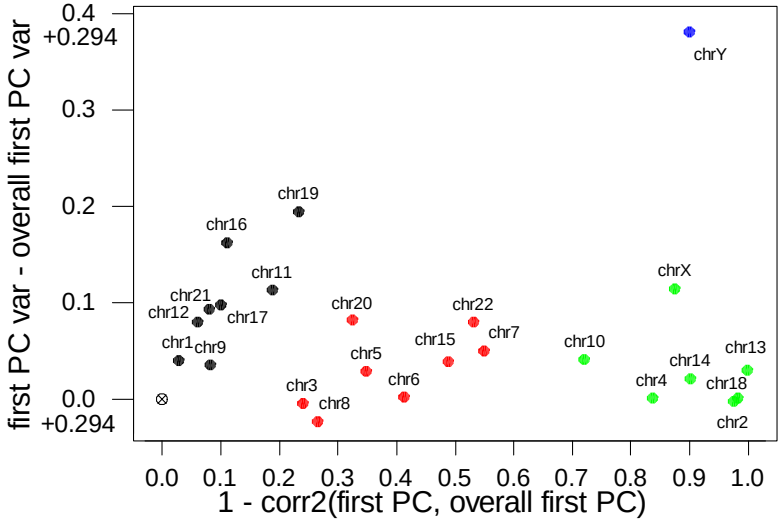
3. Their resemblance of the overall behavior, in terms of both first PC, and first PC share of explained variability (visual inspection of a 2D plot).



Agglomeration (complete linkage) of chromosomes based on euclidean distance between their first PC's – norm one vectors in a 7D space



Agglomeration (complete linkage) of chromosomes based on difference between their first PC shares of explained variability.



Chromosomes located in terms of (Horiz) discrepancy between their first PC and the overall one, as measured by 1 minus the squared correlation; and (Vert) difference between their first PC share of explained variability and the overall one, which is 0.294. Here (0,0) (circled black cross) is the position of the overall first PC. There seems to be three natural groups, and chrY stands alone (color-coding above)