Adjusting for Population Stratification with Principal Components and Sequencing Data

Wei Pan

[1]Division of Biostatistics, School of Public Health

University of Minnesota

JSM, August 2013

Joint work with Yiwei Zhang and Xiaotong Shen.

Problem

- Population stratification: GWAS with CVs
- Now seq data with RVs
- Price et al (2010, Nat Rev Genet):
  "exome resequencing projects will aim to identify genes in which individuals with extreme phenotypes have an aggregate excess or deficiency of rare nonsynonymous variants [42]. Differences in allele frequency spectrum across ancestral populations make stratification a potential concern,"

  "Finally, the advent of whole-exome or whole-genome resequencing raises the question of whether rare variants can be used to infer genetic ancestry with greater precision, perhaps using different methods than the methods currently applied to common variants."

- Henn et al (2010, Hum Mol Genet): *"Rare variants are likely to have recently arisen and segregate between populations and are informative markers of ancestry"*
- Some existing studies using genotyping or simulated data, or Zhang and Pan (2013, Genet Epi) used seq data, but not in a fine scale, not much use of RVs, ...

Data

- 1000 Genomes Project data (1000 Genomes Project Consortium 2010):
  low-coverage seq data, released Aug 2010.
- 283 European samples: 90 CEU, 92 TSI, 43 GBR, 36 FIN, 17 MXL, 5 PUR.
- 174 African samples: 78 YRI, 67 LWK, 24 ASW, 5 PUR2.
- 6,227,535 CVs (MAF $> 5\%$), 1,849,693 LFVs (MAF $1 - 5\%$), 854,921 RVs (MAF $< 1\%$);
- After pruned by Plink (Purcell et al 2007): sliding window of size 50, shifted by 5 and $r^2 \leq .05$.
  880,426 All, 148,324 CVs, 384,751 LFVs, 328,713 RVs.
- We took a random subset of 10,000 from each pruned set.

Methods

- X: $n \times p$ standardized genotype score matrix
- PCA: use $A = XX'$ as similarity matrix, PCs are in the directions of its eigen-vectors;
- Spectral dimension reduction (SDR) Lee et al 2009): use a normalized similarity matrix $W = (W_{ij})$ with

$$W_{ij} = \sqrt{X'_{i.}X_{j.}} \text{ if } X'_{i.}X_{j.} \geq 0; = 0 \text{ o/w .}$$

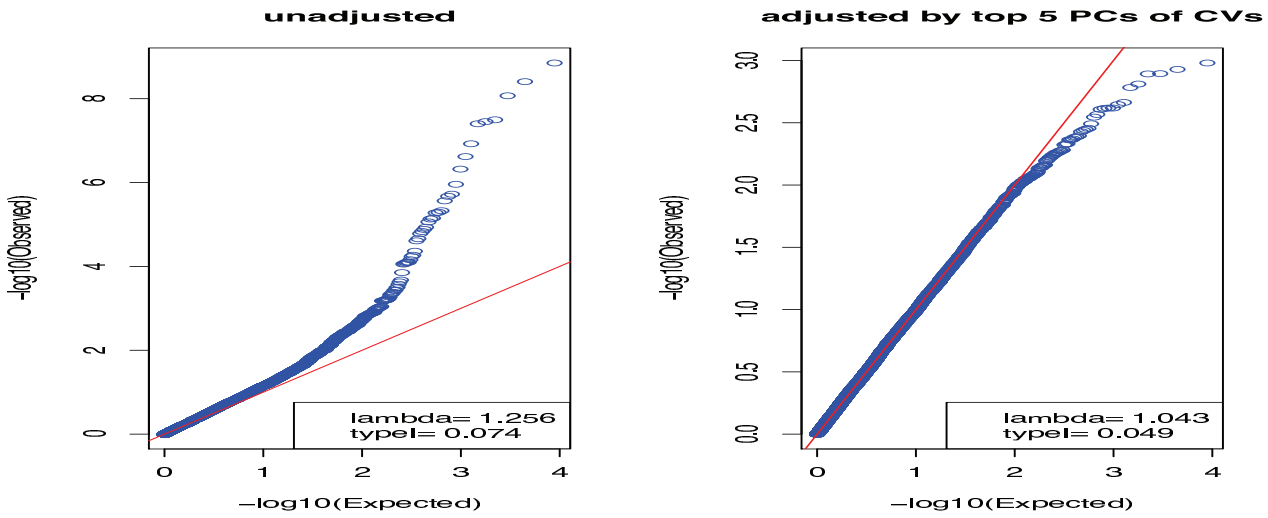
- Use the Tracy-Widom test (Patterson et al 2006) or a heuristic method (:ee et al 2008), 10-30 "significant" eigen-vectors.
- Association testing: on pruned variants from chr 1&2,
  10,848 CVs (MAF$>$ 0.2), Score test;
  61,279 LFVs, 50,476 RVs; sliding windows of size 20, moving step 5; T1 and Fp tests in SCORE-Seq (Lin and Tang 2011).

- Assigned each subgroup to “Case” or “Control” group.
  almost balanced; 3 diff ways; also tried other ways.

Results

- Is it necessary to adjust for PS?

Figure 1: Q-Q plots of the p-values for the score test in the simulated case-control study where CEUs are "cases" and GBRs are "controls".



- PC plots:

Figure 2: The top 2 PCs of SDR.

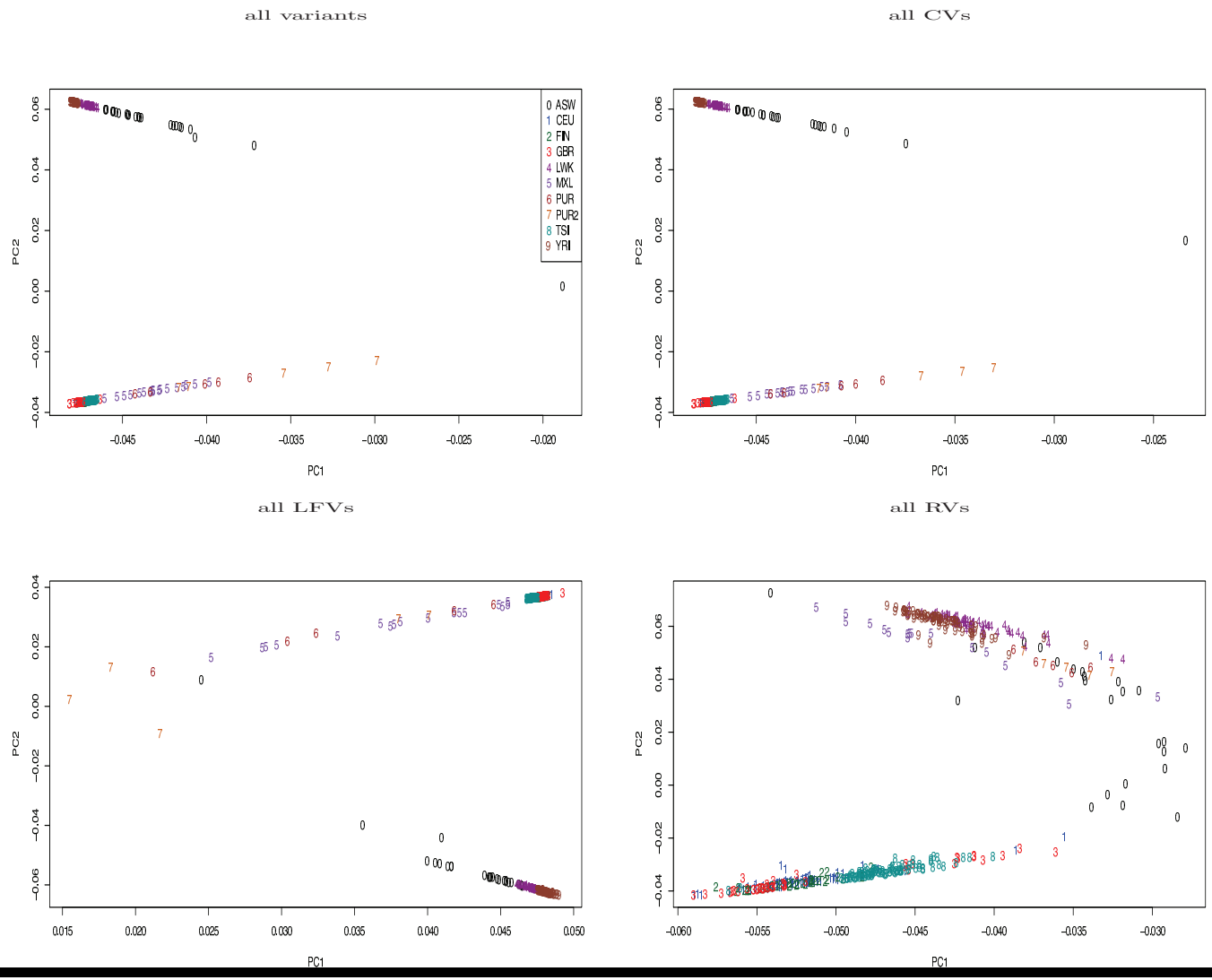
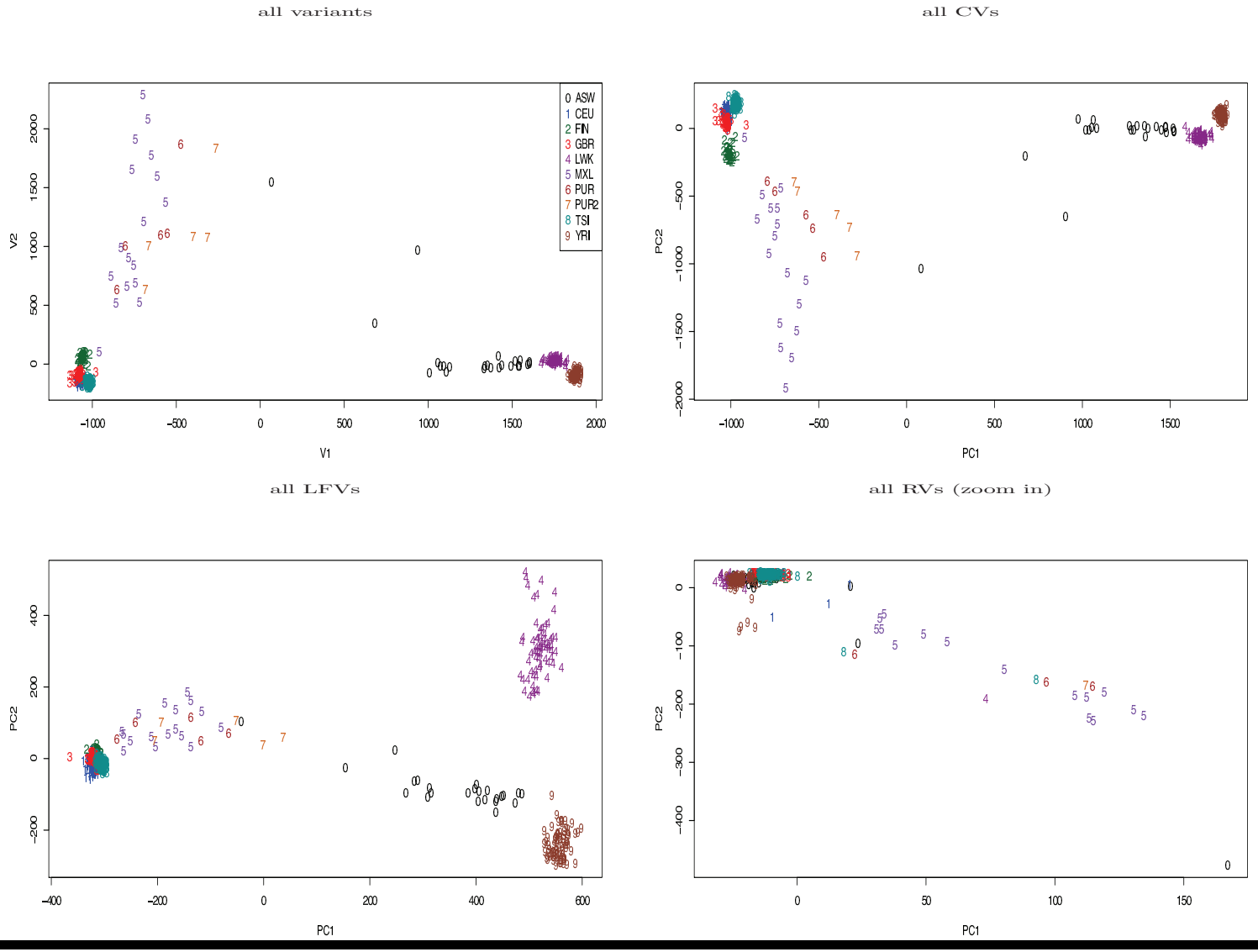


Figure 3: The top 2 PCs of PCA.



- Testing on CVs: inflation factor λ

	Method	#PCs	all	CVs	LFVs	RVs
w/o pruning	SDR	0	23.755	23.755	23.755	23.755
		10	1.126	1.136	1.302	1.349
		15	1.164	1.098	1.319	1.391
		20	1.174	1.150	1.312	1.434
	PCA	10	1.318	1.340	1.325	1.273
		15	1.308	1.167	1.317	1.347
		20	1.360	1.210	1.368	1.443
		#PCs	all	CVs	LFVs	RVs
10000 pruned	SDR	0	23.755	23.755	23.755	23.755
		10	1.307	1.410	1.304	1.414
		15	1.383	1.355	1.360	1.426
		20	1.290	1.233	1.440	1.427
	PCA	10	1.361	1.350	1.283	1.323
		15	1.297	1.276	1.327	1.324
		20	1.340	1.241	1.375	1.357

- Testing on RVs: inflation factor λ

			w/o pruning				10000 pruned			
Method	#PCs	Test	all	CV	LFVs	RVs	all	CVs	LFVs	RVs
SDR	0	T1	6.114	6.114	6.114	6.114	6.114	6.114	6.114	6.114
		Fp	5.665	5.665	5.665	5.665	5.665	5.665	5.665	5.665
	10	T1	1.009	1.025	1.122	1.246	1.237	1.098	1.309	1.669
		Fp	1.004	1.006	1.137	1.253	1.222	1.107	1.308	1.705
	25	T1	1.463	1.206	1.214	1.431	1.257	1.147	1.228	1.693
		Fp	1.459	1.198	1.202	1.440	1.212	1.170	1.238	1.701
PCA	10	T1	1.699	1.708	1.854	1.530	2.002	1.610	1.311	1.454
		Fp	1.774	1.763	1.892	1.556	2.027	1.690	1.339	1.479
	25	T1	1.191	1.482	1.342	1.199	1.350	1.781	1.308	1.254
		Fp	1.218	1.487	1.368	1.199	1.343	1.786	1.305	1.276

			w/o pruning				with pruning			
Method	#PCs	Test	all	CV	LFVs	RVs	all	CVs	LFVs	RVs
SDR	0	T1	6.114	6.114	6.114	6.114	6.114	6.114	6.114	6.114
		Fp	5.665	5.665	5.665	5.665	5.665	5.665	5.665	5.665
	10	T1	1.009	1.025	1.122	1.246	1.161	0.988	1.297	1.470
		Fp	1.004	1.006	1.137	1.253	1.187	0.986	1.289	1.472
	25	T1	1.463	1.206	1.214	1.431	1.293	1.336	1.841	1.707
		Fp	1.459	1.198	1.202	1.440	1.284	1.334	1.844	1.731
PCA	10	T1	1.699	1.708	1.854	1.530	1.491	1.656	1.314	1.205
		Fp	1.774	1.763	1.892	1.556	1.525	1.687	1.347	1.192
	25	T1	1.191	1.482	1.342	1.199	1.230	1.624	1.456	1.213
		Fp	1.218	1.487	1.368	1.199	1.241	1.623	1.465	1.191

- why not RVs?
  - Fst statistics: those based on CVs and LFVs were close to each other; those based on RVs were much smaller.
  - Testing subgroup-specific variants: significant at 5% level, 82.53% CVs vs 74.28% LFVs vs only 25.62% RVs
  - due to low-coverage?
- Finally, for a quantitative trait in a real seq dataset (60x coverage),
  unadjusted $\lambda = 1.14$,
  using CVs $\lambda = 1.068$,
  using RVs $\lambda = 1.121$.

Acknowledgment: This research was supported by NIH.

You can download our papers from
http://sph.umn.edu/ex/biostatistics/techreports.php?

Thank you!