1KG project BAMs : Sample identity verification in ILLUMINA and LS454 low coverage data

Aug 24th, 2010

Hyun Min Kang

Identity Verification Metrics

- P_{IBD}
  - estimated % of reads in IBD w/ genotyped sample
  - %Contam ≈ $1-P_{IBD}$
- %altBase@refGeno (altB@refG)
  - Fraction of non-ref bases at reference genotype
- Excessive Heterozygosity
  - Excessive fraction of heterozygous sites than expected by population allele frequency
  - Does not require genotype data
- Criteria for flagging contamination
  - ([P_{IBD}<1] AND [altB@refG>0.005]) OR EX_HET > 1.1

Summary of low coverage data (1KG Project July'10 BAMs)

- HapMap3 genotyped samples (462)
  - Contaminated samples
    - All lanes : 7 samples
    - Some lanes : 6 samples, 19 lanes
  - Swapped lanes : 7 samples, 15 lanes
- Ungenotyped samples
  - 8/67 samples with excessive heterozygosity

Contaminated samples – all lanes

SAMPLE ID	POP	CENTER	PLAT-FORM	P_{IBD}	%altBase @REFGeno	DEPTH @REFGeno	EX_HET
NA06985	CEU	SC	ILLUM	0.33	0.117	1.42	1.810
NA11881	CEU	SC	ILLUM	0.86	0.020	2.86	1.221
NA12043	CEU	454MSC	LS454	0.96	0.005	1.70	1.042
NA12234	CEU	SC	ILLUM	0.42	0.094	2.42	1.736
NA18488	YRI	BI	ILLUM	0.93	0.099	10.58	1.215
NA18533	CHB	BI	ILLUM	0.94	0.007	7.13	1.134
NA18991	JPT	SC	ILLUM	0.95	0.008	5.14	1.107

Contaminated samples – some lanes

SAMPLE ID	POP	CENTER	PLAT-FORM	SM P_{IBD}	CONTAM LANES	LANE P_{IBD}	%altB @refG	LANE EX_HET
NA07346	CEU	454MSC	LS454	0.91	SRR005992	0.50	0.091	1.390
					SRR005993	0.74	0.048	1.234
					SRR005996	0.81	0.035	1.192
					SRR005998	0.96	0.008	0.989
NA12043	CEU	454MSC	LS454	0.96	SRR006315	0.91	0.018	0.984
					SRR006316	0.93	0.015	1.041
					SRR006317	0.91	0.017	1.059
					SRR006318	0.97	0.006	0.880
					SRR006326	0.97	0.005	0.971
					SRR006327	0.97	0.005	0.976
					SRR006329	0.97	0.005	0.982
NA12045	CEU	454MSC	LS454	0.98	SRR006234	0.93	0.013	0.975
					SRR006240	0.97	0.007	0.947

Contaminated samples – some lanes

SAMPLE ID	POP	CENTER	PLAT-FORM	SM P_{IBD}	CONTAM LANES	LANE P_{IBD}	%altB @refG	LANE EX_HET
NA12234	CEU	454MSC	LS454	0.97	SRR006250	0.96	0.008	1.020
					SRR006256	0.76	0.045	1.285
NA12282	CEU	MPIMG	ILLUM	0.98	ERR005841	0.98	0.006	1.053
					ERR006226	0.98	0.005	1.381
					ERR006259	0.98	0.006	1.022
NA19711	ASW	BI	ILLUM	0.90	SRR035429	0.20	0.234	0.598

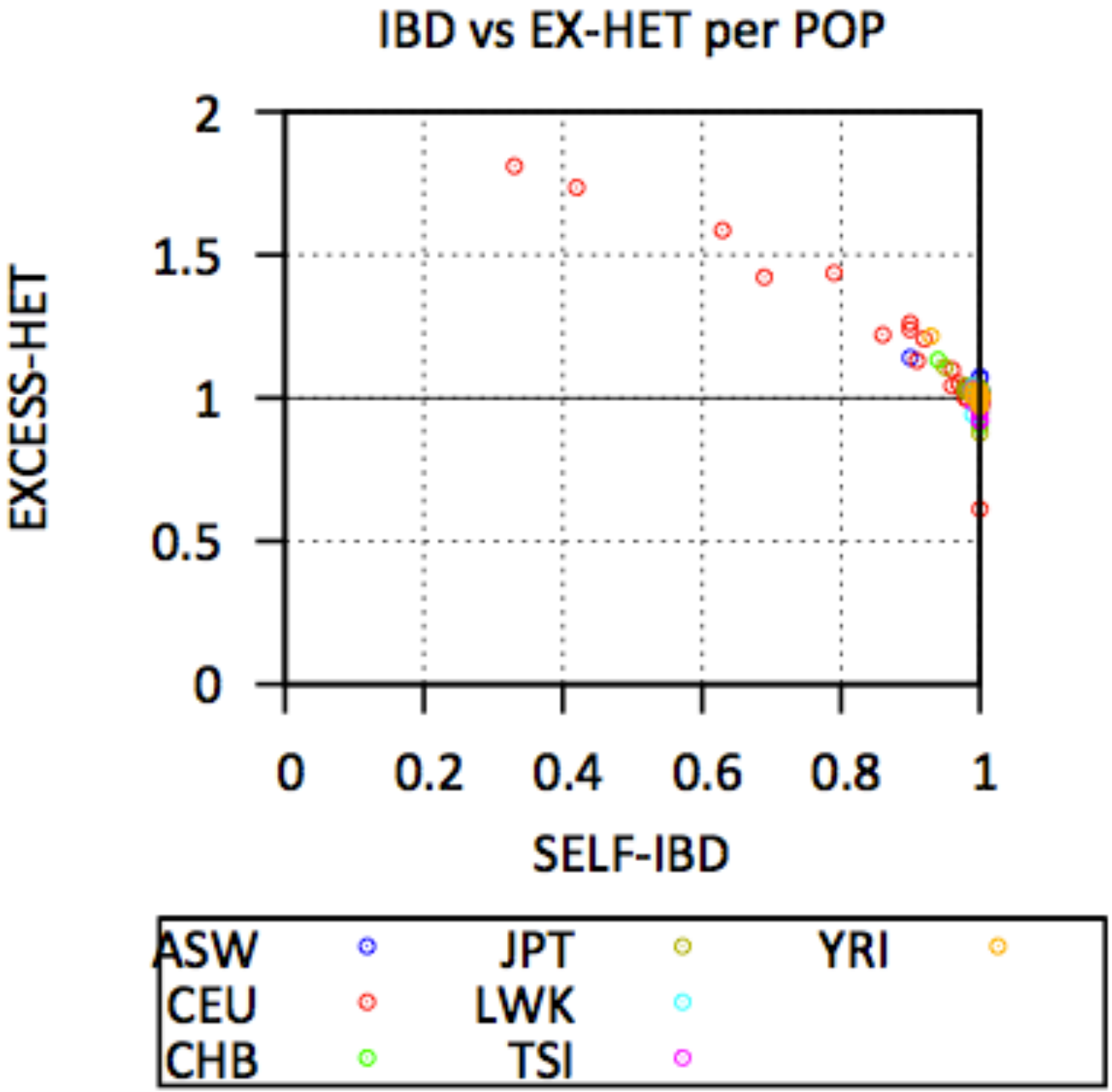
Swapped lanes

SM_ID LABELED	POP	CENTER	PLAT-FORM	SELF P_{IBD}	LANES	BEST P_{IBD}	BEST SM_ID	DP-REF
NA12283	CEU	MPIMG	ILLUM	0.03	ERR006237	1.00	NA12286	0.145
NA12286	CEU	MPIMG	ILLUM	0.04	ERR006238	1.00	NA12283	0.218
					ERR006239	1.00	NA12283	0.214
NA12748	CEU	MPIMG	ILLUM	0.08	ERR006439	1.00	NA12827	0.411
				0.07	ERR006440	1.00	NA12829	0.403
				0.07	ERR006441	1.00	NA12829	0.413
				0.08	ERR006442	1.00	NA12829	0.442
NA12827	CEU	MPIMG	ILLUM	0.07	ERR006446	1.00	NA12748	0.377
NA12829	CEU	MPIMG	ILLUM	0.06	ERR006443	1.00	NA12748	0.362
				0.06	ERR006444	1.00	NA12748	0.403
				0.06	ERR006445	1.00	NA12748	0.398
NA12842	CEU	MPIMG	ILLUM	0.03	ERR006352	1.00	NA12830	0.081
				0.08	ERR006390	1.00	NA12830	0.313
				0.06	ERR006398	1.00	NA12830	0.225
NA12843	CEU	MPIMG	ILLUM	0.12	ERR008576	1.00	NA12889	0.729

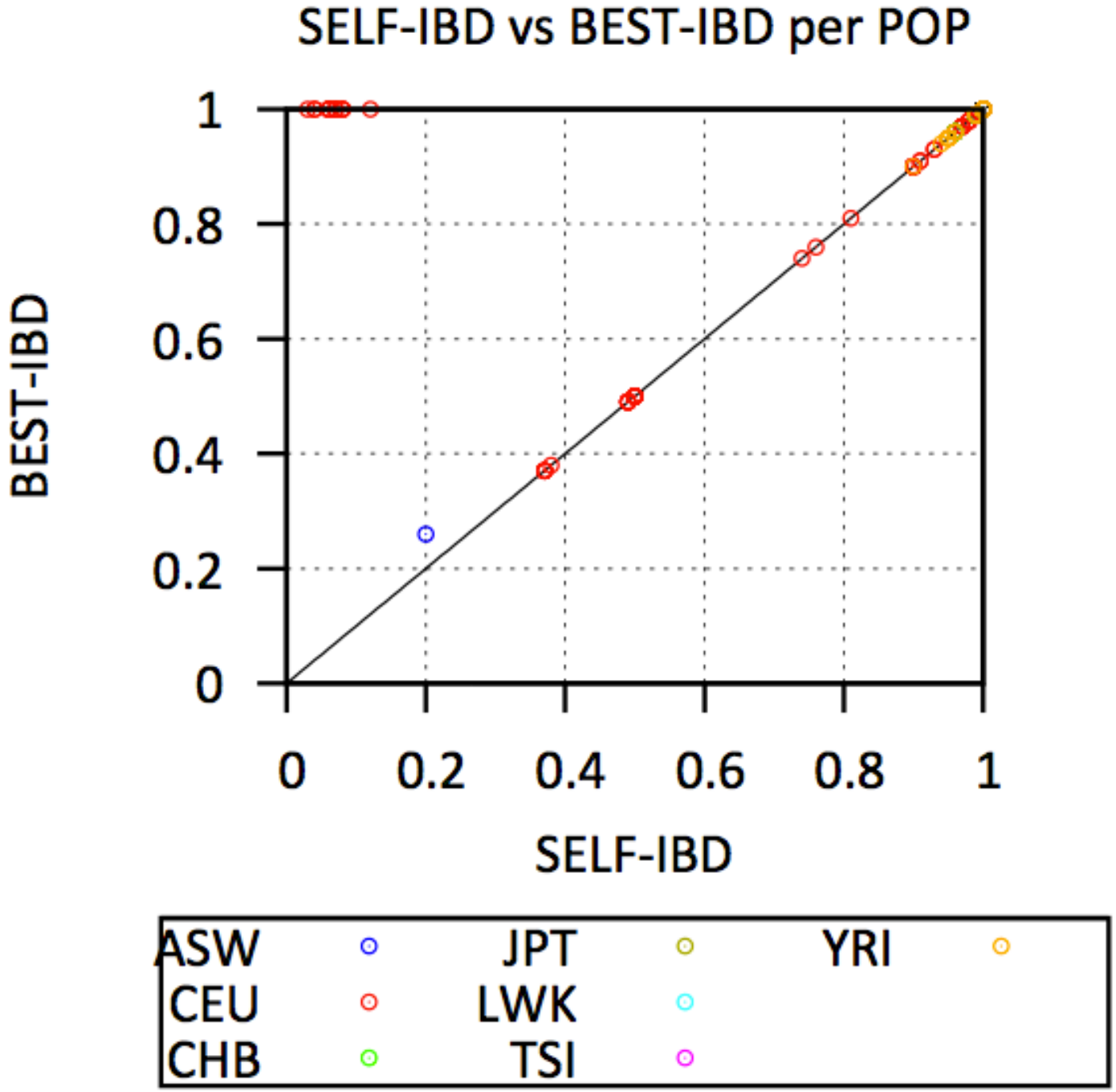
Ungenotyped samples with excessive heterozygosity

SM_ID	POP	CENTER	PLAT-FORM	#Markers DP>2	EX_HET
HG00133	GBR	BI	ILLUM	1.20M	1.817
NA20513	TSI	MULTIPLE	ILLUM	1.02M	1.571
HG00157	GBR	BI	ILLUM	1.17M	1.560
NA20507	TSI	MULTIPLE	ILLUM	1.10M	1.533
NA20503	TSI	MULTIPLE	ILLUM	0.97M	1.529
HG00155	GBR	BI	ILLUM	1.13M	1.504
NA20514	TSI	MULTIPLE	ILLUM	1.05M	1.239
NA19092	YRI	BI	ILLUM	1.19M	1.208

All samples / P_{IBD} & EX_HET



All lanes / sample swaps



Samples with lacking heterozygosity

- Some samples have EX_HET << 1
  - NA12005 (0.613), NA18969 (0.877)
- These also have high % duplicated reads
  - NA12005 (52%), NA18969 (29%)
- Lack of heterozygosity could be due to small library size, and they may need to be resequenced.

Summary

- 1KG low coverage ILLUMINA & 454 analysis
  - 13 / 462 (2.8%) of genotyped samples are fully or partially contaminated
  - 7 / 462 (1.5%) contain lane-level sample swaps, and can be fixed
  - 7 / 67 (10%) of ungenotyped samples show excessive heterozygosity
  - 2 samples lacking heterozygosity with high duplication rate

Acknowledgements

- Goncalo Abecasis
- Melinda Curran
- Paul Anderson
- Mary Kate Trost