The American Journal of Human Genetics, Volume *94*

Supplemental Data

A Statistical Framework to Guide Sequencing Choices in Pedigrees

Charles Y.K. Cheung, Elizabeth Marchani Blue, and Ellen M. Wijsman

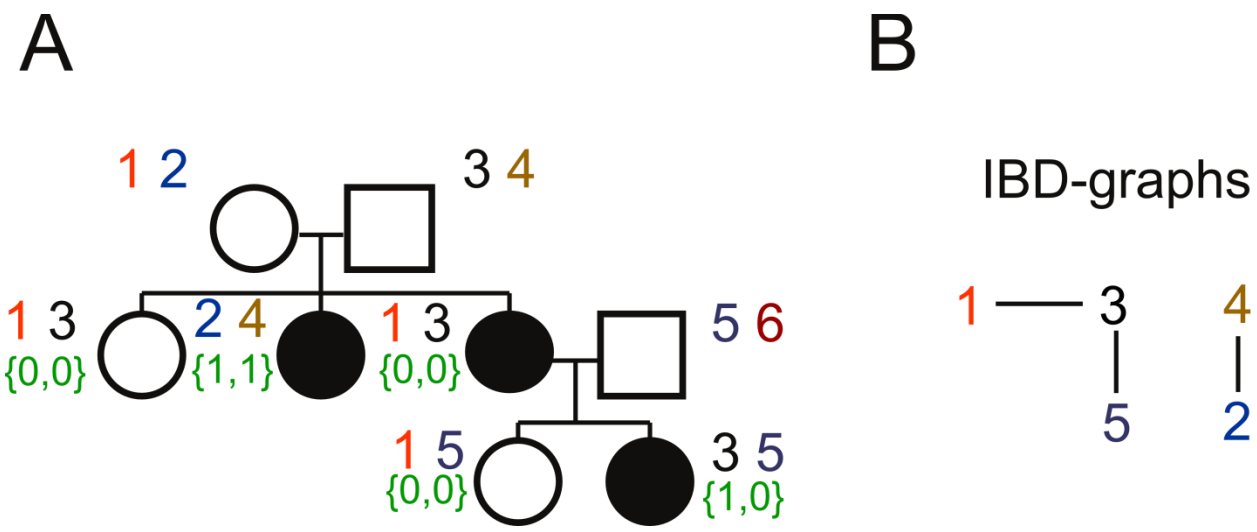

Figure S1. Inheritance Vector, Founder Genome Labels, and IBD-graphs.

A. The inheritance vector (IV) at a locus is composed of meiosis indicators (binary numbers in brackets). Each non-founder has a pair of meiosis indicators: the first/second number represents the transmission of DNA from the subject's mother/father. For each meiosis indicator, 0/1 indicates that the maternal/paternal copy of the parental DNA is transmitted. For example, the first 0 in the leftmost bracket of the leftmost child represents that the maternal DNA of the subject's mother is transmitted and the second 0 in the leftmost bracket represents that the maternal DNA of the subject's father is transmitted at this locus. The transmission of chromosome can also be summarized as founder genome labels (FGLs), which represent the copies of distinct chromosomes each subject inherited, as denoted by numbers not in brackets: by convention, the smaller/larger FGL in each founder represents the DNA inherited from the founder's mother/father. For example, FGL 1 corresponds to the chromosome that the mother from the first generation inherited from her mother. Non-founders inherited the appropriate copy of the distinct chromosomes matching those specified by the meiosis indicators. For example, because the leftmost child inherited the maternal DNA of the subject's mother, the leftmost child inherited FGL 1. Subjects who are observed for genotypes are shaded. B. IBD-graphs of observed subjects are constructed using the FGLs in A. In the process of constructing the IBD-graph(s), we list all FGLs from the observed subjects and connect each pair of FGLs that each observed subject inherited. In this example, the collection of unique FGLs in the set of observed subjects are 1, 2, 3, 4, 5. Three connections are drawn, with the first connection from the observed subject who has FGLs 2 and 4, the second connection from the observed subject who has FGLs 1 and 3, and the third connection from the observed subject who has FGLs 3 and 5. After this process, two disjoint IBD-graphs are created. If the father with FGLs 3 and 4 from the first generation is also observed, a new edge connecting FGLs 3 and 4 would be formed, and the two disjoint IBD-graphs in this original diagram would then be merged into one IBD-graph.

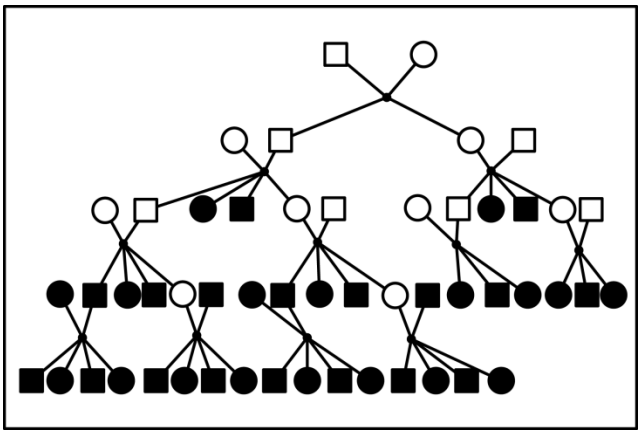

Figure S2. A 52-member Pedigree used in Simulation Study
46 subjects from the bottom 3 generations are available for sequencing; subjects who are shaded represent subjects who have framework markers used to infer IVs.

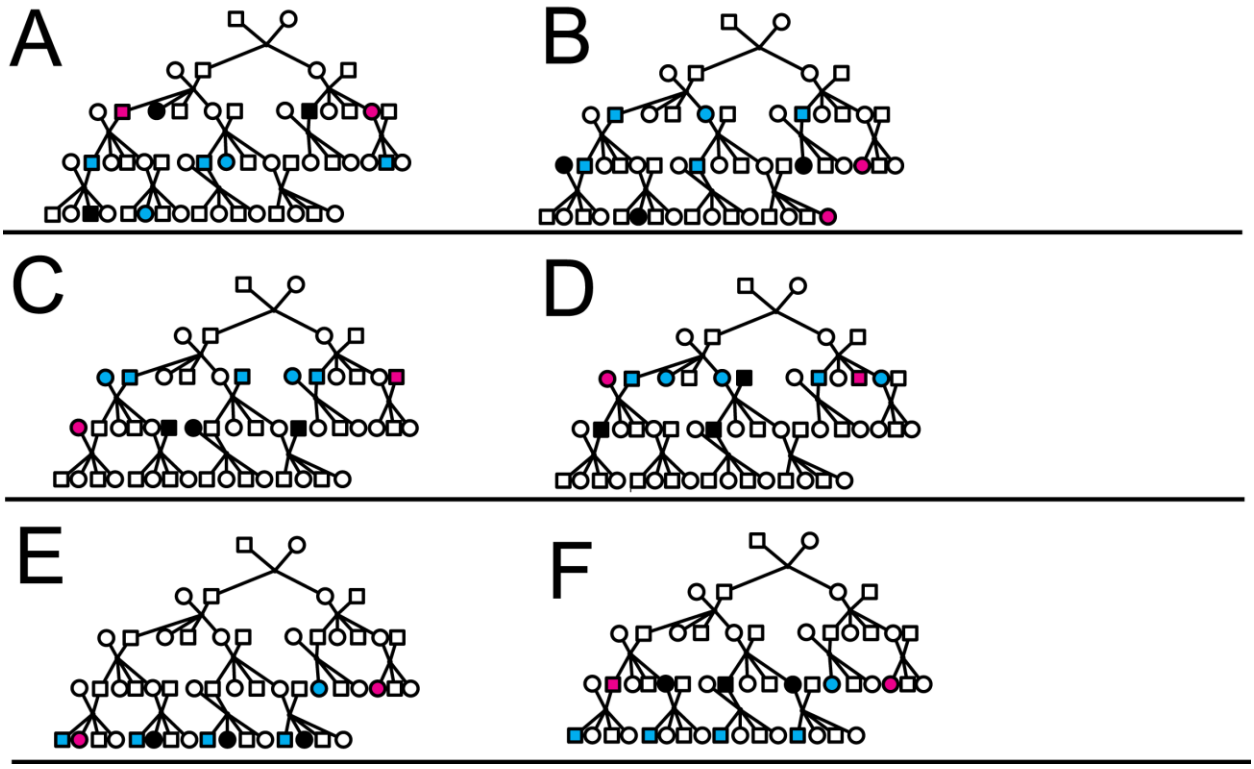

Figure S3. Subjects Selected from Various Methods from the 52-member Simulated Pedigree

Methods: A. GIGI (local) from dataset #2, B. GIGI (GW), C. PRIMUS, D. ExomePicks, E. Bottom-only, F. Bottom & parents. The first 5 selections are in cyan, the next 2 selections are in magenta, and the last 3 selections are in black.

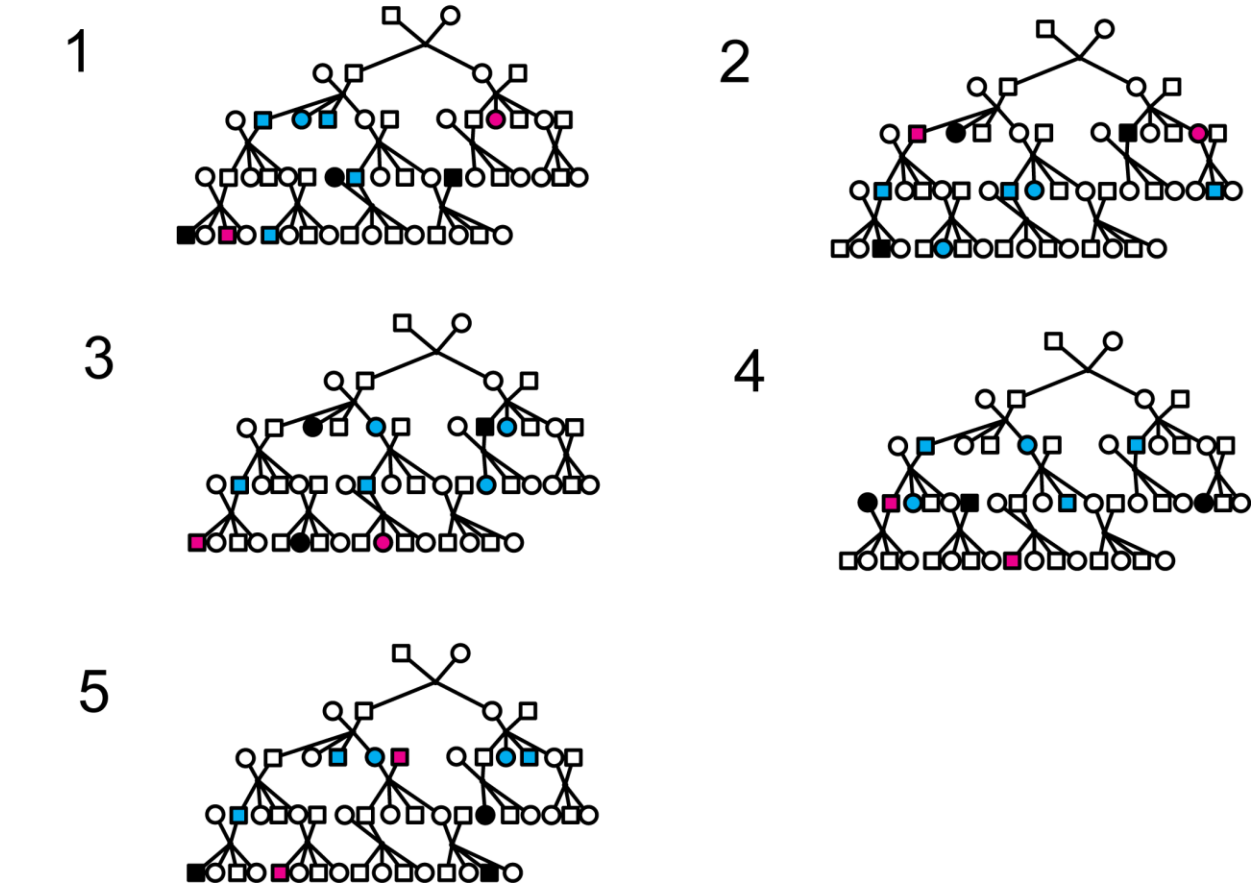

Figure S4. The Selected Subjects from GIGI-Pick (local) from the first 5 Simulated Datasets, illustrating that the Method Adapts to the Sampled IVs in each Dataset
The first 5 selections are in cyan, the next 2 selections are in magenta, and the last 3 selections are in black.

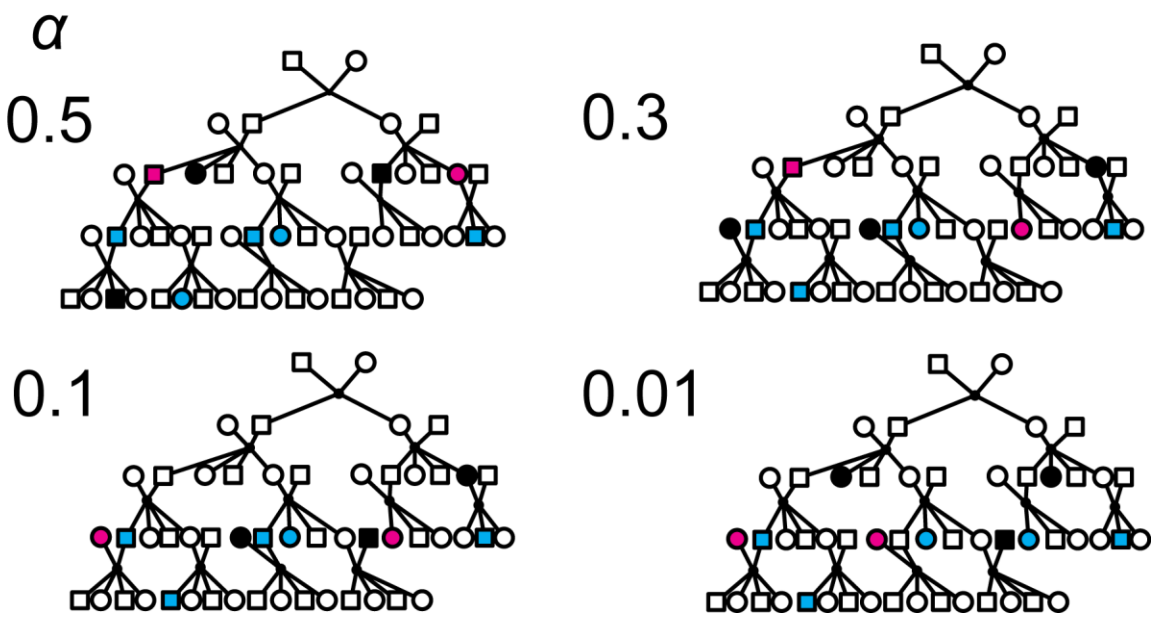

Figure S5. Varying α influences Subject Selection from GIGI-Pick (local) with Subjects Selected from Dataset #2.
The first 5 selections are in cyan, the next 2 selections are in magenta, and the last 3 selections are in black.

Table S1. Sensitivity computed on rare variants by the number of subjects selected under Methods of Subject Selection

Method of Subject Selection[a]	Number of Subjects Selected					
	5		7		10	
	Sensitivity (%)	Percentile[b] (%)	Sensitivity (%)	Percentile[b] (%)	Sensitivity (%)	Percentile[b] (%)
GIGI-Pick (local)	48.3	100	58.4	100	69.2	100
GIGI-Pick (GW)	43.8	99.5	54.2	99.5	65.7	98.0
PRIMUS	15.7	0.5	19.6	0	25.2	0
Exome Pick	36.0	84.0	41.3	39.5	60.9	80.0
Bottom-only	25.3	21.5	35.0	10.5	54.1	25.5
Bottom & parents	25.3	21.5	44.4	65.0	64.0	97.0

a. Results were averaged across all 10 simulated datasets. Refer to Figure S3 and S4 for actual subjects selected.

b. Relative to 200 random selections of subjects for sequencing

Table S2. Accuracy computed on SNPs as a function of Methods of Subject Selection and the Number of Subjects Selected

Method of Subject Selection[a]	Number of Subjects Selected					
	5		7		10	
	Accuracy (%)	Percentile[b] (%)	Accuracy (%)	Percentile[b] (%)	Accuracy (%)	Percentile[b] (%)
GIGI-Pick (local)	77.9	100	81.9	100	87.0	100
GIGI-Pick (GW)	76.6	99.0	80.4	99.5	85.5	99.5
PRIMUS	72.9	14.0	75.1	0.5	78.5	0
Exome Pick	74.8	73.5	77.4	29.0	82.9	50.5
Bottom-only	74.0	46.0	77.3	27.5	82.7	46.0
Bottom & parents	74.0	46.0	79.1	84.5	84.7	96.0

a. Results were averaged across all 10 simulated datasets. Refer to Figure S3 and S4 for actual subjects selected.

b. Relative to 200 random selections of subjects for sequencing

Table S3. Number of times each Subject Selection Method achieved a particular Rank in 10 datasets, by the Number of Subjects Selected

# Subjects	Method	Rank[a]					
		1	2	3	4	5	6
5	GIGI-Pick (local)	8	1	1	-[b]	-	-
	GIGI-Pick (GW)	1	8	1	-	-	-
	PRIMUS	-	-	-	2	-	8
	ExomePicks	1	1	5	-	3	-
	Bottom-only	-	-	3	5	2	-
	Bottom & Parents	-	-	3	5	2	-
7	GIGI-Pick (local)	9	-	1	-	-	-
	GIGI-Pick (GW)	1	8	1	-	-	-
	PRIMUS	-	-	-	-	1	9
	ExomePicks	-	-	5	3	2	-
	Bottom-only	-	-	-	3	6	1
	Bottom & Parents	-	2	3	5	-	-
10	GIGI-Pick (local)	8	-	2	-	-	-
	GIGI-Pick (GW)	1	6	1	1	1	-
	PRIMUS	-	-	-	-	-	10
	ExomePicks	-	1	2	4	3	-
	Bottom-only	-	1	-	3	6	-
	Bottom & Parents	1	2	5	2	-	-

a. Lower rank reflects higher sensitivity. e.g. at 5 subjects selected, GIGI-Pick (local) has the highest sensitivity (rank=1) in 8 of the 10 datasets;

b. - is substituted in place of 0 to improve readability

Table S4. Effect of varying α with 7 subjects selected under GIGI-Pick (local)

α	Rare variants		SNPs	
	Sensitivity (%)	Percentile[a] (%)	Accuracy (%)	Percentile[a] (%)
0.5	58.4	100	81.9	100
0.3	56.6	100	81.7	100
0.1	44.2	64.5	79.8	95.0
0.01	20.1	0	76.4	8.5

a. Comparison with 200 random selections of subjects for sequencing