Supplementary Materials

Supplementary Table 1. Incidence rate has little impact on VAAST's ability to identify rare recessive diseases. Background file consisted of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with 9 additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen set(Reese et al. 2010). Target files are as described in tables 2 and 3. Data was generated on the intersection by variant position of described exomes using VAAST's fully penetrant and monogenic recessive model. Causative allele incidence was set as indicated, and amino acid substitution frequency along with masking. For the Miller's Syndrome and CCD analyses in the main text, we estimated the prevalence of Miller's Syndrome at 120/1 billion(Ng et al. 2010) and CCD at 1/5500(Kagalwalla 1994). Assuming a recessive model with complete penetrance, we set the threshold on p_i^U such that it could not exceed the square root of the prevalence.

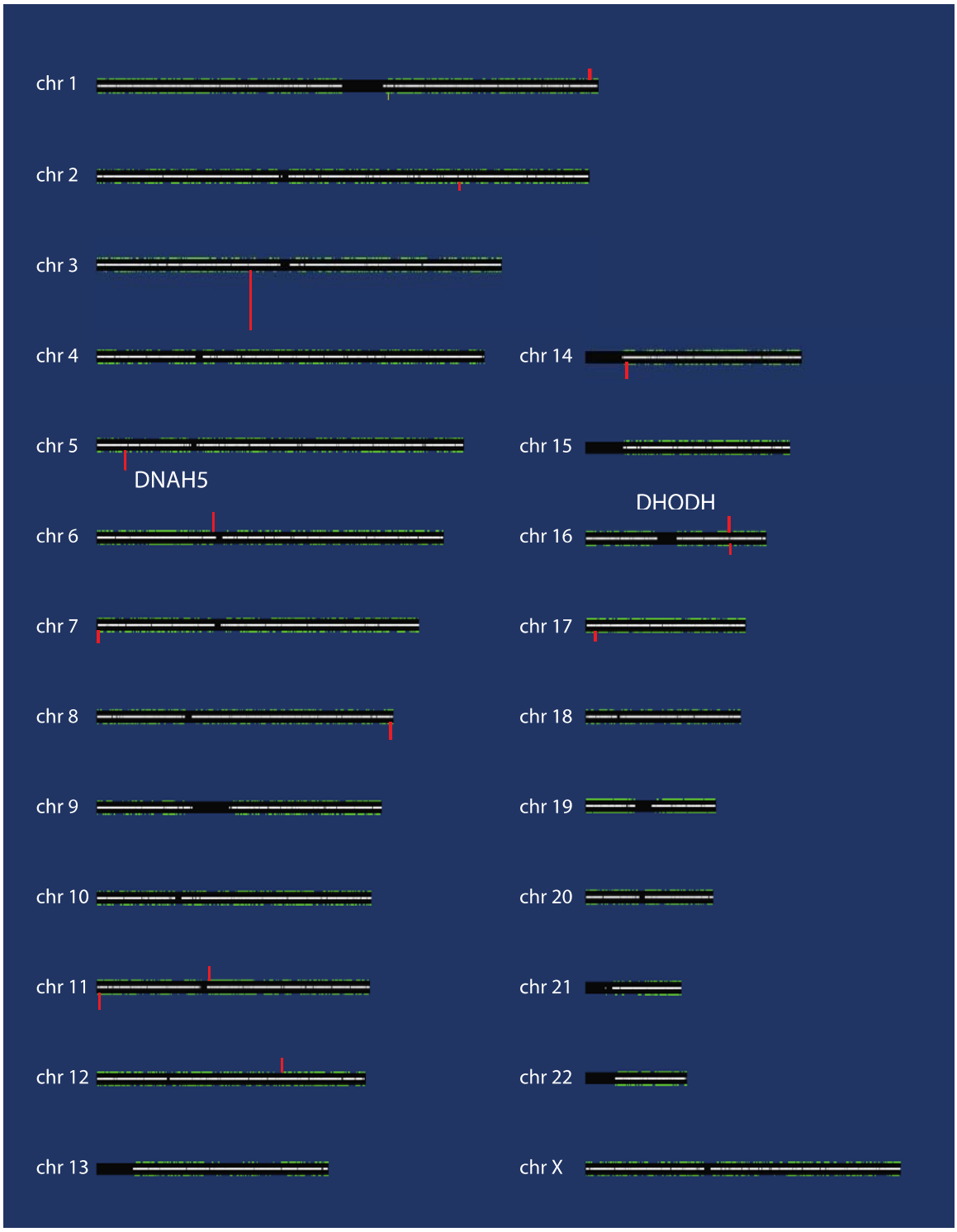
Disease	Target Genome(s)	Disease Incidence	Genome Wide			Causative Gene		
			Genes Scored	Significant Genes		Rank	P-Value	
				Non LD Corrected	LD Corrected		Non LD Corrected	LD Corrected
CCD (SLC25A2)	Homozygous Affected	0.013	127	66	0	46	1.51E-05	5.26E-03
		0.01	123	69	0	24	1.31E-05	5.26E-03
		0.001	94	66	0	37	1.44E-05	5.26E-03
	2 Unrelated Affected	0.013	3	3	0	1	1.49E-09	5.51E-05
		0.01	3	3	0	1	1.10E-09	5.51E-05
		0.001	2	2	0	1	1.13E-09	5.51E-05
	3 Unrelated Affected	0.013	1	1	1	1	3.10E-13	8.61E-07
		0.01	1	1	1	1	3.00E-13	8.61E-07
		0.001	1	1	1	1	3.34E-13	8.61E-07
Miller's Syndrome (DHODH)	Compound Heterozygous Affected	0.010	101	66	0	88	1.65E-04	5.26E-03
		0.00	93	67	0	83	1.18E-04	5.26E-03
		0.000	82	67	0	81	8.70E-05	5.26E-03
	2 Unrelated Affected	0.010	4	4	0	2	1.37E-08	5.51E-05
		0.00	4	4	0	3	3.89E-08	5.51E-05
		0.000	2	2	0	2	1.71E-08	5.51E-05
	3 Unrelated Affected	0.010	3	3	1	1	3.34E-10	8.61E-07
		0.00	3	3	2	1	1.29E-10	8.61E-07
		0.000	1	1	1	1	2.60E-11	8.61E-07

Supplementary Table 2. Impact of recessive disease modeling on VAAST Miller syndrome Accuracy. Background and target files are as described in Table 2. Genes scored refers to the number of genes in the genome with scores greater than zero and meeting the recessive model criteria, if applied.

Target Genome(s)	Recessive Modeling Employed	Genome Wide			DHODH Rank		
		Genes Scored	Significant Genes		Rank	P-Value	
			Non LD Corrected	LD Corrected		Non LD Corrected	LD Corrected
1 Complex Heterozygote	No	305	72	0	87	7.35E-05	5.26E-03
	Yes	92	67	0	86	2.36E-04	5.26E-03
2 Complex Heterozygotes	No	544	58	0	9	1.66E-07	5.51E-05
	Yes	4	3	0	2	2.81E-08	5.51E-05
3 Complex Heterozygotes	No	849	111	1	1	6.65E-11	8.61E-07
	Yes	2	2	1	1	2.61E-11	8.61E-07
4 Complex Heterozygotes	No	1069	144	6	1	1.55E-16	1.78E-08
	Yes	1	1	1	1	1.99E-15	1.78E-08
5 Complex Heterozygotes	No	1277	170	8	1	2.42E-18	4.60E-10
	Yes	1	1	1	1	6.95E-15	4.60E-10
6 Complex Heterozygotes	No	1481	215	11	1	2.85E-19	1.42E-11
	Yes	1	1	1	1	5.79E-17	1.42E-11

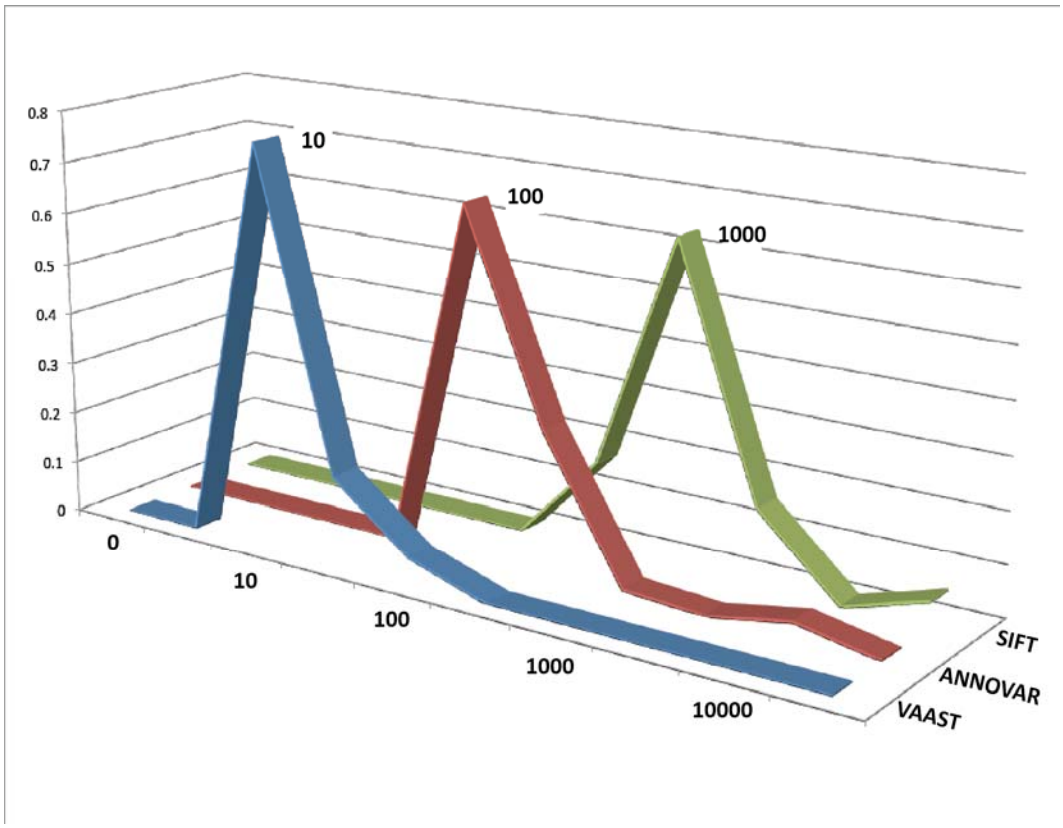
Supplementary Table 3. Impact of recessive disease modeling on VAAST CCD Accuracy. Background and target files are as described in Table 4. Genes scored refers to the number of genes in the genome with scores greater than zero and meeting the recessive model criteria, if applied.

Target Genome(s)	Recessive Modeling Employed	Genome Wide			SLC26A3		
		Genes Scored	Significant Genes		Rank	P-Value	
			Non LD Corrected	LD Corrected		Non LD Corrected	LD Corrected
Homozygous Affected	Yes	127	69	0	21	1.22E-05	5.26E-03
	No	431	71	0	15	1.04E-05	5.26E-03
Union 2-Unrelated Homozygotes	Yes	7	7	0	3	4.74E-10	5.51E-05
	No	769	81	0	3	7.06E-10	5.51E-05
Intersection 2-Unrelated Homozygotes	Yes	3	3	0	1	7.47E-10	5.51E-05
	No	36	10	0	2	6.92E-10	5.51E-05
Union 3-Unrelated Homozygotes	Yes	2	2	2	1	2.83E-13	8.61E-07
	No	1067	122	7	1	3.52E-13	8.61E-07
Intersection 3-Unrelated Homozygotes	Yes	1	1	1	1	1.29E-13	8.61E-07
	No	32	11	2	1	3.17E-13	8.61E-07

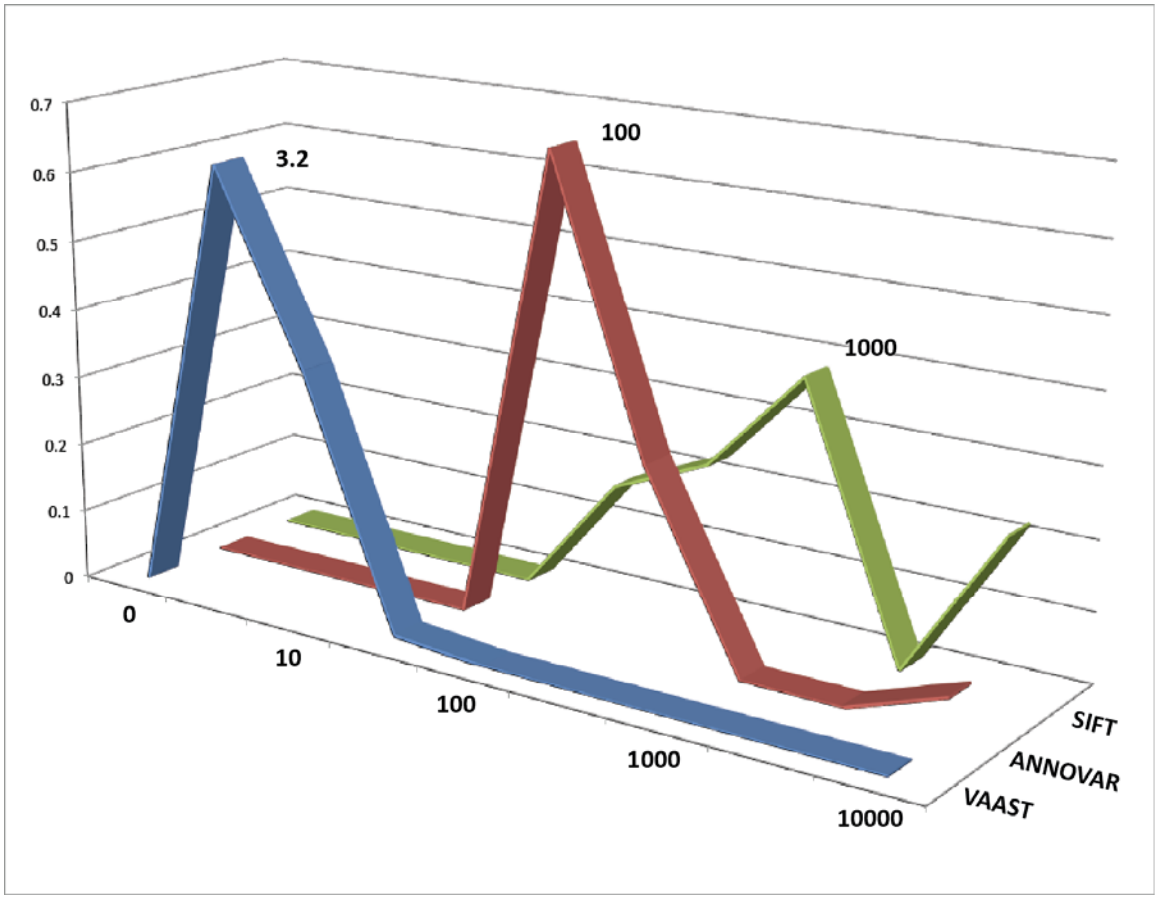


Supplemental Figure 1. Genome-wide VAAST analysis of the Utah Miller Syndrome Siblings. VAAST was run using the genomes the two affected Siblings. Grey bars along the center of each chromosome show proportion of unique sequence along the chromosome arms, with white denoting completely unique sequence; black regions thus outline highly repetitively or centromeric regions. Colored bars above and below the chromosomes (mostly green) represent each

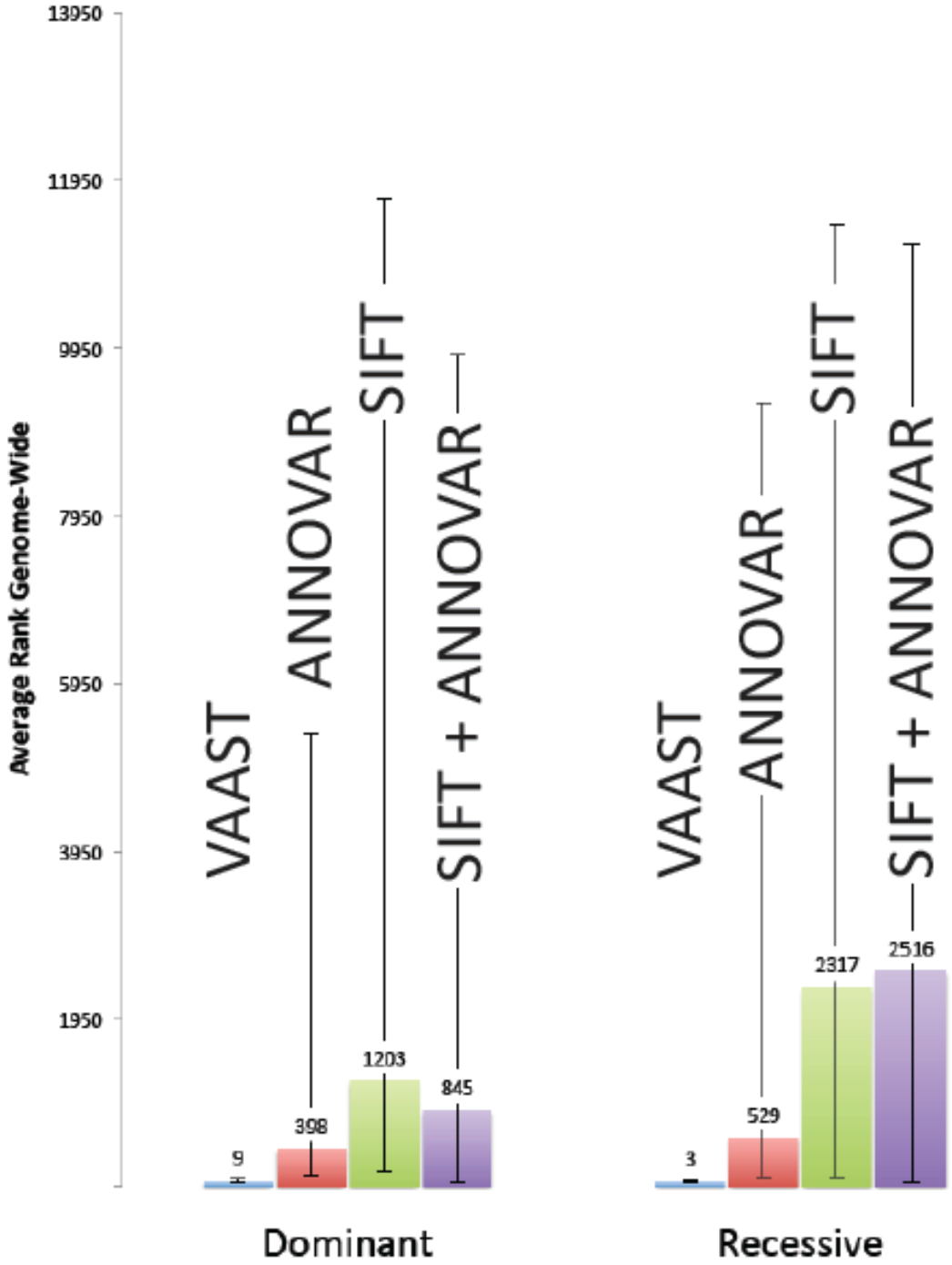
annotated gene; plus strand genes are shown above and minus strand genes below; their width is proportional their length; height of bar, their VAAST score. Genes colored red are candidates identified by VAAST achieving genome-wide statistical significance (non-LD corrected). 14 genes are identified. Causative allele incidence was set to 0.00035, and amino acid substitution frequency was used along with variant-masking. This view was generated using the VAAST report viewer. This software tool allows the visualization of a genome-wide search in easily interpretable form, graphically displaying chromosomes, genes and their VAAST scores. For comparison, the corresponding figure, with pedigree information, is provided as Figure 4.



Supplemental figure 2. Rank distributions for the dominant analyses shown in figure 5C. y-axis: frequency of ranks. X-axis: $\log_{10}$(ranks). The numbers on the peaks denote the mode rank for each tool.



Supplemental figure 3. Rank distributions for the recessive analyses shown in figure 5C. y-axis: frequency of ranks. X-axis: $\log_{10}$(ranks). The numbers on the peaks denote the mode rank for each tool.



Supplemental figure 4. Evaluation of one method for combining SIFT and ANNOVAR results for the analyses show in figure 5C. The category SIFT+ANNOVAR represents a ranking system that takes the intersection of variants that are judged to be potentially disease-causing in both SIFT and ANNOVAR. Due to SIFT's relatively low sensitivity, this method performs worse than using ANNOVAR alone. Note that we did not exhaustively explore all possible methods for combining information from SIFT and ANNOVAR. Therefore, other methods may produce different results.

Supplementary Methods

Bootstrap procedure used in Figure 6

The publicly-available datasets for the NOD2 (Lesage et al. 2002) and LPL (Johansen et al. 2010) provide the site-frequency spectra for non-synonymous variants in both cases and controls. Because these datasets do not contain individual genotype information, we were unable to randomly sample from individuals in these datasets. Therefore, to resample from the original datasets, we generated each bootstrapped individual by randomly sampling (with replacement) from each non-synonymous site in the site-frequency spectrum. Because of the relatively small control sample size in the CARD15 dataset (103 cases vs. 453 controls), we merged the NOD2 controls with the sample of 60 CEU individuals from the 1000 Genomes project(Durbin et al. 2010). Without the CEU data, the power of VAAST to detect a significant difference (α=2.38x10^{-6}) between cases and controls plateaued at approximately 0.72.

Command-line parameters

This section documents the VAAST command-line parameters that can be used to reproduce the results in Figure 6, Tables 2-5, and Table S1:

Figure 6

AAS with OMIM command-line: VAAST -q <degree of parallelization> -f -c <chromosome_number> -g 4 -k -d 100000000 -j 2.38E-6 -m lrt -o <output_file> <feature_file > <189genome.cdr> <target>

AAS with blosum62 command-line: VAAST -q <degree of parallelization> -f -c <chromosome_number> -g 4 -k -b blosum62.matrix -d 100000000 -j 2.38E-6 -m lrt -o <output_file> <feature_file> <189genome.cdr> <target>

No AAS command-line: VAAST -q <degree of parallelization> -f -c <chromosome_number> -g 4 -d 100000000 -j 2.38E-6 -m lrt -o <output_file> <feature_file> <189genome.cdr> <target>

WSS command-line: VAAST -q <degree of parallelization> -c <chromosome_number> -d 100000000 -j 2.38E-6 -m wss -o <output_file> <feature_file> <189genome.cdr> <target>

Table 2

Description: Effect of background size and heterogeneity on accuracy. Results of searching the intersection of two Utah Miller Syndrome affected genomes against two different background files, with and without masking. Caucasians Only: 65 genomes. 6 different sequencing/alignment/variant calling platforms. Mixed Ethnicities: 189 genome equivalents (62 YRI, 65 CAUC, 62 ASIAN). UMSK: unmasked; MSK: masked. Genome-wide Significant Genes: number of genes genome-wide attaining a significant non-LD corrected p-value. Rank: Gene rank of DHODH & DNAH5 among all scored genes; P-Value: non-LD corrected p-value; Genome-wide significant alpha is 2.4E-6. Data was generated using a fully penetrant and monogenic recessive model with the

likelihood ratio testing mode of VAAST. Causative allele incidence was set to 0.00035, and amino acid substitution frequency was used along with rare variant grouping and masking of variants (if applicable). Scoring was evaluated by permutation by sites and permutation by genome.

Command-line: VAAST -m lrt –iht r –lh n –pnt c -r 0.00035 -p 4 -w 100000 -d 100000 -k -g 4 -x 35bp_se.wig -o <output_file> <feature_file> <189genome.cdr> <target>

Table 3

Description: Impact of exome number on VAAST's ability to identify rare disease caused by compound heterozygous alleles. Background file consisted of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with 9 additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen genome set. Causative alleles reported in Ng et al., 2010 were added to unrelated exomes from re-sequenced individuals from Denmark reported in Li et al., 2010. Searches were carried out using unions of the exomes by variant position. Data was generated using a fully penetrant and monogenic recessive model with the likelihood ratio testing mode of VAAST. Causative allele incidence was set to 0.00035, and amino acid substitution frequency was used along with rare variant grouping and masking of variants. Scoring was evaluated by permutation by sites and permutation by genome.

Command-line: VAAST -m lrt -iht r –lh n –pnt c -r 0.00035 -p 4 -w 100000 -d 100000 -k -g 4 -x 35bp_se.wig -o <output_file> <feature_file> <189genome.cdr> <target>

Table 4

Description: Variant contribution to gene scoring of DHODH. Background file consisted of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with 9 additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen set. Target file contains 6 unrelated individuals with compound heterozygous variants described in table 3. Data was generated on the intersection of described exomes using a fully penetrant and monogenic recessive model with the likelihood ratio testing mode of VAAST. Causative allele incidence was set to 0.00035, and amino acid substitution frequency was used along with rare variant grouping and masking of variants (if applicable). Scoring was evaluated by permutation by sites and permutation by genome.

Command-line: VAAST -m lrt –iht r –lh n –pnt c -r 0.00035 -p 4 -w 100000 -d 100000 -k -g 4 -x 35bp_se.wig -o <output_file> <feature_file> <189genome.cdr> <target>

Table 5.

Description: Impact of exome numbers on VAAST's ability to identify a rare recessive disease. Background file consisted of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with 9 additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen set. Targets: 1st Homozygote affected is the single CCD affected exome reported in Choi et al., 2009; remaining affecteds: unrelated exomes from re-sequenced individuals from Denmark reported in Li et al., 2010 with the causative allele added. Data was generated on either the union or intersection of affecteds using the fully penetrant and monogenic recessive model with

the likelihood ratio testing mode of VAAST. Causative allele incidence was set to 0.013, and amino acid substitution frequency was used along with rare variant grouping and masking of variants (if applicable). Scoring was evaluated by permutation by sites and permutation by genome.

Command-line: VAAST -m lrt –iht r –lh n –pnt c -r 0.013 -p 4 -w 100000 -d 100000 -k -g 4 -x 35bp_se.wig -o <output_file> <feature_file> <189genome.cdr> <target>

Table S1

Description: Impact of disease incidence on VAAST's ability to identify rare recessive disease. Background file consisted of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with 9 additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen set. Target files are described in tables 3 and 5. Data was generated on the intersection of described exomes using a fully penetrant and monogenic recessive model with the likelihood ratio testing mode of VAAST. Causative allele incidence was set as indicated, and amino acid substitution frequency was used along with rare variant grouping and masking of variants (if applicable). Scoring was evaluated by permutation by sites and permutation by genome.

Command-line: VAAST -m lrt –iht r –lh n –pnt c -r <as indicated> -p 4 -w 100000 -d 100000 -k -g 4 -x 35bp_se.wig -o <output_file> <feature_file> <189genome.cdr> <target>

References

Durbin, RM, GR Abecasis, DL Altshuler, A Auton, LD Brooks, RA Gibbs, ME Hurles, GA McVean. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Johansen, CT, J Wang, MB Lanktree, et al. 2010. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 42:684-687.

Kagalwalla, AF. 1994. Congenital chloride diarrhea. A study in Arab children. J Clin Gastroenterol 19:36-40.

Lesage, S, H Zouali, JP Cezard, et al. 2002. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am J Hum Genet 70:845-857.

Ng, SB, KJ Buckingham, C Lee, et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30-35.

Reese, MG, B Moore, C Batchelor, F Salas, F Cunningham, GT Marth, L Stein, P Flicek, M Yandell, K Eilbeck. 2010. A standard variation file format for human genome sequences. Genome Biol 11:R88.

Roach, JC, G Glusman, AF Smit, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636-639.

Vuong, QH. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57:307-333.