

DETECTING STATISTICAL INTERACTION BETWEEN SOMATIC MUTATIONAL EVENTS AND GERMLINE VARIATION FROM NEXT-GENERATION SEQUENCE DATA

HAO HU

*Department of Epidemiology
The University of Texas MD Anderson Cancer Center
1155 Pressler Street
Houston, TX, 77030, USA
Email: hhu1@mdanderson.org*

CHAD D. HUFF

*Department of Epidemiology
The University of Texas MD Anderson Cancer Center
1155 Pressler Street
Houston, TX, 77030, USA
Email: chuff1@mdanderson.org*

The two-hit model of carcinogenesis provides a valuable framework for understanding the role of DNA repair and tumor suppressor genes in cancer development and progression. Under this model, tumor development can initiate from a single somatic mutation in individuals that inherit an inactivating germline variant. Although the two-hit model can be an overgeneralization, the tendency for the pattern of somatic mutations to differ in cancer patients that inherit predisposition alleles is a signal that can be used to identify and validate germline susceptibility variants. Here, we present the Somatic-Germline Interaction (SGI) tool, which is designed to identify statistical interaction between germline variants and somatic mutational events from next-generation sequence data. SGI interfaces with rare-variant association tests and variant classifiers to identify candidate germline susceptibility variants from case-control sequencing data. SGI then analyzes tumor-normal pair next-generation sequence data to evaluate evidence for somatic-germline interaction in each gene or pathway using two tests: the Allelic Imbalance Rank Sum (AIRS) test and the Somatic Mutation Interaction Test (SMIT). AIRS tests for preferential allelic imbalance to evaluate whether somatic mutational events tend to amplify candidate germline variants. SMIT evaluates whether somatic point mutations and small indels occur more or less frequently than expected in the presence of candidate germline variants. Both AIRS and SMIT control for heterogeneity in the mutational process resulting from regional variation in mutation rates and inter-sample variation in background mutation rates. The SGI test combines AIRS and SMIT to provide a single, unified measure of statistical interaction between somatic mutational events and germline variation. We show that the tests implemented in SGI have high power with relatively

modest sample sizes in a wide variety of scenarios. We demonstrate the utility of SGI to increase the power of rare variant association studies in cancer and to validate the potential role in cancer causation of germline susceptibility variants.

1. Introduction

In 1971, Alfred Knudson proposed the two-hit hypothesis for retinoblastoma, demonstrating that the distribution of age-of-onset for familial retinoblastoma cases was consistent with inheritance of a germline variant followed by a somatic mutation, while age-of-onset for sporadic cases was consistent with two independent somatic mutations¹. The gene responsible for this process was identified 15 years later as *RBI*, the first tumor suppressor gene^{2,3}. The two-hit hypothesis is now the classic model for DNA repair and tumor suppressor genes, which follow a dominant mode of inheritance but are typically recessive at the cellular level. This model provides a useful framework for understanding cancer predisposition, although DNA repair and tumor suppressor genes can be either dominant or recessive at the cellular level, depending on the context. Germline mutations in the tumor suppressor gene *TP53* follow both one- and two-hit models in Li-Fraumeni syndrome, with some inherited genetic causes resulting from cellular recessive loss-of-function nonsense variants and others resulting from dominant gain-of-function missense variants⁴. The DNA repair genes *BRCA1* and *BRCA2* variants are also either recessive or dominant at the cellular level depending on the type of cancer, with complete loss of the wild type allele in ovarian cancer but occasional haplo-insufficiency in breast cancer⁵. In general, inherited variants in the tumor suppressor gene *APC* are recessive at the cellular level in colorectal cancer⁶, but can exert dominant effects that can lead to chromosomal instability⁷. In contrast to DNA repair and tumor suppressor genes, oncogenes are generally dominant at both the germline and cellular levels, and thus tend to follow a one-hit model. Nonetheless, there are a number of examples of oncogenes that follow a two-hit model⁸. Thus, although one- and two-hit models are sometimes overgeneralizations, many genes display a pattern of somatic mutational events in tumors that occur more or less frequently than expected among individuals that carry particular germline susceptibility variants.

Next-generation sequencing now provides efficient, high-coverage interrogation of nearly the entire genome and is revolutionizing our understanding of somatic mutational events that drive tumorigenesis⁹⁻¹¹. The use of next-generation sequencing to identify rare germline variants that influence cancer risk also holds great promise but is fundamentally a more difficult problem given that purifying selection ensures that intermediate-penetrance germline variants are usually very rare. A number of rare variant association tests have been developed recently to identify disease-susceptibility genes from case-control next-generation sequence data. The primary advantage of these methods over traditional approaches is that they aggregate rare variants to perform a single statistical test for each gene, which greatly increases power while reducing the multiple testing burden. However, as we have previously shown, although rare variant association tests greatly improve statistical power, studies involving thousands of cases and controls will likely be needed to identify novel gene associations for common cancers¹²⁻¹⁴. The tendency for somatic mutational events to occur more or less frequently than expected given the presence of a germline

susceptibility variant is an additional piece of evidence that can aid in the search for novel gene-cancer susceptibility associations or in the validation and characterization of candidate germline susceptibility variants. The primary motivation of this work is to provide a framework for identifying these statistical interactions between somatic and germline variation in a high-throughput manner that takes advantage of available bioinformatic tools and existing next-generation sequencing capacity. The methods we present are implemented in the Somatic Germline Interaction (SGI) tool.

SGI analyzes next-generation sequencing data from tumor-normal tissue pairs and normal tissue in matched controls to determine whether germline variation in a gene or pathway statistically interacts with the occurrence of somatic events. The two-hit model describes one process that can result in statistical interaction, in which two damaged copies of a gene are required to initiate tumorigenesis. If the two-hits model holds, then the tumors of cancer patients with a deleterious germline variant in a driver gene are likely to have a second somatic mutation event in the same gene. Another process that can result in statistical interaction involves *cis*-acting germline variants that can greatly increase the somatic mutation rate in the local genomic region¹⁵⁻¹⁸. SGI identifies candidate germline susceptibility variants by interfacing with the Variant Annotation, Analysis and Search Tool (VAAST)¹⁹. The rare variant association test in VAAST incorporates amino acid substitution severities, phylogenetic conservation, and the distribution of allele frequencies in cases and controls to variants and genes that are likely to influence disease susceptibility¹². After identifying individuals in the study with candidate germline variants, SGI then analyzes tumor-normal pair sequence data to evaluate whether somatic mutational events occur more or less frequently than expected by testing the null hypothesis that the occurrence of somatic events is independent of the presence or absence of germline variation.

We divide somatic mutational events into two categories: somatic mutations and preferential allelic imbalance. SGI implements the Allelic Imbalance Rank Sum (AIRS) test to evaluate evidence for preferential allelic imbalance. Specifically, within each gene or pathway, AIRS tests whether the chromosomes harboring putatively deleterious germline mutations are preferentially amplified in tumor tissues. Allelic imbalance is an important signal of somatic mutations resulting from copy number variants (CNVs) or loss-of-heterozygosity (LOH) that has been used to identify and validate modest penetrance germline-cancer associations in both humans^{15,20,21} and mice²²⁻²⁴. In addition to allelic imbalance, SGI also evaluates whether somatic mutations occur more or less frequently than expected in the tumors of individuals that harbor putatively deleterious germline mutations using the Somatic Mutation Interaction Test (SMIT). SMIT only considers single nucleotide and small indel somatic mutations that do not result in LOH or CNVs in a large genomic region, as these larger somatic events are evaluated by allelic imbalance evidence. SGI also combines AIRS and SMIT to provide a single unified framework to detect statistical interaction between germline and somatic variation.

SGI has a number of potential applications. For known germline-susceptibility genes, SGI can validate germline variants of unknown significance. For genes that are known to be significantly mutated in tumors but not known to play a role in cancer predisposition, SGI can search for novel

germline variant associations. SGI can also identify novel cancer-associated genes that would be much more difficult to detect than germline case-control studies or somatic mutational analysis alone due to rarity and/or effect size. Here, we present the methods implemented in SGI and evaluate the performance of the tool in a wide variety of scenarios.

2. Methods

2.1. Identifying candidate germline variants

SGI processes VAAST output files to identify individuals with candidate germline susceptibility variants. For each gene, any variant that has a VAAST score of greater than 0 is identified as a candidate. SGI then performs the AIRS and SMIT tests based on the binary classification of individuals with and without candidate germline susceptibility variants. The VAAST score threshold is a tunable parameter. Other association tests can be supported, but require combining the association test results with a variant classifier – such as SIFT^{25,26}, PolyPhen-2²⁶, Align-GVGD^{27,28}, or VAAST 2.0¹² – to identify candidate susceptibility variants. For the AIRS and SMIT tests below, set A contains the affected individuals with candidate germline susceptibility variants, and set B contains all other affected individuals.

2.2. AIRS

AIRS evaluates candidate germline susceptibility variants to test for preferential allelic imbalance. For each individual i at site j , we use the raw somatic read counts for the reference and non-reference allele for each germline heterozygous to calculate the binomial one-tail probability, p_{ij} , that the allele frequency of the non-reference allele is greater than 0.5. To control for inter-sample variation in the distribution of allelic imbalance throughout the genome, we transform p_{ij} to the percentile rank, f_{ij} , using the empirical distribution function of binomial p-values among all variant sites throughout the genome for each individual. This transformation does not necessarily require whole-genome data and should effectively control for inter-sample variation in genome-wide levels of allelic imbalance in targeted gene panels that include as few as 50 genes. To control for differences in the level and distribution of allelic imbalance throughout the genome, we restrict the test to variants in or around the gene of interest (by default, all variants between the beginning of the first and the end of the last exon). Let G equal the set of variants around the gene, and let C equal the subset of candidate germline susceptibility variants. Our test statistic is a Wilcoxon-Mann-Whitney U that compares values of f_{ij} for candidate variants to all other variants in the gene among individuals that do not carry a candidate germline variant:

$$U = \sum_{i \in A} \sum_{j \in C \cap \{f_i\}} \sum_{k \in B} \sum_{l \in G \cap \{f_k\}} I(f_{ij} > f_{kl}) - \frac{v_A(v_A + 1)}{2}, \quad (1)$$

where v_A is the total number of candidate alleles. When the sample size of either group is under 20, the exact one-tail null probability is calculated. Otherwise, a normal

approximation is assumed.

Although we include only candidate germline alleles from individuals in set A, we include all heterozygous germline alleles from individuals in set B. Including multiple variants from an individual is a violation of the independence assumption in the U test, given that the observation of allelic imbalance in one variant would alter the expected distribution of read counts for other variants in the region. However, in our tests, we observed a modest increase in power and no inflation in Type I error by including all variants from B for sample sizes as small as 40. The inclusion of all heterozygous germline alleles is designed to detect subtle signals of allelic imbalance resulting from low levels of tumor purity or multiclonality. If the allelic imbalance signals are infrequent yet unambiguous, a more powerful alternative is to only include alleles in the rank sum test that are on the tails of the binomial distribution (e.g., p_{ij} less than 0.05 or greater than 0.95). These thresholds can be set as optional parameters.

We evaluated two allelic imbalance metrics other than the binomial, the proportion of non-reference alleles and a one-sided Fisher's exact test comparing read counts between normal and somatic tissue. The proportion of non-reference alleles suffered from an inability to account for differences in coverage depth. The Fisher's exact test had the advantage of controlling for allele-specific read count biases that are present in both the normal and somatic data, but this was offset by a modest reduction in power. More sophisticated methods that incorporate haplotype information to test for allelic imbalance, such as Haplotype Amplification in Tumor Sequences (HATS)²⁹ or Haplotype LOH (hapLOH)³⁰, may provide a replacement to the binomial in the future. In all cases, the raw allelic imbalance metric should be transformed using the empirical distribution function for each individual to control for inter-sample variation in the level of allelic imbalance throughout the genome.

2.3. SMIT

SMIT is designed to evaluate whether somatic mutations occur more or less frequently than expected for individuals with a candidate germline susceptibility variant in a gene or pathway of interest. More generally, SMIT tests for statistical interaction between somatic mutation frequencies and any binary classifier in a defined genomic feature. SMIT addresses the same general question as the Clinical Correlation Test (CCT) in the Mutational Significance of Cancer package (MuSiC)³¹, but provides the additional advantage of controlling for inter-sample variation in the somatic background mutation rate. Because the same genomic regions are evaluated in the two sample groups, the method is robust to heterogeneity in the mutational process between genomic regions, which is a major potential source of false-positives when searching for cancer-associated genes¹⁰.

Let M equal the set of individuals with at least one somatic mutation observed in the genomic feature (typically gene). Let t_i equal the total number of somatic mutations throughout the genome for sample i , and let l equal the proportional length of the gene in base pairs relative to the total sequenced region of the genome. For each sample i , we estimate the background mutation rate at the gene by the approximation $r_i = t_i \times l$. Let s_A and s_B equal the probability for sets A (affected individuals with candidate germline susceptibility variants) and B (all other individuals),

respectively, that a somatic mutation occurs in the gene through a process that is unrelated to the background mutation rate, which approximates the somatic driver mutation rate. SMIT tests the null hypothesis that $s_A = s_B$ against the alternative hypothesis that $s_A \neq s_B$ using a likelihood ratio test:

$$\Lambda = \frac{\prod_{i \in M} r_i + (1-r_i)\hat{s} \prod_{i \notin M} (1-r_i)(1-\hat{s})}{\prod_{i \in A \cap M} r_i + (1-r_i)\hat{s}_A \prod_{i \in \{x \in A | x \notin M\}} (1-r_i)(1-\hat{s}_A) \prod_{i \in B \cap M} r_i + (1-r_i)\hat{s}_B \prod_{i \in \{x \in B | x \notin M\}} (1-r_i)(1-\hat{s}_B)}. \quad (2)$$

We estimate the maximum likelihood of s , s_A , and s_B using a grid search. Note that when r_i does not vary between samples and the maximum likelihood of s , s_A , and s_B are all greater than 0, Eq. 2 collapses to a multinomial likelihood ratio test. We estimate the significance level of the two-tailed test using a chi-square approximation ($-2\ln\Lambda \sim \chi^2_1$). We also implement one-sided tests by applying the appropriate transformations to the significance levels of the two-sided test. The one-tailed test $s_A > s_B$ evaluates a cellular recessive (or partially recessive) two-hit hypothesis and the one-tailed test $s_B > s_A$ evaluates a cellular dominant (or partially dominant) one-hit hypothesis.

2.4. Somatic-Germline Interaction (SGI) Tool

SGI implements both AIRS and SMIT, and also combines the two tests to evaluate two- and three-hit hypotheses using a Fishers Combined Probability Test (FCPT). We refer to the combined AIRS-SMIT test as the SGI test. We also use the FCPT to perform the VAAST-AIRS, VAAST-SMIT, and VAAST-SGI tests in Figure 4.

2.5. Datasets

The breast cancer samples used in Figures 1 and 3 are from Complete Genomics (CG) whole-genome sequence data of a tumor-normal pair³². This sample exhibited high levels of allelic imbalance throughout the genome, with 77% of heterozygous germline SNPs having a somatic allele frequency significantly different from 0.5 at the 0.05 level. In Figures 1 and 4, we used the breast cancer sequence data to establish a distribution of read counts to represent next-generation sequence data in tumors. For individuals without candidate germline susceptibility alleles (group B), we sampled 50 Kb segments with replacement from the breast cancer whole-genome data. To represent the marker density of whole-exome data (approximately 2% of the genome), we performed rejection sampling on each heterozygous germline variant, rejecting each variant with probability 0.98. The top half of Figure 1 was based on the tumor tissue data and represents loci with relatively high levels of allelic imbalance in group B. The bottom half of Figure 1 was based on the normal tissue data and represents loci with very low levels of allelic imbalance in group B. For candidate germline susceptibility alleles in Figure 1 (group A), we simulated the distribution of read counts using the following procedure: For each candidate germline variant in each individual, we first designated it as preferentially amplified with probability q (between 0.1 and 1). Note that the proportion of samples with higher frequency of the preferred allele is approximately

$q+(1-q)b$, where b is the proportion of variants with a higher frequency for the preferred allele in group B (Figure 1). The read counts of alleles not designated as preferentially amplified were randomly sampled from breast cancer whole-genome data. For the remaining variants, we set the total number of reads to a Poisson random variate, t , with mean equal to 52 to match the mean read count in the normal tissue whole-genome data. We then set the expected proportion of the preferred allele, w , to between 0.6 and 1 and the number of non-reference to a binomial random variate with parameters t and w . In Figure 4, the breast cancer *ATM* case-control sequence data in Figure 4 is from a meta-analysis described in¹³. The genomic variants in group B were simulated by sampling 50 Kb segments from the breast cancer whole-genome data, and the variants in group A were simulated using the same protocol as Figure 1, with w equal to 1.

3. Results

We evaluated the performance of AIRS, SMIT, and SGI across a range of parameter values using a combination of simulated data and bootstrapped next-generation sequencing datasets (see Methods). In each comparison, we divide the cases into two groups, the normal group and the candidate germline group, representing individuals with and without candidate germline susceptibility variants, respectively.

To benchmark AIRS, we simulated the distribution of read counts according to the parameters in Figure 1 for the candidate germline group. For the normal group, we sampled whole-genome sequence data from the breast cancer tumor-normal pair. We evaluated two scenarios for the normal group, one with very low rates of allelic imbalance and one with relatively high levels of imbalance (see Figure 1). When the level of allelic imbalance in the normal group is low, preferential allelic imbalance in the candidate germline group is easier to detect, but AIRS performs well in both scenarios when 40 or more individuals are included in the candidate germline group or when the proportion imbalanced reads for the preferred allele is high. For example, with complete amplification of the preferred allele, AIRS has approximately 99% power

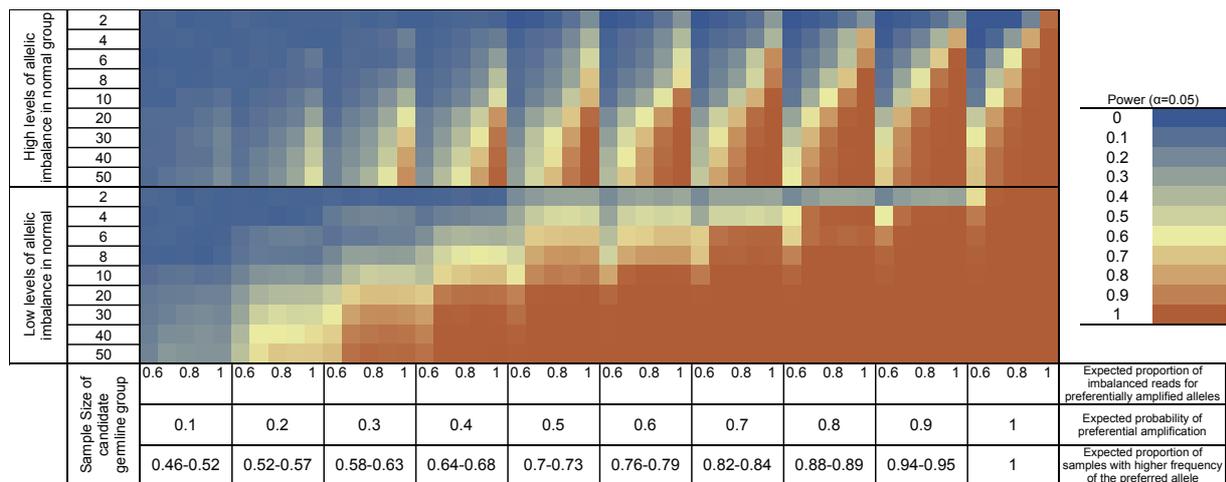


Figure 1. Power of AIRS to detect preferential allelic imbalance with α of 0.05. Sample size for group A ranged from 2 to 50. Sample size for group B was 200. The expected proportion of preferential amplification summarizes information about both the proportion of true positive germline susceptibility variants and the proportion of true positives that are preferentially amplified.

at the significance level (α) of 0.05 with a sample size of only two individuals in the candidate germline group for both scenarios. Because AIRS is designed to detect preferential allelic imbalance, it cannot be used to search for genes that tend to follow a one-hit model.

The performance of SMIT depends heavily on the frequency of somatic mutations in the gene. When the mutation frequency is high in the normal group (e.g. 0.5 for *APC* and colorectal cancer)¹¹, SMIT can detect both relative increases and decreases in the candidate germline group (Figure 2). In contrast, when the mutation frequency is very low in the normal group, SMIT can only detect mutation frequency increases in the candidate germline group. Thus, genes that follow a one-hit model can only be detected if somatic mutations are common or if the sample sizes are large. In contrast, genes that strictly follow a 2-hit model can be detected with nearly 100% power at $\alpha = 0.05$ with sample sizes of just 10 individuals in the candidate germline group, although the detection of subtle increases in mutation frequency require substantially larger sample sizes.

SMIT is designed to control for inter-sample variation in background mutation rates between samples, which can vary by three orders of magnitude¹⁰. Systematic differences in background mutation rates between the candidate germline group and control group can result from random sampling or differences in sample collection strategies. To investigate this problem, we performed simulations with identical somatic driver mutation rates but highly differentiated background somatic mutation rates between the candidate germline group and the control group. We found SMIT properly controlled for Type I error (Figure 3A). In comparison, a Fisher exact test (e.g. CCT in MuSiC³¹) exhibited a highly inflated Type I error rate (Figure 3B).

SIGI is designed to interface with VAAST to increase the power of a rare variant association study by combining case-control and tumor-normal pair sequence data. To demonstrate the utility of this approach, we analyzed a breast cancer case-control sequencing dataset of the gene *ATM* in VAAST, and then applied SIGI to evaluate the potential change in performance. We set the number of individuals in the candidate germline group equal to the number of individual variants that had a positive VAAST score from the *ATM* results. We set the frequency of somatic mutations in the normal group to 5%, which is the reported frequency of *ATM* mutations in basal-like breast cancer⁹. For the candidate germline group, we varied the frequency of somatic mutations in the candidate germline group from 0 to 0.5 and set the frequency of preferential allelic imbalance

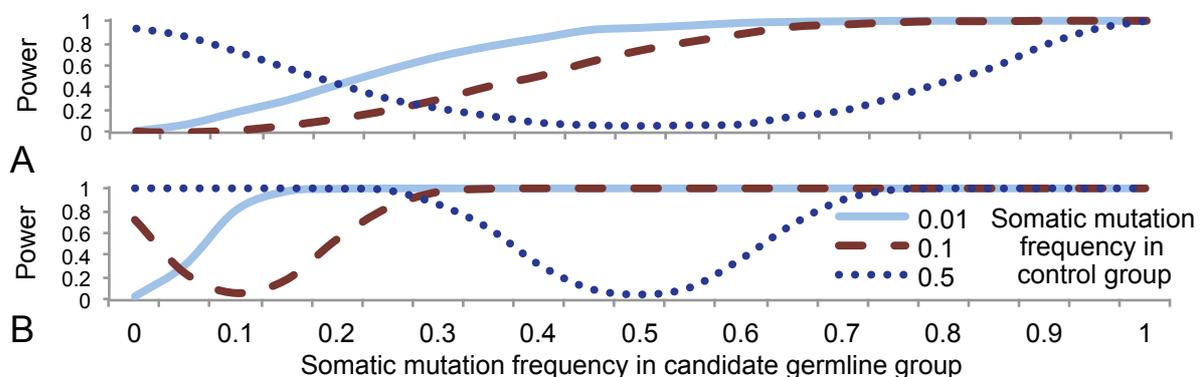


Figure 2. Power of SMIT to detect statistical interaction between germline variation and somatic small indel and point mutations at $\alpha = 0.05$. **A)** Sample size of group A is 10 individuals, and **B)** sample size of group A is 100 individuals. Sample size for group B was 200.

equal to the somatic mutation frequency. For each individual, preferential allelic imbalance and somatic mutation were mutually exclusive events. Figure 4 reports the sample size needed to achieve 80% power using SGI alone (4A-4B) and in combined VAAST-SGI analyses (4C). When somatic mutational events are common, combining VAAST with SGI can result in dramatic reductions in required sample sizes.

4. Discussion

SGI incorporates several measures to avoid artifactual findings that can result from studies of somatic mutational events due to heterogeneity in the mutational process¹⁰. Because all comparisons are restricted to the same genomic regions, we avoid issues resulting from regional variation in mutation rates across the genome, which is the most critical source of mutational heterogeneity¹⁰. The transformation of binomial probabilities to empirical probabilities for each individual in the AIRS test allows subtle signals from low purity tumor samples to be combined with stronger signals from pure tumor samples while preserving power and controlling for inter-sample variation in genome-wide levels of allelic imbalance. AIRS is comparable to the Amplification Distortion Test (ADT) in that both tests are designed to detect preferential allelic imbalance, with AIRS designed for next-generation sequence data and ADT designed for high-density SNP microarray data³³. SMIT tests for differences in the frequency of somatic mutational events between two groups at the same locus. SMIT performs the same role as the CCT test in MuSiC³¹, but additionally controls for inter-sample variation by incorporating sample-specific background mutation rates.

The tests we present here are well powered for a broad range of realistic scenarios. Studies of preferential allelic imbalance have reported the proportion of samples with higher frequency of the preferred allele of over 60% in colorectal cancer for a common susceptibility SNP at 8q24.21²¹, 70% in colorectal cancer tumors for a familial susceptibility variant in *AURKA*²⁰, over 80% in glioblastoma for common susceptibility SNPs in the *LHFPL3* gene¹⁵, and 80%, 90%, and 100%, respectively, for skin tumor susceptibility haplotypes in *Skts6*, *Skts1*, and *Skts2* in mice²²⁻²⁴. Figure

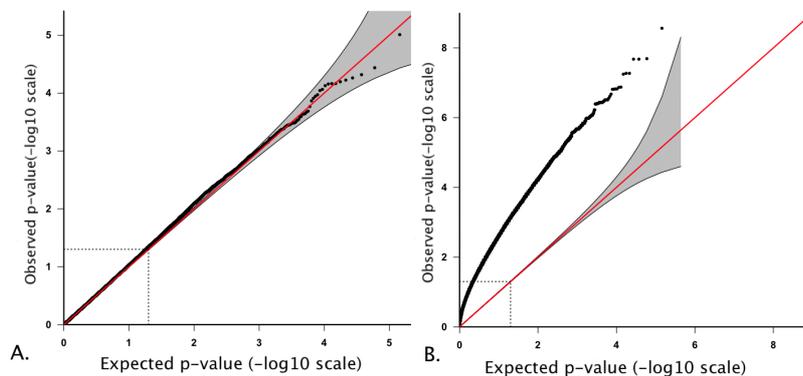


Figure 3. Observed versus expected p-values of two tests of germline-somatic interaction: **A)** SMIT and **B)** Fisher exact test (e.g. CCT³¹). Expected background mutation rate was 0.01 in the control group and 0.1 in the candidate germline group. Somatic driver mutation rate was 0.1 in both groups. Results generated from 100,000 simulations.

1 demonstrates that the sample sizes needed for AIRS to detect such signals are generally modest. For example, when 70% of samples have a higher frequency of the preferred allele, AIRS can detect preferential allelic imbalance with over 85% power from a sample of 20 individuals with germline susceptibility variants and a comparison group of 200 individuals. Unlike allelic imbalance, which can be detected from SNP microarray data, most somatic mutations can only be detected with sequence data, and thus, fewer studies of somatic mutation-germline interaction have been conducted. However, promising examples include a 10-fold increase in somatic mutations (from approximately 5% to approximately 50%) in a specific region of *APC* among carriers of a particular germline susceptibility variant in human colorectal cancer³⁴, and an 88% somatic mutation rate in carriers of the *Skt2* susceptibility haplotype in mice²³. Both scenarios could be detected by SMIT with greater than 80% power with a sample of only 10 individuals with candidate germline variants and a comparison group of 200 individuals (Figure 2).

The example of *ATM* and breast cancer in Figure 4 provides an illustration of how SGI can be combined with VAAST to identify novel cancer-gene associations and to yield new insights for known associations. *ATM* is not a classic two-hit tumor suppressor gene. Some rare missense germline variants have a dominant gain-of-function effect, and nonsense germline variants are reported to primarily increase the risk of breast cancer via haplo-insufficiency^{13,35}. However, reports of rare *ATM* germline mutations and loss of the wild-type allele in tumors³⁵ is suggestive of potential germline-somatic interaction with *ATM* and breast cancer³⁶. Figure 4 illustrates the sample size needed to conclusively detect a somatic-germline interaction effect using AIRS, SMIT, or SGI given a range of possible effect sizes. Figure 4 also demonstrates how SGI can be combined with VAAST to reduce the sample size needed to identify novel cancer-gene associations from next-generation sequence data for genes that have patterns of variation similar to *ATM* in breast cancer.

In a number of reported scenarios, preferential amplification tends to occur in conjunction with somatic mutations in a three-hit model, involving a germline susceptibility variant, a somatic point mutation on the same haplotype, and a subsequent CNV or LOH event that amplifies both the germline and somatic variants^{16-18,23}. *JAK2* and myeloproliferative neoplasms provides one such example. Somatic mutations preferentially occur on haplotypes with germline risk variants in *JAK2* 80% of the time, and frequent third-hit somatic events result in homozygosity for both the

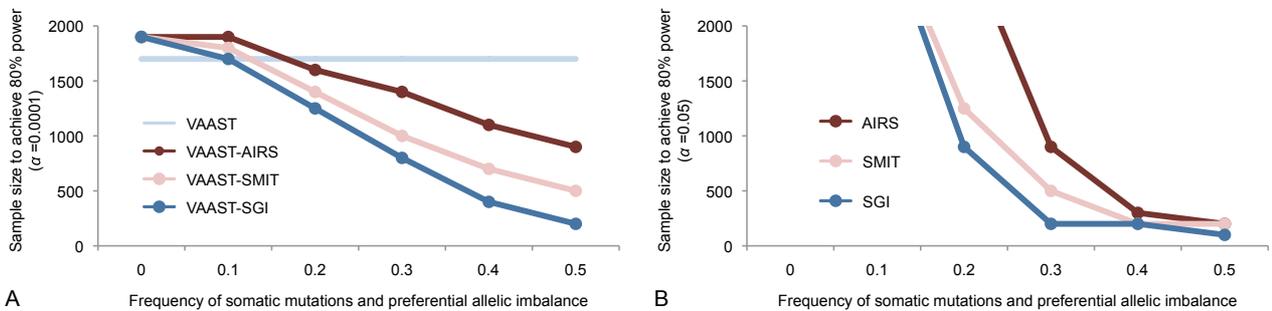


Figure 4. Sample size needed to achieve 80% power to detect **A)** cancer-germline association at $\alpha = 1 \times 10^{-4}$ and **B)** somatic-germline interaction at $\alpha = 0.05$. Sample size indicates the number of controls, normal tissue samples from cases, and paired tumor tissue samples from cases.

germline risk allele and the somatic mutation¹⁶⁻¹⁸. One potential explanation for this three-hit model is that *cis*-acting germline variants create a hypermutable region of the gene and that the subsequent somatic mutations are then amplified by the second somatic event, after which the somatic mutation (driver) and germline variant (passenger) increase in frequency together by selection¹⁵⁻¹⁸. This mechanism has been demonstrated experimentally in mice²³, and the T->A germline variant at *APC* nucleotide position 3920 has been reported as an example of a *cis*-acting hypermutable phenotype that leads to colorectal cancer in humans³⁴. A second explanation for this three-hit model is that somatic mutations functionally interact in *cis* with specific germline variants and require the presence of a germline variant to promote tumorigenesis¹⁵⁻¹⁸. By combining evidence for preferential allelic imbalance and the occurrence of somatic mutations, SGI is well suited for detecting genes that follow a three-hit model.

SGI is designed to detect statistical interaction between somatic mutational events and germline variation from next-generation sequence data. SGI is compatible with existing variant call formats (vcf and CG tsv) and interfaces with VAAST and other variant classifiers to identify candidate germline susceptibility variants in a high-throughput manner. The AIRS test evaluates evidence for preferential allelic imbalance from next-generation sequence data and allows for combined testing of multiple variants in a gene while controlling for inter-sample variation in tumor purity and genome-wide levels of allelic imbalance. SMIT evaluates evidence for statistical interaction between candidate germline susceptibility variants and somatic SNVs and small indels while controlling for inter-sample variation in background mutation rates. SGI combines AIRS and SMIT to provide a unified measure of statistical interaction between candidate germline susceptibility variants and the occurrence of somatic mutational events. SGI can be used to help demonstrate a causal role for candidate germline susceptibility variants or can be combined with rare-variant association tests to increase the power to identify cancer-gene associations.

5. Software

SGI can be found at www.hufflab.org/software/#sgi and is freely available for academic use.

Acknowledgments

We thank Sean Tavtigian for providing the *ATM* case-control data. HH was supported by the MD Anderson Cancer Center Odyssey Program. An allocation of computer time on the UT MD Anderson Research Computing High Performance Computing (HPC) facility is gratefully acknowledged.

References

1. A. G. Knudson, Jr., *Proc Natl Acad Sci U S A.* **68**, 820-3 (1971).
2. S. H. Friend, R. Bernards, S. Rogelj, R. A. Weinberg, et al., *Nature.* **323**, 643-6 (1986).
3. W. H. Lee, R. Bookstein, F. Hong, L. J. Young, et al., *Science.* **235**, 1394-9 (1987).
4. D. Malkin, *Genes Cancer.* **2**, 475-84 (2011).
5. T. A. King, W. Li, E. Brogi, C. J. Yee, et al., *Ann Surg Oncol.* **14**, 2510-8 (2007).

6. C. J. Ceol, D. Pellman and L. I. Zon, *Nat Med.* **13**, 1286-7 (2007).
7. A. Tighe, V. L. Johnson and S. S. Taylor, *J Cell Sci.* **117**, 6339-53 (2004).
8. J. Soh, N. Okumura, W. W. Lockwood, H. Yamamoto, et al., *PLoS One.* **4**, e7464 (2009).
9. *Nature.* **490**, 61-70 (2012).
10. M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, et al., *Nature.* (2013).
11. *Nature.* **487**, 330-7 (2012).
12. H. Hu, Huff, C.D., Moore, B., Flygare, S., Reese, M.G., Yandell, M, *Genetic Epidemiology.* (in press).
13. S. V. Tavtigian, P. J. Oefner, D. Babikyan, A. Hartmann, et al., *Am J Hum Genet.* **85**, 427-46 (2009).
14. F. Le Calvez-Kelm, F. Lesueur, F. Damiola, M. Vallee, et al., *Breast Cancer Res.* **13**, R6 (2011).
15. T. LaFramboise, N. Dewal, K. Wilkins, I. Pe'er, et al., *PLoS Genet.* **6**, e1001086 (2010).
16. D. Olcaydu, A. Harutyunyan, R. Jager, T. Berg, et al., *Nat Genet.* **41**, 450-4 (2009).
17. O. Kilpivaara, S. Mukherjee, A. M. Schram, M. Wadleigh, et al., *Nat Genet.* **41**, 455-9 (2009).
18. A. V. Jones, A. Chase, R. T. Silver, D. Oscier, et al., *Nat Genet.* **41**, 446-9 (2009).
19. M. Yandell, C. Huff, H. Hu, M. Singleton, et al., *Genome Res.* **21**, 1529-42 (2011).
20. T. Hienonen, R. Salovaara, J. P. Mecklin, H. Jarvinen, et al., *Int J Cancer.* **118**, 505-8 (2006).
21. S. Tuupanen, I. Niittymaki, K. Nousiainen, S. Vanharanta, et al., *Cancer Res.* **68**, 14-7 (2008).
22. J. P. de Koning, Y. Wakabayashi, H. Nagase, J. H. Mao, et al., *Oncogene.* **26**, 4171-8 (2007).
23. H. Nagase, J. H. Mao and A. Balmain, *Cancer Res.* **63**, 4849-53 (2003).
24. A. Ewart-Toland, P. Briassouli, J. P. de Koning, J. H. Mao, et al., *Nat Genet.* **34**, 403-12 (2003).
25. P. C. Ng and S. Henikoff, *Annu Rev Genomics Hum Genet.* **7**, 61-80 (2006).
26. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, et al., *Nat Methods.* **7**, 248-9 (2010).
27. S. V. Tavtigian, A. M. Deffenbaugh, L. Yin, T. Judkins, et al., *J Med Genet.* **43**, 295-305 (2006).
28. E. Mathe, M. Olivier, S. Kato, C. Ishioka, et al., *Nucleic Acids Res.* **34**, 1317-25 (2006).
29. N. Dewal, Y. Hu, M. L. Freedman, T. Laframboise, et al., *Genome Res.* **22**, 362-74 (2012).
30. S. Vattathil and P. Scheet, *Genome Res.* **23**, 152-8 (2013).
31. N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, et al., *Genome Res.* **22**, 1589-98 (2012).
32. R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, et al., *Science.* **327**, 78-81 (2010).
33. N. Dewal, M. L. Freedman, T. LaFramboise and I. Pe'er, *Bioinformatics.* **26**, 518-28 (2010).
34. S. J. Laken, G. M. Petersen, S. B. Gruber, C. Oddoux, et al., *Nat Genet.* **17**, 79-83 (1997).
35. J. O. Bay, N. Uhrhammer, D. Pernin, N. Presneau, et al., *Hum Mutat.* **14**, 485-92 (1999).
36. C. Schaffner, I. Idler, S. Stilgenbauer, H. Dohner, et al., *Proc Natl Acad Sci U S A.* **97**, 2773-8 (2000).