2012

VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing

Daniel C. Koboldt
Washington University School of Medicine in St. Louis

Qunyuan Zhang
Washington University School of Medicine in St. Louis

David E. Larson
Washington University School of Medicine in St. Louis

Dong Shen
Washington University School of Medicine in St. Louis

Michael D. McLellan
Washington University School of Medicine in St. Louis

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
Koboldt, Daniel C.; Zhang, Qunyuan; Larson, David E.; Shen, Dong; McLellan, Michael D.; Lin, Ling; Miller, Christopher A.; Mardis, Elaine R.; Ding, Li; and Wilson, Richard K., "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing." Genome Research.22,. 568-576. (2012).
https://digitalcommons.wustl.edu/open_access_pubs/1885

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing

Daniel C. Koboldt, Quanyuan Zhang, David E. Larson, et al.

Genome Res. 2012 22: 568-576 originally published online February 2, 2012
Access the most recent version at doi:10.1101/gr.129684.111

Supplemental Material
http://genome.cshlp.org/content/suppl/2012/01/11/gr.129684.111.DC1.html

References
This article cites 28 articles, 8 of which can be accessed free at:
http://genome.cshlp.org/content/22/3/568.full.html#ref-list-1

Creative Commons License
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions

Copyright © 2012 by Cold Spring Harbor Laboratory Press
VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing

Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson

Cancer is a disease driven by genetic variation and mutation. Exome sequencing can be utilized for discovering these variants and mutations across hundreds of tumors. Here we present an analysis tool, VarScan 2, for the detection of somatic mutations and copy number alterations (CNAs) in exome data from tumor-normal pairs. Unlike most current approaches, our algorithm reads data from both samples simultaneously; a heuristic and statistical algorithm detects sequence variants and classifies them by somatic status (germline, somatic, or LOH); while a comparison of normalized read depth delineates relative copy number changes. We apply these methods to the analysis of exome sequence data from 151 high-grade ovarian tumors characterized as part of the Cancer Genome Atlas (TCGA). We validated some 7790 somatic coding mutations, achieving 93% sensitivity and 85% precision for single nucleotide variant (SNV) detection. Exome-based CNA analysis identified 29 large-scale alterations and 619 focal events per tumor on average. As in our previous analysis of these data, we observed frequent amplification of oncogenes (e.g., CCNE1, MYC, and CDKN2A). We searched for additional recurrent focal CNAs using the correlation matrix diagonal segmentation (CMDS) algorithm, which identified 424 significant events affecting 582 genes. Taken together, our results demonstrate the robust performance of VarScan 2 for somatic mutation and CNA detection and shed new light on the landscape of genetic alterations in ovarian cancer.

[Supplemental material is available for this article.]
cinoma (HGS-OVCa) that were initially characterized within the Cancer Genome Atlas (TCGA) project (Cancer Genome Atlas Research Network 2011). We present a robust pipeline for the detection of both germline (inherited) and somatic (acquired) mutations by exome sequencing and describe filtering approaches for detecting variants with high sensitivity and specificity. To evaluate the performance of our SCNA detection algorithm, we compare our results to copy number data from high-density SNP array and WGS approaches. Our results demonstrate the accuracy of VarScan 2 for somatic mutation and CNA detection and enable a new survey of the genetic landscape in ovarian carcinoma.

Results

The VarScan 2 algorithm reads SAM tools pileup or mpileup output from tumor and normal samples simultaneously, performing pairwise comparisons of base calls and normalized sequence depth at each position (Fig. 1). For variant detection, a heuristic algorithm determines the genotype for normal and tumor samples independently based on adjustable minimum thresholds for coverage, base quality, variant allele frequency, and statistical significance. In single samples, the latter value is computed by Fisher's exact test of the read counts supporting each allele (reference and variant) compared to the expected distribution based on sequencing error alone. By default, VarScan 2 requires a minimum coverage of $3 \times$, minimum $p$-value of 20, allele frequency of at least 8%, and a $P$-value of $<0.05$. Variants with a variant allele frequency of $>75\%$ are called homozygous. These represent the initially recommended parameters, and they are fully adjustable by the user.

At every position where one or both samples had a variant, VarScan performs a direct comparison between tumor and normal genotypes (heuristic) and supporting read counts (Fisher's exact test) to determine the somatic status. Variants present in both samples are classified as somatic (acquired), variants heterozygous in the normal but homozygous in the tumor are classified as LOH, and variants shared between samples are classified as germline (inherited). To further refine these predictions, we developed a false-positive filter that removes likely false positives due to sequencing- or alignment-related artifacts. The filter evaluates each variant for nine empirically derived criteria to distinguish true variants from probable artifactual calls (Table 1; Supplemental Fig. 1).

To identify SCNAs, VarScan 2 compares Q20 (base quality $\geq 20$) read depths between tumor and normal samples for contiguous regions of coverage. After normalizing for the amount of input data (unique bases mapped), the relative copy number change is inferred as the log$_2$ of the ratio of tumor depth to normal depth for each contiguous region. The output of this algorithm—a set of regions, each with defined start and stop positions and a log$_2$ ratio representing the copy number change in the tumor—is similar to hybridization-based copy number data and amenable to the same segmentation methods. Therefore, we apply a circular binary segmentation (CBS) algorithm (Seshan and Olshen 2010) to delineate segments by copy number and identify significant change-points. A subsequent joining procedure merged adjacent segments of similar copy number and classified them as either large-scale ($>25\%$ of chromosome arm) or focal events (see Methods).

Application to 151 ovarian cancer tumor–normal pairs

To evaluate our methods, we applied them to exome data for tumor samples and matched normals from 151 serous ovarian carcinomas that we previously characterized (Cancer Genome Atlas Research Network 2011) as part of the Cancer Genome Atlas (Table 2). On average, we identified 18,462 coding SNVs per tumor, of which 16,340 (88.5%) were germline variants, 2013 (10.9%) were LOH events, and 109 (0.59%) were somatic mutations (Fig. 1). Gapped alignments of the relatively long (76–100 bp), paired-end reads in our data set also permitted the identification of small indels ranging in size from 1–55 bp. On average, we detected 418 coding indels per exome, of which 387 (92.6%) were germline variants, 23 (5.5%) were LOH, and eight (1.91%) were somatic mutations.

We also applied VarScan 2 to detect SCNAs. To minimize the effect of variable coverage between tumor and normal samples, we excluded nine samples with $<50\%$ of target CDS bases covered at greater than 20-fold, and focused our analysis on 142 tumor–normal pairs with sufficient sequence coverage (Supplemental Information; Supplemental Table 1). On average across 142 patient samples, we detected 29 large-scale events (eight gains and 21 losses) and 619 focal events (301 gains and 318 losses) per tumor exome (Fig. 1; Supplemental Table 5).

Comparison of germline variants to high-density SNP array genotypes

To evaluate the performance of our mutation detection approach, we utilized orthogonal validation data from two sources. First, to evaluate the accuracy of germline variant detection, we compared VarScan 2 consensus genotypes for exome data to high-density SNP array data made available by TCGA (Cancer Genome Atlas Research Network 2011). To minimize the influence of sequence
Such discrepancies (615, or 48%) had less than 20 coverage and are likely true homozygotes, but roughly half of all
coverage, we compared only sites that achieved eightfold coverage or higher in either the tumor or matched normal sample, and ex-
amined only those within the ~33-Mbp CDS target region. On average, each sample had 5425 germ line SNPs with informative
array genotypes and sufficient coverage in the exome. Genotype
concordance between VarScan 2 consensus genotypes and array
genotypes was 99.56% (Supplemental Information; Supplemental
Fig. 2A), supporting a high accuracy for germ line variant detection
by VarScan 2.
To better understand concordance metrics, we investigated
the 2854 discrepancies between array and exome data (Supple-
mental Fig. 2B). Of these, 27% of these were genotype as refer-
cence (wild type) by array but called heterozygous (21%) or homozygous
variant (6%) in the sequencing data. On average, these sites
achieved high depth (206× and 129×, respectively) and variant
allele frequencies (46.7% and 97.0%) by exome sequencing (Sup-
plemental Fig. 3). This suggests that many are true variants missed
by the array, possibly due to misclustering or allele dropout
(Koboldt et al. 2006). Another 17% of discrepancies were hetero-
ygous on both platforms, but the variant allele observed in se-
quencing data was different from that reported by SNP array. An
examination of these revealed that the majority (83.3%) were re-
verse-complementary allele combinations (e.g., G/A variant in
exome data genotyped as C/T); most likely, the strand orientation
reported for the SNP array genotype was incorrect. The most
common discrepancy (45%) occurred at sites called heterozygous
by array but homozygous variant using exome data. Some 105 of
these (8.2%) had 100% variant allele frequency with 20× or more
coverage and are likely true homozygotes, but roughly half of all
such discrepancies (615, or 48%) had less than 20× coverage in
exome data and likely reflect an allelic bias favoring the variant. We
conclude that germ line variants called using exome data are highly
accurate (99.56%), and a significant fraction of the discrepancies
can be attributed to array genotyping error or imbalanced allelic
representation in the sequence data.
Finally, to investigate the portion of the exome that is callable
by our method, we computed the fraction of CDS bases covered by
at least 20 reads (Supplemental Table 1) in each sample. As an esti-
mate of overall sensitivity for coding variants, we also determined
the proportion of heterozygous SNPs (by array genotype) in the
CDS target that were detected by VarScan 2 as germ line variants,
regardless of the coverage of those positions (Supplemental Fig. 4).

Table 1. Empirically derived filtering parameters for putative somatic mutations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read position</td>
<td>Average variant position in supporting reads, relative to read length</td>
<td>Between 10 and 90</td>
</tr>
<tr>
<td>Strandedness</td>
<td>Fraction of supporting reads from the forward strand</td>
<td>Between 1%–99%</td>
</tr>
<tr>
<td>Variant reads</td>
<td>Total number of reads supporting the variant</td>
<td>At least four</td>
</tr>
<tr>
<td>Variant frequency</td>
<td>Variant allele frequency inferred from read counts</td>
<td>At least 5%</td>
</tr>
<tr>
<td>Distance to 3′</td>
<td>Average distance to effective 3′ end of variant position in supporting reads</td>
<td>At least 20</td>
</tr>
<tr>
<td>Homopolymer</td>
<td>Number of bases in a flanking homopolymer matching one allele</td>
<td>Less than five</td>
</tr>
<tr>
<td>Map quality difference</td>
<td>Difference in average mapping quality between reference and variant reads</td>
<td>Less than 30</td>
</tr>
<tr>
<td>Read length difference</td>
<td>Difference in average trimmed read length between reference and variant reads</td>
<td>Less than 25</td>
</tr>
<tr>
<td>MMQS difference</td>
<td>Difference in average mismatch quality sum between variant and reference reads</td>
<td>Less than 100</td>
</tr>
</tbody>
</table>

Table 2. Exome sequencing data set summary

<table>
<thead>
<tr>
<th></th>
<th>WU</th>
<th>BI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exome platform</td>
<td>SureSelect (Agilent)</td>
<td>SureSelect (Agilent)</td>
</tr>
<tr>
<td>Genes targeted</td>
<td>18,568</td>
<td>18,568</td>
</tr>
<tr>
<td>Exons targeted</td>
<td>188,260</td>
<td>188,260</td>
</tr>
<tr>
<td>CDS target size</td>
<td>33 Mbp</td>
<td>33 Mbp</td>
</tr>
<tr>
<td>Sequencing platform</td>
<td>Illumina GAIIx</td>
<td>Illumina GAIIx</td>
</tr>
<tr>
<td>Number of patients</td>
<td>91</td>
<td>60</td>
</tr>
<tr>
<td>Number of samples</td>
<td>182</td>
<td>120</td>
</tr>
<tr>
<td>Read length</td>
<td>2 × 100 bp</td>
<td>2 × 76 bp</td>
</tr>
<tr>
<td>Sequence per sample</td>
<td>15.2 Gbp</td>
<td>21.9 Gbp</td>
</tr>
<tr>
<td>Average mapping rate</td>
<td>98.34%</td>
<td>81.06%</td>
</tr>
<tr>
<td>Average duplication rate</td>
<td>12.58%</td>
<td>18.50%</td>
</tr>
</tbody>
</table>

There were nine outlier samples that achieved relatively poor CDS coverage with correspondingly poor detection of heterozygous SNPs. Excluding these, we find that, on average, 79.62% of CDS bases are covered 20× or more in each sample and that 81.42% of heterozygous SNPs are detected by VarScan 2. From these metrics, we conclude that ~80% of coding sequences are “callable” by Agilent SureSelect exome sequencing and VarScan 2 analysis.

Orthogonal validation
of somatic mutations
To assess the specificity of somatic muta-
tion detection, we validated putative so-
matic coding SNVs by PCR and deep
resequencing (Table 3). Of 5871 muta-
tions for which we obtained validation data, 5225 (89.00%) were
confirmed as true somatic mutations, 572 (9.74%) were refuted as
wild type, 63 (1.07%) were germ line variants, and 11 (0.19%) were
due to LOH in the tumor. In general, exome sequence depth was
lower and observed variant allele frequency markedly reduced
among predicted SNVs that were refuted as wild type, compared
with validated somatic mutations (Supplemental Fig. 5). We also
attempted to validate 2458 putative mutations that were removed
by the false-positive filter. Of these, 292 were confirmed as valid
somatic mutations (11.88%) while 2073 (84.34%) were refuted as
wild type. Thus, our filtering strategy retained 94.71% of valid
mutations while removing 78.37% of false positives. We conclude
that this approach dramatically increases the true-positive rate of
mutation detection, with a relatively small reduction (5.3%) in
overall sensitivity.
To further assess the sensitivity of mutation detection, we
compared our predictions to validated somatic mutations reported
for 60 tumor–normal pairs that were analyzed externally (Cancer
Genome Atlas Research Network 2011). These cases harbored a
total of 3065 valid somatic mutations, of which we detected 2565
(83.7%). When we investigated the 500 valid mutations missed by
our approach, we found that 93 had been detected but deemed low
confidence (LC), 51 were high confidence (HC) but removed by
the filter, and 298 had less than fourfold coverage in one or both
samples.

There were 91 tumor–normal pairs sequenced at Washington University (WU); BAM files for an additional 60 tumor–normal pairs sequenced at the
Broad Institute of MIT and Harvard (BI) were downloaded from dbGaP.
Table 3. Estimated sensitivity and precision of mutation detection with VarScan 2 based upon orthogonal validation data

<table>
<thead>
<tr>
<th>Experimental validation (91 cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total with validation data 5871</td>
</tr>
<tr>
<td>Validated somatic 5225</td>
</tr>
<tr>
<td>Validated germline 63</td>
</tr>
<tr>
<td>Validated loss of heterozygosity 11</td>
</tr>
<tr>
<td>Validated wild type 572</td>
</tr>
<tr>
<td>Sensitivity for valid somatic mutations 92.30%</td>
</tr>
<tr>
<td>Precision of mutation calls 89.00%</td>
</tr>
<tr>
<td>Detection of reported mutations (60 external cases)</td>
</tr>
<tr>
<td>Valid somatic mutations reported 3065</td>
</tr>
<tr>
<td>Valid somatic mutations with coverage 2773</td>
</tr>
<tr>
<td>Detected by VarScan 2565</td>
</tr>
<tr>
<td>Detection sensitivity 83.69%</td>
</tr>
<tr>
<td>Adjusted sensitivity 92.50%</td>
</tr>
</tbody>
</table>

Sensitivity refers to the fraction of known, validated somatic mutations that were detected, whereas precision reflects the proportion of detected SNVs that were validated. In externally analyzed samples, we also compute the “adjusted sensitivity,” which reflects the detection of known, validated somatic mutations with sufficient coverage in the BAM files analyzed.

For somatic mutation calling, we attempted to validate 141 HC somatic indels detected by VarScan 2. Of these, 85 (60.28%) were confirmed as somatic, 30 (21.28%) were refuted as wild type, and 26 (18.4%) were found to be germline or LOH events (Supplemental Table 8). There were also 80 validated somatic indels among the externally analyzed tumor–normal pairs. Some 73 had coverage in the BAM files that we downloaded; of these, 65 (89.04%) were detected as HC somatic mutations by VarScan 2. While more comprehensive evaluations are needed, these results suggest that our method detects somatic indels with high sensitivity (89%) but a moderate true-positive rate (60%).

Comparison to single-sample methods for somatic mutation detection

We next sought to demonstrate the superiority of our method, which compares tumor and normal samples simultaneously, to more simplistic approaches for somatic mutation calling. For this analysis, we selected five ovarian cancer cases for which both exome and whole-genome sequencing (WGS) data were available (marked in Supplemental Table 1). By using the exome data, we identified tumor-specific SNPs by a simple subtraction method (see Supplemental Methods), which yielded 152,708 candidate mutations per tumor (on average). In contrast, VarScan 2 detected 508 somatic mutations per tumor (on average) for the same data set (Supplemental Table 9).

Given the large number of calls generated by the subtraction method, it was possible that a significant fraction of these were valid somatic mutations that had not been part of our validation experiment. To investigate this possibility, we determined the fraction of mutations called in each set that were also detected in the WGS data by a different algorithm, named SomaticSniper (Larson et al. 2011). Of the 2538 HC mutations called by VarScan in exome data, 1716 (67.61%) were called by SomaticSniper in WGS data. In contrast, only 736 of 763,539 calls (0.99%) made by the subtraction method were supported by SomaticSniper calls.

Next, we evaluated the sensitivity of each method to detect the known somatic mutations described above. A total of 290 validated somatic mutations had been reported for these five cases in the TCGA study. Of these, 247 (85.17%) were found by the subtraction method, while 264 (91.03%) were detected by VarScan 2 (Supplemental Table 9). Surprisingly, these results suggest that a subtraction method may also suffer slightly lower sensitivity for valid somatic mutations, possibly due to false-positive calls in the matched normal sample. We conclude that our method for somatic mutation calling delivers comparable sensitivity and dramatically higher precision, than do simple subtraction-based approaches.

Orthogonal validation of SCNAs

To evaluate the accuracy of SCNA detection, we compared copy number data for five ovarian tumors that were assessed by high-density SNP array, exome, and WGS (see Supplemental Methods). Strikingly, the exome-based copy number estimates from our algorithm were remarkably consistent with those of array and WGS data and demonstrated an ability to detect both large-scale and focal events (see example in Fig. 2A). A systematic comparison of these three approaches is more difficult, since both array and exome data are limited to a very small fraction of the genome (SNPs and exons, respectively); only WGS yields unbiased genome-wide copy number estimates. All three approaches, however, should be able to detect large-scale gains and losses of chromosome arms because these events typically span several megabases. Thus, we compared the overlap of large-scale events from exome, array, and WGS data sets for the five cases (Fig. 2B; Supplemental Table 3). A total of 206 large-scale CNAs were detected, of which 165 (80.10%) were detected by all three approaches, suggesting that most of these represent real events. Our exome-based method predicted 185 large-scale events (90% of the total); nearly all were supported by array or WGS data sets, and 89.2% were supported by both. A visual review of events not detected by WGS or array revealed that most were present but did not meet thresholds for calling an amplification (log2 ratio > 0.20) or deletion (log2 ratio < −0.10). In contrast, the −10% of large-scale events missed by our exome method were largely due to oversegmentation in sparsely targeted regions of the genome. WGS data sets yielded the most calls overall, likely reflecting a wider and more unbiased coverage of the genome.

A similar comparison for focal copy number events is challenging, since exome and SNP array data sets survey different, noncontiguous portions of the genome. To address this, we performed a three-platform comparison of copy number events affecting coding sequences. At every exon, we determined a copy number status (amplification, deletion, or neutral) based upon the best-overlapping segment from SNP array, exome, or WGS copy number data sets. There were 677,434 copy-number–altered exons (about 135,000 per case) in the five cases at which we could make this comparison. Of these, 72.1% were detected by two platforms and 44.49% were detected by all three (Fig. 2C), suggesting that this comparison strategy is reasonably accurate given the different portions of the genome surveyed by each platform.

Our exome-based method detected 488,721 focal events (72.14% of the total), achieving higher sensitivity than the SNP array (65.39%) but lower than WGS (79.05%). This result is
somewhat unsurprising, given the limited resolution of SNP arrays and the superior coverage breadth and uniformity offered by WGS. If we consider the intersection of events supported by both the array and WGS platforms to be a gold standard, there were 395,868 such focal events, of which our method detected 301,376 (76.13%). There were 161,944 events (23.91%) detected by exome or WGS methods but not by SNP arrays, consistent with the expectation of limited resolution for the array platform. A significant portion of the focal events that we detected using exome data (80.60%) were supported by at least one other platform, suggesting that the majority are likely to be real events. This is particularly promising because the exome method detected a higher fraction of platform-specific calls (94,819), which are likely to include small focal copy number changes missed by other platforms.

Encouraged by these results, we next compared recurrent large-scale gains and losses in the 142 exome cases included in this study to those of a larger data set (489 cases, array data) analyzed by TCGA (Cancer Genome Atlas Research Network 2011). Most of the cases we studied were part of the TCGA analysis, which identified 30 recurrent large-scale alterations (eight gains and 22 losses), all of which had been reported previously. Our method identified all recurrent gains and losses reported by TCGA (Fig. 3; Supplemental Information; Supplemental Table 4; Cancer Genome Atlas Research Network 2011). Further, the frequencies of arm-level events detected in our data set and the TCGA data set were highly correlated ($r^2 = 0.84$ for gains, $0.86$ for losses), suggesting that our exome-based approach was sufficiently robust to recapitulate the results of our previous array-based findings. Taken together, the results suggest that our method identifies somatic CNAs with an accuracy comparable to array-based and WGS approaches and that our set of 142 cases is representative of the larger cohort ($n = 489$) studied by TCGA.

**Figure 2.** Detection of large-scale and focal copy number alterations by sequencing- and array-based approaches. (A) Deletions and focal amplifications of chromosome 4 in sample TCGA-24-1103. Copy number estimates from array (gray), WGS (light blue), and exome (dark blue) indicate two regions of deletion as well as a focal amplification (window). Red lines indicate segmented exome CBS calls. (Below) Variant allele frequencies in the normal (blue) and tumor (green) indicate regions of loss of heterozygosity (LOH) in deleted segments. (B) Intersection of large-scale copy number alterations detected by SNP array, whole-genome sequencing, and exome sequencing approaches for five HGS-OvCa cases. For details, see Supplemental Table 3. (C) Intersection of gene-level (focal) copy number alterations detected by SNP array, whole-genome sequencing, and exome sequencing approaches for five HGS-OvCa cases.
Identification of recurrent CNAs with CMDS

High-grade serous ovarian tumors possess highly rearranged genomes, owing in part to defects in homologous recombination pathways (Patel et al. 1998; Xu et al. 1999; Bowtell 2010; Cancer Genome Atlas Research Network 2011). Genetic alterations that promote carcinogenesis, e.g., the amplification of oncogenes and deletion of tumor-suppressor genes, are likely to be recurrent across multiple tumors. Indeed, when we analyzed the mean copy number change across 142 cases genome-wide (Fig. 4), we observed a striking pattern of recurrent copy number events. Many of these correspond to large-scale gains and losses shown in Figure 3. Focal amplifications and deletions are also apparent; the majority of these have already been reported in ovarian cancer (Cancer Genome Atlas Research Network 2011). Notable examples include amplifications MYC, CCNE1, and EVI1 (also called MECOM), as well as deletions of tumor suppressors NF1, PTEN, and CDKN2A/CDKN2B (Fig. 5).

Intriguingly, we also observed tight focal amplifications of EPH receptors EPHB3 (3q27) and EPHB4 (7q22) (Figs. 4, 5), which are known to be overexpressed in ovarian carcinoma (Alam et al. 2008) but were not significant in SNP array data for 489 cases analyzed by TCGA. We reasoned that the high-resolution, exome-centric nature of our data set might enable identification of new recurrent CNAs (RCNAs) in ovarian cancer. To identify such regions, we applied the correlation matrix diagonal segmentation (CMDS) algorithm (Zhang et al. 2010) to segmented exome-based copy number data for 142 cases. CMDS employs a population-based approach to identify statistically significant RCNAs and is particularly sensitive for focal events. Our analysis identified 424 significant focal RCNAs targeting 582 known genes (Supplemental Table 6). Gene set analysis of these 582 genes revealed 10 significantly enriched pathways \( (P < 0.0005) \) (Supplemental Table 7). Focal adhesion (23 genes, \( P = 1.22 \times 10^{-13} \)) and ECM-receptor interaction (15 genes, \( P = 7.87 \times 10^{-12} \)) were the most significant pathways, suggesting that cell–cell and cell–matrix adhesion molecules are often dysregulated in high-grade ovarian carcinoma.

Discussion

In summary, we have developed an approach for simultaneous detection of germline variants, somatic mutations, LOH, and SCNAs using exome sequence data from matched tumor and normal samples. Unlike other methods for mutation detection, our algorithm reads data from tumor and normal samples simultaneously, enabling direct pairwise comparisons of base calls at each position. For germline variant detection, we observed a high genotype concordance (99.56%) between the VarScan results and high-density SNP arrays, with a significant fraction of discordant sites attributed to imbalanced allelic representation or errors in the array data.

For somatic mutation detection, we demonstrated that our method provides similar sensitivity and a dramatically higher true-positive rate compared with more simplistic approaches that analyze tumor and normal samples independently. We also demonstrated that our filtering strategy removes the vast majority of false positives due to sequencing or alignment artifacts, while preserving sensitivity for true mutations. Indeed, in exome sequence data for 151 ovarian tumors characterized by TCGA, our approach identified 5225 valid somatic mutations with 94.71% sensitivity and an 89.00% true-positive rate. We observed a comparable sensitivity (89%) and a moderate true-positive rate (60%) for validated somatic indels, though the number of such variants in our data set...
By use of the CMDS algorithm, we identified 424 significant focal RCNAs containing some 582 genes. Gene set analysis identified focal adhesion and ECM-receptor interaction as significantly altered pathways. These findings are consistent with current knowledge of ovarian carcinoma and other epithelial cancers, in which dysregulation of cell–cell and cell–matrix signaling represents a key step in tumor development, growth, and invasion. Taken together, these results suggest that our method for exome-based CNA detection both confirms and extends the results of traditional approaches.

Importantly, the samples studied here all were processed using similar hybrid capture protocols (Agilent SureSelect) and were sequenced on the same platform (Illumina GAIIx). By comparing samples from the same individual sequenced under identical conditions, our approach to CNA detection avoids GC content and mapping biases that complicate traditional sequence-based methods. However, the CNA calling could be confounded by paired samples that were sequenced under different conditions or on different sequencing platforms. Further, although we normalized for data input (unique bases mapped), it is possible that fluctuations in capture specificity and/or sequence representation could influence sequence depth between sample pairs, which might affect our results. One possible strategy to address this would be to normalize for unique on target bases (i.e., capture specificity) using the results of coverage reporting software such as RefCov (T. Wylie and J. Walker, http://gmt.genome.wustl.edu/gmt-refcov/current). Discordant read-pairs may also offer a source of supporting evidence for CNAs caused by structural variation (SV). We are currently evaluating both strategies to improve sensitivity and specificity.

Exome sequencing has the potential to rapidly screen the coding regions of tumor samples for somatic alterations. The analysis methods described here will help realize that potential by enabling the simultaneous identification of germline variants, mutations, and SCNs in matched tumor–normal pairs. By design, an exome sequencing strategy prioritizes protein-altering mutations, which are not only easier to interpret in the context of pathways and biological processes but may encode “druggable” targets. It is important to realize, however, that exome sequencing surveys only ~1% of the tumor genome. This strategy will miss many noncoding mutations, as well as larger events (e.g., SV) that may contribute to tumor development and progression. Ultimately, WGS will be required to identify the full spectrum of somatic alterations in tumor genomes.

Methods

Mutation detection algorithm

Given pileup input for a tumor sample and matched normal control, the mutation detection algorithm performs several steps at each position. First, it determines if both samples meet the minimum coverage requirement (by default, three reads with base quality ≥20) and determines a genotype for each sample individually based upon the read bases observed. By default, a variant allele must be supported by at least two independent reads and at least 8% of all reads. If no variant allele meets the criteria, the position is called wild type (homozygous reference) in that sample. If multiple variant alleles are observed, the most-supported (by read count, and then by base quality) variant allele is chosen. Variants are called homozygous if supported by 75% or more of all reads at a position; otherwise they are called heterozygous. Positions where neither sample is determined to be variant are
excluded unless the --validation flag is set to 1. Next, at positions where one or both samples have a variant, the algorithm performs a direct comparison between normal and tumor as follows:

If the genotypes do not match, then their read counts are evaluated by one-tailed Fisher's exact test in a two-by-two table (see Supplemental Fig. 6), comparing the number of reference-supporting reads (outcome 1) and variant-supporting reads (outcome 2) observed in tumor (category 1) to the numbers that were observed in normal (category 2). If the resulting P-value meets the significance threshold (default 0.10), then the variant is called somatic (if the normal matches the reference) or LOH (if the normal is heterozygous). If the difference does not meet the significance threshold, the variant is called germline and processed as described below.

If the genotypes match, the variant is called germline. The variant P-value is computed by one-tailed Fisher's exact test (FET) in a two-by-two table, comparing the total number of reference-supporting reads and the total number of variant supporting reads (normal and tumor values are combined) to the expected distribution for a nonvariant position due to sequencing error (0.01%).

For example, the expected read distribution for a nonvariant position with 500× coverage in each sample would be 999 reference-supporting reads, and one variant-supporting read due to sequencing error.

Germline, LOH, and somatic mutations are further categorized as HC or LC by the VarScan processesomatic command. By default, somatic mutations are deemed HC if the variant allele frequency is at least 10% in tumor, <5% in normal, and the FET P-value is less than 0.07. Germline variants are deemed HC if they have at least 10% variant allele frequency in both normal and tumor samples. LOH variants are deemed HC if the variant allele frequency is at least 10% in the normal sample and the FET P-value is less than 0.07. Any variant not meeting the HC criteria is deemed LC. Positions that are homozygous in normal but heterozygous in tumor (gain of heterozygosity) or where the variant allele is not the same (e.g., a SNP and an indel) are presumed to be sequencing/alignment artifacts and are discarded.

CNA detection algorithm

Given pileup input for a tumor sample and matched normal, the CNA detection algorithm first determines that at least one of the samples meets the minimum coverage requirement. To reduce noise from spurious differences at low coverage, the default setting for this parameter (20) is higher than that of mutation detection. Next, the algorithm computes the depth of high-quality bases (phred base quality ≥20) individually for tumor and normal samples. These depths are recorded for each consecutive position until (1) a gap in minimum coverage is encountered, (2) the end of the chromosome is reached, or (3) the ratio of tumor depth to normal depth changes significantly, as computed by the variant's exact test. For each contiguous region, the relative copy number change (C) in the tumor is inferred by the log base 2 of the normalized depth ratio:

\[ C = \log_2 \left( \frac{D_T}{D_N} \right) + \left( I_T / I_T \right) \]

Here \( D_T \) is the average tumor depth, \( D_N \) is the average normal depth, \( I_T \) is the number of uniquely mapped bases in the normal BAM, and \( I_T \) is the number of uniquely mapped bases in the tumor BAM. The number of uniquely mapped bases is computed using SAMtools flagstat information for each BAM file, specifically as

\[ I = R_0 \times (1 - Dup) \times L, \]

where \( R_0 \) is the number of reads mapped, Dup is the proportion of mapped reads marked as duplicates, and \( L \) is the average read length. Raw copy number regions with chromosome, start position, stop position, and log₂ value underwent CBS in the DNAcopy package (Seshan and Olshen 2010) to produce segmented calls delineated by significant change-points of at least three standard deviations (Supplemental Methods). Adjacent segments of similar copy number from the CBS algorithm were merged by an internally developed Perl script (MergeSegments), and classified by size. Events encompassing >25% of a chromosome arm were classified as large-scale; all others were considered focal events.

Software implementation

The VarScan 2 core software was developed in Java; the false-positive filter was implemented in Perl. Binary executables, scripts, and source code are free for noncommercial use and available at http://varscan.sourceforge.net. The false-positive filter requires the bam-readcount utility (D. Larson et al., https://github.com/genome/bam-readcount), which is written and compiled in C.

Ovarian cancer data

The ovarian cancer data set, including exome sequence data, SNP array data, and validated somatic mutations, was generated and published by the Cancer Genome Atlas Research Network (Cancer Genome Atlas Research Network 2011). The WGS data for the five cases utilized in the cross-platform copy number comparison will be described in a separate publication. Exome and WGS sequence data are available in BAM format at the dbGaP database (http://www.ncbi.nlm.nih.gov/gap). Identifiers for samples in this study are in Supplemental Table 1.

Mutations were called in exome data for 151 tumor–normal pairs by the VarScan somatic command with the following parameters: --min-coverage 4, --min-var-freq 0.08, -p-value 0.05, --strand-filter 1--min-avg-qual 20. HC mutations were filtered to remove false positives using the criteria described in Table 1 (see Supplemental Methods). Filter-passed somatic mutations were annotated using gene structure and UCSC (Karolchik et al. 2003) annotation information, assigning each mutation to one of four tiers as previously described (Ley et al. 2008; Mardis et al. 2009). Only tier 1 mutations, which alter coding sequence (nonsynonymous, synonymous, splice site, or noncoding RNA), were reported in Figure 1 or selected for orthogonal validation. CNAs were called in exome data for 142 tumor–normal pairs (nine poor-coverage tumors were excluded) by the VarScan copynumber command with the following parameters: --min-coverage 20--min-region-size 100. Raw CNA calls underwent CBS and a subsequent merging procedure as described in Supplemental Methods.

RCNA identification, annotation, and pathway analysis

The CMDs algorithm (Zhang et al. 2010) was applied to identify regions of statistically significant RCNAs. For each tumor sample, the merged segmented copy number events (see Supplemental Methods) were cross-referenced with the coordinates of about 200,000 protein-coding exons to obtain the mean log₂ of copy number change for the start position and stop position of each exon. CMDs was configured to run with a minimum of 20 markers (exon starts or stops), corresponding to roughly one region tested per gene. Regions meeting the significance threshold (\( P < 0.0001 \)) were merged if within 100 kb of one another, yielding a set of 520 candidate RCNA regions. These were visually reviewed to identify target genes, and remove peaks encompassing six or more unrelated genes, as the target of these nonfocal events was unclear.

The cytoBand.txt and refGene.txt files from the UCSC Genome Browser Database (Karolchik et al. 2003) version hg18 were used to annotate CNA events with cytogenetic band and RefSeq gene in-
formation, respectively, using a customized Perl script. Information on specific genes from the RefSeq and KEGG databases was retrieved using GeneCards (Safran et al. 2002) version 3.0. Pathway-based analysis of 582 RCNA genes was performed using KEGG and GO database information using WebGestalt Gene Set Analysis Toolkit version 2.0 (http://bioinfo.vanderbilt.edu/webgestalt/) with the default settings (hypergeometric test, BH correction, at least two genes per category).

Acknowledgments

We thank Heather Schmidt and Joelle Veizer for extensive manual review of predicted mutations, John Wallis for insightful comments on the manuscript, and Mike Wendl for statistical discussions. We also thank the medical genomics, analysis pipeline, and technology development groups of the Genome Institute at Washington University in St. Louis, as well as the members of the Cancer Genome Atlas research consortium, for their support.

References


Received July 27, 2011; accepted in revised form January 11, 2012.