APPLICATION NOTE

Identifying Pathogenic Somatic Variants in Complex Cancer Genomes

Complete Genomics Whole Genome Cancer Sequencing Service

Complete Genomics' Cancer Sequencing Service provides:

- Comprehensive results spanning all variation types including: SNPs, indels, block substitutions, CNVs, SVs, and MEIs
- Sensitive detection of somatic variants in the presence of aneuploidy and tumor heterogeneity
- A dedicated workflow and pipeline addressing cancer study designs and challenges

Overview

Cancer is a disease of the genome, generally driven by the accumulation of mutations which come in the form of sequence variations, copy number changes, or structural variants. Research aims to better understand cancer as a disease, to aid in diagnosis, to better predict prognosis, or to enable more effective treatment after diagnosis. Essential to reaching these goals is the identification of pathogenic features of the cancer genome.

Genomic studies using cancer samples contend with several challenges that often make them more difficult to characterize at the DNA level than non-cancer samples. These features include (1) sample impurity due to the presence of non-cancerous cells in the collection, (2) tumor heterogeneity, due to the progressive accumulation of mutations as the cancer evolves or different mutations in different branches of the tumor's evolutionary path, and (3) widespread copy number changes (aneuploidy). The impact of these common cancer characteristics is that important somatic variants can have variable allele fractions within a sample. For example, a heterozygous event in a triploid region of the genome will be present at 33% or 67% allele fraction. Because of the common features of cancer genomes, and their impact, research projects focused on understanding cancer genomes must be sensitive to detecting variants at low allele fraction within a sample.

The Complete Genomics Cancer Sequencing Service offers whole human genome sequencing of cancer pairs and trios as an end-to-end service. Customers ship DNA samples and in return, they receive comprehensive summaries of all germline and somatic variations detected in each sample, along with scores, annotations, and raw supporting data (Table 1). Advanced algorithms ensure sensitive detection of mutations at low allele fractions and effective scoring and annotations provide a powerful means to filter for causative events.

Complete Genomics' Cancer Sequencing Service Reflects Common Cancer Study Designs

Cancer genome studies often focus on identifying the somatic mutations, in which the variant is present in the tumor but not the matched normal, within a cancer sample or across a set of cancer samples most likely to be causing or promoting disease. In general, these studies involve tumor samples matched to non-cancerous DNA from the same patient. The Cancer Sequencing Service accepts samples as pairs and trios, confirming before sequencing that each sample within a group is in fact derived from the same individual.

Once a set of samples have been accepted, their processing is synchronized throughout the sequencing operation. Complete Genomics offers the choice of two coverage levels: standard coverage (an average of ≥ 40x coverage across the reference genome) or high coverage (an average of ≥ 80x coverage across the reference genome). High coverage is the recommended choice for cancer samples, as it increases the chances of detecting alleles present in low fractions of the sample, as discussed below. With high coverage, double the amount of sequencing output is generated for each genome. The sequencing reads are generated in tandem, so there is no impact on turnaround time when high coverage is selected.

Once sequencing is complete, raw data is fed into the Complete Genomics Analysis Pipeline. At this phase, reads are mapped to the human genome reference and then analyses are performed to identify all variation types: SNVs and indels, copy number variations (CNVs), structural variations (SVs), and mobile element insertions (MEIs). For the Cancer Sequencing Service, each genome is compared to the human reference. Additionally, each tumor is compared to the normal match within the pair or trio, resulting in the identification, scoring, and annotation of somatic events.

In addition to comparing small variants, CNVs, and SVs between tumor and normal pairs for the identification of somatic events, Complete Genomics provides the Lesser Allele Fraction (LAF) for all tumors within cancer pairs and trios. The LAF reflects the fraction of the tumor that contains the less abundant allele at a heterozygous locus in the matched normal and is useful for identifying Loss of Heterozygosity (LOH) events. LAF ranges between 0-0.5. If the tumor sample is also heterozygous, the LAF will be approximately 0.5 (assuming the sample is pure). If the tumor sample is homozygous at this site, the LAF will be reduced down to 0 (again, assuming the sample is pure). Regions where the LAF is equal to 0 are highly suggestive of LOH, a somatic event that may correlate with the loss of function of a tumor suppressor gene. The LAF is a useful approach to identify LOH even in the absence of copy number change (copy neutral LOH).

	STANDARD SEQUENCING SERVICE	CANCER SEQUENCING SERVICE
Sample Acceptance	Samples treated as individuals	Samples treated as pairs and trios
Sample QC	• Sample quality • Gender matching • Sample matching between samples on the same plate or Sales Order • Quality Control (QC) Report provided	*Features of the Standard Sequencing Service, plus:* • Sample matching also confirms that tumor-normal pairs and trios come from a single individual
Sample Tracking	Sample processing and delivery occurs for each sample as soon as possible	Pairs and trios regrouped prior to acceptance, sequencing, analysis, and delivery
Data Results	• Germline small variants, CNVs, SVs, and MEIs • Variant annotations (dbSNP, RefSeq genes, ncRNA) • Annotation of small variant events (frame-shifting mutation, missense mutations,…) • Annotation of SV events (gene fusions, translocations,…) • Circos plots	*Features of the Standard Sequencing Service, plus:* • Somatic small variants, CNVs, and SVs • Paired-sample VCF file • Somatic Circos plot • Allele read counts for related sample • Evidence directory for related samples
Data Delivery	One data set = one sample	Group packaged as a single data set

Table 1. *Complete Genomics offers two sequencing services, the Standard Sequencing Service for individual samples and the Cancer Sequencing Service for tumor-normal pairs and trios. All features of the Standard Sequencing Service are provided in the Cancer Sequencing Service, with additional features to address the unique aspects of a tumor-normal comparison.*

Sensitive Detection of all Variants

The Analysis Pipeline Small Variant Caller was recently optimized to provide greater sensitivity to SNPs and indels present at variable allele fractions. The model employed in Analysis Pipeline Version 2.0 allows an extra term that represents allele fraction, and does not constrain heterozygous variants to the diploid assumption (i.e., equal allele fraction). The score for a heterozygous call is then based on the log likelihood ratio for the optimal allele fraction, compared to the best homozygous hypothesis. Figure 1 provides an example where 20 reads support the reference genotype and five reads support the non-reference genotype. Support for a heterozygous call is low if the sample is assumed to contain heterozygous alleles at equal fractions, but it is higher when considering variable allele fractions.

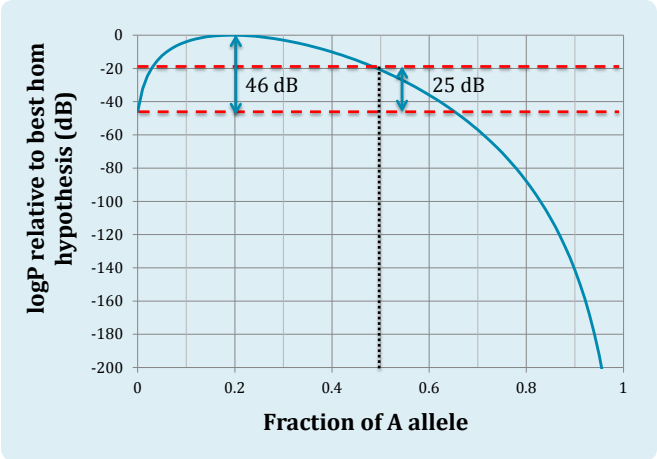


Figure 1. *Variation calling allowing for allelic imbalance. In the example shown, 20 reads support the reference genotype and five reads support an alternate genotype. Support for a heterozygous SNP using the diploid caller is only 25 dB, or very low confidence. Support for a heterozygous SNP using the variable fraction caller increases to 46 dB, indicating higher confidence, given the underlying assumption that the alternate genotype may be present in only ~ 20% of the sample.*

The gain in sensitivity to heterozygous alleles is a function of allele fraction (Figure 2). There is minimal difference in sensitivity to SNPs with a 50% allele fraction (i.e., a heterozygous SNP in a pure, diploid sample). As the allele fraction reduces, the gain in sensitivity for SNPs relative to the diploid caller increases significantly by leveraging the updated algorithms, available in Analysis Pipeline 2.0. Discovery of variants present at lower allele fractions is further enhanced by high coverage. The reasoning for this is intuitive; with more reads available to genotype a given locus, the more likely it is that a

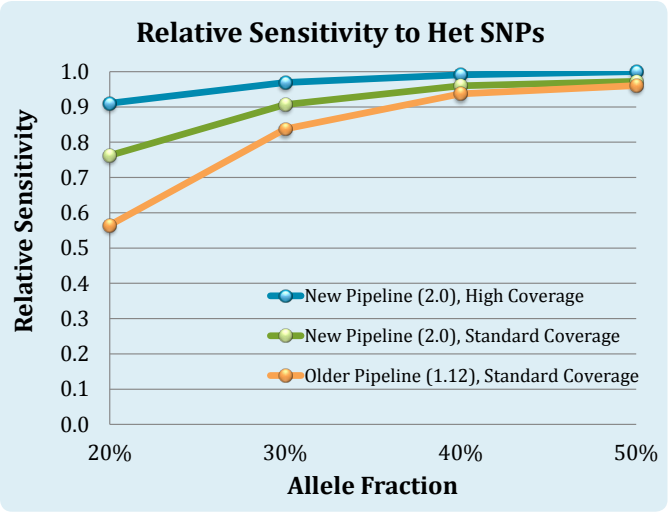


Figure 2. *Sensitivity to heterozygous SNPs at low allele fractions improves with updated algorithms and deeper sequencing. An* in silico *titration of two HapMap samples, NA19240 and NA12878, was performed using the previous Analysis Pipeline (version 1.12) and the new Analysis Pipeline (version 2.0) at both standard coverage (a genome-wide average of 50x coverage) and high coverage (a genome-wide average of 100x coverage). Sensitivity to SNVs is reported relative to the number of SNVs detected in NA19240 using the new pipeline and at high coverage. A score threshold was applied for all samples, filtering for variants assigned "VQHIGH" confidence, which translates to ~ 0.5 FP/Mb for 50% allele fraction and ~ 1.5 FP/Mb for lower allele fractions for this sample set.*

hidden variant can be detected. With a combination of variable allele fraction modeling in the small variant caller and the ability to perform deeper sequencing, Complete Genomics' Cancer Sequencing Service maximizes the ability to detect somatic mutations that can help characterize the disease.

A comparison of somatic events detected in well-characterized tumor-normal cell line pairs to those that have been previously reported supports the observation that both the algorithms and higher coverage combine to provide highly sensitive detection. Sensitivity of somatic events depends on accurate detection of genotypes in both samples within the pair. The COSMIC* database includes 84 and 183 somatic small variants (SNVs and indels) confirmed in the exome for the HCC1187 and HCC2218** cell-line pairs, respectively. Filtering for both high confidence-only somatic events and exome-only calls, Complete Genomics identified greater than 90% of these at high coverage, with even more identified when

* Catalogue of Mutations in Cancer
** DNA purchased from ATCC.

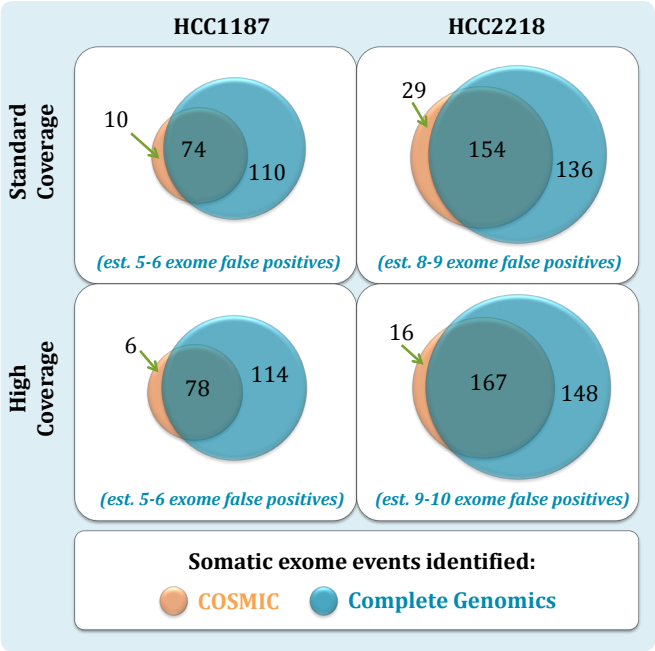


Figure 3. *Overlap between somatic events detected (exome only) between COSMIC (orange) and Complete Genomics (blue). Standard coverage is 60x to 63x; high coverage is 123x (except the HCC2218 Normal, which is 92x). Complete Genomics events were filtered to a somatic score of ≥ -10, equivalent to SQHIGH. Exome false positives based on estimated 3% FDR at SQHIGH, estimated based on technical replicate analysis, see Table 3.*

including events identified at low confidence (Figure 3).

Using the same filters for high confidence, Complete Genomics identified approximately twice as many novel somatic events in the exome than COSMIC, greatly expanding the number of events detected in these well-characterized samples. These additional variants had similar score characteristics to those confirmed in COSMIC and are considered likely to be true. In summary, both a non-diploid calling algorithm and deeper sequencing combine to provide great sensitivity to variants in tumor samples, which likely exist in variable fraction amounts of the sample.

We Report, You Discover

One of the greatest challenges of whole human genome sequencing is the identification of changes in the genome that are both real and relevant to the disease.

Narrowing down to what is 'truth'

The haploid human genome consists of approximately three billion bases, meaning that even with a low error rate of 1 x 10^{-5} base pairs, there would be 30,000 errors expected genome-wide. While cancer genomes often accumulate a large number of mutations, effective mutation detection would be unrealistic amidst the reporting of so many errors without tools to determine which result is likely true and which result is likely false. Further, some cancer samples contain only a small number of somatic events, resulting in a larger impact from minimal errors.

Complete Genomics provides scores that allow the researcher to appropriately balance sensitivity and specificity for a given study. Scores are provided for variants called in comparison to the human genome

>locus	ploidy	chromosome	begin	end	zygosity	varType	reference	allele1Seq	allele2Seq	allele1VarScoreVAF	allele2VarScoreVAF	allele1VarScoreEAF	allele2VarScoreEAF	allele1VarQuality	allele2VarQuality	allele1ReadCount	allele2ReadCount	allele1ReadCount-N1	allele2ReadCount-N1	somaticCategory	somaticRank	somaticScore	somaticQuality
16551	2	chr1	1038975	1038976	hom	snp	C	T	T	79	718	79	718	VQHIGH	VQHIGH	35	35	1	1	snp	0.831	14	SQHIGH
16552	2	chr1	1038976	1040885	hom	ref	=	=	=														
16553	2	chr1	1040885	1040885	het-ref	ins		A		26	26	3	3	VQLOW	VQLOW	3	19	0	17	ins	0.012	-28	

Table 2. *Example of a Tumor Master Variations file. The variant is described by the varType (e.g. snp), zygosity (homozygous or heterozygous), chromosome, position, and sequence columns. Scores are provided whenever a variant is detected, using both the variable allele fraction model (varScoreVAF) and the equal allele fraction model (varScoreEAF), for each allele. The quality of the variant call is also assigned to one of two possible categories, VQLOW or VQHIGH. The allele read count for each allele is also provided for both samples within the comparison. If the variant is a potential somatic event, the somatic columns will also be filled in. These include somaticRank, which ranks all identified somatic events within a somaticCategory and somatic score. Any variant determined to have a somatic score > -10 is assigned SQHIGH in the somaticQuality category.*

reference using either the diploid model or the non-diploid model (varScoreEAF and varScoreVAF, respectively; Table 2). Prior knowledge of the sample ploidy or purity can be used to inform which score to use in the prioritization of events identified.

Identification of somatic events is susceptible to a larger error rate because it relies on accurate genotyping of both samples involved in the comparison. Either a false positive in the tumor or a false negative in the normal could result in a spurious somatic call for the tumor genome. The prioritization of called somatic events is made easier with somatic-specific annotation and scoring provided by the Cancer Sequencing Service. Each putative somatic event detected is assigned to a somaticCategory (SNP, insertion, deletion, or block substitution), ranked within the category in terms of likelihood of being a true event, and assigned a somaticScore reflecting the probability that the event is true and not false (Table 2). Additionally, any somatic event that earns a somaticScore of reasonable confidence (≥ -10) is assigned to the SQHIGH category for somaticQuality, making it easy to apply one simple filter to significantly reduce the error rate with a disproportionately lower reduction in sensitivity.

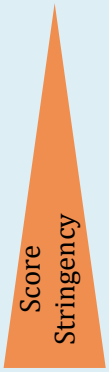
All somatic variants identified include a comprehensive list of putative somatic events, including those with low confidence indicating a high error rate. Applying incrementally more stringent scores effectively reduces the number of errors, without initially causing a significant impact on the overall number of somatic events detected (Table 3). By selecting a somatic score that maintains good sensitivity while reducing the number of spurious calls, it is possible to narrow down the search of variants to those that are most likely to be true.

Narrowing down to what is 'relevant'

The relevance of any detected event will depend on the study and its goals. Common criteria for relevance in cancer genome studies include: what is somatic, what is pathogenic, and what is a driver mutation.

What is somatic? As discussed above, identifying the variants that are likely to be somatic with high confidence is straightforward based on the variant annotations, including somaticScore and somaticQuality. These are based on probabilistic comparisons of the variant or reference calls in the tumor and matched normal samples. Beyond small variants, the Cancer Sequencing Service also highlights and scores somatic CNVs and somatic SVs, and provides the LAF estimates that are useful in identifying regions of loss of heterozygosity, including those without copy number change.

For an event to be somatic, it must be present in the tumor sample and absent in the matched normal. Somatic small variants identified may be a true positive, or false positives as a result of either a false positive in the tumor itself or a false negative in the normal. To provide confidence that an event is in



SOMATIC SCORE THRESHOLD	GENOME-WIDE			EXOME-WIDE	
	CANDIDATE SOMATIC SNVS	FALSE POSITIVES	ESTIMATED FDR	CANDIDATE SOMATIC SNVS	FALSE POSITIVES
-10	21,932	666	3.04%	239	5
-5	18,158	171	0.94%	219	1
0	15,131	72	0.48%	186	1
5	11,918	29	0.24%	145	1
10	6,763	7	0.10%	76	0
15	1,690	1	0.06%	35	0

Table 3. *Somatic events were identified in which the event was present in the tumor and absent in the normal for HCC1143* paired cell lines, sequenced at high coverage. Somatic event counts at each somatic threshold applied are listed, genome-wide and exome-wide. Lower somatic score thresholds correlate with lower score stringency and more events detected. Higher somatic score thresholds correlate with higher score stringency, fewer events detected, and a lower false detection rate.*

** DNA purchased from ATCC.*

fact truly somatic, Complete Genomics computes the somatic score that takes into consideration both the likelihood of the variant called in the tumor and the likelihood of reference call in the matched normal. Additionally, Complete Genomics performs a realignment to the variant sequence in the matched sample whenever a small variant is identified in one sample and not in the other. The realigned allele reads counts are calculated and the reads are provided in a sub-folder for further interrogation and visualization. Therefore, for any somatic event of interest, it is possible to look in the matched normal sample to confirm that the evidence based on realignment does in fact suggest that the variant is absent in the normal.

What is pathogenic? Categorization of a variant as 'pathogenic' requires making inferences based on common characteristics of a pathogenic variant. These characteristics may include that it is not common in the population, that it has a predictable impact at the protein level including disregulation, and that it exists in a class of previously-defined cancer genes.

Filtering for pathogenic somatic events generally relies on filtering against annotations that support these characteristics. Annotations provided by Complete Genomics can be used for this purpose, as follows:

- *Is the event common?* Variation identifiers from dbSNP, minor allele frequency reported by 1000 Genomes Project.
- *Does the event have a predictable impact?* Entrez gene identifier, RefSeq mRNA accession number (versioned), RefSeq protein accession number (versioned), NCBI Gene Symbol, region of the gene impacted (eg, intron, splice site donor), the functional impact (eg, no-change, synonymous, missense), Pfam domain information, list of ncRNA annotations.
- *Is the event in a known cancer locus?* Variation identifers from COSMIC.

What is a driver mutation? Cancer cells tend to accumulate a large number of mutations, but only a subset of them may be related to the initiation or progression of the disease. These are considered driver mutations, versus passenger mutations that arise due to the instability of the genome.

The best way to confirm that a mutation is driving the disease is through functional studies, but this is typically impossible to do for the full range of mutations identified in a genome. A distinguishing characteristic between driver mutations and passenger mutations is that a driver mutation is likely to be recurring in a specific cancer or across cancer types, whereas passenger mutations arise sporadically and randomly and therefore are less likely to be recurring. The mutation itself may not always be the same; it may be the gene affected or the pathway or network that is impacted. Therefore, the most common approach to identifying driver mutations is by searching for loci, genes, or pathways that are consistently impacted across a set of samples. Complete Genomics offers tools to compare sets of samples for this purpose. The masterVarBeta file links locus, gene, and somatic scoring information all in a single file for easy filtering. The somaticVcfBeta file combines small variant, CNV, and SV data all into a single source in a format commonly used for further evaluating sequencing data. Complementing the Analysis Pipeline is a set of open source CGA™ Tools. These include two tools - listvariants and testvariants - which, together, list all variants detected in any sample within a set of samples and then report whether or not that variant is present in each sample of the set. The output of these simple tools provides a straightforward mechanism for identifying those variants that are recurring across many cancer samples or for categorizing variants between primary tumors, relapse or metastatic tumors, and their normal match.

Conclusion

Complete Genomics offers whole human genome sequencing results that are cost-effective, informative, and comprehensive. Complete Genomics now provides a customized service and enhanced results for cancer samples. The Cancer Sequencing Service includes synchronized sequencing of cancer pairs and trios that have been confirmed to originate from the same patient plus identification, scoring, and annotation of somatic events present in the tumor and absent in the normal. Variation calling accommodates low allele fractions using a variable allele fraction caller to ensure sensitive detection of important somatic events in challenging genomes. Complete Genomics delivers comprehensive reports pointing researchers toward the true and relevant mutations present in their cancer samples.

www.completegenomics.com info@completegenomics.com
2071 Stierlin Court, Mountain View, CA 94043 USA Tel 650.943.2800

Published in U.S.A., March 2012, AN_CS-01