# Accurate IGN Cancer Sequencing with a Combined-Calling Method

Christopher Saunders, Lisa Murray, Wendy Wong, Sajani Swamy, Jennifer Becq, Keira Cheetham, Brad Sickler, Marc Laurent, Dipesh Risal, and Han-Yu Chuang

Illumina Inc., 5200 Illumina Way, San Diego, CA and Chesterford Research Park, Cambridge, UK

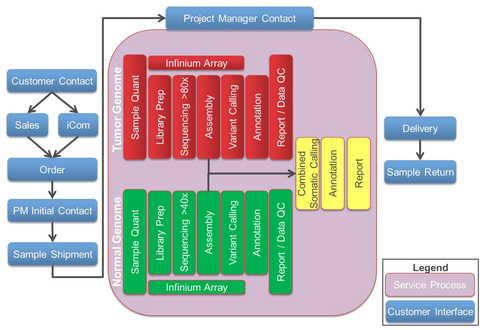
## Cancer Analysis Service from IGN

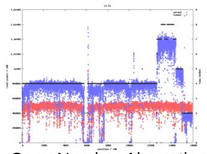
The Illumina Genome Network (IGN) employs the latest Illumina sequencing technology, the **HiSeq® 2000 system with TruSeq®** SBS chemistry, ensuring the most effective and accurate human genomes for large-scale whole-genome projects.

The IGN Cancer Analysis Service delivers **>40x coverage** whole genome sequencing (WGS) of the normal sample and **>80x coverage** WGS of the tumor sample(s) (other coverages available), and leverages the same high-quality service process available from the standard IGN service offering. IGN's Bayesian combined calling method is used to detect somatic variants. The sample input requirement is 3 μg for each tumor and normal samples.

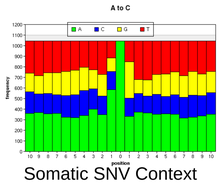
The deliverables include

- Somatic SNVs, indels, structural variants, and copy number aberrations
- Circos plot that displays the various somatic calls graphically
- PDF summary reports for somatic SNVs, indels, structural variants and copy number variation
- Individual sequence, variant, and genotyping data for both normal and tumor genomes
  - Aligned and non-aligned reads in archival BAM format
  - SNPs, indels, CNV, and structural variants in VCF format
  - Omni2.5M genotyping array raw intensity and calls

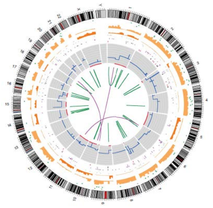


Copy Number Aberrations

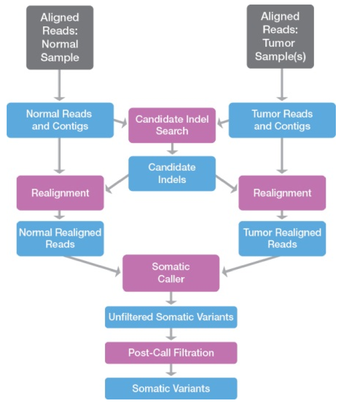



Somatic SNV Context



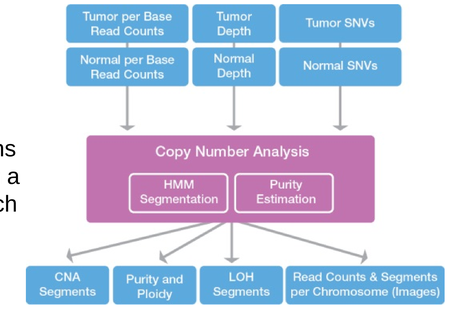
Circos plot for visualizing genome-wide somatic variations

## Somatic Variant Analysis Pipeline




**Combined calling for somatic SNVs and indels**
IGN's robust somatic caller combines both the tumor and matched normal data to detect somatic SNVs and indels (method submitted for publication). The method models the normal sample as a mixture of diploid germ-line variation and noise, while the tumor sample is modeled as a combination of the normal sample and somatic variation. The combined analysis of the two genomes assumes that the somatic variation and the normal noise can occur at any allele frequency ratio. The method therefore is optimal for real-life tumor samples which can possess a varying amount of normal contamination.

**Hidden Markov Model for somatic copy number aberrations (CNAs)**
CNASeg, based on a published method[1] for detecting germ-line copy number changes, estimates copy number aberrations from matched normal and tumor data. CNAs are derived from a Hidden Markov Model fitted to read depth estimation from each genome for calculating purity and ploidy of the tumor sample.

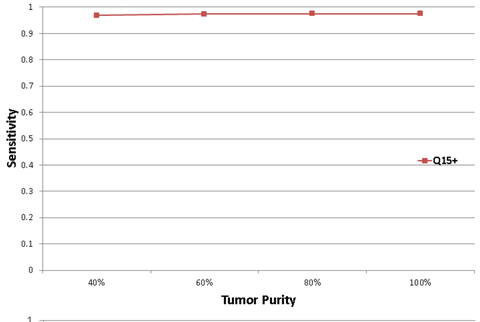



**Somatic structural variations (SVs)**
Somatic SVs, such as large indels (>300bp) and inversions, are obtained by comparing the SVs identified in the tumor and the SVs in the matched normal by a guided reassembly of unaligned reads. After subtraction, dedicated filtering is performed to reduce the false positive rate.

## TruSeq Technology and Combined Calling Deliver Robust Performance for Impure Tumor Samples

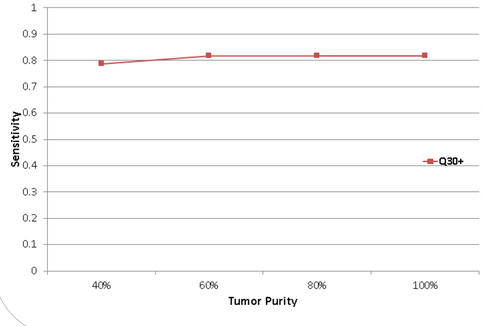
**Sequencing data**
The tumor (sequenced to 80x) and matched normal (40x) samples used in this study are the metastatic melanoma cell line COLO-829 and COLO-829BL, a lymphoblastoid line derived from the same patient. Tumor purity was simulated by mixing tumor and normal reads (from a separate batch).

### Sensitivity Evaluation




Sensitivity rate = fraction of the 454 CE-confirmed colon cancer cell line SNVs[2] recovered in IGN data

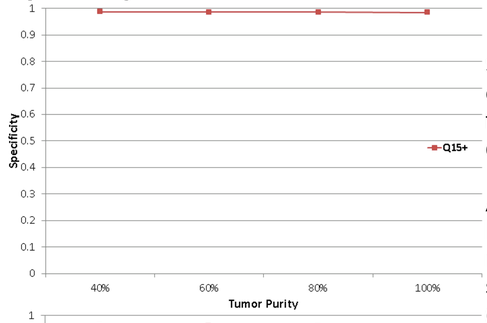
At 40% tumor purity, IGN can recover **97%** of the 454 SNV's with a Q15+ (default) quality cutoff




Sensitivity rate = fraction of the 66 CE-confirmed metastatic melanoma cell line indels[2] recovered in IGN data

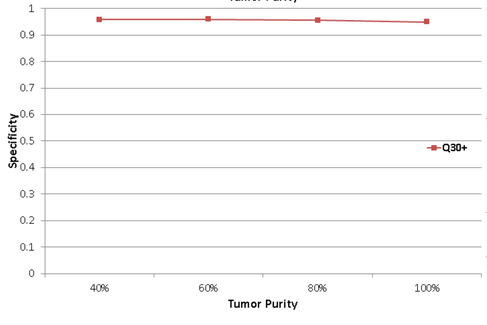
At 40% tumor purity, IGN can recover **79%** of the 66 SNV's with a Q30+ (default) quality cutoff

### Specificity Evaluation




Specificity rate = fraction of the called SNVs in IGN data not found as a common SNPs in dbSNP 132[3]

At all tested tumor purity, IGN has only a **1.5%** chance of making a false positive call for somatic SNVs with a Q15+ (default) quality cutoff

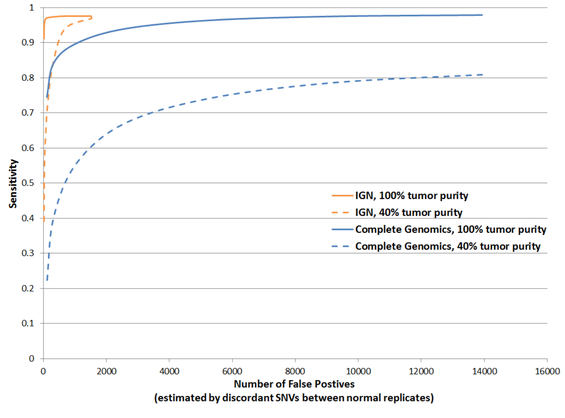



Specificity rate = fraction of the called indels in IGN data not found as a common indel in 1000 genomes data[4]

At all tested tumor purity, IGN has only **5%** chance to make a false positive call for somatic indels with a Q30+ (default) quality cutoff

### Competitive Analysis

A pseudo-ROC plot was generated for IGN and Complete Genomics datasets to compare the false positive calls and sensitivity rates of somatic calls.
X-axis: Two separate datasets for the same sample were used to generate "somatic" calls (false positives) for IGN (NA18506 at 40x vs. 80x; 9 Qscore cutoffs) and Complete Genomics[5] (NA12878 at 55x vs. 55x; 7 Somatic Score cutoffs).
Y-axis: Sensitivity rates for SNVs for IGN (Colo-829 and matched normal at 40x vs. 80x) and Complete Genomics[5] (NA19240 vs. NA12878 at 55x vs. 55x.)




For the most tolerant scoring cutoffs, IGN data generates far fewer false positive calls (@1500 total vs. @14K for Complete Genomics.)

At a fixed false positive level, IGN calls are more sensitive than those of Complete Genomics.

1. Ivakhno S *et al.* Bioinformatics. 2010 Dec 15;26(24):3051-8.
2. Pleasance ED *et al.* Nature. 2010 Jan 14;463(7278):191-6.
3. dbSNP build 132 has a "Common SNPs" track for uniquely mapped variants that appear in at least 1% of the population; see http://www.ncbi.nlm.nih.gov/projects/SNP/.
4. Common indels across populations can be found on 1000 genomes website at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/low_coverage/indels/.
5. Sensitivity and discordant SNV counts are obtained from Table 20 in Complete Genomics' user guide downloaded from http://cgatools.sourceforge.net/docs/1.5.0/cgatools-user-guide.pdf