MAKING AN I.M.P.A.C.T
ADVANCING THE COMPUTATION OF NEXT-GENERATION SEQUENCING DATA

Krishan Kania

Part I: My Story

After competing in the Intel Science Talent Search, I am sometimes asked about what I have gained or learned, aside from specific academic knowledge. Well, in the process, I believe I have gained clarity pertaining to science in general.

Clarity: I've always thought of biology as the qualitative science-- that biology is applied chemistry, chemistry is applied physics, and physics is applied mathematics. And yes, from my research experience this past summer, I've experienced first hand that the tertiary and quaternary principles that explain many of our observations in biology do rest in these quantitative disciplines. However, I've also realized that these "degrees of separation" between biology and math are meaningless.

In order to accomplish my project goals, I needed to combine number theory, statistics, genetics, and computer science. My project was never to just observe tissue samples, nor was it to crunch numbers and perform calculations. It was to do both in a way that allows me to approach a question in a more comprehensive way than I have ever been exposed to in any of my previous academic experiences. My research experience has clarified the value (and necessity) of uniting all scientific and quantitative disciplines when answering a question, responding to a worthy cause.

Part II: The Science

INTRODUCTION:

Next-generation sequencing (NGS) has allowed substantial advances in cancer genomics. In fact, large-scale discovery efforts have propelled the identification of hundreds of cancer-related genes in recent years. To be truly transforming, however, key cancer-associated mutations must be profiled systematically in the clinical and translational arena to guide rational cancer therapeutics. This aim has yet to be achieved on a large scale, mainly because many methodologies cannot be applied efficiently and reliably on formalin-fixed paraffin embedded (FFPE) tumor samples that are routinely encountered in the clinic and in archived tumor banks. This project is a part of the computational effort to develop and apply a robust and cost-effective methodology, empowered by solution-phase exon capture and massively parallel next-generation sequencing, by which any FFPE tumor may be characterized for somatic base mutations and copy number changes in all known cancer genes. With the programming language "R," the computational analysis of NGS data for assays running clinical samples has been redeveloped, automated, and graphically represented. Moreover, such analysis, such as copy-number graphs or QC metrics, can be computed at a speed that is 568 times as fast as the traditional, and manual, computational techniques of alternative methodologies. Furthermore, the program is built with careful considerations to make an even more comprehensive analysis than before, collectively addressing the two most important concerns for translating NGS to patient care: (1) speed and (2) confidence.

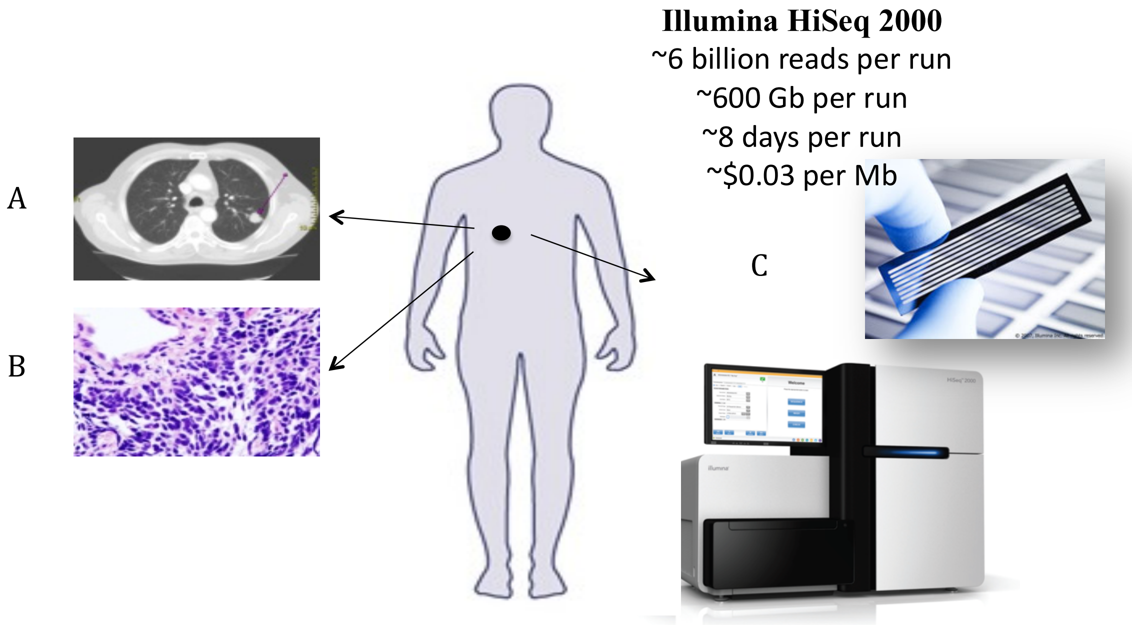
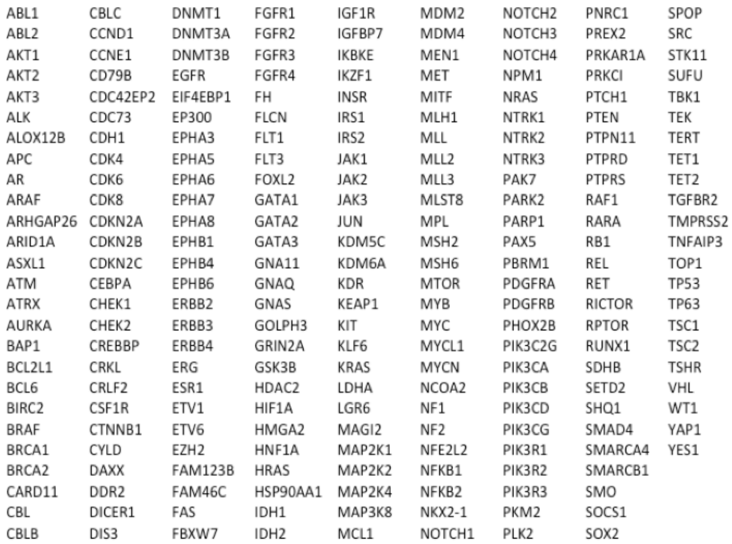


Figure 1: The paradigm of IMPACT Personalized Medicine[2]
(A) Radiology, (B) Pathology, (C) IMPACT

Integrated Mutation Profiling of Actionable Cancer Targets (IMPACT)

IMPACT refers to the assay performed in this study. While NGS assays are generally similar, there are important differences between IMPACT and competing assays of other labs. IMPACT takes advantage of capture-based sequencing that targets a subset of the genome using "baits" that select for specific DNA sequences. Scientists can lightly sequence the entire genome or exome; this is useful in discovery projects. IMPACT, however, uses baits (Roche NimbleGen) to only "pull down" the 275 proto-oncogenes, tumor-suppressor genes, and any other genes involved in cell growth or division (Figure 2). This allows for deep coverage of the genes that other labs have defined as most relevant to cancer20. By sequencing only targeted regions of the genome, this technique not only allows for deep coverage of key genes, but also, can detect low frequency mutations that occur in heterogeneous tumors or impure samples21-24. Furthermore, capture-based sequencing data can be used to identify structural rearrangements when at least one of the breakpoints is located in a targeted region25. On average, IMPACT provides 700-1200 reads of coverage.



ABL1
ABL2
AKT1
AKT2
AKT3
ALK
ALOX12B
APC
AR
ARAF
ARHGAP26
ARID1A
ASXL1
ATM
ATRX
AURKA
BAP1
BCL2L1
BCL6
BIRC2
BRAF
BRCA1
BRCA2
CARD11
CBL
CBLB

CBLC
CCND1
CCNE1
CD79B
CDC42EP2
CDC73
CDH1
CDK4
CDK6
CDK8
CDKN2A
CDKN2B
CDKN2C
CEBPA
CHEK1
CHEK2
CREBBP
CRKL
CRLF2
CSF1R
CTNNB1
CYLD
DAXX
DDR2
DICER1
DIS3

DNMT1
DNMT3A
DNMT3B
EGFR
EIF4EBP1
EP300
EPHA3
EPHA5
EPHA6
EPHA7
EPHA8
EPHB1
EPHB4
EPHB6
ERBB2
ERBB3
ERBB4
ERG
ESR1
ETV1
ETV6
EZH2
FAM123B
FAM46C
FAS
FBXW7

FGFR1
FGFR2
FGFR3
FGFR4
FH
FLCN
FLT1
FLT3
FOXL2
GATA1
GATA2
GATA3
GNA11
GNAQ
GNAS
GOLPH3
GRIN2A
GSK3B
HDAC2
HIF1A
HMGA2
HNF1A
HRAS
HSP90AA1
IDH1
IDH2

IGF1R
IGFBP7
IKBKE
IKZF1
INSR
IRS1
IRS2
JAK1
JAK2
JAK3
JUN
KDM5C
KDM6A
KDR
KEAP1
KIT
KLF6
KRAS
LDHA
LGR6
MAGI2
MAP2K1
MAP2K2
MAP2K4
MAP3K8
MCL1

MDM2
MDM4
MEN1
MET
MITF
MLH1
MLL
MLL2
MLL3
MLST8
MPL
MSH2
MSH6
MTOR
MYB
MYC
MYCL1
MYCN
NCOA2
NF1
NF2
NFE2L2
NFKB1
NFKB2
NKX2-1
NOTCH1

NOTCH2
NOTCH3
NOTCH4
NPM1
NRAS
NTRK1
NTRK2
NTRK3
PAK7
PARK2
PARP1
PAX5
PBRM1
PDGFRA
PDGFRB
PHOX2B
PIK3C2G
PIK3CA
PIK3CB
PIK3CD
PIK3CG
PIK3R1
PIK3R2
PIK3R3
PKM2
PLK2

PNRC1
PREX2
PRKAR1A
PRKCI
PTCH1
PTEN
PTPN11
PTPRD
PTPRS
RAF1
RARA
RB1
REL
RET
RICTOR
RPTOR
RUNX1
SDHB
SETD2
SHQ1
SMAD4
SMARCA4
SMARCB1
SMO
SOCS1
SOX2

SPOP
SRC
STK11
SUFU
TBK1
TEK
TERT
TET1
TET2
TGFBR2
TMPRSS2
TNFAIP3
TOP1
TP53
TP63
TSC1
TSC2
TSHR
VHL
WT1
YAP1
YES1

Figure 2: IMPACT panel of captured genes.

MATERIALS & METHODS:

2.1 Targeted Sequencing Methodology in IMPACT

Once Formalin-fixed paraffin embedded (FFPE) tumor tissue is obtained, the genomic DNA is extracted and sheared to a mean fragment length of 200-300 base pairs. Adaptors containing sequencing primer sites and a unique barcode are ligated to the ends of DNA fragments to create a sequencing library (approximately 24 barcoded libraries are combined in an equimolar pool). These libraries are hybridized in solution to biotinylated capture oligonucleotides (baits) complementary to the exons of 275 cancer genes. Captured DNA is enriched via streptavidin-coated magnetic beads and eluted. The DNA is then sequenced on one lane of an Illumina HiSeq 2000. After QC metrics and other metric analysis of Illumina's fastQ file, the sequence reads are aligned to the reference human genome, and target genes are examined for mutations, InDels, copy number alterations, and rearrangements (figure 3).

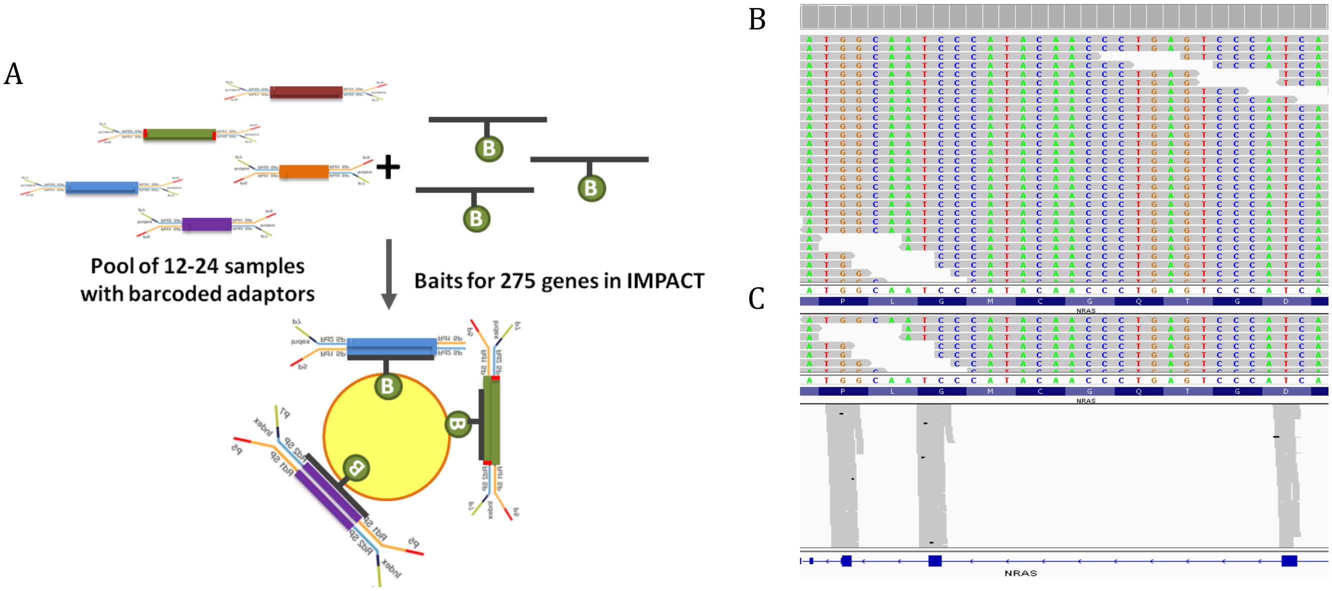


Figure 3: Overview of IMPACT
(A) IMPACT uses a hybrid capture method to sequence multiplexed libraries of 12-24 samples. (B) Reads are aligned to the reference genome and can be visualized by the Integrated Genomics Viewer. (C) This method allows

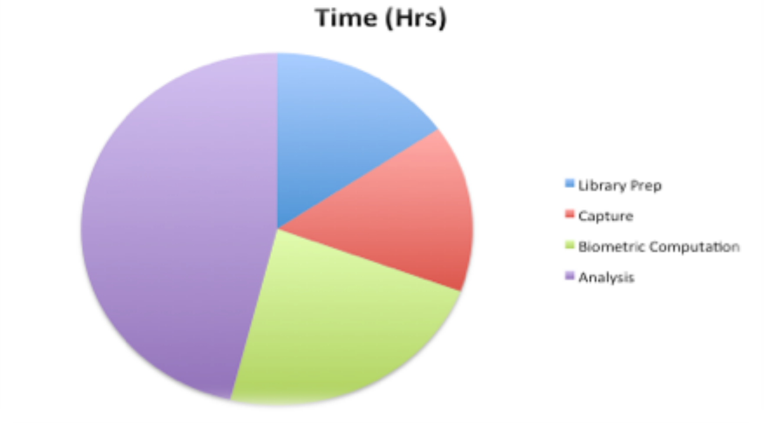
2.2 Metrics & Quality Control (QC):

To become familiar with the computational side of IMPACT, with the instruction of my mentor, I performed the following calculations from raw data for ovarian, colorectal, and melanoma projects of collaborating labs. These experiences would later expand into this project, where I will go on to develop these metrics into an R-scripted program that relates each calculation with graphical support:

Metrics:

Cluster Density & Alignment Rate
Base Quality Scores
Insert Size Distributions
Fingerprints
Contamination
Capture Specificity
Library Complexity
Mean Target Coverage
GC Loess Normalization

Figure 5:



Context:

All metrics are derived from fundamental concepts in biology, chemistry, physics, and statistics. For example, “Fingerprints” is a metric that checks the alleles of a tumor/normal pair at 42 sites of single-nucleotide polymorphism (SNP). SNP is a single nucleotide base-pair site where variation is found in at least 1% of the population[26]. Specifically, the SNP’s we capture are in tiling regions, regions of the genome very close to the 275 cancer genes. When 38 out of 42 SNP sites (arbitrary threshold) match between tumor and normal tissue, there is confidence that the tumor is paired with its matched normal. “Fingerprints” allows an investigator to identify contamination, sample swamps, and even loss of heterozygosity. GC content is the percentage of nitrogenous bases on a DNA molecule in a particular region that are either guanine or cytosine.

$$GC\ Content = \frac{G+C}{A+T+G+C} \times 100$$

A GC pair is bound by three hydrogen bonds, compared to AT pairs, which are bound by only two hydrogen bonds. A high GC-content implies a higher annealing temperature and higher melting temperature in PCR experiments. Moreover, high GC-content often implies major technology-related artifacts and biases due to the weakness of sequencing technology. Since levels GC bias is varied across the genome, the GC effect can be hard to tell apart from the true signal[27]. Even more challenging, the effect is not consistent between repeated experiments, or even libraries within the same experiment. Unsurprisingly, estimating and directly correcting for this effect has become a well-established step in protocol design. Normalization is therefore essential to ensure accuracy, particularly in GC rich regions; the statistical method of choice is the Loess model. GC Loess Normalization will be incorporated into the R-Scripted program.

Traditionally, the two biggest limitations of metric computation in IMPACT and similar assays are speed and confidence. In a typical project, these metrics require hours of valuable technician time to do rudimentary calculations or manipulations. As a corollary, this process is not as comprehensive as it can be, resulting in more scenarios where mutation calls are questionable and thus subject to primer evaluation. Speed and confidence, arguably the two most important concerns of a clinical setting, will be better addressed with the R-scripted program.

2.3 Mutation Calling & Primer Evaluation

Following bench work, sequencing, and metric computation, scientists are now ready to confidently discuss genomic mutations:

Copy-Number Alterations:

IMPACT can determine the pattern of copy-number alterations: the gain/loss of chromosome arms or focal amplifications and deletions that might range from tens of kilobases to tens of megabases in size. In a normal genome, across the autosomes, there should be two copies of every gene, one maternal and one paternal. However, in the cancerous genome there is usually a gain or loss of one or several exons. Of course, these events have important therapeutic implications. Copy-number was traditionally visualized in Microsoft Excel in manually generated graphs depicting the tumor/normal ratio of exon reads between a given tumor and its matched normal[28]. The limitations of this strategy will be revisited in results.

Somatic nucleotide substitutions and small insertion and deletion mutations:

Nucleotide substitution mutations are the most frequent somatic genomic alteration in cancer, occurring at the rate of about one somatic nucleotide substitution per million nucleotides; insertion and deletion events are approximately tenfold less common[29]. The detection of somatic mutations in cancer requires mutation calling in both the tumor DNA and the matched normal DNA, coupled with comparison to a reference genome and an assessment of the statistical significance of the number of counts of the mutation in the cancer sequence and its absence in the matched normal sequence[30].

Primer Evaluation:

The validation of questionable mutation calls require PCR amplification of the region surrounding the mutation in question followed by sanger sequencing[30].

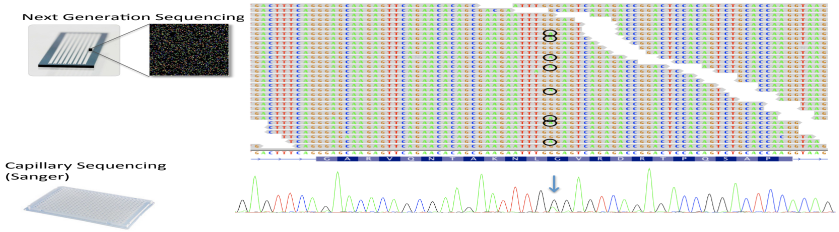
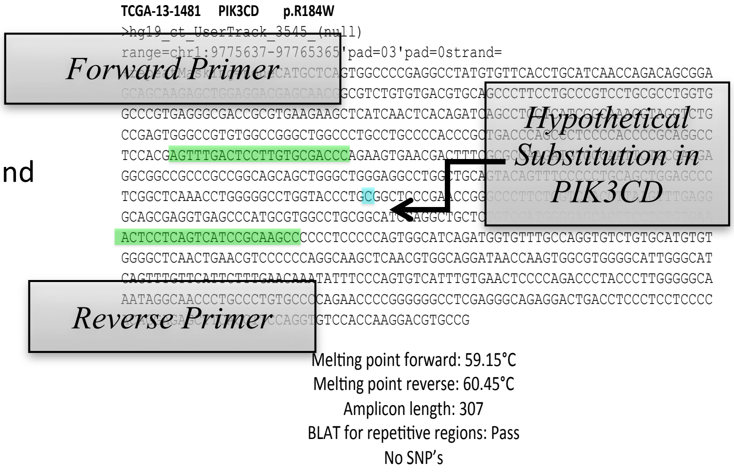


Figure 6: Suspected substitution event in PIK3CD shown in Integrated Genomic Viewer (IGV) is confirmed with primer evaluation.

Primer Development:

A length of 18-25 bases
%GC content between 40 and 60%
Tm (melting temperature) between 50 and 60°C

No secondary priming sites (BLAT)
No dimerization capability
No significant hairpins (> 3 bp)



TCGA-13-1481 PIK3CD p.R184W

```
>hg19_ct_UserTrack_3545_(null)
range=chr1:9775637-97765365'pad=03'pad=0strand=
repeatMasking=noneCATGCTCAGTGGCCCCGAGGCCTATGTGTTCACCTGCATCAACCAGACAGCGGA
GCAGCAAGAGCTGGAGGACGAGCAACGGCGTCTGTGTGACGTGCAGCCCTTCCTGCCCGTCCTGCGCCTGGTG
GCCCGTGAGGGCGACCGCGTGAAGAAGCTCATCAACTCACAGATCAGCCTGCTCATCGGCAAAGCTACGTCTG
CCGAGTGGGCCGTGTGGCCGGGCTGGCCCTGCCTGCCCCACCCGCTGACCCAGCCCTCCCCACCCCGCAGGCC
TCCACGAGTTTGACTCCTTGTGCGACCCAGAAGTGAACGACTTTCGCGCCAAGATGTGCCAATTCTGCGAGGA
GGCGGCCGCCCGCCGGCAGCAGCTGGGCTGGGAGGCCTGGCTGCAGTACAGTTTCCCCCTGCAGCTGGAGCCC
TCGGCTCAAACCTGGGGGCCTGGTACCCTGCGGCTCCCGAACCGGGCCCTTCTCGTCAACCTTATGTTTGAGG
GCAGCGAGGTGAGCCCATGCGTGGCCTGCGGCATCCAGGCTGCTCTGTCCATGGGGAGCACTTCCTCTGTGAA
ACTCCTCAGTCATCCGCAAGCCCCCCTCCCCCAGTGGCATCAGATGGTGTTTGCCAGGTGTCTGTGCATGTGT
GGGGCTCAACTGAACGTCCCCCCAGGCAAGCTCAACGTGGCAGGATAACCAAGTGGCGTGGGGCATTGGGCAT
CAGTTTGTTCATTCTTTGAACAAATATTTCCCAGTGTCATTTGTGAACTCCCCAGACCCTACCCTTGGGGGCA
AATAGGCAACCCTGCCCTGTGCCCCAGAACCCCGGGGGGCCTCGAGGGCAGAGGACTGACCTCCCTCCTCCCC
TCAGGCAGCTCTGACCTCCAGGTGTCCACCAAGGACGTGCCG
```

Melting point forward: 59.15°C
Melting point reverse: 60.45°C
Amplicon length: 307
BLAT for repetitive regions: Pass
No SNP's

Figure 7: Choosing the best primer to evaluate the substitution in PIK3CD

2.4 "R" Package & R-Scripted Programming

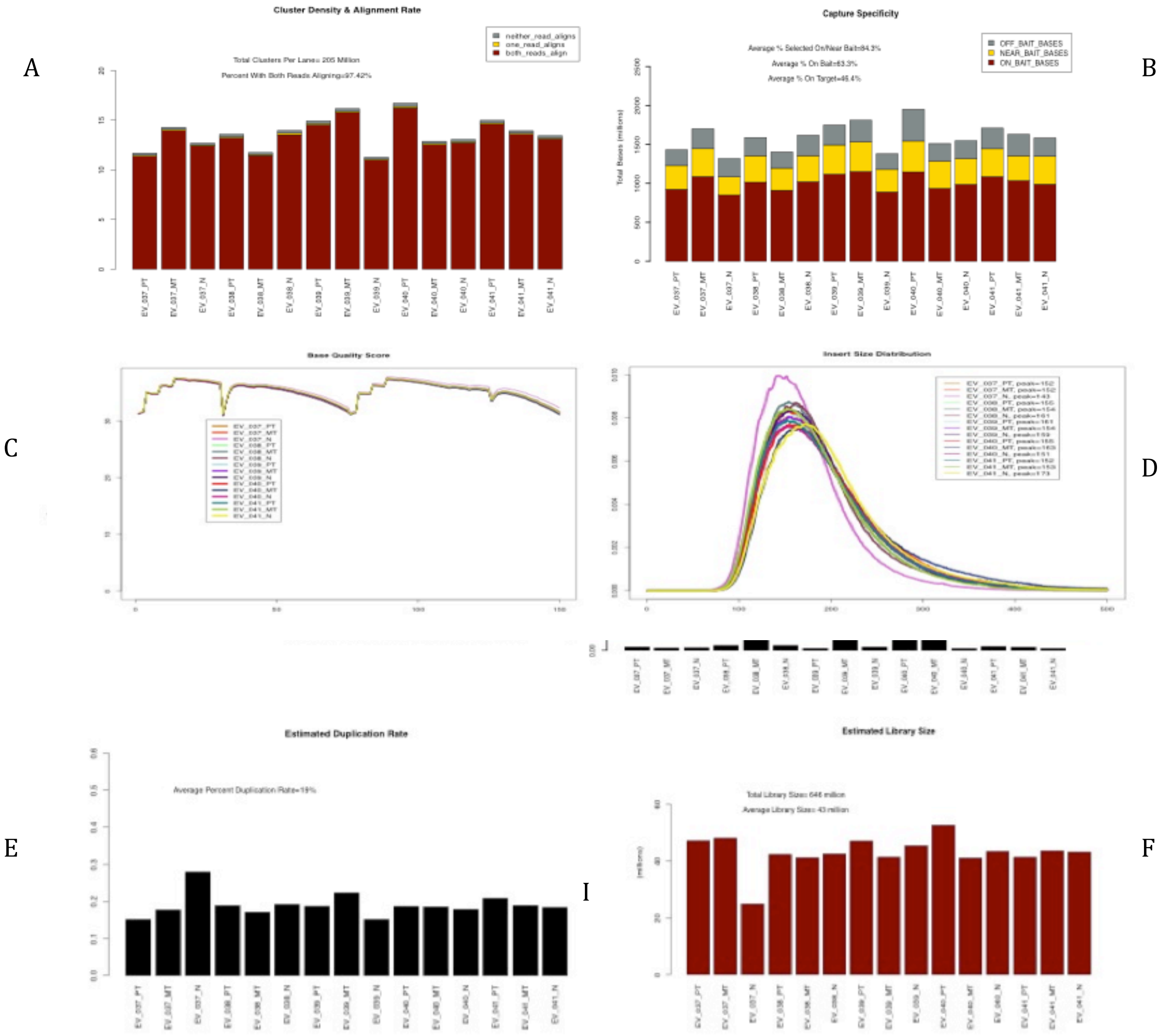
R is a programming language widely used among biostaticians and is highly relevant in bioinformatics (gene expression data, serial analysis of gene expression, etc.). R is a useful tool for plotting graphics, analyzing data, and fitting data to statistical models. It is open source, free, and maintained by a team of developers around the world. Traditionally, the metrics, QC metrics, and copy-number analysis of IMPACT were done on Microsoft Excel; however, they are now all redeveloped, improved, and automated in R-Studio. *R Cookbook*, *R Graphics Cookbook,* and *R in a Nutshell* have all been very helpful in writing code and integrating it into Next-Generation Sequencing assays, including IMPACT[31,32,33].

RESULTS:

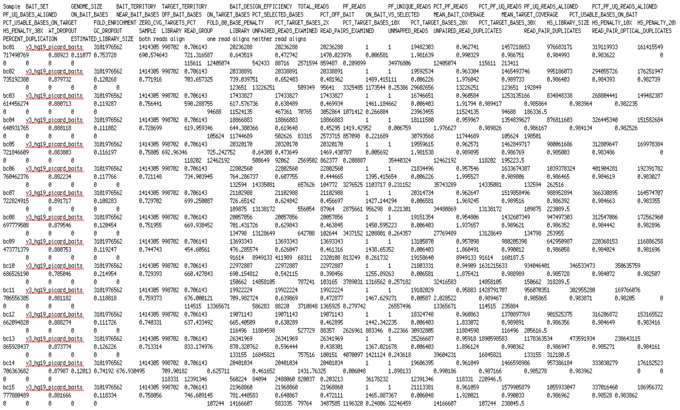
To illustrate the performance of the R-scripted program to compute metrics and copy number alterations, results will be presented for a project to identify biomarkers of metastasis in colorectal cancer. DNA was sequenced from primary tumors, metastatic tumors, and matched normal blood from 36 patients. Nevertheless, this program can and should be applied to enable discovery across virtually all cancers studied with NGS.

3.1 Visualization, Automation, and Development of Metric Calculations

The R-scripted program relates each metric computation to a comprehensible graph that will quickly allow investigators to match tissues, identify potential contamination, evaluate the performance of the sequencing technology, and document quality control (figure 8).

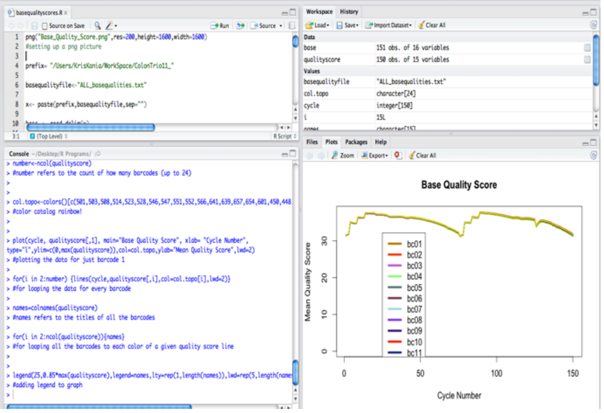


G



```
Sample  BAIT_SET        GENOME_SIZE     BAIT_TERRITORY  TARGET_TERRITORY        BAIT_DESIGN_EFFICIENCY  TOTAL_READS     PF_READS        PF_UNIQUE_READS PCT_PF_READS    PCT_PF_UQ_READS PF_UQ_READS_ALIGNED     PCT_PF_UQ_READS_ALIGNED
PF_UQ_BASES_ALIGNED     ON_BAIT_BASES   NEAR_BAIT_BASES OFF_BAIT_BASES  ON_TARGET_BASES PCT_SELECTED_BASES      PCT_OFF_BAIT    ON_BAIT_VS_SELECTED     MEAN_BAIT_COVERAGE      MEAN_TARGET_COVERAGE    PCT_USABLE_BASES_ON_BAIT
PCT_USABLE_BASES_ON_TARGET      FOLD_ENRICHMENT ZERO_CVG_TARGETS_PCT    FOLD_80_BASE_PENALTY    PCT_TARGET_BASES_2X     PCT_TARGET_BASES_10X    PCT_TARGET_BASES_20X    PCT_TARGET_BASES_30X    HS_LIBRARY_SIZE HS_PENALTY_10X  HS_PENALTY_20X
HS_PENALTY_30X  AT_DROPOUT      GC_DROPOUT      SAMPLE  LIBRARY READ_GROUP      LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED     UNMAPPED_READS  UNPAIRED_READ_DUPLICATES        READ_PAIR_DUPLICATES    READ_PAIR_OPTICAL_DUPLICATES
PERCENT_DUPLICATION     ESTIMATED_LIBRARY_SIZE  both reads align        one read aligns neither read aligns
bc01    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        20236288        20236288        20236288        1       1       19482303        0.962741        1457218653      976683171       319119933       161415549
717490769       0.88923 0.11077 0.753728        690.574643      721.316587      0.643519        0.472742        1470.023976     0.006581        1.981639        0.990329        0.986751        0.984993        0.983622                0
0       0       0       0                               115611  12405074        542433  88716   2571594 859487  0.209899        34976806        12405074        115611  213411
bc02    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        20338891        20338891        20338891        1       1       19592534        0.963304        1465493746      995186073       294055726       176251947
735192380       0.879732        0.120268        0.771916        703.657325      739.039751      0.652403        0.481962        1489.415111     0.006226        1.976042        0.989733        0.986403        0.984393        0.982739
        0       0       0       0       0                       123651  13226251                509349  95641   3325485 1173544 0.25386 29602656        13226251        123651  192849
bc03    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        17433827        17433827        17433827        1       1       16746651        0.960584        1253135166      834848338       268804441       149482387
614456274       0.880713        0.119287        0.756441        590.288755      617.576736      0.638489        0.469934        1461.184662     0.006403        1.91794 0.989417        0.985864        0.983964        0.982235
0       0       0       0       0                       94688   11524135        467361  70765   3052864 1071412 0.266884        23963455        11524135        94688   186336.5
bc04    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        18866883        18866883        18866883        1       1       18111580        0.959967        1354839627      876811603       326445340       151582684
640931765       0.888118        0.111882        0.728699        619.959346      644.380366      0.619648        0.45295 1419.42952      0.006759        1.976627        0.989826        0.986167        0.984134        0.982526
0       0       0       0       0                       105624  11744689        502626  83315   2573715 857098  0.221689        30793568        11744689        105624  198501
bc05    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        20320170        20320170        20320170        1       1       19559615        0.962571        1462849717      980061686       312809647       169978384
721846609       0.883803        0.116197        0.75805 692.96346       725.242752      0.64308 0.473649        1469.430787     0.005692        1.981538        0.989895        0.986769        0.985003        0.983486                0
0       0       0       0                               118202  12462192        508649  92062   2569502 862377  0.208887        35440324        12462192        118202  195223.5
bc06    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        22802560        22802560        22802560        1       1       21834496        0.957546        1633674387      1039378324      401904281       192391782
760462376       0.882234        0.117766        0.721148        734.903945      764.286737      0.607755        0.444665        1395.415654     0.006226        1.995527        0.989806        0.986465        0.984619        0.983027
        0       0       0       0       0                       132594  14335081                657626  104772  3276525 1103717 0.231152        35743289        14335081        132594  262516
bc07    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        21102988        21102988        21102988        1       1       20314734        0.962647        1519858496      988952894       366330895       164574707
722824915       0.891717        0.108283        0.729702        699.250087      726.65142       0.624842        0.456697        1427.144294     0.006581        1.969245        0.989516        0.986392        0.984663        0.983355
        0       0       0       0       0                       109875  13138172                556054  87964   2875661 956298  0.221301        34480869        13138172        109875  223089.5
bc08    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        20057856        20057856        20057856        1       1       19151354        0.954806        1432607349      947497303       312547086       172562960
697779508       0.879546        0.120454        0.751955        669.938452      701.431726      0.629843        0.463845        1450.595223     0.006403        1.937657        0.989621        0.986352        0.984442        0.982896
        0       0       0       0       0                       134798  13128649                642708  102644  3437152 1208801 0.264357        27769489        13128649        134798  253955
bc09    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        13693343        13693343        13693343        1       1       13105870        0.957098        980205398       642950987       220368153       116886258
473771379       0.880753        0.119247        0.744743        454.60561       476.285574      0.626047        0.461316        1438.65352      0.006403        1.860491        0.990012        0.986058        0.984024        0.981696
        0       0       0       0       0                       91614   8949133 411989  68311   2320108 813249  0.261732        19158640        8949133 91614   160187.5
bc10    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        22972887        22972887        22972887        1       1       21803331        0.94909 1631215633      934046401       346533473       350635759
686526190       0.785046        0.214954        0.729393        660.427843      690.154812      0.542115        0.398456        1255.89263      0.006581        1.875421        0.988989        0.985728        0.984072        0.982507
        0       0       0       0       0                       150662  14858105                787241  103165  3789031 1316562 0.257182        32416583        14858105        150662  318289.5
bc11    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        19922224        19922224        19922224        1       1       19102029        0.95883 1428791707      956070351       302955280       169766076
706556385       0.881182        0.118818        0.759373        676.000121      709.982724      0.639869        0.472877        1467.629271     0.00587 2.028522        0.989467        0.985865        0.983871        0.98205         0
0       0       0       0                               114515  13365671        586283  88220   3710848 1365925 0.279742        26557496        13365671        114515  235884
bc12    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        19071143        19071143        19071143        1       1       18324748        0.960863        1370897769      901525375       316206872       153165522
662094820       0.888274        0.111726        0.740331        637.433492      665.40509       0.630289        0.462895        1442.342235     0.006403        1.833072        0.989891        0.986356        0.984649        0.983416
        0       0       0       0       0                       116496  11884598                527729  88357   2626961 883346  0.22366 30932805        11884598        116496  205616.5
bc13    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        26341969        26341969        26341969        1       1       25266687        0.95918 1890598583      1178363534      473591934       238643115
865928437       0.873774        0.126226        0.713314        833.174976      870.320762      0.596444        0.438301        1367.021678     0.006403        1.896124        0.990362        0.986947        0.985271        0.984161
        0       0       0       0       0                       133155  16845821                757516  100151  4070097 1421124 0.243618        39604231        16845821        133155  312180.5
bc14    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        20401034        20401034        20401034        1       1       19606395        0.961049        1466598986      957386184       333030279       176182523
706363682       0.87987 0.12013 0.74192 676.930495      709.90182       0.625711        0.461652        1431.76325      0.006048        1.898133        0.990186        0.987166        0.985278        0.983962                0       0
0       0       0                               118331  12391346        560224  84094   2488060 820037  0.203213        36178232        12391346        118331  220946.5
bc15    v3_hg19_picard_baits    3101976562      1414305 998702  0.706143        21968860        21968860        21968860        1       1       21113381        0.961059        1579905879      1055933047      337016460       186956372
777880489       0.881666        0.118334        0.758056        746.609145      781.448583      0.640867        0.472111        1465.887367     0.006048        1.920021        0.990033        0.986962        0.98528 0.983862
0       0       0       0       0                       107244  14166607        583335  79764   3487585 1196320 0.24806 32246459        14166607        107244  238045.5
```

H



```
png("Base_Quality_Score.png",res=200,height=1600,width=1600)
#setting up a png picture

prefix= "/Users/KrisKania/WorkSpace/ColonTrio11_"

basequalityfile<-"ALL_basequalities.txt"

x<- paste(prefix,basequalityfile,sep="")

base<-read.delim(x)
```

```
> number<-ncol(qualityscore)
> #number refers to the count of how many barcodes (up to 24)
>
>
> col.topo<-colors()[c(501,503,508,514,523,528,546,547,551,552,566,641,639,657,654,601,450,448
> #color catalog rainbow!
>
>
> plot(cycle, qualityscore[,1], main="Base Quality Score", xlab= "Cycle Number",
type="l",ylim=c(0,max(qualityscore)),col=col.topo,ylab="Mean Quality Score",lwd=2)
> #plotting the data for just barcode 1
>
> for(i in 2:number) {lines(cycle,qualityscore[,i],col=col.topo[i],lwd=2)}
> #for looping the data for every barcode
>
> names=colnames(qualityscore)
> #names refers to the titles of all the barcodes
>
> for(i in 2:ncol(qualityscore)){names}
> #for looping all the barcodes to each color of a given quality score line
>
>
> legend(25,0.85*max(qualityscore),legend=names,lty=rep(1,length(names)),lwd=rep(5,length(name
> #adding legend to graph
>
```

Data	
base	151 obs. of 16 variables
qualityscore	150 obs. of 15 variables
Values	
basequalityfile	"ALL_basequalities.txt"
col.topo	character[24]
cycle	integer[150]
i	15L
names	character[15]

Figure 8: An Automated and more developed QC Metric Computation on colorectal cancer NGS data

(A) Cluster Density & Alignment Score, (B) Capture Specificity, (C) Base Quality Score, (D) Insert Size Distribution, (E) Fingerprint Matching, (F) Contamination, (G) Estimated Duplication Rate, (H) Estimated Library Size. These calculations (and graphs) are now automatically computed from raw data with the R-Scripted program (I). This program also offers previously uncalculated metrics such as mean exon coverage (in addition to mean target coverage). Most importantly, the program is universally functional across any permutation or combination of up to 24 tumors/normals/etc. prioritizing speed and automation across virtually all laboratory scenarios encountered in IMPACT or the other targeted cancer assays in published literature.

3.2 Automated Copy-Number Plotting with "Normal to Normal", "Tumor to all Normals", and "Tumor to Tumor-Median" Capability

IMPACT is run on a diverse set of projects, which include the conventional tumor to normal comparisons, but also, primary tumor to metastatic tumor to normal comparisons (for discovery projects on tumor evolution), tumor to all tumor comparisons, and other variations, which have severely limited automation programs in the past. However, this project, with the tools of R, is built with the flexibility to approach virtually all variations of NGS on cancerous tissue.

Furthermore, with the computational power of R, copy-number graphs exist for each normal compared to all normals, and each tumor compared to all normals. On Microsoft Excel, it would be possible to make these comparisons, however, this would make an already long process even longer.

The advantage of comparing normals to all normals is that one can easily identify contamination or germ line copy-number alterations in the normals. The advantage of comparing tumors to every normal exists when the tumor to matched normal comparison

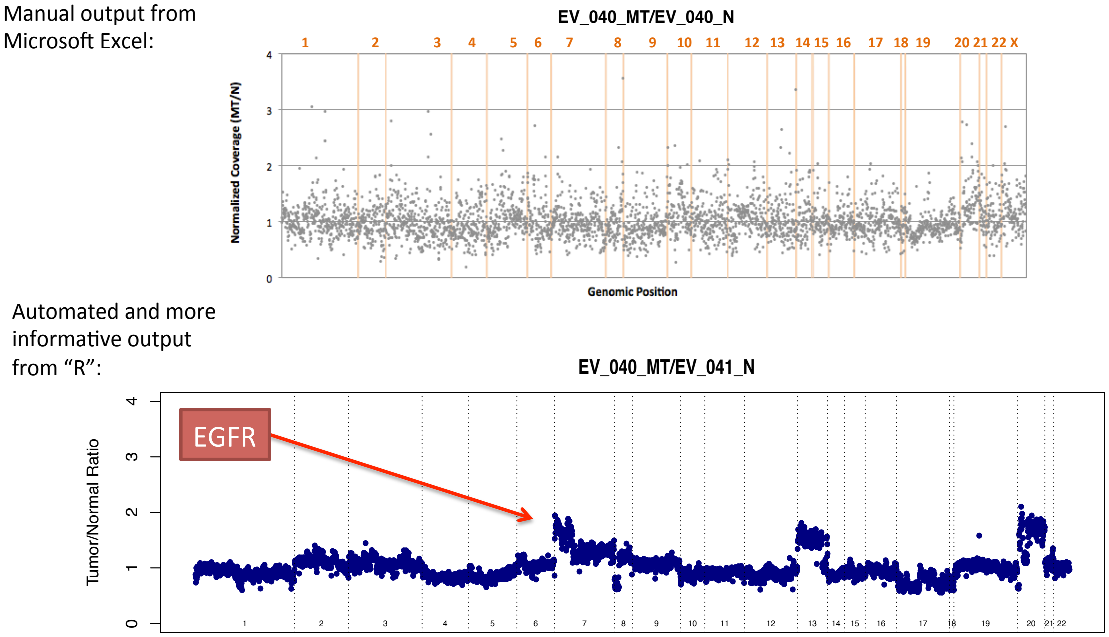


Figure 9: Advantage of Copy-Number Plotting with R compared to competing methodologies

results in a "messy" copy-number graph, where amplifications or deletions are not clear. (This usually occurs due to machine artifacts during sequencing, differences in DNA quality, or differences in input size during library preparation). With the R-scripted program, tumor tissue is automatically compared to not only the matched normal, but also, all other normals, often times resulting in a cleaner tumor/normal copy-number graph. It is scientifically sound to compare a given tumor to an unmatched normal because in theory, all normal samples should have two copies of every exon in the twenty-two autosomes sequenced in IMPACT. If there is suspicion that there is a germ-line amplification or any contamination in the normal compared to the given tumor, the investigator can always consult the normal to all normal comparisons.

In figure 9, a traditional pipeline using Microsoft Excel was too messy to provide any meaningful conclusions for this patient with colorectal cancer. However, with R, it is clear that there is a EGFR amplification, an actionable cancer target, on the 7th chromosome. The copy-number is cleaner on R because the program is actually comparing the tumor to all normals in the pool, rather than just the matched normal, allowing the investigator to choose the most informative graph. After comparing this "miracle" normal to other normals in the pool, the normal-to-normal comparison indicates that even though the normal does not match the tumor, it does not suffer from any autosomal bias, and is perfectly legitimate (Figure 10). In a prospective setting, detecting the EGFR amplification, with the tumor to all normals comparison and the added confidence of the normal to all normals comparison, would open important doors to targeted therapies such as cetuximab, gefitinib, erlotinib, and panitumumab. In addition, in the event that there are no normals in the pool, the program is prepared to compare the given tumor to the median exon values across all tumors.

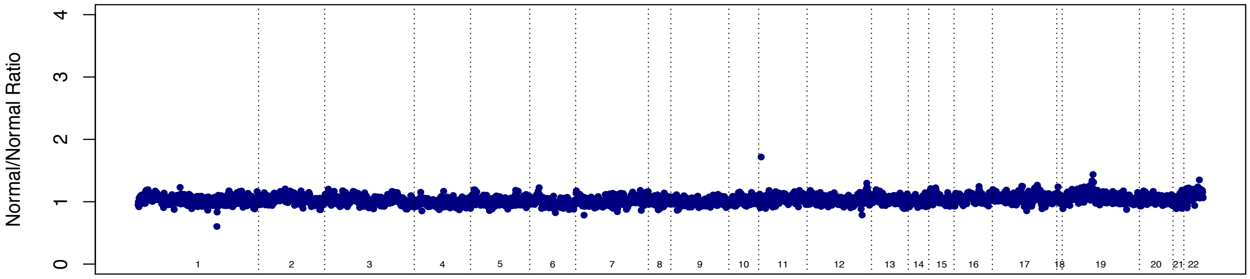
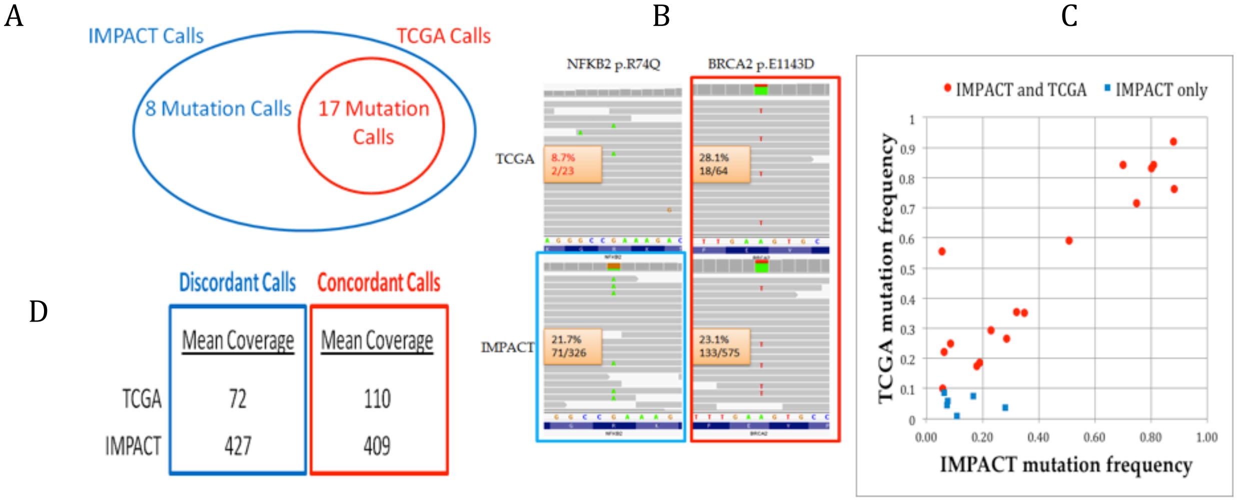


Figure 10: Relatively quiet copy-number comparison between the normal used in Figure 7 and another normal in the given pool.

3.3 Comparison to The Cancer Genome Atlas (TCGA)

Interestingly, these computational efforts, compounded with the existing experimental design of IMPACT, has enabled mutation calling that does not only match, but occasionally, even surpasses the NGS assay of The Cancer Genome Atlas (TCGA). When frozen and FFPE tissue first screened by the TCGA[34] was run on IMPACT, the investigators using IMPACT called all 17 mutations found by TCGA, and 8 additional mutations not found by TCGA. These mutations were all confirmed with sanger sequencing (primer evaluation).



	Discordant Calls	Concordant Calls
	Mean Coverage	Mean Coverage
TCGA	72	110
IMPACT	427	409

Figure 11: Comparison with TCGA
A comparison of 6 frozen ovarian tumors sequenced by IMPACT and TCGA revealed that (A) all 17 mutations found by TCGA in IMPACT genes were detected by IMPACT. Additionally, 8 mutations not found by TCGA were detected by IMPACT, as seen in (B) IGV screenshots. (C) These mutations were at low frequency in both tumors but (D) detected due to higher coverage.

DISCUSSION:

Automating the metric computation of NGS has provided an easier transition from the raw data output of sequencing (Illumina HiSeq 2000 FastQ file) to mutation calling. The programing language R has served to bridge the gap in this transition, replacing the traditional, and less robust, approach of Microsoft Excel and the other competing programs. While many labs performing NGS are not core-facilities, they can still benefit from the computational and statistical power of a well-scripted program.

Speed: To begin with, for IMPACT, as well as other assays, this project guarantees speed. On average, running 12 patients (24 tissue samples) on IMPACT implies at least 3 hours behind the computer or calculator performing metric computation or plotting copy-number graphs. This process, on average, now takes just 19 seconds with the R-Scripted program. Thus, an investigator can shift laboratory resources from computation to interpretation, while performing critical quality controls checks to identify artifacts that could lead to false positive mutation calls or spurious conclusions.

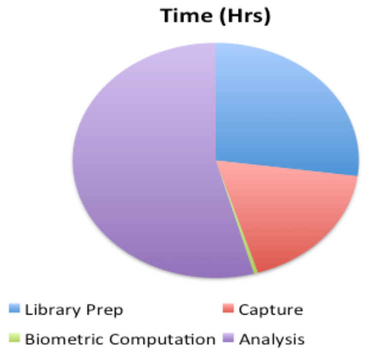


Figure 12: Average time spent on metric computation with R $\approx \mathbf{19\ seconds\ (CPU = 2.2\ GHz, RAM = 4\ Gb)}$

Confidence: The R-Scripted program has not only made the assay faster, but also, more informative. In addition to the traditional metric calculations of the existing pipeline, the R-Script provides 100% bar plots, average exon coverage, normal to all normal copy number, and tumor to all normal copy number data. With these additional resources, labs can be more confident in their calls as they move into Integrated Genomic Viewer (IGV) and rate mutations. This will also allow for more informed choices when contemplating primer evaluation on a questionable mutation.

FUTURE WORK:

One important limitation of the R-Scripted program is that the investigator must visually inspect the copy number graphs and then manually choose the most informative one. The program produces graphs that compare one tumor to every normal in a given pool. Currently, all graphs are displayed in the PDF, including those that are not particularly informative. For example, in a pool of 12 patient samples, the investigator would be looking for the 12 cleanest tumors to normal graphs from a PDF of 78 graphs. While it is important to preserve this element of human intuition, I am experimenting with computational methods that can output graphs that would match visual inspection.

One method that is currently being tested is a square-adjacent exon calculation: the addition of the square of the difference in tumor/normal ratio between adjacent exons across all exons, where the smallest sum relates to the most informative graph.

$$\sum_{n=1}^{4655} (exon_{n+1} - exon_n)^2$$

*There are 4656 exons captured with the current version of baits in IMPACT

Taking the square, will exaggerate the sum of graphs with particularly messy copy numbers more than simply taking the sum of an absolute value of the difference between adjacent exons. The "R" script could then easily select to output (PDF) only the graph with the smallest sum in a for-loop for each tumor.

Another option is to segment the copy number data using circular binary segmentation via DNAcopy, an "R" package developed by BioConductor that has resolved similar issues.

CONCLUSION:

The establishment of the experimental and computational efforts of IMPACT will have immediate, far-reaching benefits for translational and clinical research and will provide the foundation for personalized cancer medicine. Systematic profiling of every cancer gene in tumor DNA from every cancer patient would improve diagnosis and reveal the spectrum of alterations across tumor types, the presence of mutations with potential therapeutic implications in unexpected contexts, and their patterns of co-occurrence that might direct treatment choice. Profiling these same genes retrospectively across a vast collection of clinically annotated FFPE tumors would enable the discovery of significant oncogenic mutations in rare or understudied tumor types and the identification of genomic biomarkers exhibiting correlations with clinical outcomes or phenotypes in every cancer. These efforts collectively embody the goal to produce better outcomes in cancer patients and make cancer a more manageable disease.

ACKNOWLEDGEMENTS:

This research was conducted in the Human Oncology & Pathology department of Memorial Sloan-Kettering throughout the summer of 2012. I would like to thank everyone who has made this project possible. First, I would like to thank Dr. Michael Berger for allowing me to conduct my research at Zuckerman Research Center and for providing me invaluable insights into targeted cancer therapy. Second, I would like to thank Dr. A Rose Brannon, for guiding me through the programming languages of R and MATLAB, and for her indispensible mentorship. Last, I would like to thank Sasinya Scott, Gregory McDermott, Helen Won, for their endless help, the coordinators of HOPP for this opportunity, and my fellow HOPP summer students for a fantastic eight weeks.

REFERENCES

[1] MacConaill, L. E., & Garraway, L. A. (2010). Clinical implications of the cancer genome. *Journal of Clinical Oncology*, 5219-5228.

[2] Wagle, N., & Berger, M. F. (2012). High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discovery*, 83-93.

[3] Bentley, D., Balasubramanian, S., & Swerdlow, H. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 53-59.

[4] Drmanac, R. *et al.* (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 78-81.

[5] Marguilies, M. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 376-380.

[6] Shendure, J. *et al.* (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 1728-1732.

[7] Wheeler, D. *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 872-876.

[8] Ley, T. J. *et al.* (2008). DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. *Nature*, 66-72.

[9] Mardis, E. R. *et al.* (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. *New England Journal of Medicine*, 1058-1066.

[10] Pleasance, E. D. *et al.* (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 191-196.

[11] Weir, B., Zhao, X., & Meyerson, M. (2004). Somatic alterations in the human cancer genome. *Cancer Cell*, 433-438.

[12] Mitsudomi, T. *et al.* (2009). Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harboring mutations of the epidermal growth factor receptor: an open label, randomized phase 3 trial. *Lancet Oncology*, 121-128.

[13] Rosell, R. *et al.* (2009). Screening for epidermal growth factor receptor mutations in lung cancer. *New England Journal of Medicine*, 958-967.

[14] Campbell, P. J. *et al.* (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.*, 722-729.

[15] The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 1061-1068.

[16] Wood, H.M. *et al.* (2010). Using next-generation sequencing for high-resolution multiplex analysis of copy number variation from nanograms quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids*, 38.

[17] Gallegos, R.M. *et al.* (2007). EGFR and K-Ras mutation analysis in non-small cell lung cancer: comparison of paraffin embedded versus frozen specimens. *Cell Oncology*, 257-264.

[18] Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews: Genetics*, 685-691.

[19] Carter, H. *et al.* (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, 6660-6667.

[20] Kerick, M., Isau, M., & Timmermann, B. (2011). Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *Medical Genomics*, 68-81.

[21] Blumenstiel, B., Kristian, C., & Fisher, S. (2010). Targeted exon sequencing by in-solution hybrid selection. *Current Protocols in Human Genetics*, 66-84.

[22] Fisher, S., Barry, A., & Abreu, J. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology*, 12-27.

[23]Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *CSH Protocols*, 1-9.

[24]Parkinson, N., Maslau, S., & Ferneyhough, B. (2012). Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Research*, 125-133.

[25] Campbell, P. J. *et al.* (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.*, 722-729.

[26]Ramensky, V., & Bork, V. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids*, 3894-3900.

[27]Aird, D., Ross, M., & Chen, W. (2011). Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biology*, 1-14.

[28] Beroukhim, R. *et al.* (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, 899-905.

[29] Stephens, P.J. *et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 1005-1010.

[30] Thomas, R.K. *et al.* (2007). High-throughput oncogene mutation profiling in human cancer. *Nature Genetics*, 347-348.

[31] Teetor, P. (2011). *R cookbook*. (1st ed.). San Francisco, CA: O'Reily.

[32] Adler, J. (2009). *R in a nutshell*. (1st ed.). San Francisco, CA: O'Reily.

[33] Mittal, H. (2011). *R graphs cookbook*. (1st ed.). Birmingham, UK: Packt Publishing

[34] Cancer Genome Atlas Research Network. (2011). Integrated genomic analysis of ovarian carcinoma. *Nature*, 609-615.