Targeted re-sequencing of cancer-related genes from matched FFPE and Fresh-Frozen tumor samples using the Illumina sequencing platform

AACR 2012
# 3179

Marina Bibikova[1], Jeremy Chien[2], Vincent Ho[1], Craig April[1], Sarah Munchel[1], Joseph Cottrell[1], Samantha Cooper[1], Russell Grocock[1], Fiona Nielsen[1], Yaman Tarabishy[2], Daniel Visscher[2], Megan Manion[3], Jonathan Liu[3], Eric Wieben[2], Lynn Hartmann[2], Kim Kalli[2], Viji Shridhar [2], and Jian-Bing Fan[1].

[1]Illumina, Inc., San Diego, California; [2]Mayo Clinic, Rochester, Minnesota; [3]Softgenetics, LLC, State College, Pennsylvania

1. Introduction

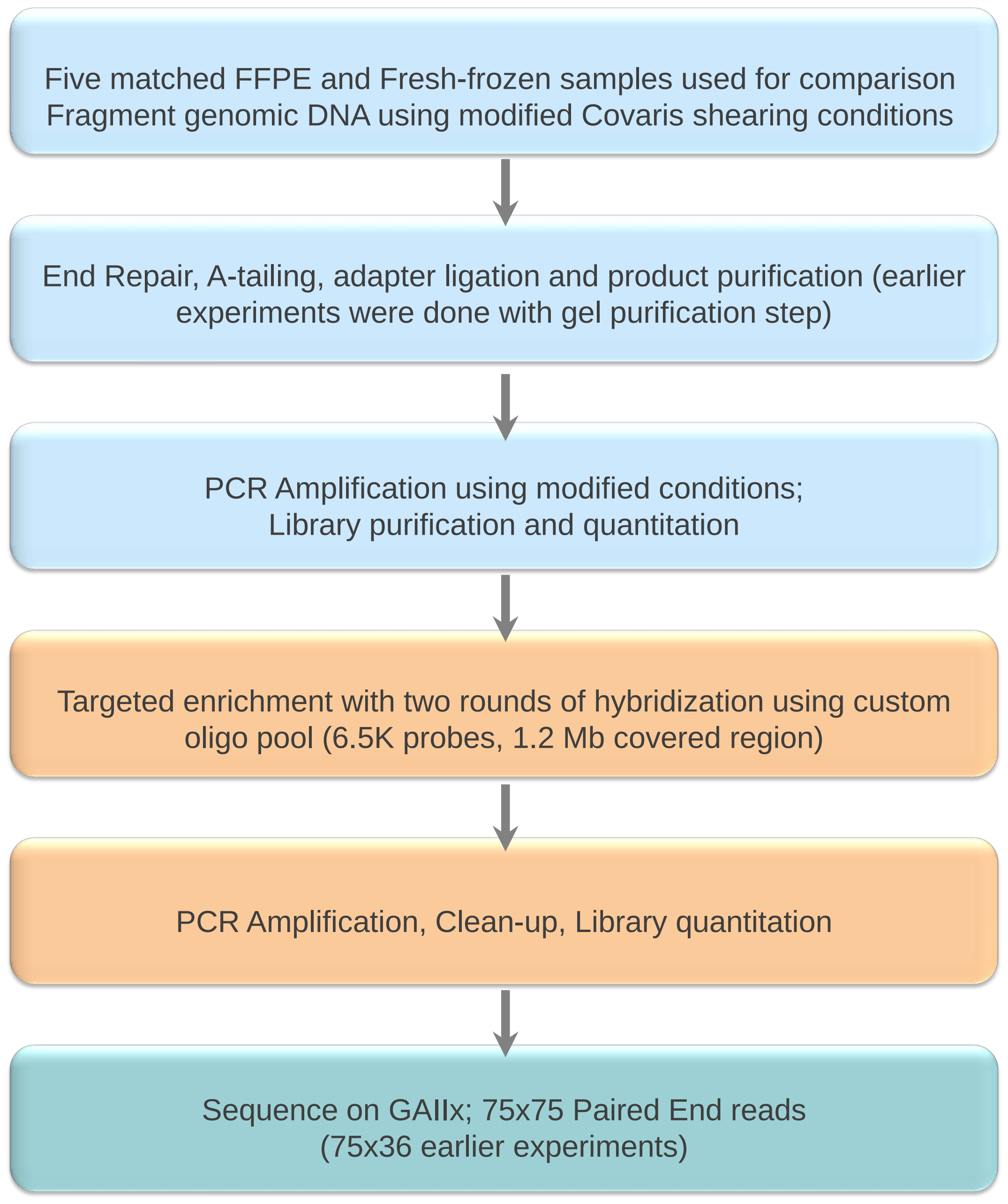
High throughput sequencing technologies open up a new dimension in cancer genomics by enabling the characterization of cancer genomes at base-pair resolution. The large-scale genomics projects are mostly utilizing fresh-frozen biospecimens in their studies to characterize cancer genomes. Unlike fresh-frozen samples, archived formalin-fixed paraffin-embedded (FFPE) tissues are more readily accessible, and are often associated with known clinical outcomes and more complete clinical annotations. However, sequencing library preparation methods need to be further optimized with regard to applicability to FFPE samples.

To test whether next-generation sequencing technologies could report accurate sequencing results and overcome previously reported artifacts associated with formalin fixation, we compared whole-genome and targeted enrichment DNA sequencing data obtained from five FFPE tumor samples for which matching frozen tissues were available. Using 1 µg of genomic DNA as input material and a modified TruSeq® sample preparation protocol, we successfully generated sequencing libraries. We prepared a custom enrichment pool that targeted exons from 215 cancer-related genes and utilized this pool to perform the pull-down and sequencing of approximately 1.3 Mb region with average 400x coverage depth to test if deep sequencing would help in validation of tumor-specific somatic mutations.

We used the Illumina sequence analysis pipeline and the Windows-based second generation DNA sequencing software NextGENe to analyze the data and identify sequence variations that are different from the human reference genome. CNV detection was overall higher among FFPE samples compared with fresh-frozen samples, demonstrating that tissue processing can impact sequencing data quality. We obtained good concordance in variant calls between matched FFPE and fresh-frozen samples. Discordant variant calls were mainly due to low depth of coverage in the regions where variant calls were made.

Improvements to DNA sequencing methods for archived samples will significantly enhance cancer research and will result in more reliable prediction of individual cancer therapies.

2. Targeted Enrichment Overview



Five matched FFPE and Fresh-frozen samples used for comparison
Fragment genomic DNA using modified Covaris shearing conditions

End Repair, A-tailing, adapter ligation and product purification (earlier experiments were done with gel purification step)

PCR Amplification using modified conditions;
Library purification and quantitation

Targeted enrichment with two rounds of hybridization using custom oligo pool (6.5K probes, 1.2 Mb covered region)

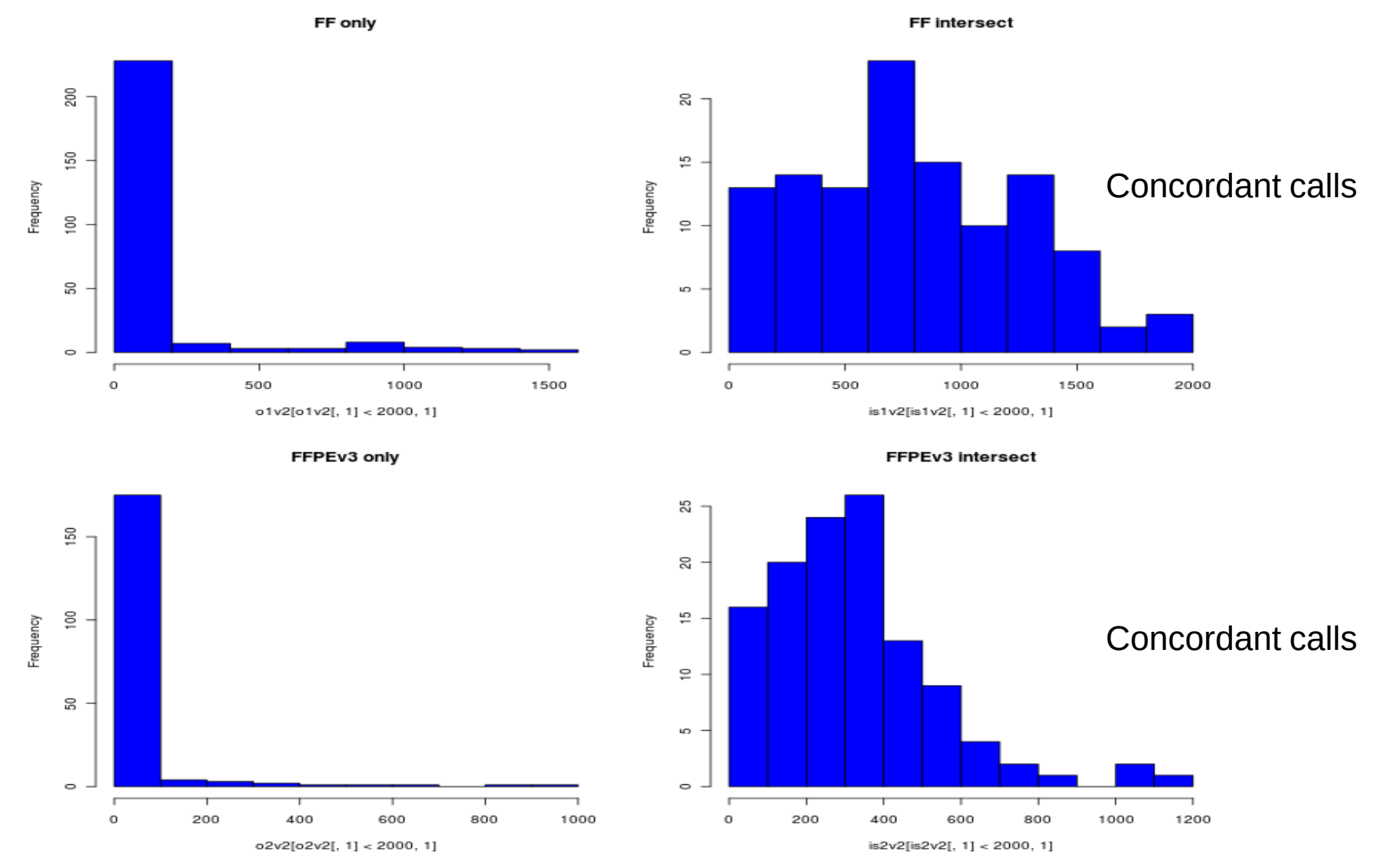
PCR Amplification, Clean-up, Library quantitation

Sequence on GAIIx; 75x75 Paired End reads
(75x36 earlier experiments)

3. Data Analysis

- In addition to CASAVA 1.8 data analysis pipeline, we used Windows-based NextGENe DNA sequencing software to analyze the data and identify sequence variations that are different from human reference genome.
- Fastq files from CASAVA 1.8 output were processed into FASTA format by NextGENe software. During this conversion, reads not meeting preset QC criteria were filtered out and not used in read mapping to the human reference genome.
- SNP/ Indel Discovery Application was used for further analysis.
- Mutation reports were generated in the NextGENe Viewer and analyzed in pairs through the JMP statistical analysis program, as well as PERL and MATLAB scripts to determine the number of concordant and discordant calls between matched samples.
- Concordant calls were defined as calls with matching genotypes in matched samples.
- Discordant calls were made where there was a difference in variant call between the matched samples at the same reference position.

4. Effect of coverage on variant call concordance



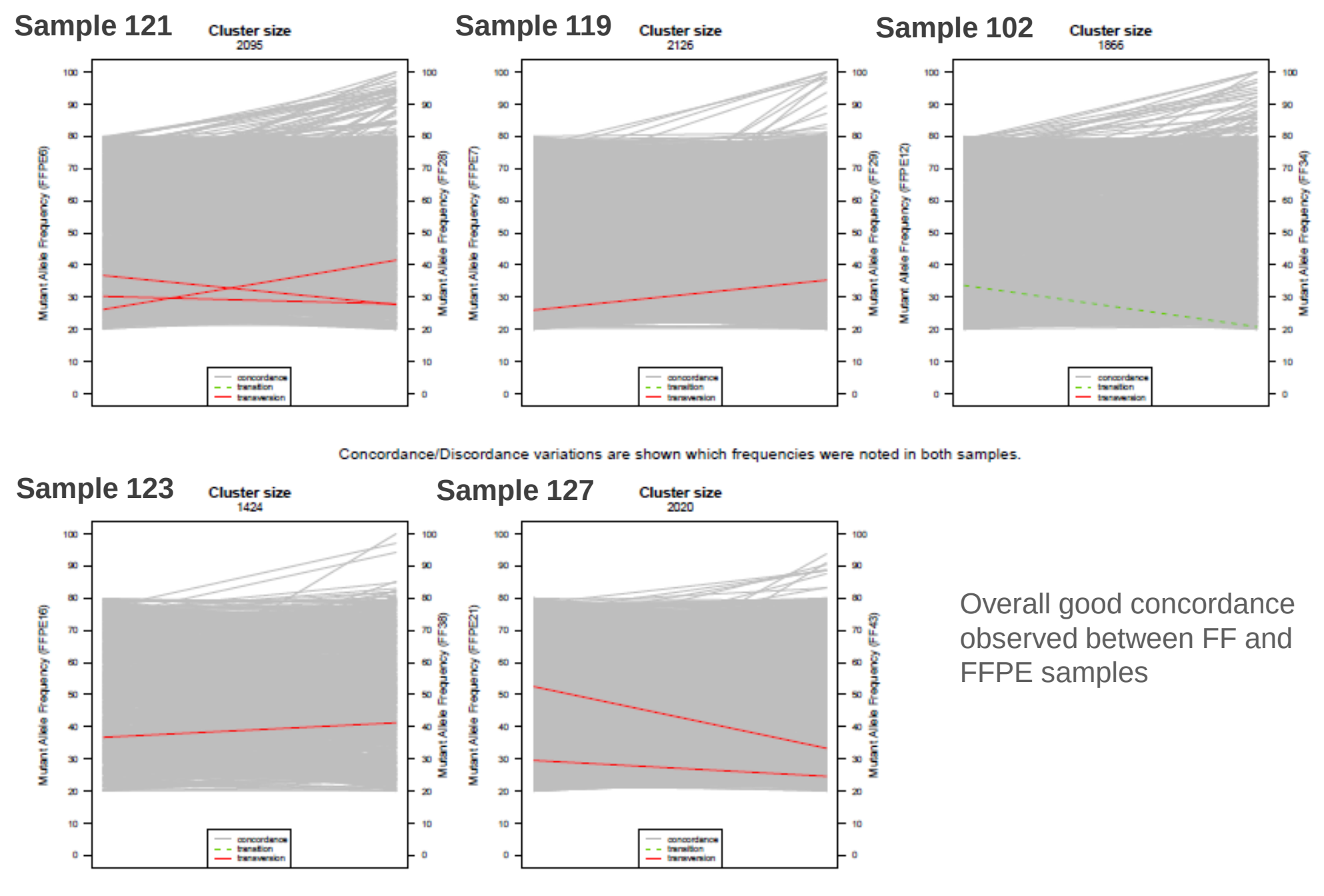
Concordant calls

Concordant calls

Sample ID	Concordant calls	Discordant calls	% Concordance
Sample_121	6498	76	**98.84**
Sample_119	6456	63	**99.03**
Sample_102	6088	78	**98.73**
Sample_123	6418	87	**98.66**
Sample_127	6491	57	**99.13**

- Discordant variant calls are mainly due to low coverage of reads in the region where variant calls are being made.

5. Concordance analysis

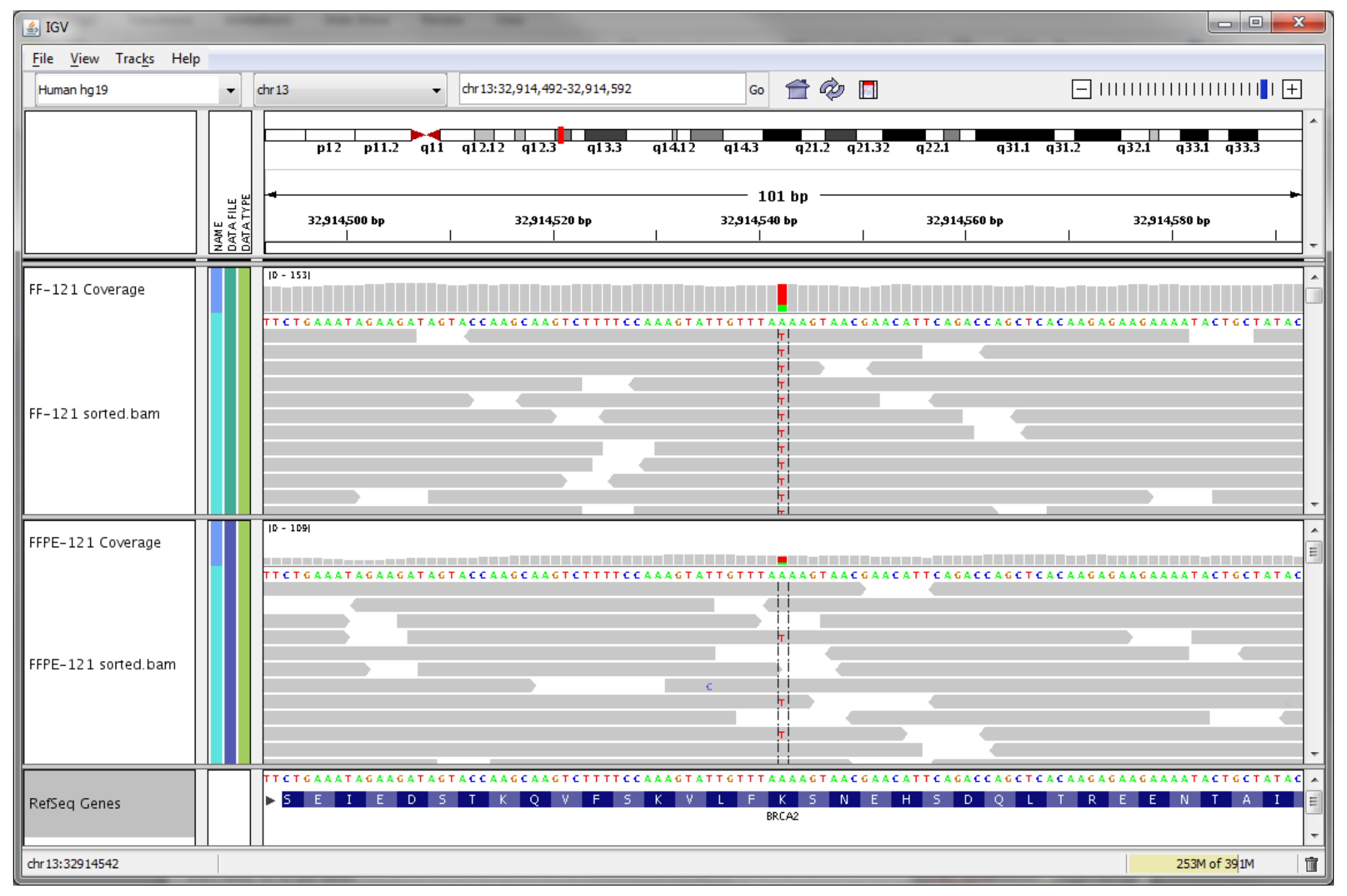


Concordance/Discordance variations are shown which frequencies were noted in both samples.

Overall good concordance observed between FF and FFPE samples

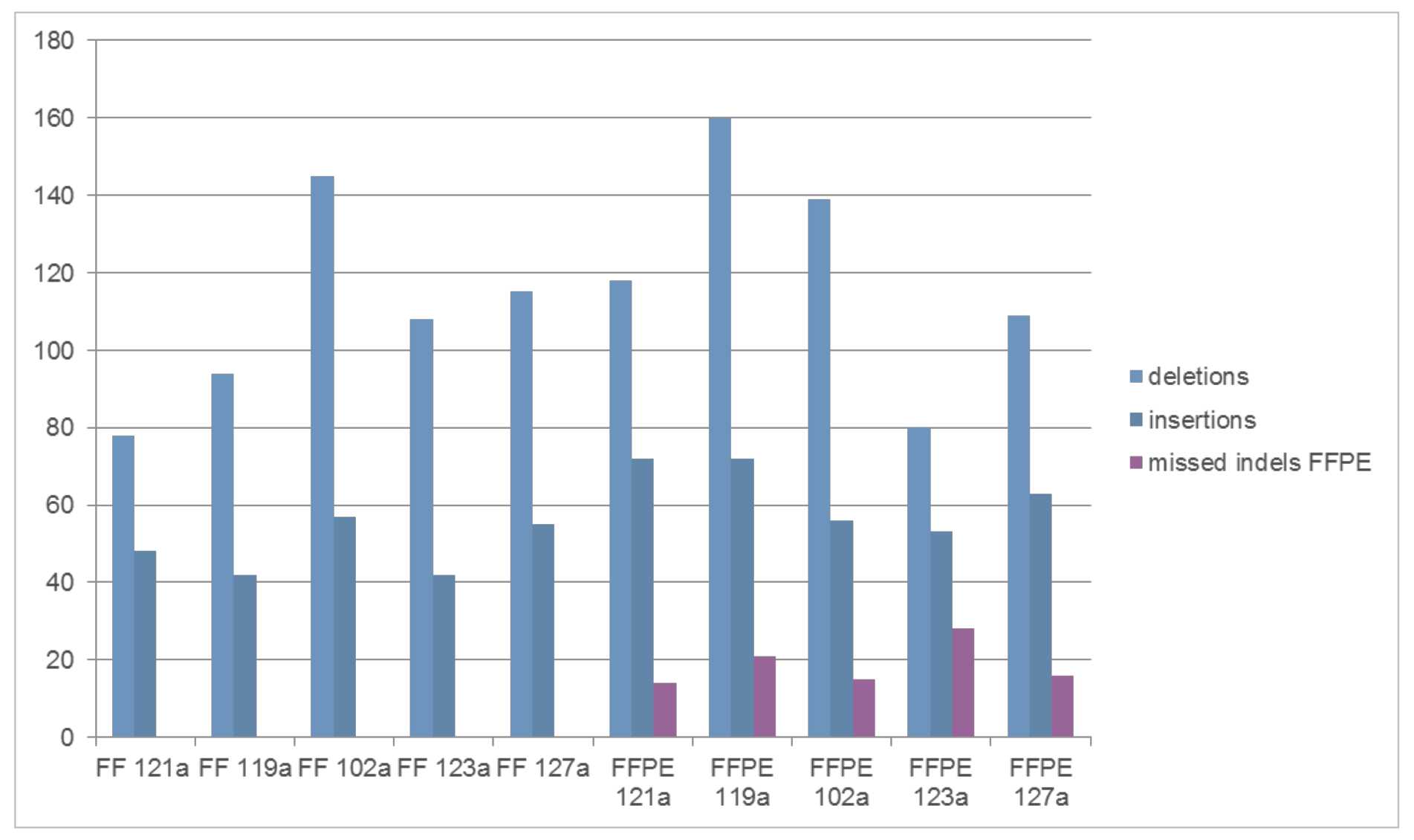
- We called variants if variant frequency was > 20% in FF or FFPE samples.
- Variant frequency < 20% is not included in the graphs.

6. BRCA2 mutation in FF-121 and matched FFPE samples



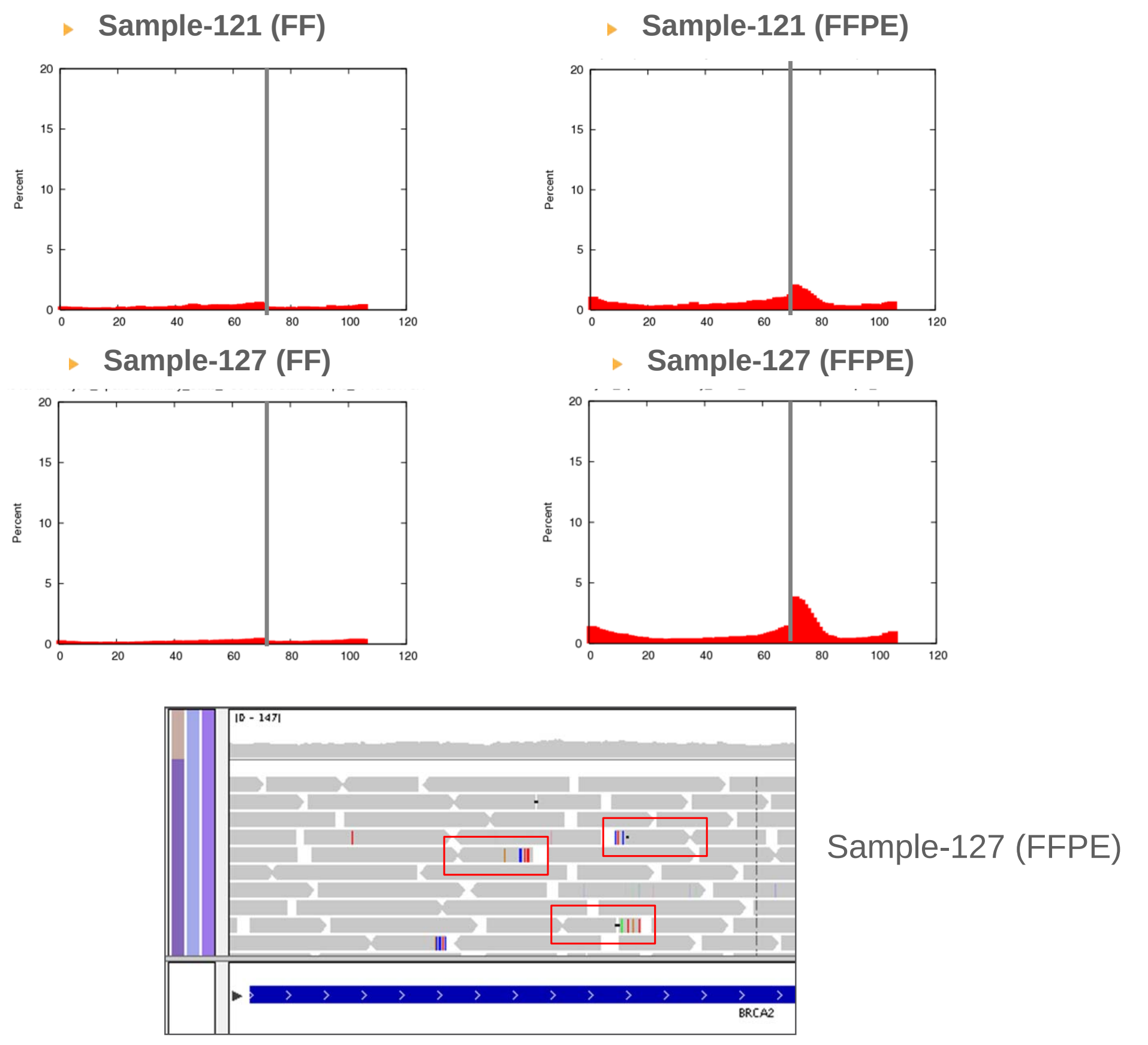
- An example of SNV discovered through WGS of the FF-121 sample.
- The same SNV can be detected in both FF and FFPE samples after targeted enrichment

7. Challenges with structural variant detection



- Insert size is shorter and more variable in FFPE samples
- Accurate indel detection in FFPE samples is challenging

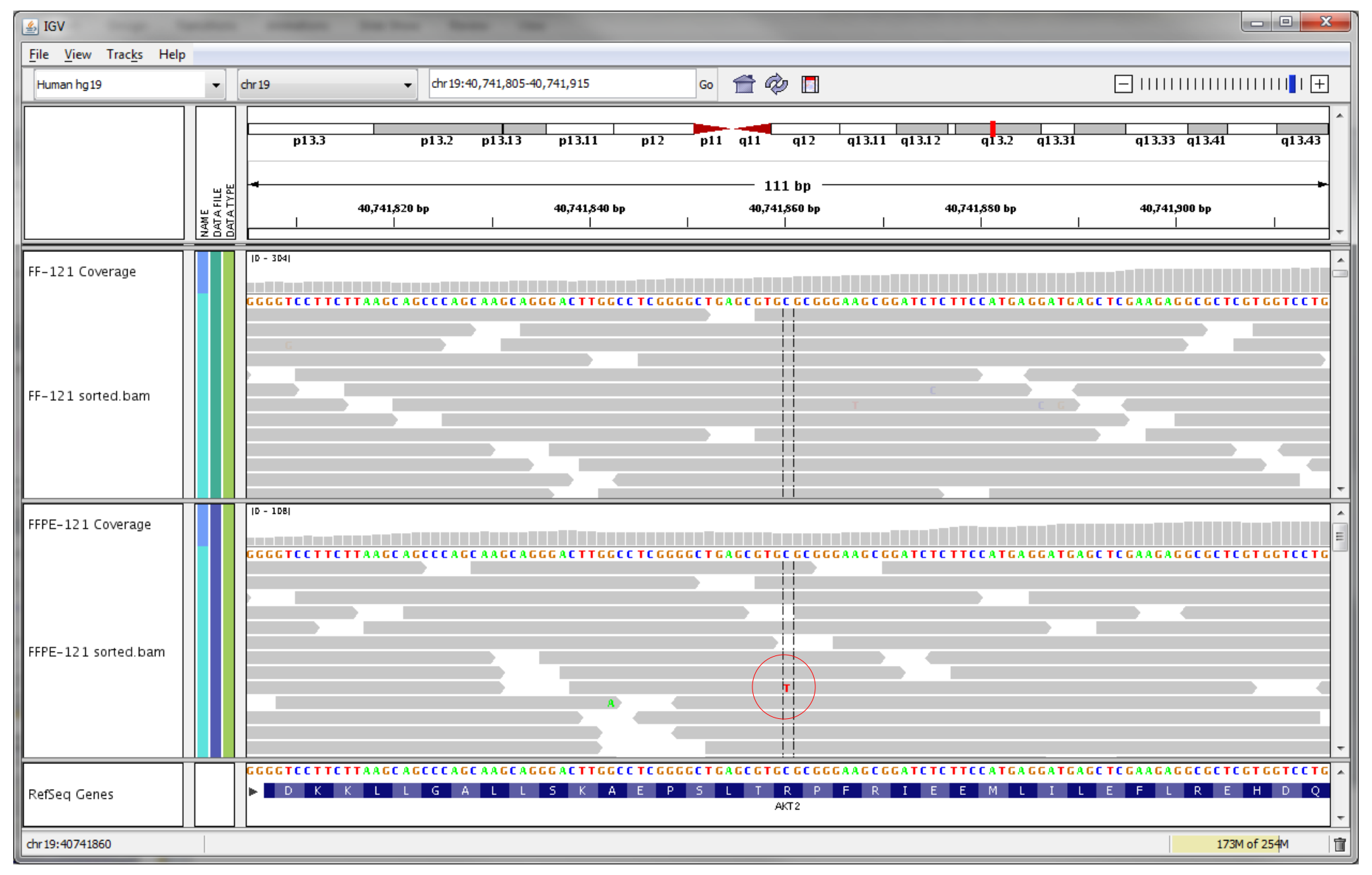
8. Elevated error rate at the start of Read 2



Sample-127 (FFPE)

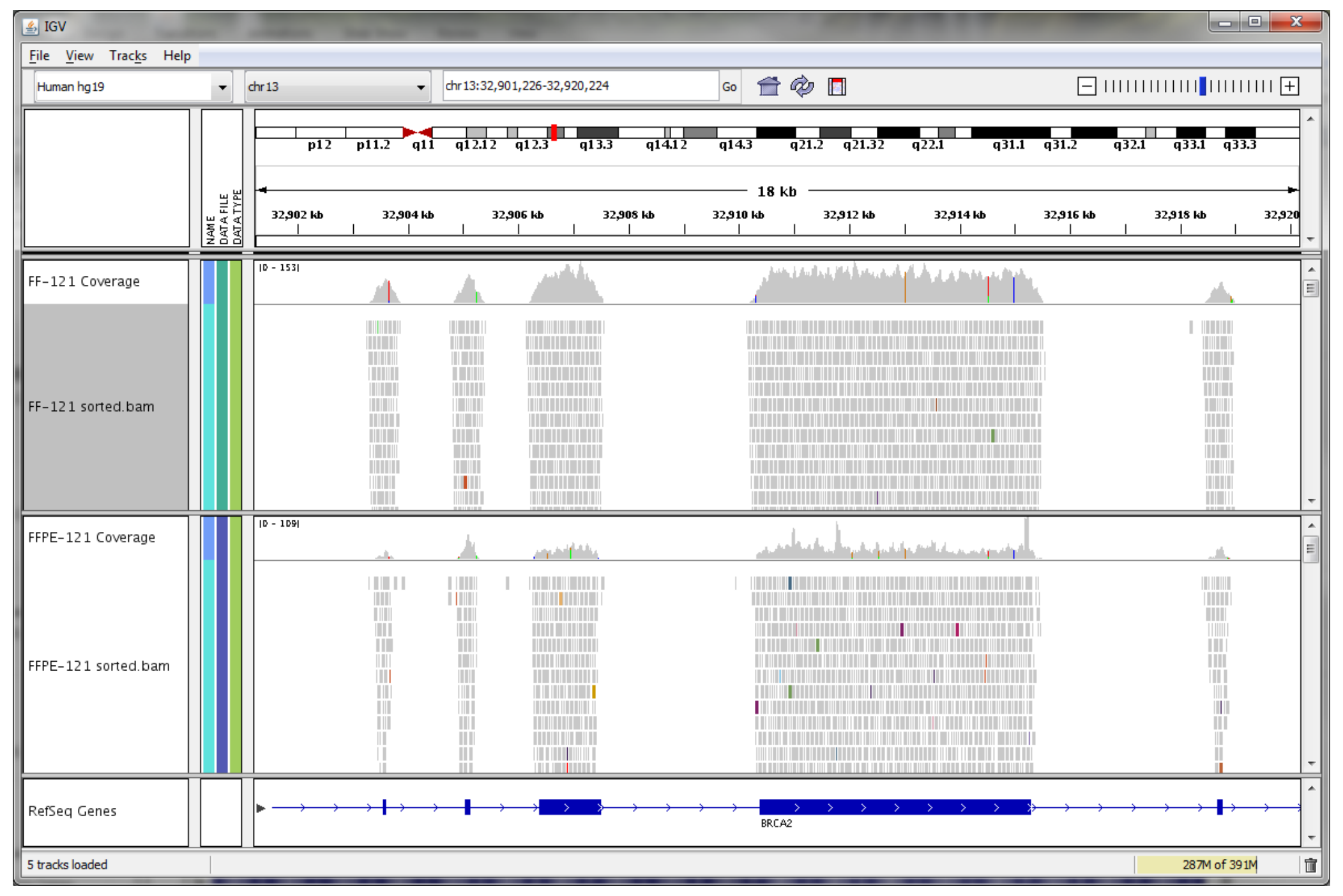
- Elevated error rate is often detected at the beginning of Read 2 in FFPE samples only

9. Effect of Cytosine Deamination



- Earlier reports indicate that most of the artificial mutations recorded in DNA extracted from FFPE are either C-T or G-A transitions.
- C->T transition (result of Cytosine deamination) is present in FFPE but not in the matched FF sample

10. Depth and uniformity of coverage



- After targeted pool down, samples were sequenced to 400x depth on average
- Large number of PCR duplicates result in overall lower unique coverage in FFPE samples

11. Conclusions

- Overall good concordance in variant calls was observed between FF and FFPE samples
- FFPE samples can be used for biomarker validation in clinical studies
- Challenges with FFPE analysis:
  - Increased number of false positives due to C->T deamination
  - Lower coverage uniformity
  - Elevated error rate at the start of read 2
- Protocol optimization may further improve data quality
- Data analysis pipeline optimized for FFPE samples is needed for robust variant calling

illumina®