**Ovarian cancer subtypes discovery with functional data based on next generation RNA-SEQ technology**

**Abstract**

**Background**

Ovarian cancer (OvCa) is the fifth-leading cause of cancer death among women in the United States which mainly caused by late diagnosis and great progess of the ovarian cancer. Treatment for advanced-stage, high-grade serous ovarian cancer has improved over the past 20 years with the introduction of the platinums and, more recently, taxane-based chemotherapy. However, the majority of patients eventually the majority of ovarian cancer patients will go on to develop chemotherapy-resistant disease, mainly due to the emergence of chemotherapy resistance[[1](#_ENREF_1)] which is suggest to be since for the high heterogeneity. Under such condition, replace with suitable or better chemotherapy drug for potential chemotherapy-resistant patients is becoming urgent according to different subtype of the Ovarian cancer.

Prediction of response to platinum-based chemotherapy based on gene expression array data has been focused on for many years since its importance and significance [[2-4](#_ENREF_2)]. However, none of these prediction model have been successfully used in clinical utilize which maybe caused the insufficiency information usage of the microarray technology in which splicing and isoform informations are mixed and can’t be recognized. And interesting, In Liu’s article, the expression of 227 genes[[4](#_ENREF_4)] from 232 TCGA ovarian cancer patients were selected to train the prediction model but actually there is just one gene which is statistically differentially expressed after the multiple comparison adjustment, such as bonferroni or fdr, between chemo-resistant and chemo-sensitive patients. So we hypothesis that the dramatic useful information which can be used to distinguish chemo-resistant/sensitive is not the level of the gene expression but the splice and isoform information which have been ignored in the long history of gene expression analysis.

**Result**

OvCa subtype analysis base on RNA-seq dataset with Functional data analysis

OvCa subtype analysis base on DNA methylation dataset with KNN

OvCa subtype analysis base on gene expression microarray dataset with KNN

**Material and Method**

412 OvCa patients in the TCGA database with raw Bam dataset [[5](#_ENREF_5)] are obtained to generate the expression curve profile, among which 163 patients are sensitive to chemotherapy, and 70 are chemoresistant. Platinum status is defined as resistant if the patient recurred within six months or else sensitive if the platinum free interval is six months or greater, there is no evidence of progression or recurrence, and the follow-up interval is at least six months from the date of last primary platinum treatment defined by TCGA committee[[5](#_ENREF_5)]. [table, clinical information]. Then we get gene expression[233], methylation[233] dataset of these 233 patients.

412 OvCa patients in the TCGA database with raw Bam dataset [[5](#_ENREF_5)] are obtained to generate the expression curve profile, among which 163 patients are sensitive to chemotherapy, and 70 are chemoresistant. Platinum status is defined as resistant if the patient recurred within six months or else sensitive if the platinum free interval is six months or greater, there is no evidence of progression or recurrence, and the follow-up interval is at least six months from the date of last primary platinum treatment defined by TCGA committee[[5](#_ENREF_5)]. [table, clinical information]. Then we get gene expression[233], methylation[233] dataset of these 233 patients.

Minimum non-overlapping transcript units (MNOTU) was defined as genome regions among which there is no overlap transcripts. We download exon data from latest UCSC GENECODE V14[[6](#_ENREF_6), [7](#_ENREF_7)], RefGene[[8-10](#_ENREF_8)] and merge overlapping region and then 27,768 units for encode V14 and 282,027 units for refGene. statistic the number of the genes in this region. Note, sort bed file first then merge the bed file with the above principle.

**Alignment**

Illumina paired-end RNA sequencing reads were aligned to GRCh37-lite genome-plus-junctions reference (chromosomes 1-22, MT, X and Y) using BWA version 0.5.7. This reference combined genomic sequences in the GRCh37-lite assembly and exon-exon junction sequences whose corresponding coordinates were defined based on annotations of any transcripts in Ensembl (v59), Refseq and known genes from the UCSC genome browser, which was downloaded on August 19 2010, August 8 2010, and August 19 2010, respectively. Reads that mapped to junction regions were then repositioned back to the genome, and were marked with 'ZJ:Z' tags. BWA is run using default parameters, except that the option (-s) is included to disable Smith-Waterman alignment. Finally, reads failing the Illumina chastity filter are flagged with a custom script, and duplicated reads were flagged with Picard's MarkDuplicates.

**cDNA Library Preparation**

2-3 ug Total RNA samples were arrayed into a 96-well plate and PolyA+ RNA was purified using the 96-well MultiMACS mRNA isolation kit on the MultiMACS 96 separator (Miltenyi Biotec, Germany) with on column DNaseI-treatment as per the manufacturer's instructions. The eluted PolyA+ RNA was ethanol precipitated and resuspended in 10uL of DEPC treated water with 1:20 SuperaseIN (Life Technologies, USA). Double-stranded cDNA was synthesized from the purified polyA+RNA using the Superscript Double-Stranded cDNA Synthesis kit (Life Technologies, USA) and random hexamer primers at a concentration of 5uM. The cDNA was quantified in a 96-well format using PicoGreen (Life Technologies, USA) and VICTOR3V Spectrophotometer (PerkinElmer, Inc. USA). The quality was checked on a random sampling on the NULL using the High Sensitivity DNA chip Assay. cDNA was fragmented by Covaris E210 (Covaris, USA) for 55 seconds, a "Duty cycle" of 20% and "Intensity" of 5. Plate-based libraries were prepared following the BC Cancer Agency, Genome Sciences Centre paired-end (PE) protocol on a Biomek FX robot (Beckman-Coulter, USA). Briefly, the cDNA was purified in 96-well format using Ampure XP SPRI beads, and was subject to end-repair and phosphorylation by T4 DNA polymerase, Klenow DNA Polymerase, and T4 polynucleotide kinase respectively in a single reaction, followed by cleanup using Ampure XP SPRI beads and 3' A-tailling by Klenow fragment (3' to 5' exo minus). After cleanup using Ampure XP SPRI beads, picogreen quantification was performed to determine the amount of Illumina PE adapters used in the next step of adapter ligation reaction. The adapter-ligated products were purified using Ampure XP SPRI beads, then PCR-amplified with Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) using Illumina PE primer set, with cycle condition 98 degrees C 30sec followed by 10-15 cycles of 98 degrees C 10 sec, 65 degrees C 30 sec and 72 degrees C 30 sec, and then 72 degrees C 5 min. The PCR products were purified using Ampure XP SPRI beads, and checked with Caliper LabChip GX for DNA samples using the High Sensitivity Assay (PerkinElmer, Inc. USA). PCR product of desired size range was purified using an in-house 96-channel size selection robot, and the DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay and Quant-iT dsDNA HS Assay Kit using Qubit fluorometer (Invitrogen), then diluted to 8nM. The final concentration was verified by Quant-iT dsDNA HS Assay prior to Illumina HiSeq 2000 PE 75 base sequencing.

Mapping quality threshold (mirror Phred scores=10)

Most tools report a quality score for the mapping of a read to the reference genome. These scores mirror Phred scores (Ewing and Green, 1998); they represent the log-scaled probability that the mapping is incorrect. Meaningful scores typically range from 0 to 60 , where 0 corresponds to very low quality mapping and scores of 30 are considered to be very good. A score of 30 denotes a 10−3 chance that the mapping is incorrect; as the score increases to 40, the chance of an incorrect mapping theoretically drops to 10−4. As Ruffalo’s estimation, for majority mapping tools, such as bwa, Novoalign, SOAP, most incorrect mappings can be discarded simply by considering mapping qualities of at least 1[[11](#_ENREF_11)].

1. Agarwal, R. and S.B. Kaye, *Ovarian cancer: strategies for overcoming resistance to chemotherapy.* Nat Rev Cancer, 2003. **3**(7): p. 502-16.

2. Hartmann, L.C., et al., *Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy.* Clin Cancer Res, 2005. **11**(6): p. 2149-55.

3. Selvanayagam, Z.E., et al., *Prediction of chemotherapeutic response in ovarian cancer with DNA microarray expression profiling.* Cancer Genet Cytogenet, 2004. **154**(1): p. 63-6.

4. Liu, Y., et al., *Integrated analysis of gene expression and tumor nuclear image profiles associated with chemotherapy response in serous ovarian carcinoma.* PLoS One, 2012. **7**(5): p. e36383.

5. *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.