OTKA PD104604

Molecular oncogenesis in human chordoma: targeting early diagnosis and effective therapy

Final report

30 Sep 2015

Summary

In this research project, our hypothesis was that by performing complex transcriptional profiling studies on human chordoma we will be able to detect the key elements of the oncogenesis and the possible molecular targets of a new, systemic therapy. To explore the molecular mechanism of oncogenesis in chordoma, we planned to perform a comparative transcriptomics study, where the differences in the transcriptome profile of the human notochord derivates (chordoma and nucleus pulposus) would indicate the genes of interest in the pathogenesis of chordoma development. In case of such a complex transcriptomics research, the most important issue is the quality of the molecular material, actually the isolated RNA. The notochordal derivates (chordoma and nucleus pulposus (NP)) are very rich in mucinous extracellular matrix making their processing difficult. Moreover in case of the NP tissue, the cell density is very low. There is not any paper published in the international literature containing a precisely described and validated RNA isolation method from chordoma or NP so we were not able to simply apply a well working methodology or the previously published methods failed the quality control. On the other hand, the recent technological development resulted in the availability of advanced molecular methods like as Next Generation Sequencing (NGS). This technique gives the possibility for the deeper analysis of the transcriptom (mRNA and miRNA expression) than the conventional array techniques. In the recent years, NGS platforms have become available and cost-effective in Hungary, so it has been reasonable for us to prepare such a final transcriptom sequencing step for the 3rd year of the project, when both mRNA and miRNA expression of all the samples will be sequenced on an Illumina HiSeq 2000 platform at the Clinical Genomics Center in Debrecen. These changes in the final methodology will provide a more detailed dataset about the transcriptom of the human chordoma however the quality and amount of the input RNA in case of NGS is crucial without the possibility of any compromise. That is why, we had to develop a new method for the RNA isolation from chordoma tissues what was successful even if it meant a significant delay in the preparation period. Our final sample set contained high quality RNA isolates from 8-8 individual chordoma and NP samples (we exceeded the planned sample number with 33.3% to improve the reliability of the comparative transcriptomic studies).

After the meticulous preparation work, a successful mRNA NGS project has been performed. The quality control showed very good quality of the sequencing project and the first level of bioinformatic analysis confirmed most of the previously published gene expression findings in chordoma and revealed a number of previously not know gene and gene clusters having a possible role in the oncogenesis. To our knowledge this is the first successful NGS mRNA project in chordoma and among the firsts in human tumors.

Despite the fact that NGS results regularly show good correlation with RT-PCR studies, we have decided to do a validation study on the final set of gene-of-interests to improve the level of evidence originated from our research. We have prepared for the RT-PCR studies but it will be implemented together with the miRNA data validation measurements which have been delayed because of an unanticipated technical error. The NGS miRNA project has been also implemented but the post-run quality control showed an unreliable read in a certain position what was likely caused by an air microbubble in the system. It only affected one position in the reads, but we decided to reply the whole measurement because it can significantly reduce the reliability of the whole sequencing run. This repeated run could not be done by the end of the official project period, but it is fully prepared and covered. Analysis of the miRNA data and validation of the whole transcriptomic work can be done after the repeated miRNA NGS run. The whole planned project will be technically finished in three months, but this delay does not mean any significant failure, because everything is prepared for the measurements and the human power is also available to do that. Beyond the technical results of this OTKA grant, a new scientific consortium has been established. A productive partnership in chordoma research has been proved between the National Center for Spinal Disorders, the Debrecen Clinical Genomics Center and the 1st Department of Pathology and Experimental Cancer Research at the Semmelweis University. The members of the consortium are fully engaged to continuing the chordoma molecular research and have already sought for new national and international funding possibilities to cover the next – in vitro – validation phase of the possible molecular targets. The present OTKA PD grant have given the possibility for the collection of the valuable raw data and for starting its analysis, however, with the completion of the project an enormous database with transcriptomic data is developed. During the grant period, journal papers and conference presentations have been published related to the research. Further scientific papers and one PhD thesis are under preparation based on the molecular results.

Main results of the project

1. Development of appropriate technique for the sample collection and high-quality RNA isolation from human chordoma surgical samples

2. Establishing of two new chordoma cell-lines

3. Genotyping and immunohistochemistry study on brachyury candidate SNP

4. Successful application of a next generation sequencing technique onto human chordoma and nucleus pulposus RNA

5. Successful application of multi-level bioinformatic data-analsyis methods onto RNA sequencing data confirming the previously published results on altered mRNA expression in chordoma

Original goals not accomplished within the project timeline

1. Next generation sequencing of isolated miRNA samples has to be repeated because of an unanticipated technical error

2. RT-PCR validation of NGS data

The reasons of the above mentioned deviations from the original project plan are the difficulties arisen during the RNA isolation from human chordoma and nucleus pulposus samples and a technical error during the next generation sequencing (NGS) run. The new NGS run is planned to perform in October 2015 and just after that all the required validation steps will be done. The technical, human and financial background of this research work is ensured.

Detailed technical report

1. RNA isolation, quantification and quality control

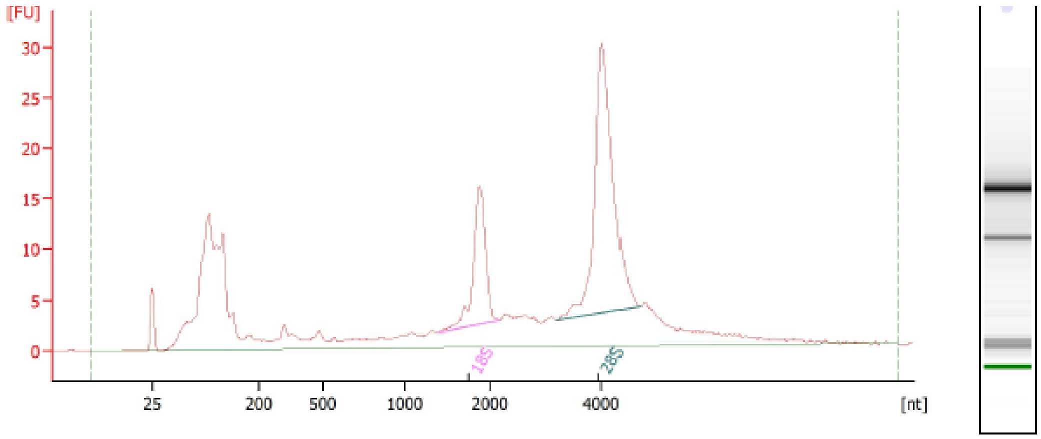
LN_2 snap frozen chordoma tissue sample were ground to a fine powder while frozen, the tissue powder was added to TRIzolate Reagent in a 50 ml centrifuge tube (50 mg sample per 500 uL reagent), and the samples were homogenized with a rotor-stator homogenizer. Equal volume chloroform was added to the homogenizates and vortexed, then the samples were loaded into Phase Lock Gel™ (PLG 15 ml Heavy) tubes. The organic and aqueous phases were thoroughly mixed without vortexing to form a transiently homogeneous suspension. Samples were centrifuged for 5 minutes at 1500 x g to separate the phases. The PLG formed a barrier between the aqueous and organic phases. The upper aqueous phase was transferred to a fresh tube, and RNA was isolated from the samples with the Direct-Zol RNA Miniprep kit, according to the manufacturer's instructions, without DNase I treatment. Cell lysates and the upper aqueous phases containing RNA were prepared from the nucleus pulposus cell lines with TRIzolate Reagent, according to the manufacturer's instructions. RNA was isolated from the aqueous phases with the Direct-Zol RNA Miniprep kit, according to the manufacturer's instructions, without DNase I treatment. The developed multi-step RNA extraciton method was not suitable for surgical nucleus pulposus samples because of its enourmus low cell/extracellular matrix ratio (any publication has not been reported using a suitable method for high-quality RNA isolation from surgical nucleus pulposus tissue yet). Here, we had to use another model and high-quality RNA was isolated from in vitro monolayer nucleus pulposus cell cultures. All cell cultures were started from individual surgical samples (Grade III discs from patients in the age of 35-45y old) without any enzymatic digestions only by the explantation of the cut tissue. Total RNA from the monolayer cultures was isolated using commercial RNA isolation kits.

The NanoDrop spectrophotometer was used for quality control and quantification, and the Agilent 2100 platform was used for checking the integrity of the RNA samples. Table 1. summarizes the characteristics of the chordoma and NP RNA samples. Based on these analyses, all RNA samples were of high quality, and suitable for next generation sequencing.

Table 1. Quality control and quantification of the 16 RNA samples selected for the next generation sequencing

Sample ID	Sample origin	ng/uL	260/280	260/230	Agilent RIN
CH4	chordoma tumor. surgical specimen	223.19	2.08	2.32	8.2
CH6		468.01	1.95	1.95	8.2
CH7		462.8	1.94	1.27	8.3
CH8		3262.68	1.93	1.96	8.8
CH10		578.77	2.03	1.43	8.6
CH11		1357.92	2	1.82	7.3
CH12		938.61	2.03	1.78	8.0
CH13		1045.25	2.06	1.54	7.9
NP5	Nucleus pulposus cell culture	233.58	2.07	2	8.1
NP6		218.29	2.07	1.99	9.2
NP7		311.51	2.06	1.91	9.1
NP9		900.41	2.15	2.12	10
NP10		832.77	2.15	2.11	10
NP11		599.07	2.13	1.36	10
NP12		209.48	2.09	2.07	7.9
NP13		687.45	2.13	2.06	9.6

Figure 1. Agilent Bioanalyzer 2100 run of an RNA sample from chordoma tissue (RIN: 8.6)



2. New chordoma cell-lines

Surgical chordoma samples have been explanted and cultured in vitro. In two cases a long-term cultivation and proliferation of the cells have been observed. In the maintenance of the cell-lines we have followed the instructions available in the literature. Based on the morphology and the brachyury positivity, the cells are chordoma cells. The two cell-lines are more than 10 months old after the 50 doubling periods. The international registration at the Chordoma Foundation of the new chordoma cell-lines are under process. These cell–lines will be very important in our future in vitro studies two when potential molecular targets will be in vitro studied.

3. The brachyury SNP and immunohistochemistry study

During the sample processing period, we performed a substudy on FFPE chordoma samples. The brachyury gene has been recently published as one of the possible key elements of the oncogenesis in chordoma and a SNP (rs2305089) was identified in association with the tumor. In a multinational consortium, we had genotyped this SNP in 109 tumors (largest published dataset so far) and found that the GG genotype of the SNP has been significantly associated with poorer survival in chordoma patients (Figure 2). Our center did a brachyury immunohistology study on our sample-set (48 samples, Figure 3) but not found any association between the rs2305089 genotype and the brachyury expression. The molecular effect of the polymorphism can be related to the altered brachyury function what should be further investigated. The paper about results of this study has been submitted to the Clinical Cancer Research.

Figure 2. Association of rs2305089 genotype with survival of chordoma patients (p<0.005)

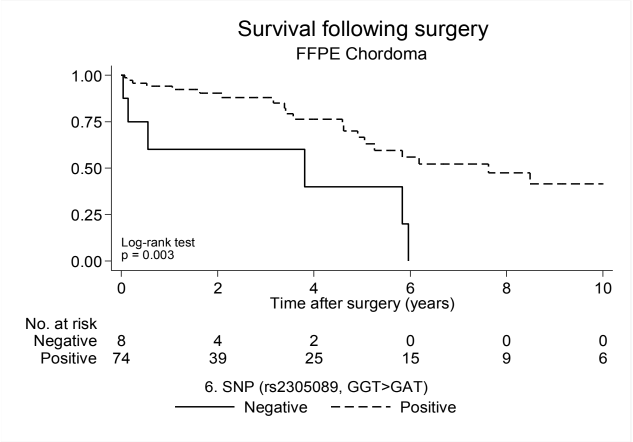
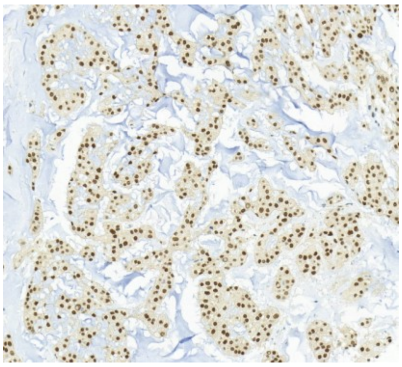


Figure 3. Brachyury positivity in chordoma cells



4. Next generation sequencing of the RNA samples

Sequencing libraries were prepared with the Illumina TruSeq RNA SamplePrep v2 kit, according to the manufacturer's instructions. All the RNA libraries were indexed, allowing multiplexed sequencing. Libraries were sequenced as 2x100 bp paired-end reads on the Illumina HiSeq 2000. Raw data was preprocessed with the CASAVA 1.8.2 software (converts*.bcl files into*.fastq.gz files=compressed FASTQ files), and resulted in the demultiplexed, fastq.gz files used in all subsequent analyses. Sequencing for one sample, CH4 (tumor) was unsuccessful, likely due to an index sequence mixup or mislabeling during library preparation, resulting in a missing file after demultiplexing. Sequencing of this sample will be repeated at a later date.

5. Bioinformatic analysis of RNA next generation sequencing data – RNA-seq

5.1. Quality control

For all bioinformatic analyses the public Galaxy server was used: https://usegalaxy.org/. First, all fastq.gz files containing the sequence data for the 15 samples were uploaded to Galaxy, then subjected to quality control analysis using the FASTQC program.

On the QC analysis, all RNA-seq sequences were of high quality, as evidenced e.g. by the „per_base_quality" analysis, showing that all bases of the fragments had higher than 20 quality scores. Other analysis metrics were also consistent with high quality sequences derived from RNA-seq. Results for sample CH6 (forward read) is shown below in Fig. 4A-K. as an example - all other sequence files had similar characteristics. Due to the high quality of the sequences, no trimming of the fragments were necessary, and the sequences were subjected to further analytical steps without any manipulation.

Figure 4. Graphical representation of the results of the FASTQC analysis for sample CH6, RNA-seq, forward reads.

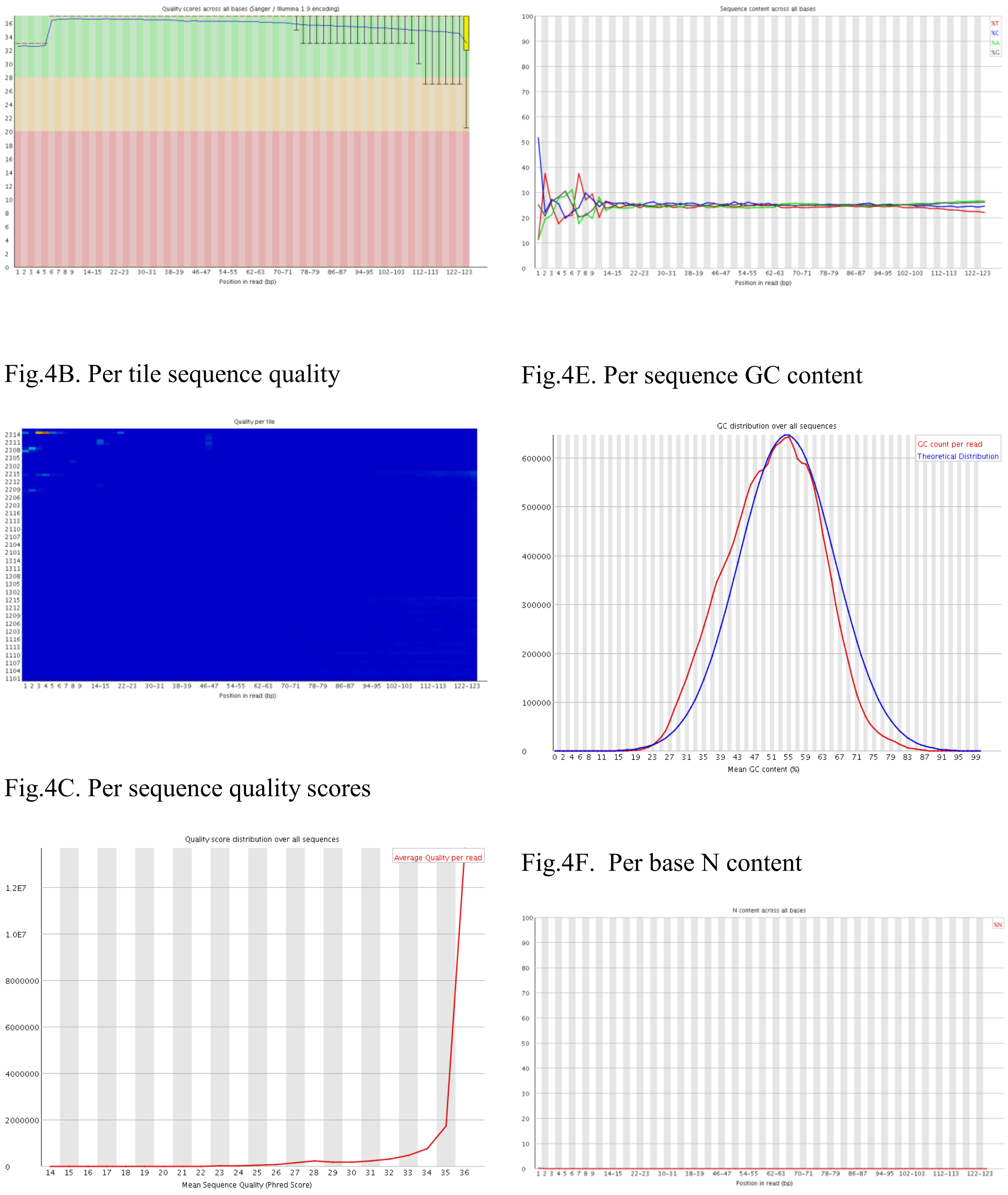


Fig.4A: Per base sequence quality

Fig.4B. Per tile sequence quality

Fig.4C. Per sequence quality scores

Fig.4D. Per base sequence content

Fig.4E. Per sequence GC content

Fig.4F. Per base N content

Fig.4G. Sequence Length Distribution

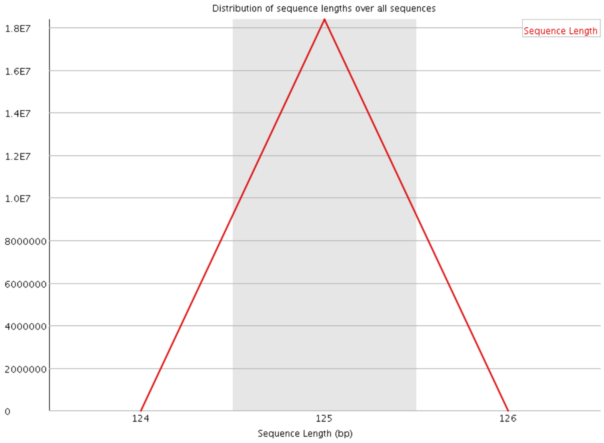


Fig.4H. Sequence Duplication Levels

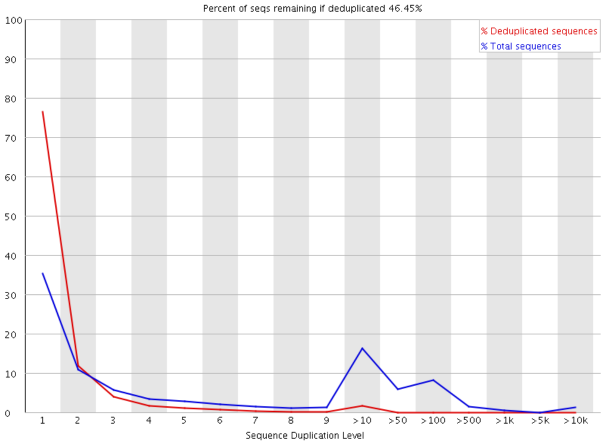
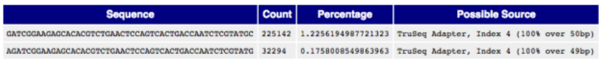


Fig.4I. Overrepresented sequences. Index4 was used for labelling this sample during library preparation.



Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGC	225142	1.2256194987721323	TruSeq Adapter, Index 4 (100% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATG	32294	0.1758008549863963	TruSeq Adapter, Index 4 (100% over 49bp)

Fig. 4J. Adapter Content

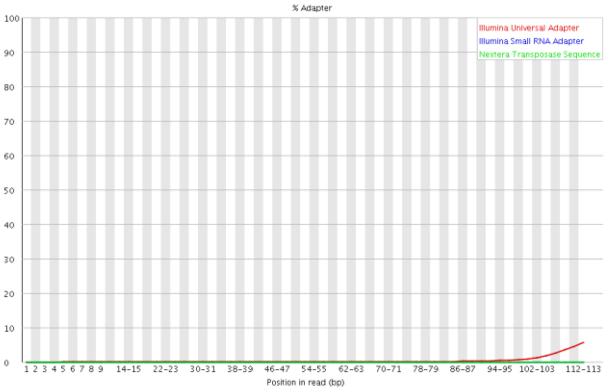
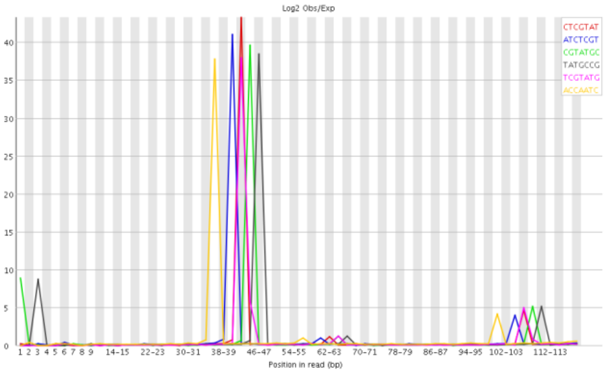


Fig.4K. Kmer Content



On the other hand, based on the QC analysis, there was a technical error in the miRNA-seq at the 24th cycle, and the base could not be read – hence, a sharp drop in the „per_base_quality" scores, and high „per base N content values at the 24th basepair (Fig. 2A-B). Since this position is within the miRNA sequence, miRNAs could not be identified reliably from the data. All miRNA-seq runs will be repeated at a later date.

Figure 5. Graphical representation of the results of the FASTQC analysis for sample CH4, miRNA-seq.

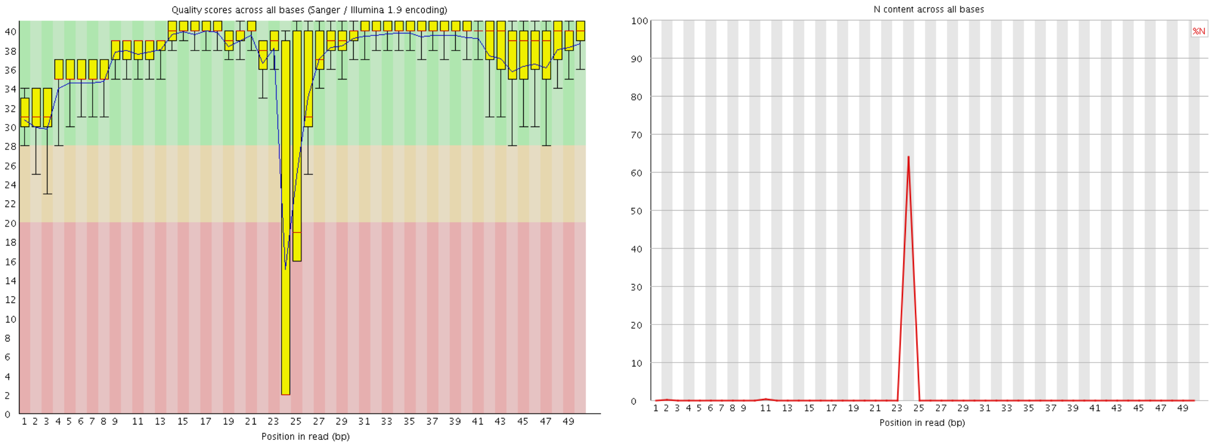


Fig.5A: Per base sequence quality

Fig. 5B. Per base N content

5.2. The bioinformatic analysis pipeline.

5.2.1. Steps of the pipeline and version numbers of the software tools are listed below in Table 2.

Table 2.

Step	Function in Galaxy	Galaxy tool version Tool version	File format (input)	Other info	Task performed
1.	**Input dataset** hs_CH-6_1_sequence.txt	1.1.4.	fastq.gz	Forward sequence	Uploaded sequence files into Galaxy server
2.	**Input dataset** hs_CH-6_2_sequence.txt	1.1.4.	fastq.gz	Reverse sequence	Uploaded sequence files into Galaxy server
3.	**Input dataset** For other tumor samples	1.1.4.	fastq.gz files	Forward and reverse sequences	Uploaded sequence files into Galaxy server
4.	**Input dataset** For all normal samples	1.1.4.	fastq.gz files	Forward and reverse sequences	Uploaded sequence files into Galaxy server
5.	**FASTQC**	0.63	fastqsanger	On all sequence files uploaded in Steps 1-4.	Quality control analysis of sequencing
6.	**TopHat2 (no fusion)** hs_CH-6_1_sequence.txt hs_CH-6_2_sequence.txt	0.9 TopHat v2.0.14	fastqsanger fastqsanger	Forward sequence Reverse sequence	Alignment of paired reads per sample to the human genomic sequence
7.	**TopHat2 (no fusion)** Performed the same as in Step 6. for all samples	TopHat v2.0.14	fastqsanger		Alignment of paired reads per sample to the human genomic sequence
8.	**Cufflinks** Accepted hits TopHat file for CH6	2.2.1.0 cufflinks v2.2.1	bam		Assembles transcripts, guided by reference genome annotation
9.	**Cufflinks** Performed the same as in Step 8. for all samples	2.2.1.0 cufflinks v2.2.1	bam		Assembles transcripts, guided by reference genome annotation
10.	**Cuffcompare** Assembled transcript files from Cufflinks for all tumor samples Assembled transcript files from Cufflinks for all normal samples	2.2.1.0 cuffcompare v2.2.1 (4237)	gtf gtf		Prepares data for CuffDiff analysis
11.	**CuffDiff** Combined transcript files from Cuffcompare			geometric normalization	Differential gene expression between tumor and normal

5.2.3. Parameter settings for TopHat2, Cufflinks, Cuffcompare and Cuffdiff analyses.

Parameters are listed below in Tables 3-6. Settings for TopHat2 and Cufflinks are shown for specific samples, but the same settings were used for all samples.

Table 3. Parameter settings for TopHat2 analysis

Input Parameter	**Value**
Is this single-end or paired-end data?	paired
RNA-Seq FASTQ file, forward reads	2: hs_CH-6_1_sequence.txt
RNA-Seq FASTQ file, reverse reads	3: hs_CH-6_2_sequence.txt
Mean Inner Distance between Mate Pairs	20
Std. Dev for Distance between Mate Pairs	20
Report discordant pair alignments?	Yes
Use a built in reference genome or own from your history	indexed
Select a reference genome	hg19
TopHat settings to use	full
Max realign edit distance	0
Max edit distance	2
Library Type	FR Unstranded
Final read mismatches	2
Use bowtie -n mode	No
Anchor length (at least 3)	8
Maximum number of mismatches that can appear in the anchor region of spliced alignment	0
The minimum intron length	70
The maximum intron length	500000
Allow indel search	Yes
Max insertion length.	3
Max deletion length.	3
Maximum number of alignments to be allowed	20
Minimum intron length that may be found during split-segment (default) search	50
Maximum intron length that may be found during split-segment (default) search	500000
Number of mismatches allowed in each segment alignment for reads mapped independently	2
Minimum length of read segments	25
Output unmapped reads	False
Do you want to supply your own junction data	Yes
Use Gene Annotation Model	Yes
Gene Model Annotations	168: UCSC Main on Human: knownGene (genome)
Use Raw Junctions	No
Only look for supplied junctions	No
Use Coverage Search	No
Use Microexon Search	No
Do Fusion Search	No
Set Bowtie2 settings	No
Specify read group?	no
Job Resource Parameters	no

Table 4. Parameter settings for Cufflinks

Input Parameter	**Value**
SAM or BAM file of aligned RNA-Seq reads	183: Tophat on data 168, data 7, and data 6: accepted_hits
Max Intron Length	300000
Min Isoform Fraction	0,1
Pre MRNA Fraction	0,15
Use Reference Annotation	Use reference annotation
Reference Annotation	168: UCSC Main on Human: knownGene (genome)
Count hits compatible with reference RNAs only	Yes
Perform Bias Correction	Yes
Reference sequence data	cached
Using reference genome	hg19
Use multi-read correct	Yes
Apply length correction	Cufflinks Effective Length Correction
Global model (for use in Trackster)	No dataset
Set advanced Cufflinks options	No
Job Resource Parameters	no

Table 5. Parameter settings for Cuffcompare

Input Parameter	**Value**
GTF file(s) produced by Cufflinks	340: Cufflinks chord Normal
GTF file(s) produced by Cufflinks	338: Cufflinks chord tumor
Use Reference Annotation	Yes
Reference Annotation	168: UCSC Main on Human: knownGene (genome)
Ignore reference transcripts that are not overlapped by any input transfrags	False
Ignore input transcripts that are not overlapped by any reference transcripts	False
Use Sequence Data	No
discard (ignore) single-exon transcripts	No
Max. Distance for assessing exon accuracy	100
Max.Distance for transcript grouping	100
discard intron-redundant transfrags sharing 5'	FALSE

Table 6. Parameter settings for Cuffdiff

Input Parameter	**Value**	**Sample**
Transcripts	363: Cuffcompare on data 168, data 252, and others: combined transcripts	
Omit Tabular Datasets	False	
Generate SQLite	True	
Input data type	BAM	
Name	Normal	
Replicates	243: Tophat on data 168, data 51, and data 50: accepted_hits, 238: Tophat on data 168, data 49, and data 48: accepted_hits, 233: Tophat on data 168, data 47, and data 46: accepted_hits, 228: Tophat on data 168, data 45, and data 44: accepted_hits, 223: Tophat on data 168, data 43, and data 42: accepted_hits, 218: Tophat on data 168, data 41, and data 40: accepted_hits, 213: Tophat on data 168, data 39, and data 38: accepted_hits, 208: Tophat on data 168, data 37, and data 36: accepted_hits	NP13 NP12 NP11 NP10 NP9 NP7 NP6 NP5
Name	Tumor	
Replicates	332: Tophat on data 168, data 35, and data 34: accepted_hits, 198: Tophat on data 168, data 33, and data 32: accepted_hits, 193: Tophat on data 168, data 11, and data 10: accepted_hits, 188: Tophat on data 168, data 9, and data 8: accepted_hits, 183: Tophat on data 168, data 7, and data 6: accepted_hits, 178: Tophat on data 168, data 5, and data 4: accepted_hits, 173: TopHat_Ch6_accepted_hits.bam	CH13 CH12 CH11 CH10 CH8 CH7 CH6
Library normalization method	geometric	
Dispersion estimation method	pooled	
False Discovery Rate	0.05	
Min Alignment Count	10	
Use multi-read correct	True	
Perform Bias Correction	Yes	
Reference sequence data	cached	
Using reference genome	hg19	
Include Read Group Datasets	No	
Include Count Based output files	No	
apply length correction	cufflinks effective length correction	
Set Additional Parameters for single end reads? (not recommended for paired-end reads)	No	
Set Advanced Cuffdiff parameters?	No	
Job Resource Parameters	no	

5.2.4. Summary statistics of TopHat2 alignments.

Statistical analysis of TopHat2 alignments of paired read files demonstrated that the sequence files contained on average 18330864 accepted reads, with 77% overall read mapping rate, and on average 13325254 aligned pairs, with 72.7% concordant pair alignment rate (Tables 7A-B). The number of accepted reads per sample show that the coverage of the transcriptome for the tumor samples is from 24 to 26 fold, whereas for the normal samples it is from 20 to 27 fold, as intended. The lower coverage for some of the normal samples (NP7, NP11 and NP13) is acceptable, since we expect less gene expression variation from these samples, and very low rate of gene mutations or chromosomal abnormalities. The mapping rates for left and right reads are all above the required 60% - the slightly lower mapping rates of the right (reverse) reads is as expected, due to the usually lower sequence quality of the reverse reads. The concordant pair alignment rates are also above the required 60%.

Table 7A. Read statistics for forward (left) and reverse (right) sequences

		Left reads		Right reads	
Sample ID	Tissue type	All accepted reads	Mapped reads (% of input)	All accepted reads	Mapped reads (% of input)
CH4	tumor	sequencing was unsuccessful			
CH6	tumor	18369649	14849045⍰(80.8%)	18369649	14021174⍰(76.3%)
CH7	tumor	18284510	14765809⍰(80.8%)	18284510	14008803⍰(76.6%
CH8	tumor	18238291	14136624⍰(77.5%⍰)	18238291	⍰13421589⍰(73.6%
CH10	tumor	18797429	15037662⍰(80.0%)	18797429	14274028⍰(75.9%)
CH11	tumor	19349383	14411471⍰(74.5%⍰)	19349383	13664557⍰(70.6%
CH12	tumor	19855625	15353196⍰(77.3%)	19855625	14563183⍰(73.3%⍰
CH13	tumor	18048073	⍰13578229⍰(75.2%)	18048073	12884589⍰(71.4%
NP5	normal	18567461	14739372⍰(79.4%)	18567461	13925793⍰(75.0%⍰
NP6	normal	19416597	15762699⍰(81.2%)	19416597	14891436⍰(76.7%⍰
NP7	normal	16709134	13206963⍰(79.0%)	16709134	⍰12500064⍰(74.8%
NP9	normal	19065870	15363582⍰(80.6%⍰)	19065870	14395381⍰(75.5%⍰
NP10	normal	20370354	16418370⍰(80.6%)	20370354	15411488⍰(75.7%
NP11	normal	16482810	12674606⍰(76.9%)	16482810	11880093⍰(72.1%⍰
NP12	normal	18152445	14539912⍰(80.1%)	18152445	13628977⍰(75.1%⍰
NP13	normal	15255331	⍰12866219⍰(84.3%⍰)	15255331	12034558⍰(78.9%

Table 7B. Read statistics for paired reads

Sample ID	Tissue type	Overall read mapping rate	Aligned pairs	Concordant pair alignment rate
CH4	tumor	-	-	-
CH6	tumor	78.6%	13646400	74.2%
CH7	tumor	78.7%	13627713	74.4%
CH8	tumor	75.6%	13004960	71.2%
CH10	tumor	78%	13872992	73.7%
CH11	tumor	72.6%	13253182	68.4%
CH12	tumor	75.3%	14160732	71.2%
CH13	tumor	73.3%	12529207	69.4%
NP5	normal	77.2%	13560750	73%
NP6	normal	78.9%	14496195	74.6%
NP7	normal	76.9%	12170942	72.8%
NP9	normal	78%	13995591	73.4%
NP10	normal	78.1%	14975557	73.5%
NP11	normal	74.5%	11551344	70.1%
NP12	normal	77.6%	13280481	73.1%
NP13	normal	81.6%	11752767	77%

5.3. Results of the comparative gene expression analysis

Initial differential gene expression analysis was done by Cuffdiff. It should be noted that this is a preliminary analysis, since Cuffdiff is known for the high false positive rate. Differential gene expression analysis using the DeSeq2 program has very high specificity, but lower sensitivity, whereas the EdgeR program has good specificity and sensitivity as well. Therefore, pathway analysis and target selection for gene expression validation will be based mostly on the EdgeR analysis, although overlapping gene sets from two or three analyses will also be considered. These analyses are ongoing. Differential splicing analysis will be done based on the CuffDiff analysis.

5.3.1. Expression of known chordoma marker genes

Most previously published genes with altered expression in chordoma have been confirmed by our study.

5.3.1.1. Scheil-Bertram S et al. Molecular profiling of chordoma. Int J Oncol. 2014 Apr;44(4):1041-55.

„T brachyury (mouse) homolog (T), CD24 antigen (CD24), insulin-like growth factor binding protein 2 (IGFBP2), retinoic acid receptor responder 2 (RARRES2), esophageal cancer-related gene 4 protein (ECRG4) and keratin 18 (KRT18)] with increased expression and one gene (T1A-2 lung type-I cell membrane-associated glycoprotein T1A2) with reduced expression compared to control and chondrosarcoma.“

Overexpression was confirmed for CD24, KRT18, ECGR4 (c2orf40), and suppression was confirmed for T1A2 (PDPN) in our dataset. Other keratin genes (KRT80, KRT7 and KRT8) were also overexpressed (Table 8). Brachyury and IGFBP2 overexpression was not confirmed, although both transcripts were present in chordoma and nucleus pulposus samples, as evidenced by visualization of Cufflinks transcript assembly in IGB viewer.

Table 8. Altered expression of keratin genes, ECGR4 and T1A-2 in the chordoma samples

Gene symbol	Locus	log2 fold change
KRT18	chr12:53290970-53346685	10,3016
KRT7	chr12:52626953-52642709	2,5419
KRT8	chr12:53290970-53346685	3,3858
KRT80	chr12:52562779-52585784	5,10391
ECGR4	chr2:106682112-106694609	6,72475
T1A-2	chr1:13910251-13944452	-5,18271

5.3.2. Expression of genes hypermethylated in chordoma

5.3.2.1. Alholle A et al. Genome-wide DNA methylation profiling of recurrent and non-recurrent chordomas. Epigenetics 2015, 10:213. FAM181B, KANK2, NPR3, PON3, RAB32, RAI1, SLC16A5 and ZNF397OS are hypermethylated in recurrent chordoma.

We confirmed suppressed expression of NPR3 and NPR2 – their neighboring genes also appeared to be suppressed (Table 9). The other genes, with the exception of RAB32 (overexpressed), were not among the CuffDiff differentially expressed gene set.

Table 9. Altered expression of NPR3 and NPR2 in the chordoma samples

Gene symbol	Locus	log2 fold change
ZFR	chr5:32354455-32444844	-4,53965
NPR3	chr5:32710742-32791830	-2,84018
TARS	chr5:33440801-33468196	-0,827534
MSMP	chr9:35697333-35754274	-7,95238
RGP1	chr9:35697333-35754274	-3,82363
NPR2	chr9:35792405-35812259	-2,04956

***5.3.2.2.** Rinner B et al. Chordoma characterization of significant changes of the DNA methylation pattern. PLoS One : HIC1, CTCFL, ACTB, RASSF1, CDX1, GBP2, IRF4, DLEC1, COL21A1, GNAS, KL, C3, SRGN, S100A9 were hypermethylated in the chordoma vs. blood comparison.*

We could confirm suppression of only COL21A1, which was in a larger genomic region containing mostly down-regulated genes, from 6p12.3-p11.2. RASSF1, C3 and SRGN were actually overexpressed, with RASSF1 and C3 being located in a genomic region containing highly upregulated genes. Such coordinated changes in the gene expression of larger genomic regions may result from regulatory changes, or, more likely, copy number alterations and chromosomal rearrangements.

5.3.3. Regional gene suppression in the chordoma genome: potential genomic regions with deletions or hypermethylation

The usual approach to identify focal amplifications and deletions starts with global analyses on genomic DNA template, but the putative copy number changes should correlate with gene expression changes for the affected genes in that region. Similarly, methylation changes are first detected by specialized global approaches, then validated through gene expression studies. Our primary interest in these regions is that they frequently contain important driver genes for oncogenesis. Although our analysis is far from complete, the NGS-based gene expression data shows several examples of genomic regions where most or all genes are similarly up- or down-regulated – further experiments are needed to validate the molecular mechanisms behind these changes.

One example is the protocadherin gamma gene cluster on chromosome 5: simultaneous suppression of 10 protocadherin gamma transcripts is in line with hypermethylation of this cluster, which has been observed for the alpha and beta gene clusters as well, in different cell types.

Table 10. Suppression of the protocadherin gamma gene cluster

Gene ID	log2 fold change
PCDHGA10	-4,72996
PCDHGA11	-3,91882
PCDHGA6	-3,50589
PCDHGA7	-2,99119
PCDHGA8	-2,95992
PCDHGB2	-3,31718
PCDHGB5	-4,3468
PCDHGB6	-3,28179
PCDHGB7	-5,33641
PCDHGC3	-1,97042

Genes in the 3q26.32 genomic region are also suppressed, including PIK3CA – this region is frequently deleted in chordomas.

Table 11. Putative microdeletion and the affected genes in the 3q26.32 chromosomal region

Gene ID	Locus	log2 fold change
NCEH1	chr3:172348434-172429008	-3,93385
ECT2	chr3:172468474-172539264	-6,58931
NAALADL2	chr3:174577110-175523428	-4,75923
TBL1XR1	chr3:176738541-176915048	-1,33549
ZMAT3	chr3:178735010-178789656	-2,3725
PIK3CA	chr3:178866310-178952497	-2,1177

The 1p36-p11 region also harbors several deletions in chordoma – a similar signature in our dataset is in the 1p31.1 region, containing the adenylate kinase 5 (AK5) gene, and the USP33 gene. Interestingly, USP33 is required for SLIT/ROBO signalling, which inhibits cancer cell migration. Therefore, deletion of this region may contribute to the aggressive nature of chordomas. Two other adenylate kinases – AK3 (chr9) and AK9 (chr6) are also suppressed. Starting from the STRING protein interaction analysis for the interacting partners of AK5, we found altered expression of ectonucleoside triphosphate diphosphohydrolases (ENTPD4-5), cyclic nucleotide phosphodiesterases (9 PDE genes), cytosolic 5',3'-nucleotidases (4 NT5 genes) and nucleoside diphosphate kinases (NME2-4). These genes are involved in different steps of purine metabolism – the gene expression changes suggest significant alterations in this metabolic pathway.

Table 12. Putative microdeletion and the affected genes in the 1p31.1 chromosomal region

Gene ID	Locus	log2 fold change
PIGK	chr1:77554666-77685132	-2,56645
AK5	chr1:77747661-78025654	-5,39117
ZZZ3	chr1:78030189-78149104	-7,64646
USP33	chr1:78161673-78225564	-4,40011
FAM73A	chr1:78245308-78345225	-1,83815

The 10q26.11 region harbors microdeletions in certain developmental syndromes – the possible function of the genes in this region in oncogenesis is unclear, although FGFR2 overexpression is more common in different tumor types.

Table 13. Putative microdeletion and the affected genes in the 10q26.11 chromosomal region

Gene ID	Locus	log2 fold change
WDR11	chr10:122521323-122669038	-3,51142
FGFR2	chr10:123237843-123357972	-2,55595
ATE1	chr10:123502624-123688217	-4,89868
TACC2	chr10:123748688-124014057	-5,74034
PLEKHA1	chr10:124134093-124191871	-1,18076

5.3.4. Highly suppressed or overexpressed genes: functional analysis

Frequent suppression of certain genes in different cancer types is usually associated with their tumor suppressor function. There are several (significantly down-regulated) tumor suppressor genes in our dataset, including REST, APC, CDKN2B (p15) or PAX1. Downregulation of Brain-Derived Neurotrophic Factor (BDNF, log2FC= -7.47) and its receptor, NTRK2 (log2FC= -5.35) was also evident, suggesting that the BDNF pathway is suppressed in chordoma.

Table 14. Down-regulation of tumor suppressor genes in the chordoma samples

Gene ID	log2 fold change	Gene ID	log2 fold change
APC	-4,89177	GREM1	-8,37107
CACNA1A	-10,498	HIPK1	-7,25676
CCBE1	-8,99152	HOXB6	-6,29727
CDKN2B	-2,71106	LSAMP	-3,01414
DLG1	-6,18826	**PAX1**	-7,38545
FAM107A	-20,7002	**REST**	-10,7139
FOXN3	-8,2789		

Panther analysis identified several protein classes in both the overexpressed and down-regulated gene sets. The receptor protein is class is particularly interesting, since these proteins, or the pathways activated by them may offer druggable targets in chordoma.

An interesting overexpressed receptor in our dataset is „patched homolog 1“ (PTCH1), the receptor for hedgehog signalling – this is in accordance with previous immunohistochemistry studies in chordoma (*Cates JMM et al. The sonic hedgehog pathway in chordoid tumors, Histopathology 2010, DOI: 10.1111/j.1365-2559.2010.03572.x).* Both GLI1 and GLI4, downstream effectors of the hedgehog pathway are overexpressed in our dataset, although SHH1 and SHH2 are both suppressed. Despite GLI1 overexpression, known targets of this transcription factor appear to be mostly down-regulated in our dataset, suggesting that the sonic hedgehog pathway may not be active in these samples.

Table 15. Overexpressed receptor genes – Panther analysis

Gene ID	Gene name
ADORA3	Adenosine receptor A3
ASPN	Asporin
CD37	Leukocyte antigen CD37
CFD	Complement factor D
CHAD	Chondroadherin
CHRM2	Muscarinic acetylcholine receptor M2
CNR1	Cannabinoid receptor 1
COL10A1	Collagen alpha-1(X) chain
COL9A3	Collagen alpha-3(IX) chain
ESAM	Endothelial cell-selective adhesion molecule
FCGRT	IgG receptor FcRn large subunit p51
FMOD	Fibromodulin
HAVCR2	Hepatitis A virus cellular receptor 2
HPN	Serine protease hepsin
IL10RA	Interleukin-10 receptor subunit alpha
IL12RB1	Interleukin-12 receptor subunit beta-1
LILRB3	Leukocyte immunoglobulin-like receptor subfamily B member 3
LILRB5	Leukocyte immunoglobulin-like receptor subfamily B member 5
MSR1	Macrophage scavenger receptor types I and II
MYOT	Myotilin
OLFML2B	Olfactomedin-like protein 2B
PODN	Podocan
PTCH1	Protein patched homolog 1
PTGER3	Prostaglandin E2 receptor EP3 subtype
PTPRB	Receptor-type tyrosine-protein phosphatase beta
RORC	Nuclear receptor ROR-gamma
S1PR4	Sphingosine 1-phosphate receptor 4
SCTR	Secretin receptor
SFRP2	Secreted frizzled-related protein 2
TF	Tissue factor
TLR2	Toll-like receptor 2
TLR7	Toll-like receptor 7
TNFRSF18	Tumor necrosis factor receptor superfamily member 18
TPSB2	Tryptase beta-2
TRIL	TLR4 interactor with leucine rich repeats
TRPM2	Transient receptor potential cation channel subfamily M member 2
TSPAN33	Tetraspanin-33
VAC14	Protein VAC14 homolog

Table 16. Gene expression in the sonic hedgehog pathway

	Gene ID	log2 fold change
Ligands	SHH1	-5,41424
	SHH2	-1,0271
Receptors	PTCH1	9,37378
	PTCH2	4,01205
Transcription factors	GLI1	4,2857
	GLI4	2,51892
Target genes	CCND1	-1,97529
	CCND2	-1,14564
	ZEB1	-2,83738
	TWIST1	-1,48027
	TWIST2	-1,70815
	FOXC2	-1,67203
	ANGPT1	-7,34205
	ANGPT2	3,4723

Another overexpressed receptor is asporin (ASPN), which, together with the similarly overexpressed SPARC may activate the epithelial-mesenchymal transition (EMT) in chordoma cells. In addition, down-regulation of CMTM8 (a candidate tumor suppressor gene in osteosarcoma and a negative regulator of c-met) may also contribute to the EMT through the increased activity of the c-met pathway. C-met (MET) is also overexpressed in our chordoma samples, similar to previous immunohistochemical studies.

Table 17. Activation of the EMT process, and the c-met pathway

Gene ID	log2 fold change
MET	2,26657
CMTM8	-2,0434
ASPN	9,74441
SPARC	4,98074

Overexpression of two other gene sets may also reflect adaptive changes in the chordoma cells. Several troponin subunit genes are overexpressed, and based on the interacting partner analysis of STRING, the majority of their interacting proteins, such as myosin heavy chains and others, are also overexpressed. The significance of this is unclear, but maybe connected with cellular motility. On the other hand, overexpression of the metallothionein gene cluster (5 genes) on chromosome 16 may protect the chordoma cells against oxidative stress, and may contribute to their cisplatin resistance.

Table 18. Overexpression of the troponin complex and interacting proteins

	Gene ID	log2 fold change
Troponin subunits	TNNC1	7,51689
	TNNC2	10,867
	TNNI3	8,49085
	TNNT1	6,79757
	TNNT3	9,04854
	TNNT3	12,4446
Myosin heavy chains	MYH11	5,28683
	MYH14	11,7127
	MYH2	12,4988
	MYH3	4,38774
Other partners in the STRING network	TMOD4	8,19351
	TTN	5,43022
	DES	6,89626

Table 19. Overexpression of the metallothionein gene cluster

Gene ID	**Locus**	**log2 fold change**
MT1A	chr16:56642477-56673999	3,88898
MT1M	chr16:56642477-56673999	5,88449
MT1F	chr16:56691854-56693215	5,37215
MT1G	chr16:56700652-56701977	6,59408
MT1X	chr16:56716381-56718108	1,96232

Summary of results and future goals

In summary, bioinformatic analysis of the RNA-seq data from chordoma and nucleus pulposus samples demonstrated important similarities in the gene expression patterns between our data and previously published results. In addition, novel genomic regions, gene sets and pathways with altered activities could also be identified in the partial analysis, which were not described before in chordoma. The bioinformatic analysis is ongoing, with the following immediate goals:

1. Establish the *complete list of putative microdeleted/methylated or amplified chromosomal regions* in the chordoma samples, based on the RNA-seq data. The genes in the affected regions will be functionally characterized based on literature data, to identify the regions that are likely to harbor *driver genes of oncogenesis*. We plan to use real-time quantitative RT-PCR (*R-qPCR*) to validate the gene expression changes in selected regions, followed by qPCR on the genomic DNA to validate copy number changes.

2. *Expand the differential gene expression analysis*, using DefSeq2 and EdgeR packages in Galaxy. Summary of the different approaches.

3. *Continue the pathway analysis,* supported by the additonal differential gene expression analyses and Gene Set Enrichment Analysis (GSEA), to identify key pathways contributing to the aggressive phenotype of chordoma. Validate the findings with R-qPCR.

4. Perform *mutation analysis* and search for potential *gene fusion products* in the RNA-seq data, to identify driver genes and druggable targets in chordoma. Validate our findings with traditional sequencing of genomic DNA.

5. Similar *analysis pipeline* will be performed on the *miRNA* sequencing data after the correction of the unanticipated technical error.

6. Validated RNA and miRNA sequencing results will be merged and bioinformatically analyzed to explore the *miRNA related regulatory elements* in the chordoma development.

Acknowledgement

Principal Investigator of this project wants to thank the efficient collaboration of all research partners, namely Beata Scholtz PhD, Gergely Papp PhD, Arpad Bozsodi MD and Peter Pal Varga MD. Without their work, we have not been able to take these significant steps forward in chordoma research.

Aron Lazary MD PhD

Budapest, 30 Sep 2015