Identifying potential cancer vaccine targets with high-throughput sequencing

E. Aronesty[1,2], K. Robasky[2], W. D. Jones[2,3]

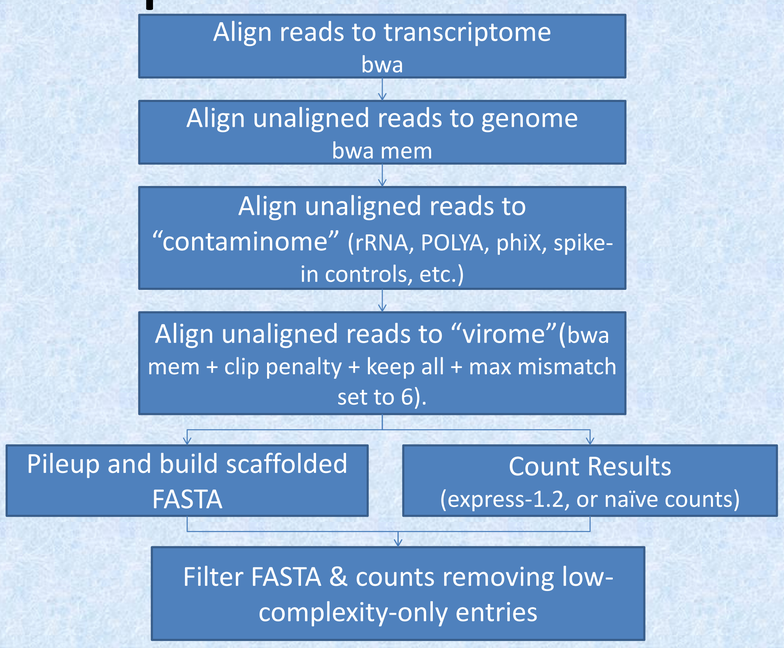
[1] Bioinformatics, Johns Hopkins University; [2] Expression Analysis, Durham, NC; [3] Pathology, UNC School of Medicine

Here, we use RNA-Seq to detect viral homologs in tumor sample expression data obtained from The Cancer Genome Atlas (TCGA) [1]. We compared normal tissue viral expression to cancer tissue expression using sequence alignment to a repeat-masked virome, quantification and differential expression. We validated the viral homologs using pileup+scaffold assembly and BLAST of high-complexity contigs.

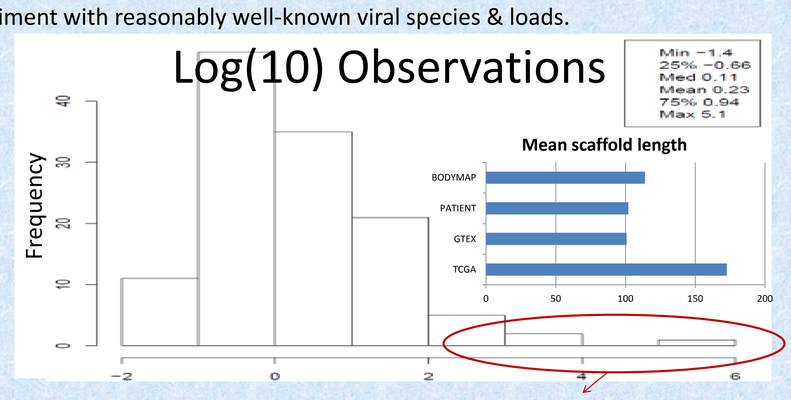
Samples included public TCGA large-b-cell lymphoma samples, Illumina Body Map lymph node pooled sample, and a Hodgkin's cancer patient's whole node. All samples were fresh-frozen, shipped on dry ice before extraction and RNA sequencing. Although not preferred, Poly-A selection was used.

In 1991 it was estimated that 15% of cancers were caused by viruses [2]. By 2012 this estimate was 30% [3]. TCGA gastric and colorectal cancer data have been screened for viral presence [4,5] however these experiments lacked normal controls. The Genome Tissue Expression (GTEX) project [6] can serve an adequate control for viral screening when no other controls are available. The use of normal controls enables differential expression methods to be used on viral quantification matrices.

Viral Quantification Pipeline Overview



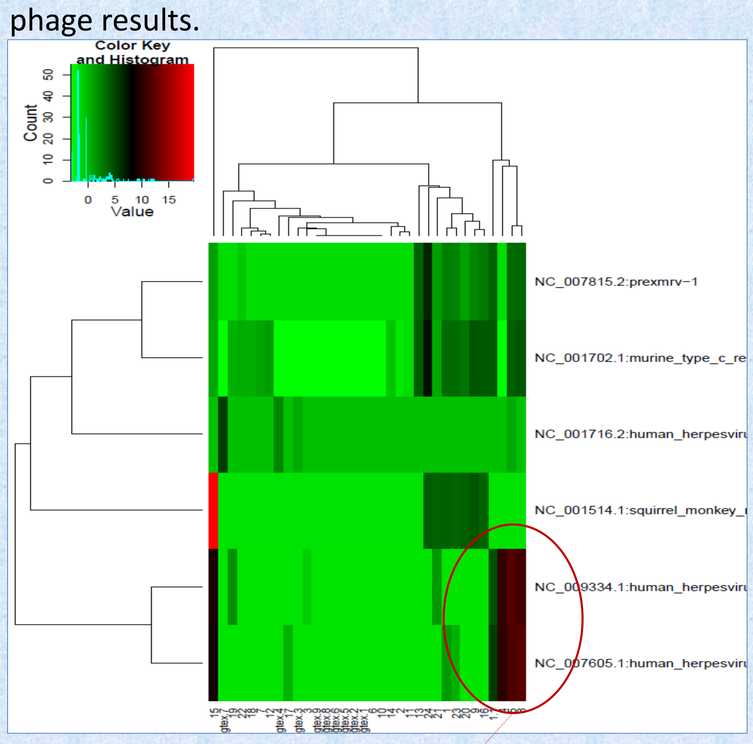
Programs like express & RSEM *should* be able to better disambiguate species by using expectation maximization. Demonstrating this would require a controlled experiment with reasonably well-known viral species & loads.



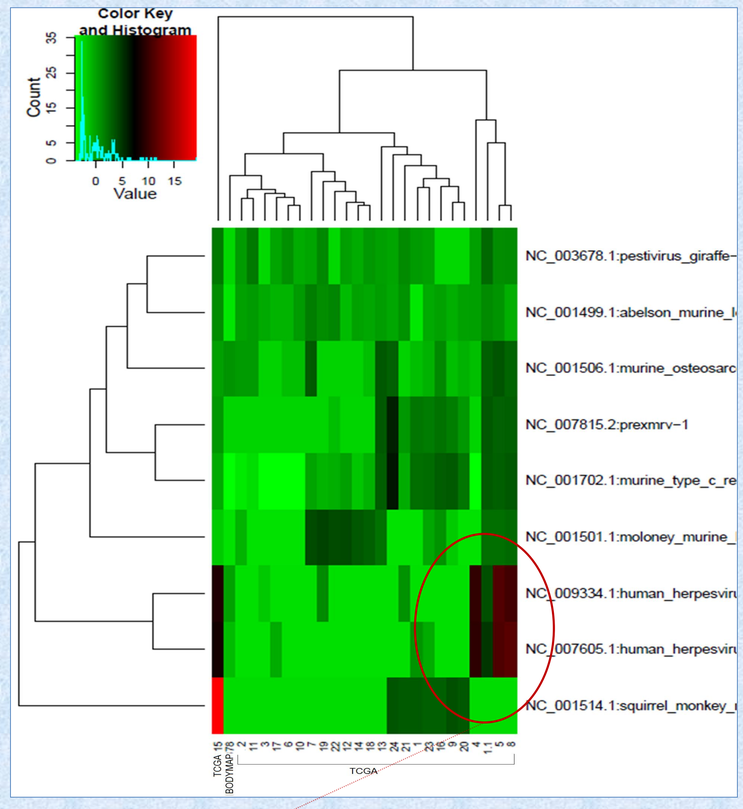
The vast majority of alignments were low-numbered & likely artifactual. However, there are a some viral alignments with sufficient reads (>100) to build a scaffold and demonstrate a very low eVal using tools like BLAST. TCGA viral scaffolds were generally longer, and had more specific BLAST annotation results.

TCGA vs Gtex*

edgeR + FDR adjustment were used on control (normal) vs exp (cancer). Heatmaps include significant, non-phage results.



TCGA vs Bodymap*



* Before plotting, viral count matrixes were filtered for significant (q<.05) viral fragments first using edgeR to evaluate whether the viral presence in the TCGA samples were greater than the controls.

Despise some significant results, these are a suboptimal comparisons. Poly-A selection was used, so some kinds of viral RNAs could not be observed. Also, GTEX has no lymph node samples, so thyroid was substituted. Finally, both GTEX and TCGA do not provide raw, unmodified FASTQs, and there are, likely, artifacts resulting from post-processing. Worse, some TCGA samples, like thyroid cancer, appear to have had unaligned reads *discarded*, making a complete analysis of this type impossible.

Illumina's body map while providing raw data for reference, has only a single, pooled sample for lymph node, so it is impossible to estimate normal variation.

Lack of access to normal reference samples is a systemic issue that is not, yet, adequately addressed by GTEX/bodymap.

- EBV Presence in 32% of TCGA lymphoma samples vs bodymap or GTEX.
- XMRV Presence in 80% of TCGA lymphoma samples vs bodymap or GTEX.

Virus : Genus or Name	TCGA qVal	GTEX	Bodymap Lymph	TCGA Lymph	Patient Lymph	% TCGA Obs	Longest Contig	BLAST eVal	BLAST ID (Human Preferred 100x)
NC_001514.1:betaretrovirus	0.00	0.00	0.00	121090.64	0.00	0.32	8776.00	0.00	M23385.1 (SMRV)
NC_009334.1:human_herpesvirus_4_type_2 (EBV)	0.00	0.00	7.00	1366.64	959.59	0.32	6503.00	0.00	V01555.2 (EBV)
NC_007605.1:human_herpesvirus_4 (EBV)	0.00	0.00	3.00	1545.68	742.86	0.36	8920.00	0.00	V01555.2 (EBV)
NC_001702.1:gammaretrovirus	0.00	0.00	0.00	120.96	28.98	0.80	3732.00	0.00	JF908816 (XMRV)
NC_001408.1:alpharetrovirus	1.00	0.00	0.00	0.00	113.06	0.00	123.00	4.00E-55	D10652.1 (Sarcoma Virus)
NC_008094.1:alpharetrovirus	0.58	1.44	0.00	0.36	103.27	0.16	105.00	6.00E-31	NM_198291.2 (SRC Viral Homolog)
NC_001407.1:alpharetrovirus	0.65	0.22	0.00	0.00	99.59	0.00	76.00	2.00E-16	D10652.1 (Sarcoma Virus)

1. Running BLAST on the contigs is important to filter out viral homologs like this one. Not certain why this was not removed during alignment to genome + transcriptome.
2. Long contigs for SMRV, EBV were unambiguous indicators of viral presence in TCGA lymph node samples. These were not observed in GTEX or Illumina Bodymap samples.
3. In addition to EBV, Patient A had significant presence of a Sarcoma Virus that were not observed in control samples (either GTEX or Bodymap). All samples were fresh-frozen, however as with all findings, these could have been the results of sample handling or laboratory contamination.

>CONTIG1 – BLAST 4e-55, 99% Identity, Rous Sarcoma or Avian Leukosis Virus
AATAGTGGTCGGCCACAGACGGCGTGGCGATCCTGTCTCCATCCGTTTCGTCTATCGGGAGGCGAGTTCGATGACCCTGGTGGAGGGGGCTGCGGCTTAGGGAGGCAGAAGCTGAGTACCGTC
>CONTIG2 – BLAST 3e-16, 100% Identity , Rous Sarcoma or Avian Leukosis Virus
GACACAACGCTAAACAGTGTAGGAAGCGGGATGGCAACCAGGGCCAACGC

Conclusions

Roughly 30% of lymphoma cancer samples are associated with EBV, and the majority are associated with EBV, XMRV or SMRV.

The use of controls is essential for removing ubiquitous viruses and viral homologs from results.

It was difficult obtaining information about public samples. Each TCGA contributor used different processing methods. Many samples (thyroid, ovarian) did not have whole FASTQ's (an artifact of using BAM as the primary submission format). Although there are often stringent sample submission guidelines for large-scale projects, there appear to be insufficient guidelines for sequencing and bioinformatics.

References

1. *The Cancer Genome Atlas initiative*, National Human Genome Research Institute, USA. http://cancergenome.nih.gov
2. zur Hausen H.. "Viruses in human cancers" (1991) Science. 1991;254(5035):1167–1173.
3. "Global burden of cancers attributable to infections in 2008: a review and synthetic analysis" The Lancet Oncology, Volume 13, Issue 6, Pages 607 - 615, June 2012
4. Stevenson, "Vaccination against a hit-and-run viral cancer" J Gen Virol. 2010 September; 91(Pt 9): 2176–2185.
5. Moore et al, "The Sensitivity of Massively Parallel Sequencing for Detecting Candidate Infectious Agents Associated with Human Tissue" (2011) PLoS ONE 6.5
6. Strong, "Differences in Gastric Carcinoma Microenvironment Stratify According to EBV Infection Intensity: Implications for Possible Immune Adjuvant Therapy" (2103) PLoS Pathogens