DepMap is interested in extracting new features from existing DepMap omics data and releasing them on a quarterly basis. This is particularly of interest to improve drug target discovery, to fill in the 'gaps' in our existing predictive models of dependencies, to increase the interpretability of the models, and to find better associations between cell lines and patient tumors.

The following omics features are candidates for further evaluation:
Extracted feature
Additional feature
Non directly useful for dependency prediction
Being worked on now

From DNAseq (WES/WGS)

1. ***SV rate status***
2. **SV caused gene disruptions** (e.g. caused by an in-frame SV event)
3. **Improved somatic mutations:** using a modified version of the CGA pipeline with updated annotations and filters for cell lines
4. **Germline mutations:** using HaplotypeCaller, defining any mutations not in our somatic mutation set as being germline.
5. ***Tumor mutational burden***
6. ***DNA repair deficiency scores*:** (HRD, TAI). See some description here. These are included in Xena for TCGA.
7. ***COSMIC mutational signatures*:** using CGA's SignatureAnalyzer
8. ***Improved estimation of mutation function*:** using GATK's Funcotator and EMBL's VEP (with plugins: ExACpLI, dbNSFP, FunMotifs, Mastermind, SpliceAI)
9. **Phased Germline mutations:** Important for many other downstream analysis. Can be backed by paired end RNA sequencing using whatshap
10. **Allele-specific copy number**
11. ***mutation LOF/GOF***: Maybe existing tools are doing enough on this front, but there might be more we could do. For example, looking for expression signatures associated with functional mutations. Using information about allelic fraction and allele-specific expression.
12. **Exonic copy number counts**
13. **Cancer cell fractions:** to capture the amount of subclonality in each the sample. We can use tools such as PyClone and ..
14. **LOH**
15. **Evolutionary/subclonal tree of cell lines** (using PyClone, PhyloWGS, etc)

From WGS only

16. **Whole genome doubling (WGD)**

17. **Tandem duplications status**
18. **MSI status**
19. **MSI motif:** Determine short repeats (shorter than 6 bp) across the WGS data and calculate their length/motifs. using tool such as : with annotations using : (e.g. based on closest gene)
20. STR: compute a pseudo STR profile for a cell line using its WGS, adding a layer of QC to our internal data
21. **ecDNA**
22. **Telomere lengths**
23. **Mitochondrial DNA features:** mutations/CNV/SVs
24. **Splicing disruption**: e.g. if a SNP/ indel has changed the splicing of a gene (using VEP's spliceAI tool or similar)
25. ***TF disruption:***
    a. whether a SNP/indel caused the appearance/disappearance of a TF binding site. We can use VEP's FunMotif or Expecto's chromatin predictor,
    b. TF / TSS location change caused by a SV
26. ***Inferred chromatin accessibility*(?): From mutation rates (Gaddy postdoc working on this). Or inferred from expression data/methylation data?
27. **Inferred Expression effect:** We can use the Expecto model to find the impact of the mutations (SV, indels, SNP) on the nearby gene expression
28.

From RNAseq

29. **Circular RNA (ciRNA)**
30. ***New pathway/network enrichment scores*:** Trying/implementing other methods beyond the usual GSEA for assigning function to pathways.
31. **Allele-specific expression**
32. **Exon expression**
33. **TF->gene regulatory network: inferelator**

From Multiple Omics sources (WES/WGS/RNA/RRBS)

34. **Two-hit features:** e.g. a gene that has both methylation and a mutation (on different copies) → *MAKE OUR OWN TOOL*
35. *MethylClock: get access to the metabolic age of a sample from analyzing its methylation pattern using the methylClock algorithm*
36. ***Multi-omic features/clusters*** (like icluster+ clusters or MOFA)
37. **Detecting cancer-causing viruses:** EBV, HPV, etc (some work has already been done here using RNAseq)
38. ***Inferred protein/TF activity*:** (e.g. Califano et al methods). [Califano's group is presumably generating VIPER based scores now, and we can check in with them about

this]. Transcription factor activity using RABIT. Julio Saez-Rodriguez TF inference method (DoRothEA).

39. ***Omics gene level functional states*:** Combine DNA, RNA, (and protein level where available) features to estimate single gene-level activity states. Related to PARADIGM modeling approach. This could include the idea of estimating a 'null state' explicitly (incorporating damaging mutation, deletion, methylation, etc).
40. **Batch-corrected expression/CNV/etc**
41. **Global vs local information**: Maybe similar things could be true of splicing dysregulation, high rates of structural variants, methylation/epigenetic signatures, etc.