Relative frequency measurements: Metrics for sample quality, sequencing integrity, and batch effects in targeted NGS

Bonnie LaFleur, Dominic LaRoche, Kurt Michels, Shripad Sinari, and Dean Billheimer

(16 May 2016)

The views and opinions expressed in this talk are those of the authors, and not necessarily those of HTG Molecular Diagnostics, Inc. or The University of Arizona

THE UNIVERSITY OF ARIZONA

Outline and Strategy

- Introduction to RNASeq and HTG workflow
- Description of two target assays
- Framework for data evaluation through a series of propositions
  - Each proposition is demonstrated through a series of examples
  - Mathematical development is referenced when possible
- Recommendations
- Future directions
  - Rethink differential expression in terms of difference in compositions

NGS Workflow

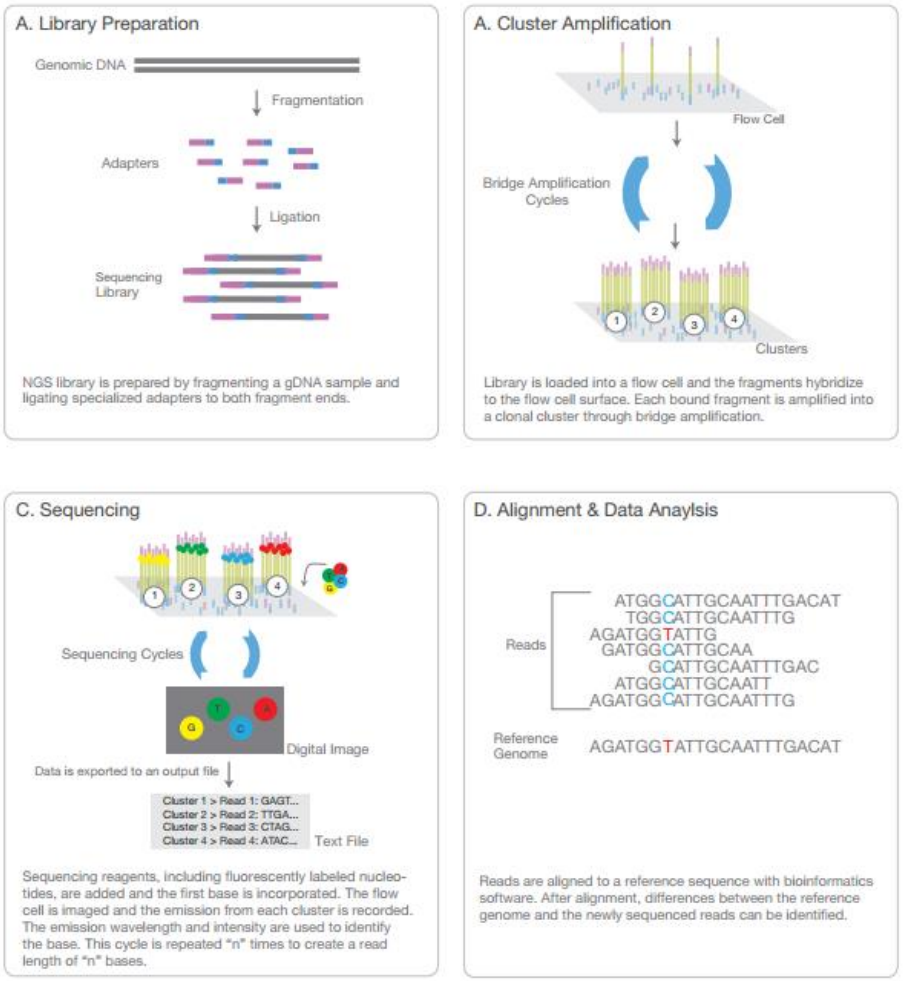
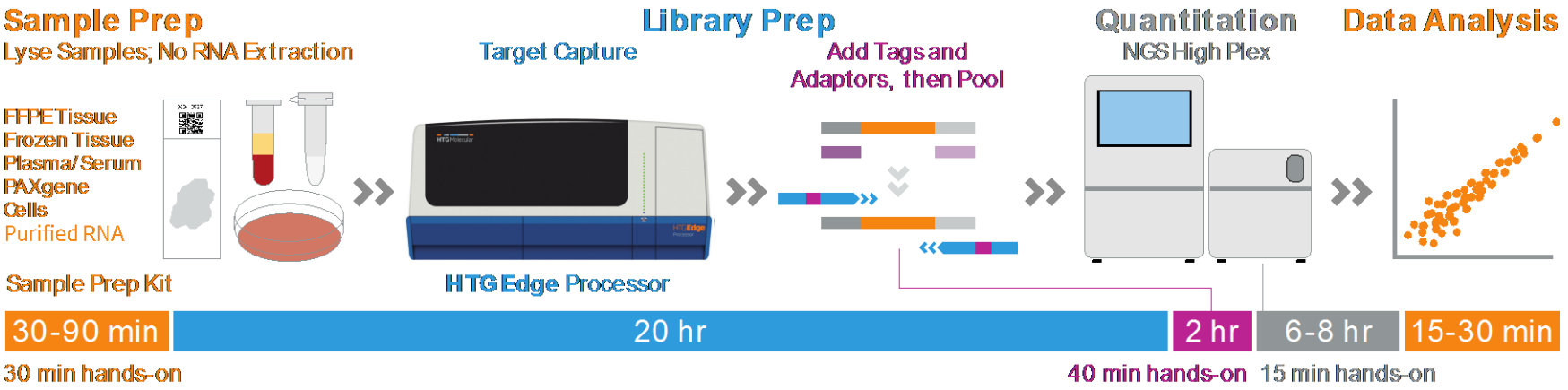


Figure 3: Next-Generation Sequencing Chemistry Overview.

Source: http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Workflow Synergy | HTG and NGS

HTG's Edge chemistry is optimized for NGS workflow automation



HTG Molecular

HTG EdgeSeq | Immuno-Oncology Assay

Immuno-Oncology drug response and immune response

549 genes, 10 major groups and pathways

- Drug / therapeutic targets
- Lymphocyte lineage markers
- Mechanisms of B and T cell activation
- Mechanisms of B and T cell response
- Cell adhesion molecules (integrins, adhesins, cadhesins)
- Inflammation activators and effectors
- Chemokines
- TNFs
- Ubiquitin and the Proteosome
- Toll-like receptors

Research Use Only

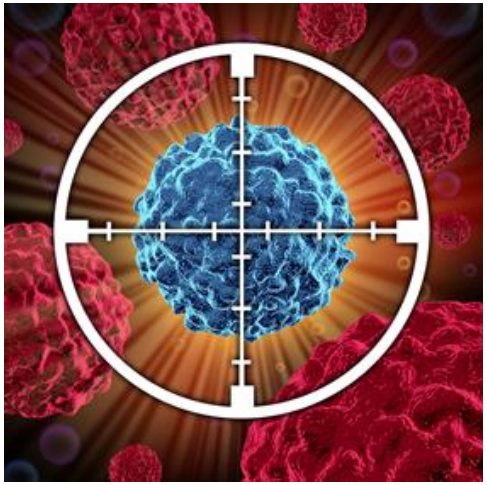
HTG Molecular

HTG EdgeSeq miRNA Whole Transcriptome Assay

Noncoding RNA

2,083 human miRNA transcripts

Sample Type	HTG EdgeSeq chemistry
FFPE Tissue	0.8-10 mm^2 area - Single 5 µm section
Frozen Tissue	10 µg
Cell Lines	250-5,000 cells
Plasma/Serum	15 µl
PAXgene	32 µl
Purified RNA	1.5-10 ng



Research Use Only

HTG Molecular

HTG Reproducibility Studies

HTG EdgeSeq assays used as examples

Sequencing plates for reproducibility studies

	Day 1	Day 2	Day 3
Processor 1	Plate 1	Plate 4	Plate 5
Processor 2	Plate 2		
Processor 3	Plate 3		

HTG miRNA WTA

- Study Design

  Multiple sample types and technical replicates are processed on five (5) quarter plates and then individually tagged, cleaned and quantitated to form five (5), 24-sample libraries sequenced on the Illumina MiSeq

- Samples
  - 3 sample types: plasma, FFPE & Brain RNA
  - 1 biological samples per sample type
  - 8 technical replicates per sample plate

  24 total wells per plate randomized across quadrant 1

HTG EdgeSeq Immuno-Oncology

Study Design

Single technical replicate of uRNA lysates over (5) quarter plates are tagged, cleaned as a pool, and quantitated to form five (5), 24-sample libraries sequenced on the Illumina MiSeq

- Samples

  24 total wells of uRNA lysate per plate randomized across quadrant 1

HTG Molecular

Proposition 1

Data that arise as measurements of relative frequency can be evaluated as compositional data

- Introduction to compositions
- Properties and forms of compositions

Targeted RNASeq is an example of inherently compositional data

Compositional Data

$\mathbf{x} = (x_1, x_2, \ldots, x_k)'$ vector of proportions

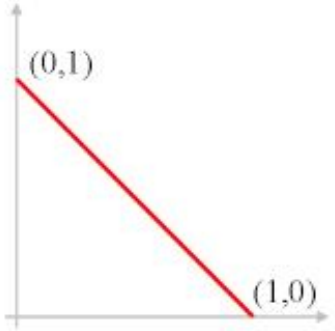
$0 < x_i < T$ all components positive

$\sum_{i=1}^{k} x_i = T$ sum to a constant (often T=1)

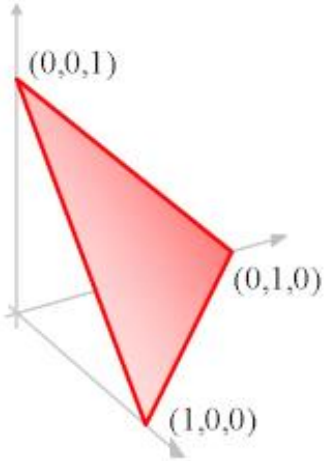
- Positivity and summation constraint complicate analysis
- Complicated covariance structure (Aitchison, 1982)
- As one component increases some other(s) must decrease

"Spurious correlation" (Pearson, 1897) - "fraught with difficulty and danger"

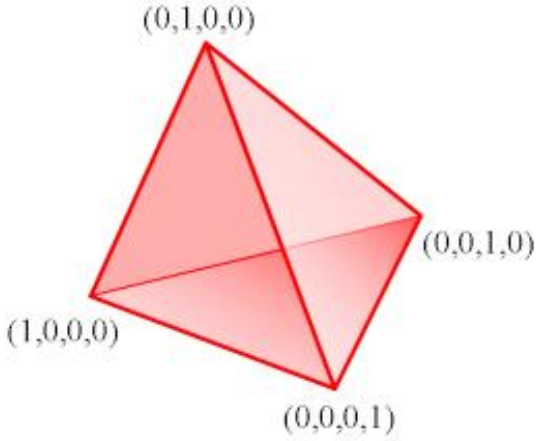
Geometry of Compositions



Any two part composition must lie on $x_1 + x_2 = 1$



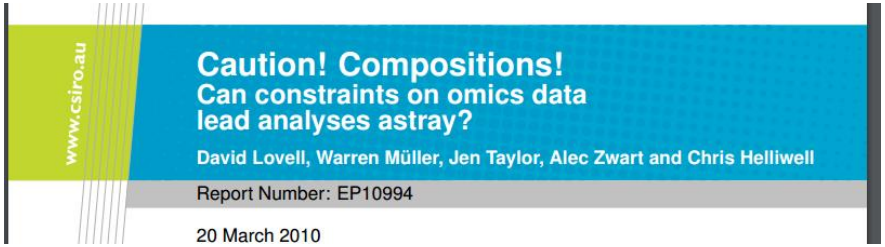
Any three part composition must lie on $x_1 + x_2 + x_3 = 1$



Any four part composition must lie on $x_1 + x_2 + x_3 + x_4 = 1$

Each figure represents a “standard simplex”

Source:



HTG Molecular

Mathematics of Compositions

Aitchison 1982, 1986

- Compositions lie in the *k*-1 dimensional simplex (S^{k-1})
- Use transformations to mitigate effects of constraints (multiple transformations to achieve different goals)
- One such transformation is the centered log ratio (clr)

$$\mathrm{clr}(\mathbf{x}) = log_2(\frac{\mathbf{x}}{g(\mathbf{x})})$$

where $g(\mathbf{x})$ is the geometric mean

- Resulting data in R^{k-1} (sums to 0), but angles between components are interpretable

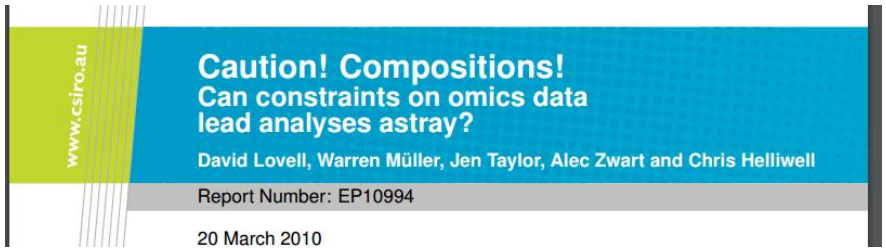
Example

	a	b	c	d	e	z	
basis:	$\mathbf{w} = (1001$	809	488	352	211	$100)$	
size:	$t = 1001 +$	$809 +$	$488 +$	$352 +$	$211 +$	100	$= 2961$
composition:	$\mathbf{x} = (\frac{1001}{2961}$	$\frac{809}{2961}$	$\frac{488}{2961}$	$\frac{352}{2961}$	$\frac{211}{2961}$	$\frac{100}{2961})$	
	$= (0.340$	0.270	0.160	0.120	0.071	$0.034)$	
geometric mean:	$g_m = (0.340 \times$	$0.270 \times$	$0.160 \times$	$0.120 \times$	$0.071 \times$	$0.034)^{1/6}$	$= 0.128$

clr(**x**)=	1.41	1.08	0.32	-.009	-0.85	-1.91

Counts Per Million (CPM) is similar to $\boldsymbol{x}$ - e.g., a composition. The compositional operations can be leveraged for use on this scale.

Source of table:



HTG Molecular

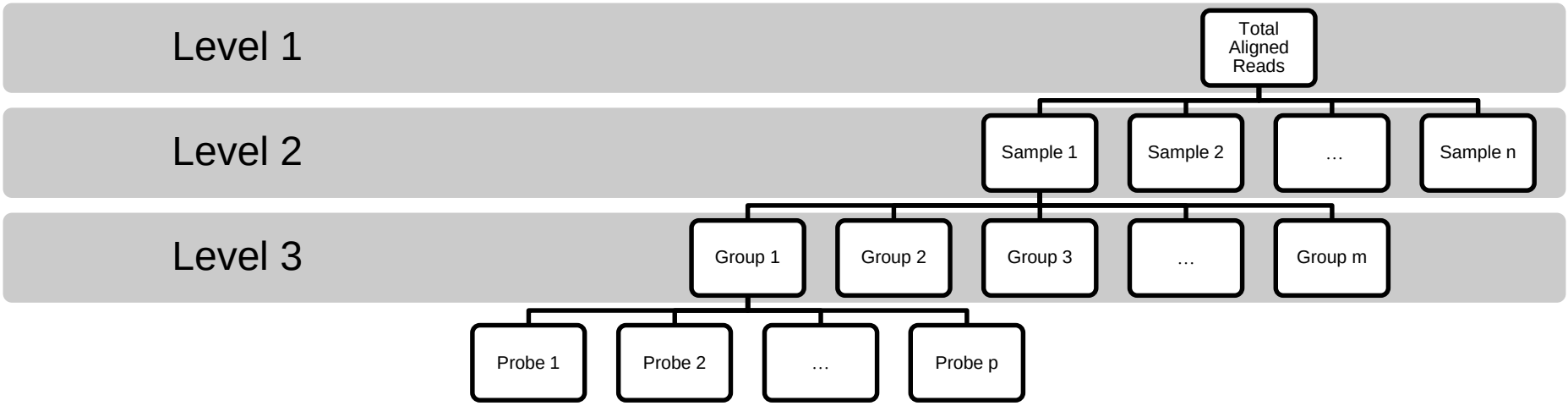
Operations on Compositional Geometry

- Amalgamation - can group/split components to work across hierarchical levels

- Subcompositional coherence - can omit unneeded components, and still retain coherent inference

Need to retain compositional structure at each level

Hierarchical Amalgamation



Example of a group would translational/functional category, like a GO or KEGG classification

Example Data

Probe Set	WT-miRNA				
Sample ID	1	2	3	4	5
Well	A1	B1	C1	D1	E1
Sample Name	run228-Plasma_4_1	run228-FFPE_5_1	run228-Plasma_8_1	run228-Brain_6_1	run228-FFPE_2_1
Total Reads	607503	482904	502930	534275	591505
Aligned Reads	472621	454161	396747	508749	553588
CTRL_ANT1	19	0	0	0	0
CTRL_ANT2	15	0	4	0	0
CTRL_ANT3	26	1	3	0	3
CTRL_ANT4	12	0	7	0	0
CTRL_ANT5	4	0	7	0	0
CTRL_miR_POS1	31230	1103	30389	3164	1190
CTRL_miR_POS2	21932	711	20076	2031	705
CTRL_miR_POS3	30824	1069	29763	3245	1111
CTRL_miR_POS4	25986	977	24757	2593	955
CTRL_miR_POS5	31259	1101	29123	3074	1051
CTRL_miR_POS6	28501	961	26477	2752	911
HK_ACTB	12	518	18	28	735
HK_B2M	135	1391	110	78	1879
HK_GAPDH	379	425	171	406	548
HK_PPIA	21	368	12	16	418
HK_RNU47	19	1792	4	648	1818
HK_RNU75	32	6519	5	307	8267
HK_RNY3	536	576	362	716	600
HK_RPL19	39	428	18	38	504
HK_RPL27	20	390	2	75	463
HK_RPS12	9	399	6	66	628
HK_RPS20	12	398	8	37	470
HK_SNORA66	18	1384	11	104	1682
HK_YWHAZ	33	474	20	126	659
let-7a-2-3p	19	6	2	5	3
let-7a-3p	13	0	1	0	1
let-7a-5p	1974	5685	1612	26055	7780
let-7b-5p	545	4746	438	17057	6437
let-7c-3p	14	0	5	1	1
let-7c-5p	434	2699	362	15988	3775

Example of Hierarchy:

- Total reads over entire run
- Sample level reads
- Functional group of probe reads (Control, HK, oncogenes, etc.)

HTG Molecular

Discussion

- Value in using the compositional framework for relative measurements
  - Leverages inherent structure of the data
  - Mathematical properties are well characterized
  - Convenient representation to examine sub-compositions

Proposition 2

Quality control metrics can be viewed as detection of unexpected data features

- Number of aligned reads example

Aligned Reads / Total Reads

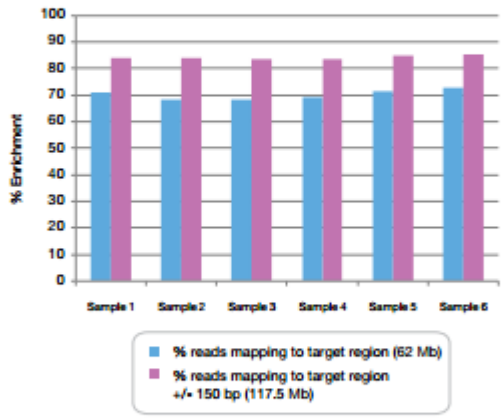
Number of aligned reads / total reads is compositional – that is it constrained by the available reads within a sequencing run

- Interested in how many sequencing counts have been allocated
- Measured on the sample level
- Ultimately impacts relative frequency at the probe level
- Most important contributor to success of differential expression/prediction

Distribution of Coverage Depth for Targeted Regions

Determining the distribution of coverage depth for targeted regions requires the generation of normalized coverage plots. Simply calculating the mean sequencing coverage will provide only a summary of the average read depth across the bases targeted in the enriched sample. The most commonly used methods report a given percentage of targeted bases covered at a particular depth (e.g., 90% of targeted bases covered at 10× read depth). It is possible to increase the total

Figure 4: High Target Specificity



Six pooled samples were analyzed on the HiSeq™ 2000 to demonstrate the specificity obtained in an optimized TruSeq exome enrichment experiment. The percent enrichment (y-axis) shows a high proportion of total reads mapping to the target regions (blue bars). The target region of +/- 150 bp represents percentage of total reads within 150 bases of the defined target regions (purple bars).

Source:
http://support.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_optimizing_coverage_for_targeted_resequencing.pdf

HTG Reproducibility Studies

HTG EdgeSeq assays used as examples

Sequencing plates for reproducibility studies

	Day 1	Day 2	Day 3
Processor 1	Plate 1	Plate 4	Plate 5
Processor 2	Plate 2		
Processor 3	Plate 3		

HTG miRNA WTA

- Study Design

  Multiple sample types and technical replicates are processed on five (5) quarter plates and then individually tagged, cleaned and quantitated to form five (5), 24-sample libraries sequenced on the Illumina MiSeq

- Samples
  - 3 sample types: plasma, FFPE & Brain RNA
  - 1 biological samples per sample type
  - 8 technical replicates per sample plate

  24 total wells per plate randomized across quadrant 1

HTG EdgeSeq Immuno-Oncology

Study Design

Single technical replicate of uRNA lysates over (5) quarter plates are tagged, cleaned as a pool, and quantitated to form five (5), 24-sample libraries sequenced on the Illumina MiSeq

- Samples

  24 total wells of uRNA lysate per plate randomized across quadrant 1

HTG Molecular

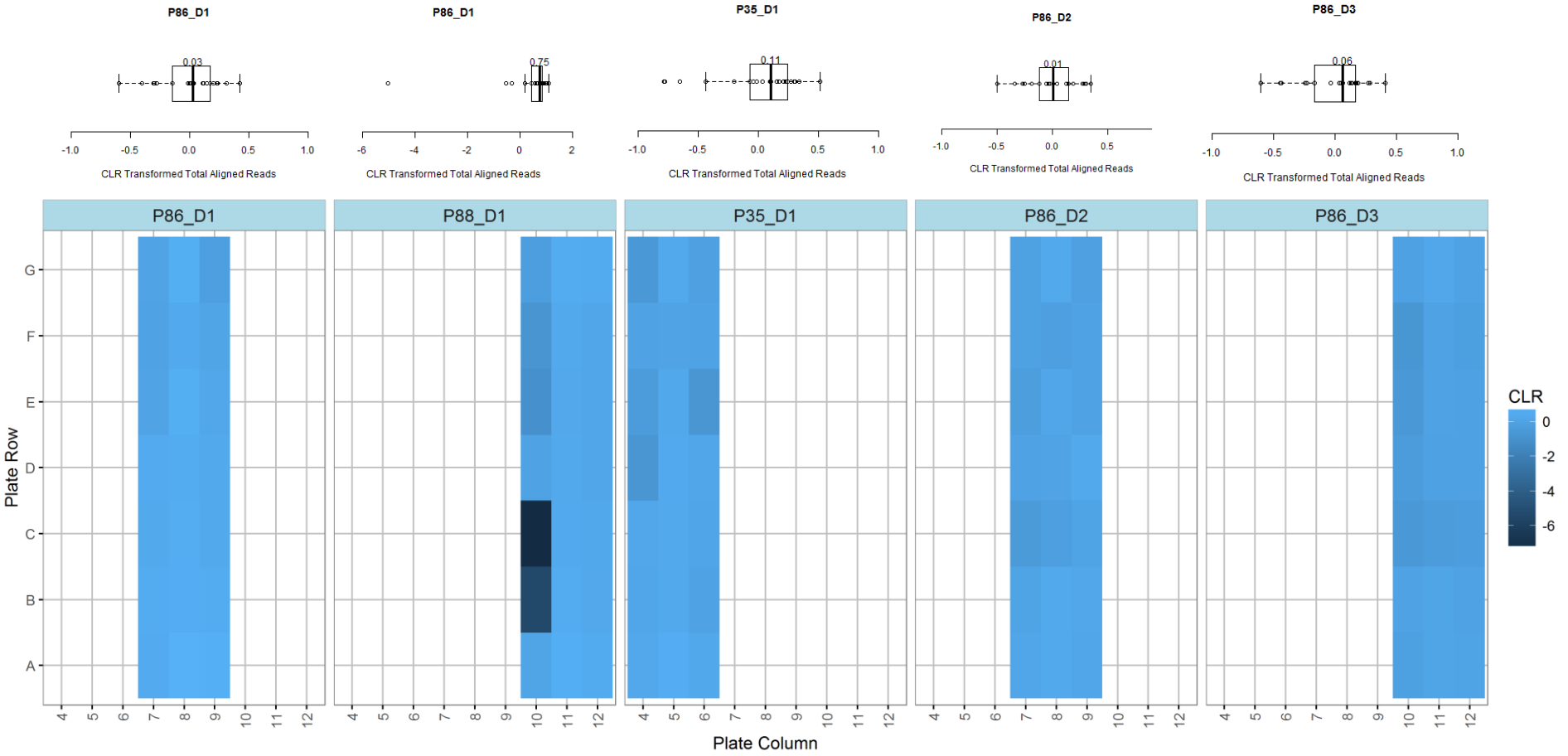
Example of Read Depth / # Aligned Reads

HTG EdgeSeq miRNA reproducibility study

- Visual display of sample level clr transformed total aligned reads
- Transformation occurs at the plate level – this retains hierarchical compositional structure on the plate
- Idea: use extreme values using residuals under normal theory assumptions to detect “outliers”

Example of Read Depth / # Aligned Reads

HTG EdgeSeq Immuno-Oncology reproducibility study



Test identifies 6 samples with lower than expected # aligned reads – indicates possible loss of sequencing integrity

HTG Molecular

Discussion

- This simple example shows how exploiting the inherent compositional nature of RNASeq data can be used to detect outliers
- This can be extended to other sequencing-based QC metrics (% passing Q30 score)
- Detection of sample or run level failure is critical for diagnostic assays

Proposition 3

Compositional geometry enhances multivariate feature evaluation

- Exploratory data analysis for batch effects

Evaluation of Batch Effects

- Definition: batch effects are technical variation that can possibly confound biologic variation

- Typical methods for detection of batch effects
  - Multivariate methods – Principal Components Analysis (PCA)
  - Visual inspection of expression differences (not useful for diagnostic appliations)

Correlations and Distances

- clr() covariances are interpretable in R^k
  - Useful for PCA and other dimension reduction
  - Compute usual (Euclidean) covariances and correlations on clr transformed data
- New distance metric - Aitchison distance (1986)
  - Accounts for compositional simplex structure

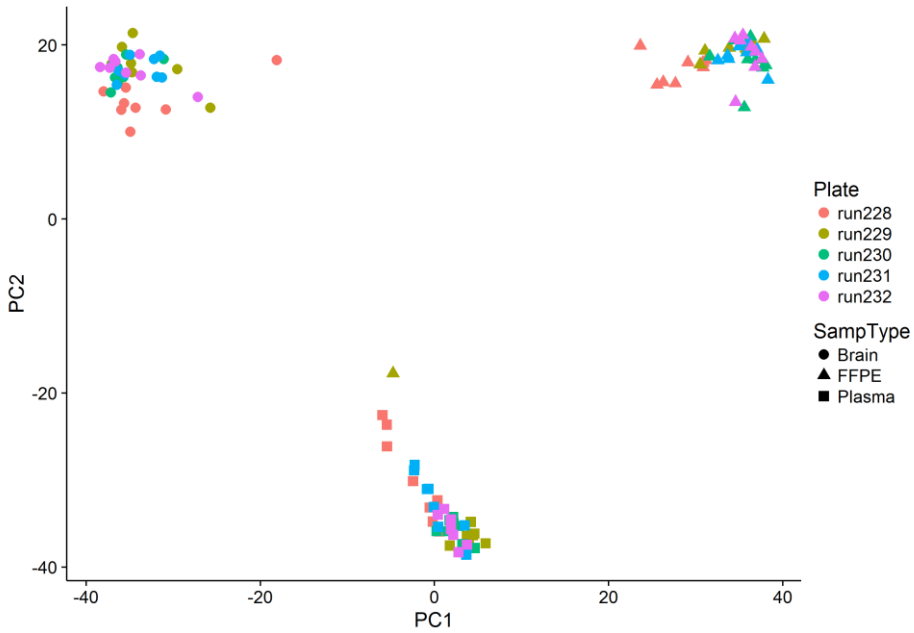
$$d_A(\mathbf{x}, \mathbf{y}) = ||\mathrm{clr}(\mathbf{x}) - \mathrm{clr}(\mathbf{y})||_2$$

- Statistical methods using correlations and distances are most affected by compositional structure
  - principal components, clustering, outlier detection

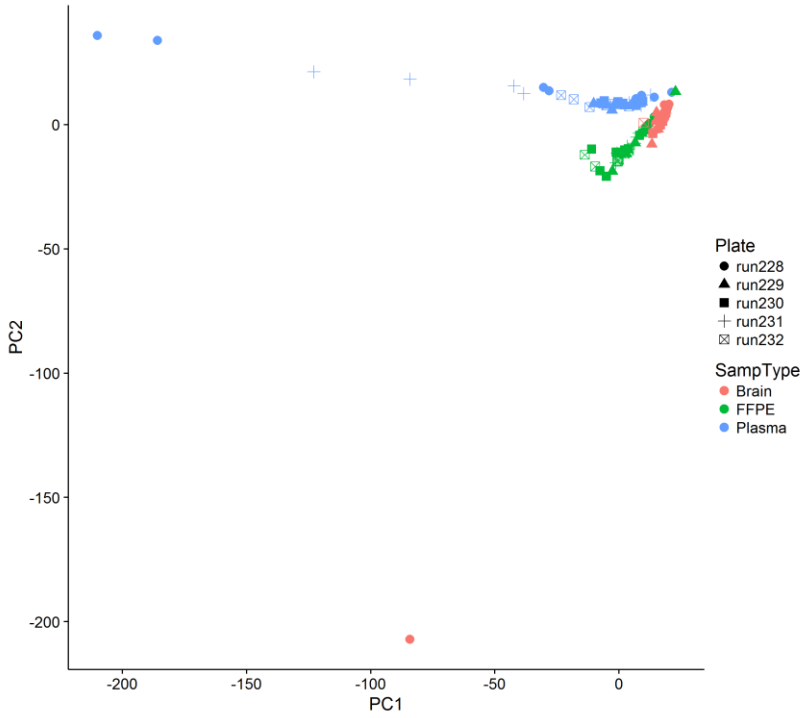
PCA Of Compositions

HTG EdgeSeq miRNA reproducibility study

clr Transformed



Raw Data



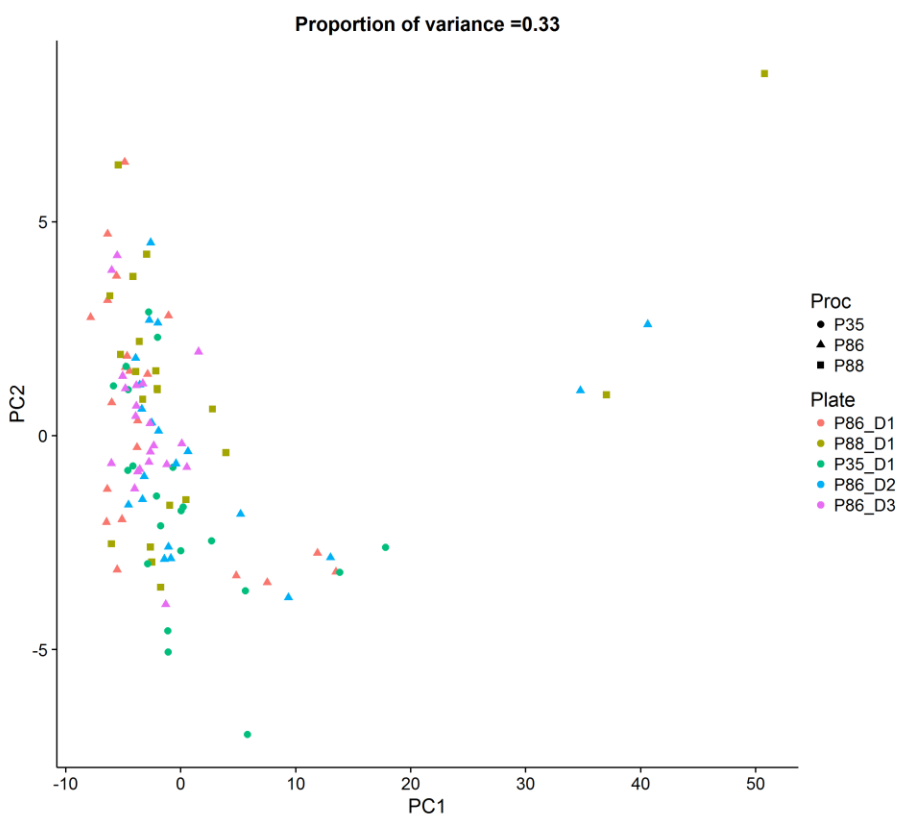
Neither method identifies a batch effect - clr transformation results in more meaningful evaluation of sample effects

HTG Molecular

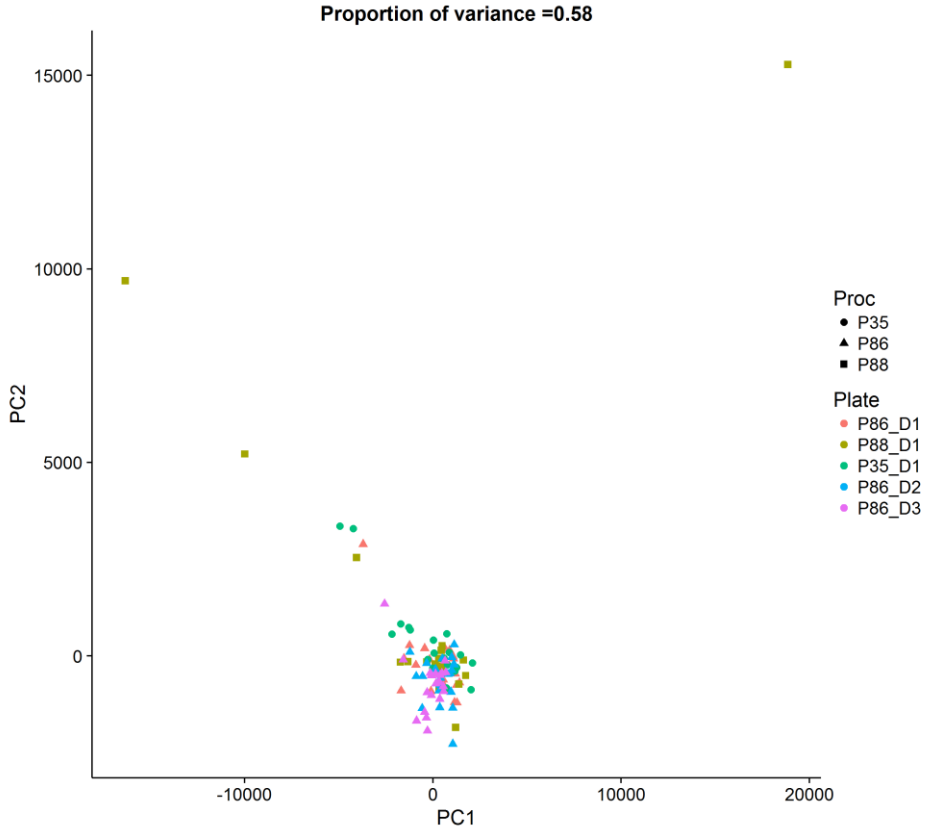
PCA Of Compositions

HTG EdgeSeq Immuno-Oncology reproducibility study

clr Transformation



Quantile Normalization



HTG Molecular

Discussion

- Aitchison distance and other compositional transformations provide more accurate measures of distance in multivariate space
  - Compositional geometry adds analytic benefit when data are inherently compositional

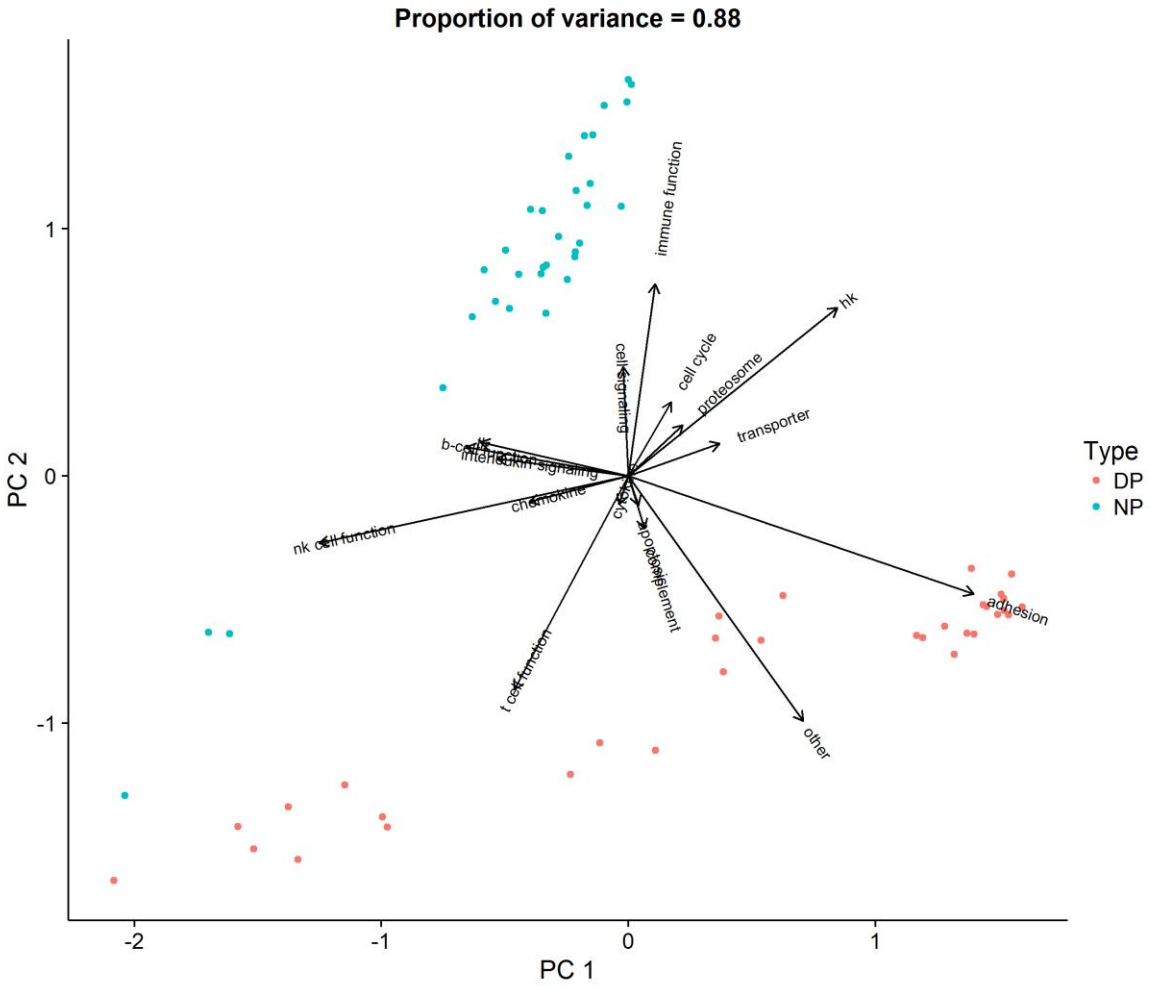
- Can construct these tests at the sample level
  - Avoiding group-level normalization methods that require renormalization as new cases are added
  - More appropriate for single sample diagnostic evaluation

Future Directions

- Use simplex geometry to evaluate patterns between biologically related groups of probes
- Process level QC metrics

Covariance Biplot Of Compositions

HTG EdgeSeq Immuno-Onocology Assay with Control Samples



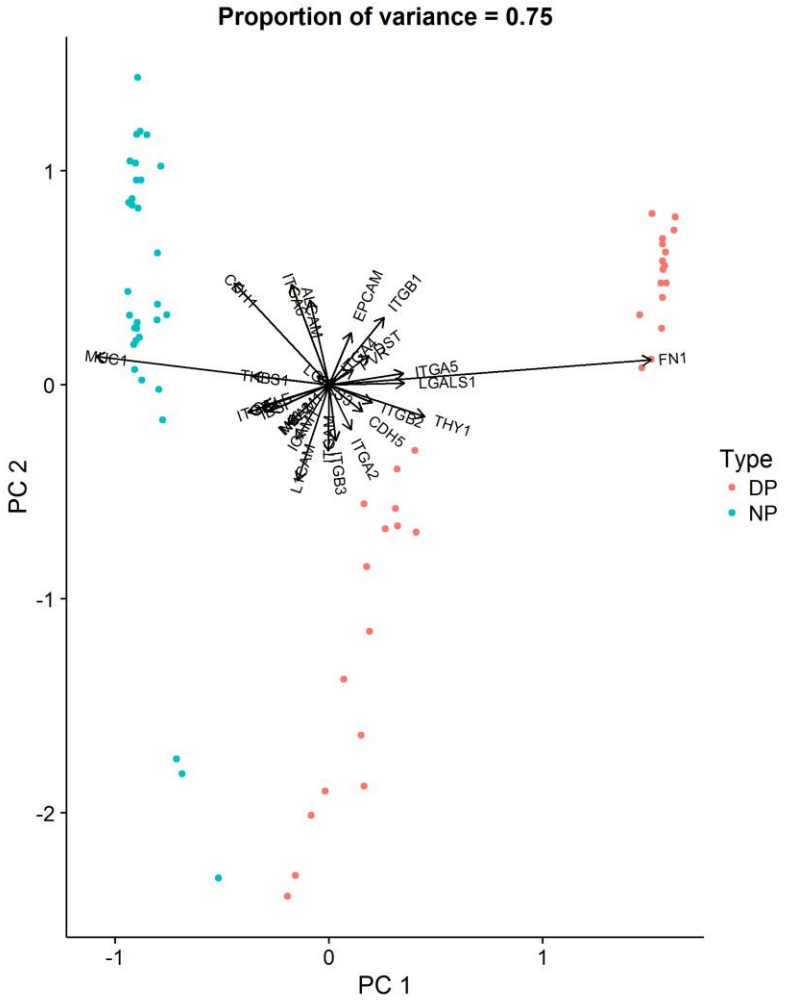
- Major grouping/pathway between normal pancreas (NP) and cancer (diseased pancreas = DP)
- Compositional structure is maintained
- Adhesion and immune function groups are contributing most to the discrimination between DP and NP
- We can further amalgamate down to the probe level with the groupings

HTG Molecular

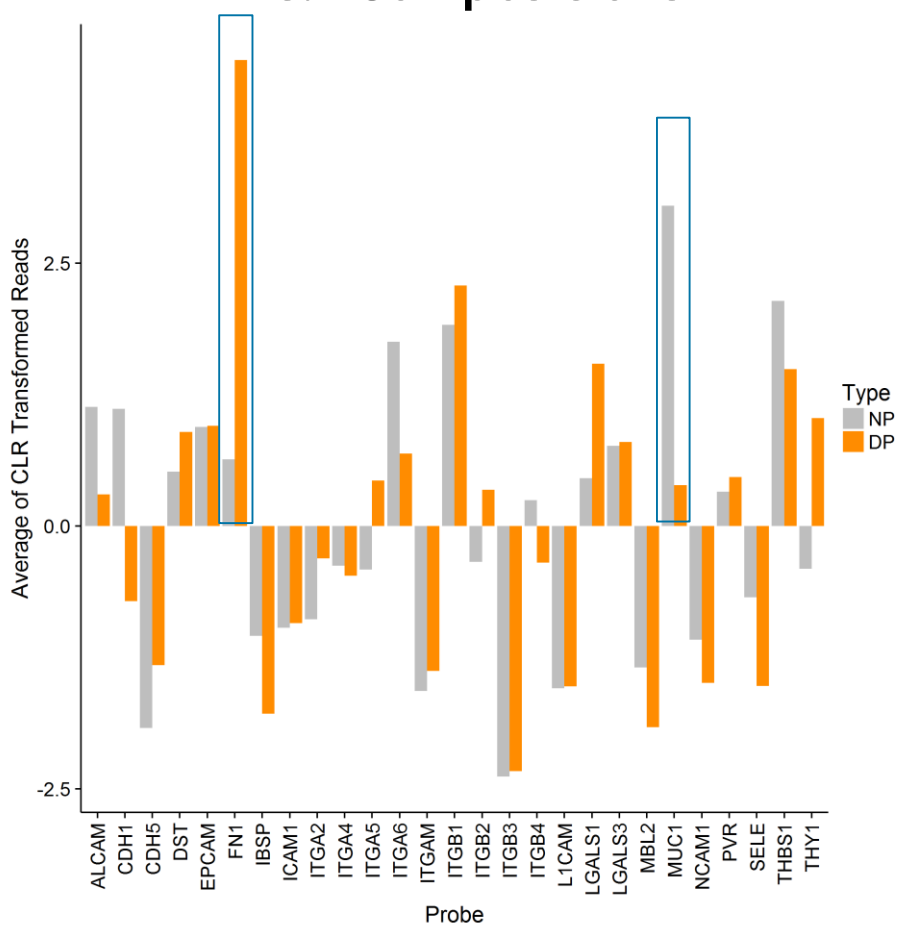
Probes Within Group Compositions

Cell Adhesion Probes

Biplot



clr Compositions

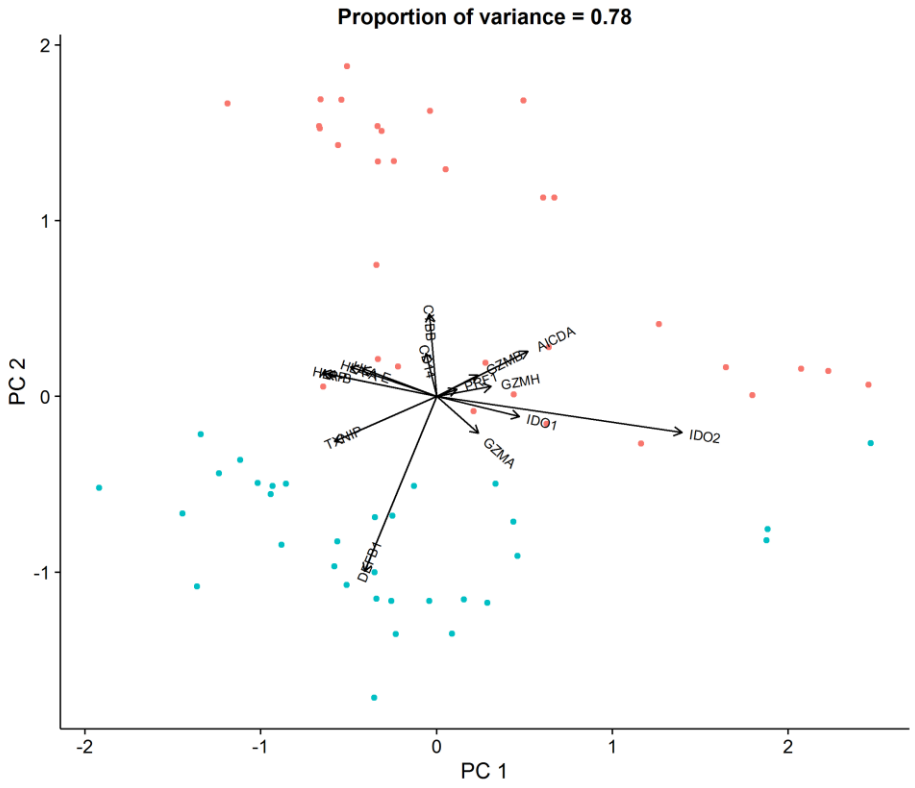


HTG Molecular

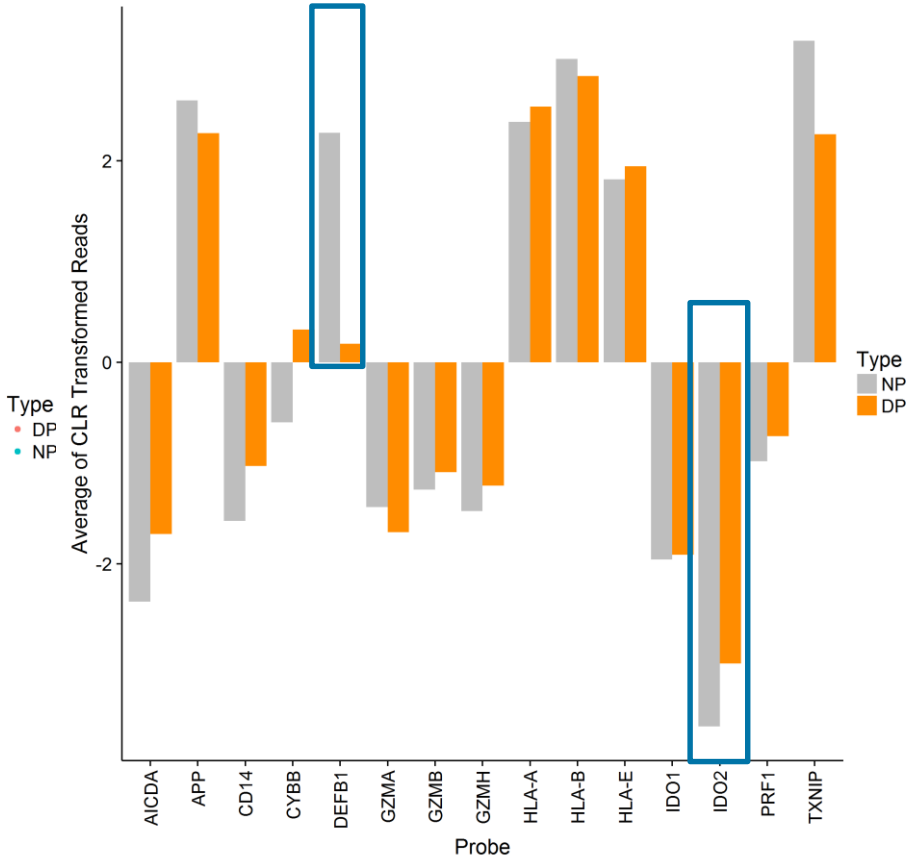
Probes Within Group Compositions

Immune Function Probes

Biplot



clr Compositions



HTG Molecular

Process Quality Control

- Current methods of process level (not sequencing) QC involves characterizing expected performance in advance
- Expected probe level expression (and variance) is determined over several sequencing runs
- Unexpected behavior identifies pre-sequencing issues (e.g., un-interesting amplification)

- The compositional framework can be used to identify “uniform” distributed sample compositions as process failures without defining “expected” behavior

Summary

- Evaluation of features in RNASeq (targeted and de novo) can be viewed as compositional data
  - Mathematical properties of compositional data are well established
  - CPM transformation is a composition
- Quality control metrics can be viewed as detection of unexpected data features
  - Outlier and influential sample features can be identified using well-established “normal theory” metrics on transformed data
- Compositional geometry enhances multivariate feature evaluation
  - Aitchison distance is equivalent to Euclidean distance when applied to clr transformed data

References

- John Aitchison, 2003 (2nd ed.). The Statistical Analysis Of Compositional Data. The Blackburn Press, Caldwell, NJ (USA).

- Vera Pawlowsky-Glahn, Antonella Buccianti (Editors), 2011. Compositional Data Analysis: Theory and Applications. Wiley, NY (USA)

- David Lovell, Jen Taylor, Alec Zwart, Chris Helliwell, 2010. Caution! Compositions! Can constraints on omics data lead analyses astray? CSIRO Technical Report, EP10994.

- Shripad Sinari, Dean Billheimer, to appear. The Analysis of Human Serum Albumin Proteoforms Using Compositional Framework. Statistical Analysis of Spectrometry based Proteomics and Metabolomics Data. Springer