Linear Models for Expression Profiling

Generalized linear models (GLMs) provide a framework for analyzing "counting"-based sequencing data from more than two conditions. The simplest case that really illustrates the power of GLMs is a two-by-two condition experiment. These conditions may be two independent treatments (+/- drug, +/- hypoxia, etc.), or one condition with ribosome and mRNA abundance profiling in parallel.

Below are two sample 2x2 experimental designs. The rows and columns are individual conditions and the table entries are the experimental samples.
The GLM analysis estimates expression levels and expression changes from the full data set. Different GLMs can be compared to test whether different conditions affect expression significantly -- including whether the combination of the two conditions can be explained as the sum of each condition individually, or whether they "interact".

	no ISRIB	**+ISRIB**
no Tm	ribo_untr	ribo_isrib
+Tm	ribo_tm	ribo_tmisrib

	Normox	**Hypox**
mRNA	mrna_norm	mrna_hypo
Ribo Prof	ribo_norm	ribo_hypo

Consider three genes showing different patterns of expression change in response to hypoxia, with replicated ribosome profiling and mRNA-Seq measurements in each:

Gene	**Effect**	**mrna_norm**	**mrna_hypo**	**ribo_norm**	**ribo_hypo**
A	no change	195 and 200	210 and 205	305 and 310	290 and 315
B	transcription	200 and 210	95 and 100	295 and 310	150 and 155
C	translation	310 and 295	295 and 300	105 and 100	395 and 410

Here's an R data frame of those counts. In the real example, there are some dummy genes included to keep the size factors and dispersions from getting weird.

```
> rawCounts[1:3,]
  mrna_normox_a mrna_normox_b mrna_hypox_a mrna_hypox_b ribo_normox_a ribo_normox_b ribo_hypox_a ribo_hypox_b
A           195           200          210          205           305           310          290          315
B           200           210           95          100           295           310          150          155
C           310           295          295          300           105           100          395          410
```

Here's an R data frame for the condition matrix in the data set. There are two factors, the `biol` factor, which is `mrna` or `ribo`, and the `oxia` factor, which is `norm` or `hypo`. These are created as factors using `factor (..., levels=...)` so we can control the order of the factor levels and make sure that the defaults are `mrna` and `norm`.

```
> conditions
              biol oxia
mrna_normox_a mrna norm
mrna_normox_b mrna norm
mrna_hypox_a  mrna hypo
mrna_hypox_b  mrna hypo
ribo_normox_a ribo norm
ribo_normox_b ribo norm
ribo_hypox_a  ribo hypo
ribo_hypox_b  ribo hypo
```

The simplest model is no gene expression change at all from hypoxia. In this model, the read count depends only on the "type" of sample (i.e., mRNA-Seq or ribosome profiling) and nothing else. There are two parameters, one giving the mRNA expression level and one giving the protein synthesis expression level.

	Normox	**Hypox**
mRNA	mrna_norm = biolmrna	mrna_hypo = biolmrna
Ribo Prof	ribo_norm = biolribo	ribo_hypo = biolribo

Here is the GLM for that model. The two parameters are estimated for each gene and log2-scaled. The deviance is also computed -- this is a measure of how well the optimized GLM fits the actual data. You can think of it as the probability of generating the real count data, assuming this GLM is true.

```
> glmNoChg  <- fitNbinomGLMs( countData, count ~ biol - 1 )
> format(glmNoChg[1:3,])
  biolmrna biolribo deviance converged
A    7.662    8.253  0.05282      TRUE
B    7.241    7.830  6.17859      TRUE
C    8.229    7.980 14.11188      TRUE
```

Because the GLM parameters are log2-scaled, it's hard to see how they line up with our real counts. We can compute two new columns, each of which reverses the log scaling. Once we do this, we can see how the parameters for gene A are good estimates of the actual read counts, whereas in gene B, the parameters split the difference between the hypoxia and normoxia value (i.e., gene B mrnaCounts is ~150, whereas normoxic mRNA is ~200 and hypoxic mRNA is ~100).

```
> glmNoChg$mrnaCounts <- 2**(glmNoChg$biolmrna)
> glmNoChg$riboCounts <- 2**(glmNoChg$biolribo)
> format(glmNoChg[1:3,])
  biolmrna biolribo deviance converged mrnaCounts riboCounts
A    7.662    8.253  0.05282      TRUE      202.5      305.0
B    7.241    7.830  6.17859      TRUE      151.3      227.5
C    8.229    7.980 14.11188      TRUE      300.0      252.5
```

We next try a model where hypoxia can cause an expression change. However, this change is the same in the mRNA abundance and protein synthesis samples. It's an extra parameter that's 0 for the "default" `oxia` condition, `norm`, and adds a contribution of `oxiahypo` in `hypo`.

	Normox	Hypox
mRNA	mrna_norm = biolmrna	mrna_hypo = biolmrna + oxiahypo
Ribo Prof	ribo_norm = biolribo	ribo_hypo = biolribo + oxiahypo

```
> glmNoTrl <- fitNbinomGLMs( countData, count ~ biol + oxia - 1 )
```

Add columns to this GLM that compute the read count values according to the formula above, as well as the non-log-scaled hypoxia effect.

```
> glmNoTrl$mrnaNormC <- 2**(glmNoTrl$biolmrna)
> glmNoTrl$mrnaHypoC <- 2**(glmNoTrl$biolmrna + glmNoTrl$oxiahypo)
> glmNoTrl$riboNormC <- 2**(glmNoTrl$biolribo)
> glmNoTrl$riboHypoC <- 2**(glmNoTrl$biolribo + glmNoTrl$oxiahypo)
> glmNoTrl$hypoxChange <- 2**(glmNoTrl$oxiahypo)
> format(glmNoTrl[1:3,])
  biolmrna biolribo oxiahypo deviance converged mrnaNormC mrnaHypoC riboNormC riboHypoC hypoxChange
A    7.650    8.241   0.0235  0.04875      TRUE     200.8    204.12     302.6     307.5      1.0164
B    7.659    8.262  -1.0298  0.03665      TRUE     202.1     98.98     306.9     150.3      0.4898
C    7.826    7.277   0.9640  7.65938      TRUE     226.8    442.50     155.1     302.5      1.9507
```

This model fits the data for gene B much better than the glmNoChg model, and the count estimates match the data very closely. It can't improve much on the fit of gene A. It's possible to test whether the decrease in deviance is "big enough", i.e., statistically significant.

```
> pNoTrlVsNoChg <- nbinomGLMTest( glmNoTrl, glmNoChg )
> pNoTrlVsNoChg[1:3]
[1] 0.94909 0.01320 0.01108
```

Here the p values for gene A (the first in the list) is quite high, whereas those for genes B and C are both quite low (p ~ 0.01). The model where expression changes in hypoxia explains the data for genes B & C much better than the model where expression depends only on whether the sample is mRNA-Seq or ribosome profiling.

The fit of gene C is better (lower deviance) as well, though it can't predict expression levels right. For instance, the actual mRNA abundance is ~300 in all samples, but glmNoTrl estimates ~225 for normoxic mRNA and ~440 for hypoxic mRNA. The model has only a single `oxiahypo` parameter and so it can't capture a change in ribosome profiling data that doesn't show up in mRNA abundance.

In order to capture this, we could add 2 extra factors, one for mRNA change in hypoxia and one for ribosome profiling change in hypoxia. Alternately, we could keep `oxiahypo` and add a 3rd factor corresponding to the change in translational efficiency in hypoxia -- that is, the additional change in ribosome profiling in hypoxia, on top of the change in mRNA abundance. This second alternative is closer to the biology we want to study.

The extra factor appears only in hypoxia ribosome profiling. In linear models, it's called an "interaction" term because it captures the interaction between the sample type (ribosome profiling, i.e., `biol ribo`) and the treatment (hypoxia, i.e., oxia hypo). R can create these interaction terms automatically if we combine individual factors using "*" rather than "+".

	Normox	**Hypox**
mRNA	mrna_norm = biolmrna	mrna_hypo = biolmrna + oxiahypo
Ribo Prof	ribo_norm = biolribo	ribo_hypo = biolribo + oxiahypo + biolribo:oxiahypo

Mathematically speaking, we now have 4 parameters (biolmrna, biolribo, oxiahypo, and biolribo:oxiahypo) that we're using to represent 4 different conditions. This is the fully saturated ("full") model, as we couldn't add any other parameter to it. We could choose a different set of 4 parameters (e.g., in place of biolribo:oxiahypo, we could instead add a 3rd factor, oxiatranslation that took on the value hypo only in ribo_hypo, and then ribo_hypo = biolribo + oxiatranslationhypo as discussed above) but they could be computed from these 4 parameters by simple arithmatic.

```
> glmFull  <- fitNbinomGLMs( countData, count ~ biol * oxia - 1 )
```

Here we calculate the counts for each condition using the formula above.

```
> glmFull$mrnaNormC <- 2**(glmFull$biolmrna)
> glmFull$mrnaHypoC <- 2**(glmFull$biolmrna + glmFull$oxiahypo)
> glmFull$riboNormC <- 2**(glmFull$biolribo)
> glmFull$riboHypoC <- 2**(glmFull$biolribo + glmFull$oxiahypo
     + glmFull$"biolribo:oxiahypo")
> glmFull$hypoxMrnaChg <- 2**(glmFull$oxiahypo)
> glmFull$hypoxTEChg <- 2**(glmFull$"biolribo:oxiahypo")
> format(glmFull[1:3,])
  biolmrna biolribo oxiahypo biolribo:oxiahypo deviance converged
A    7.626    8.264  0.07126          -0.09491  0.03212      TRUE
B    7.679    8.241 -1.07215           0.08402  0.02623      TRUE
C    8.241    6.679 -0.02405           1.99741  0.02624      TRUE
  mrnaNormC mrnaHypoC riboNormC riboHypoC hypoxMrnaChg hypoxTEChg
A     197.5     207.5     307.5     302.5       1.0506     0.9363
B     205.0      97.5     302.5     152.5       0.4756     1.0600
C     302.5     297.5     102.5     402.5       0.9835     3.9928
```

These counts all fit the actual data very well, and the mRNA and TE fold-changes match the values I picked when making up the data. We can compare this model, in which hypoxia affects mRNA abundance and translation, to the other two.

```
> pFullVsNoChg  <- nbinomGLMTest( glmFull, glmNoChg )
> pFullVsNoTrl  <- nbinomGLMTest( glmFull, glmNoTrl )
> pFullVsNoChg[1:3]
[1] 0.9897002 0.0461353 0.0008737
```

```
> pFullVsNoTrl[1:3]
[1] 0.897395 0.918724 0.005731
```

The p values here tell us that this model improves on glmNoChg for genes B and C both, but only improves on glmNoTrl for gene C. That is, adding a term for hypoxia affecting translation helps explain the gene C data better, whereas a single term for hypoxia impacting mRNA is enough to explain the gene B data.

When actually testing thousands of genes in parallel, it’s important to correct for multiple hypothesis testing (the p value adjudstment).