Statistical challenges in the "*Omics*"

Octavio Martínez de la Vega
Computational Biology Lab
"**From Data to Knowledge**"

National Laboratory of Genomics for Biodiversity
(Langebio - '*Unidad de Genómica Avanzada*')
Cinvestav - Irapuato

March 19, 2015

"Omics"?

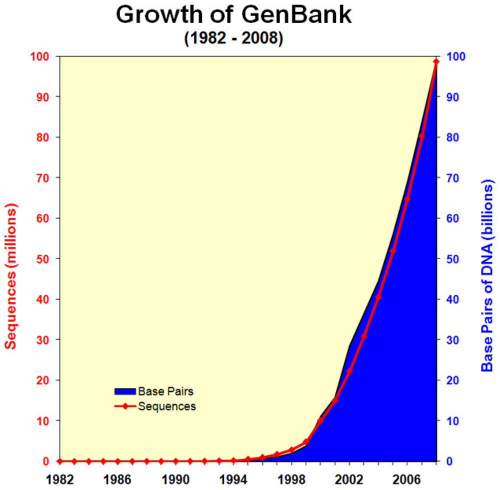
Informally refers to the study of 'whole' sets of molecules,

- Genomics - Study the whole genome (in contrast to studying a single gene).
- Transcriptomics - Study all genes expressed in a tissue under given conditions.
- Proteomics - Same for all proteins.
- Metabolomics - All chemical compounds produced
- ...

The main change with traditional molecular biology:
The very large nature of datasets

GeneBank 2015

187,893,826,750 bases, from 181,336,445 sequences



How many genomes are there?

Sequenced *versus* Existent (both estimated):

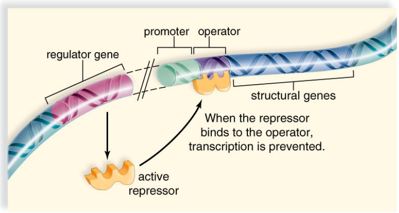
- Microorganisms: 18,000 / 10×10^6 ($< 0.2\%$)
- Fungi: 356 / 1.5×10^6 ($< 0.03\%$)
- Insects: 98 / 10×10^6 ($< 0.001\%$)
- Plants 150 / 435,000 ($< 0.04\%$)
- Terrestrial vertebrates and fish: 235 / 80,500 ($< 0.3\%$)
- Marine invertebrates: 60 / 6.5×10^6 ($< .001\%$)
- Other (nematodes, ...): 17 / 1×10^6 ($< .001\%$)

Genomics pipeline / challenges

- Select one (or a few) individual(s) / Who?.
- Sequence tens (to hundreds) of millions of small DNA sequences / keep and order these data.
- Solve this gigantic puzzle (obtain the 'genome') / Assembling.
- Where are the genes? / find (models) for genes.
- Make sense of all this / Annotate, Annotate, Annotate, ...

Genes: complex constructs

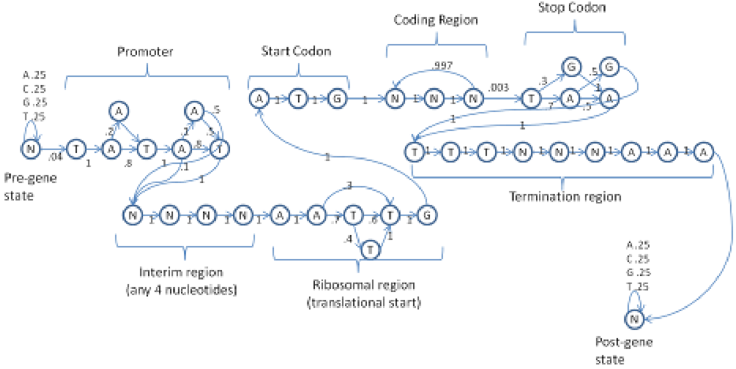
Genes are complex ‘data structures’ They include code for transcription and translation but also fuzzy signals for its processing (promotors, enhancers, exon / intron borders, methylation patterns, etc.)


promoter
operator
regulator gene
structural genes
When the repressor binds to the operator, transcription is prevented.
active repressor

Finding the ‘genes’ in a newly sequenced genome is a non-trivial exercise.

Statistical challenge: Model genes by HMM

Hidden Markov models (HMM): A finite model describing the probability distribution over an infinite set of sequences.



Statistical challenge: Model genes by HMM

To predict genes in a new genome by HMM we need:

- A good model with accurate estimates of transition probabilities.
- This can only be obtained and contrasted using empirical evidence on related genomes.
- Even when some of the states are well defines (p.e., ‘coding’ *vs.* ‘non-coding’ or repeated DNA), other are more fuzzy signals (p.e., ‘intron’ *vs.* ‘exon’ regions, etc.).
- Signals (code) between and within genes are not as well conserved as the usual ‘genetic code’ -in fact, there are many meta-codes that are taxa-specific.

Statistical challenge: Gene Identification

Even when model organisms (Human, mouse, rat among mammalians; Arabidopsis among plants, etc.) have well identified genes in other organisms we do not have experimental evidence to identify the genes with particular peptides

- We can compare similarity between DNA segments to look for an ‘orthologous’ gene.
- However, different gene families evolve (diverge) at different speeds.
- For many classes of non-protein coding genes there is a high degree of uncertainty about function.

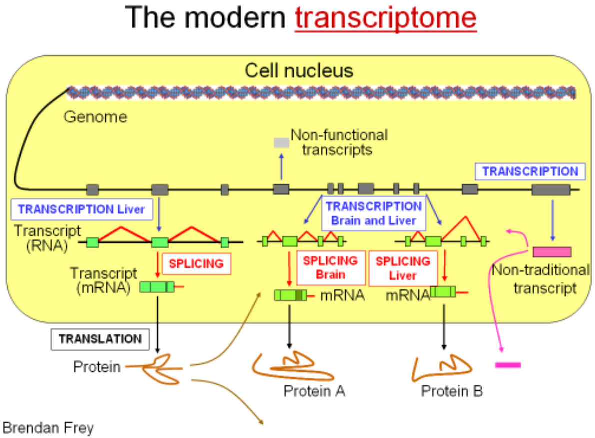
Statistical challenge: Gene annotation

To ‘annotate’ a gene means to describe its molecular function, cellular place and conditions of expression, etc. As for identification this constitute a statistical challenge because

- Different levels of noise are involved.
- Errors in the annotation of a gene are ‘inherited’ by all genes using this information.
- In many organisms there is lack of direct experimental evidence about gene function, thus the researcher must use annoyation inherited from model organisms with a high risk of error.

Transcriptomics: The genome in action

Genes are ‘active’ only at particular times and tissues. This is controlled trough a complex network: signals go from the environment to inter and intra cellular places activating and repressing gene expression.



Transcriptomics: The genome in action

A transcriptome experiment (RNA-seq):

- Select organism / organ/ tissues / time / conditions ⇒ Experimental design
- Isolate mRNA, convert to cDNA ⇒ construct genetic 'libraries'
- If genome is unknown assembly the transcriptome ⇒ core transcriptome
- Re-map the reads to the transcriptome ⇒ counts for each gene
- Statistical analyses of the counts from each library

Transcriptomics: Statistical challenges

In an RNA-seq we obtain counts for tens of thousand of genes from each library (the genetic library is the experimental unit, representing a particular replicate for each treatment). Researchers are interested in answering:

1. Which genes are expressed at each treatment?
2. Which genes are 'differentially expressed' between treatments?

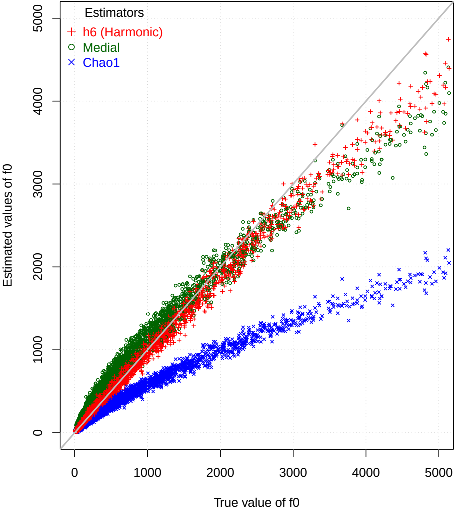
Statistical questions:

- How many replicates per treatment?
- How deep do we need to be the sampling (number of gene tags per replicate)?
- Which is the 'best' method for the analysis of this kind of data?

Transcriptomics: How many genes are expressed?

- We are sampling with replacement from a finite population (the expressed genes).
- Each expressed gene is represented by one or more mRNA molecules.
- The number of expressed genes, k, is unknown.
- The same problem exist in Ecology, when estimating the number of species in a community.
- The vector of frequencies of frequencies, $(f_1, f_2, f_3, \cdots,)$ gives information agout k (f_0)
- We have obtained better estimators for k, functions of $(f_1, f_2, \cdots, f_6)$ which give better results than the one in the literature.

The problem of the missing genes



Transcriptomics: Differential gene expression

- Counts at each library are the result of a multinomial distribution with unknown number of classes (k) and unknown probabilities, $(p_1, p_2, \cdots, p_k)$.
- This can also be modeled as a set of independent Poisson or Negative Binomial (NB) variables.
- In particular NB is attractive because it allows for the estimation of 'extra dispersion' that could be present between replicates.
- Classical analysis methods include the ones for 'contingency tables' (Pearson's χ^2, Likelihood Ratio Test for independence (also called G-test), Fisher exact test (suitable for 2×2 tables), etc.
- Other possibility is to use Generalized Lineal Models (GLM), in particle log-linear models.
- Because tens of thousands of tests will be performed, there is a need to correct for multi-testing.

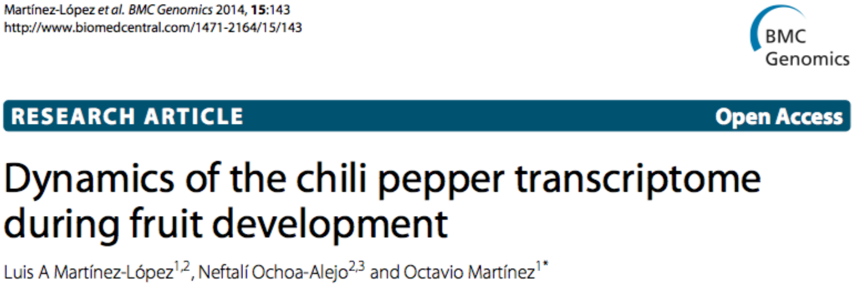
TRANOVA: a method for DGE in transcriptomics

‘TRanscriptome Analysis of Variance’ (TRANOVA) consist in measure the departure from independence (variance) within and between treatments through the Likelihood Ratio Test (or G-test):

$$G = \sum_i O_i \log_e \left(\frac{O_i}{E_i} \right)$$

- Values of G are calculated for the counts between treatments (G_b) to test the hypothesis of equality of expression between treatments.
- The same test is performed within treatments (G_w) to test the influence of replicates in expression.
- An F test, $F = (G_b/df_b)/(G_w/df_w)$ test the hypothesis of equality of variance between and within treatments.
- Combining by conditional probabilities the evidence from these test, we obtain a probability, P_T, that summarizes the evidence of DGE

Transcriptome of chili pepper fruit during development



Martínez-López *et al. BMC Genomics* 2014, **15**:143
http://www.biomedcentral.com/1471-2164/15/143

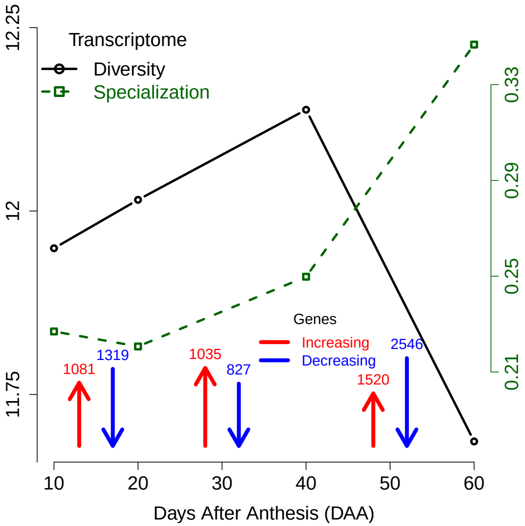
BMC Genomics

RESEARCH ARTICLE **Open Access**

Dynamics of the chili pepper transcriptome during fruit development

Luis A Martínez-López[1,2], Neftalí Ochoa-Alejo[2,3] and Octavio Martínez[1*]

Transcriptome of chili pepper fruit during development



Bias in RNA-Seq: A serious and unsolved problem

- RNA-Seq: Sequencing large number of gene tags from transcriptomes
- Main aim: Detecting differences in the expression of genes depending on treatments
- Problem: Only the *relative* expression of the genes can be estimated
- Proposed Solutions:
- a) Use internal controls (genes assumed to have the same expression)
- b) Use external evidence (qRT-PCR, microarrays, ... ?)

In other ‘omics’...

- Proteomics: One gene produces more than one peptide (same problems than in genomics for identification, quantification and annotation)
- Metabolomics: Thousands of biological compounds are not yet well described (problems for identification, quantification and annotation)
- Nascent fields: Methiloma, interactoma, ...
- In all cases the quantity of data is very large and the availability of methods to analyze them is still in development.

“**From Data to Knowledge**”

Thank you for your attention

http : //computational.biology.langebio.cinvestav.mx/

omartine@langebio.cinvestav.mx