Next-Generation Annotation for Personal Genomics: Gene Fossils & Integrative Models

Mark Gerstein
Yale

Slides at
Lectures.GersteinLab.org
(See Last Slide for References & More Info.)
**THE SEQUENCE EXPLOSION**

- Automated Sanger Sequencing: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

- Third-Generation Sequencing: Companies such as Helicos BioSciences already read sequences from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent, say they can read from longer molecules as they pass through a pore.

- Sequencing by Ligation: This technique employed in SOLiD and Polonator Instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

- The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.6 trillion base pairs.

- 454 Pyrosequencing Released: This technique uses a technique called Pyrosequencing to produce hundreds of base pairs in a single run.


- Navigenics, 23andMe, deCODEme: Companies involved in genetic sequencing and analysis.

- International HapMap Project: A large-scale project to map human genetic variation.
Personal genomes will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.
Importance of Dark Matter of the Genome

- Non-coding regions contain the control elements for coding regions.
- Some non-coding regions are functional & are pervasively transcribed.
- “Molecular Fossils” in the non-coding genome represent a historical record of the genome
- Most disease-associated mutations (e.g. GWAS hits) are in non-coding regions.
How might we annotate a human text?

## The Semicolon Wars

**Brian Hayes**

If you want to be a thorough-going world traveler, you need to learn 6,912 ways to say “Where is the toilet, please?” That’s the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot, you also have quite a challenge ahead of you, learning all the ways to say:

```c
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diarmuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we’ve been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity.

Every programmer knows there is one true programming language. A new one every week.

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they’ll concede, but it’s built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will decide which end of a boiled egg to crack. This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that’s not what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

This modern echo of Swift’s Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, “On holy wars and a plea for peace.” The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, in

```cpp
x := 0; y := x+1; z := 2
```

the semicolons tell the compiler where one statement ends and the next begins. C
Sources of Annotation: Comparative & Functional

Large-scale sequence similarity comparison

Identify large blocks of repeated and deleted sequence:
- Within the human reference genome
- Within the human population
- Between closely related mammalian genomes

Identify smaller-scale repeated blocks using statistical models

Signal processing of raw experimental data:
- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks
Human Genome Analysis: Classic Approach v Future Direction

- Annotating Pseudogenes
  - Identifying Pseudogenes in intergenic DNA & potentially ascribing regulatory function to them
  - ENCODE & 1000G approach to current Human analysis
  - “Junk to Part”

- Modeling Gene Expr.
  - Using His. Mods. to predict gene expression
  - Comparing this with TF binding
ENCODE Production Project

Consortium
Comprises 
~50 Labs

Subprojects:
Transcriptome 
+ Chromatin 
+ TFs
Pseudogenes are among the most interesting intergenic elements

• Formal Properties of Pseudogenes ($\Psi G$)
  – Inheritable
  – Homologous to a functioning element
  – Non-functional
    • No selection pressure so free to accumulate mutations
      – Frameshifts & stops
      – Small Indels
      – Inserted repeats (LINE/Alu)

• What does this mean? no transcription, no translation?…

[Mighell et al. FEBS Letts, 2000]
Identifiable Features of a Pseudogene (ψRPL21)

Genome-wide Annotation of Pseudogenes

PseudoPipe

- Protein Sequence
- Reference Genome

Six-frame blast

- Eliminate redundant hits
- Remove hits overlapping exon

Merge hits and identify parents

FASTA re-alignment

Processed Pseudogenes
Duplicated Pseudogenes

Pseudogene Information Pool

- 18,046 PseudoPipe
- 13,644 RetroFinder
- 11,240 HAVANA

2-way consensus
9,093

~14K total

Δ2-way
1,907
7,186
4,054

Level 1
Level 2

Pseudogene Decoration Resource psiDR

Polymorphic Pseudogenes
24

Surveyed Set
11,216

7,183 level 1
4,033 level 2

1000G
ENCODE

[Pei et al., GenomeBiology ('12, submitted)]
Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Human</th>
<th>Chimp</th>
<th>Mouse</th>
<th>Rat</th>
<th>Chicken</th>
<th>Zebrafish</th>
<th>Pufferfish</th>
<th>Fruitfly</th>
<th>Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>1/0</td>
<td>1/2</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GPI</td>
<td>-</td>
<td>-</td>
<td>1/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PFK</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ALDO</td>
<td>1/1</td>
<td>1/1</td>
<td>11/0</td>
<td>7/0</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TPI</td>
<td>3/0</td>
<td>2/1</td>
<td>6/1</td>
<td>3/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>GAPDH</strong></td>
<td><strong>60/2</strong></td>
<td><strong>47/3</strong></td>
<td><strong>285/46</strong></td>
<td><strong>329/35</strong></td>
<td><strong>0/1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PGK</td>
<td>1/1</td>
<td>1/2</td>
<td>2/0</td>
<td>12/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PGM</td>
<td>12/0</td>
<td>13/1</td>
<td>9/0</td>
<td>3/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ENO</td>
<td>1/0</td>
<td>1/2</td>
<td>12/1</td>
<td>36/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PK</td>
<td>2/0</td>
<td>3/0</td>
<td>10/3</td>
<td>4/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LDH</td>
<td>10/2</td>
<td>9/1</td>
<td>27/7</td>
<td>25/4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>97</td>
<td>91</td>
<td>422</td>
<td>463</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.

Processed/Duplicated

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Human</th>
<th>Chimp</th>
<th>Mouse</th>
<th>Rat</th>
<th>Chicken</th>
<th>Zebrafish</th>
<th>Pufferfish</th>
<th>Fruitfly</th>
<th>Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>1/0</td>
<td>1/2</td>
<td>0/1</td>
<td>-</td>
<td>0/2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GPI</td>
<td>-</td>
<td>-</td>
<td>1/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PFK</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ALDO</td>
<td>1/1</td>
<td>1/1</td>
<td>11/0</td>
<td>7/0</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TPI</td>
<td>3/0</td>
<td>2/1</td>
<td>6/1</td>
<td>3/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GAPDH</td>
<td>47/3</td>
<td>285/46</td>
<td>329/35</td>
<td>0/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PGK</td>
<td>1/1</td>
<td>1/2</td>
<td>2/0</td>
<td>12/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PGM</td>
<td>12/0</td>
<td>13/1</td>
<td>9/0</td>
<td>3/0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ENO</td>
<td>1/0</td>
<td>1/2</td>
<td>12/1</td>
<td>36/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PK</td>
<td>2/0</td>
<td>3/0</td>
<td>10/3</td>
<td>4/1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LDH</td>
<td>10/2</td>
<td>9/1</td>
<td>27/7</td>
<td>25/4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>97</td>
<td>91</td>
<td>422</td>
<td>463</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Distribution of 62 human GAPDH pseudogenes

[Liu et al. BMC Genomics ('09, in press)]
Age calculated based on Kimura-2 parameter model of nucleotide substitution

**Human GAPDH**

**Dating of GAPDH pseudogenes, highlights their creation ~50 Mya**

[Liu et al. BMC Genomics (’09)]

**Burst of Retrotranspositional Activity**
- Unprocessed pseudogenes with no functional counterparts in the same genome

76 Unitary pseudogenes

[Graph showing evolutionary tree and gene nomenclature]

[Zhang et al. Genome Biology (2010)]
11 Polymorphic pseudogenes: Ex of SerpinB11

G_{aa} (E) => T_{aa} (stop)

Freq. of this Mutation in Various Human Populations

[Zhang et al. ('10) GenomeBiology]
LOF variants: creation of “pseudogenes” from annotated genes

[Balasubramanian et al., Genes Dev., '11]
VAT – a robust software for functionally annotating variants

---

Habegger et al. ('12) Bioinformatics
Cloud-enabled (AWS) workflows with VAT

- Operates on 1000G data in the cloud (S3)

[Habegger et al. ('12) Bioinformatics]
### Number of LOF Variants in an Individual

[MacArthur et al., *Science* (‘12, in press)]

<table>
<thead>
<tr>
<th>variant type</th>
<th>total</th>
<th>1000G low-coverage average per individual</th>
<th>NA12878 (high coverage European)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CEU</td>
<td>CHB+JPT</td>
</tr>
<tr>
<td>stop</td>
<td>565</td>
<td>26.2 (5.2)</td>
<td>27.4 (6.9)</td>
</tr>
<tr>
<td>splice</td>
<td>267</td>
<td>11.2 (1.9)</td>
<td>13.2 (2.5)</td>
</tr>
<tr>
<td>frameshift</td>
<td>337</td>
<td>38.2 (9.2)</td>
<td>36.2 (9.0)</td>
</tr>
<tr>
<td>indel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>large deletion</td>
<td>116</td>
<td>28.3 (6.2)</td>
<td>26.7 (5.9)</td>
</tr>
<tr>
<td>total</td>
<td>1285</td>
<td>103.9 (22.5)</td>
<td>103.5 (24.3)</td>
</tr>
</tbody>
</table>

---

**Diagram:**
- **Red:** high-confidence LoF
- **Purple:** candidate LoF
- **Orange:** missense
- **Yellow:** synonymous

**X-axis:** Proportion
**Y-axis:** Proportion

[Graph showing distribution of variant types by proportion.]
Impact of Polymorphic Pseudogenes & LoF Events on Gene Sets

• What should be the reference gene set?
  – Single individual
  – Ancestral individual
  – Current reference
  – Union of genes found in any individual
  – Intersection of genes found in everyone

[Balasubramanian et al., *Genes Dev.*, ’11]
IncRNA: Identification of many candidate ncRNAs through evidence integration

- No single feature (e.g. expr. expts., conservation, or sec. struc.) finds all known ncRNAs => combine features in stat. model
- 90% PPV, 13 of 15 tested validate

Pseudogene Transcription

Total transcribed pseudogene: 863
RTPCR Experimental validated: 57 of 76

Pseudogene: ENSG00000232553.2
Parent: ENSG00000176444.13

Simple & Complex
Ex of Pseudogene
Transcription

[Pei et al., GenomeBiology ('12, submitted)]
Partially Active Chromatin Around Transcribed Pseudogenes
Partial Pseudogene Activity

Transcribed with Additional Activity

Partially Active

Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE

H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE

H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE

Digital DNaseI Hypersensitivity Clusters from ENCODE

Transcription Factor ChIP-seq from ENCODE

Mapability or Uniqueness of Reference Genome from ENCODE

[Pei et al., GenomeBiology ('12, submitted)]
- **Ka/Ks**: conventional measure of selection for genes, shows no signal for pgenes.
- Signature for selection on some SD pgenes (16%), derived from intersecting with UW SD DB & looking for differential conservation of neighborhood vs. center of pgene.
- Weak signature for greater selection on transcribed pgenes using 1000G polymorphisms.

### Signatures of Selection on Some Pseudogenes

**Comparison of nucleotide substitutions per site (N) in pgenes and SDs**

- **Negative selection**: $N_{(p gene)} < N_{(SD)}$
- **Neutral evolution**: $N_{(p gene)} \approx N_{(SD)}$
- **Positive selection**: $N_{(p gene)} > N_{(SD)}$

- **4** negative selection
- **139** neutral evolution
- **24** positive selection

---

**Figure Description**

- Segmental duplication:
  - Parent gene
  - Duplicated pgene

- Comparison of nucleotide substitutions per site (N) in pgenes and SDs:
  - $N_{(p gene)} < N_{(SD)}$
  - $N_{(p gene)} \approx N_{(SD)}$
  - $N_{(p gene)} > N_{(SD)}$

- **Mu et al., NAR 39: 7058**
Partial Activity for ~2.5K Pseudogenes

[Pei et al., GenomeBiology ('12, submitted)]
Examples & speculation on the function of pseudogene ncRNAs:

Regulating their parents

- via acting as endo-siRNAs [Recent ex. in fly & mouse, '08 refs.]
- via acting as miRNA decoys [PTEN]
- via inhibiting degradation of parent’s mRNA [makorin]

Alternatively, just last gasps of a dying gene

Poliseno et al. Nature 465:1033 ('10)
Concepts:

Raw Tracks to Networks & Relating Genomic Inputs to Outputs

Signal processing of raw experimental data:
- Removing artefacts
- Normalization
- Window smoothing

Segmentation of processed data into active regions:
- Binding sites
- Transcriptionally active regions

Group active regions into larger annotation blocks

Further analysis: Building regulatory networks

Inputs (mostly Chip-seq of TFs & Chromatin)

Outputs (mostly RNA-seq)


[Science 330:6012]
Modeling Transcription: Connecting Inputs & Outputs

- Models connecting various types of high-level data
  - HMs+TFs => gene expression
  - HMs => TFs

Histone Modification (HM) model

Chromatin features: Histone modifications

Predictors

RNA-Seq data

Prediction target: Gene expression level

HM1, 2, 3, …

Bin 1
Bin 2
Bin 160

[Cheng et al. ('11) Genome Biol. 12; R15]
His. mods around TSS & TTS are clearly related to level of gene expression, in a position-dependent fashion.

H3K4me2

[Science 330:6012] [Related work: Ouyang et al. ('09) PNAS; Karlic et al. ('10) PNAS]
Integrate all histone modifications to predict gene expression levels

Classify H/L genes (SVM)

Predict expression values

Magnitude of Prediction from a "bin" around the TSS

Cheng et al. ('11) Genome Biol. 12:R15
HM models are tissue specific

Best prediction is achieved by using histone modification and expression data from the same developmental stage.

[Cheng et al. (’11) Genome Biol. 12: R15]
Scale up to Human
Application of chromatin model in 5 species: Consistent Performance

>50% of variation of expression levels can be explained by HMs
TFs => Expr
Doing a Model with TFs:
Positive and negative regulators from correlating TF signal at TSS with gene expression

[Cheng et al. (’11) PLOS CB]
Predictor v2: 2-levels, now with TFs

[Cheng et al. NAR ('11)]
Mouse ESC Models Illuminates Different Regions of Influence for TFs vs HMs

- Datasets
  - ChIP-Seq for 12 TFs (Chen et al. 2008)
  - ChIP-Seq for 7 HMs (Meissner et al.'08; Mikkelsen et al. ’07)
  - RNA-Seq (Cloonan et al. 2008)

A TF+HM model that combine TF and HM features does NOT improve accuracy!

A Model with only a Few of the Thousands of Mouse TFs is able to Predict Well

[Cheng et al. NAR ('11, in press)]
Scale up to Human: TFs

CAGE PolyA+ K562 Whole Cell

Pearson’s $r=0.81$; RMSE=2.57
Classification: AUC = 0.89
Regression: $r = 0.62$; RMSE = 3.06

Different types of TFs have different correlation with gene expression

Prediction of Differential Expression

TF model– differential TF binding signals are predictive of differential expression levels between two human cell lines

TFs active in a particular cell type are the strongest predictors in that cell

Functionally Interpreting Personal Genomes through Annotation
A Traditional Annotation Story: Human Pseudogenes

- Converting what was considered “junk” into potentially functional elements on the genomic “parts” list
- Large scale assignment of pgenes (~11K hi qual. of ~14K tot.)
  - Polymorphic pgenes & ~100 LOF events per person call into question ref. gene set
- Overlap with functional annotation
  - Many pgenes are transcribed ncRNAs (~900)
  - Many are under selection but not as proteins
  - Partial activity be signature of a dying gene or of a regulatory ncRNA
- Pseudogene.org & VAT.gersteinlab.org
Models of Transcription Relating Various Genome Tracks of Information – His. Mods, TF Binding & Gene Expression

• Predictive models of gene expression
  – Applicable in many contexts
  – Work for miRNAs as well as genes
  – Show variable importance of regions around genes for chromatin & TFs
  – Show TF & HM signals are redundant for ‘predicting’ gene expression
  – Surprisingly, a few TFs, particularly TFSSs, are quite predictive
Acknowledgements

ENCODE, 1000G

GENCODE (Not Linked): B Pei, C Sisu, D Chen
Transcription Models

Acknowledgements

TF-v-expr:

worm-HM:

encode:
Chao Cheng, Roger Alexander, Renqiang Min, Kevin Y. Yip, Jing Leng, Joel Rozowsky, Koon-kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie A Davis, Piero Carninci, Timo Lassman, Thomas R. Gingeras, Roderic Guigó Serra, Ewan Birney, Zhiping Weng, Michael Snyder
Default Theme

• Default Outline Level 1
  – Level 2
Info about content in this slide pack

• Some of the material in this presentation is unpublished, marked by in press or submitted citations on various slides. Please treat it as confidential material and do not redistribute.

• PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2010. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

• Paper references in the talk were mostly from Papers.GersteinLab.org.

• PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as kwpotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt