The Gene Concept

Sonja Prohaska

Computational EvoDevo
Universitaet Leipzig

January 3, 2015

What is a gene?

"I can't tell but I recognize a gene when I see one."
a biologist

"Something is a gene when a biologist says it is one."
a bioinformatician

"A gene is a database entry with an Ensembl gene ID."
a computer scientist

"A gene is what Wikipedia says it is."
a student

"A gene is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions." Wikipedia

Historical view – really short

In the beginning...

- a *phenotype* has characteristics
- some characteristics are independent
- some characteristics are heritable
- all heritable characteristics need to go through a single cell (gamete)

How to put (all) characteristics of a phenotype into a gamete?

- miniature organism within gamete?
- gemmule, shed by the organs accumulated in gametes? (Darwin 1868)
- distinct, discrete entities that specify characteristics (Mendel 1866)

"special conditions, foundations and determiners which are present [in the gametes] in unique, separate and thereby independent ways [by which] many characteristics of the organism are specified" by Johannsen (1909)

... the **gene** is a (unknown) substance **representing a characteristic**.

Historical view – really short

linkage of genes

- Morgan (1915)
- segregation experiments and crossbreeding
- the observed linkage of genes best fitted a model of a linear arrangement
- size of genes and distance between genes could be inferred
- the model had predictive power in breeding

How did this change the understanding of a gene?

- genes are continuous
- genes are nonoverlapping
- distinct genes have distinct dimensions
- genes are linked to verying degrees

A gene is an abstract entity whose existance is reflected in the way a phenotypeis transmitted between generations.

Historical view – really short

- **1941** Beadle and Tatum: *"one gene, one enzyme"*
  The gene is the information behind the individual molecule.
- **1955** Hershey and Chase: the substance for genes is DNA
- **1955** Benzer: a cistron (gene) is a region of DNA defined by mutations that in *trans* could not genetically complement each other.
- **1953** Watson and Crick: how DNA could function as a molecule of heredity
- **1958** Crick: flow of information from DNA $\rightarrow$ RNA $\rightarrow$ protein
- **1970 – 1980** Fiers: RNA and DNA sequencing
- understanding of how genes are expressed, discovery of splicing
- development of computational tools
- **the "nominal gene"** is defined by its **predicted sequence** rather than a genetic locus
- **1986** the gene effectively became identified as an annotated ORF

pre-ENCODE: the birth of the structural gene

a gene is...

"... a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology." Human Genome Nomenclature Organization

"... a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions" Sequence Ontology Consortium

" ... the entire nucleic acid sequence that is necessary for the synthesis of a functional polypeptide (or RNA)" by Lodish (2000)

Problematic issues with the gene concept

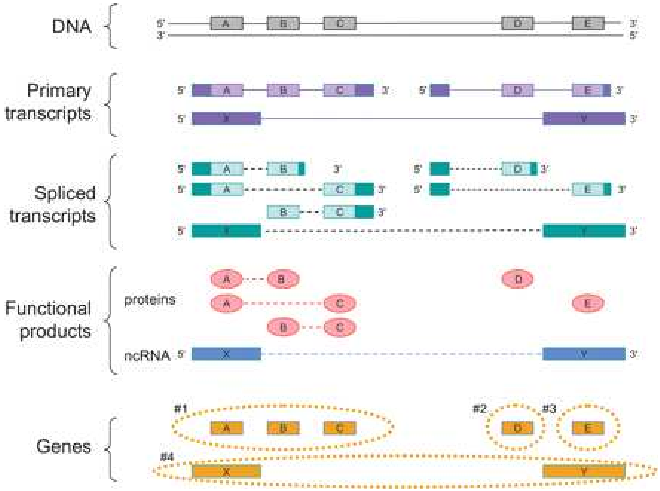
- **regulatory sequence**: part of a gene or associated with a gene?
- **overlapping genes**: same strand different reading frame or readingframes on opposite strands
- **splicing**: open reading frame is segmented
- **alternative splicing**: multiple different transcripts with different function
- **trans-splicing**: distinct transcripts can be joint
  the gene as a single locus no longer applies
- **run-through transcripts and fusion proteins**
- **parasitic and mobile elements**

A gene is a set of connected transcripts where "connected" means sharing of exons.

How ENCODE ruined/challenged the gene concept

- functional non-coding RNAs
- unannotated transcription: only 50% of spliced transcripts are annotated
- transcription from (distal) alternative transcription start sites (TSS)
- alternative 3'UTRs
- transcription at regulatory elements
- dispersed regulation and elements (upstream, downstream, within the first exon, within the first intron, anywhere else)
- blurring of the destinction between genic and intergenic, exonic and intronic
- act of transcription of functional importance, transcript irrelevant
- pseudogenes
- highly conserved elements, only 20% in annotated regions

The Gerstein-Snyder gene definition



The Gerstein-Snyder gene definition – in words

- a gene is a genomic sequence on DNA (or RNA)
- it encodes (**one or many**) functional product molecules (RNA or ppotein)
- functional products sharing overlapping genomic regions are **united**
- the union must be **coherent**
- i.e. union built separately for RNA and protein products, plus and minus
- does **not** require that all products necessarily share a common subsequence

  Example: Three functional protein products built from genomic elements A,B,C: A+B, A+C, C only belong to the same gene even though A+B and C only do not share a common subsequence.

  Notice: sharing of UTRs or regulatory regions is not sufficient (see D,E).

"The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products."

The Gerstein-Snyder gene definition – problemes

- a container term
- “a genomic region” versus “an ordered set of genomic sequences”
  - region = intervall $[x_1, x_2]$ where $x_1 <= x_2$
  - what the authors mean: a gene is “an set of genomic sequences”
- “gene” = concatenation of the “oriented and ordered set of genomic sequences”
  - results in a sequence that **does not exist** in the genome as such (hint: introns)
  - conceptual translation of the “gene” does not necessarely result in an **existing** functional product (example: A+B+C does not exist)
- “overlapping” versus “sequence in common”
  - one genomic region but two unrelated protein sequences due to frame-shifted ORFs

Is everything that makes a functional gene product encoded in the gene?

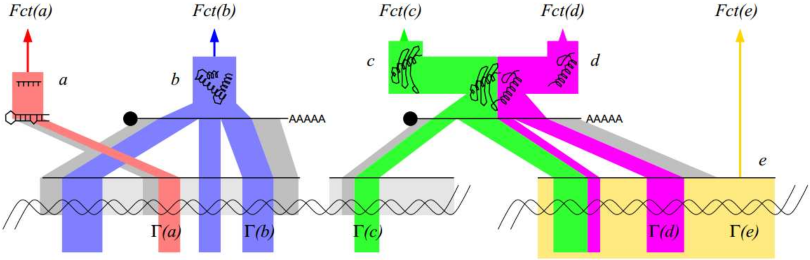
For proteins this would mean:

- each aa can be mapped onto a nucleotide triplet/codon on the DNA
- key: genetic code
- these triplets might be parted in two (introns)
- and put together by gene expression
- no addition or modification of amino acids
- counter examples: selenoproteins (stop codon UGA is mapped onto selenocystein in the presence of SECIS), cleavage, deamination, deimination, racemization,...

For RNAs this means:

- each RNA nucleotide can be mapped to a single continuous locus
- key: transcription
- may counter examples: splicing, polyadenylation, cleavage, ligation, poly-adenylation, CCA-addition, pseudouridinylation,...

The Stadler-Prohaska gene definition



in red ... function *Fct*(*a*) of miRNA *a* is inhibition of translation of a particular set of mRNA → miRNA *a* → derived from its precursor hairpin → ... → genomic footprint Γ(*a*) of miRNA *a*); *b* – classic eukaryotic protein Γ(*b*)*isidenticaltotheCDSsof* b;c,d− − *proteolyticallycleavedproteinsfromatrans* − *splicedmRNA*;e− − *functionalprimarytranscript*;

The Stadler-Prohaska gene definition

- a function $Fct(a)$ is carried out by a biomolecule a
- project the sequence of the molecule a down onto the original genomic sequence from which it was derived
- projection rules are specified by conceptual revers gene expression
- from protein to RNA: genetic code
- from RNA to DNA: error-free transcription
- result: genomic footprint $\Gamma(a)$ of the functional biomolecule

A gene represents the duality of a **functional product** and its **genomic footprint**.

Literature

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S and Snyder M (2007). *What is a gene, post-ENCODE? History and updated definition.* Genome Res. 17:669-681

Griffiths PE (2002). *Lost: One Gene Concept. Reward to Finder.* Biology and Philosophy 17:271-283

Prohaska SJ and Stadler PF (2008) *"Genes"* Theory Biosci. 127: 215-221

Stadler PF, Prohaska SJ, Forst CV and Krakauer DC (2009). *Defining genes: a computational framework.* Theory in Biosciences 128:165-170

Engelhardt, Kirsten T, Stadler PF and Prohaska S (2010). *Genome Annotation without Genes.* Technical report.