



# Cancer Genome Anatomy Project (CGAP) Data

Carl F. Schaefer  
March 12, 2004



## Overview

- ❖ Genes & EST libraries
- ❖ Mammalian Gene Collection (MGC)
- ❖ SAGE
- ❖ Gene Ontology (& prot. xrefs)
- ❖ Pathway
- ❖ Chromosome aberrations  
(Mitelman)



# Genes & EST Libs -- Data Sources

- ◆ "hierarchy.txt" – NCBI keyword hierarchy
- ◆ "library.report"
  - custom dump of EST library info from NCBI
- ◆ UniGene
  - Hs.data, Mm.data
  - Hs.seq.all, Mm.seq.all
  - Hs.lib.info, Mm.lib.info (reconcile naming problems w/ library.report)
- ◆ LocusLink (LL\_tmpl)
- ◆ HomoloGene (hmlg.ftp [uh-uh... now XML])



# Genes & EST Libs -- Processing

## lib info

- reconcile name problems
- propagate redundant keywords
  - ◆ E.g. "cerebrum" implies "brain"
- determine unique tissue
  - ◆ E.g. pool of cerebrum & cerebellum implies "brain"



# Genes & EST Libs -- Processing

## ◆ gene info

- pick representative seqs (RefSeq, longest mRNA, ...)
- build UG/LL correspondence (in theory, simple!)
- build (a) UG/seq, (b) UG/library, (c) UG/tissue/hist freq, (d) UG/tissue, (e) UG/alias relations
- build Hs/Mm, Mm/Hs ortholog relation
- build abbreviated UG blast libraries



## MGC -- Inputs

- ◆ Custom dumps from NCBI
  - “Trace.Info” (1 M EST/trace id)
  - “Picked.Clones” (95 K selected clones)
- ◆ UG data for other orgs (Rn, Dr, XI, Str)
- ◆ GenBank for sequences, def lines, cds
- ◆ Plate info from LLNL (IMAGE)
- ◆ Trace files from Incyte and Agencourt



## MGC -- Outputs

- ◆ Relate IMAGE clone id to:
  - MGC id, accession
  - Status
  - UG, LL, etc.
- ◆ MGC-only blast libraries
- ◆ Dumps of traces by plate for full-length sequencers
- ◆ Finding identical CDS, 5'UTR, 3'UTR



## SAGE

- ◆ “sageedit” system for lib info
- ◆ manual load of new of tag/lib/freq data
- ◆ occasional rebuilding of ranked tag-to-acc map
- ◆ auto rebuilding of best-tag-for-gene (kicked off by new UG data; uses tag-to-acc map)
- ◆ future complications:
  - Mm data
  - “long” sage



## Gene Ontology (GO) -- Inputs

- ◆ go\_termdb.xml
- ◆ gene\_association.goa\_sptr
- ◆ nr (NCBI) (for 100% identities)
- ◆ hum.dat, rod.dat (SP/TrEMBL)
- ◆ embltosp.txt (mRNA link SP/LL)
- ◆ NREF.xml (PIR) (SP/Tr/NP/GenPept)
- ◆ LL\_tmpl (NCBI)



# Gene Ontology (GO) -- Outputs

- ◆ go\_name

- (go id, name, class)

- ◆ go\_parent

- (child go id, parent go id, relation)

- ◆ go\_ancestor

- (child go id, ancestor go id)

- ◆ ll\_go

- (locus id, organism, go id, evidence)



## Pathway

- ◆ Periodic download of pathways from KEGG, BioCarta
- ◆ Associate enzymes/genes with LocusLink ids
  - KEGG (automated)
  - BioCarta (partially automated)
- ◆ Newer pathway stuff is far more complicated



## Chromosome Aberrations

- ◆ Original db design from Lund, little changed
- ◆ Quarterly delivery of data tables from U. Lund
- ◆ Run scripts to update HTML select lists
- ◆ Associate new gene symbols with LL ids
- ◆ Provide dumps to NCBI (LocusLink; SKY/CGH)



# Mechanics

- ◆ cron kicks off make
- ◆ make does most of it: from ftp to load
- ◆ preprocessing with perl and Unix utilities  
(egrep, cut, join, sort)
- ◆ Oracle: drop idx; trunc; sqllldr; re-idx; alz
- ◆ Schema switch
  - auto load new data into schema cgap2
  - switch production from cgap → cgap2
  - load new data into cgap
  - switch production from cgap2 → cgap