

Genomics

Tom Petty
latex: Dan Simola

21 January 2005

Overview

- GCB 531 ala Harold Riethman - Sequence Analysis
- Genome Sequencing / Databases
- Experimental organisms and humans

Aims of Genome Science

- integrated data base and research interface
- physical and genetic maps
- annotate set of genes
- atlases of gene expression
- functional data - biochemical, phenotypic
- sequence diversity
- comparative genomics

Brief Review

1. Crazy fast really brief overview of things
2. ~3 Billion bases in human genome
 - Should that be made public?
 - New ethical and legal issues to consider, new regulations needed, like human cloning, etc
3. 1st physical map done in drosophila
 - 1972: splice DNA into plasmid > bacteria used to Xerox DNA
 - 1989: this allowed first genetic linkage studies between people
4. 1980s were critical to genome science
 - Changed the way everyone thought about DNA
5. Hard to separate 40 kb molecules (or bigger)
 - Pulsed-field electrophoresis has directional electric charge
 - Now possible to increase distance you could analyze DNA
 - Allowed entire genome mapping

6. Mapping (in general):

- Look at multiple markers to build genetic maps, Much better than older, pairwise approaches
- Study the effects of radiation on DNA
 - 1st proposed by US Department of Energy (DOE)
- YACS (Yeast Artificial Chromosomes)
 - Advanced cloneable size from 40,000 to 1000 Kb (1 Mb)
 - So now only ~3,000 molecules end to end to reconstruct genome
- NIH convinced congress to make genome public

7. Model organisms:

- Needed sequences of model organisms to serve as test bed for new technology
- Important, even though sequences were not human

8. By 1994, 1 cM map $\dot{\iota}$ 52,000 STSs mapped

- Human variation sequence SNPs added in 1998
- Addition of ESTs (Expresses sequence tags)
 - Using cDNA made from mRNA as probes

Maps

Characterization

- Landmarks
 - DNA probes
 - STS (Sequence Tagged Site) STS markers are helpful in integrating Physical, genetic and RH maps.
STS (1989): Replaces cloned DNA probe mapping with PCR Each STS unique in that: Pair of oligonucleotides (primers) Product size PCR reaction conditions Stored, distributed electronically
- Fragment Ends
 - Restriction Enzyme
 - Meiosis: Mapping meiotic breaks in chromosomes
 - Radiation
 - * Radiation Hybrid Map (X-ray breaks)
 - * Donor cells (Human Chroms), Recipient Hamster Cells
 - * Combine them, rescue by fusion and selection
 - * Unfused human cells lethally irradiated, check linked markers on living cells
 - Clone ends

Types

- Genetic Map
 - defined by relative order of genetic markers in linkage groups
 - distance defined by centiMorgan (cM), units of recombination
 - * definition: percent of progeny in which a recombination event has occurred between two markers
 - * 1cM \approx 0.01 recombination frequency (rf), about 1000kb in human, 500k in fly
 - * markers across chromosomes have a 1/2 chance of co-segregating \implies 50cM distance
 - * problem of interference - one crossover reduces likelihood of another crossover nearby
 - fully connected, 2-5 centiMorgan resolution, each marker identified by STS markers

- markers often physical, such as repeats, restriction cut sites, sequence tags
- RH maps often used with recombinant inbred lines (see above)
- Physical Map
 - assembly of contiguous stretches of native genomic DNA (contigs)
 - distance units are kilobases
 - Clone-based (contig), made from a large library of cloned fragments
 - localized and oriented relative to the radiation hybrid and genetic linkage maps
 - two ways to assemble
 - * alignment by restriction fragment profiles - similar profiles indicate neighboring clones
 - * hybridization approaches
 - use common probes to identify contiguous clones
 - iteration of this is called chromosome walking
 - STS markers are used as hybridization probes
 - 100 kb resolution STS map
 - 2 Mb contigs for most of genome
 - * Cytogenetic: 2 Mb res
 - * Long-range restriction maps 10 Mb 40 kb res
 - * Interphase FISH 500 kb 50 kb
- Cytological Map
 - defined by banding patterns of stained chromosomes
 - aligns to physical map using in situ hybridization of clones
- Goal is to merge mapping data obtained from different labs using three mapping methods into a single consensus map of landmarks along the chromosomes

Human Genome Project

- first 5-year plan (1991-1995)
- genetic map, 2-5cM
- physical map, 100kb complete STS map
- Sequencing
 - Improve current technology to \sim \$0.50 per base
 - Generate total of 10 Mb human sequence in large stretches
- Model Organisms
 - Genetic map of mouse
 - Generate 20 Mb from models, focus on continuous stretches of 1 Mb
- Informatics
 - Develop software and database designs to support large scale sequencing mapping
 - Create Dbase tools for easy access to latest data
 - Develop algorithms and analytical tools to interpret info
- Nice point (Im sure it wont be tested but) 3-5% of the entire budget was dedicated to ethics. Take home: be a good scientist when you grow up
- Added Gene ID after the first 5-year plan by 1993 realized importance of finding the functionality of genome

Nuclear Biology

Bio 101

- Key difference: Meiosis, homologous chromosomes have crossing-over events
- Crossing over occurs at prophase in meiosis
- Mitosis, no crossing over!!!
- Phases
 - G1: Growth 1 Cell grows
 - S: synthesis (DNA duplication)
 - G2: Growth 2, fattening up getting ready to split to 2 cells
 - M: Mitosis splitting of cell into two.
- Meiosis
 - Prophase I
 - * Each homolog lines up form bridges
 - * This is where crossing over occurs critical point in genetic mapping
 - Metaphase I
 - * Chromosomes pulled apart
 - Finish up with gametes that contain 1 copy of each chromosome

Chromosomes

- Naked DNA wraps around histones ζ forms extended nucleosomes ζ then condensed nucleosomes
- Interphase: DNA in fibers or extended nucleosome
- Chromosomes are only compact during mitosis, meiosis
- metaphase chromosomes
 - Human Karyotypes made from these Light and dark bands correspond to GC rich and poor (few Mb resolution)
 - Microtubules attach to centromeres during division

Euchromatin

- Human Genome Project focuses on this stuff!
- Also has tandem repeats 100 to 10s of Kb in size
- Mini-satellite, each block of 20-200 bases repeated over and over
Length of repeats highly variable among humans
- Genes form a small fraction of euchromatin, most of this is non-coding

Heterochromatin

- bright staining, large arrays of tandem repeats, difficult to clone and sequence
- Still, extremely important for stability and replication
- Highly compact, yet influences other features
 - Rate of replication
 - Gene transcription
 - Folds differently than Euchromatin
 - Recombination typically suppressed near regions of heterochromatin

3 main chromosomal elements

- Telomeres
 - prevent chromosomes from fusing with each other
 - ~220 bp in yeast, necessary and sufficient to allow DNA to be properly segregated and maintained via meiosis and mitosis
 - human centromeres span Mbs not as highly conserved in yeast
 - Simple repeat element different than centromeres, G-rich
 - Maintained by Ribonucleoprotein - RNA serves as telomere sequence
 - Low gene-copy number
 - Single-stranded overhangs involved in folding into loop
 - Regulatory function:
 - This loop can bind various proteins and also regulate cell separation and ability of the cell to continue to replicate involved in lots of cellular systems
- Centromeres
 - allow proteins to bind spindle and segregate chromosomes
 - 95-97% similarity of all repeats, so difficult to map these regions
 - Euchromatic-like DNA also found in centromeres
- Replication Origins
 - specific sequences distributed widely on each chrom
 - but only a subset of these origins used at a given time
- Miscellany
 - Micro-satellites - Only 1,2,3 bases influence transcription of adjacent genes
 - Sources of variability are due to repeats, recombination occurring in tandem arrays
 - G-rich DNA can fold up within itself, 4 Gs can form a planar structure, a very stable conformation

Yeast modeling

- Using Yeast to study features of the Human Genome
- Origins of replication
 - 200-400 in the yeast genome
 - most are intergenic
 - Only a subset chosen for use in any given S phase
- Can biochemically detect origins on 2D gels (look this up)
- Most origins fire mid-S phase
 - Origins fire at similar times the closer they are to each other
- No correlation with firing time and AT versus GC richness
- Firing Bias, fire early near centromere, late near telomere
- Lots of different replication rates
- Take home: with complete yeast genome and a few tools, got all new information that is easy to use and gives nice detail about chromosome activities

Linkage mapping

- accuracy decreases with distance - best to take large number of closely linked markers for higher accuracy
- Easier to study in model organisms such as yeast, etc, than in humans, because of the following:
 - Cant control matings
 - No uniform genetic context (no inbred strains...)
 - Need to collect families and look at inheritance of markers
 - * BUT this is LIMITED to MEIOTIC amounts (very few)
 - * Reduces the precision of maps
 - Combine data from many families until you are confident that a pair of markers is really linked
 - Out of 20 meioses, expect around 10 recombination events

LOD score analysis - assessment of linkage

- LOD score
 - Are events likely to be from linkage or chance?
 - Likelihood of observed data / likelihood of data given by chance λ converted to log scale
 - Odds $> 1000:1$ is needed (lod of 3), to conclude significant evidence for linkage
 - Continuously add families, to hopefully increase LOD score until you eventually reach lod of 3 (or 10^3 or $1000:1$)
- CEPH: Collected large numbers of families in order to immortalize lines to have many meiosis
- In humans, many meiosis dont distinguish between homologous chromosomes, how do we know theyre recombined?
 - Big assumption that you can tell difference of each homolog, not the case in real life
 - Polymorphic markers are key, because one can tell apart a locus in one chromosome from a locus at same location on different chromosomes
- RFLP (Restriction Fragment Length Polymorphism)
 - Look at where a hybridized probe is located different in all people, but can also help visualize related due to some pattern overlap (common “DNA fingerprint” technique)
- Genetic Linkage Markers
 - STRP: Short Tandem Repeats - Do PCR on these repeats, use length of repeats to compare
 - STSs: Sequence Tagged Site - To mark positions of sequences along a chromosome
- Markers used average 77% heterozygosity.
- Most heterozygous alleles must be compared separately, depends what youre looking at. Tandem repeats? SNPs? Etc.
- percent heterozygosity changes locus to locus.
- Can come up with a continuous map via statistics and computational approaches that consider pairwise LODs of markers along a chromosome.

Linkage maps

- Statistical error involved, including experimental error in determining markers
- Will more seriously affect the linkage maps

Radiation Hybrid maps

- Instead of meiotic recombination
 - Make breaks via X-rays
 - After cell fusion step, look at how frequently markers are associated.
 - Compare this to what you would expect randomly
- Retention Bias near centromeres tend to be retained more often in hamster cells
- Increase Radiation = more breaks > smaller fragment size > increased resolution
- Significant differences in male and female meiosis - vary across chromosome
 - Females: all eggs ever made already exist- arrested in Prohpase I
 - * One egg at a time triggered to complete meiosis
 - Male: constantly changing

Clone-based physical maps

- Clone Library Preparation
- Clone Overlap Detection and Contig Construction
- Contig placement along chromosome
- Different cloning systems have different properties, eg cosmids are good for yeast and elegans, not for human
- Cloning Systems
 - YAC Yeast Artificial Chromosome (highest capacity) 200 kb 2 Mb
 - BAC Bacterial artificial chromosome, 150 kb 250 kb
 - Cosmid/Fosmid 40 kb
- How many clones would we need?
 - For 1X coverage (end to end no overlaps)
 - 3,300,000 1 Kb Yac clones
 - 33,000,000 for 10X coverage
- YAC
 - DNA must be intact chromosomal DNA lightly digested to segment
 - After digest, use pulsed-field electrophoresis to select proper-sized inserts for library
 - * Has central centromere, telomeres, ori of rep, etc.
 - * Array yeast via micro-well plates, then lyse cells and fix dna on a filter
 - * Then screen via PCR to ID clones that you want
 - * If you have markers, can screen libraries with multiple markers,
 - Then ID clones with each marker
 - Can generate a clone-based STS map
 - No precise physical distance but get an ordered map
 - * Problems: Chimeric YACs (genome inserts from two genomes ligated together)
 - Advantages:
 - * Very large clones
 - * Excellent contiguity of maps
 - * Large genes cloned on single fragment
 - Disadvantages:

- * Chimera and deletions common (30-50%!!!!)
- * Cannot easily purify human from yeast DNA
- * Hard to use directly for DNA sequencing
- BAC
 - Better for high-throughput applications
 - F1 plasmid based, maintained in single copy
 - BAC really a misnomer, really just a BIG plasmid with replication origin
 - Screen libraries in similar manner to YACs
 - * Create restriction endonuclease map
 - * Can automate capture of these RE profiles for each clone
 - * Able to connect overlapping clones, but needs user interaction
 - Now take BAC clones and hybridize to meiotic chromosomes
 - BAC Advantages:
 - * Chimer and deletions rare
 - * Easily purify away from bacterial DNA
 - * Use directly in DNA sequencing
 - * Easy to fingerprint and end-sequence
 - BAC Disadvantages:
 - * Much smaller than YACs
 - * Less contiguity and coverage
 - * Large genes not cloned on single fragment
- Physical Map:
 - Instead of starting with markers, took entire library and fingerprinted all
 - Let computers try to sort out contigs
 - Worked quite well! But still required manual editing

Assembly

Automatic BAC map assembly

- Edit fingerprint data
- Remove bands ≤ 600 bp
- Collapse multiplets into single band
- Optimize overlap and tolerance params
- Overlap about 75%, tolerance ~ 0.7 mm

Singleton clones

- Didn't match any other patterns in database
 - Vectors that lost insert, poor dna quality, etc
 - Regions that don't clone well have high amount of singleton clones

Achieving map continuity

- Manual editing of contigs
- New contig ends assembled with 50% overlap, but checked for consistency with STS maps
- Draft sequence used to detect overlaps

Integration with other maps All of these should agree well

- STS maps on chromosomes
- RH and Genetic linkage maps
- YAC-based STS maps
- FISH-mapped BAC clones
- Other regional BAC clones

Problems

- Duplications Pericentromeric and Subtelomeric regions
- Heterochromatin regions
- Gaps nature unknown (no clones to fill in the holes in the maps)

Review

- STS based on sequence, need sequence first - sequence defines the STS!!
- Go to NCBI website, ePCR and put in sequence of choice, it will return all STSs
 - STS should produce a SINGLE band and be SPECIFIC for a chromosome
- If present on more than one chromosome, its no good
- STS maps were important for a lot of the groundwork of HGP
 - STS on same locus in different people.
 - So, if STS flanks some tandem repeats, STS can change in size
These size variations are critical, allows to distinguish homologs in people to see who they inherited it from
- Were all 99.9% similar, only about 0.1% variation between human genomes. So single locus can be found via PCR assay or hybridization.
 - If STS amplifies as polymorphism, it can be a genetic marker
 - If it is the same, then a physical marker
- Key point to all this is looking at the ORDER of genetic markers, how markers are defined on a given chromosome. Positions determined by patterns of meiotic recombination frequencies.
- On RH maps, break apart human chromosomes, make a human-rodent line, and map connected markers via frequencies that they appear together in same cell.
- STS maps central to genetic mapping

Large-scale DNA sequencing

History

- First 5-6 years of HGP dedicated to improving technology
- Landmark: Smith, automated sequencing
- Fred Sanger developed in mid 70's
 - start the reaction with primer, polymerase
 - mix in ddNTPs to terminate reaction randomly

- dye-primer: has separate reactions per dye
- dye terminator -all in one
- First 10 years were big, old-school polyacrylamide slab gels (1986-1998)
- Labor intensive = slow data generation, 4-12 hour runs, etc
- Capillary DNA sequencers (circa 1998)
 - More automated, smaller reagent volumes, better sensitivity, cheaper
 - Directly coupled to thermal cyclers doing the sequencing reactions
 - 2- hour runs
 - Much more cost-effective, less labor involved, etc.
- Read length of sequences from 100-400 High quality base calls

Phred, by Phil Green

- Want a way to look at how good sequence is at any particular position
- So, this is a way of attaching quality values to data
- Locate predicted peaks (Fourier methods to best fit distribution)
- Locate observed peaks, for which area under peak exceeds 10
- Match observed and predicted peaks using three-stage shifting algo
- Find missing peaks
- Assess error probs of each peak according to four-parameter model
- (So this method also takes into account the spacing of the peaks)
- Phred error probabilities came from empirical investigation of raw and sequence data
 1. Variance in peak spacing over 7-peak window centered on base
 2. ratio of largest uncalled peak to smallest called peak in window
 3. ratio of largest uncalled peak to smallest called peak in 3-peak window
 4. number of bases between the current base and the nearest unresolved one
- Phred q-values
 - Base-by-base measure of sequence quality: $q = -10\log(p)$
 - where p = estimated error probability for that base-call
 - 1/100 prob of incorrect base call: q = 20
 - 1/1000 prob of incorrect base call: q = 30
 - Higher q value = Better Quality
 - q = 40 required to call a sequence “finished” in HGP
- Software that was used mainly in HGP
 - Phred: Base Call
 - Phrap: Assembly
 - Consed: Editing
 - Bottom line is that small insert sequencing → Highly Automated

Templates for sequencing

Types

- M13: Single-stranded, easy to prepare shotgun libraries and sequence
- Plasmid: double-stranded, can obtain reads from both ends of insert, “double-barrel shotgun”
- cDNA: plasmid containing partial sequence of mRNA, EST
- BAC: more difficult and expensive to obtain direct sequence
- End-seqs extremely variable for connecting BAC contigs to chromosomes

Preparation of Templates and Sequencing Reactions highly automated

1. Growth of clones
2. colonies picked robotically, 96-well format
3. Purification of DNA
4. DNA sequencing reactions
5. Everything performed robotically
6. Primers still expensive however. Used towards the end of HGP to “primer walk” to close gaps in the maps

Shotgun Cloning BACterial

1. Purify BAC DNA, mechanically shear to 1-2 kb
2. Clone fragments with plasmid of M13, transform E. coli
3. Robotically pick colonies, array in 96-well plates for growth
4. Purify template DNA in 96-well format

Sequencing BAC

- At each step, sequences deposited into GenBank with accession number
- Phase 0: 96 reactions, sample sequences, overlap detection
- Phase 1: 3-5x coverage, assembled draft
- Phase 2: 8-10x coverage, high-quality draft assemblies
- Phase 3: finished sequence, q-vals ≥ 40 , no gaps in sequence

Two sequencing approaches (see also sequencing review)

- Hierarchical Sequencing
 - Chromosomes
 - Generate and align large BAC or P1 clones
 - Fragment and sequence a subset of the clones
- Shotgun sequencing
 - Chromosomes
 - Fragment and sequence entire genome
 - Put all the pieces back together...