Comparative genomics

Luce Skrabanek
ICB, WMC
25 April 2013
What does it encompass?

- **Genome conservation**
  - transfer knowledge gained from model organisms to non-model organisms

- **Genome evolution**
  - understand how genomes change over time in order to identify evolutionary processes and constraints

- **Genome variation**
  - understand how genomes vary within a species to identify genes central to particular processes
Main uses

• Whole genome comparisons
  – Genome evolution

• Coding regions comparisons
  – Gene prediction
  – Gene structure (exon-intron) prediction
  – Function prediction

• Non-coding region comparisons
  – Regulatory region discovery, siRNAs

• Protein-protein interaction prediction
Genome alignment

• Very different from single gene or protein alignments
• Standard DPAs are too expensive
• Made complicated by extensive rearrangements of large homologous segments, as well as duplications
• Aligning at the nucleotide level
Problems

- Looking for orthologous regions
- Rearrangements disrupting orthology
  - Insertions
  - Deletions
  - Inversions
  - Translocations
  - Duplications
Rearrangement rate

• Can estimate the number of rearrangements from cytological comparisons

• Two very different rates:
  – Very slow rate of rearrangement (1 or 2 exchanges per 10 MYR)
    • ~7 rearrangements between the human genome from the hypothetical primate ancestor (60 MYA)
    • 13 rearrangements between cat and human
Rearrangement rate

• Punctuated by abrupt global genome rearrangement in some lineages
  – Gibbons and siamangs rearranged 3 or 4 times more extensively than human or other great apes
  – Dogs have highly rearranged genomes compared to the ancestral Carnivora genome
  – Rodent species exhibit very rapid patterns of chromosome change
    • ~180 conserved segments between mouse and human
    • ~100 conserved segments between rat and human
Fig. 1. Comparison of the relative order of X chromosome type 1 coding gene homologs between human, feline, and mouse X chromosomes shows six rearranged (by intrachromosomal inversion) segments conserved between mouse and human or between mouse and cat. The same genes have an identical order across the entire feline and human X chromosomes (72). ~ indicates type II STRs used to build the integrated cat map (19, 51, 72). Arrows indicate the polarity of mouse gene order; for example, toward the chromosome terminus in mouse.

O’Brien et al, Science 1999
Genome comparison across species
gross changes in chromosome number

O’Brien et al, Science 1999
Assumptions

- The two genomes to be aligned derived from a common ancestor
- There remains sufficient similarity between the genomes to enable “easy” identification of homologous regions
- For the alignment to be informative, there has to have been time for the genomes to diverge and for selection to have occurred
Algorithm requirements

• Genome alignment algorithms must
  – Scale linearly (computationally)
  – Be robust (not too many parameters)
  – Be memory-efficient
  – Be able to handle rearrangements, gene duplications, repetitive elements

• Smith-Waterman, Needleman Wunsch
  – Time to do calculation of the order of $O(n^2)$
    • Not feasible for sequence length > 10,000 bp
  – Cannot handle rearrangements or inversions
Alignment methods

- Seeding methods (e.g., BLASTZ, BLAT, EXONERATE)
  - Produce “local” alignments
  - All matches found (including all paralogs)
    - Very sensitive, not very specific
    - Very fast
- Anchor-based methods (e.g., MUMmer, AVID, WABA)
  - Produce “global” alignments
  - Specific
  - Difficulty with rearrangements
- Multiple sequence aligners (e.g., Mauve, GRIMM-Synteny, Shuffle-LAGAN, Mercator)
  - Designed to deal with rearrangements
Local aligner: LASTZ

- Remove lineage-specific interspersed repeat regions
- As in BLAST, find seeds
  - Seeds are determined by a 12of19 weighted-spaced seeds strategy where the positions for strict matches are specified
  - $1110100110010101111$ combination shown to be the most sensitive (and more sensitive than the 11-consecutive match seed strategy used by BLAST)
  - Some strategies also allow one transition in $1/12$ strict match positions
  - Seeds with many matches masked out (assumed to be repetitive regions)

LASTZ

- Initial gap-free extension
- Further DPA gapped extension of extended matches over a certain threshold score
  - Low complexity regions have their scores specifically down-weighted
- Repeat above steps (using a more sensitive seed, e.g., 7-mer) for regions that lie between matches, that share the same order and orientation, and are separated by <50 kb
- Post-processing to remove multiple hits to the same region (attempt to find single best orthologous match)
- When aligning human and mouse genomes, can achieve 98% alignment of known coding regions
MULTIZ

- M and N are “reference blocksets”, aligned using LASTZ
- Reference blocksets are sequences aligned to a reference sequence, where each position in the reference can occur only once (i.e., be part of only one block)
- Blocks in M and G are sorted by starting position
- Find segments of a block in M and a block in N that are predicted to align to each other, according to some block in G
- Align segments by DPA
“Glocal” aligner: Shuffle-LAGAN

- 2-genome aligner
- First one to explicitly deal with rearrangements
- CHAOS: local aligner
  - Finds matching seeds, and chains them together (within a cutoff distance)
  - Extends chains with ungapped BLAST
- Create a map where the anchors have to be collinear in one sequence only (1-monotonic conservation map)
- Align consistent subsegments
  - Sort all local alignments by their start coordinates, identify alignable subsegments, expand to borders of adjacent subsegments

Brudno et al, Bioinformatics, 2003
Global aligner: Mauve

- Multiple species aligner
  1. Find local alignments (multi-MUMs)
     - seed-and-extend hashing method
  2. Create a phylogenetic guide tree (NJ)
     a) Calculate a similarity score from the ratio of shared base pairs between two genomes to their average genome length
     b) Account for overlapping multi-MUMs
  3. Select a subset of multi-MUMs to use as anchors
     a) Partition into LCBs (locally collinear blocks)
     b) Remove low weight LCBs to maximize adjacent LCB lengths
  4. Recursively identify additional anchors around and within each LCB, using smaller (more sensitive) seeds
  5. Progressively align each LCB using the guide tree.

Darling et al, Genome Research, 2004
A) The initial set of matching regions:

B) Minimum partitioning into collinear blocks:

C) After removing block 3:
LCBs in enterobacteria genomes

Darling et al, Genome Research, 2004
Breakpoint identification: GR-Aligner

- Identify and merge conserved sequence pairs
  - BL2Seq used to generate local non-overlapping matches
  - Matches are sorted by start position
  - Merge adjacent matches if they are either both direct or inverse and if the combined score is over some threshold (interval aligned with NW)

- Identify breakpoints in inversions
  - Identify all inverse matches flanked by direct matches
  - Align homologous sequence between the inverse and direct match and find the optimal score

Chu et al, Bioinformatics, 2009
GR-Aligner

- Identify breakpoints in transpositions
  - Defined as direct matches that cross only one other direct match, and where the crossed matches are flanked by direct matches that do not cross any other matches
  - Align sequence between the transposed match with the sequences outside the transposed match and find the optimal score

Chu et al, Bioinformatics, 2009
Genome alignment parameters

- **Repeat-masking, E-value**
  - E-values estimated by aligning reversed genome
  - TRF (tandem repeats finder): only repeat-finder that eliminated many spurious alignments (non-standard parameters (found by trial and error), hard-masking)

- **Scoring matrix**

- **Gap costs**
  - 495 combinations of above 3 parameters (using hard-masking with TRF and an E-value of 1)
  - For structural RNA alignments, 3:3:4:24:1 and 4:4:5:24:1

- **X-drop parameter**
  - Used to terminate extension of gapped alignments (terminates if the score drops by more than X below the previously seen maximum score)
  - Seems to find fewer false-positives when X-drop is not greater than alignment score cutoff
Visualization tools: VISTA vs. PipMaker

- **VISTA (VISualization Tools for Alignments)**
  - Uses AVID to generate alignments
  - Uses a sliding window approach
    - Plots percent identity within a fixed window size, at regular intervals

- **PipMaker (Percent Identity Plot)**
  - Uses BLASTZ
  - X-axis is the reference sequence; horizontal lines represent gap-free alignments
2-D view: VISTA-dot

http://genome.jgi-psf.org/synteny/
Circos

http://mkweb.bcgsc.ca/circos/
Identification of functional elements

• Coding sequences
  – Relatively easy to identify
    • Many gene prediction programs available
    • General gene structure known, e.g.,
      – TATA boxes
      – Splice donor-acceptor sites
  – ESTs and cDNAs available to aid gene prediction

• Non-coding functional sequences
  – Much harder to identify
  – No common structure of regulatory regions
  – TF binding sites are short and ubiquitous

• Comparative genomics
  – A genomic sequence that provides a function that is under selection and tends to be conserved between species is called a “functional sequence” (e.g., a protein-coding region or a transcription factor binding site)
Coding region comparison

- Discover new genes
  - Annotate gene structure
    - Exon-intron structure

- Compare gene content
  - Find genes common to sets of organisms
  - Find genes unique to an organism

- We can study how genomes vary within a species to identify genes central to particular processes
  - Can also compare subspecies e.g., *E. coli* K12 and O157:H7 (pathogenic)

- Discover evolution of gene function
  - HAR1F
Predicting structure of ‘new’ genes

• Many gene finding programs
  – Look for start sites, termination sites, splice sites
  – Analyze codon usage, exon length

• Comparative method A
  – Find syntenic regions
  – Predict genes using conventional gene finding techniques in both species
  – Genes predicted in both species are probable
Gene prediction

- Comparative method B
  - Find syntenic regions
  - Perform pattern filtering
    - Coding exons tend to be well conserved
    - Conservation higher in first and second positions of codons
  - Advantage: can deal with sequencing errors
Identification of paralogues and orthologues in other species

- Best Reciprocal Hit (BRH)
- Reciprocal Hit by Synteny (RHS)
  - Identification of adjacent orthologues
- Domain checking - internal quality control

http://www.nbn.ac.za/
Identification of orthologues

BRH → Others → Orphan

RHS → Others → Orphan

Matches to some other chromosome

Human → BRH → Human

Human → Others → Human

Human → Orphan → Human

Human → Mouse

Mouse → Others → Mouse

Mouse → Orphan → Mouse

Mouse → BRH → Mouse

Mouse → RHS → Mouse

Mouse → Human

http://www.nbn.ac.za/
Finding all homologs, and only homologs

Find most similar vertebrate gene (here M1) to the *Ciona* gene. Other vertebrate genes are added to the cluster if they are more similar to M1 than M1 is to the *Ciona* gene.

Duplicates may have arisen by speciation (lineage-splitting) or by gene duplication events specific to one or more vertebrates.
Genome sequences of 3 strains of *E. coli*

MG1655 (K-12) non-pathogenic

- 4,288 genes
  - 585 (7.6%) in all 3
  - 514 (6.7%) in 2 out of 3
  - 3554 (46.5%) in 1 out of 3

EDL933 (O157:H7) enterohaemorrhagic

- 5,063 genes
  - 2996 (39.2%) in all 3
  - 911 (11.9%) in 2 out of 3

CFT073 uropathogenic

- 5,016 genes
  - 193 (2.5%) in all 3
  - 1623 (21.2%) in 2 out of 3

Total proteins = 7461

Rich in genes encoding potential fimbrial adhesins, autotransporters, iron-sequestration systems, phase-switch recombinases.

Includes genes encoding fimbrial adhesins, iron uptake systems, as well as toxins and toxin-transport systems and integrases.

Welch et al, PNAS 2002
Conserved non-coding elements

- PhastCons – phylo-HMM
- Alignment of five vertebrate genomes, using MULTIZ
- 3-8% of the human genome is conserved in other vertebrates – about 1.18 million elements
  - Distinguish from “ultra-conserved” coding elements (exons)
  - Some highly conserved HCEs are 1000s of bases long
  - Many HCEs overlap with 3’ UTRs, particularly of genes that regulate other genes,
  - Many HCEs show potential to have local RNA secondary structure
    - Possible function – post-transcriptional modifications/ regulation or code for functional RNAs
  - Also many HCEs are in gene deserts – possible distal regulatory elements?

Siepel et al, Genome Research 2005
RNA gene evolution

• Identify regions which exhibit an accelerated lineage-specific rate of evolution
  – Look for regions in the chimpanzee genome over 100bp long that are ≥ 96% identical with mouse or rat
  – Identify orthologous human regions using MULTIZ, and identify those regions which have increased numbers of changes
  – 49 HARs/34498 orthologous regions, most in non-coding regions
    • Adjacent to genes involved in DNA-binding or transcriptional regulation
  – Most striking example: HAR1

HAR1

- 118bp region on chromosome 20
  - Region not present in frogs or fish
    - Arose < 400 MYA
  - 18 substitutions compared with an expected 0.27 (based on rate of change in other amniotes)
  - Strong bias of “weak” to “strong” substitutions
  - Changes occurred in the human lineage only, and are fixed
  - Part of the HAR1R/HAR1F genes, which exhibit low levels of conservation except around the HAR1 region
  - HAR1F and HAR1R are novel non-coding RNA genes
  - Expressed in developing neocortex – may be involved in human neurodevelopment

Genome evolution

- Once we have a macro-alignment, we can also study the evolution of genomes between species, and also can trace the evolutionary history of the structure of each genome itself
- Genome structure rearrangement
- Gene duplication, chromosomal duplication and polyploidization (whole genome duplication)
  - New genes
Phylogenetic tree – shared gene content

http://www.bork.embl-heidelberg.de/~korbel/SHOT_v2/
Phylogenetic tree – gene order

Unravelling the history of the genome

- Plants
  - Wheat
    - Allohexaploid (AABBCC)
  - Maize
    - Diploidized allotetraploid
      - Has grown 12x in the past 5 MY due to increased numbers of transposable elements
  - Arabidopsis
    - Haploid, 5N

- Yeast
  - *Saccharomyces cerevisiae* vs. *Kluyveromyces lactis*

- Human
  - Genome duplication?
Decipher history of lineages

- **Genome size changes**
  - Compaction
    - Large scale deletions - Fugu
    - Intron loss
  - Expansion
    - Duplication
    - Transposable element insertions

- **Transposons - Alus, LINEs**
  - Ancient insertions prior to eutherian radiation
  - More recent insertions - maize

Baxendale et al, Nature Genet, 1995
Why gen(om)e duplication?

• Duplicated genes provide a source for genetic novelty during evolution
  – Either member of a duplicated gene pair can diverge to either
    • Acquire a new function which may be positively selected for
    • Subfunctionalize
    • Be differently regulated (e.g., tissue-specific)
    • Become a pseudogene
    • Be deleted

• Whole genome duplication allows for the duplication (and subsequent divergence) of whole pathways at a time
Duplication events

• Resolution of polyploidy
  – Non-disjunction of chromosomes (form multivalents instead of divalents)
  • Sterility
  – Duplicated genes do not start to diverge (or get deleted) until disomic inheritance resolved
Rearrangements

• Are genome rearrangements the cause or consequence of diploidization?
  – Most widely accepted hypothesis is that diploidization proceeds by structural divergence of chromosomes
  – Some loci appear disomic, others tetrasomic
  • Stage 1: pairing between similar chromosomes allowed
    – Loci near the centromeres can display residual tetrasomy
  • Stage 2: non-homologous chromosome pairing resolved
Effects of polyploidization

Figure 4 | Computer simulations of the dot-matrix patterns expected from one or two rounds of genome duplication. a | One round of genome duplication, and b | two rounds of genome duplication; in both cases, after a small number of chromosomal rearrangements and assuming no gene deletion. In a, a chromosome containing 200 genes was duplicated, and five interchromosomal translocations were made; each gene appears twice in the genome. In b, a chromosome containing 100 genes was duplicated twice, and then five interchromosomal translocations were made; each gene appears four times in the genome.
Saccharomyces cerevisiae (baker's yeast)
Whole Genome Duplication (WGD)

Number of chromosomes

- Saccharomyces cerevisiae: 16
- Saccharomyces paradoxus: 16
- Saccharomyces mikatae: 16
- Saccharomyces kudriavzevii: 16
- Saccharomyces bayanus: 16
- Saccharomyces castellii: ~9
- Candida glabrata: 13

- Kluyveromyces waltii: 8
- Saccharomyces kluyveri: 8
- Kluyveromyces lactis: 6
- Ashbya gossypii: 7
- Debaryomyces hansenii: 7
- Candida albicans: 8

Yeast

- *Saccharomyces cerevisiae*
  - Degenerate tetraploid
  - Polyploidy followed by extensive deletion and (70-100) reciprocal translocations
  - 8% of genes duplicated in 55 blocks (plus many missed smaller blocks)
  - Relative orientation of genes in blocks conserved with respect to the centromere

Seoighe and Wolfe, PNAS, 1998
Duplicated blocks in yeast

Wolfe and Shields, Nature 1997
Estimation of time of polyploidy event

- Diverged from *Kluyveromyces lactis* (unduplicated) ~150 MYA
  - Comparison of gene sequences and gene order reveals conservation
    - 59% of adjacent gene pairs in *K. lactis* or *K. marxianus* are also adjacent in *S. cerevisiae*
    - 16% of *Kluyveromyces* neighbors can be explained in terms of inferred ancestral gene order
  - Phylogenetic analyses of duplicated genes where both the *Kluyveromyces* orthologue and an outgroup orthologue were available, deduced that the polyploidization event in *S. cerevisiae* occurred around 100 MYA (used only 12 gene pairs)

10% retention

Different subsets retained

Evidence from conserved order of a very few genes

Evidence from interleaving genes from sister segments
Each region of *K. waltii* matches two regions of *S. cerevisiae*. We don't even need any remaining two-copy genes to infer the ancestral order; 47/253 DCS contain no duplicated genes.

Evidence of accelerated evolution
In 17% of gene pairs
Chlamydia evolution

• Chlamydia psittaci
  – Virulent potentially life-threatening pathogen
• Multiple genome alignment of 20 genomes, using MAUVE
  – 9 human serotypes, other animal host infections
  – 1.1Mb genome, 8kb plasmid
  – 911 core genes (shared by all genomes)

1.682x10^{-4} substitutions/year/site (mutations and recombinations)

Read et al, mBio, 2013
Reconstruction of plasticity region

Read et al, mBio, 2013
ClonalFrame

- Bayesian inference of bacterial clonal relationships
- Distinguish point mutations from homologous recombination
- Differs from allele-based inference methods
  - Distinguishes between “alleles” that are similar versus those that are highly divergent
  - Allows for breakpoints to occur within “alleles”
- Does not model origin of recombinant sequence import

Didelot and Falush, Genetics, 2007
Recombination events

Read et al, mBio, 2013
Origins of recombinant fragments

3 ancestral populations required to explain the current population structure

come from a non-C. psittaci source

Read et al, mBio, 2013
Future challenges

• Number of genomes available growing rapidly
  – Need novel ways of storing and representing 1000s of genomes
  – Algorithms need to be able to handle large numbers of genomes (not just very long sequences)

• Need more rigorous methods of defining parameter sets (currently, many are chosen heuristically)
  – Recognize that rate of evolution may vary in different parts of the genome