Reviews in Computational Biology

4. Figures & Tables in Reviews

Christophe Dessimoz

March 15th, 2010

ETH Zürich inf | Informatik Computer Science

Conceptual overview: flowchart

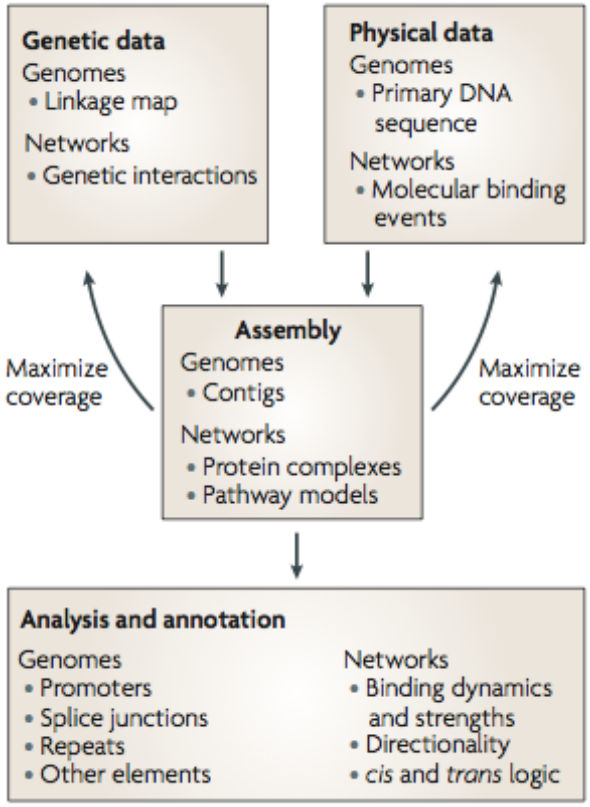

Figure 1 | **Genetic and physical mapping for networks and genomes. a** | The assembly and analysis of genetic and physical interaction networks runs parallel to the procedures that were previously developed for assembly and analysis of DNA sequences. **b** | An integrated map of human chromosome X. Markers are listed in the centre column,

Beyer et al. Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet (2007) vol. 8 (9) pp. 699-710

More concrete overview: Schematic

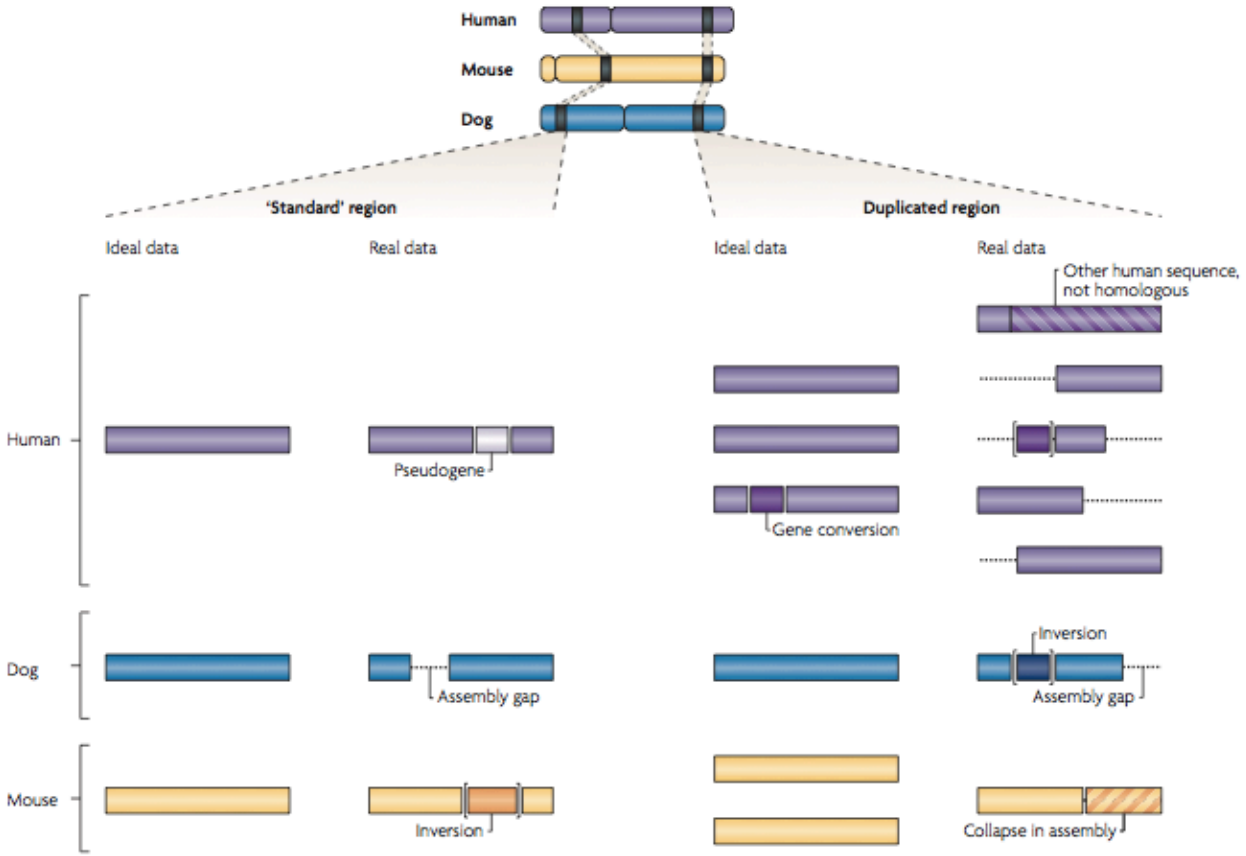

Figure 3 | **Challenges in the reconstruction of homologous collinear regions.** The left-hand panel shows a 'standard' region that is expected to be collinear along its entire length. However, the presence of retrotransposed pseudogenes, assembly gaps and inversions complicates the problem. The right-hand panel shows a duplicated region for which, even when there is ideal data, challenges are presented by aspects such as gene conversion between duplicated copies that make the relationship between different copies in different genomes complex. In the example shown, there are three copies of the duplication in human, one in dog, and two in mouse, which ideally should all be aligned in a collinear manner. In addition to the challenge of gene conversion, these regions usually contain the most complex assembly artefacts, including artificial expansions of duplicated regions into partially overlapping contigs with inappropriate neighbouring data interspersed with the genuinely homologous regions (shown in human), or collapses in assembly where there is an apparent single region that is actually a chimaera of two regions (shown in mouse). The normal assembly issues of gaps and inversions also occur (shown in dog), usually at a higher rate than in standard regions.

Margulies and Birney. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. Nat Rev Genet (2008) vol. 9 (4) pp. 303-13

Enumerate relevant structures using figures…

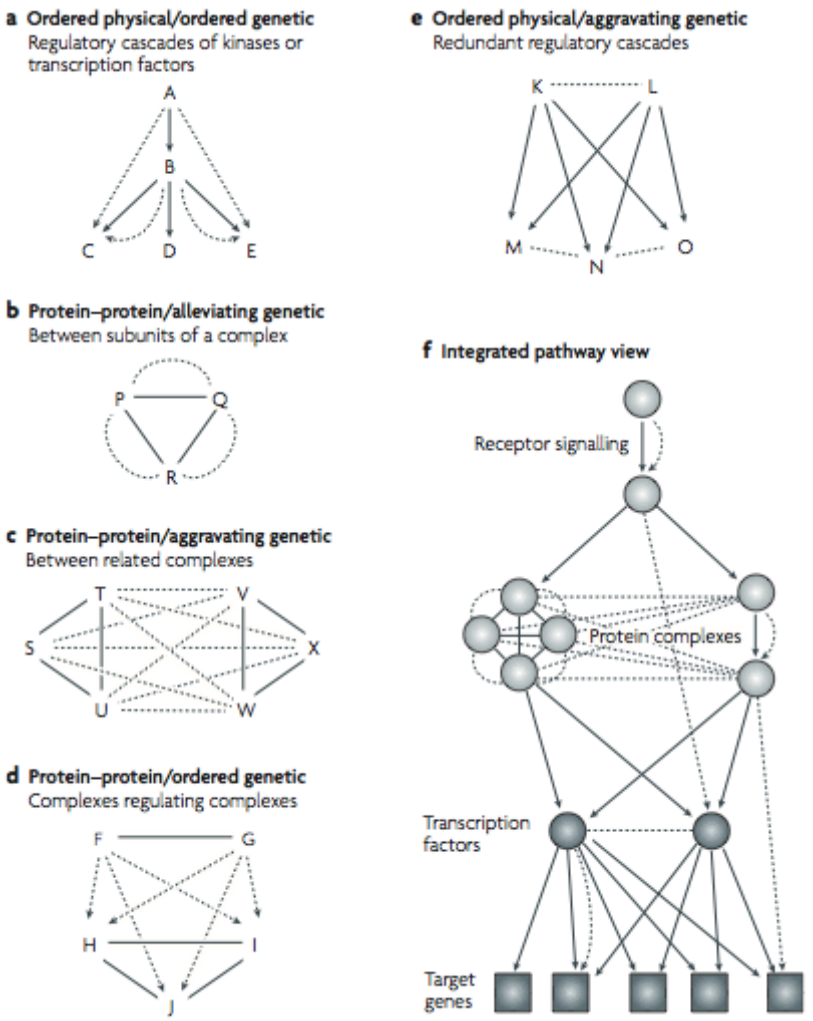

Figure 4 | **Network motifs assembled from different combinations of interaction measurements.** Physical interactions are shown as solid lines and arrows, and genetic interactions are shown as dashed lines and arrows. Part **a** shows an example of integrating ordered physical versus ordered genetic interactions, in which knockout of A or B results in changes in the activity of C, D and E (ordered genetic interactions), which are brought about because of changes in transcriptional activity or kinase–substrate binding (ordered physical interactions)[52,57,110,111]. Members of protein complexes (protein–protein interactions) can be connected by genetic interactions either within complexes (shown in part **b**) or between complexes (shown in part **c**) [37–39,42,46,65]. In part **d**, members of a complex made up of F–G (protein–protein interactions) operate upstream of or epistatically to (ordered genetic) the complex H–I–J[42,109]. In part **e**, regulatory factors K and L cooperate to activate targets M, N and O (ordered physical) which function in parallel pathways (alleviating genetic[46,51,66]). Part **f** shows how the motifs of previous panels might combine within a still larger network, starting at a receptor protein and ending at transcription factors modulating the expression of target genes. Note that the motifs in each panel are summarized from the literature (see references provided) and are not intended as an exhaustive catalogue of all ways of integrating interactions.

Beyer et al. Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet (2007) vol. 8 (9) pp. 699-710

...or a table

Table 1 | **Evidence codes used by GO**

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

*October 2007 release

Rhee et al. Use and misuse of the gene ontology annotations.
Nat Rev Genet (2008) vol. 9 (7) pp. 509-15

e.g. list of tools reviewed

Table 1 | **List of population genetics programs examined in this review**

Name	Version	Platform	Graphical interface	Accepted data type	Handled data format	References
Multi-purpose packages						
Arlequin	3.01	Win	Yes	DNA, SNP, STR, MULT, FREQ	Specific, GENEPOP	49
DnaSP	4.10	Win	Yes	DNA, SNP	In — MEGA, NEXUS, FASTA, PHYLIP; out — MEGA, NEXUS, FASTA, PHYLIP, Arlequin	50
FSTAT	2.93	Win	Yes	STR, MULT	Specific, GENPOP	51
GDA	1.1	Win	Yes	AFLP, MULT	In — NEXUS, BIOSYS, GeneStrut; out — NEXUS, BIOSYS, GeneStrut, GENESTAT-PC, SAS	See Boxes 1,2
Genepop	3.4	DOS	No	STR, MULT	Specific	52
GENETIX	4.05	Win	Yes	MULT	In — specific, FSTAT, Genepop; out — specific, FSTAT, Genepop, BIOSYS, Arlequin	See Box 1
MEGA	3.1	Win	Yes	DNA, DIST	In — specific, CLUSTAL, NEXUS, PHYLIP, GCG, FASTA, NBRF/PIR, MSF, IG; out — specific, PHYLIP, NEXUS	53
MSA	4.0	DOS, MacOS, Linux	No	STR, MULT	In — EXCEL; out — Genepop, MSVAR, Structure, Arlequin, Migrate	54
SPAGeDi	1.2	DOS	No	STR, MULT	Specific, FSTAT, Genepop	55
Individual-centred programs						
BayesAss+	1.3	Win, MacOS, Linux	Yes	MULT	Specific, IMMANC	56
BAPS	3.2	Win	Yes	MULT	Specific, Genepop	57
GeneClass	2.0g	Win	Yes	MULT	Genepop, FSTAT, GENETIX	58
Geneland	1.05	R	No	MULT	Specific	46,47
NewHybrids	1.1b3	Win, Linux	Yes	MULT	Specific	59
Structure	2.1	Java	Yes	MULT	Specific	60,61

Excoffier and Heckel. Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet (2006) vol. 7 (10) pp. 745-58

Here with more details

Table 3 | **Program functionalities and assumptions: individual-centred programs**

Name	Short description of functionalities	Special features	Inference framework	Assumptions and issues
BayesAss+	Estimates recent migration rates between populations from multilocus genotype data	Estimates each individual's immigrant ancestry, the generation in which immigration occurred, and inbreeding levels within populations	MCMC, Bayesian	Assumes co-dominant unlinked markers, and sampling of source populations of the immigrants; allows for missing data
BAPS	Assigns individuals to genetic clusters by either considering them as immigrants (mixture analysis) or as descendents from immigrants (admixture analysis)	Estimates the number of genetic clusters; provides the proportion of the genome of each individual that can be assigned to the inferred clusters (admixture analysis)	Bayesian	Assumes HWE within clusters and unlinked markers; partially uses information on the sampling origin of the individuals
GeneClass	Detects immigrants from multilocus genotypes, assignment of individuals to populations	Assesses whether a given genotype can be excluded from a given population	Bayesian, likelihood	Assumes HWE within populations; assignment to sampled populations only; no attempt to reconstruct virtual populations
Geneland	R package to detect population subdivisions that explicitly take into account the spatial position of sampled multilocus genotypes; computes *F*-statistics between inferred virtual populations	Determines the best number of subdivisions, and assigns geo-referenced individuals to a subdivision; provides graphical output of the spatial distribution of the subdivisions	MCMC, Bayesian	Assumes HWE and no linkage within subdivisions; immigrant genes are supposed to be present only in new immigrants

Excoffier and Heckel. Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet (2006) vol. 7 (10) pp. 745-58

Use a table to define terminology

Table 1 Homology: terms and definitions

Homologs	**Genes sharing a common origin**
Orthologs	Genes originating from a single ancestral gene in the last common ancestor of the compared genomes.
Pseudoorthologs	Genes that actually are paralogs but appear to be orthologous due to differential, lineage-specific gene loss.
Xenologs	Homologous genes acquired via XGD by one or both of the compared species but appearing to be orthologous in pairwise genome comparisons.
Co-orthologs	Two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage due to a lineage-specific duplication(s). Members of a co-orthologous gene set are inparalogs relative to the respective speciation event.
Paralogs	**Genes related by duplication**
Inparalogs (symparalogs)	Paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event (defined only relative to a speciation event, no absolute meaning).
Outparalogs (alloparalogs)	Paralogous genes resulting from a duplication(s) preceding a given speciation event (defined only relative to a speciation event, no absolute meaning).
Pseudoparalogs	Homologous genes that come out as paralogs in a single-genome analysis but actually ended up in the given genome as a result of a combination of vertical inheritance and HGT.

Koonin. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet (2005) vol. 39 pp. 309-38

Table to summarize results (note the references)

Table 1. Prevalence of gene duplication in all three domains of life[a]

	Total number of genes	Number of duplicate genes (% of duplicate genes)	Refs
Bacteria			
Mycoplasma pneumoniae	677	298 (44)	[65]
Helicobacter pylori	1590	266 (17)	[66]
Haemophilus influenzae	1709	284 (17)	[67]
Archaea			
Archaeoglobus fulgidus	2436	719 (30)	[68]
Eukarya			
Saccharomyces cerevisiae	6241	1858 (30)	[67]
Caenorhabditis elegans	18 424	8971 (49)	[67]
Drosophila melanogaster	13 601	5536 (41)	[67]
Arabidopsis thaliana	25 498	16 574 (65)	[69]
Homo sapiens	40 580[b]	15 343 (38)	[11]

[a]Use of different computational methods or criteria results in slightly different estimates of the number of duplicated genes [12].
[b]The most recent estimate is ~30 000 [61].

Zhang. Evolution by gene duplication: an update. Trends in Ecology and Evolution (2003) vol. 18 (6) pp. 292-298