

A Bayesian Framework for Data Integration: Application to Comparative Genomics

Madhu Bhattacharjee

**School of Mathematics and Statistics
University of St Andrews
UK**

Acknowledgement  ComparaGRID

Comparative Genomics

- In many situations comparison between two organisms can greatly benefit if information from multiple sources could be combined.
- Depending on the hypothesis of interest these sources can be various.
- In order to integrate and compare data from the different sources, say e.g. maps, a statistical model is needed which explains all of the relevant entities and their variability, e.g. markers and their locations on maps.

Bayesian Inference

- Bayesian statistical inference is ideal for data fusion and data synthesis.
- Allows us to utilise available prior information and develop models integrating diverse data sources.
- Models are quite often analytically intractable but sophisticated computational algorithms are available for implementation.
- Models allow propagation of error coherently making final inference more robust.

Map Integration

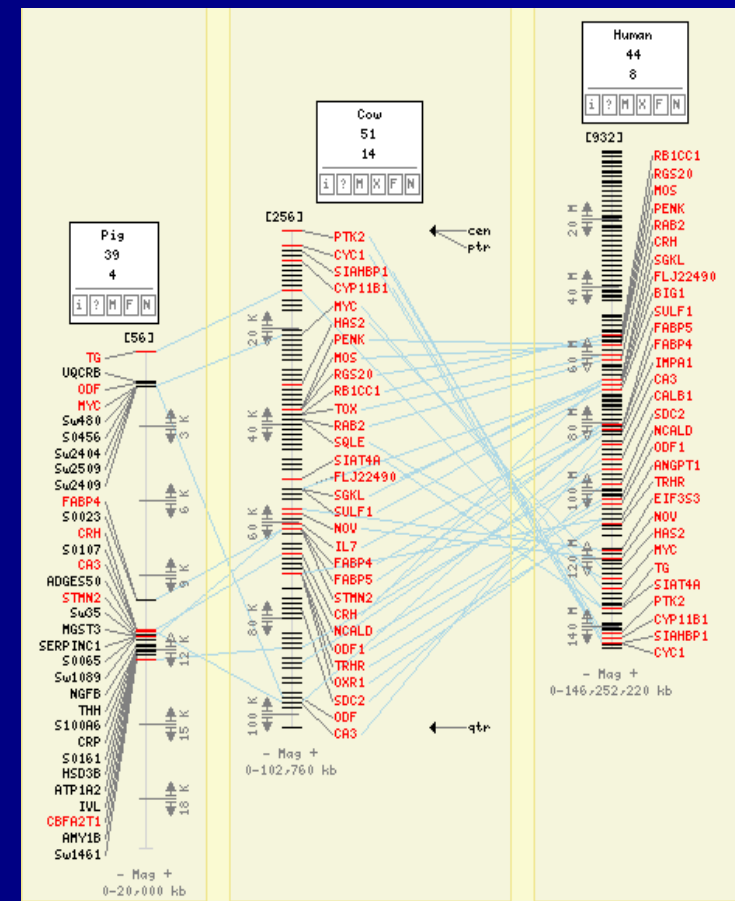
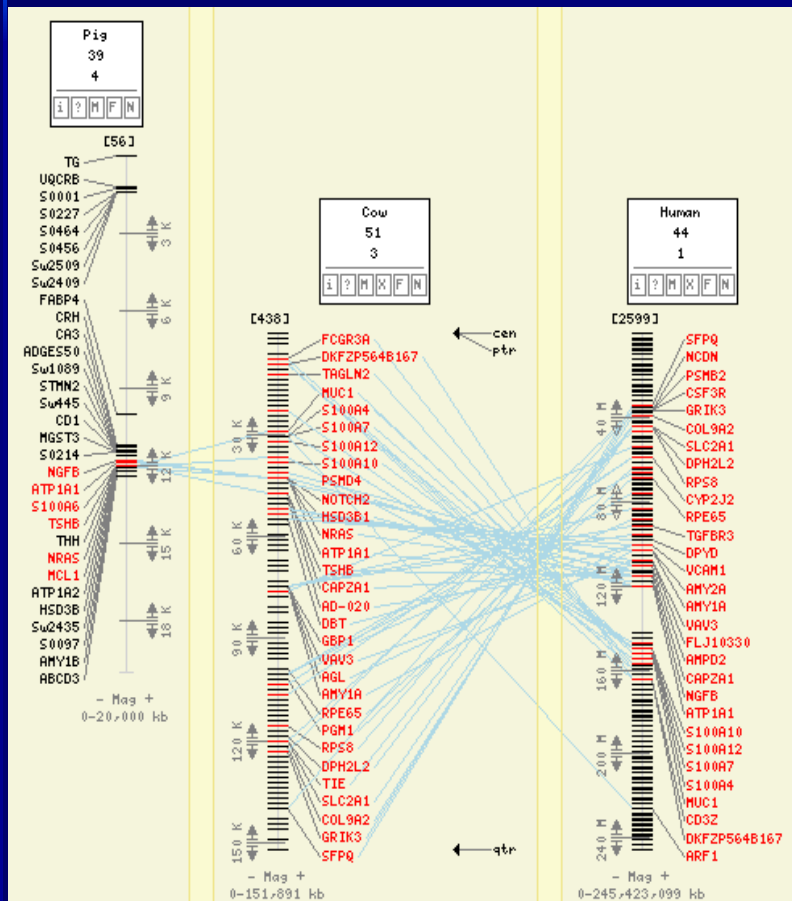
- We illustrate how we can integrate disparate information from maps to form a comprehensive and higher resolution view of a genome.
- Map integration is very important for any species for which an annotated complete genome sequence is not available.
- For organisms that are currently being sequenced a pre-sequence comprehensive map is essential to provide a backbone for assembly.
- Integration also facilitates the identification and resolution of discrepancies among different maps.

Integration of Maps - Existing literature

- Typical solutions are visual:
 - graphical representation of multiple source maps are placed side-by-side (in their original units)
 - highlighting the commonality as well as discrepancy amongst these.
- Notable feature of majority of these attempts:
 - attempt to produce a single map, possibly linear
 - all the uncertainties involved are ignored
- Softwares like Joinmap (Stam 1993) or Carthagene (Schiex and Gaspin 1997) do use a statistical framework, but are not free from some limitations.
- CLDB and ICCARE projects are relevant and interesting although not very statistical.

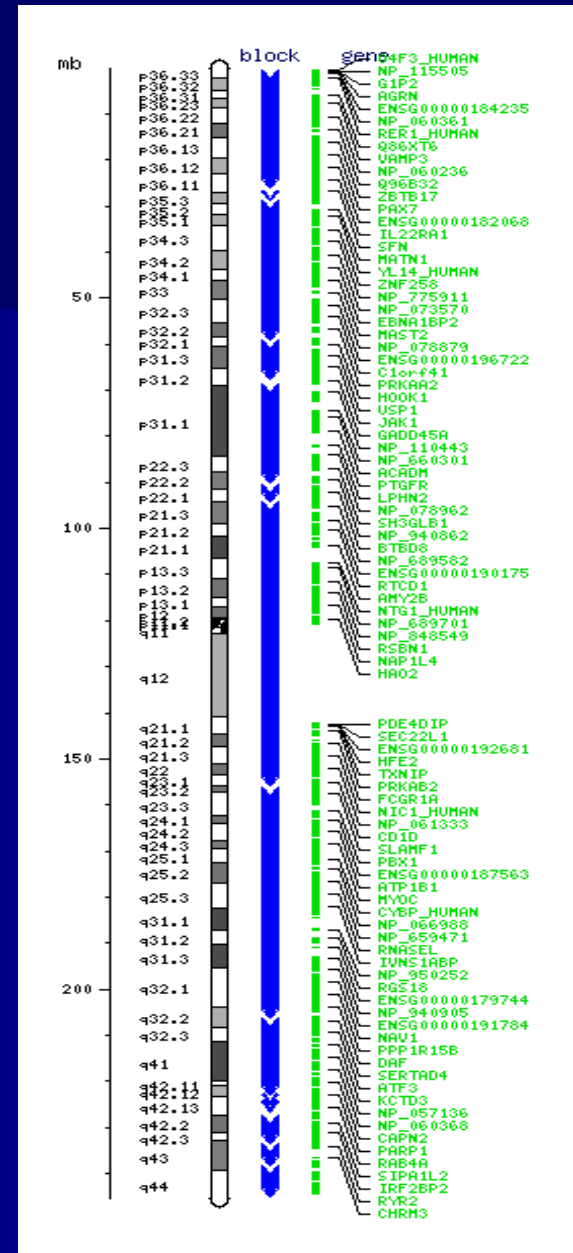
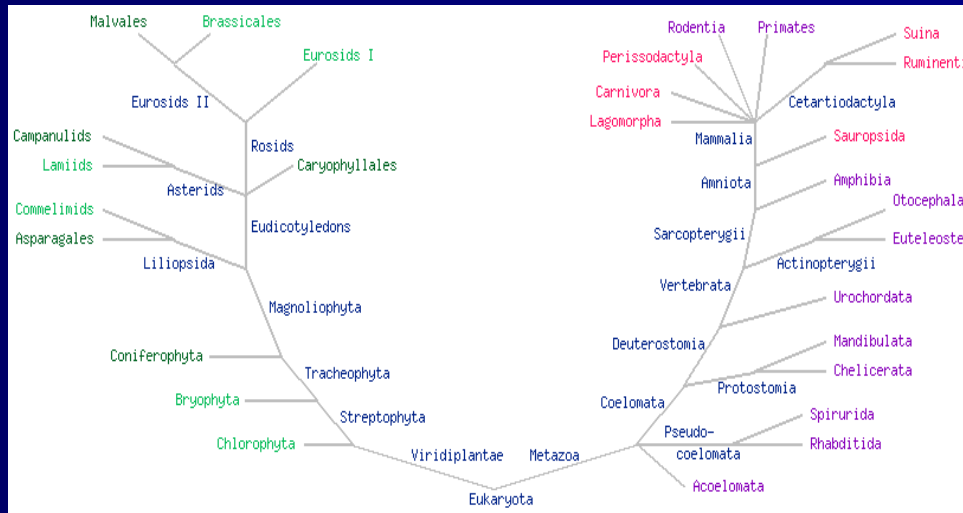
Integration of Maps - Existing literature

Comparative maps from CLDB



Integration of Maps - Existing literature

Comparative maps from ICCARE



■ Homo sapiens gene with candidate Sus scrofa orthologous EST

Application to comparative genomics

- Refine estimates of loci for species of interest loci using other (e.g. model) organism loci location information
- Predict (for one species) physical locations of interesting genes which are present on another genome.
- Predict (for one species) locations on relevant genetic map of interesting genes which are present on the human genome.
- Incorporate information such as zoo-FISH data and create a “virtual map”.

Bayesian Integration of Maps

- Utilises Bayesian graphical network representation of the problem.
- The ideology behind application of such a model to the genomic data is that there is a true physical map and all the partial maps provide a view of this true map.
- Neither the observed loci ordering nor their (estimated) positions need to be assumed to be true
- Few relevant manuscripts:
 - Liao, Colins, Hobs, Khatkar, Luo and Nicholas 2007
 - Yap et al. 2003
 - Stassen and Scharfetter 2000

Partial Maps

- For a large number of organisms several partial maps of various types are available,
e.g. genetic, landmark based physical maps, clone based maps.
- Typically these are created by various groups over a period of time, created for a wide range of purpose, based on diverse mapping populations.
- Each of which may contain valuable information for that particular species, but possibly none which is complete enough to form a comprehensive and reliable basis for genome study of that species.

Partial Maps: Orientation & origin

- Default orientation of maps would be from the p-telomere to the q-telomere.
- However occasionally maps are found inverted with respect to others, partially or fully.
- May span only a segment of a chromosome and thus have an origin either at zero or at a location within the chromosome.
 - e.g. Radiation hybrid maps typically consist of such segments

Partial Maps: Relationship with physical map

- RH maps are typically expressed in “centiray (cR)” measurements. cR is assumed to have a linear relationship with physical map.
- Cytogenetic maps provide crude location information for loci. However either the locations or the band positions can be converted in “flpter” format. Using genome size and arm length informations.
- Genetic maps require careful consideration before integration. For one type of species (e.g. animals) a linear approximation with physical map may be reasonable but not so for another (e.g. plants)

Non-linearity of Maps

- Majority of the observed non-linearity can be attributed to the characteristics of map construction algorithms that involve maximum likelihood estimates from incomplete data.
- Stassen and Scharfetter (2000) noted that observed non-linearity between maps are predominantly local effects at a good overall linearity.
- Occasional telomeric abnormalities that were also observed, which were postulated to be caused by sparse markers and missing flanking information.

Loci

- A “locus” is taken in a broad sense to include any category of mappable object.
- The type of loci information may carry useful information about their quality (i.e. precision of location information).
- Aliases should be investigated and distinct loci list would be preferable for map integration.
- For comparative purposes orthologous loci in another species may be identified before hand.

Bayesian Integration of Maps

The observed value y_{ij} from the i^{th} map and j^{th} data is assumed to be

$$y_{ij} \sim \text{Normal}(\alpha_i + \beta_i * \mu(d_{ij}), 1/\tau_{ij}), \text{ where}$$

$i = 1, \dots, N$: number of maps used for analysis

$j = 1, \dots, n_i$: number of data points on i^{th} map

d_{ij} = the distinct marker no. corresponding to the j^{th} data from i^{th} map

$\mu(k)$ = true location of k^{th} distinct marker

τ_{ij} = precision associated with j^{th} data from i^{th} map

The parameters involved are modelled utilising prior information about the maps and/or loci.

Possible priors for latent true location of loci :

A prior distribution for latent true locations may be given by

$$\mu(k) = \mu^0(c^0(k), k) \text{ where}$$

$$\mu^0(j, k) \sim \text{Uniform}(L(k, j), U(k, j)); j = 1, 2, \dots, nc_k,$$

nc_k = number of intervals given by cytogenetic data for k-th locus

$$\mu^0(nc_k+1, k) \sim \text{Uniform}(0, L(C_k)),$$

$$c^0(k) \sim \text{Categorical}(p^1(k)),$$

$$p^1(k) \sim \text{Dirichlet}(\mathbf{I}),$$

$$\mathbf{I} = (1/nc_k, 1/nc_k, \dots, 1/nc_k, 1); \text{ dimension } nc_k+1,$$

$L(k, j)$ = For the k-th locus the lower bound from the j-th cytogenetic intervals for this particular locus,

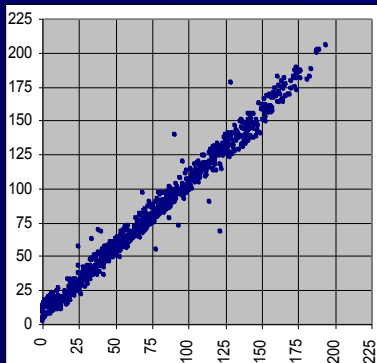
$U(k, j)$ = For the k-th locus the upper bound from the j-th cytogenetic intervals for this particular locus,

C_k = chromosome number to which k-th distinct marker is located,

$L(l)$ = length of l-th chromosome.

Barley : Bayesian Integrated Map

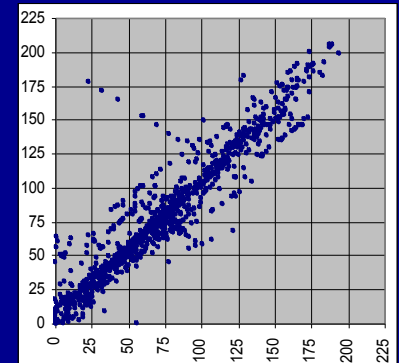
- Barley map data on 14 map population from Graingene database
<http://wheat.pw.usda.gov/GG2/index.shtml>
- Original number of data: 3990, data used: 3517, 2856 distinct loci
- Softwares like JoinMap has limitations of working with large number of loci.



BI using 14 map populations (prelim.)

vs.

← BI using 3 map populations
JoinMap for 3 map populations →



Barley : Bayesian Integrated Map

Bayesian estimated location and posterior interval for loci on Barley chromosome 1

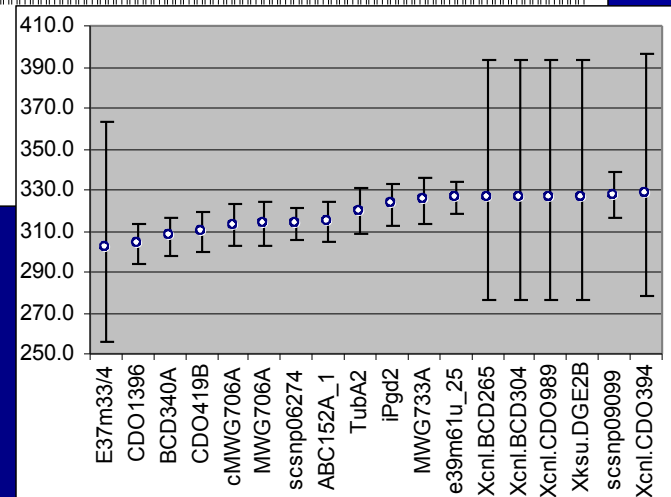
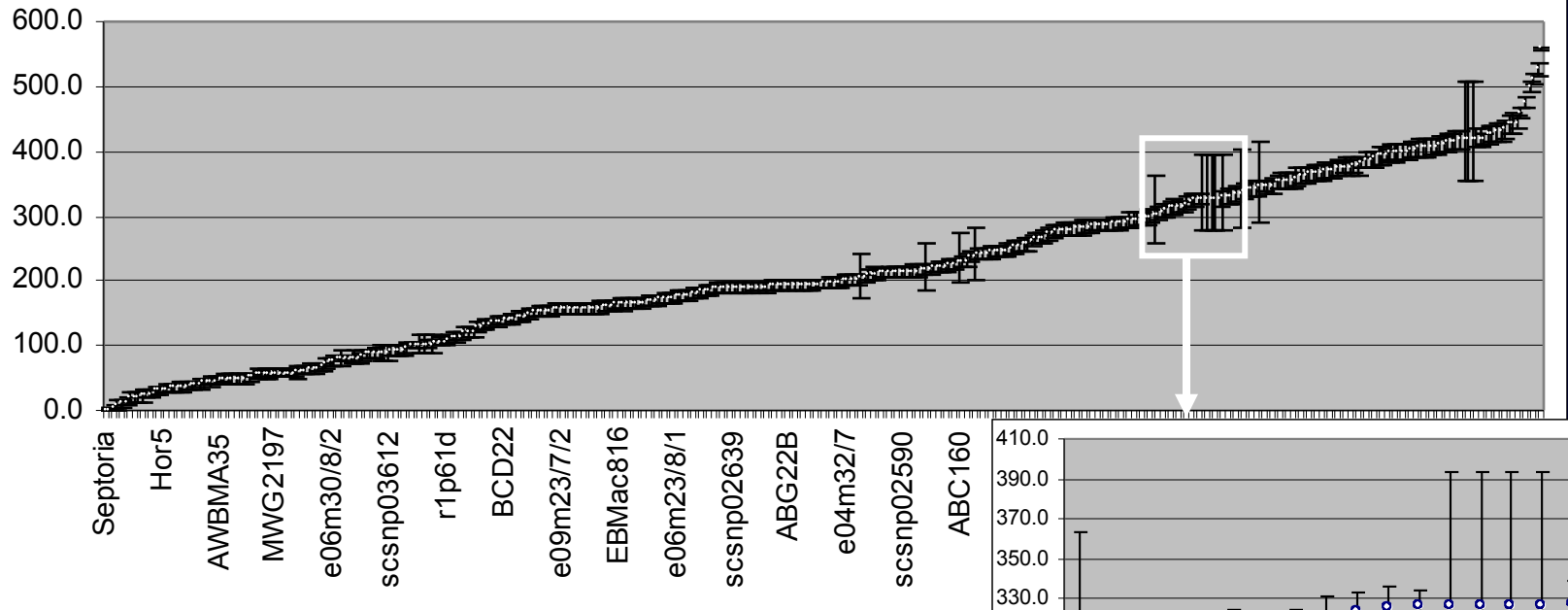


Fig : Bayesian Integration of Linkage, RH & Cytogenetic Maps

- Cytogenetic map data, in cytoband form, obtained from INRA & CLDB.
- Cytogenetic map data, in flpter form, obtained with help of
Andy Law <andy.law@bbsrc.ac.uk>
Trevor Patterson <trevor.paterson@bbsrc.ac.uk>
Phil Devall <phil.devall@bbsrc.ac.uk>

Pig : Bayesian Integration of Linkage, RH & Cytogenetic Maps

- All chromosomes: Obtained linkage map data with help of
Andy Law <andy.law@bbsrc.ac.uk>
Trevor Patterson <trevor.paterson@bbsrc.ac.uk>
Phil Devall <phil.devall@bbsrc.ac.uk>
- Data consisted of:
 - 86 map populations with position info
 - 336 individual maps
 - 5108 data entries
 - 1383 distinct loci,
 - Chromosome lengths from Schmitz et al.'s (Cytometry 1992).

Pig : Bayesian Integration of Linkage, RH & Cytogenetic Maps

- Obtained porcine RH map data from:
Univ. of Minnesota, INRA & CLDB
- 177 partial maps covering the 20 chromosomes
766 data entries
59 of these pertained to additional distinct loci
- Several of these were singleton maps

Modelling aspects: Estimating Location

- Location of a distinct locus can be modelled as a distribution over an interval. This would allow model any data coming in the interval form (e.g. cytogenetic data) or point information form (e.g. linkage, RH data).
- (Partial) ordering of loci (or maps) could be utilised if available. This typically would mean change in the interval specified above as a stochastic interval.
- An example of that would be use of the “bin-markers” in plants, like Barley in our example. Bin-markers are relatively better studied and hence the data is more reliable.

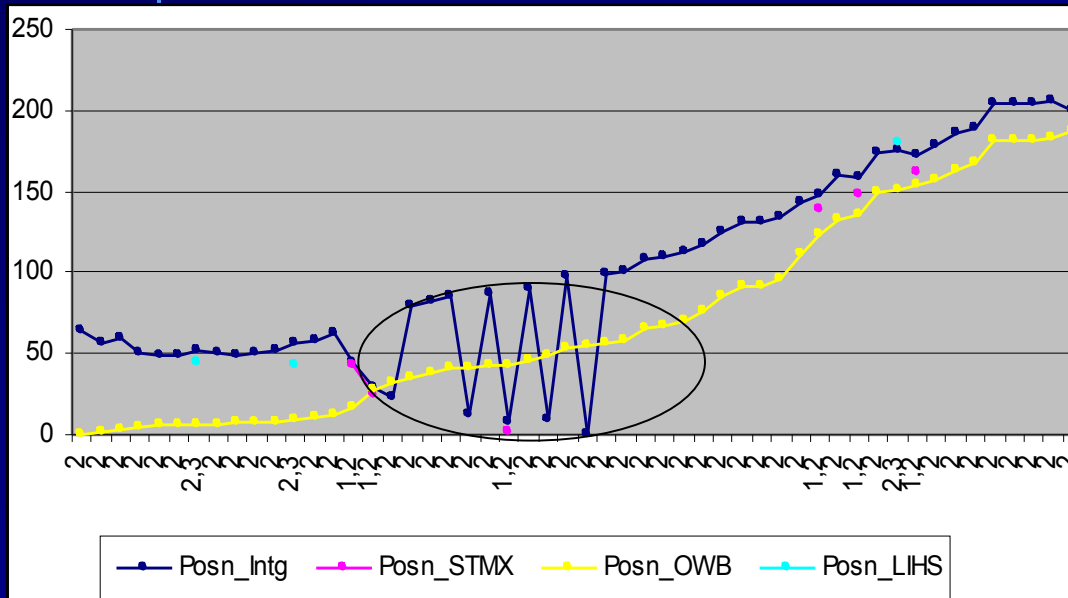
Modelling aspects: Precision of location

- Reliability or precision of data is another aspect that can be incorporated in the model if available.
- We have used different choices for precision parameters,
 - fixed
 - map-type specific
 - map specific
 - locus-type specific
- Markers identifying centromere were selected, these were modelled as location ordered. However this region is also known to be difficult to map. This motivates usage of region specific precision.
- However region for a locus may not be known before hand so in effect we will be defining a mixture model to account to this.

Modelling aspects: Inversion

- Model allows correction of inversion of a map with respect to others (and hence with respect to the integrated map).
- Use of landmarks, like bin-markers, allows the model to explore possible local/partial inversions
- Markers identifying centromere were selected, these were modelled as location ordered.

Barley map integration revisited:



X axis:

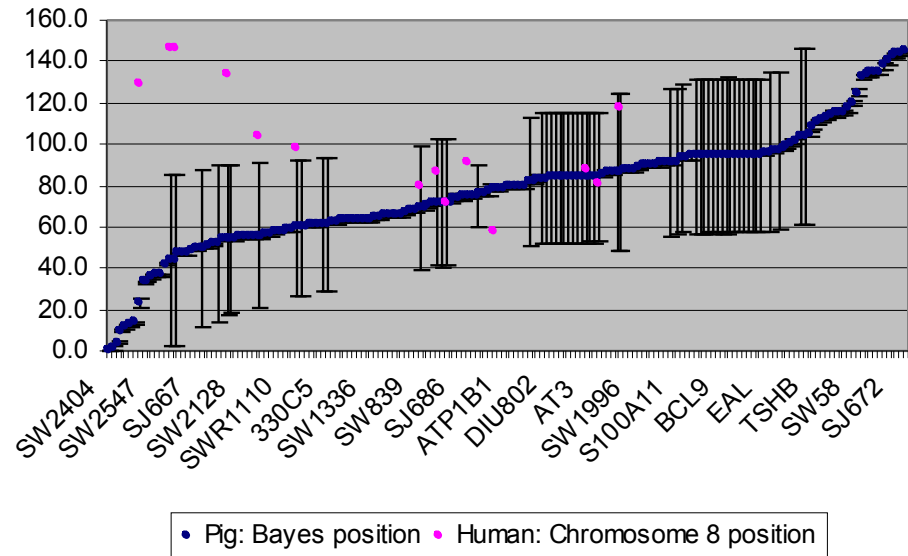
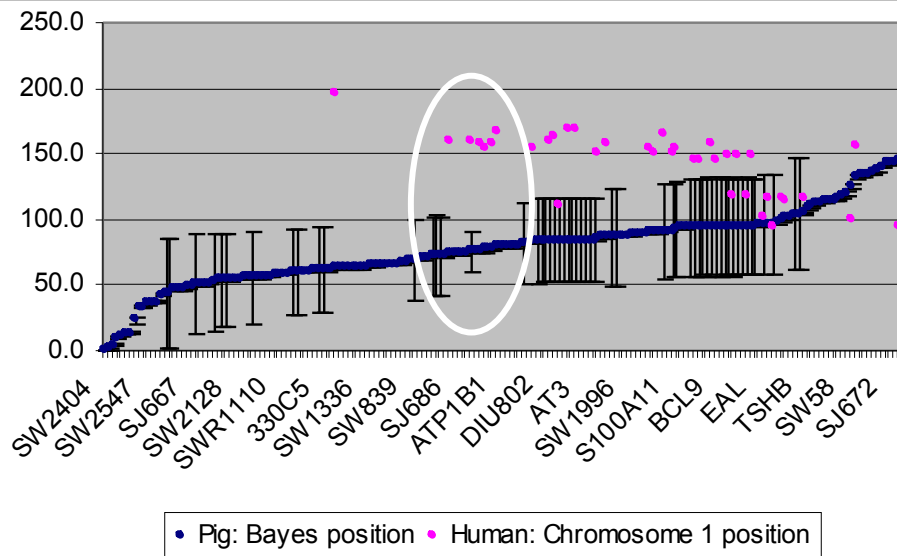
- Markers from OWB map chr 5
- Ordered acc. to position on OWB map,
- X-axis label indicates which map(s) share the same marker
 - 1,2 => STMX & OWB
 - 1,3 => STMX & LIHS
 - 2,3 => OWB & LIHS

Y axis: Positions on maps

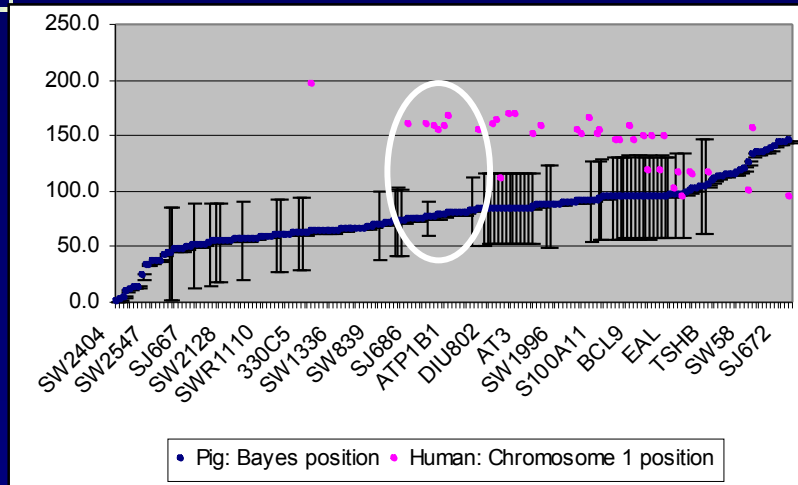
Map based on LIHS population is possible partially inverted with respect to the other two maps.

Comparative genomics & Bayesian Integrated maps: An example

- Consider Bayesian integrated Pig genome using linkage & RH map data.
- Obtained Human sequence based map data (28,974 loci)
- Comparative analysis revealed that gene order conserved across SSC4 compared to HSA1 and HSA8.



Comparative genomics & Bayesian Integrated maps: An example



- The region highlighted in white contains some well-known markers of interest

- A major QTL for fatness and growth, denoted FAT1, has previously been detected on pig chromosome 4q.
- Subsequently the critical region for FAT1 has been reduced to an interval between the RXRG and SDHC loci.
- The orthologous region of FAT1 in the human genome is located on HSA1q23.3 (159.6-163.7 Mb) and harbours many genes.

Comparative genomics & Bayesian Integrated maps: An example

- Possible inferences

Refine estimates of pig loci using human location information

Predict (for Pig) physical locations of interesting genes which are present on the human genome.

Predict (for Pig) locations on relevant genetic map of interesting genes which are present on the human genome.

Comparative genomics: Location information

- In the proposed approach it would be desirable to identify orthologs (between the species of interest) as a separate step
- Conserved regions may be identified subsequently through the model, by
 - firstly identifying regions on reference species with reasonable number of common loci within a narrow range,
 - followed by estimation of probability that these genes appear in similar order based on the posterior distribution,
 - for the regions with high probability of conservation, additional loci from the reference species can be mapped treating the data from that species as a partial map for the target species

Comparative genomics: Additional information

- A variety of data could prove useful for comparison purposes.
- Depending on the underlying hypothesis this could be (fragmented) sequence information, functional information, etc.
- The usefulness of the Bayesian framework are:
 - a variety of data type can be used to model
 - very complex hypotheses can be assessed easily using the posterior distributions from the model

References:

Bayesian Integrated Modelling:

<http://www.mcs.st-andrews.ac.uk/~madhu/>

Other works:

2. W, Collins A, Hobbs M, Khatkar MS, Luo J and Nicholas FW (2007), Mamm Genome, 18, 287-299.
3. Stam P (1993) The Plant Journal, 3, 739-744.
4. Schiex T, Gaspin C (1997) Proceedings of the International Conference on Intelligent Systems for Molecular Biology; ISMB 5, 258-67.
5. Stassen, HH and Scharfetter, C (2000) American Journal of Medical Genetics (Neuropsychiatric Genetics), 96, 108-113.
6. Yap IV, Schneider D, Kleinberg J, Matthews D, Cartinhour S and McCouch SR (2003) Genetics, 165, 2235-2247.