Paper Presentation: Justin Ho (jho@dgp)
Paper Authors: Korbel, et al.

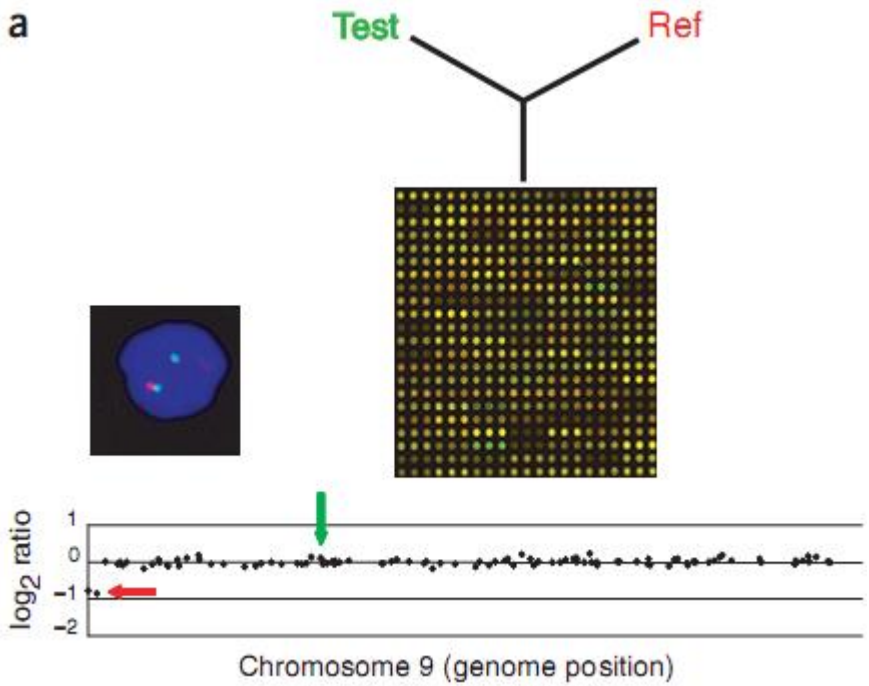
Systematic prediction and validation of breakpoints associated with CNVs in the human genome

http://www.pnas.org/cgi/content/full/104/24/10110

CNVs are important

- **C**opy-**n**umber **v**ariants are form of genetic variation in population
- Genotype-phenotype studies want to know about CNVs
- We only have SOME approximate genomic coordinates of CNV breakpoints

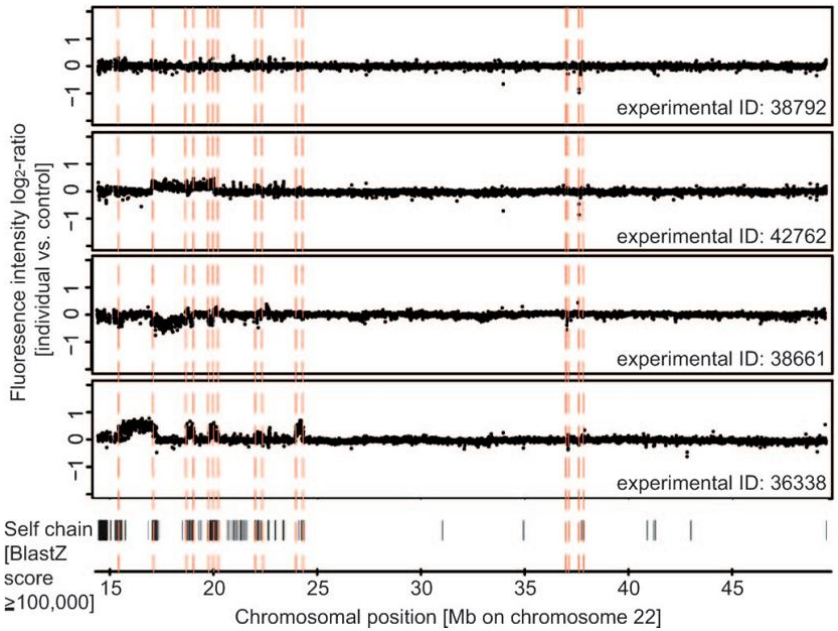
CGHs help detect CNVs



http://www.nature.com/ng/journal/v37/n6s/full/ng1569.html

- Comparative Genome Hybridization shows where two sequences are alike
- Comparing reference and experimental via CGH shows where repeats are

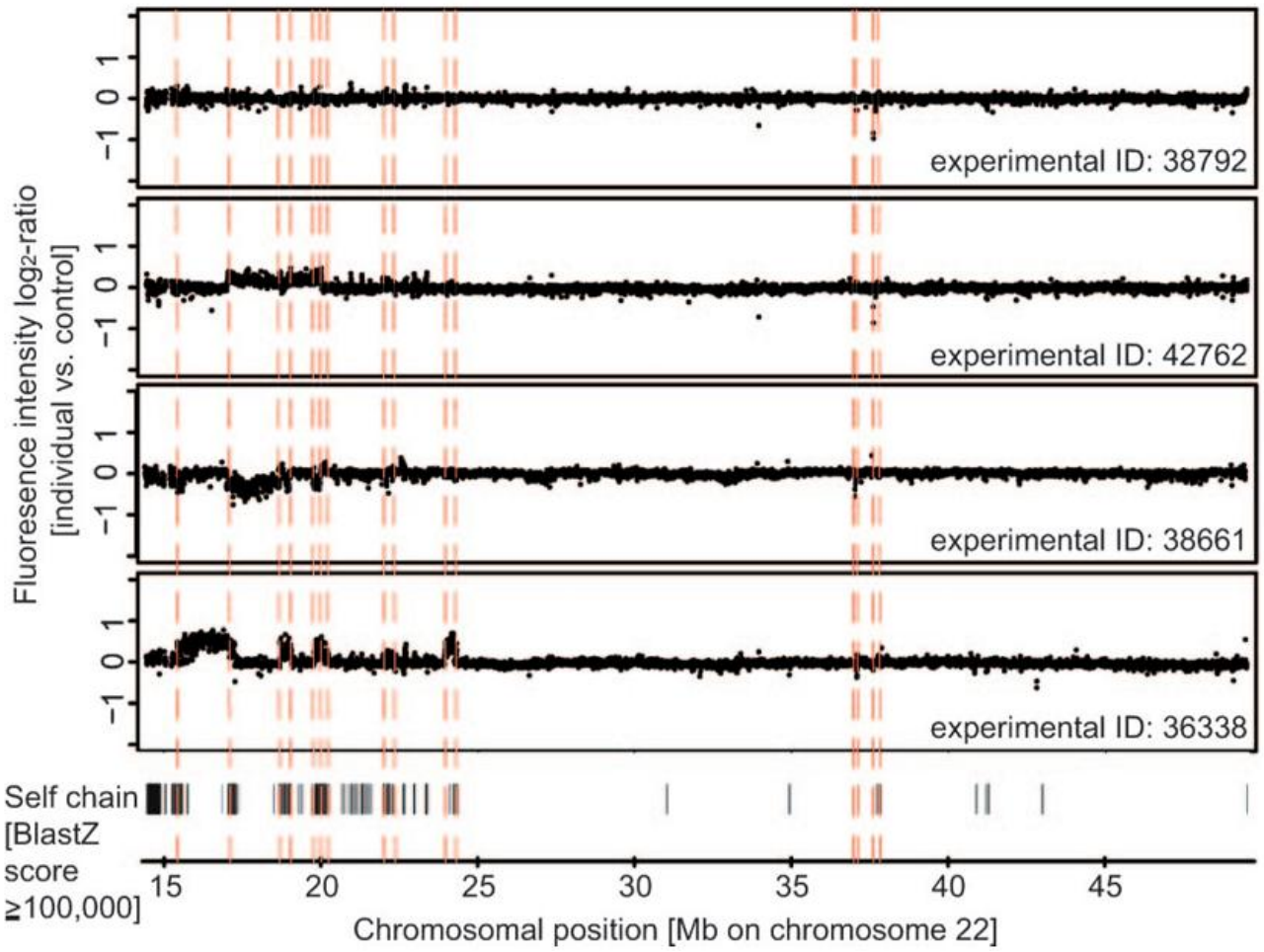
The Observation



- Regions flanked by SDs are susceptible to rearrangements
- Hotspots – prone to CNVs
- "Visible correlation"
  - HighRes-CGH data & genomic sequence features

SDs and Breakpoints are Related

Breakpoints



SDs

HMMs Try To Understand Relationships

Friend lives far away	• Talk about his day on phone
Friend does 3 activities, based ONLY on weather	• Walk, shop, clean
You don't have information about weather	• But you have general trends

HMMs Try To Understand Relationships

Based on what he tells you he did each day, you guess weather

Weather is hidden from you

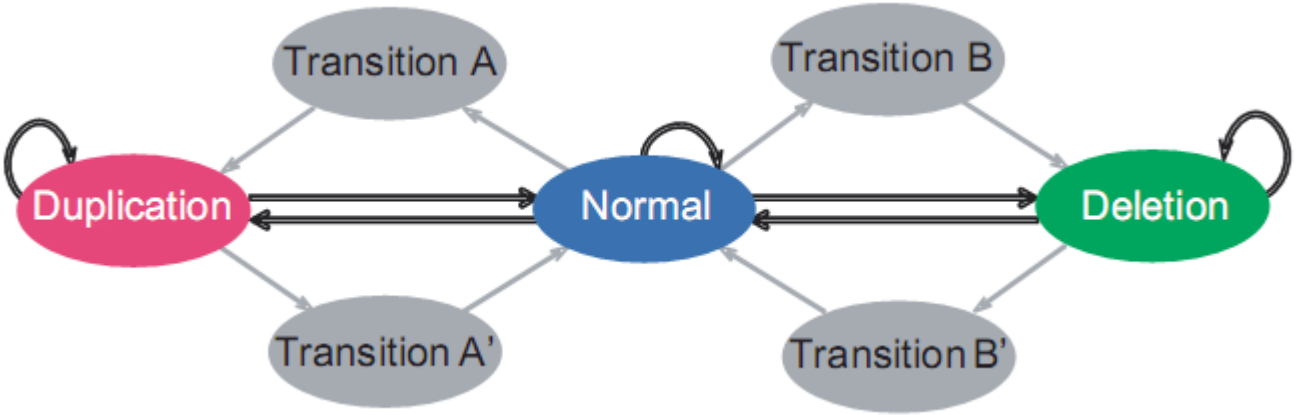
- Either "rainy" or "sunny"

Friend's behaviour is based on weather and chance

HMMs Try To Understand Relationships

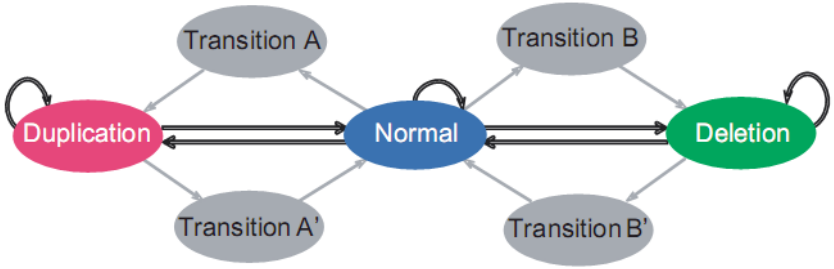
Observables	• His activity (walk, shop, clean)
Hidden state	• Weather (sunny, run)
Parameters	• General weather trends • What he usually likes to do (transition probabilities)

Discrete Bivariate Hidden Markov Model (dbHMM)

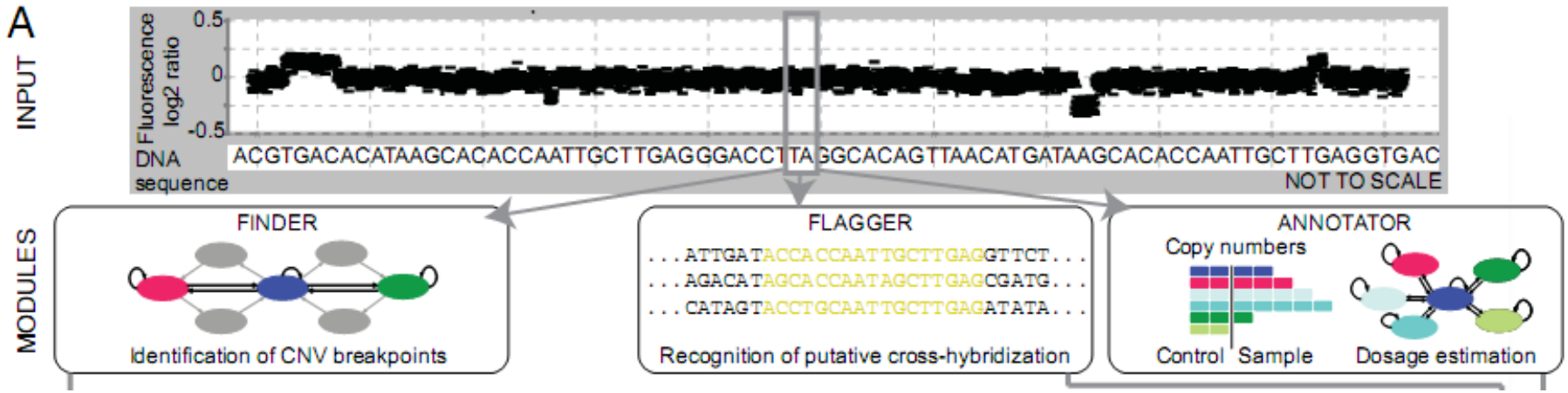


Core dbHMM Considers CGH Data Only

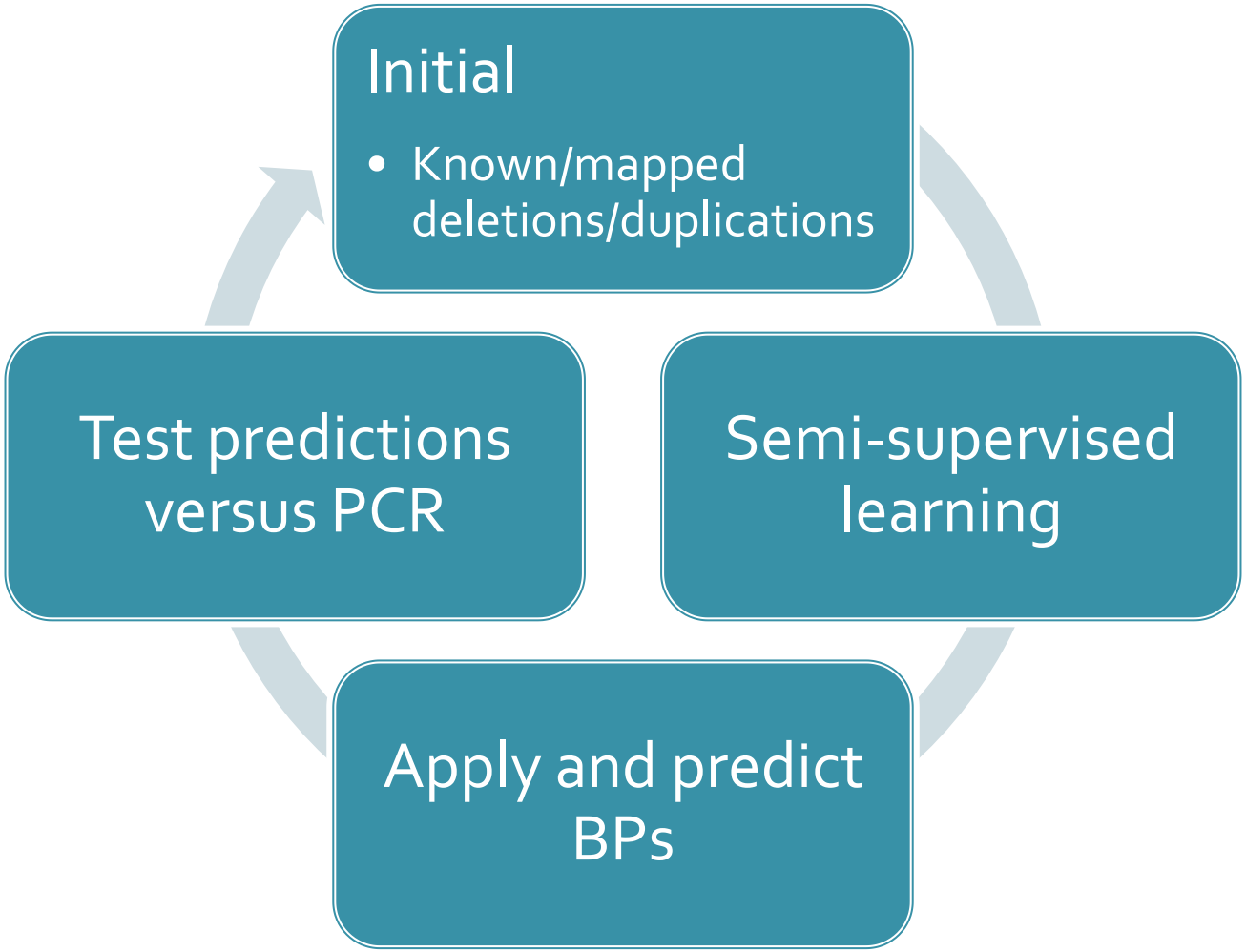
- Core Model == CGH data only
- States
  - Unaffected genomic regions
  - Deletions
  - Duplications
- Transition between states == breakpoint!



Full dbHMM Model Considers CGH + Reference Genome



The Process


Initial
• Known/mapped deletions/duplications
Semi-supervised learning
Apply and predict BPs
Test predictions versus PCR

Paper Results May Not Generalize

232 putative CNVs identified	• 464 breakpoints, flanking 121 duplications and 111 deletions
"More Gold Standard Will Help"	• Agreement: 29% overlap with previously reported CNVs, over 4 individuals • Using Full model: 31% overlap
Refined breakpoints for 36 of the 108 previously mapped locations	• We overlapped, therefore, we can refine existing mappings too!

Paper Results May Not Generalize

Additional "gold standard data" / more data on this model may not improve performance	• 29% to 31% may not generalize • More data != improve accuracy
10 subject pool may be small	• 8 with known defects / 2 "normal" • 91/210 genes did not overlap – may be valid
29% is over POOLED subjects	• What about average per subject?
Over training HMM?	• What is a suitable endpoint?