Identification of regions with common copy-number variations using SNP array

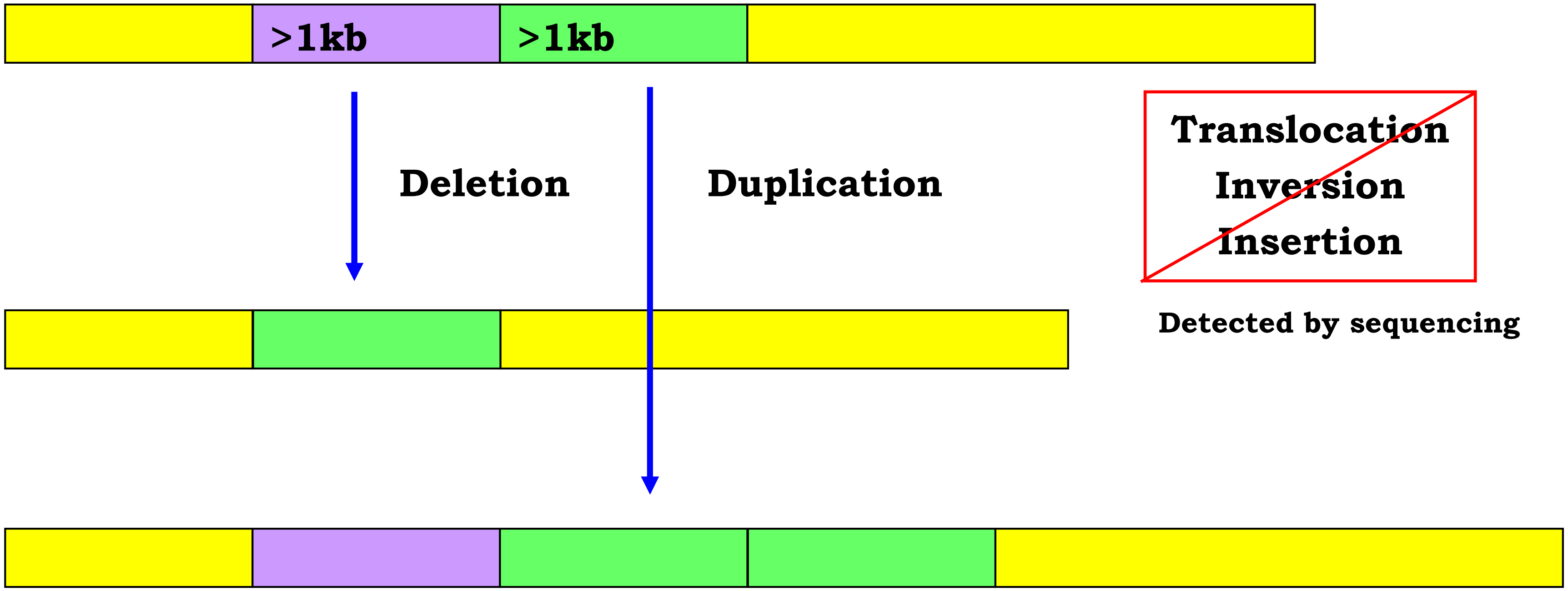
Agus Salim

Epidemiology and Public Health

National University of Singapore

Copy Number Variation (CNV)

- Copy number alteration of a segment of DNA sequence >1kb



~~Translocation~~
~~Inversion~~
~~Insertion~~

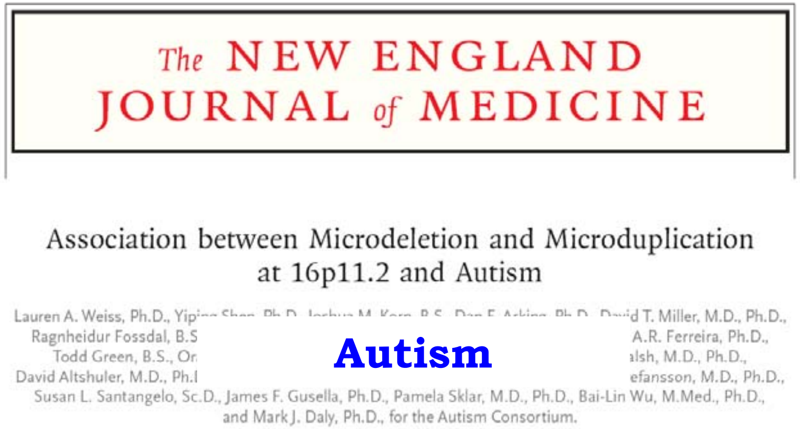
Detected by sequencing

- Vast majority of CNV regions are inherited (Locke, 2006)

The link of CNVs with complex diseases

- **Increasing evidence to link CNVs with complex diseases**

Autism

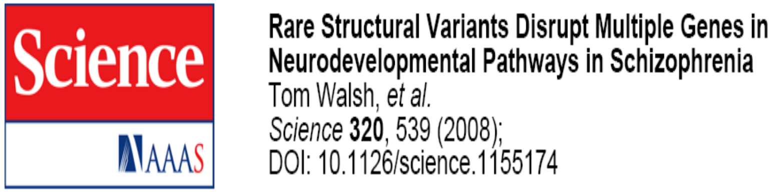

The NEW ENGLAND JOURNAL of MEDICINE

Association between Microdeletion and Microduplication at 16p11.2 and Autism

Lauren A. Weiss, Ph.D., Yiping Shen, Ph.D., Joshua M. Korn, B.S., Dan E. Arking, Ph.D., David T. Miller, M.D., Ph.D., Ragnheidur Fossdal, B.S ... A.R. Ferreira, Ph.D., Todd Green, B.S., Or ... alsh, M.D., Ph.D., David Altshuler, M.D., Ph.I ... efansson, M.D., Ph.D., Susan L. Santangelo, Sc.D., James F. Gusella, Ph.D., Pamela Sklar, M.D., Ph.D., Bai-Lin Wu, M.Med., Ph.D., and Mark J. Daly, Ph.D., for the Autism Consortium.

Schizophrenia

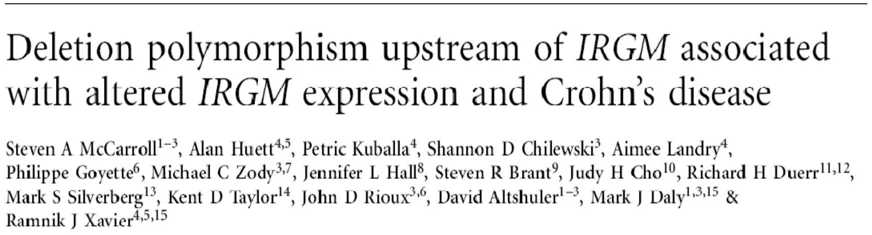
EXTENDED PDF FORMAT SPONSORED BY usb www.usbweb.com


Science AAAS

Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia
Tom Walsh, *et al.*
Science **320**, 539 (2008);
DOI: 10.1126/science.1155174

nature genetics

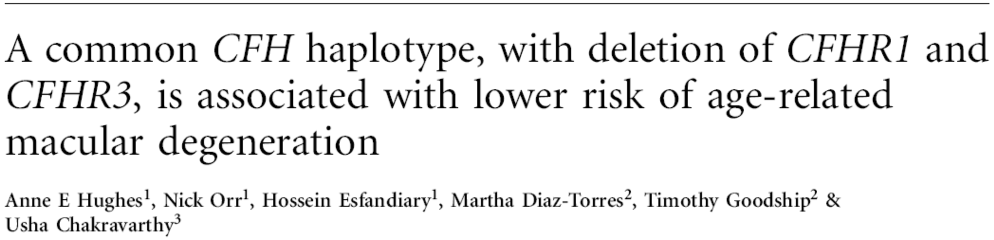
Crohn’s disease


Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn’s disease

Steven A McCarroll[1–3], Alan Huett[4,5], Petric Kuballa[4], Shannon D Chilewski[3], Aimee Landry[4], Philippe Goyette[6], Michael C Zody[3,7], Jennifer L Hall[8], Steven R Brant[9], Judy H Cho[10], Richard H Duerr[11,12], Mark S Silverberg[13], Kent D Taylor[14], John D Rioux[3,6], David Altshuler[1–3], Mark J Daly[1,3,15] & Ramnik J Xavier[4,5,15]

nature genetics

Age related macular degeneration


A common *CFH* haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration

Anne E Hughes[1], Nick Orr[1], Hossein Esfandiary[1], Martha Diaz-Torres[2], Timothy Goodship[2] & Usha Chakravarthy[3]

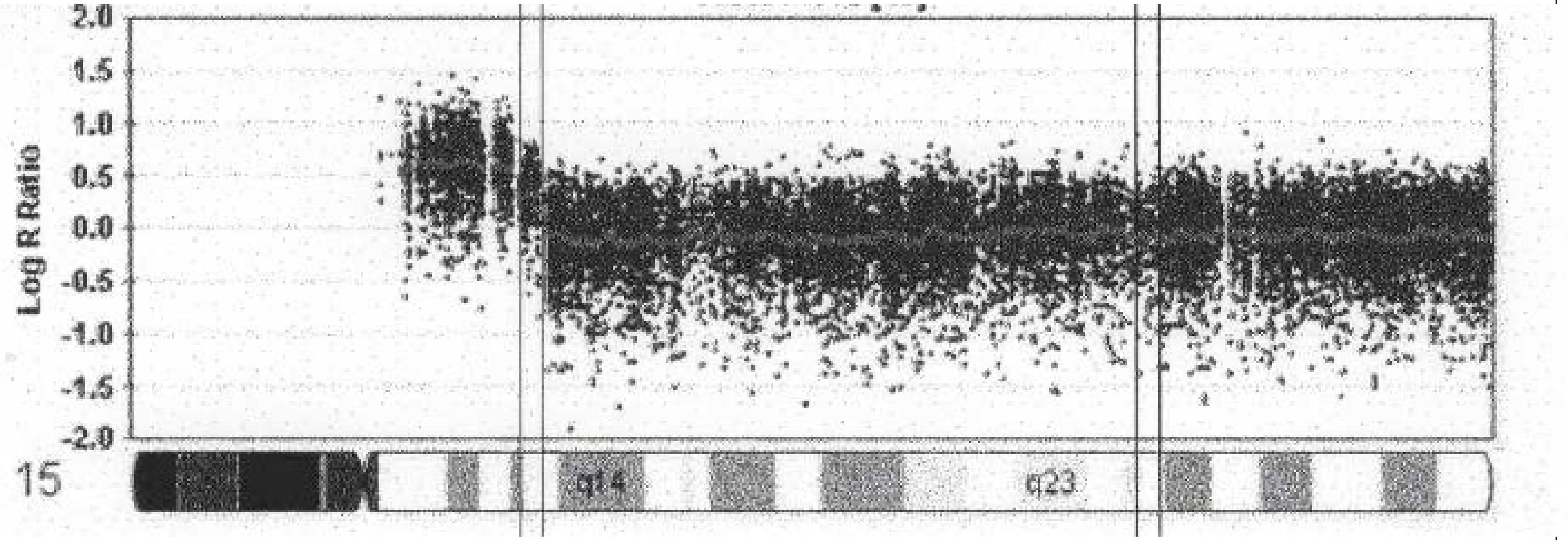
Platforms for CNV detection

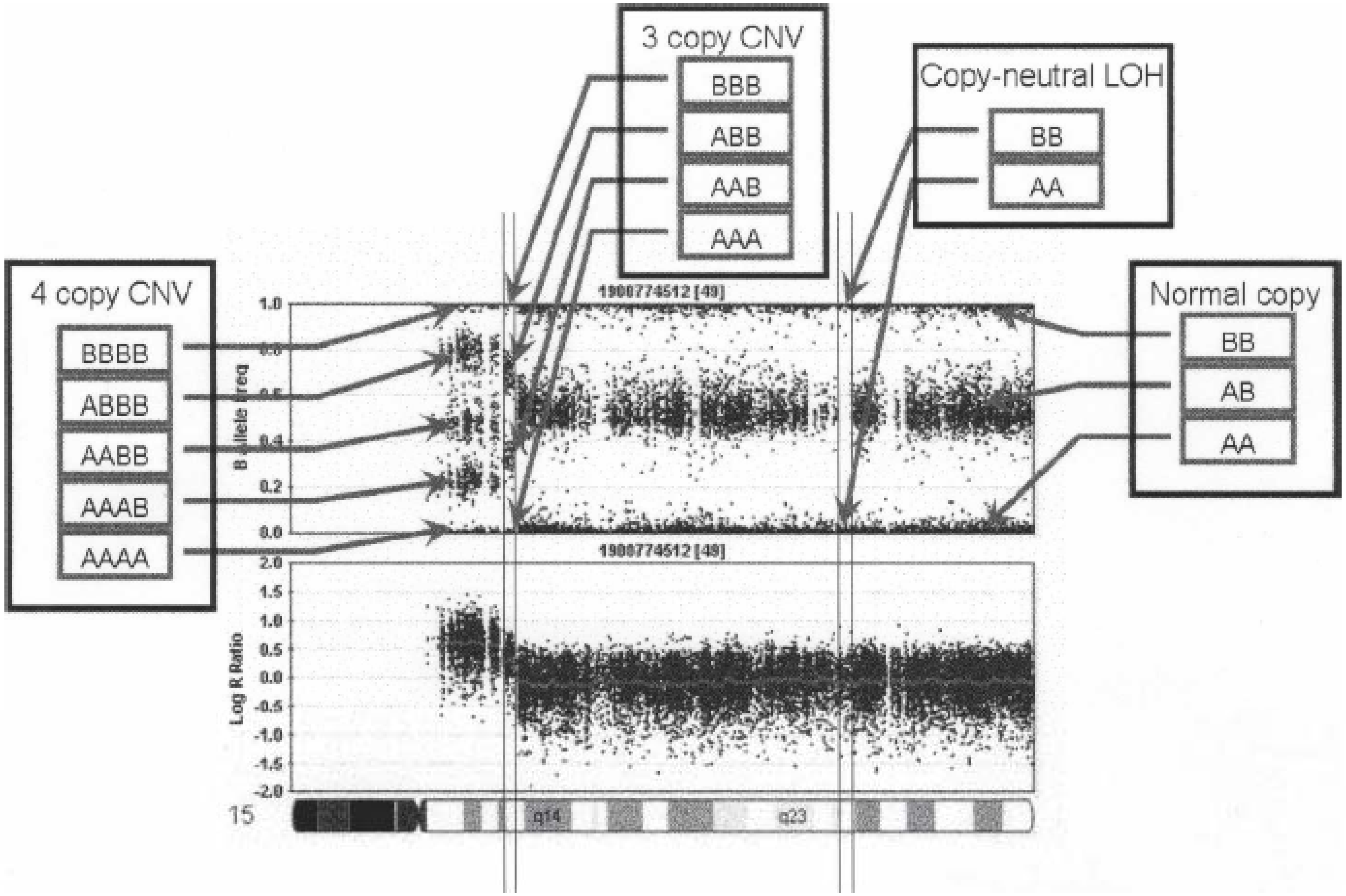
- Array CGH
  - NimbleGen (2.1M probes), Agilent (250K?)
- *SNP array*
  - Illumina 1M, Affy 6.0 (1.8M probes)
- Sequencing

SNP array

- Originally built to detect Single Nucleotide Polymorphism.
- Main output: total signal intensity at each probe.
- Total intensity (R) = signal intensity A + signal intensity B
- Log R ratio = $\log_2$ (observed R/ Expected R)
  - Measures copy-number changes relative to the reference genome.
- BAF = normalized measure of relative signal intensity (B/A)

Log R Ratio





4 copy CNV

BBBB

ABBB

AABB

AAAB

AAAA

3 copy CNV

BBB

ABB

AAB

AAA

Copy-neutral LOH

BB

AA

Normal copy

BB

AB

AA

Detection of CNV regions

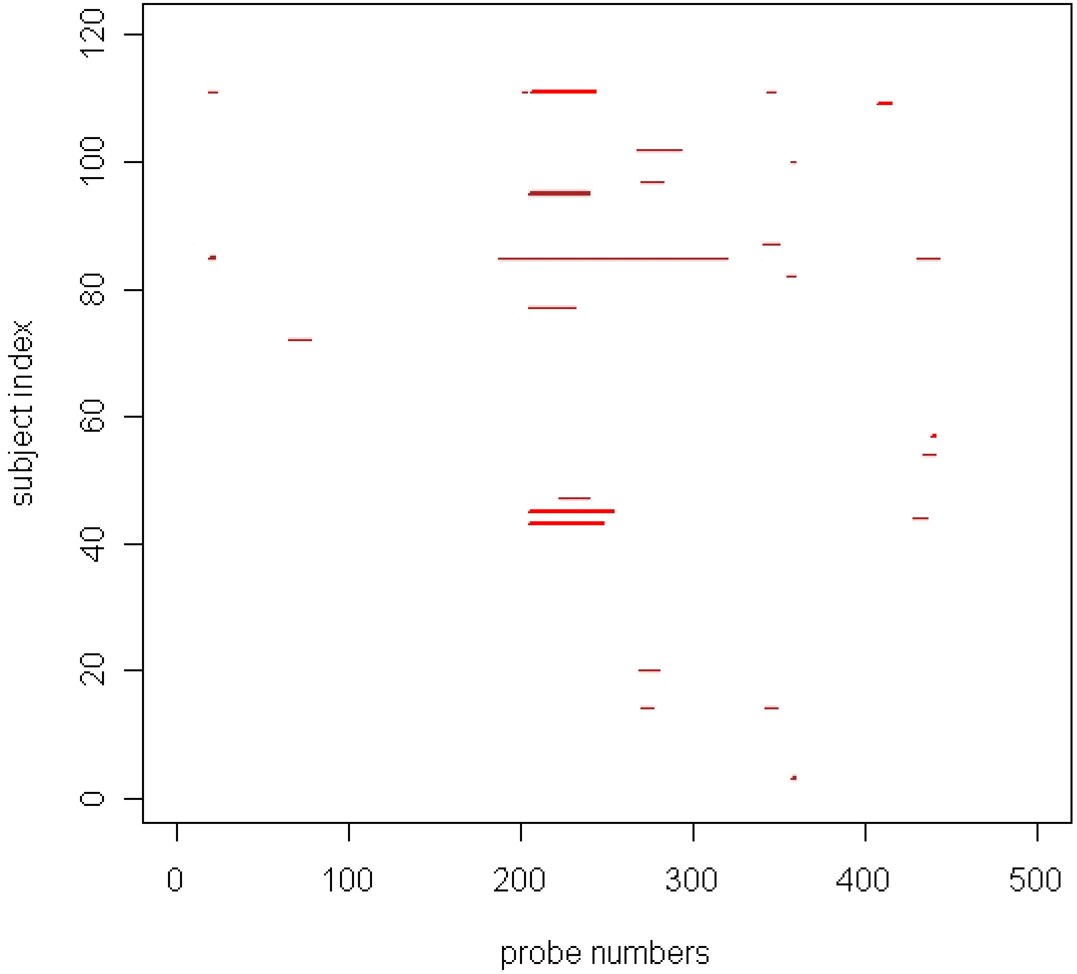
- PennCNV, QuantiSNPs
  - Use Log R ratio and B allele frequency (BAF)
  - Based on hidden Markov model (HMM)
  - Hidden state = copy number (0,1,2,3,4)
- DNACopy (CBS algorithm)
  - Find change points in intensity data (Log R ratio)
  - Segments genome into multiple regions defined by the change points.
  - No copy number calls.

Sample-by-sample approach

- Using HMM or CBS algorithm, CNV regions are detected for each individual separately.
- Sample-by-sample detection is reasonable for rare CNV where individual-specific regions are expected.
- However, common CNV regions occur at relatively the same genomic regions across individuals (e.g, McCaroll et al., 2008).

Individual CNV regions

- For each region we have:
  - Start, end, confidence score
- Confidence score = a score to reflect confidence that the detected region truly exists.
- How to determine common boundaries for common CNV regions?



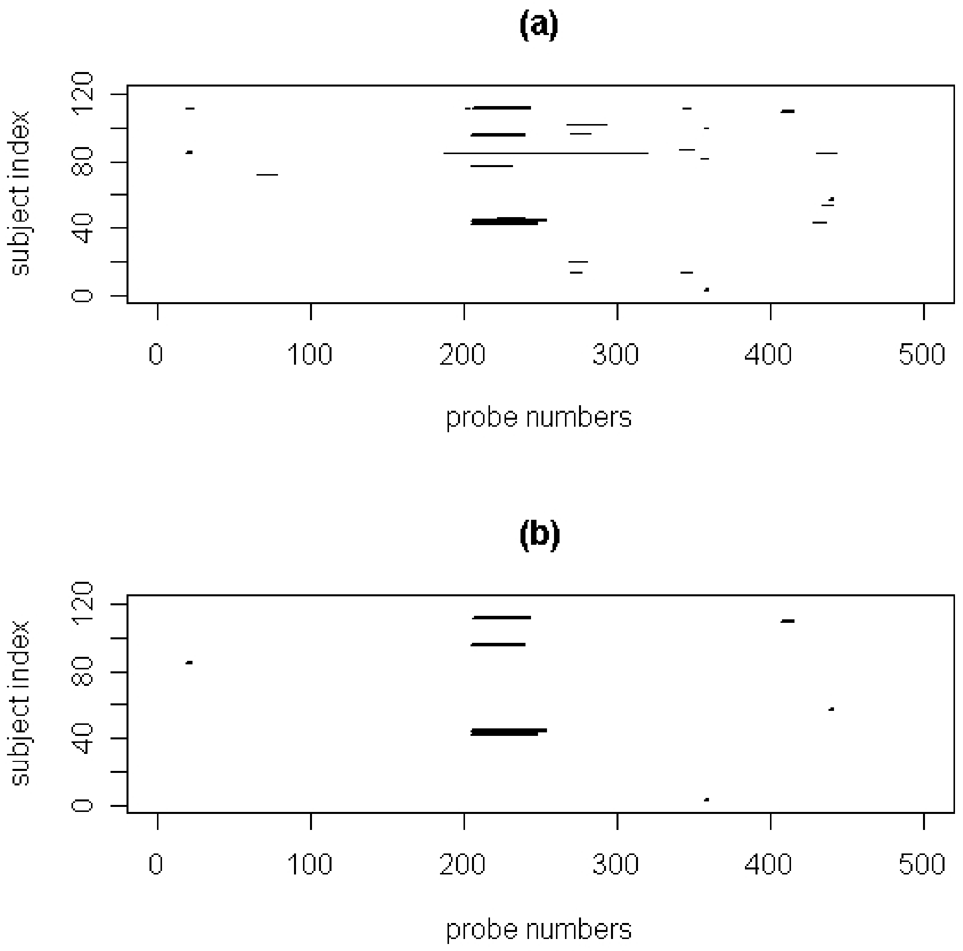
Within and between individuals reliability

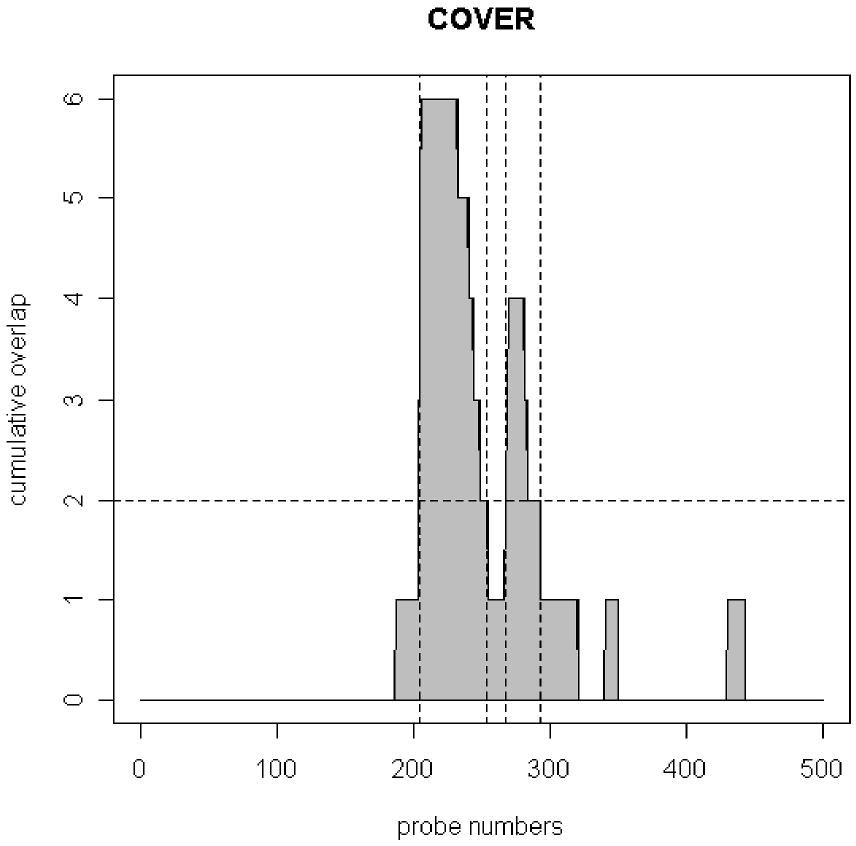
- Reliability of (within) an individual regions is reflected by the confidence scores.
- Between-individual reliability of the region is reflected by how many other individuals have CNV in the region.

Cumulative Overlap of Very Reliable Regions (COVER)

- Key idea: Common CNV regions occur at almost the same genomic regions.
- Choose regions where multiple individuals have CNV.
- But need to do something about the unreliable calls with low confidence scores.

COVER

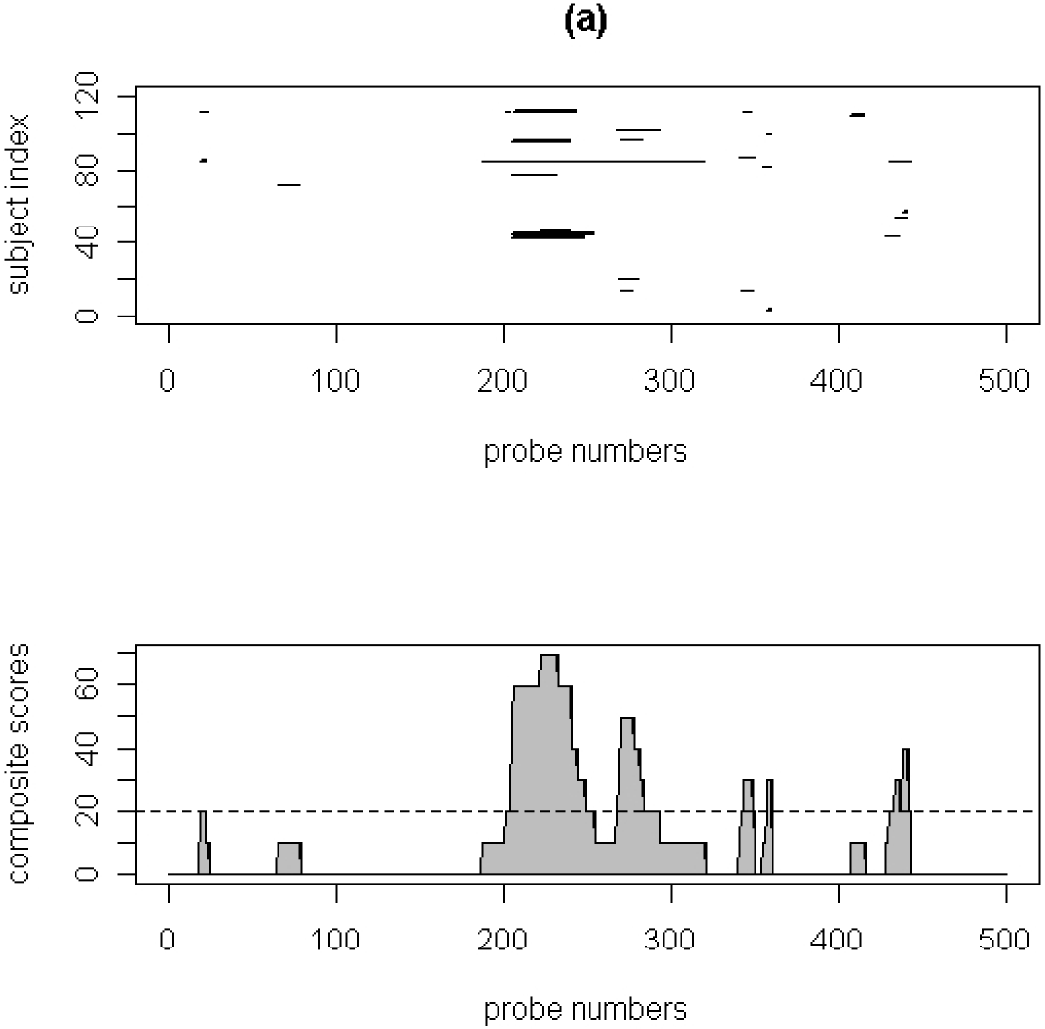

(a)
subject index
0
40
80
120
0
100
200
300
400
500
probe numbers
(b)
subject index
0
40
80
120
0
100
200
300
400
500
probe numbers


COVER
cumulative overlap
0
1
2
3
4
5
6
0
100
200
300
400
500
probe numbers

Composite Scores (COMPOSITE)

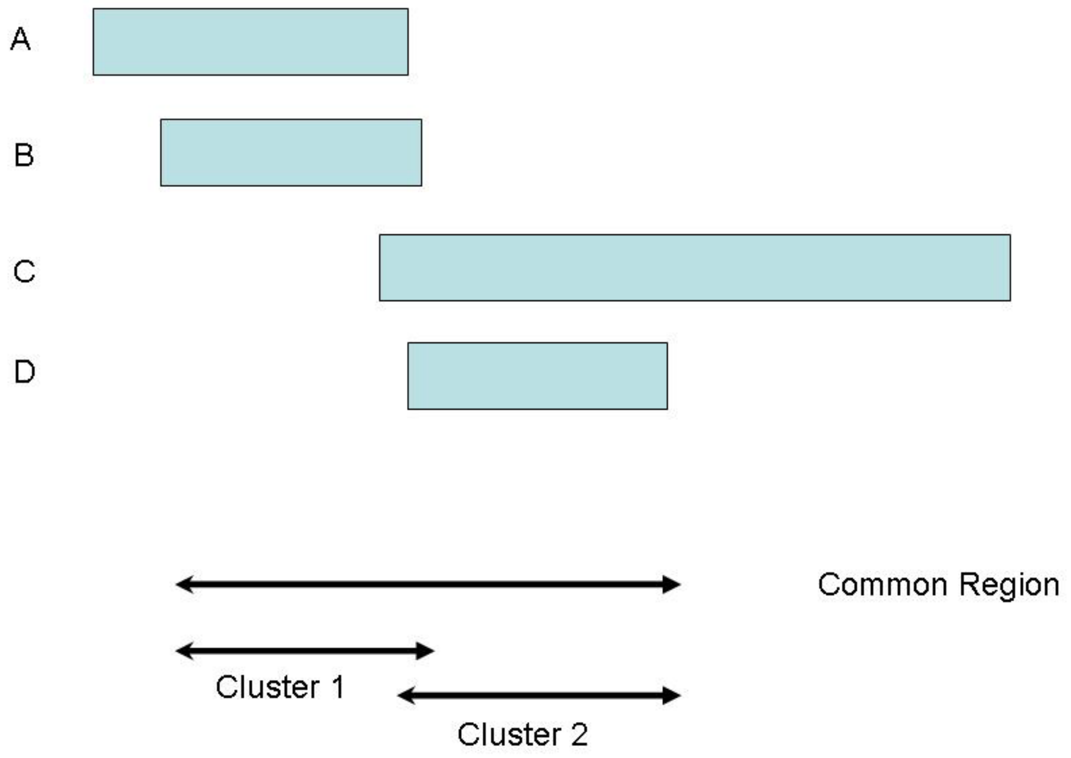
- Using COVER, only ‘very reliable’ regions are selected. What about regions with ‘average’ reliability but yet occur in many individuals?
- COVER put more emphasis on Within-individual reliability.
- Composite Scores = Sum of confidence scores at each probe, give more emphasis on between-individual reliability.

COMPOSITE


(a)
subject index
0
40
80
120
0
100
200
300
400
500
probe numbers
composite scores
0
20
40
60
0
100
200
300
400
500
probe numbers

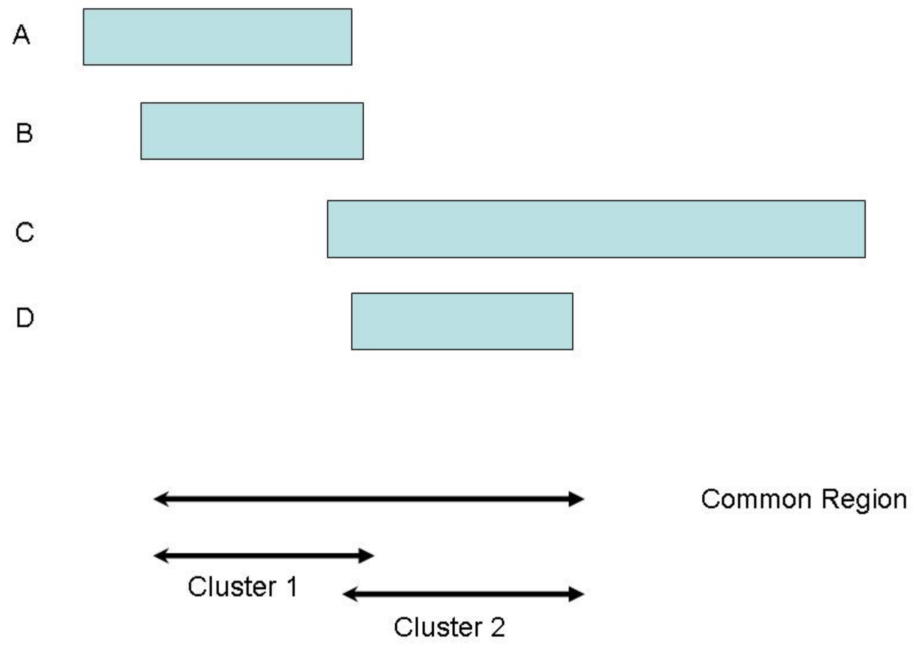
Clustering of Individual Regions

- Two clusters of region form one common region.

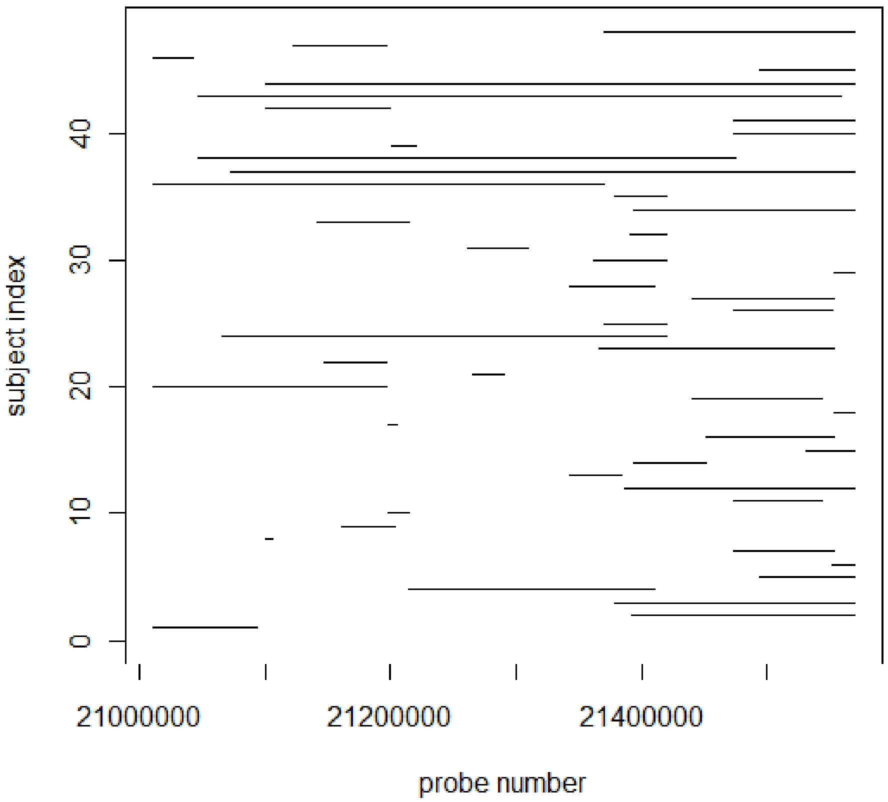


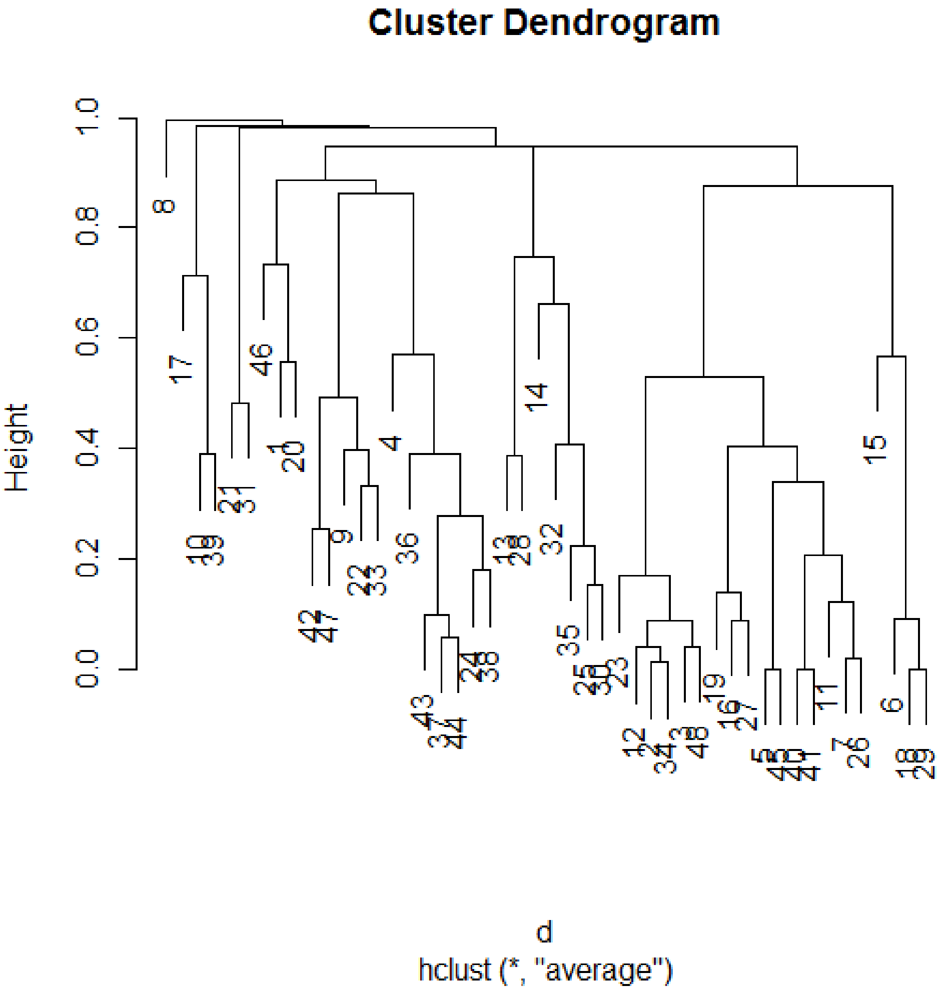
Hierarchical Clustering of Individual Regions

- Jaccard Coefficient
  - Similarity (A,B) = number of bases shared/length of union of A and B (bases).
  - Linkage: single, average, complete.

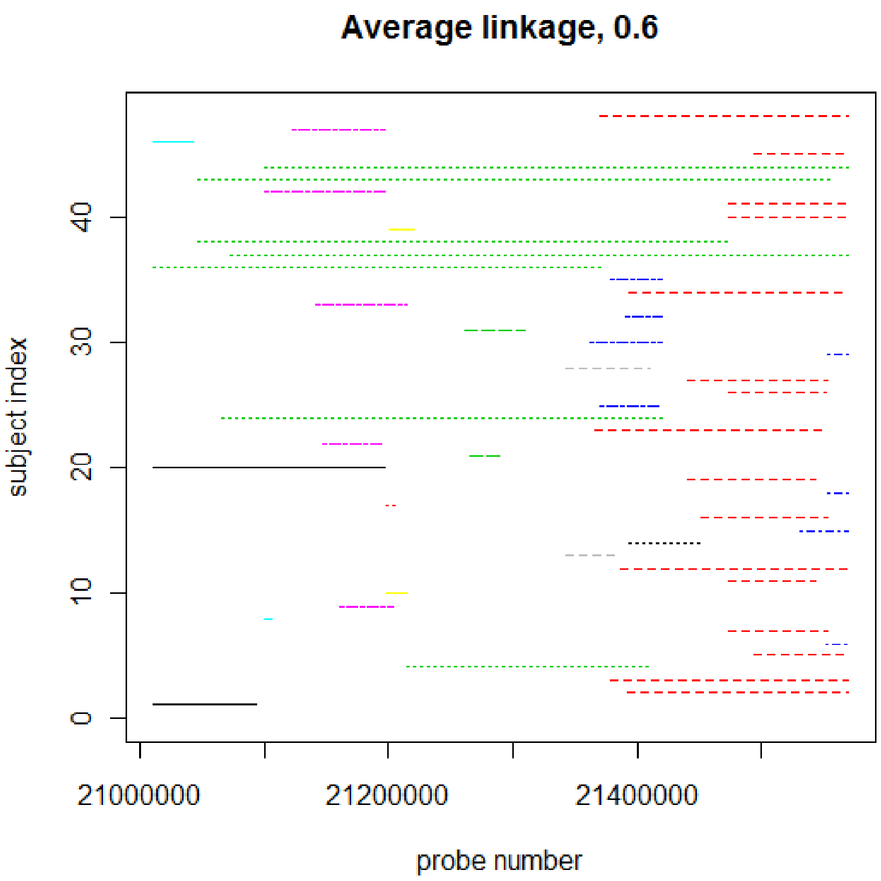


CLUSTER examples

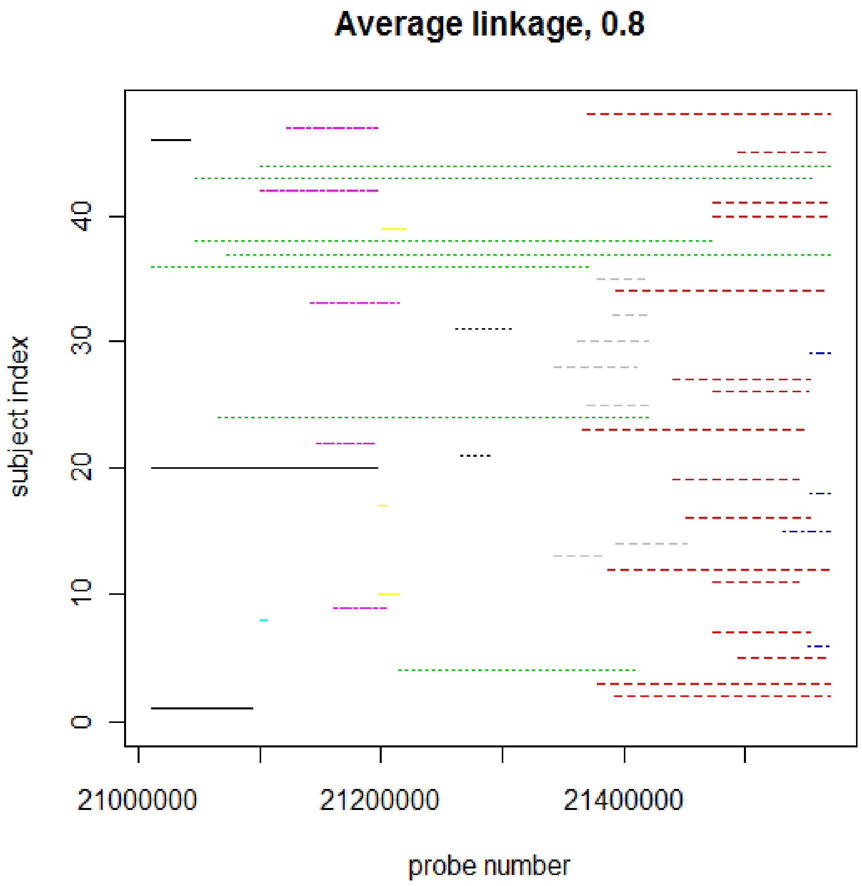




Average linkage



13 groups



9 groups

Cut off @	Complete linkage	Single linkage	Average linkage
	No. of clusters		
0.3	28	24	27
0.4	23	16	21
0.5	21	11	17
0.6	17	6	13
0.7	14	4	12
0.8	12	2	9
0.9	10	2	6
0.95	9	2	5
0.99	9	1	2

Comparison of Methods

- COVER
- COMPOSITE
- Both can be refined using CLUSTER.

- What are we comparing?
  - Proportion of identified CNV regions violate HWE
  - Discordant rates between the identified CNV regions and sequencing results (when available)

HWE of common CNV regions

- The number of copies in offspring = number of copies in father + number of copies in mother.
- For diallelic CNV with deletion only (CN=0,1,2),
  - Population frequency of 0 copy allele = p
  - Population frequency of 1 copy = 1-p = q
- Under random mating, the frequency of subjects with:
  - ‘0’ copy = p^2
  - ‘1’ copy = 2pq
  - ‘2’ copies = q^2

- For diallelic CNV with duplication only (CN=2,3,4),
  - Population frequency of 1 copy allele = p
  - Population frequency of 2 copy = 1-p = q
- Under random mating, the frequency of subjects with:
  - '2' copies = p^2
  - '3' copies = 2pq
  - '4' copies = q^2
- For multiallelic CNV, HWE test cannot be applied to the unphased copy number calls.

Discordant rates

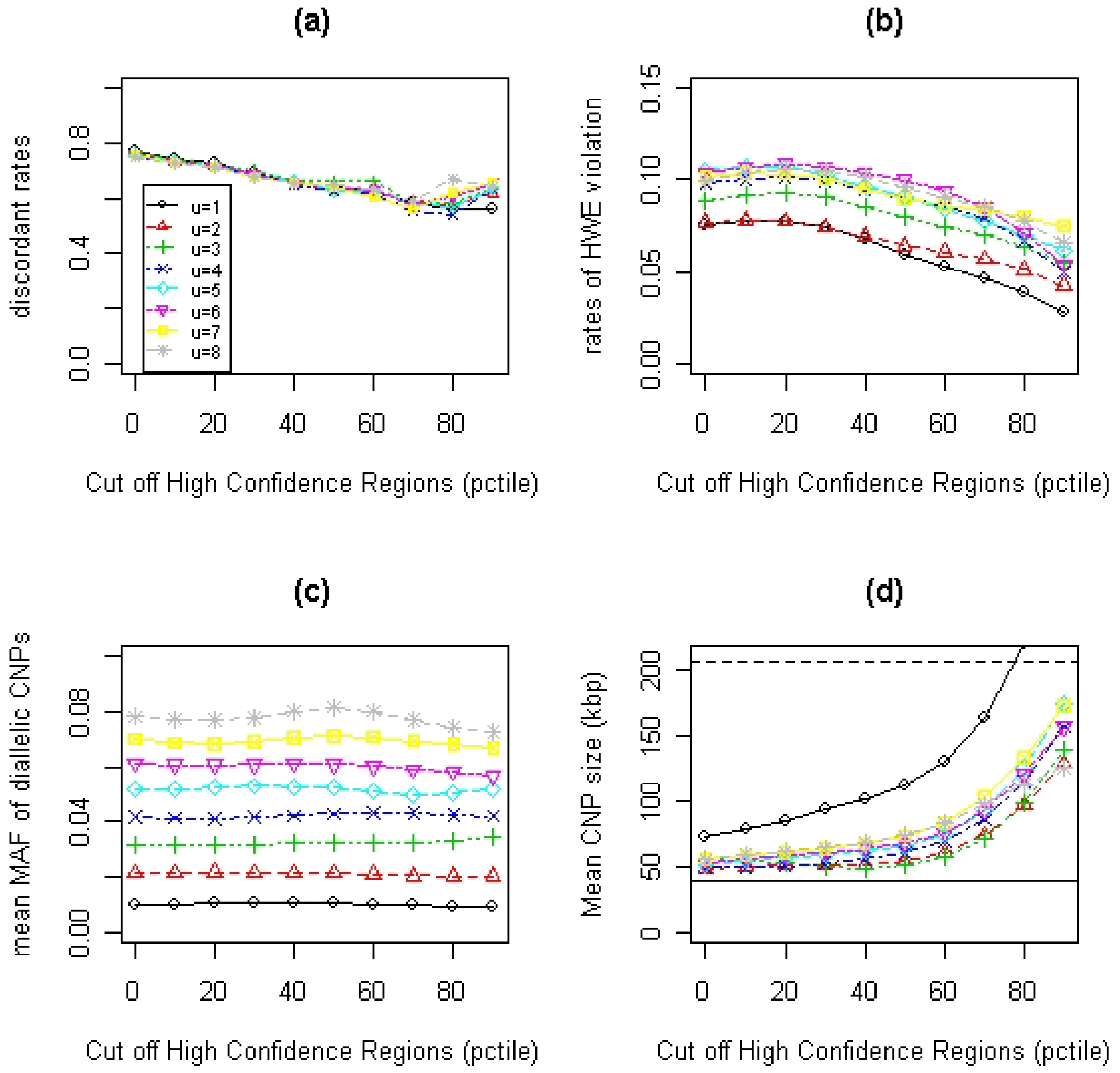
- What is the proportion of CNV regions identified in a sample that 'overlap' with sequencing result of the same sample?
- 'overlap' = the two regions overlap above certain threshold.
  - E.g, Jaccard coefficient >= 0.5

Application

- 112 HapMap samples (CEU, CHBJPT, YRI) on Illumina 1M platform;
- 83 of them are unrelated individuals;
- 8 of them were sequenced by Kidd et al (2008)
- Statistics of interest:
  - average discordant rates in the eight sample,
  - Prop of diallelic CNV do not follow HWE (among unrelated individuals)
  - Average minor allele frequency of diallelic CNVs
  - Average size of CNV regions

COVER results

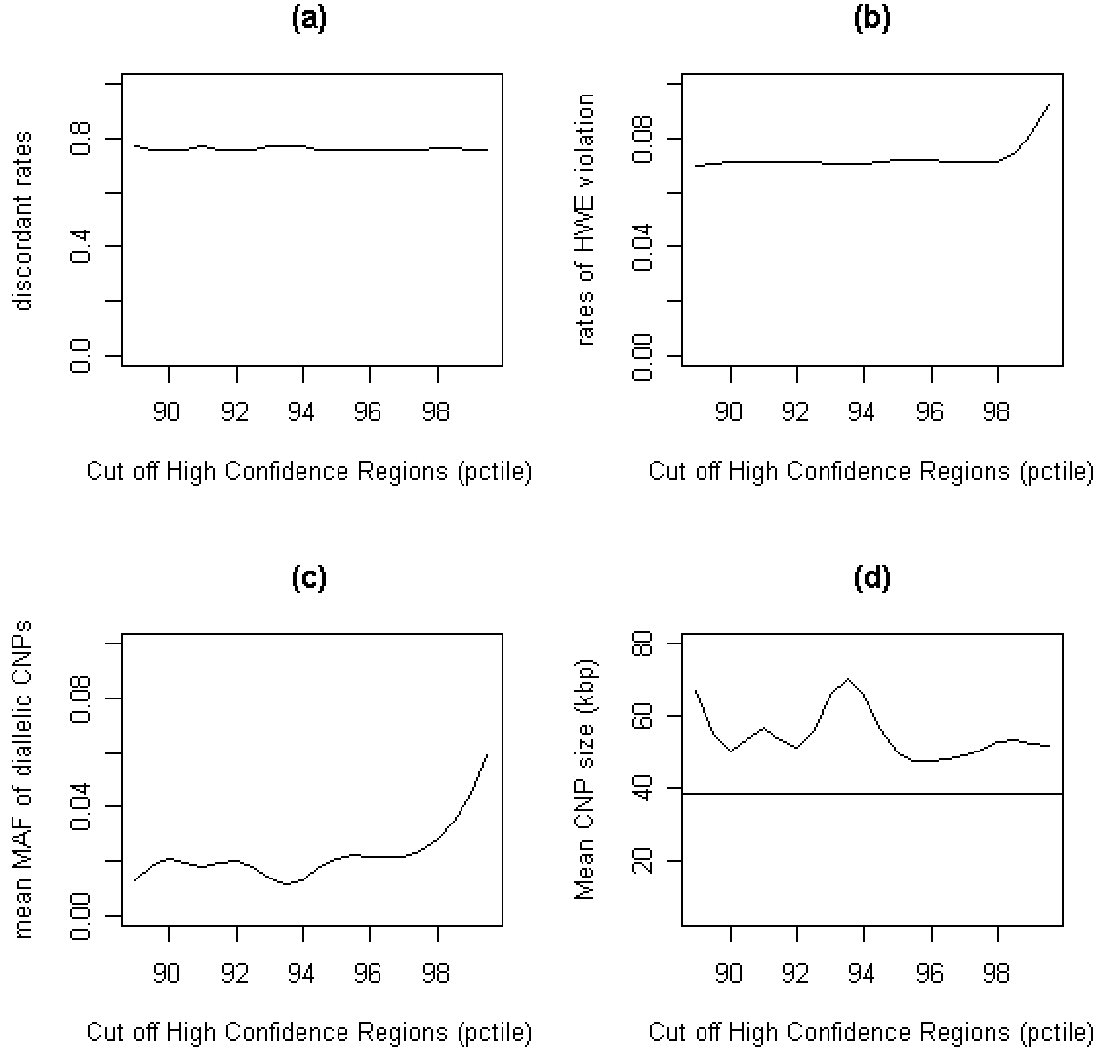
a) Discordant Rates
b) HWE violation
c) MAF
d) Avg Size



- Discordant rates can be lowered by choosing individual regions with higher confidence scores.
- > 90% of identified regions follow HWE.

COMPOSITE Results

a) Discordant Rates
b) HWE violation
c) MAF
d) Avg Size



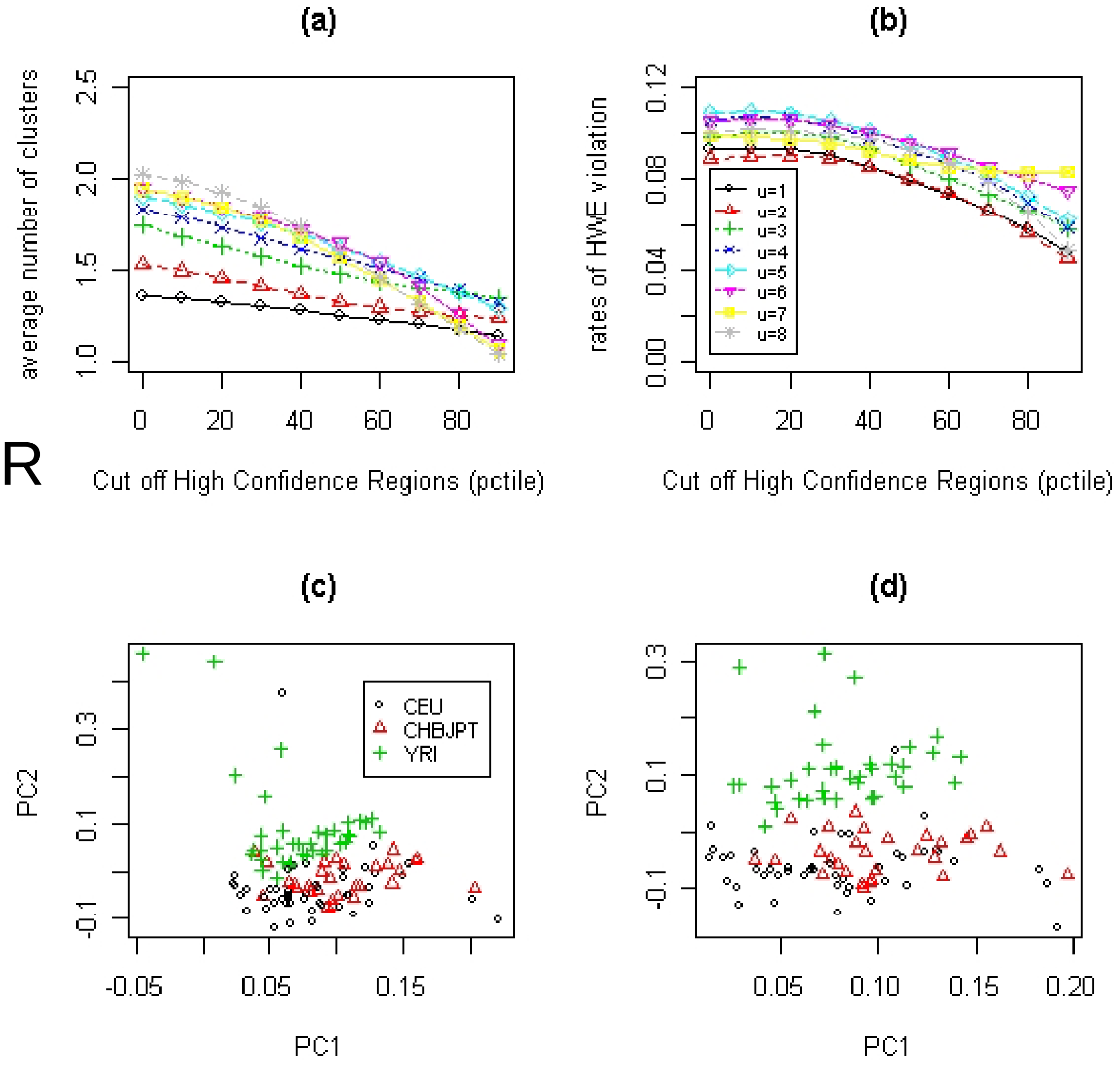
- COVER performs better than COMPOSITE; in particular it provides more control on discordant rates and HWE.
- Within-individual reliability is more useful when identifying common CNV regions.
- Discordant rates is high.
  - Discordant rates between McCaroll et al (2008) results (Affy 6.0) and sequencing is also around 50-70%.

CLUSTER refinement of COVER

- Cluster

cut-off = 0.6

a) Avg num clusters
b) HWE violation
c) PCA without CLUSTER
d) PCA after CLUSTER



Software

- cnvpack R package
  - available at http://www.meb.ki.se/~yudpaw

Conclusion

- COVER performs better than COMPOSITE; in particular it provides more control on discordant rates and HWE.
- CLUSTER provides potential refinement to COVER.
- Our approach only requires: start, end and confidence scores of individual CNV regions.
  - These are output of CNV detection software.

Collaborators

- National University of Singapore
  - Teo Shu Mei
  - Chia Kee Seng
  - Ku Chee Seng
- Karolinska Institute, Sweden
  - Yudi Pawitan
- Universita Milano Bicocca, Italy
  - Stefano Calza