

Colon CFR Whole Genome Association Study

Joint Breast-Colon CFR Steering Committee
Meeting

Bethesda Dec 2005

Graham Casey, Dave Duggan (TGen), Ellen
Goode, Duncan Thomas, John Potter, John Hopper,
Robert Haile on behalf of the Colon CFR

Colon CFR WGA Study Design

- **Two Stage design:**

Stage I: genome-wide scan of 500,000+ SNPs in population-based CRC cases and unrelated controls

Stage II: genotype only “significant” SNPs from stage I in CRC cases and familial controls

Colon CFR WGA Study Design

- **Stage I Study Population:**

Drawn from population-based centers

Available DNA and Epidemiology Questionnaire

Non-hispanic whites

Reduce clinical and genetic heterogeneity

(known MSS or MSI-L cases only- exclude
MMR mutation carriers and MSI-H cases)

Genetically enriched (FH, <50)

Study Population

Methods

Aim 1

Population-based case-control

952 CRC probands and 1020 controls
FH+ and <50 yr over represented

Genotyping:

500K Affymetrix SNP Chip

Statistical analysis:

SNP- and Haplotype-based
Hierarchical modeling incorporating
prior genomic data

- **Population-based case-control study:**

Disadvantage : possible population stratification

Advantage: greater power to detect differences

Affymetrix 500K GeneChip

- It is estimated that LD in the north European population of the USA typically extends 60kb from common alleles. The 500K chip has marker spacing every 5.8kb 30% average heterozygosity and 21% average minor allele frequency
- Approx 80% genome coverage in caucasians ($r^2 > 0.8$: 90%, $r^2 > 0.5$) when using a multi marker approach

Genome-Wide Genotyping Technologies.

	Affymetrix 500K GeneChips*	Illumina 250K Infinium 2 (Expected Jan 2006)	Perlegen 275K LD Panel α	ParAllele 20K cSNP Panel \grave{a}
Selection Strategy	LD & HapMap prioritized; extra probes for select regions	Haplotype tagSNPs	LD tagSNPs	Coding SNPs
% of genome coverage	63% ($r^2 > 0.8$)** 81% ($r^2 > 0.5$)	65% ($r^2 > 0.8$) 80% ($r^2 > 0.5$)	70% ($r^2 > 0.8$) 86% ($r^2 > 0.5$)	nd
No. of SNPs	504,040	261,000 (over 100,000 singletons \wedge)	275,960	20,000
SNPs in genes	28,021 \dagger	11,000	nd	20,000
Average MAF	0.22	0.26	nd	nd
Average heterozygosity	30%	nd	nd	nd
Mean spacing	5.8kb	10kb	nd	nd
Median spacing	2.8kb	nd	nd	nd
DNA required	500ng	750ng	10ug	4ug
Cost (per sample)	\$800	Not available	(IP issues)	\$440

*Greg Marcus, Ph.D. (Affymetrix), personal communication; Sa rah Murray, Ph.D. (Illumina), as presented at SNP2005 (UK, September 22-24, 2005); These are expected values. The product was still in development at the time of this presentation; α Hinds et al. 2005; \grave{a} MegAllele Genotyping Human 10K cSNP Panel I (www.parallelebio.com/products-services/genotyping-products.html). Panel II will be available prior to the start of this proposal (Jay Kaufman, Affymetrix/ParAllele, personnel communication); **Using a multi-marker approach, ~80% genome coverage ($r^2 > 0.8$; $> 90\%$, $r^2 > 0.5$). Manuscript in preparation (Greg Marcus, Affymetrix, personal communication, in collaboration with the Broad Institute). \wedge Singletons = SNPs not in LD with any other SNP(s) and thus are not truly tagSNPs; \dagger includes exons, predicted exons, or ESTs; 225,090 SNPs in or within 10kb of a gene; 30,349 of the 33,653 genes in Ensembl (90%) have ≥ 1 SNP within 10kb, 18,852 ≥ 5 SNPs within 10kb (56%).

Colon CFR WGA Study Design

- **Stage I Study Population:**

Drawn from population-based centers

Available DNA and Epidemiology Questionnaire

Non-hispanic whites

Reduce clinical and genetic heterogeneity

(known MSS or MSI-L cases only- exclude
MMR mutation carriers and MSI-H cases)

Genetically enriched (FH, <50)

**Type additional markers around “significant
SNPs**

Study Population

Methods

Aim 1

Population-based case-control

952 CRC probands and 1020 controls
FH+ and <50 yr over represented

Genotyping:
500K Affymetrix SNP Chip

Statistical analysis:
SNP- and Haplotype-based
Hierarchical modeling incorporating
prior genomic data

Aim 2



Genotyping:

1000 significant SNPs
plus 4000 adjacent additional SNPs (include Functional data)

Statistical analysis:

As above

Additional SNPs would be chosen based on:

functional data
haplotype data
prior linkage data

Study Population

Methods

Aim 1

Population-based case-control

952 CRC probands and 1020 controls
FH+ and <50 yr over represented

Genotyping:
500K Affymetrix SNP Chip

Statistical analysis:
SNP- and Haplotype-based
Hierarchical modeling incorporating
prior genomic data

Aim 2

Genotyping:
1000 significant SNPs
plus 4000 adjacent additional SNPs (include Functional data)

Statistical analysis:
As above

Stage II

Colon CFR WGA Study Design

- **Stage II Study Population:**

Family-based case-control population

- drawn from all centers

Available DNA and Epidemiology Questionnaire

Non-hispanic whites

Reduce clinical and genetic heterogeneity

(known MSS or MSI-L cases only- exclude
MMR mutation carriers and MSI-H cases)

Genetically enriched (FH, <50)

Study Population

Methods

Aim 1

Population-based case-control

952 CRC probands and 1020 controls
FH+ and <50 yr over represented

Genotyping:
500K Affymetrix SNP Chip

Statistical analysis:
SNP- and Haplotype-based
Hierarchical modeling incorporating
prior genomic data

Aim 2

Genotyping:

1000 significant SNPs
plus 4000 adjacent additional SNPs (include Functional data)

Statistical analysis:

As above

Aim 3

Family-based case-control

612 CRC probands and 950 controls
Controls = same generation unaffecteds
FH+ and <50 yr over represented



Study Population

Methods

Aim 1

Population-based case-control

952 CRC probands and 1020 controls
FH+ and <50 yr over represented

Genotyping:

500K Affymetrix SNP Chip

Statistical analysis:

SNP- and Haplotype-based
Hierarchical modeling incorporating
prior genomic data

Aim 2

Genotyping:

1000 significant SNPs
plus 4000 adjacent additional SNPs (include Functional data)

Statistical analysis:

As above

Aim 3

Family-based case-control

612 CRC probands and 950 controls
Controls = same generation unaffecteds
FH+ and <50 yr over represented

This stage is limited to the most informative family-based case-control pairs and will help address the issue of possible population stratification biases in Stage I.

Study Population

Methods

Aim 1

Population-based case-control

952 CRC probands and 1020 controls
FH+ and <50 yr over represented

Genotyping:

500K Affymetrix SNP Chip

Statistical analysis:

SNP- and Haplotype-based
Hierarchical modeling incorporating
prior genomic data

Aim 2

Genotyping:

1000 significant SNPs
plus 4000 adjacent additional SNPs (include Functional data)

Statistical analysis:

As above

Aim 3

Family-based case-control

612 CRC probands and 950 controls
Controls = same generation unaffecteds
FH+ and <50 yr over represented

Aim 4

Family-based case-control

~3000 remaining CRC cases and same
generation controls

Genotyping:

All remaining significant SNPs

Statistical analysis:

As above incorporating Stage I and II data

Study Population

Methods

Aim 1

Population-based case-control

952 CRC probands and 1020 controls
FH+ and <50 yr over represented

Genotyping:

500K Affymetrix SNP Chip

Statistical analysis:

SNP- and Haplotype-based
Hierarchical modeling incorporating
prior genomic data

Aim 2

Genotyping:

1000 significant SNPs
plus 4000 adjacent additional SNPs (include Functional data)

Statistical analysis:

As above

Aim 3

Family-based case-control

612 CRC probands and 950 controls
Controls = same generation unaffecteds
FH+ and <50 yr over represented

Aim 4

Family-based case-control

~3000 remaining CRC cases and same
generation controls

Genotyping:

All remaining significant SNPs

Statistical analysis:

As above incorporating Stage I and II data

To fully exploit the Colon CFR, we will validate most significant SNPs in all remaining family members

Future Directions

- Validate significant SNPs in independent populations
- Confirm biological functionality of significant SNPs
- Extend genome SNP coverage

Genome-Wide Genotyping Technologies.

	Affymetrix 500K GeneChips*	Illumina 250K Infinium 2 (Expected Jan 2006)	Perlegen 275K LD Panel[⌘]	ParAllele 20K cSNP Panel^à
Selection Strategy	LD & HapMap prioritized; extra probes for select regions	Haplotype tagSNPs	LD tagSNPs	Coding SNPs
% of genome coverage	63% ($r^2 > 0.8$)** 81% ($r^2 > 0.5$)	65% ($r^2 > 0.8$) 80% ($r^2 > 0.5$)	70% ($r^2 > 0.8$) 86% ($r^2 > 0.5$)	nd
No. of SNPs	504,040	261,000 (over 100,000 singletons [^])	275,960	20,000
SNPs in genes	28,021 [!]	11,000	nd	20,000
Average MAF	0.22	0.26	nd	nd
Average heterozygosity	30%	nd	nd	nd
Mean spacing	5.8kb	10kb	nd	nd
Median spacing	2.8kb	nd	nd	nd
DNA required	500ng	750ng	10ug	4ug
Cost (per sample)	\$800	Not available	(IP issues)	\$440

*Greg Marcus, Ph.D. (Affymetrix), personal communication; Sa rah Murray, Ph.D. (Illumina), as presented at SNP2005 (UK, September 22-24, 2005); These are expected values. The product was still in development at the time of this presentation; [⌘]Hinds et al. 2005; ^àMegAllele Genotyping Human 10K cSNP Panel I (www.parallelebio.com/products-services/genotyping-products.html). Panel II will be available prior to the start of this proposal (Jay Kaufman, Affymetrix/ParAllele, personnel communication); **Using a multi-marker approach, ~80% genome coverage ($r^2 > 0.8$; $> 90\%$, $r^2 > 0.5$). Manuscript in preparation (Greg Marcus, Affymetrix, personal communication, in collaboration with the Broad Institute). [^]Singletons = SNPs not in LD with any other SNP(s) and thus are not truly tagSNPs; [!] includes exons, predicted exons, or ESTs; 225,090 SNPs in or within 10kb of a gene; 30,349 of the 33,653 genes in Ensembl (90%) have ≥ 1 SNP within 10kb, 18,852 ≥ 5 SNPs within 10kb (56%).

Genetics of Colorectal Cancer

- **High penetrance “rare” mutations: known familial CRC syndromes:**
 - Familial Adenomatous Polyposis (FAP)
 - Hereditary Non-Polyposis Colorectal Cancer (HNPCC)
 - Peutz Jeghers syndrome
 - Juvenile Polyposis
 - 9p22.2-p31.2 (Markowitz)
- **Low penetrance, “common” mutations**

Microsatellite Instability (MSI)

- Hallmark of HNPCC tumors
- Also seen in ^15% of non-HNPCC CRCs
Typically associated with older age of onset
Mucinous phenotype / proximal location
- Example:

normal allele 1

CACACACA (CA)₂₀

normal allele 2

CACACA (CA)₁₈

aberrant allele 3 (Tu)

CACACACACACA (CA)₂₄