Πανεπιστήμιο Θεσσαλίας

Γενετική Ποικιλότητα, HapMap και 1,000 Genomes Project

Γιάννης Βασιλόπουλος, PhD

Ποικιλομορφία στην μοναδικότητα και στην ευπάθεια

Φαινοτυπική Ποικιλομορφία

↕

Γενετική Ποικιλομορφία

↓

Γονίδια

↙ ↘

Coding DNA, 1.5% Non-coding DNA, 98.5%

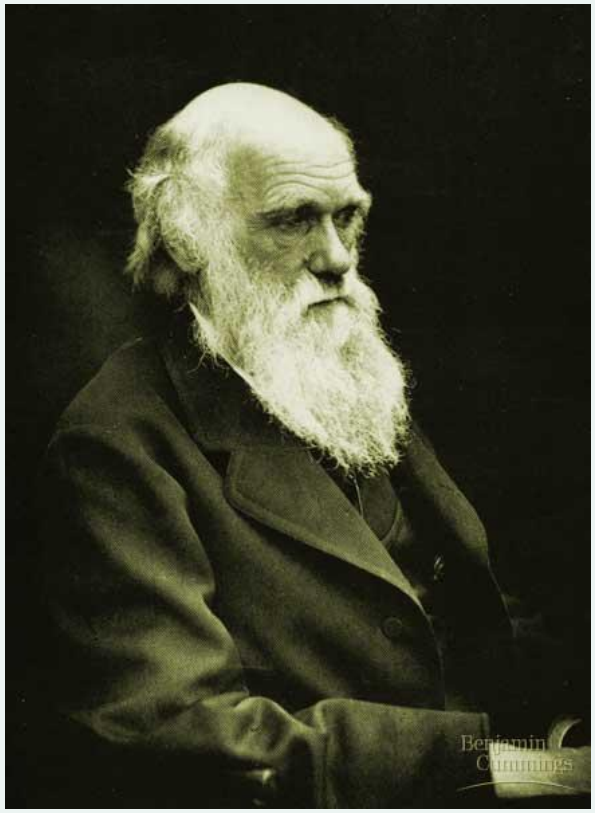
Πολυμορφισμός → συχνότητα σπάνιου αλληλόμορφου ≥ 0.01%, ετεροζυγώτες ≥ 2%

Πεπρωμένο, Ποικιλομορφία και Πυρηνικά Οξέα

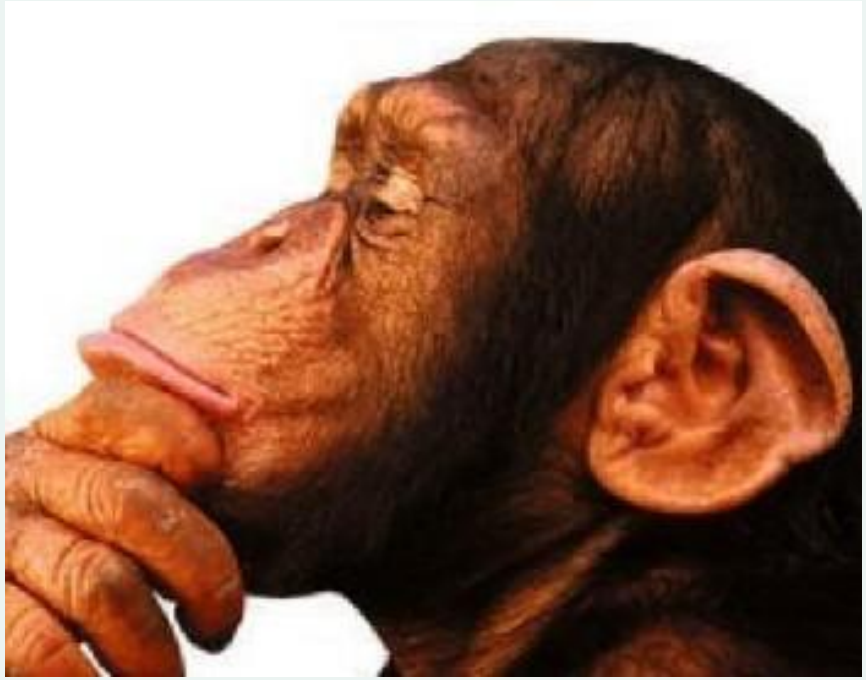
99.9%

***Πόσο διαφέρουμε μεταξύ μας?*!!!!!!**

Πεπρωμένο, Ποικιλομορφία και Πυρηνικά Οξέα



98.5%



***Πόσο διαφέρουμε με άλλους οργανισμούς?*!!!!!!**

Πεπρωμένο, Ποικιλομορφία και Πυρηνικά Οξέα

Πεπρωμένο……..……Type I Diabetes

Ποικιλομορφία……Type II Diabetes

Πυρηνικά Οξέα……..DNA, RNA

Πεπρωμένο → Ένα βιβλίο με κενά που καλούμαστε να συμπληρώσουμε εμείς!!!!!!

Jim Fixx

5’10”, 150 lbs
Marathon runner
Healthy lifestyle promoter
Died MI, age 52 (while running)
Father died MI age 43

Winston Churchill

5’8”, 270 lbs
Legendary gluttony
Smoker
Slothful
Died age 90

2001…Αποκωδικοποίηση του Γενετικού Κώδικα του ανθρώπου

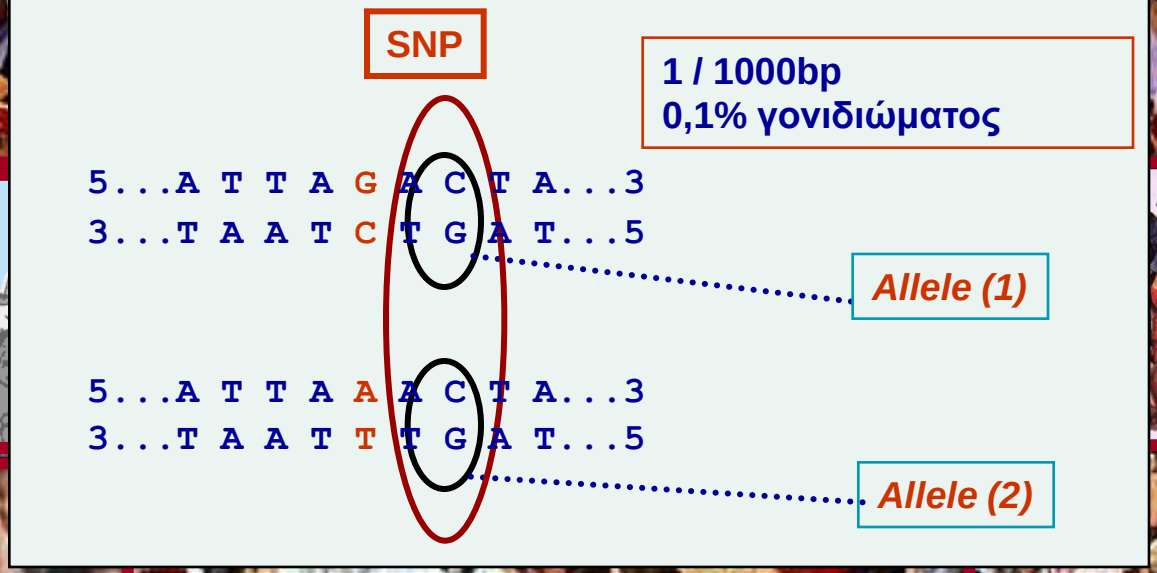
GAATTCTCTTTGGTATCCAATGAAGAAATCGAATCCATACCCATAGCTATAAAAAACAT
TTCAGGAGAAAATAAGACCGAAGCTGCTCAATTAGGCGCAATTGATTCGTTTCAAAAAT
GTGAAACTTGCCAGCTTACTTCGGCATGTCCTGGTCATTTTGGAAAATTTCATCTTACT
CAACCATTATTTAAAGTCGCATTTAAAAAACTTGTTGAAAATATTTTTAAATATACTTG
TTCTTTCTGTGGTGCTTTACAAAATCTTGAACTTCTGGAATTGATCAAGCAGATAGACG
AACGAAATACTGGAATAACAGTTAAAGATCGTGCTGCTTTTAAAAAAATTTTAGAAGCT
ACCAAACAAAGCAAATTCAAGTGTATTGCACCTAATTGCCAAAAACAAGTCTCTCCTTT
ACAATATTCGAAAAATAATAACTTTATATATAATTCGGGTACTACAAAGGGTATAGTTT
TGGATAACAGGCATGTGTTTAATATCTTACAAAATCTTCCACAAACGTTTAAATTATTG
TTAACCCCTTCGAATGCTCATCAAATCGTATCTCCCGAAAATGTCTTTTATGCTAATAG
TATCTTACTTCCACCACATAATCTACGAACTATCAATGTTTATGATGGTCAGGTTACGA
GTTTGTTAACAAGTGATTTGAATCTGATAATGCGAAGAGTTGCTAATAATGAGACAAAT
GCAAAAATACAAAAAATCTTGGATTCTATCGATAACAGCCGAGGTGCCAATCCATATGC
TACAAATAAAAAGCTTACTTTGGATACTTTGACAGGTGGACACTCAAAAGAATCTTATT
TGCGAAGTTATATTAATGGCAAACGTATTCCTGAGACTGCCAGAGCTGTAATCGAACCC
TCTATGAATAAAACTGGCTTTATTGAAGTACCATCTTACATTTTAAACAAGTTAAGAGA
TGTTGTCTTTTTATAATCACGTTACGAAAGATAACATACTCAAAAGTCTTCAAAACGAAC
AAGCTTTTCTAACATATATCAAAAGTGATCATAATTCTGAAAATCCTTATATGGTTTAT
GATTTAGCACAGAAGAATGGATATTTAACCTTGGCTCCTAATTTCGGTGATATTTTCGA
AAAAAGGAAAGAGGAAGGTGGTTTTGTAACTATTTGCAGACATCCATCTATCTGGTTAA
CTAATATCCAATCTGGTATAATAAAAAGATCAGAAGGGTTTACTATTAACATCCCAACC
ACAATTTGCACATCTTTTAATGCTGATTTTGATGGAGATGAGATGACAATATATTCTTT
CAAATCCCCATGTGCCAATCTCGAACAAGCTTTGATTATGAACTCACGAAATCTCTTCA
AAAATTCTATAACAAGCAATCCAATGTTCGGCTTGGTCCAAGATCAAATACCAGCCTTG
AATAAGTTATATAGACGACAAAATTATACATATAACGATGCGTTGGTGATTTTAGGACA
ATTCGGATTTCTGTTAACACCTGGAAAAGATAATTATACCGGAAAAGATATACTTTCTT
GTGTATTCCCAAAACATTATACACTCAAAGGAATTGTTGAAAATGGCGAACTTATTTTG
GAGAATTTTACAAATAAACTCGTTTCCGCAAATTCCTCAAAGTCCATCTTTGGGCATCT
TGTTTTATTTTATGGACAAGAGTATGGTTTGACTATATTGGATACAATGCGAGATATTG
TTCAAAATTTTATTACACATTTTGGTTTCAGTGTAAAAATCCGAGATATGATCCCAAGC
CCAAAAATTTTGGATATTCTAGAAAAGATCGTAGACCAAGAAGTGGATAAAATTGATAA
ACAAACAAAACTTCTATATGACGATATCGAACAAGGTAAGGTTATAATCAACTCTTATG
ATGATATTTCTGAGTTCAGATTAAAAAATGTGGCTATTATGAAAAAGAAACTAGAAAGC
AAACTTTTGGAACTTTTGGATGAATATTATGATGAAGACAATAATTTCCTAGAGATGTA
TAGAACGGGATATAAGGTCAACATTAACGAACTTCTCTCTATTATGTGTTTCTCGGGTT
TTAAAAATTATGGAAATATCGAAATGATTACACCGGGTCTTAATGGTAAAACATCTTTG
TTTAGCTTACCAGATTCTATAAACTTACAAGATTATGGGTTCATCAAAAGCTCTATTGC
CAAAGGGTTAACGTTTGAAGAATATGCTACAATCGTAAAACAAGAAGCTTTTCCACAAA
TTGTTAATGTTACAACTGGTACTTCACAAACAGGATTTTTGGGGAAAAAAATGGTTAAA
ATGGCTTCTGAATTC

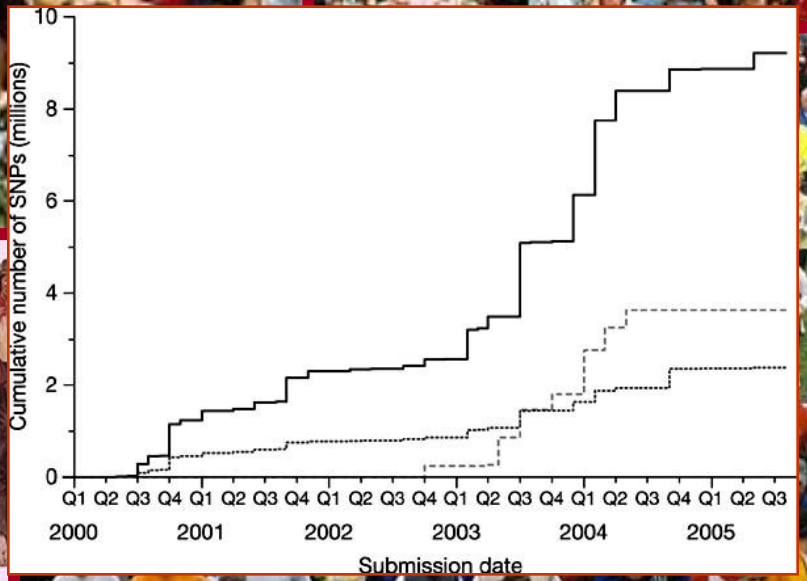
3 Δισεκατομμύρια βάσεις!!!

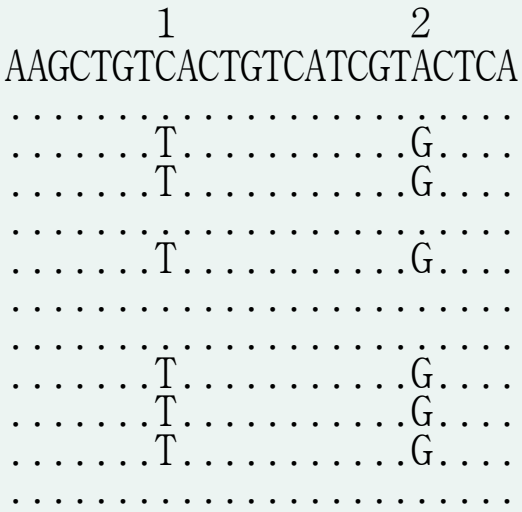
400 Τόμοι Βιβλίων!!!

ΓΕΝΕΤΙΚΗ ΠΟΙΚΙΛΟΤΗΤΑ

SNPs (single nucleotide polymorphisms)





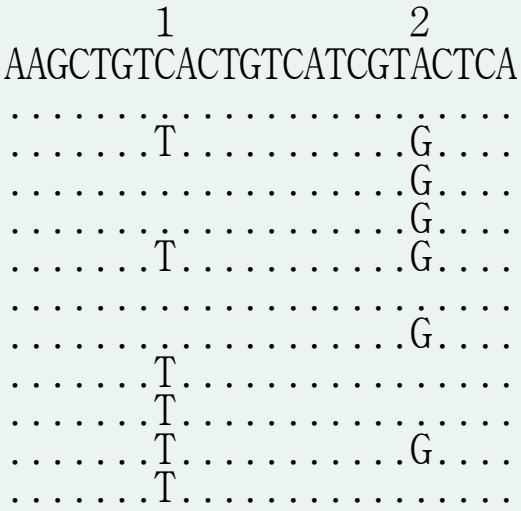


Site 2

Site 1	A	G
C	6	0
T	0	6

Complete LD (D’ = 1)

Fisher’s Exact Test P = 0.002



Site 2

Site 1	A	G
C	3	3
T	3	3

No LD (D’ = 0)

Fisher’s Exact Test P = 1.00

Παράδειγμα ανισορροπίας σύνδεσης και εμφάνισης απλοτύπων

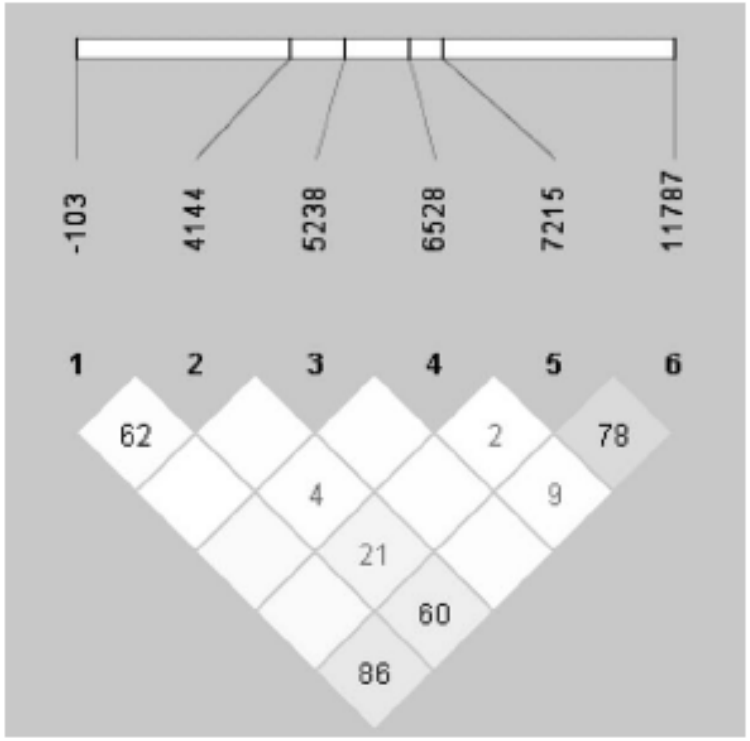


Figure 1 Haploview output showing pairwise coefficients of disequilibrium for the *AIRE* single nucleotide polymorphisms at positions −103, 4144, 5238, 6528, 7215 and 11787, respectively. D′ values are not shown where there is complete LD.

Table 4 *P* values for case–control analysis on *AIRE* C-103T, G6528A, T7215C, T11787C haplotypes[a]

All patients	Controls	Frequency	Cases	Frequency	χ^2	*P* value	OR (95% CI)
CGTT	136	0.404	193	0.352	2.377	0.123	0.80 (0.61–1.06)
CGTC	96	0.283	98	0.179	13.73	2.10×10^{-4}	0.55 (0.40–0.75)
CGCT	3	0.008	51	0.092	25.77	3.84×10^{-7}	11.4 (3.53–36.9)
CGCC	32	0.096	100	0.183	12.56	3.94×10^{-4}	2.13 (1.39–3.24)
CATT	20	0.060	25	0.045	0.819	0.365	0.76 (0.41–1.39)
TGTT	39	0.115	62	0.112	0.015	0.903	0.97 (0.64–1.50)

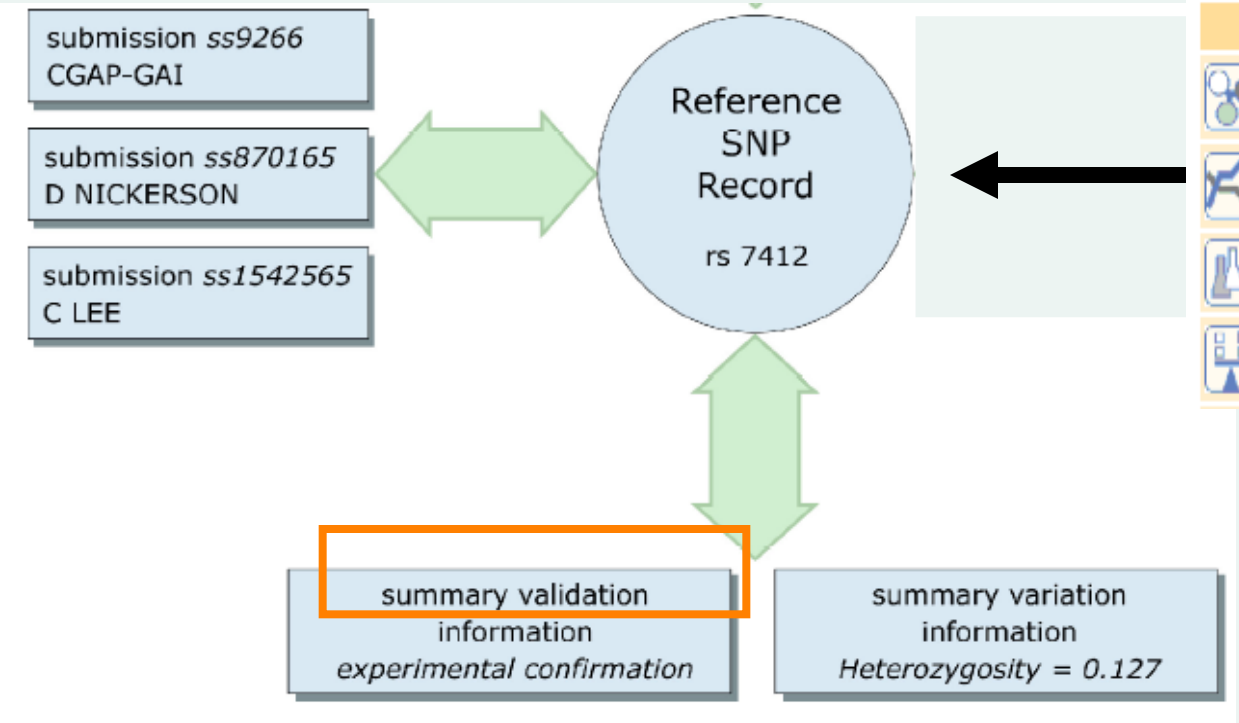
[a] The susceptibility haplotypes showing significant association are in bold and the haplotype with negative association is underlined.

Πχ για 4 SNPs εδώ εμφανίζονται μόνο 6 από τους 16 απλότυπους

Υποβολή SNPs στο dbSNP

SNPs submitted
By research communtiy
(**s**ubmitted SNPs = ss#)

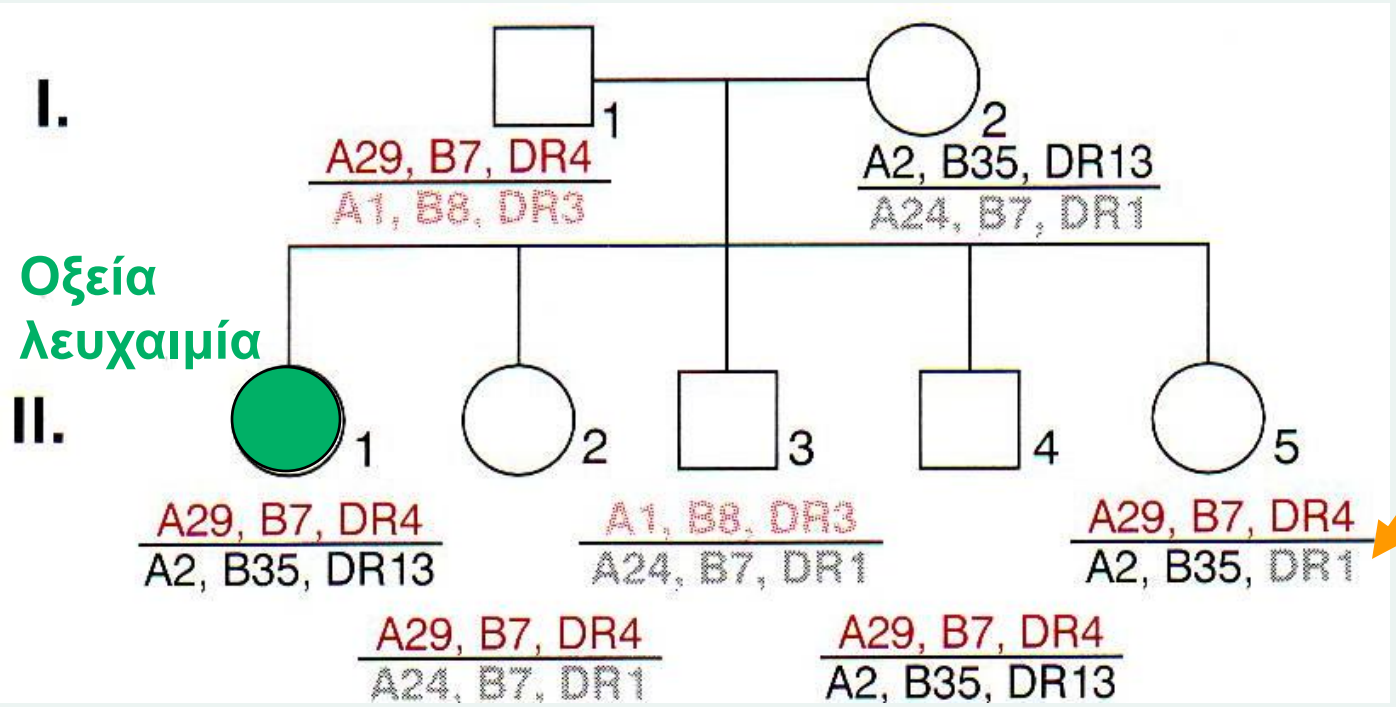
Unique mapping
to a genome location
(**r**eference **S**NP = rs#)



Validation status description	
	validated by multiple, independent submissions to the refSNP cluster
	validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
	validated by submitter confirmation
	all alleles have been observed in at least two chromosomes apiece **(by 2hit-2allele)**

Απλότυπος

- Ομάδα πολυμορφικών συνδεδεμένων αλληλόμορφων που κληρονομούνται μαζί

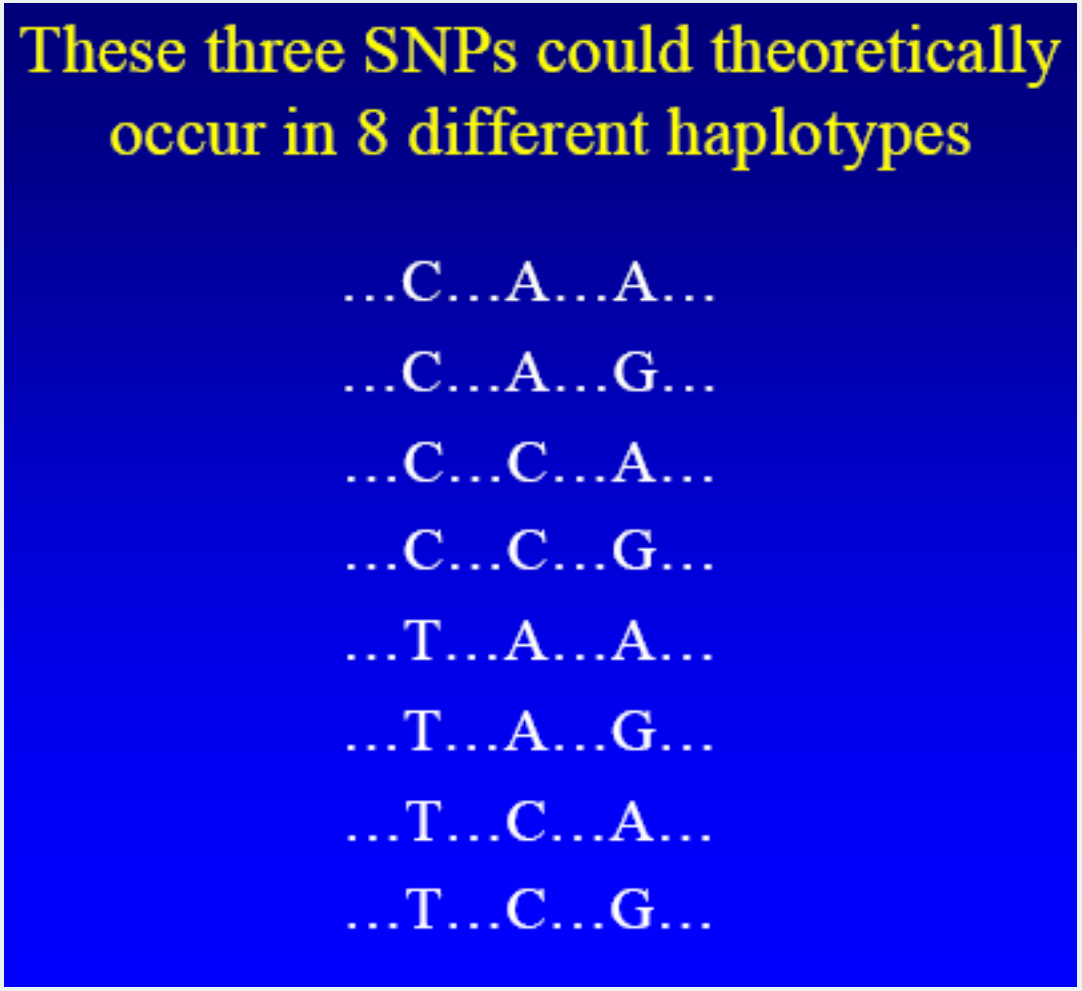


Ανταλλαγή υλικού μεταξύ δυο μητρικών χρωμοσωμάτων

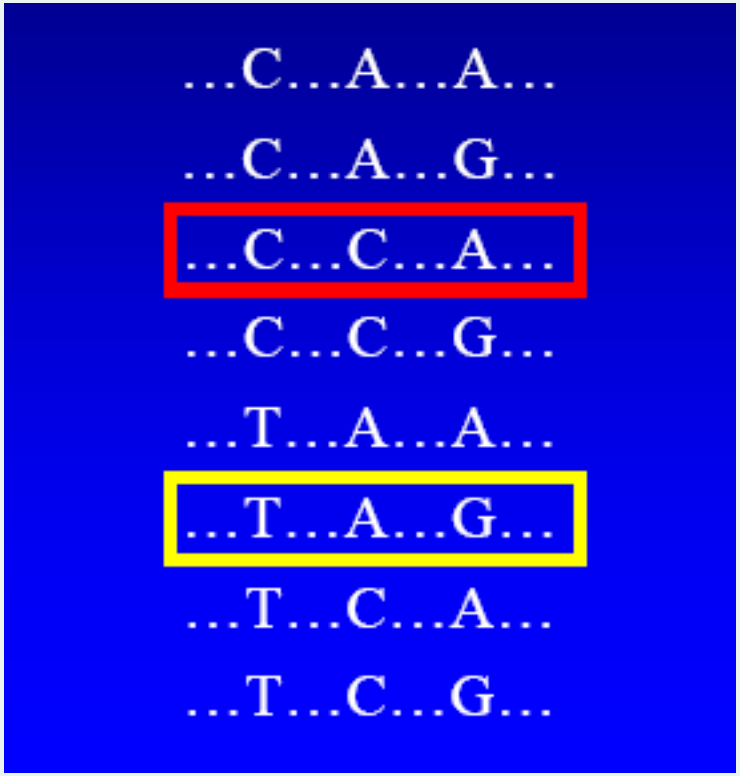
Απλοταυτόσημη

Ας υποθέσουμε ότι έχουμε 3 SNPs……

Θεωρητικά προκύπτουν 8 απλότυποι.....

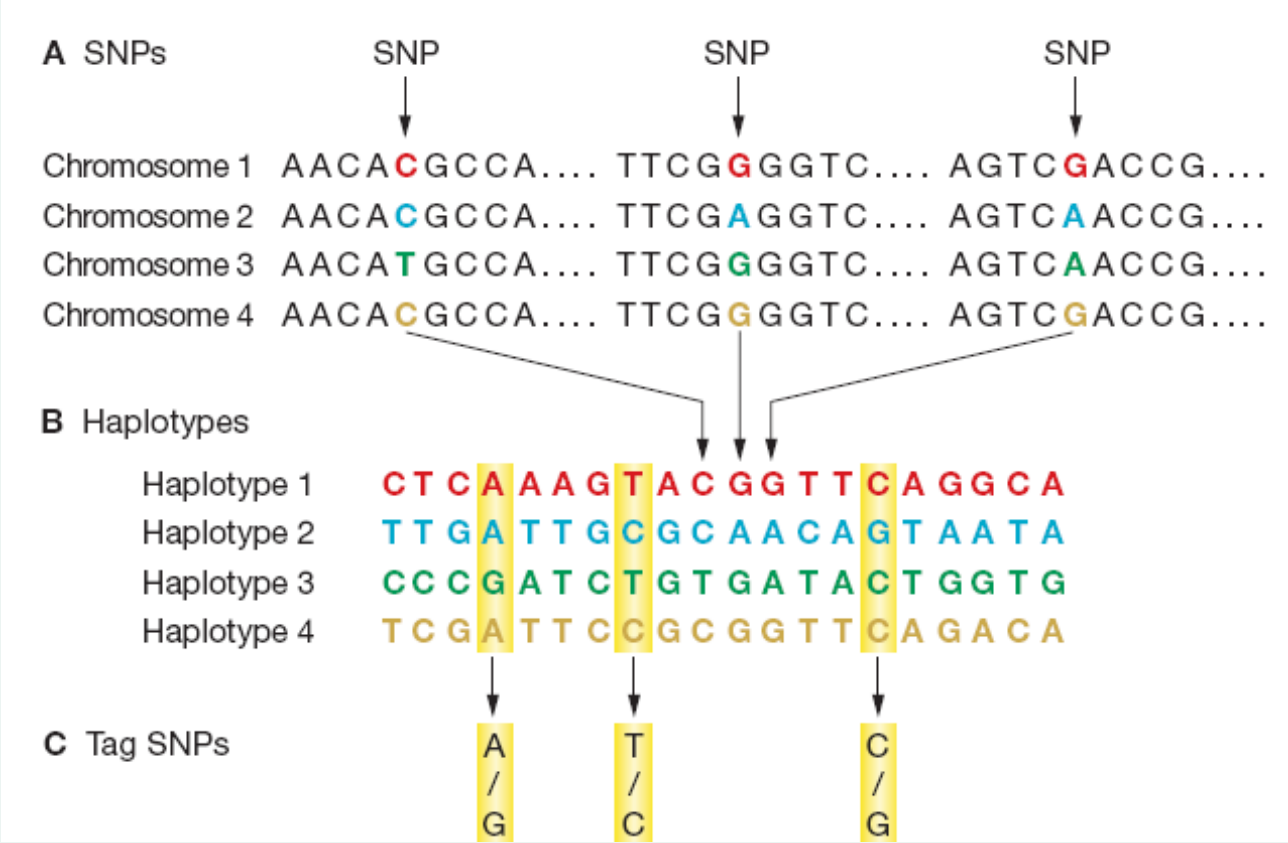


Όμως πρακτικά εμφανίζονται δύο.....



Tag-SNPs

Τα Tag SNP's μπορούν να προσδιορίσουν κοινούς απλότυπους



Η γονοτύπηση των 3 αυτών μόνο Tag-SNPs (από σύνολο 20) μπορούν να δώσουν τους παραπάνω 4 απλότυπους

Απλότυπος: ένας συγκεκριμένος συνδυασμός αλληλομόρφων σε ένα χρωμόσωμα

Συγκρίνοντας απλοτύπους από πολλά άτομα, παρατηρούμε κοινά πρότυπα

- **Ομάδες γειτονικών SNPs στο ίδιο χρωμόσωμα κληρονομούνται μαζί (blocks) - ανισορροπία σύνδεσης**
- **Το πρότυπο των SNPs σε ένα block είναι ο απλότυπος**
- **Είναι δυνατόν να επιλεγούν και να ελεγχθούν συγκεκριμένα SNPs ώστε να γίνουν αναλύσεις συσχέτισης με συγκεκριμένο φαινότυπο.**
- **Για άτομα που έχουν ένα συγκεκριμένο SNP σε μία θέση, μπορούμε να προβλέψουμε τα SNPs σε γειτονικές θέσεις**
- **Ο HapMap είναι ο χάρτης των blocks των SNPs που συν-κληρονομούνται καθώς και των επιλεγμένων SNPs (tag SNPs) που ταυτοποιούν αυτά τα blocks**

articles

A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms

The International SNP Map Working Group*

* *A full list of authors appears at the end of this paper.*

We describe a map of 1.42 million single nucleotide polymorphisms (SNPs) distributed throughout the human genome, providing an average density on available sequence of one SNP every 1.9 kilobases. These SNPs were primarily discovered by two projects: The SNP Consortium and the analysis of clone overlaps by the International Human Genome Sequencing Consortium. The map integrates all publicly available SNPs with described genes and other genomic features. We estimate that 60,000 SNPs fall within exon (coding and untranslated regions), and 85% of exons are within 5 kb of the nearest SNP. Nucleotide diversity varies greatly across the genome, in a manner broadly consistent with a standard population genetic model of human history. This high-density SNP map provides a public resource for defining haplotype variation across the genome, and should help to identify biomedically important genes for diagnosis and therapy.

International HapMap Project

(Χάρτης απλοτύπων)

www.hapmap.org

HapMap Project

Nature 437: 1299-320, 2005

Participating Groups

Baylor College of Medicine (USA)
Beijing Genomics Institute (China)
Beijing Normal University (China)
Broad Institute of Harvard and MIT (USA)
Center for Statistical Genetics, University of Michigan (USA)
Chinese National Human Genome Center at Beijing (China)
Chinese National Human Genome Center at Shanghai (China)
Cold Spring Harbor Laboratory (USA)
Eubios Ethics Institute (Japan)
Health Sciences University of Hokkaido (Japan)
Hong Kong University of Science and Technology (China)
Howard University (USA)
Illumina (USA)
Johns Hopkins School of Medicine (USA)
McGill University & Génome Québec Innovation Centre (Canada)
ParAllele BioScience (USA)
Perlegen Science (USA)
RIKEN (Japan)
The Chinese University of Hong Kong (China)
The University of Hong Kong (China)
University of California, San Francisco (USA)
University of Ibadan (Nigeria)
University of Oxford (UK)
University of Oxford / Wellcome Trust Centre for Human Genetics (UK)
University of Tokyo (Japan)
University of Utah (USA)
Washington University, St. Louis (USA)
Wellcome Trust Sanger Institute (UK)

<u>Σκοπός</u>: Η δημιουργεία ενός χάρτη κοινών γενετικών μεταλλάξεων και απλοτύπων – 600,000 κοινά SNP's από 4 διαφορετικούς πληθυσμούς.

- CEPH (CEU) (Europe - n = 90, trios)
- Yoruban (YRI) (Africa - n = 90, trios)
- Japanese (JPT) (Asian - n = 45)
- Chinese (HCB) (Asian - n =45)

1 SNP/ 5 Kb
MAF>5%
Φάση I: 1M SNPs
Φάση II: 4.6M SNPs
Φάση III: .....!...

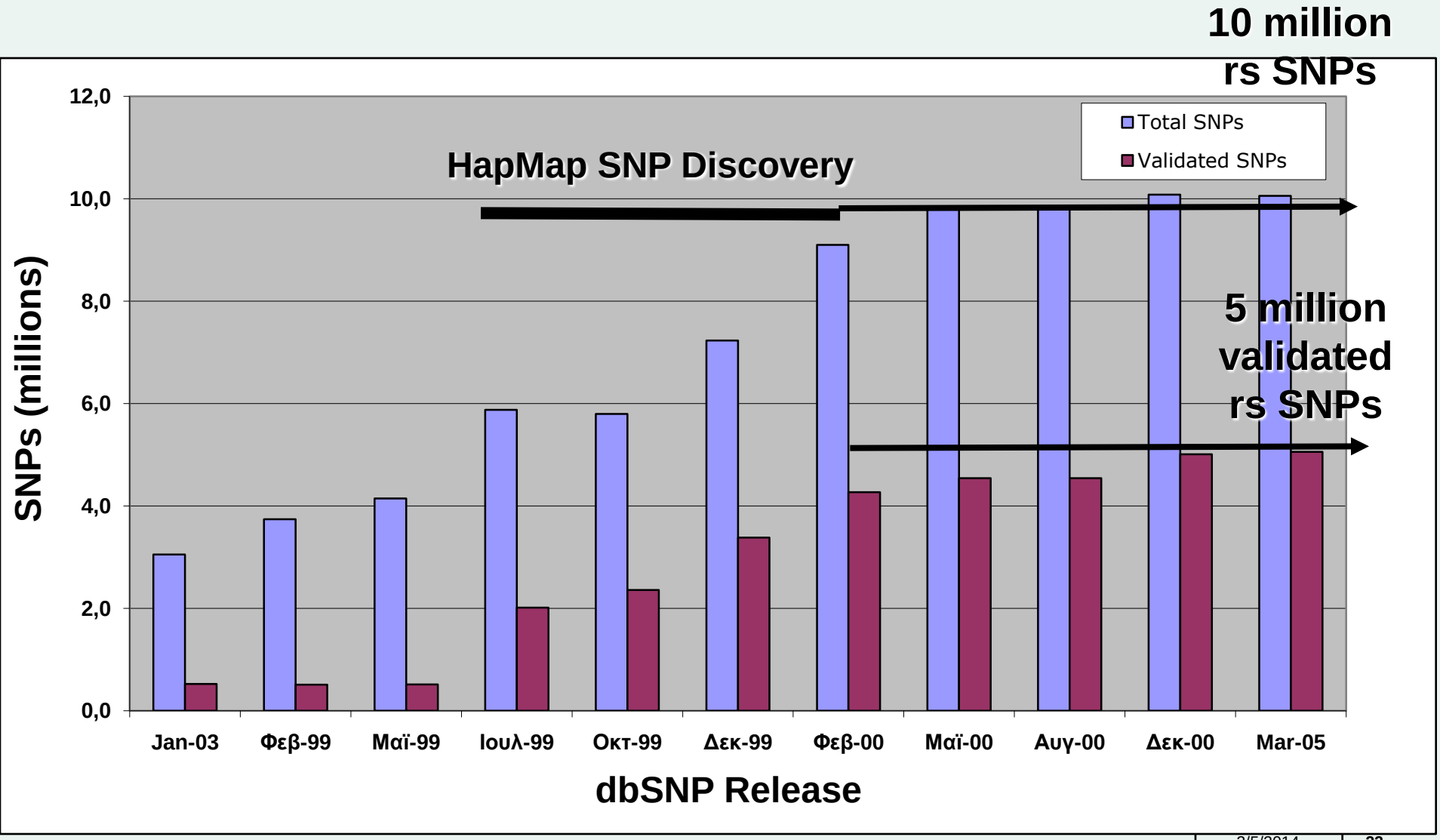
HapMap Project – Φάση III

Ενίσχυση των αρχικών HapMap δειγμάτων με ακόμη 1,115 δείγματα τα οποία προέρχονται από 11 διαφορετικούς πληθυσμούς ανα τον κόσμο για περισσότερη ποικιλότητα: Αφρική, Ασία, Ευρώπη και Μεξικό, άτομα που μένουν σε διαφορετικές περιοχές – Ιδανικό εργαλείο για μελέτη γονότυπου-φαινότυπου, σε ασθένειες αλλά και στην εξέλιξη.

label	population sample	# samples	QC+ Draft 1
ASW*	African ancestry in Southwest USA	90	71
CEU*	Utah residents with Northern and Western European ancestry from the CEPH collection	180	162
CHB	Han Chinese in Beijing, China	90	82
CHD	Chinese in Metropolitan Denver, Colorado	100	70
GIH	Gujarati Indians in Houston, Texas	100	83
JPT	Japanese in Tokyo, Japan	91	82
LWK	Luhya in Webuye, Kenya	100	83
MEX*	Mexican ancestry in Los Angeles, California	90	71
MKK*	Maasai in Kinyawa, Kenya	180	171
TSI	Toscans in Italy	100	77
YRI*	Yoruba in Ibadan, Nigeria	180	163
		1,301	**1,115**

HapMap Project

1 x 10^9 γονότυποι → $800,000 για κάθε ασθένεια.



Παράδειγμα μελέτης χρησιμοποιώντας το HapMap

nature genetics

HapMap leads to a new diabetes gene discovery

Variant of transcription factor 7-like 2 (*TCF7L2*) confers risk of type 2 diabetes

Struan F A Grant[1], Gudmar Thorleifsson[1], Inga Reynisdottir[1], Rafn Benediktsson[2,3], Andrei Manolescu[1], Jesus Sainz[1], Agnar Helgason[1], Hreinn Stefansson[1], Valur Emilsson[1], Anna Helgadottir[1], Unnur Styrkarsdottir[1], Kristinn P Magnusson[1], G Bragi Walters[1], Ebba Palsdottir[1], Thorbjorg Jonsdottir[1], Thorunn Gudmundsdottir[1], Arnaldur Gylfason[1], Jona Saemundsdottir[1], Robert L Wilensky[4], Muredach P Reilly[4], Daniel J Rader[4], Yu Bagger[5], Claus Christiansen[5], Vilmundur Gudnason[2], Gunnar Sigurdsson[2,3], Unnur Thorsteinsdottir[1], Jeffrey R Gulcher[1], Augustine Kong[1] & Kari Stefansson[1]

Published on line January 15, 2006
Already confirmed by multiple groups

Διαβήτης τύπου Ι

- Αυτοάνοση νόσος (Αυτοάνοση νόσος or viral infection)
- Καταστροφή pancreatic β cells → failure to produce insulin
- 15 loci – PTPN22, CTLA4, IL2RA, IL2, IL7R.
- 25,000 genes → 15-20 genes in T1D
- Βλαστοκύτταρα ή ουσίες που διεγείρουν την αναγέννηση των λίγων κυττάρων β που έχουν απομείνει

Nature. 2007 Aug 2;448(7153):591-4. Epub 2007 Jul 15.
A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene.
Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC, Lawson ML, Robinson LJ, Skraban R, Lu Y, Chiavacci RM, Stanley CA, Kirsch SE, Rappaport EF, Orange JS, Monos DS, Devoto M, Qu HQ, Polychronakos C.
Center for Applied Genomics, Abramson Research Center, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. hakonarson@chop.edu

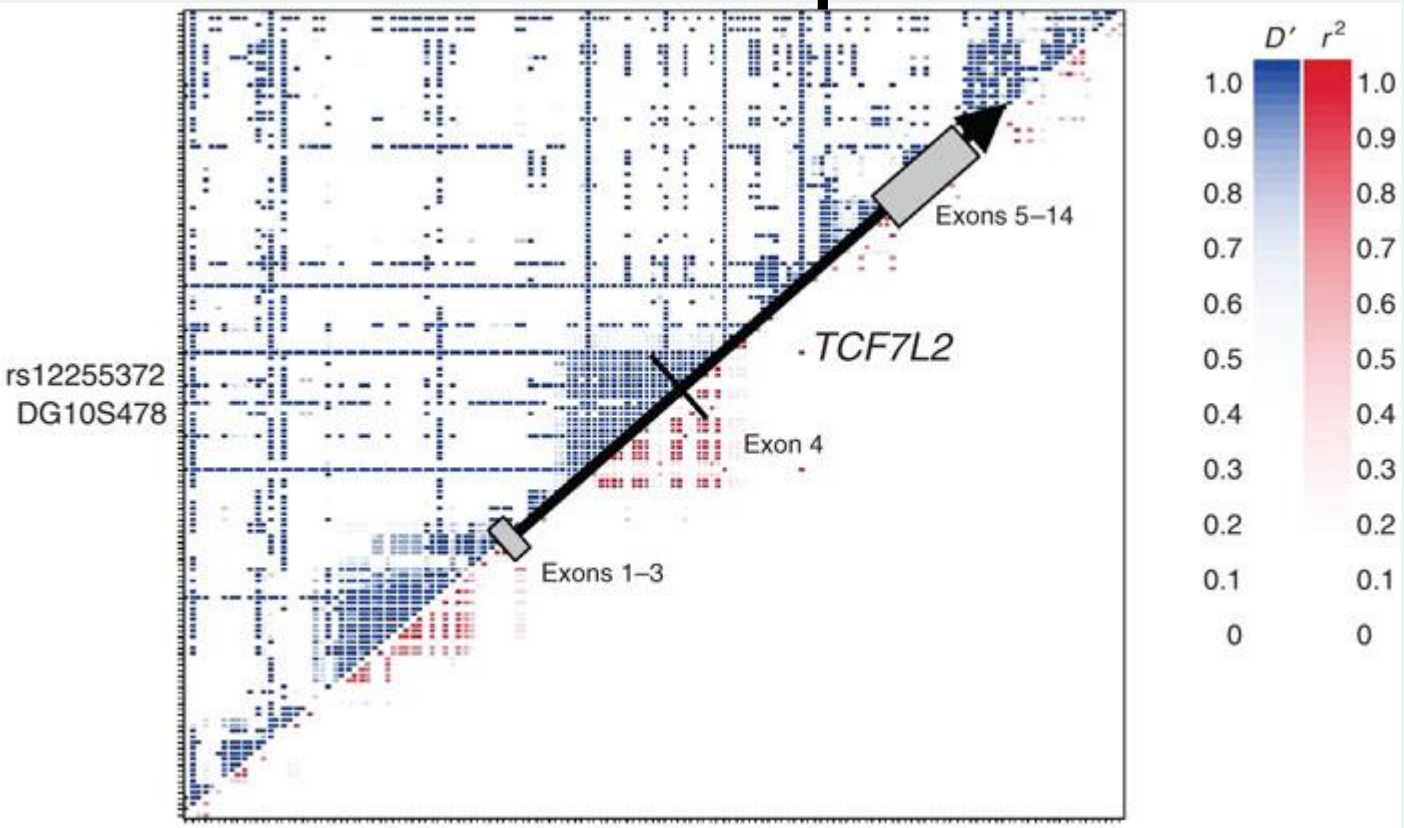
Διαβήτης τύπου ΙΙ

- Χαρακτηρίζεται από υπεργλυκαιμία, πιθανόν λόγω βλάβης στους μηχανισμούς έκκρισης της ινσουλίνης ή/και αυξημένη παραγωγή γλυκόζης από το συκώτι.
- Επιπολασμός ~60% αλλά τείνει να μεγαλώνει λόγω αύξησης της μέσης ηλικίας και παχυσαρκίας.
- Σημαντικός γενετικός παράγοντας: $\lambda_s = 3.5$.

- Προηγούμενη μελέτη σύνδεσης → χρ. 5q, 10q, 12q στον Ισλανδικό πληθυσμό, και επαλήθευση του 10q σε Μεξικανο-Αφρικανούς.

- Μελέτη συσχέτισης στο χρ. 10q καλύπτοντας μια περιοχή 10.5Mb (228 μικροδορυφορικοί δείκτες σε 1,185 ασθενείς τύπου ΙΙ και 931 μη-σχετιζόμενους controls.

Αποτελέσματα



- Συσχέτιση του δείκτη DG10S478 και της ασθένειας αλλά και σε υψηλή LD με τον rs12255372 από τα δείγματα του HapMap CEPH Utah

- Ο DG10S478 βρίσκεται στο ιντρόνιο 3 του παράγοντα μεταγραφής TCF7L2 μέσα σε ένα τμήμα LD 92.1 Kb που περιλαμβάνει μέρος του ιντρονίου 3, όλο το εξώνιο 4 και τμήμα του ιντρονίου 4 του TCF7L2

- Ο TCF7L2 μετέχει στο μονοπάτι σηματοδότησης του Wnt και ρυθμίζει τα επίπεδα της ορμόνης GLP-1 η οποία επιδρά στην ομοιόσταση της γλυκόζης του αίματος

Φλεγμονώδης νόσος του εντέρου

- Νόσος του Crohn – ασυνεχής, διατοιχωματική φλεγμονή που προσβάλλει τον γαστρεντερικό σωλήνα και συχνά τον ειλεό (συχνότητα 1:5000)
- Ελκώδη κολίτιδα – προσβάλλει κυρίως το ορθό (συχνότητα 1:1500)
- Αύξηση του ρίσκου x13 και x15 για ΝΚ και ΕΚ, αντίστοιχα για συγγενείς 1^{ου} βαθμού
- Ανοσολογική προέλευση – διήθηση του βλεννογόνου απο λεμφοκύτταρα και παρουσία Abs ενάντια στα επιθηλιακά κύτταρα του παχέος εντέρου → βλεννογόνια βλάβη
- Περιβάλλον – φαίνεται από την τεχνολογική ανάπτυξη στην Ασία και την παράλληλη αύξηση των περιστατικών

Τα γονίδια που έχουν συσχετιστεί εώς σήμερα αντιστοιχούν στο ~20% του γενετικού παράγοντα

Table 2 | Gene associations in Crohn's disease and ulcerative colitis

Chromosome	Location (Mb)	Genes of interest	Associated with Crohn's disease	Associated with ulcerative colitis
1p31	67	*IL23R*	Yes	Yes
2q37	231	*ATG16L1*	Yes	No
3p21	49	Multiple, including *MST1*	Yes	Yes
5p13	40	Intergenic, *PTGER4*	Yes	No
5q31	131	Multiple, including *SLC22A5*	Yes	Unclear
5q33	150	Multiple, including *IRGM*	Yes	No
5q33	158	*IL12B* (p40)	Yes	Yes
10q21	64	ZNF365	Yes	Unclear
10q24	101	*NKX2-3*	Yes	Yes
16q12	49	*NOD2*	Yes	No
17q21	37	Multiple, including *STAT3*	Yes	Yes
18p11	12	*PTPN2*	Yes	Unclear

ATG16L1, autophagy related 16-like protein 1; *IL12B*, interleukin-12β; *IL23R*, interleukin-23 receptor; *IRGM*, immunity-related GTPase family, M; *NKX2-3*, NK2 transcription factor related, locus 3; *NOD2*, nucleotide-binding oligomerization domain protein 2; *PTGER4*, prostaglandin receptor, EP4; *PTPN2*, protein tyrosine phosphatase, non-receptor type 2; *SLC22A5*, solute carrier family 22, member 5; *STAT3*, signal transducer and activator of transcription 3; *ZNF365*, zinc-finger protein 365.

Αυτοφάγωση ενδοκυτταρικών βακτηρίων και αντιγονοπαρουσίαση

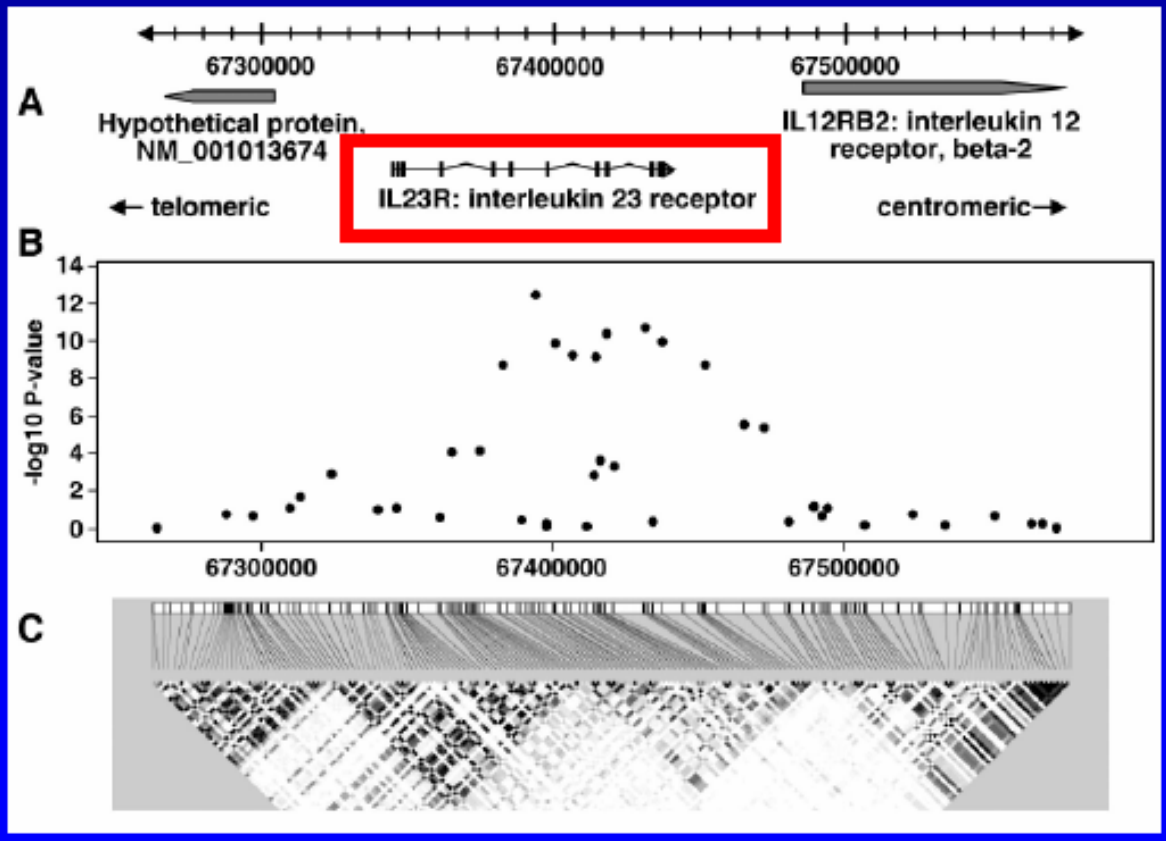
Εντοπισμό των λεμφοκυττάρων στο σπλήνα για ωρίμανση των Β κυττάρων και Τ ειδική απάντηση

Υποδοχέας αναγνώρισης της πεπτιδογλυκάνης των βακτηρίων → NfKB και MAP κινάσες → έμφυτη ανοσία

Sciencexpress Report

A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr,[1,2] Kent D. Taylor,[3,4] Steven R. Brant,[5,6] John D. Rioux,[7,8] Mark S. Silverberg,[9] Mark J. Daly,[8,10] A. Hillary Steinhart,[9] Clara Abraham,[11] Miguel Regueiro,[1] Anne Griffiths,[12] Themos Dassopoulos,[5] Alain Bitton,[13] Huiying Yang,[3,4] Stephan Targan,[4,14] Lisa W. Datta,[5] Emily O. Kistner,[15] L. Philip Schumm,[15] Annette Lee,[16] Peter K. Gregersen,[16] M. Michael Barmada,[2] Jerome I. Rotter,[3,4] Dan L. Nicolae,[11,17] Judy H. Cho[18]*



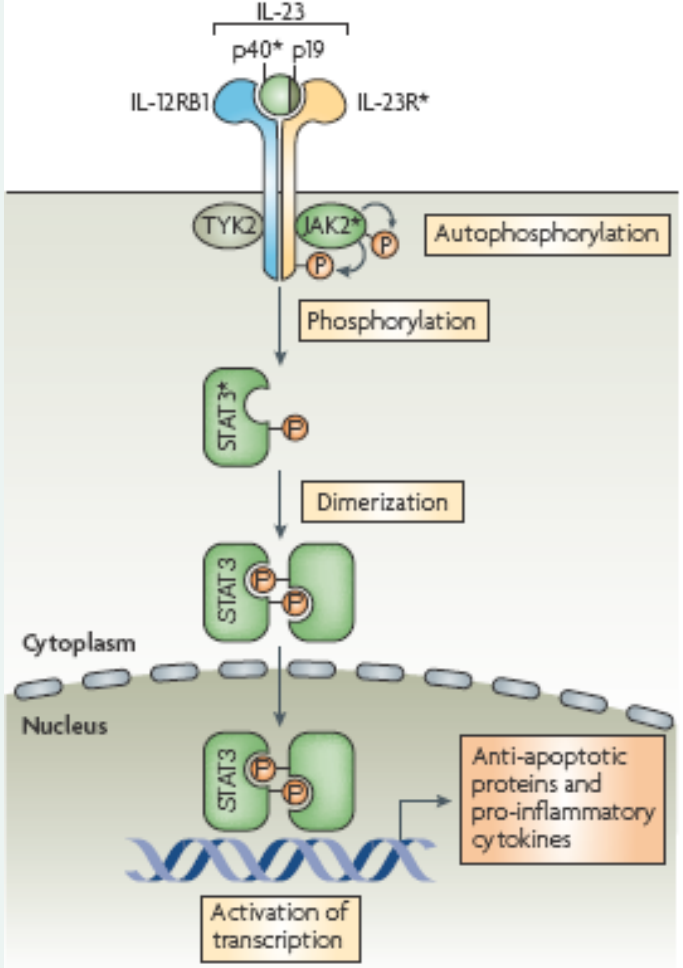
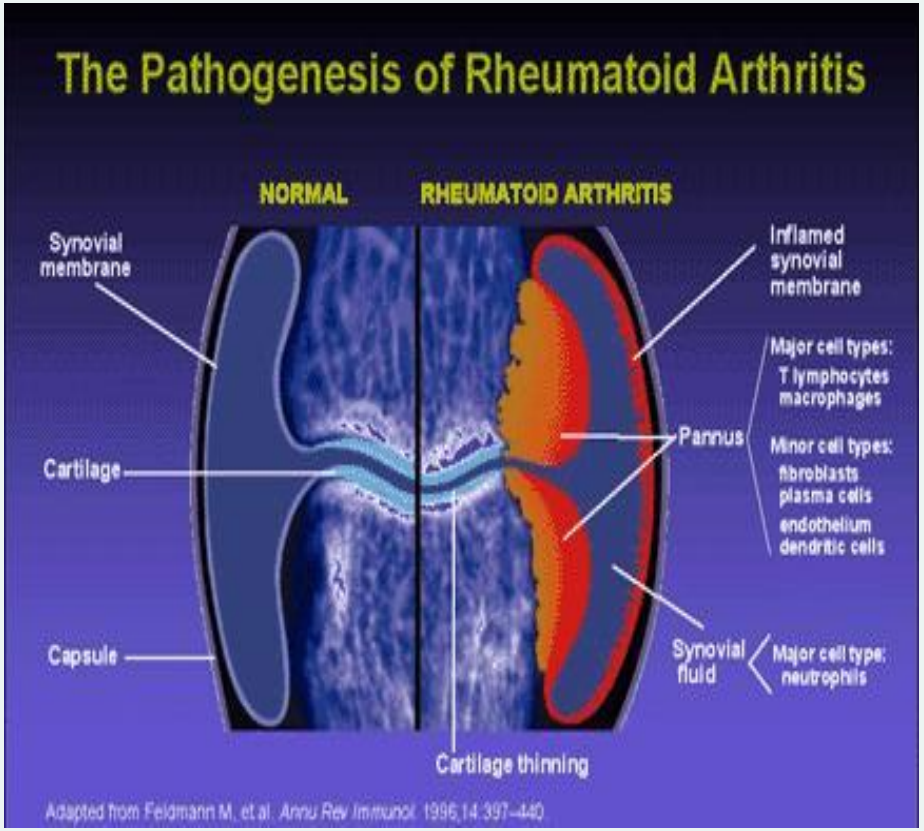


Figure 2 | **Variation in multiple genes in the IL-23R pathway is associated with Crohn's disease.** Functional IL-23R (interleukin-23 receptor) signalling results from the engagement of a heterodimeric cytokine (comprised of p40 and p19 subunits) with a heterodimeric receptor (comprised of IL-23R and IL-12RB1 subunits). On engagement of IL-23 with its receptor, Janus kinase 2 (JAK2) is activated, resulting in JAK2 autophosphorylation and tyrosine phosphorylation of IL-23R. This in turn results in the recruitment, phosphorylation, homodimerization and nuclear translocation of signal transducer and activator of transcription 3 (STAT3). Asterisks denote genes proven to be associated with Crohn's disease. TYK2, tyrosine kinase 2.

- Η IL-23R σε mRNA επίπεδο εκφράζεται από NK κύτταρα, CD4 και CD8 Τ κύτταρα
- Καθοριστικό ρόλο στην διαφοροποίηση των Th κυττάρων, π.χ. IFN-γ που ρυθμίζει την TH1 αντίδραση μειώνει τα επίπεδα mRNA της IL23R
- Μεταλλάξεις στο μονοπάτι της IL23R σε πολλά γονίδια που επηρεάζουν την ρύθμισή της π.χ. STAT3 ένας ισχυρός ενεργοποιητής της μεταγραφής και ρυθμιστής της TH17 και TH1 διαφοροποίησης

Ρευματοειδής αρθρίτιδα

- Η πιο κοινή φλεγμονώδης αρθροπάθεια – επιπολασμός 0.5-1%
- Συχνότητα εμφάνισης νέων περιστατικών ~0.03% παγκοσμίως
- Παρουσιάζεται συχνότερα στις ηλικίες 40-50, προσβάλλοντας x3 περισσότερες γυναίκες από άντρες
- Αυξάνει 4-15 φορές την αναπηρία καθώς και μειώνει την διάρκεια ζωής κατά 3-10 χρόνια (σε σχέση με τον γενικό πληθυσμό)
- Αιτιολογία άγνωστη – πολυπαραγοντική ασθένεια



Feldman et al, Ann Rev Immun 1996; 14:337-440.

Alamanos and Drosos. Autoimmunity Rev 2005; 4:130.

Ρευματοειδής αρθρίτιδα

TABLE 1. **Loci With Significant Linkage With Rheumatoid Arthritis, Replicated in Independent Samples* in Genome-Wide Screening of Multicase Families†**

Reference	Locus	Suggested candidate genes
Cornelis et al,[21] 1998 Jawaheer et al,[24] 2001 MacKay et al,[23] 2002 Jawaheer et al,[25] 2003 Osorio et al,[22] 2004	6p21.3	*HLA-DRB1*, other MHC genes
Cornelis et al,[21] 1998 Jawaheer et al,[24] 2001 Jawaheer et al,[25] 2003	18q21	RANK
Cornelis et al,[21] 1998 Jawaheer et al,[24] 2001 MacKay et al,[23] 2002 Jawaheer et al,[25] 2003	1q43	Unknown, also linkage with SLE
Jawaheer et al,[24] 2001 MacKay et al,[23] 2002 Jawaheer et al,[25] 2003	6q21	Unknown
Jawaheer et al,[24] 2001 Jawaheer et al,[25] 2003	1p13	Unknown

*Replicated in at least 2 independent samples; $P<.005$ in at least 1 sample.
†MHC = major histocompatibility complex; RANK = receptor activator of nuclear factor κB; SLE = systemic lupus erythematosus.

- MZ: 20-30% , ΔZ: 5-15%,
- Κληρονομησιμότητα: 60%
- **HLA-DRB1 - 1/3 της γενετικής βάσης της αρθρίτιδας**

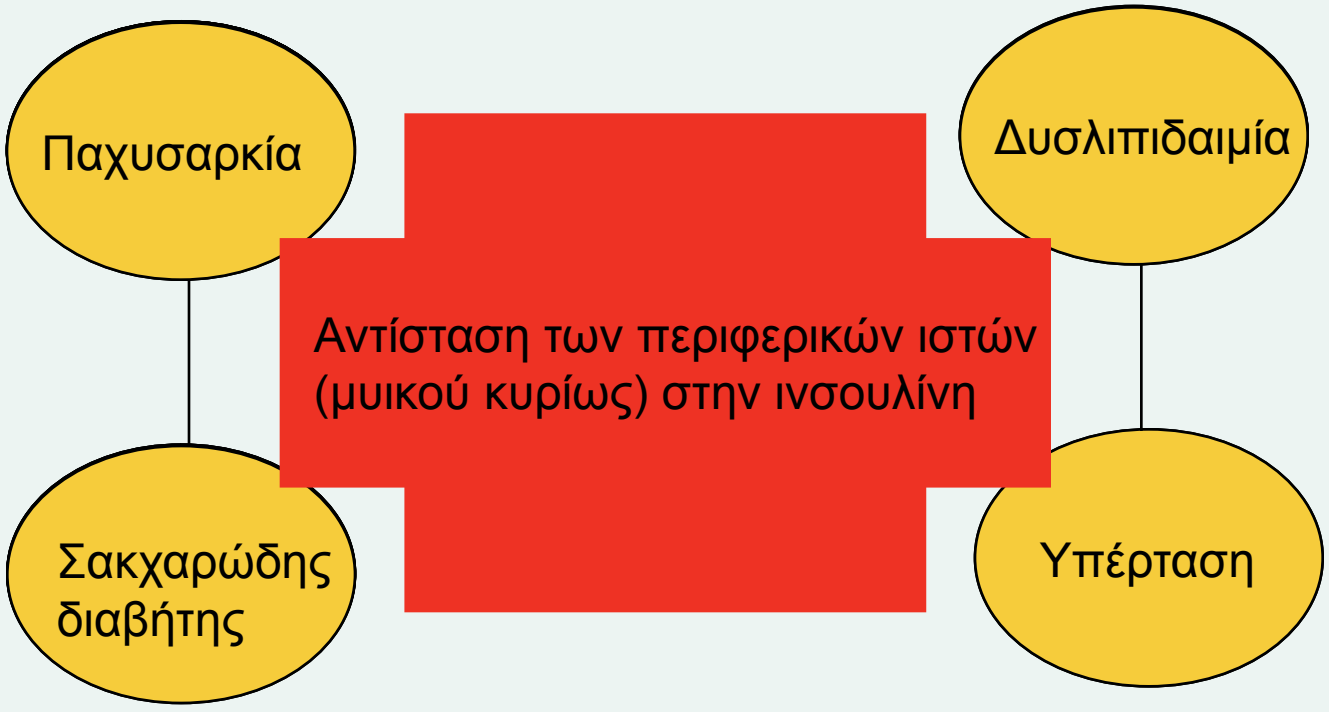
TABLE 2. **Rheumatoid Arthritis Susceptibility Alleles Identified in Association Studies and Suggested Underlying Mechanisms***

Disease susceptibility allele	Gene product	Suggested mechanism	Reference
HLA-DRB1–shared epitope alleles	HLA-DRβ chain	T-cell selection and maturation Immune response to specific peptides	47, 48 29
TNFSR11A	RANK	Osteoclast differentiation	38, 39
CRHA2	CRH	Defective HPA response to inflammation	43-45
Slc2F2T	SCL22A4 organic cation transporter	Regulates lymphocyte activation in secondary lymphoid organs and/or contributes to local inflammation	46
Runx1	RUNX1 (Runt-related transcription factor)	Regulates expression of SCL22A4	46

*CRH = corticotropin-releasing hormone; HPA = hypothalamus-pituitary axis; RANK = receptor activator of nuclear factor kB.

Μεταβολικό Σύνδρομο

Το μεταβολικό σύνδρομο είναι ένα πάζλ μεταβολικών διαταραχών που οδηγούν αναπόφευκτα στην αθηρωμάτωση και την καρδιαγγειακή νόσο



Επιπολασμός: 25 και 18% στους άντρες και γυναίκες αντίστοιχα

Μεταβολικό Σύνδρομο

- Σήμερα υπάρχουν 1 δισεκατομμύριο υπέρβαρα άτομα στον κόσμο, ενώ τα παχύσαρκα άτομα φτάνουν τα 300 εκατομμύρια.
- Η Ελλάδα είναι η πρώτη χώρα της ΕΕ σε συχνότητα παχυσαρκίας ενηλίκων και δεύτερη (μετά την Ιταλία) σε συχνότητα παιδικής παχυσαρκίας.
- Τα καρδιαγγειακά νοσήματα αποτελούν την πρώτη αιτία θανάτου τόσο στους άνδρες όσο και στις γυναίκες στον δυτικό κόσμο (ακολουθεί ο καρκίνος).

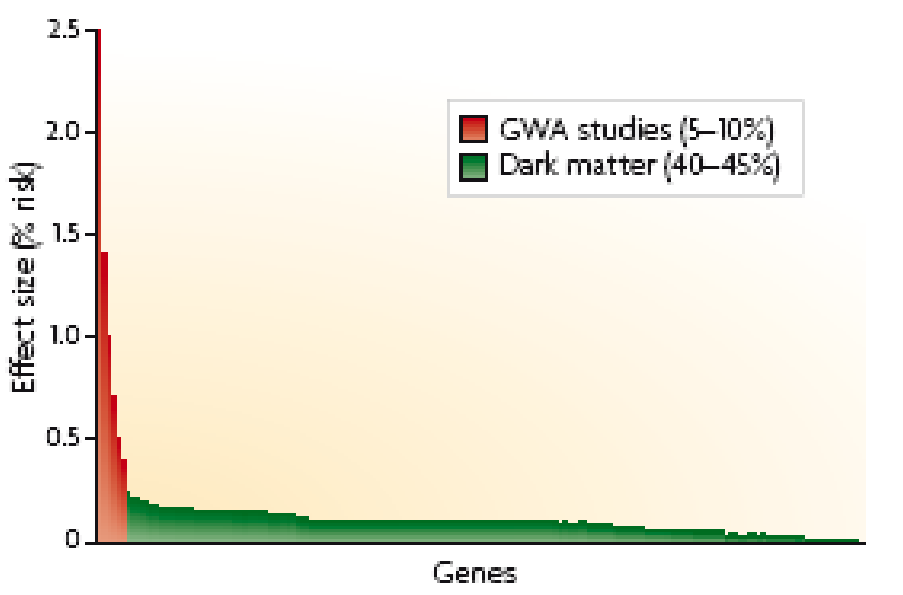
- <u>Υπόβαθρο</u>

- Κύριος πυροδοτητής του μεταβολικού συνδρόμου θεωρείται η κοιλιακή παχυσαρκία, καθώς το μεγάλο κοιλιακό κύτταρο εκκρίνει κυτταροκίνες, όπως η πρωτεΐνη TNFα που επιδρά στον υποδοχέα της ινσουλίνης και συμβάλλει στη δημιουργία ινσουλινοαντίστασης.
- Επιπρόσθετα, η αυξημένη απελευθέρωση λιπαρών οξέων επί κοιλιακής παχυσαρκίας οδηγεί, αφενός μεν, σε αυξημένη σύνθεση τριγλυκεριδίων, αφετέρου δε, σε αυξημένη σύνθεση ειδικών αθηρωματογόνων μορίων (μικρές πυκνές LDL) και σε μείωση της προστατευτικής χοληστερίνης (HDL). Οι δυο αυτές προσεγγίσεις καταλήγουν και στη δημιουργία δυο οντοτήτων, οι οποίες αμφότερες διευκολύνουν την αθηρωματική διαδικασία: στον διαβήτη με μεταβολικό σύνδρομο και στο μεταβολικό σύνδρομο χωρίς διαβήτη., οι οποίες οξύνονται μέσα σε ένα ‘τοξικό’ διατροφικό περιβάλλον.
- Έτσι ιδιαίτερη έμφαση αποδίδεται τελευταία στην σύνδεση φλεγμονής και ινσουλινοαντίστασης και τη σημασία εκτίμησης ειδικών πρωτεϊνών φλεγμονής, όπως η CRP.

Μεταβολικό Σύνδρομο - Γενετική

- 2058 άνδρες διδύμους: MZ 31.6% vs 63%, για υπέρταση, διαβήτη, παχυσαρκία
- 236 γυναίκες διδύμους: Κληρονομησιμότητα: Παχυσαρκία=0.61, Ινσουλίνη=0.87, Δυσλιπιδαιμία=0.25
- Northern Manhattan Family Study (803 άτομα από 89 οικογένειες Carribean-Hispanic) → κληρ. Του MetS: 24%, ενώ Λιπίδια/Γλυκόζη/Παχυσαρκία:44% - Υπέρταση 20%
- Μελέτες σύνδεσης: 3q27, 17p12, 1q23-31, chr. 6 (D6S403-D6S264), chr.7 (D7S479-D7S471)……..!
- Μη επαλήθευση σε άλλους πληθυσμούς!!!

Ενδείξεις από GWAS για το ΜΣ



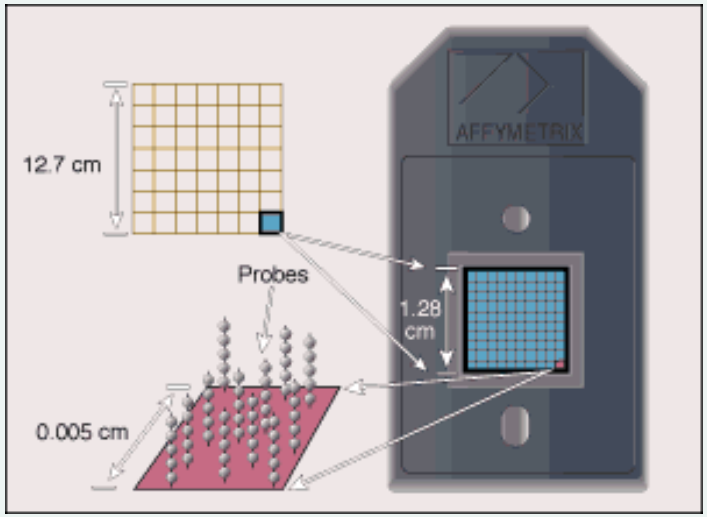
- Γονίδια από GWAS θα έχουν την μεγαλύτερη επίδραση αλλά αντιπροσωπεύουν ένα 5-10% της ποικιλότητας των χαρακτηριστικών του ΜΣ
- Το υπόλειπο 40-45% οφείλεται σε γονίδια όπως αυτά πού ρυθμίζουν τα επίπεδα των λιπιδίων και της πίεσης του αίματος: MAF<5%

- Υποδοχέας της μελανοκορτίνης 4 – σε πληθυσμούς της ευρώπης → ρυθμιστής της ισορροπίας της ενέργειας (υπερφαγία και παχυσαρκία)
- FTO (γονίδιο μάζας λίπους και παχυσαρκίας)
- MLX1L – μεταγραφικός ρυθμιστής ενζύμων που σχετίζονται με επίπεδα τριγλυκεριδίων του πλάσματος
- TCF7L2 – φαίνεται να επηρεάζει την λειτουργία των παγκρεατικών Β κυττάρων

- Σε ερευνητικό ακόμη επίπεδο: Λεπτίνη, Λιποπονεκτίνη, TNFα, Ιντερλευκίνη-6, CRP, SAA
- Μελέτες γονιδίων: melanocortin-3 receptor (MC3R), melanocortin-4 receptor (MC4R), leptin (LEP), leptin receptor (LEPR), tumor necrosis factor-alpha (TNF-alpha), interleukin-6 (IL-6), Agouti-related protein (AGRP), peroxisome proliferator-activated receptor-γ (PPAR gamma), insulin receptor (IR), glucocorticoid receptor (GR)

Τεχνολογική εξέλιξη - DNA chips και Πλατφόρμες ανάλυσης

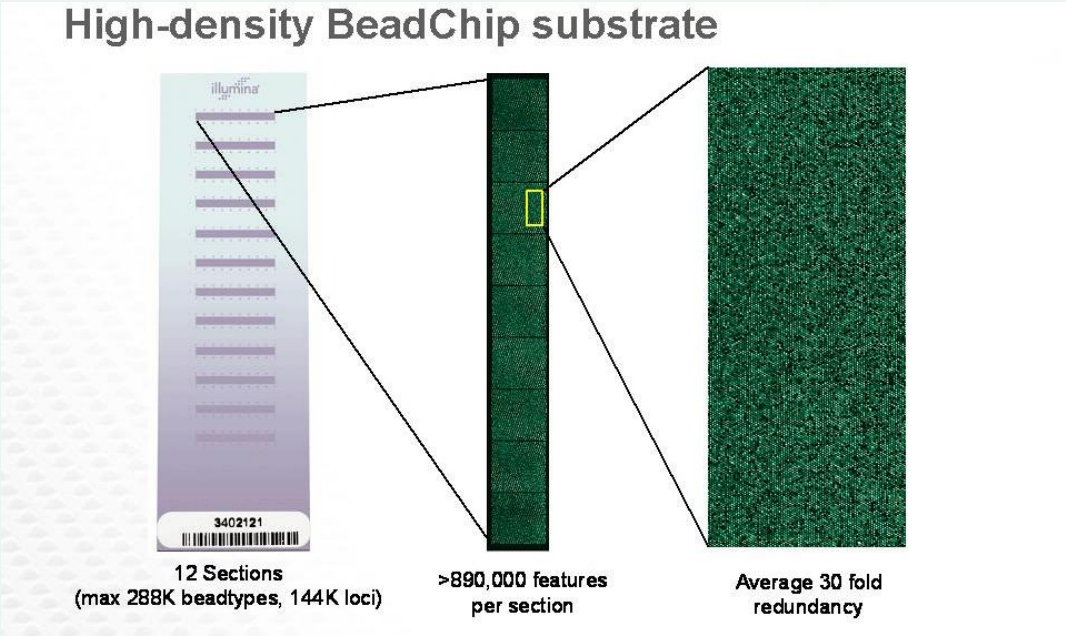
Affymetrix



100,000 or 500,000 Quasi-Random SNPs

65% των κοινών μεταλλάξεων

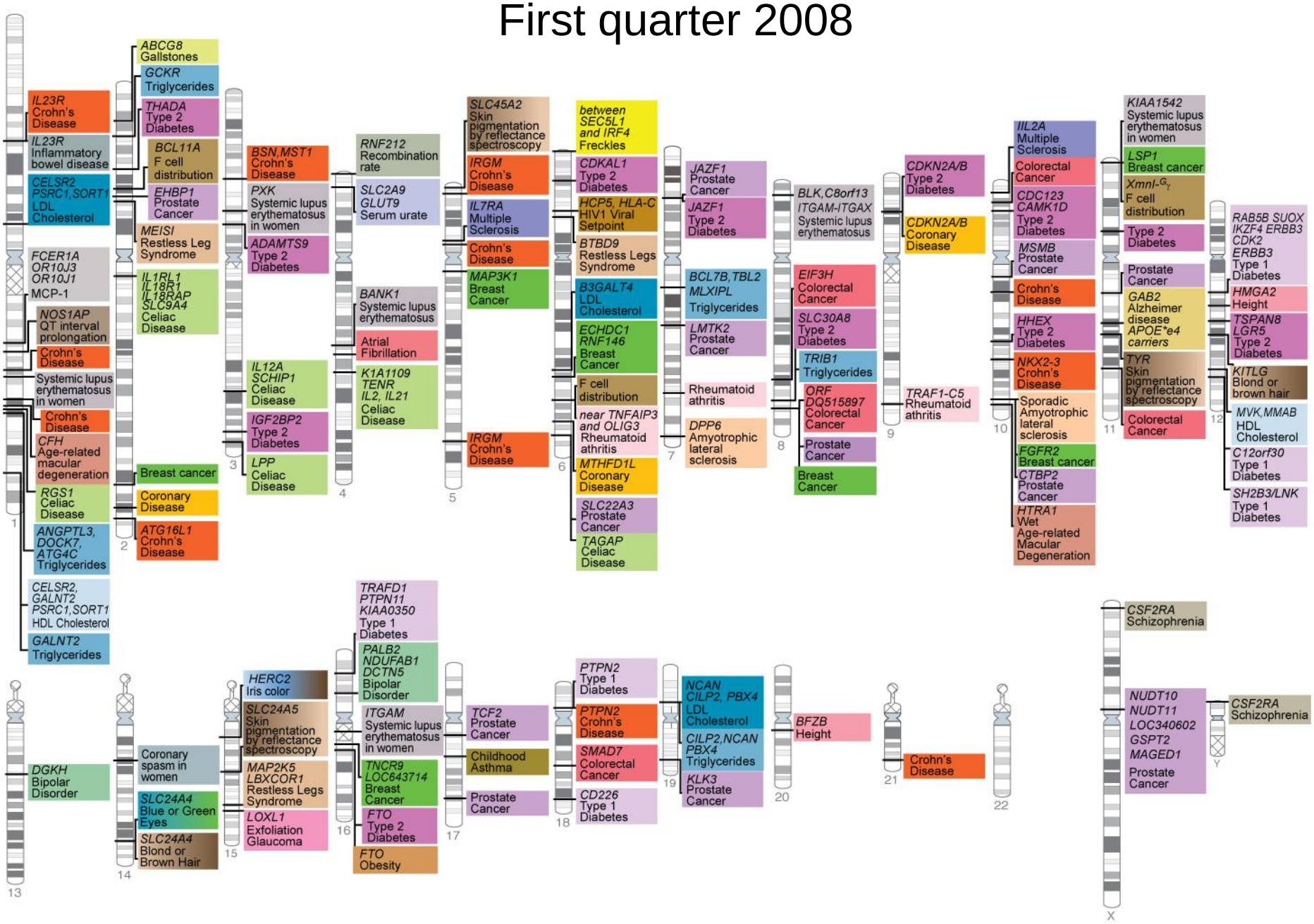
Illumina



100,000, 317,000, 550,000, 650,000Y SNPs

75% των κοινών μεταλλάξεων

First quarter 2008



Manolio, Brooks, Collins, J. Clin. Invest., May 2008

ΔΙΑΛΕΙΜΑ!!!

deCODE me

genetic profile

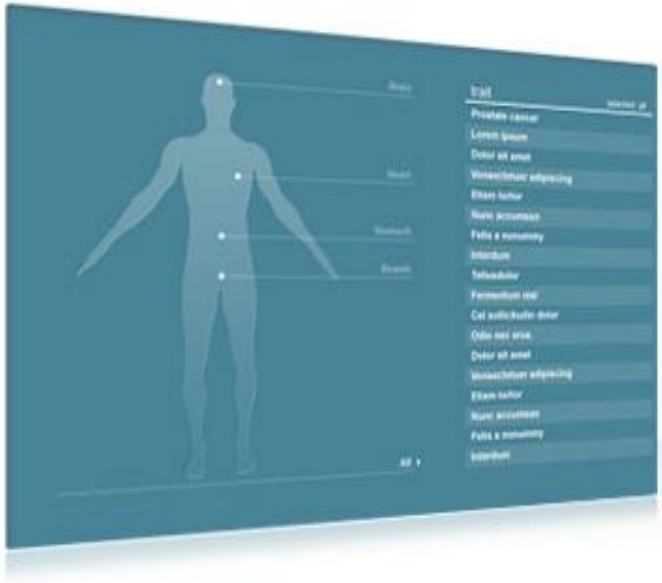
what is decodeme?

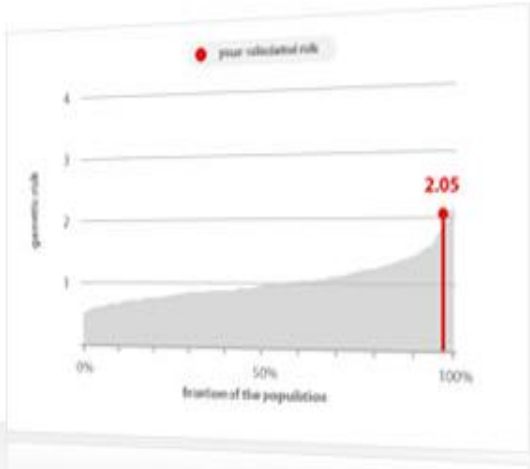
Login to myCODE

genetic profile

Concept Ancestry

- Home
- What is deCODEme?
  - deCODEme concept
  - Genetic profile
  - Ancestry
  - Comparisons
  - Physical attributes
  - About settings
  - How to order
- About deCODE
- Signup
- Login to myCODE





900,000 SNPs < $1,000

discover your genetic profile

deCODEme provides you with an introduction to your genome under expert guidance by a world leader in human genetics. We will analyze your genetic information, store it securely and give you up-to-date information about your genetic profile.

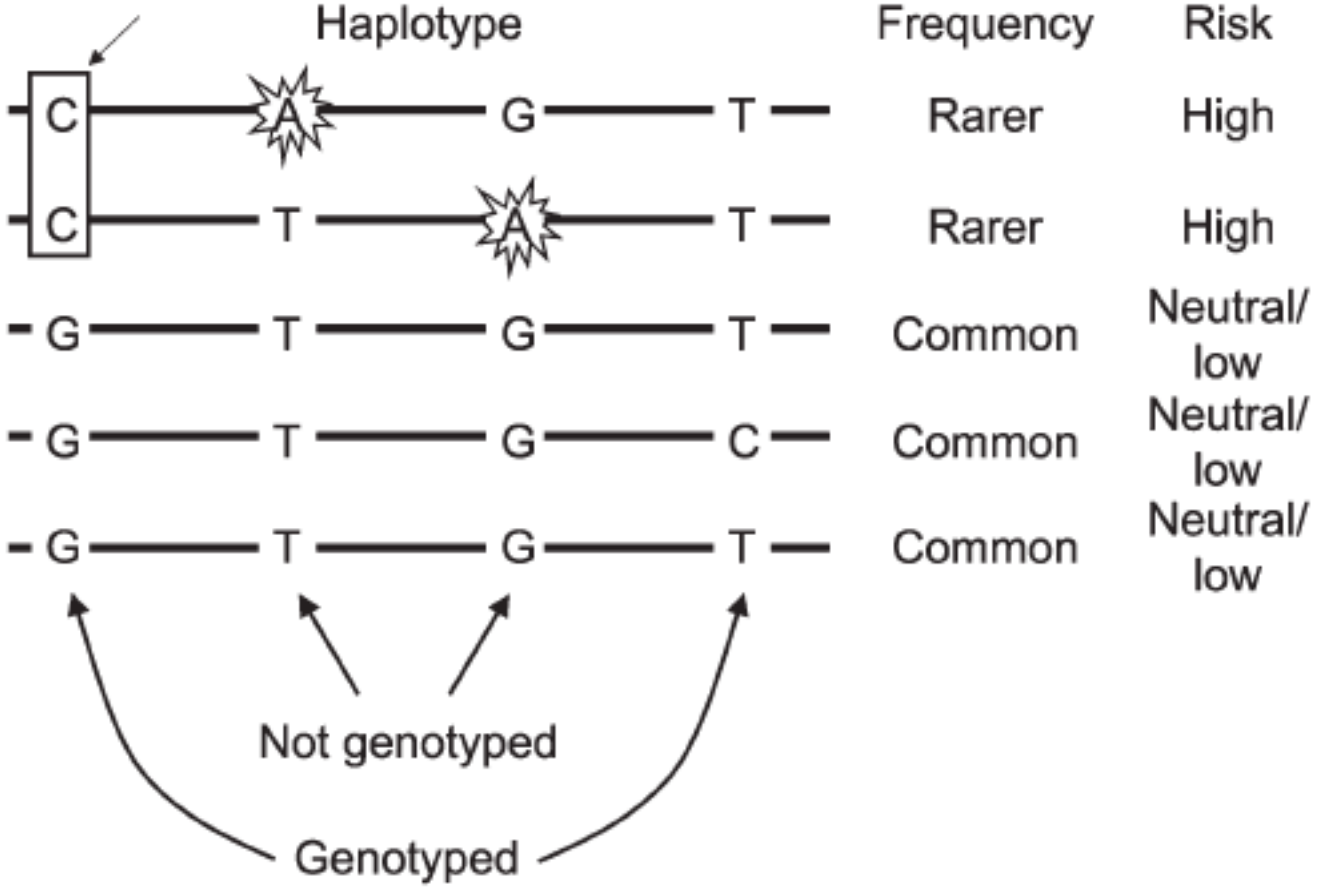
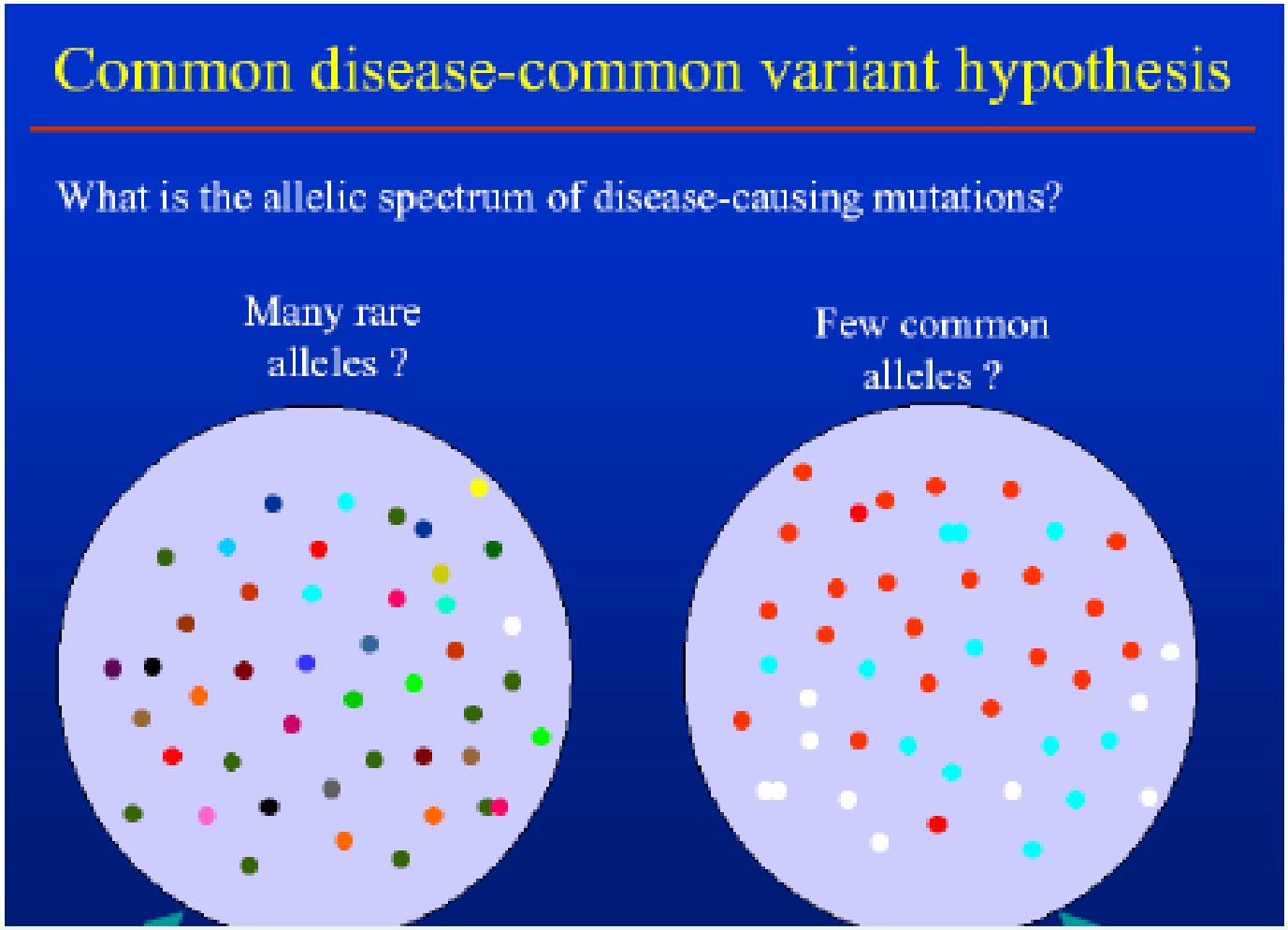


Figure 2. A common SNP may be strongly associated because it tags multiple rarer causal variants. In this hypothetical example, the C allele of the genotyped SNP on the left (indicated by the box) is strongly associated with disease risk because it tags a combination of two rarer causal variants which are themselves only weakly correlated with the associated SNP. Sequencing in affected individuals carrying high-risk haplotypes might be required to uncover the actual causal variants, which in this example have not been genotyped.

Η υπόθεση της κοινής ασθένειας = κοινή μετάλλαξης



Οι μελέτες συσχέτισης μάλλον θα αποτύχουν σε αυτήν την περίπτωση

Πιθανή επιτυχία

Η υπόθεση της κοινής ασθένειας = κοινή μετάλλαξης

Common disease, polygenic effects

Genotype	*Risk of disease*
AA	*0.01*
AG	*0.012*
GG	*0.0144*

Disease prevalence ~1 in 100

Each extra *G* allele increases risk by ~1.2 times

Frequency of *G* in controls ~ 5%
Frequency of *G* in cases ~ 6%

Η υπόθεση της κοινής ασθένειας = κοινή μετάλλαξης

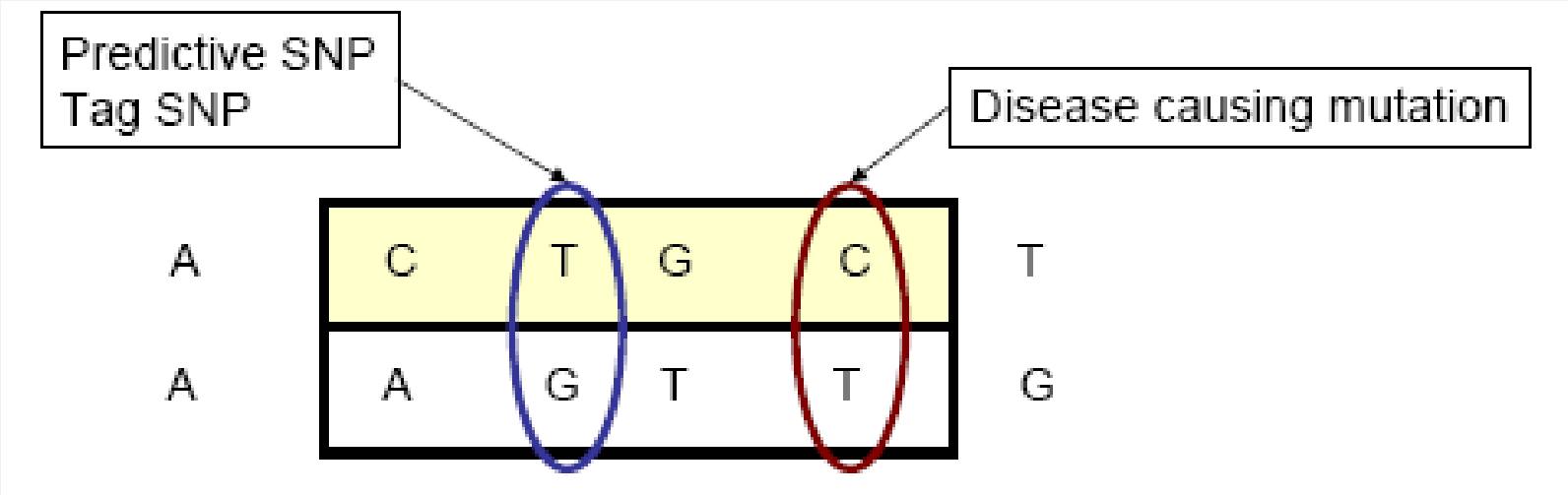
Rare disease, major gene effect

Genotype	*Risk of disease*
AA	*0.001*
AG	*0.001*
GG	*0.95*

Disease prevalence	~1 in 1000
Individuals with *GG* are ~1000 times more likely to get disease	
Frequency of *G* in controls	~ 5%
Frequency of *G* in cases	~ 96%

~10 εκατομμύρια SNPs στο ανθρώπινο γονιδίωμα

~ 500,000 tag SNPs

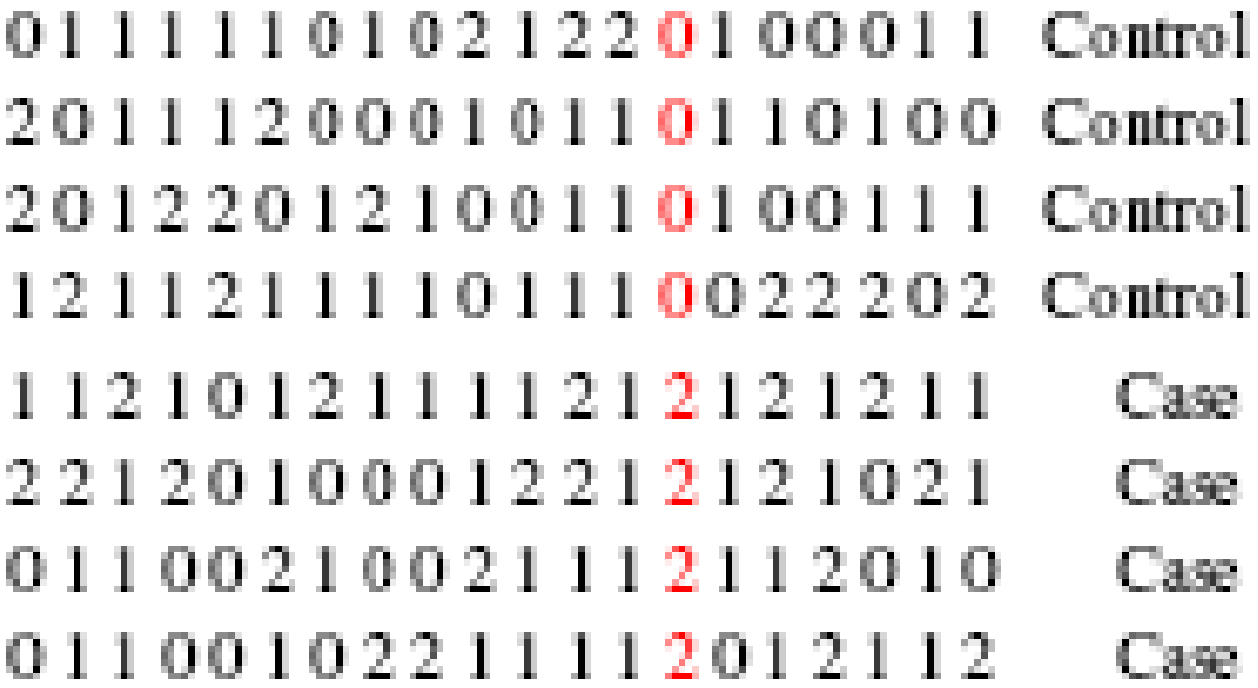


Μελέτες συσχέτισης σε επίπεδο γονιδιώματος (χωρίς επιλογή υποψηφίων γονιδίων)

Μελέτες Σάρωσης του γονιδιώματος (Genome wide scans, GWAS)

Οι μελέτες σάρωσης του γονιδώματος είναι μελέτες αναζήτησης κοινών γενετικών μεταλλάξεων, σε όλο το γονιδίωμα, και συσχέτισης αυτών με φαινοτυπικά χαρακτηριστικά ασθενειών.

- **Απαιτούν μεγάλο αριθμό ασθενών/controls (>1,000).**
- **Απαιτούν μεγάλο αριθμό SNPs (>100,000).**
- **Υποβάλλονται σε αυστηρή στατιστική διόρθωση (Bonferroni correction).**



Η βασική ιδέα πίσω από τις μελέτες σάρωσης του γονιδιώματος (GWAS)

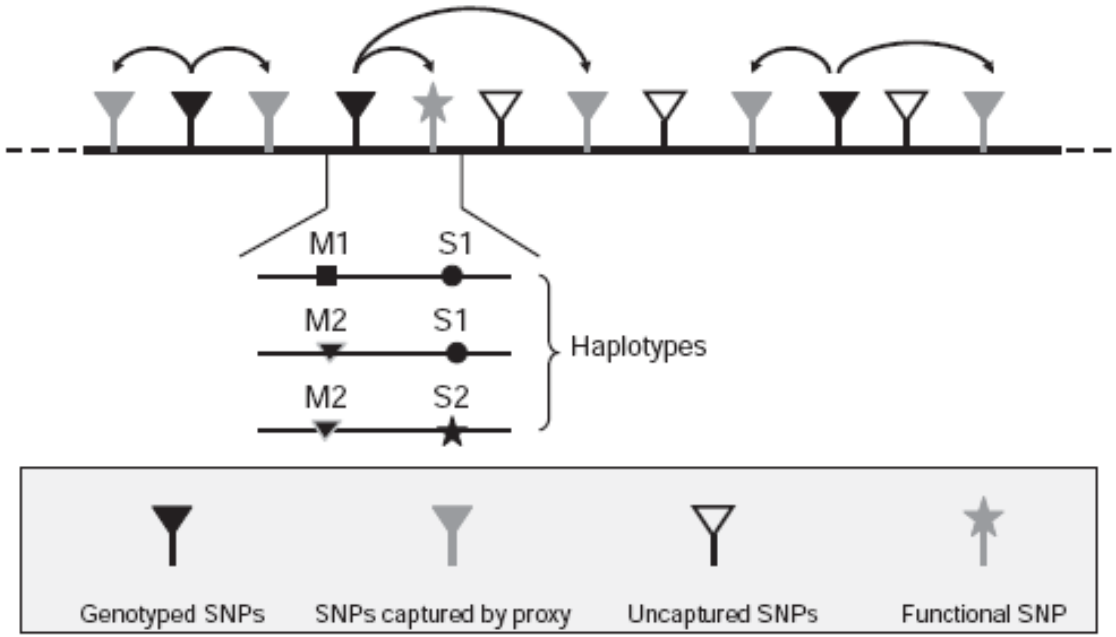
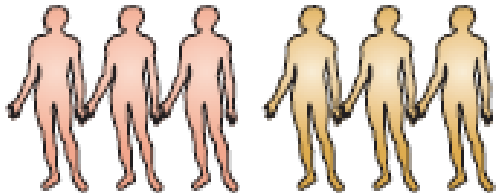
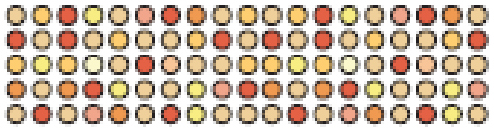


Figure 1: This figure from [CT07] gives an example of the underlying idea behind genome association studies. The black triangles represent tagged SNPs, which are sequenced. Shaded markers are SNPs associated with tagged SNPs by linkage disequilibrium. The unfilled markers are not associated with tagged SNPs and thus can't be typed by the study. The shaded star represents a SNP that has a causal relationship with the disease. The haplotypes shown give an example of linkage disequilibrium. Each SNP has two possible alleles. $\{M_1, M_2\}$ for the triangle and $\{S_1, S_2\}$ for the star. Only 3 of the 4 possible haplotypes (SNP combinations) are shown to occur. This means linkage disequilibrium is in action and the tagged SNP is associated with the causal SNP.

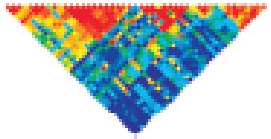
Large cohort of cases and controls ($n > 1{,}000$)
- Matched for confounding variables, such as race, ethnicity and sex
- Stratified in order to maximize signals



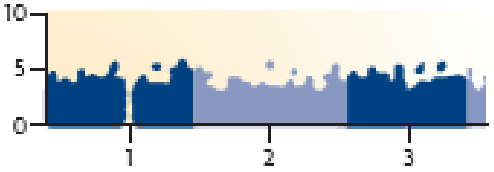
Microarray-based SNP genotyping
- ~1 million random marker SNPs or ~25,000 risk-enhancing SNPs (for example, nsSNPs)



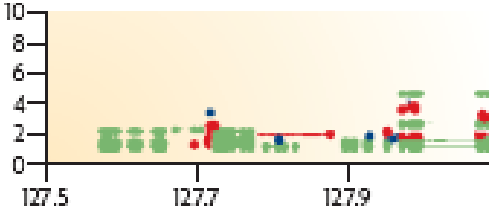
Derivation of haplotypes
- Predicated on International HapMap



Detection of association signals
- χ^2 or similar test
- Uncorrected $P < 10^{-7}$ or false discovery rate-like correction



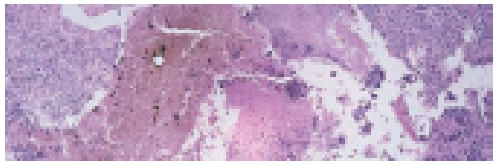
Fine mapping of association signal (see FIG. 2)
- Directed genotyping of additional SNPs in region
- Fine mapping of LD in region of association
- Empirical derivation of haplotypes
- Examination of effect of stratification, if available



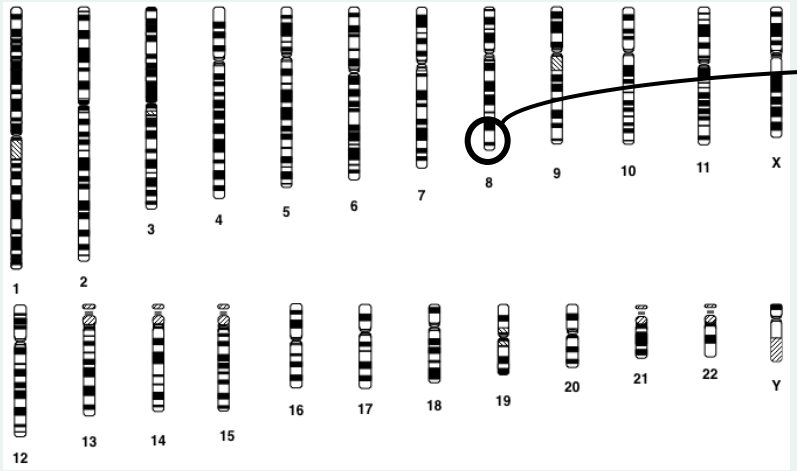
Replication of association
- Large independent cohort of cases and controls ($n > 1{,}000$)
- Genotyping of nominated candidate SNPs (<20)
- χ^2 or similar test; replication of initial signal

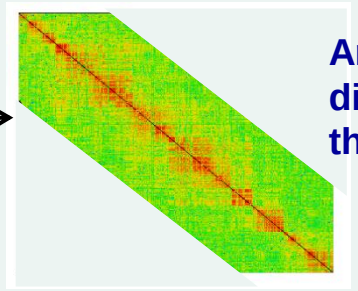
Genotypes	CC	AA	CA	Total
Cases observed	59	27	98	184
Controls observed	60	89	36	185
Total	119	116	134	369

Biological validation of association
- Identification of risk-enhancing variant
- Examination of functional consequence of variant
- Determination of mechanism of risk-enhancement



10 million SNPs across the genome

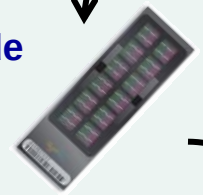




Areas of linkage disequilibrium across the genome

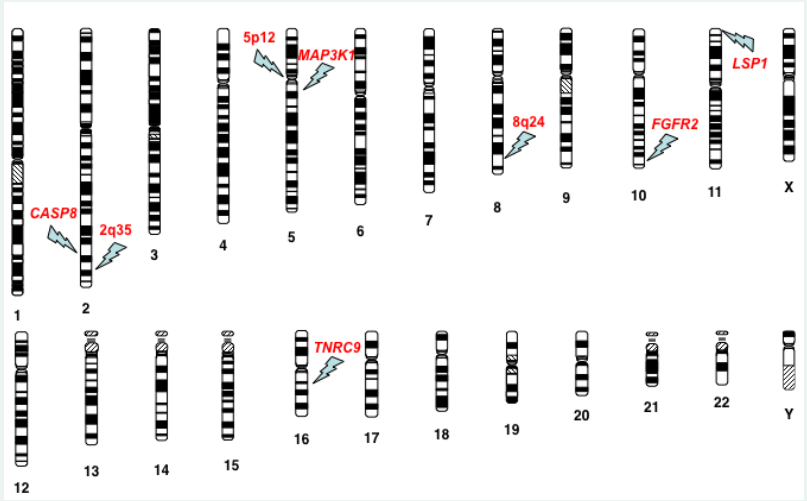
Selection of tagSNPs to capture common genetic variation in population under study

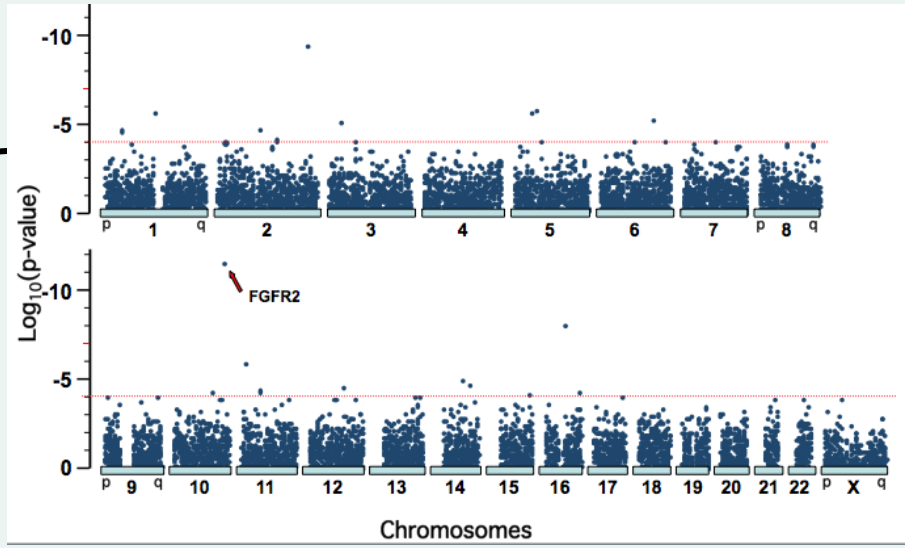
Genome wide SNP chips



Association with disease risk in discovery and replication studies

Mapping of susceptibility loci





Τύποι μελετών σάρωσης του γονιδιώματος (GWAS)

Table 1. Study Designs Used in Genome-wide Association Studies

	Case-Control	Cohort	Trio
Assumptions	Case and control participants are drawn from the same population Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified Genomic and epidemiologic data are collected similarly in cases and controls Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls	Participants under study are more representative of the population from which they are drawn Diseases and traits are ascertained similarly in individuals with and without the gene variant	Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents
Advantages	Short time frame Large numbers of case and control participants can be assembled Optimal epidemiologic design for studying rare diseases	Cases are incident (developing during observation) and free of survival bias Direct measure of risk Fewer biases than case-control studies Continuum of health-related measures available in population samples not selected for presence of disease	Controls for population structure; immune to population stratification Allows checks for Mendelian inheritance patterns in genotyping quality control Logistically simpler for studies of children's conditions Does not require phenotyping of parents
Disadvantages	Prone to a number of biases including population stratification Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases Overestimate relative risk for common diseases	Large sample size needed for genotyping if incidence is low Expensive and lengthy follow-up Existing consent may be insufficient for GWA genotyping or data sharing Requires variation in trait being studied Poorly suited for studying rare diseases	May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset Highly sensitive to genotyping error

GWAS στον Ηλικιακά-σχετιζόμενο εκφυλισμό της ωχράς κηλίδας

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

www.sciencemag.org SCIENCE VOL 308 15 APRIL 2005

- **96 ασθενείς και 50 controls**
- **116,204 SNPs**
- **Μετάλλαξη στο γονίδιο του παράγοντα του συμπληρώματος H (CFH)**
- **Αναστολέας της ενεργοποίησης του C3 στην φλεγμονή**
- **Ομοζυγώτες έχουν x7.4 ρίσκο για την ασθένεια**

- **Επιτυχία στην συσχέτιση λόγω σημαντικού ποσοστού της συμμετοχής γενετικού παράγοντα**

Vol 447|7 June 2007|doi:10.1038/nature05911

ARTICLES

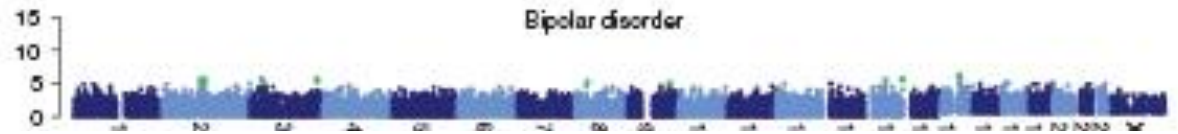
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

WTCCC

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point P values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple disease phenotypes; has generated a genome-wide genotype database for future studies of common diseases in the British population; and shown that, provided individuals with non-European ancestry are excluded, the extent of population stratification in the British population is generally modest. Our findings offer new avenues for exploring the pathophysiology of these important disorders. We anticipate that our data, results and software, which will be widely available to other investigators, will provide a powerful resource for human genetics research.

WTCCC αποτελέσματα (εντοπισμός SNPs με MAF >5% είναι 43% με RR 1.3 και 80% για RR 1.5, για $P<5x10^{-7}$)



- 24 ανεξάρτητες ενδείξεις ισχυρής συσχέτισης

Table 3 | Regions of the genome showing the strongest association signals

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF), additive	$\log_{10}$(BF), general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
					Standard analysis								
BD	16p12	23.3–23.62	rs420259	2.19×10^{-04}	6.29×10^{-08}	1.96	4.79	A	G	2.08 (1.60–2.71)	2.07 (1.6–2.69)	0.282	0.248
CAD	9p21	21.93–22.12	rs1333049	1.79×10^{-14}	1.16×10^{-13}	11.66	11.19	C	C	1.47 (1.27–1.70)	1.9 (1.61–2.24)	0.474	0.554
CD	1p31	67.3–67.48	rs11805303	6.45×10^{-13}	5.85×10^{-12}	10.07	9.41	T	T	1.39 (1.22–1.58)	1.86 (1.54–2.24)	0.317	0.391
CD	2q37	233.92–234	rs10210302	7.10×10^{-14}	5.26×10^{-14}	11.11	11.28	T	C	1.19 (1.01–1.41)	1.85 (1.56–2.21)	0.481	0.402
CD	3p21	49.3–49.87	rs9858542	7.71×10^{-07}	3.58×10^{-08}	4.24	5.22	A	A	1.09 (0.96–1.24)	1.84 (1.49–2.26)	0.282	0.331
CD	5p13	40.32–40.66	rs17234657	2.13×10^{-13}	1.99×10^{-12}	10.41	9.89	G	G	1.54 (1.34–1.76)	2.32 (1.59–3.39)	0.125	0.181
CD	5q33	150.15–150.31	rs1000113	5.10×10^{-08}	3.15×10^{-07}	5.36	5.01	T	T	1.54 (1.31–1.82)	1.92 (0.92–4.00)	0.067	0.098
CD	10q21	64.06–64.31	rs10761659	2.68×10^{-07}	1.75×10^{-06}	4.69	4.13	G	A	1.23 (1.05–1.45)	1.55 (1.3–1.84)	0.461	0.406
CD	10q24	101.26–101.32	rs10883365	1.41×10^{-08}	5.82×10^{-08}	5.91	5.48	G	G	1.2 (1.03–1.39)	1.62 (1.37–1.92)	0.477	0.537
CD	16q12	49.02–49.4	rs17221417	9.36×10^{-12}	3.98×10^{-11}	8.93	8.47	G	G	1.29 (1.13–1.46)	1.92 (1.58–2.34)	0.287	0.356
CD	18p11	12.76–12.91	rs2542151	4.56×10^{-08}	2.03×10^{-07}	5.42	5.00	G	G	1.3 (1.14–1.48)	2.01 (1.46–2.76)	0.163	0.208
RA	1p13	113.54–114.16	rs6679677	4.90×10^{-26}	5.55×10^{-25}	22.36	21.99	A	A	1.98 (1.72–2.27)	3.32 (1.93–5.69)	0.096	0.168
RA	6	MHC	rs6457617*	3.44×10^{-76}	5.18×10^{-75}	74.84	73.18	T	T	2.36 (1.97–2.84)	5.21 (4.31–6.30)	0.489	0.685
T1D	1p13	113.54–114.16	rs6679677	1.17×10^{-26}	5.43×10^{-26}	23.07	22.83	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
T1D	6	MHC	rs9272346*	2.42×10^{-134}	5.47×10^{-134}	141.9	142.2	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.150
T1D	12q13	54.64–55.09	rs11171739	1.14×10^{-11}	9.71×10^{-11}	8.89	8.24	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
T1D	12q24	109.82–111.49	rs17696736	2.17×10^{-15}	1.51×10^{-14}	12.53	11.88	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
T1D	16p13	10.93–11.37	rs12708716	9.24×10^{-08}	4.92×10^{-07}	5.15	4.70	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.350	0.297
T2D	6p22	20.63–20.84	rs9465871	1.02×10^{-06}	3.34×10^{-07}	4.15	3.98	C	C	1.18 (1.04–1.34)	2.17 (1.6–2.95)	0.178	0.218
T2D	10q25	114.71–114.81	rs4506565	5.68×10^{-13}	5.05×10^{-12}	10.14	9.43	T	T	1.36 (1.2–1.54)	1.88 (1.56–2.27)	0.324	0.395
T2D	16q12	52.36–52.41	rs9939609	5.24×10^{-08}	1.91×10^{-07}	5.35	5.05	A	A	1.34 (1.17–1.52)	1.55 (1.3–1.84)	0.398	0.453
					Multi-locus analysis								
T1D	4q27	123.26–123.92	rs6534347	4.48×10^{-07}	1.83×10^{-06}	5.15	4.69	A	A	1.30 (1.10–1.55)	1.49 (1.25–1.78)	0.351	0.402
T1D	12p13	9.71–9.86	rs3764021	7.19×10^{-05}	5.08×10^{-08}	2.12	4.55	C	T	1.57 (1.38–1.79)	1.48 (1.25–1.75)	0.467	0.426
					Sex differentiated analysis								
RA	7q32	130.80–130.84	rs11761231	3.91×10^{-07}	1.37×10^{-06}	-	-	G	A	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
					Combined cases								
RA+T1D	10p15	6.07–6.17	rs2104286	5.92×10^{-08}	2.52×10^{-07}	5.26	4.45	T	C	1.35 (1.11–1.65)	1.62 (1.34–1.97)	0.286	0.245

Μελέτες Σάρωσης του γονιδιώματος (Genome wide scans) – Το μέγεθος πληθυσμού είναι σημαντικό (Μετά-ανάλυση)

Results. Twenty-nine studies with 43 comparisons including 13 rheumatoid arthritis (RA), six systemic lupus erythematosus (SLE), six type-1 DM (T1D), three Grave's disease (GD), four inflammatory bowel diseases (IBD), three juvenile idiopathic arthritis (JIA), two psoriasis, two multiple sclerosis, two Addison's disease and two Celiac disease were available for the meta-analysis. The overall odds ratios (ORS) for T-allele, T/T and T/T + C/T genotypes were significantly increased in RA, SLE, GD and T1D (OR for T-allele = 1.58, 1.49, 1.85, 1.61, respectively, $P < 0.00001$). This meta-analysis showed the association between the T-allele and the T/T genotype and JIA (OR = 1.34, $P = 0.03$; OR = 1.97, $P = 0.02$) but did not reveal the

A

Review: RA
Comparison: 01 RA and PTPN22 C1858T SNP
Outcome: 01 T- *vs* C-allele

Study or sub-category	RA n/N	Control n/N	OR (random) 95% CI	Weight %	OR (random) 95% CI
Seldin	436/2060	433/2800		15.51	1.47 [1.27, 1.70]
Begovich-1	131/950	84/950		7.36	1.65 [1.23, 2.20]
Begovich-2	283/1680	161/1852		11.33	2.13 [1.73, 2.61]
Gomez	33/596	27/616		2.85	1.28 [0.76, 2.15]
Hinks	316/1772	123/1190		10.37	1.88 [1.51, 2.35]
Oene-1	255/1812	100/1206		9.29	1.81 [1.42, 2.31]
Oene-2	112/656	36/376		4.51	1.94 [1.30, 2.90]
Orozco	171/1652	153/2072		10.07	1.45 [1.15, 1.82]
Simkins	262/1738	112/1126		9.71	1.61 [1.27, 2.03]
Steer	96/604	63/748		5.87	2.05 [1.47, 2.88]
Wesoly	99/832	162/1782		8.34	1.35 [1.04, 1.76]
Zhernakova	44/302	92/1046		4.79	1.77 [1.20, 2.60]
Total (95% CI)	14654	15764		100.00	1.68 [1.53, 1.84]

Total events: 2238 (RA), 1546 (Control)
Test for heterogeneity: $\chi^2 = 16.97$, df = 11 ($P = 0.11$), $I^2 = 35.2\%$
Test for overall effect: Z = 10.85 ($P < 0.00001$)

0.1 0.2 0.5 1 2 5 10
Control RA

nature genetics

Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer

COGENT Study[1]

Genome-wide association (GWA) studies have identified multiple loci at which common variants modestly influence the risk of developing colorectal cancer (CRC). To enhance power to identify additional loci with similar effect sizes, we conducted a meta-analysis of two GWA studies, comprising 13,315 individuals genotyped for 38,710 common tagging SNPs. We undertook replication testing in up to eight independent case-control series comprising 27,418 subjects. We identified four previously unreported CRC risk loci at 14q22.2 (rs4444235, ***BMP4***; $P = 8.1 \times 10^{-10}$), 16q22.1 (rs9929218, ***CDH1***; $P = 1.2 \times 10^{-8}$), 19q13.1 (rs10411210, ***RHPN2***; $P = 4.6 \times 10^{-9}$) and 20p12.3 (rs961253; $P = 2.0 \times 10^{-10}$). These findings underscore the value of large sample series for discovery and follow-up of genetic variants contributing to the etiology of CRC.

Whereas inherited susceptibility is responsible for ~35% of all CRC[1], high-risk germline mutations in *APC*, the mismatch repair (MMR) genes, *MUTYH* (*MYH*), *SMAD4*, *BMPR1A* and *STK11/LKB1* account for <6% of all cases[2]. Recent GWA studies have validated the hypothesis that part of the heritable risk is caused by common, low-risk variants, identifying CRC susceptibility loci mapping to 8q24 (rs6983267)[3,4], 8q23.3 (rs16892766, *EIF3H*)[5], 10p14 (rs10795668)[5], 11q23 (rs3802842)[6], 15q13 (rs4779584)[7] and 18q21 (rs4939827, *SMAD7*)[6,8].

GWA studies are not contingent on prior information concerning candidate genes or pathways, and thereby have the ability to identify important variants in hitherto unstudied genes. However, the effect sizes of individual variants, the need for stringent thresholds for establishing statistical significance, and financial constraints on numbers of variants that can be followed up inevitably constrain study power. We recently published two separate GWA studies for CRC. To augment the power to detect additional CRC risk loci, we have conducted a meta-analysis of data from these studies and followed up the best supported associations in large sample sets. This analysis, in conjunction with a replication study using eight independent case-control series, has enabled us to identify four new loci predisposing to CRC. This brings to ten the number of independent loci conclusively associated with CRC risk, and provides additional insight into the genetic architecture of inherited susceptibility to CRC.

RESULTS

Meta-analysis of genome-wide association scans

The GWA studies were both conducted by centers in London and Edinburgh, and were both based on designs involving two-phase strategies and using samples from UK populations (Table 1 and Supplementary Table 1 online). The London phase 1 was based on genotyping 940 cases with familial colorectal neoplasia and 965 controls ascertained through the Colorectal Tumour Gene Identification (CoRGI) consortium for 555,352 SNPs using the Illumina HumanHap550 BeadChip Array. Phase 1 in the Edinburgh study consisted of genotyping 1,012 early-onset (aged ≤55 years) Scottish CRC cases and 1,012 controls for 555,510 SNPs using the Illumina HumanHap300 and HumanHap240S arrays. After applying quality control filters, the following data were available: London phase 1, 547,487 SNP genotypes from 922 familial neoplasia cases (614 with CRC and 308 with high-risk colorectal adenomas) and 927 controls; Edinburgh phase 1, 548,586 SNP genotypes from 980 CRC cases and 1,002 controls.

London phase 2 was based on genotyping 2,873 CRC cases and 2,871 controls ascertained through the National Study of Colorectal Cancer Genetics (NSCCG), whereas Edinburgh phase 2 was based on genotyping 2,057 cases and 2,111 controls. For phase 2, the London and Edinburgh samples were genotyped for a common set of SNPs: the 14,982 SNPs most strongly associated with colorectal neoplasia from London phase 1; the 14,972 most strongly associated SNPs from Edinburgh phase 1 (432 of these SNPs were common to both the London and Edinburgh lists of most strongly associated SNPs); and 13,186 SNPs showing the strongest association with CRC risk from a joint analysis of all CRC cases and controls from both phase 1 data sets (that were not already included in any of the preceding categories). Therefore, phase 2 was based on genotyping 42,708 SNPs in total. After applying quality control filters, the following data were available: London phase 2, 38,715 polymorphic SNPs in 2,854 cases and 2,822 controls; Edinburgh phase 2, 38,710 polymorphic SNPs in 2,024 cases and 2,092 controls. Overall, there were 38,710 polymorphic SNPs common to all four data sets (phases 1 and 2 in London and Edinburgh).

Prior to undertaking the meta-analysis of phases 1 and 2, we searched for potential errors and biases in the four case-control series.

[1]A full list of authors and affiliations is provided at the end of this paper.

Received 6 August; accepted 17 September; published online 16 November 2008; doi:10.1038/ng.262

Πριν την Μετα-ανάλυση

- 8q24
  - Also Prostate Region 3
  - Colorectal Adenoma
- 8q23.3 (*EIF3H*)
- 10p14
- 11q23
- 15q13
- 18q21 (*SMAD7*)

Νέοι γενετικοί τόποι

- 14q22 (*BMP4*)
- 16q22.1 (*CDH1*)
- 19q13.1 (*RHPN2*)
- 20p12.3

Κατασκευή ενός SNP χάρτη όλου του γονιδιώματος: Πόσα SNPs;

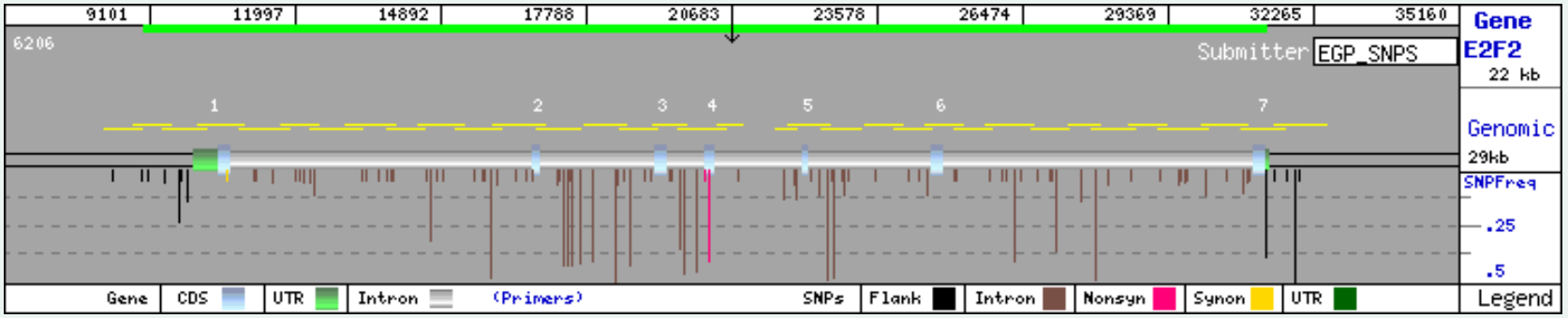
Table 1 • Occurrence of SNPs in the human population

Minimal allele frequency	Expected SNP number (millions)	Expected SNP frequency (bp)
1%	11.0	290
5%	7.1	450
10%	5.3	600
20%	3.3	960
30%	2.0	1,570
40%	0.97	3,280

Nickerson and Kruglyak, Nature Genetics, 2001

~ 10 M κοινά SNPs (> 1- 5% MAF) - 1/300 bp

Αναζήτηση SNPs για Μελέτες σάρωσης του γονιδιώματος



Συνολικό Γονιδίωμα:

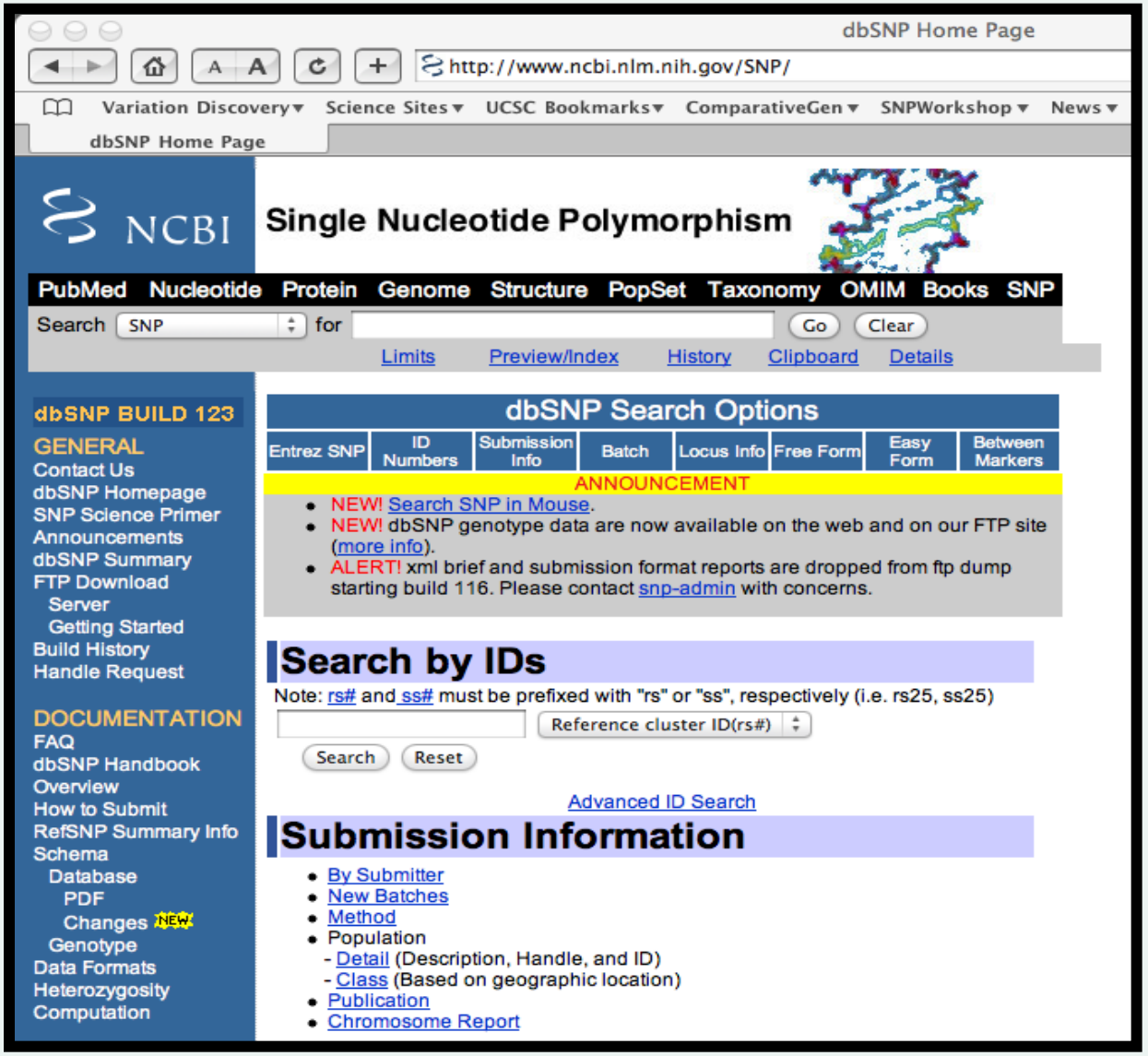
- **20,000,000 SNPs**
- **10,000,000 SNPs ≥ 5% MAF**

Μέσο Γονίδιο:

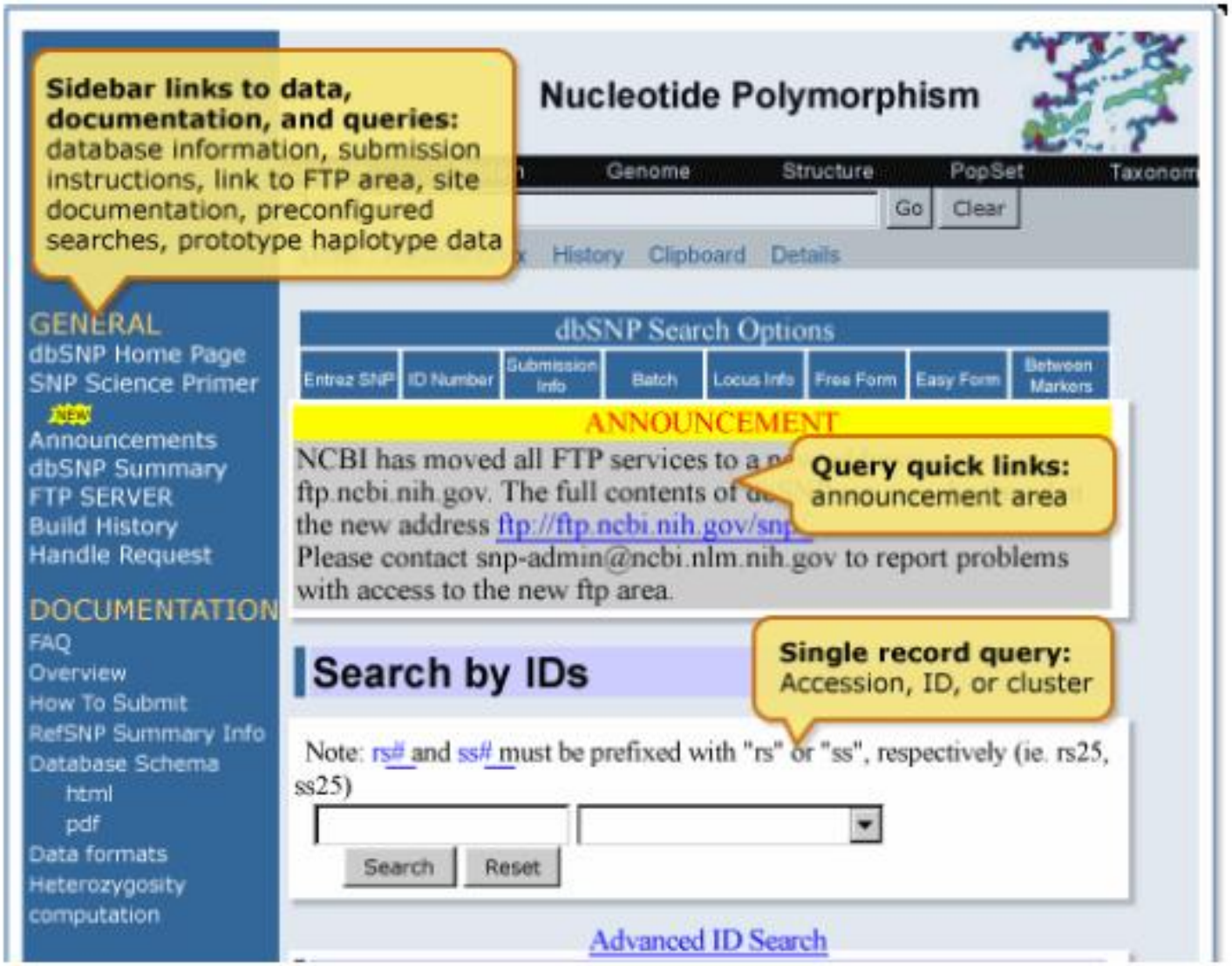
- **26.5 kb**
- **130 SNPs**
- **44 SNPs ≥5% MAF**

Υπερβολικά μεγάλος αριθμός SNPs → 20 x 10^9 γονότυποι → $10 x 10^9 για κάθε ασθένεια!!!!!!!!!!!!!!!!!! Λύση: Tag-SNPs

Αναζήτηση SNPs - dbSNP βάση δεδομένων



http://www.ncbi.nlm.nih.gov/SNP/



What is a copy number variant?

Human DNA has one copy of autosomal regions on each chromosome. However, as discovered by the Human Genome Project, many genetic regions display a variation in the number of copies (more or less than two copies). Alleles containing 0–13 gene copies have been reported across the human population [13]. These genetic variants are termed CNVs and are defined as DNA segments ranging in size from one kilobase to several megabases among individuals owing to deletion, insertion, inversion, duplication, or complex recombination [14] (Figure 1). Many groups have

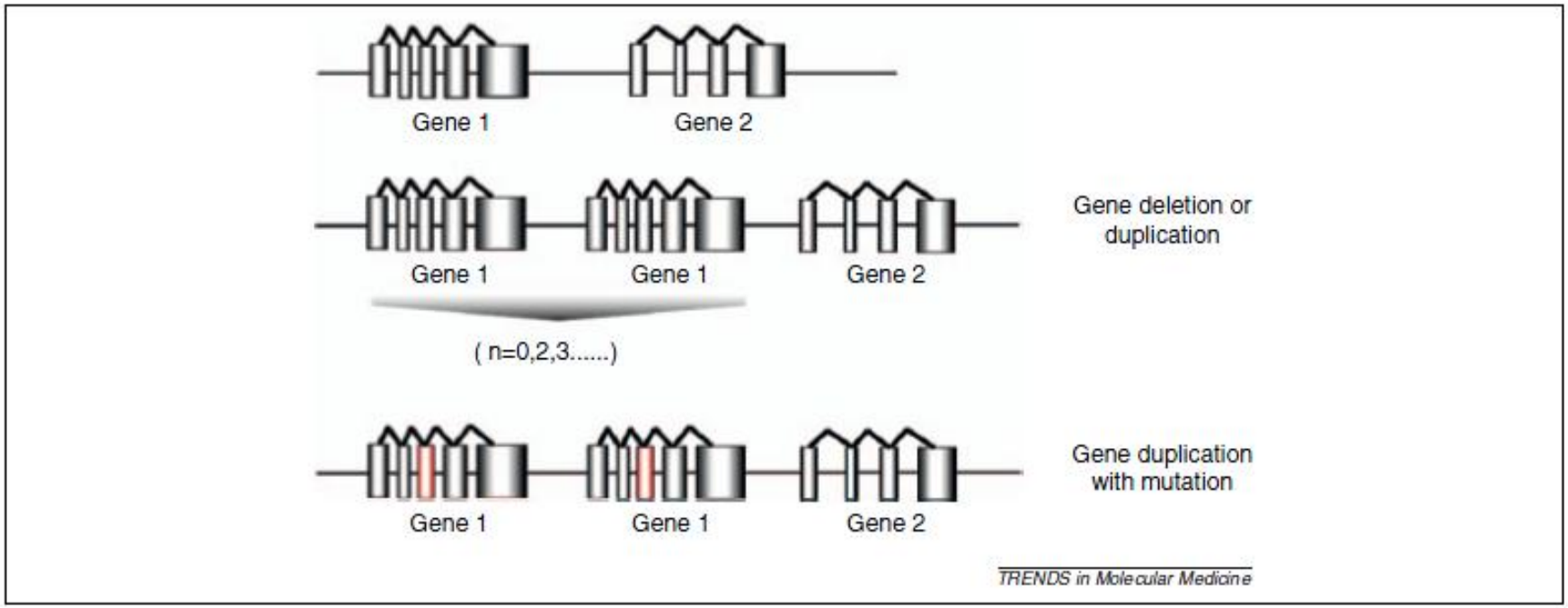


Figure 1. A diagram for copy number variants in the human genome. If the gene recombination event occurs between two genes, a gene duplication or multiplication (n=2, 3. . .) or a gene deletion (n=0) could occur. A duplication of a gene could also carry mutations from the original copy (show in red column).

Box 1. Currently available methodologies for CNV detection

Many PCR based methods have been developed for the detection of CNVs. However, the low-throughput and low reproductively rate limit the use of these methods in the clinic. Recently, new methods have been developed based on microarray hybridization technology.

Conventional methods

Long distance and multiplex PCR: this is the first method developed for the detection of *CYP2D6* CNV. Long PCR is used to amplify a >4000 bp fragment of the *CYP2D6* gene. Specific primers for *CYP2D6*5* have been designed. Two separate PCR reactions for either wild-type or mutated allele primers are run for each sample using the long PCR product of the *CYP2D6* gene. Following up with gel electrophoresis, the *CYP2D6*5* can be identified [64].

TaqMan real-time PCR: this is a PCR quantification method. TaqMan probes emit a specific report fluorophore (usually a short-wavelength colored dye, such as green) during the elongation process of PCR reactions. The fluorophore can be detected by a corresponding detection system and compared with the signal of reference genes, such as *albumin*, to calculate the amount of specific genes in each sample. This is a low-throughput detection method for CNVs [65].

Microarrays hybridization-based methods

Array comparative genomic hybridization (aCGH): this is a genome-wide screening technique for CNVs [66] with higher resolution than chromosome-based comparative genomic hybridization [67]. It allows detection of copy number changes of 5–10 kb of DNA sequence.

Roche AmpliChip CYP450 system: this is an oligonucleotide microarray hybridization method for genotyping 27 *CYP2D6* variants (including CNVs) and two *CYP2C19* variants. It has been developed by Roche based on Affymetrix microarray technology [68]. A logarithmic scale is used to predict CYP2D6 and CYP2C19 phenotypes [69]. This is the first FDA-approved pharmacogenetics test for clinic use.

Genome-wide association SNP microarrays: these arrays cover 300 K to more than 1 million genetic markers including thousands of probes for the detection of CNVs [70].

Affymetrix DMET plus microarray: this is a novel genotyping tool customized for pharmacogenetics research. It covers 225 essential pharmacogenetics genes and 1936 common or rare variants including five CNVs belonging to *CYP2D6*, *CYP2A6*, *UGT2B17*, *GSTM1* and *GSTT1* [71].

Prevalence of CNVs in drug-related genes in human populations

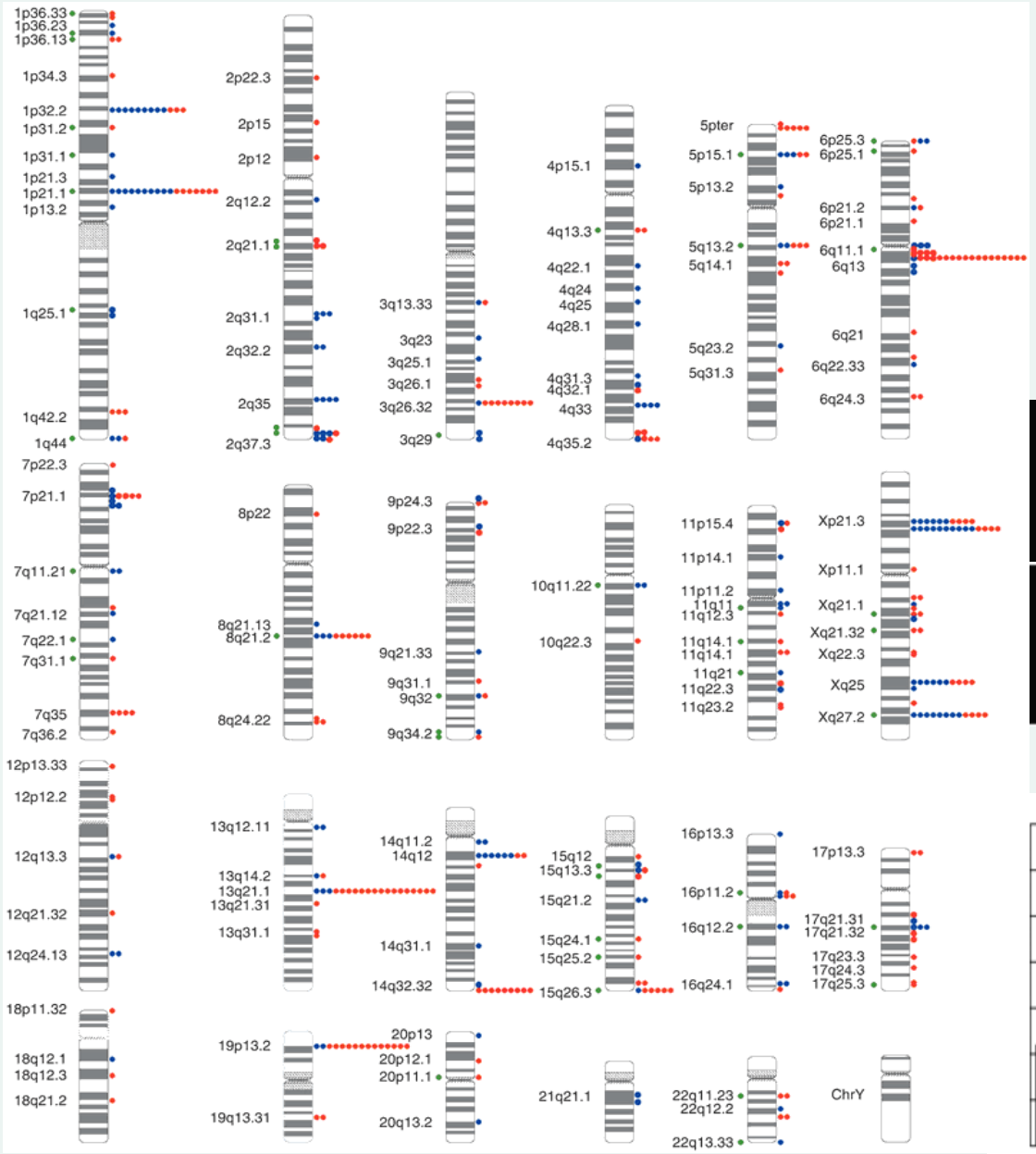
Understanding the prevalence of major genetic variations related to drug efficacy and toxicity is crucial for both health providers and patients. Both the Ministry of Health in each nation and physicians in clinics can use this knowledge to maximize benefit and minimize harm for patients prior to and during drug therapy [61], which is widely accepted as the base for personalized genomic medicine [62].

The frequency of CNVs is shown in Table 1. These data can help each population to select the most appropriate drugs and proper doses for specific treatments. For example, in the United States, over 50% and 30% of Caucasian populations carry either *GSTM1* or *GSTT1* gene deletions [50], which lead to increased susceptibility to colorectal cancer, acute myeloid leukemia and chemotherapy-induced toxicities [53]. Using the CNVs frequency data, physiologic, toxicological, pharmacologic, or clinical significance of the test results". Drugs for which therapeutic response or toxicity is affected by these markers were recorded in this list including one CNV marker: *CYP2D6*2×2*. The package insert for codeine sulfate and other drugs containing codeine as an ingredient, such as Fiorinal® with codeine (butalbital, aspirin, caffeine and codeine phosphate) and Fioricet® with codeine (butalbital, acetaminophen, caffeine and codeine phosphate), warns of elevated risk of codeine-induced toxicity in individuals with more copies of *CYP2D6* functional genes. However, none of the other drugs (tamoxifen, tacrine, etc.) or other CNVs (*CYP2D6*5*, *GSTM1*, *GSTT1* deletions) discussed in this review are included in this list yet. Meanwhile, an FDA-approved analytical test system for *GSTM1* and *GSTT1* deletions is not available at this time. The only approved genotyping platform for CNV markers diagnosis in the clinic is the AmpliChip CYP450 chip [3], which detects *CYP2D6* CNVs. There is still a gap between the

Table 1. A summary of CNV frequencies in pharmacogenetic genes in different ethnic populations

Population	CNV frequencies (%)				
	*CYP2D6 *1×N*	*CYP2D6 *2×N*	*CYP2D6*5*	*GSTM1* deletion	*GSTT1* deletion
North America					
USA (Caucasian)	0.2 [72]	0.7 [72]	6.2 [72]	54.3 [50]	27.6 [50]
USA (African American)	1.2 [72]	1.6 [72]	4.0 [72]	23.7 [73]	17.5 [73]
Mexico (Mestizo)			1.3–2.7 [74,75]	33.5 [76]	12.1 [76]
Africa					
Cameroon	3.3 [26]	3.3 [26]		27.8 [51]	46.8 [51]
Ethiopia		10–16 [21,77]	3.3 [77]	43.8 [51]	37.3 [51]
Ghana		1.6 [27]	6 [27]	19.3 [78]	73.7 [78]
East Asia					
China (Han)	2.2 [25]	0–2 [21]	7.2 [79]	52 [80]	38.7 [80]
Japan	0.5 [81]	0.5 [81]	6.2 [82]	50.8 [83]	45.8 [83]
Korea	0.13 [84]	0.5 [84]	6.1 [84]	51.4 [85]	51.6 [85]
Europe					
France	2.0 [25]		4.0 [25]	49 [86]	26 [86]
Germany	0.51 [87]	1.34 [87]	1.95 [87]	51.6 [50]	19.5 [50]
Middle East					
Saudi Arabia		10.4 [88]	1.0 [88]	8.3 [89]	4.2 [89]

CNPs (copy number variation)

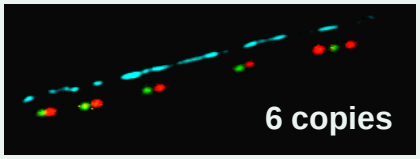


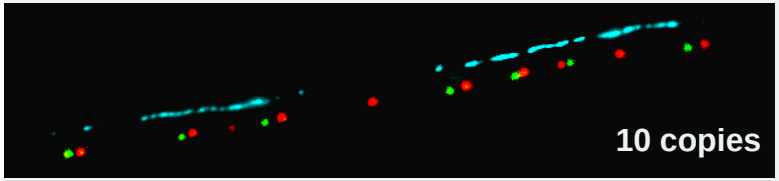
Iafrate et al Nat Genet 2004, Aldred et al Human Mol Genet 2005

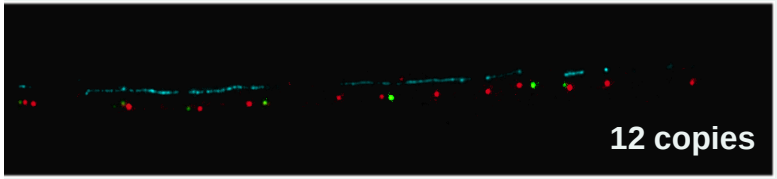
~12 CNPs

0,1-3Mb

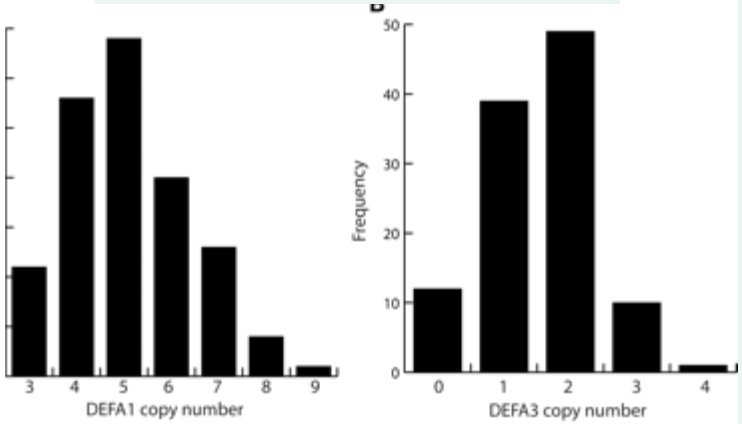
AMYLASE GENES







DEFENSIN GENES



nature Vol 444|23 November 2006|doi:10.1038/nature05329

ARTICLES

Global variation in copy number in the human genome

Richard Redon[1], Shumpei Ishikawa[2,3], Karen R. Fitch[4], Lars Feuk[5,6], George H. Perry[7], T. Daniel Andrews[1], Heike Fiegler[1], Michael H. Shapero[4], Andrew R. Carson[5,6], Wenwei Chen[4], Eun Kyung Cho[7], Stephanie Dallaire[7], Jennifer L. Freeman[7], Juan R. González[8], Mònica Gratacòs[8], Jing Huang[4], Dimitrios Kalaitzopoulos[1], Daisuke Komura[3], Jeffrey R. MacDonald[5], Christian R. Marshall[5,6], Rui Mei[4], Lyndal Montgomery[1], Kunihiro Nishimura[2], Kohji Okamura[5,6], Fan Shen[4], Martin J. Somerville[9], Joelle Tchinda[7], Armand Valsesia[1], Cara Woodwark[1], Fengtang Yang[1], Junjun Zhang[5], Tatiana Zerjal[1], Jane Zhang[4], Lluis Armengol[8], Donald F. Conrad[10], Xavier Estivill[8,11], Chris Tyler-Smith[1], Nigel P. Carter[1], Hiroyuki Aburatani[2,12], Charles Lee[7,13], Keith W. Jones[4], Stephen W. Scherer[5,6] & Matthew E. Hurles[1]

Copy number variation (CNV) of DNA sequences is functionally significant but has yet to be fully ascertained. We have constructed a first-generation CNV map of the human genome through the study of 270 individuals from four populations with ancestry in Europe, Africa or Asia (the HapMap collection). DNA from these individuals was screened for CNV using two complementary technologies: single-nucleotide polymorphism (SNP) genotyping arrays, and clone-based comparative genomic hybridization. A total of 1,447 copy number variable regions (CNVRs), which can encompass overlapping or adjacent gains or losses, covering 360 megabases (12% of the genome) were identified in these populations. These CNVRs contained hundreds of genes, disease loci, functional elements and segmental duplications. Notably, the CNVRs encompassed more nucleotide content per genome than SNPs, underscoring the importance of CNV in genetic diversity and evolution. The data obtained delineate linkage disequilibrium patterns for many CNVs, and reveal marked variation in copy number among populations. We also demonstrate the utility of this resource for genetic disease studies.

LETTERS

Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans

Timothy J. Aitman[1], Rong Dong[1]*, Timothy J. Vyse[2]*, Penny J. Norsworthy[1]*, Michelle D. Johnson[1], Jennifer Smith[3], Jonathan Mangion[1], Cheri Roberton-Lowe[1,2], Amy J. Marshall[1], Enrico Petretto[1], Matthew D. Hodges[1], Gurjeet Bhangal[3], Sheetal G. Patel[1], Kelly Sheehan-Rooney[1], Mark Duda[1,3], Paul R. Cook[1,3], David J. Evans[3], Jan Domin[3], Jonathan Flint[4], Joseph J. Boyle[5], Charles D. Pusey[3] & H. Terence Cook[5]

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

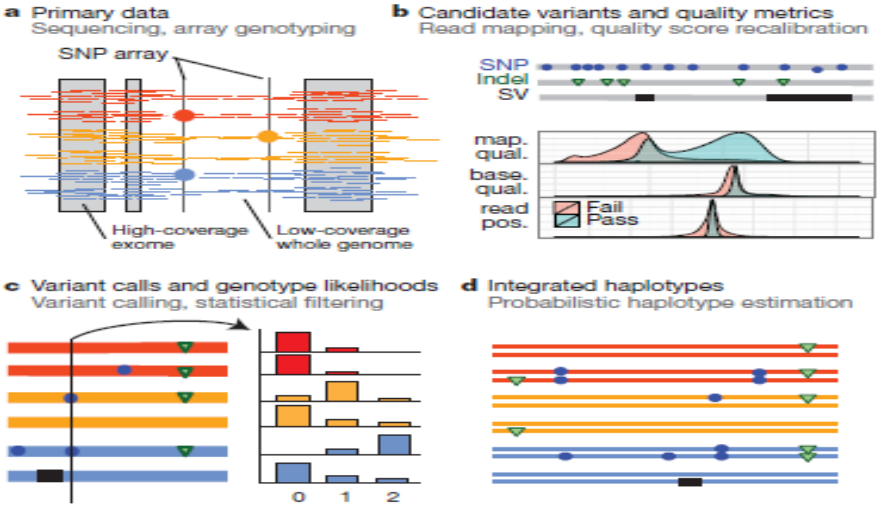
Recent efforts to map human genetic variation by sequencing exomes[1] and whole genomes[2–4] have characterized the vast majority of common single nucleotide polymorphisms (SNPs) and many structural variants across the genome. However, although more than 95% of common (>5% frequency) variants were discovered in the pilot phase of the 1000 Genomes Project, lower-frequency variants, particularly those outside the coding exome, remain poorly characterized. Low-frequency variants are enriched for potentially functional mutations, for example, protein-changing variants, under weak purifying selection[1,5,6]. Furthermore, because low-frequency variants tend to be recent in origin, they exhibit increased levels of population differentiation[6–8]. Characterizing such variants, for both point mutations and structural changes, across a range of populations is thus likely to identify many variants of functional importance and is crucial for interpreting individual genome sequences, to help separate shared variants from those private to families, for example.

We now report on the genomes of 1,092 individuals sampled from 14 populations drawn from Europe, East Asia, sub-Saharan Africa and the Americas (Supplementary Figs 1 and 2), analysed through a combination of low-coverage (2–6×) whole-genome sequence data, targeted deep (50–100×) exome sequence data and dense SNP genotype data (Table 1 and Supplementary Tables 1–3). This design was shown by the pilot phase[2] to be powerful and cost-effective in discovering and genotyping all but the rarest SNP and short insertion and deletion (indel) variants. Here, the approach was augmented with statistical methods for selecting higher quality variant calls from candidates obtained using multiple algorithms, and to integrate SNP, indel and larger structural variants within a single framework (see

BOX 1

Constructing an integrated map of variation

The 1,092 haplotype-resolved genomes released as phase I by the 1000 Genomes Project are the result of integrating diverse data from multiple technologies generated by several centres between 2008 and 2010. The Box 1 Figure describes the process leading from primary data production to integrated haplotypes.



a, Unrelated individuals (see Supplementary Table 10 for exceptions) were sampled in groups of up to 100 from related populations (Wright's F_{ST} typically $< 1\%$) within broader geographical or ancestry-based groups[2]. Primary data generated for each sample consist of low-coverage (average 5×) whole-genome and high-coverage (average 80× across a consensus target of 24 Mb spanning more than 15,000 genes) exome sequence data, and high density SNP array information. **b**, Following read-alignment, multiple algorithms were used to identify candidate variants. For each variant, quality metrics were obtained, including information about the uniqueness of the surrounding sequence (for example, mapping quality (map. qual)), the quality of evidence supporting the variant (for example, base quality (base. qual.) and the position of variant bases within reads (read pos.)), and the distribution of variant calls in the population (for example, inbreeding coefficient). Machine-learning approaches using this multidimensional information were trained on sets of high-quality known variants (for example, the high-density SNP array data), allowing variant sites to be ranked in confidence and subsequently thresholded to ensure low FDR. **c**, Genotype likelihoods were used to summarize the evidence for each genotype at bi-allelic sites (0, 1 or 2 copies of the variant) in each sample at every site. **d**, As the evidence for a single genotype is typically weak in the low-coverage data, and can be highly variable in the exome data, statistical methods were used to leverage information from patterns of linkage disequilibrium, allowing haplotypes (and genotypes) to be inferred.

Table 1 | Summary of 1000 Genomes Project phase I data

	Autosomes	Chromosome X	GENCODE regions*
Samples	1,092	1,092	1,092
Total raw bases (Gb)	19,049	804	327
Mean mapped depth (×)	5.1	3.9	80.3
SNPs			
No. sites overall	36.7 M	1.3 M	498 K
Novelty rate†	58%	77%	50%
No. synonymous/non-synonymous/nonsense	NA	4.7/6.5/0.097 K	199/293/6.3 K
Average no. SNPs per sample	3.60 M	105 K	24.0 K
Indels			
No. sites overall	1.38 M	59 K	1,867
Novelty rate†	62%	73%	54%
No. inframe/frameshift	NA	19/14	719/1,066
Average no. indels per sample	344 K	13 K	440
Genotyped large deletions			
No. sites overall	13.8 K	432	847
Novelty rate†	54%	54%	50%
Average no. variants per sample	717	26	39

NA, not applicable.
* Autosomal genes only.
† Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

*Lists of participants and their affiliations appear at the end of the paper.

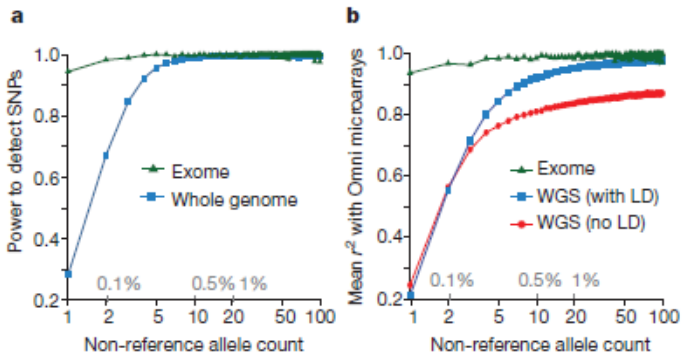


Figure 1 | Power and accuracy. **a**, Power to detect SNPs as a function of variant count (and proportion) across the entire set of samples, estimated by comparison to independent SNP array data in the exome (green) and whole genome (blue). **b**, Genotype accuracy compared with the same SNP array data as a function of variant frequency, summarized by the r^2 between true and inferred genotype (coded as 0, 1 and 2) within the exome (green), whole genome after haplotype integration (blue), and whole genome without haplotype integration (red). LD, linkage disequilibrium; WGS, whole-genome sequencing.

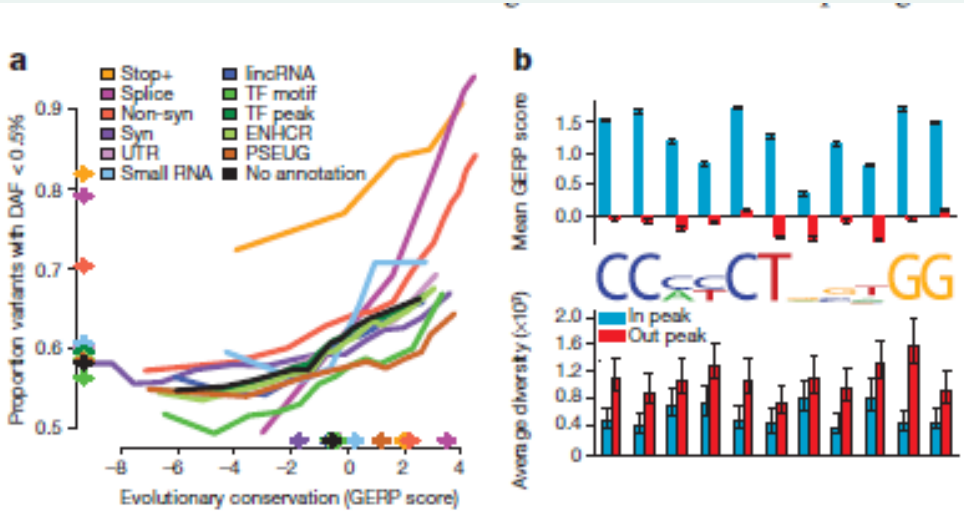


Figure 4 | Purifying selection within and between populations. **a**, The relationship between evolutionary conservation (measured by GERP score[19]) and rare variant proportion (fraction of all variants with derived allele frequency (DAF) $< 0.5\%$) for variants occurring in different functional elements and with different coding consequences. Crosses indicate the average GERP score at variant sites (*x* axis) and the proportion of rare variants (*y* axis) in each category. ENHCR, enhancer; lincRNA, large intergenic non-coding RNA; non-syn, non-synonymous; PSEUG, pseudogene; syn, synonymous; TF, transcription factor. **b**, Levels of evolutionary conservation (mean GERP score, top) and genetic diversity (per-nucleotide pairwise differences, bottom) for sequences matching the CTCF-binding motif within CTCF-binding peaks, as identified experimentally by ChIP-seq in the ENCODE project[13] (blue) and in a matched set of motifs outside peaks (red). The logo plot shows the distribution of identified motifs within peaks. Error bars represent ± 2 s.e.m.

- In conserved coding sites - 85% of non-synonymous variants and >90% of stop-gain and splice-disrupting variants are <0.5% in frequency compared with 65% of synonymous variants.

- Rare variant excess tracks the level of evolutionary conservation for variants of most functional consequence.

- However stop-gain and splice-site disrupting variants show increased rare-variant excess despite conservation due to high deleterious effect.

Estimation of segregating load arising from rare, deleterious mutations across a set of genes

- Compare the ratios of non-synonymous to synonymous variants in different frequency ranges.

- The non-synonymous to synonymous ratio among rare (<0.5%) variants is typically in the range 1-2 and among common variants in the range of 0.5-1.5
- → **suggesting that 25-30% of rare non-synonymous variants are deleterious.**

- Individuals typically carry >2500 non-synonymous variants, 20-40 damaging variants and ~150 loss-of-function variants **at conserved positions**. However most are common (>5%) or low frequency (0.5-5%).
- → They estimate that individuals carry an excess of 76-190 rare deleterious non-synonymous variants and up to 20 loss-of-function and disease-associated variants.

Table 2 | Per-individual variant load at conserved sites

Variant type	Number of derived variant sites per individual			Excess rare deleterious	Excess low-frequency deleterious
	Derived allele frequency across sample				
	<0.5%	0.5–5%	>5%		
All sites	30–150 K	120–680 K	3.6–3.9 M	ND	ND
Synonymous*	29–120	82–420	1.3–1.4 K	ND	ND
Non-synonymous*	130–400	240–910	2.3–2.7 K	76–190†	77-130†
Stop-gain*	3.9–10	5.3–19	24–28	3.4–7.5†	3.8–11†
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8	0.81–1.1†	0.80–1.0†
HGMD-DM*	2.5–5.1	4.8–17	11–18	1.6–4.7†	3.8–12†
COSMIC*	1.3–2.0	1.8–5.1	5.2–10	0.93–1.6†	1.3–2.0†
Indel frameshift	1.0–1.3	11–24	60–66	ND§	3.2–11†
Indel non-frameshift	2.1–2.3	9.5–24	67–71	ND§	0–0.73†
Splice site donor	1.7–3.6	2.4–7.2	2.6–5.2	1.6–3.3†	3.1–6.2†
Splice site acceptor	1.5–2.9	1.5–4.0	2.1–4.6	1.4–2.6†	1.2–3.3†
UTR*	120–430	300–1,400	3.5–4.0 K	0–350‡	0–1.2 K‡
Non-coding RNA*	3.9–17	14–70	180–200	0.62–2.6‡	3.4–13‡
Motif gain in TF peak*	4.7–14	23–59	170–180	0–2.6‡	3.8–15‡
Motif loss in TF peak*	18–69	71–300	580–650	7.7–22‡	37–110‡
Other conserved*	2.0–9.9 K	7.1–39 K	120–130 K	ND	ND
Total conserved	2.3–11 K	7.7–42 K	130–150 K	150–510	250–1.3 K

Only sites in which ancestral state can be assigned with high confidence are reported. The ranges reported are across populations. COSMIC, Catalogue of Somatic Mutations in Cancer; HGMD-DM, Human Gene Mutation Database (HGMD) disease-causing mutations; TF, transcription factor; ND, not determined.
* Sites with GERP >2
† Using synonymous sites as a baseline.
‡ Using 'other conserved' as a baseline.
§ Rare indels were filtered in phase I.

INVITED COMMENT

AMERICAN JOURNAL OF medical genetics PART A

The Centers for Mendelian Genomics: A New Large-Scale Initiative to Identify the Genes Underlying Rare Mendelian Conditions

Michael J. Bamshad,[1,2,3]* Jay A. Shendure,[2] David Valle,[4] Ada Hamosh,[4] James R. Lupski,[5,6,7,8] Richard A. Gibbs,[5,8] Eric Boerwinkle,[8,9] Richard P. Lifton,[10] Mark Gerstein,[11] Murat Gunel,[10,12] Shrikant Mane,[10] and Deborah A. Nickerson[2]

on behalf of the Centers for Mendelian Genomics

[1]Department of Pediatrics, University of Washington, Seattle, Washington
[2]Department of Genome Sciences, University of Washington, Seattle, Washington
[3]Seattle Children's Hospital, Seattle, Washington
[4]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland
[5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas
[6]Department of Pediatrics, Baylor College of Medicine, Houston, Texas
[7]Texas Children's Hospital, Houston, Texas
[8]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas
[9]Human Genetics Center, University of Texas Health Sciences Center at Houston, Houston, Texas
[10]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut
[11]Department of Biophysics and Biochemistry, Yale University School of Medicine, New Haven, Connecticut
[12]Department of Neurosurgery, Yale University School of Medicine, New Haven, Connecticut

Manuscript Received: 2 April 2012; Manuscript Accepted: 19 April 2012

Next generation exome sequencing (ES) and whole genome sequencing (WGS) are new powerful tools for discovering the gene(s) that underlie Mendelian disorders. To accelerate these discoveries, the National Institutes of Health has established three *Centers for Mendelian Genomics* (CMGs): the Center for Mendelian Genomics at the University of Washington; the Center for Mendelian Genomics at Yale University; and the Baylor–Johns Hopkins Center for Mendelian Genomics at Baylor College of Medicine and Johns Hopkins University. The CMGs will provide ES/WGS and extensive analysis expertise at no cost to collaborating investigators where the causal gene(s) for a Mendelian phenotype has yet to be uncovered. Over the next few years and in collaboration with the global human genetics community, the CMGs hope to facilitate the identification of the genes underlying a very large fraction of all Mendelian disorders; see http://mendelian.org.

Key words: Mendelian; exome sequencing; commentary

Uses of 1,000 Genomes project

-Use of data to screen variants discovered in exome data from individuals with genetic disorders and in cancer genome projects.

- Enhanced catalogue improves the power of such screening.

- Also it provides a null expectation for the number of rare, low frequency and common variants with different functional consequences typically found in randomly sampled individuals from different populations.

Nat Rev Genet. 2011 Sep 27;12(11):745-55. doi: 10.1038/nrg3031.

Exome sequencing as a tool for Mendelian disease gene discovery.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J.

Department of Pediatrics, University of Washington, Health Sciences Building RR349, 1959 NE Pacific Street, Seattle, Washington 98195-6320, USA. mbamshad@u.washington.edu

Abstract

Exome sequencing - the targeted sequencing of the subset of the human genome that is protein coding - is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Over the past 2 years, experimental and analytical approaches relating to exome sequencing have established a rich framework for discovering the genes underlying unsolved Mendelian disorders. Additionally, exome sequencing is being adapted to explore the extent to which rare alleles explain the heritability of complex diseases and health-related traits. These advances also set the stage for applying exome and whole-genome sequencing to facilitate clinical diagnosis and personalized disease-risk profiling.

The Genome of the Netherlands (GoNL- sequence 1000 genomes from the Dutch population)

Figure 1 The 12 provinces of the Netherlands, the 12th province (Flevoland) is a recent province (land reclaimed from water) and was not included as a separate sampling unit (Image by Wikimedia Commons user Alphathon).

- **Identifying low frequency and rare variants by re-sequencing will contribute to resolving the missing heritability**

- **rarer variants are more likely to be population specific → mandatory for translation into public health and clinical benefits**

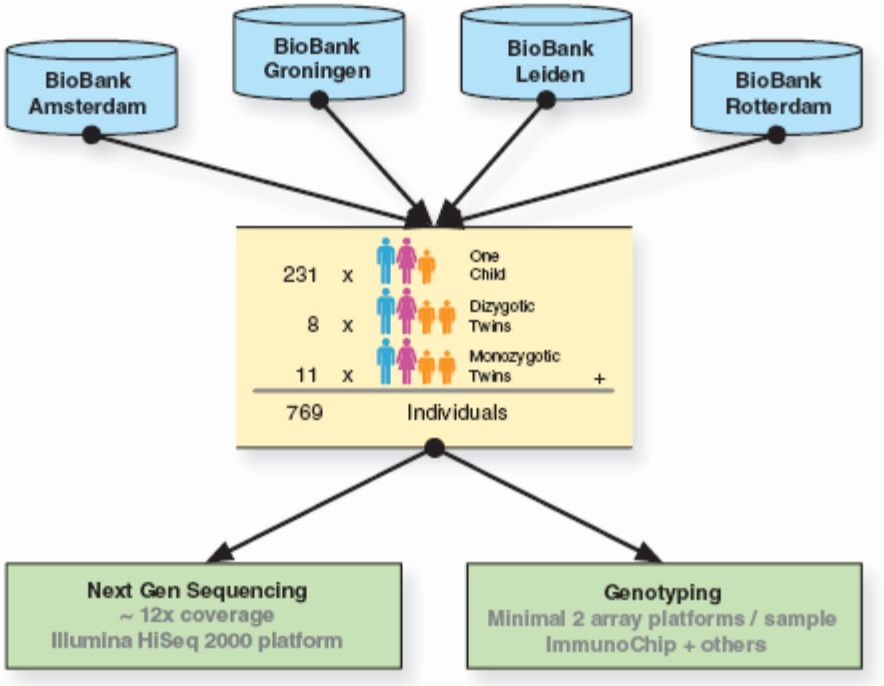


Figure 2 Sampling schedule for the GoNL: four population-based biobanks contributed samples for sequencing at the BGI (Beijing Genomics Institute).

European Journal of Human Genetics

Deciphering Disease-causing genes : Just the Start……

"This is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning."

Sir Winston Churchill @ Lord Mayor's Luncheon,
Mansion House following the victory at El Alameinin North Africa
London, 10 November 1942.

1,000 $ per genome – A new era of Genomics

Το HeSeq X Ten διαβάζει πέντε γονιδιώματα την ημέρα, υπερηφανεύεται η Illumina