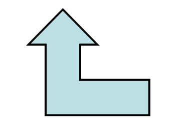
From toxicogenetics to toxicogenomics: current diagnostic tools and new technologies

Chakradhara Rao S Uppugunduri: CANSEARCH Research Laboratory, University of Geneva

OUTLINE

1) *Basic concepts*
2) *Candidate gene approach*
3) *Diagnostic tools*
4) *Common variants: HapMap and GWAS*
5) *Rare variants: 1000 Genomes Project*
6) *Evolution of sequencing technologies*

Variability in Drug Response



Prescribed dose

Compliance
Dosage mistakes

Ingested dose

Absorption
Distribution
Metabolism
Excretion

Pharmacokinetic

Physiological factors
Pathological factors
Genetic factors
Drug-drug interactions
Environmental factors (dietary)

Concentration at target site

Drug-target interaction

Pharmacodynamic

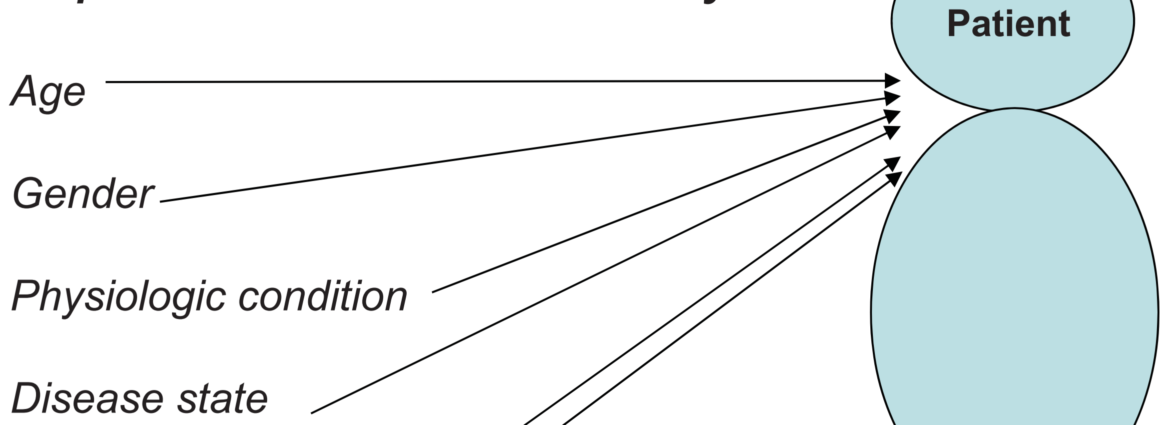
Response intensity

VARIABLE

Drug response

Drug response is influenced by

- *Age*
- *Gender*
- *Physiologic condition*
- *Disease state*
- *Environment*
- *Other drugs*
- **Genes**



Genetic variation (vs) drug response ?

Adverse drug reactions

- **ADVERSE DRUG REACTIONS**
  - *Causes hospitalization*
  - *Experienced by hospitalized patients*
  - *Economic burden*
  - *6th Leading cause of death*
- *59% of drugs causing ADRs are metabolized by polymorphic enzymes*
- *7-22% of other randomly selected drugs are substrates for polymorphic enzymes*
- *Polymorphisms occur in transporters, receptors, and other therapeutic targets are also associated with interindividual variability in drug response*

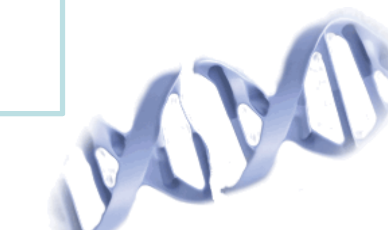
Huang SM, Goodsaid F, Rahman A, et el. Toxicology Mechanisms and Methods. 2006;(16) 89-99

Human genome project

HUMAN GENOME PROJECT

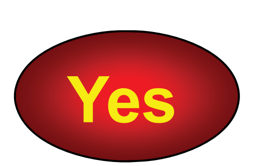
- ❖ Human genome was sequenced as a whole
- ❖ Out of 3.2 billion bases only 5% encode genes
- ❖ 32,000 genes (22,000 on list right now)
- ❖ 15,000 known genes and 17,000 predicted genes
- ❖ 6000 gaps to fill in
- ❖ 971 known disease genes (e.g., Huntington's Gene, Muscular dystrophy)
- ❖ >500 known genes encoding drug targets
- ❖ 1 in 1000 bp differences between us

Genetic variation

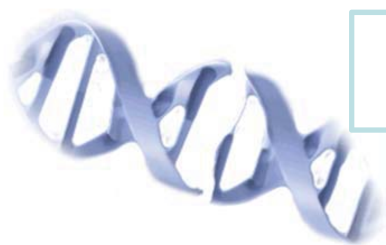


DNA sequence of all human beings is 99.9% identical

- Our DNAs differ by 0.1%.
- Does it make a difference ?



0.1% difference translates into 3 million separate "spelling" differences in a genome of 3 billion bases



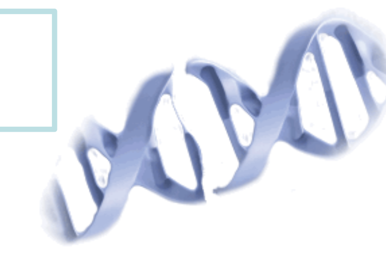
Genetic variation

'Any two unrelated people are 99.9 percent identical at the level of DNA sequence. The remaining 0.1 percent difference can help explain why one person has distinct physical features, is more susceptible to a disease or responds differently to a drug or an environmental factor than another person'

Exploiting this information for optimizing treatment is nothing but

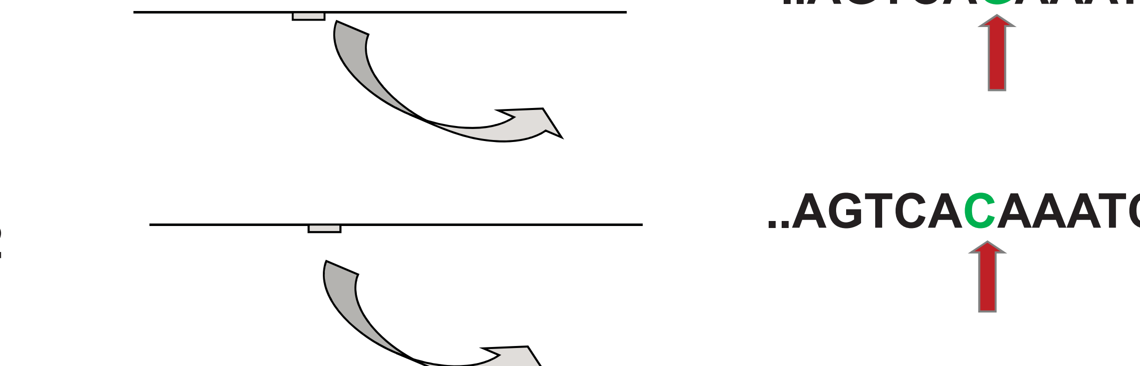
-Pharmacogenetics/P'cogenomics- in relation to toxicology- toxicogenetics/genomics

Genetic variation



SNP

- *Altered single nucleotide in the genome sequence.*
- *Found in at least 1% of the population.*
- *Occurs on an average every 1000 to 3000 bp;*
- *Bi-allelic: wild type and mutation (<1%) more neutral term variation.*

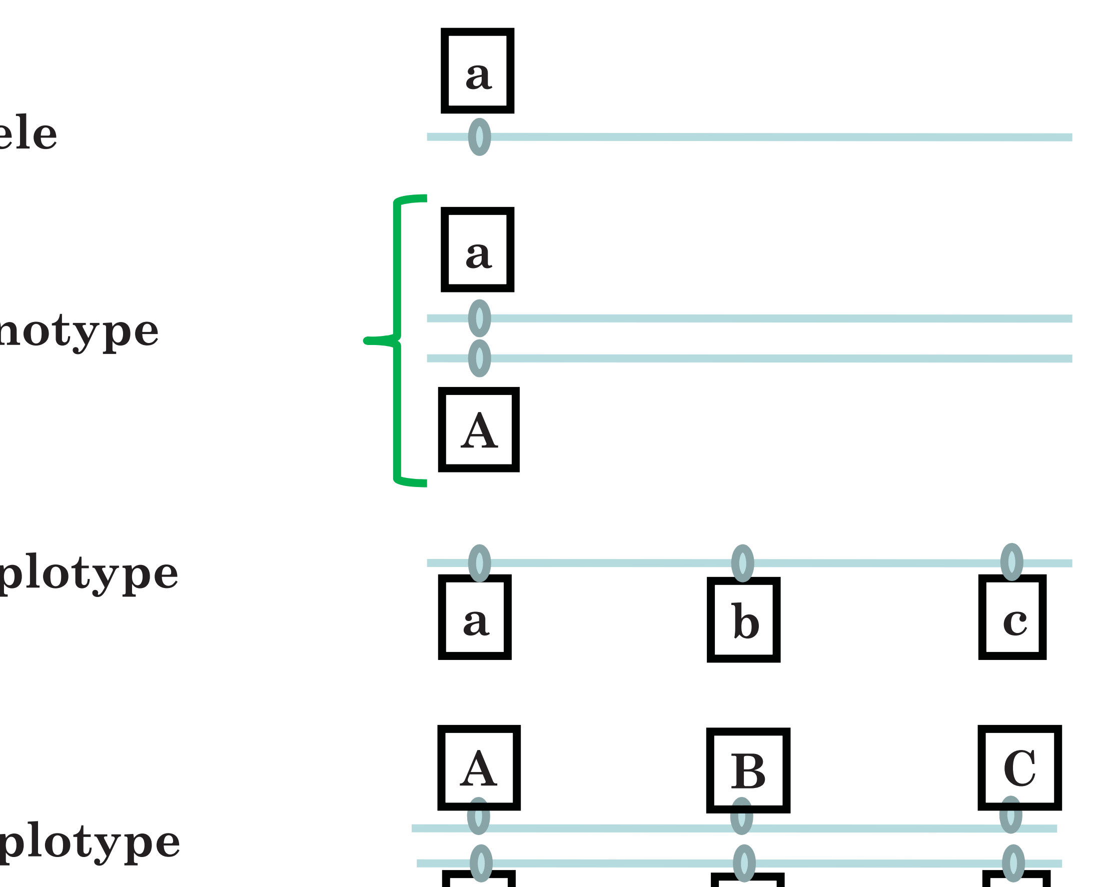


Genetic variation

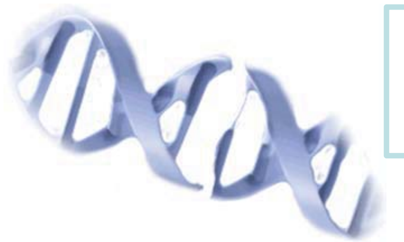
SNP

- *15 million sites in the human genome where SNPs could occur (10 million are common)*
- *3.8 million SNPs catalogued in 270 individuals (1 SNP / 1000 bases) (International HapMap Consortium, October 2007, Phase 3 in 2010);*
- *38 million SNPs, 1.4 million short insertions and deletions, and more than 14,000 larger deletions (1000 genome project, 2012)*
- *Each person has 3-5 million common SNPs (inherited) and 30 new SNPs that have arisen {mutation rate of 1.1 x 10^{-8} / base / generation}*
- *Types:*
    - *Coding SNPs (cSNPs)*
    - *(5'UTR or 3'UTR) regulatory SNPs (rSNPs)*
    - *Splice site SNPs (ssSNPs)*
    - *Intronicsnps (iSNPs)*
    - *Anonymous SNPs (aSNPs; SNPs that occur in "junk"DNA)*

Genotype/Haplotype



We will know the details later under GWAS



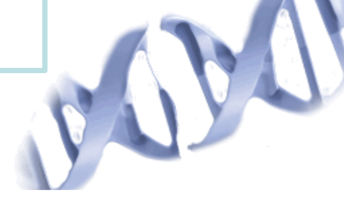
Tagged SNP

GSTA1 Haplotype	*-69*	*-513*	*-631*	*-1142*
**A1*	*C*	*T*	*T*	*C*
**A2*	*C*	*T*	*G*	*C*
**A3*	*C*	*T*	*T*	*G*
**B1*	*T*	*T*	*G*	*G*
**B1a*	*T*	*C*	*G*	*G*
**B2*	*T*	*T*	*G*	*C*

↓

Tag SNP ⟶ Aids in identification of the haplotype blocks

We will know the details later under GWAS



Genetic causes of inter-individual variability

CCTGGTTGACTGATCGCGGGATTTATATATATGG

CCTCGTTGAC..ATCGCGGGATTTATATATATATATATATGG

SNPs

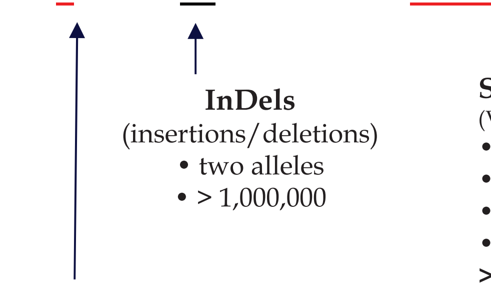
InDels
(insertions/deletions)
- two alleles
- > 1,000,000

SSR – short sequence repeats
(VNTR - variable number tandem repeats)
- **many alleles**
- **microsatellites (1-5)**
- **minisatellites (6-100)**
- **...**

> 1,000,000

- **Inversions**
- **Duplications**
- **Translocations**
- **Transposon insertions**

Variations exceeding 1000bp - STRUCTURAL VARIATIONS
- **less than 3 million bp - submicroscopic; larger– microscopic**
- **InDels and duplications are called CNVs (copy number variations)**



Single nucleotide
- • Base change – substitution – point mutation
- → Insertion-deletions ("indels")
- • SNPs – tagSNPs

2 bp to 1,000 bp
- • Microsatellites, minisatellites
- → Indels
- • Inversions
- • Di-, tri-, tetranucleotide repeats
- • VNTRs

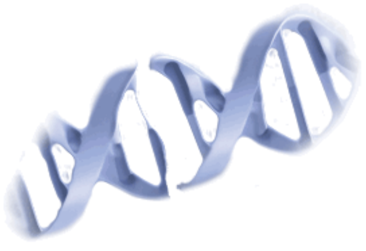
1 kb to submicroscopic
- → Copy number variants (CNVs)
- → Segmental duplications
- • Inversions, translocations
- → CNV regions (CNVRs)
- • Microdeletions, microduplications

Microscopic to subchromosomal
- → Segmental aneusomy
- • Chromosomal deletions – losses
- • Chromosomal insertions – gains
- • Chromosomal inversions
- • Intrachromosomal translocations
- • Chromosomal abnormality
- → Heteromorphisms
- • Fragile sites

Whole chromosomal to whole genome
- • Interchromosomal translocations
- • Ring chromosomes, isochromosomes
- • Marker chromosomes
- → Aneuploidy
- → Aneusomy

Scherer et al., Nat Genet Suppl 39:7;2007

Pharmaco/Toxico genetics & Genomics



- *Pharmacogenetics (Single or few gene effects)*
  - *Study of single gene or few gene variants and response to a drug*
  - *Genes may encode drug metabolizing enzymes, drug targets, or transporter proteins etc.*
- *Pharmacogenomics*
  - *Whole genomic approach to study the response to a drug or occurrence of any disease*
  - *Involves the study of multiple genes /whole genome*

  - *"The molecular study of genetic factors that determine drug efficacy and toxicity"- in relation to toxicology-toxicogenetics and genomics.*

Toxicogenomics/Genetics

- *Toxicogenetics and toxicogenomics aim to study the effects of genetic variations in drug response.*
- *Toxicogenomics is often described as the whole-genome application of toxicogenetics (overlapping disciplines).*
- *Not only limited to the study of genetic variations, but also studying epigenetic pattern, gene expression pattern in relation to drug response and toxicity.*

Toxicogenetics versus toxicogenomics

Gene (DNA)

Transcript (mRNA)

Protein

Metabolite

Clinical Response

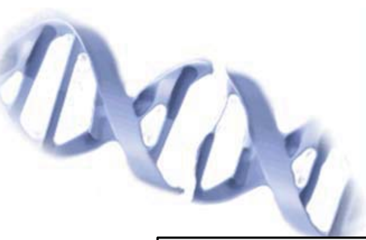
Genome

Transcriptome

Proteome

Metabolome

Clinical Response



PERSONALIZED MEDICINE
Role of toxicogenetics

"Personalized Medicine: the use of marker assisted diagnosis and targeted therapies designed from an individual molecular profile"
Ginsburg and McCarthy 2001

Personalized medicine: using information about a person's biological and genetic make up to tailor strategies for the prevention, detection and treatment of disease

Trial & Error Medicine

DRUG A
→ Adverse response/ No response
→ Adjust the dose
→ Side effects/ No response
→ Drug B

Personalized (Precision, Predictive) medicine

Genetic test (PGx)
→ Poor GSTA conjugator
→ Drug B with specific dose

Toxicogenetics and toxicogenomics goals

Clinical:

avoid adverse drug reactions
optimize drug efficacy
select responsive patients

Scientific:

link genetic variations to phenotypic variations to
understand the mechanisms responsible for that link
translate the knowledge in improving therapies

The candidate genes approach

- ✓ *Until mid 2000s, in toxicogenetics and toxicogenomics association of particular phenotypes to specific genotypes has been carried out through candidate genes studies.*

- ✓ *Valuables data has been obtained for adverse drug reactions associated to alleles of a single gene often involved in the metabolism of the drug.*

- ✓ *FDA recommends testing for: TPMT (bone-marrow suppression by azathioprine), UGT1A1*28 (neutropenia by irinotecan), HLA-B*5701 (hypersensitivity reactions by abacavir).*

The candidate genes approach (2)

- ✓ *This approach was less successful in understanding drug response overall probably because several, less "obvious", genes are involved.*

- ✓ *Exceptions: CYP2C9 and VKORC1 for coumarins drugs, CYP2C19 for clopidogrel and CYP2D6 for tamoxifen.*

- ✓ *Useful mostly in identifying patients with higher risk of adverse effects especially when they are prone for immunological reactions / organ toxicity.*

Genotyping in toxicogenetics

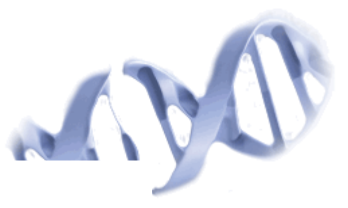
Genotyping: identification of individual genetic variants

Utility in toxicogenetics:

- *Select the best drug*
- *Choose the initial dose (narrow therapeutic index)*

- *Absence of therapeutic effects*
- *Adverse drug reactions*

Public database dbSNP



NCBI Linked Resources

- Entrez Databases: Gene, Nucleotide, Protein, Structure, Taxonomy
- Genome Assembly
- Homologene
- Mapviewer
- OMIM
- Probe
- UniGene
- UniSTS

Submission *ss9266*
CGAP-GAI

Submission *ss870165*
D NICKERSON

Submission *ss1542565*
C LEE

Submission *ss4429119*
C LEE

Submission *ss12568612*
CUORCGL

Submission *ss15840556*
SC_SNP

Submission *ss16231140*
CGAP-GAI

Reference SNP Record
rs7412

Linkout

External Databases
LEE
CGAP-GAI
Ensembl
PolyPhen
TaqMan
ALFRED

Summary Validation Information
experimental confirmation

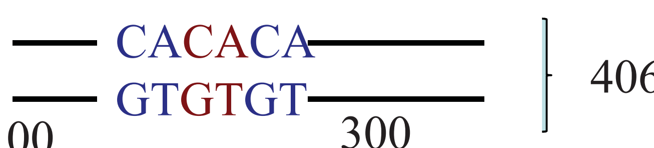
Summary Variation Information
Heterozygosity = 0.183

www.ncbi.nlm.nih.gov/SNP

Principle of genotyping methods

DS DNA 3' 5' 5' 3'
DENATURATION 95°C
ANNEALING P1 P2 ~50°C
EXTENSION 72°C
(A)

• Short tandem repeats → length differences



• SNPs → only sequence difference

G
C
A
T

- Destruction restriction site (RFLP)
- Alle-Specific PCR
- Hybridization differences (TaqMan)
- One base-pair sequencing reaction-primer extension (Sequenom, Orchid)
- Ligation assay (Illumina)

Fluorescent probes in real-time quantitative PCR

- *Taqman or 5'-nuclease assay*
- *FRET assays*

LightCycler (Roche)

Thermal chamber

Filters

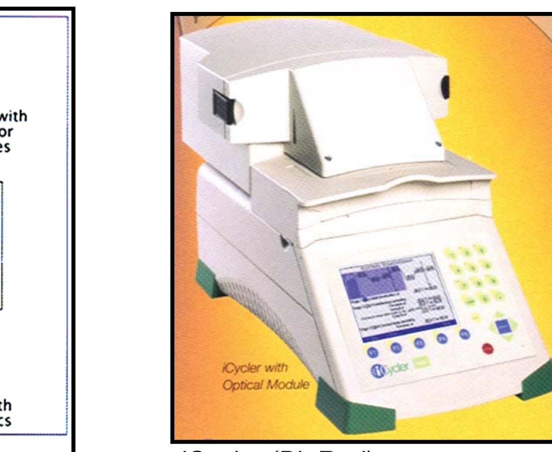
Maintenance-free LED light source

Carousel with capacity for 32 samples

Photohybrids

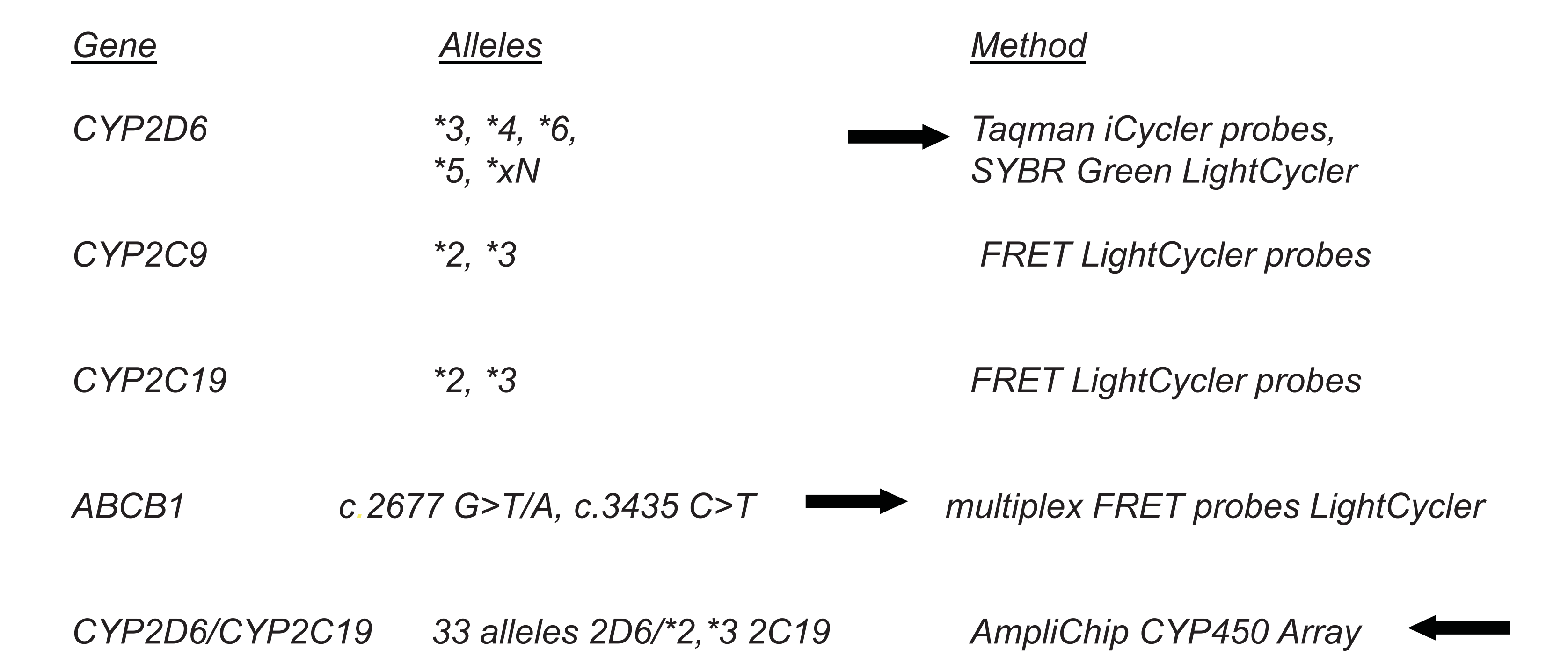
Microvolume fluorimeter with Rodenstock quality optics

Figure 1: LightCycler, thermal chamber and optical unit.



iCycler (BioRad)

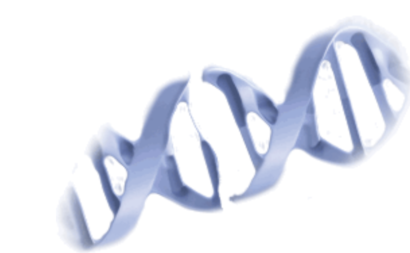
Examples of genotyping methods

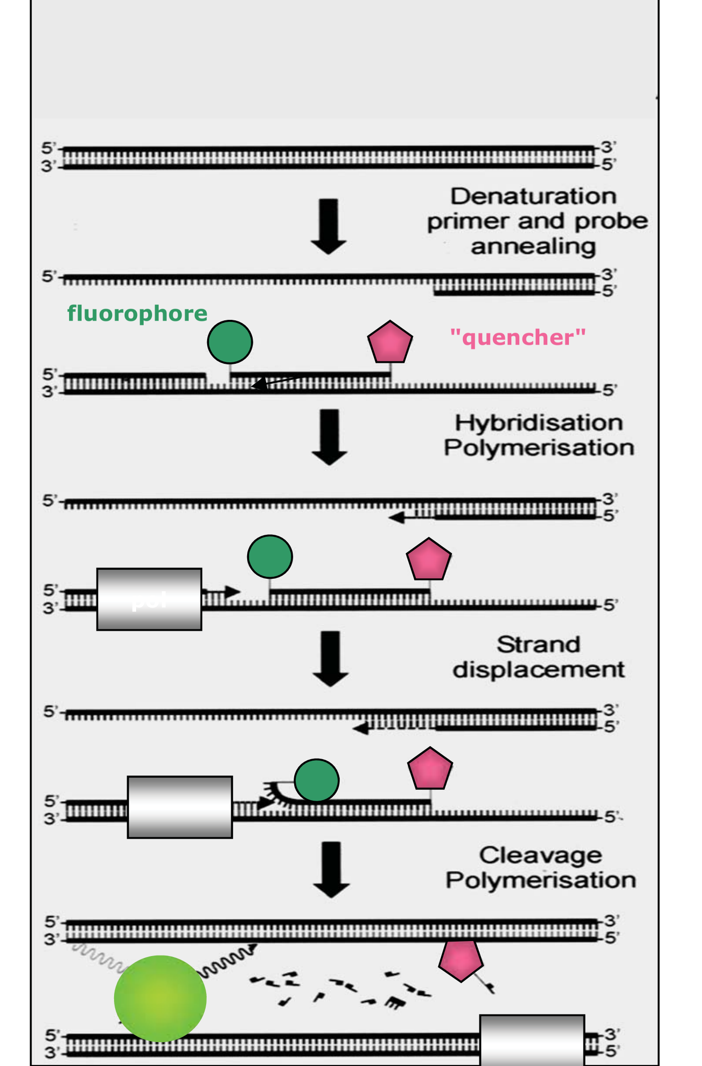


Gene	Alleles	Method
CYP2D6	*3, *4, *6, *5, *xN	Taqman iCycler probes, SYBR Green LightCycler
CYP2C9	*2, *3	FRET LightCycler probes
CYP2C19	*2, *3	FRET LightCycler probes
ABCB1	c.2677 G>T/A, c.3435 C>T	multiplex FRET probes LightCycler
CYP2D6/CYP2C19	33 alleles 2D6/*2,*3 2C19	AmpliChip CYP450 Array

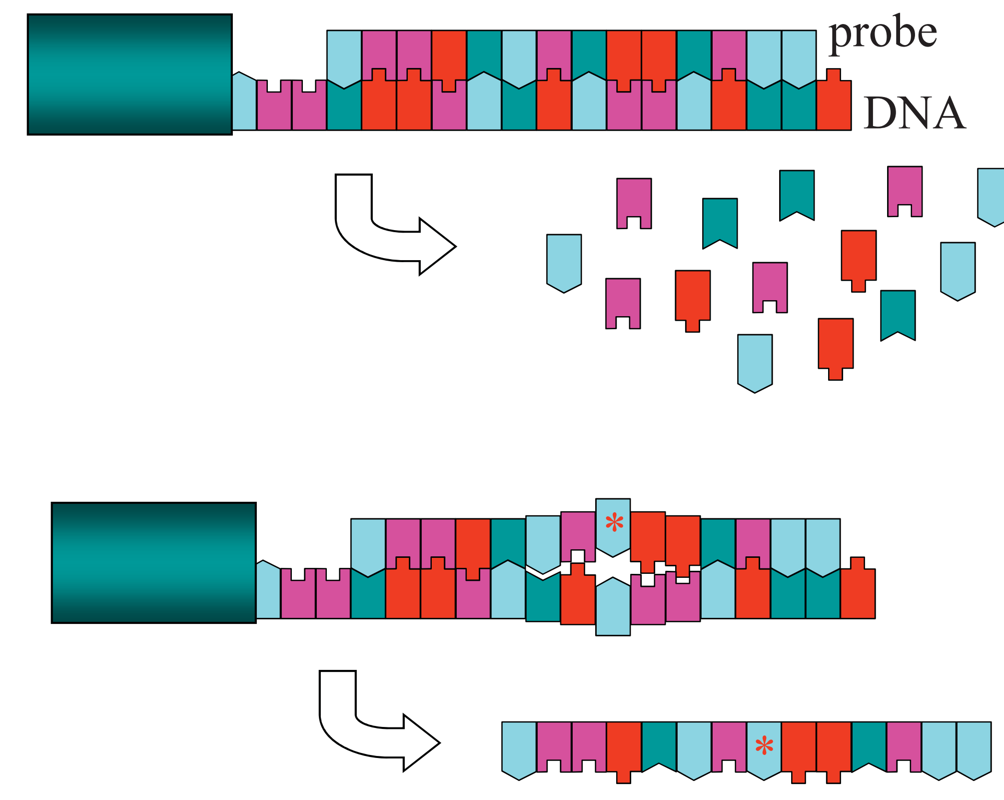
✓ Methodology used in diagnostics shall have been validated with few false positives and greater sensitivity. Cross validation also shall be performed periodically

Genotyping by Taqman assays (5'-exonuclease assay)





Exonuclease function is effective
only when complementarity is 100 %



Allele specific detection by Taqman® assays

Allèle 1

complém. 100%

complém. imparfaite

Allèle 2

complém. 100%

complém. imparfaite

V VIC

F FAM

TaqGold DNA Polymérase

Homozygous major(C)-allele

FAM

Heterozygote

Homozygous minor(G)-allele

TET

Homozygous C-allele

Heterozygote

Homozygous G-allele

Fluorescence

1.0

0,5

0

20 40

20 40

20 40

Thermal Cycles

Allelic Discrimination Plot

Allele 2

4.1
3.6
3.1
2.6
2.1
1.6
1.1
0.6
0.1

0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6

Allele 1

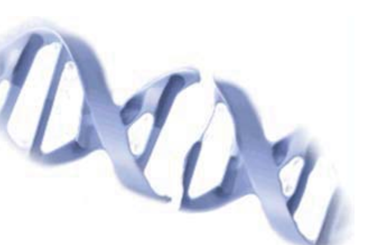
Legend

Homozygous Allele 1/Allele 1

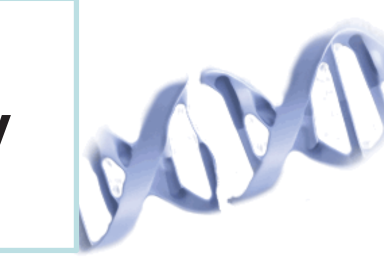
Homozygous Allele 2/Allele 2

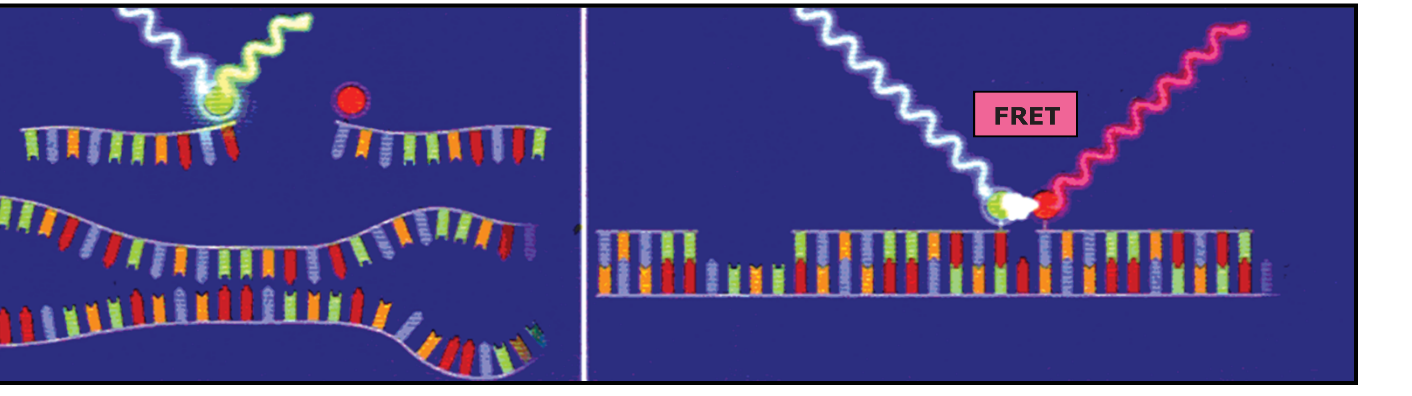
Heterozygous Allele 1/Allele 2

Undetermined

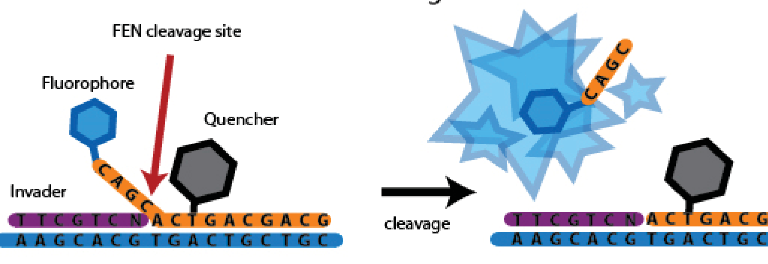


Genotyping by FRET (Fluorescence Resonance Energy Transfer)

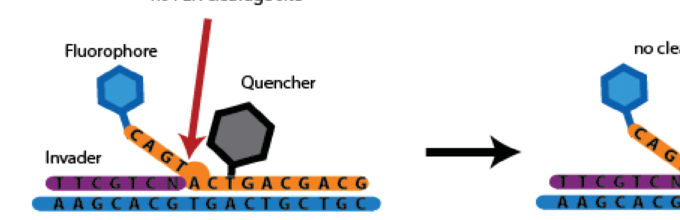




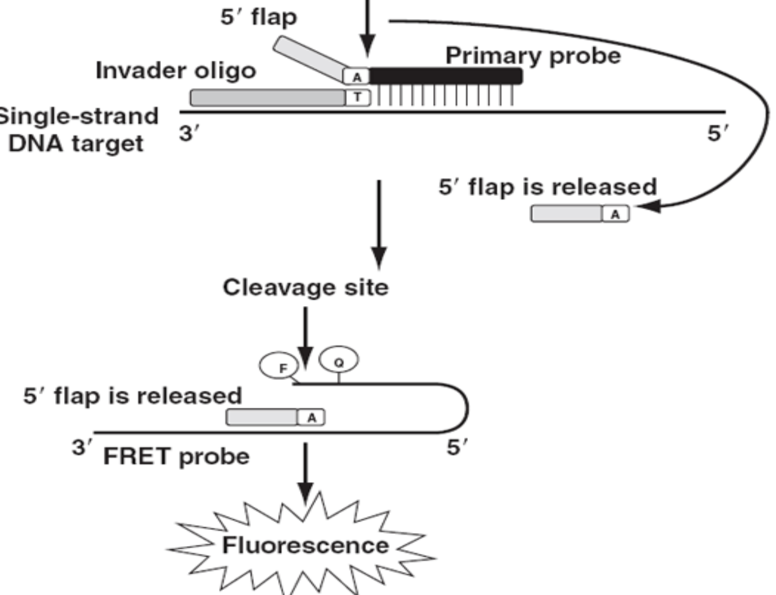
Invader assay in which probe complements the SNP resulting in fluroescene



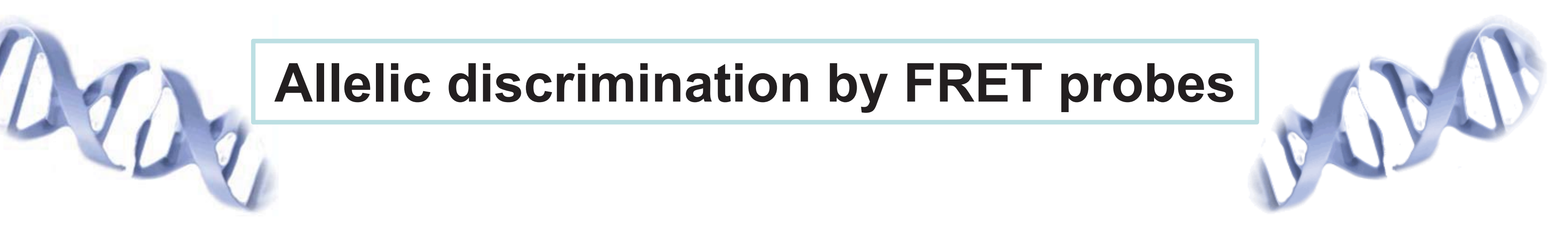
Invader assay in which probe mismatches at the SNP location preventing cleavage from occuring

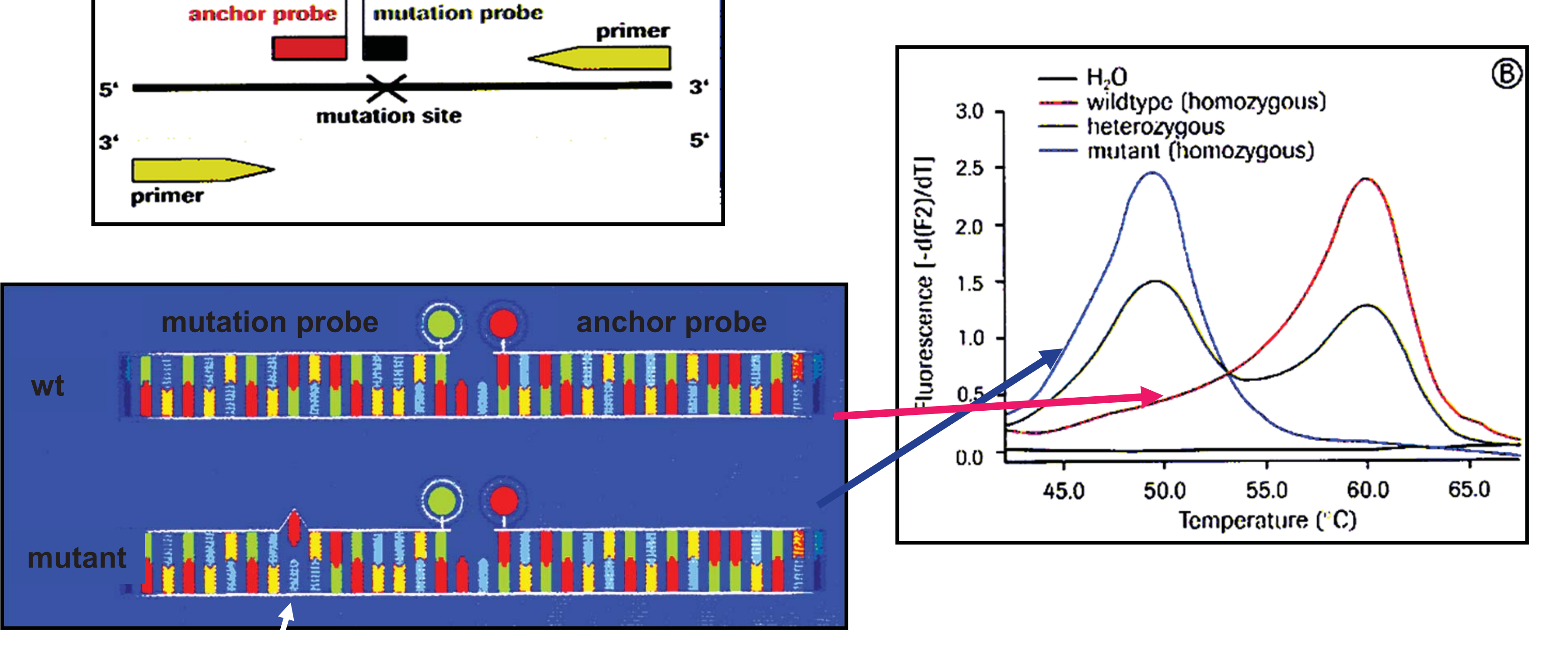


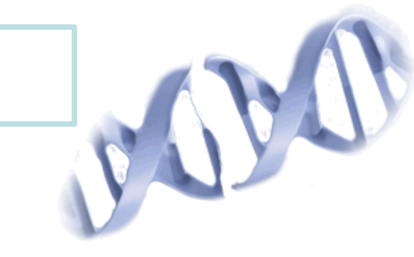
Reference: Based on Olivier M. 2005. The Invader assay for SNP genotyping.



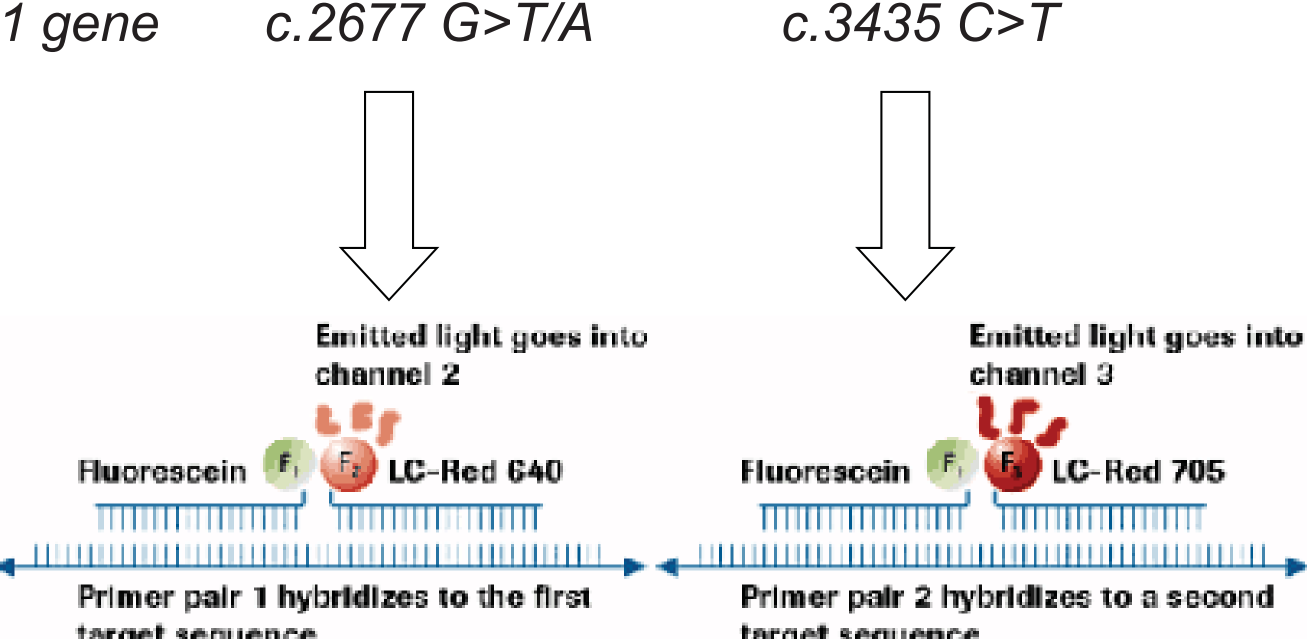
Allelic discrimination by FRET probes







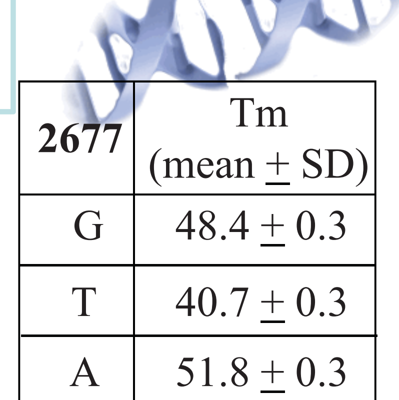
Multiplex PCR using FRET probes



Melting curves

Genotyping of c.3435 C>T et c.2677 G>T/A SNPs of ABCB1 gene

3435	Tm (mean ± SD)
C	65.8 ± 0.2
T	57.7 ± 0.3



2677	Tm (mean ± SD)
G	48.4 ± 0.3
T	40.7 ± 0.3
A	51.8 ± 0.3

Fluorescence –d(F3)/dT

0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05

H_2O

36.0 38.0 40.0 42.0 44.0 46.0 48.0 50.0 52.0 54.0 56.0 58.0 60.0 62.0 64.0 66.0 68.0 70.0 72.0 74.0

Temperature (°C)

Fluorescence –d(F2)/dT

0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05

H_2O

36.0 38.0 40.0 42.0 44.0 46.0 48.0 50.0 52.0 54.0 56.0 58.0 60.0 62.0

Temperature (°C)

Genotyping of Presence or absence of gene copies or copy number using melt curve analysis- StepOne Real Time PCR with SYBR green chemistry

GSTM1 : 81.4± 0.6

Bcl GSTM1 GSTT1

Bcl2 : 78.1± 0.4

Bcl

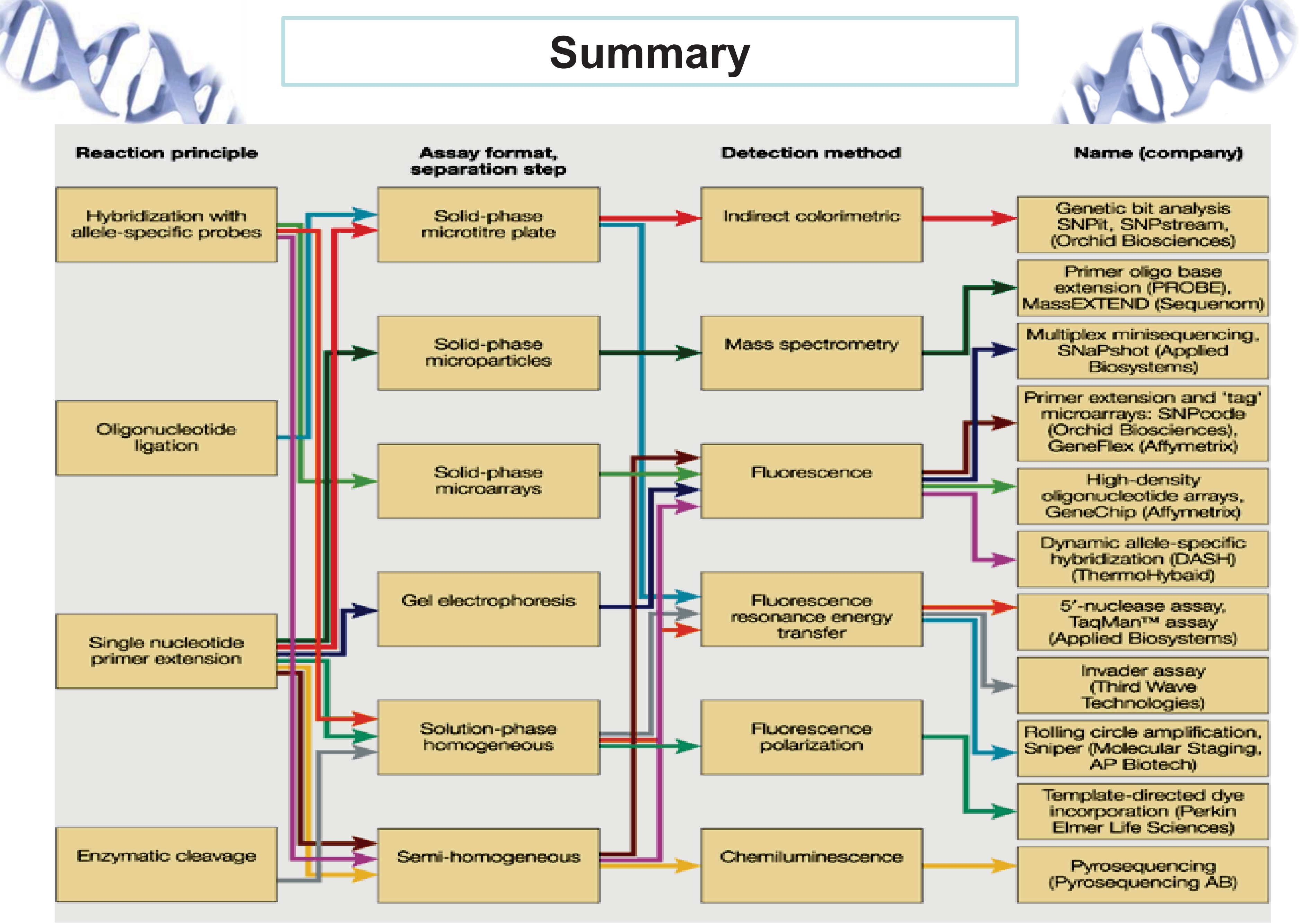
GSTT1 : 86.8± 0.2

Bcl GSTT1

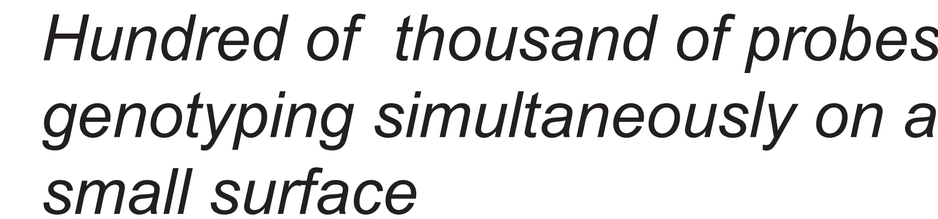
Bcl GSTM1

Low resolution- Validate by Gel electophoresis and always use positive controls.
HRM analysis possible with dyes such as eva green and others with bright ds DNA binding dyes unlike SYBR green

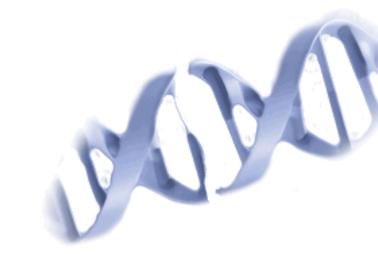
Summary



DNA arrays



Hundred of thousand of probes genotyping simultaneously on a small surface

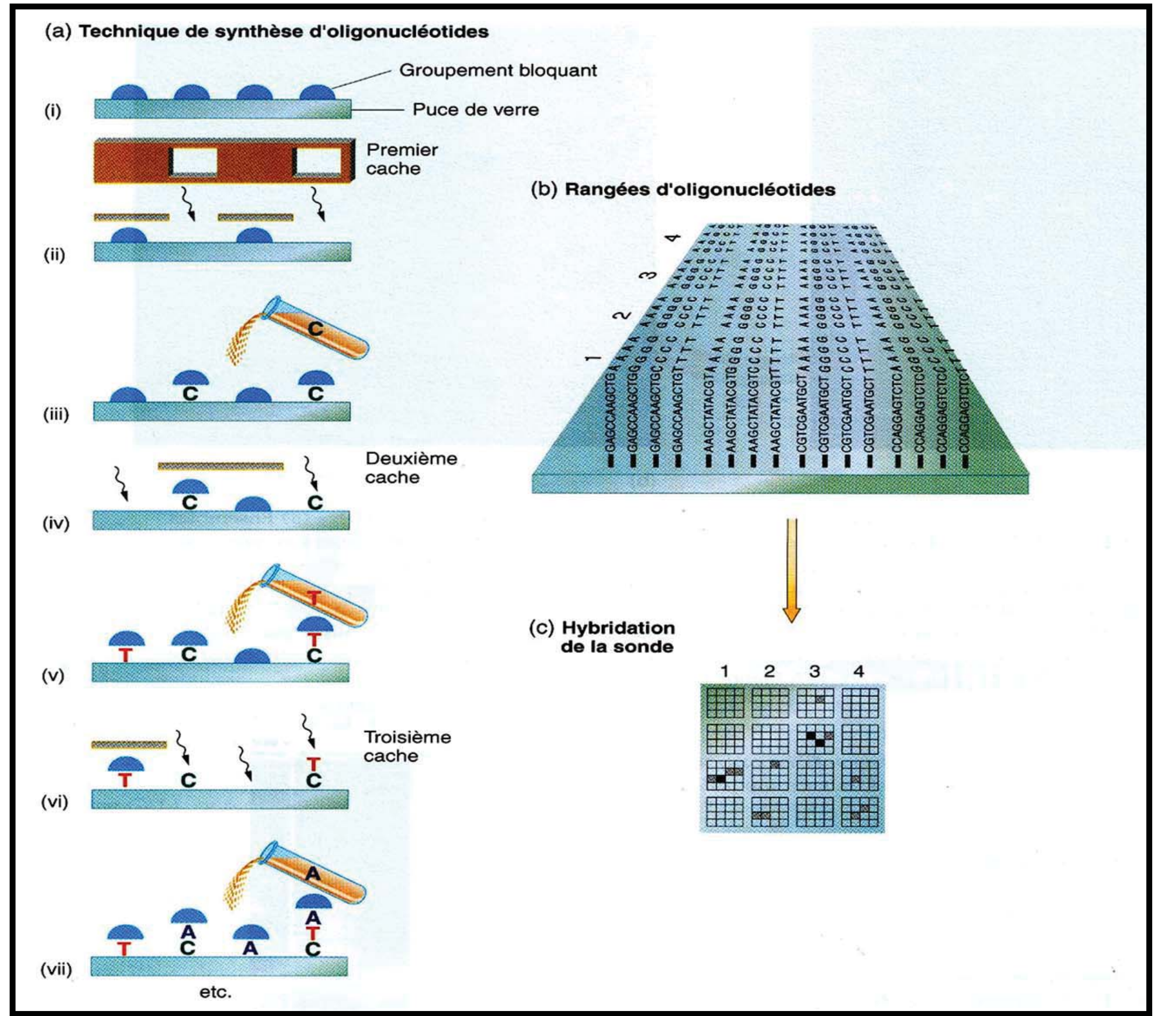


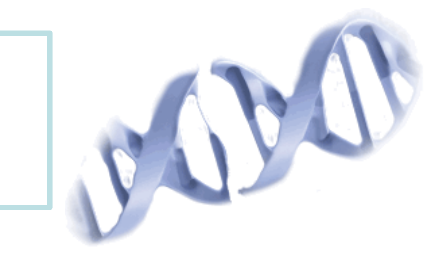
DNA arrays production

Nanotechnologies:
- injectors (printers)
- covers

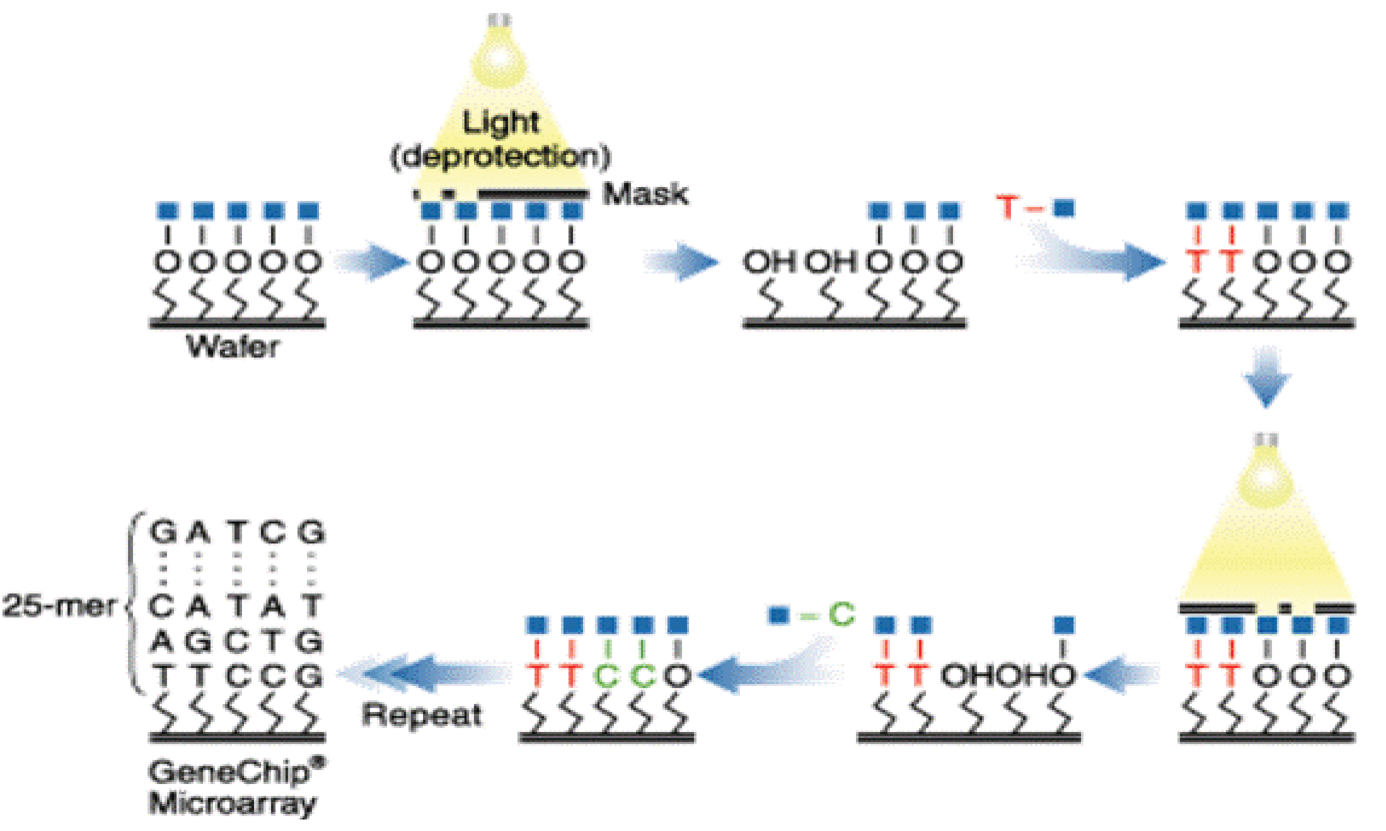
Locations:
- 10-50 mm
- 10-500'000/array
(13-25 mm)

Oligonucleotides:
- 12-25mer
- 10^6 copies/location

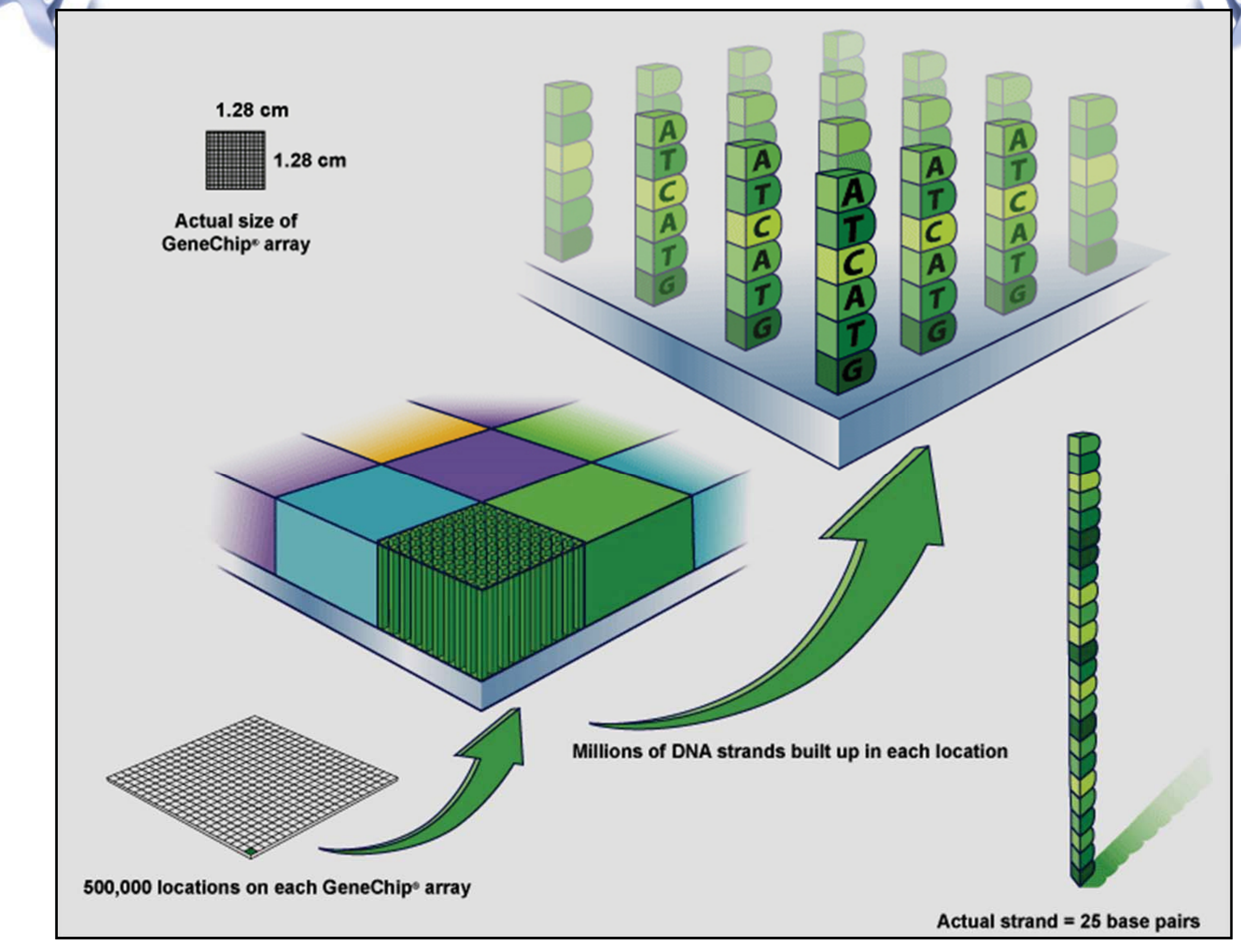




DNA arrays production by photolithography

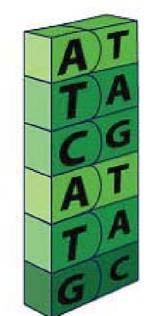


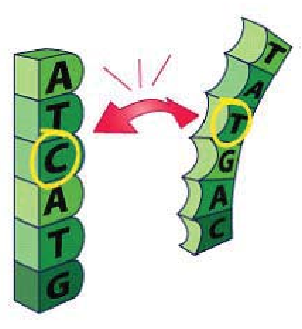
Organisation of DNA arrays

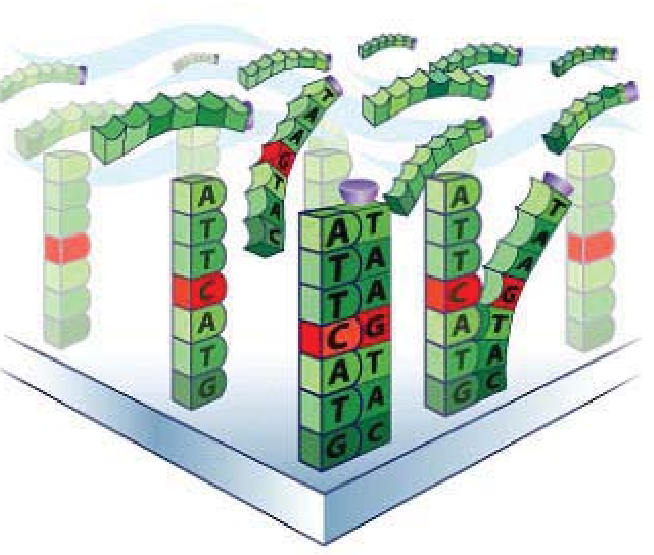


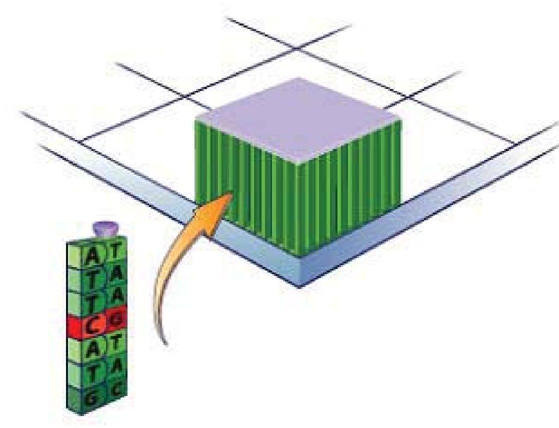
www.affymetrix.com

Hybridization on DNA arrays





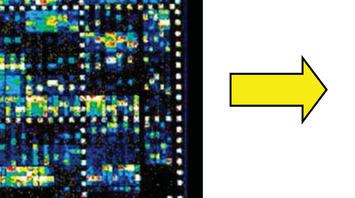




www.affymetrix.com

Data Analysis-DNA arrays

Scanner



Bio-informatics

CYP2D6 polymorphisms

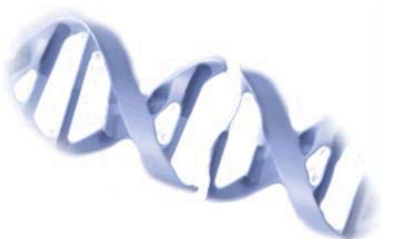
More than 100 allelic variants of CYP2D6 have been characterized and are responsible for the genetic heterogeneity in different populations.
A large interethnic variability is also observed.

Allele CYP2D6		Caucasians (%)	Asians (%)	Africans (%)	other (%)
*2xN	gene duplication (increased activity)	1-10	0-2	1-3	11-29 (Ethiop.)
*3	2549del (inactive enzyme)	1-2	-	-	-
*4	splicing defect (inactive enzyme)	12-21	1	1-9	-
*5	gene deletion (no enzyme)	2-7	6	4-6	10-13 (Japan.)
*6	1707del (inactive enzyme)	1	-	-	-
*10	Pro34Ser, Ser486Thr (instable enzyme)	1-2	33-60	4-6	-
*17	Thr107Ile, Arg296Cys, Ser486Thr (reduced affinity)	0	-	9-34	19 (Korean)

These allelic variants results in 4 metabolic phenotypes : poor metabolisers (PM), intermediate (IM), extensive (EM) and ultra-rapid (UM).

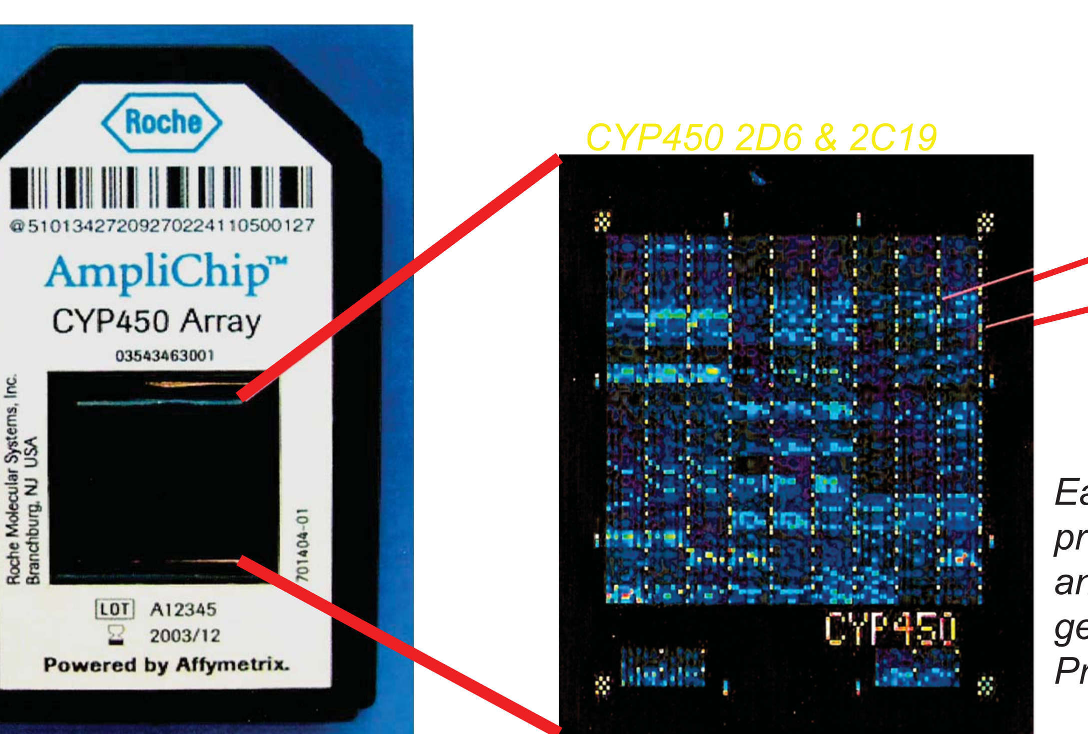
M.Ingelman-Sundberg, 2001, J Int Med 250:186
L. DiAnne Bradford, 2002, Pharmacogenomics 3:229

http://www.cypalleles.ki.se/cyp2d6.htm



AmpliChip CYP450 test (Roche-Affymetrix)

Hybridization of target DNA to labeled complementary probes

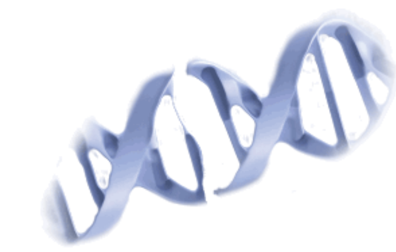


Each array contains >10'000 different probes complementary to the sens and antisens CYP2D6/2C19 genomic DNA.
Probes size: 18mer to 22mer

33 allelic variants of CYP2D6 (including 7 duplications and the deletion); 2 CYP2C19 variants

Redundancy of the information leads to an excellent sensitivity and specificity (240 probes/polymorphism, blocs with coding and non-coding sequences of different sizes and polymorphism position)

AmpliChip CYP450 test



Blood

DNA extraction

Amplification by PCR

Fragmentation / Labeling (biotin)

AmpliChip CYP 450

Hybridization

Scanning

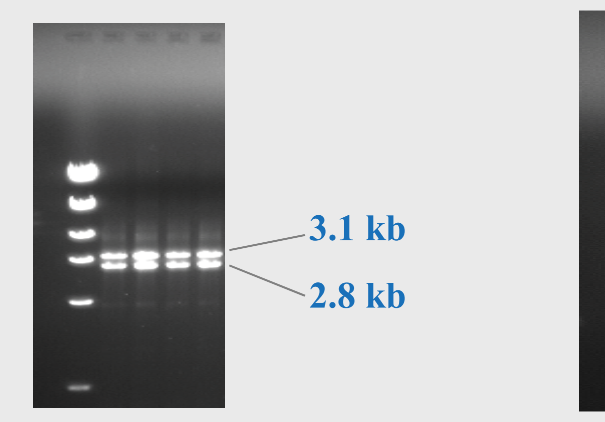
Analyse

Report

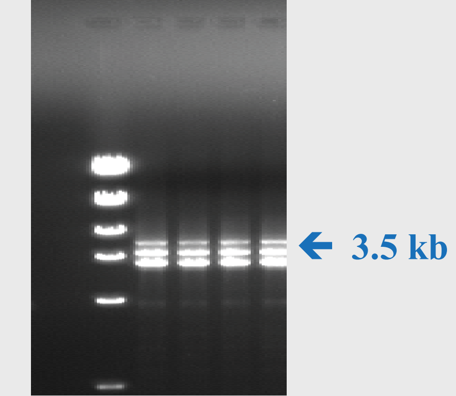
AmpliChip CYP450: PCR amplification

Primer MixA: 2D6

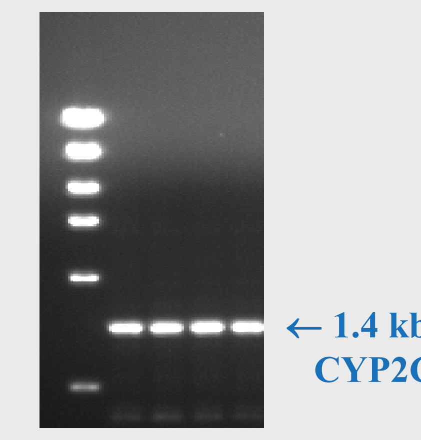
No deletion, no duplication of CYP2D6



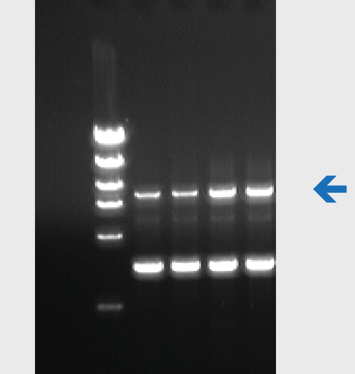
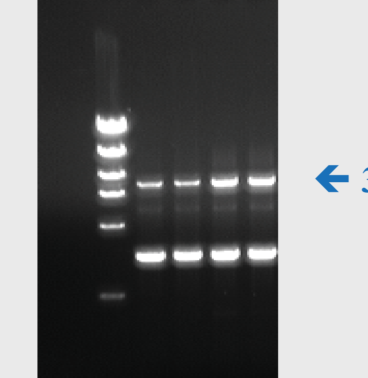
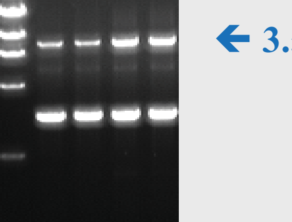
CYP2D6 duplication

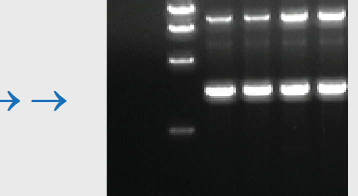


Primer Mix B: CYP2C19 and CYP2D6 deletion



CYP2D6 deletion



DMET array

DMET array
(Drug Metabolizing Enzyme and Transporters)

DMET array allows simultaneous genotyping of 225 genes potentially involved in drug absorption, distribution, metabolism and excretion (ADME).

It analyses 1’936 markers including common and rare variants, short insertion or deletion alleles, and analysis of tri allelic SNPs. In addition to known biomarkers such as common variants in CYP2D6, CYP2C19 and other CYP450 genes, DMET array contains over 1’000 variants in drug transporters that can be used in clinical research studies to discover novel genetic associations. DMET array also performs quantitative assessment of genes with whole-gene deletions (including GSTT1, GSTM1, CYP2D6, CYP2A6, and UGT2B17) and reports allele names in both genotyping reports and translation reports.

www.affymetrix.com

DMET array results

CHP Filename	Gene	Known Call	Common Name	Basecall	Call	Haplotype Marker	Change for Variant	cDNA Change	Genome Position	dbSNP RS ID
00002904	ABCB1	*14/*14	ABCB1*14_61A>G(N21D)	G/G	Var/Var	Y	N21D	61A>G	Ch7:87067376	rs9282564
00002904	ABCB1	*14/*14	ABCB1_68883G>T(S893A)	T/T	Var/Var	Y	A893S,A8	2677G>T	Ch7:86998554	rs2032582
00002904	CYP1A2	*1A/*1F	CYP1A2*1K_-163C>A	C/A	Ref/Var	Y	Promotor	-163C>A	Ch15:72828970	rs762551
00002904	CYP2C19	*1A/*17	CYP2C19*17_-806>(rs12248560)	C/T	Ref/Var	Y	5'UTR	-806C>T	Ch10:96511647	rs12248560
00002904	CYP2C8	*1A/*3	CYP2C8*3_2130G>A(R139K)	G/A	Ref/Var	Y	R139K	2130G>A	Ch10:96817020	rs11572080
00002904	CYP2C8	*1A/*3	CYP2C8*3_30411A>G(K399R)	A/G	Ref/Var	Y	K399R	30411A>G	Ch10:96788739	rs10509681
00002904	CYP2J2	*1/*7	CYP2J2*7_-76G>T(rs890293)	G/T	Ref/Var	Y	Promotor	-76G>T	Ch1:60165082	rs890293
00002904	GSTP1	*A/*B	GSTP1_1375A>G(I105V)	A/G	Ref/Var	Y	I105V	313A>G	Ch11:67109265	rs1695
00002904	NAT2	*4/*6A,*6B/*13	NAT2*6_590G>A(R197Q)	G/A	Ref/Var	Y	R197Q	590G>A	Ch8:18302383	rs1799930
00002904	SLCO1B1	*1B/*17B	SLCO1B1*17_-11187G>A	G/A	Ref/Var	Y	Promotor	-11187G>A	Ch12:21174589	rs4149015
00002904	UGT1A1	*1/H132	UGT1A1*76_1813C>T(rs10929303)	C/T	Ref/Var	Y	3'UTR	1813C>T	Ch2:234346155	rs10929303
00002904	UGT1A1	*1/H132	UGT1A1*78_1941C>G(rs1042640)	C/G	Ref/Var	Y	3'UTR	1941C>G	Ch2:234346283	rs1042640
00002904	UGT1A1	*1/H132	UGT1A1*79_2042C>G(rs8330)	C/G	Ref/Var	Y	3'UTR	2042C>G	Ch2:234346384	rs8330
00002904	UGT2B7	*2A/*2B	UGT2B7*2B_-327>(rs7662029)	A/G	Ref/Var	Y	Promotor	-327A>G	Ch4:69996501	rs7662029
00002904	VKORC1	B3/A4	VKORC1_(rs9923231)	A/G	Ref/Var	Y	Promotor	-1639G>A	Ch16:31015190	rs9923231
00002904	VKORC1	B3/A4	VKORC1_497(rs2884737)	T/G	Ref/Var	Y	Intron	IVS1-T>G	Ch16:31013055	rs2884737
00002904	VKORC1	B3/A4	VKORC1_1173(rs9934438)	T/C	Ref/Var	Y	Intron	IVS1-G>A	Ch16:31012379	rs9934438
00002904	VKORC1	B3/A4	VKORC1_2255(rs2359612)	T/C	Ref/Var	Y	Intron	IVS2-T>C	Ch16:31011297	rs2359612
00002904	FAAH		FAAH_10741C>A(P129T)	C/A	Ref/Var	N	P129T	10741C>A	Ch1:46643348	rs324420
00002904	GSTM1		GSTM1_2643C>G(rs1065411)	C/G	Ref/Var	N	K173N	573C>G	Ch1:110034661	rs1065411
00002904	SLC15A2		SLC15A2_34845A>G(R509K)	G/G	Var/Var	N	R509K	1563A>G	Ch3:123130858	rs1143672
00002904	SLC22A1		SLC22A1_181C>T(R61C)	C/T	Ref/Var	N	R61C	181C>T	Ch6:160463138	rs12208357
00002904	SLC22A1		SLC22A1_17878A>G(M408V)	G/A	Ref/Var	N	M408V	1222G>A	Ch6:160480835	rs628031
00002904	SLCO1A2		SLCO1A2_38T>C(I13T)	T/C	Ref/Var	N	I13T	38T>C	Ch12:21378811	rs10841795
00002904	UGT2B15		UGT2B15*2_>(rs1902023)	T/T	Var/Var	N			Ch4:69218679	rs1902023
00002904	ABCC2	*1A/*1A	All markers responsible for functional changes are Ref/Ref							
00002904	ABCG2	*1/*1	All markers responsible for functional changes are Ref/Ref							
00002904	CDA	*1/*1	All markers responsible for functional changes are Ref/Ref							
00002904	CES2	*1/*1	All markers responsible for functional changes are Ref/Ref							
00002904	CYP1A1	*1/*1	All markers responsible for functional changes are Ref/Ref							
00002904	CYP2A6	*1A/*1B	All markers responsible for functional changes are Ref/Ref							
00002904	CYP2B6	*1A/*1E	All markers responsible for functional changes are Ref/Ref							
00002904	CYP2C9	*1A/*1A	All markers responsible for functional changes are Ref/Ref							
00002904	CYP2D6	*1B/*1B	All markers responsible for functional changes are Ref/Ref							

MALDI-TOF Based genotyping

- *Primer extension reactions designed to generate different sized products*
- *Analysis by mass spectrometry*

C/T

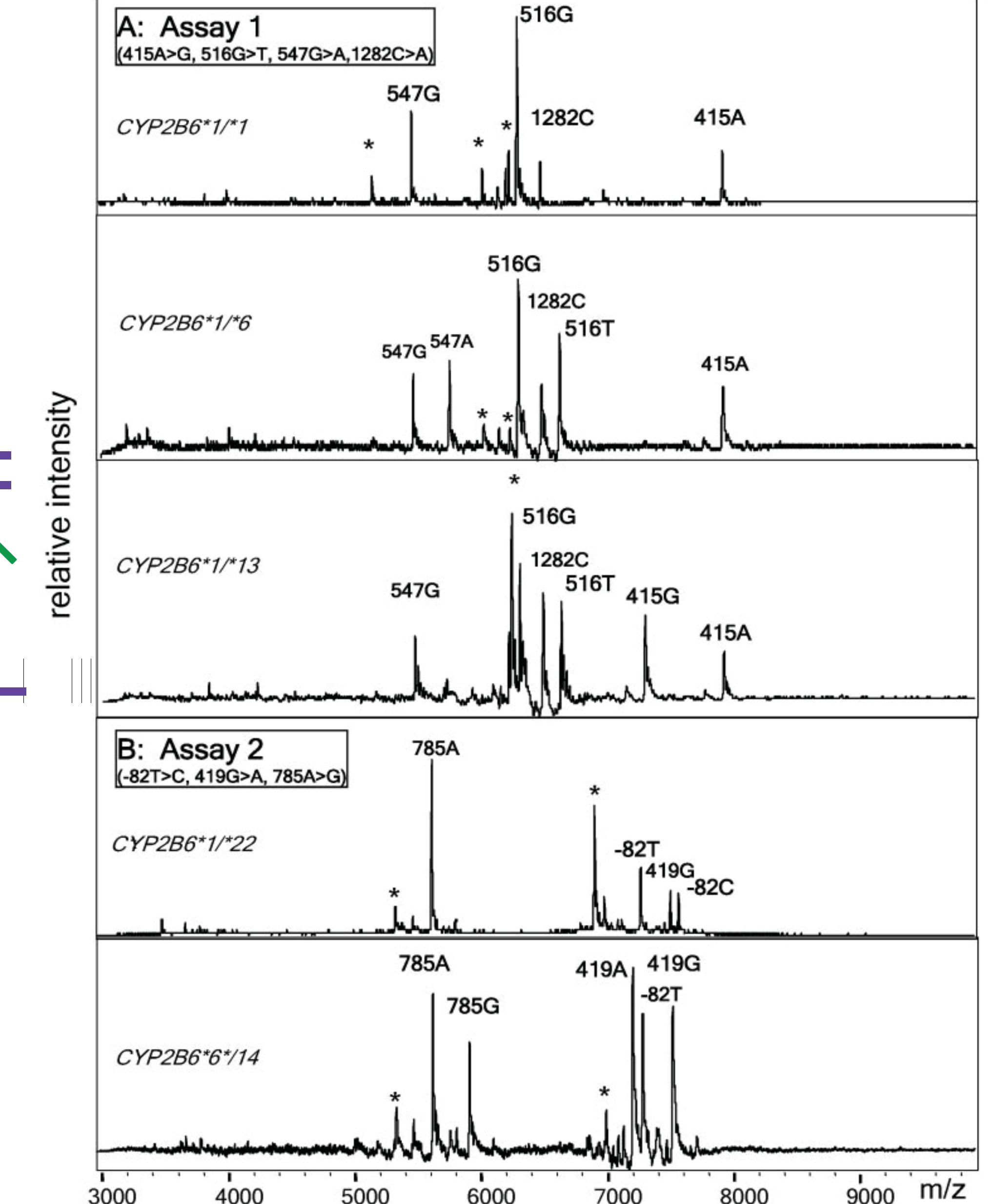
G/A

dGTP dTTP

dATP ddCTP

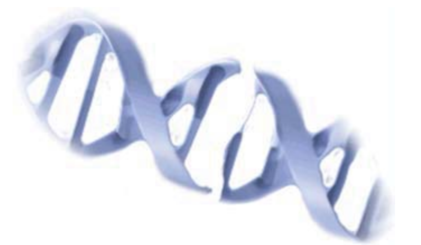
G/A

GGACCTGGAGCCCCCACC
GGACCTGGAGCCCCCACCC



Genome wide association studies (GWAS)

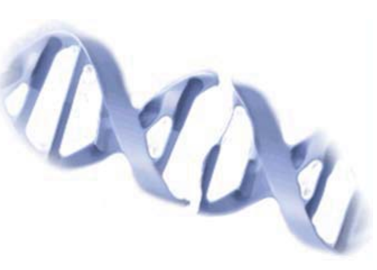
- *No knowledge of candidate gene is required.*
- *Unbiased search across the genome to find genetic polymorphisms.*
- *Utilizes arrays-based technologies to simultaneously genotype hundreds of thousand of SNP.*
- *Main information about genetic variations across the genome has been firstly provided by the HapMap project/1000 genome project*
- *Linkage disequilibrium pattern could be exploited for haplotype construction and tagging of SNPs*



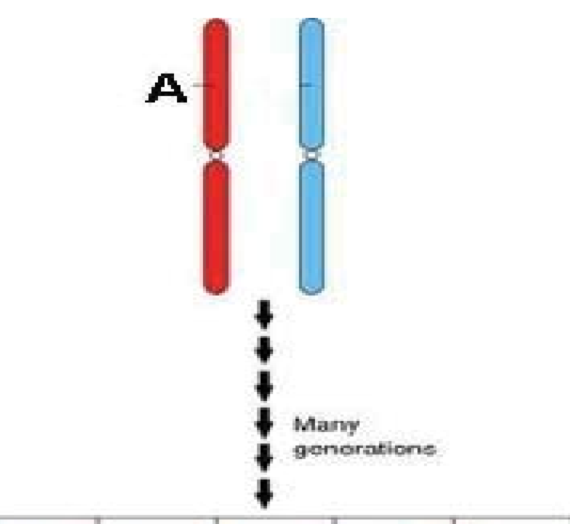
The HapMap project

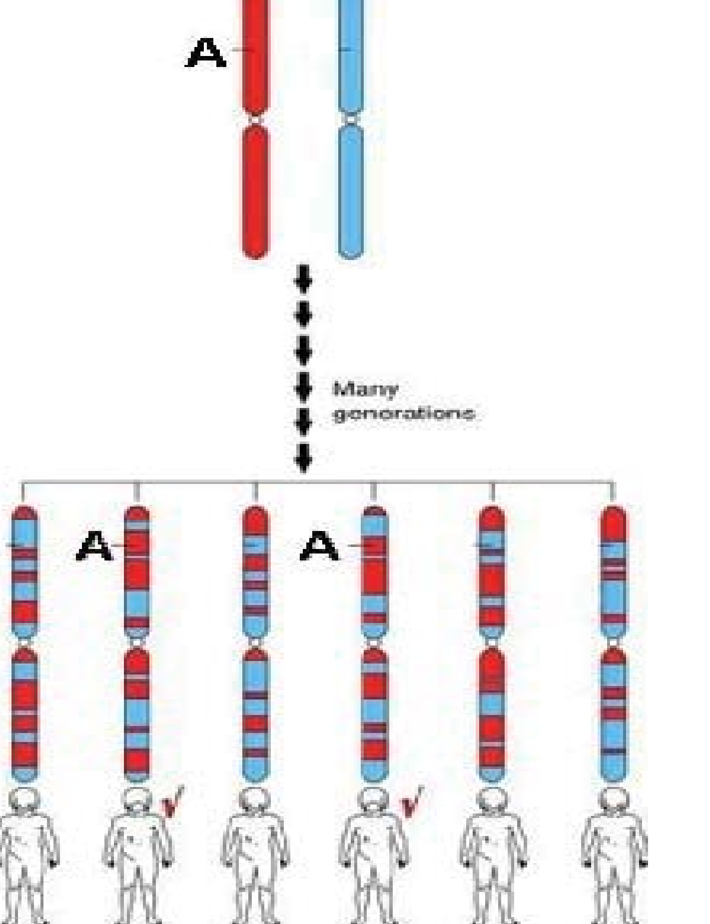
The aim of the International HapMap project is to characterize common sequence variants, their frequencies, and correlation between them in 270 DNA samples from 4 geographically different populations (Africa, Asia and Europe).

The second generation of human haplotype map (2007) characterized 3.1 millions SNPs and includes 25-35% of common SNPs (MAF > 5%) .



Origin of haplotypes

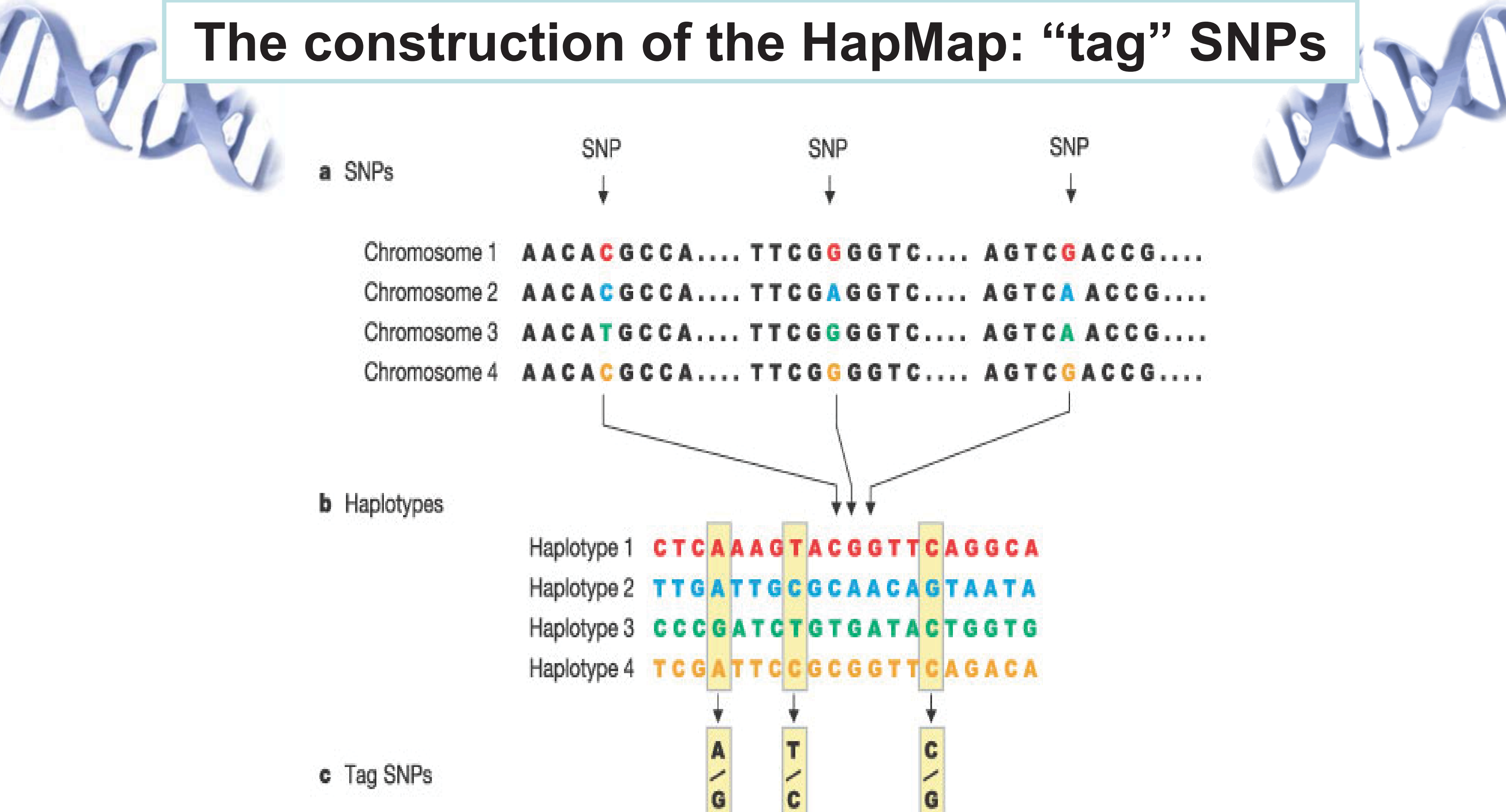




This diagram shows two ancestral chromosomes being scrambled through recombination over many generations to yield different descendant chromosomes. Adjacent to the variant marked by the A are many SNPs that can be used to identify the location of the variant.

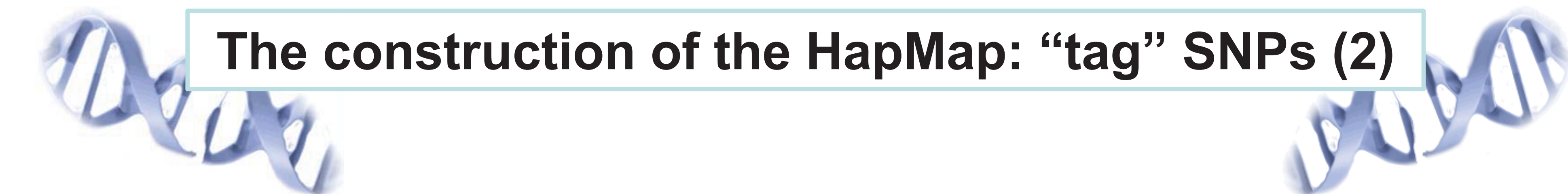
www.hapmap.ncbi.nlm.nih.gov

The construction of the HapMap: “tag” SNPs



(a) SNPs are identified in DNA samples from multiple individuals. (b)Adjacent SNPs that are inherited together are compiled into haplotypes. (c)"Tag" SNPs within haplotypes are identified that uniquely identify those haplotypes. By genotyping the three tag SNPs shown in this figure, researchers can identify which of the four haplotypes shown here are present in each individual.

www.hapmap.ncbi.nlm.nih.gov



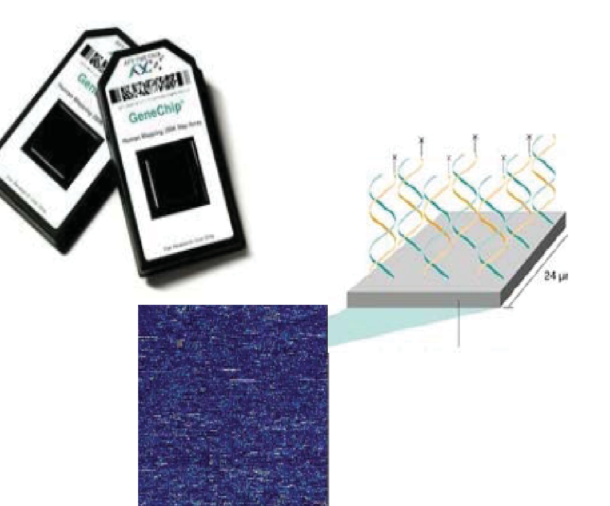
The construction of the HapMap: “tag” SNPs (2)

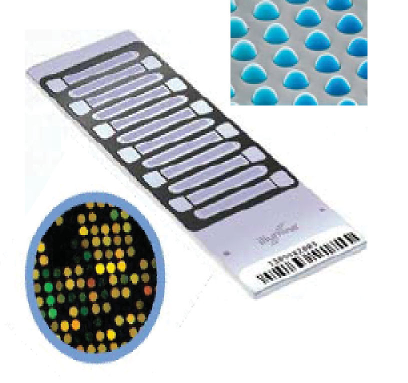
Defining haplotypes (linkage disequilibrium) and identifying “tag” SNPs allows reducing the number of SNPs to be genotyped in order to cover genetic variation.

The number of “tag” SNPs that contains most of the information about patterns of genetic variation is estimated to be about 300’000 to 600’000, which is far lower than the total 10 millions common SNPs.

Examples of GWAS arrays technologies

High-throughput array-based genotyping





Affymetrix

Human SNP Array 6.0

>1.8 million markers

906,600 SNPs

946,000 for CNVs

Illumina

Human 660W-Quad BeadChip

2.6 million markers / four samples

550,000 tag SNPs

100,000 for CNVs

5,000 common CNVs

GWAS

- *A genome-wide association study is an approach that involves genotyping markers across the genome (≈0.5-1Million) of a large number of individuals (≈2000) to find genetic variations associated with a particular disease.*

- *A large number of subjects are needed because*
  - *associations between SNPs and causal variants are expected to show* **low odds ratios**, *typically below 1.5*
  - *in order to obtain a reliable signal, given the very large number of tests that are required, associations must show a high level of significance to survive the multiple testing correction*

– *Such studies are particularly useful in finding genetic variations that*
– *contribute to common, complex diseases (diabetes, obesity,*
– *cancer…)*

GWAS (2)

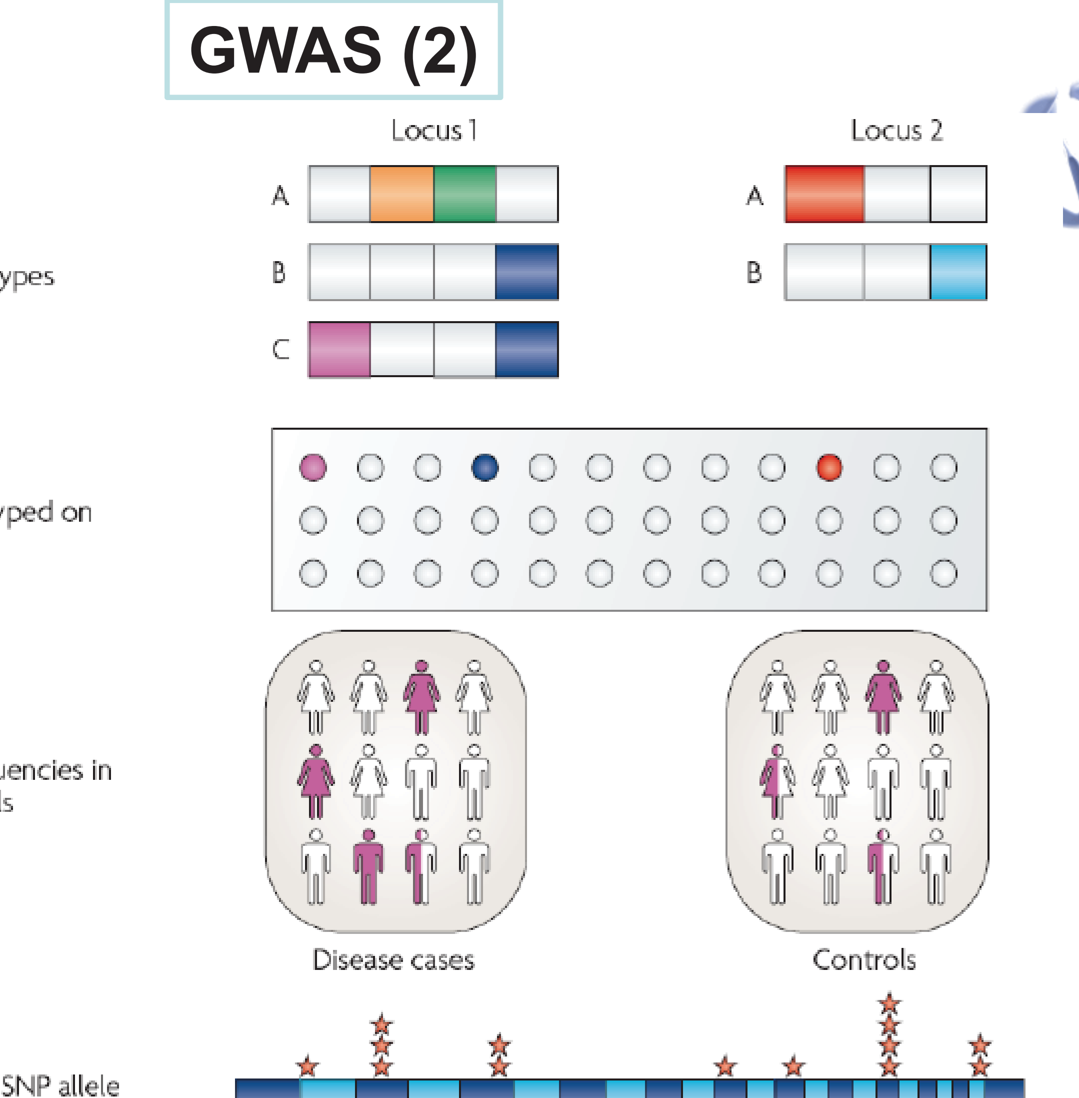
HapMap
Select SNPs to tag haplotypes

Genotyping
300,000–500,000 SNPs typed on high-density arrays

Case–control study
Compare SNP allele frequencies in disease cases and controls

Genome scan result
Significant differences in SNP allele frequencies indicate possible new disease genes and loci

Replication test
Confirm scan findings



Mathew, Nat Rev Genet, 9, 2008

GWAS (3)

1a
Study sample 1
Microarray SNP chip

Selection of best SNPs with $p<10^{-7}$

or

1b
Study sample 1
Microarray SNP chip
Study sample 2
Microarray SNP chip

Selection of best SNPs with $p<0.01$ in both scans

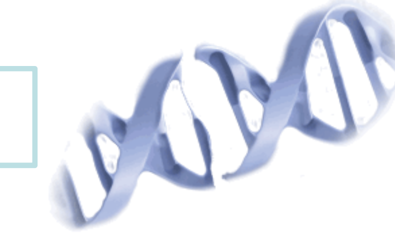
2 **Validation**
in one large study population
10-500 SNPs

3 **Replication**
in other populations
5-50 SNPs

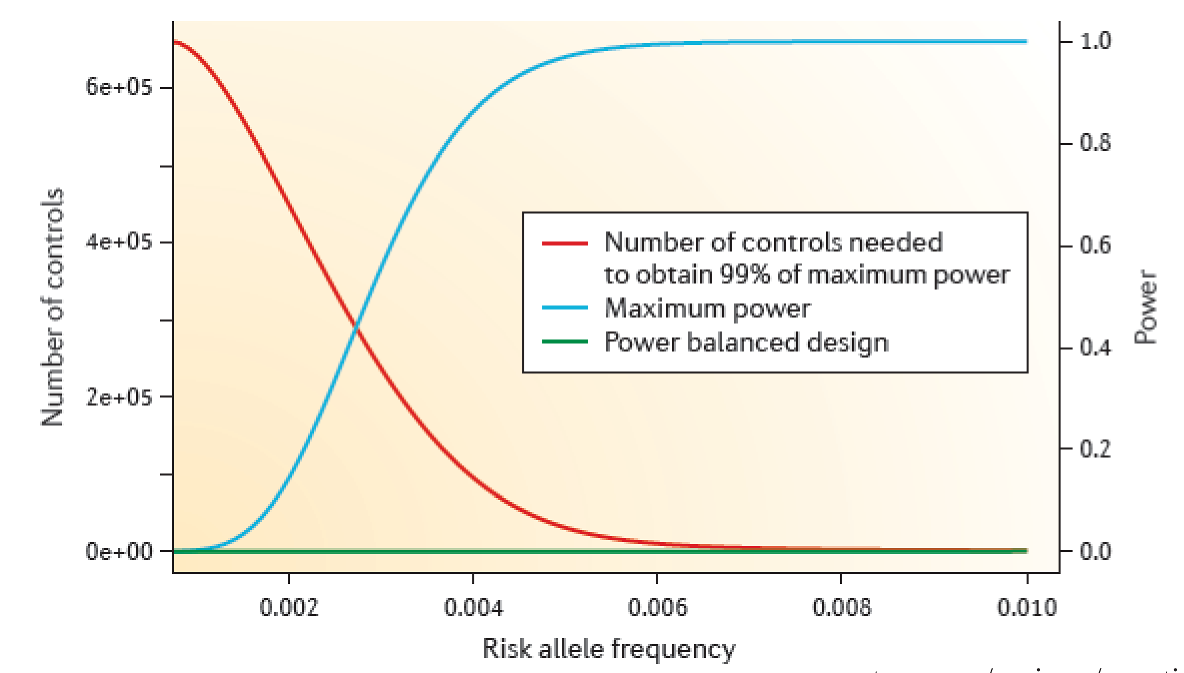
4 Exon 1 2 3 4 5
Finemapping
by sequencing and genotyping
5-20 SNPs per gene

5 **Functional studies**
(bioinformatics, cell culture, animal models, ...)
1-2 genes

Kronenberg, Exp Geront 43, 2008

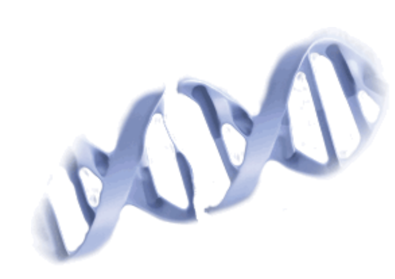


Sample Size for GWAS



www.nature.com/reviews/genetics 460 | JULY 2013 | VOLUME 14

"...the extreme phenotypes approach should be prioritized for genome-wide pharmacogenomic studies, including those exploring genomic biomarkers for treatment efficacy as an alternative to enrolling very large random patient cohorts." *Pharmacogenomics (2013) 14(4), 337–339*

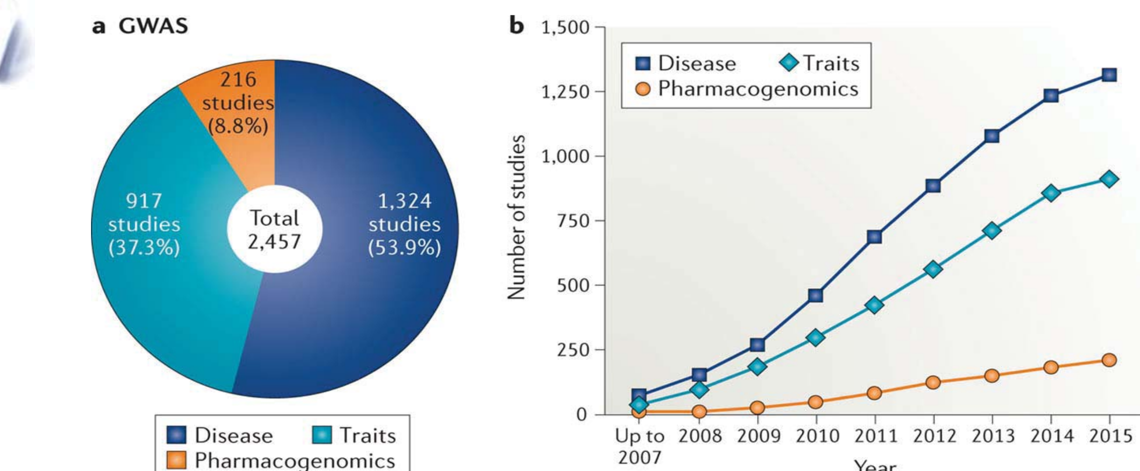


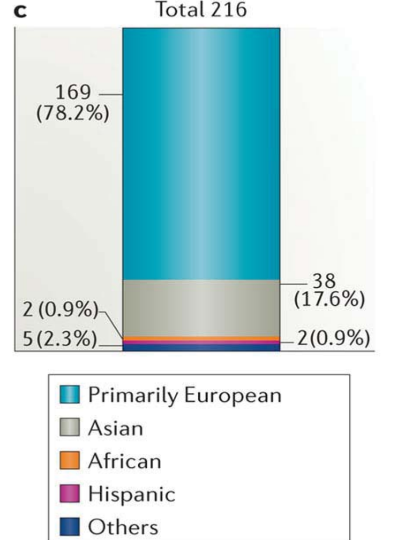
GWAS in toxicogenomics

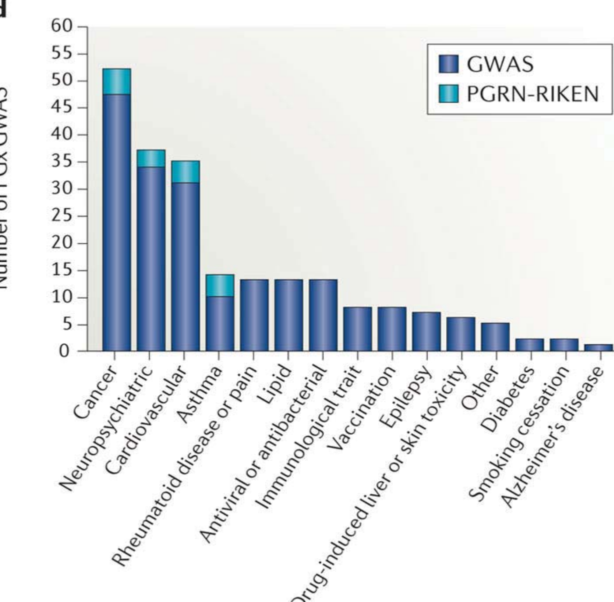
Differs from complex-diseases GWA:

- *Sample size: insufficient power due to smaller sample size. Drug non-responders are rarer than responders and severe adverse drug reactions often affect only 1/10’000-1/100’000 treated patients.*
- *Phenotypic characterization: heterogeneity in drug response phenotype.*
- *Replication: small sample size of many GWAS makes replication challenging.*
- *Effect size: several GWAS involving only 100 cases have detected significant association suggesting that some pharmacogenomic effects may be larger and involve fewer genes than complex diseases.*

GWAS in PgX







Nature Reviews Drug Discovery **16**, *70 (2017)*

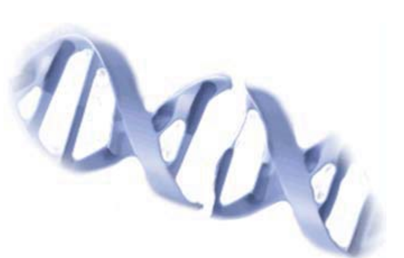
GWAS and drug response

70% of GWAS in toxicogenomics concern drug response

Drug	Response	Number of cases	Genome-wide significance/ replication	Lowest p value	SNP genotyping platform	Significant gene(s)
Warfarin	Maintenance dose	181; 1,053	Yes/Yes	6×10^{-13}; 1×10^{-78}	Illumina 550K; Illumina CNV370	*VKORC1*, *CYP2C9*, *CYP4F2*
Acenocoumarol	Maintenance dose	1,451	Yes/Yes	2×10^{-123}	Illumina 550K	*VKORC1*, *CYP2C9*, *CYP4F2*, *CYP2C18*
Interferon-α	Response in hepatitis C infection assessed by absence of viral RNA in serum	1,137; 293; 154	Yes/Yes	1×10^{-25}; 9×10^{-9}; 3×10^{-15}	Illumina 610 quad	*IL28B*
Clopidogrel	Antiplatelet effect assessed by platelet aggregometry	429	Yes/Yes	4×10^{-11}	Affymetrix 500K or 1M	*CYP2C19*
Methotrexate	Drug clearance and incidence of gastrointestinal toxicity in childhood leukaemia patients	434	Yes/Yes	1.7×10^{-10}	Affymetrix 500K	*SLCO1B1*
Thiazide	Diuretic response assessed by diastolic blood pressure change	389*	Yes/No	2×10^{-7}‡	Affymetrix 100K	Region of chromosome 12q15
Interferon-β	Response in multiple sclerosis assessed by clinical scoring	206	No/No	0.004	Affymetrix 100K	None found

Daly, Nature Reviews, Genetics,11,2010

GWAS and adverse drug reactions



Drug	ADR	Gene variant	Ref
Ribavarin	*Hemolytic anaemia*	*ITPA*	*Nature 464, 405–408 (2010)*
Thiopurine	*Leucopenia*	*NUD15*	*Nat. Genet. 46, 1017–1020 (2014).*
Phenytoin	*SJS*	*CYP2C9*	*JAMA 312, 525–534 (2014)*
Cisplatin	*Hearing loss*	*ACYP2*	*Nat Genet. 2015 Mar; 47(3): 263–266.*
Vincristine	*Peripheral neuropathy*	*CEP72*	*JAMA. 2015 Feb 24;313(8):815-23*
Statins	*myopathy*	*SLCO1B1*	*Pharmacogenet Genomics. 2011 May;21(5):280-8*
Bisphosphonates	*osteonecrosis*	*CYP2C8*	*Blood. 2008 Oct 1;112(7):2709-12.*

https://www.ebi.ac.uk/gwas/

GWAS in toxicogenomics

- *As for complex diseases, the number of GWAS in toxicogenomics is increasing rapidly.*

- *GWAS results have confirmed several association previously revealed by candidate-gene approaches. Due to small cohorts mainly genes with large effects are identified.*

- *Stratification of patients at higher risk of developing ADR helps*
  - *-- to choose alternate medication*
  - *-- to initiate prophylactic measures*
  - *--to evaluate treatment strategies in high risk population*
  - *--Increased observation /care of high risk patients*
  - *--Aiding in preventing, personalizing, and providing precise medication (P3 medicine)*

1000 Genomes Project

- *It has been suggested that common genomic variants covered in the HapMap project can explain only a fraction (<50%) of the variation in drug response.*
- *Other genetic variations (as CNV), non-genetic factors but also rare variants could explain the remaining variation.*
- *The 1000 Genome Project (KGP) has the goal to establish a detailed catalogue of human genetic variation in at least 1000 human genomes from world-wide populations using next-generation sequencing technologies.*
- *Specific aims: identify >95% of the variants with alleles frequencies >1% in the genome and >95% of variants with frequencies >0.1-0.5% in exons.*

1000 Genomes Project in toxicogenomics: finding new SNPs

Gene	Full name	Additional novel SNPs*	Known SNPs‡
ADRB1	Adrenergic, β-1-, receptor	3	2
ADRB2	Adrenergic, β-2-, receptor, surface	0	8
AHR	Aryl hydrocarbon receptor	25	29
ALOX5	Arachidonate 5-lipoxygenase	39	107
BRCA1	Breast cancer 1, early onset	22	65
COMT	Catechol-*O*-methyltransferase	8	50
CYP2A6	Cytochrome P450, family 2, subfamily A, polypeptide 6	0	1
CYP2D6	Cytochrome P450, family 2, subfamily D, polypeptide 6	1	1
DRD2	Dopamine receptor D2	36	113
GSTP1	Glutathione *S*-transferase pi 1	2	5
KCNJ11	Potassium inwardly-rectifying channel, subfamily J, member 11	1	7
MTHFR	5,10-methylenetetrahydrofolate reductase	8	54
NQO1	NAD(P)H dehydrogenase, quinone 1	7	16
P2RY1	Purinergic receptor P2Y, G-protein coupled, 1	1	4
SCN5A	Sodium channel, voltage-gated, type V, α-subunit	56	204
SLC19A1	Solute carrier family 19 (folate transporter), member 1	16	38
SULT1A1	Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	5	12
TPMT	Thiopurine *S*-methyltransferase	23	35
TYMS	Thymidylate synthetase	5	42
VDR	Vitamin D receptor	34	111
VKORC1	Vitamin K epoxide reductase complex, subunit 1	1	6

Browse SNPs : https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/

Zhang and Dolan, Pharmacogenomics, 11,2010

100,000 Genomes Project

- *The project will sequence 100,000 genomes. Participants are NHS (UK) patients with a rare disease, plus their families, and patients with cancer.*

- *The raw data from one genome is about 200GB which would occupy most of the average laptop's hard drive*

- *Data Banking, data management, analysis algorithm development, analysis pipeline management -promising avenues in near future*

Evolution of sequencing technologies

- *Sanger sequencing has the capacity to read through 1000-1200 bp.*
- *Shotgun sequencing (Human Genome Project): genomic DNA is enzymatically or mechanically broken down in small fragments and cloned into sequencing vectors to be sequenced.*
- *Complete sequence of a long DNA fragment is generated by alignment and reassembly of the sequence fragments based on partial sequence overlaps (= sequencing of the entire genome is possible).*
- *Next generation sequencing (NGS): breaking DNA in small pieces, ligating them to adapters for random read during DNA synthesis. The number of continuous sequenced bases is short, therefore coverage is a major issue (= number of short reads that overlap each other within a specific genomic region).*

Single stranded DNA to be sequenced 3′ T G A C G C T T G T C A 5′

Sequencing reactions using different fluorescent-labeled primers with each dideoxynucleotide

ddG

ddA

ddC

ddT

Pool products
Electrophoresis

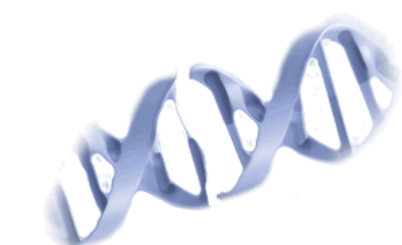
Fragment migration

Photomultiplier

eam

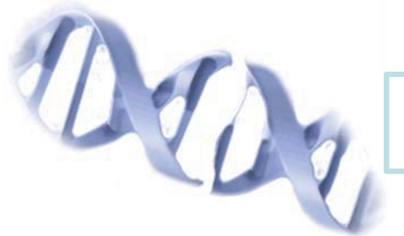
Fluorescence

Computer



Electopherogram

A T C G



Sequence Anaylsis

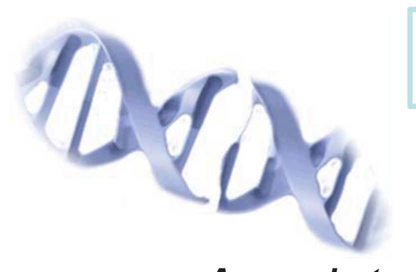
Iam I am I am I am amah amah amah mah mah mah ahuman ahuman ahuman ahuman human human human human manbe anbein anbein anbein anbein nbeing nbeing nbeing

↓

?

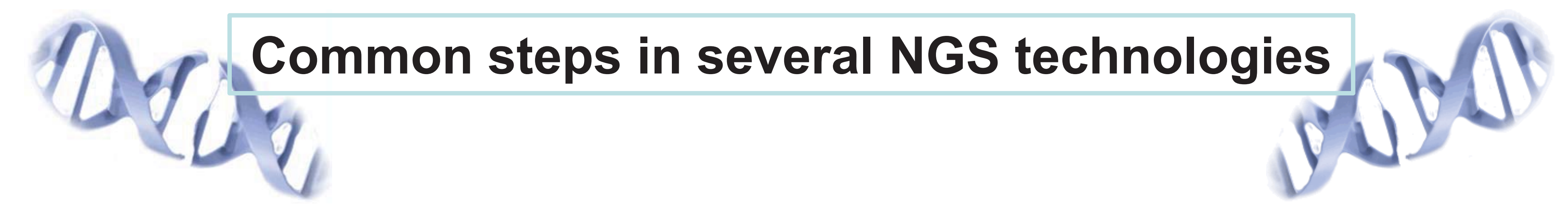
I am a human being

I ban Hinge mumaa



Next generation sequencing (NGS)

- *A variety of NGS technologies are now available.*
- *Sequential fluorescent imaging of base incorporations in large scale arrays of single DNA molecules or single molecule-derived clusters of DNA.*
- *Enzymology of cluster generation and base incorporation differ between technologies.*
- *These technologies produce shorter individual sequences with higher per base error rates than traditional Sanger sequencing but the enormous number of individual sequences generated result in high-redundant coverage of each base leading to accurate consensus sequence and calling of variants.*
- *NGS can be applied to whole genomes, exomes or can be focused on panels of candidate genes (hybridization-based sequence capture or multiplexed PCR amplification).*



Common steps in several NGS technologies

Library preparation

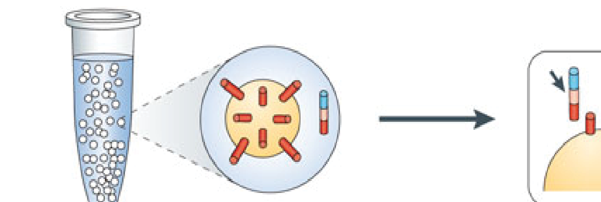
Sequence capture

Amplification

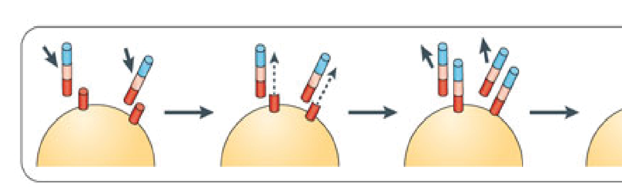
Sequencing

Template amplification

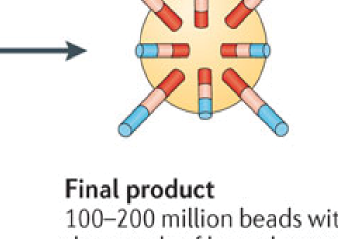
Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))



Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase

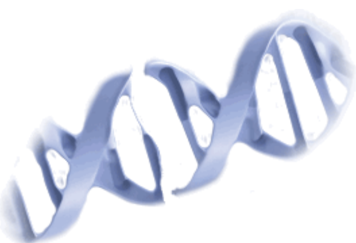


On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates



Final product
100–200 million beads with thousands of bound template

Nature Reviews Genetics 17, 333–351 (2016)

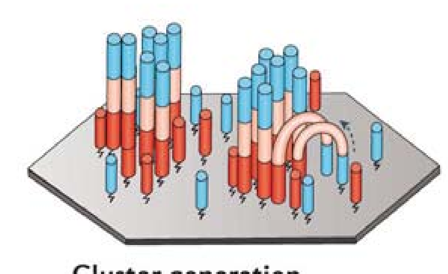


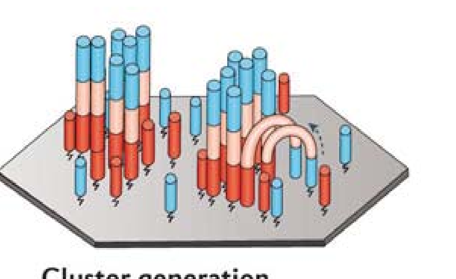
Template amplification

b **Solid-phase bridge amplification (Illumina)**

Template binding
Free templates hybridize with slide-bound adapters

Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place



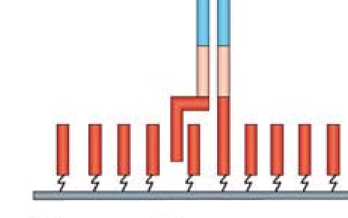


Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed

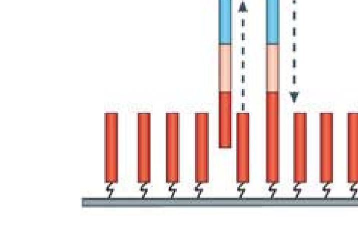
Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density

c **Solid-phase template walking (SOLiD Wildfire (Thermo Fisher))**

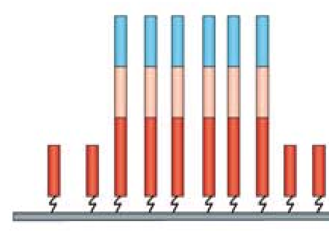
Template binding
Free DNA templates hybridize to bound primers and the second strand is amplified



Primer walking
dsDNA is partially denatured, allowing the free end to hybridize to a nearby primer



Template regeneration
Bound template is amplified to regenerate free DNA templates

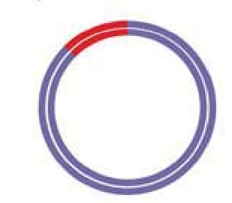


Cluster generation
After several cycles of amplification, clusters on a patterned flow cell are generated

Nature Reviews Genetics 17, 333–351 (2016)

Template amplification

d In-solution DNA nanoball generation (Complete Genomics (BGI))

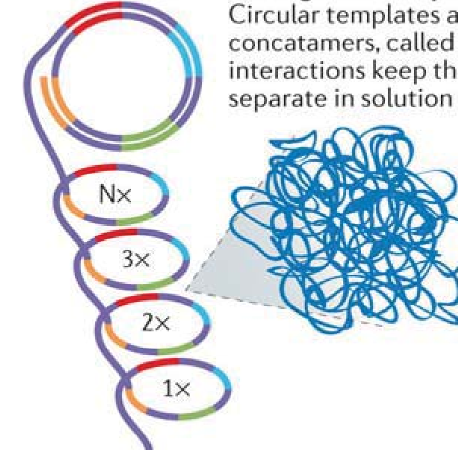


Adapter ligation
One set of adapters is ligated to either end of a DNA template, followed by template circularization

Cleavage
Circular DNA templates are cleaved downstream of the adapter sequence

Iterative ligation
Three additional rounds of ligation, circularization and cleavage generate a circular template with four different adapters

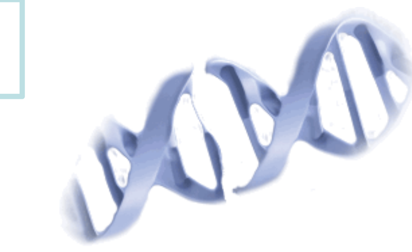
Rolling circle amplification
Circular templates are amplified to generated long concatamers, called DNA nanoballs; intermolecular interactions keep the nanoballs cohesive and separate in solution



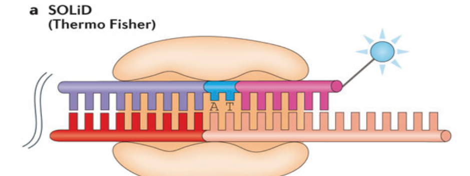
Hybridization
DNA nanoballs are immobilized on a patterned flow cell

Nature Reviews | Genetics

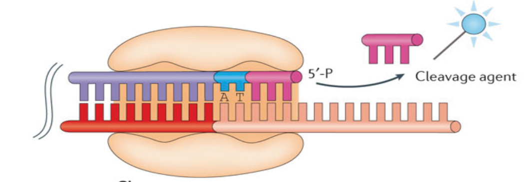
Sequencing by ligation

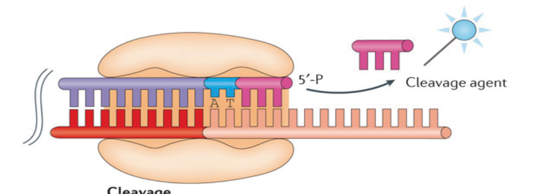


a SOLiD (Thermo Fisher)

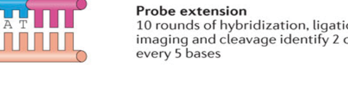


Two-base-encoded probes
Probes with two known bases followed by degenerate or universal bases hybridize to a template; ligase immobilizes the complex and the slide is imaged





Cleavage
The fluorophore is cleaved from the probe along with several bases, revealing a 5′ phosphate



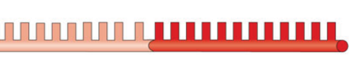
Probe extension
10 rounds of hybridization, ligation, imaging and cleavage identify 2 out of every 5 bases

Anchor with an n+2 offset

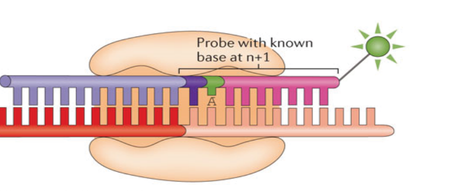
Reset
After a round of probe extension, all probes and anchors are removed and the cycle begins again with an offset anchor

b Complete Genomics (BGI)

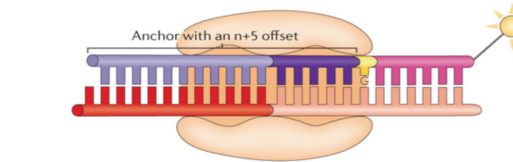
Single-base-encoded probes
A probe with a single known base and degenerate bases hybridizes to a template and is imaged



Reset
After each imaging step, both the probe and anchor are removed



Paired-end sequencing
Sequencing is performed for both the left and right sides of the adapter

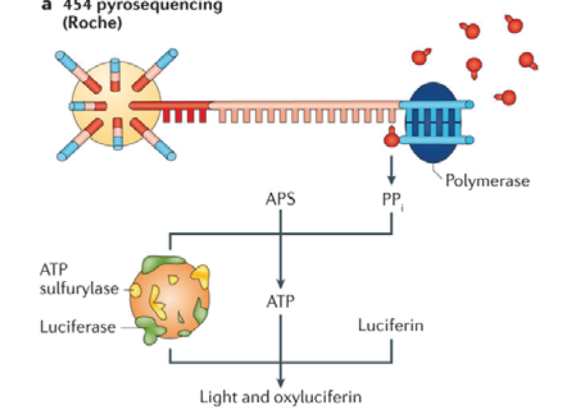


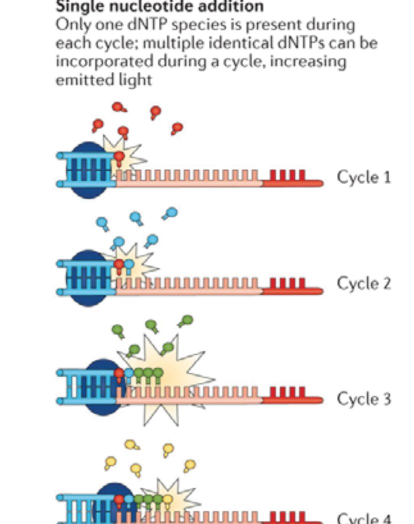
Offset anchors
Subsequent rounds of hybridization and ligation use offset anchors to sequence more-distant bases

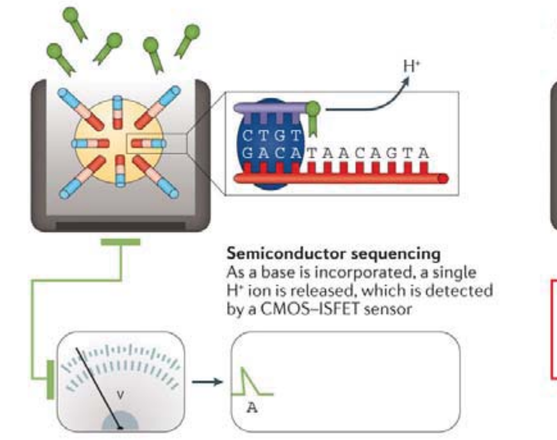
Nature Reviews | Genetics

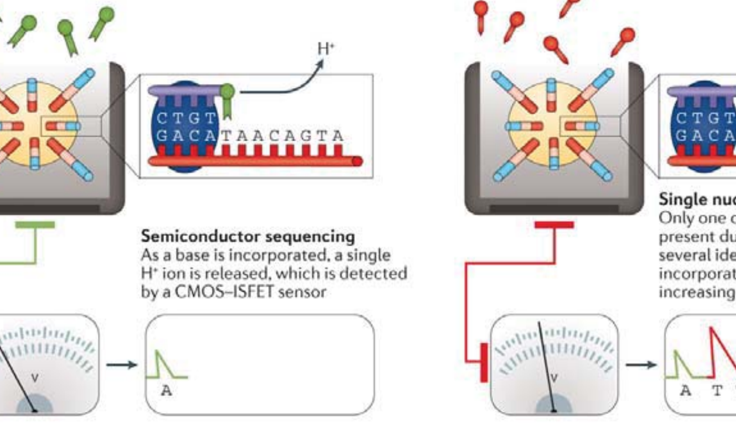
Nature Reviews Genetics 17, 333–351 (2016)

Sequencing by synthesis-SNA



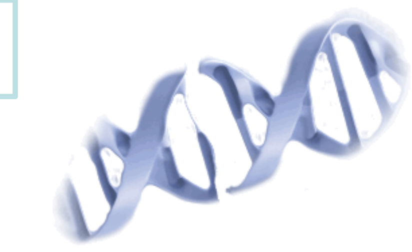


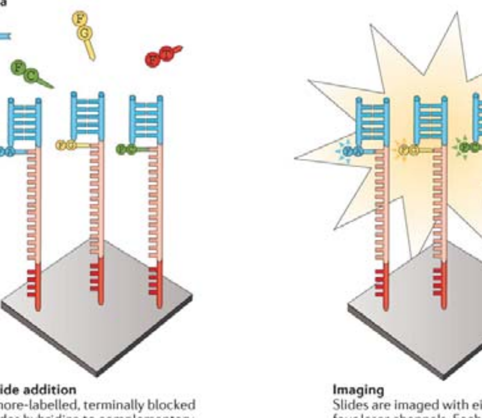
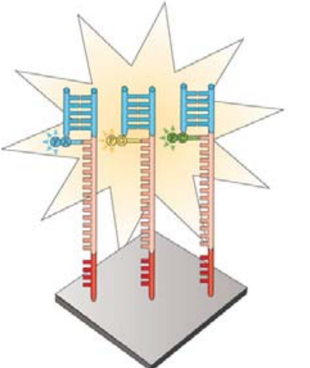
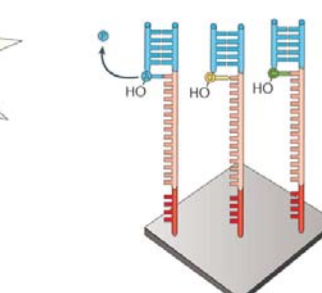
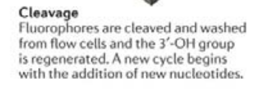
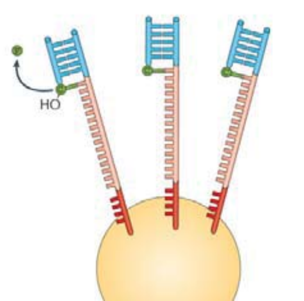
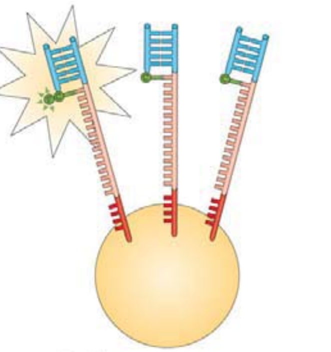
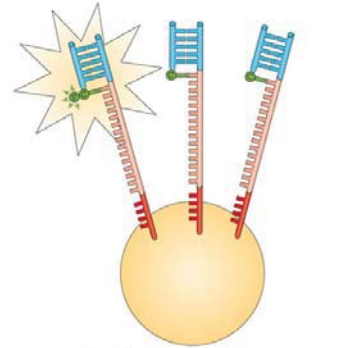
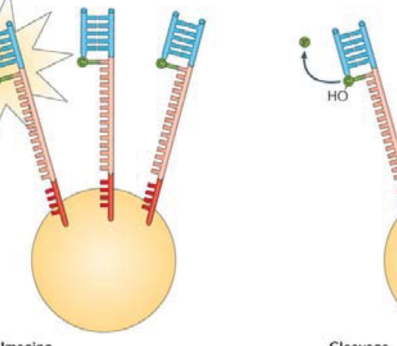
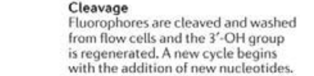




Nature Reviews Genetics 17, 333–351 (2016)

Sequencing by synthesis-CRT



Nature Reviews Genetics 17, 333–351 (2016)

Different platforms for NGS

Sequencing by ligation

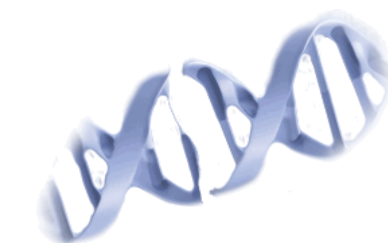
Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
SOLiD 5500 Wildfire	50 (SE) 75 (SE) 50 (SE)*	80 Gb 120 Gb 160 Gb*	~700 M*	6 d*	≤0.1%, AT bias	NA§	$130
SOLiD 5500 xl	50 (SE) 75 (SE) 50 (SE)*	160 Gb 240 Gb 320 Gb*	~1.4 B*	10 d*	≤0.1%, AT bias	$251,000	$70
BGISEQ-500 FCS155	50–100 (SE/PE)*	8–40 Gb*	NA‖	24 h*	≤0.1%, AT bias	$250	NA
BGISEQ-500 FCL155	50–100 (SE/PE)*	40–200 Gb*	NA‖	24 h*	≤0.1%, AT bias	$250,000	NA

Approx., approximate; AT, adenine and thymine; B, billion; bp, base pairs; d, days; Gb, gigabase pairs; h, hours; indel, insertions and deletions; Kb, kilobase pairs; M, million; Mb, megabase pairs; NA, not available; PE, paired-end sequencing; SBS, sequencing by synthesis; **SE, single-end sequencing**; Tb, terabase pairs.

*Manufacturer's data.

Nature Reviews Genetics 17, 333–351 (2016)

Different platforms for NGS



Sequencing by synthesis: CRT

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Illumina NextSeq 500/550 Mid output	75 (PE)	16–20 Gb	Up to 260 M (PE)*	15 h	<1%, substitution	$250	$42
	150 (PE)*	32–40 Gb*		26 h*			$40
Illumina NextSeq 500/550 High output	75 (SE)	25–30 Gb	400 M (SE)*	11 h	<1%, substitution	$250	$43
	75 (PE)	50–60 Gb	800 M (PE)*	18 h			$41
	150 (PE)*	100–120 Gb*		29 h*			$33
Illumina HiSeq2500 v2 Rapid run	36 (SE)	9–11Gb	300 M (SE)*	7 h	0.1%, substitution	$690	$230
llumina HiSeq2500 v3	36 (SE)	47–52 Gb	1.5 B (SE)	2 d	0.1%, substitution	$690	$180
	50 (PE)	135–150 Gb	3 B (PE)*	5.5 d			$78
	100 (PE)*	270–300 Gb		11 d*			$45
Illumina HiSeq2500 v4	36 (SE)	64–72 Gb	2 B (SE)	29 h	0.1%, substitutio‡	$690‡	$150
	50 (PE)	180–200 Gb	4 B (PE)*	2.5 d			$58
	100 (PE)	360–400 Gb		5 d			$45
	125 (PE)*	450–500 Gb*		6 d*			$30
Illumina HiSeq3000/4000	50 (SE)	105–125 Gb	2.5 B (SE)*	1–3.5 d*	0.1%, substitution	$740/$900	$50
	75 (PE)	325–375 Gb					$31
	150 (PE)*	650–750 Gb*					$22

*Manufacturer's data.

Approx., approximate; AT, adenine and thymine; B, billion; bp, base pairs; d, days; Gb, gigabase pairs; h, hours; indel, insertions and deletions; Kb, kilobase pairs; M, million; Mb, megabase pairs; NA, not available; PE, paired-end sequencing; SBS, sequencing by synthesis; **SE, single-end sequencing**; Tb, terabase pairs.

Different platforms for NGS

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by synthesis: SNA							
454 GS Junior	*Up to 600; 400 average (SE, PE)**	*35 Mb**	*~0.1 M**	*10 h**	*1%, indel*	*NA*	*$40,000*
454 GS Junior+	*Up to 1,000; 700 average (SE, PE)**	*70 Mb**	*~0.1 M**	*18 h**	*1%, indel*	*$108,000*	*$19,500*
454 GS FLX Titanium XLR70	*Up to 600; 450 mode (SE, PE)**	*450 Mb**	*~1 M**	*10 h**	*1%, indel*	*NA*	*$15,500*
454 GS FLX Titanium XL+	*Up to 1,000; 700 mode (SE, PE)**	*700 Mb**	*~1 M**	*23 h**	*1%, indel*	*$450,000*	*$9,500*
Ion PGM 314	*200 (SE)*	*30–50*	*400,000–550,000**	*23 h*	*1%, indel*	*$49*	*$25–3,500*
	400 (SE)	*60–100 Mb**		*3.7 h**			
Ion PGM 316	*200 (SE)*	*300–500 Mb*	*2–3 M**	*3 h*	*1%, indel*	*$49*	*$700–1,000*
	*400 (SE)**	*600 Mb–1 Gb**		*4.9 h**			
Ion PGM 318	*200 (SE)*	*600 Mb–1 Gb*	*4–5.5 M**	*4 h*	*1%, indel*	*$49*	*$450–800*
	*400 (SE)**	*1–2 Gb**		*7.3 h**			
Ion Proton	*Up to 200 (SE)*	*Up to 10 Gb**	*60–80 M**	*2–4 h**	*1%, indel*	*$224*	*$80*
Ion S5 520	*200 (SE)*	*600 Mb–1 Gb*	*3–5 M**	*2.5 h*	*1%, indel*	*$65*	*$2,400**
	*400 (SE)**	*1.2–2 Gb**		*4 h**			*$1,200**
Ion S5 530	*200 (SE)*	*3–4 Gb*	*15–20 M**	*2.5 h*	*1%, indel*	*$65*	*$950**
	*400 (SE)**	*6–8 Gb**		*4 h**			*$475**
Ion S5 540	*200 (SE)**	*10–15 Gb**	*60–80 M**	*2.5 h**	*1%, indel*	*$65*	*$300**

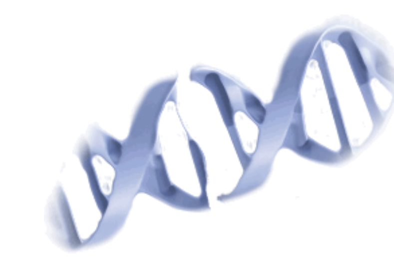
Different platforms for NGS

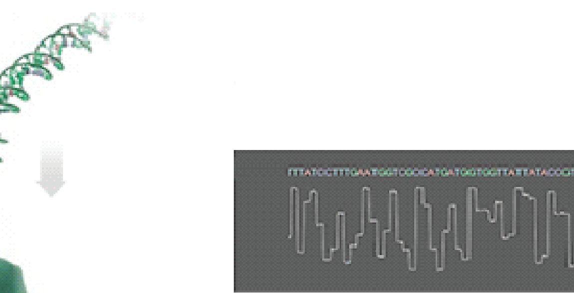
Single-molecule real-time long reads

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Pacific BioSciences RS II	*~20 Kb*	*500 Mb–1 Gb**	*~55,000**	*4 h**	*13% single pass, ≤1% circular consensus read, indel*	*$695*	*$1,000*
Pacific Biosciences Sequel	*8–12 Kb69*	*3.5–7 Gb**	*~350,000**	*0.5–6 h**	*NA*	*$350*	*NA*
Oxford Nanopore MK 1 MinION	*Up to 200 Kb159*	*Up to 1.5 Gb159*	*>100,000 (Ref. 159)*	*Up to 48 h160*	*~12%, indel159*	*$1,000**	*$750**
Oxford Nanopore PromethION	*NA*	*Up to 4 Tb**	*NA*	*NA*	*NA*	*$75**	*NA*

Nature Reviews Genetics 17, 333–351 (2016)

Nanopore technologies



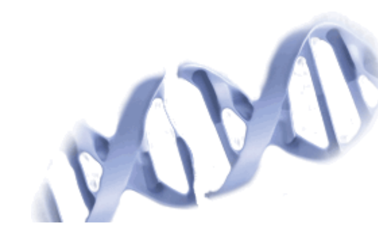


MinION GridION PromethION

Oxford Nanopore Technologies

https://nanoporetech.com/how-it-works

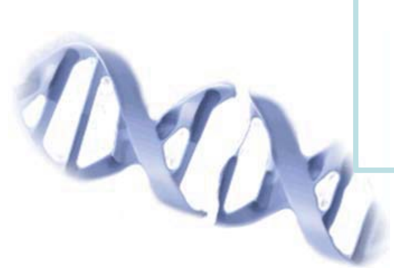
GS Junior System



First NGS system 454

Discontinued , as Roche is investing in nanopore technologies and 454 could not compete with other platforms

Roche; www.my454.com



Sequencing by Expansion Technology (SBX)

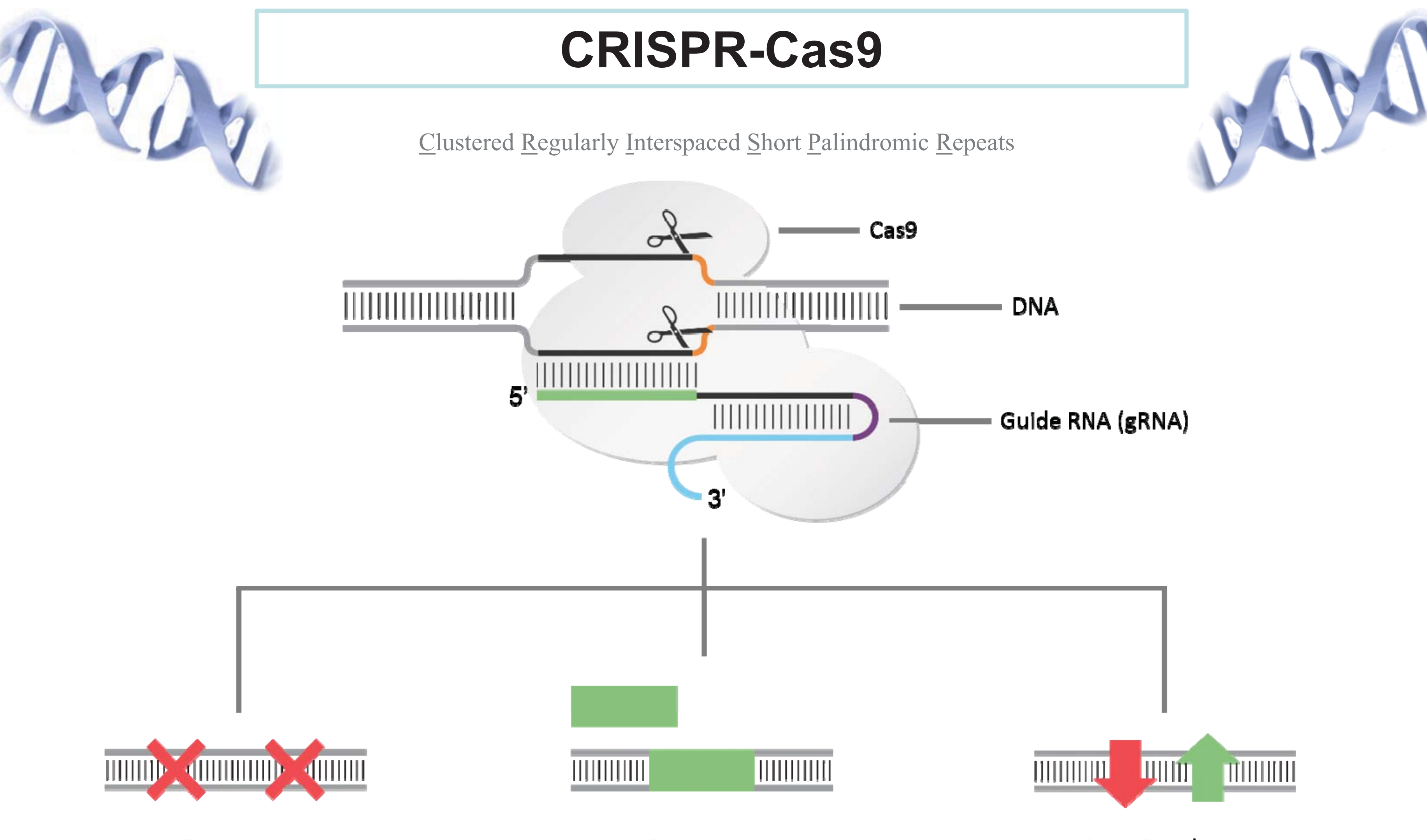
- *Sequencing by Expansion technology (SBX™) represents a new platform for a low cost and more rapid method for whole genome sequencing. In a nutshell, SBXTM "is an efficient, low-cost DNA preparation method that rescales a DNA target into a longer surrogate polymer,'' according to the company's recent press release.*
- *Sequencing By Xpansion (SBX) is a DNA sequencing method that uses a simple biochemical reaction to encode the sequence of a DNA molecule into a highly measurable surrogate called an Xpandome- nanopore technology*
- *This single molecule approach produces enough Xpandomer in a single drop reaction to sequence an entire human genome 1000X over.*
- ________ *probably cheapest technology ?*

Stratos Genomics inc

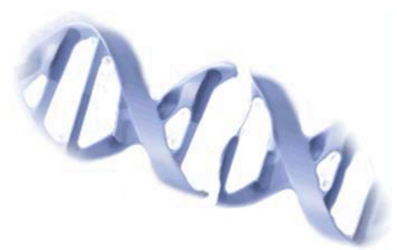
More demand, more competition, less price- but analysis , data storage, retrieval, sharing ????

CRISPR-Cas9

Clustered Regularly Interspaced Short Palindromic Repeats



"Technologies come and go. You can't get married to one,"
"You need to always think about what biological problems need to be solved."-
Patrick Hsu , Bioengineer , Salk Institute



Further Reading

- *Sara Goodwin,John D. McPherson& W. Richard Mc. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17, 333–351 (2016)*
- *Sara A Byron et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nature Reviews Genetics 17, 257–271 (2016)*
- *Kathleen M Giacomini et al. Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. Nature Reviews Drug Discovery 16, 70 (2017)*
- *William S Bush et al. Genome-Wide Association Studies. https://doi.org/10.1371/journal.pcbi.1002822*
- *Heidi ledford. CRISPR: gene editing is just the beginning. http://www.nature.com/news/crispr-gene-editing-is-just-the-beginning-1.19510 (Not covered in this lecture but this technology might transform the therapeutics in near future)*