Molecular and Genetic Epidemiology I

Lecture III:

Mendel's Laws and Molecular Genetic Measures

Jan 21, 2014

Joe Wiemels

HD 274 (Mission Bay)

514-0577

joe.wiemels@ucsf.edu

Lecture

Genetic Variation and Mendel's Laws

Types of genetic markers: SNPs and microsatellites

Assessing genetic markers

1. PCR (polymerase chain reaction) and DNA amplification methods

2. Detection of mutations and polymorphisms: low and high throughput techniques

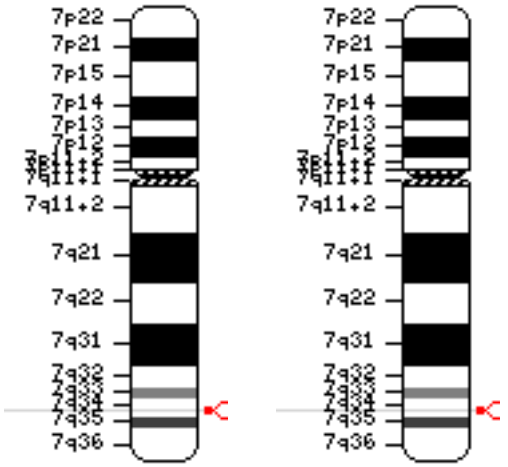
Microarray techniques: SNPs, gene expression, DNA methylation.

Next Generation Sequencing methods

Epigenetics

Genes and alleles

Each cell has two chromosomes, therefore there are two physical copies of each gene. The position of a gene is called a **locus,** and the exact form of the gene is called an **allele**. Each gene can exist in the form of two alleles



Chromosome 7

1378 genes on chromosome 7,

159,000,000 base pairs.

TAS2R38 gene (PTC TASTE RECEPTOR)

Mendel's Laws

I. Independent segregation: one allele from each parent is randomly and independently selected for the offspring

Bi-allelic loci:

Aa♂ X Aa♀ leads to 3 possible genotypes

AA, Aa, and aa. What are the offspring ratios?

Mendel's Laws

II. The alleles underlying two or more different traits are transmitted independently of each other. (independent assortment)

Note: this does not apply to traits that are on the same chromosome, which are physically and genetically "linked"

mode of inheritance/penetrance

Dominant $P(Y=1|DD) = P(Y=1|Dd) = 1;\ P(Y=1|dd) = 0$

Recessive $P(Y=1|DD) = 1;\ P(Y=1|Dd) = P(Y=1|dd) = 0$

Additive intermediate penetrance (3 phenotypes)

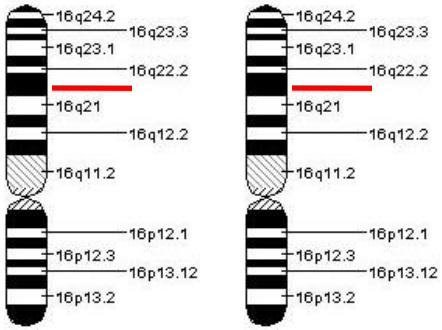
Codominant both alleles represented (ex. ABO)

P – probability, Y – expression of a trait, D and d – alleles

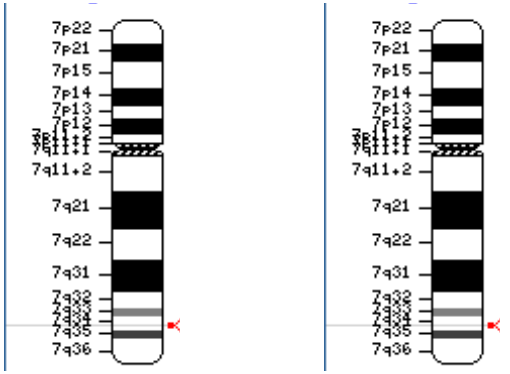
Cerumen phenotype
PTC taster phenotype

ABCC11: chromosome 16

TAS2R38: chromosome 7



ABCC11

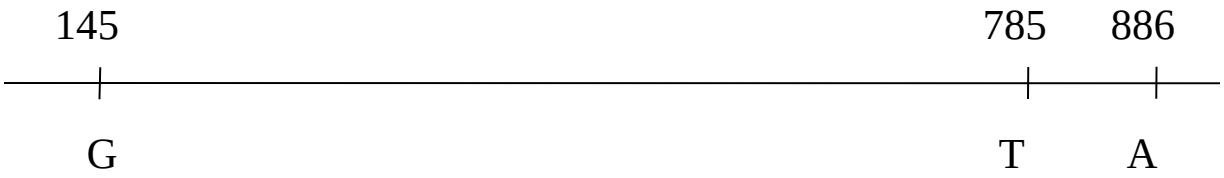


TAS2R38

TAS2R38 – 3 SNP loci

Taster haplotype

Non-taster haplotype



Nucleotide Position	TASTER		NONTASTER	
	Codon	Amino Acid	Codon	Amino Acid
145	CCA	proline	GCA	alanine
785	GCT	alanine	GTT	valine
886	GTC	valine	ATC	isoleucine

Haplotype: a set of linked genotypes, carried together

ABCC11 SNP: rs17822931

Query SNP: rs17822931 and variants with r^2 >= 0.8

chr	pos (hg19)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	GENCODE genes	dbSNP func annot
16	48258198	1	1	rs17822931	C	T	0.01	0.12	0.90	0.14				Adult_CD4_Th0			CTCF	ABCC11	missense
16	48276823	0.86	0.97	rs111915588	G	A	0.08	0.13	0.93	0.13				GM12864			17 altered motifs	ABCC11	
16	48375777	0.92	-0.98	rs6500380	A	G	0.99	0.87	0.08	0.86							Mef2	LONP2	intronic
16	48382522	0.92	-0.98	rs8052293	A	G	0.99	0.87	0.08	0.86							Smad3	LONP2	intronic
16	48438484	0.83	-0.91	rs1982782	A	G	0.90	0.85	0.08	0.85							8 altered motifs	SIAH1	

HAPLOREG – summary of functional consequence of SNPs

http://www.broadinstitute.org/mammals/haploreg

A disease has a genetic component, what do you do now?

No idea of the gene: whole genome scan of genetic markers: SNPs or microsatellites, or whole genome sequencing

Fair idea of the gene: candidate gene SNPs at “medium throughput”

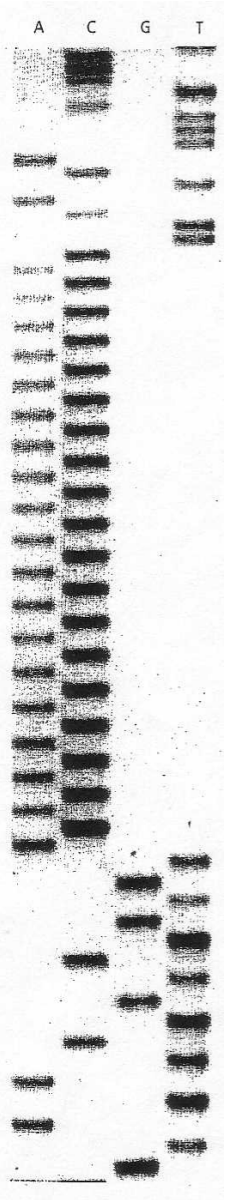
(You know what the gene is, but no idea of the genetic alteration: DNA sequencing and functional genomics)

The study of DNA polymorphisms as markers

If a marker is distributed non-randomly with a disease, it must be “linked” to the disease gene.

We can study markers through family lineages (linkage studies) and/or by associating with markers with diseased individuals in populations (association studies).

Comparison of *affected* vs *unaffected*



Microsatellite (aka STS, sequence tagged site): highly polymorphic DNA sequence feature (not functionally polymorphic).

A simple repeat sequence that invites slippage-mispair during replication, and hence many polymorphic variations in size in the population.

DNA sequence, showing alternating “ACACACAC”

6-20 or more alleles, so nearly everyone is heterozygote

Rarely associated with disease, typically used as a marker only (1000s in genome)

Single Nucleotide Polymorphism

For usefulness as a genetic marker, it should be common (>5% allele frequency)

Only two variants, so much less information, as a marker, per test than a microsatellite

Whole genome disease scan requires far more tests than microsatellite, but each test is far less expensive

13 million in genome

How do we test for genetic variants?

Many Genetic Analyses begin with PCR

Polymerase Chain Reaction (PCR) – specific amplification of a single gene sequence

2 synthetic oligonucleotides can “find” their complementary DNA sequences among 3 billion nucleotide sequence.

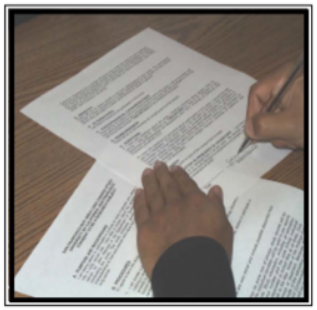
Able to faithfully amplify a specific sequence 10^{30} times.

Buccal cell collection

SAN FRANCISCO BAY AREA LUNG CANCER STUDY

STEPS FOR TAKING A CHEEK CELL SAMPLE

1. Please read, then sign and date the "Consent To Be A Research Subject" form. The blank copy of the form is for your future reference.



2. Rinse out your mouth with water.

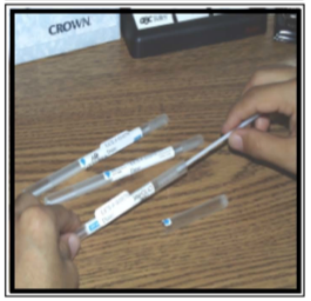
3. Pull the cap off the first tube by twisting at the blue arrow. Remove the brush from the tube.

(over, please)

C:\Lung\Forms\Misc\BuccalInstructions.doc

4. Insert the brush into your mouth and place it firmly against the middle of your cheek. Twirl the brush around for 30 seconds. This will collect the cells we need to study, and should not cause you any pain. You will not be able to see the cells.

5. Reinsert the brush into the tube and replace the cap tightly.



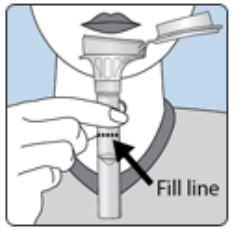
6. Repeat steps 3 to 5 using a new brush on your other cheek. Repeat these steps again using the last new brush on either cheek.

7. Put the three tubes and the signed consent form into the return envelope. Seal the envelope and place it in the mail today.

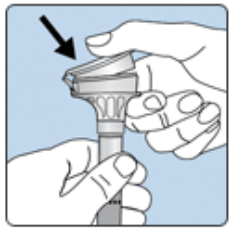
If you have any questions, please call us on our toll-free number (888) 452-3535. Thank you for helping us learn more about cancer.

Saliva collection for DNA

Oragene•DISCOVER (OGR-500)

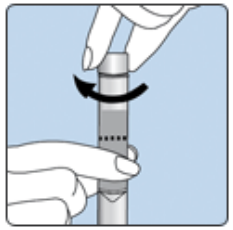


1. Spit into funnel until the amount of liquid saliva (not bubbles) reaches the fill line shown in picture #1.



2. Hold the tube upright with one hand. Close the funnel lid with the other hand (as shown) by firmly pushing the lid until you hear a loud click. The liquid in the lid will be released into the tube to mix with the saliva. Make sure that the lid is closed tightly.

3. Hold the tube upright. Unscrew the funnel from the tube.



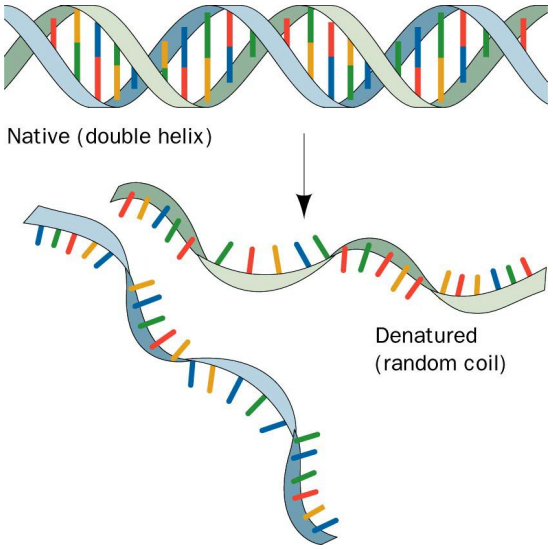
4. Use the small cap to close the tube tightly.

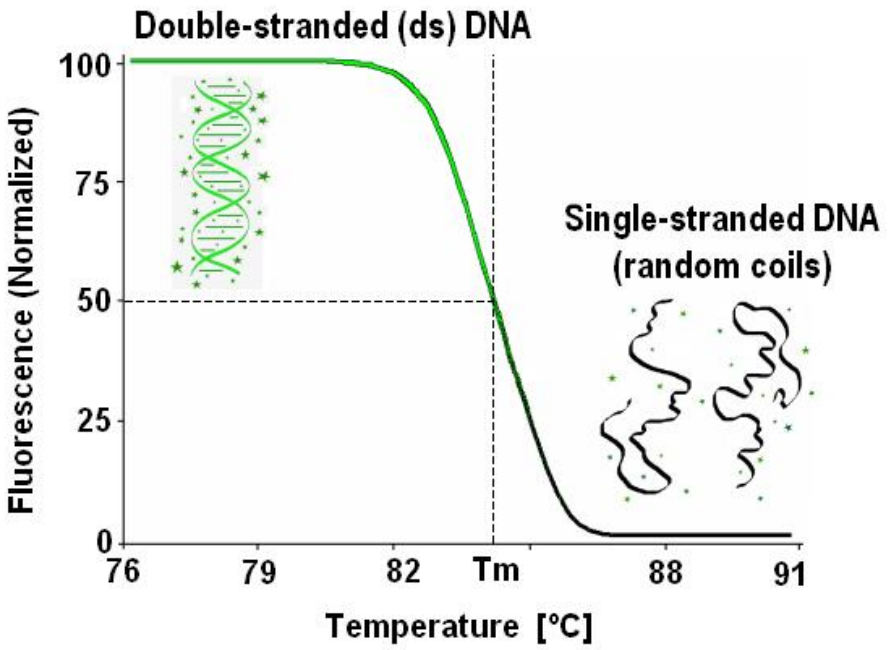
5. Shake the capped tube for 5 seconds. Discard or recycle the funnel.

Genotyping in MGE – TICR Individuals

1. Get genomic DNA from subject (buccal cell demonstration in class)
2. Isolate DNA on Autogen 3000 →
    1. Lyses cells with detergent and digests protein with Proteinase K
    2. Removes protein with Phenol
    3. Concentrates DNA using ethanol precipitation, rehydrates DNA in buffered water.

Basis of all nucleic acid techniques: DNA hybridization





Long DNA melts
around 75-85 degrees C

Tm calculation

Melting temperature of DNA dependent on:

- *length of oligonucleotide*
- *content of A,C,G,T*
- *salt content of solution*

Hybridization to specific sequence

A 15 base pair sequence would be unique in a random genome of 3 billion bases.

Hybridization is *specific* around the Tm.

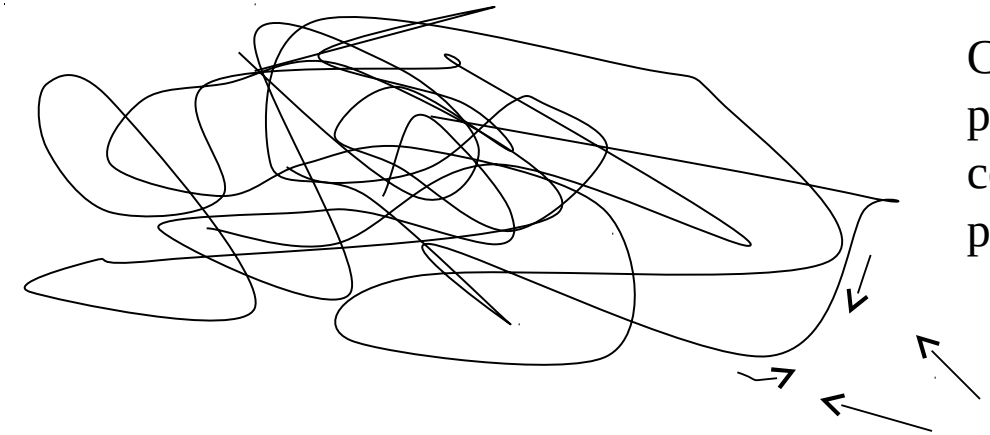
Nearly all genetic applications are dependent on this feature of DNA. The sizes of nucleotides will be adjusted for specificity and efficiency at a specific temperature.

PCR: 17-35 base pairs
Microarrays: 25-80 base pairs

Genotyping in MGE - TICR Individuals *(continued)*

Purified genomic DNA will be amplified in the region of the polymorphisms, then a “readout” performed

PCR amplification is a standard method, but there are many methods to “read” the polymorphism



Cellular DNA is 3 X 10^9 base pairs (haploid), but only a one copies of the gene of interest per genome

Two PCR primers (oligonucleotides) will be able to make billions of copies of one small segment, crowding out the rest of the genomic DNA

PCR design for one of the *TAS2R38* polymorphism

```
     LEFT PRIMER      length 20  Tm 59.56
5'-CCCACATTAAAGCCCTCAAG-3'
   RIGHT PRIMER       length 20  Tm 59.15
5'-AGGGACAAGCTGCCATTATC-3'
   PRODUCT SIZE: 150

PCR product:

5'-
CCCACATTAAAGCCCTCAAG-3'
CCCACATTAAAGCCCTCAAGTCTCTTGTCTCCTTTTTCTGCT
TCTTTGTGATATCATCCTGTGYTGCCTTCATCTCTGTGCCCC
TACTGATTCTGTGGCGCGACAAAATAGGGGTGATGGTTTGTG
TTGGGATAATGGCAGCTTGTCCCT-3'
 3'-CTATTACCGTCGAACAGGGA-5'
```

DNA polymerase synthesizes DNA directionally (5' to 3')

PCR protocol:

10 ng of DNA mixed with

10 pmoles each PCR primer

1 pmoles each probe

2.5 umoles each dNTP

Reaction buffer (salts including $MgCl_2$)

Taq polymerase (thermostable DNA polymerase)

The temperature of the mixture is cycled 35 times:

60 degrees 30 seconds

72 degrees 30 seconds

94 degrees 15 seconds

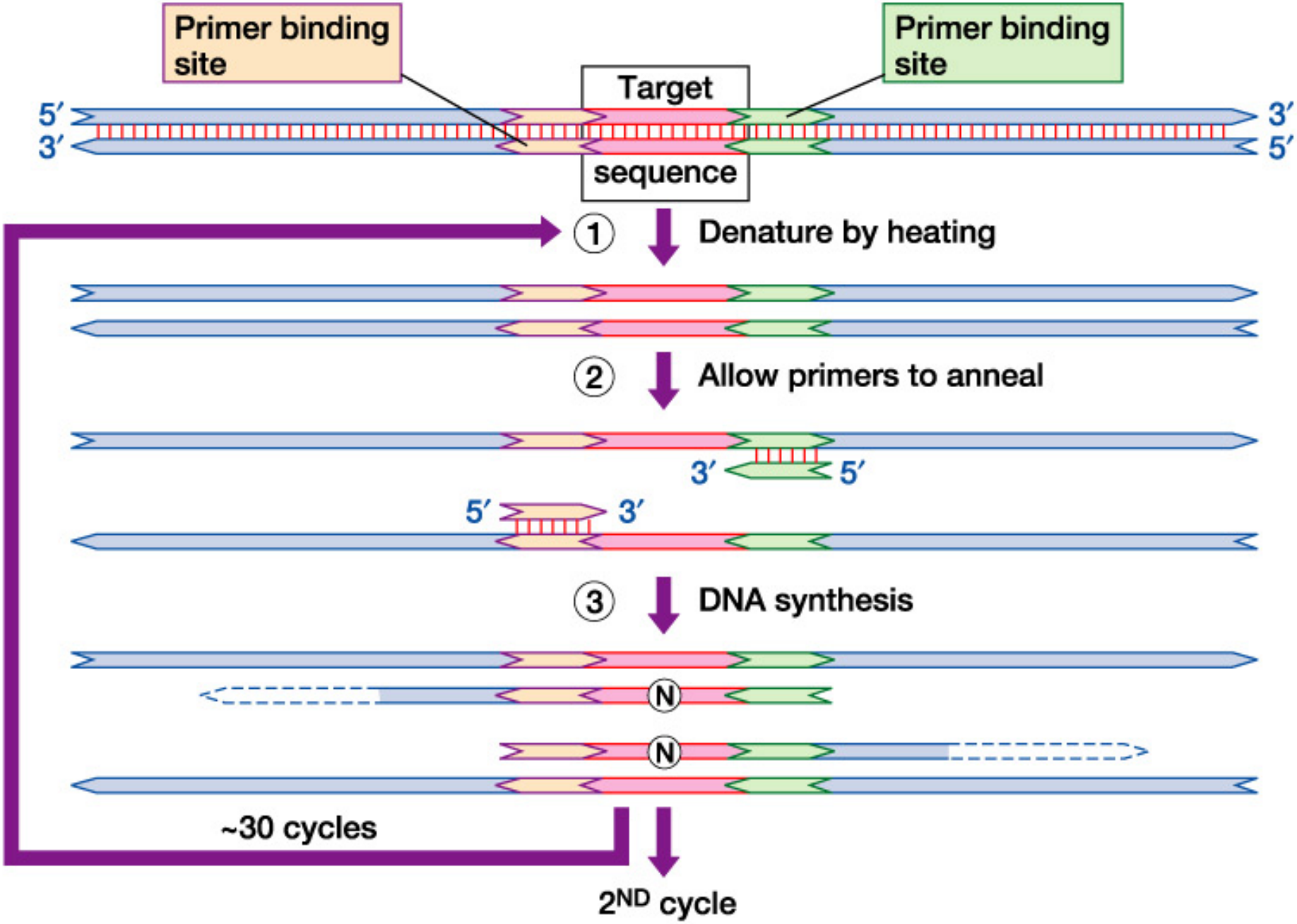


Figure 5-2 part 1 of 3 Human Molecular Genetics, 3/e. (© Garland Science 2004)

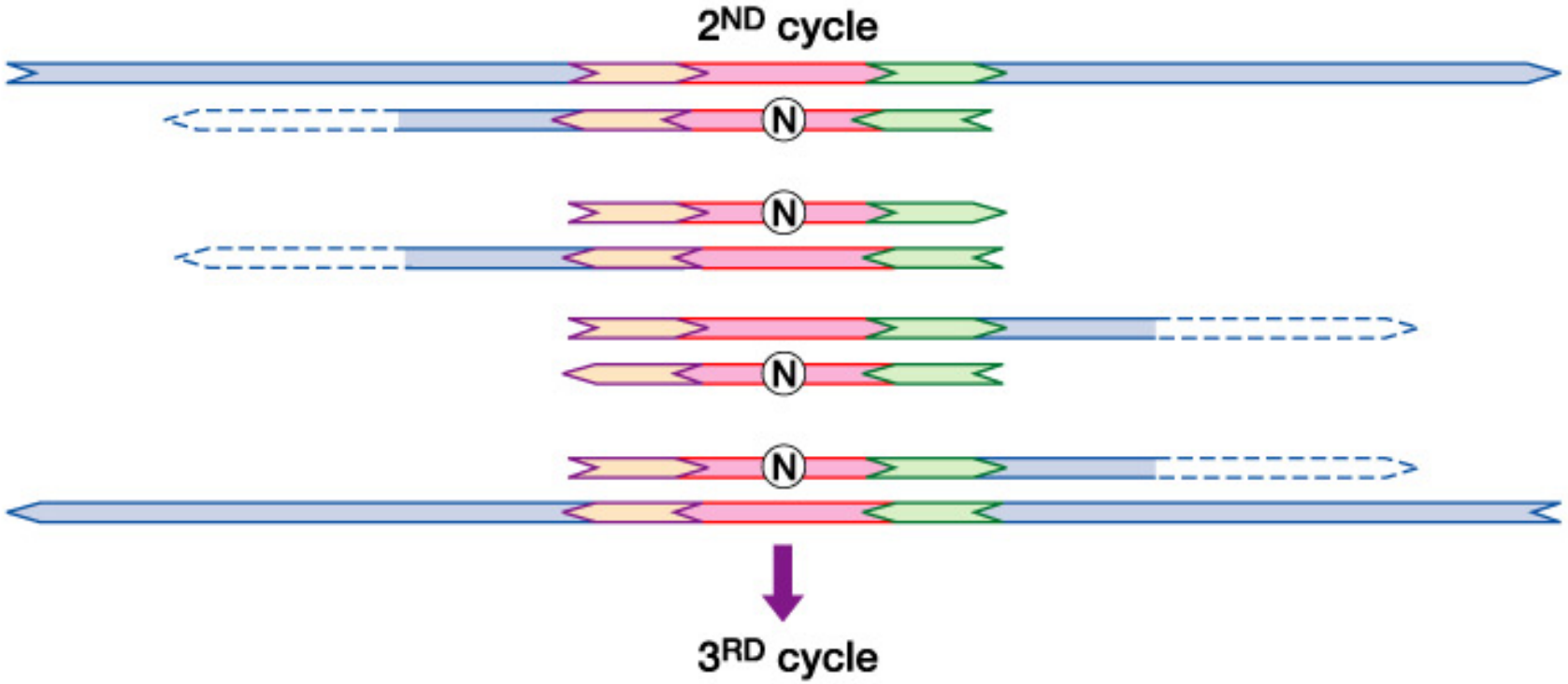


Figure 5-2 part 2 of 3 Human Molecular Genetics, 3/e. (© Garland Science 2004)

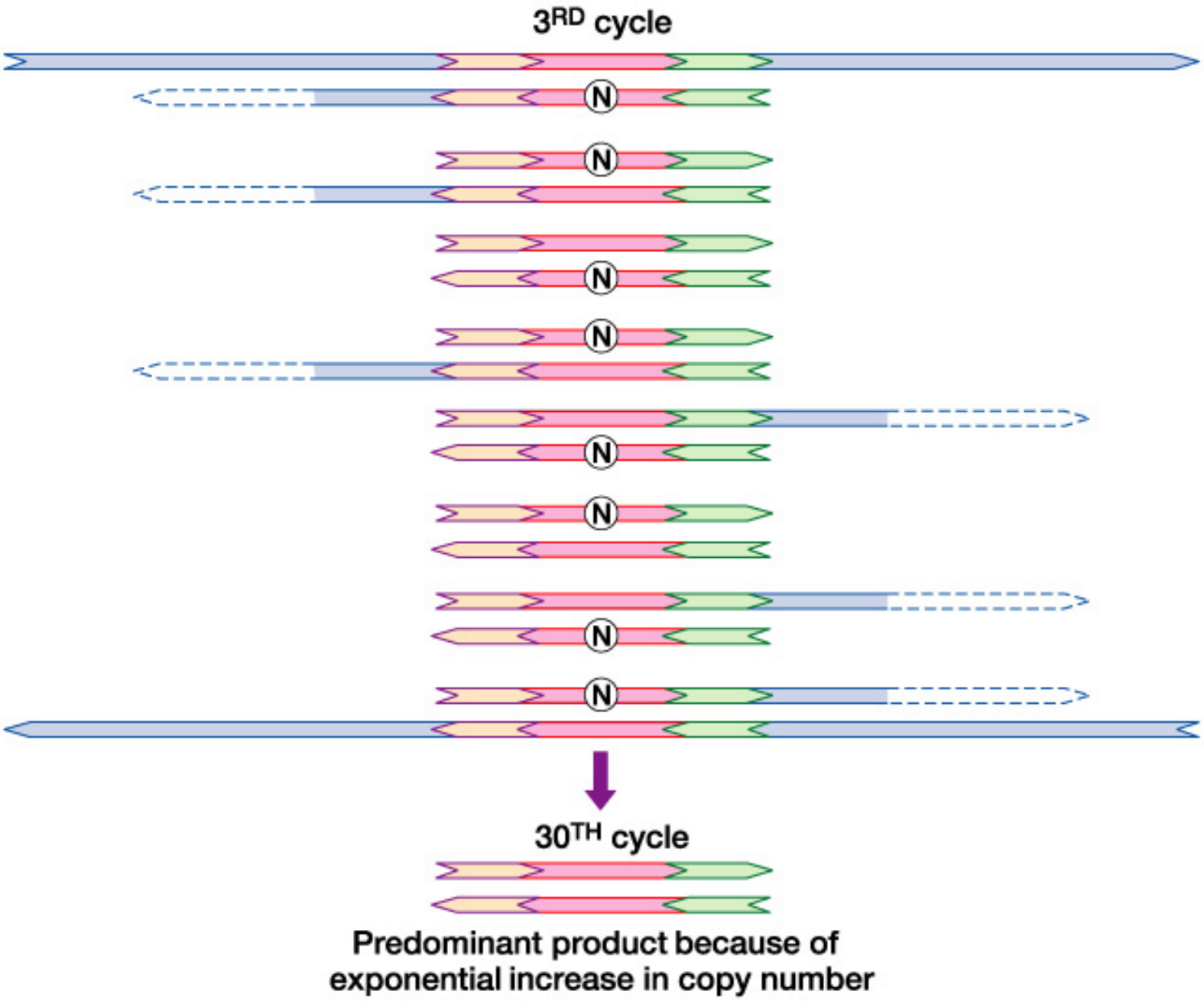
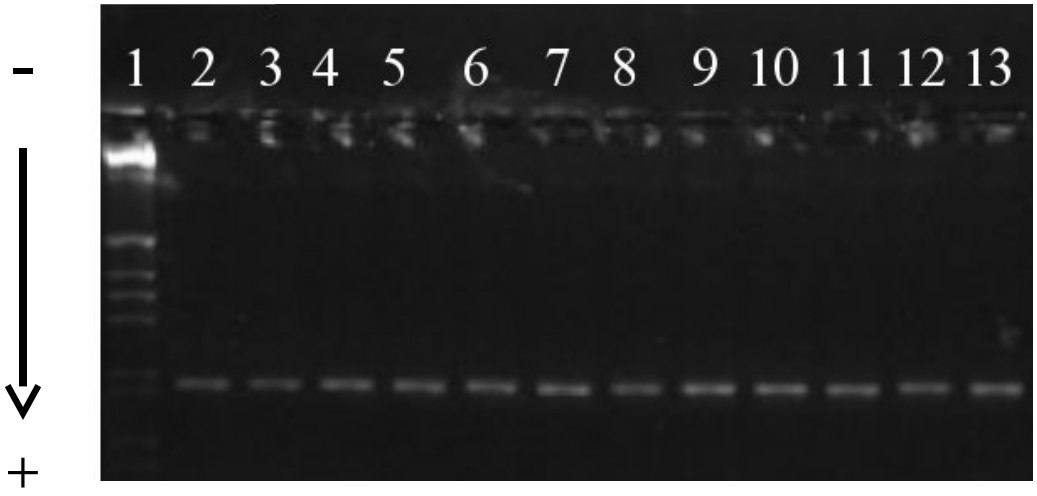


Figure 5-2 part 3 of 3 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Detection of PCR products using Electrophoresis gel.

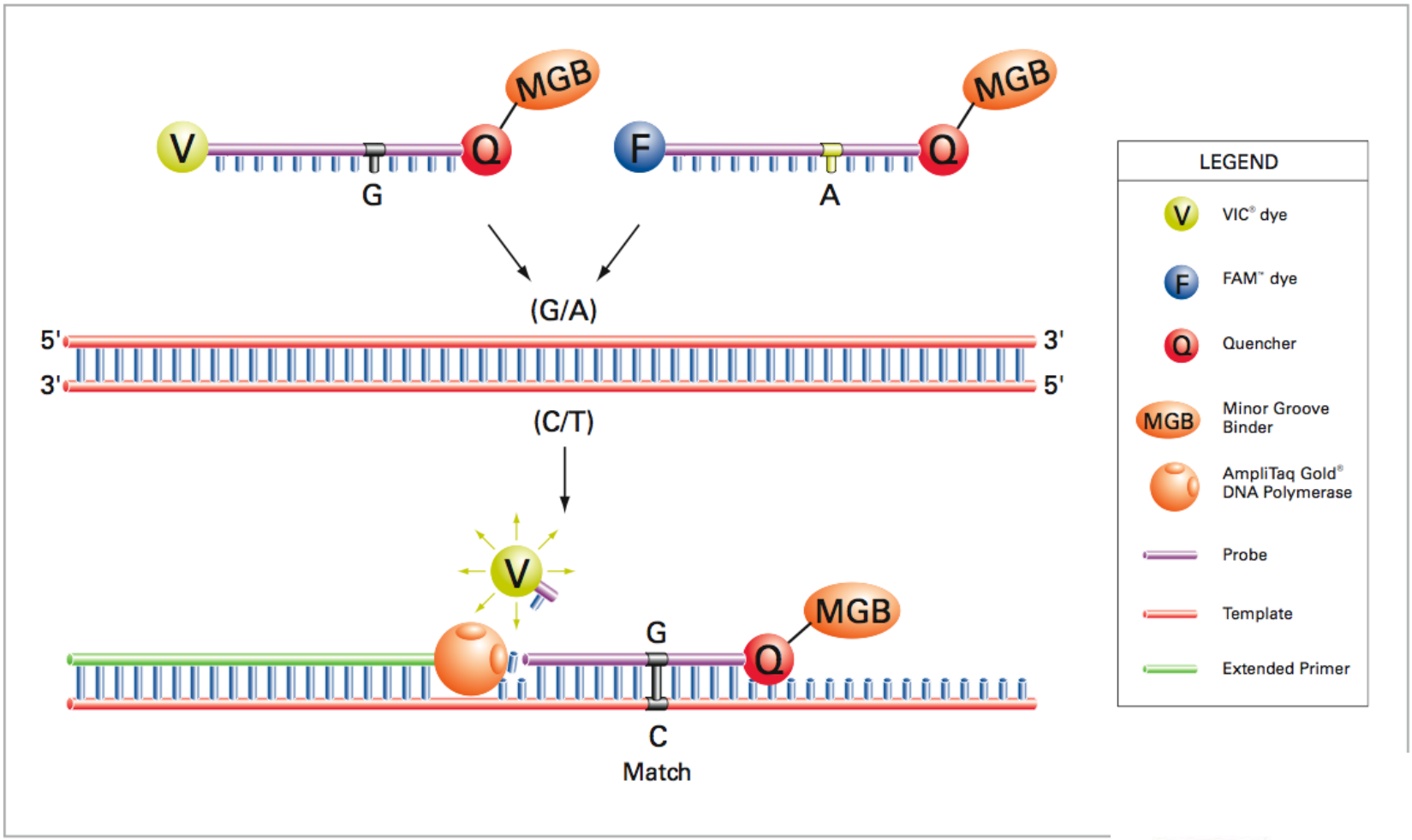


Individual lane markers

PCR product

PCR products for a SNP are all the same size;
this “gel” is not diagnostic for the SNPs

Taqman allelic discrimination genotyping (for *taste receptor TASR32*)



There are four oligonucleotides in the reaction mix -- two PCR primers and two “probes” each labeled different color and each matching different SNP allele.

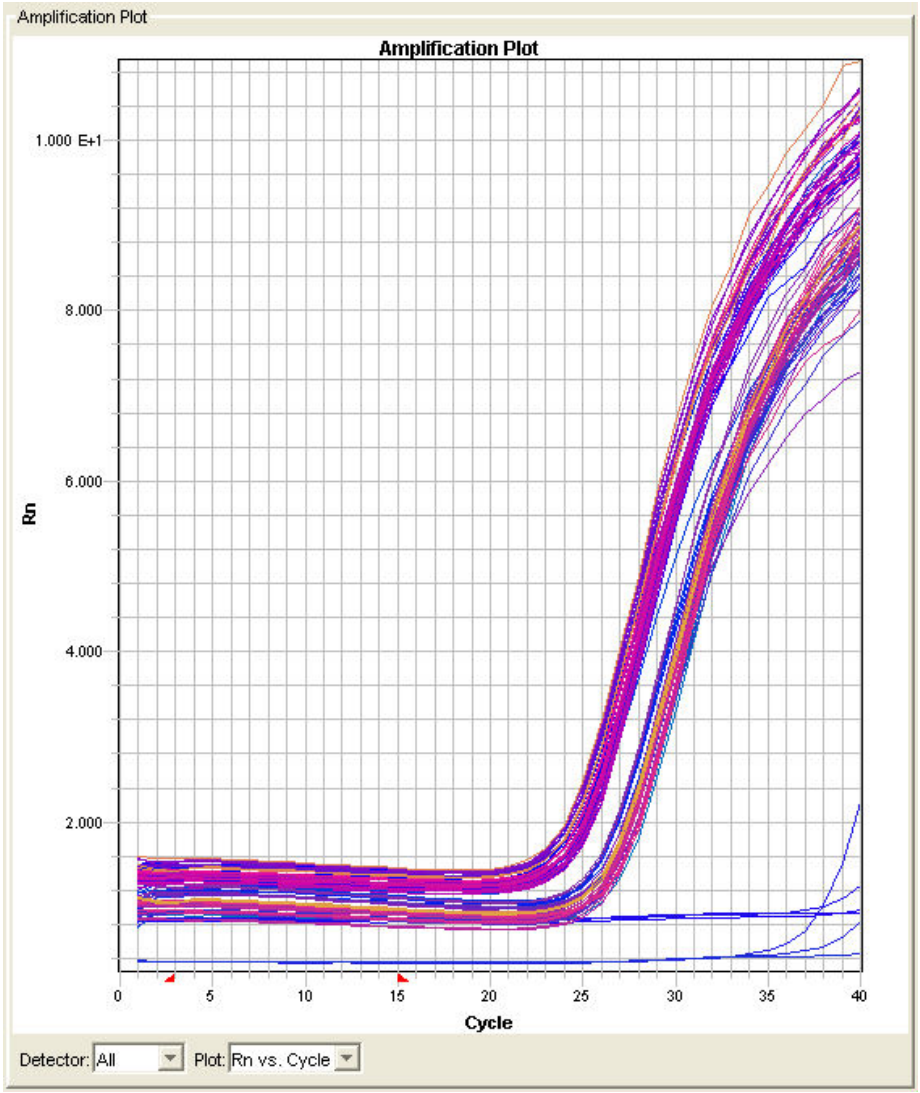
PCR design for *TAS2R38* polymorphism

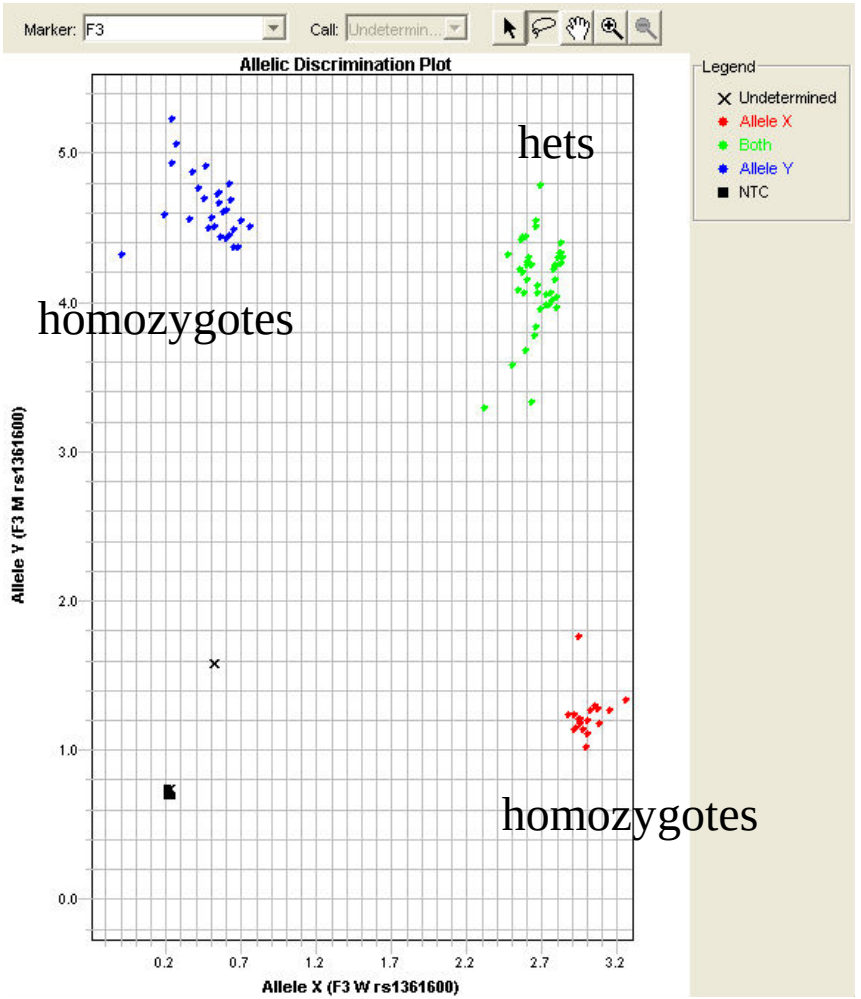
```
     LEFT PRIMER      length 20  Tm 59.56
5'-CCCACATTAAAGCCCTCAAG-3'
   RIGHT PRIMER       length 20  Tm 59.15
5'-AGGGACAAGCTGCCATTATC-3'
   PRODUCT SIZE: 150

PCR product:

5'-
CCCACATTAAAGCCCTCAAG-3'
CCCACATTAAAGCCCTCAAGTCTCTTGTCTCCTTTTTCTGCT
TCTTTGTGATATCATCCTGTGYTGCCTTCATCTCTGTGCCCC
TACTGATTCTGTGGCGCGACAAAATAGGGGTGATGGTTTGTG
TTGGGATAATGGCAGCTTGTCCCT-3'
 3'-CTATTACCGTCGAACAGGGA-5'
```

```
probes for Taqman analysis
TCCTGTGCTGCCTTC-FAM
TCCTGTGTTGCCTTC-VIC
```

These probes are used to diagnose the SNP.

Taqman Genotyping - Real-time PCR (Q-PCR)





Many genotyping platforms available today

Taqman genotyping: Low throughput (1-10)

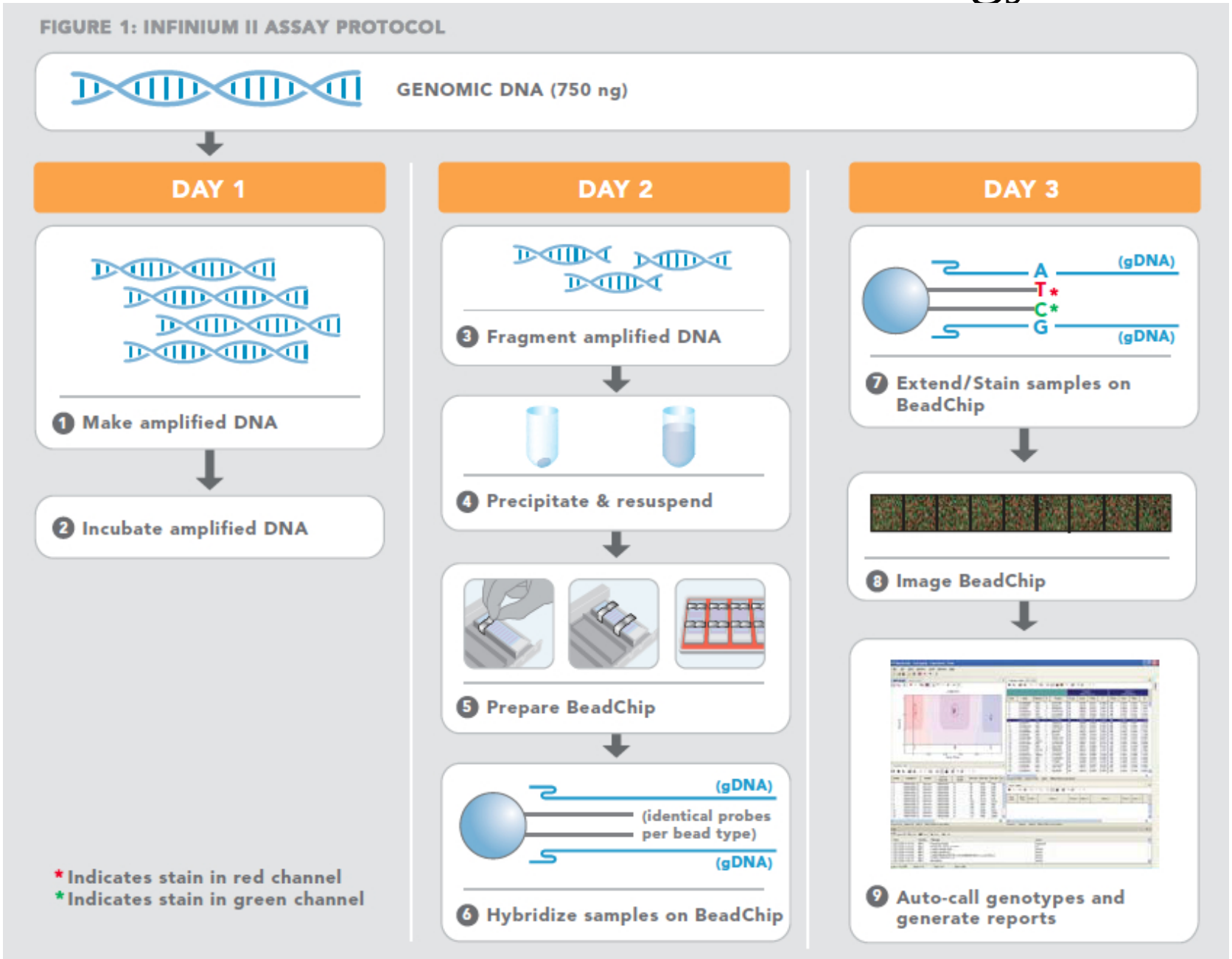
Sequenom (10-50), Illumina GoldenGate: medium

Massive parallel genotyping: High throughput, useful for whole genome scans:

- Affymetrix (Kaiser RPEG)
- Illumina

“deep” or “next generation” sequencing: Illumina , Applied Biosystems Solid, 454 (Roche), Ion Torrent

Illumina Infinium technology



Illumina prices

Arrays and reagents. UCSF core costs around $45 per sample

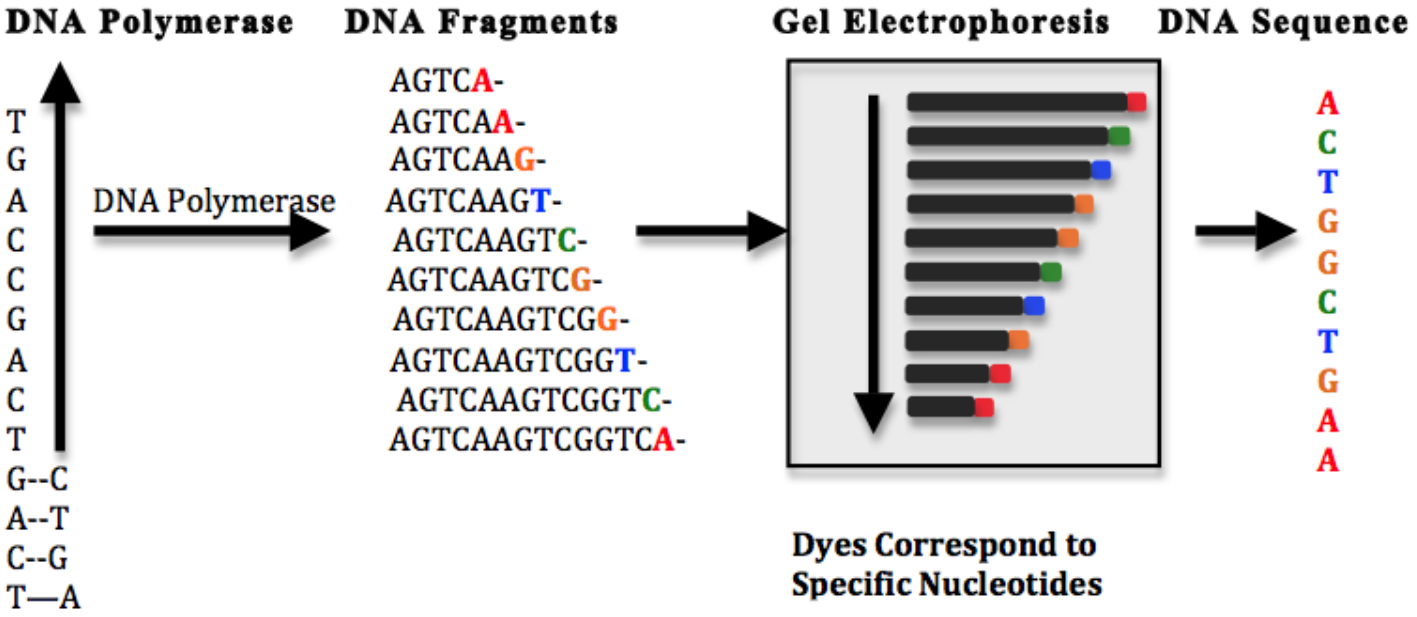
	Array product			Custom content	
	OncoArray	HumanCore	HumanOmniExpress	50,000	100,000
Price/Sample	$65	$45	$83	$70	$140

Price for full service – just send DNA to Illumina

	FastTrack Service		
	OncoArray	HumanCore	HumanOmniExpress
Price/Sample	N/A	$79	$151

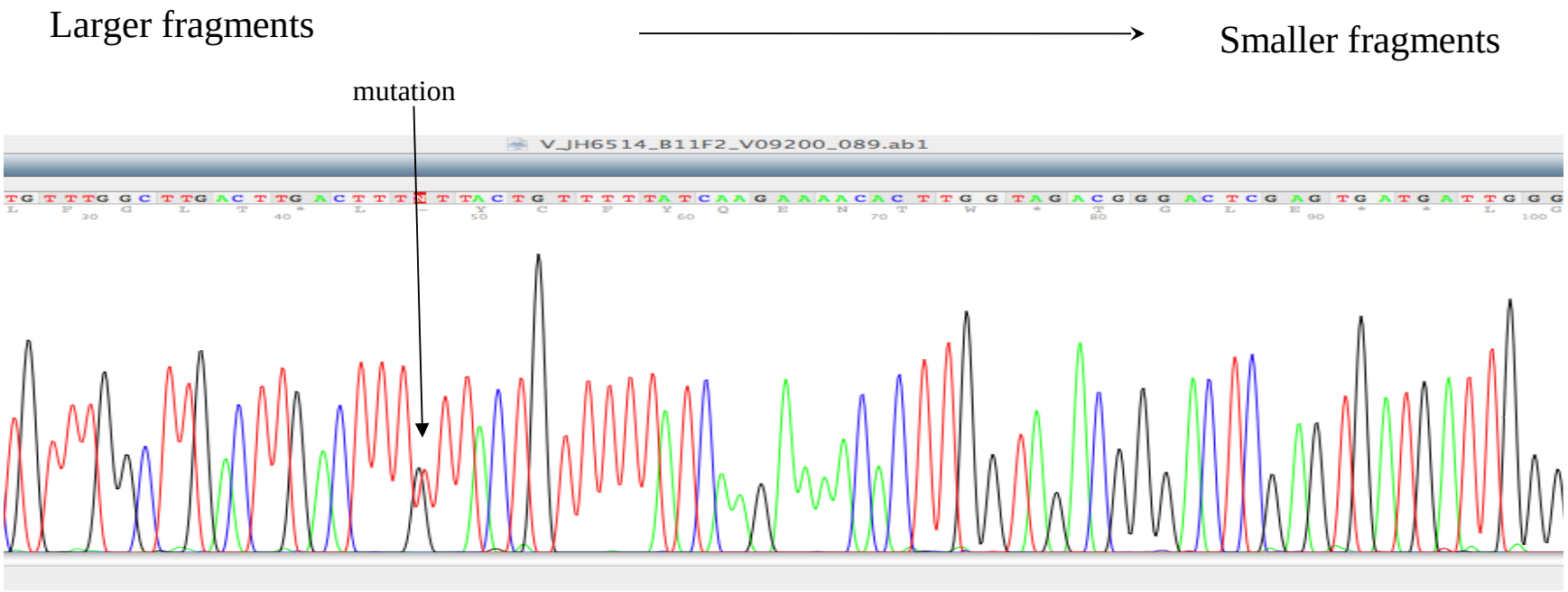
DNA sequencing: the method to obtain the genotype of a new mutation/polymorphism

Prior to sequencing, one first amplifies a sequence by PCR or cloning in a bacterial vector. Then, using ONE primer, adds fluorescent labeled dideoxy chain terminators and DNA polymerase. ddNTPs will "cap" the sequence.



DNA sequencing

The products of the sequencing reaction are separated on a gel mixture that can separate fragments by one base pair.



Useful when you suspect a gene, but don't know the variant. This one is *BRAF* gene in leukemia

Microarray basics

All nucleic acid microarray experiments involve four steps:

1. Labeling sample (fluorescent or chemiluminescent)

2. Hybridization of a sample to immobilized nucleic acid probe

3. Scanning using a high definition optical device

4. Conversion of the raw image to data, followed by normalization steps and subsequent analysis and interpretation.

Microarray basics

Some Applications for Microarray:

1. SNP genotyping (eg Affymetrix, Illumina)
2. Gene expression patterns - comparing one tissue to another (Affymetrix, Superarray, etc)
3. Gene deletion or amplification: arrayCGH (for cancer applications, Albertson and Pinkel, UCSF)
4. microRNA (UCSF Gladstone, Ambion)
5. Pathogen identification (DeRisi, UCSF)
6. DNA methylation

Types of Microarrays

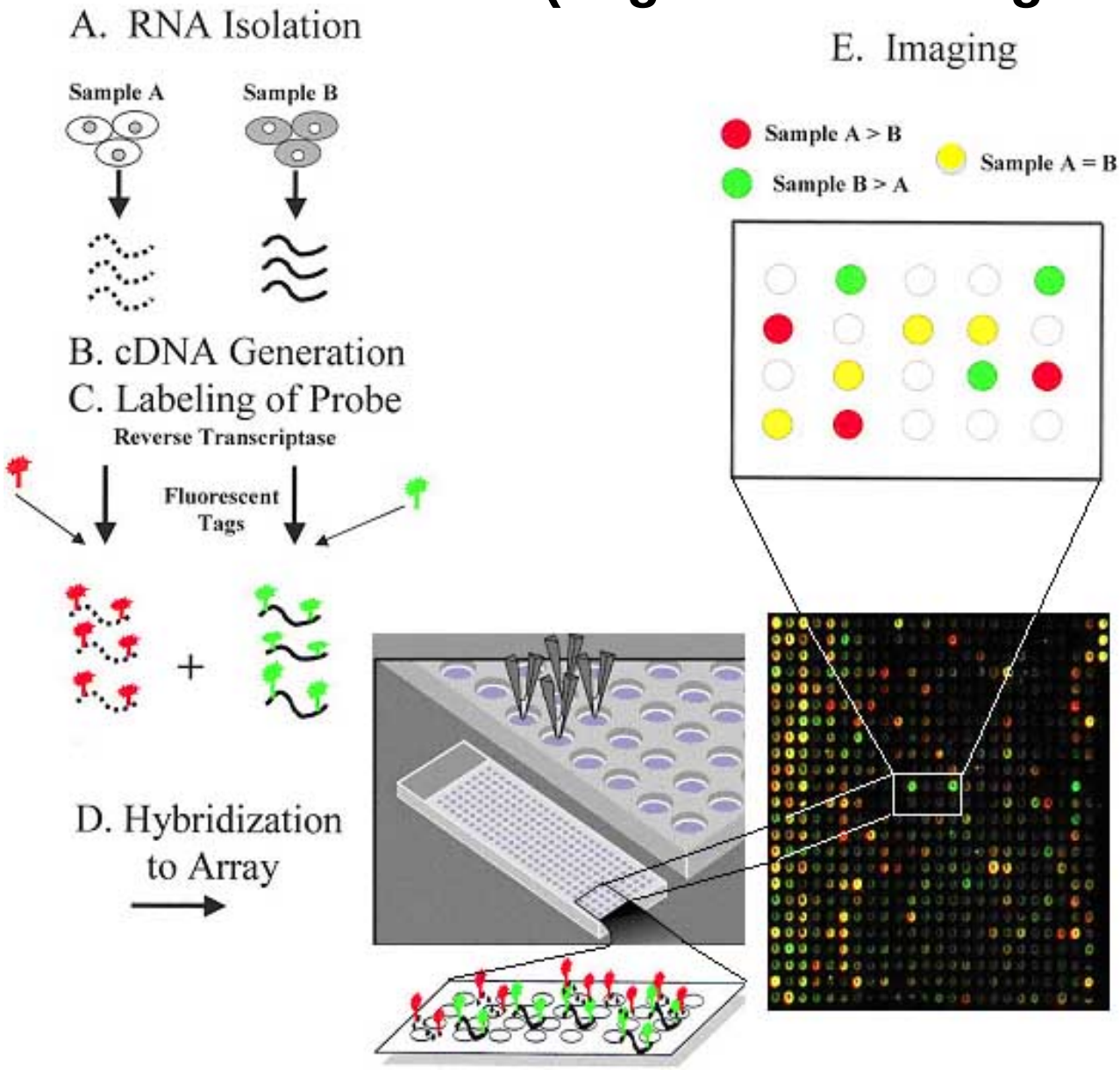
Spotted (early technology)

- cDNA (for expression, 100s - 1000s bases)
- oligonucleotide (less than 100 bp)

Chemically synthesized oligonucleotides (Affymetrix [Gladstone, Inst Hum Genet],Illumina [Inst Hum Genet, Langley Porter], NimbleGen, Agilent[Sandler Rock Hall])

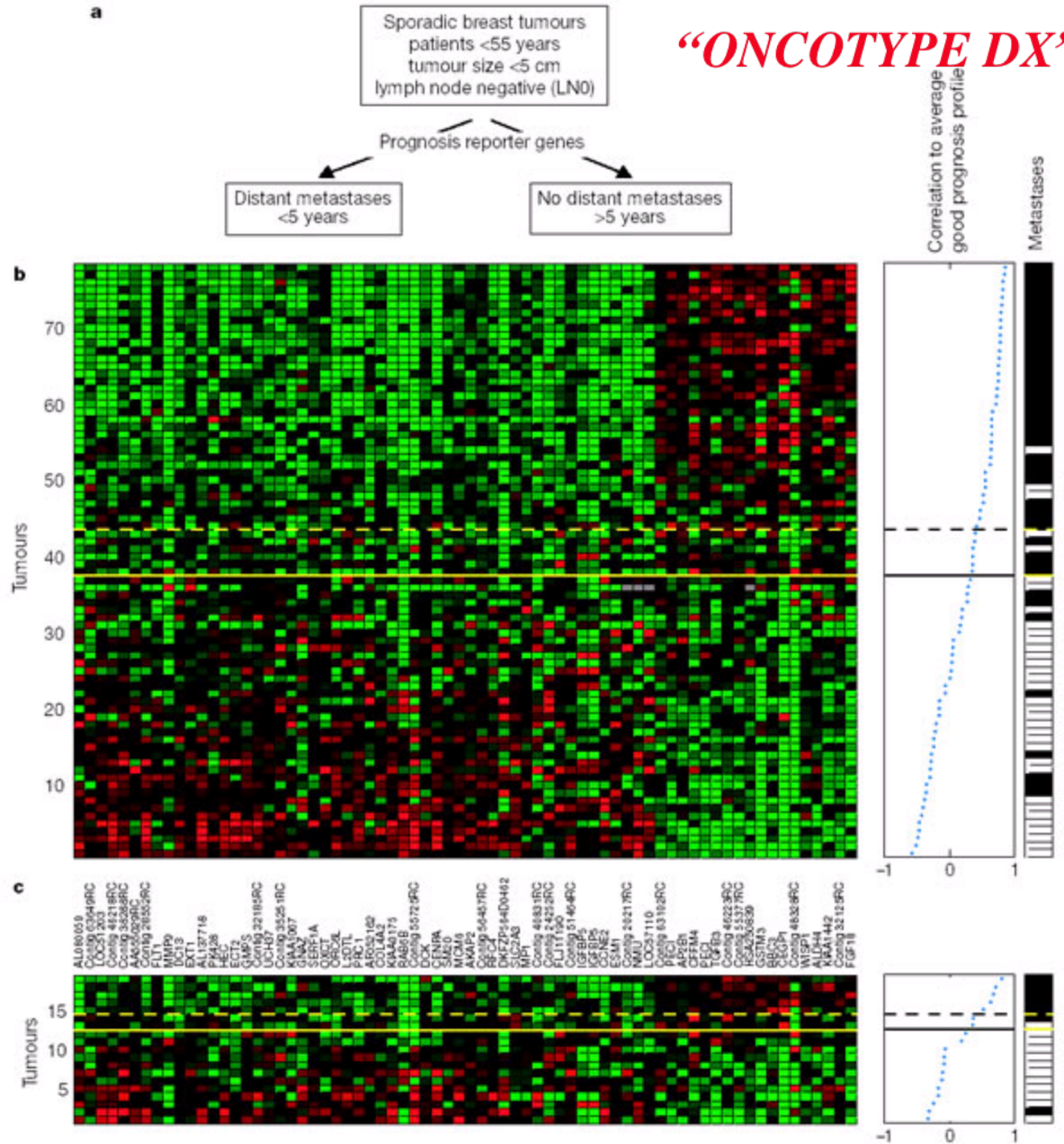
- expression
- gene resequencing
- SNP genotyping
- array-based CGH

Spotted microarray for gene expression (oligos or cloned genes)



The microrarray may have immobilized oligonucleotides (eg., virochip, UCSF) or cloned genes

Gene Expression of Breast Cancer predicts disease-free outsome (Nature 2002 Friend et al)

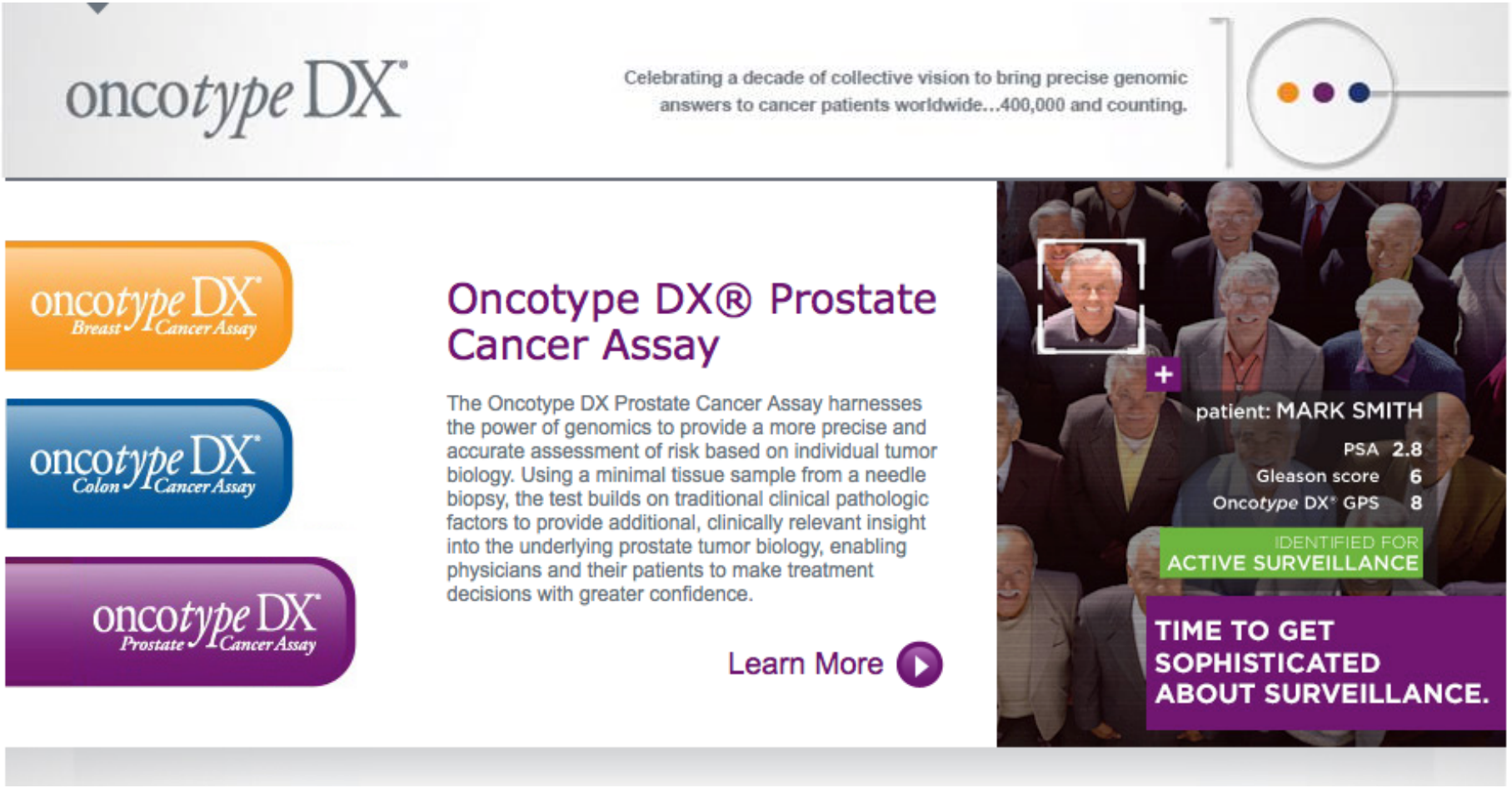


"ONCOTYPE DX"

Figure 2 Supervised classification on prognosis signatures. a, Use of prognostic reporter genes to identify optimally two types of disease outcome from 78 sporadic breast tumours into a poor prognosis and good prognosis group (for patient data see Supplementary Information Table S1). b, Expression data matrix of 70 prognostic marker genes from tumours of 78 breast cancer patients (left panel). Each row represents a tumour and each column a gene, whose name is labelled between b and c. Genes are ordered according to their correlation coefficient with the two prognostic groups. Tumours are ordered by the correlation to the average profile of the good prognosis group (middle panel). Solid line, prognostic classifier with optimal accuracy; dashed line, with optimized sensitivity. Above the dashed line patients have a good prognosis signature, below the dashed line the prognosis signature is poor. The metastasis status for each patient is shown in the right panel: white indicates patients who developed distant metastases within 5 years after the primary diagnosis; black indicates patients who continued to be disease-free for at least 5 years. c, Same as for b, but the expression data matrix is for tumours of 19 additional breast cancer patients using the same 70 optimal prognostic marker genes. Thresholds in the classifier (solid and dashed line) are the same as b. (See Fig. 1 for colour scheme.)

Agilent arrays

Gene expression used as a clinical assay



Deep sequencing

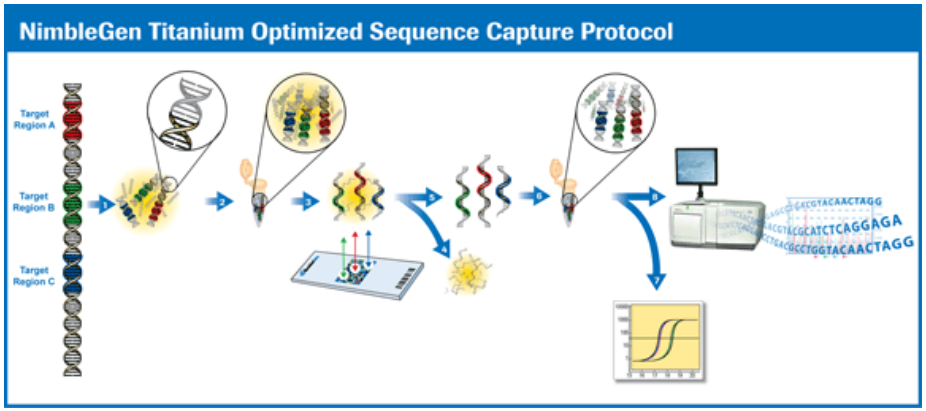
Deep or “next generation” sequencing involves three steps

1.Creation of a **library** (a mixture of DNA that may be a whole genome, or part of a genome from many individuals, or a whole “transcriptome” etc)

2. Parallel sequencing of all the DNA molecules in a single reaction.

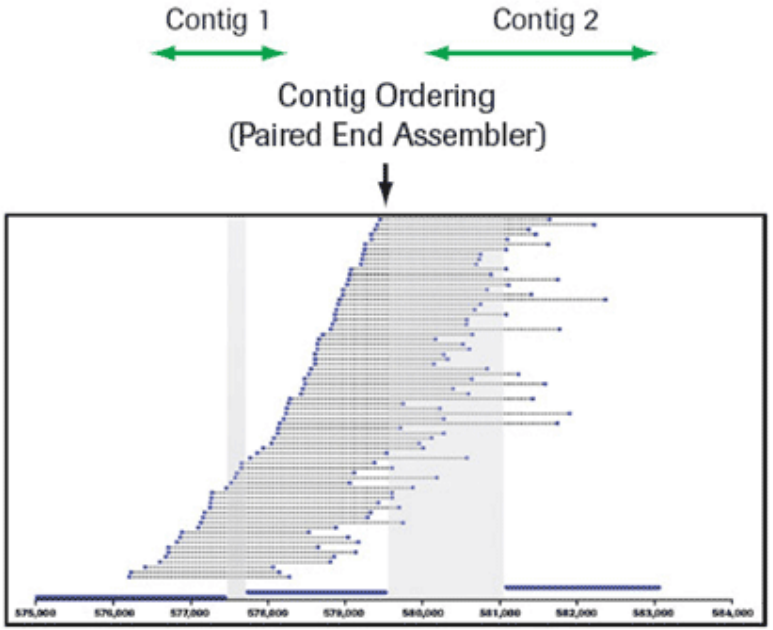
Applications of Next Generation Sequencing

Sequence capture protocol: resequence a region hundreds of times

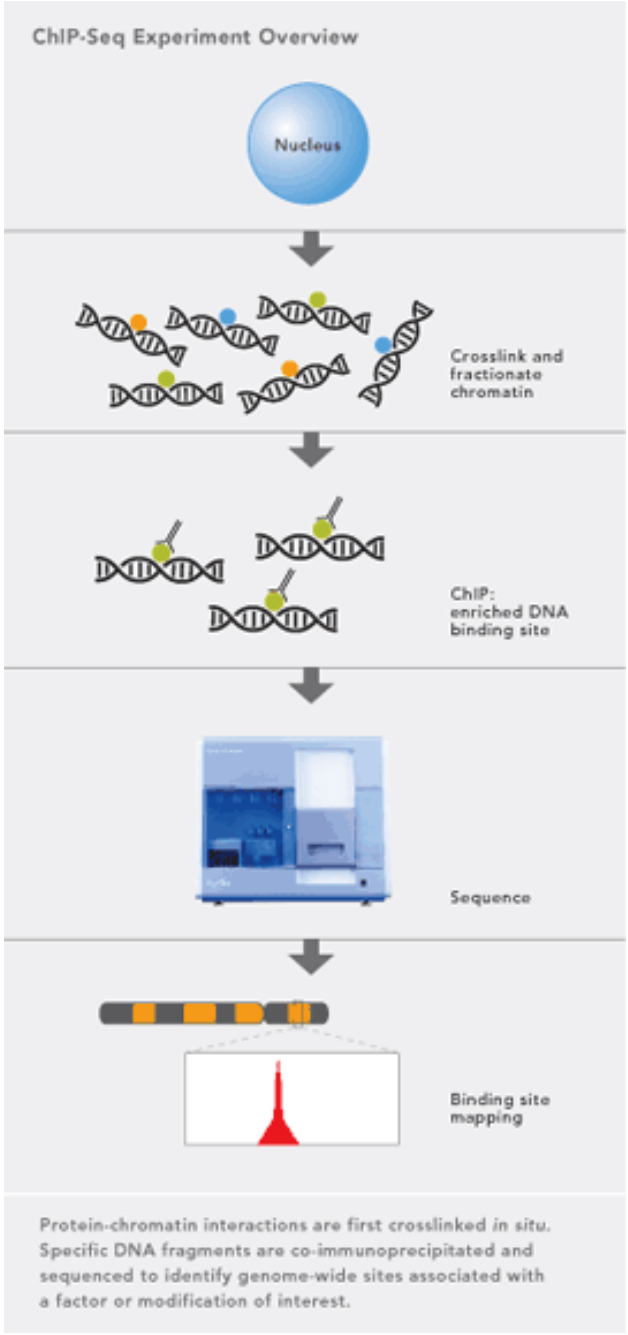


One can use DNA pools to completely describe variation in a region

Whole genome shotgun sequencing



Cut a single genome to bits, sequence the pieces, and use a bioinformatic program to “reassemble” the genome

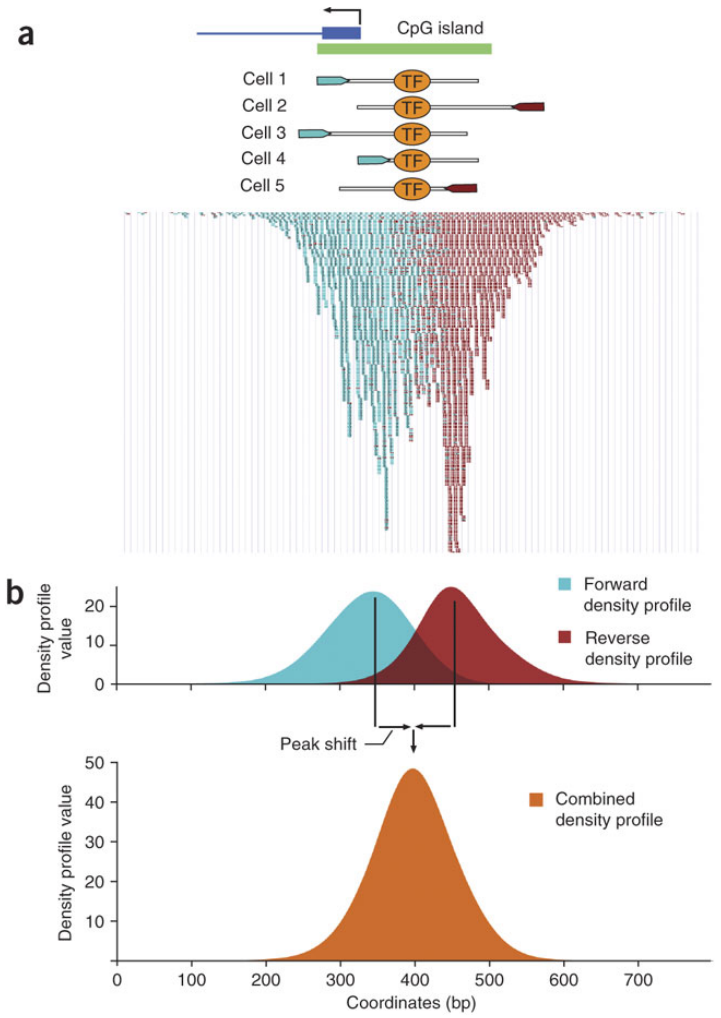


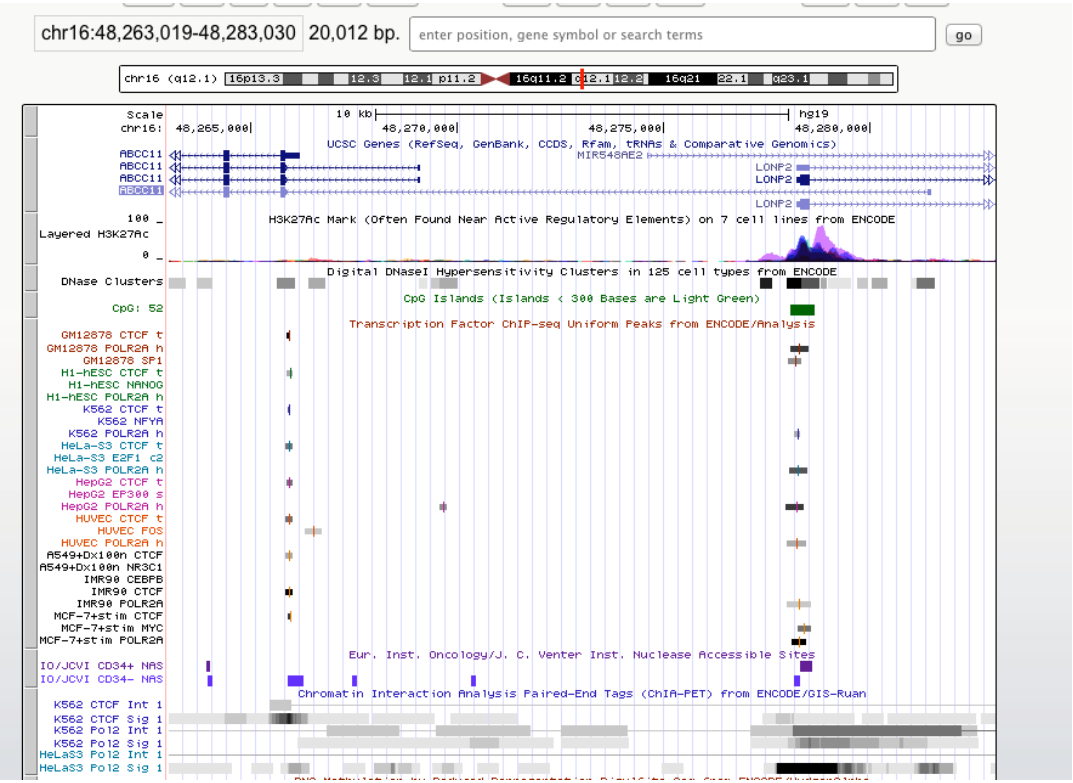
Applications of Next Generation Sequencing

ChIP-Seq: *Chromatin Immunoprecipitation sequencing*. Use an antibody to "pull-down" a DNA binding protein. Sequence the DNA fragments that are attached to the protein.

RNA-Seq, DNA-Seq, Sequence capture-Seq (ChIP-Seq etc), pathogen discovery, etc.

ChIP – Seq for annotation of genome

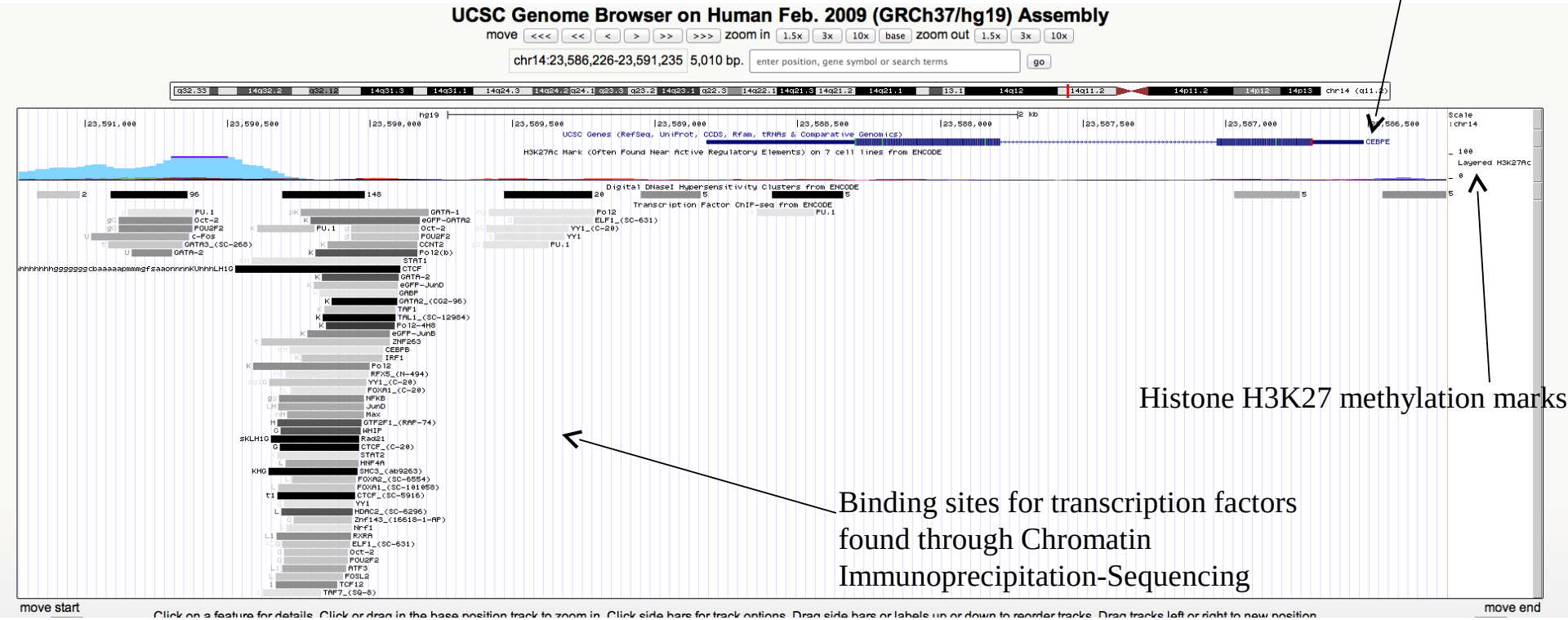




ENCODE data on epigenetic regulation

Encyclopedia of DNA elements (ENCODE)

A gene involved in leukemia, *CEBPE*



Epigenetics

the study of heritable changes in gene expression or cellular phenotype caused by mechanisms other than changes in the underlying DNA sequence

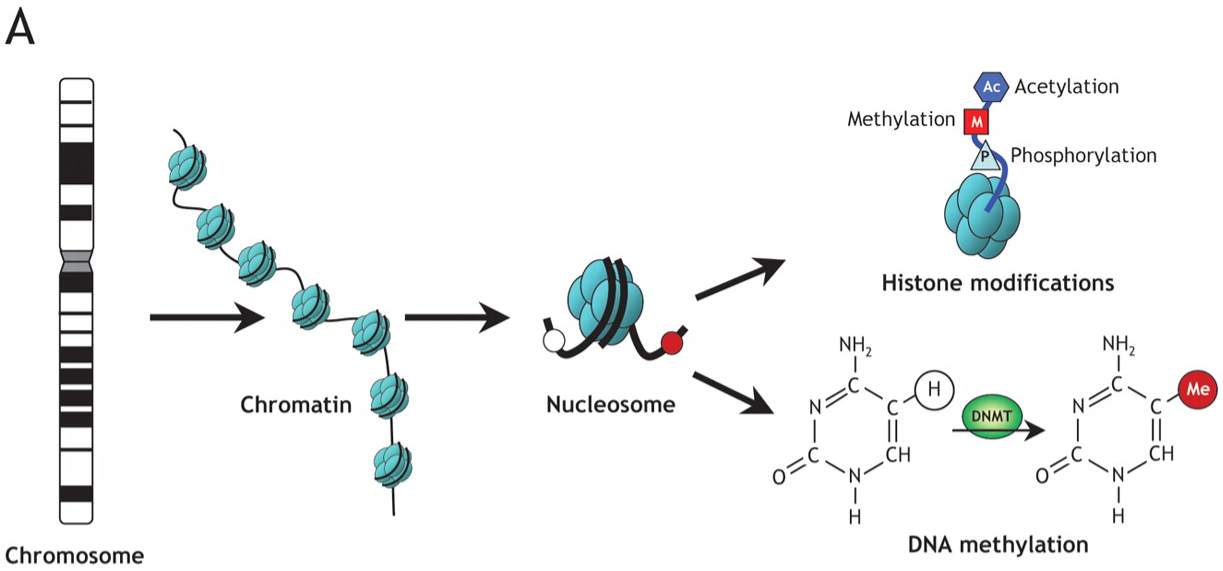
DNA methylation, histone modifications, accessible chromatin, chromosome structure

Epigenetics

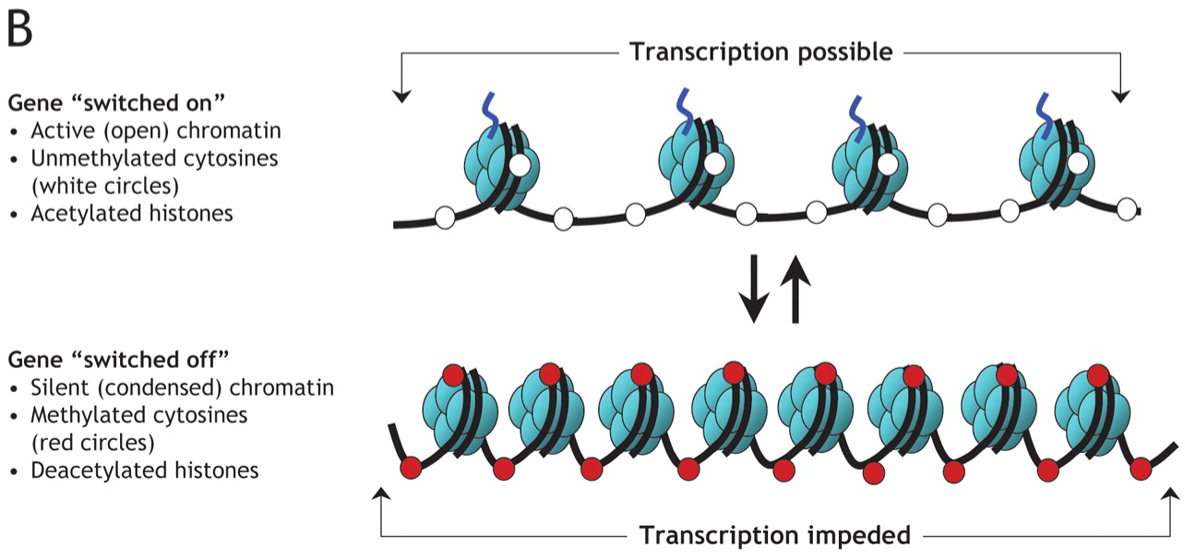
"Your DNA is not your destiny"

- DNA is the same in all body cells
- DNA is packaged and marked differently in different cell types
- This marking is "on top" of the genes, or "epi-genetic"

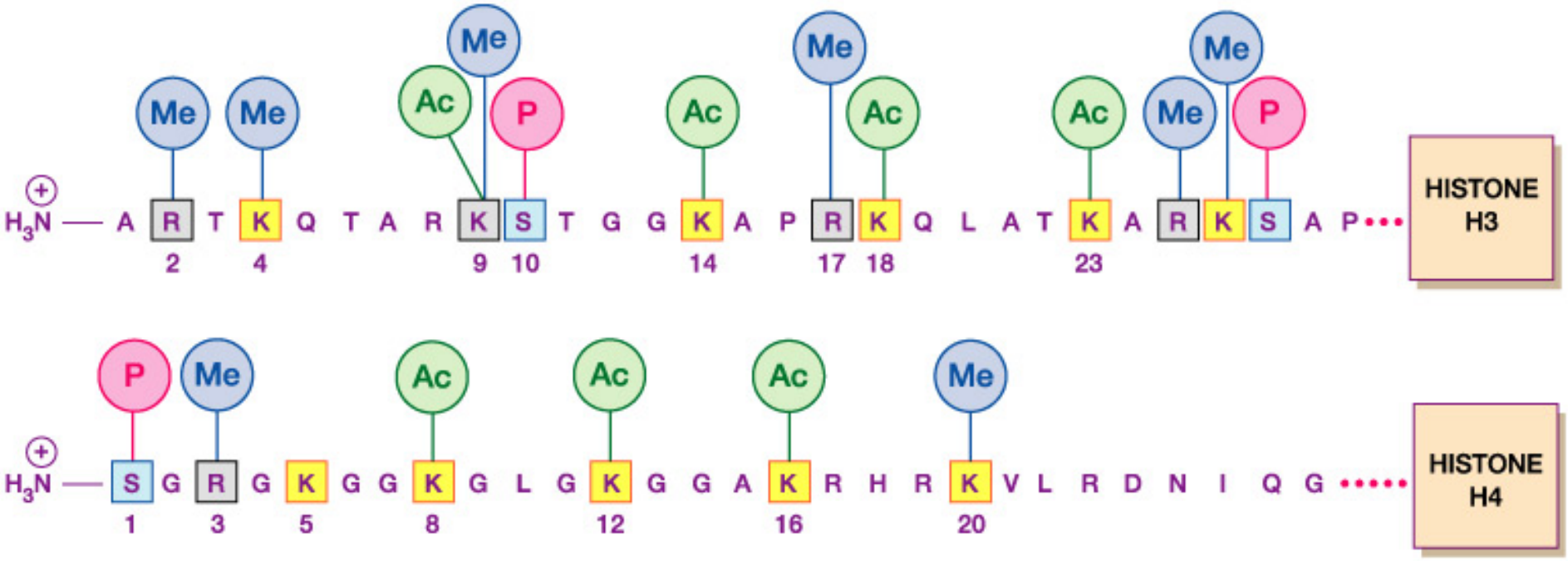
Epigenetics operates via DNA methylation and histone modification



Gene organization



Gene expression



Histone modification: controls the status of chromatin and gene promoters

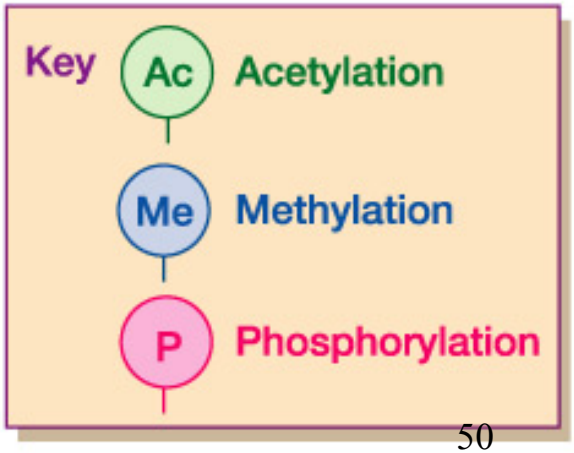
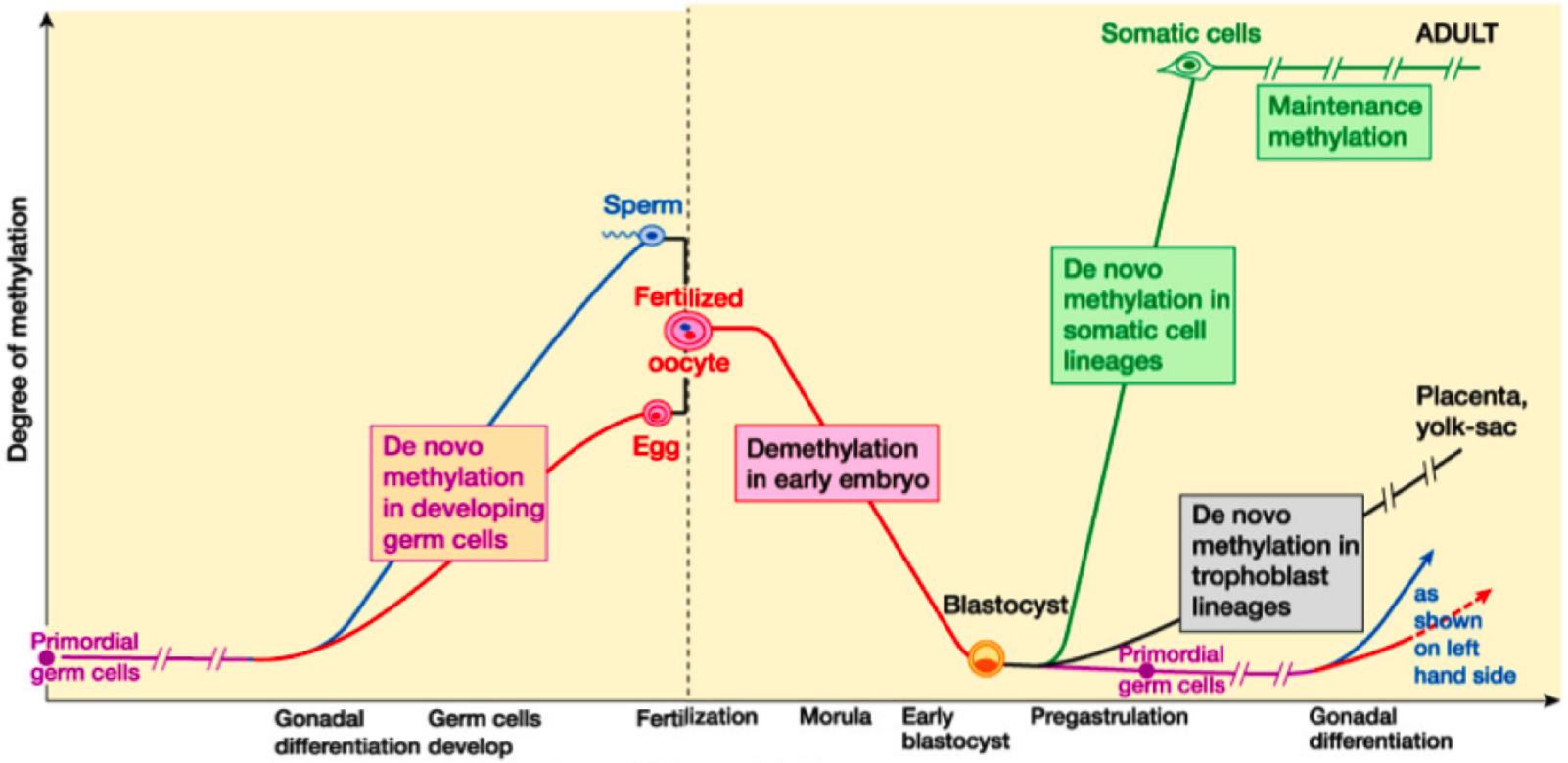


Figure 10-1 Human Molecular Genetics, 3/e. (© Garland Science 2004)

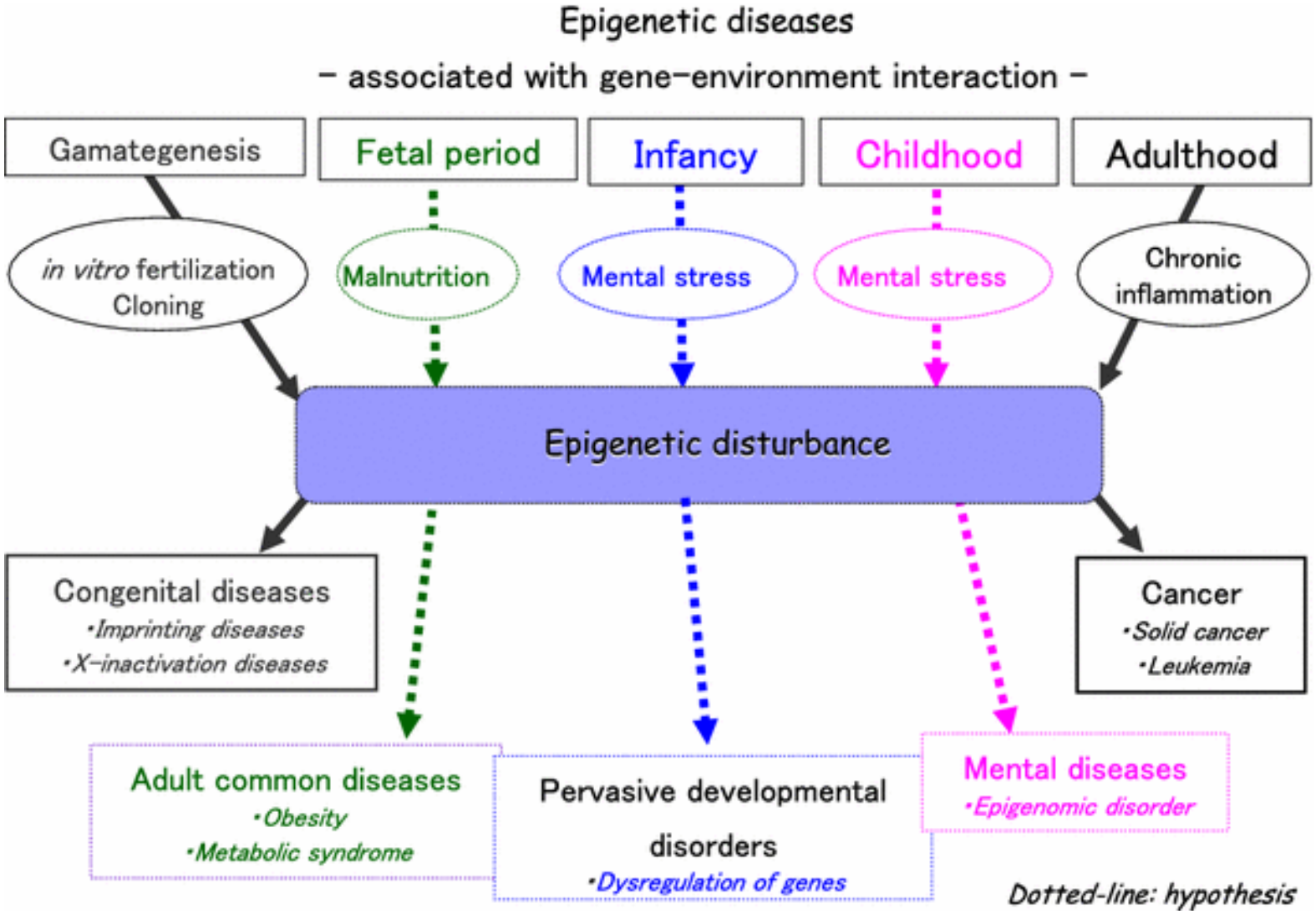
Most epigenetic variation is established in development



Gamete formation

Fetal and childhood

Diseases linked to epigenetics



Kubota, EHPM 2007

Drugs with epigenetic mechanisms