Using genetic data in the Add Health Sample

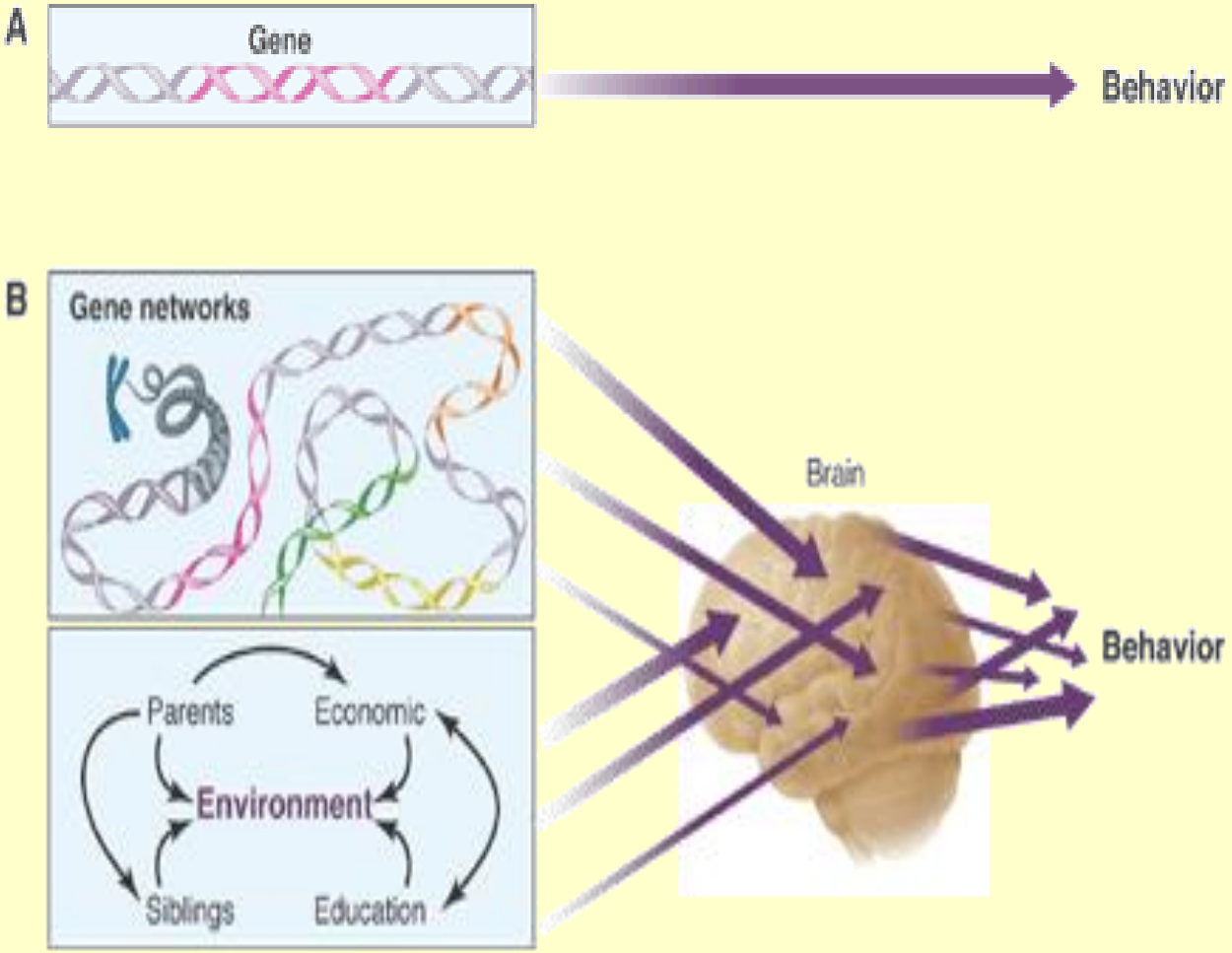
Andrew Smolen and Brett Haberstick

Institute for Behavioral Genetics, University of Colorado at Boulder

Add Health Users Workshop, July 2010

Bethesda, Maryland

Genetic influences can be direct or indirect

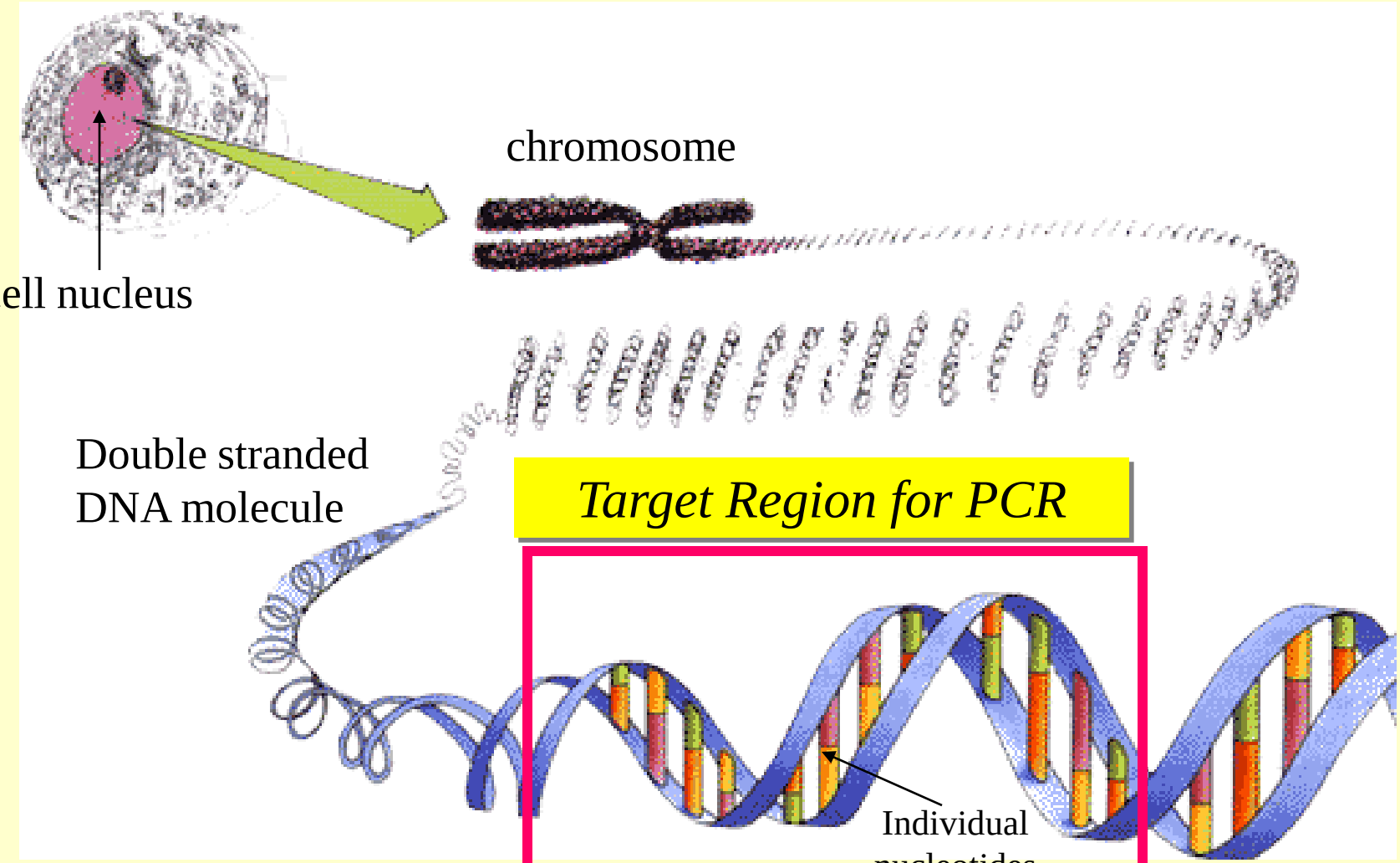


= Mendelian trait:
a single gene with a large effect

= Complex trait:
many genes, each with a small effect

Hamer, 2002; Dr. Ursula M. D'Souza

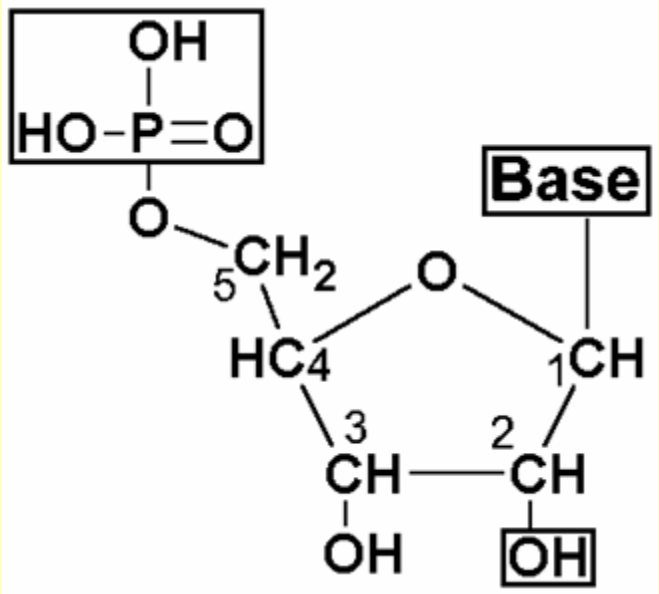
DNA in the Cell



from John Butler - NIST

Nucleotides

- **Nucleotides are the building blocks of nucleic acids**
- **ribose + base + 5' phosphate**
- **5 different bases**
  - **adenine (A)**
  - **guanine (G)**
  - **cytosine (C)**
  - **thymine (T)**
  - **uracil (U)**

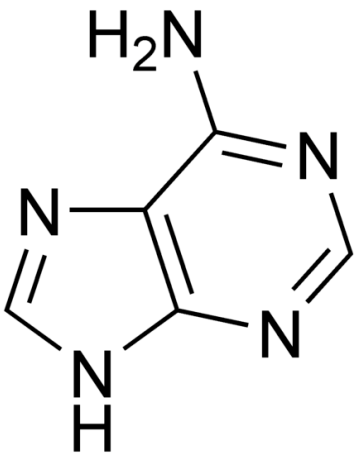


DNA = deoxyribonucleic acid
RNA = ribonucleic acid

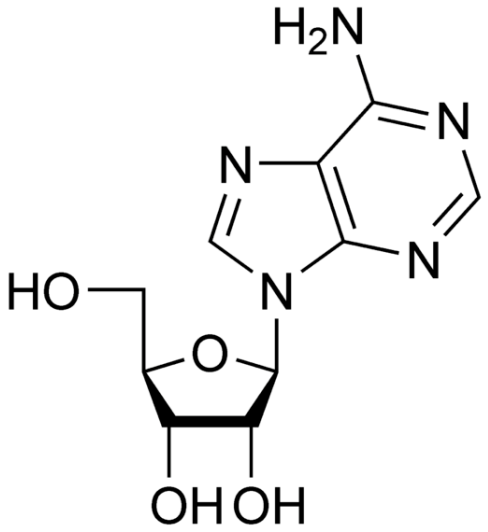
DNA	vs	RNA
		2' oxygen
thymine (T)		uracil (U)
double stranded		single stranded

from Mark F. Wiser - Tulane

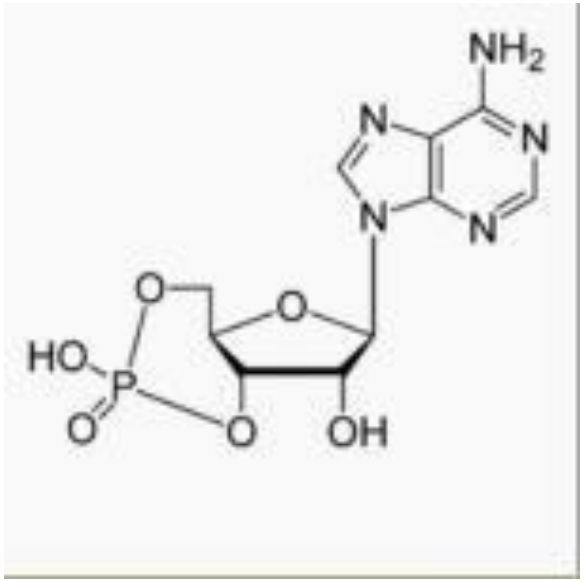
Adenine



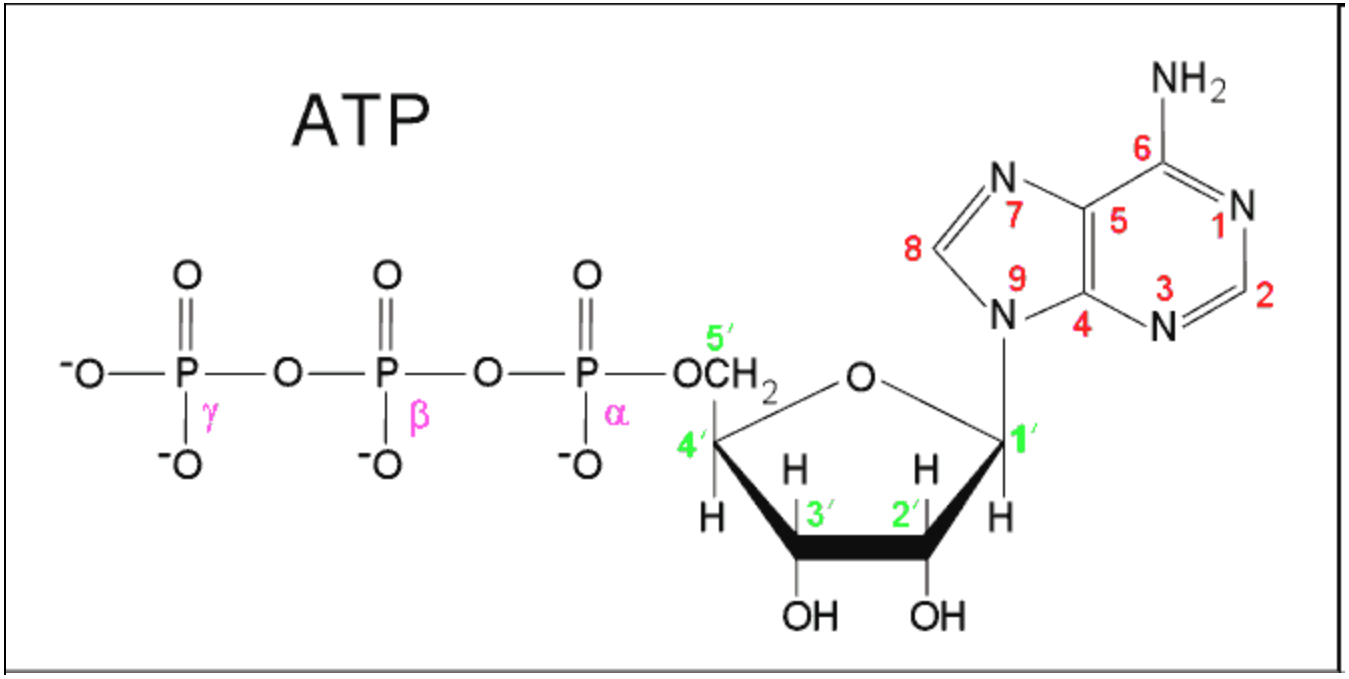
Adenosine

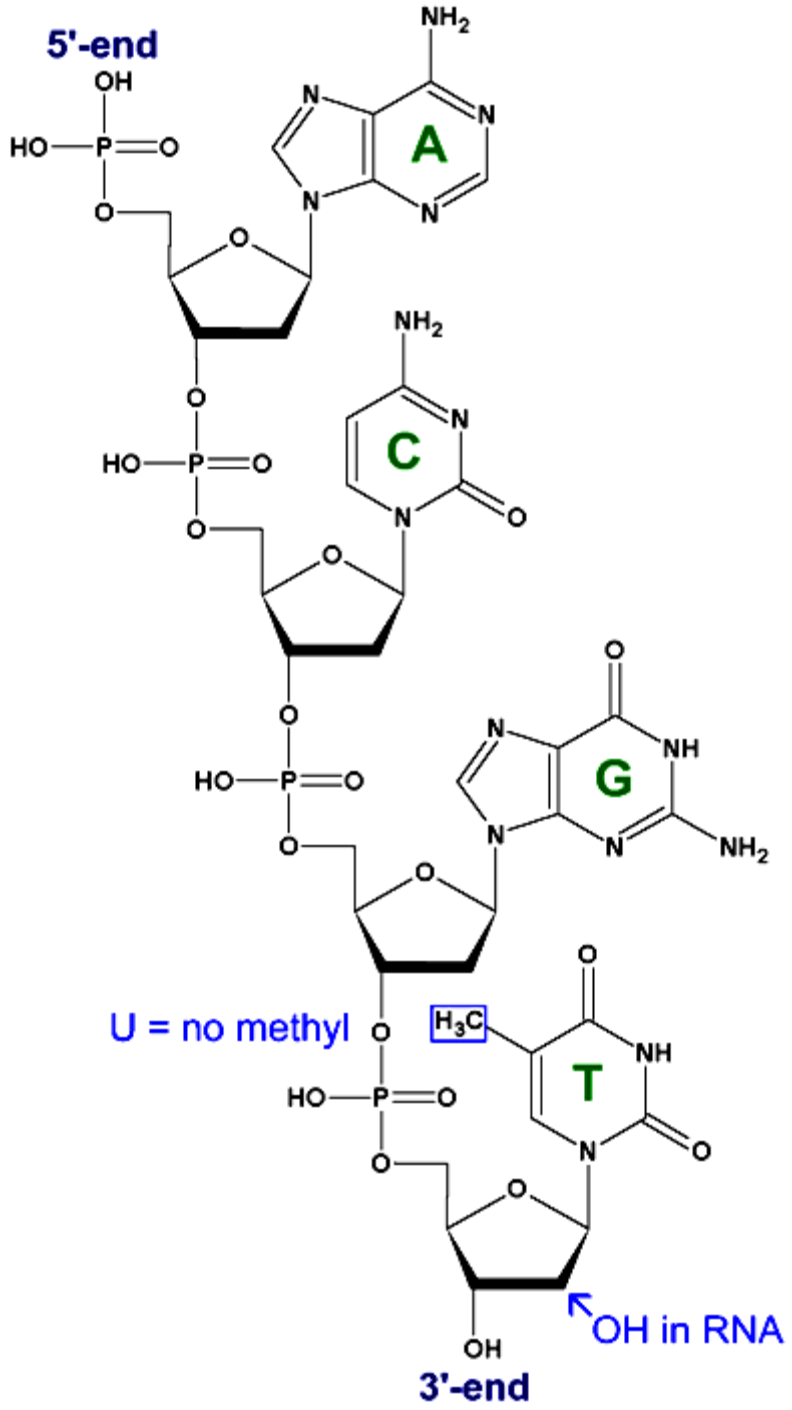


cAMP



ATP





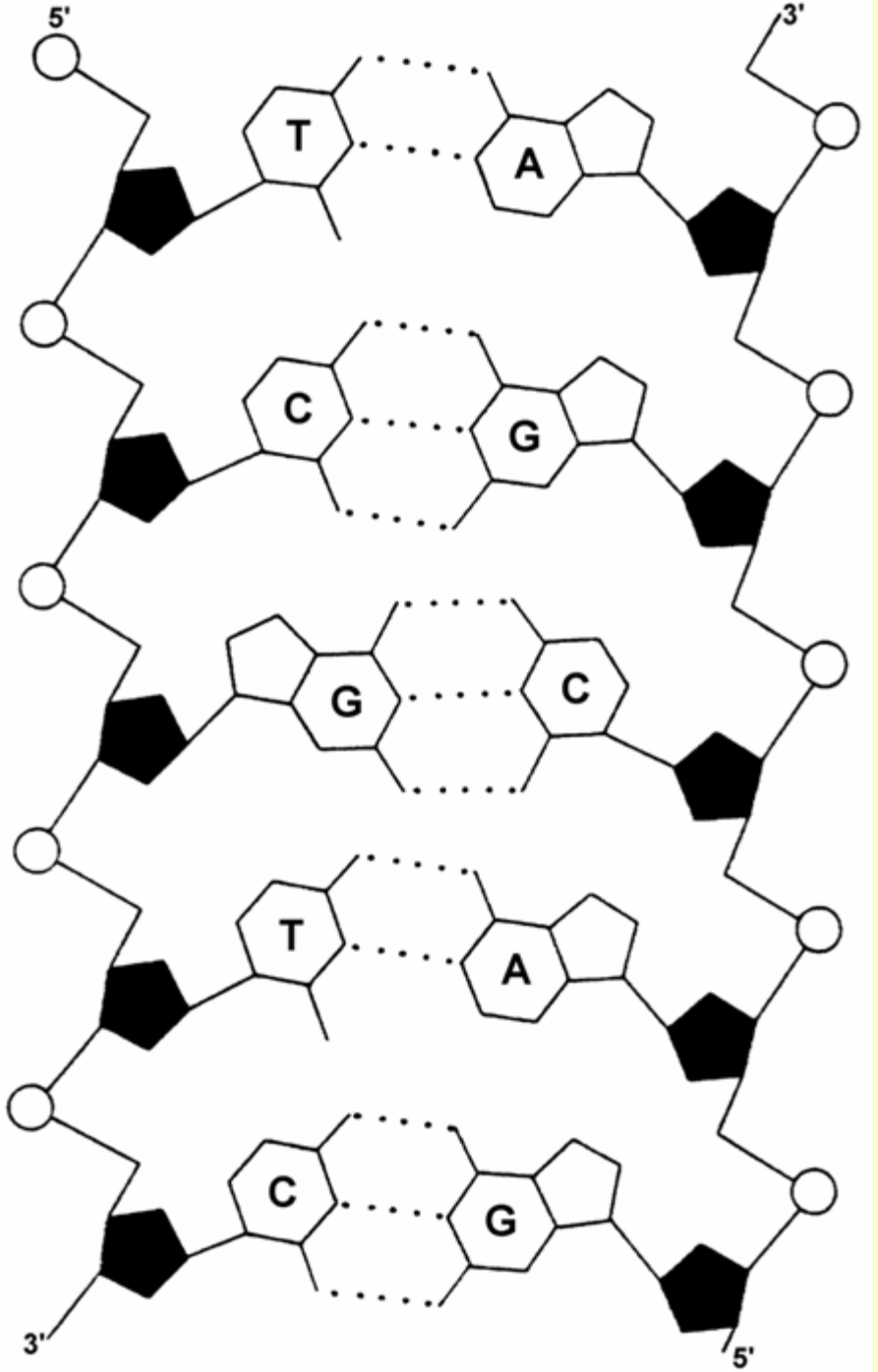
Nucleic Acid Structure

- nucleic acids are polymers of nucleotides
- nucleotides linked through phosphate bonds
- strand has polarity—5'-PO_4 and 3'-OH ends
- order of the nucleotides defines a sequence
- 4 letter alphabet (T or U)

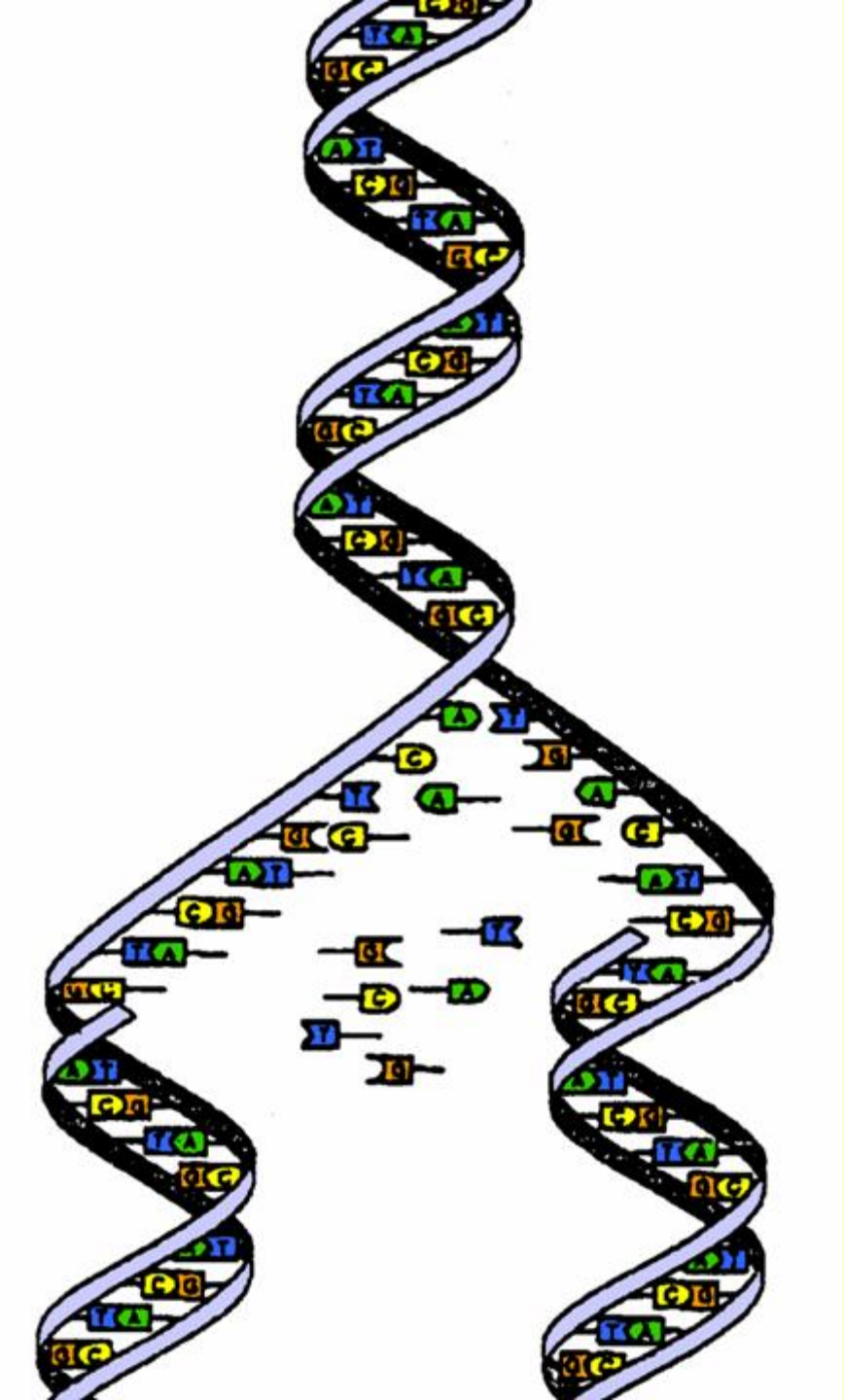
from Mark F. Wiser - Tulane

Double Strands

- specific base pairing
  - A and T
  - G and C
  - hydrogen bonds
- anti-parallel (5'-3')
- complementary strands
- template for synthesis



from Mark F. Wiser - Tulane

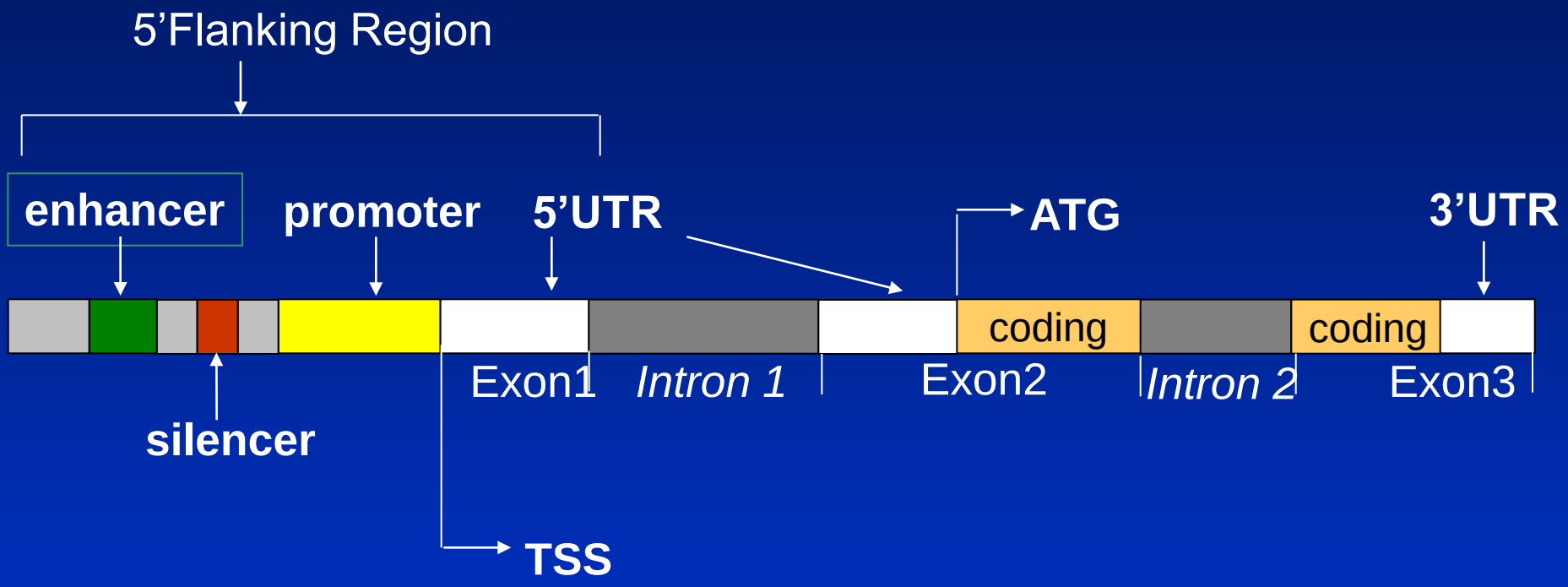


DNA Replication

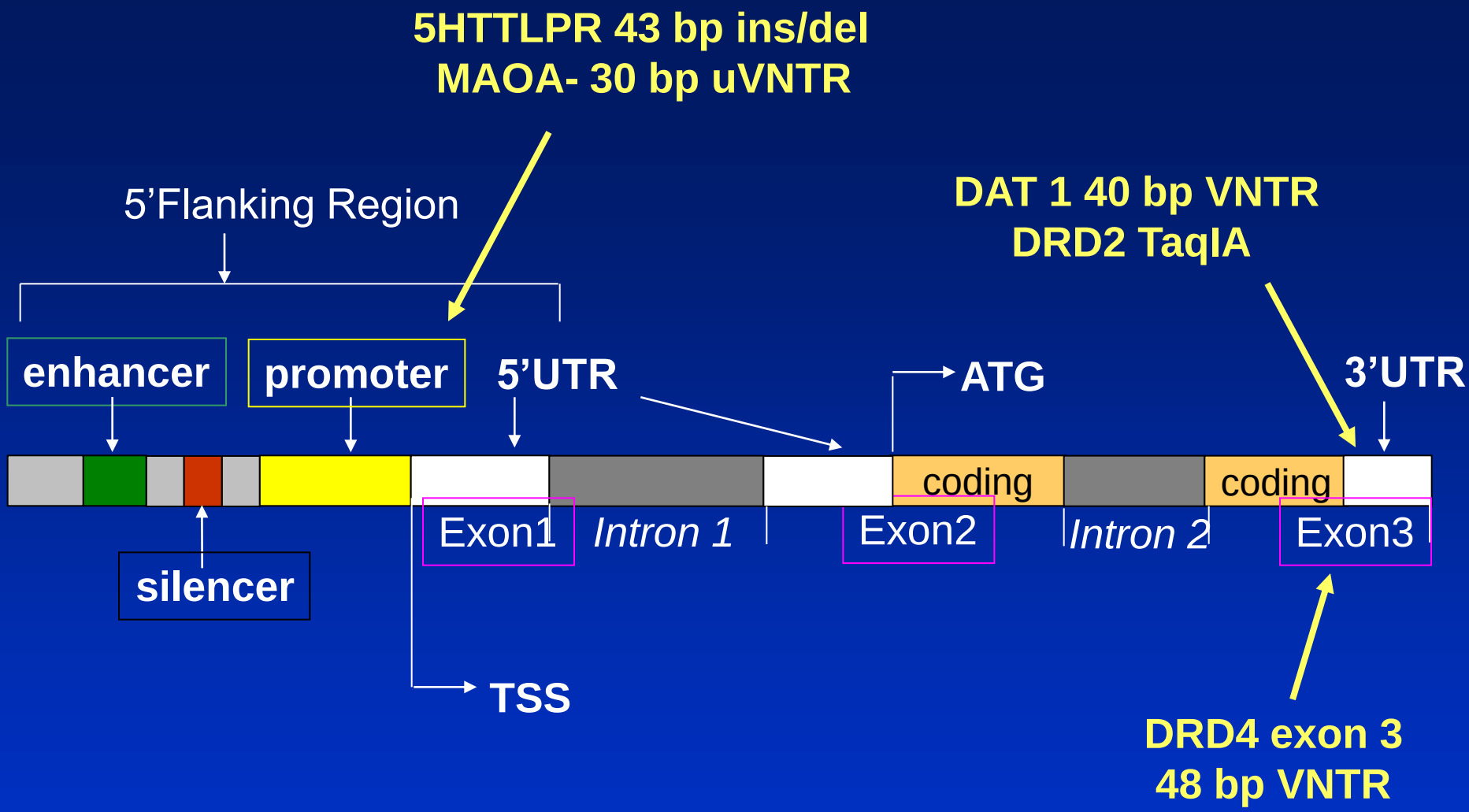
- the DNA strands are separated
- each strand serves as template
- complementary strands are synthesized (5'→3')
- yields 2 identical DNA molecules
- semi-conservative replication
  - carried out by cellular proteins (= enzymes)
  - DNA polymerase
    - complex process involving many proteins

from Mark F. Wiser - Tulane

The generalized structure of a gene



Adapted from Dr. Ursula. M. D’Souza



For AddHealth we analyze loci across the entire gene

Types of genetic differences between people

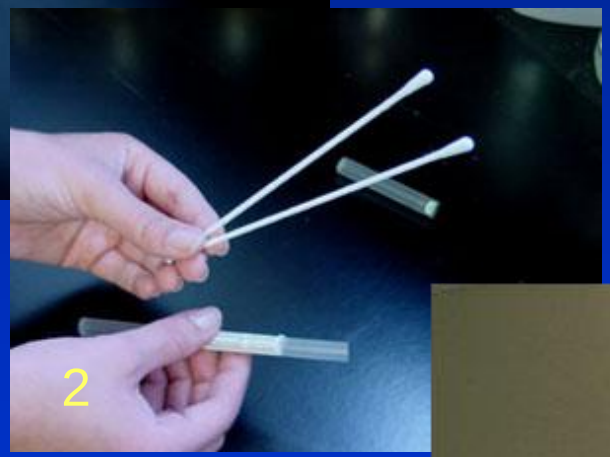
Any differences in DNA segments between any two persons can be used as genetic markers.
Fragments of DNA can be distinguished from one another because of differences in their nucleotide sequences.

Types:

1. Single Nucleotide Polymorphisms SNPs. A single base pair change one strand of DNA. The most prevalent form of differences between any two individuals.
2. Minisatellites 10-100 nucleotides repeated several times in tandem; bordered by unique DNA sequences. Variable Number Tandem Repeats (VNTR) is an example. There are about 50,000 VNTRs in the human genome.
3. Microsatellites or Short Tandem Repeats (STRs). Smaller sequence repeats than minisatellites. Di- and tetra-nucleotide sequence repeats are common.

DNA collection

Any tissue can be used.
Blood is the gold standard.
Buccal cells are more convenient.



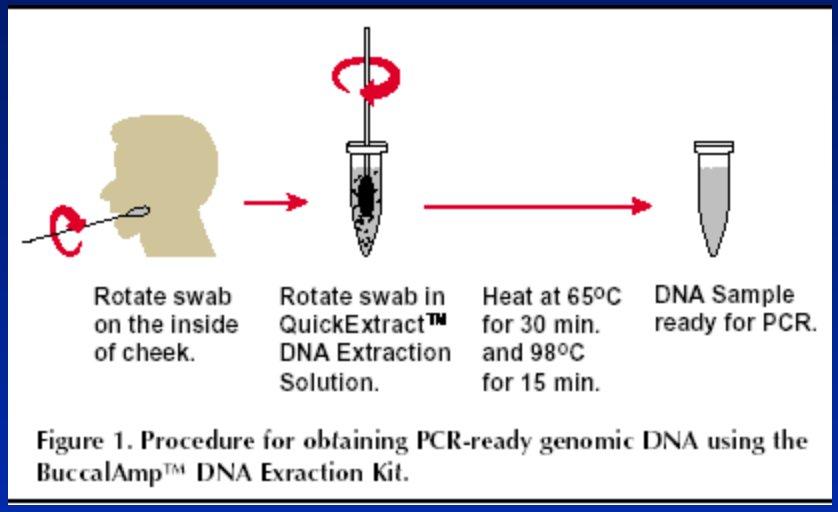
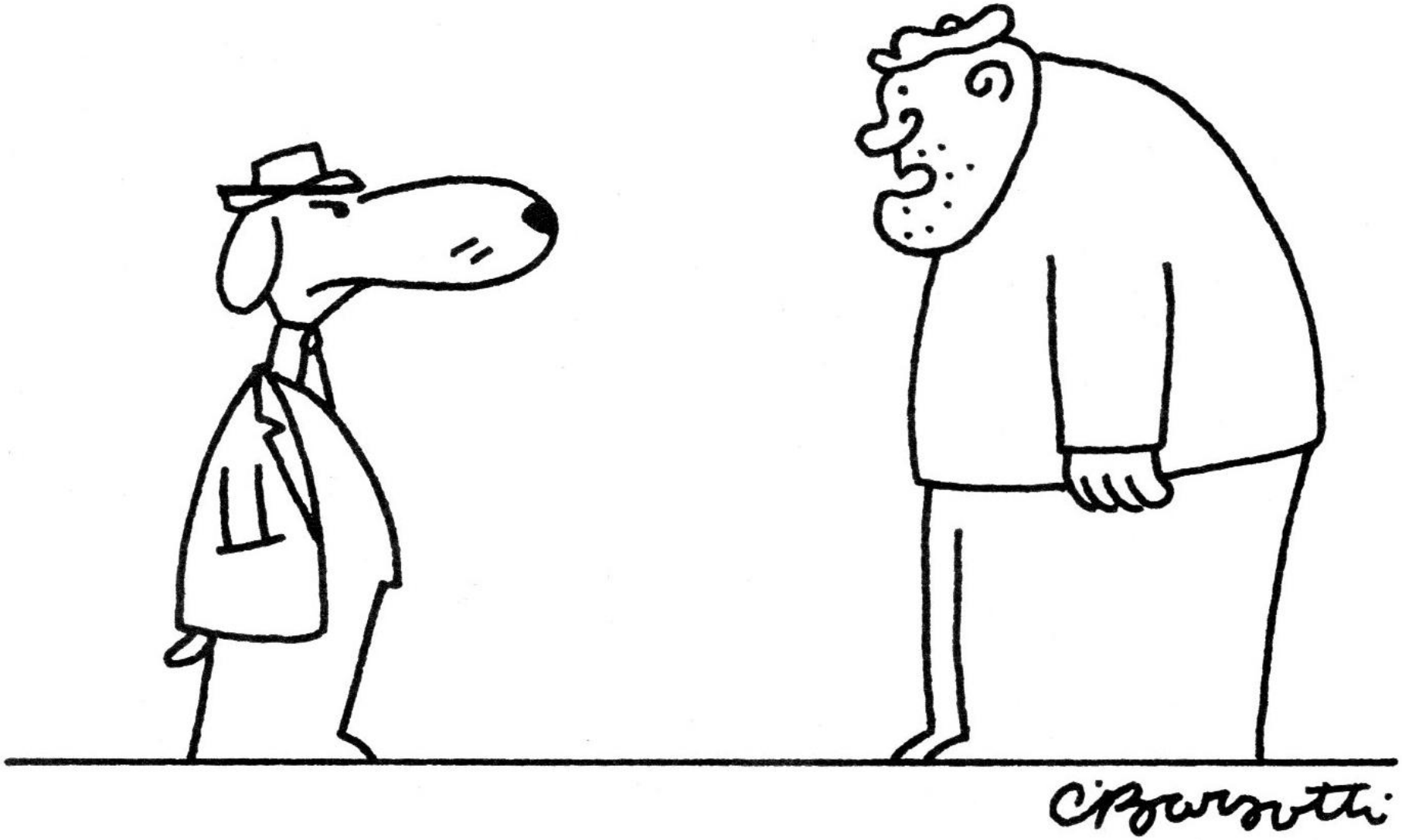
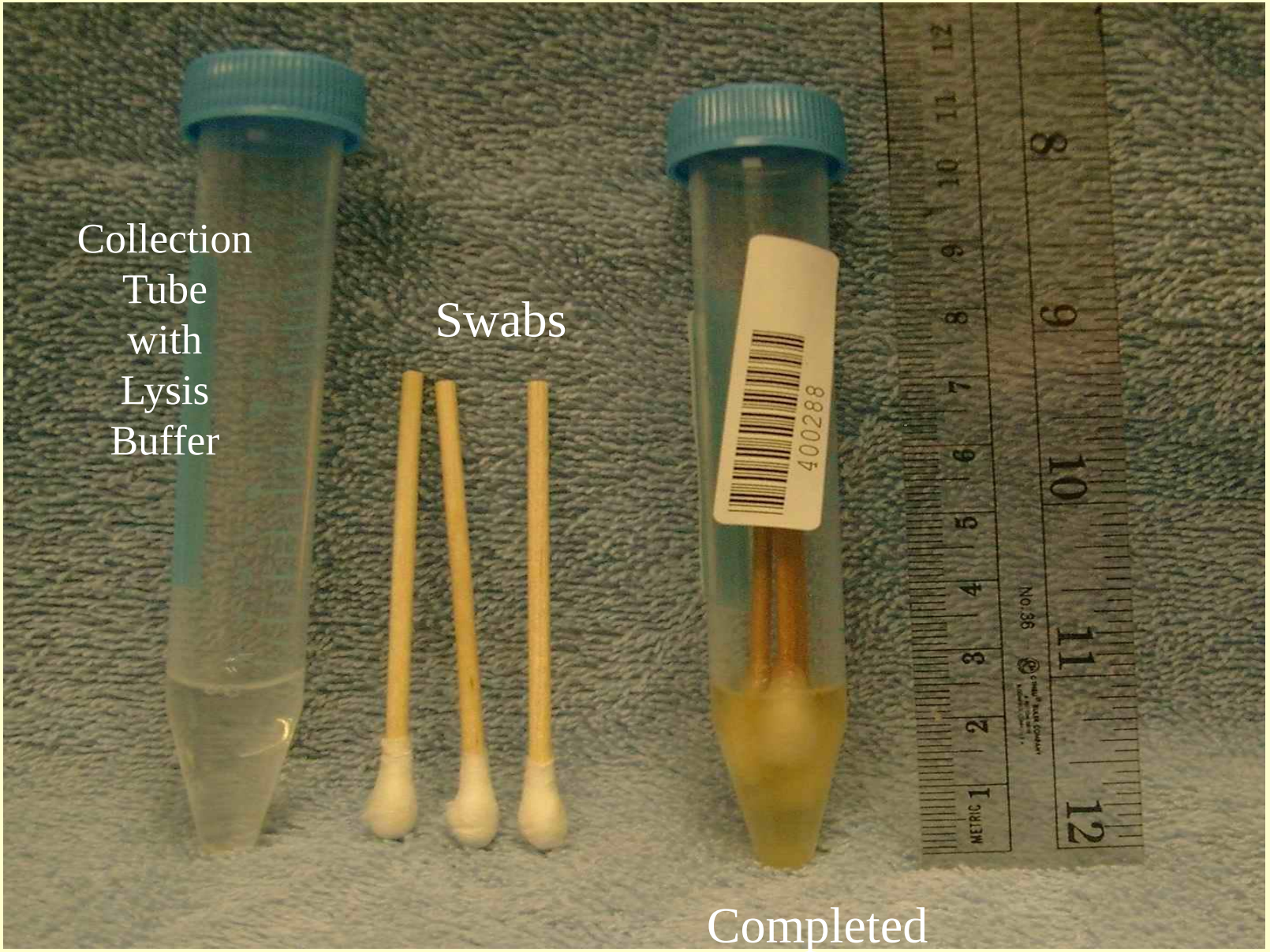


Figure 1. Procedure for obtaining PCR-ready genomic DNA using the BuccalAmp™ DNA Exraction Kit.



"It looks bad, boss—they got your DNA off an old tennis ball."


Collection Tube with Lysis Buffer
Swabs
Completed
400288
METRIC
No. 36

Oragene®-DNA Self Collection Products

DISC Format
OG-250

Compact & robust for easy mailing

TUBE Format
OG-300

Standardized format for high-throughput processing

VIAL Format
OG-100

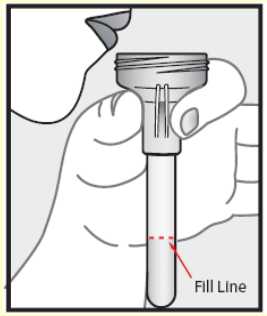
Improved classic format

Courtesy of DNA Genotek, Inc.

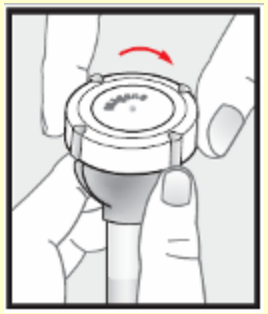
DNA genotek

www.dnagenotek.com

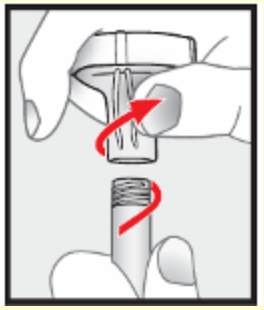
Oragene®•DNA Self-Collection Kit User Instructions (OG-300 Tube Format)



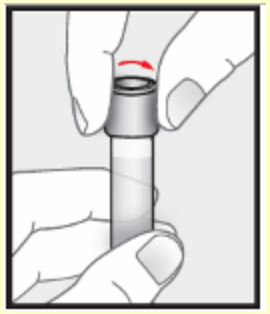
Spit until liquid saliva reaches the fill line.



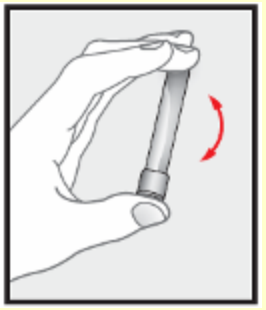
Close funnel. Liquid in big cap will mix with saliva.



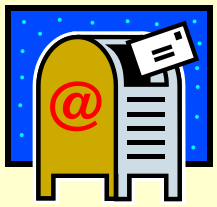
Unscrew funnel from tube.



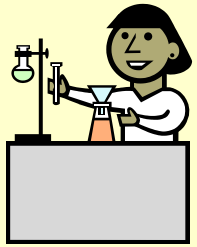
Close tube with small cap.



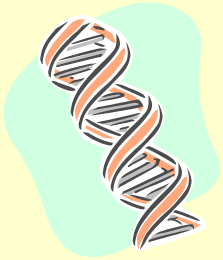
Mix 5 times. Throw out funnel and big cap.



Mail or FedEx



Laboratory



Extract DNA

Courtesy of DNA Genotek, Inc.

DNA genotek

www.dnagenotek.com

DNA Collection and Genotyping

for Wave IV

of Add Health

DNA Collection and Genotyping

The Function of the IBG Lab in Program Project

1) Extract, catalogue, and store DNA from all subjects.
   Buccal Cells.
   ~15,000 individuals.

2) Perform genotyping on all of the individuals.

3) Provide DNA to other Add Health investigators.

Collection of DNA from Buccal Cells or Saliva

MAJOR ADVANTAGES

Simple, non-invasive and non-threatening
Requires no special training
Avoids blood-borne pathogen issues

MAJOR DISADVANTAGE

Less DNA compare to blood

Archiving the DNA

Samples arrive by post or overnight carrier

Samples are visually inspected

Samples are scanned-in using barcode reader

Each sample is assigned a lab id number and position in a storage box

DNA is extracted, quantified and stored.

Samples are delivered 20 per bag

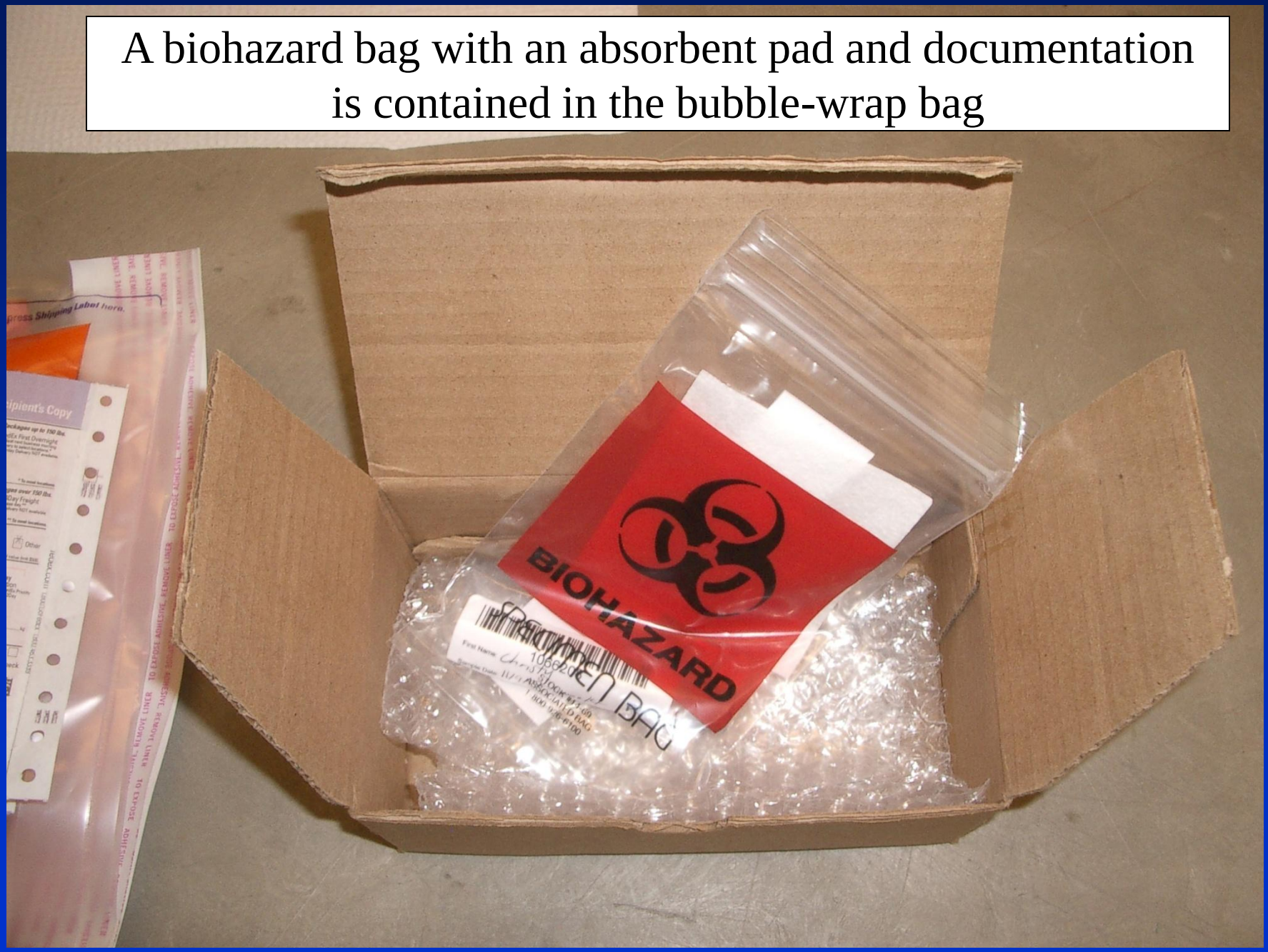
Outer bag


EXEMPT HUMAN SPECIMENS
FedEx Express
fedex.com 1.800.GoFedEx 1.800.463.3339
RECIPIENT: PEEL HERE
FedEx US Airbill
8630 7344 2124
1 From
FedEx Tracking Number 863073442124
Sender's Name
Phone
Company
Address
City
State
ZIP
2 Your Internal Billing Reference
3 To
Recipient's Name ANDY SMOLEN
Phone 303 492-8639
Company UNIV OF CO-INST OF BEHAV GENETIC
Recipient's Address 1480 30TH STREET
Address
City BOULDER
State CO
ZIP 80303
8630 7344 2124
0215
Recipient's Copy
4a Express Package Service
4b Express Freight Service
5 Packaging
6 Special Handling
7 Payment
8 Residential Delivery Signature Options
519
Align top of FedEx Express Shipping Label here.
Align bottom of Pouch here.

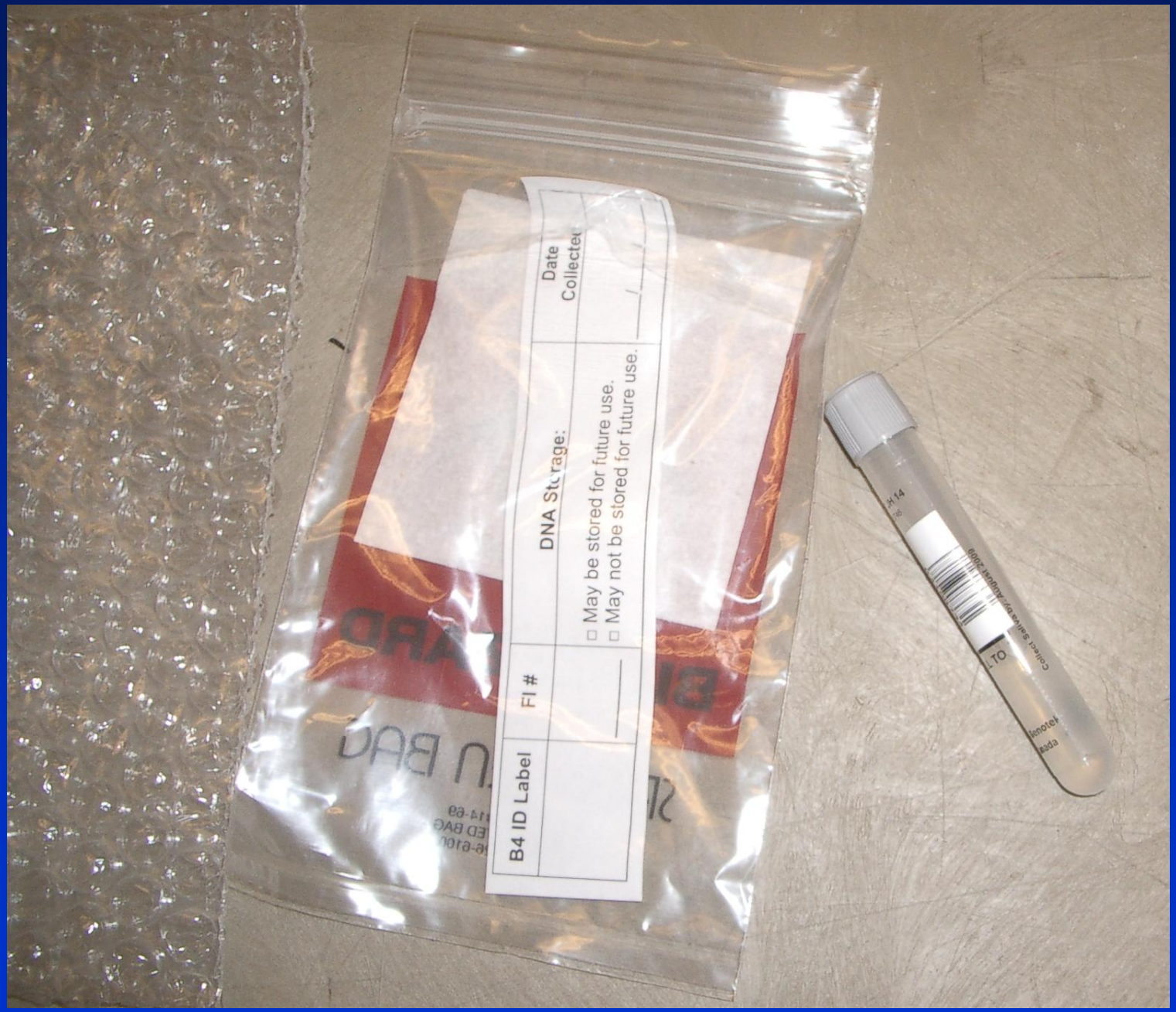
A box is the bag

In the box is a bubble-wrap bag with the sample

A biohazard bag with an absorbent pad and documentation is contained in the bubble-wrap bag



The Oragene tube is removed from the biohazard bag



The whole package

This generates an impressive amount of garbage.

Samples are annotated and scanned into the data base

Lined up and prepared to be scanned

Archiving the DNA

Samples arrive by post or overnight carrier

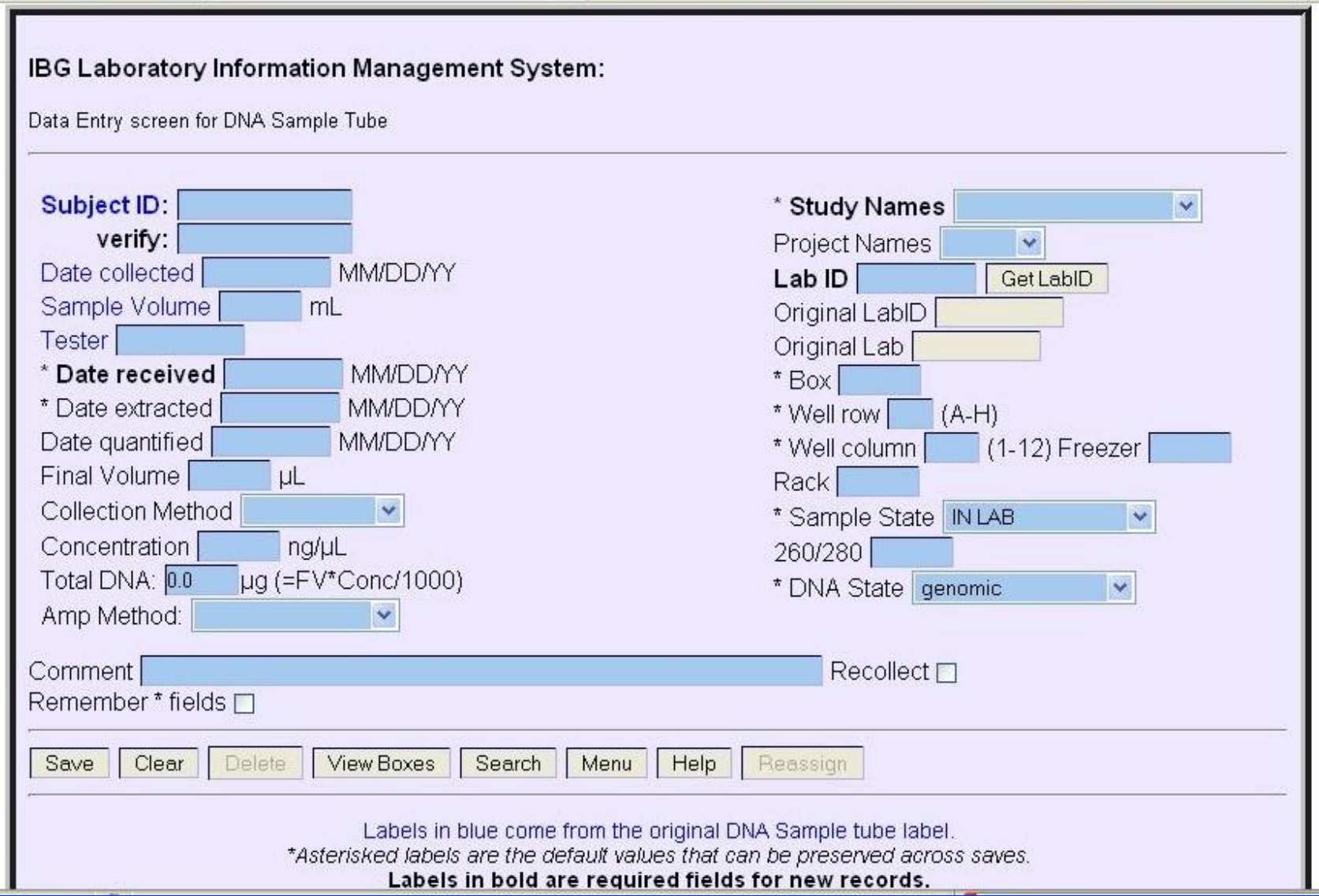
Samples are visually inspected

Samples are scanned-in using barcode reader

Each sample is assigned a lab id number and position in a storage box

DNA is extracted, quantified and stored.

IBG Laboratory Information Management System (LIMS)
a secure, password-protected web site
Sample Input Screen



IBG Laboratory Information Management System:

Data Entry screen for DNA Sample Tube

Subject ID:
verify:
Date collected MM/DD/YY
Sample Volume mL
Tester
* **Date received** MM/DD/YY
* Date extracted MM/DD/YY
Date quantified MM/DD/YY
Final Volume µL
Collection Method
Concentration ng/µL
Total DNA: 0.0 µg (=FV*Conc/1000)
Amp Method:

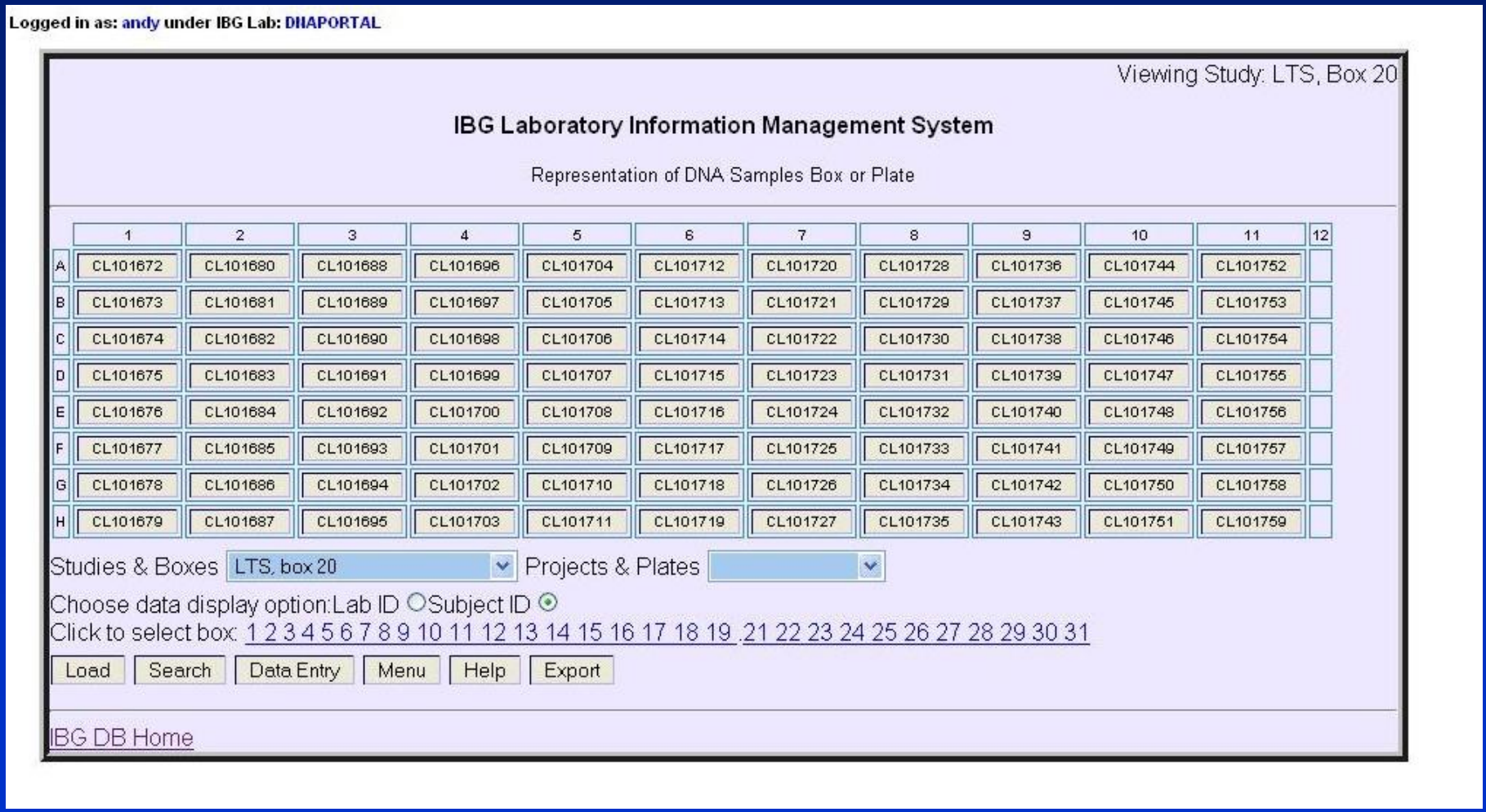
* **Study Names**
Project Names
Lab ID Get LabID
Original LabID
Original Lab
* Box
* Well row (A-H)
* Well column (1-12) Freezer
Rack
* Sample State IN LAB
260/280
* DNA State genomic

Comment Recollect
Remember * fields

Save | Clear | Delete | View Boxes | Search | Menu | Help | Reassign

Labels in blue come from the original DNA Sample tube label.
**Asterisked labels are the default values that can be preserved across saves.*
Labels in bold are required fields for new records.

IBG LIMS – Tool to Locate Archived Samples
Note that only lab IDs are displayed.
Subject IDs are hidden.



Logged in as: andy under IBG Lab: DNAPORTAL

Viewing Study: LTS, Box 20

IBG Laboratory Information Management System

Representation of DNA Samples Box or Plate

	1	2	3	4	5	6	7	8	9	10	11	12
A	CL101672	CL101680	CL101688	CL101696	CL101704	CL101712	CL101720	CL101728	CL101736	CL101744	CL101752	
B	CL101673	CL101681	CL101689	CL101697	CL101705	CL101713	CL101721	CL101729	CL101737	CL101745	CL101753	
C	CL101674	CL101682	CL101690	CL101698	CL101706	CL101714	CL101722	CL101730	CL101738	CL101746	CL101754	
D	CL101675	CL101683	CL101691	CL101699	CL101707	CL101715	CL101723	CL101731	CL101739	CL101747	CL101755	
E	CL101676	CL101684	CL101692	CL101700	CL101708	CL101716	CL101724	CL101732	CL101740	CL101748	CL101756	
F	CL101677	CL101685	CL101693	CL101701	CL101709	CL101717	CL101725	CL101733	CL101741	CL101749	CL101757	
G	CL101678	CL101686	CL101694	CL101702	CL101710	CL101718	CL101726	CL101734	CL101742	CL101750	CL101758	
H	CL101679	CL101687	CL101695	CL101703	CL101711	CL101719	CL101727	CL101735	CL101743	CL101751	CL101759	

Studies & Boxes LTS, box 20 Projects & Plates

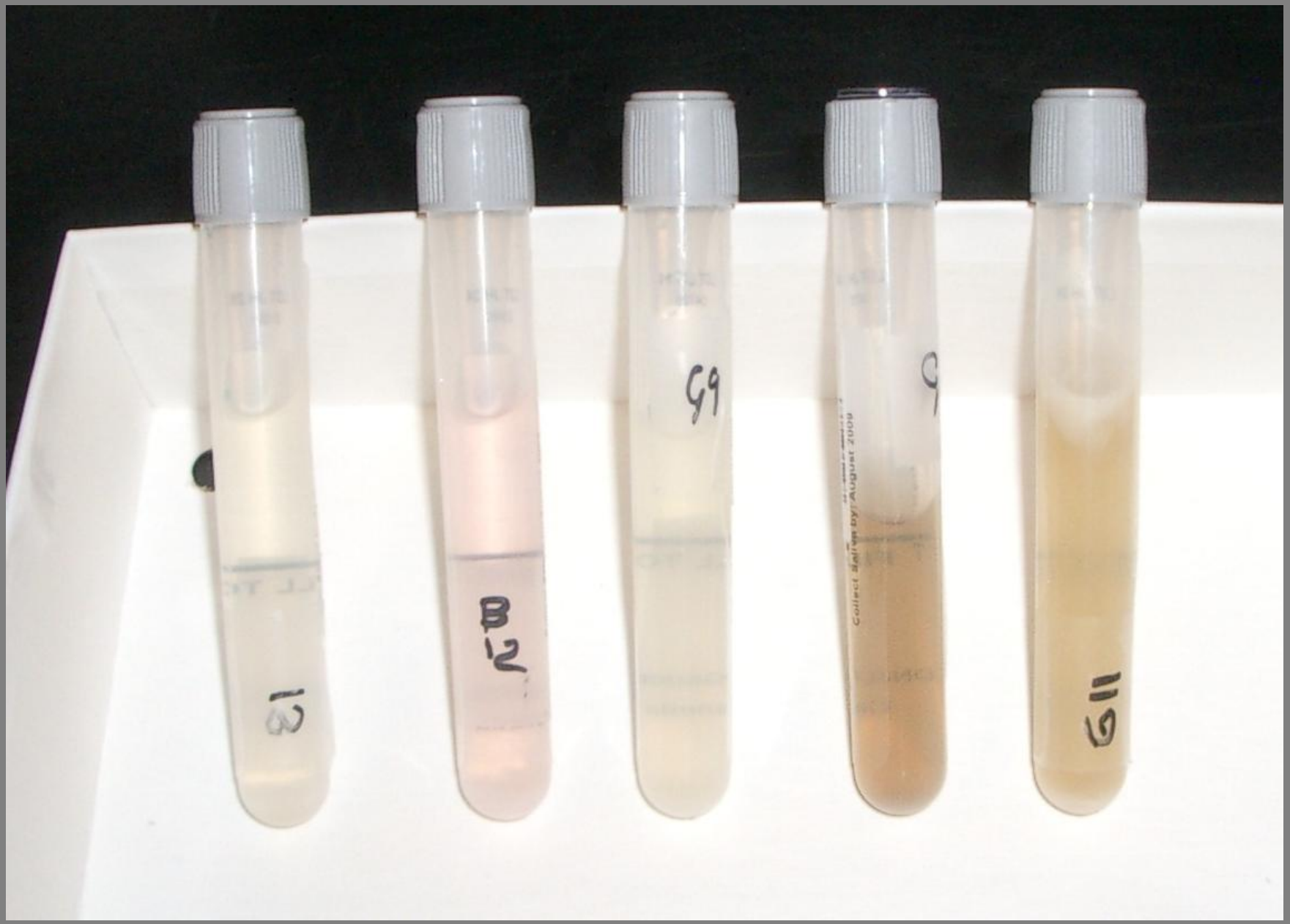
Choose data display option: Lab ID ○ Subject ID ◉

Click to select box: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23 24 25 26 27 28 29 30 31

Load | Search | Data Entry | Menu | Help | Export

IBG DB Home

Samples are placed in boxes awaiting extraction


13
B2
69
9
11

Archiving the DNA

Samples arrive by post or overnight carrier

Samples are visually inspected

Samples are scanned-in using barcode reader

Each sample is assigned a lab id number and position in a storage box

DNA is extracted, quantified, assessed for quality and stored.

From January 21, 2008 to February 4, 2009

Samples logged into database:

15,249 Samples

12,328 May be retained for further genotyping

2,921 Limited genotyping (19%)

450 No usable DNA (3%)

14,800 N for initial genotyping of Wave IV

11,960 N for future genotoyping of Wave IV

DNA Extracted, quantified and agarose gels run

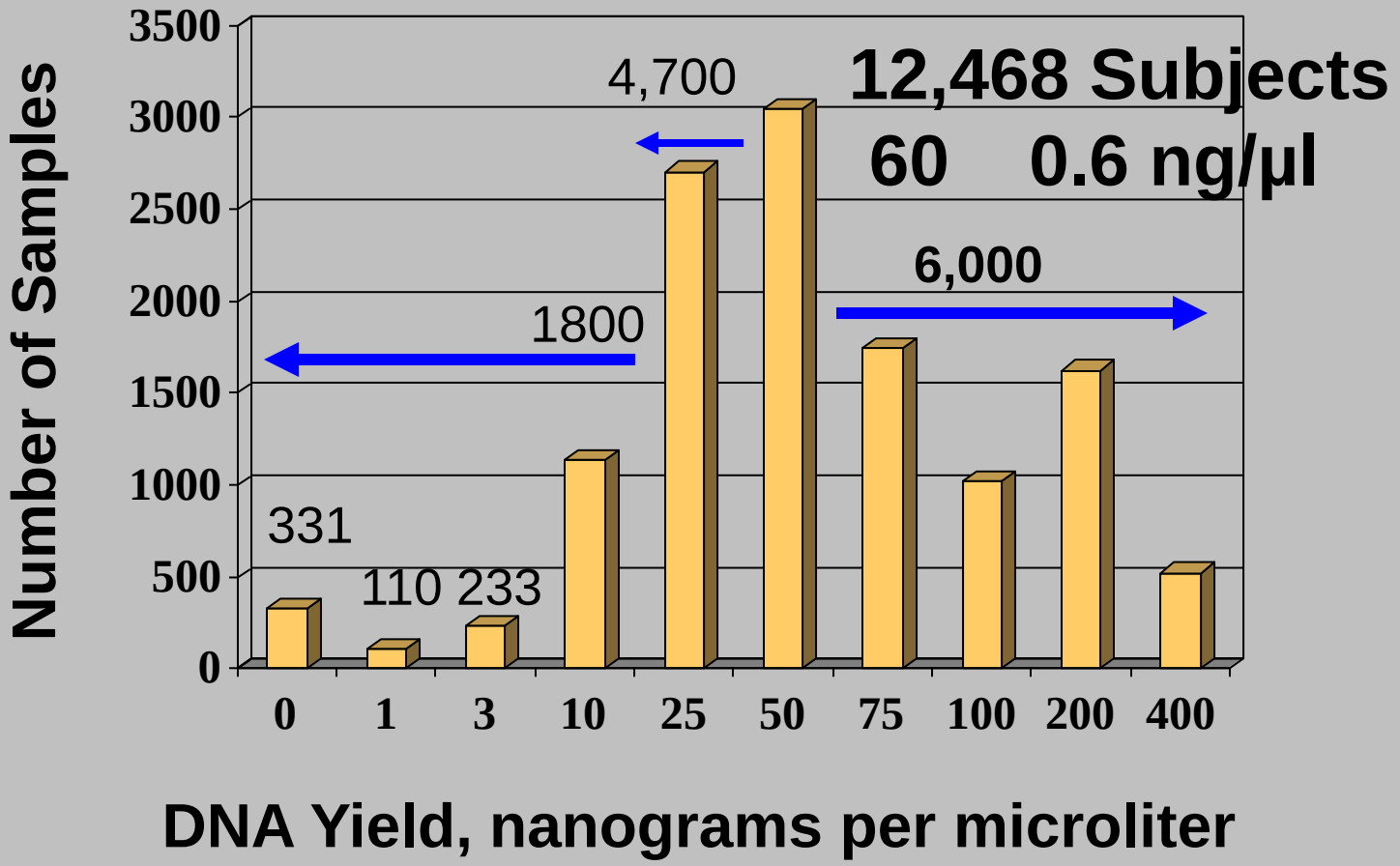
15,249 samples

DNA extraction was done using Zymo Research ZR-96 plates

Quantification by PicoGreen® Fluorescence

Standard agarose gels visualized with SYBR®Safe

DNA Yield for Oragene™ Saliva Method



DNA Yield for Oragene™ Saliva Method

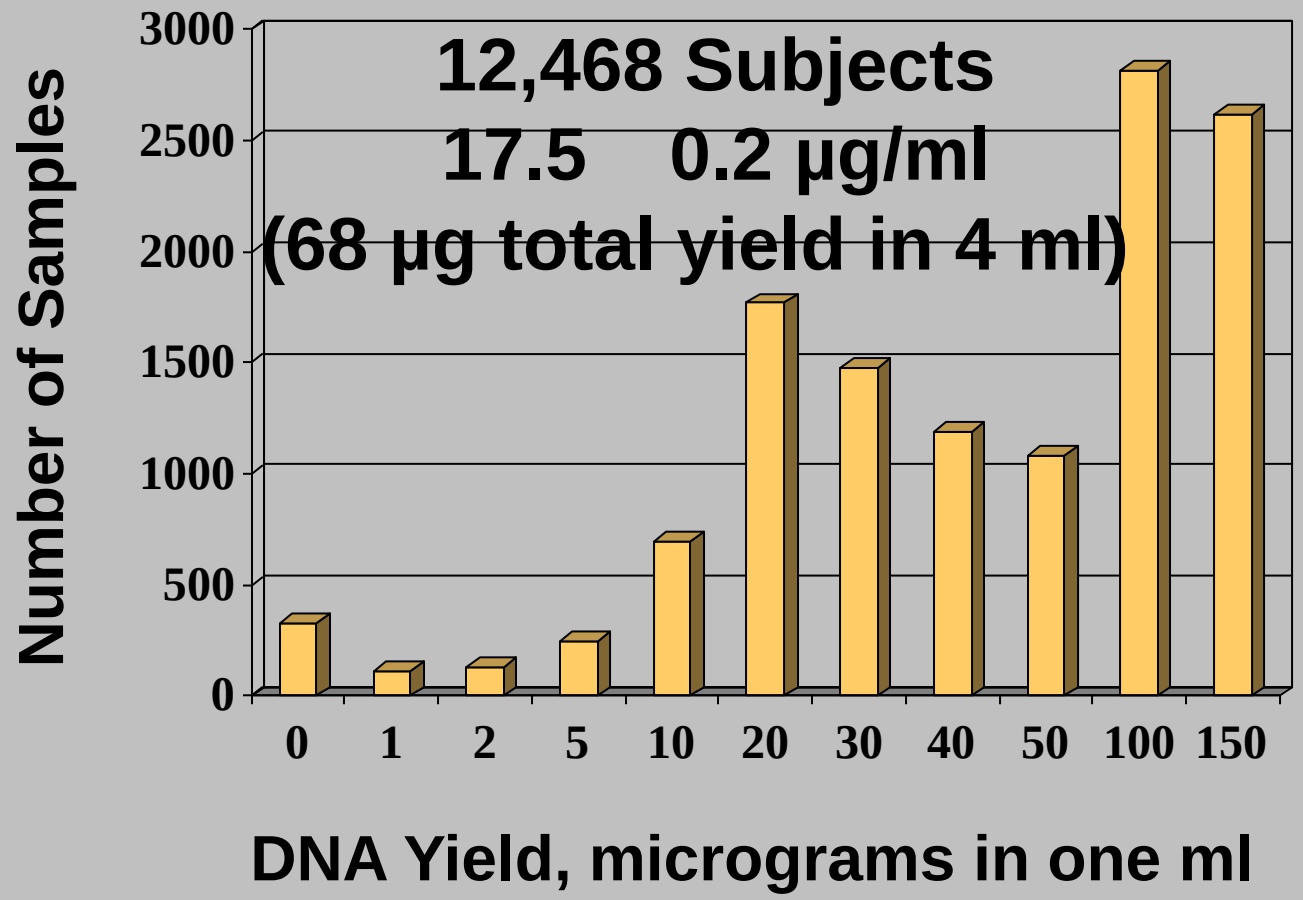
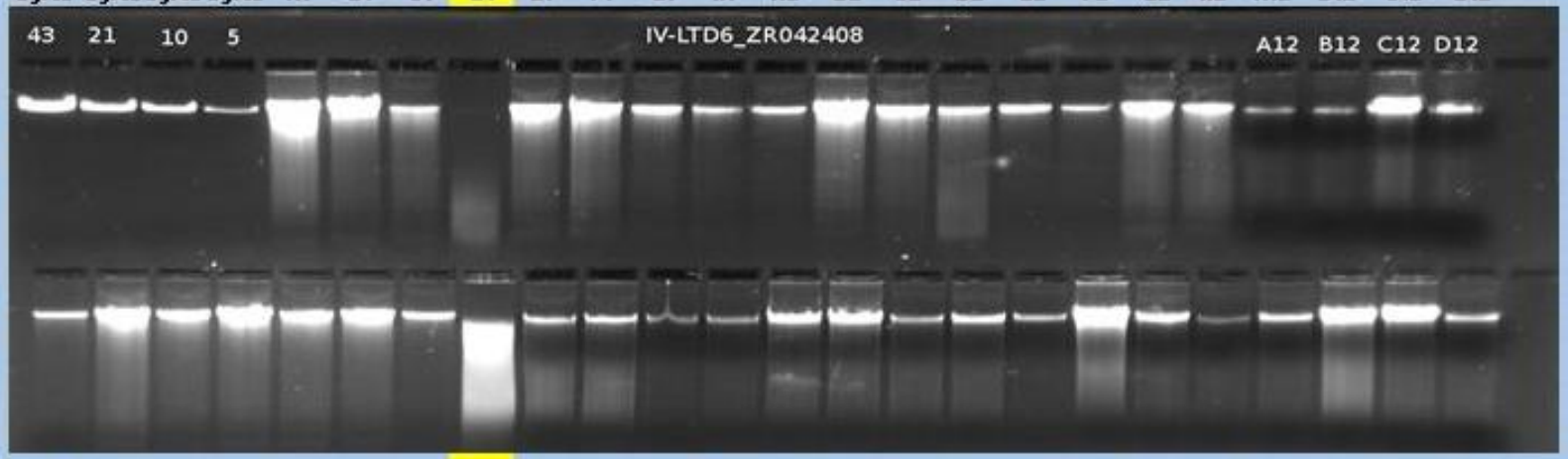
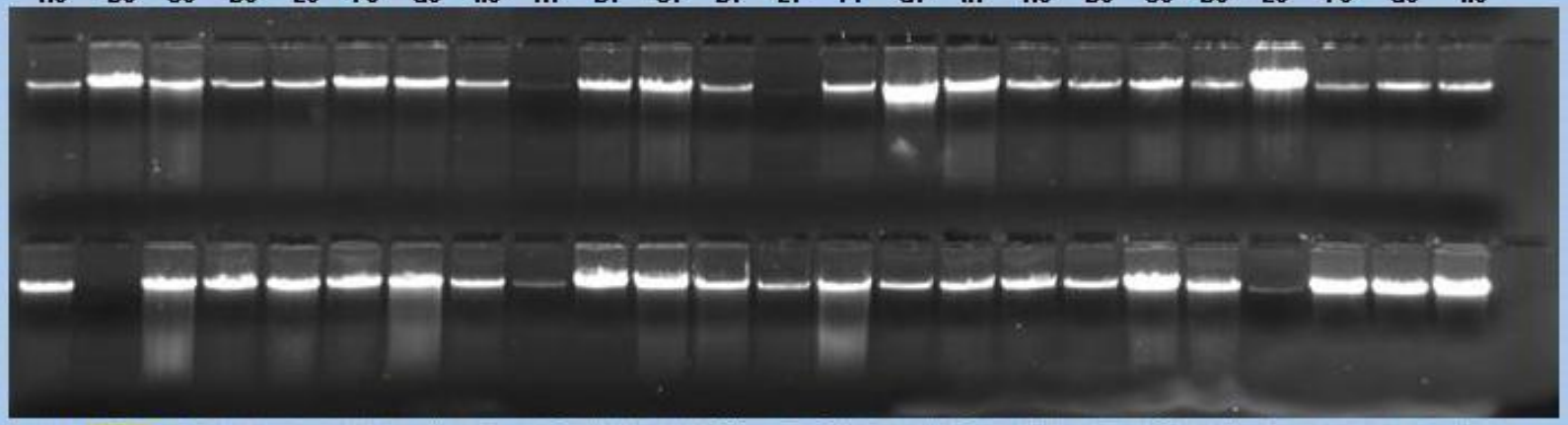


Plate Record for Ltd-06

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	DNA	Vol-uL									<2mL																
2	RepliG	100	ug/uL					96	177	115	29	123	119	143	78	149	194	264	186	144	131	254	215	93	45	82	77
3	genomic	150	ug/uL					128	119	42	39	91	64	30	26	29	111	65	67	35	21	90	46	28	25	90	93
4				ug/uL	ug/uL	ug/uL	ug/uL	A1	B1	C1	D1	E1	F1	G1	H1	A2	B2	C2	D2	E2	F2	G2	H2	A12	B12	C12	D12
5																											
6																											
7																											
8																											
9																											
10																											
11																											
12																											
13																											
14																											
15																											
16																											
17																											
18																											
19																											
20																											
21																											
22																											
23																											
24																											
25																											
26																											
27																											
28																											
29																											
30																											
31																											
32																											
33																											
34																											
35																											
36																											
37																											
38																											
39																											
40																											
41																											
42																											
43																											
44																											
45																											
46																											
47																											
48																											
49																											
50																											
51																											
52																											
53																											
54																											
55																											
56																											
57																											
58																											
59																											
60																											
61																											
62																											
63																											
64																											
65																											



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
66	DNA	Vol-uL		A3	B3	C3	D3	E3	F3	G3	H3	A4	B4	C4	D4	E4	F4	G4	H4	A5	B5	C5	D5	E5	F5	G5	H5
67	RepliG	100	ug/uL	153	139	111	165	160	176	107	52	117	78	111	91	113	157	39	70	283	318	275	##	280	211	179	218
68	genomic	150	ug/uL	27	92	51	64	71	58	24	78	36	61	23	29	91	84	43	38	23	121	46	21	34	136	89	36
69											<2mL																
70	RepliG	100	ug/uL	179	159	154	185	153	221	135	179	51	71	116	87	53	186	6	79	172	133	123	144	129	156	136	123
71	genomic	150	ug/uL	29	106	38	29	29	30	47	43	14	51	66	30	13	48	114	111	47	34	42	35	220	26	35	40
72				A6	B6	C6	D6	E6	F6	G6	H6	A7	B7	C7	D7	E7	F7	G7	H7	A8	B8	C8	D8	E8	F8	G8	H8
73																											
74																											
75																											
76																											
77																											
78																											
79																											
80																											
81																											
82																											
83																											
84																											
85																											
86																											
87																											
88																											
89																											
90																											
91																											
92																											
93																											
94																											
95																											
96																											
97																											
98																											
99																											
100																											
101																											
102																											
103																											
104																											
105																											
106																											
107																											
108																											
109																											
110																											
111																											
112																											
113																											
114																											
115																											
116																											
117																											
118																											
119																											
120																											
121																											
122																											
123																											
124																											
125																											
126																											
127																											
128																											
129																											
130																											



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
131	DNA			A9	B9	C9	D9	E9	F9	G9	H9	A10	B10	C10	D10	E10	F10	G10	H10	A11	B11	C11	D11	E11	F11	G11	H11
132	RepliG	100	ug/uL	250	27	123	131	126	89	152	77	159	97	166	175	111	146	187	155	143	77	142	103	47	132	98	72
133	genomic	150	ug/uL	41	16	36	45	85	59	77	50	22	80	40	86	43	87	35	46	63	39	71	62	21	78	74	158
134					<2mL																						

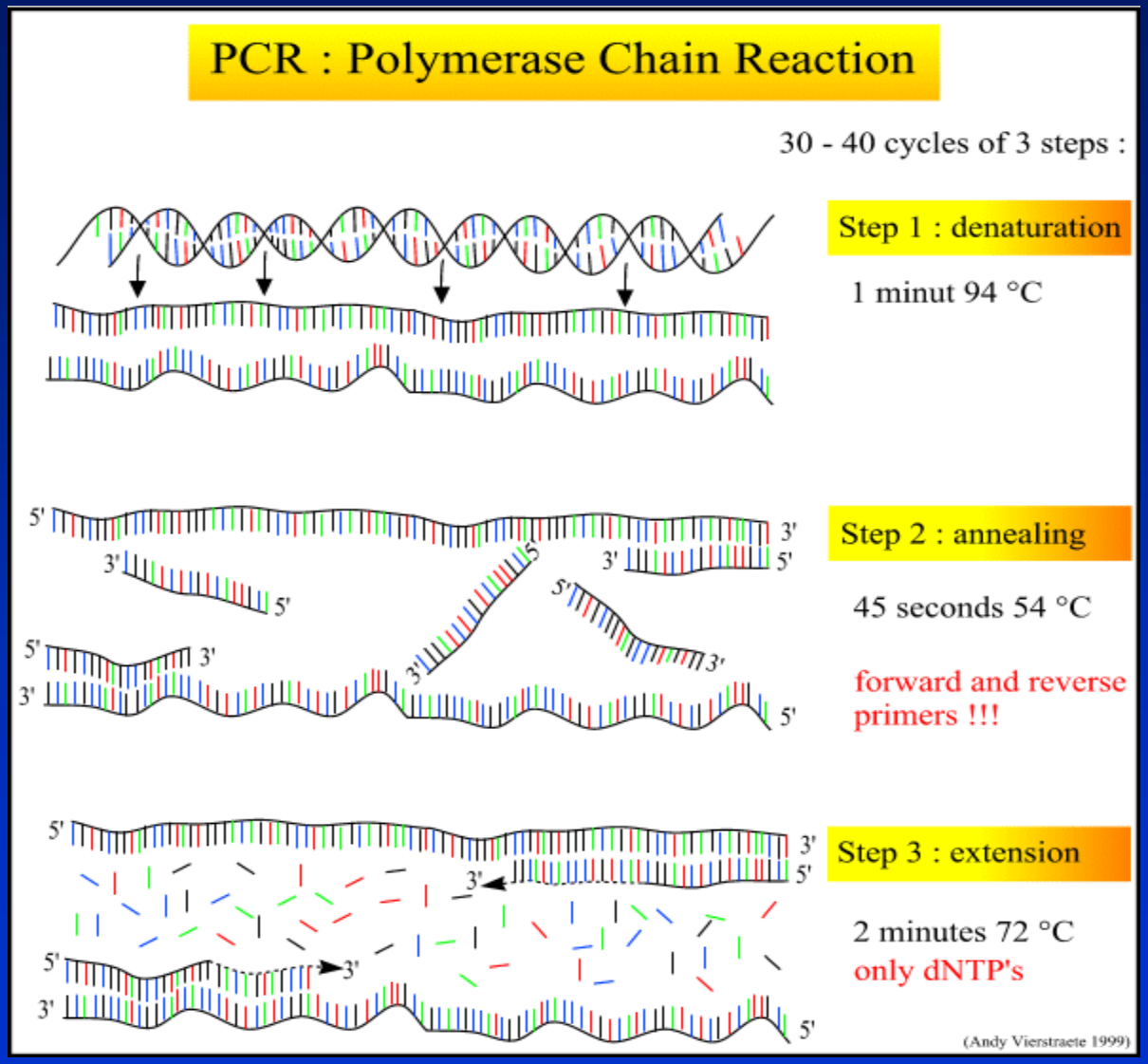
Genotyping via PCR amplification of Specific Targets

genomic DNA
+
primers
+
Taq DNA polymerase
+
dNTPs (ACGT)
+
buffer

From: Applied Biosystems

Thermocyclers (PCR)

PCR: Polymerase Chain Reaction



A 3 stage process that is temperature dependent:

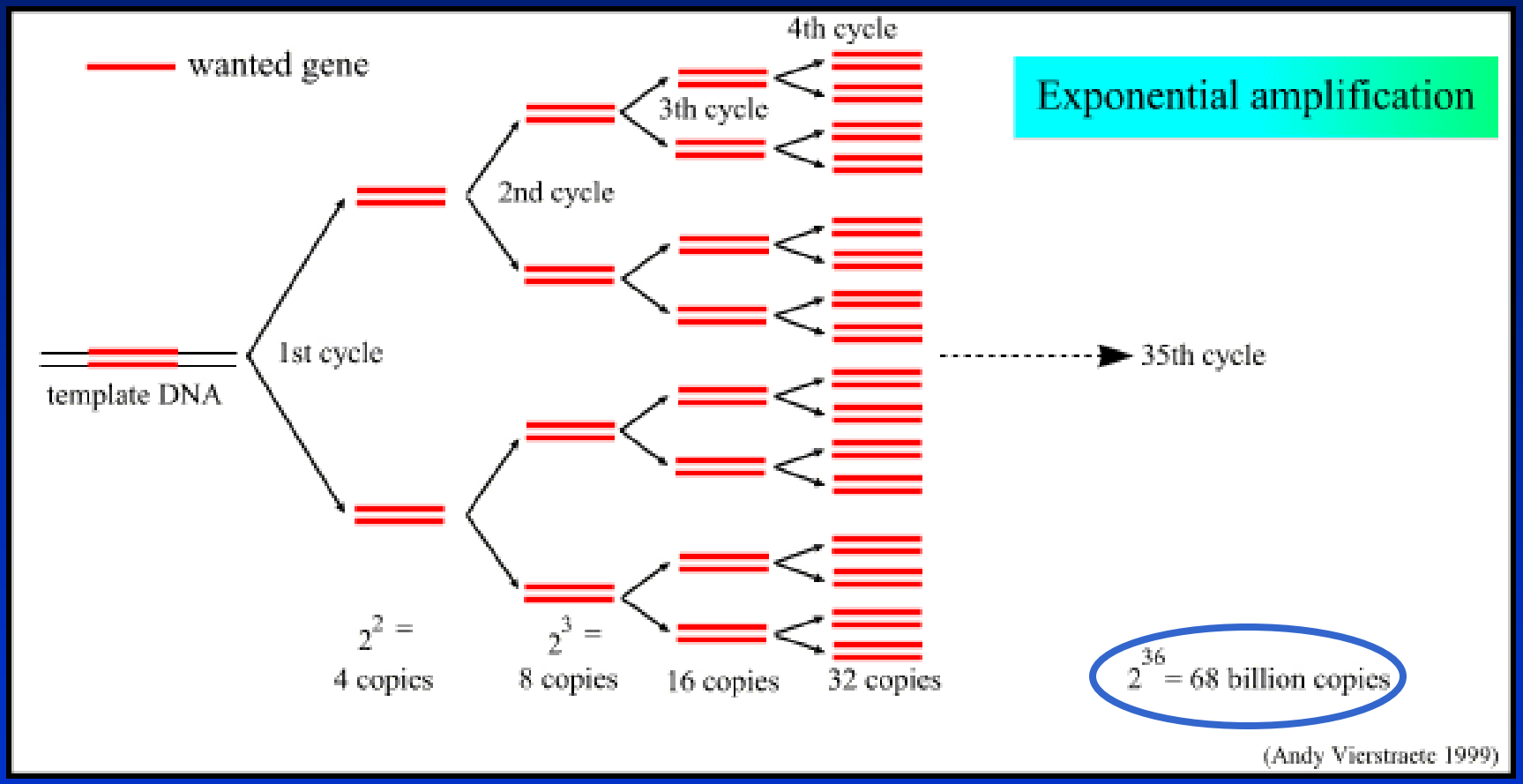
Step 1: with heat, the two DNA stands separate

Step 2: with cooling, primers bind to the DNA in the desired location

Step 3: with more heat, new nucleotides are added at the ends of the primers making 2 new DNA strands

http://allserv.rug.ac.be/~avierstr/principles/pcr.html

PCR amplification: A rags to riches story. . .

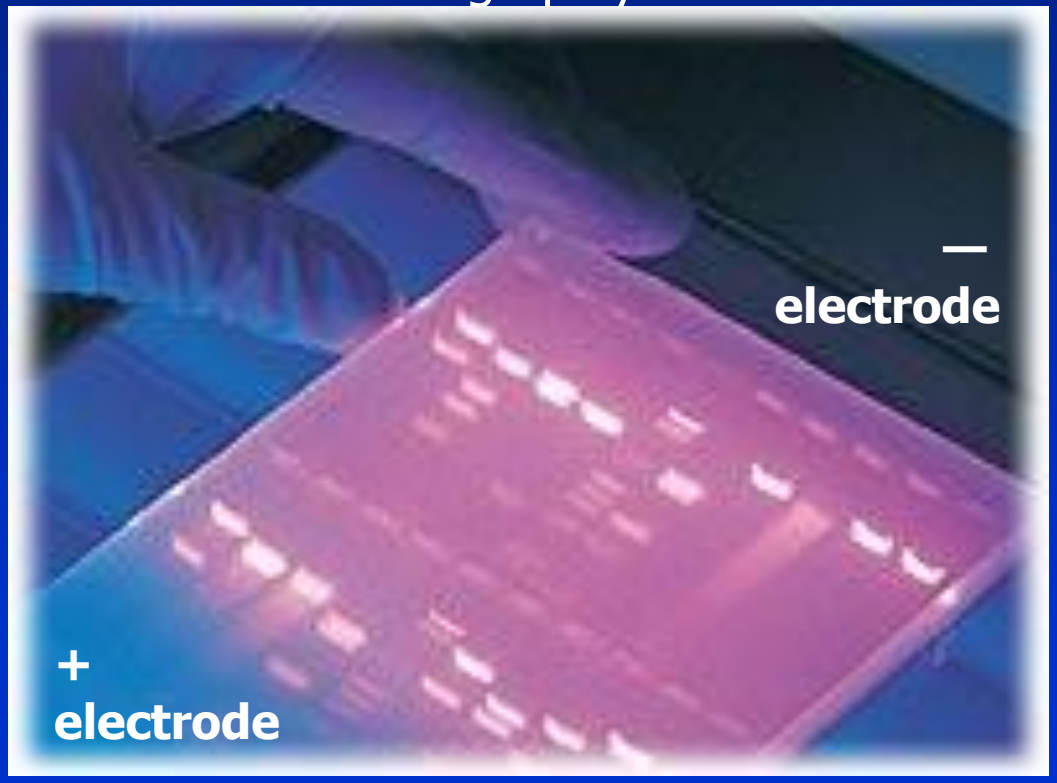


http://allserv.rug.ac.be/~avierstr/principles/pcr.html

Genotyping: Electrophoresis is used to Detect Length Differences

Agarose or polyacrylamide slab gel

- DNA is negatively charged
- Longer fragments migrate slower than shorter ones through polymer network.



DNA fragments are visualized by staining with ethidium bromide and UV light

Types of genetic differences between people

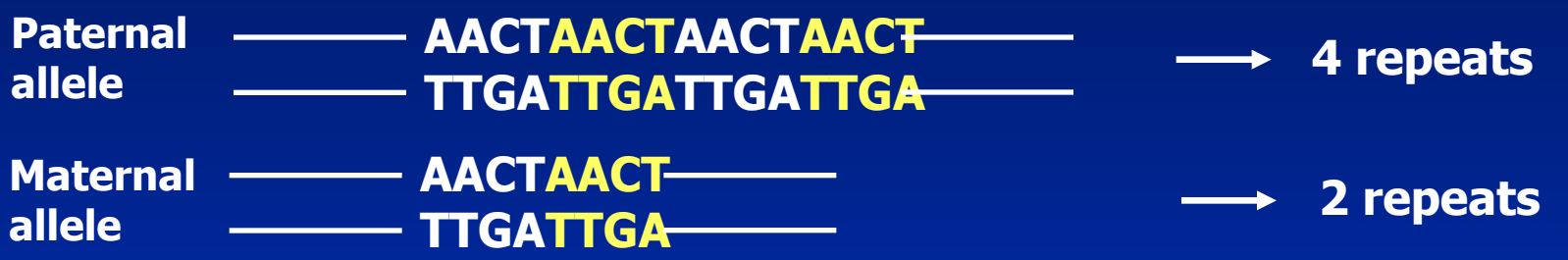
Any differences in DNA segments between any two persons can be used as genetic markers.
Fragments of DNA can be distinguished from one another because of differences in their nucleotide sequences.

Types:

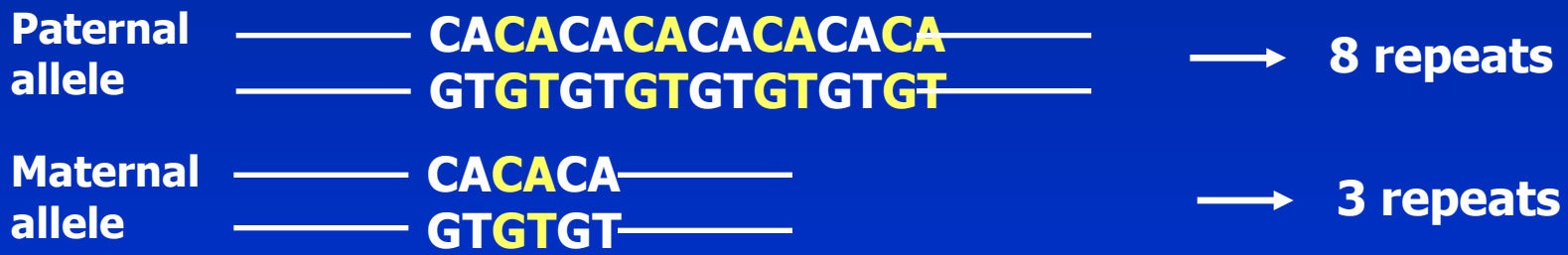
1. Single Nucleotide Polymorphisms SNPs. A single base pair change one strand of DNA. The most prevalent form of differences between any two individuals.
2. Minisatellites 10-100 nucleotides repeated several times in tandem; bordered by unique DNA sequences. Variable Number Tandem Repeats (VNTR) is an example. There are about 50,000 VNTRs in the human genome.
3. Microsatellites or Short Tandem Repeats (STRs). Smaller sequence repeats than minisatellites. Di- and tetra-nucleotide sequence repeats are common.

Two most common types of microsatellites

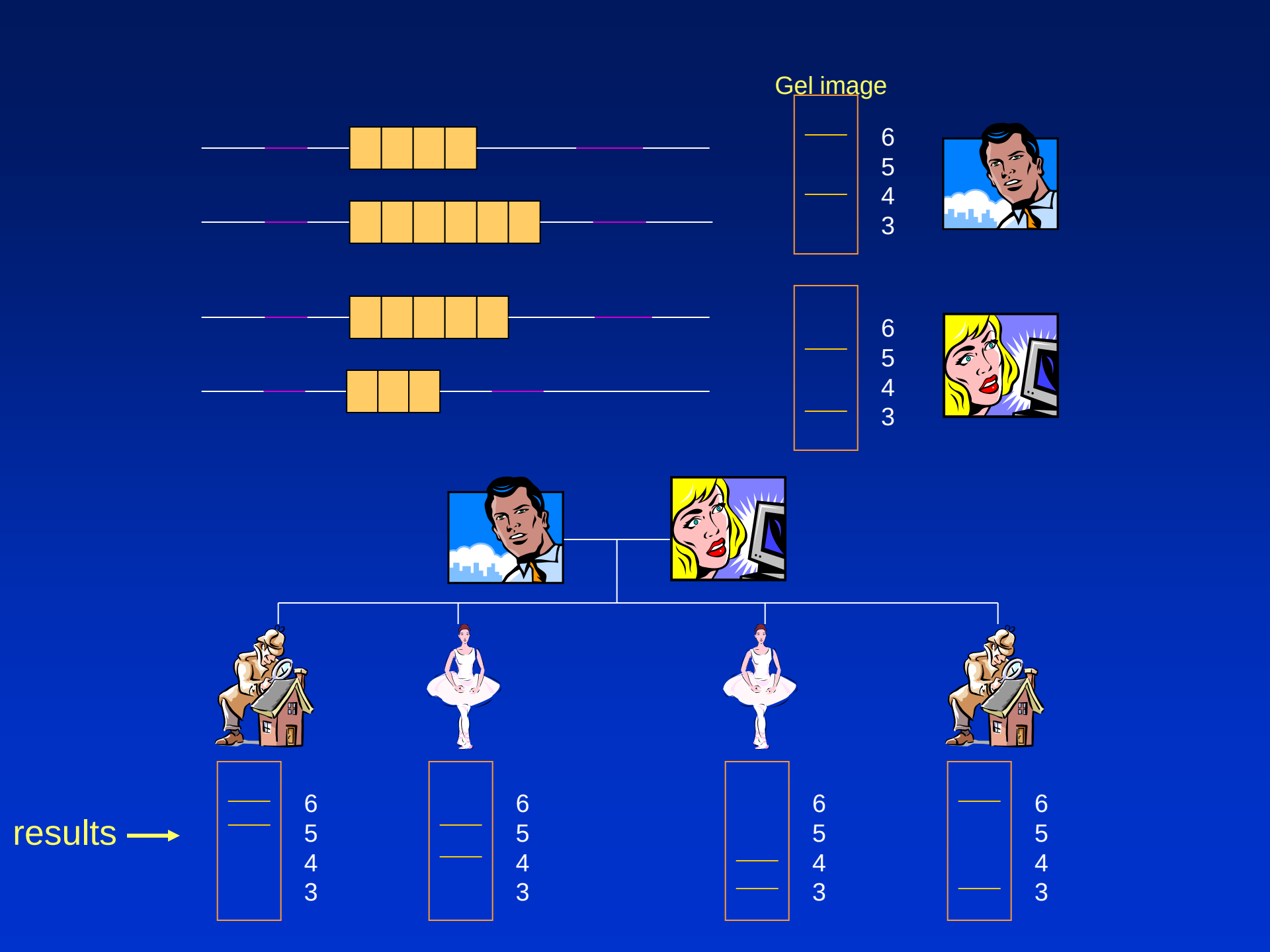
Tetranucleotide repeat (used in forensic applications and paternity):



Dinucleotide repeat (used extensively in gene mapping):



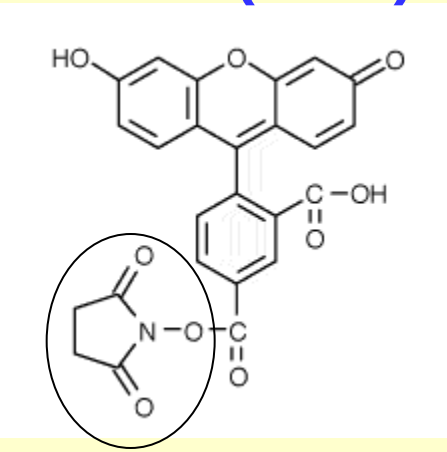
Minisatellites or VNTRs are fundamentally the same as microsatellites. The only difference being the length of the repeated nucleotides


Gel image
6
5
4
3
6
5
4
3
results
6
5
4
3
6
5
4
3
6
5
4
3
6
5
4
3

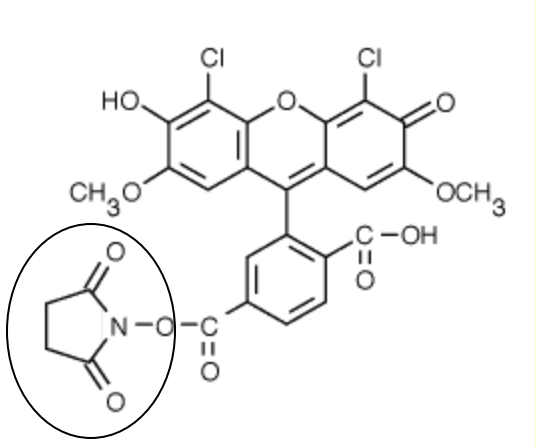
Fluorescent Dyes are used to visualize DNA fragments in automated sequencers

FL

FAM (Blue)

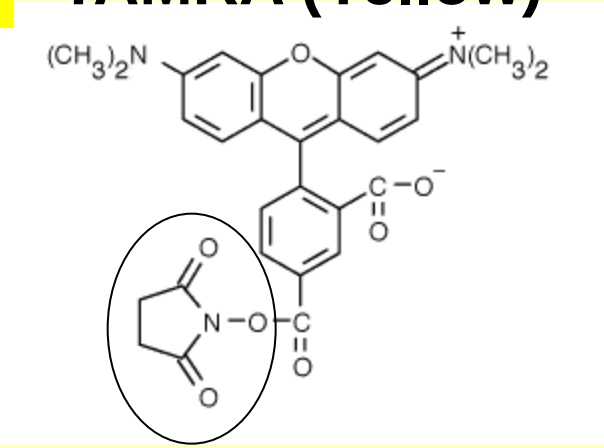


JOE (Green)

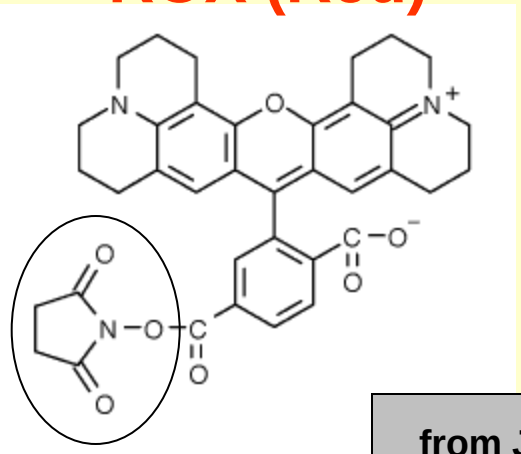


NED

TAMRA (Yellow)



ROX (Red)

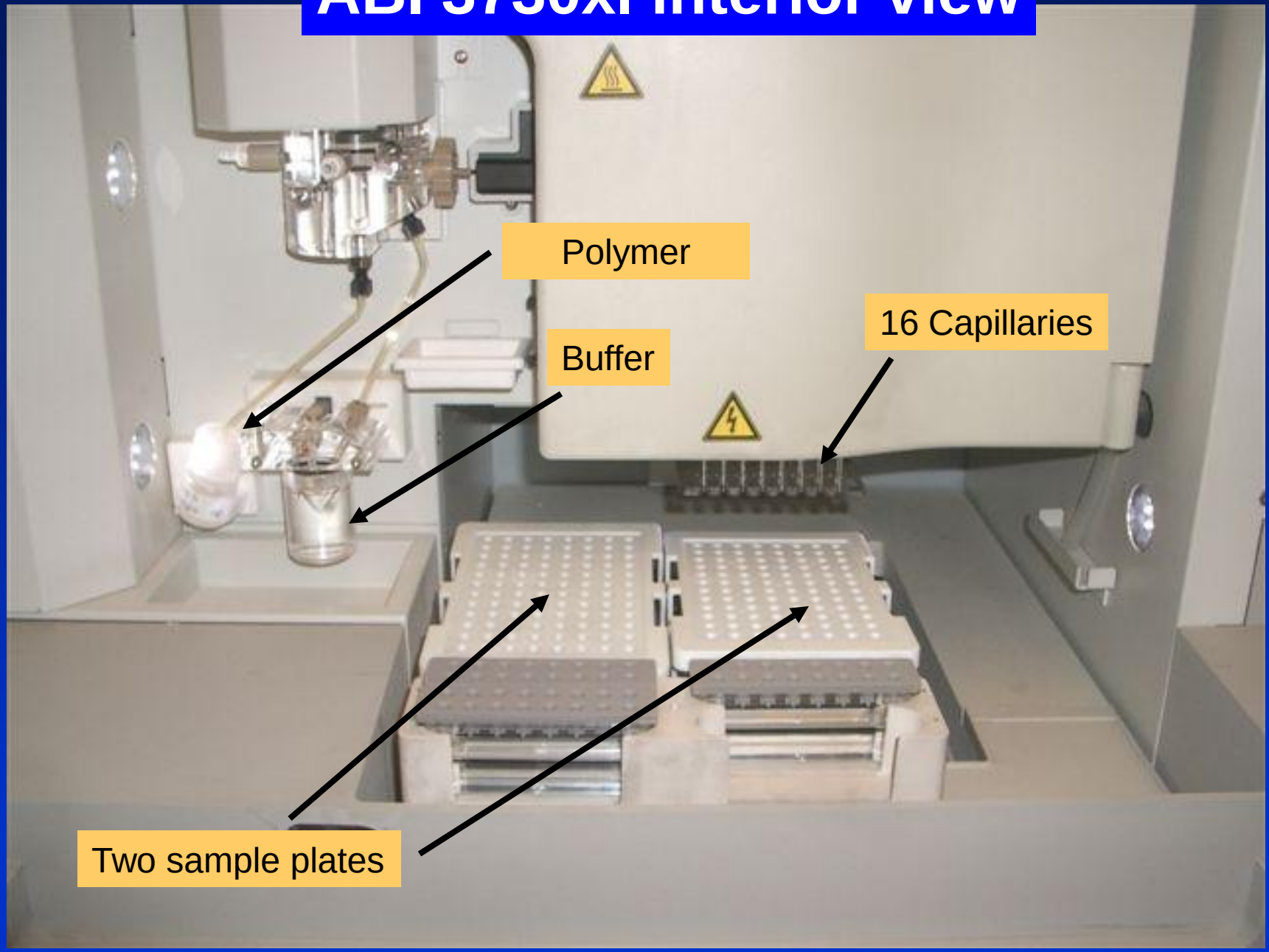


CXR

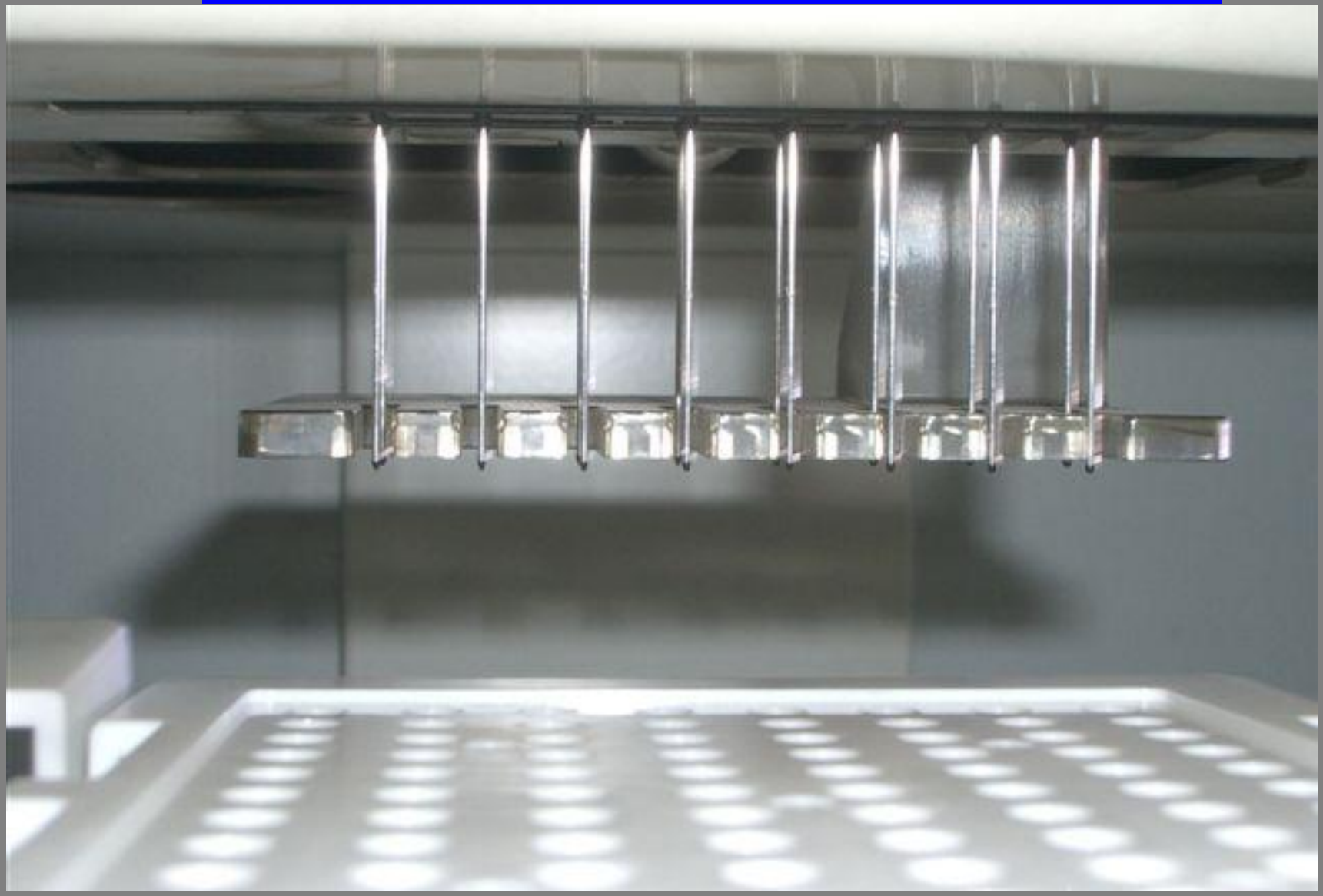
from John Butler - NIST

ABI 3730xl external view

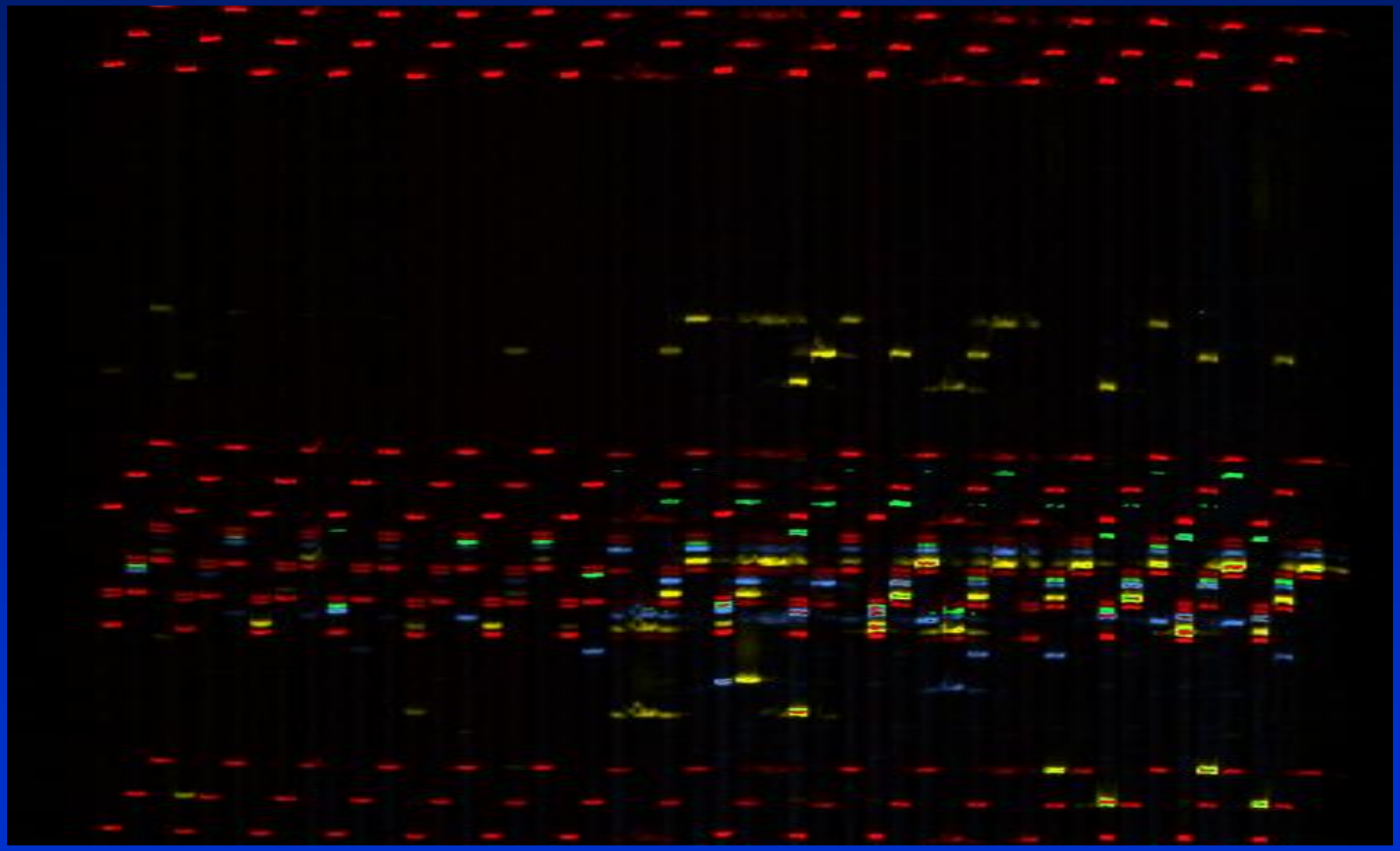
ABI 3730xl interior view



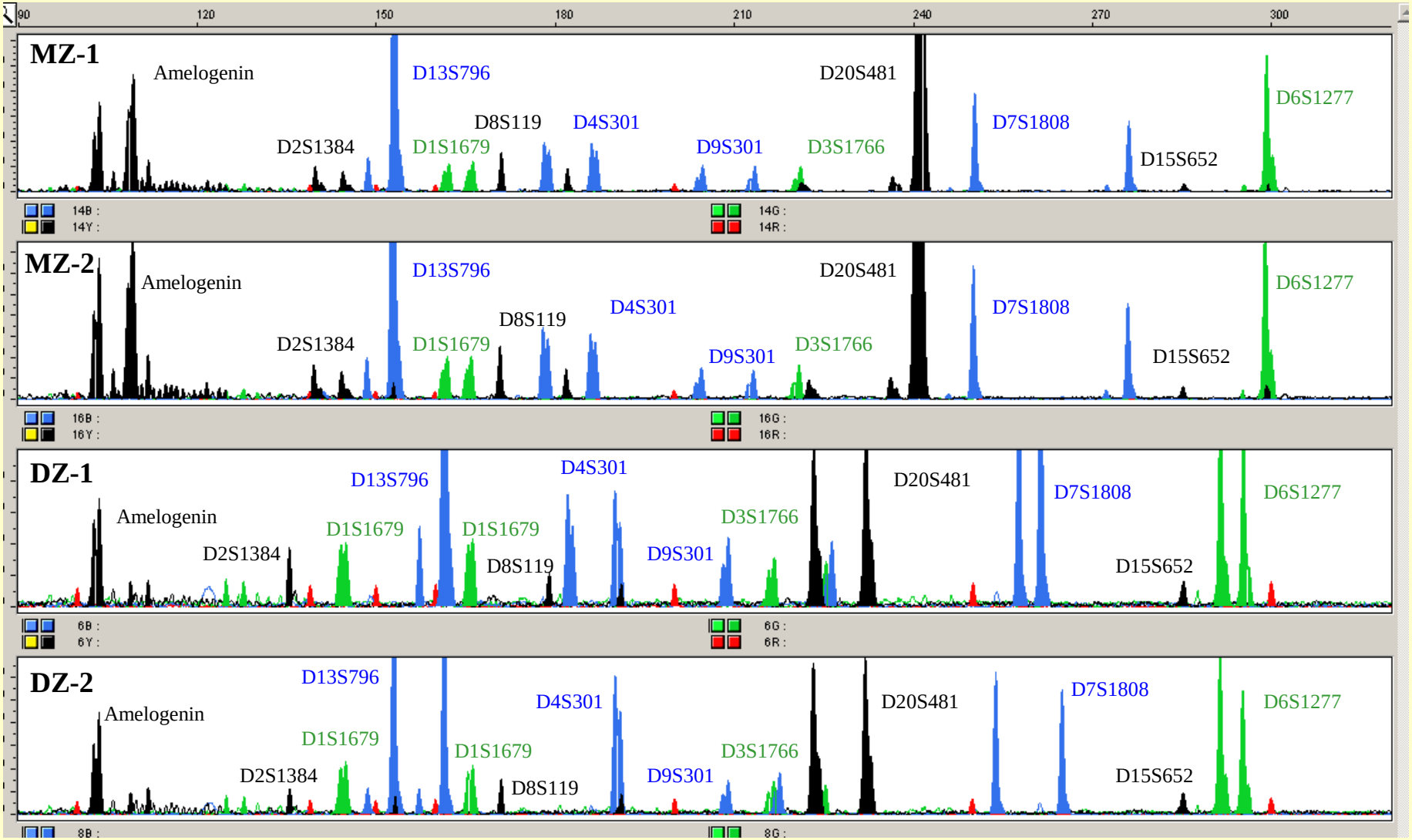
ABI 3130xl Closeup of 16 Capillaries



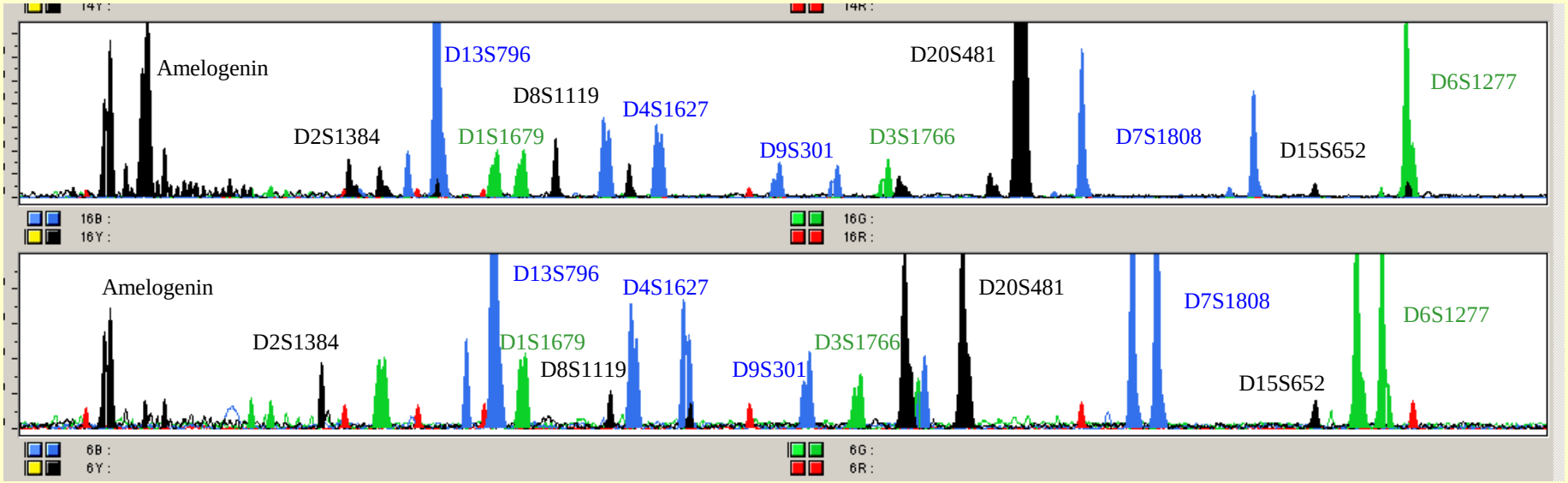
Genotyping STRs length differences: polyacrylamide gel



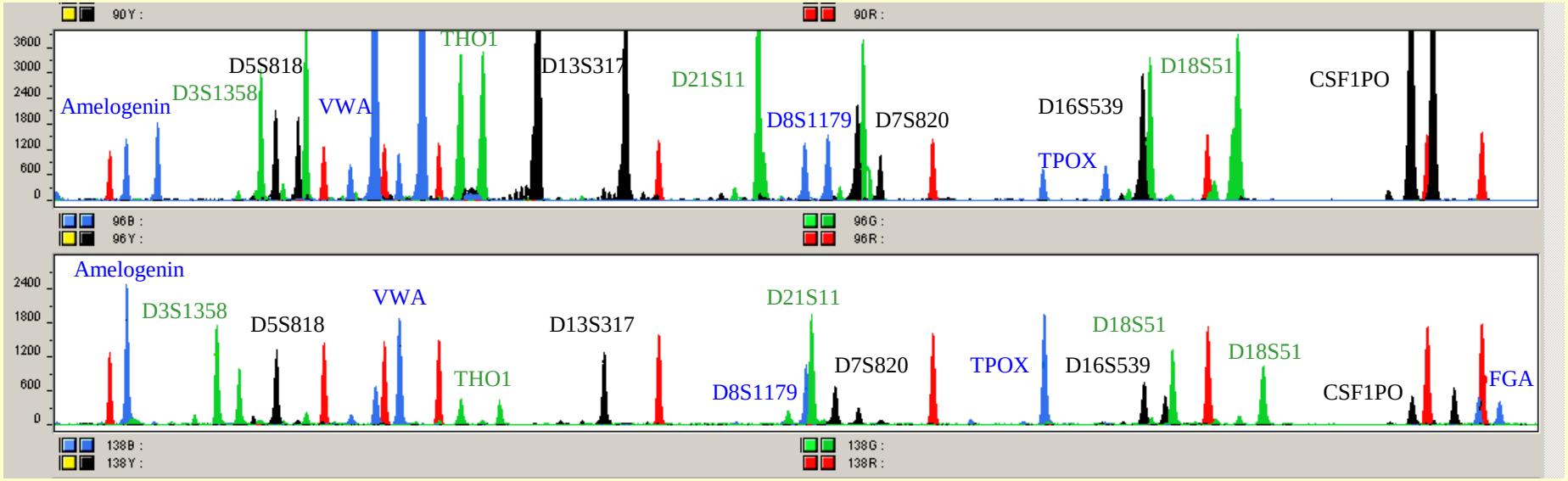
IBG-Hvar1


90
120
150
180
210
240
270
300
MZ-1
Amelogenin
D2S1384
D13S796
D1S1679
D8S119
D4S301
D9S301
D3S1766
D20S481
D7S1808
D15S652
D6S1277
14B :
14Y :
14G :
14R :
MZ-2
Amelogenin
D2S1384
D13S796
D1S1679
D8S119
D4S301
D9S301
D3S1766
D20S481
D7S1808
D15S652
D6S1277
16B :
16Y :
16G :
16R :
DZ-1
Amelogenin
D2S1384
D13S796
D1S1679
D1S1679
D8S119
D4S301
D9S301
D3S1766
D20S481
D7S1808
D15S652
D6S1277
6B :
6Y :
6G :
6R :
DZ-2
Amelogenin
D2S1384
D13S796
D1S1679
D1S1679
D8S119
D4S301
D9S301
D3S1766
D20S481
D7S1808
D15S652
D6S1277
8B :
8G :

IBG-Hvar1 (a zygosity panel)



IBG-Hvar2 (a CODIS panel)



Determination of Individual Identity

Locus Name	Expected Heterozygosity (Hexp)	**Probability for sharing both alleles, P(IBS=2)** **0.1013 + 2.1431*Hexp - 4.7086*Hexp2 +2.4723*Hexp3**
D3S1358	0.789	0.075241909
D5S818	0.698	0.143816331
VWA	0.810	0.061699124
THO1	0.756	0.098507981
D13S317	0.786	0.077262116
D8S1179	0.816	0.058034202
D21S11	0.835	0.047082087
D7S820	0.816	0.058034202
TPOX	0.637	0.194814466
D16S539	0.754	0.099986311
D18S51	0.880	0.025603386
CSF1PO	0.724	0.122943202
FGA	0.857	0.035746410

Determination of Individual Identity

Probability of a random match
(2 unrelated individuals sharing both alleles at all markers) =
0.00000000000023472

Percent of unrelated individuals expected to share
both alleles at all markers =
0.00000000002347173

Odds against one random individual being mistakenly identified
as another =
42,604,448,465,045

or more than 1 in 42 trillion.

This far exceeds the population of the earth,

which is somewhere around 6.5 billion persons.

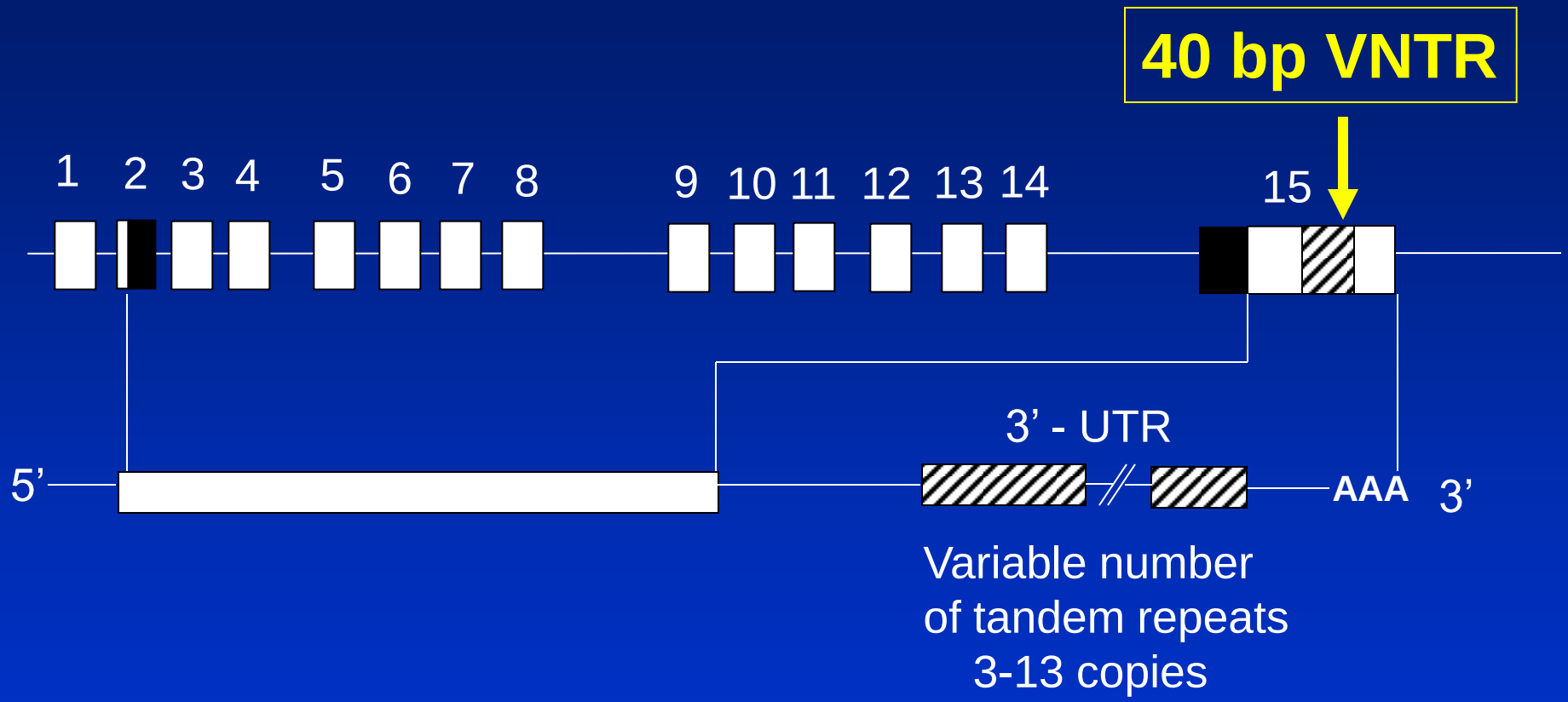
Types of genetic differences between people

Any differences in DNA segments between any two persons can be used as genetic markers.
Fragments of DNA can be distinguished from one another because of differences in their nucleotide sequences.

Types:

1. Single Nucleotide Polymorphisms SNPs. A single base pair change one strand of DNA. The most prevalent form of differences between any two individuals.
2. Minisatellites 10-100 nucleotides repeated several times in tandem; bordered by unique DNA sequences. Variable Number Tandem Repeats (VNTR) is an example. There are about 50,000 VNTRs in the human genome.
3. Microsatellites or Short Tandem Repeats (STRs). Smaller sequence repeats than minisatellites. Di- and tetra-nucleotide sequence repeats are common.

VNTR polymorphism of Dopamine Transporter



Adapted from Miller & Madras, 2002; Fuke et al., (2001), Vandenberg et al., 1992