Next Generation Sequencing Activities at NIST

NGS Workshop

Mid-Atlantic Association of Forensic Science

May 19, 2014

Katherine Butler Gettings, Ph.D.
Research Biologist, Applied Genetics Group
National Institute of Standards and Technology

Disclaimer

I will mention commercial STR kit names and information, but I am in no way attempting to endorse any specific products.

NIST Disclaimer: Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Points of view are mine and do not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Justice. **Our group receives or has received funding from the FBI Laboratory and the National Institute of Justice.**

Outline

Background

NGS of Forensic DNA markers

- STRs
- mtDNA
- Single Nucleotide Polymorphisms (SNPs)

NGS on the PGM- Ampliseq workflow

Experimental data

- HID-Ion Ampliseq Identity Panel
- HID-Ion Ampliseq Ancestry Panel

What's in a name???

Massively parallel sequencing

NGS

Second-generation sequencing

Next-generation sequencing

Whole-genome sequencing

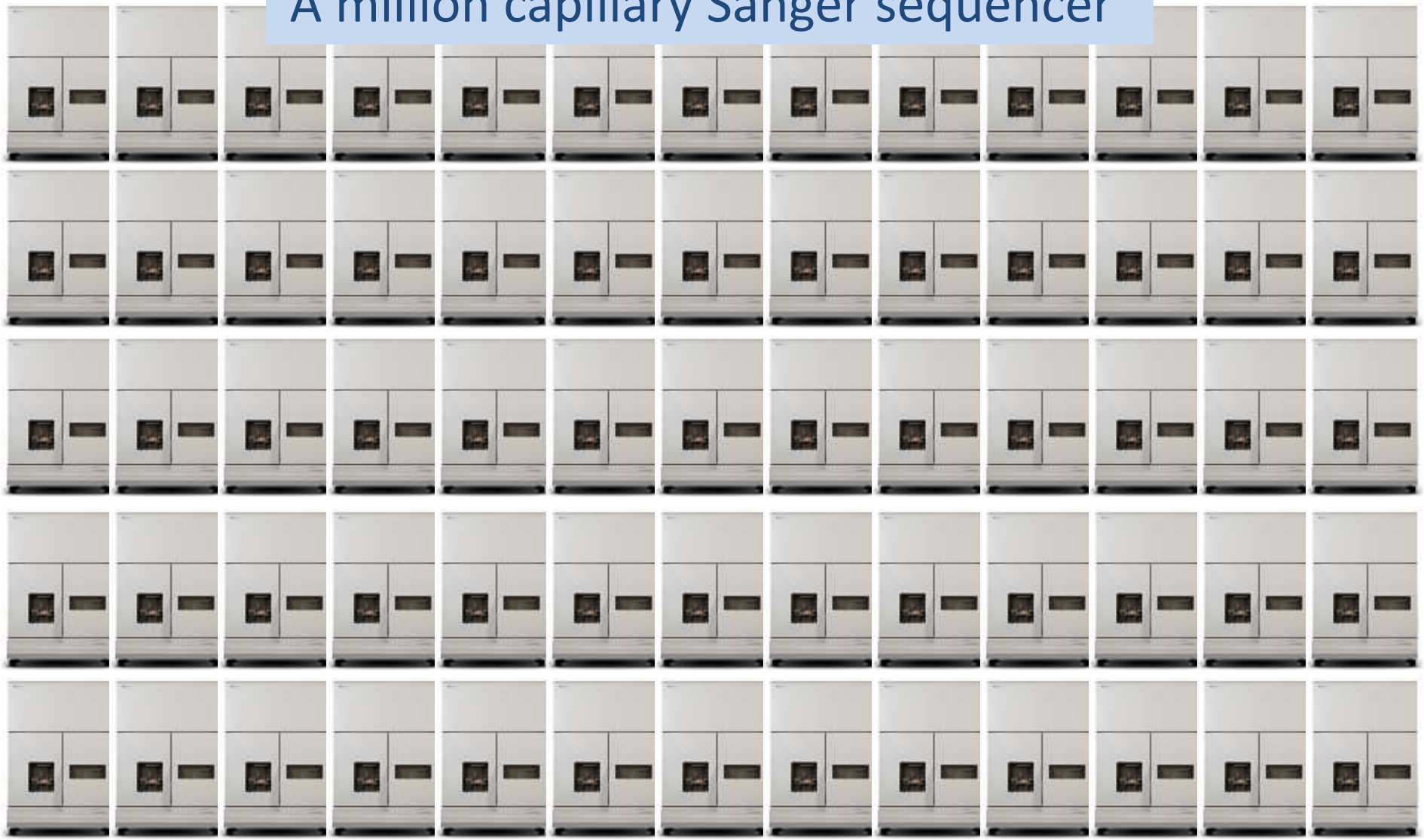
Third-generation sequencing

HIGH-THROUGHPUT SEQUENCING

Next-generation genomics

Parallel Sequencing

‘A million capillary Sanger sequencer’

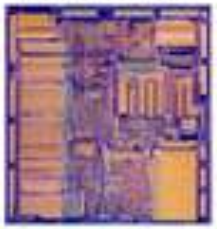


Parallel Sequencing

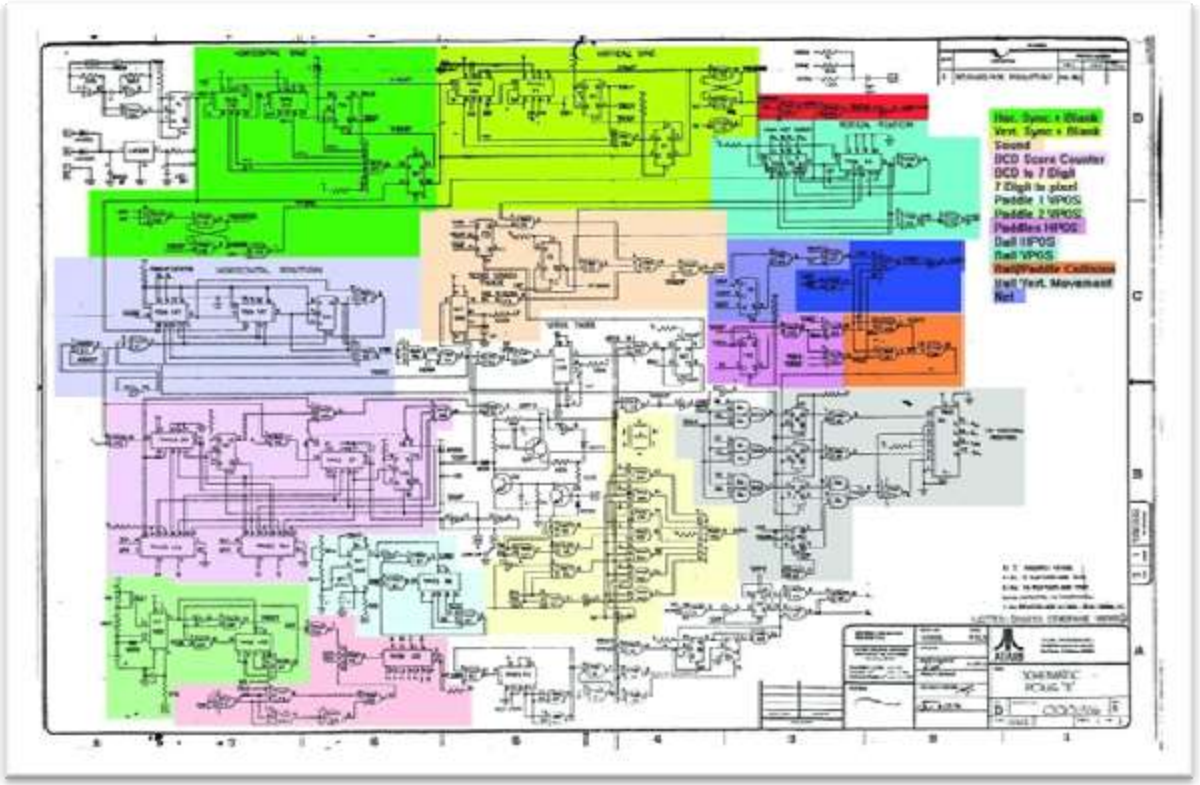
‘A million capillary Sanger sequencer’

- Clonal vs population amplification
- Shorter reads (Range 75 to 400)
- Errors are more ‘detectable’
- High coverage 100 – 1000 - 10,000x
- **Rely more on informatics to assemble millions of short reads**

MOORE'S LAW "Transistor density on integrated circuits doubles about every two years." *

1950s	1960s	1970s	1980s	1990s	2000s
Silicon Transistor	TTL Quad Gate	8-bit Microprocessor	32-bit Microprocessor	32-bit Microprocessor	64-bit Microprocessor
					
1 Transistor	**16 Transistors**	**4500 Transistors**	**275,000 Transistors**	**3,100,000 Transistors**	**592,000,000 Transistors**

http://electronicsbyexamples.blogspot.com/2013/03/milestones-in-digital-electronics.html


Hor. Sync + Blank
Vert. Sync + Blank
Sound
BCD Score Counter
BCD to 7 Digit
7 Digit to pixel
Paddle 1 VPOS
Paddle 2 VPOS
Paddles HPOS
Ball HPOS
Ball VPOS
Ball/Paddle Collision
Ball Vert. Movement
Net

VISUALIZING PROGRESS

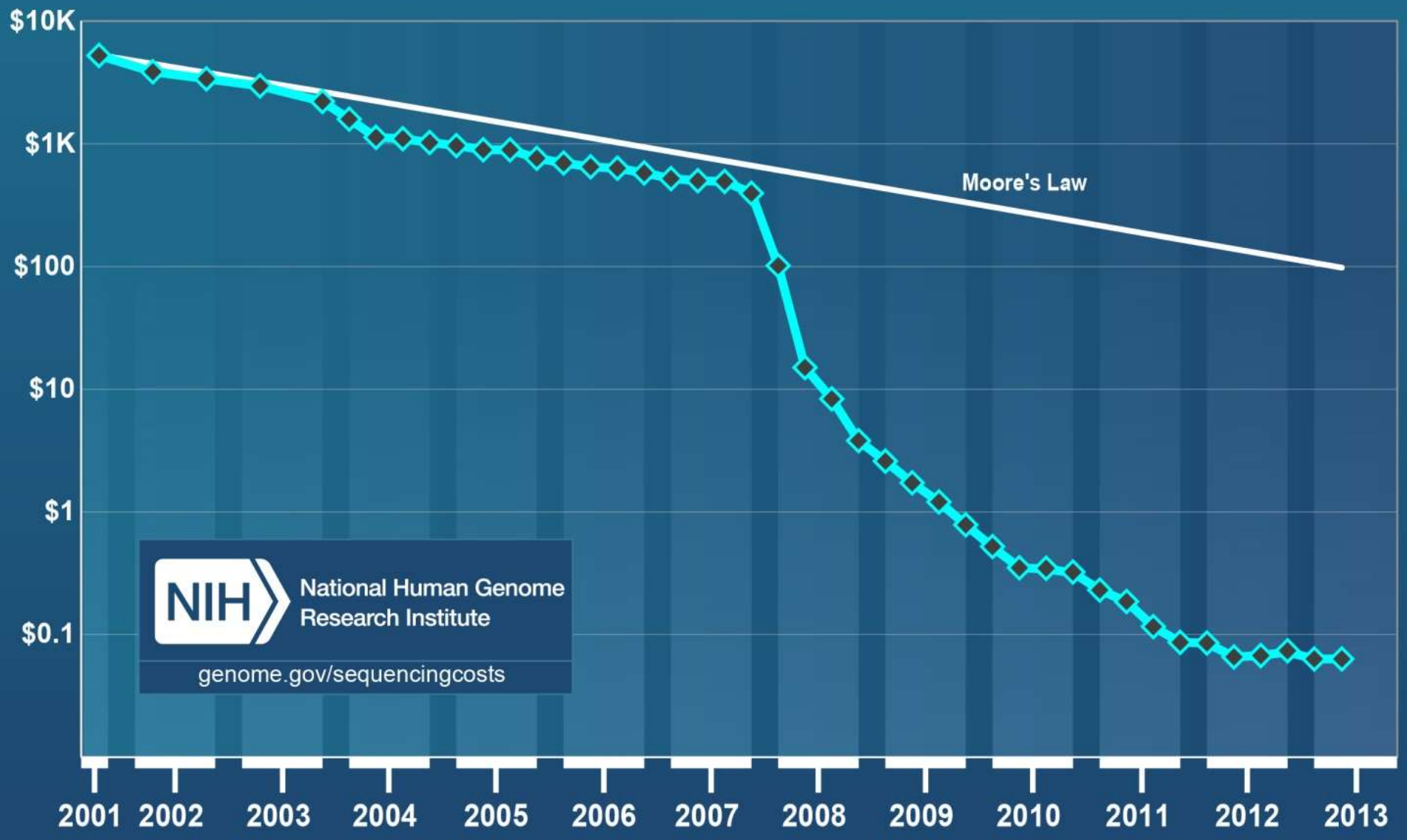
If transistors were people

If the transistors in a microprocessor were represented by people, the following timeline gives an idea of the pace of Moore's Law.

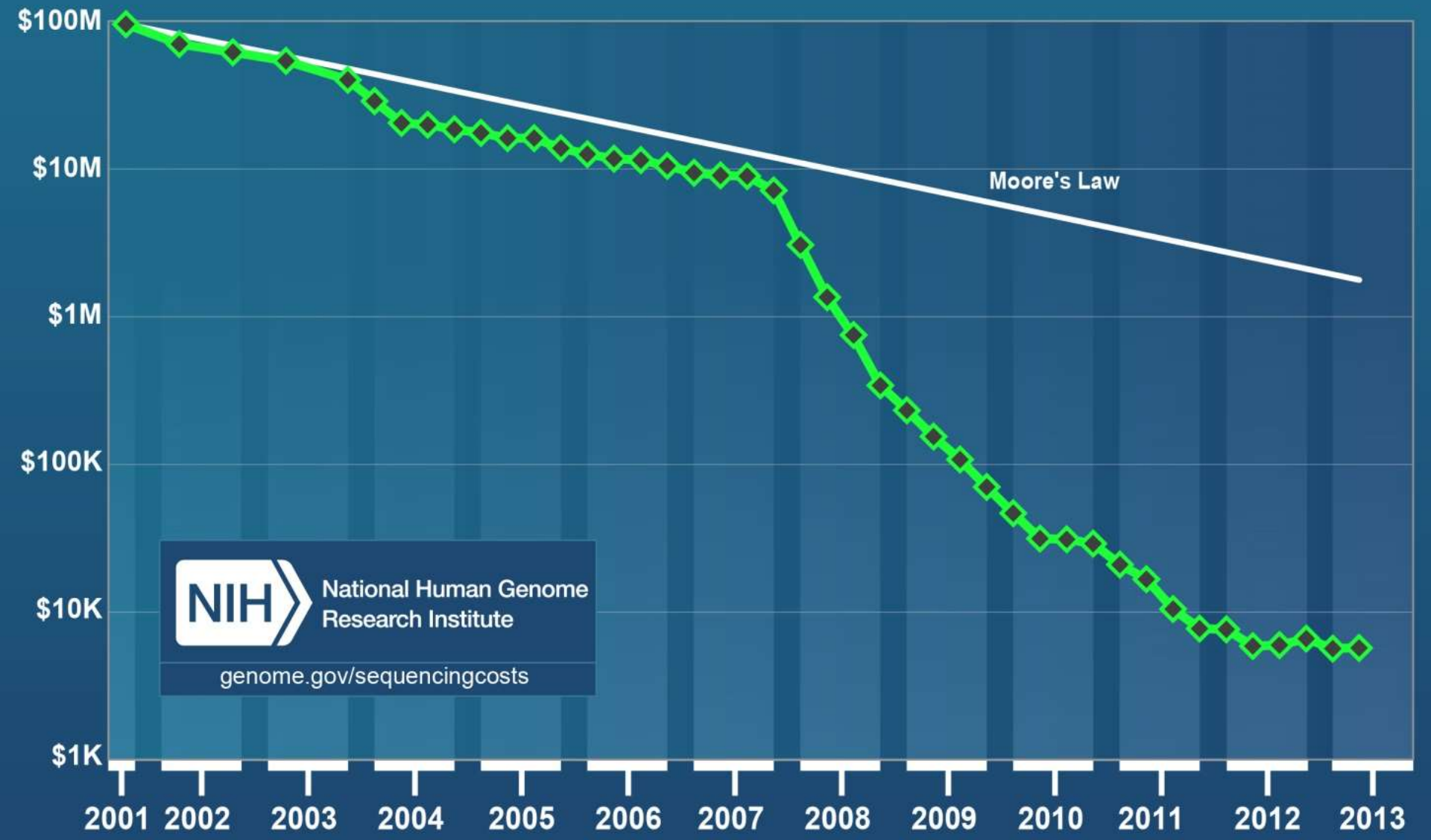
Now imagine that those 1.3 billion people could fit onstage in the original music hall. That's the scale of Moore's Law.

http://blog.trentonsystems.com/moores-law-pushing-processor-technology-to-14-nanometers/

Cost per Raw Megabase of DNA Sequence



Cost per Genome

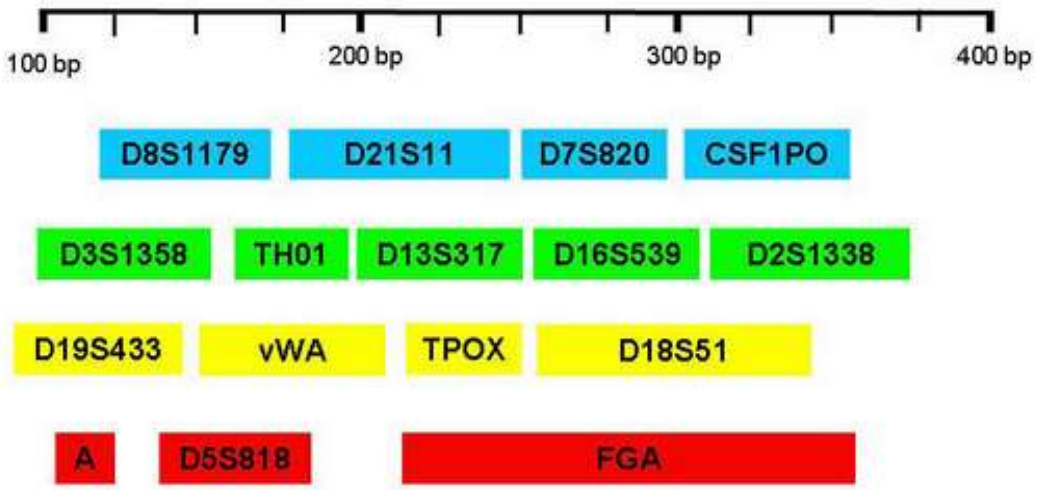


Forensic NGS Applications

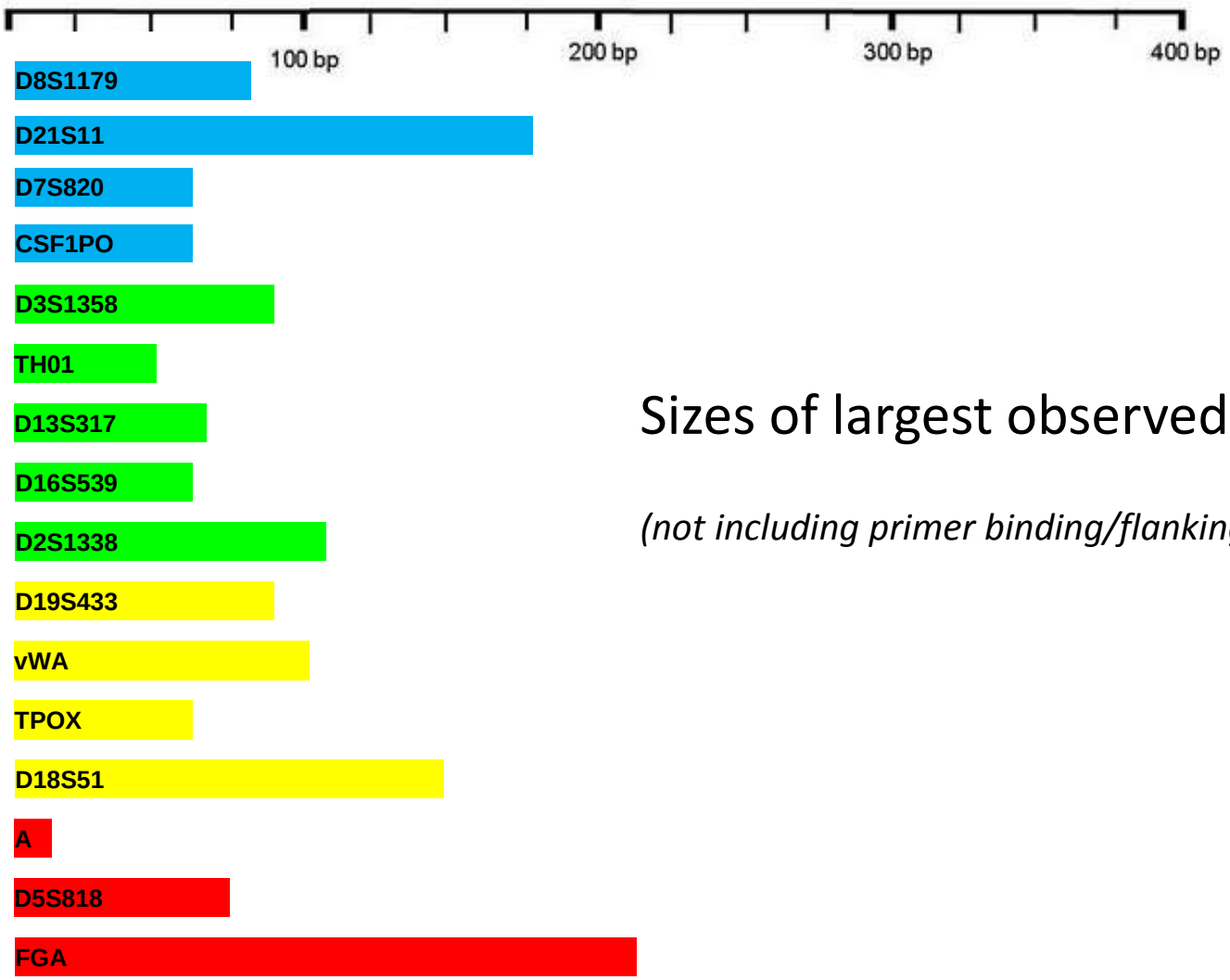
- Short Tandem Repeats (STRs)
  - PCR fragment-length polymorphisms
- Mitochondrial DNA (mtDNA)
  - Sanger sequencing
- Single Nucleotide Polymorphisms (SNPs)

Capillary electrophoresis electropherogram

NGS of Forensic STR Loci



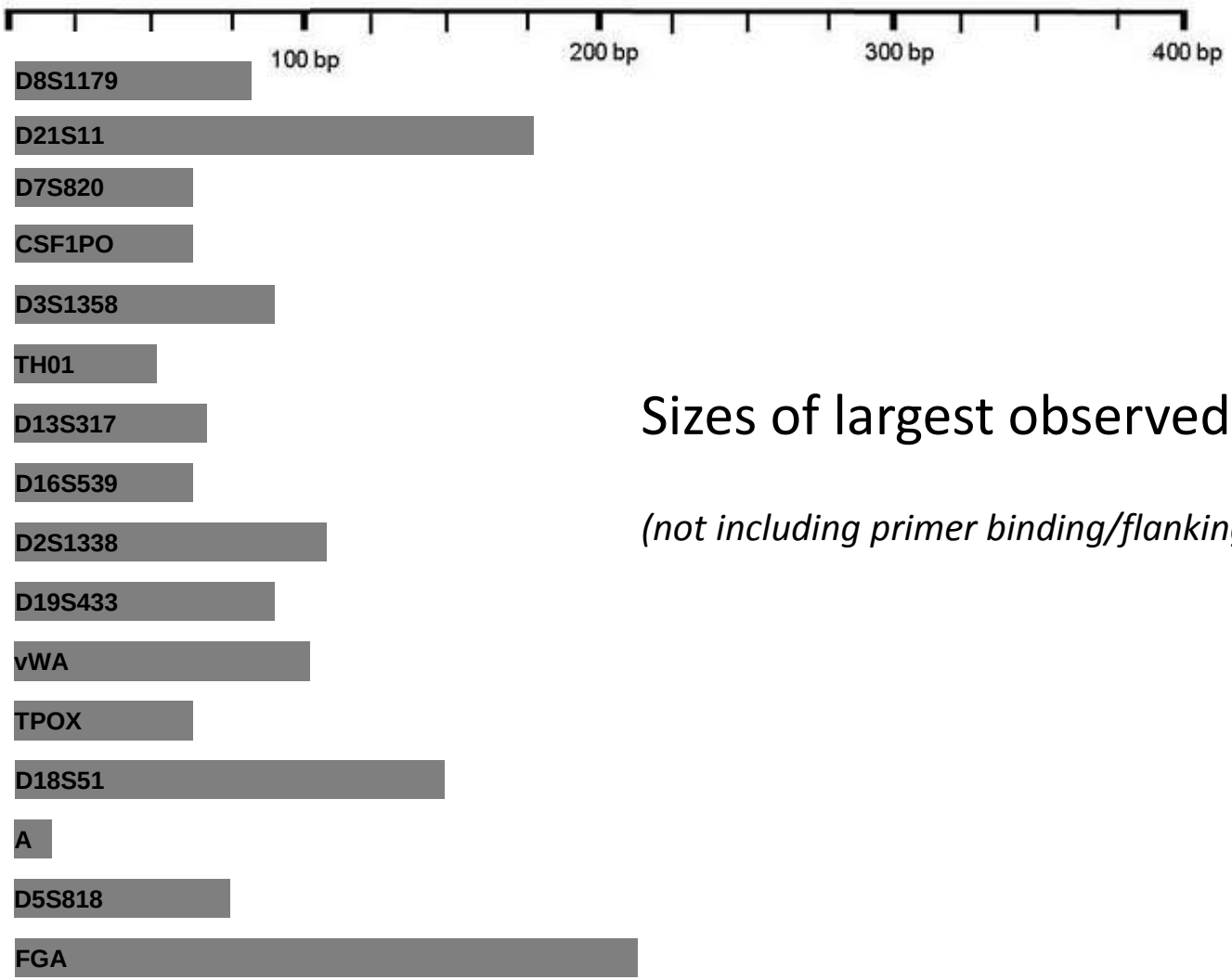
NGS of Forensic STR Loci



Sizes of largest observed alleles

(not including primer binding/flanking region)

NGS of Forensic STR Loci



Sizes of largest observed alleles

(not including primer binding/flanking region)

NGS of Forensic STR Loci

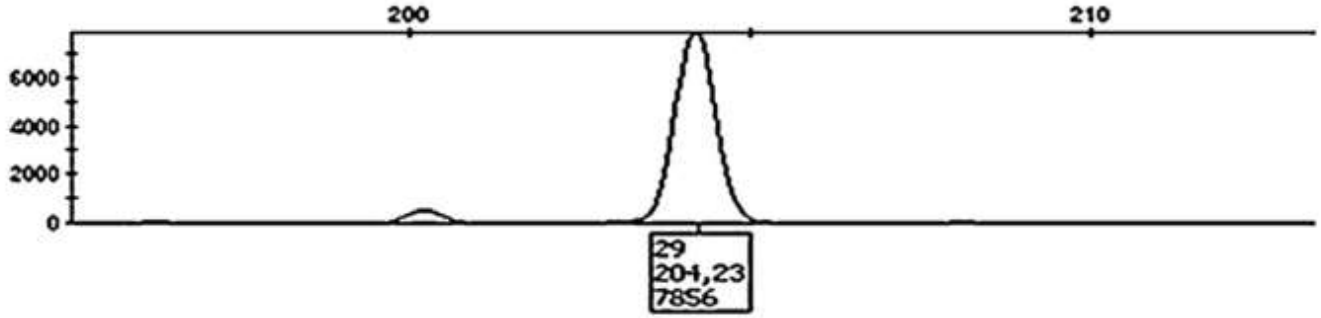
Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing

Eszter Rockenbauer [a,1,*], Stine Hansen [b,1], Martin Mikkelsen [a], Claus Børsting [a], Niels Morling [a]

Allele sequences	Allele length
[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	(29)
[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	(29)



D21S11: Individual appears homozygous by CE but different sequencing composition shown with NGS.

NGS of Forensic STR Loci

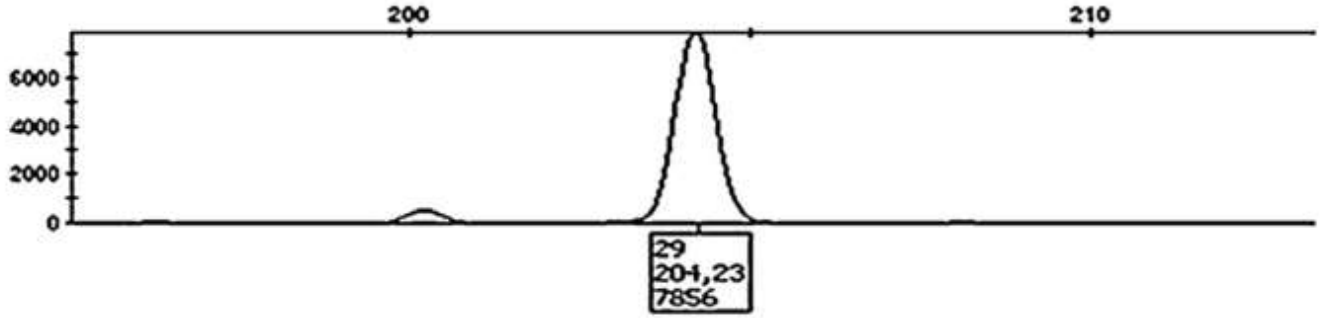
Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing

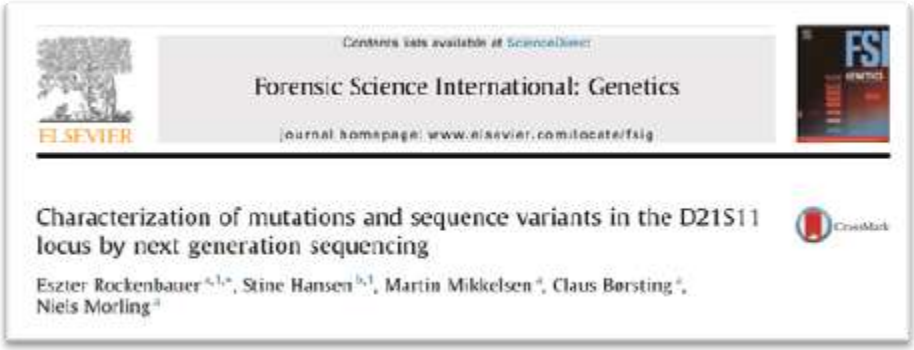
Eszter Rockenbauer [a,1,*], Stine Hansen [b,1], Martin Mikkelsen [a], Claus Børsting [a], Niels Morling [a]

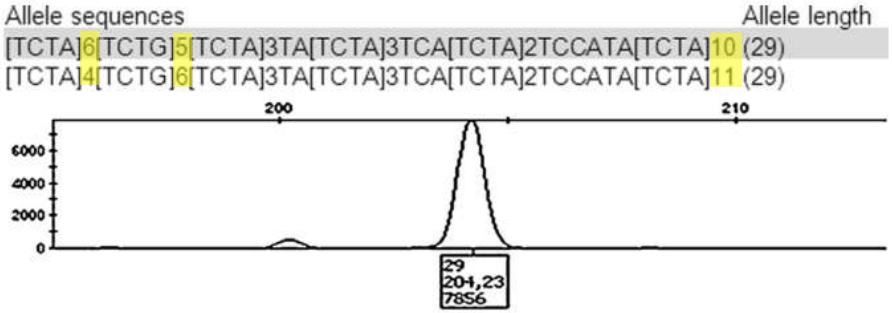
Allele sequences	Allele length
[TCTA]6[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	(29)
[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	(29)

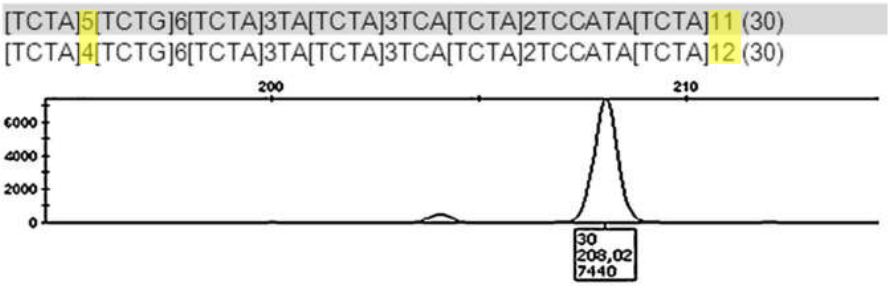


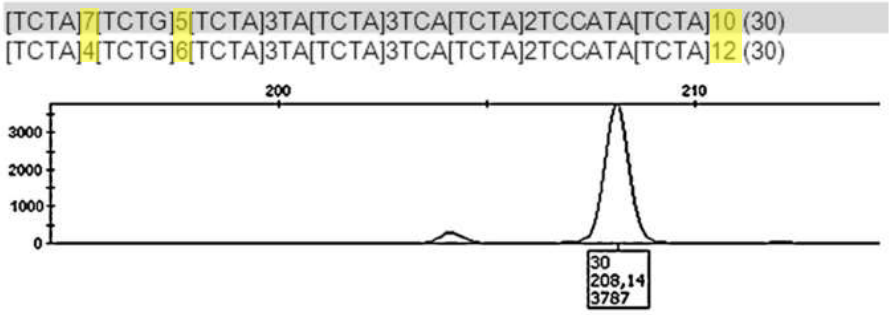
D21S11: Individual appears homozygous by CE but different sequencing composition shown with NGS.

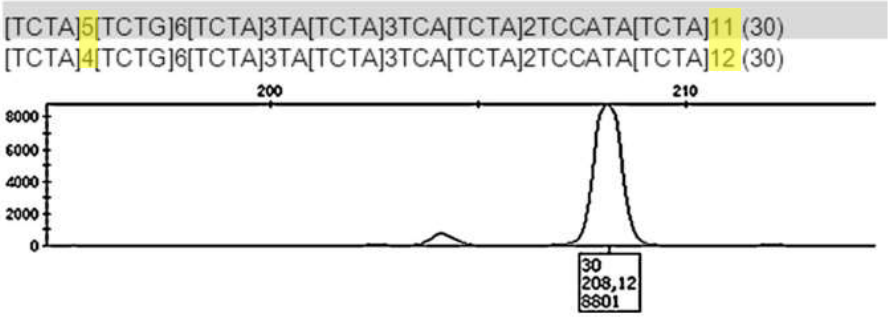
NGS of Forensic STR Loci











D21S11: Individuals appear homozygous by CE but different sequencing composition shown with NGS.

NGS of Forensic STR Loci

STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data

David H. Warshauer [a], David Lin [b], Kumar Hari [b], Ravi Jain [b], Carey Davis [a], Bobby LaRue [a], Jonathan L. King [a], Bruce Budowle [a,c,*]

AGCTCGATCGCGCTAGTTTAGAGCattgactctatcGATAGATAGATAGATAGATAGATAGATAGATAcctgatagtattGATTCTATCGATATATCCTAGCTA

~~ATAGATAGATAGATAGATAcctgatagtattGATTCTATCGATATATCCTAGCTAGAATCTCGGTGATCTAGCTATTTATCCGCTAGCCATATGTGCGATCGAATC~~

~~GCCTAGCTAGCTTAGCTATATCTAGATAGCTAGCTAGCTTTCGCGAAAGCTCGATCGCGCTAGTTTAGAGCattgactctatcGATAGATAGATAGATAGATAGA~~

ATCGCGCTAGTTTAGAGCattaactctatcGATAGATAGATAGATAGATAGATAGATAGATAcctgacagtattGATTCTATCGATATATCCTAGCTAGAATCT

* *

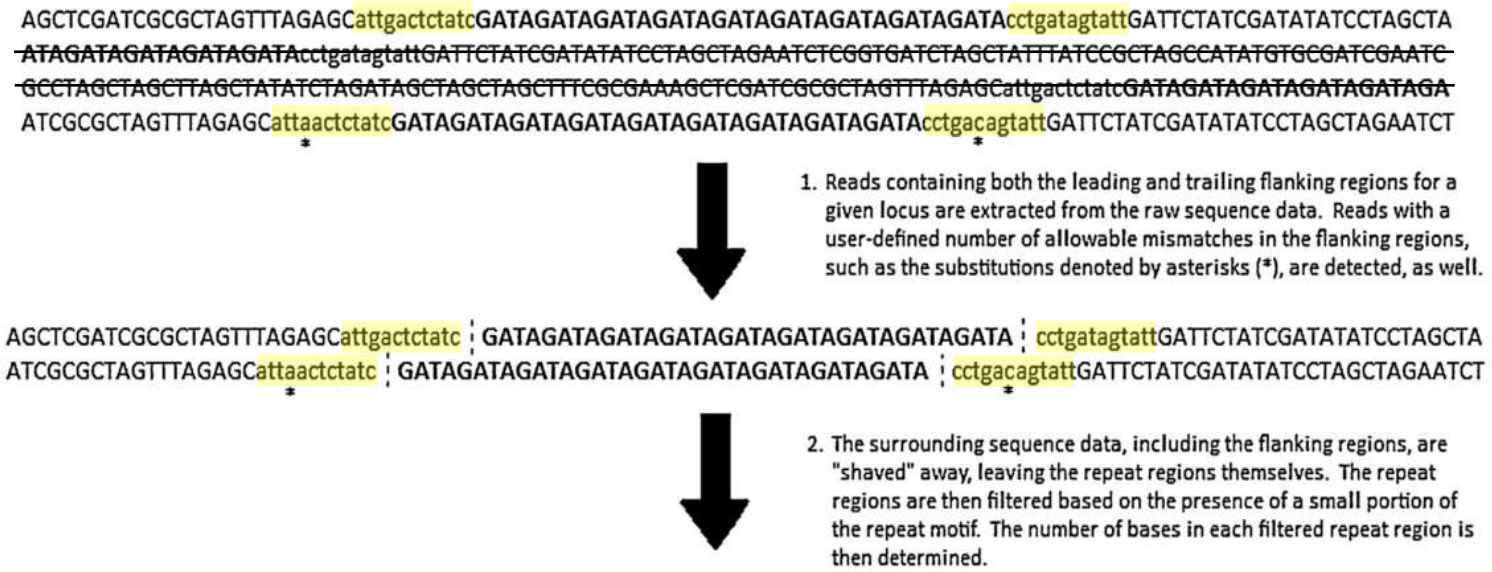
1. Reads containing both the leading and trailing flanking regions for a given locus are extracted from the raw sequence data. Reads with a user-defined number of allowable mismatches in the flanking regions, such as the substitutions denoted by asterisks (*), are detected, as well.

NGS of Forensic STR Loci

STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data

David H. Warshauer[a], David Lin[b], Kumar Hari[b], Ravi Jain[b], Carey Davis[a], Bobby LaRue[a], Jonathan L. King[a], Bruce Budowle[a,c,*]

D.H. Warshauer et al. / Forensic Science International: Genetics 7 (2013) 409–417



NGS of Forensic STR Loci

STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data

David H. Warshauer[a], David Lin[b], Kumar Hari[b], Ravi Jain[b], Carey Davis[a], Bobby LaRue[a], Jonathan L. King[a], Bruce Budowle[a,c,*]

D.H. Warshauer et al. / Forensic Science International: Genetics 7 (2013) 409–417

AGCTCGATCGCGCTAGTTTAGAGCattgactctatc**GATAGATAGATAGATAGATAGATAGATAGATAGATA**cctgatagtattGATTCTATCGATATATCCTAGCTA
ATAGATAGATAGATAGATAcctgatagtattGATTCTATCGATATATCCTAGCTAGAATCTCGGTGATCTAGCTATTTATCCGCTAGCCATATGTGCGATCGAATC
GCCTAGCTAGCTTAGCTATATCTAGATAGCTAGCTAGCTTTCGCGAAAGCTCGATCGCGCTAGTTTAGAGCattgactctatc**GATAGATAGATAGATAGATAGA**
ATCGCGCTAGTTTAGAGCattaactctatc**GATAGATAGATAGATAGATAGATAGATAGATAGATA**cctgacagtattGATTCTATCGATATATCCTAGCTAGAATCT

1. Reads containing both the leading and trailing flanking regions for a given locus are extracted from the raw sequence data. Reads with a user-defined number of allowable mismatches in the flanking regions, such as the substitutions denoted by asterisks (*), are detected, as well.

AGCTCGATCGCGCTAGTTTAGAGCattgactctatc | **GATAGATAGATAGATAGATAGATAGATAGATAGATA** | cctgatagtattGATTCTATCGATATATCCTAGCTA
ATCGCGCTAGTTTAGAGCattaactctatc | **GATAGATAGATAGATAGATAGATAGATAGATAGATA** | cctgacagtattGATTCTATCGATATATCCTAGCTAGAATCT

2. The surrounding sequence data, including the flanking regions, are "shaved" away, leaving the repeat regions themselves. The repeat regions are then filtered based on the presence of a small portion of the repeat motif. The number of bases in each filtered repeat region is then determined.

GATAGATAGATAGATAGATAGATAGATAGATAGATA
GATAGATAGATAGATAGATAGATAGATAGATAGATA

(36 bases long) (GATA repeat motif)

3. Allele determination is performed by comparing the repeat region lengths to the known repeat motif.

Allele = 9

NGS of Forensic STR Loci

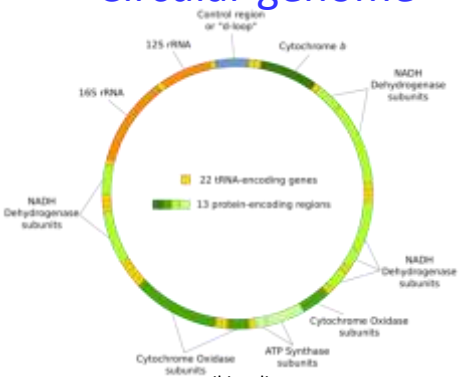
STR sequence data from 2391c, Component A:

- Truseq Library Prep
- MiSeq sequencing
- STRait Razor data parsing
- R script (NIST) data viewer

Forensic DNA Markers

- Short Tandem Repeats (STRs)
  - PCR fragment-length polymorphisms
- Mitochondrial DNA (mtDNA)
  - Sanger sequencing
- Single Nucleotide Polymorphisms (SNPs)

Circular genome

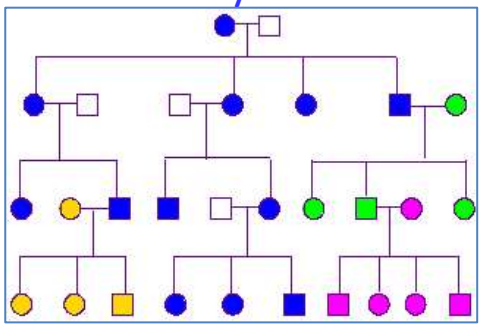


www.wikipedia.org

Mitochondria

http://remf.dartmouth.edu/images/mammalianLungTEM/source/8.html

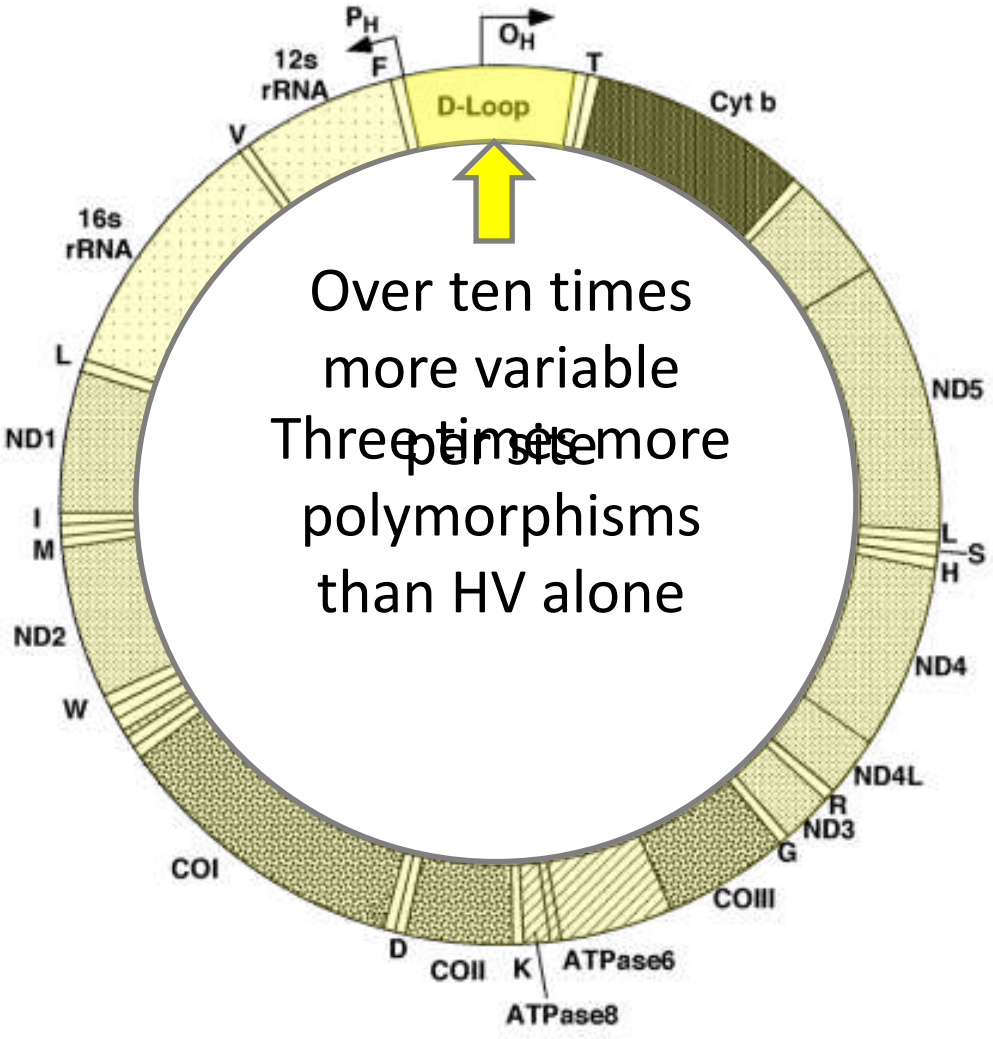
Maternally inherited



http://www.orchidcellmark.ca

mtDNA Information

- Increase in variants by whole genome analysis



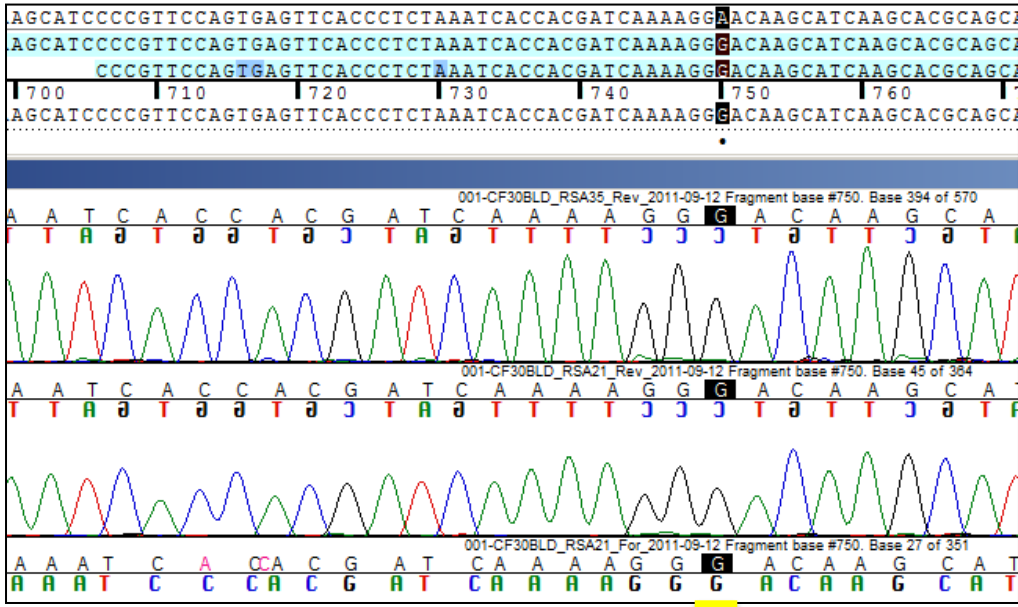
(based on analysis of 3 SRM samples)

http://www.sas.upenn.edu/~tgschurr/labwork/labwork_text.html

mtDNA Information

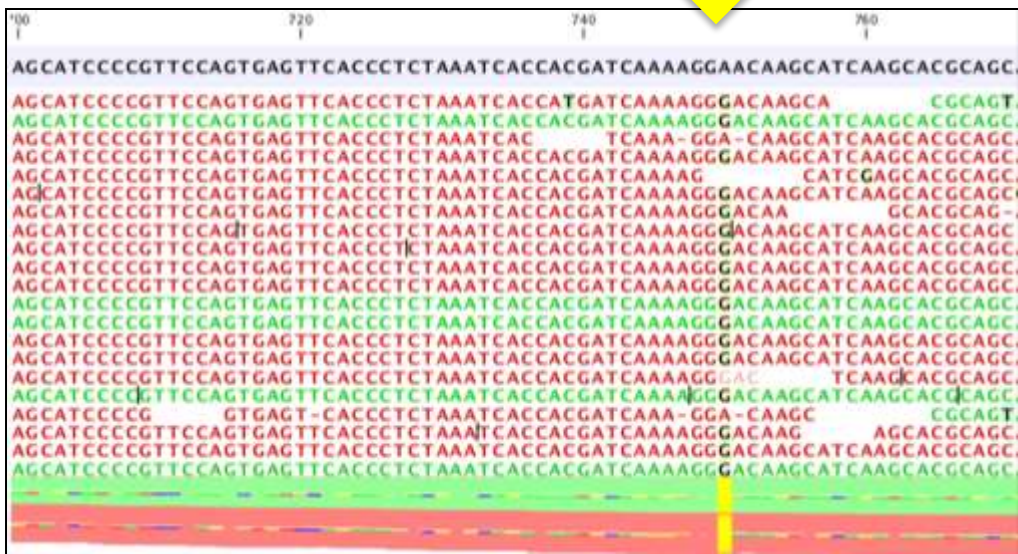
Current Method

- Sequence based on chromatogram
- Consensus of one forward and one reverse



NGS

- Sequence based on thousands of individual reads
- Improved sensitivity:
  - Mixture detection
  - Low level heteroplasmy



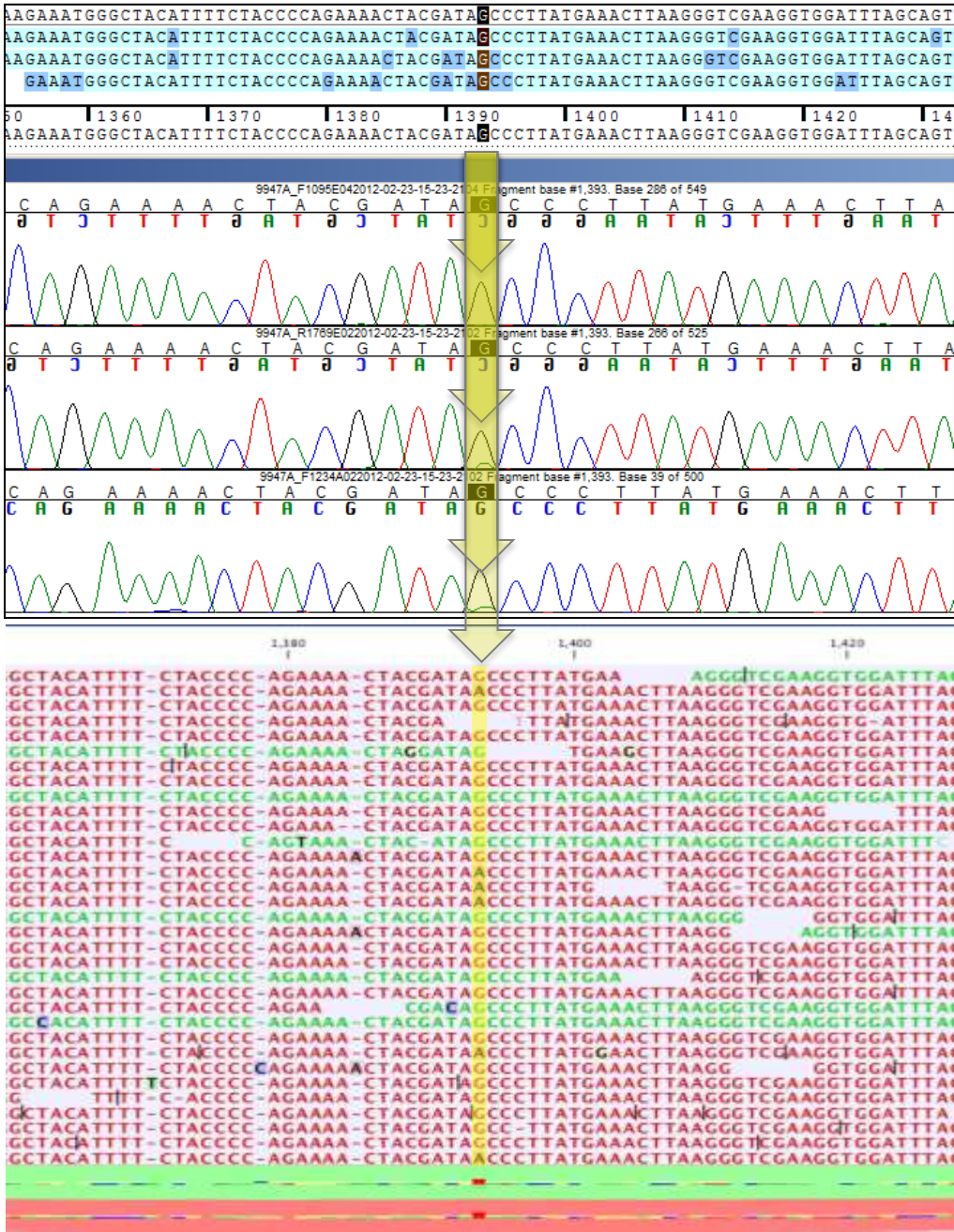
mtDNA Information

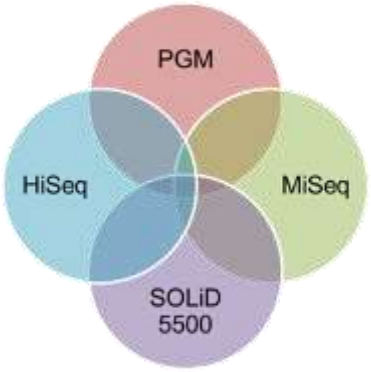
Current Method

- Minor peaks may not be reproducible
- SRM 2392 9947a, 1393 G/A heteroplasmy

NGS

- More consistent detection of minor genotypes
- Validation important
  - Variant calling thresholds
  - Characterizing noise

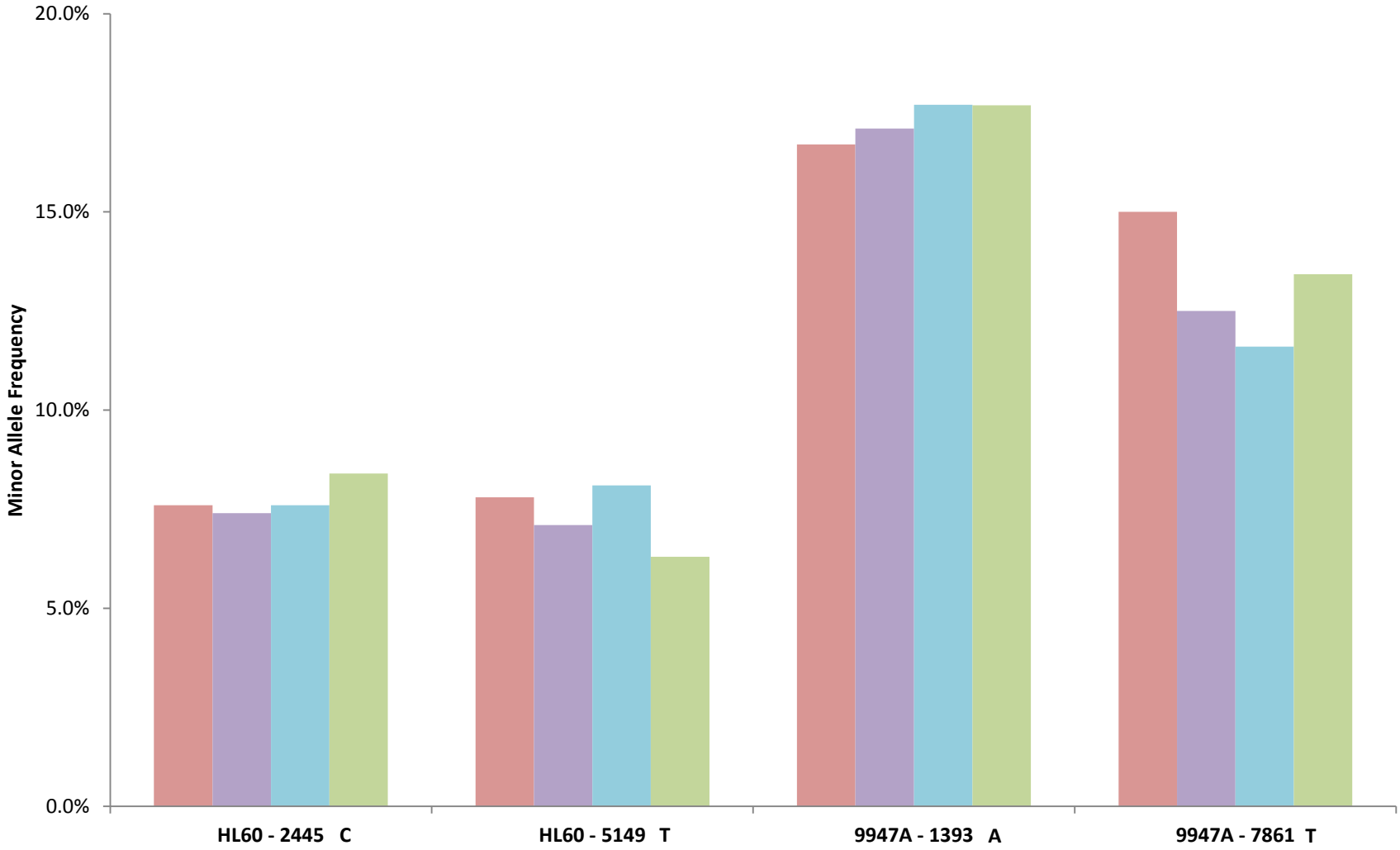


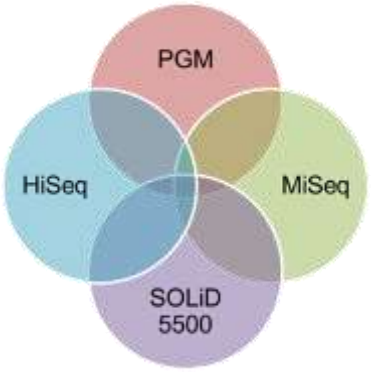


Characterization of SRM 2392 and 2392-I

Mitochondrial genome sequencing standard

Detection of low level heteroplasmy

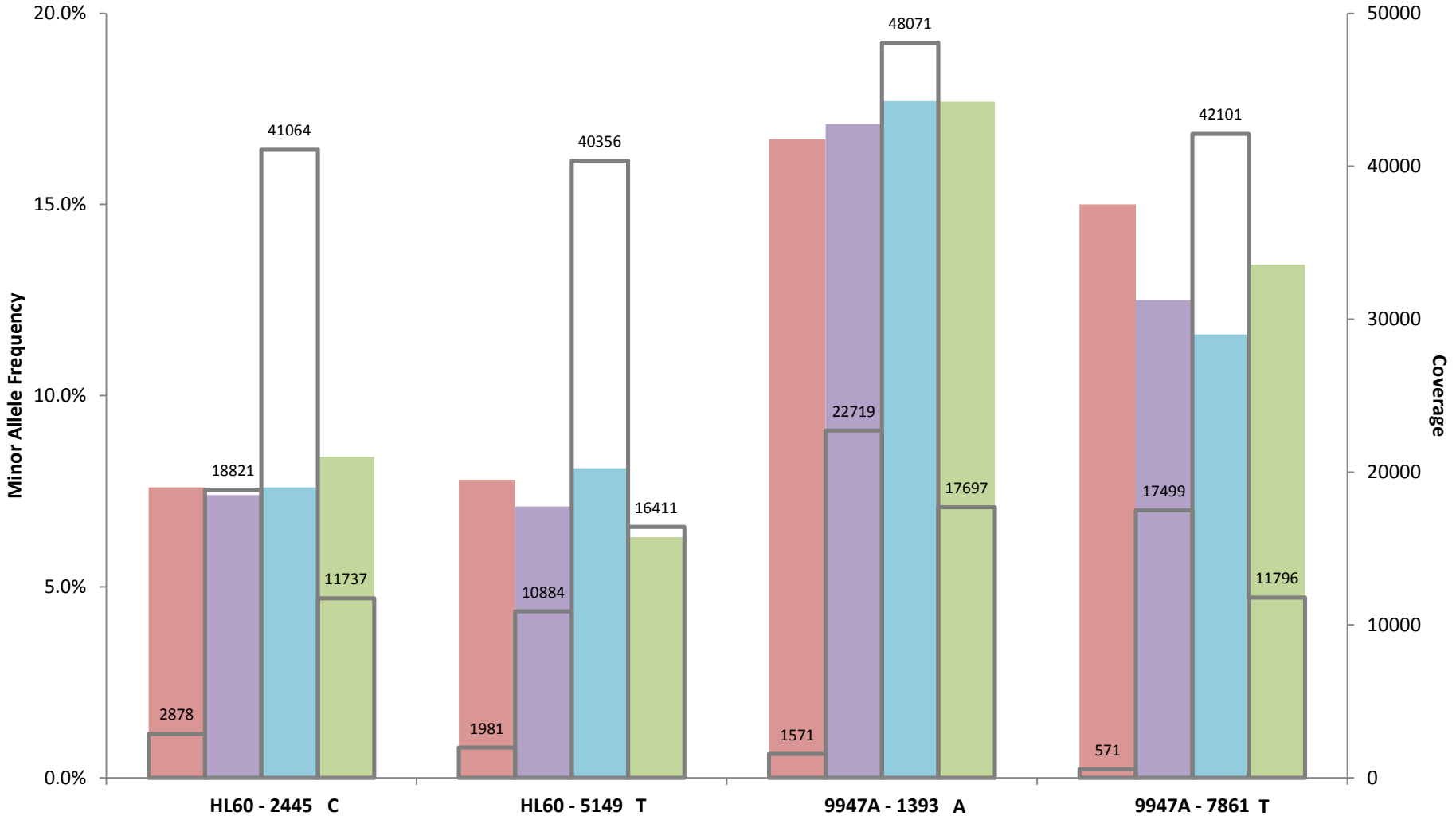




Characterization of SRM 2392 and 2392-I

Mitochondrial genome sequencing standard

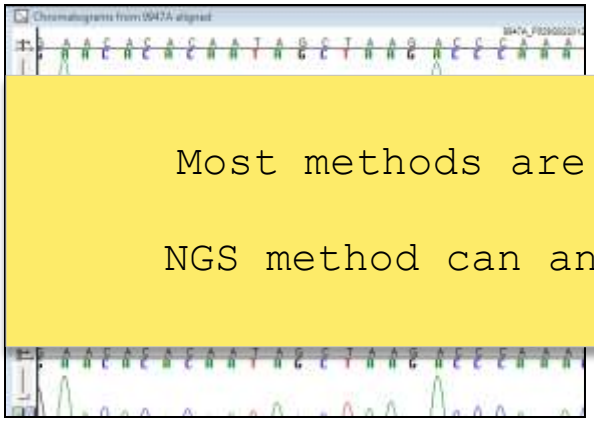
Detection of low level heteroplasmy



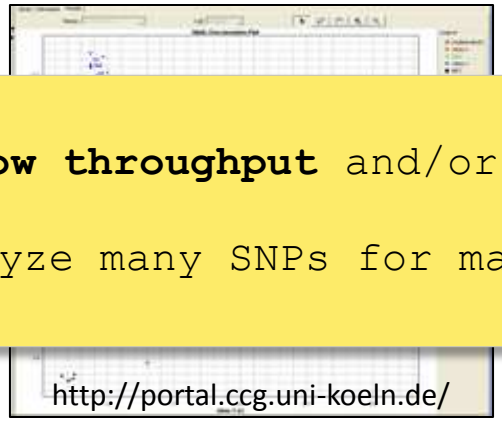
Forensic DNA Markers

- Short Tandem Repeats (STRs)
  - PCR fragment-length polymorphisms
- Mitochondrial DNA (mtDNA)
  - Sanger sequencing
- Single Nucleotide Polymorphisms (SNPs)

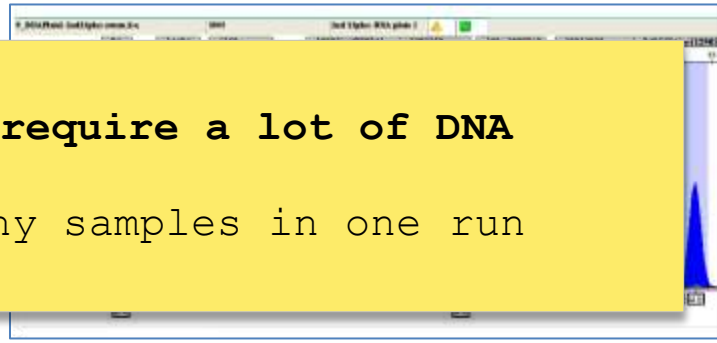
Sanger Sequencing



Allelic Discrimination (qPCR)



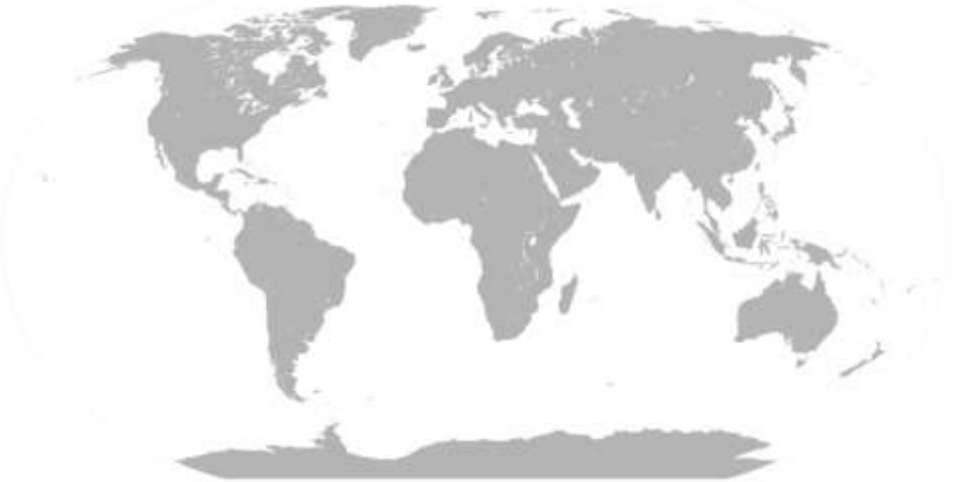
SNaPshot



Most methods are **low throughput** and/or **require a lot of DNA**

NGS method can analyze many SNPs for many samples in one run

SNP Information



- IISNP-Individual
- AISNP-Ancestry
- LISNP-Lineage
- PISNP-Phenotype

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Forensic Science International: Genetics Supplement Series 1 (2008) 471–472
FSI GENETICS
www.elsevier.com/locate/FSIGSS

Research article

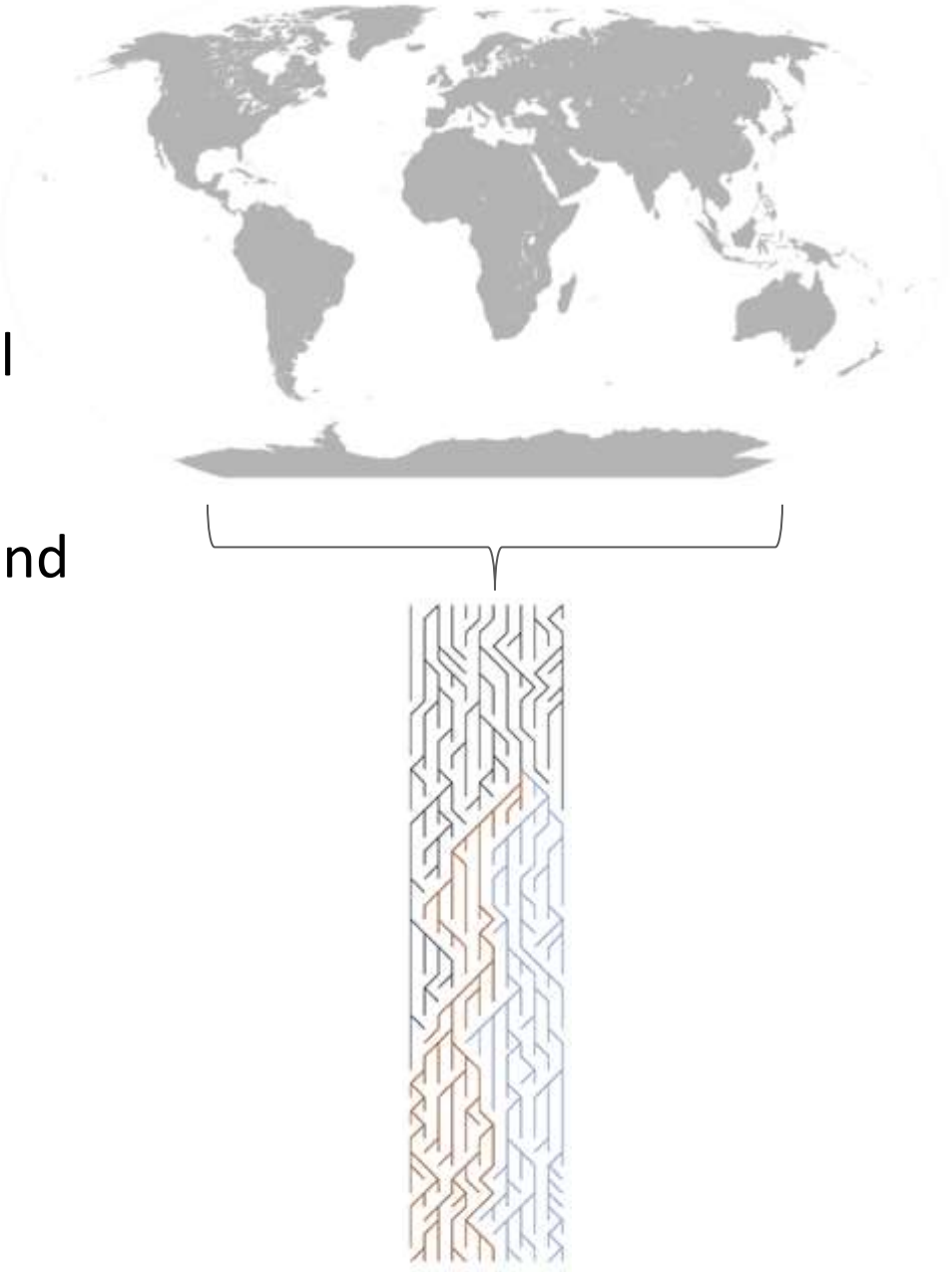
Report on ISFG SNP Panel Discussion

J.M. Butler [a,*], B. Budowle [b], P. Gill [c], K.K. Kidd [d], C. Phillips [e], P.M. Schneider [f], P.M. Vallone [a], N. Morling [g]

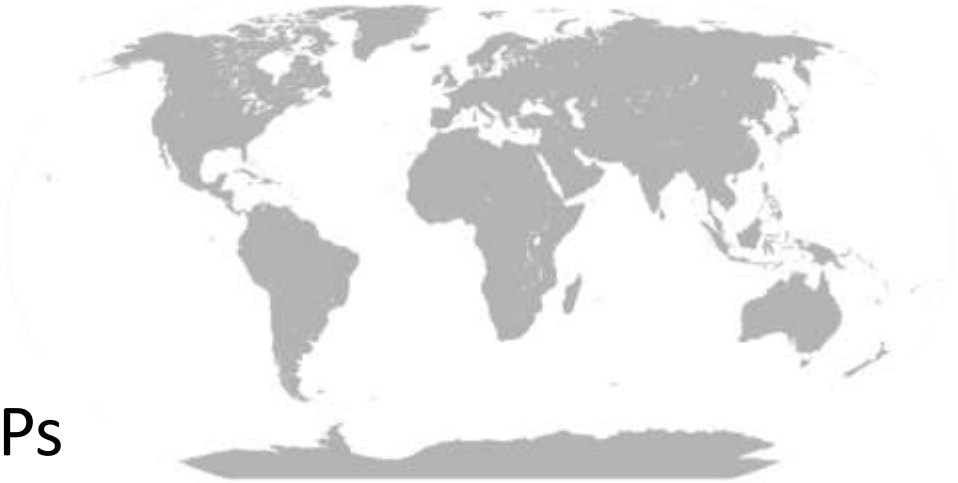
[a] National Institute of Standards and Technology, Gaithersburg, MD, USA
[b] FBI Laboratory, Quantico, VA, USA
[c] Forensic Science Service, Birmingham, UK
[d] Yale University, New Haven, CT, USA
[e] University of Santiago de Compostela, Spain
[f] University of Cologne, Germany
[g] University of Copenhagen, Denmark

SNP Information

- Individual Identification
  - Balancing has occurred in all populations
  - Low F statistics within (F_{IS}) and among (F_{ST}) populations
  - High heterozygosity



SNP Information



- Individual Identification

  Pakstis 2010, Kidd 2012

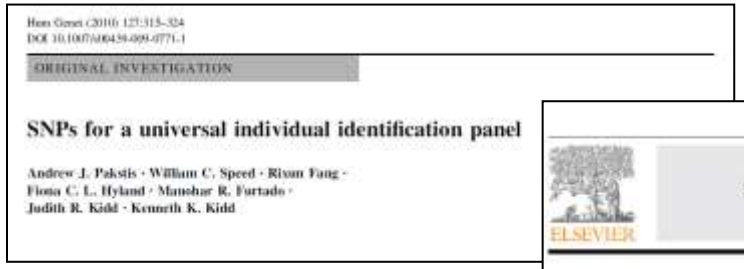
  – Panel of 45 unlinked SNPs

  – F_{ST} below ≈ 0.07

  – Avg het > 0.4

  – RMP 10^{-15} to 10^{-18}

  in 44 populations



Hum Genet (2010) 127:315–324
DOI 10.1007/s00439-009-0771-1

ORIGINAL INVESTIGATION

SNPs for a universal individual identification panel

Andrew J. Pakstis · William C. Speed · Rixun Fang · Fiona C. L. Hyland · Manohar R. Furtado · Judith R. Kidd · Kenneth K. Kidd

Forensic Science International: Genetics 6 (2012) 646–652

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

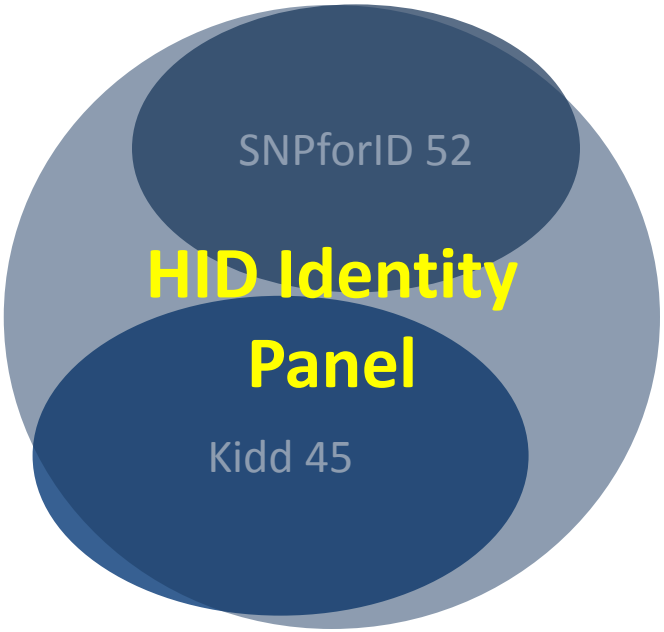
Expanding data and resources for forensic use of SNPs in individual identification

Kenneth K. Kidd [a,*], Judith R. Kidd [a], William C. Speed [a], Rixun Fang [b], Manohar R. Furtado [b], F.C.L. Hyland [b], Andrew J. Pakstis [a]

[a] Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA
[b] Applied Markets, Applied Biosystems/Life Technologies, Foster City, CA 94404, USA

SNP Information

- HID-Ion Ampliseq Identity Panel *(version 2.3)*
  - 90 autosomal SNPs
  - 30 Y-chromosome SNPs
  - RMP 10^{-35}



Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
FSI GENETICS

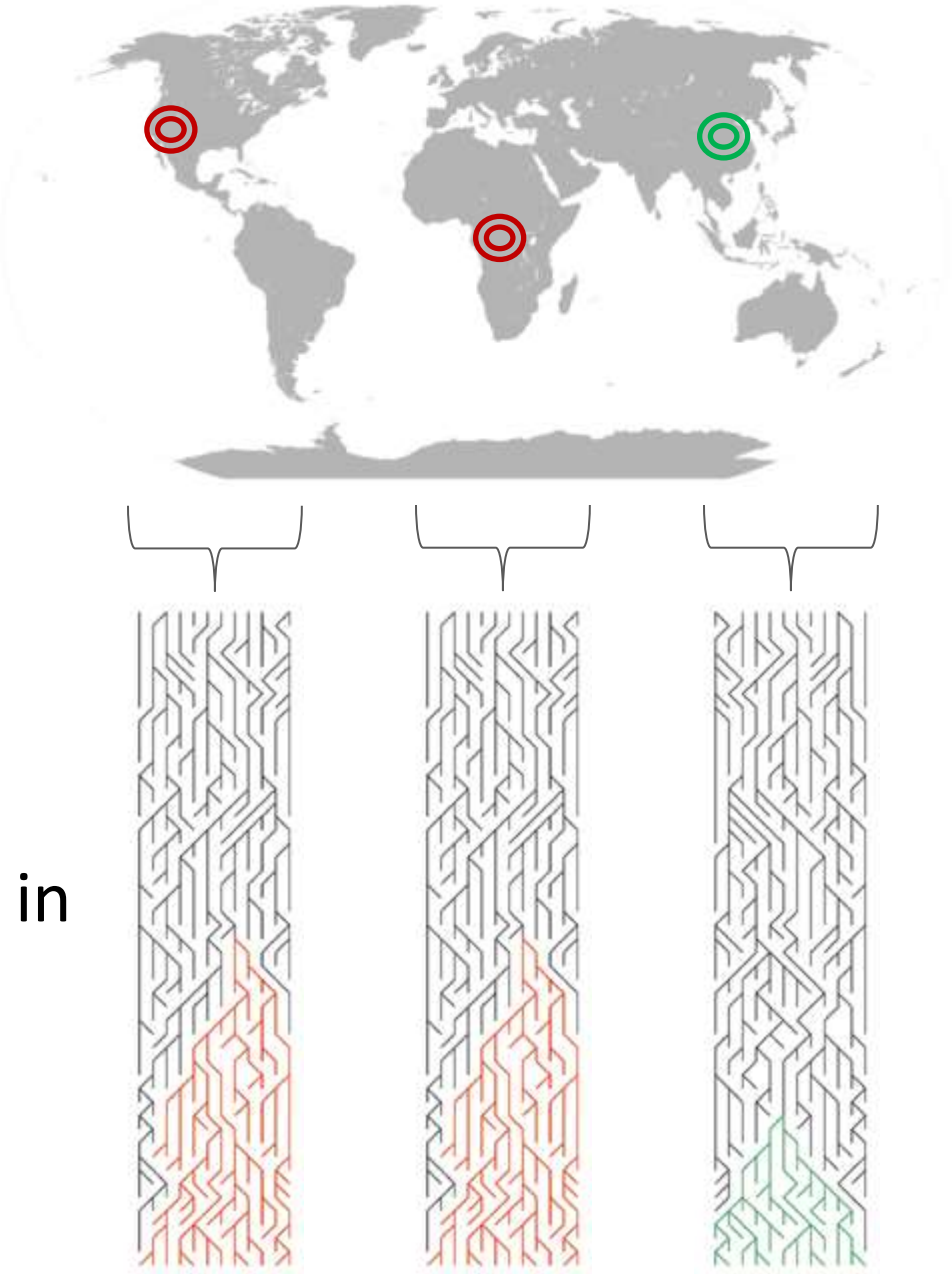
Forensic validation of the SNP*for*ID 52-plex assay

Esther Musgrave-Brown [a,*], David Ballard [a], Kinga Balogh [b], Klaus Bender [b], Burkhard Berger [c], Magdalena Bogus [b], Claus Børsting [d], María Brion [e], Manuel Fondevila [e], Cheryl Harrison [a], Ceylan Oguzturun [a], Walther Parson [c], Chris Phillips [e], Carsten Proff [f], Eva Ramos-Luis [e], Juan J. Sanchez [d], Paula Sánchez Diz [e], Bea Sobrino Rey [e], Beate Stradmann-Bellinghausen [b], Catherine Thacker [a], Angel Carracedo [e], Niels Morling [d], Richard Scheithauer [c], Peter M. Schneider [f], Denise Syndercombe Court [a]

SNP Information

- Ancestry Information
  - High Fixation Index (F_{ST})
  - Population specific fixation has occurred
  - Low heterozygosity

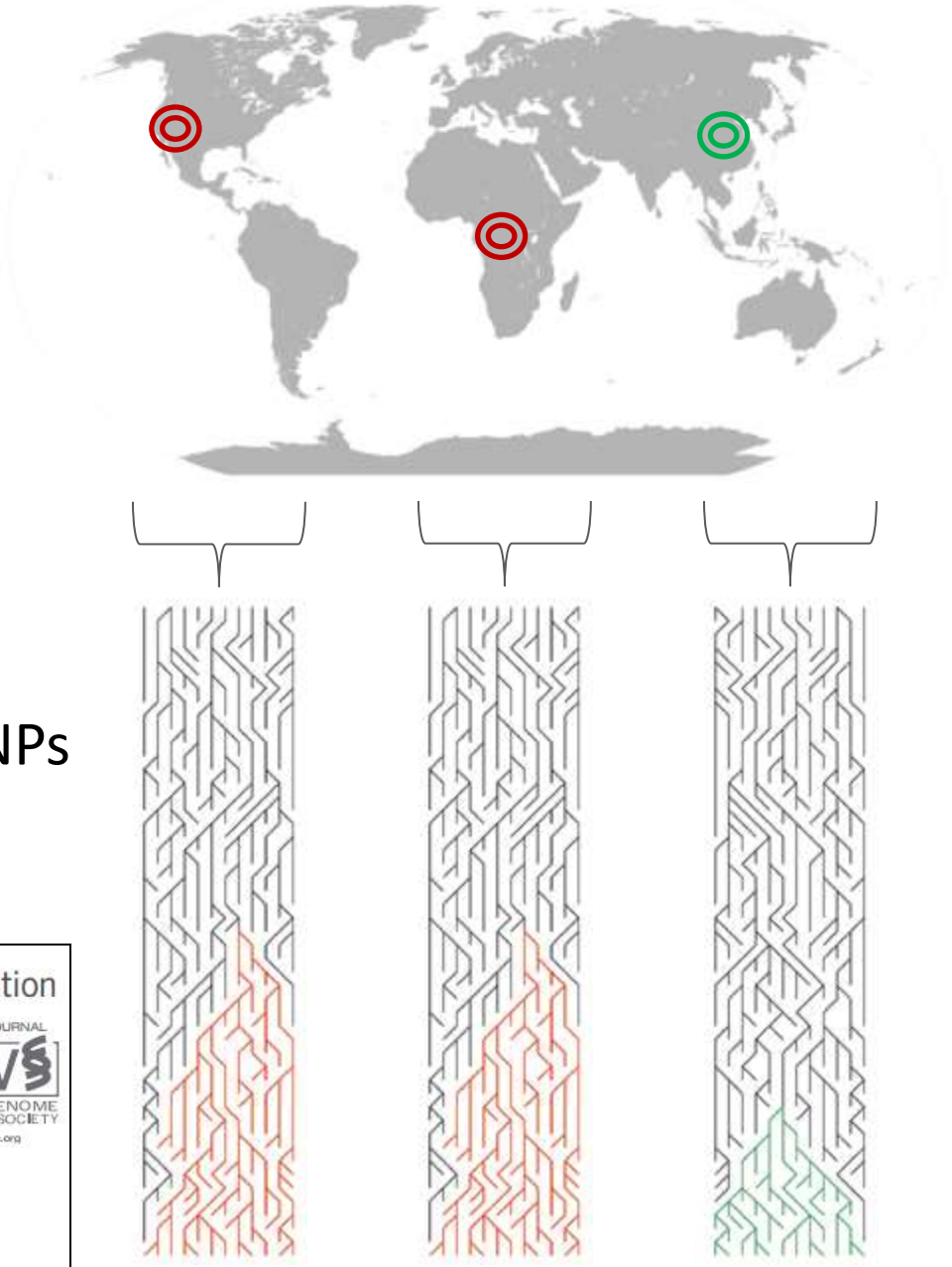
- Example
  - Malaria resistance SNPs in Sub-Saharan Africa



SNP Information

- HID Ancestry Panel
  - Beta version 3.0
  - Publicly available soon
  - 170 loci
  - Derived from

    Kosoy *et. al* (2008): 128 SNPs
    Kidd *et. al* (2014): 55 SNPs



RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America

Roman Kosoy,[1] Rami Nassir,[1] Chao Tian,[1] Phoebe A. White,[2] Lesley M. Butler,[3] Gabriel Silva,[4] Rick Kittles,[5] Marta E. Alarcon-Riquelme,[6] Peter K. Gregersen,[7] John W. Belmont,[8] Francisco M. De La Vega,[2] and Michael F. Seldin[1]*

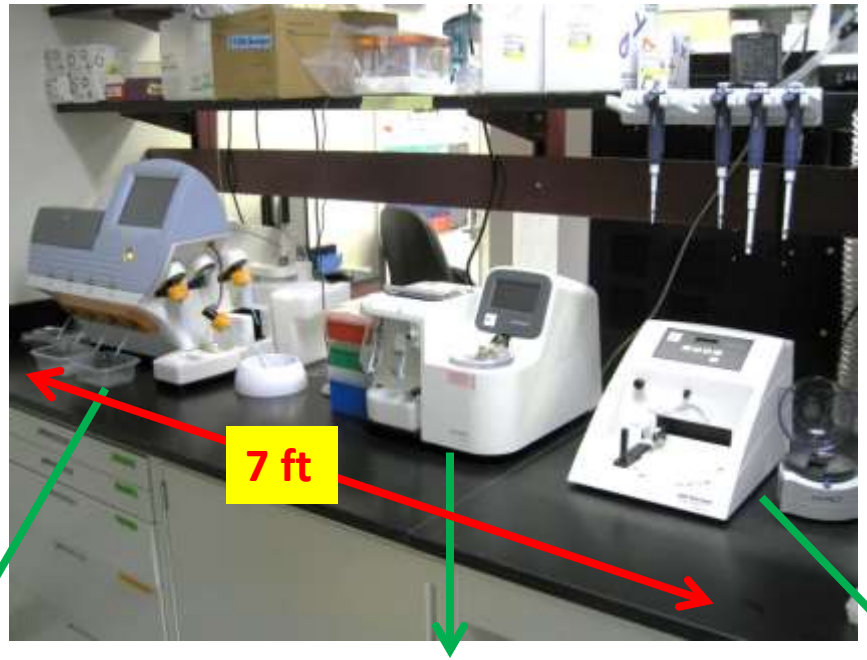
Life Tech - Ion Torrent - PGM

- Ion Torrent Personal Genome Machine (PGM)
  - Launched in 2010
- Ion Torrent sequencing:
  - Emulsion PCR for single copy reactors
  - Non-labeled nucleotide triphosphates
  - Flowed over a bead on a semiconductor surface
- Hydrogen Ion detection
  - pH change is detected
  - **No optics**

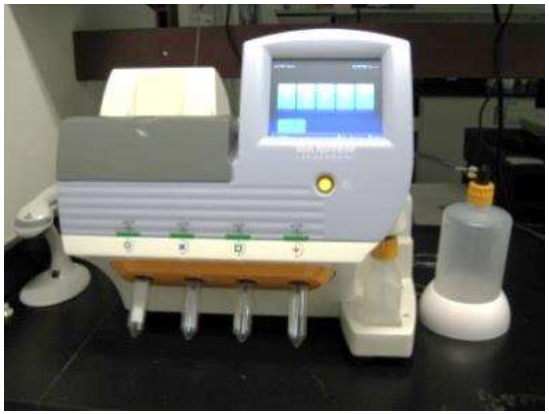
Ion Torrent PGM Workflow

http://www.youtube.com/watch?v=MxkYa9XCvBQ

The PGM Instrument at NIST



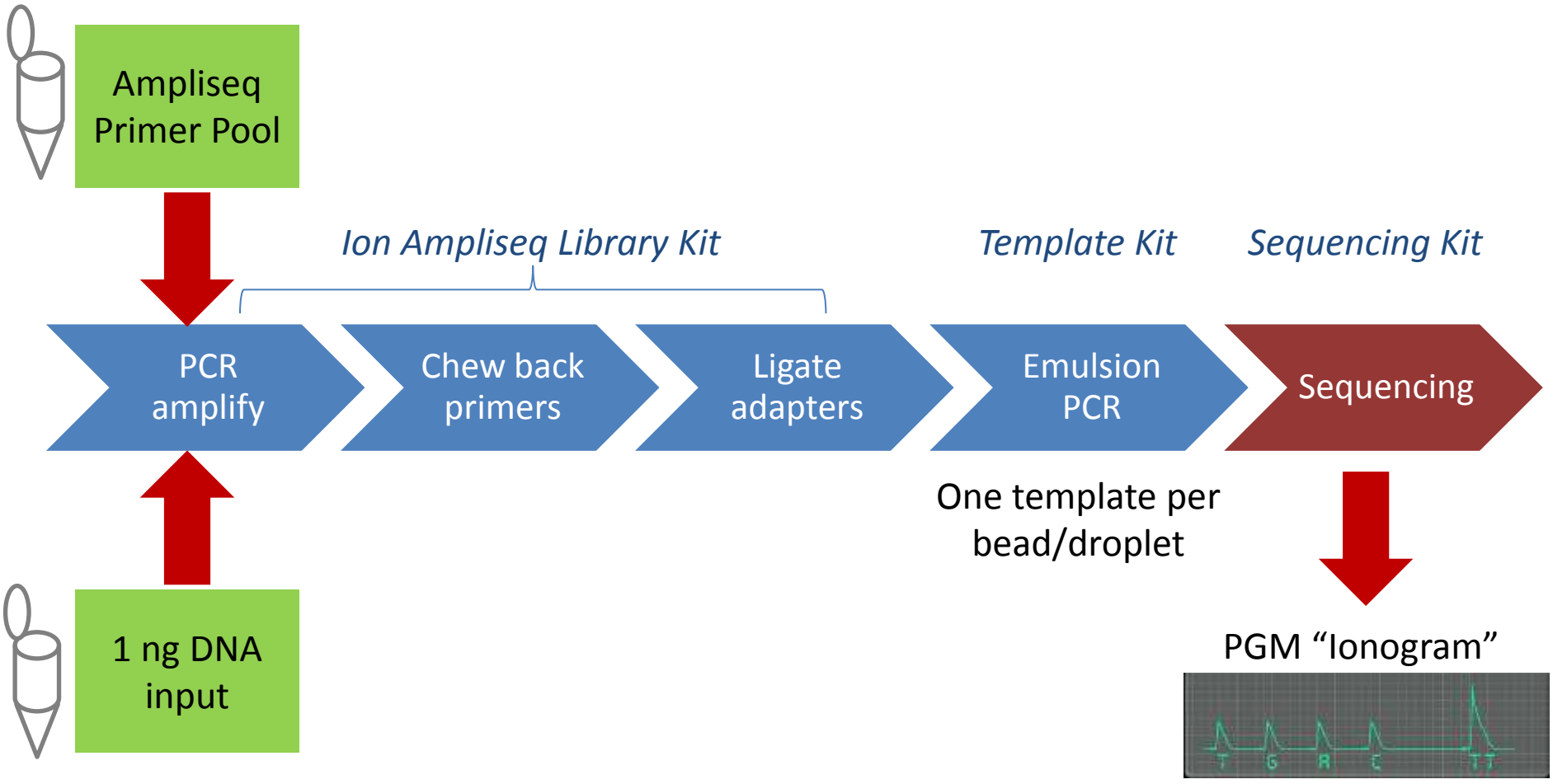
PGM Sequencer



OneTouch 2 (Emulsion PCR)

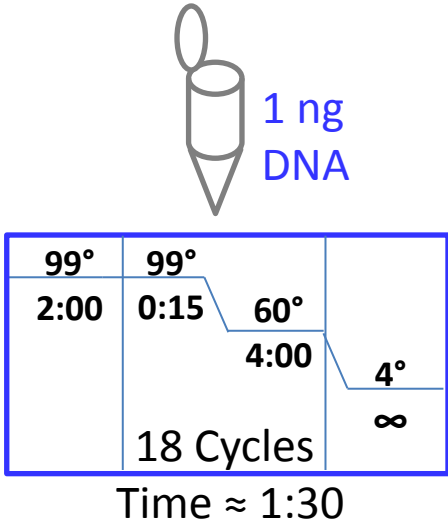
OneTouch ES (Enrichment)

Ampliseq Workflow



Front-End: Multiplex PCR

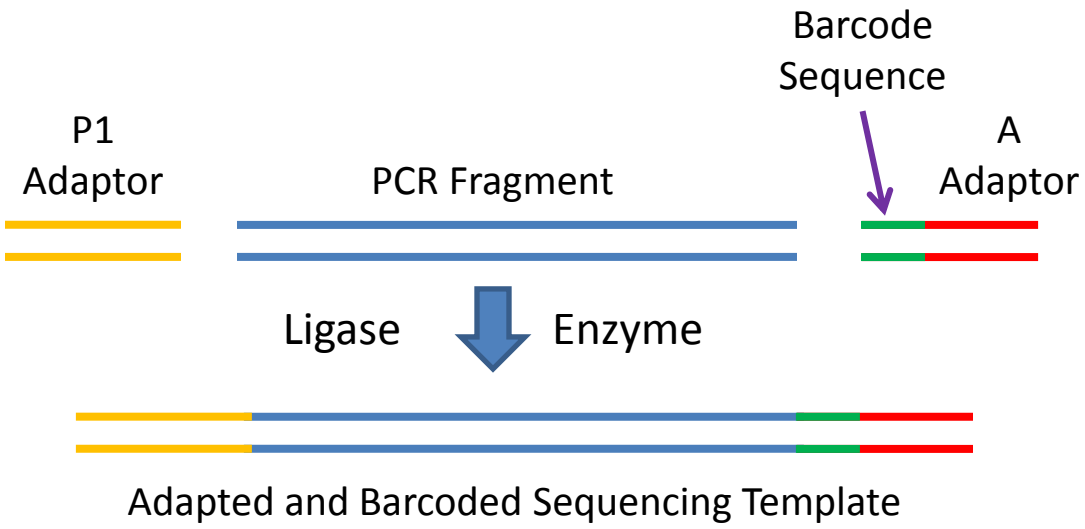
- HID-Ion Ampliseq Identity Panel (IISNP)
  - 120 markers in a single PCR reaction
  - Amplified regions 33 bp to 192 bp long
- HID-Ion Ampliseq Ancestry Panel (AISNP)
  - 170 markers in a single PCR reaction
  - Amplified regions 34 bp to 136 bp long



- Small amplicons well suited to degraded or damaged DNA

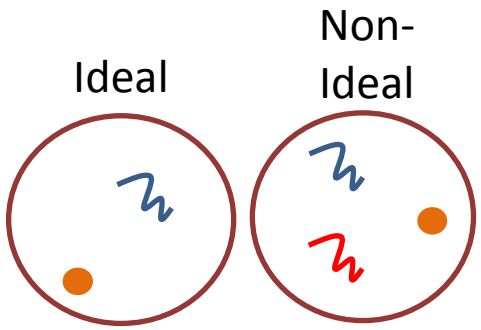
Digest Primer Regions & Ligate Adaptors

- Enzymatic digestion removes ≈ 25 bp from ends of amplicons
- Universal sequencing adaptors are ligated to DNA
  - Adaptors termed P1 and A
- Barcoded sequencing adaptors can be used in this step
  - Sequence multiple samples in one PGM run



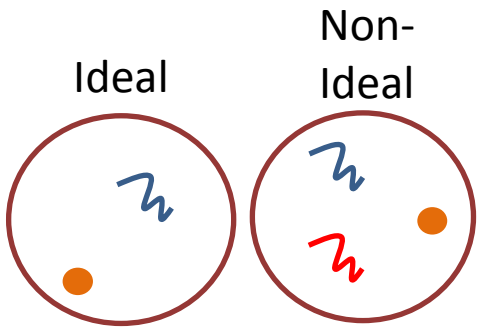
Prepare Ion Sphere Particles (ISPs)

- Libraries quantified by qPCR
  - Quantity of DNA going into emPCR is very important!
  - Goal: 10 % to 30 % template positive ISPs
    - Too much DNA → polyclonal ISPs (mixed read)



Prepare Ion Sphere Particles (ISPs)

- Libraries quantified by qPCR
  - Quantity of DNA going into emPCR is very important!
  - Goal: 10 % to 30 % template positive ISPs
    - Too much DNA → polyclonal ISPs (mixed read)

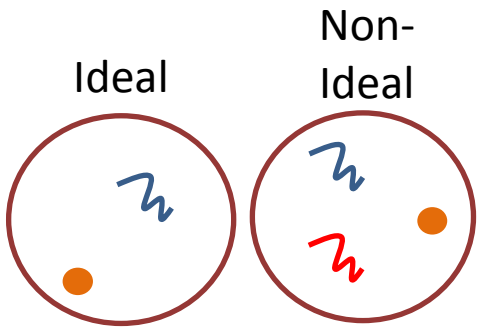


- Emulsion PCR
  - Nanoliter droplets of PCR reagents in oil
  - Attaches sequencing template to the ISP

OneTouch 2

Prepare Ion Sphere Particles (ISPs)

- Libraries quantified by qPCR
  - Quantity of DNA going into emPCR is very important!
  - Goal: 10 % to 30 % template positive ISPs
    - Too much DNA → polyclonal ISPs (mixed read)

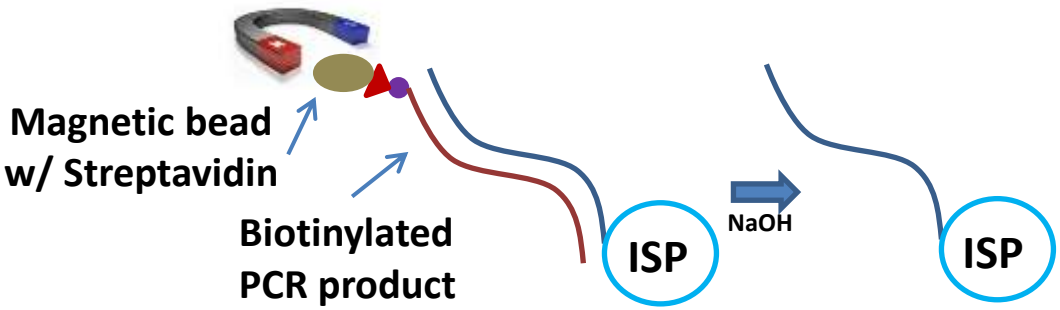


- Emulsion PCR
  - Nanoliter droplets of PCR reagents in oil
  - Attaches sequencing template to the ISP

OneTouch 2

- Enrich for positive ISPs
  - Liquid handler removes non-templated ISPs
  - Biotinylated primer/streptavidin beads

OneTouch ES



Sequencing & Data Analysis

- Library ISPs loaded onto chip
- PGM runs flows & detects pH
- Torrent Server & Torrent Suite Software
  - Processes pH signal into base calls
  - Displays run summary
  - Maps reads to reference genome

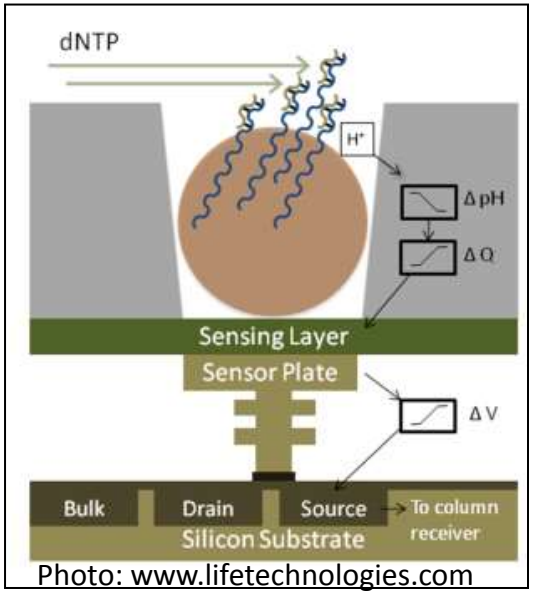
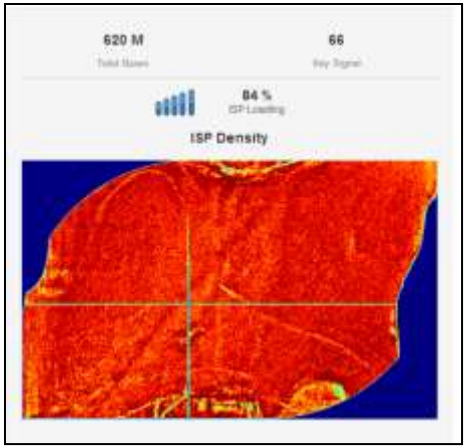


Photo: www.lifetechnologies.com

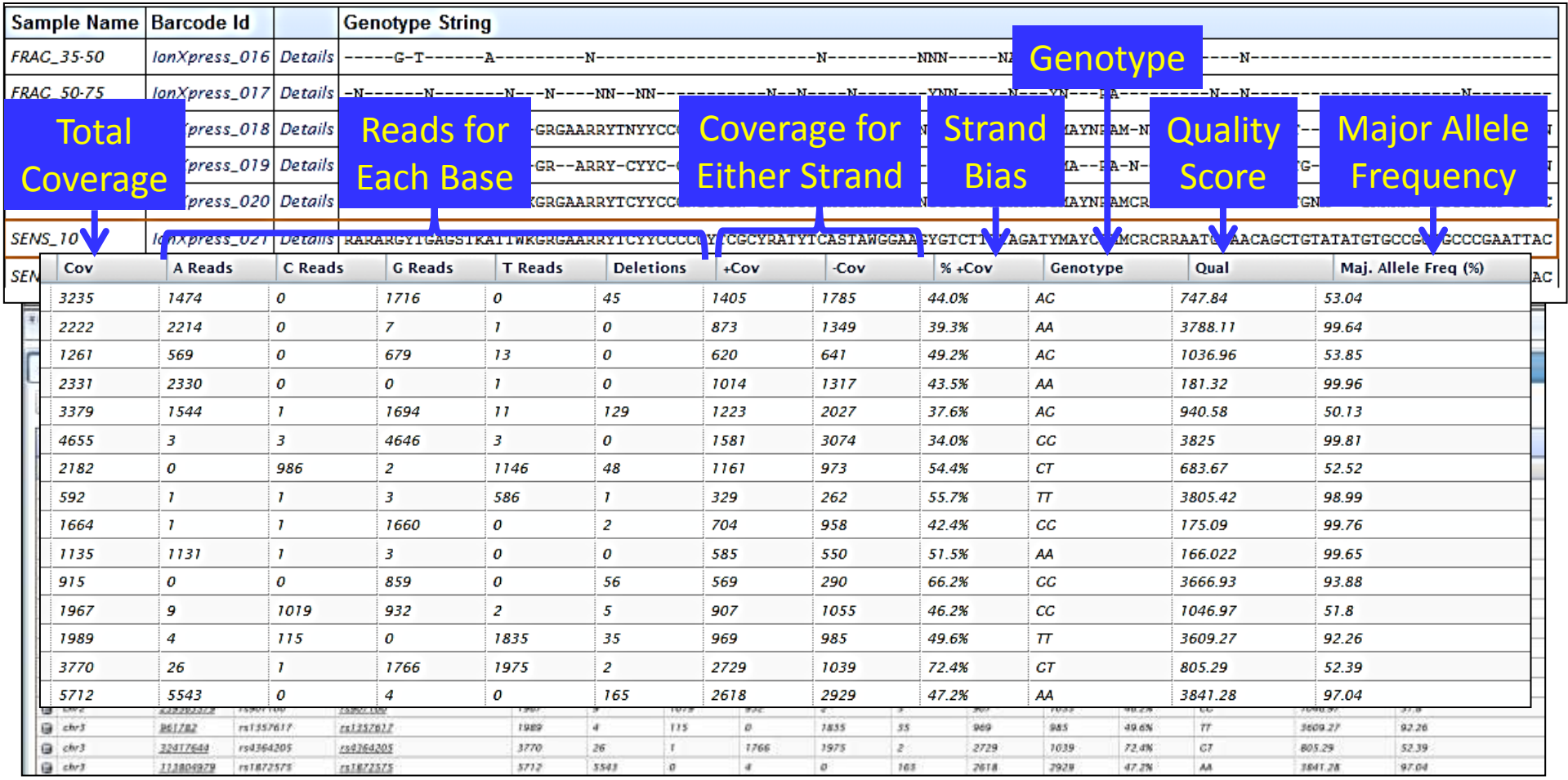


Data Analysis
HID SNP Genotyper Plugin

Allele coverage histogram

Normalized y-axis scale

Data Analysis
HID SNP Genotyper Plugin



Sample Name	Barcode Id		Genotype String
FRAC_35-50	IonXpress_016	Details	-----G-T------A---------N----------N---------NNN-----N----N----------
FRAC_50-75	IonXpress_017	Details	-N------N-------N---N----NN--NN----------N--N----N-------YNN-----N---YN---A---------N--N----------N--------
	press_018	Details	-GRGAARRYTNYYCC MAYNRAM-N
	press_019	Details	-GR--ARRY-CYYC- MA--RA-N- G-
	press_020	Details	KGRGAARRYTCYYCC MAYNRAMCR GN
SENS_10	IonXpress_021	Details	RARARGYTGAGSTKATTWKGRGAARRYTCYYCCCCYTCGCYRATYTCASTAWGGAAGYGTCTTNAGATYMAYCAMCRCRRAATTAACAGCTGTATATGTGCCGTGCCCGAATTAC
SEN			AC

Cov	A Reads	C Reads	G Reads	T Reads	Deletions	+Cov	-Cov	% +Cov	Genotype	Qual	Maj. Allele Freq (%)
3235	1474	0	1716	0	45	1405	1785	44.0%	AC	747.84	53.04
2222	2214	0	7	1	0	873	1349	39.3%	AA	3788.11	99.64
1261	569	0	679	13	0	620	641	49.2%	AC	1036.96	53.85
2331	2330	0	0	1	0	1014	1317	43.5%	AA	181.32	99.96
3379	1544	1	1694	11	129	1223	2027	37.6%	AC	940.58	50.13
4655	3	3	4646	3	0	1581	3074	34.0%	CC	3825	99.81
2182	0	986	2	1146	48	1161	973	54.4%	CT	683.67	52.52
592	1	1	3	586	1	329	262	55.7%	TT	3805.42	98.99
1664	1	1	1660	0	2	704	958	42.4%	CC	175.09	99.76
1135	1131	1	3	0	0	585	550	51.5%	AA	166.022	99.65
915	0	0	859	0	56	569	290	66.2%	CC	3666.93	93.88
1967	9	1019	932	2	5	907	1055	46.2%	CC	1046.97	51.8
1989	4	115	0	1835	35	969	985	49.6%	TT	3609.27	92.26
3770	26	1	1766	1975	2	2729	1039	72.4%	CT	805.29	52.39
5712	5543	0	4	0	165	2618	2929	47.2%	AA	3841.28	97.04

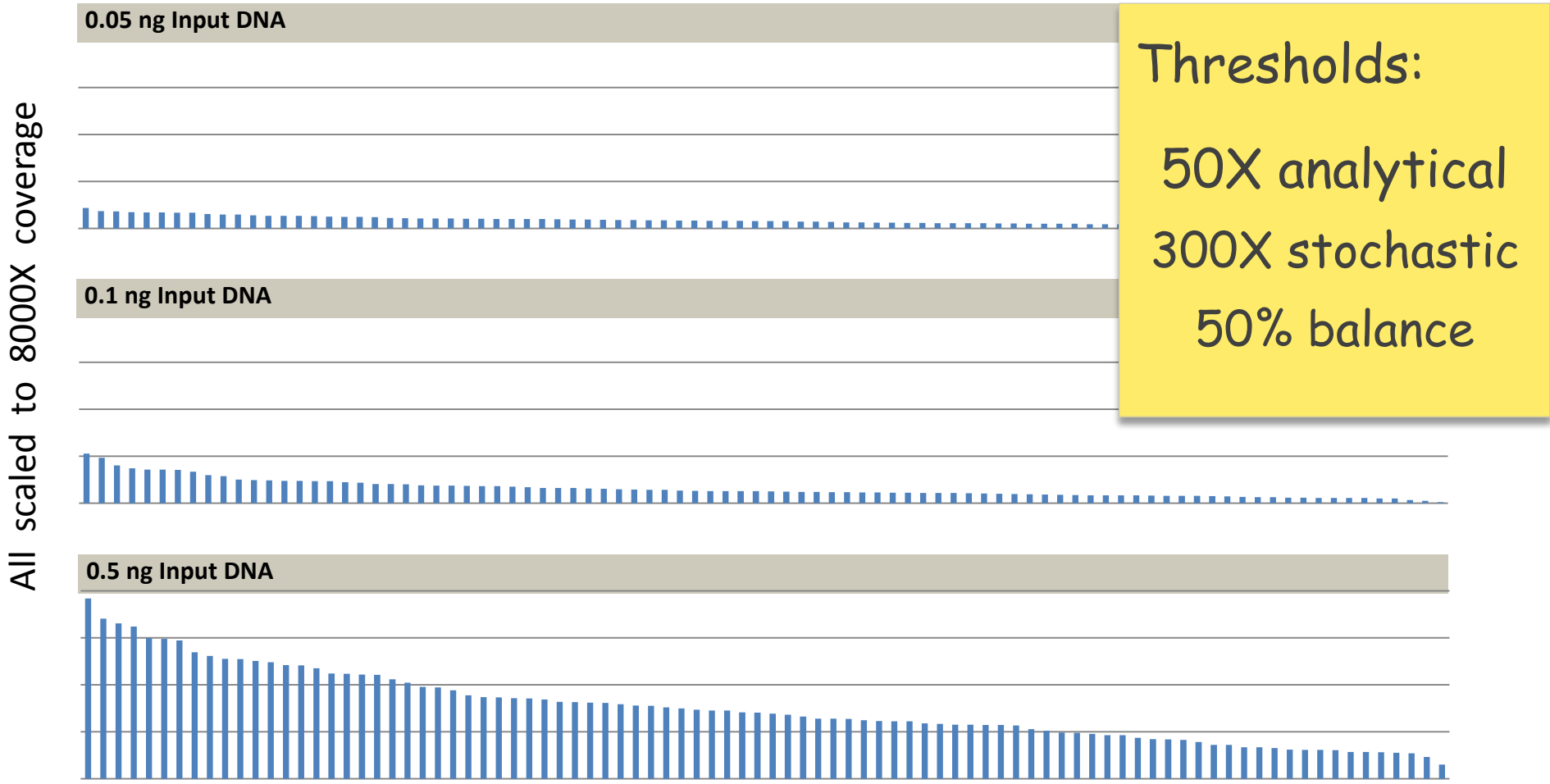
HID SNP Panel Sensitivity Study

- Dynamic range of DNA input to PCR
  - 1 ng is recommended
  - 10 ng (1 data point) – no problems were observed
  - 1 ng
  - 0.5 ng
  - 0.1 ng
  - 0.05 ng

  3 Replicates
- Libraries were generated and pooled (n = 12)
- Sequenced on PGM 318 chip (11 M wells)
  - 200 bp read chemistry

HID SNP Panel Sensitivity Study

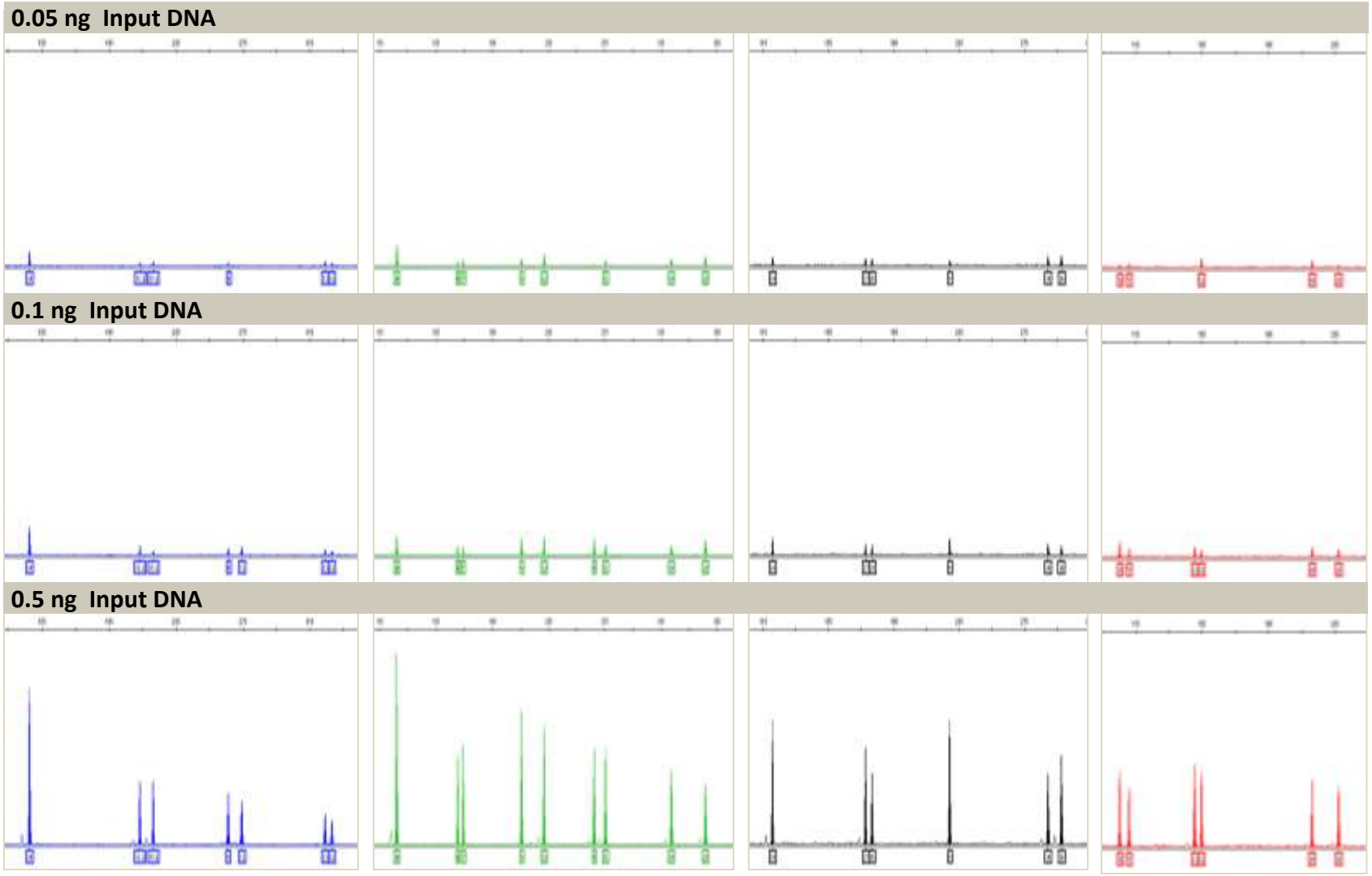
90 Autosomal SNP loci, sorted from highest to lowest coverage



Thresholds: analytical = 50 RFU, stochastic = 300 RFU, PHR = 0.5

HID SNP Panel Sensitivity Study

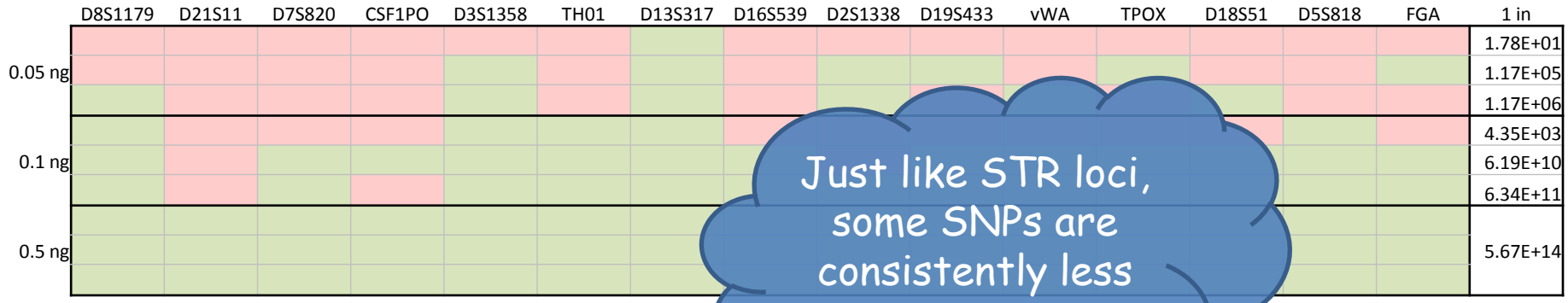
All scaled to 4000 RFU



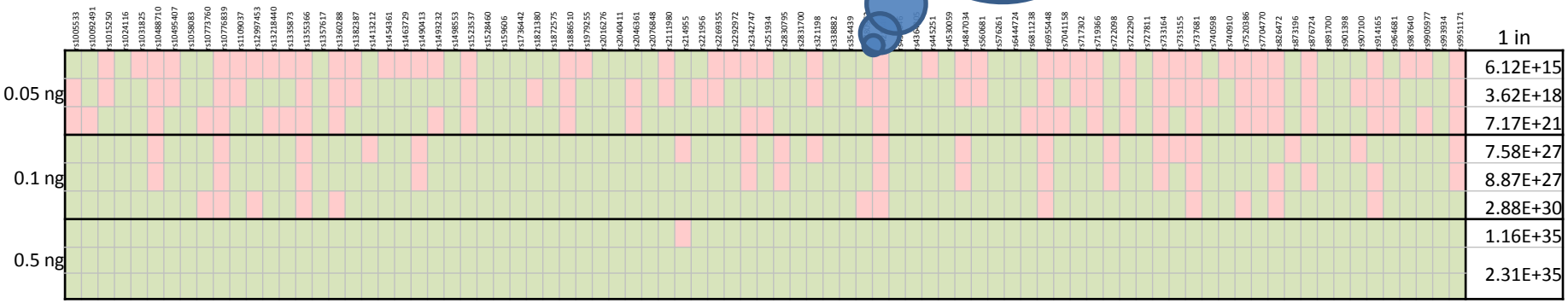
Identifiler® Plus amplification (29 cycle), 25 µl reaction , 3500*xl* electrophoresis, 1.2 kV for 8 seconds
Thresholds: analytical = 50 RFU, stochastic = 200 RFU, PHR = 0.5

HID SNP Panel Sensitivity Study

Identifiler Plus

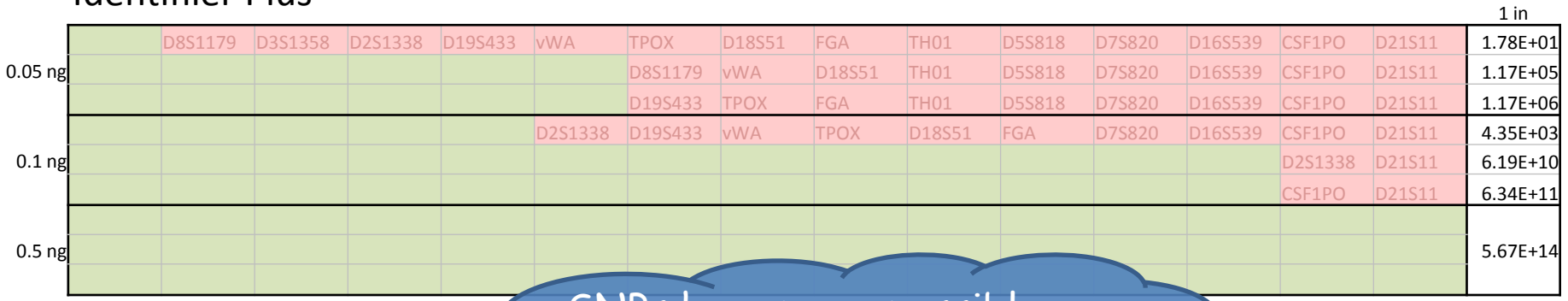


PGM HID SNP Panel v2.3



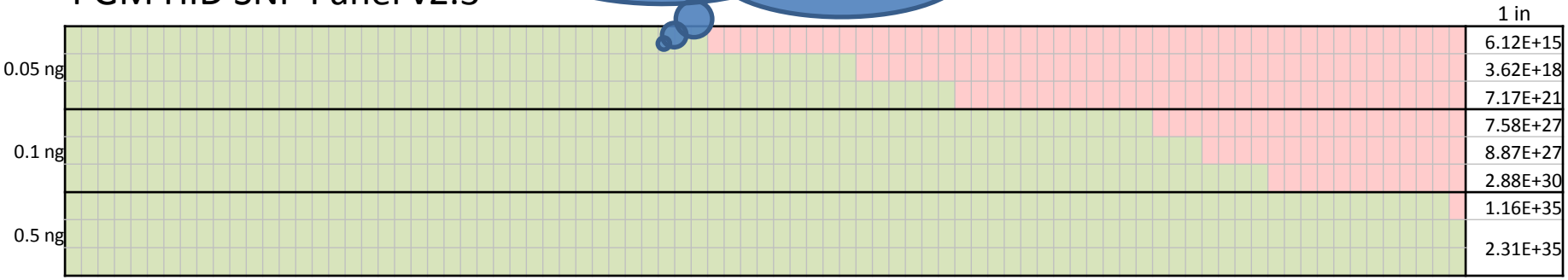
HID SNP Panel Sensitivity Study

Identifiler Plus

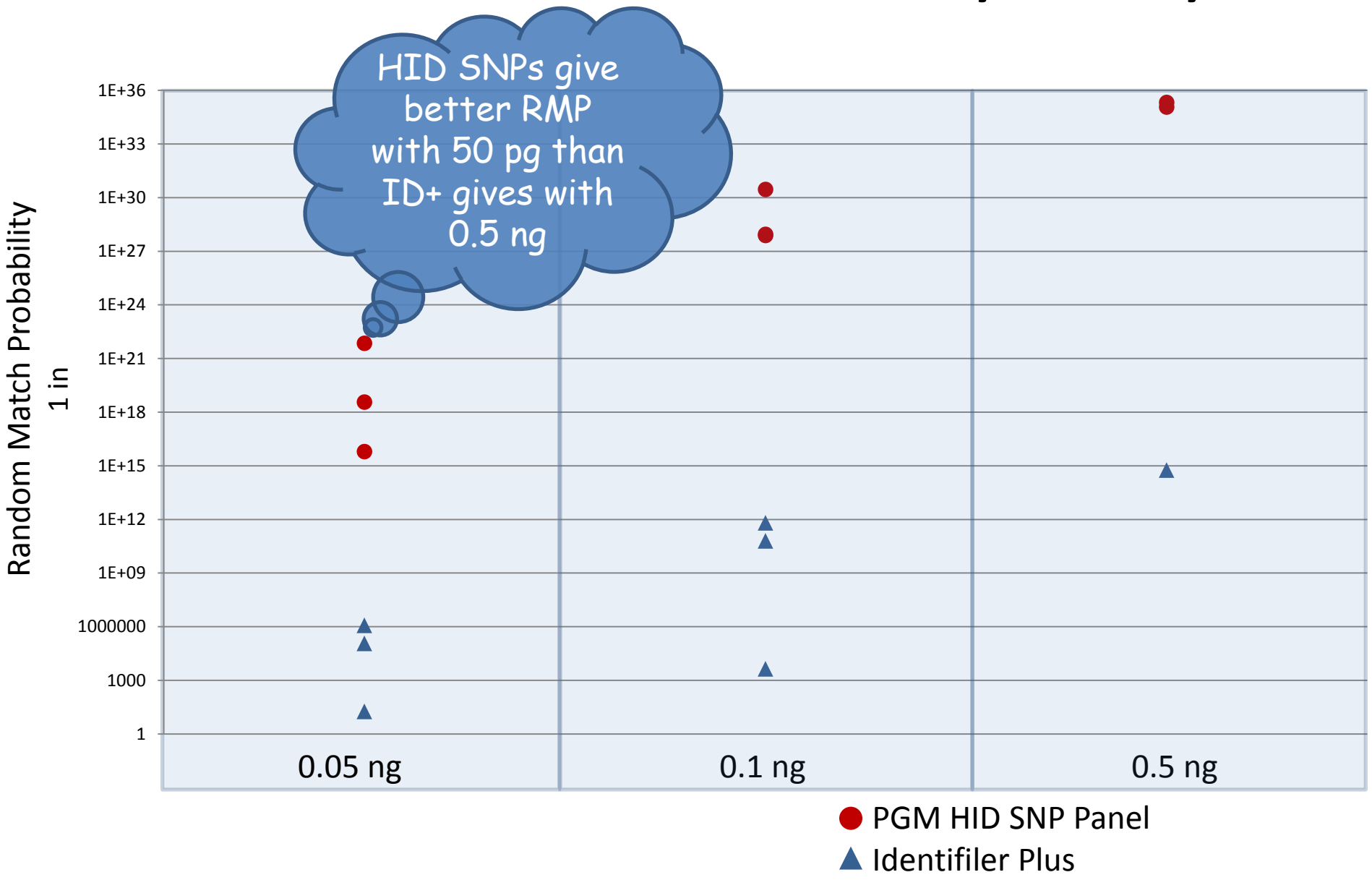


SNPs have more possible loci and better performance at low levels

PGM HID SNP Panel v2.3



HID SNP Panel Sensitivity Study



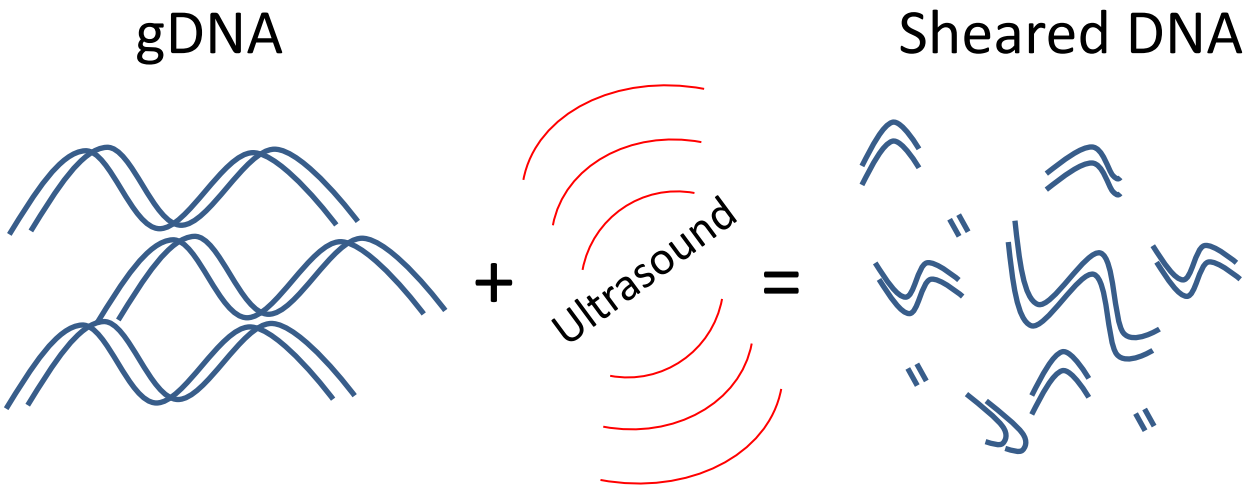
HID SNP Panel Sensitivity Study Summary

- Higher RMPs are expected for SNP panel compared to STRs due to many more loci
- Under thresholds indicated, higher % SNPs produce results than STRs also
- Better STR assays (GlobalFiler or NGS-STR) may lessen the “gap”
- Validation needed for SNP thresholds

HID SNP Panel Degraded DNA Study

Sheared genomic DNA

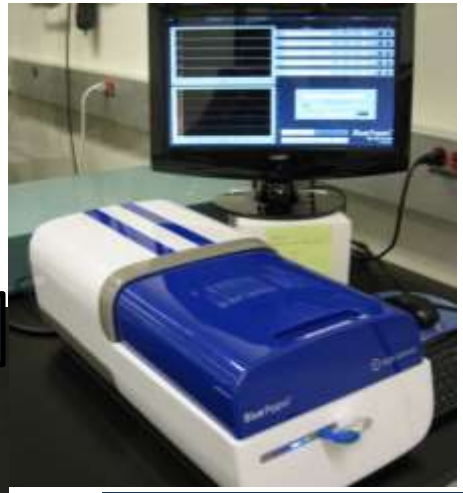
→Covaris S2 Focused Ultrasonicator

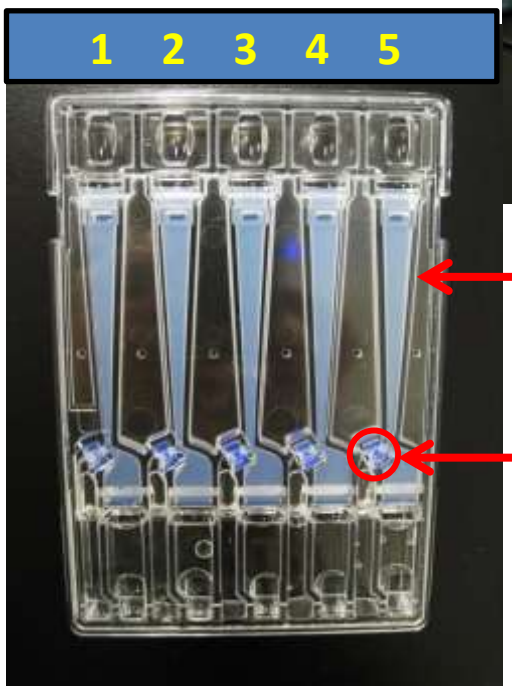


HID SNP Panel
Degraded DNA Study

Sheared DNA was fractionated by size range

- Blue Pippin system (3% Gel)
- Automated size selection
  1) 50 bp to 200 bp
  2) 50 bp to 150 bp
  3) 50 bp to 100 bp
  4) 50 bp to 75 bp
  5) 35 bp to 50 bp





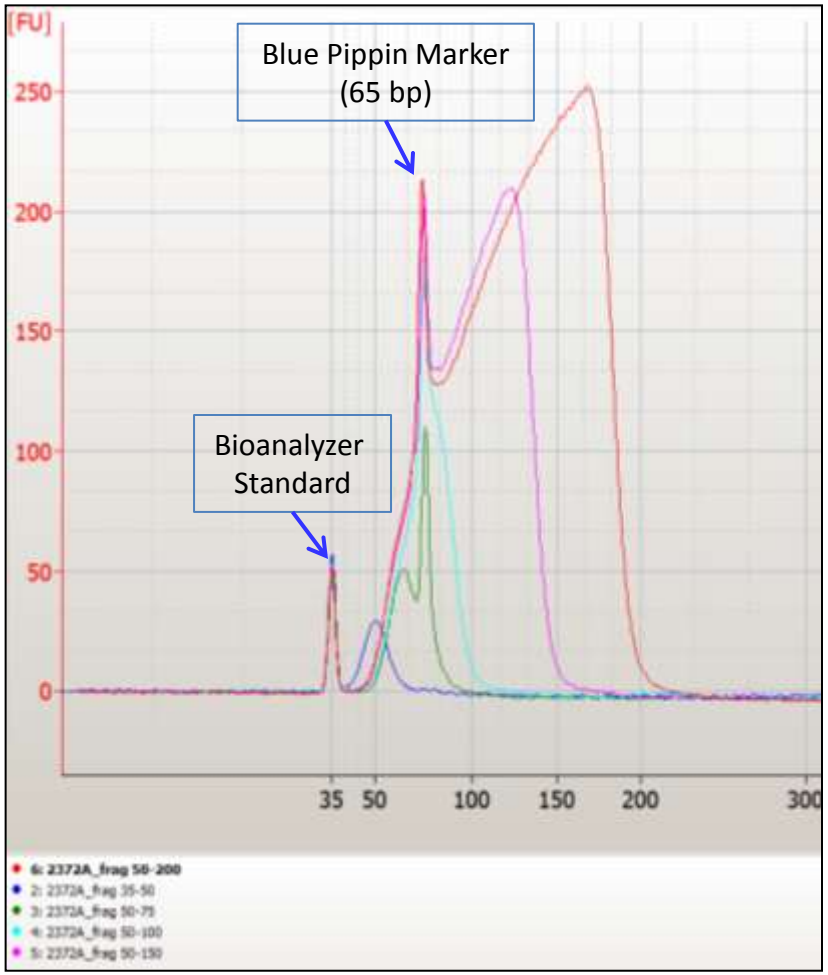
Five individual agarose columns

Size fractionated fragments collected into recovery wells

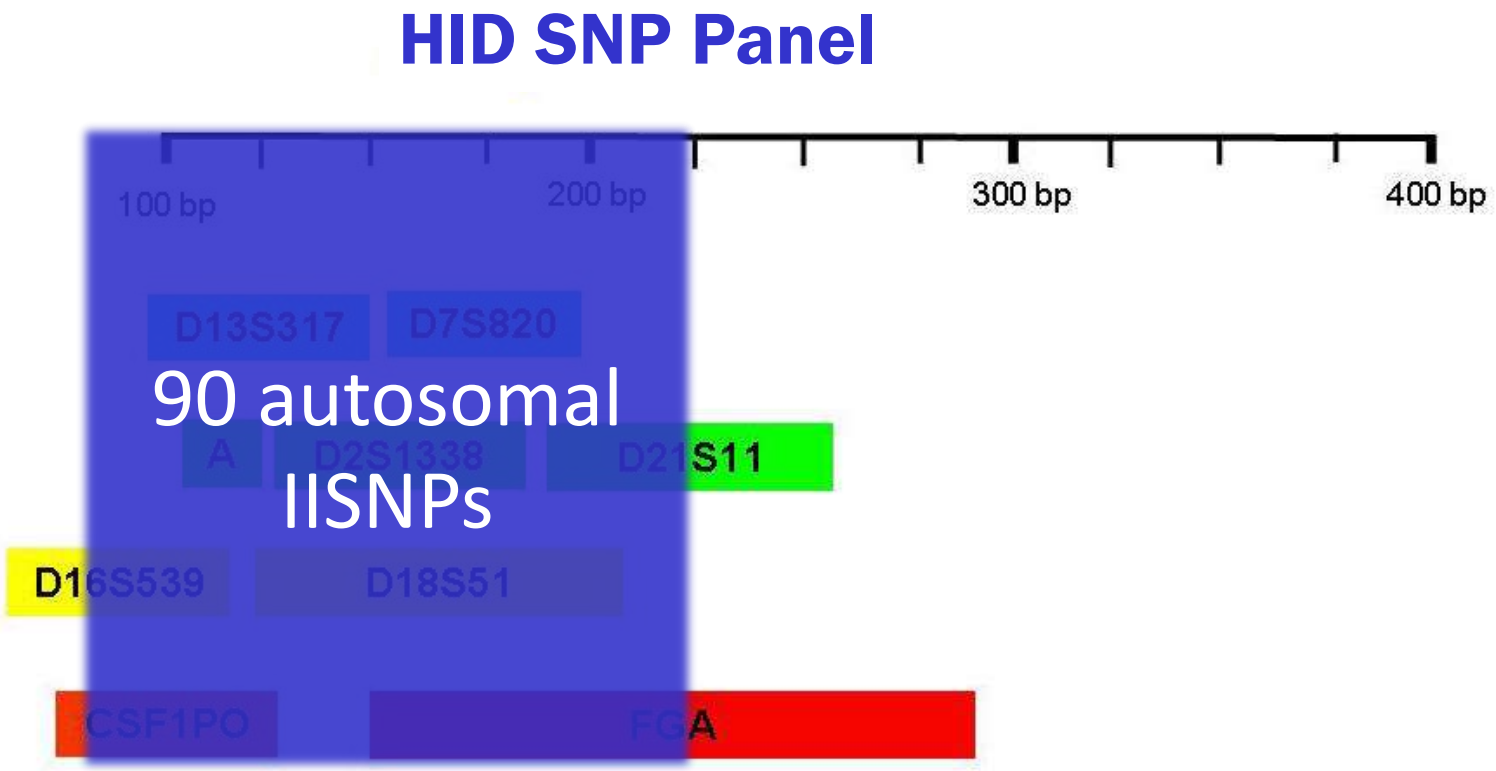
HID SNP Panel
Degraded DNA Study

Sheared DNA was fractionated by size range

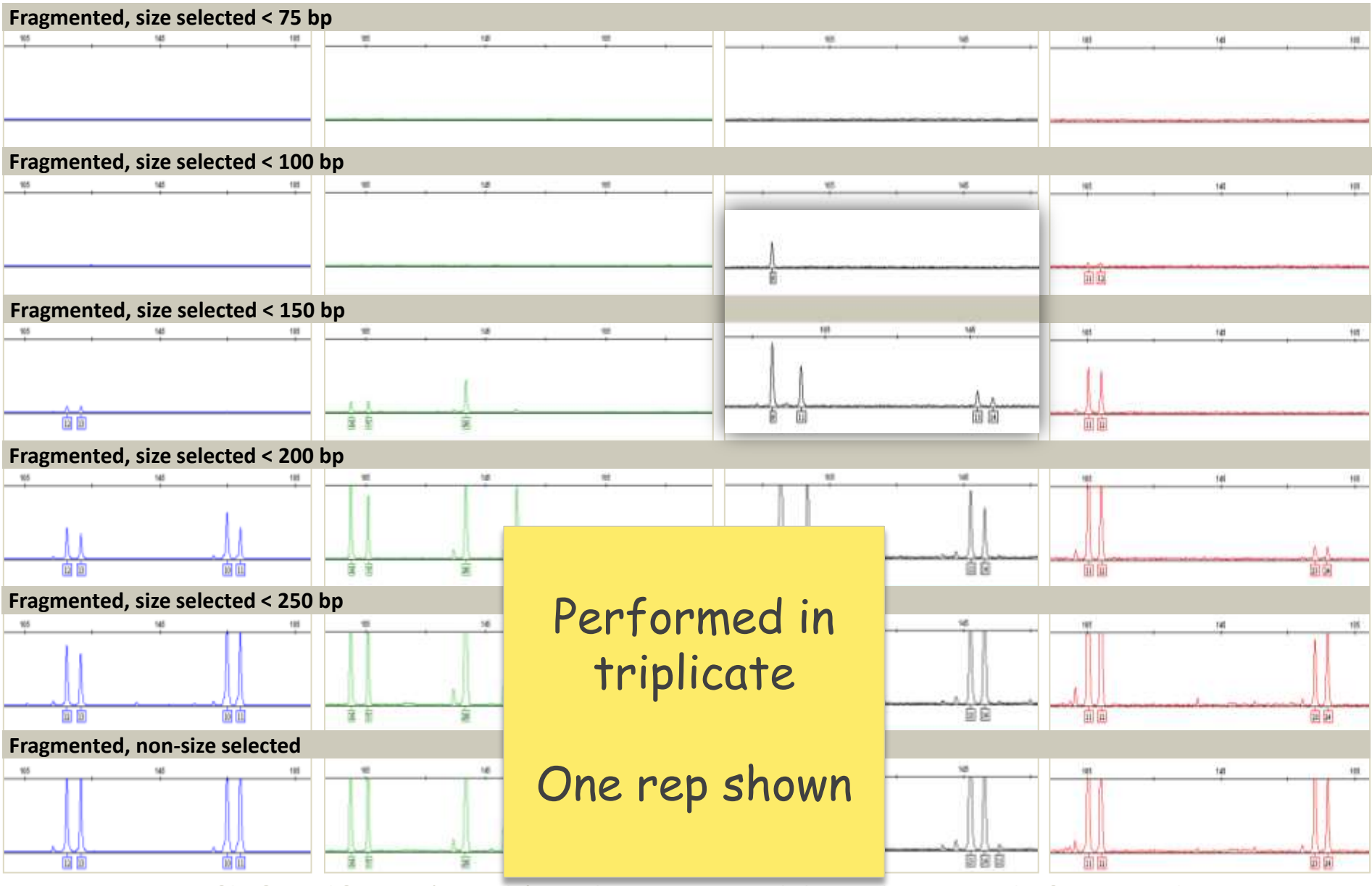
- Agilent Bioanalyzer Trace
- Size selected sheared DNA
  - 50 bp to 200 bp
  - 50 bp to 150 bp
  - 50 bp to 100 bp
  - 50 bp to 75 bp
  - 35 bp to 50 bp
- Input to HID Panel PCR
  - 1 ng DNA
  - Built libraries and sequenced



HID SNP Panel
Degraded DNA Study

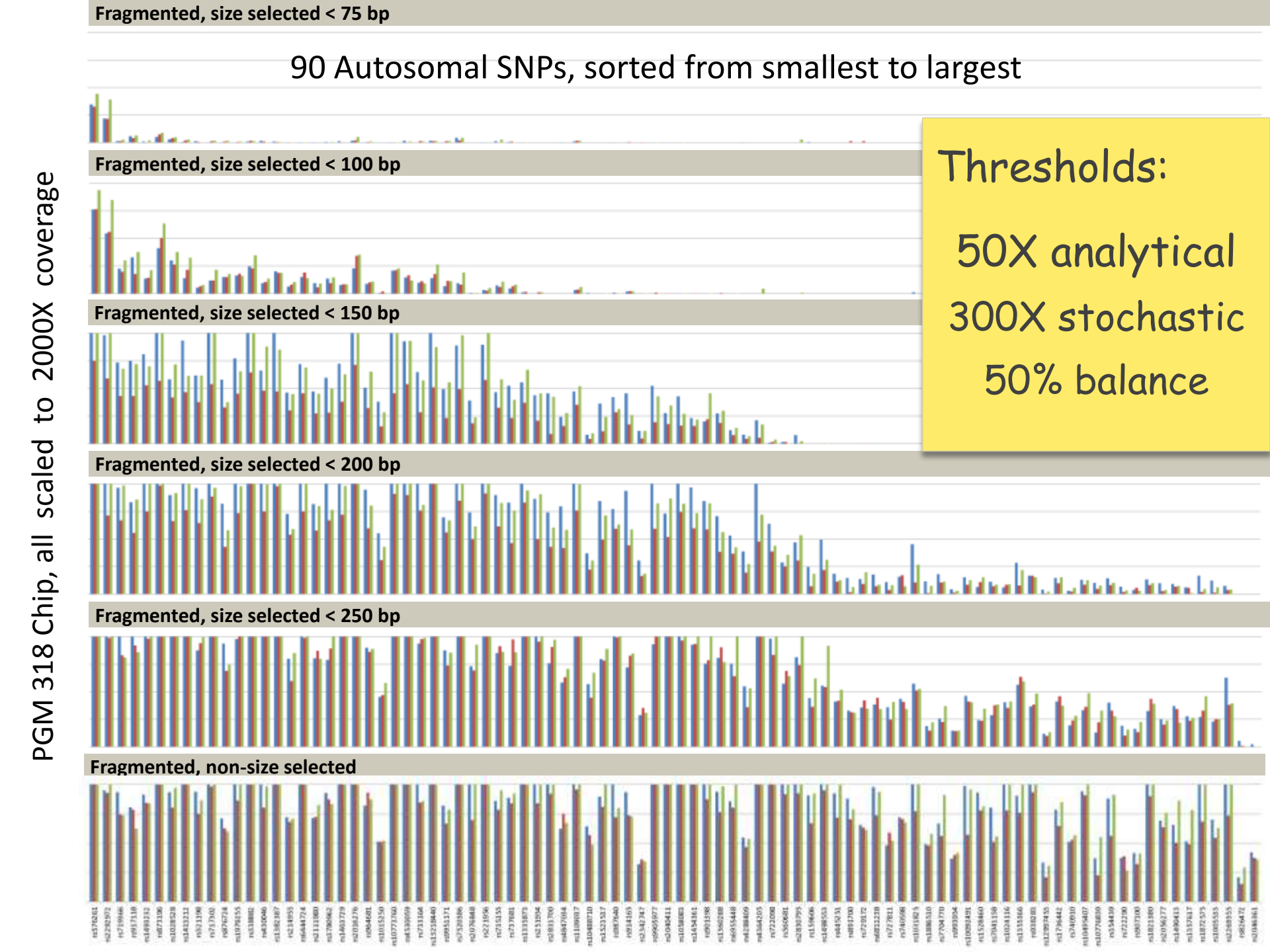


HID SNP Panel Degraded DNA Study

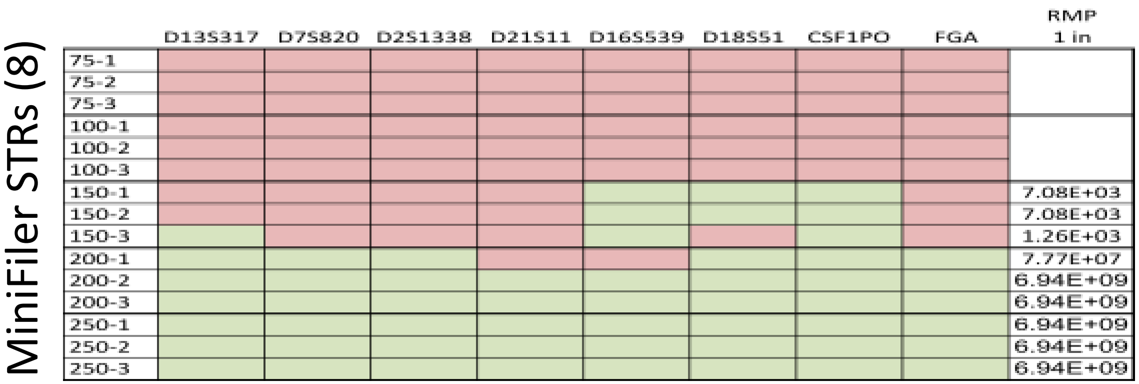


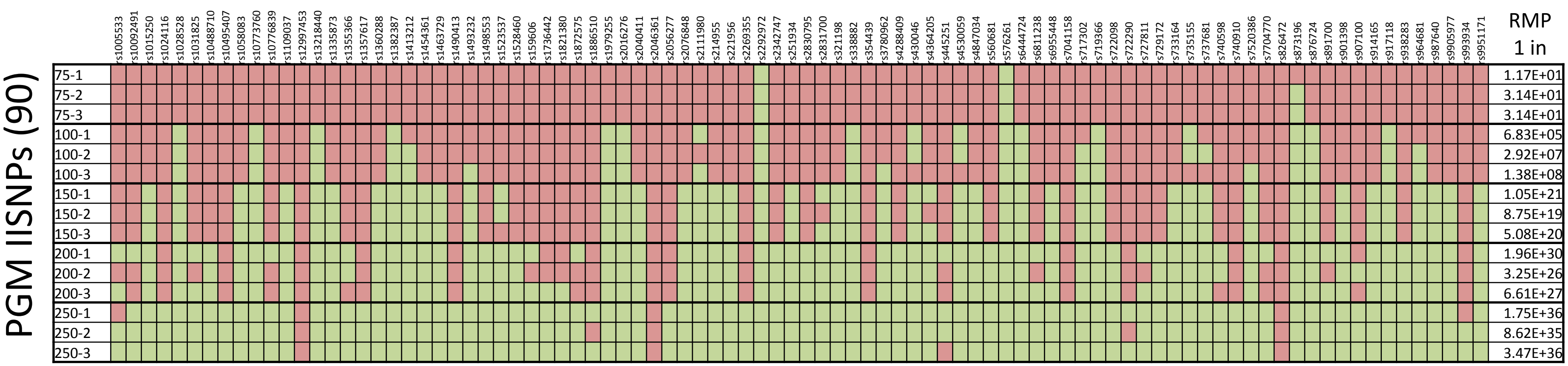
Minifiler® amplification (30 cycle), 25 µl reaction, 3500*xl* electrophoresis, 1.2 kV for 8 seconds

Thresholds: analytical = 100 RFU, PHR = 0.5; data scaled to 1000 RFU


Fragmented, size selected < 75 bp
90 Autosomal SNPs, sorted from smallest to largest
Fragmented, size selected < 100 bp
Thresholds:
50X analytical
300X stochastic
50% balance
Fragmented, size selected < 150 bp
Fragmented, size selected < 200 bp
Fragmented, size selected < 250 bp
Fragmented, non-size selected
PGM 318 Chip, all scaled to 2000X coverage

HID SNP Panel
Degraded DNA Study





HID SNP Panel
Degraded DNA Study

MiniFiler STRs (8)

	RMP 1 in
75-1	
75-2	
75-3	
100-1	
100-2	
100-3	
150-1	7.08E+03
150-2	7.08E+03
150-3	1.26E+03
200-1	7.77E+07
200-2	6.94E+09
200-3	6.94E+09
250-1	6.94E+09
250-2	6.94E+09
250-3	6.94E+09

PGM IISNPs (90)

	RMP 1 in
75-1	1.17E+01
75-2	3.14E+01
75-3	3.14E+01
100-1	6.83E+05
100-2	2.92E+07
100-3	1.38E+08
150-1	1.05E+21
150-2	8.75E+19
150-3	5.08E+20
200-1	1.96E+30
200-2	3.25E+26
200-3	6.61E+27
250-1	1.75E+36
250-2	8.62E+35
250-3	3.47E+36

SNPs have MANY more possible loci and better performance in degraded samples

HID SNP Panel Degraded DNA Study

HID SNP Panel Degraded DNA Study Summary

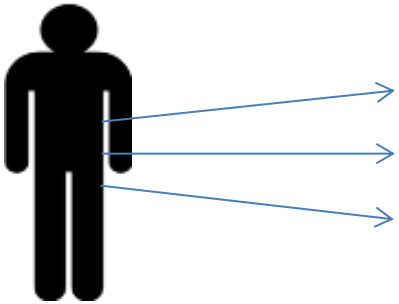
- SNPs and STRs show expected performance in each fraction based on amplicon size
- Some SNPs can still amplify in degraded samples where STRs cannot
- Due to the high number of SNPs, very high RMPs are possible
- Better STR assays (GlobalFiler or NGS-STR) may lessen the “gap”
- Validation needed for SNP thresholds

SNPs and Mixtures

At each SNP, a person is either homozygous or heterozygous.

Single source samples have 3 possible biallelic genotypes:

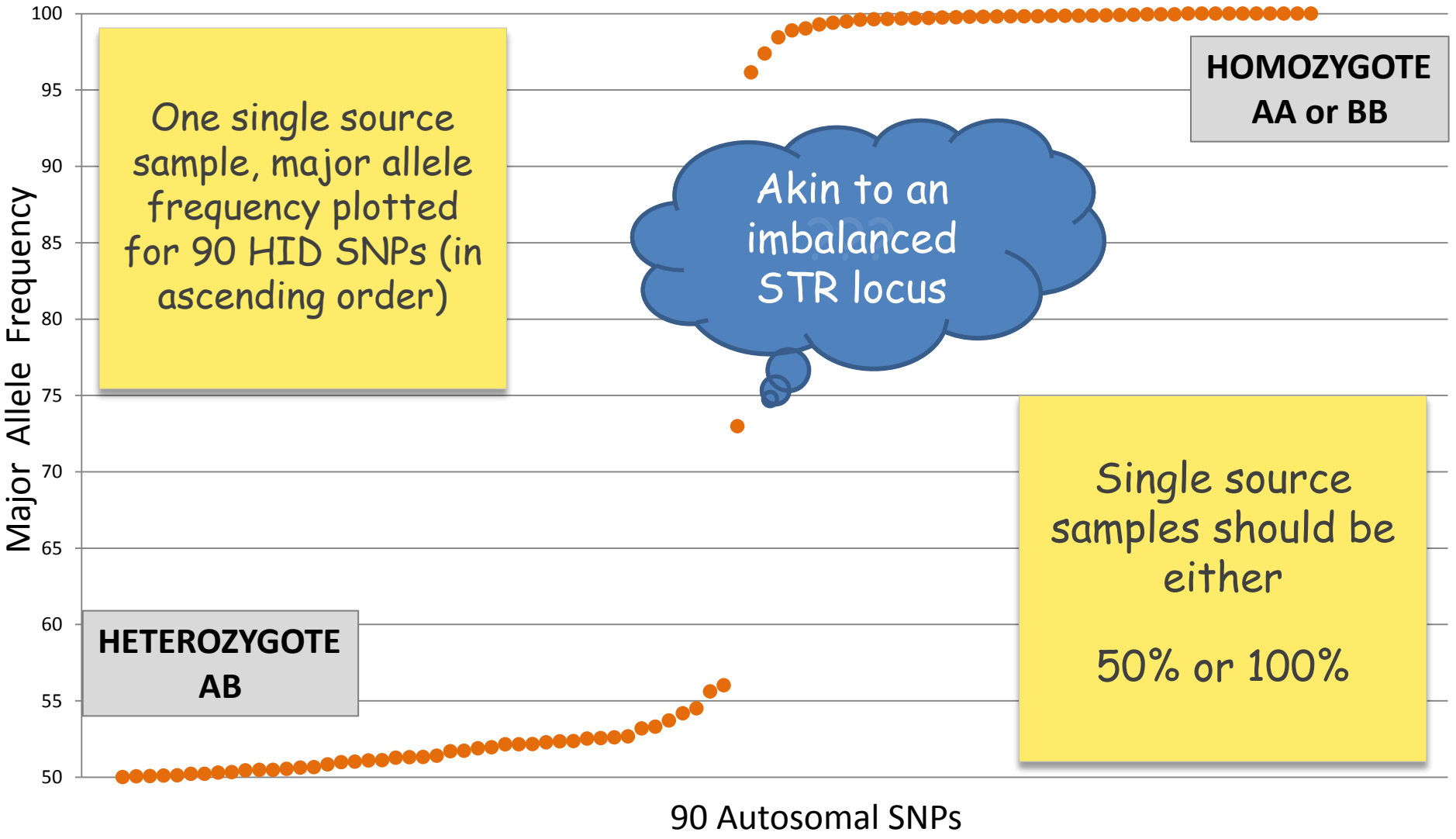
AA, BB, or AB



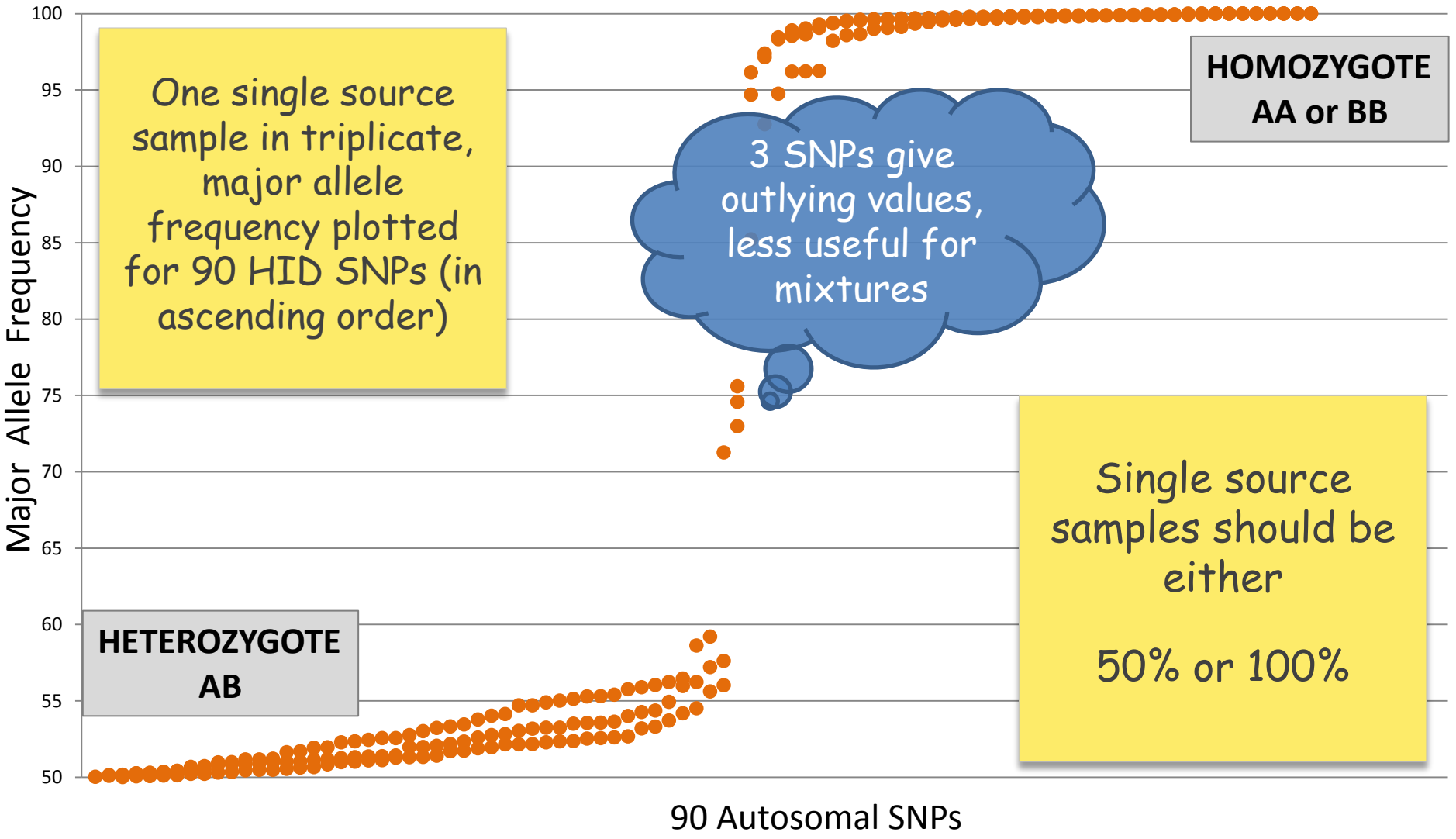
1	A	B
AA	100%	0%
AB	50%	50%
BB	0%	100%

% is coverage (like PH balance)

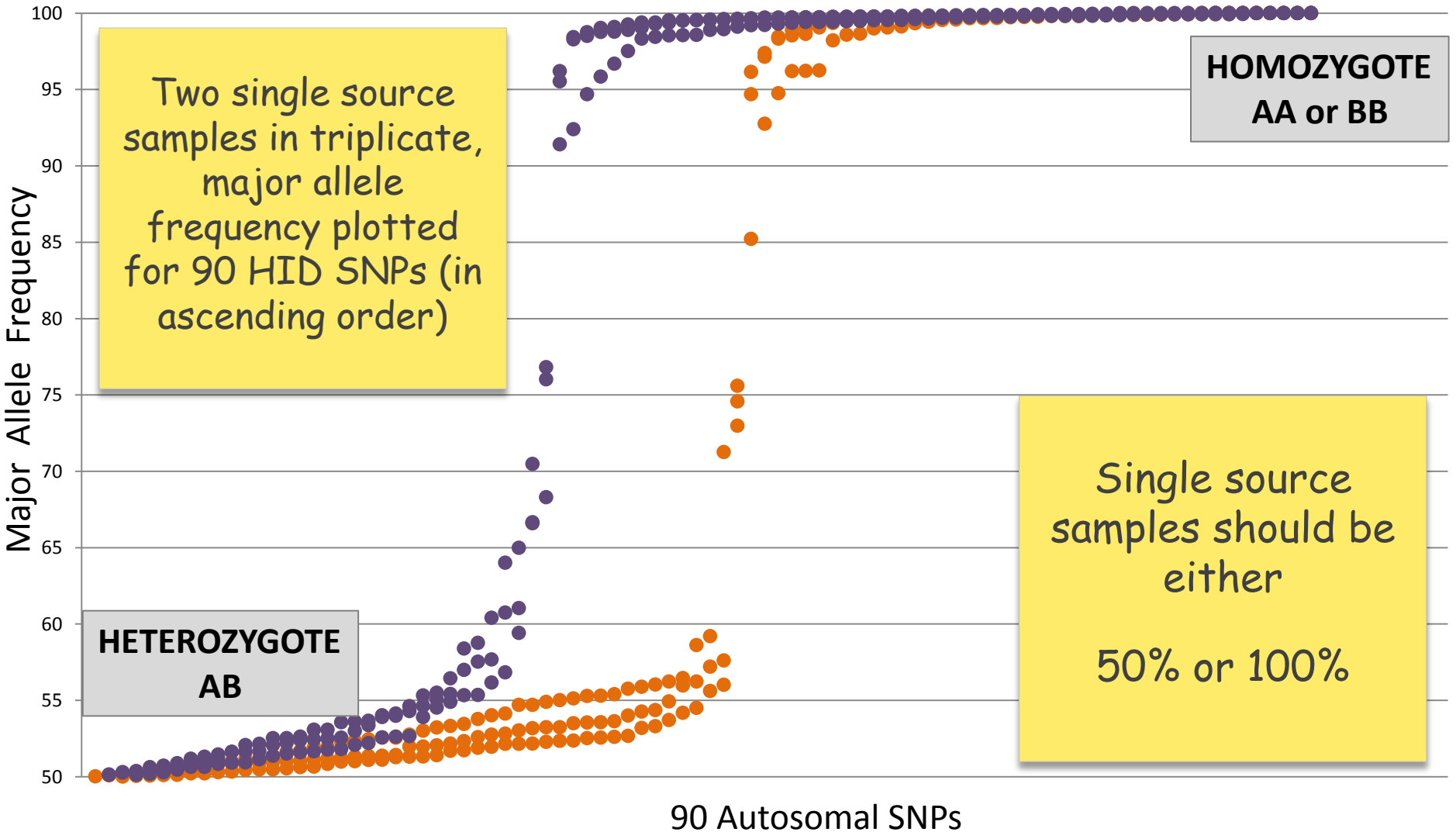
HID SNP Panel
Mixture Detection



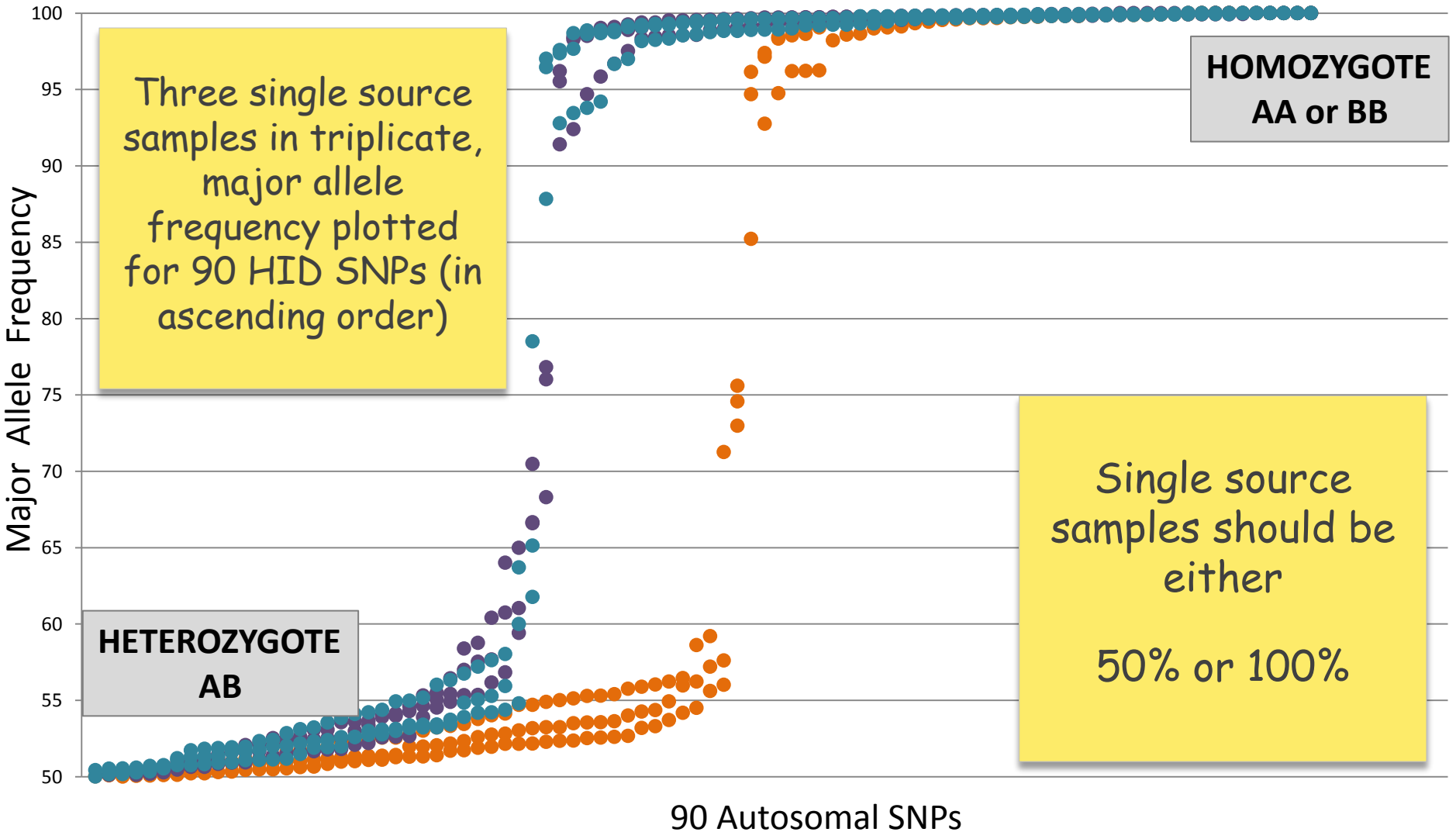
HID SNP Panel
Mixture Detection



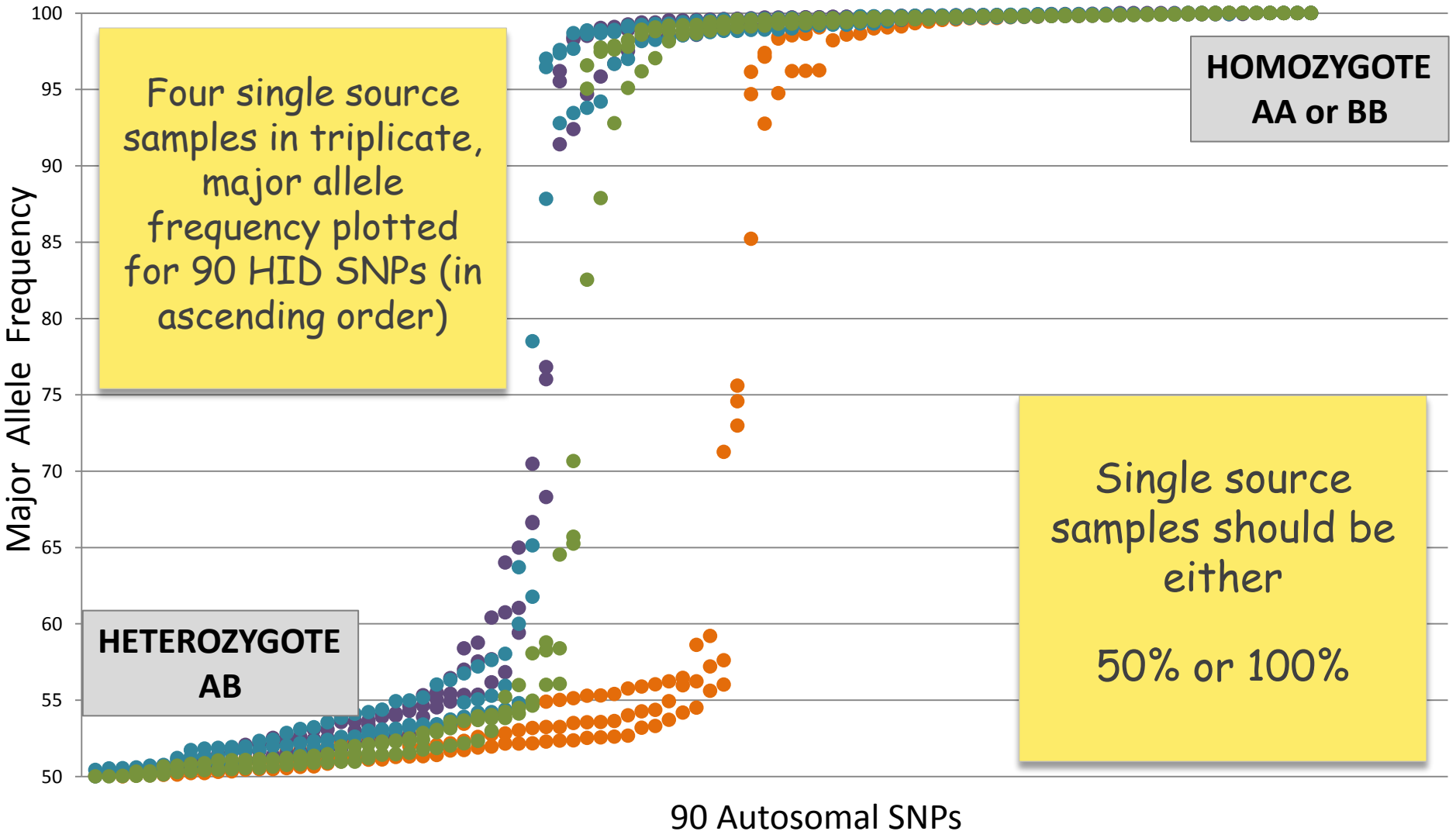
HID SNP Panel
Mixture Detection



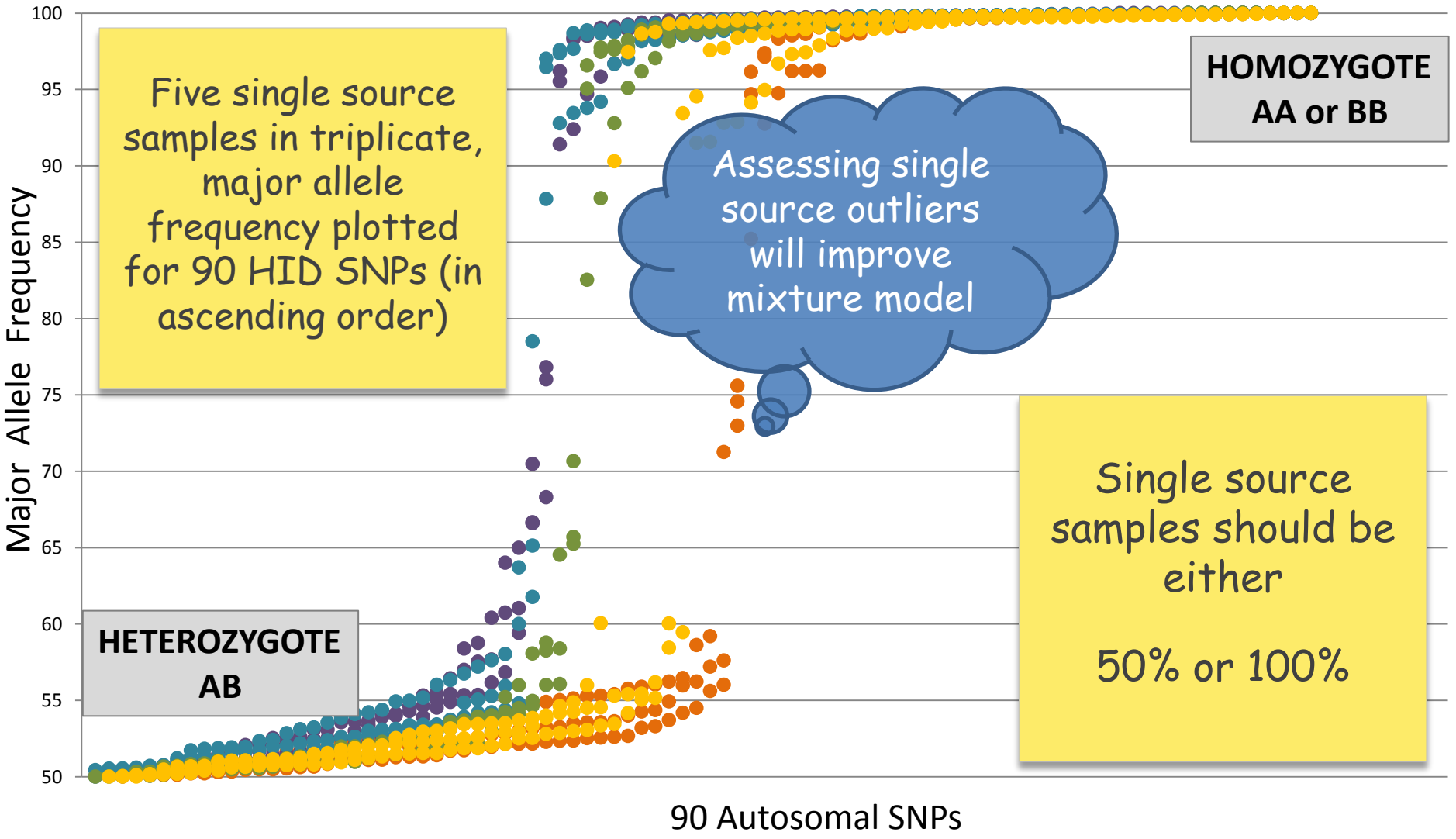
HID SNP Panel
Mixture Detection



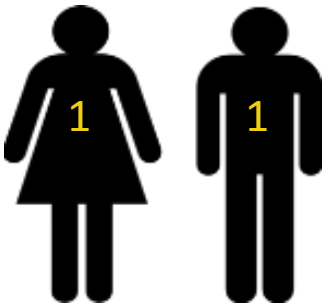
HID SNP Panel
Mixture Detection



HID SNP Panel
Mixture Detection



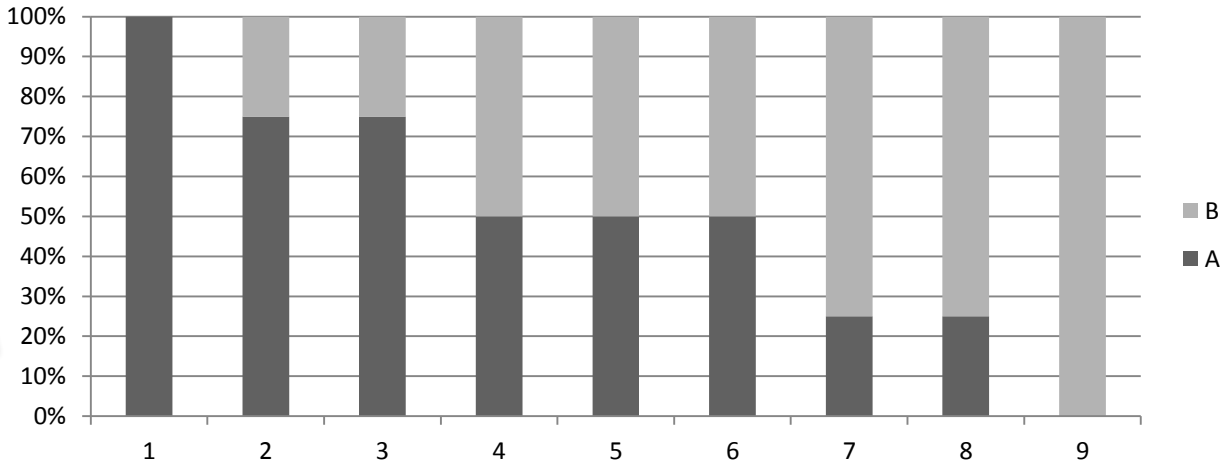
SNPs in 1:1 Mixtures



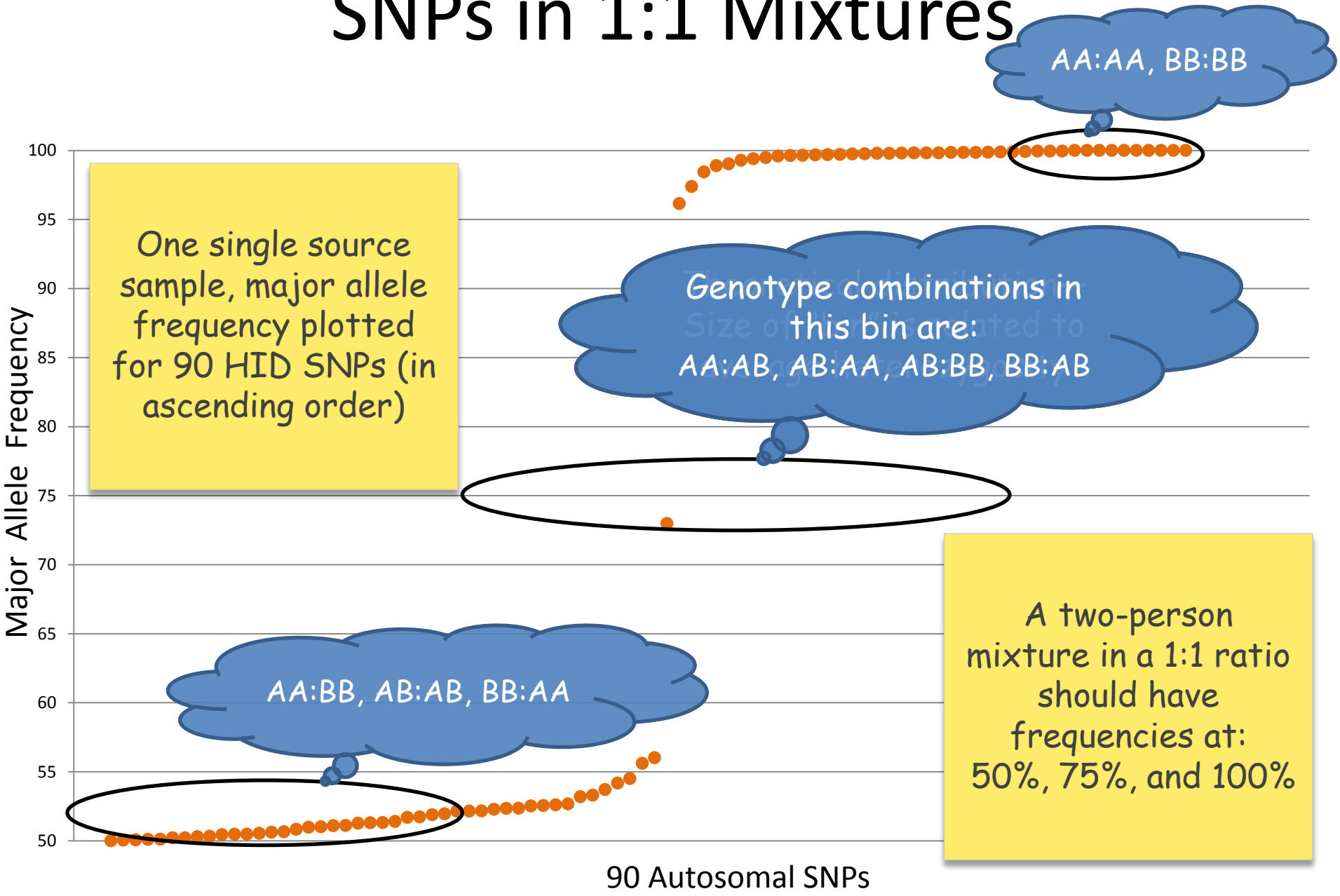
1	1	1A	1A	1B	1B	A	B
AA	AA	2	2	0	0	100%	0%
AA	AB	2	1	0	1	75%	25%
AB	AA	1	2	1	0	75%	25%
AA	BB	2	0	0	2	50%	50%
AB	AB	1	1	1	1	50%	50%
BB	AA	0	2	2	0	50%	50%
AB	BB	1	0	1	2	25%	75%
BB	AB	0	1	2	1	25%	75%
BB	BB	0	0	2	2	0%	100%

Two-person mixtures have 9 possible genotype combinations:

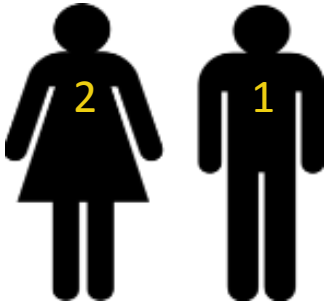
3 genotypes (Person 1)
x
3 genotypes (Person 2)



SNPs in 1:1 Mixtures



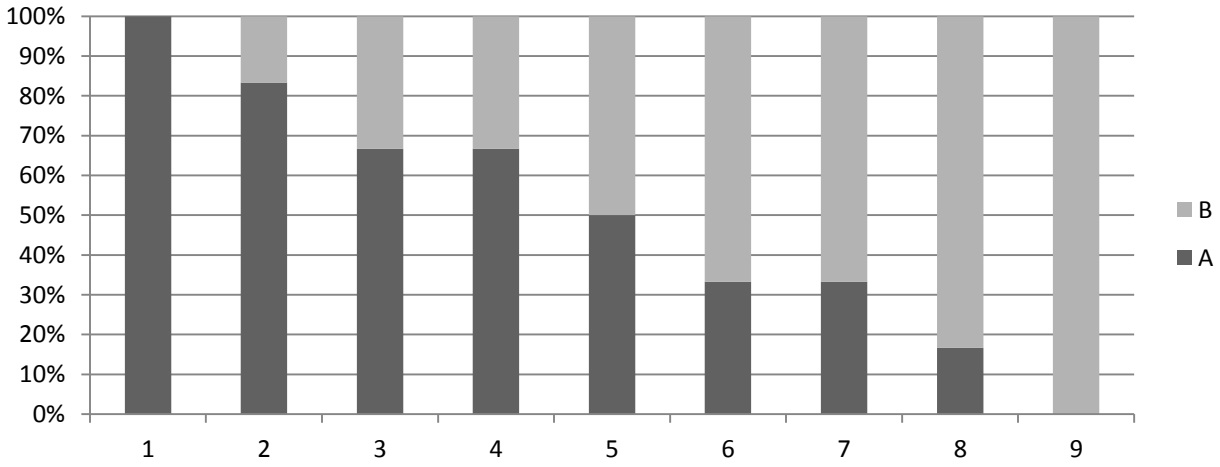
SNPs in 2:1 Mixtures



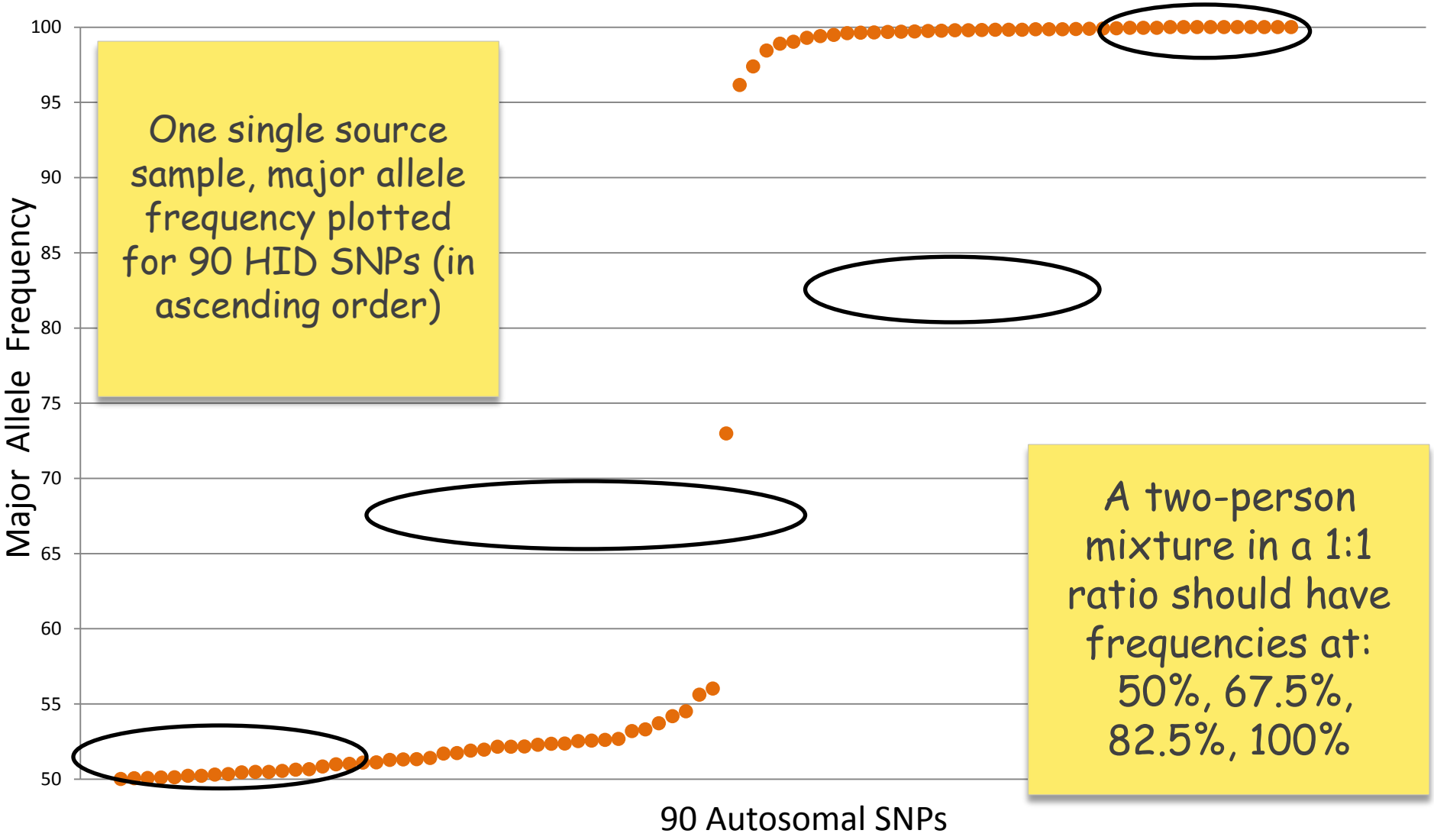
2	1	2A	1A	2B	1B	A	B
AA	AA	4	2	0	0	100%	0%
AA	AB	4	1	0	1	83%	17%
AA	BB	4	0	0	2	67%	33%
AB	AA	2	2	2	0	67%	33%
AB	AB	2	1	2	1	50%	50%
AB	BB	2	0	2	2	33%	67%
BB	AA	0	2	4	0	33%	67%
BB	AB	0	1	4	1	17%	83%
BB	BB	0	0	4	2	0%	100%

Two-person mixtures have 9 possible genotype combinations:

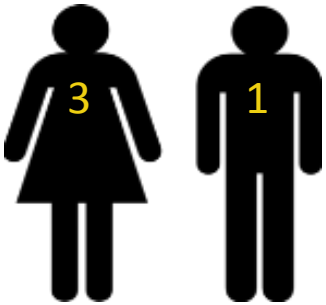
3 genotypes (Person 1)
x
3 genotypes (Person 2)



SNPs in 2:1 Mixtures



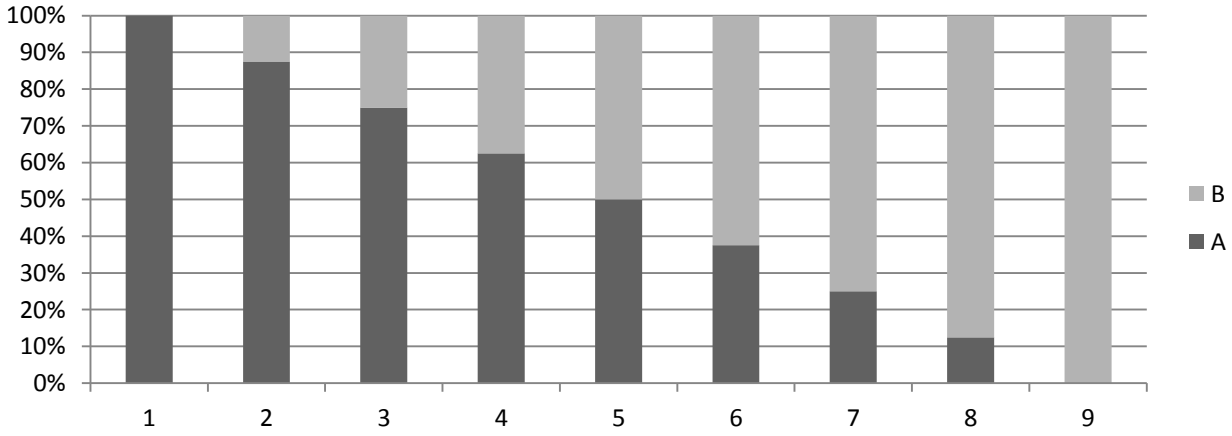
SNPs in 3:1 Mixtures



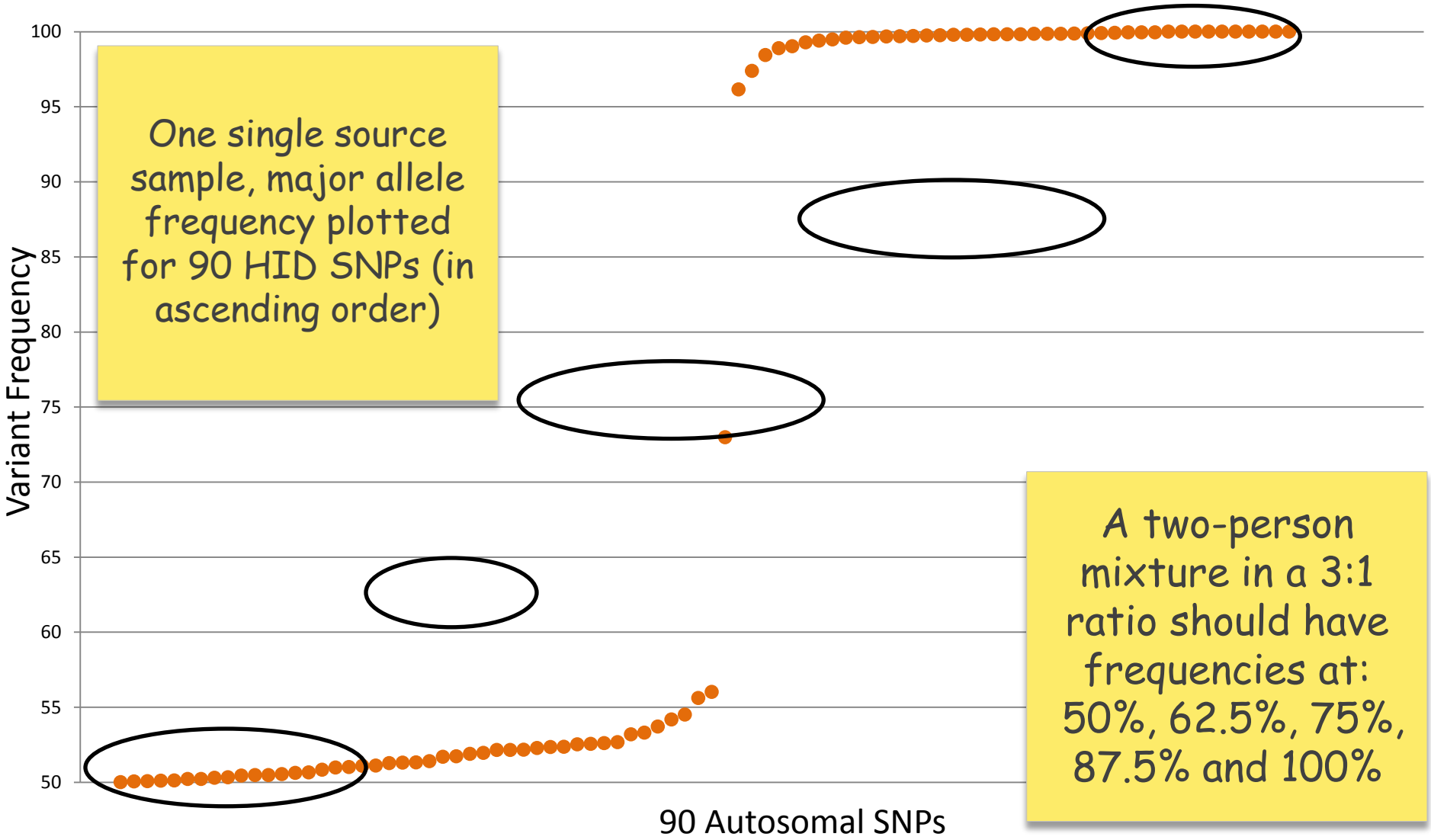
3	1	3A	1A	3B	1B	A	B
AA	AA	6	2	0	0	100%	0%
AA	AB	6	1	0	1	88%	13%
AA	BB	6	0	0	2	75%	25%
AB	AA	3	2	3	0	63%	38%
AB	AB	3	1	3	1	50%	50%
AB	BB	3	0	3	2	38%	63%
BB	AA	0	2	6	0	25%	75%
BB	AB	0	1	6	1	13%	88%
BB	BB	0	0	6	2	0%	100%

Two-person mixtures have 9 possible genotype combinations:

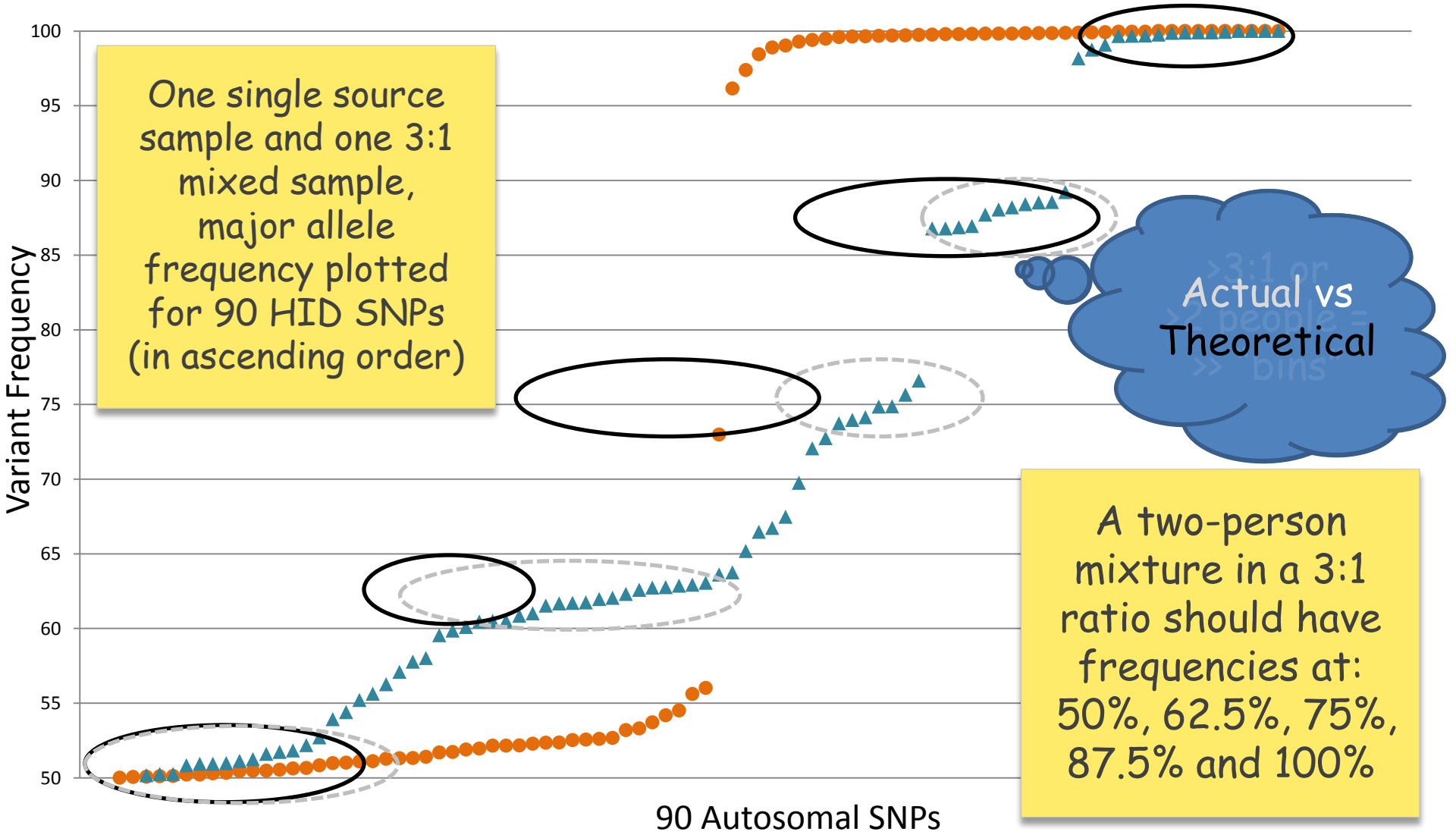
3 genotypes (Person 1)
x
3 genotypes (Person 2)



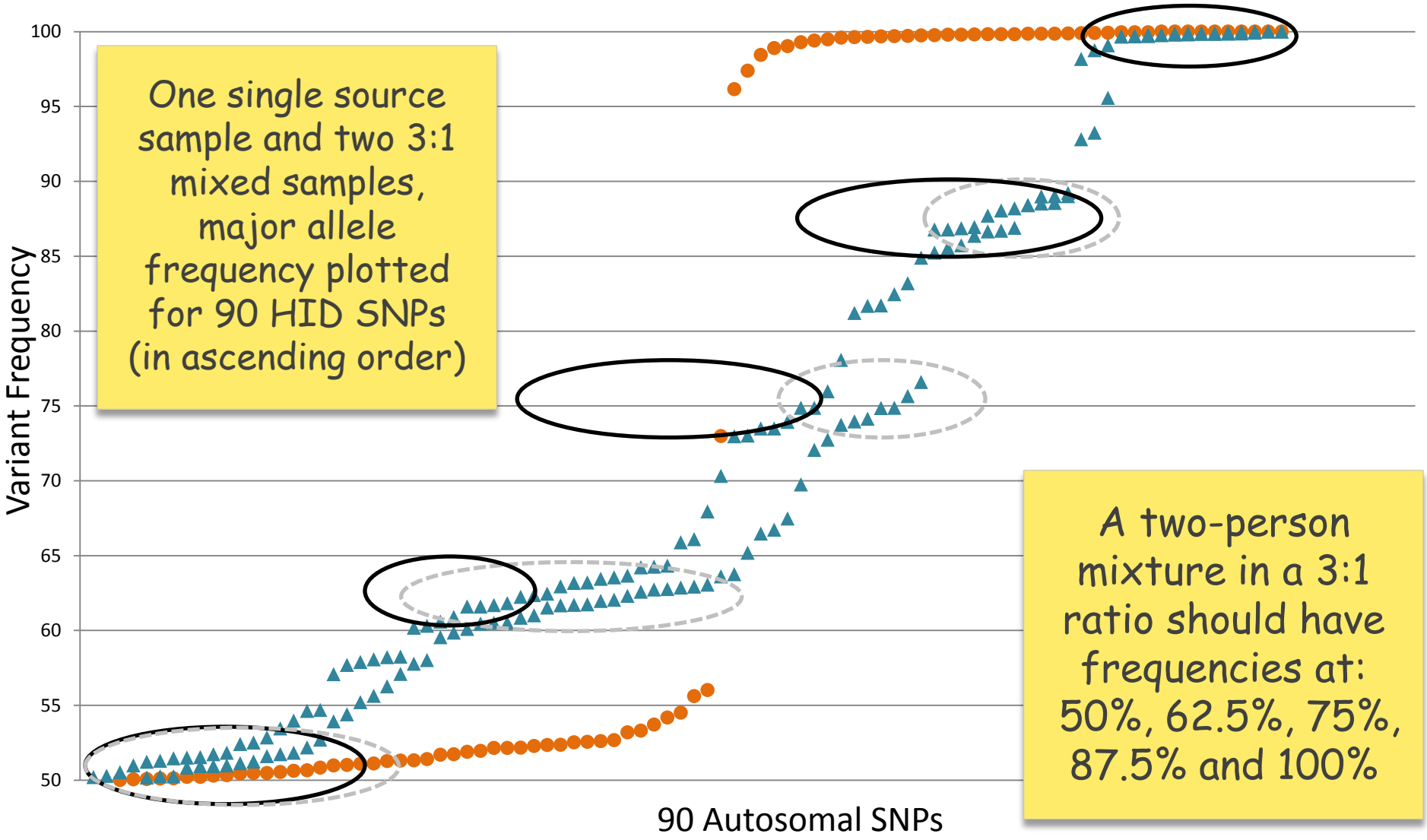
SNPs in 3:1 Mixtures



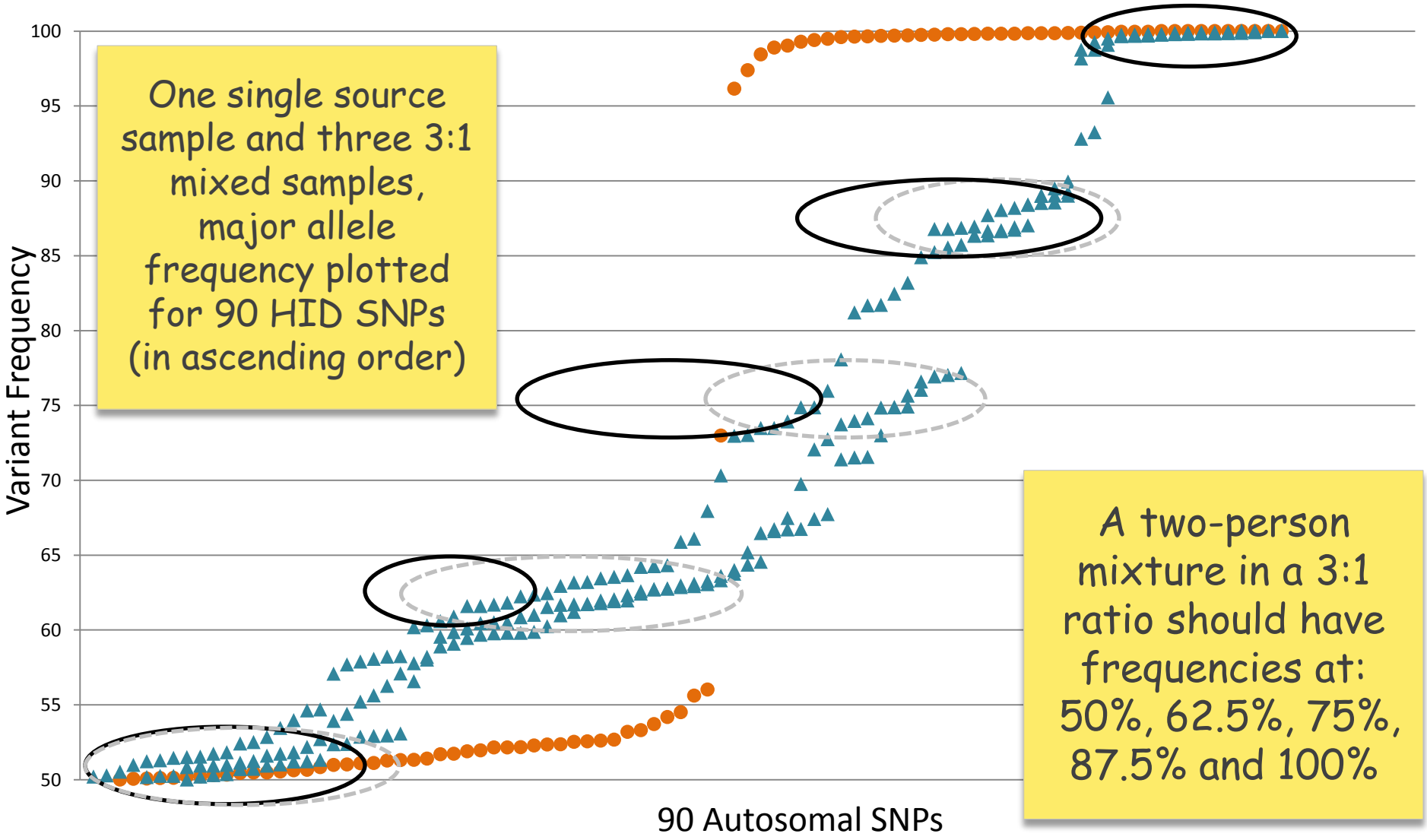
SNPs in 3:1 Mixtures



SNPs in 3:1 Mixtures

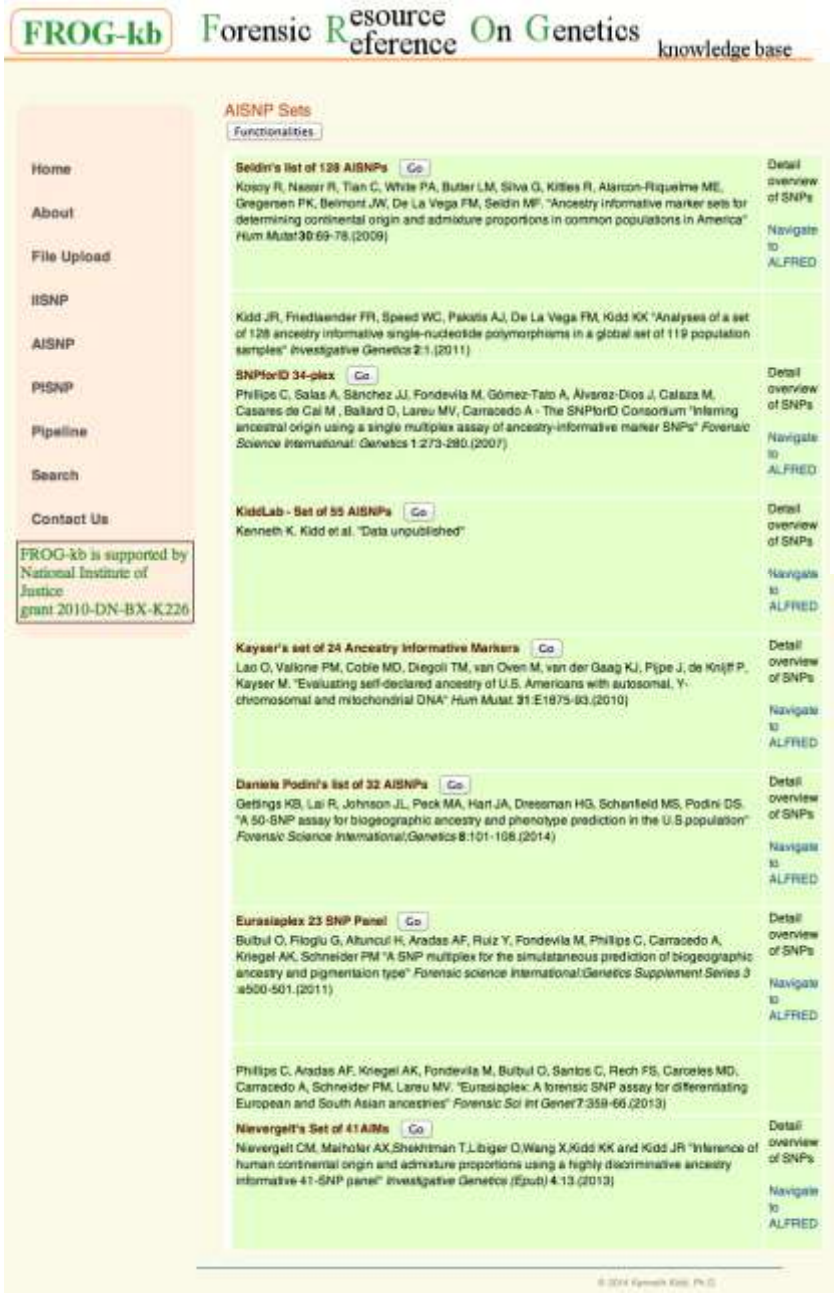


SNPs in 3:1 Mixtures



HID SNP Panel Mixtures Summary

- Mixtures can be detected in SNP data based on the coverage levels at heterozygous loci
- It may be possible to determine two-person 1:1 or 2:1 mixtures (maaaybe 3:1)
- More than two contributors or greater than 3:1 mixtures will be difficult to distinguish
- Need to determine which SNPs “behave”
- Stay tuned!



PGM AIM Panel (beta testing)

- Ampliseq library prep
- 170 SNPs
- Seldin 128
- Kidd 55
- Analysis plug-in integrates FROGkb

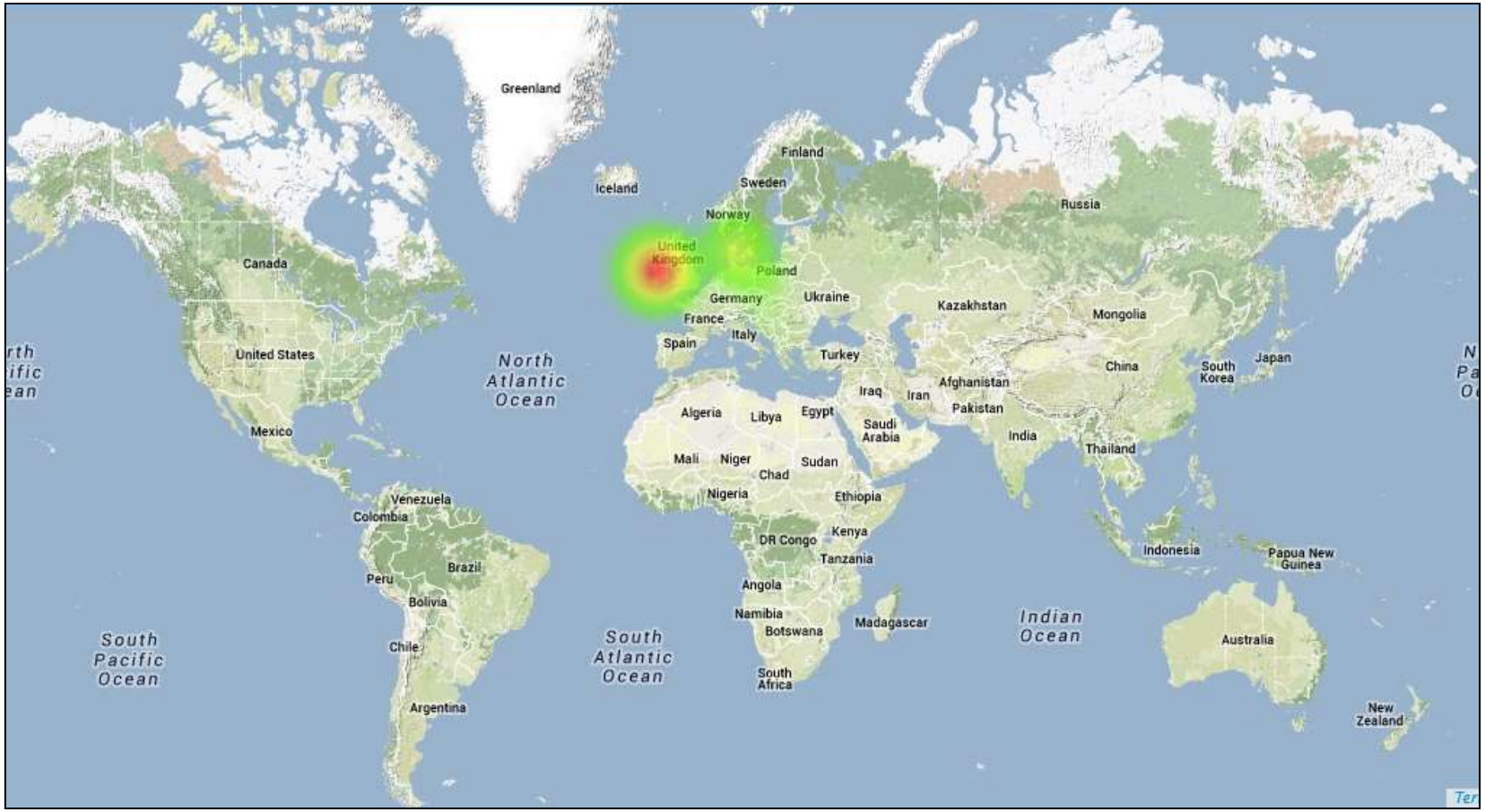
AIM Panel
Ancestry Prediction – SRM 2391c

- Likelihood Ratio calculations
  - Four categories extant in both Kidd and Seldin studies
    - Europeans, African Americans, Maya, and Han Chinese
  - Allows comparison of SNP sets' performance
  - Representative of major U.S. populations

SRM 2391c Component	**Gender**	**Ethnicity** (self declared)
A	Female	Not listed
B	Male	Mexican-American
C	Male	Melanesian
D	Female:Male	Mixed sample
E	Female	Not listed
F	Male	Caucasian

HID SNP Genotyper Plugin (v4.1 Beta) New Feature – Ancestry Map

- Heatmap of highest probability of origin



Ancestry Prediction
SRM 2391c Component A

SRM 2391c Component	Gender	Ethnicity	Kidd 55 Prediction	Seldin 128 Prediction
A	Female	Not listed	European 1.02×10^{33}	European 6.32×10^{66}

Kidd 55 SNPs

Seldin 128 SNPs

Ancestry Prediction
SRM 2391c Component B

SRM 2391c Component	Gender	Ethnicity	Kidd 55 Prediction	Seldin 128 Prediction
B	Male	Mexican-American	European 5.39×10^{12}	Han Chinese 1.48×10^{19}

Kidd 55 SNPs

Seldin 128 SNPs

Ancestry Prediction
SRM 2391c Component C

SRM 2391c Component	Gender	Ethnicity	Kidd 55 Prediction	Seldin 128 Prediction
C	Male	Melanesian	Han Chinese 1.54×10^{14}	Han Chinese 6.67×10^{28}

Kidd 55 SNPs

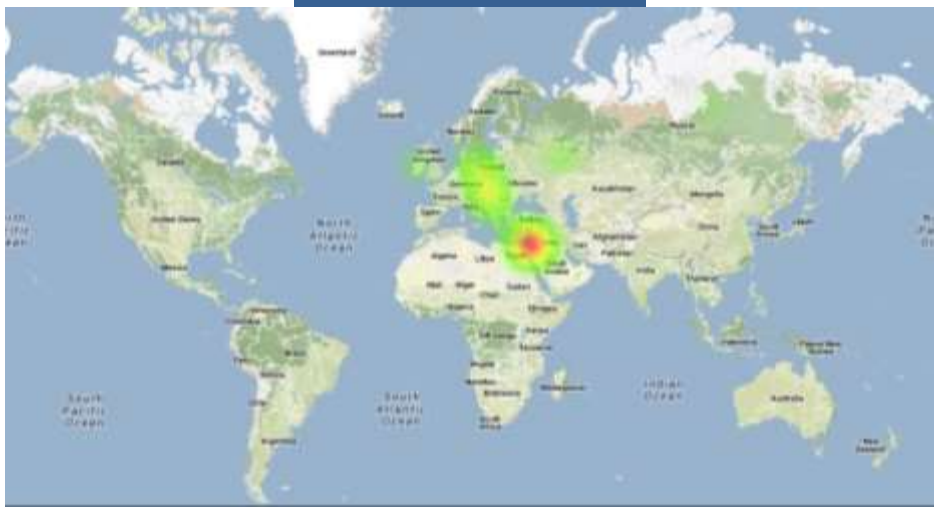
Seldin 128 SNPs

Ancestry Prediction
SRM 2391c Component E

SRM 2391c Component	Gender	Ethnicity	Kidd 55 Prediction	Seldin 128 Prediction
E	Female	Not listed	European 5.41×10^{21}	European 3.92×10^{50}

Kidd 55 SNPs

Seldin 128 SNPs



Ancestry Prediction
SRM 2391c Component F

SRM 2391c Component	Gender	Ethnicity	Kidd 55 Prediction	Seldin 128 Prediction
F	Male	Caucasian	European 2.35×10^{31}	European 1.16×10^{55}

Kidd 55 SNPs

Seldin 128 SNPs

HID SNP Panel Ancestry Summary

- 170 SNP panel containing two SNP sets that are suitable for use in U.S.
- Plug-in integrates FROG-kb (http://frog.med.yale.edu/FrogKB/)
- Heat maps give quick overview
- Interpretation tools being developed
  - Combining loci
  - Choosing/combining populations

Conclusions

- NGS can give more information on currently used forensic markers
  - More STRs and STR sequence info
  - Whole genome mtDNA
- NGS facilitates genotyping of forensic SNPs
- SNPs may help with low level & degraded samples
- SNPs may provide ancestry (and phenotype?) information
- Forensic NGS kits/methods are being developed
- Many questions to answer prior to implementation

Acknowledgements

THANK YOU

Dr. Peter Vallone
Group Leader

Kevin Kiesler
Research Biologist

Funding from the
FBI Biometrics Center of Excellence
Forensic DNA Typing as a Biometric Tool

Thermo Fisher (Life Tech):
Nnamdi Ihuegbu
Robert Lagace

NIST
Applied
Genetics

Thank you for your attention!

Contact Info:
katherine.gettings@nist.gov
301-975-6401