The Role of Microsatellite Variation in Cancer:
New Technological Approaches for Biomarker Discovery from Within Our Genomic Repetitive DNA

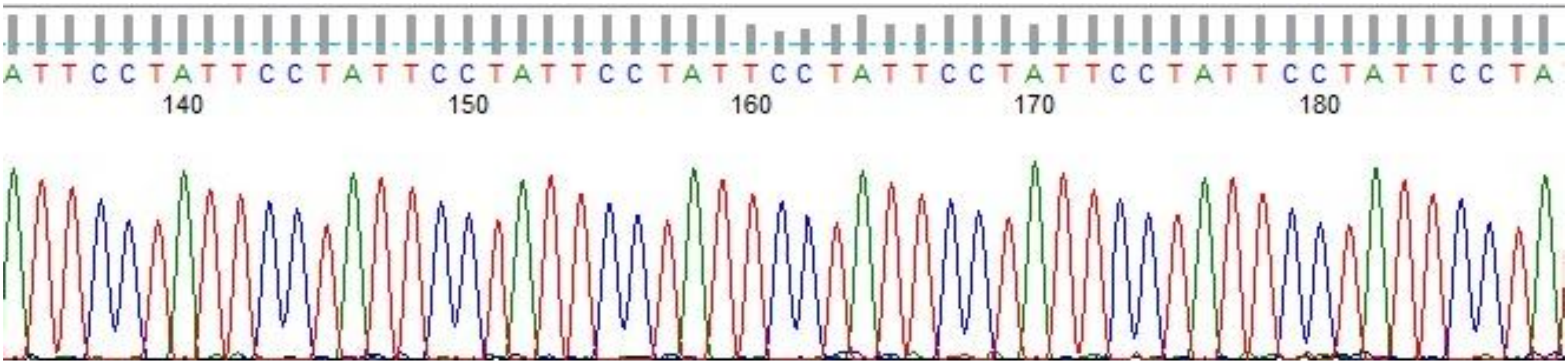
And extensions to Autism

Virginia Bioinformatics Institute
Virginia Tech

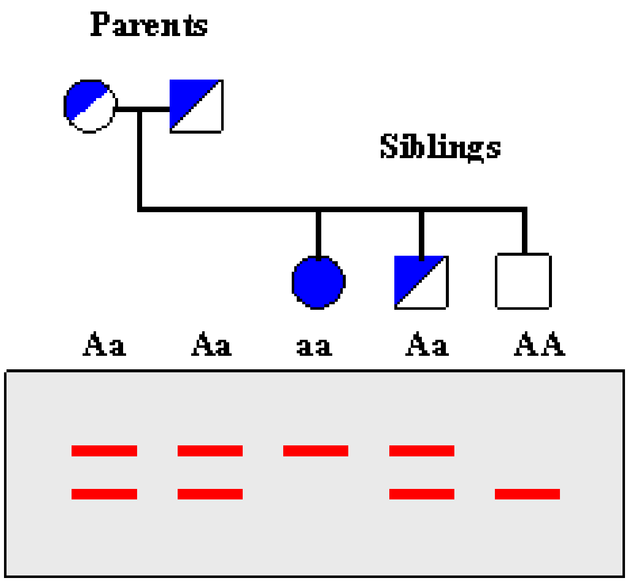
Analysis of the human genome has focused on SNPs. There is a large discrepancy between the known heritability of disease and the genetic component that can be explained by SNPs. The other variable genomic component, repeated DNA, may account for the missing genetic disease component.

Microsatellites are understudied despite playing a role in a number of diseases: Machado-Joseph (CAG repeat), Haw River Syndrome (CAG), Huntington’s Disease (CAG), some forms of Fragile-X Syndrome (CGG), Friedreich’s Ataxia (GAA), Myotonic Dystrophy (CAG), to name a few.

What are Microsatellites?



- Microsatellites are repetitive DNA sequences, typically 1-6 bases are repeated
- There are ~500,000 to 2,000,000 such repetitive regions in the human genome
- They are highly variable, much more than single nucleotide polymorphisms (SNPs)
- They are the key element in forensics and paternity testing



There are many of Microsatellite repeat loci in mammalian genomes – the human genome has about 2 million

Location	Number of Microsatellites	Number of Variable Microsatellites
Upstream	14,671	4,032
5'UTR	106,065	27,660
Intron	641,627	166,319
Exon	4,908	1,124
3'UTR	24,879	5,141
Downstream	12,789	3,609
In/near Gene Regions	804,959	207,885
Intergenic	1,101,147	299,705
Total	1,906,106	507,590

Only genes in the RefSeq database were included.

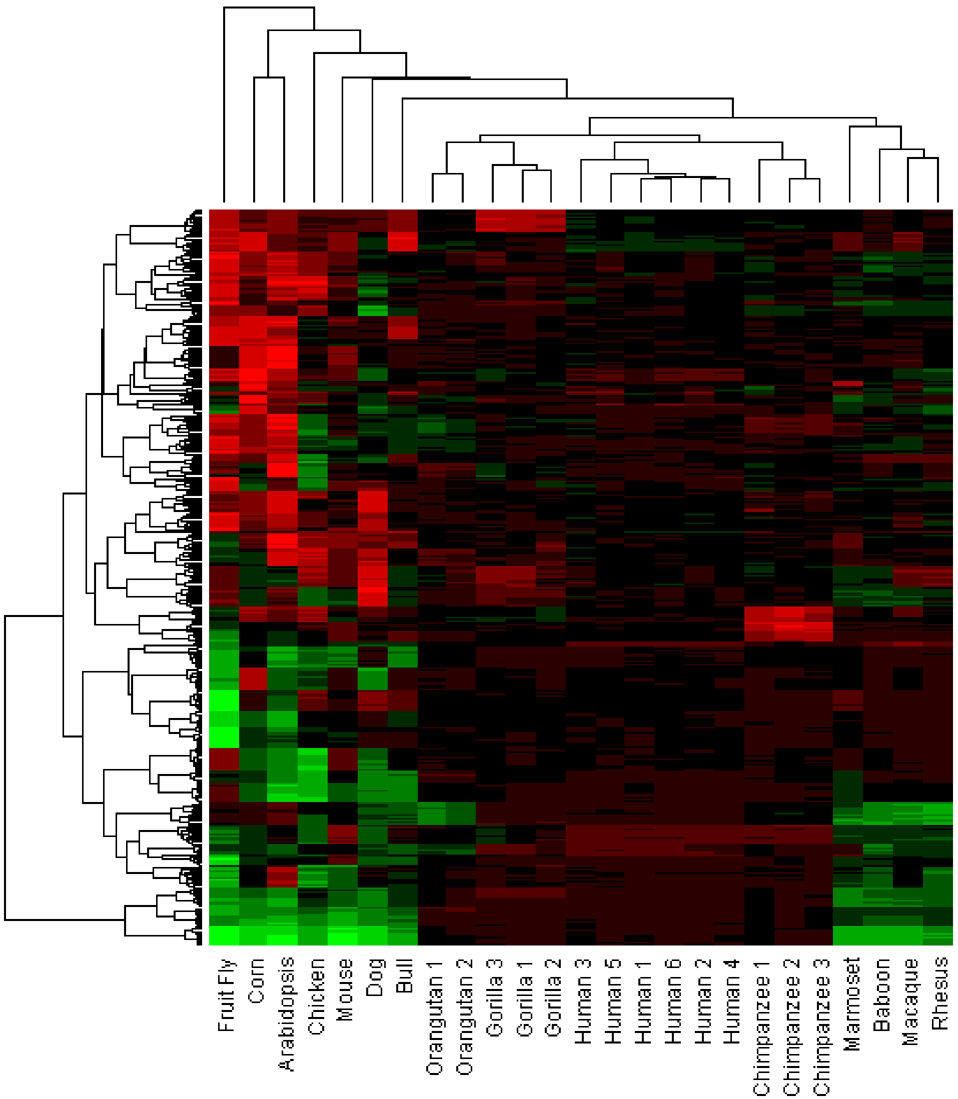
A “count” is defined as a complete tandem repeat at least 18 bp (for 3-mers and 6-mers) or 20 bp (for 1-, 2-, 4-, 5-, and 6-mers), in length.

*defined as 1,000 bp distal from the transcribed gene

Repetitive DNAs, microsatellites: do they contribute to defining a species or just provide natural variation among individuals?

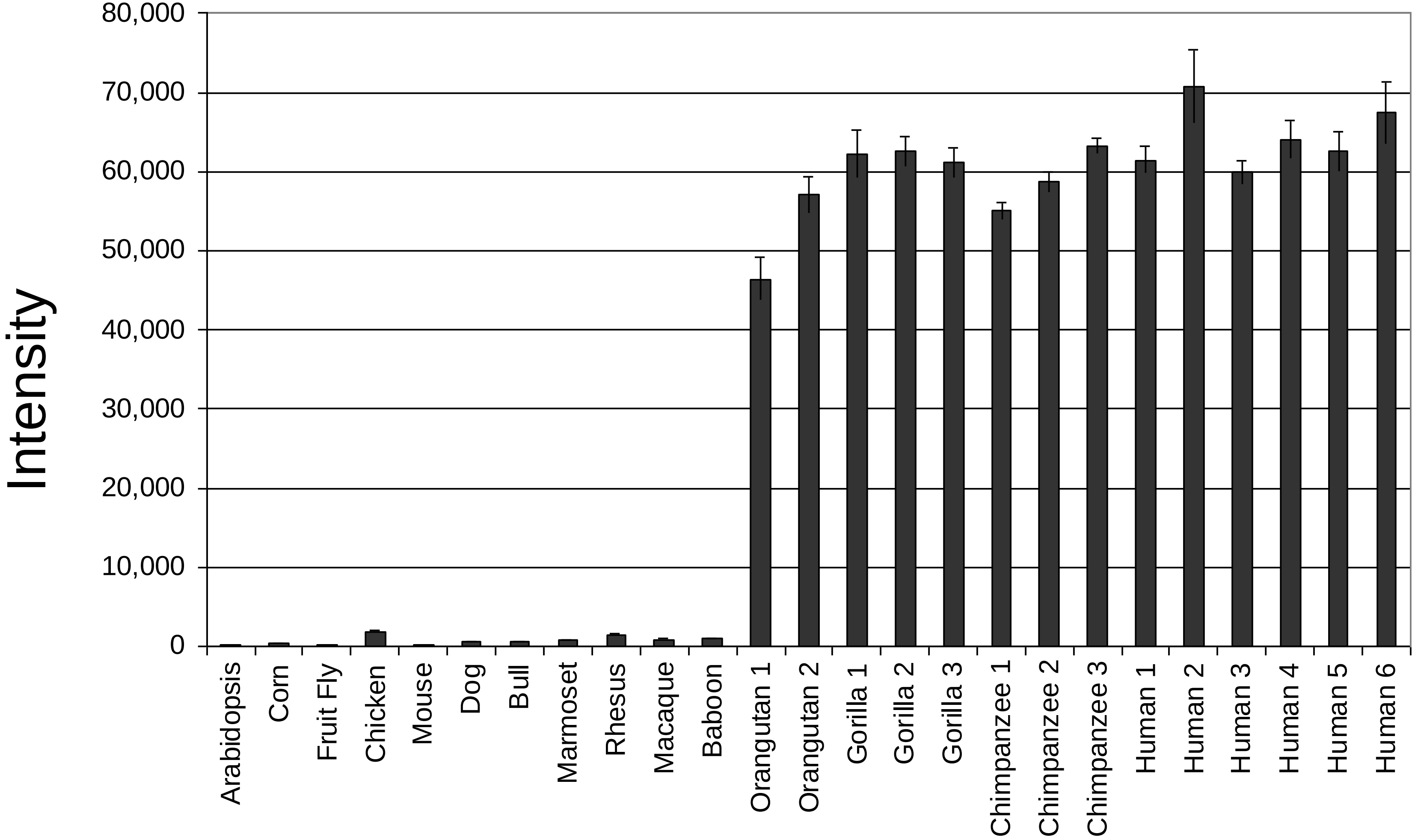
Microsatellite probes can differentiate genomes and build taxonomic relationships

- DNA extracted from various species was hybridized to the array
- Un-biased cluster analysis of hybridization patterns can easily distinguish species into accepted phylogenic relationships
- Analysis did not require previous knowledge of the species genome sequence
- A phylogenic/taxonomic tree emerged that resembles generally accepted relationships

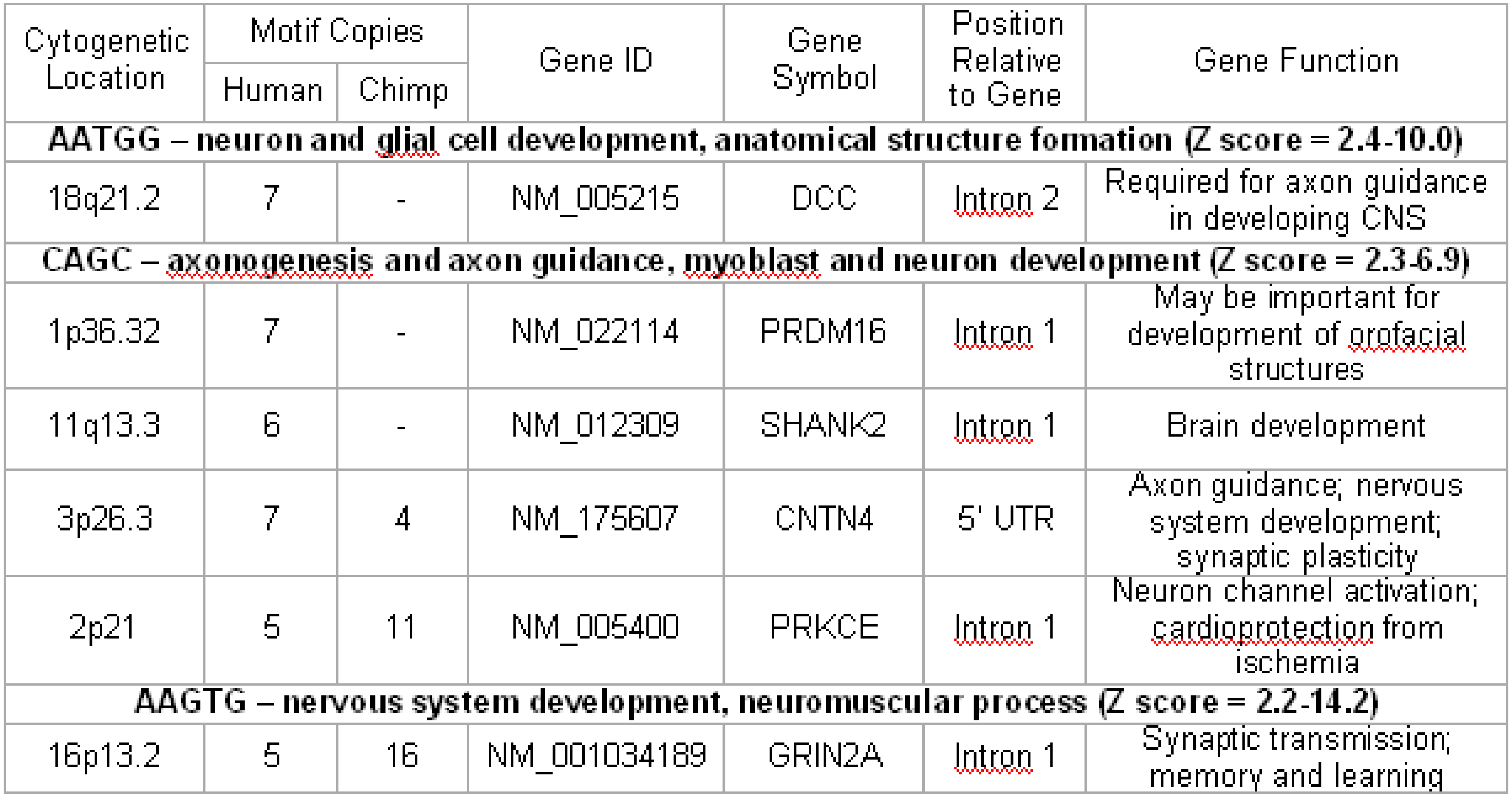


Molecular Biology and Evolution

And some motifs have emerged recently - AATGG motif is pronounced in hominids



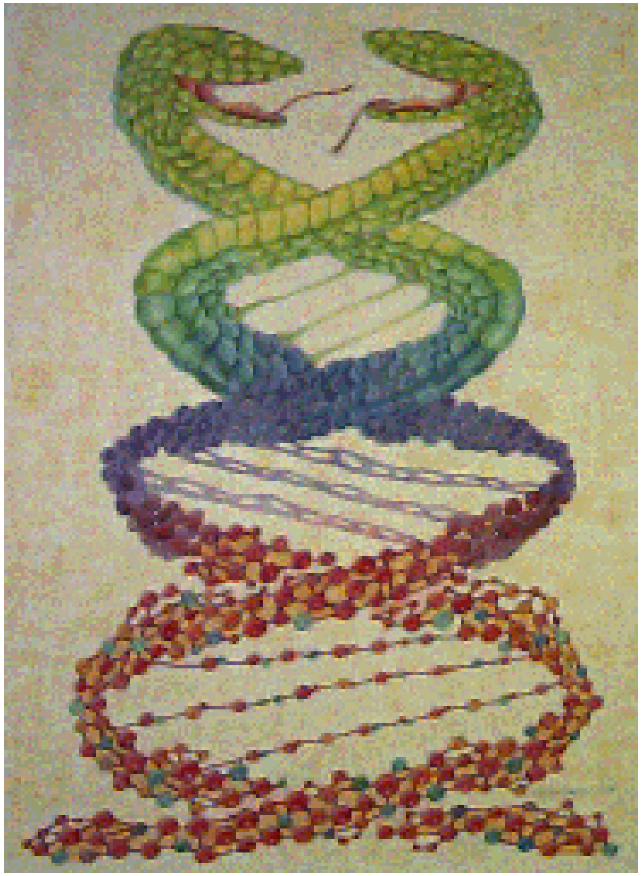
There are many, many genes containing motifs that were found to be differential, and they fall into ontological categories that could explain human/chimp differences.



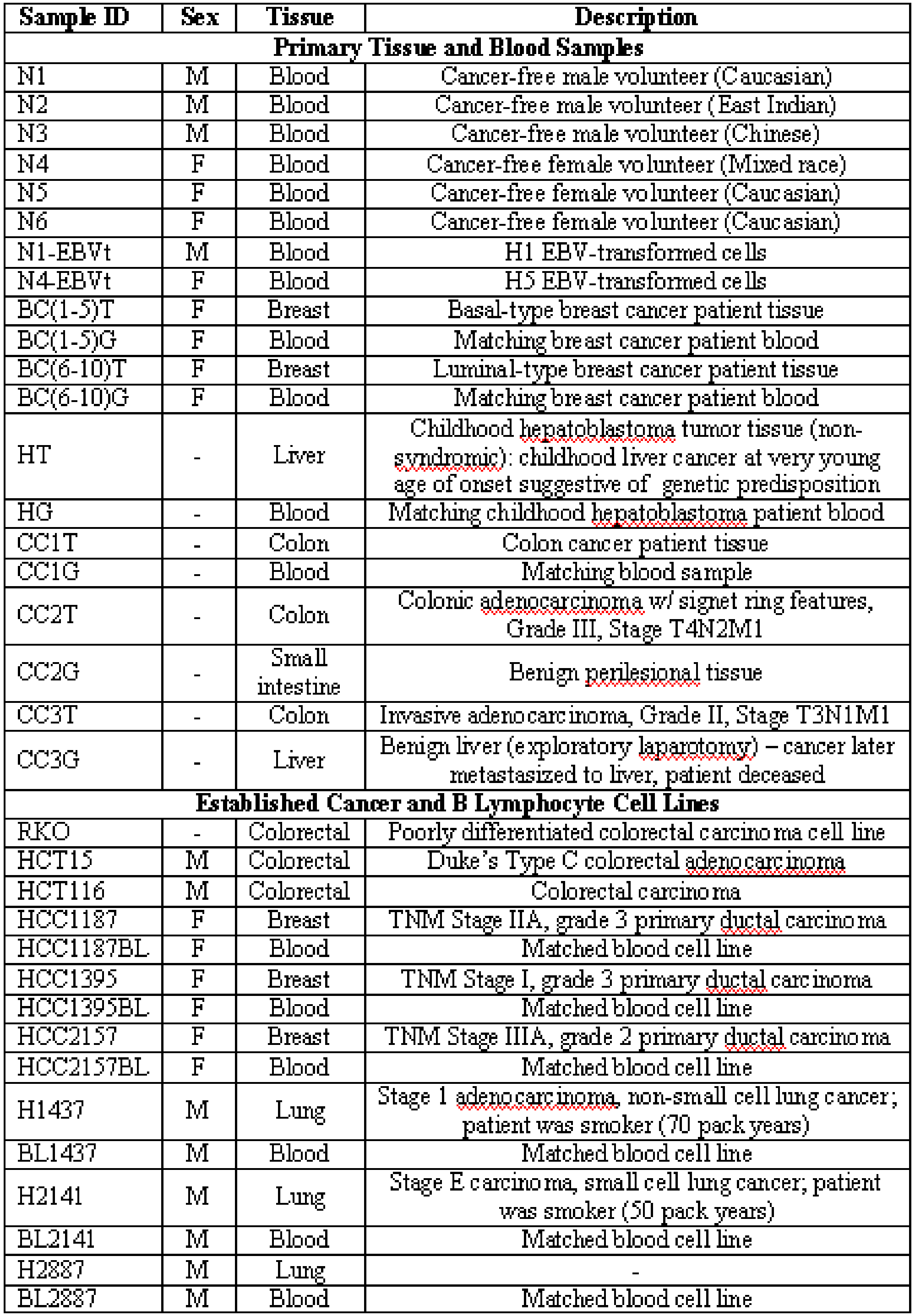
Cytogenetic Location	Motif Copies		Gene ID	Gene Symbol	Position Relative to Gene	Gene Function
	Human	Chimp				
AATGG – neuron and glial cell development, anatomical structure formation (Z score = 2.4-10.0)						
18q21.2	7	-	NM_005215	DCC	Intron 2	Required for axon guidance in developing CNS
CAGC – axonogenesis and axon guidance, myoblast and neuron development (Z score = 2.3-6.9)						
1p36.32	7	-	NM_022114	PRDM16	Intron 1	May be important for development of orofacial structures
11q13.3	6	-	NM_012309	SHANK2	Intron 1	Brain development
3p26.3	7	4	NM_175607	CNTN4	5' UTR	Axon guidance; nervous system development; synaptic plasticity
2p21	5	11	NM_005400	PRKCE	Intron 1	Neuron channel activation; cardioprotection from ischemia
AAGTG – nervous system development, neuromuscular process (Z score = 2.2-14.2)						
16p13.2	5	16	NM_001034189	GRIN2A	Intron 1	Synaptic transmission; memory and learning

So, what can we learn about cancer with this Global Microsatellite Content technique?

Each microarray also contained probes for transcription factor binding sites, ultraconserved regions and other repetitive elements, and no statistically significant and reproducible differences were observed across all samples, including probes that measure the content of ALU, SINE and LINE elements.

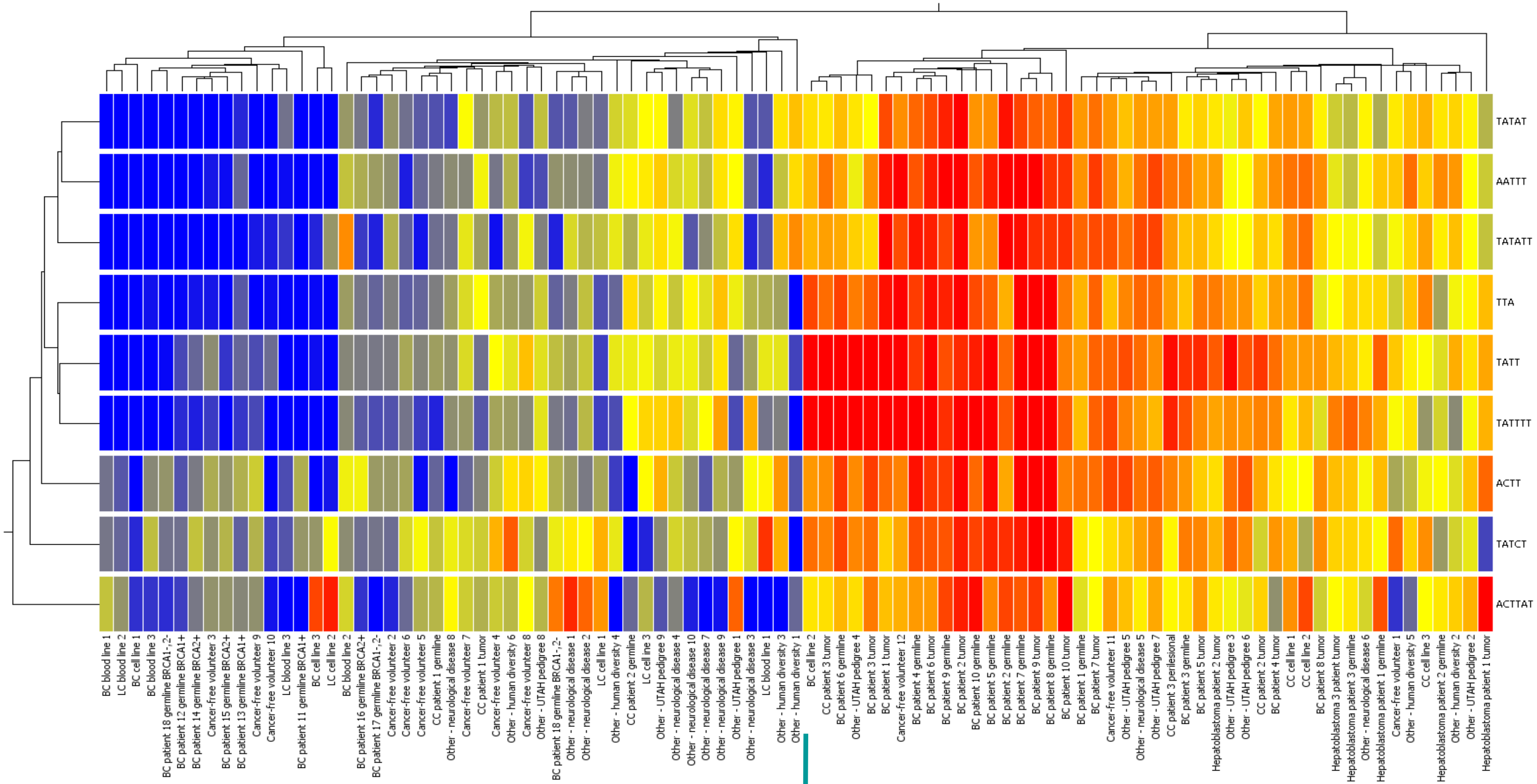


Samples included patient tumor and germline, ‘normals (cancer-free volunteers)’ and matched cell lines



Sample ID	Sex	Tissue	Description
Primary Tissue and Blood Samples			
N1	M	Blood	Cancer-free male volunteer (Caucasian)
N2	M	Blood	Cancer-free male volunteer (East Indian)
N3	M	Blood	Cancer-free male volunteer (Chinese)
N4	F	Blood	Cancer-free female volunteer (Mixed race)
N5	F	Blood	Cancer-free female volunteer (Caucasian)
N6	F	Blood	Cancer-free female volunteer (Caucasian)
N1-EBVt	M	Blood	H1 EBV-transformed cells
N4-EBVt	F	Blood	H5 EBV-transformed cells
BC(1-5)T	F	Breast	Basal-type breast cancer patient tissue
BC(1-5)G	F	Blood	Matching breast cancer patient blood
BC(6-10)T	F	Breast	Luminal-type breast cancer patient tissue
BC(6-10)G	F	Blood	Matching breast cancer patient blood
HT	-	Liver	Childhood hepatoblastoma tumor tissue (non-syndromic): childhood liver cancer at very young age of onset suggestive of genetic predisposition
HG	-	Blood	Matching childhood hepatoblastoma patient blood
CC1T	-	Colon	Colon cancer patient tissue
CC1G	-	Blood	Matching blood sample
CC2T	-	Colon	Colonic adenocarcinoma w/ signet ring features, Grade III, Stage T4N2M1
CC2G	-	Small intestine	Benign perilesional tissue
CC3T	-	Colon	Invasive adenocarcinoma, Grade II, Stage T3N1M1
CC3G	-	Liver	Benign liver (exploratory laparotomy) – cancer later metastasized to liver, patient deceased
Established Cancer and B Lymphocyte Cell Lines			
RKO	-	Colorectal	Poorly differentiated colorectal carcinoma cell line
HCT15	M	Colorectal	Duke’s Type C colorectal adenocarcinoma
HCT116	M	Colorectal	Colorectal carcinoma
HCC1187	F	Breast	TNM Stage IIA, grade 3 primary ductal carcinoma
HCC1187BL	F	Blood	Matched blood cell line
HCC1395	F	Breast	TNM Stage I, grade 3 primary ductal carcinoma
HCC1395BL	F	Blood	Matched blood cell line
HCC2157	F	Breast	TNM Stage IIIA, grade 2 primary ductal carcinoma
HCC2157BL	F	Blood	Matched blood cell line
H1437	M	Lung	Stage 1 adenocarcinoma, non-small cell lung cancer; patient was smoker (70 pack years)
BL1437	M	Blood	Matched blood cell line
H2141	M	Lung	Stage E carcinoma, small cell lung cancer; patient was smoker (50 pack years)
BL2141	M	Blood	Matched blood cell line
H2887	M	Lung	-
BL2887	M	Blood	Matched blood cell line

Cancer (tumor and germline) has a unique Microsatellite signature defined by 9 core motifs

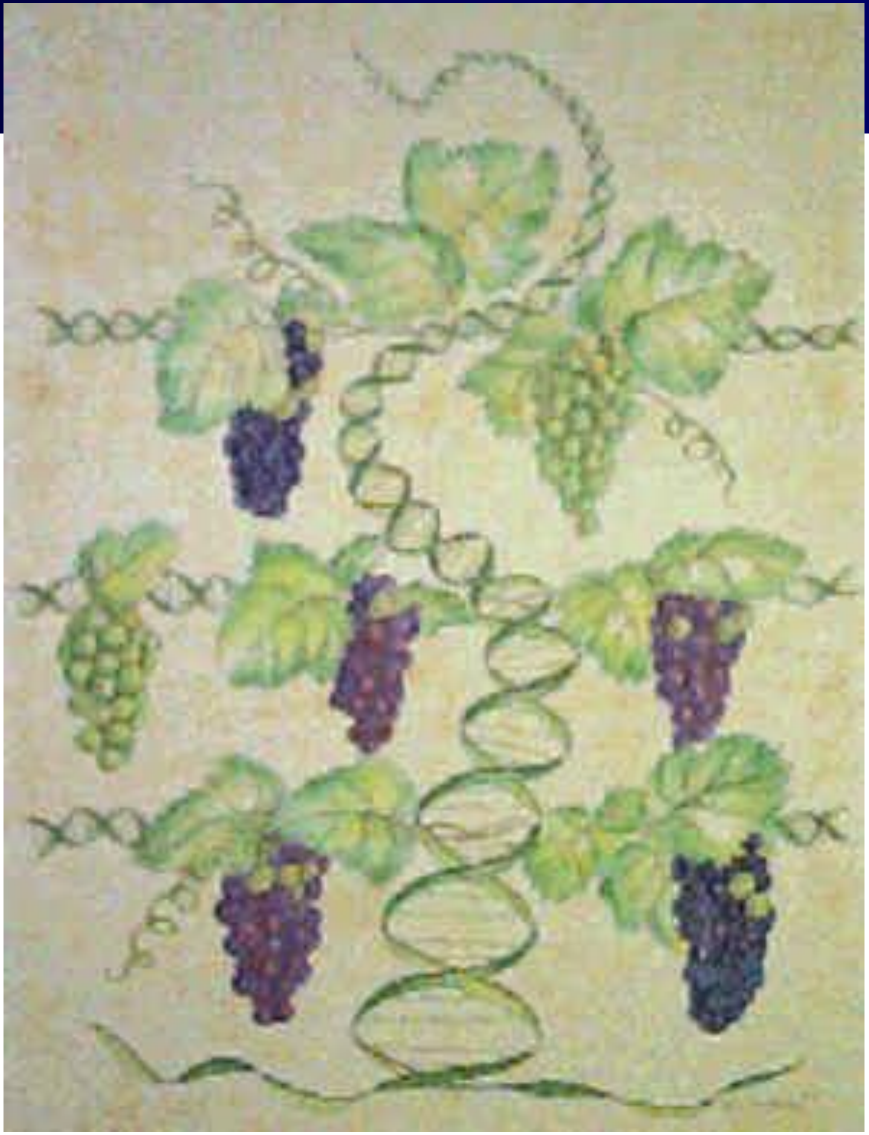


Genes,
Chromosomes
and Cancer

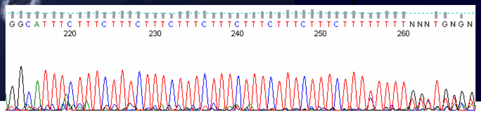
All BRCA1/2+ patients (germlines)
All Familial BC (germlines)
All BC cell lines (except triple negative)
All LC cell lines
10 Cancer-free volunteers
15 Other (4 diversity, 8 neurological, 3 UTAH)

10 BC patients (tumors and germlines)
All hepatoblastoma patients (tumors and germlines)
1 BC cell line (the only triple negative)
All 3 CC tumor cell lines
2 cancer-free volunteers
10 Other (2 diversity, 2 neurological, 6 UTAH)

The loci containing differential microsatellite motifs identified using the array can then using bioinformatics and the genome sequence be localized into candidate ‘cancer genes’ and pursued to identify mechanism and ultimately translate

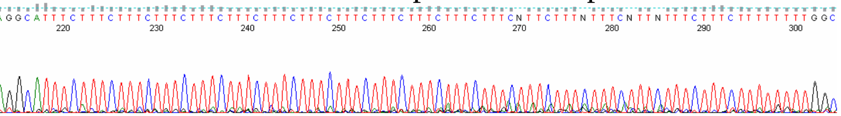


Cancer-free H1: 9 repeats

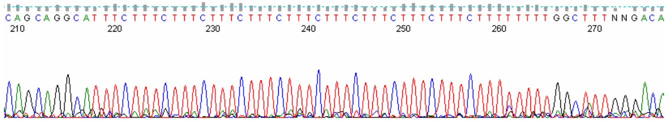


Homozygous short

Breast Cancer 2 top band: 19 repeats



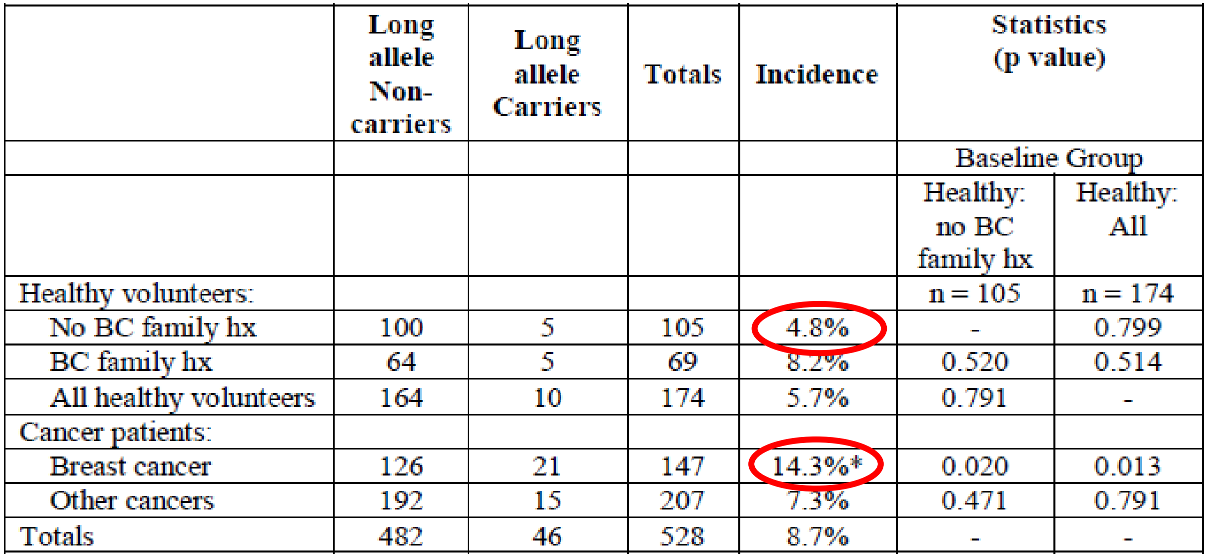
Breast Cancer 2 bottom band: 10 repeats



Hetero-zygous

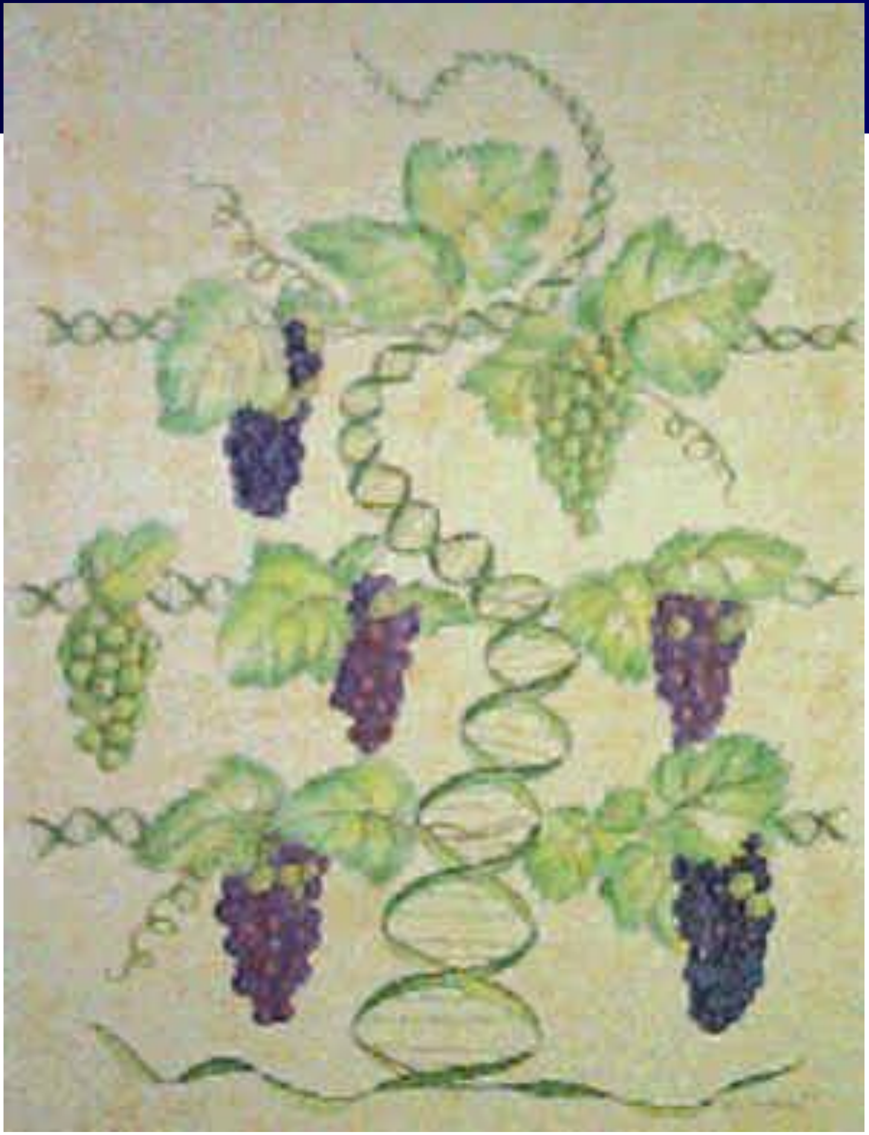
The ESRRG story: Array analysis led us to find an AAAG repeat locus in the 5’UTR of the Estrogen Related Receptor Gamma (ERR-g) gene, a gene which is differentially expressed in breast cancer tumors and has been linked to Tamoxifen resistance.

We validated that biomarker as a breast cancer risk marker



	Long allele Non-carriers	Long allele Carriers	Totals	Incidence	Statistics (p value)	
					Baseline Group	
					Healthy: no BC family hx	Healthy: All
Healthy volunteers:					n = 105	n = 174
No BC family hx	100	5	105	4.8%	-	0.799
BC family hx	64	5	69	8.2%	0.520	0.514
All healthy volunteers	164	10	174	5.7%	0.791	-
Cancer patients:						
Breast cancer	126	21	147	14.3%*	0.020	0.013
Other cancers	192	15	207	7.3%	0.471	0.791
Totals	482	46	528	8.7%	-	-

The computational analysis of the 1000 Genome Project and The Cancer Genome Atlas project data is telling us why others are not making progress correlating disease to microsatellite genotypes.



The 1000 Genomes and TCGA Projects data is illustrative of where genomics is going

- The 1000 Genomes Project, launched in January 2008, is an international research effort to establish a detailed catalogue of human genetic variation by sequencing 2,400 individuals.
- The Cancer Genome Atlas will sequence germline and tumor genomes of at least 200 forms of cancer.
- Identifying the changes in each cancer’s complete set of DNA – its genome – and understanding how such changes interact to drive the disease will lay the foundation for improving cancer prevention, early detection and treatment.

Glioblastoma Multiforme (Brain Cancer) case study

- 19 exome sequenced GBM patients were compared to 250 disease-free genomes from the 1000 Genomes Project. 80 more GBM genomes are in processing now.
- GBM affects 1-4 in 100,000 individuals
- 9 loci were found to be informative following a leave-one-out statistical comparison of affected individuals to disease-free individuals
- 32% of the GBM patient germlines had variations in at least two loci, but none of the disease-free patients.

Glioblastoma Multiforme analysis has produced informative biomarkers

								1kGP 250 normals			GM BL		
motif	ref	region	gene symbol	gene title	intron size	intron distance	TFBS	samples	consensus	alleles	samples	consensus	alleles
A	13	intergenic	-	-	-	-	-	183	14	13 (2), 14 (364)	18	14	13 (4), 14 (32)
T	15	intron	TNPO1	transportin 1	1,480	19	-	56	15	15 (112)	17	15	14 (3), 15 (31)
A	13	intron	XPO5	exportin 5	853	38	-	116	13	12 (4), 13 (228)	17	13	13 (31), 12 (3)
T	12	intron	RGS6	regulator of G-protein signaling 6	26,139	3,746	-	91	12	12 (182)	14	12	11 (6), 12 (22)
A	13	intron	DOCK4	dedicator of cytokinesis 4	10,074	7	OCT1_03, HOXA3_01,RFX1_01	84	13	13 (165), 12 (2), 14 (1)	16	13	13 (29), 12 (3)
AC	18	intron	CDC25A	cell division cycle 25 homolog A (S. pombe)	409	21	-	54	16	16 (108)	15	16	18 (3), 16 (27)
T	12	intron	RAD51AP2	RAD51 associated protein 2	1,074	254	-	182	12	11 (2), 12 (361), 13 (1)	13	12	11 (4), 12 (22)
A	12	intron	MLKL	mixed lineage kinase domain-like	16,935	6,999	-	110	12	12 (220)	12	12	11 (2), 13 (2), 12 (20)
A	13	intron	ALS2	amyotrophic lateral sclerosis 2 (juvenile)	4,090	23	-	89	13	12 (1), 13 (177)	18	13	13 (33), 12 (3)
AAAC	22	intron	COL24A1	collagen, type XXIV, alpha 1	5,269	617	-	95	22	22 (190)	4	9	9 (8)
A	14	intergenic	-	-	-	-	BRN2_01	62	14	14 (124)	17	14	13 (3), 14 (31)
A	13	intron	ULK4	unc-51-like kinase 4 (C. elegans)	47,900	10	-	60	13	16 (2), 13 (118)	12	13	16 (2), 13 (20), 15 (2)
T	15	intron	PTP4A1	protein tyrosine phosphatase type IVA, member 1	723	21	MEF2_03	29	15	14 (1), 15 (57)	15	15	14 (3), 15 (27)
AAAC	19	intron	PTPRN2	protein tyrosine phosphatase, receptor type, N polypeptide 2	12,673	83	-	206	19	19 (412)	18	19	19 (33), 20 (3)
T	12	intron	TCF7L2	transcription factor 7-like 2 (T-cell specific, HMG-box)	1,200	170	-	184	12	11 (1), 13 (4), 12 (363)	18	12	11 (4), 12 (32)
AAATA	15	intergenic	-	-	-	-	-	214	15	15 (428)	2	14	14 (4)
A	14	intron	CAPN6	calpain 6	260	14	COMP1_01	83	14	14 (166)	13	14	13 (3), 14 (23)
AT	23	intron	MANBA	mannosidase, beta A, lysosomal	23,665	23	-	140	23	21 (1), 23 (279)	5	23	23 (6), 17 (4)

Autism data is becoming available to study

- De novo mutations revealed by whole-exome sequencing are strongly associated with autism -- whole-exome sequencing of 928 individuals, including 200 pheno- typically discordant sibling pairs -- still can't find data in the SRA
- Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations -- total of 677 individual exomes from 209 families -- couldn't find data in dbGaP
- Patterns and rates of exonic de novo mutations in autism spectrum disorders -- exome sequencing of 175 ASD probands and their parents -- data is available in dbGaP
- A total of 2,130 samples are becoming available for us to repeat the analysis that we did for Cancer in Autism to find new biomarkers that could be informative for risk of developing autism, intensity, response to therapy, etc.

Thank you for your interest. This work represents the efforts of many in the lab and collaborators

http://innovation.vbi.vt.edu

- The work of many in the lab
  - John McCormick
  - Johnny Sun
  - Lauren McIver
  - Jasmin Bavarva
  - Hongseok Tae
  - Natalie Fonville
  - Wyatt McMahon
  - Shamira Shallom
  - Iccha Sethi
  - Mara Gibson
  - Heather Lewenczuk
  - Hadas Fletcher
- Important and sustained collaborations with
  - John Minna, Cheryl Lewis, David Boothman, Gail Tomlinson, Cristi L. Galindo
  - Chuck Perou, UNC
- Supported by Virginia Tech, the NIH, the NSF, the P. O’B. Montgomery Endowed Chair and the Hudson Foundation
- Thanks to Nimblegen/Roche for array production, and VBI core for sequencing and Dell and Convey Computing for the supercomputer horsepower