SESSION 14. PERSONAL GENOMES

The differences between you and me

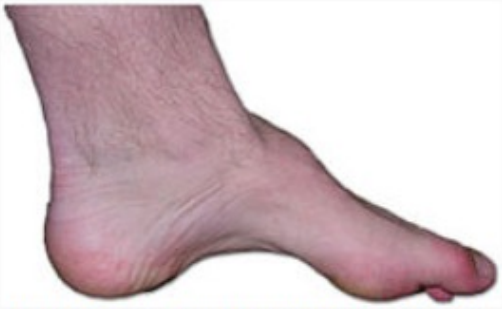
Personal genomes

- Human genome sequence ($3 billions grant to HGP /$300 million to Celera) – 1990~2003
- $600 / personal genome (30X), 1 week → $200/personal genome (30X)
- BGI produces tens terabytes of DNA per day
- **Personal genome era:**
  - 1000 genome project, 100,000 genome UK
  - Korean BioBigData project: 300,000 genomes + multiomics
  - USA AllOfUs project: 1M genomes
  - 1 million genomes for precision medicine (China)
- TCGA/ICGA cancer genomes (thousands of cancer genomes)
- Thousands of Korean genomes are sequenced

A selection of first personal genomes

- First version of human genome (mixture of anonymous individuals) 2003
- Craig venter 2007
- James Watson 2008
- AML patient (normal and cancer) 2008
- Yoruba, Ibadan, Nigeria (anonymous) 2008
- YanHuang (Han Chinese) 2009
- Stephen Quake (Standford) 2009
- Seong-Jin Kim 2009
- James Lupski 2010(CMT disease)

Charcot–Marie–Tooth disease



The foot of a person with Charcot–Marie–Tooth disease. The lack of muscle, a high arch, and claw toes are signs of this genetic disease.

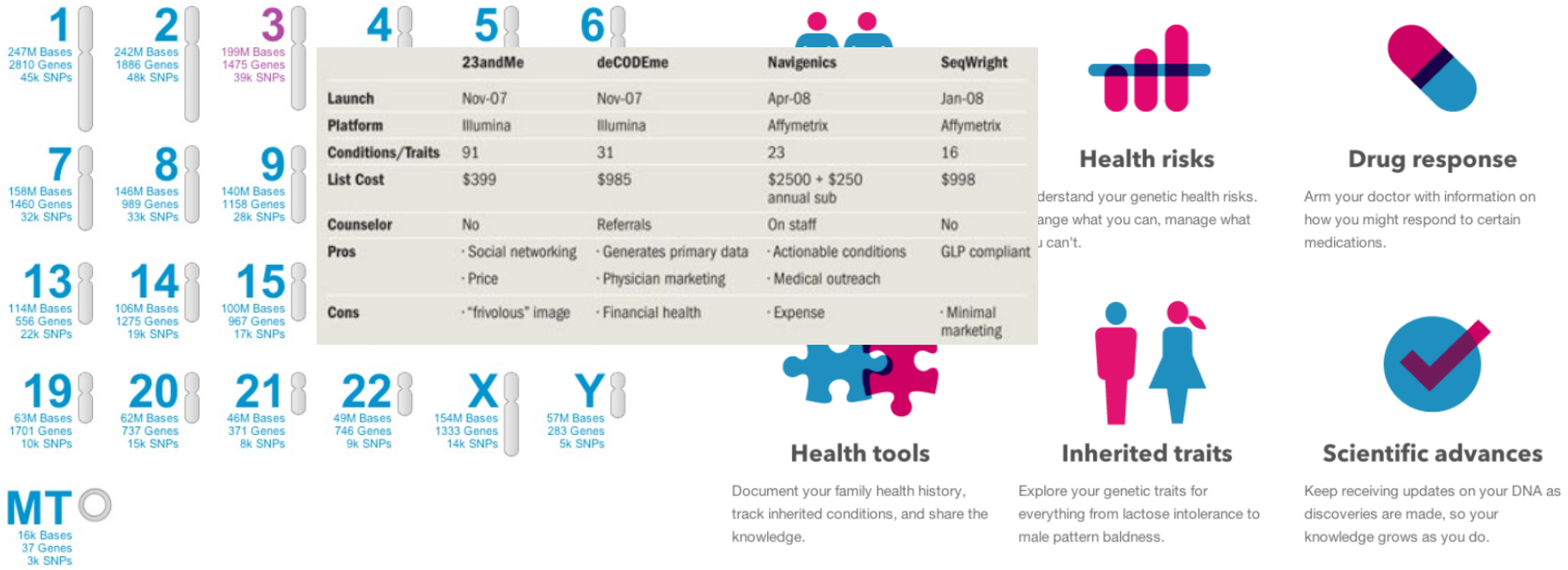
Individual variation and SNPs

- Single nucleotide polymorphism (SNP, i.e., C→T)
- Small insertions and deletions (i.e., G→GAC)
- Copy number variations (i.e., CAG repeats)
- Large structural variations

→ Medical and forensic applications

Individual variation and SNPs

- Some SNPs in noncoding could affect the expression of genes
- But, SNPs in coding of genes are particularly interesting.
  - Synonymous vs nonsynonymous
- SNPs are sometimes associated with diseases → i.e., blood coagulation



Individual variation and SNPs

- NCBI SNP database (2012) – 60 million SNPs
- Between two random individuals – 3 million SNPs

- Two copy of genomes
  - Heterozygote: G-C A-T → Two different alleles
  - Homozygote: G-C G-C → One allele
- SNP positions → at least one allele is different

- Most common alleles ? Mutations vs SNPs
- Human reference genome (2001) – does not mean that it includes the most common alleles among humans at SNP positions.

Counting SNPs

- Using the table browser at the UCSC genome database (http://genome.ucsc.edu/cig-bin/hgTables?command=start)
- Comparing chr4 of eight different human individuals
  - YanHuang (Han Chinese)
  - Seong-Jin Kim
  - Jame Watson
  - Craig Ventor
  - YRI (Yaruba, one of 1000 genomes project)
  - NA12891 (Central European origin, one of 1000 genomes)
  - ABT, Demond Tutu
  - KB1, Bushman individual

Counting SNPs

- SNP.txt (input data)
- 38357 T T Y T T T Y Y T
- The first column is position
- Nucleotides from eight individuals
- The last column is the base of chimpanzee
- Exclude positions where at least one genome has an unknown base
- Exclude positions containing the same nucleotide in all nine genomes have been removed

IUPAC code

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

snp.txt

```
3263 A A A A A A A A T
3351 T W W W T T W W A
3544 T T T T T T T Y T
3567 T T T T T T T Y T
3774 K G T T T T T T T
4131 G K G G G G G G T
4190 A A A R A A A R A
4306 T T T T T T T T C
4371 C Y Y Y C C C C C
4489 G R A A A A A A A
6394 T T T T T T T T C
6523 G R A R A A R A A
7764 C C Y Y C C C C C
7836 T T K T T T T T T
8171 C C C M C C C C C
8294 A W A A A A A A A
8395 T T Y T Y T Y C C
8584 G G R G G G G G G
8648 R A A A A R R A G
8675 R R A A A A A A A
8751 G G S G G G G G G
11280 G G G G G G G G T
11284 A A A A A A A A T
13060 A A A A A R A A A
13098 C C C C C C C Y C
15231 C C C C C C C C G
```

snp.py

- Create a distance matrix from SNPs of 9 genomes

```
#!/usr/bin/python
# obtain pairwise distances from snp data,
# counting sites where at least one allele is different
import re
humans = [
    # SNPs appear in the SNP data file in columns in this order
    'YH',          # Han chinese
    'SJK',         # Seong-Jin Kim
    'JW',          # James Watson
    'CV',          # Craig Venter
    'NA18507',     # Yoruban of 1000 Genomes project
    'NA12891',     # Of Central European origin
    'ABT',         # Archbishop Desmond Tutu
    'KB1',         # Bushmen individual
    'chimp'        # chimpanzee
]
# 1 #
# initialize the distance matrix with zero values
# for the diagonal cells
diff = [[] * 10]
for j in range(0, 10):
    diff[0].append(0)

for i in range(1, 10):
    diff.append([])
    for j in range(0, 10):
        diff[i].append(0)
```

```
# read the snp data from file
for line in open('snp.txt'):
    line = line.rstrip()
    columns = re.split(' ', line)
    # 2 #
    for i in range(1, 9):
        for j in range(i + 1, 10):
            # 3 #
            if columns[i] != columns[j]:
                diff[i][j] += 1
                # 4 #
                # to produce a symmetric matrix
                diff[j][i] += 1
# 5 #
# print a header for PHYLIP format
# with the number of species
print '    ', '9'
# print the matrix data
for i in range(1, 10):
    # 6 #
    txt = humans[i - 1]
    txt = txt[0:7]
    print txt,
    length = 10 - len(txt)
    short = ' ' * (length - 2)
    print short,
    for j in range(1, 10):
        print diff[i][j],
    print ''
```

snp.py

- Create a distance matrix from SNPs of 9 genomes
  - #1 : diff is 2D-array to store the counts of pairwise distances and initialized with zeros
  - #2 : use two for loops to go through all pairs of 9 genomes

```
# 2 #
   for i in range(1, 9):
       for j in range(i + 1, 10):
```

  - #3 : Test whether two genome at a specific position is equal or not and, if it's not same, then count +1
  - #4 : Making a symmetric matrix
  - #5 : print out a distance matrix
  - #6: truncate the name to seven characters

snp.py

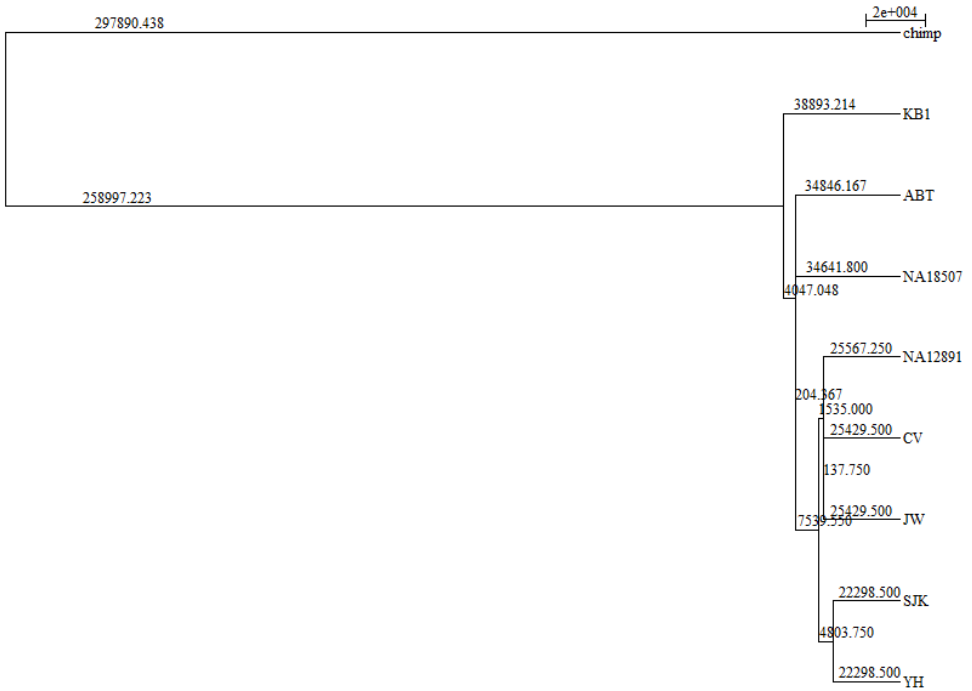
```
[jwnam@biglab-master Session13]$ python snp.py
    9
YH          0 44597 53594 53913 67914 53710 68837 77272 593367
SJK         44597 0 54192 54537 68826 55281 69404 76929 593496
JW          53594 54192 0 50859 70284 51260 70256 77590 592751
CV          53913 54537 50859 0 70149 51009 69659 77369 592632
NA18507     67914 68826 70284 70149 0 69245 70057 79508 599102
NA12891     53710 55281 51260 51009 69245 0 69941 78130 594831
ABT         68837 69404 70256 69659 70057 69941 0 77707 599292
KB1         77272 76929 77590 77369 79508 78130 77707 0 600776
chimp       593367 593496 592751 592632 599102 594831 599292 600776 0
```

Phylip package - neighbor

neighbor

```
C:\Users\Jin-Wu\Downloads\phylip-3.695\phylip-3.695\exe\neighbor.exe

Neighbor-Joining/UPGMA method version 3.695

Settings for this run:
  N       Neighbor-joining or UPGMA tree?  UPGMA
  L         Lower-triangular data matrix?  No
  R         Upper-triangular data matrix?  No
  S                       Subreplicates?  No
  J     Randomize input order of species?  No. Use input order
  M           Analyze multiple data sets?  No
  0   Terminal type (IBM PC, ANSI, none)?  IBM PC
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                       Print out tree  Yes
  4       Write out trees onto tree file?  Yes


  Y to accept these or type the letter for one to change
Y

Cycle   8: species 1 (22298.50000) joins species 2 (22298.50000)
Cycle   7: species 3 (25429.50000) joins species 4 (25429.50000)
Cycle   6: node 3 ( 137.75000) joins species 6 (25567.25000)
Cycle   5: node 1 (4803.75000) joins node 3 (1535.00000)
```

NJPlot




```
(((((YH:22298.50000,SJK:22298.50000):4803.75000,((JW:25429.50000,
CV:25429.50000):137.75000,NA12891:25567.25000):1535.00000):7539.55000,
NA18507:34641.80000):204.36667,ABT:34846.16667):4047.04762,
KB1:38893.21429):258997.22321,chimp:297890.43750);
```