Trait-o-matic : How To (Part I)

Alexander Wait Zaranek
awaitz@post.harvard.edu

Biophysics 101 seminar
Thursday, October 15th, 2009

What would you do with twenty-five individual human genomes?

Trait-o-matic

http://snp.med.harvard.edu

Analysis of individual genomes

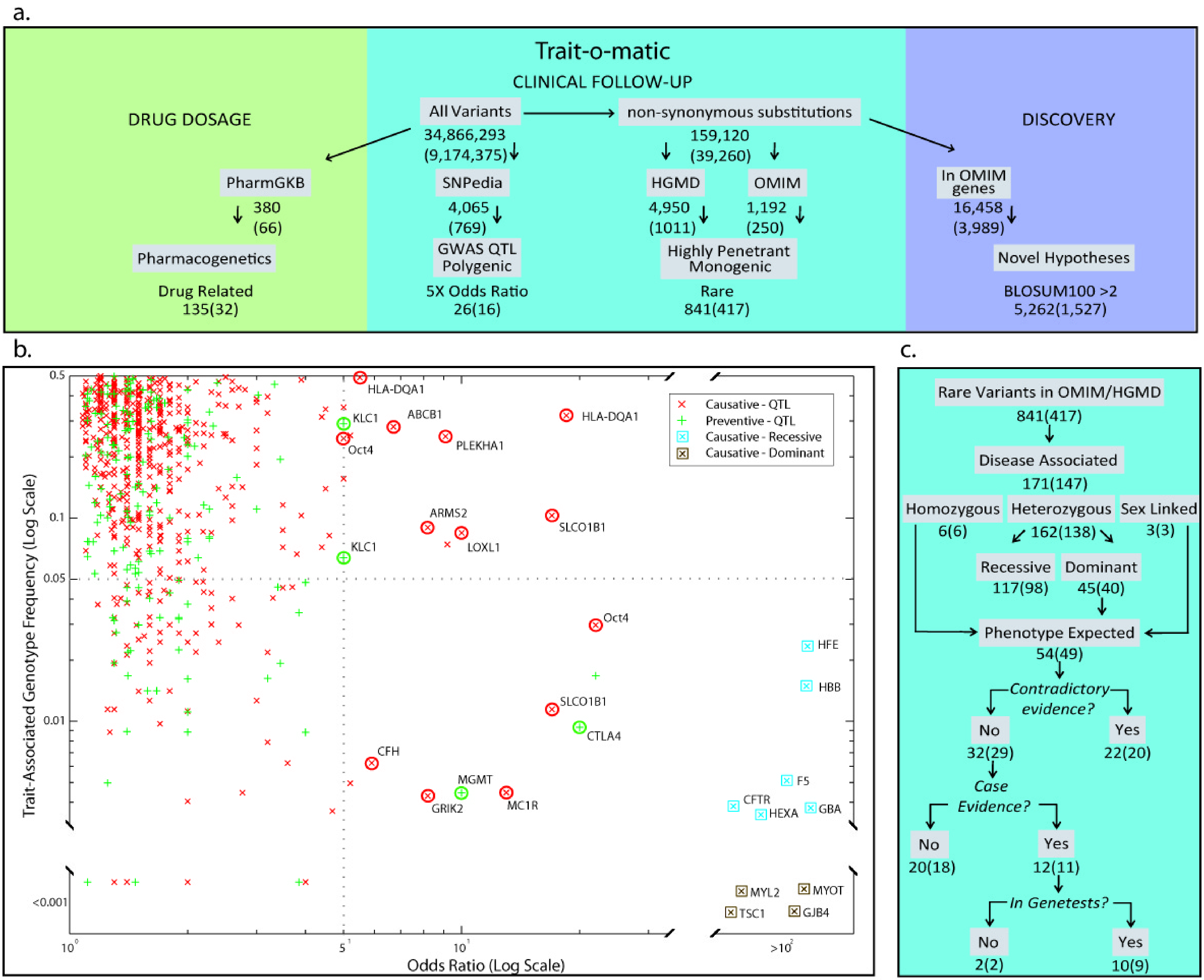
Use GeneTests to focus on genes where clinical action is already taken

Convert variants in HG18 coordinates into gene/ protein coordinates

Cross-reference with OMIM/HGMD/SNPedia/PharmGKB to obtain a list of known variants with pointers into the literature

Obtain allele frequencies when available (typically not available for rare variants)

Rosenbaum, Thakuria, Wu, Zaranek et al


a.
Trait-o-matic
CLINICAL FOLLOW-UP
DRUG DOSAGE
DISCOVERY
All Variants
34,866,293
(9,174,375)
non-synonymous substitutions
159,120
(39,260)
PharmGKB
380
(66)
Pharmacogenetics
Drug Related
135(32)
SNPedia
4,065
(769)
GWAS QTL
Polygenic
5X Odds Ratio
26(16)
HGMD
4,950
(1011)
OMIM
1,192
(250)
Highly Penetrant
Monogenic
Rare
841(417)
In OMIM genes
16,458
(3,989)
Novel Hypotheses
BLOSUM100 >2
5,262(1,527)
b.
Trait-Associated Genotype Frequency (Log Scale)
Odds Ratio (Log Scale)
0.5
0.1
0.05
0.01
<0.001
10⁰
5¹
10¹
>10²
Causative - QTL
Preventive - QTL
Causative - Recessive
Causative - Dominant
HLA-DQA1
KLC1
ABCB1
Oct4
PLEKHA1
HLA-DQA1
ARMS2
SLCO1B1
LOXL1
KLC1
Oct4
HFE
HBB
SLCO1B1
CTLA4
CFH
MGMT
GRIK2
MC1R
F5
CFTR
HEXA
GBA
MYL2
MYOT
TSC1
GJB4
c.
Rare Variants in OMIM/HGMD
841(417)
Disease Associated
171(147)
Homozygous
6(6)
Heterozygous
162(138)
Sex Linked
3(3)
Recessive
117(98)
Dominant
45(40)
Phenotype Expected
54(49)
Contradictory evidence?
No
32(29)
Yes
22(20)
Case Evidence?
No
20(18)
Yes
12(11)
In Genetests?
No
2(2)
Yes
10(9)

PersonalGenomes.org

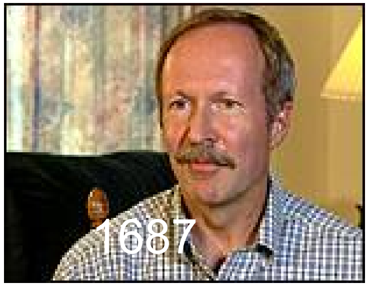
Subject & public access (not just research elite)

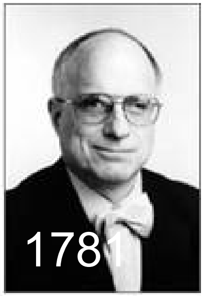
Entrance exam to ensure highly informed consent

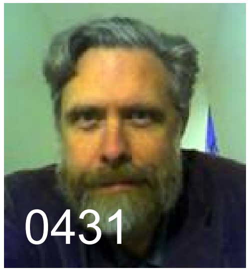
Scalable to millions of research subjects, budget $1,000/person for DNA & trait data

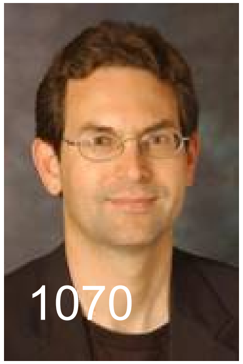
Highly integrated, holistic, systems-biology

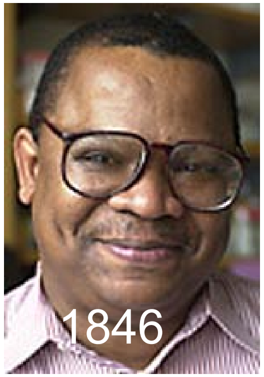
Cells available for personal functional genomics











What would you do with a hundred thousand individual human genomes?

To get an answer – ask a different question!

How do we organize computational resources to serve the combined needs of scientists, physicians and the general public?

Many commercial organizations aim to answer this type of question in other domains—Amazon Web Services is a leading provider

CYCLECOMPUTING

Appistry

amazon webservices™ Amazon Elastic Compute Cloud (Amazon EC2) - *Beta*

GOGRID beta
A ServePath Company

rPath®

VMOps

Sun microsystems TAP INTO THE POWER OF NETWORK.COM

3tera™

b-hive

MOSSO® the hosting cloud

Engine Yard™

citrix®

Microsoft®

vmware®

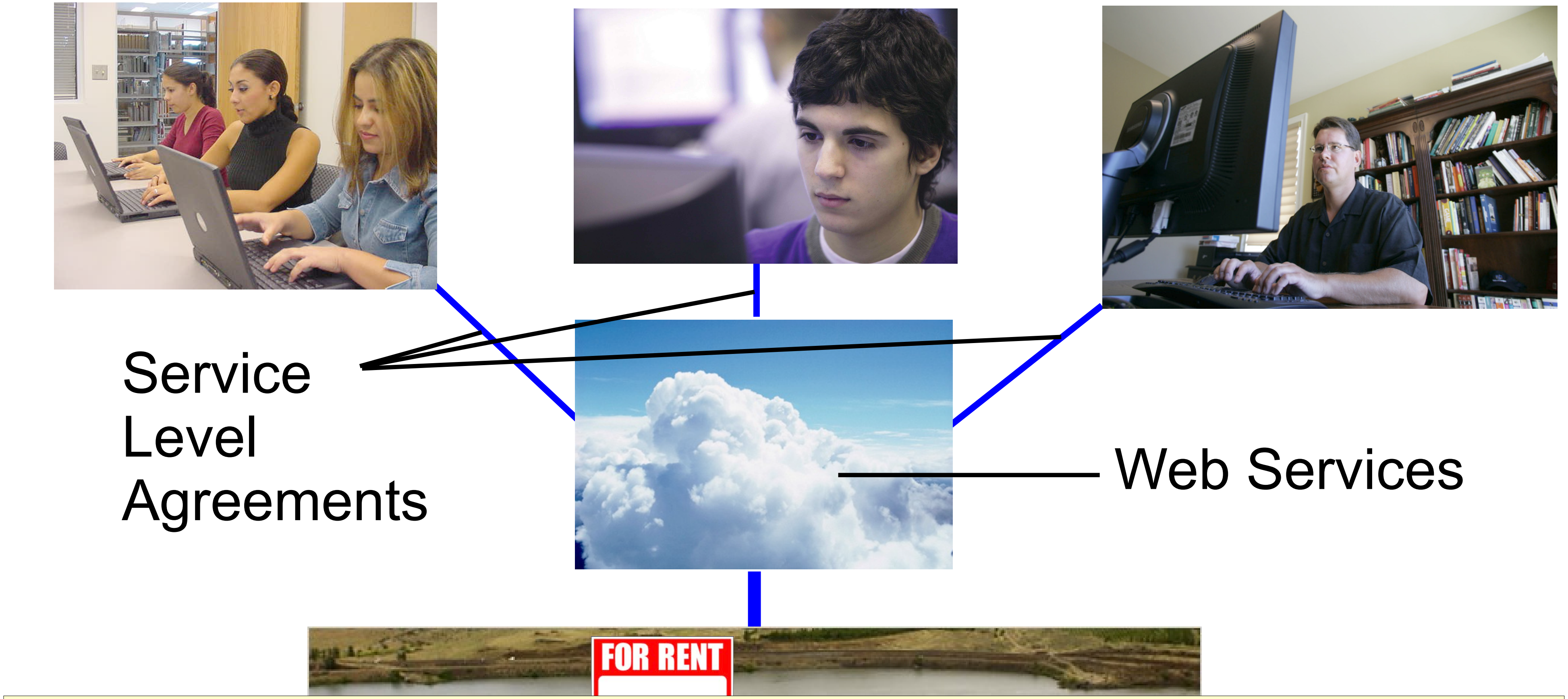
Q-layer

Joyent

VERIO
An NTT Communications Company

hp

How does a “cloud” work for 2nd generation sequencing?



Service Level Agreements

Web Services

Abstract away users (with a simple web browser) from massive, physical computational resources and highly parallel data acquisition instruments via standard internet protocols and Service Level Agreements

Modified from Eucalyptus, Cloud Computing Expo 2009

A Free Factory is inspired by Free Software and embodies a special case of the “cloud” paradigm

Free Software is a matter of the users' freedom to run, copy, distribute, study, change and improve the software.

(http://www.gnu.org/philosophy/free-sw.html)

A Free Factory should protect the freedom of its user community to:

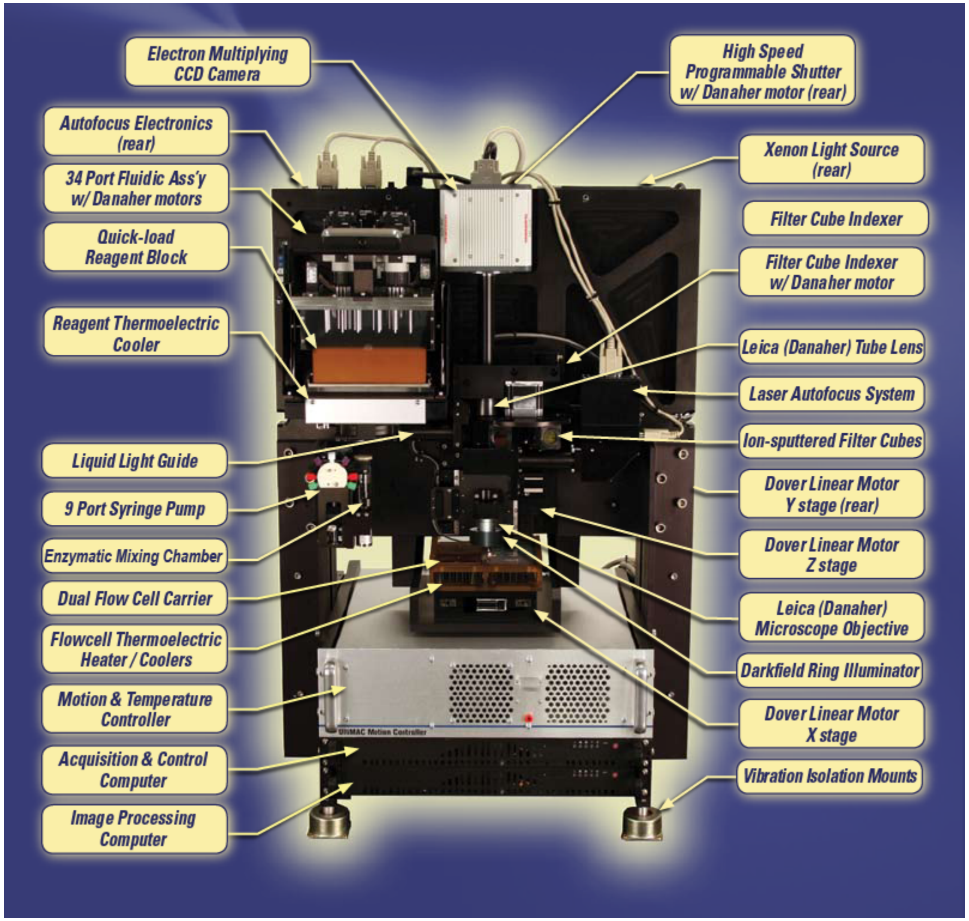
1) operate their own identical factory;

2) operate a modified factory;

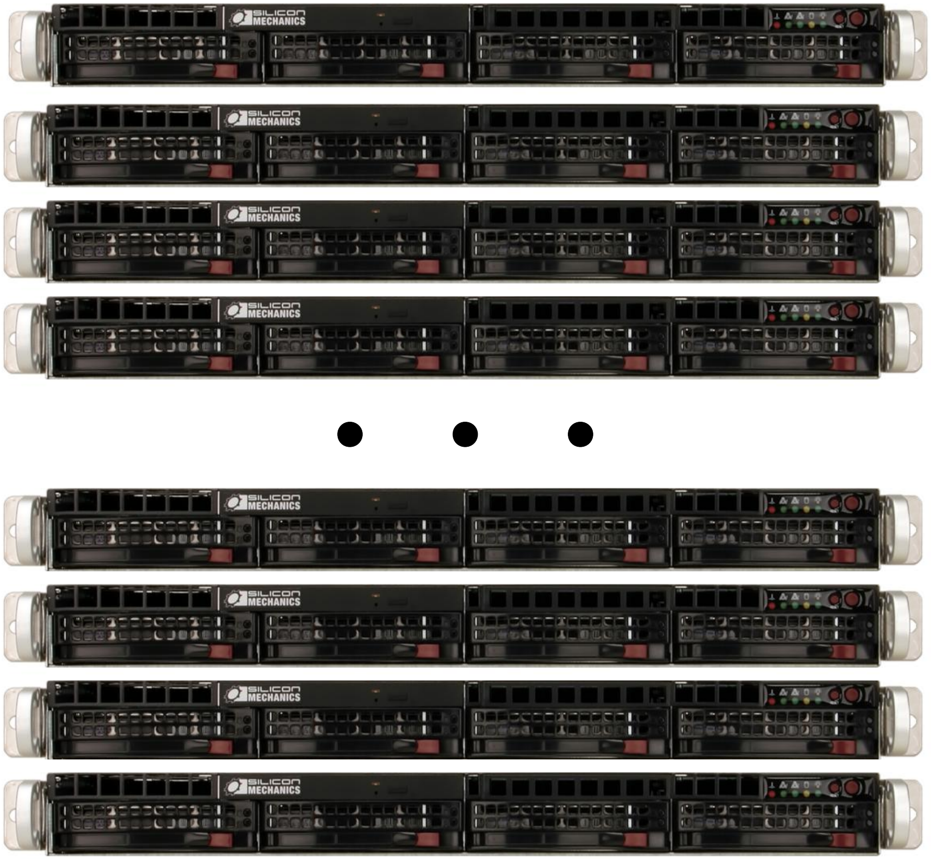
3) distribute the information required to operate and modify the factory to others, and;

4) study and improve all factory equipment, methods, software, raw materials, and so on.

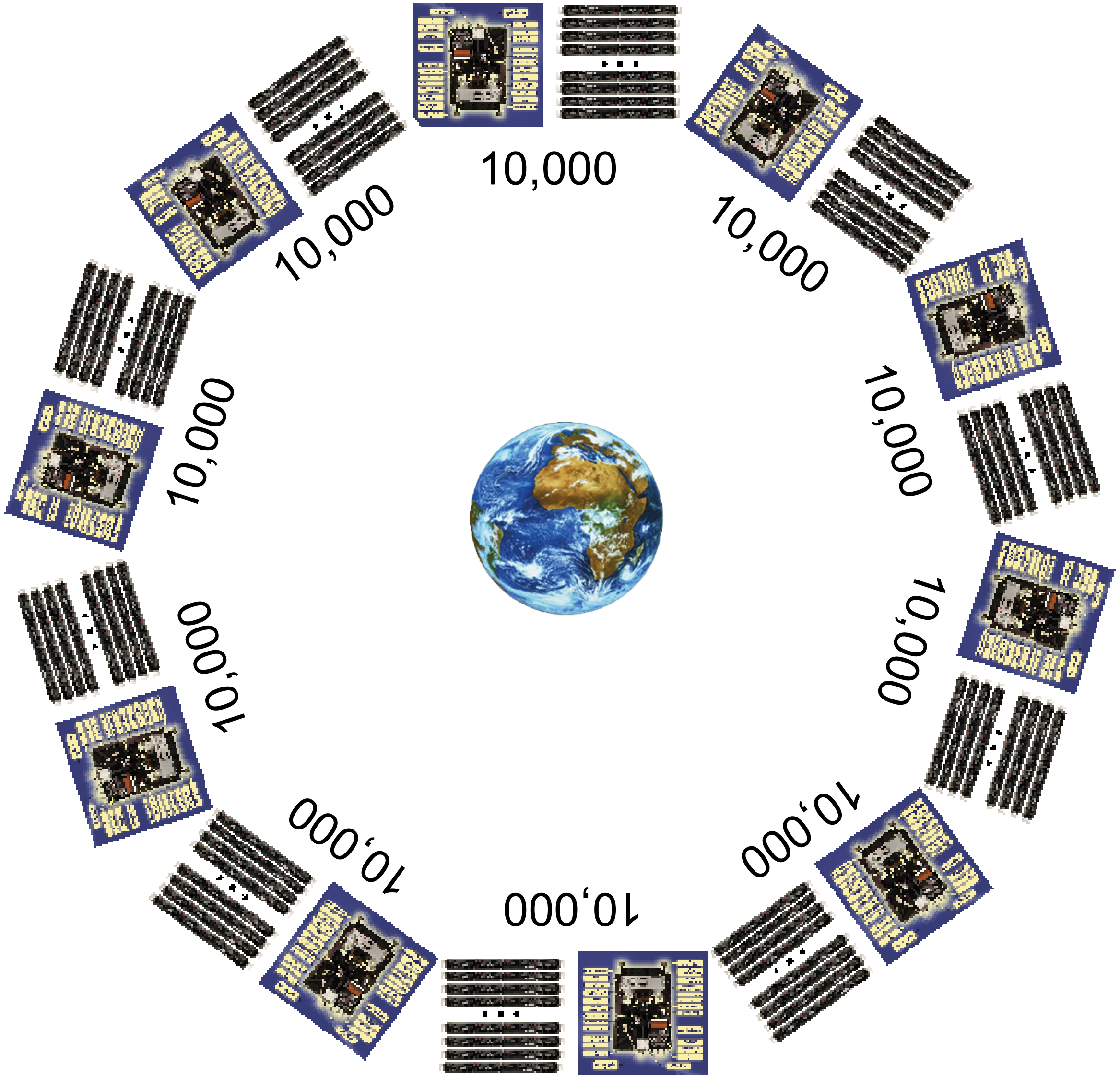
A Free DNA Sequencing Factory could be built by combining the “Polonator” with commodity computers running Free and Open Source Software


Electron Multiplying CCD Camera
High Speed Programmable Shutter w/ Danaher motor (rear)
Autofocus Electronics (rear)
Xenon Light Source (rear)
34 Port Fluidic Ass'y w/ Danaher motors
Filter Cube Indexer
Quick-load Reagent Block
Filter Cube Indexer w/ Danaher motor
Reagent Thermoelectric Cooler
Leica (Danaher) Tube Lens
Laser Autofocus System
Liquid Light Guide
Ion-sputtered Filter Cubes
9 Port Syringe Pump
Dover Linear Motor Y stage (rear)
Enzymatic Mixing Chamber
Dover Linear Motor Z stage
Dual Flow Cell Carrier
Leica (Danaher) Microscope Objective
Flowcell Thermoelectric Heater / Coolers
Darkfield Ring Illuminator
Motion & Temperature Controller
Dover Linear Motor X stage
Acquisition & Control Computer
Vibration Isolation Mounts
Image Processing Computer

Courtesy — Rich Terry and Greg Porreca



Scalable Infrastructure for 100,000 people



Maintain infrastructure close to participants

Add sequencing instruments, computational clusters, and storage independently

Freegols can use storage and compute resources from any Free Factory

Fault-tolerant to hardware and software failures

Built-in provenance tracking

The Idea

Public

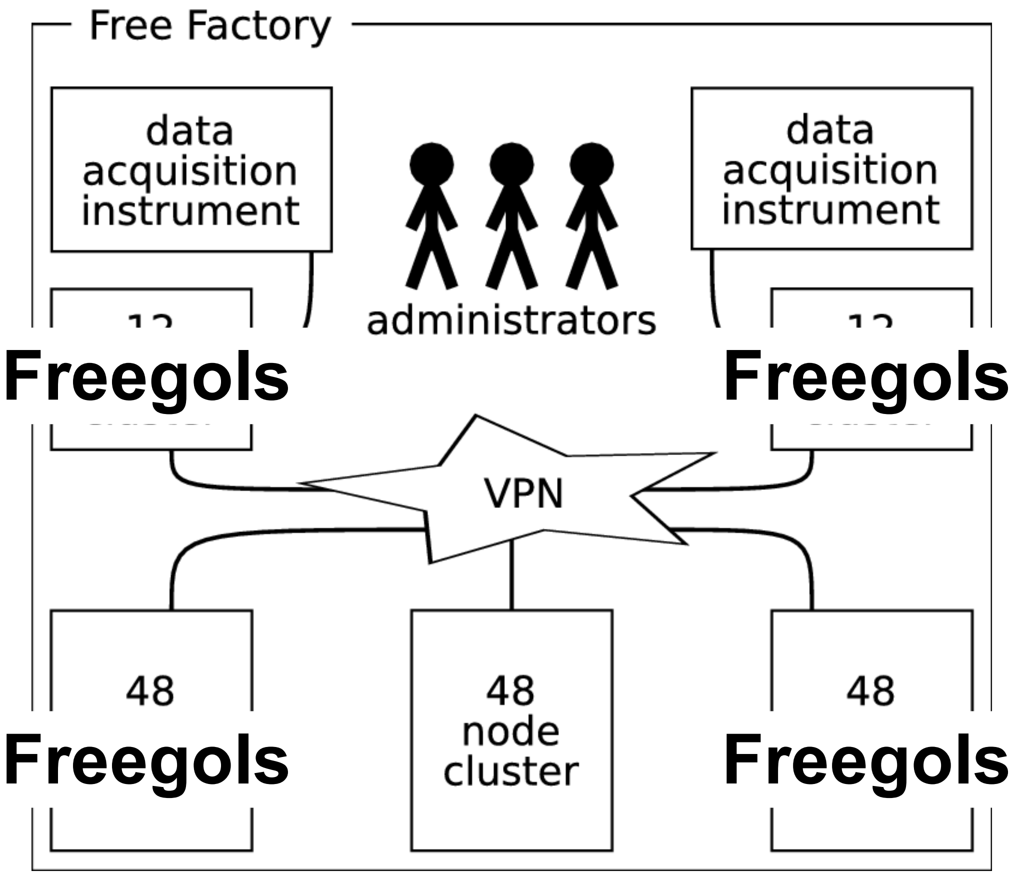
Scientists

Physicians

Public

Scientists

Physicians



Freegols

Freegols

Freegols

Freegols

Public

Scientists

Physicians

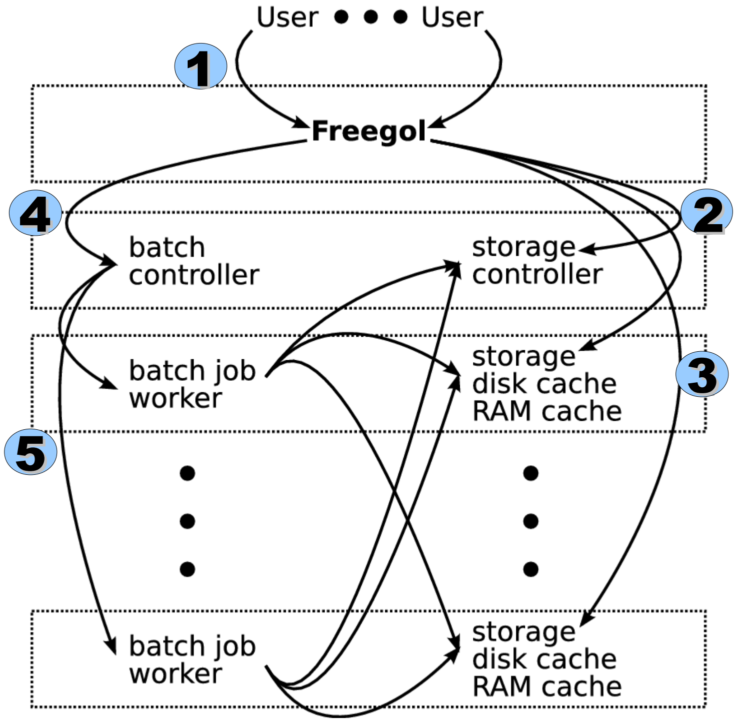
Public

Scientists

Physicians

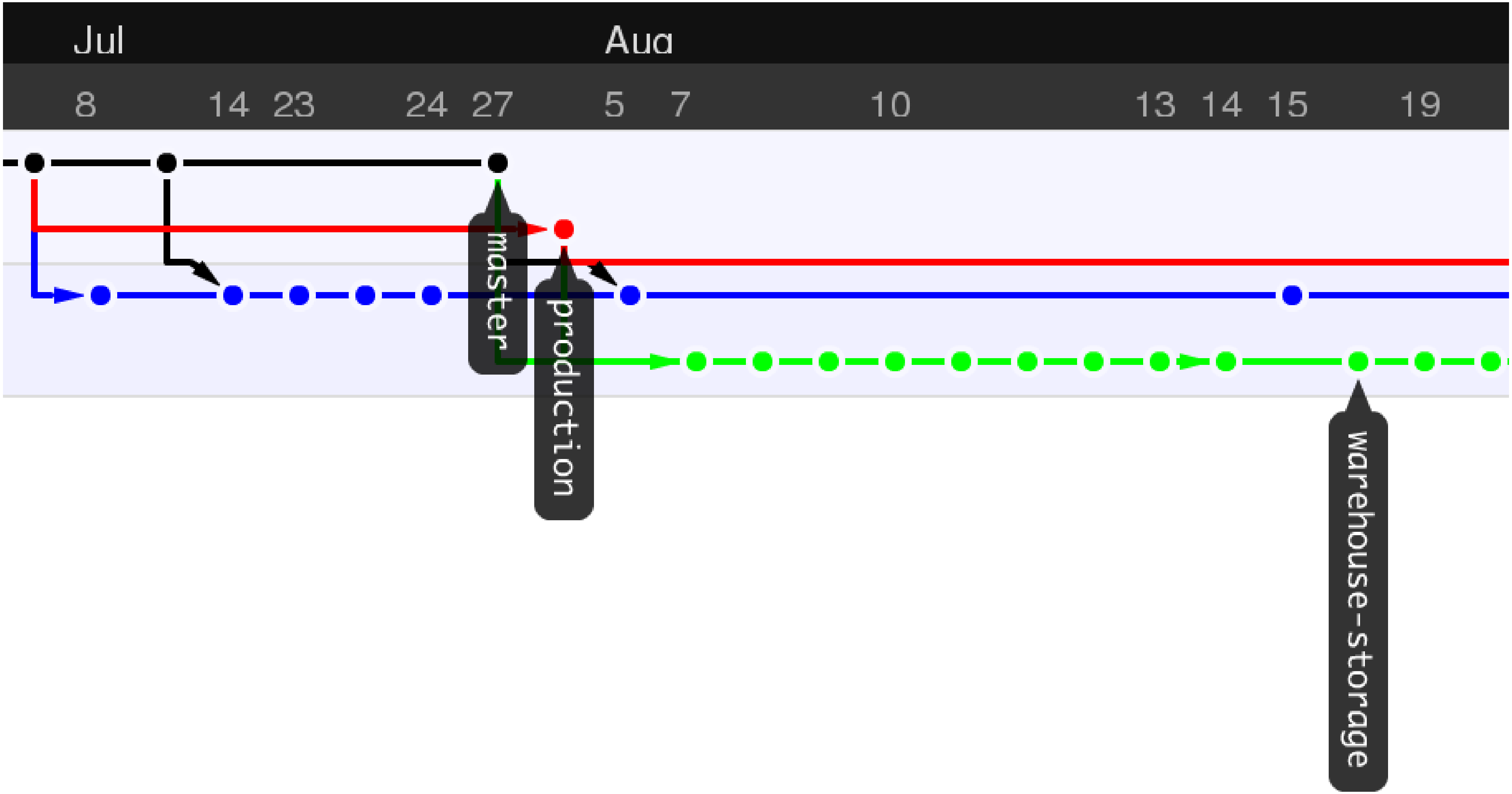
A shared infrastructure for web service virtual machines, which I call “Freegols”.

Freegols—or Free Golems (another word for robot)—operate in independent virtual machines running on the Free Factories infrastructure.



As a Freegol services many simultaneous user requests, it continually supervises "workflows" that process terabytes of data and consume many thousands of CPU hours

Trait-o-Matic is the archetypal "Freegol" and maintained using the distributed development paradigm

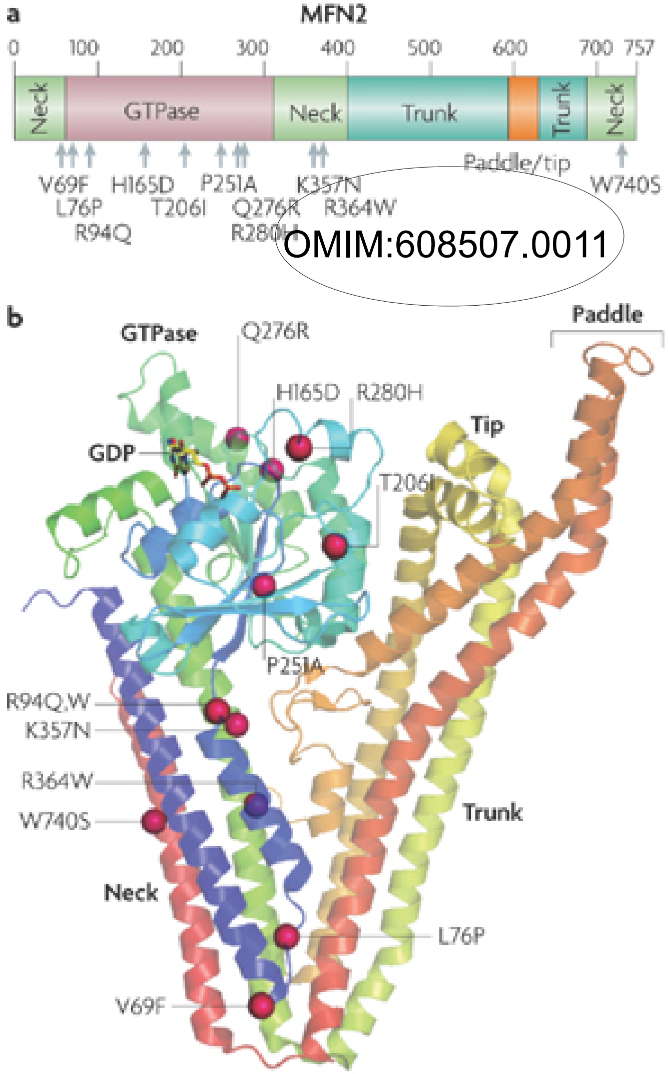


Class projects can use the lab "cloud" or Trait-o-matic as a platform for further development

Trait-o-matic cross-references variants with major databases and looks for damaging coding changes

PGP2 – MFN2 R364W – HEREDITARY MOTOR AND SENSORY NEUROPATHY VI

?

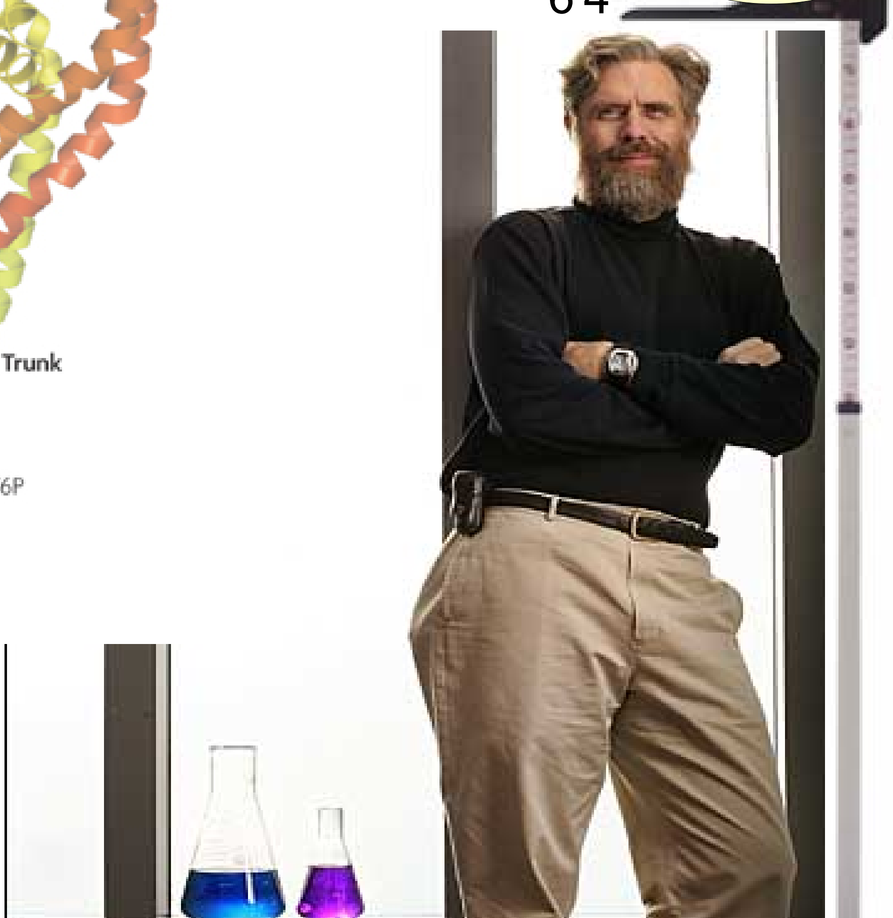


OMIM:608507.0011

PGP1 – GHR G186* – (associated with) INCREASED RESPONSIVENESS TO GROWTH HORMONE

?

6'4”



PGP1 HGR Mutation

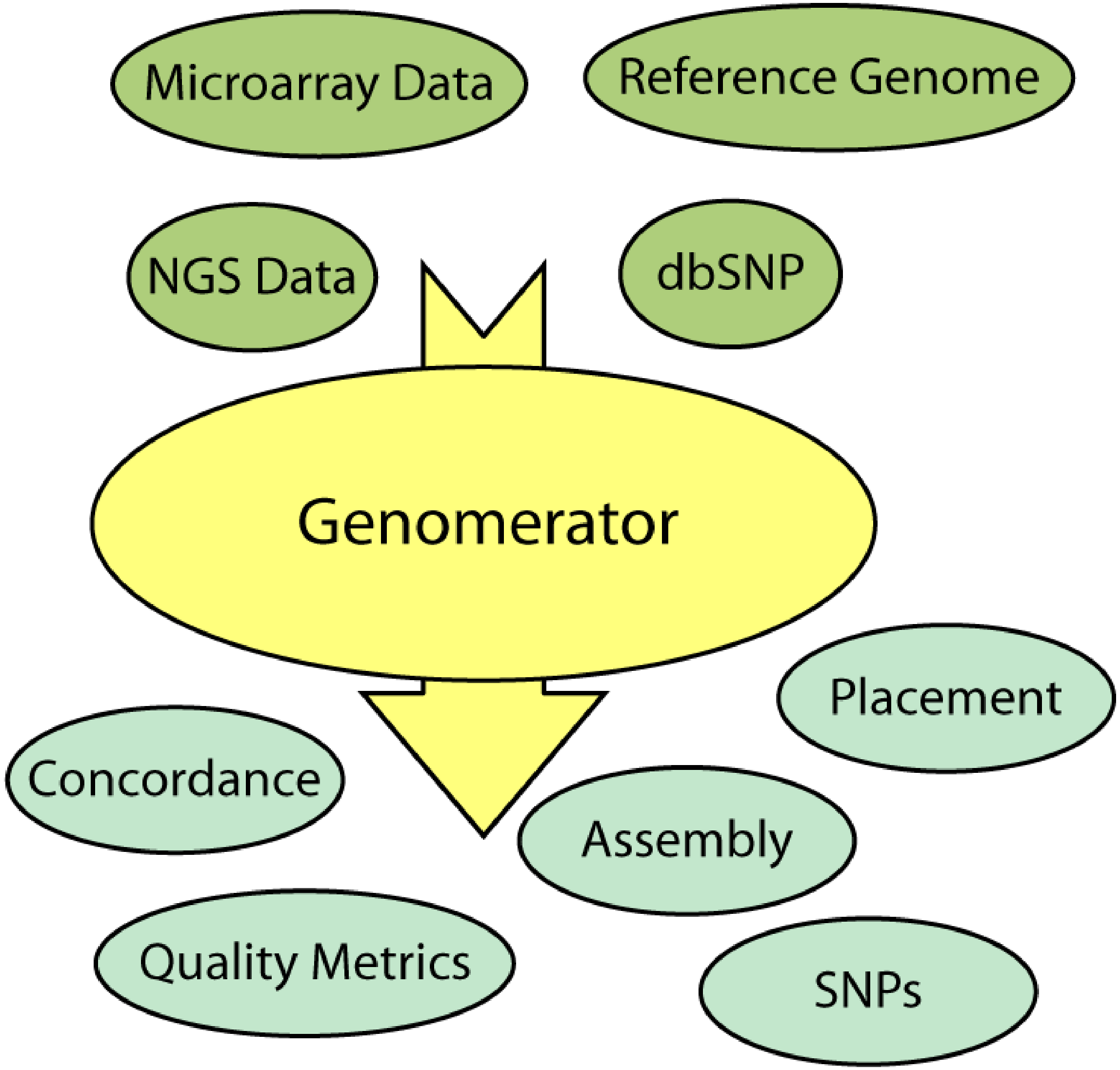
```
chr5 42735769 42735805     GAAGCACCACGcAaTGCAGATaTTcaGAAaGGAtGG
chr5 42735776 42735812            CaCgcAATgCaGaTaTtCagaaAtgATggAtggttc
chr5 42735776 42735812            CacGCaaTGCaGatATTcaGaaATGaTggATggtTc
chr5 42735776 42735812            CAcGCAATGCAGaTaTTcagaAATgatggatggtTc
chr5 42735776 42735812            CACGCAATGCAGATATTCAGAAATGATGgATGGtTc
chr5 42735790 42735826                          ATTcAGAAAGGATGGATGGTTcTGGAGTATGAACTT
chr5 42735790 42735826                          AttcAgAAAGGATGGAtGGTtCtGGAGTATGAACtT
                                                         *
```

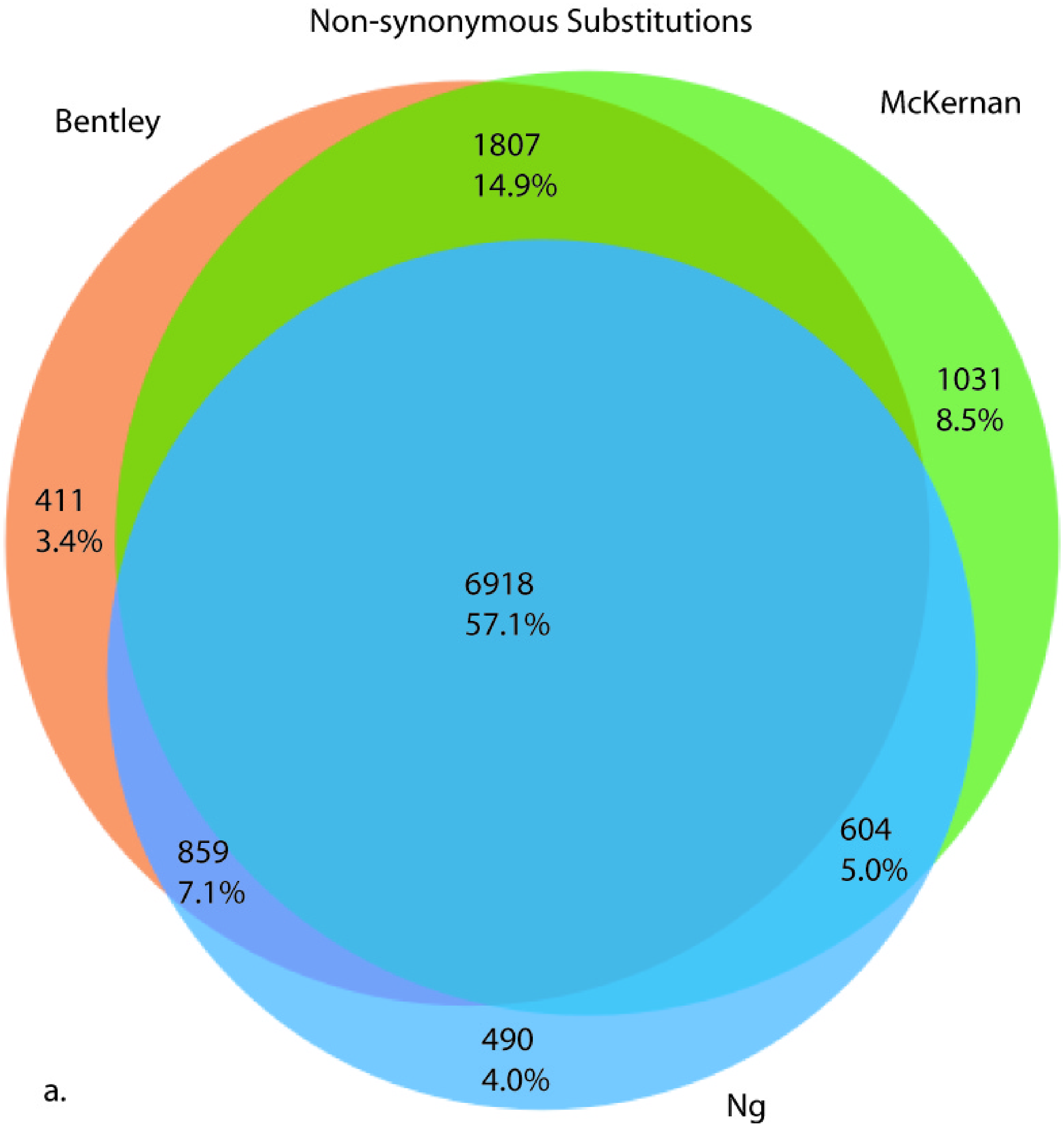
PGP2 MFN2 Mutation

```
chr1 11984646 11984682 AGTGAAGACCAAGTTTGAGCAGCACACGGTCCGGGC
chr1 11984658 11984694             GTTTGAGCAGCACACGGTCCGGGCCAAGCAGATTGC
chr1 11984658 11984694             GTTTGAGCAGCACAcGGTCCGGGCcAAGCaGATTGC
chr1 11984658 11984694             GTTTGAGCAGCACACGGTCCGGGCCAAGCAGATTGC
chr1 11984658 11984694             GTTTGAGCAGCACACGGTCCGGGCCAAGCAGATTGC
chr1 11984658 11984694             GTTTGAGCAGCACACGGTCCGGGCCAAGCaGATTGC
chr1 11984658 11984694             GTTTGAGCAGCACACgGTCCgGGCCaaGCAGATTgC
chr1 11984658 11984694             GTTTGAGCAGCACACGGTCCGGGCCAAGCAGATTGC
chr1 11984662 11984698                 GAGCAGCACACGGTCCGGGCCAAGCAGATTGCAGAG
chr1 11984662 11984698                 GAGCAGCACACGGTCCGGgCCAagCAgATTgCAGAg
chr1 11984662 11984698                 GAGCAGCACACGGTCCGGGCCAAGCAGaTTGCAgAG
chr1 11984662 11984698                 gAgCAGCACACgGTCCGGGCCaAGCAGATTGCAGAG
chr1 11984665 11984701                    CAGCACACGGTCCGGGCCAAGCAGATTGCAGAGGCG
chr1 11984667 11984703                      GCACACGGTCTGGGCCAAGCAGATTGCAGAGGCGGg
chr1 11984667 11984703                      GCACACGGTCTGGGCCAaGCAGATTGCAGAGGCGGg
chr1 11984667 11984703                      GCACACGGTCTGGGCCAAGCAGATTGCAGAGGCGGg
chr1 11984667 11984703                      GCACACGGTCTGGGCCAAGCAGATTGCAGAGGCGGt
chr1 11984668 11984704                       CACACGGTCCGGGCCAAGCAGATTGCAGAGGCGGTT
chr1 11984668 11984704                       CACACGGTCCGGGCCAAGCAGATTGCAGAGGCGGTT
                                                      *
```

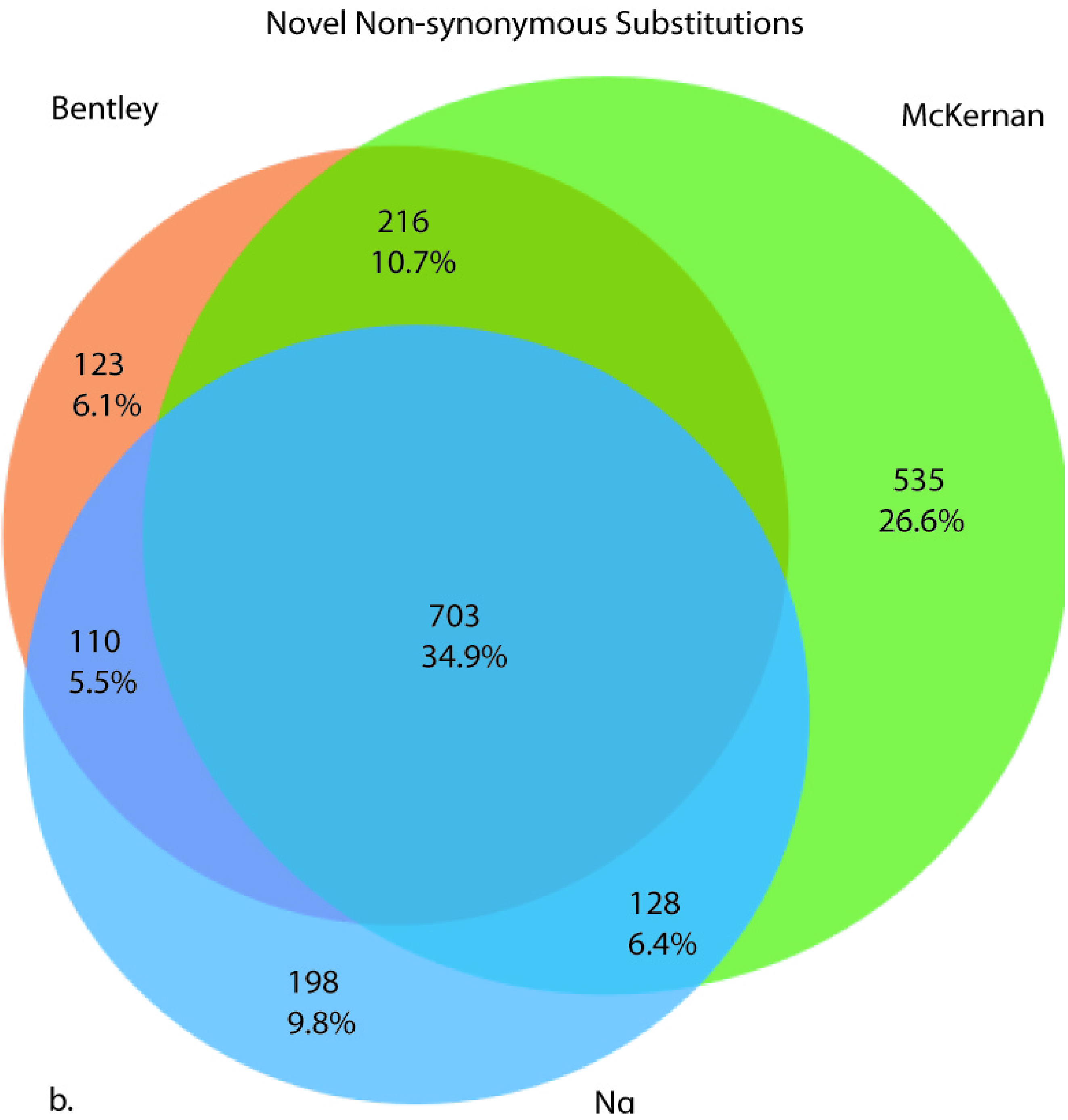
So what went wrong? The error probably occurs in an amplification step required by the capture process.

On our cloud, Genomerator manages NGS data, launches workflows, and, generates quality metrics – without high quality variant calls and data-sources Trait-o-matic is useless!



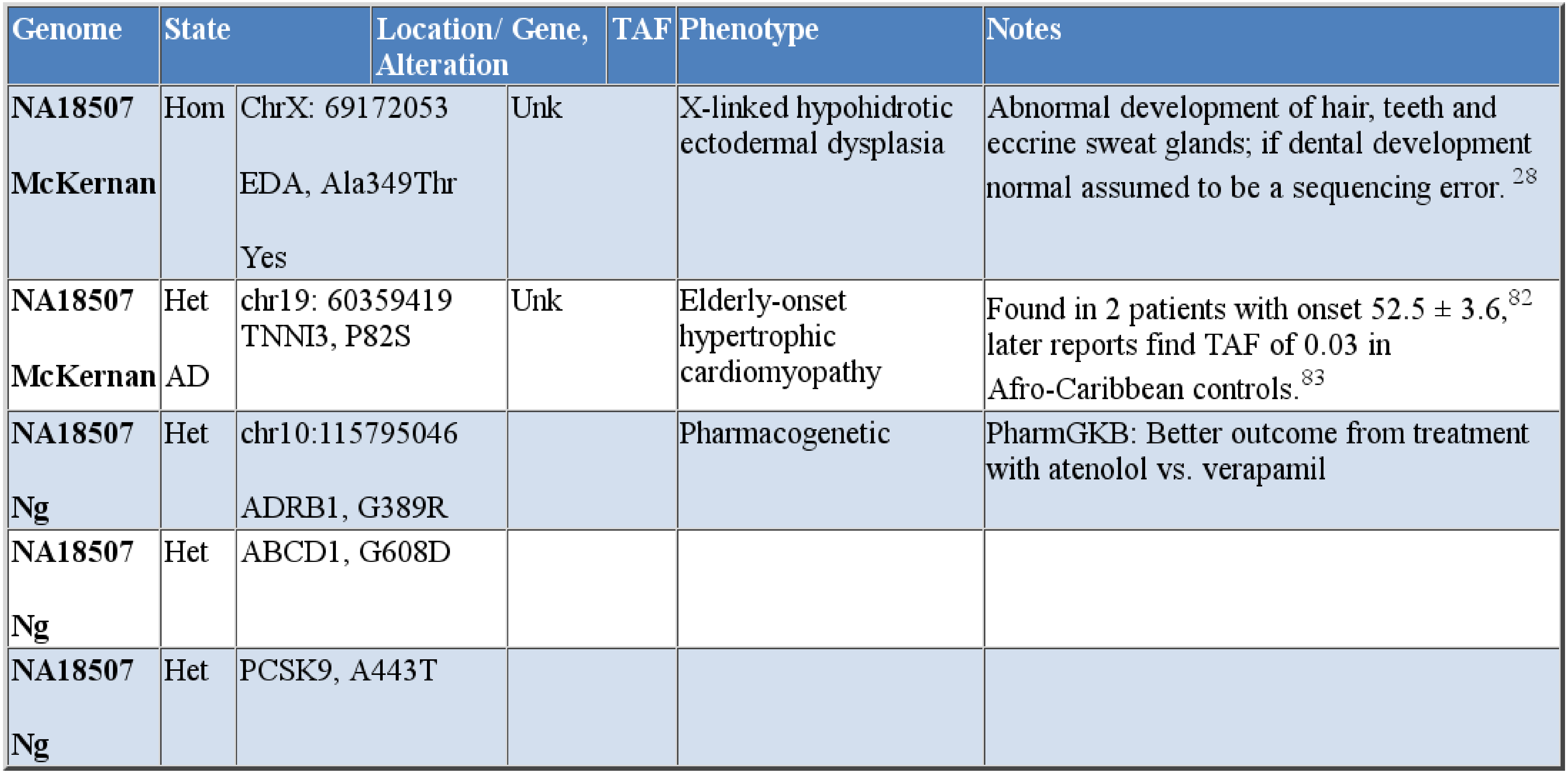


Comparison of non-synonymous substitutions from three independent experiments on the same HapMap sample, NA18507, indicates relatively poor concordance between all three samples.



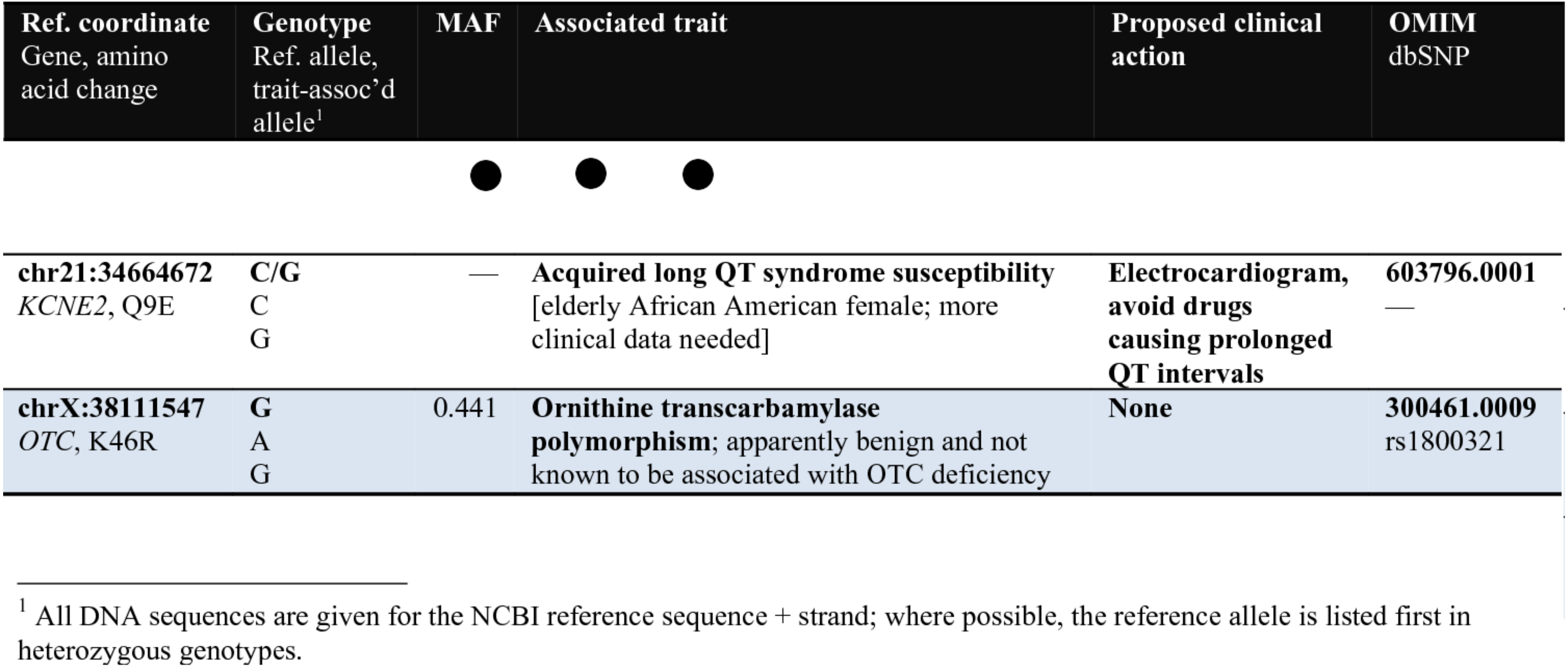
Comparison of non-synonymous substitutions from three independent experiments on the same HapMap sample, NA18507, novel variants have even worse concordance.

NA18507 Variants Found by Only One Group



Genome	State	Location/ Gene, Alteration	TAF	Phenotype	Notes
NA18507 McKernan	Hom	ChrX: 69172053 EDA, Ala349Thr Yes	Unk	X-linked hypohidrotic ectodermal dysplasia	Abnormal development of hair, teeth and eccrine sweat glands; if dental development normal assumed to be a sequencing error. [28]
NA18507 McKernan	Het AD	chr19: 60359419 TNNI3, P82S	Unk	Elderly-onset hypertrophic cardiomyopathy	Found in 2 patients with onset 52.5 ± 3.6,[82] later reports find TAF of 0.03 in Afro-Caribbean controls.[83]
NA18507 Ng	Het	chr10:115795046 ADRB1, G389R		Pharmacogenetic	PharmGKB: Better outcome from treatment with atenolol vs. verapamil
NA18507 Ng	Het	ABCD1, G608D			
NA18507 Ng	Het	PCSK9, A443T			

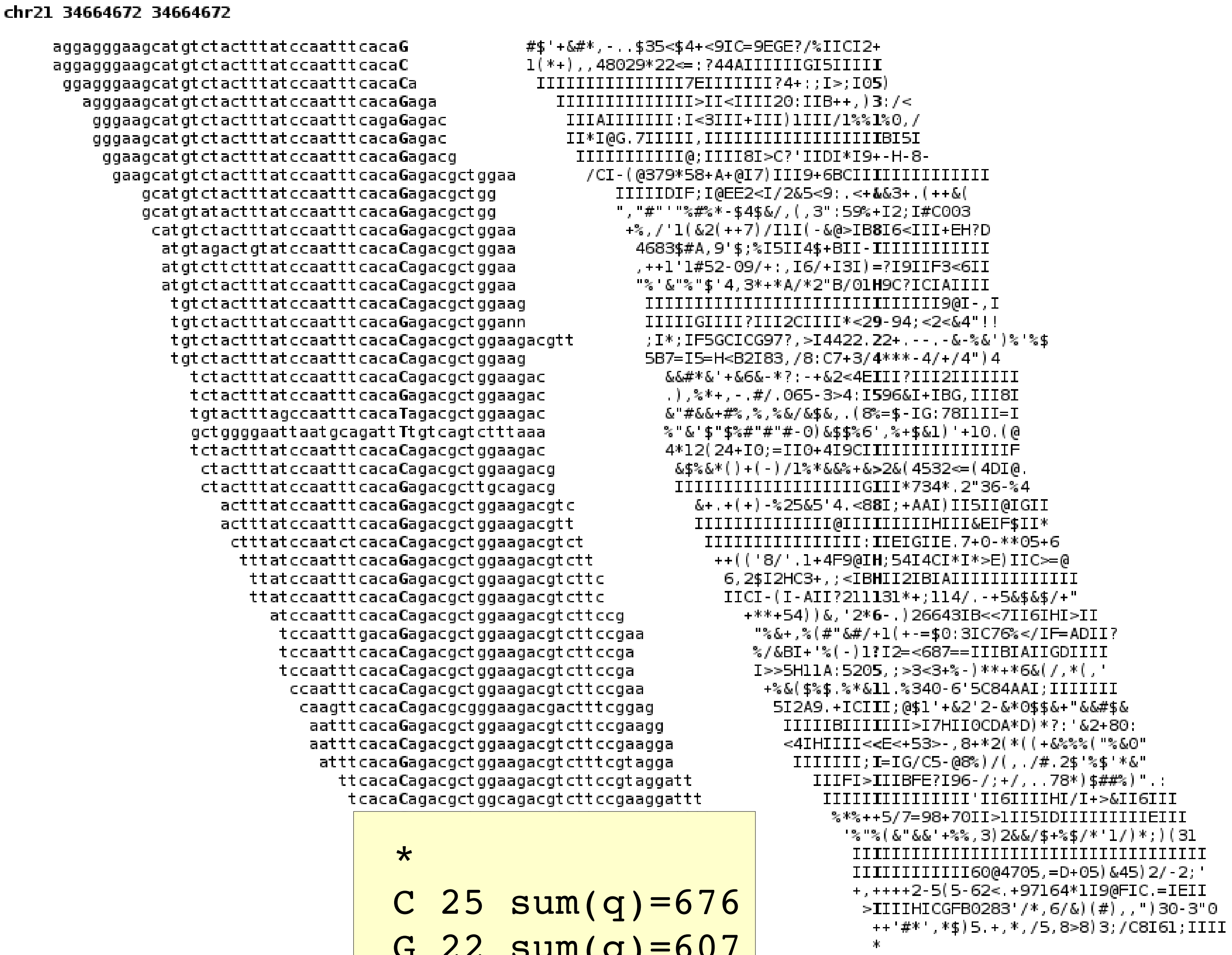
These variants could be sequencing errors that are easily seen in the consensus alignments or even the underlying images. It's also possible, however, that the raw data will support these consensus calls as “real” while the poor replication across three experiments suggests the opposite.



Ref. coordinate Gene, amino acid change	**Genotype** Ref. allele, trait-assoc'd allele[1]	**MAF**	**Associated trait**	**Proposed clinical action**	**OMIM** dbSNP
		●	● ●		
chr21:34664672 *KCNE2*, Q9E	**C/G** C G	—	**Acquired long QT syndrome susceptibility** [elderly African American female; more clinical data needed]	**Electrocardiogram, avoid drugs causing prolonged QT intervals**	**603796.0001** —
chrX:38111547 *OTC*, K46R	**G** A G	0.441	**Ornithine transcarbamylase polymorphism**; apparently benign and not known to be associated with OTC deficiency	**None**	**300461.0009** rs1800321

[1] All DNA sequences are given for the NCBI reference sequence + strand; where possible, the reference allele is listed first in heterozygous genotypes.

Analysis of an individual African genome reveals a rare mutation—KCNE2 Q9E—not present in dbSNP. Is this variant real?



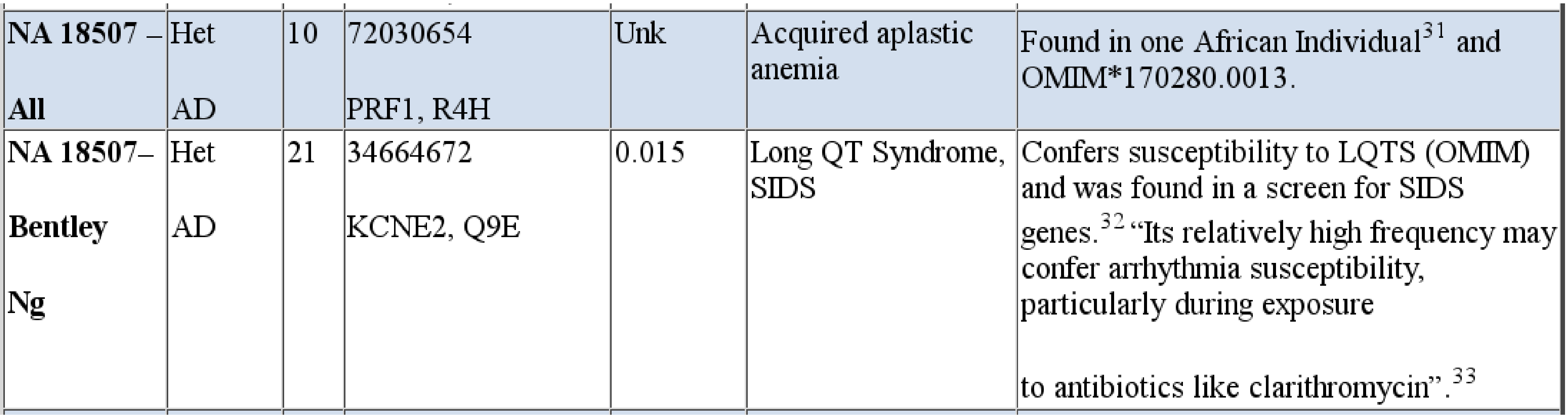
```
chr21 34664672 34664672

     aggagggaagcatgtctactttatccaatttcacaG            #$'+&#*,-..$35<$4+<9IC=9EGE?/%IICI2+
     aggagggaagcatgtctactttatccaatttcacaC            1(*+),,48029*22<=:?44AIIIIIIGI5IIIII
      ggagggaagcatgtctactttatccaatttcacaCa            IIIIIIIIIIIIIII7EIIIIIII?4+:;I>;I05)
        agggaagcatgtctactttatccaatttcacaGaga            IIIIIIIIIIIIII>II<IIII20:IIB++,)3:/<
         gggaagcatgtctactttatccaatttcagaGagac            IIIAIIIIIII:I<3III+III)1III/1%%1%0,/
         gggaagcatgtctactttatccaatttcacaGagac            II*I@G.7IIIII,IIIIIIIIIIIIIIIIIIBI5I
          ggaagcatgtctactttatccaatttcacaGagacg            IIIIIIIIIII@;IIII8I>C?'IIDI*I9+-H-8-
           gaagcatgtctactttatccaatttcacaGagacgctggaa       /CI-(@379*58+A+@I7)III9+6BCIIIIIIIIIIIIII
              gcatgtctactttatccaatttcacaGagacgctgg            IIIIIDIF;I@EE2<I/2&5<9:.<+&&3+.(++&(
              gcatgtatactttatccaatttcacaGagacgctgg            ","#"'"%#%*-$4$&/,(,3":59%+I2;I#C003
               catgtctactttatccaatttcacaGagacgctggaa           +%,/'1(&2(++7)/I1I(-&@>IB8I6<III+EH?D
                atgtagactgtatccaatttcacaCagacgctggaa            4683$#A,9'$;%I5II4$+BII-IIIIIIIIIIII
                atgtcttctttatccaatttcacaCagacgctggaa            ,++1'1#52-09/+:,I6/+I3I)=?I9IIF3<6II
                atgtctactttatccaatttcacaCagacgctggaa            "%'&"%"$'4,3*+*A/*2"B/01H9C?ICIAIIII
                 tgtctactttatccaatttcacaCagacgctggaag            IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9@I-,I
                 tgtctactttatccaatttcacaGagacgctggann            IIIIIGIIII?III2CIIII*<29-94;<2<&4"!!
                 tgtctactttatccaatttcacaCagacgctggaagacgtt       ;I*;IF5GCICG97?,>I4422.22+.--.-&-%&')%'%$
                 tgtctactttatccaatttcacaCagacgctggaag            5B7=I5=H<B2I83,/8:C7+3/4***-4/+/4")4
                   tctactttatccaatttcacaCagacgctggaagac            &&#*&'+&6&-*?:-+&2<4EIII?III2IIIIIII
                   tctactttatccaatttcacaGagacgctggaagac            .),%*+,-.#/.065-3>4:I596&I+IBG,III8I
                   tgtactttagccaatttcacaTagacgctggaagac            &"#&&+#%,%,%&/&$&,.(8%=$-IG:78I1II=I
                   gctggggaattaatgcagattTtgtcagtctttaaa            %"&'$"$%#"#"#-0)&$$%6',%+$&1)'+10.(@
                   tctactttatccaatttcacaCagacgctggaagac            4*12(24+I0;=II0+4I9CIIIIIIIIIIIIIIIF
                    ctactttatccaatttcacaCagacgctggaagacg            &$%&*()+(-)/1%*&&%+&>2&(4532<=(4DI@.
                    ctactttatccaatttcacaGagacgcttgcagacg            IIIIIIIIIIIIIIIIIIIGIII*734*.2"36-%4
                      actttatccaatttcacaGagacgctggaagacgtc            &+.+(+)-%25&5'4.<88I;+AAI)II5II@IGII
                      actttatccaatttcacaGagacgctggaagacgtt            IIIIIIIIIIIIII@IIIIIIIIIHIII&EIF$II*
                       ctttatccaatctcacaCagacgctggaagacgtct            IIIIIIIIIIIIIIII:IIEIGIIE.7+0-**05+6
                        tttatccaatttcacaGagacgctggaagacgtctt            ++(('8/'.1+4F9@IH;54I4CI*I*>E)IIC>=@
                         ttatccaatttcacaGagacgctggaagacgtcttc            6,2$I2HC3+,;<IBHII2IBIAIIIIIIIIIIIII
                         ttatccaatttcacaCagacgctggaagacgtcttc            IICI-(I-AII?211131*+;114/.-+5&$&$/+"
                           atccaatttcacaCagacgctggaagacgtcttccg            +**+54))&,'2*6-.)26643IB<<7II6IHI>II
                            tccaatttgacaGagacgctggaagacgtcttccgaa           "%&+,%(#"&#/+1(+-=$0:3IC76%</IF=ADII?
                            tccaatttcacaCagacgctggaagacgtcttccga            %/&BI+'%(-)1?I2=<687==IIIBIAIIGDIIII
                            tccaatttcacaCagacgctggaagacgtcttccga            I>>5H11A:5205,;>3<3+%-)**+*6&(/,*(,'
                             ccaatttcacaCagacgctggaagacgtcttccgaa            +%&($%$.%*&11.%340-6'5C84AAI;IIIIIII
                              caagttcacaCagacgcgggaagacgactttcggag            5I2A9.+ICIII;@$1'+&2'2-&*0$$&+"&&#$&
                               aatttcacaGagacgctggaagacgtcttccgaagg            IIIIIBIIIIIII>I7HII0CDA*D)*?:'&2+80:
                               aatttcacaCagacgctggaagacgtcttccgaagga           <4IHIIII<<E<+53>-,8+*2(*((+&%%%("%&0"
                                atttcacaGagacgctggaagacgtctttcgtagga            IIIIIII;I=IG/C5-@8%)/(,./#.2$'%$'*&"
                                  ttcacaCagacgctggaagacgtcttccgtaggatt            IIIFI>IIIBFE?I96-/;+/,..78*)$##%)".:
                                   tcacaCagacgctggcagacgtcttccgaaggattt            IIIIIIIIIIIIIII'II6IIIIHI/I+>&II6III
                                                                                    %*%++5/7=98+70II>1II5IDIIIIIIIIIEIII
                                                                                     '%"%(&"&&'+%%,3)2&&/$+%$/*'1/)*;)(31
                                                                                      IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
                                                                                      IIIIIIIIIIII60@4705,=D+05)&45)2/-2;'
                                                                                      +,++++2-5(5-62<.+97164*1I9@FIC.=IEII
                                                                                       >IIIIHICGFB0283'/*,6/&)(#),,")30-3"0
                                                                                        ++'#*',*$)5.+,*,/5,8>8)3;/C8I61;IIII
                                                                                        *
```

```
*
C 25 sum(q)=676
G 22 sum(q)=607
T 2 sum(q)=10
```

Our cloud infrastructure was used to assemble the raw reads—120 gigabases—from HapMap NA18507. The alignment for KCNE2 Q9E is shown above. Manually assembled from data in Bentley et al. (2008) Nature.

Further literature search brings into question the importance of KCNE2 Q9E

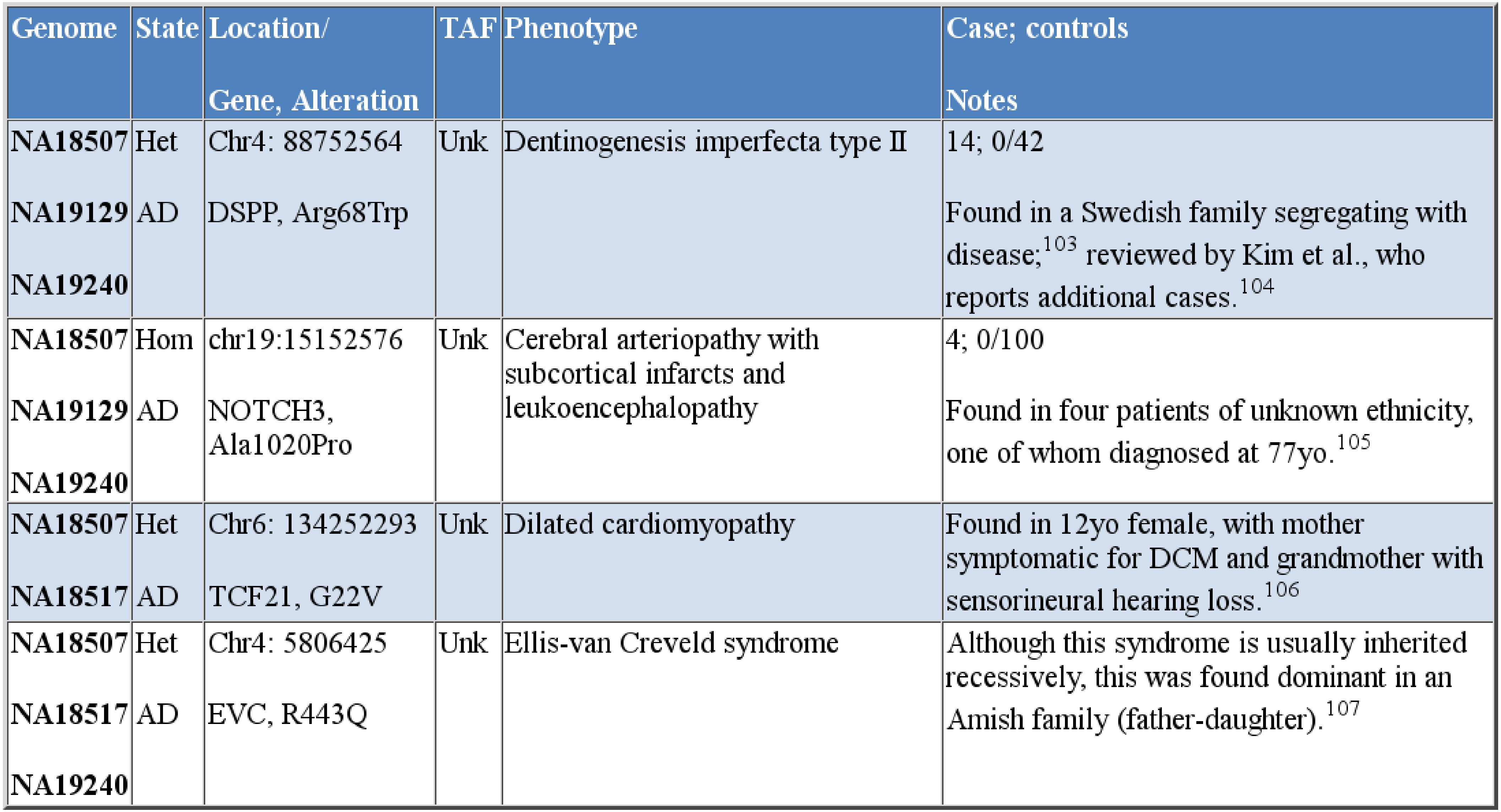


NA 18507 – **All**	Het AD	10	72030654 PRF1, R4H	Unk	Acquired aplastic anemia	Found in one African Individual[31] and OMIM*170280.0013.
NA 18507– **Bentley** **Ng**	Het AD	21	34664672 KCNE2, Q9E	0.015	Long QT Syndrome, SIDS	Confers susceptibility to LQTS (OMIM) and was found in a screen for SIDS genes.[32] “Its relatively high frequency may confer arrhythmia susceptibility, particularly during exposure to antibiotics like clarithromycin”.[33]

Can clinical genetic labs share (some of) these data which are typically proprietary?

Without comprehensive and accurate genotype to phenotype databases—good variant calls are not clinically useful.

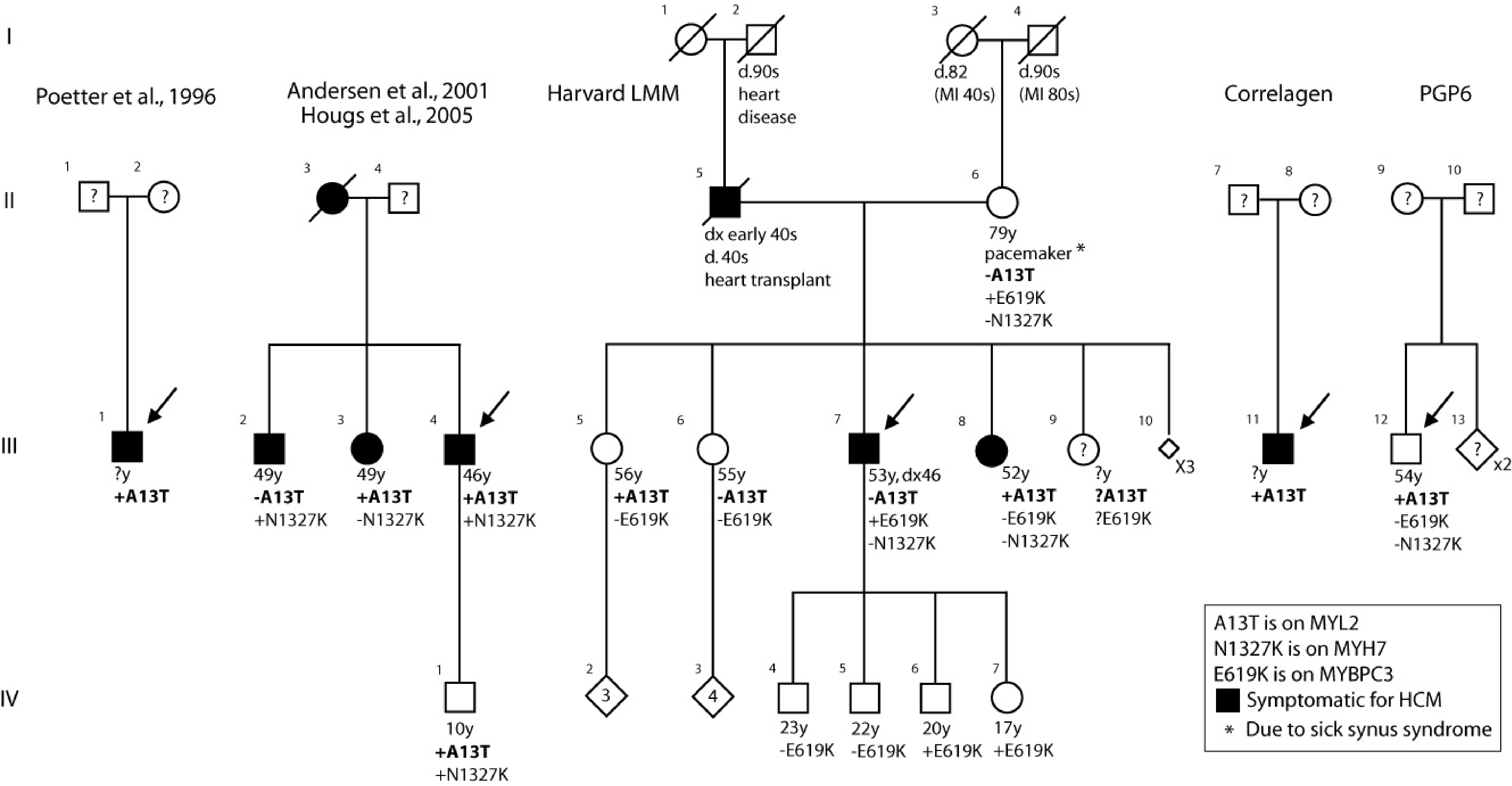
Variants Implicated in Disease with Unreported Frequencies, but Appearing Frequently in YRI Genomes.



Genome	State	Location/ Gene, Alteration	TAF	Phenotype	Case; controls Notes
NA18507 **NA19129** **NA19240**	Het AD	Chr4: 88752564 DSPP, Arg68Trp	Unk	Dentinogenesis imperfecta type II	14; 0/42 Found in a Swedish family segregating with disease;[103] reviewed by Kim et al., who reports additional cases.[104]
NA18507 **NA19129** **NA19240**	Hom AD	chr19:15152576 NOTCH3, Ala1020Pro	Unk	Cerebral arteriopathy with subcortical infarcts and leukoencephalopathy	4; 0/100 Found in four patients of unknown ethnicity, one of whom diagnosed at 77yo.[105]
NA18507 **NA18517**	Het AD	Chr6: 134252293 TCF21, G22V	Unk	Dilated cardiomyopathy	Found in 12yo female, with mother symptomatic for DCM and grandmother with sensorineural hearing loss.[106]
NA18507 **NA18517** **NA19240**	Het AD	Chr4: 5806425 EVC, R443Q	Unk	Ellis-van Creveld syndrome	Although this syndrome is usually inherited recessively, this was found dominant in an Amish family (father-daughter).[107]

Interpreting 2nd generation sequencing results goes far beyond accurate variant calls but requires a worldwide effort to develop accessible databases of cases and controls; without such databases clinical interpretation will remain elusive!

Personal Genomics has arrived but it will take significant community effort to achieve its potential—you can help!



A13T is on MYL2
N1327K is on MYH7
E619K is on MYBPC3
■ Symptomatic for HCM
* Due to sick synus syndrome

Acknowledgments

George Church
James Hogle
Zak Kohane
Jon Seidman
Jack Szostak
Abraham Rosenbaum
Xiaodi Wu
Mike Chou
Billy Li
Wendy Chung
Heidi Rehm
John Aach
Joe Thakuria
Erez Levanon
Jason Bobe

Harris Wang
Tom Clegg
Ward Vandewege
Nava Whiteford
Chris Archibald
Andy Chute
Andrea Loehr
Irwin Jungreis
Miron Cuperman
Erik Garrison
Paras Doshi
Church lab members
Harvard Biophysics
Many others

Thank-you!