Practical: *De novo* mutation identification and analysis

Instructor: Daniel P Howrigan
Massachusetts General Hospital, Boston, MA
Broad Institute, Cambridge, MA

2015 International workshop on statistical genetic methods for human complex traits

Overview

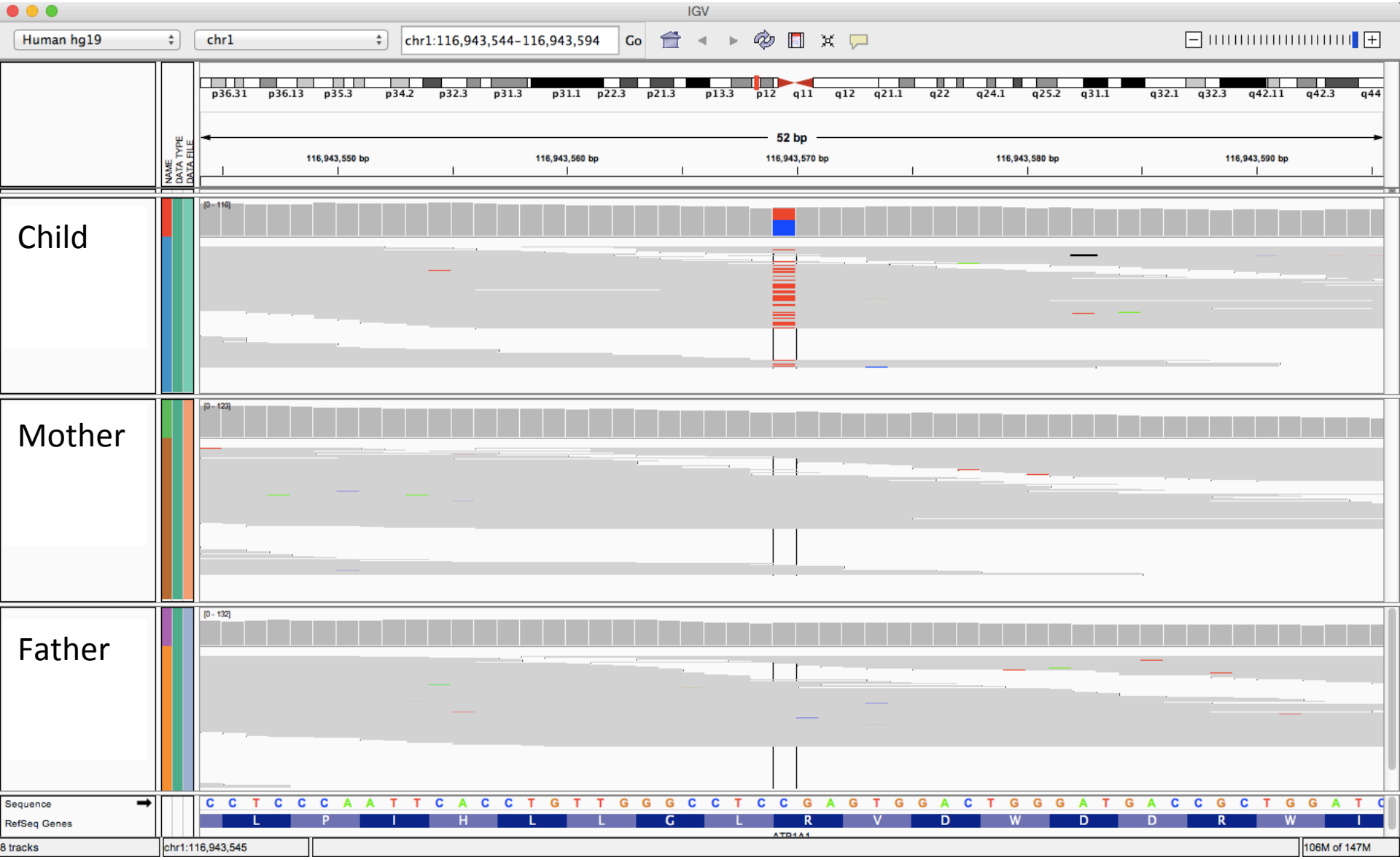
De novo identification

- Visualizing a *de novo* variant
- Using genotype information from the VCF
- Assessing potential errors in de novo identification

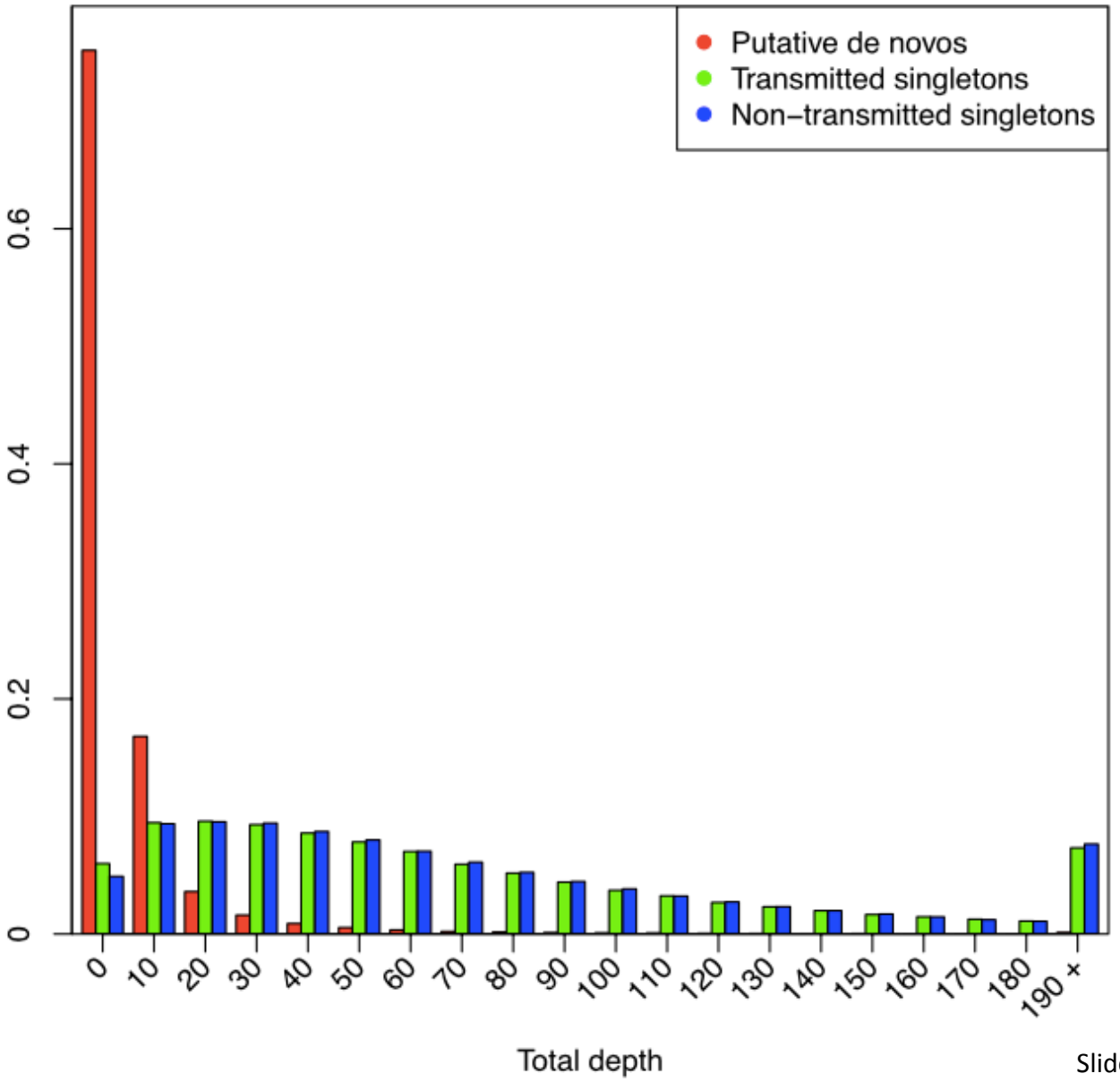
De novo analysis

- Modeling the expectation of *de novo* mutations
- Testing individual genes
- Testing for enrichment in gene sets

Visualizing a *de novo* variant



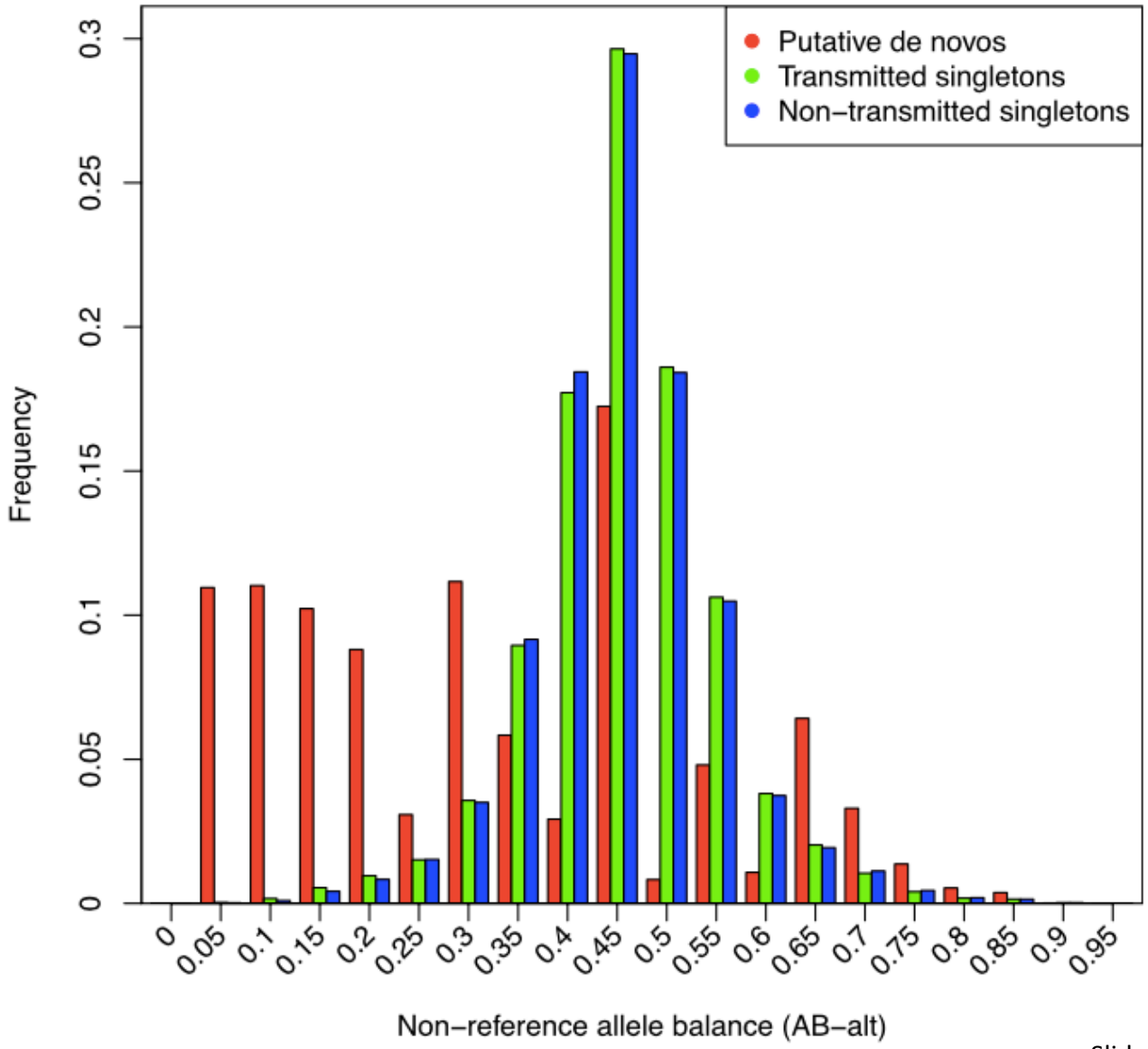
De novo calling is highly susceptible to sequencing errors



All variants have passed quality control in GATK

Slide courtesy of Fromer et al. 2014 *Nature*

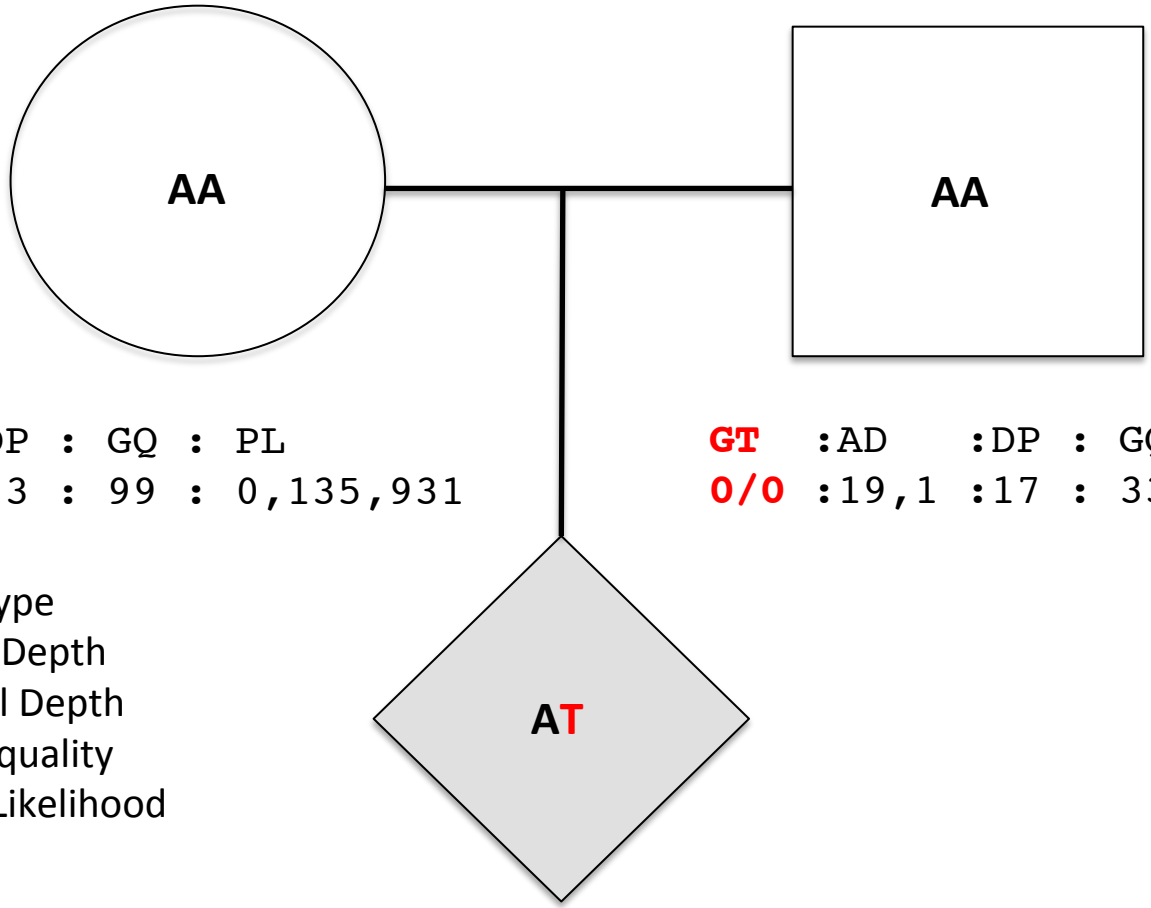
De novo calling is highly susceptible to sequencing errors



All variants have passed quality control in GATK

Slide courtesy of Fromer et al. 2014 *Nature*

Calling *De Novo* Variants



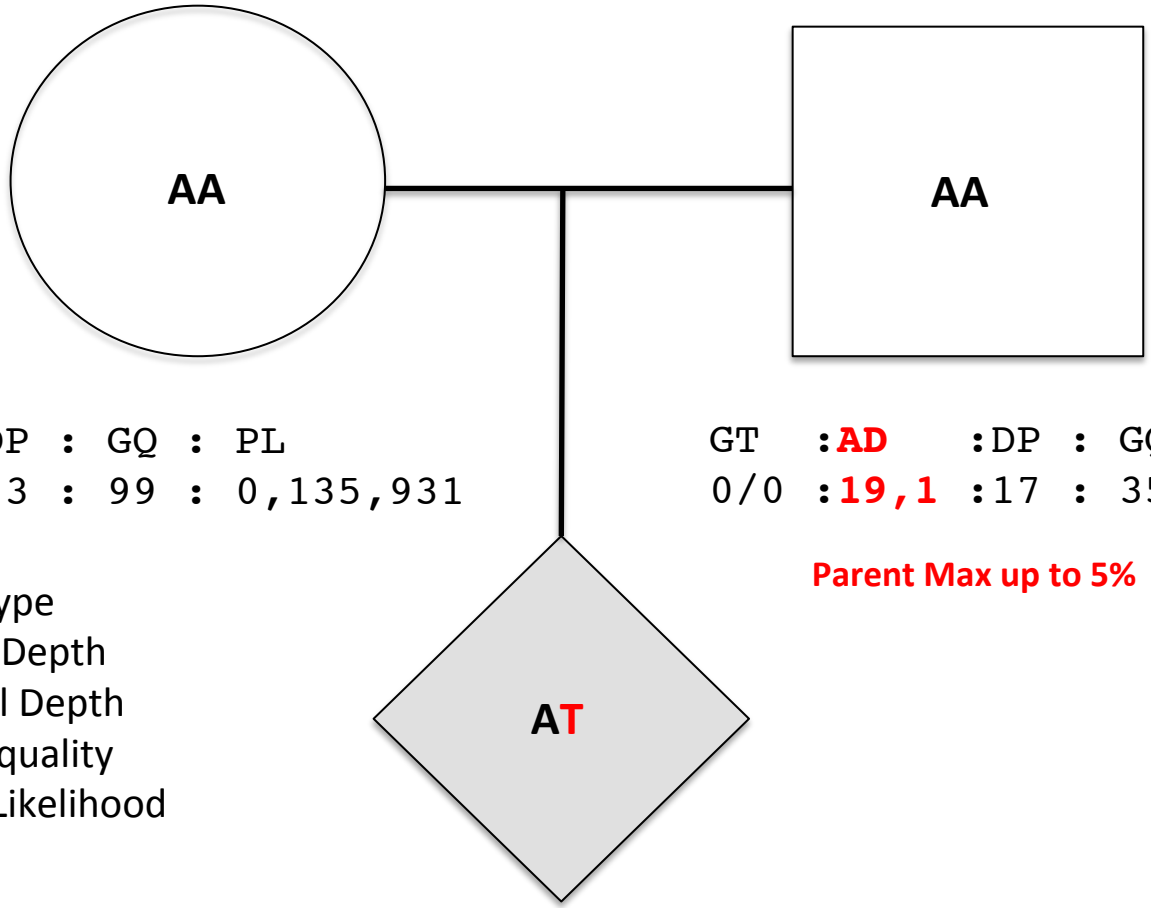
GT	:AD	:DP	: GQ	: PL
0/0	:25,0	:23	: 99	: 0,135,931

GT	:AD	:DP	: GQ	: PL
0/0	:19,1	:17	: 33	: 0,35,121

GT = Genotype
AD = Allelic Depth
DP = Overall Depth
GQ = Geno quality
PL = Phred Likelihood

GT	:AD	:DP	: GQ	: PL
0/1	:6,9	:13	: 34	: 0,34,81

Calling *De Novo* Variants




```
GT   :AD    :DP : GQ : PL
0/0 :25,0 :23 : 99 : 0,135,931
```

```
GT   :AD    :DP : GQ : PL
0/0 :19,1 :17 : 35 : 0,35,121
```

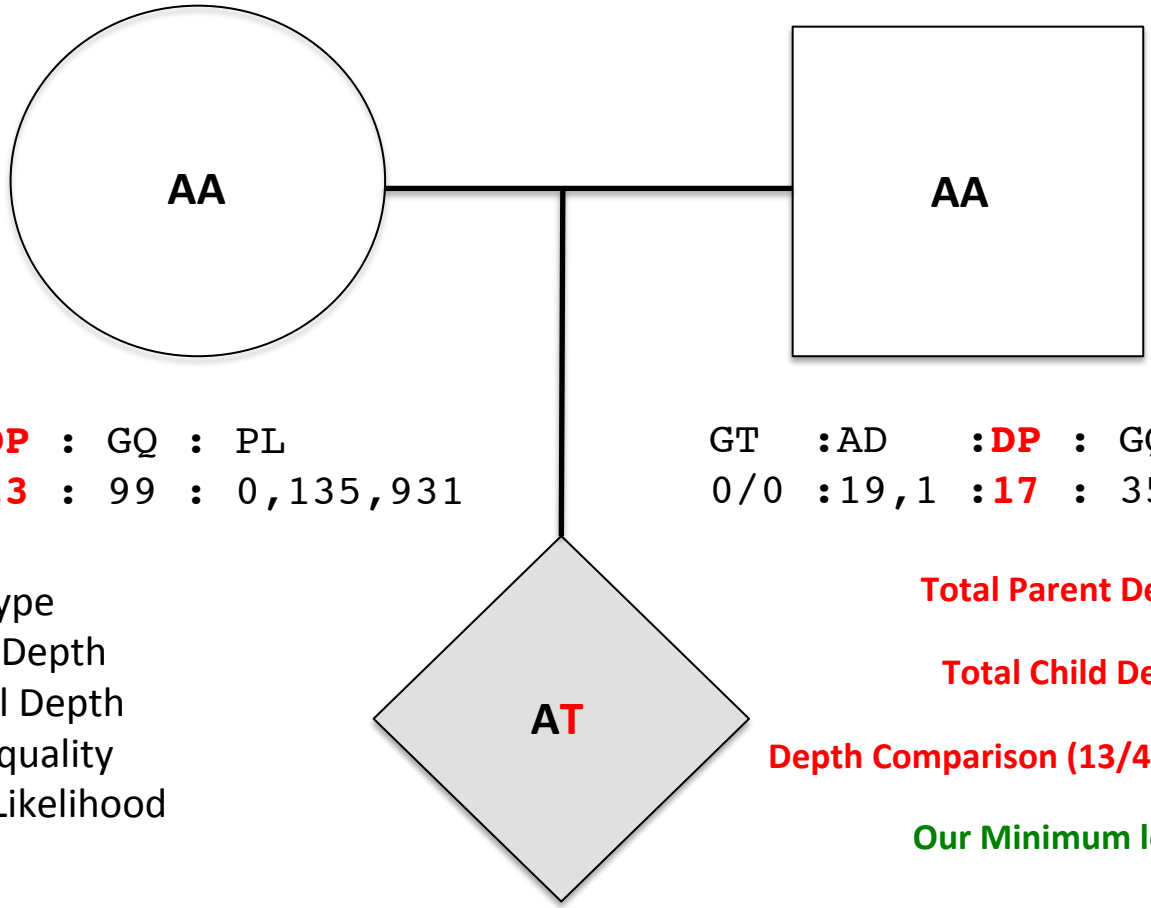
Parent Max up to 5%

GT = Genotype
AD = Allelic Depth
DP = Overall Depth
GQ = Geno quality
PL = Phred Likelihood

```
GT   :AD   :DP : GQ : PL
0/1 :6,9 :13 : 34 : 34,0,81
```

Child Min down to 20%

Calling *De Novo* Variants




```
GT   :AD    :DP : GQ : PL
0/0 :25,0 :23 : 99 : 0,135,931
```

```
GT   :AD    :DP : GQ : PL
0/0 :19,1 :17 : 35 : 0,35,121
```

GT = Genotype
AD = Allelic Depth
DP = Overall Depth
GQ = Geno quality
PL = Phred Likelihood

Total Parent Depth = 40

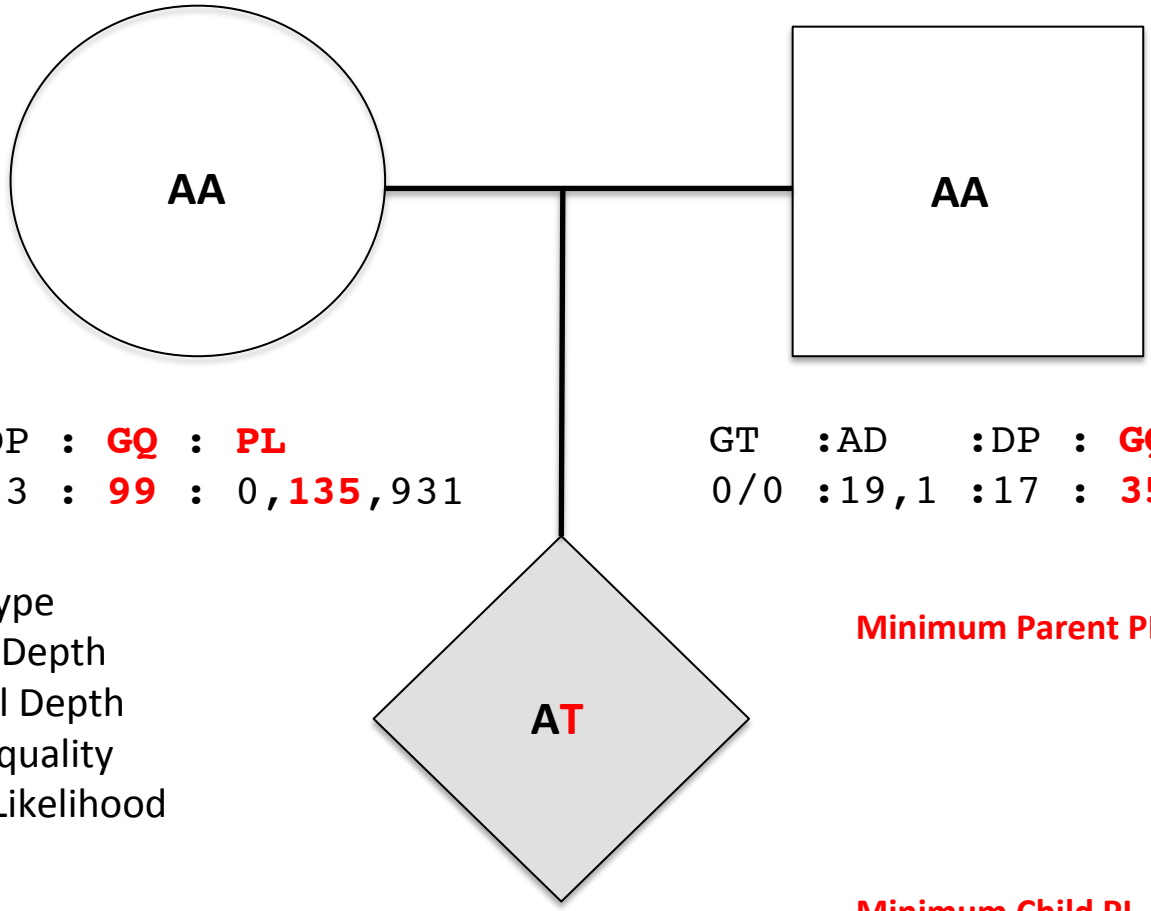
Total Child Depth = 13

Depth Comparison (13/40) = 32.5%

Our Minimum level = 10%

```
GT   :AD   :DP : GQ : PL
0/1 :6,9 :13 : 34 : 34,0,81
```

Calling *De Novo* Variants




```
GT   :AD    :DP : GQ : PL
0/0  :25,0  :23 : 99 : 0,135,931
```

```
GT   :AD    :DP : GQ : PL
0/0  :19,1  :17 : 35 : 0,35,121
```

GT = Genotype
AD = Allelic Depth
DP = Overall Depth
GQ = Geno quality
PL = Phred Likelihood

Minimum Parent PL *AB* : NONE

Minimum Child PL *AA* >= 20

```
GT   :AD   :DP : GQ : PL
0/1  :6,9  :13 : 34 : 34,0,81
```

practical: *de novo* identification

data directory: /faculty/dan/practical_2015/

Family file: denovo-example.fam

- 2 families
- Child/proband diagnosed with disorder and parents are unaffected
- File format identical to PLINK

VCF file: denovo-example.vcf

- Header has a lot of "meta-information" on the file
- 1st 853 variants in Chromosome 1

```
Command line:

mkdir denovo
cp -r /faculty/dan/practical_2015/* denovo
less denovo-example.fam
less -S denovo-example.vcf
```

press q to exit the 'less' program

practical: *de novo* identification

data directory: /faculty/dan/practical_2015/

Python program: de_novo_finder_3.py

- Scans through VCF for *de novo* variants that pass thresholds
- Returns a tab delimited list of *de novo* variants and genotype information

Using MAF as an additional QC parameter: all_ESP_counts_5.28.13.txt

- Exonic variants and allele frequencies from 5K+ exomes

```
Command line:

less -S de_novo_finder_3.py
less -S all_ESP_counts_5.28.13.txt
```

press q to exit the 'less' program

Combining call quality and population frequency for better calls

Prob of DNM: **P(true DNM | data) /**
(P(true DNM | data) + P(one parent het | data))

- **P(true DNM | data) = P(data | true DNM) * P(true DNM)**
- **P(data | true DNM)** = Pdad_ref * Pmom_ref * Pchild_het *(our observed DNM call quality)*
- **P(true DNM)** = 1/30 Mb *(theoretical DNM rate)*
- We want High numbers for this probability

- **P(one parent het | data)** = (Pdad_ref*Pmom_het + Pdad_het*Pmom_ref) * Pchild_het
- **P(one parent het)** = 1 - (1-F)^4 *(population MAF applied to parents)*
- **F** = Maximum MAF in either ESP or current data
- We want Low numbers for this probability

Combining call quality and population frequency for better calls

Prob of DNM: **P(true DNM | data) /**
(P(true DNM | data) + P(one parent het | data))

- **P(true DNM | data) = P(data | true DNM) * P(true DNM)**
- **P(data | true DNM)** = Pdad_ref * Pmom_ref * Pchild_het *(our observed DNM call quality)*
- **P(true DNM)** = 1/30 Mb *(theoretical DNM rate)*
- We want High numbers for this probability

- **P(one parent het | data)** = (Pdad_ref*Pmom_het + Pdad_het*Pmom_ref) * Pchild_het
- **P(one parent het)** = 1 - (1-F)^4 *(population MAF applied to parents)*
- **F** = Maximum MAF in either ESP or current data
- We want Low numbers for this probability

Combining call quality and population frequency for better calls

Prob of DNM: **P(true DNM | data) /**
(P(true DNM | data) + P(one parent het | data))

- **P(true DNM | data) = P(data | true DNM) * P(true DNM)**
- **P(data | true DNM)** = Pdad_ref * Pmom_ref * Pchild_het *(our observed DNM call quality)*
- **P(true DNM)** = 1/30 Mb *(theoretical DNM rate)*
- We want High numbers for this probability

- **P(one parent het | data)** = (Pdad_ref*Pmom_het + Pdad_het*Pmom_ref) * Pchild_het
- **P(one parent het)** = 1 - (1-F)^4 *(population MAF applied to parents)*
- **F** = Maximum MAF in either ESP or current data
- We want Low numbers for this probability

practical: *de novo* identification

running de_novo_finder_3.py

```
Command line:

python de_novo_finder_3.py \
denovo-example.vcf \
denovo-example.fam \
all_ESP_counts_5.28.13.txt -q > example.denovo.txt
```

```
Command line:

less -S example.denovo.txt
column -t example.denovo.txt | less -S
```

Overview

***De novo* identification**

- Visualizing a *de novo* variant
- Using genotype information from the VCF
- Assessing potential errors in de novo identification

***De novo* analysis**

- Modeling the expectation of *de novo* mutations
- Testing individual genes
- Testing for enrichment in gene sets / pathways

A model for interpreting *de novo* mutation

Patterns and rates of exonic *de novo* mutations in autism spectrum disorders

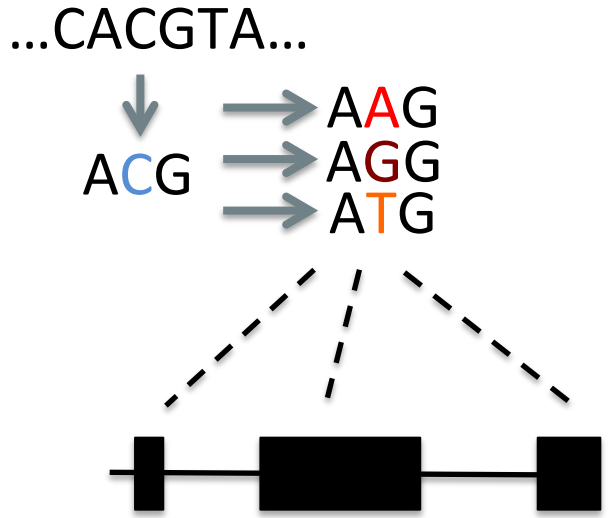
Benjamin M. Neale[1,2], Yan Kou[3,4], Li Liu[5], Avi Ma'ayan[3], Kaitlin E. Samocha[1,2], Aniko Sabo[6], Chiao-Feng Lin[7], Christine Stevens[2], Li-San Wang[7], Vladimir Makarov[4,8], Paz Polak[2,9], Seungtai Yoon[4,8], Jared Maguire[2], Emily L. Crawford[10], Nicholas G. Campbell[10], Evan T. Geller[7], Otto Valladares[7], Chad Scha
Omar Jabado[12], Zuleyma Peralta[12], Uma Nag
Lora Lewis[6], Yi Han[6], Benjamin F. Voight[2,13]
Menachem Fromer[1,2], Khalid Shakir[2], Tim Fe
Jack R. Wimbish[14], Braden E. Boone[14], Shaw
Joseph D. Buxbaum[4,8,12,17], Edwin H. Cook J
James S. Sutcliffe[10] & Mark J. Daly[1,2]

A framework for the interpretation of *de novo* mutation in human disease

Kaitlin E Samocha[1–4], Elise B Robinson[1–3], Stephan J Sanders[5,6], Christine Stevens[2,3], Aniko Sabo[7], Lauren M McGrath[8], Jack A Kosmicki[1,9,10], Karola Rehnström[11,12], Swapan Mallick[13], Andrew Kirby[1,2], Dennis P Wall[9,10], Daniel G MacArthur[1,2], Stacey B Gabriel[2], Mark DePristo[14], Shaun M Purcell[1,2,8,15–17], Aarno Palotie[8,11,12], Eric Boerwinkle[7,18], Joseph D Buxbaum[15–17,19–21], Edwin H Cook Jr[22], Richard A Gibbs[7], Gerard D Schellenberg[23], James S Sutcliffe[24], Bernie Devlin[25], Kathryn Roeder[26,27], Benjamin M Neale[1–3] & Mark J Daly[1–3]

mutation probability estimated at each base position

- **Tri-nucleotide context of mutation**
- **Aggregate probabilities across various contexts**
  - Whole exome
  - Annotation classes (synonymous, missense, etc..)
  - Individual genes and gene sets
- **Utilize a Poisson model informed by trio size**



practical: single gene *de novo* enrichment

running multiple_hits_onelist.py and overlap2mutprobs_1.2.py

```
less -S fixed_mut_prob_fs_adjdepdiv.txt
```

view the gene model

```
python multiple_hits_onelist.py \
Neale_2012_denovo.txt > Neale_2012_genes.txt
```

select genes with recurrent mutations

```
python overlap2mutprobs_1.2.py \
Neale_2012_genes.txt \
fixed_mut_prob_fs_adjdepdiv.txt \
175 > Neale_2012_gene_results.txt
```

Test genes against model

```
perl -pe 's{, }{:}g' Neale_2012_gene_results.txt \
| column -t | less -S
```

view the results

practical: gene set *de novo* enrichment

running listcrusher_3.5.py

General framework

- The **[mutation model]** tests for an enrichment of our **[observed *de novos*]** in a given **[gene set of interest]**

- Enrichment dependent not on the number of trios, but on the number of *de novo* mutations

```
python list_crusher3_5.py \
fixed_mut_prob_fs_adjdepdiv.txt \
Neale_2012_denovo.txt \
JOINT_CONSTRAINT_829.set -p
```