

# Exploring genetic variation in in family studies with GEMINI

Aaron Quinlan  
University of Utah



USTAR Center for  
Genetic Discovery

[quinlanlab.org](http://quinlanlab.org)

Commands for this session:

<https://gist.github.com/arq5x/9e1928638397ba45da2e#file-gemini-intro-sh>

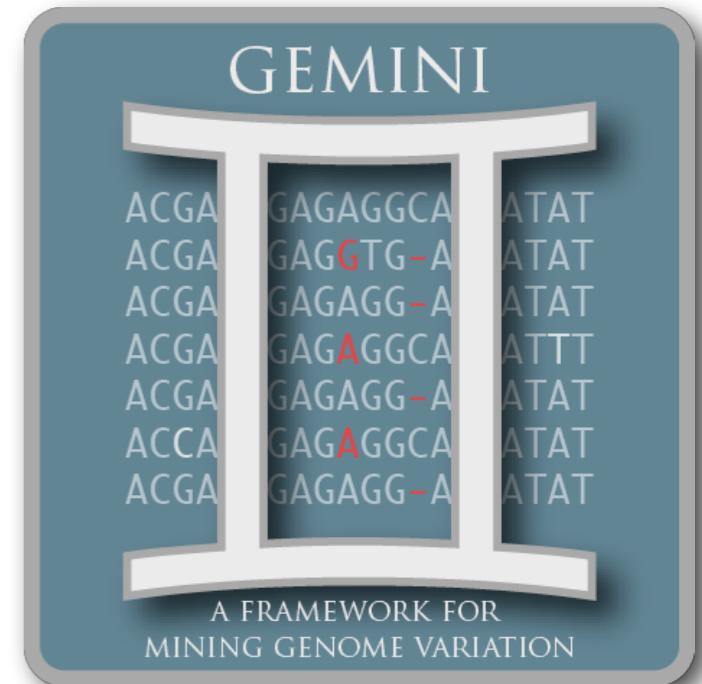
# What is GEMINI?

## Software package for exploring genetic variation

- Integrates annotations from many different sources (ClinVar, dbSNP, ENCODE, UCSC, 1000 Genomes, ESP, KEGG, etc.)

## What can you do with Gemini?

- Load a VCF into an “easy to use” database
- Query (fetch data) from database based on annotations or subject genotypes
- Analyze simple genetic models
- More advanced pathway, protein-protein interaction analyses



[github.com/arq5x/gemini](https://github.com/arq5x/gemini)



Uma Paila

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

**GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations**

Umadevi Paila<sup>1</sup>, Brad A. Chapman<sup>2</sup>, Rory K. Wilson<sup>3</sup>, Brent Pedersen<sup>4</sup>, Daniel J. Murphy<sup>5</sup>, Michael A. Tabor<sup>6</sup>, Daniel R. Goldstein<sup>7</sup>, Daniel E. Haussler<sup>8</sup>, Richard M. Hardison<sup>9</sup>, Mark A. Gerold<sup>10</sup>, Michael A. Patti<sup>11</sup>, Michael S. Gershoff<sup>12</sup>, Michael A. Hinds<sup>13</sup>, Mark D. Sherry<sup>14</sup>, Daniel J. Schadt<sup>15</sup>, David C. Schwartz<sup>16</sup>, Michael A. Flicek<sup>17</sup>, Mark A. Gerold<sup>18</sup>, Michael A. Patti<sup>19</sup>, Michael S. Gershoff<sup>20</sup>, Michael A. Hinds<sup>21</sup>, Mark D. Sherry<sup>22</sup>, Daniel J. Schadt<sup>23</sup>, David C. Schwartz<sup>24</sup>, Michael C. Schatz<sup>25</sup>, Michael A. Flicek<sup>26</sup>, Mark A. Gerold<sup>27</sup>, Michael A. Patti<sup>28</sup>, Michael S. Gershoff<sup>29</sup>, Michael A. Hinds<sup>30</sup>, Mark D. Sherry<sup>31</sup>, Daniel J. Schadt<sup>32</sup>, David C. Schwartz<sup>33</sup>, Michael C. Schatz<sup>34</sup>, Michael A. Flicek<sup>35</sup>, Mark A. Gerold<sup>36</sup>, Michael A. Patti<sup>37</sup>, Michael S. Gershoff<sup>38</sup>, Michael A. Hinds<sup>39</sup>, Mark D. Sherry<sup>40</sup>, Daniel J. Schadt<sup>41</sup>, David C. Schwartz<sup>42</sup>, Michael C. Schatz<sup>43</sup>, Michael A. Flicek<sup>44</sup>, Mark A. Gerold<sup>45</sup>, Michael A. Patti<sup>46</sup>, Michael S. Gershoff<sup>47</sup>, Michael A. Hinds<sup>48</sup>, Mark D. Sherry<sup>49</sup>, Daniel J. Schadt<sup>50</sup>, David C. Schwartz<sup>51</sup>, Michael C. Schatz<sup>52</sup>, Michael A. Flicek<sup>53</sup>, Mark A. Gerold<sup>54</sup>, Michael A. Patti<sup>55</sup>, Michael S. Gershoff<sup>56</sup>, Michael A. Hinds<sup>57</sup>, Mark D. Sherry<sup>58</sup>, Daniel J. Schadt<sup>59</sup>, David C. Schwartz<sup>60</sup>, Michael C. Schatz<sup>61</sup>, Michael A. Flicek<sup>62</sup>, Mark A. Gerold<sup>63</sup>, Michael A. Patti<sup>64</sup>, Michael S. Gershoff<sup>65</sup>, Michael A. Hinds<sup>66</sup>, Mark D. Sherry<sup>67</sup>, Daniel J. Schadt<sup>68</sup>, David C. Schwartz<sup>69</sup>, Michael C. Schatz<sup>70</sup>, Michael A. Flicek<sup>71</sup>, Mark A. Gerold<sup>72</sup>, Michael A. Patti<sup>73</sup>, Michael S. Gershoff<sup>74</sup>, Michael A. Hinds<sup>75</sup>, Mark D. Sherry<sup>76</sup>, Daniel J. Schadt<sup>77</sup>, David C. Schwartz<sup>78</sup>, Michael C. Schatz<sup>79</sup>, Michael A. Flicek<sup>80</sup>, Mark A. Gerold<sup>81</sup>, Michael A. Patti<sup>82</sup>, Michael S. Gershoff<sup>83</sup>, Michael A. Hinds<sup>84</sup>, Mark D. Sherry<sup>85</sup>, Daniel J. Schadt<sup>86</sup>, David C. Schwartz<sup>87</sup>, Michael C. Schatz<sup>88</sup>, Michael A. Flicek<sup>89</sup>, Mark A. Gerold<sup>90</sup>, Michael A. Patti<sup>91</sup>, Michael S. Gershoff<sup>92</sup>, Michael A. Hinds<sup>93</sup>, Mark D. Sherry<sup>94</sup>, Daniel J. Schadt<sup>95</sup>, David C. Schwartz<sup>96</sup>, Michael C. Schatz<sup>97</sup>, Michael A. Flicek<sup>98</sup>, Mark A. Gerold<sup>99</sup>, Michael A. Patti<sup>100</sup>, Michael S. Gershoff<sup>101</sup>, Michael A. Hinds<sup>102</sup>, Mark D. Sherry<sup>103</sup>, Daniel J. Schadt<sup>104</sup>, David C. Schwartz<sup>105</sup>, Michael C. Schatz<sup>106</sup>, Michael A. Flicek<sup>107</sup>, Mark A. Gerold<sup>108</sup>, Michael A. Patti<sup>109</sup>, Michael S. Gershoff<sup>110</sup>, Michael A. Hinds<sup>111</sup>, Mark D. Sherry<sup>112</sup>, Daniel J. Schadt<sup>113</sup>, David C. Schwartz<sup>114</sup>, Michael C. Schatz<sup>115</sup>, Michael A. Flicek<sup>116</sup>, Mark A. Gerold<sup>117</sup>, Michael A. Patti<sup>118</sup>, Michael S. Gershoff<sup>119</sup>, Michael A. Hinds<sup>120</sup>, Mark D. Sherry<sup>121</sup>, Daniel J. Schadt<sup>122</sup>, David C. Schwartz<sup>123</sup>, Michael C. Schatz<sup>124</sup>, Michael A. Flicek<sup>125</sup>, Mark A. Gerold<sup>126</sup>, Michael A. Patti<sup>127</sup>, Michael S. Gershoff<sup>128</sup>, Michael A. Hinds<sup>129</sup>, Mark D. Sherry<sup>130</sup>, Daniel J. Schadt<sup>131</sup>, David C. Schwartz<sup>132</sup>, Michael C. Schatz<sup>133</sup>, Michael A. Flicek<sup>134</sup>, Mark A. Gerold<sup>135</sup>, Michael A. Patti<sup>136</sup>, Michael S. Gershoff<sup>137</sup>, Michael A. Hinds<sup>138</sup>, Mark D. Sherry<sup>139</sup>, Daniel J. Schadt<sup>140</sup>, David C. Schwartz<sup>141</sup>, Michael C. Schatz<sup>142</sup>, Michael A. Flicek<sup>143</sup>, Mark A. Gerold<sup>144</sup>, Michael A. Patti<sup>145</sup>, Michael S. Gershoff<sup>146</sup>, Michael A. Hinds<sup>147</sup>, Mark D. Sherry<sup>148</sup>, Daniel J. Schadt<sup>149</sup>, David C. Schwartz<sup>150</sup>, Michael C. Schatz<sup>151</sup>, Michael A. Flicek<sup>152</sup>, Mark A. Gerold<sup>153</sup>, Michael A. Patti<sup>154</sup>, Michael S. Gershoff<sup>155</sup>, Michael A. Hinds<sup>156</sup>, Mark D. Sherry<sup>157</sup>, Daniel J. Schadt<sup>158</sup>, David C. Schwartz<sup>159</sup>, Michael C. Schatz<sup>160</sup>, Michael A. Flicek<sup>161</sup>, Mark A. Gerold<sup>162</sup>, Michael A. Patti<sup>163</sup>, Michael S. Gershoff<sup>164</sup>, Michael A. Hinds<sup>165</sup>, Mark D. Sherry<sup>166</sup>, Daniel J. Schadt<sup>167</sup>, David C. Schwartz<sup>168</sup>, Michael C. Schatz<sup>169</sup>, Michael A. Flicek<sup>170</sup>, Mark A. Gerold<sup>171</sup>, Michael A. Patti<sup>172</sup>, Michael S. Gershoff<sup>173</sup>, Michael A. Hinds<sup>174</sup>, Mark D. Sherry<sup>175</sup>, Daniel J. Schadt<sup>176</sup>, David C. Schwartz<sup>177</sup>, Michael C. Schatz<sup>178</sup>, Michael A. Flicek<sup>179</sup>, Mark A. Gerold<sup>180</sup>, Michael A. Patti<sup>181</sup>, Michael S. Gershoff<sup>182</sup>, Michael A. Hinds<sup>183</sup>, Mark D. Sherry<sup>184</sup>, Daniel J. Schadt<sup>185</sup>, David C. Schwartz<sup>186</sup>, Michael C. Schatz<sup>187</sup>, Michael A. Flicek<sup>188</sup>, Mark A. Gerold<sup>189</sup>, Michael A. Patti<sup>190</sup>, Michael S. Gershoff<sup>191</sup>, Michael A. Hinds<sup>192</sup>, Mark D. Sherry<sup>193</sup>, Daniel J. Schadt<sup>194</sup>, David C. Schwartz<sup>195</sup>, Michael C. Schatz<sup>196</sup>, Michael A. Flicek<sup>197</sup>, Mark A. Gerold<sup>198</sup>, Michael A. Patti<sup>199</sup>, Michael S. Gershoff<sup>200</sup>, Michael A. Hinds<sup>201</sup>, Mark D. Sherry<sup>202</sup>, Daniel J. Schadt<sup>203</sup>, David C. Schwartz<sup>204</sup>, Michael C. Schatz<sup>205</sup>, Michael A. Flicek<sup>206</sup>, Mark A. Gerold<sup>207</sup>, Michael A. Patti<sup>208</sup>, Michael S. Gershoff<sup>209</sup>, Michael A. Hinds<sup>210</sup>, Mark D. Sherry<sup>211</sup>, Daniel J. Schadt<sup>212</sup>, David C. Schwartz<sup>213</sup>, Michael C. Schatz<sup>214</sup>, Michael A. Flicek<sup>215</sup>, Mark A. Gerold<sup>216</sup>, Michael A. Patti<sup>217</sup>, Michael S. Gershoff<sup>218</sup>, Michael A. Hinds<sup>219</sup>, Mark D. Sherry<sup>220</sup>, Daniel J. Schadt<sup>221</sup>, David C. Schwartz<sup>222</sup>, Michael C. Schatz<sup>223</sup>, Michael A. Flicek<sup>224</sup>, Mark A. Gerold<sup>225</sup>, Michael A. Patti<sup>226</sup>, Michael S. Gershoff<sup>227</sup>, Michael A. Hinds<sup>228</sup>, Mark D. Sherry<sup>229</sup>, Daniel J. Schadt<sup>230</sup>, David C. Schwartz<sup>231</sup>, Michael C. Schatz<sup>232</sup>, Michael A. Flicek<sup>233</sup>, Mark A. Gerold<sup>234</sup>, Michael A. Patti<sup>235</sup>, Michael S. Gershoff<sup>236</sup>, Michael A. Hinds<sup>237</sup>, Mark D. Sherry<sup>238</sup>, Daniel J. Schadt<sup>239</sup>, David C. Schwartz<sup>240</sup>, Michael C. Schatz<sup>241</sup>, Michael A. Flicek<sup>242</sup>, Mark A. Gerold<sup>243</sup>, Michael A. Patti<sup>244</sup>, Michael S. Gershoff<sup>245</sup>, Michael A. Hinds<sup>246</sup>, Mark D. Sherry<sup>247</sup>, Daniel J. Schadt<sup>248</sup>, David C. Schwartz<sup>249</sup>, Michael C. Schatz<sup>250</sup>, Michael A. Flicek<sup>251</sup>, Mark A. Gerold<sup>252</sup>, Michael A. Patti<sup>253</sup>, Michael S. Gershoff<sup>254</sup>, Michael A. Hinds<sup>255</sup>, Mark D. Sherry<sup>256</sup>, Daniel J. Schadt<sup>257</sup>, David C. Schwartz<sup>258</sup>, Michael C. Schatz<sup>259</sup>, Michael A. Flicek<sup>260</sup>, Mark A. Gerold<sup>261</sup>, Michael A. Patti<sup>262</sup>, Michael S. Gershoff<sup>263</sup>, Michael A. Hinds<sup>264</sup>, Mark D. Sherry<sup>265</sup>, Daniel J. Schadt<sup>266</sup>, David C. Schwartz<sup>267</sup>, Michael C. Schatz<sup>268</sup>, Michael A. Flicek<sup>269</sup>, Mark A. Gerold<sup>270</sup>, Michael A. Patti<sup>271</sup>, Michael S. Gershoff<sup>272</sup>, Michael A. Hinds<sup>273</sup>, Mark D. Sherry<sup>274</sup>, Daniel J. Schadt<sup>275</sup>, David C. Schwartz<sup>276</sup>, Michael C. Schatz<sup>277</sup>, Michael A. Flicek<sup>278</sup>, Mark A. Gerold<sup>279</sup>, Michael A. Patti<sup>280</sup>, Michael S. Gershoff<sup>281</sup>, Michael A. Hinds<sup>282</sup>, Mark D. Sherry<sup>283</sup>, Daniel J. Schadt<sup>284</sup>, David C. Schwartz<sup>285</sup>, Michael C. Schatz<sup>286</sup>, Michael A. Flicek<sup>287</sup>, Mark A. Gerold<sup>288</sup>, Michael A. Patti<sup>289</sup>, Michael S. Gershoff<sup>290</sup>, Michael A. Hinds<sup>291</sup>, Mark D. Sherry<sup>292</sup>, Daniel J. Schadt<sup>293</sup>, David C. Schwartz<sup>294</sup>, Michael C. Schatz<sup>295</sup>, Michael A. Flicek<sup>296</sup>, Mark A. Gerold<sup>297</sup>, Michael A. Patti<sup>298</sup>, Michael S. Gershoff<sup>299</sup>, Michael A. Hinds<sup>300</sup>, Mark D. Sherry<sup>301</sup>, Daniel J. Schadt<sup>302</sup>, David C. Schwartz<sup>303</sup>, Michael C. Schatz<sup>304</sup>, Michael A. Flicek<sup>305</sup>, Mark A. Gerold<sup>306</sup>, Michael A. Patti<sup>307</sup>, Michael S. Gershoff<sup>308</sup>, Michael A. Hinds<sup>309</sup>, Mark D. Sherry<sup>310</sup>, Daniel J. Schadt<sup>311</sup>, David C. Schwartz<sup>312</sup>, Michael C. Schatz<sup>313</sup>, Michael A. Flicek<sup>314</sup>, Mark A. Gerold<sup>315</sup>, Michael A. Patti<sup>316</sup>, Michael S. Gershoff<sup>317</sup>, Michael A. Hinds<sup>318</sup>, Mark D. Sherry<sup>319</sup>, Daniel J. Schadt<sup>320</sup>, David C. Schwartz<sup>321</sup>, Michael C. Schatz<sup>322</sup>, Michael A. Flicek<sup>324</sup>, Mark A. Gerold<sup>325</sup>, Michael A. Patti<sup>326</sup>, Michael S. Gershoff<sup>327</sup>, Michael A. Hinds<sup>328</sup>, Mark D. Sherry<sup>329</sup>, Daniel J. Schadt<sup>330</sup>, David C. Schwartz<sup>331</sup>, Michael C. Schatz<sup>332</sup>, Michael A. Flicek<sup>334</sup>, Mark A. Gerold<sup>335</sup>, Michael A. Patti<sup>336</sup>, Michael S. Gershoff<sup>337</sup>, Michael A. Hinds<sup>338</sup>, Mark D. Sherry<sup>339</sup>, Daniel J. Schadt<sup>340</sup>, David C. Schwartz<sup>341</sup>, Michael C. Schatz<sup>342</sup>, Michael A. Flicek<sup>344</sup>, Mark A. Gerold<sup>345</sup>, Michael A. Patti<sup>346</sup>, Michael S. Gershoff<sup>347</sup>, Michael A. Hinds<sup>348</sup>, Mark D. Sherry<sup>349</sup>, Daniel J. Schadt<sup>350</sup>, David C. Schwartz<sup>351</sup>, Michael C. Schatz<sup>352</sup>, Michael A. Flicek<sup>354</sup>, Mark A. Gerold<sup>355</sup>, Michael A. Patti<sup>356</sup>, Michael S. Gershoff<sup>357</sup>, Michael A. Hinds<sup>358</sup>, Mark D. Sherry<sup>359</sup>, Daniel J. Schadt<sup>360</sup>, David C. Schwartz<sup>361</sup>, Michael C. Schatz<sup>362</sup>, Michael A. Flicek<sup>364</sup>, Mark A. Gerold<sup>365</sup>, Michael A. Patti<sup>366</sup>, Michael S. Gershoff<sup>367</sup>, Michael A. Hinds<sup>368</sup>, Mark D. Sherry<sup>369</sup>, Daniel J. Schadt<sup>370</sup>, David C. Schwartz<sup>371</sup>, Michael C. Schatz<sup>372</sup>, Michael A. Flicek<sup>374</sup>, Mark A. Gerold<sup>375</sup>, Michael A. Patti<sup>376</sup>, Michael S. Gershoff<sup>377</sup>, Michael A. Hinds<sup>378</sup>, Mark D. Sherry<sup>379</sup>, Daniel J. Schadt<sup>380</sup>, David C. Schwartz<sup>381</sup>, Michael C. Schatz<sup>382</sup>, Michael A. Flicek<sup>384</sup>, Mark A. Gerold<sup>385</sup>, Michael A. Patti<sup>386</sup>, Michael S. Gershoff<sup>387</sup>, Michael A. Hinds<sup>388</sup>, Mark D. Sherry<sup>389</sup>, Daniel J. Schadt<sup>390</sup>, David C. Schwartz<sup>391</sup>, Michael C. Schatz<sup>392</sup>, Michael A. Flicek<sup>394</sup>, Mark A. Gerold<sup>395</sup>, Michael A. Patti<sup>396</sup>, Michael S. Gershoff<sup>397</sup>, Michael A. Hinds<sup>398</sup>, Mark D. Sherry<sup>399</sup>, Daniel J. Schadt<sup>400</sup>, David C. Schwartz<sup>401</sup>, Michael C. Schatz<sup>402</sup>, Michael A. Flicek<sup>404</sup>, Mark A. Gerold<sup>405</sup>, Michael A. Patti<sup>406</sup>, Michael S. Gershoff<sup>407</sup>, Michael A. Hinds<sup>408</sup>, Mark D. Sherry<sup>409</sup>, Daniel J. Schadt<sup>410</sup>, David C. Schwartz<sup>411</sup>, Michael C. Schatz<sup>412</sup>, Michael A. Flicek<sup>414</sup>, Mark A. Gerold<sup>415</sup>, Michael A. Patti<sup>416</sup>, Michael S. Gershoff<sup>417</sup>, Michael A. Hinds<sup>418</sup>, Mark D. Sherry<sup>419</sup>, Daniel J. Schadt<sup>420</sup>, David C. Schwartz<sup>421</sup>, Michael C. Schatz<sup>422</sup>, Michael A. Flicek<sup>424</sup>, Mark A. Gerold<sup>425</sup>, Michael A. Patti<sup>426</sup>, Michael S. Gershoff<sup>427</sup>, Michael A. Hinds<sup>428</sup>, Mark D. Sherry<sup>429</sup>, Daniel J. Schadt<sup>430</sup>, David C. Schwartz<sup>431</sup>, Michael C. Schatz<sup>432</sup>, Michael A. Flicek<sup>434</sup>, Mark A. Gerold<sup>435</sup>, Michael A. Patti<sup>436</sup>, Michael S. Gershoff<sup>437</sup>, Michael A. Hinds<sup>438</sup>, Mark D. Sherry<sup>439</sup>, Daniel J. Schadt<sup>440</sup>, David C. Schwartz<sup>441</sup>, Michael C. Schatz<sup>442</sup>, Michael A. Flicek<sup>444</sup>, Mark A. Gerold<sup>445</sup>, Michael A. Patti<sup>446</sup>, Michael S. Gershoff<sup>447</sup>, Michael A. Hinds<sup>448</sup>, Mark D. Sherry<sup>449</sup>, Daniel J. Schadt<sup>450</sup>, David C. Schwartz<sup>451</sup>, Michael C. Schatz<sup>452</sup>, Michael A. Flicek<sup>454</sup>, Mark A. Gerold<sup>455</sup>, Michael A. Patti<sup>456</sup>, Michael S. Gershoff<sup>457</sup>, Michael A. Hinds<sup>458</sup>, Mark D. Sherry<sup>459</sup>, Daniel J. Schadt<sup>460</sup>, David C. Schwartz<sup>461</sup>, Michael C. Schatz<sup>462</sup>, Michael A. Flicek<sup>464</sup>, Mark A. Gerold<sup>465</sup>, Michael A. Patti<sup>466</sup>, Michael S. Gershoff<sup>467</sup>, Michael A. Hinds<sup>468</sup>, Mark D. Sherry<sup>469</sup>, Daniel J. Schadt<sup>470</sup>, David C. Schwartz<sup>471</sup>, Michael C. Schatz<sup>472</sup>, Michael A. Flicek<sup>474</sup>, Mark A. Gerold<sup>475</sup>, Michael A. Patti<sup>476</sup>, Michael S. Gershoff<sup>477</sup>, Michael A. Hinds<sup>478</sup>, Mark D. Sherry<sup>479</sup>, Daniel J. Schadt<sup>480</sup>, David C. Schwartz<sup>481</sup>, Michael C. Schatz<sup>482</sup>, Michael A. Flicek<sup>484</sup>, Mark A. Gerold<sup>485</sup>, Michael A. Patti<sup>486</sup>, Michael S. Gershoff<sup>487</sup>, Michael A. Hinds<sup>488</sup>, Mark D. Sherry<sup>489</sup>, Daniel J. Schadt<sup>490</sup>, David C. Schwartz<sup>491</sup>, Michael C. Schatz<sup>492</sup>, Michael A. Flicek<sup>494</sup>, Mark A. Gerold<sup>495</sup>, Michael A. Patti<sup>496</sup>, Michael S. Gershoff<sup>497</sup>, Michael A. Hinds<sup>498</sup>, Mark D. Sherry<sup>499</sup>, Daniel J. Schadt<sup>500</sup>, David C. Schwartz<sup>501</sup>, Michael C. Schatz<sup>502</sup>, Michael A. Flicek<sup>504</sup>, Mark A. Gerold<sup>505</sup>, Michael A. Patti<sup>506</sup>, Michael S. Gershoff<sup>507</sup>, Michael A. Hinds<sup>508</sup>, Mark D. Sherry<sup>509</sup>, Daniel J. Schadt<sup>510</sup>, David C. Schwartz<sup>511</sup>, Michael C. Schatz<sup>512</sup>, Michael A. Flicek<sup>514</sup>, Mark A. Gerold<sup>515</sup>, Michael A. Patti<sup>516</sup>, Michael S. Gershoff<sup>517</sup>, Michael A. Hinds<sup>518</sup>, Mark D. Sherry<sup>519</sup>, Daniel J. Schadt<sup>520</sup>, David C. Schwartz<sup>521</sup>, Michael C. Schatz<sup>522</sup>, Michael A. Flicek<sup>524</sup>, Mark A. Gerold<sup>525</sup>, Michael A. Patti<sup>526</sup>, Michael S. Gershoff<sup>527</sup>, Michael A. Hinds<sup>528</sup>, Mark D. Sherry<sup>529</sup>, Daniel J. Schadt<sup>530</sup>, David C. Schwartz<sup>531</sup>, Michael C. Schatz<sup>532</sup>, Michael A. Flicek<sup>534</sup>, Mark A. Gerold<sup>535</sup>, Michael A. Patti<sup>536</sup>, Michael S. Gershoff<sup>537</sup>, Michael A. Hinds<sup>538</sup>, Mark D. Sherry<sup>539</sup>, Daniel J. Schadt<sup>540</sup>, David C. Schwartz<sup>541</sup>, Michael C. Schatz<sup>542</sup>, Michael A. Flicek<sup>544</sup>, Mark A. Gerold<sup>545</sup>, Michael A. Patti<sup>546</sup>, Michael S. Gershoff<sup>547</sup>, Michael A. Hinds<sup>548</sup>, Mark D. Sherry<sup>549</sup>, Daniel J. Schadt<sup>550</sup>, David C. Schwartz<sup>551</sup>, Michael C. Schatz<sup>552</sup>, Michael A. Flicek<sup>554</sup>, Mark A. Gerold<sup>555</sup>, Michael A. Patti<sup>556</sup>, Michael S. Gershoff<sup>557</sup>, Michael A. Hinds<sup>558</sup>, Mark D. Sherry<sup>559</sup>, Daniel J. Schadt<sup>560</sup>, David C. Schwartz<sup>561</sup>, Michael C. Schatz<sup>562</sup>, Michael A. Flicek<sup>564</sup>, Mark A. Gerold<sup>565</sup>, Michael A. Patti<sup>566</sup>, Michael S. Gershoff<sup>567</sup>, Michael A. Hinds<sup>568</sup>, Mark D. Sherry<sup>569</sup>, Daniel J. Schadt<sup>570</sup>, David C. Schwartz<sup>571</sup>, Michael C. Schatz<sup>572</sup>, Michael A. Flicek<sup>574</sup>, Mark A. Gerold<sup>575</sup>, Michael A. Patti<sup>576</sup>, Michael S. Gershoff<sup>577</sup>, Michael A. Hinds<sup>578</sup>, Mark D. Sherry<sup>579</sup>, Daniel J. Schadt<sup>580</sup>, David C. Schwartz<sup>581</sup>, Michael C. Schatz<sup>582</sup>, Michael A. Flicek<sup>584</sup>, Mark A. Gerold<sup>585</sup>, Michael A. Patti<sup>586</sup>, Michael S. Gershoff<sup>587</sup>, Michael A. Hinds<sup>588</sup>, Mark D. Sherry<sup>589</sup>, Daniel J. Schadt<sup>590</sup>, David C. Schwartz<sup>591</sup>, Michael C. Schatz<sup>592</sup>, Michael A. Flicek<sup>594</sup>, Mark A. Gerold<sup>595</sup>, Michael A. Patti<sup>596</sup>, Michael S. Gershoff<sup>597</sup>, Michael A. Hinds<sup>598</sup>, Mark D. Sherry<sup>599</sup>, Daniel J. Schadt<sup>600</sup>, David C. Schwartz<sup>601</sup>, Michael C. Schatz<

# What is GEMINI?

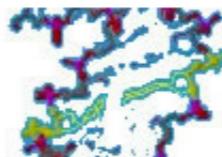


Conservation  
Repeat elements  
Genome Gaps  
Cytobands  
Gene annotations  
“Mappability”  
DeCIPHER  
ISGA

Pfam



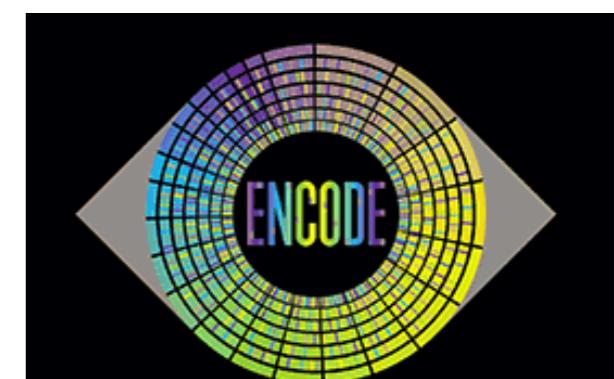
**dbSNP**  
Short Genetic Variations



**ClinVar**

**OMIM**  
Online Mendelian Inheritance in Man

**1000 Genomes**  
A Deep Catalog of Human Genetic Variation



**Genetic variation**

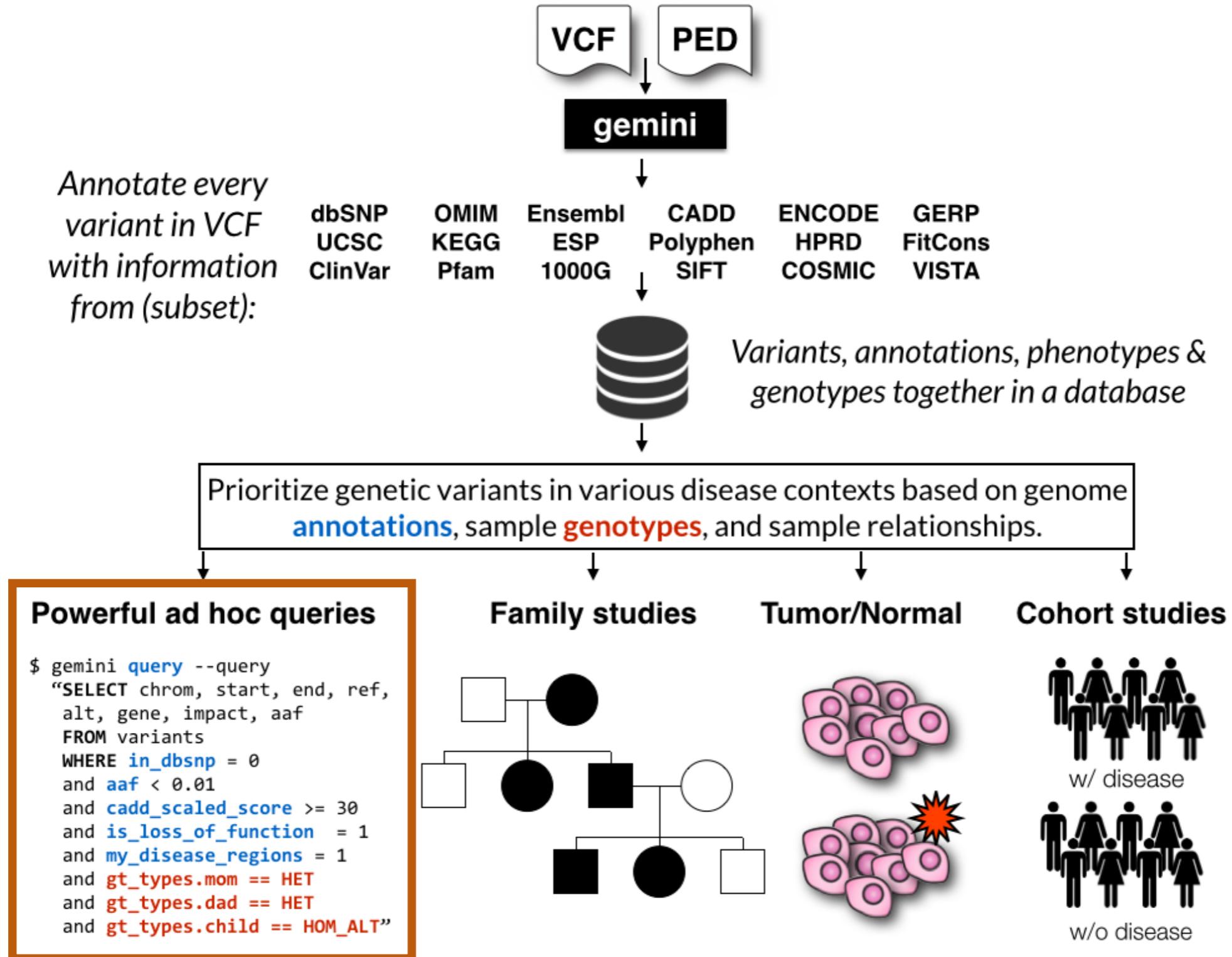
...CCTCATG**C**ATGGAAA...  
...CCTCATG**T**ATGGAAA...  
...CCTCATG**C**ATGGAAA...  
...CCTCATG**C**ATGGAAA...  
...CCTCATG**T**ATGGAAA...  
...CCTCATG**C**ATGGAAA...  
...CCTCATG**T**ATGGAAA...



**Human Protein Reference Database**

Chromatin marks  
DNA methylation  
RNA expression  
TF binding

# GEMINI Framework uses databases and SQL



The screenshot shows a web browser window displaying the GEMINI documentation at <http://gemini.readthedocs.org/en/latest/#>. The page title is "gemini v0.10.0a". The main content area features a large image of the GEMINI logo, which includes a stylized DNA helix and the text "GEMINI A FRAMEWORK FOR MINING GENOME VARIATION". Below the logo, a section titled "GEMINI is a flexible framework for exploring genome variation" is present. To the left of the main content, there are sections for "GEMINI links" (Issue Tracker, Source @ GitHub, Mailing list @ Google Groups, Quinlan lab @ UVa) and "Sources" (Browse source @ GitHub). A yellow callout box titled "This Page" provides instructions for editing the document on GitHub. The right side of the page contains a detailed overview of GEMINI's purpose and functionality, a note about its limitations, and a citation for the original manuscript. A "Table of contents" sidebar is visible on the far left.

**GEMINI: a flexible framework for exploring genome variation**

## Overview

GEMINI (GEnome MINIng) is designed to be a flexible framework for exploring genetic variation in the context of the wealth of genome annotations available for the human genome. By placing genetic variants, sample genotypes, and useful genome annotations into an integrated database framework, GEMINI provides a simple, flexible, yet very powerful system for exploring genetic variation for disease and population genetics.

Using the GEMINI framework begins by loading a VCF file into a database. Each variant is automatically annotated by comparing it to several genome annotations from sources such as ENCODE tracks, UCSC tracks, OMIM, dbSNP, KEGG, and HPRD. All of this information is stored in portable SQLite databases that allow one to explore and interpret both coding and non-coding variation using “off-the-shelf” tools or an enhanced SQL engine.

Please also see the [original manuscript](#).

This [video](#) provides more details about GEMINI’s aims and utility.

**Note**

1. GEMINI solely supports human genetic variation mapped to build 37 (aka hg19) of the human genome.
2. GEMINI is very strict about adherence to VCF format 4.1.
3. For best performance, load and query GEMINI databases on the fastest hard drive to which you have access.

## Citation

If you use GEMINI in your research, please cite the following manuscript:

Paila U, Chapman BA, Kirchner R, Quinlan AR (2013)  
GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations  
PLoS Comput Biol 9(7): e1003153. doi:10.1371/journal.pcbi.1003153

## Table of contents

# Setup GEMINI exercise

---

```
# login to AWS

$ mkdir wed

$ cd wed

$ mkdir mydata

$ cd mydata

$ curl https://s3.amazonaws.com/gemini-tutorials/learnSQL.db
> learnSQL.db

$ curl https://s3.amazonaws.com/gemini-tutorials/learnSQL2.db
> learnSQL2.db

$ curl https://s3.amazonaws.com/gemini-tutorials/
chr22.VEP.vcf > chr22.VEP.vcf

$ curl https://s3.amazonaws.com/gemini-tutorials/trio.ped >
trio.ped
```

# GEMINI uses a database. Uses SQL to “talk” to it.

- SQL databases to organize genotypes and annotations
- SQL = Structured Query Language
- Data stored in tables

samples							
sample_id	name	family_id	paternal_id	maternal_id	sex	phenotype	ethnicity
1	John	1	Bob	Sue	1	2	CEU
2	Bob	1	0	0	1	1	CEU
3	Mary	1	0	0	2	1	CEU
4	Sue	2	0	0	2	2	YRI

- Query (ask the database) to report data matching your requirements
- Can combine queries, act on results from previous query

# GEMINI **query** allows one to use SQL to query

samples								
sample_id	name	family_id	paternal_id	maternal_id	sex	phenotype	ethnicity	
1	John	1	Bob	Sue	1	2	CEU	
2	Bob	1	0	0	1	1	CEU	
3	Mary	1	0	0	2	1	CEU	
4	Sue	2	0	0	2	2	YRI	



```
gemini query -q "SELECT name FROM samples" learnSQL.db
```



John
Bob
Mary
Sue

# Filtering rows with the WHERE (==) clause

samples								
sample_id	name	family_id	paternal_id	maternal_id	sex	phenotype	ethnicity	
1	John	1	Bob	Sue	1	2	CEU	→
2	Bob	1	0	0	1	1	CEU	
3	Mary	1	0	0	2	1	CEU	
4	Sue	2	0	0	2	2	YRI	→

gemini query -q "SELECT name FROM samples WHERE phenotype == 2"  
learnSQL.db



- phenotype == 2 gets rows where the phenotype value is equal to 2

# Filtering rows with the WHERE (<>) clause

samples								
sample_id	name	family_id	paternal_id	maternal_id	sex	phenotype	ethnicity	
1	John	1	Bob	Sue	1	2	CEU	
2	Bob	1	0	0	1	1	CEU	
3	Mary	1	0	0	2	1	CEU	
4	Sue	2	0	0	2	2	YRI	

→ → |

```
gemini query -q "SELECT name FROM samples WHERE phenotype <> 2"  
learnSQL.db
```



- `phenotype <> 2` gets rows where the phenotype value is NOT equal to 2

# Filtering rows with the WHERE (<) clause

samples								
sample_id	name	family_id	paternal_id	maternal_id	sex	phenotype	ethnicity	
1	John	1	Bob	Sue	1	2	CEU	→
2	Bob	1	0	0	1	1	CEU	→
3	Mary	1	0	0	2	1	CEU	
4	Sue	2	0	0	2	2	YRI	



```
gemini query -q "SELECT name FROM samples WHERE sample_id < 3"  
learnSQL.db
```



- Many numeric comparisons possible:  $\leq$ ,  $\geq$ ,  $>$ ,  $<$

# Filtering rows with the WHERE (NULL) clause

samples								
sample_id	name	family_id	paternal_id	maternal_id	sex	phenotype	ethnicity	
1	John	1	Bob	Sue	1	2	CEU	
2	Bob	1	0	0	1	1	CEU	
3	Mary	1	0	0	2	1	CEU	
4	Sue	2	0	0	2	2	YRI	
5	Greg	3	0	0	1	2		

```
gemini query -q "SELECT name FROM samples WHERE ethnicity IS NULL" learnSQL2.db
```

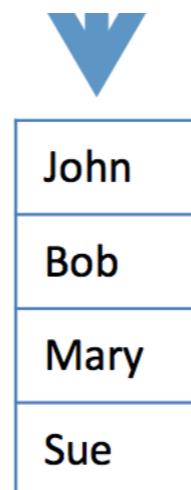


- When value is missing (empty), it is "NULL"

# Filtering rows with the WHERE (NOT NULL) clause

samples								
sample_id	name	family_id	paternal_id	maternal_id	sex	phenotype	ethnicity	
1	John	1	Bob	Sue	1	2	CEU	
2	Bob	1	0	0	1	1	CEU	
3	Mary	1	0	0	2	1	CEU	
4	Sue	2	0	0	2	2	YRI	
5	Greg	3	0	0	1	2		

```
gemini query -q "SELECT name FROM samples WHERE ethnicity IS NOT NULL" learnSQL2.db
```



# The \* wildcard

fakevariants				
chrom	start	end	rsid	in_dbsnp
chr1	1	2	rs123	1
chr1	3	4		0
chr2	1	2	rs456	1
chr2	3	4		0



```
gemini query -q "SELECT * FROM fakevariants" learnSQL2.db
```



chr1	1	2	rs123	1
chr1	3	4		0
chr2	1	2	rs456	1
chr2	3	4		0

# Booleans

---

- Boolean values are TRUE or FALSE
  - TRUE is equivalent to a value of 1
  - FALSE is equivalent to value of 0

fakevariants				
chrom	start	end	rsid	in_dbsnp
chr1	1	2	rs123	1
chr1	3	4		0
chr2	1	2	rs456	1
chr2	3	4		0

# Booleans (True)

fakevariants				
chrom	start	end	rsid	in_dbsnp
chr1	1	2	rs123	1
chr1	3	4		0
chr2	1	2	rs456	1
chr2	3	4		0

```
gemini query -q "SELECT chrom,start,end FROM fakevariants  
WHERE in_dbsnp == 1" learnSQL2.db
```

**OR**

```
gemini query -q "SELECT chrom,start,end FROM fakevariants  
WHERE in_dbsnp" learnSQL2.db
```



chr1	1	2
chr2	1	2

# The COUNT() operation

fakevariants				
chrom	start	end	rsid	in_dbsnp
chr1	1	2	rs123	1
chr1	3	4		0
chr2	1	2	rs456	1
chr2	3	4		0

```
gemini query -q "SELECT COUNT(*) FROM fakevariants  
WHERE chrom == 'chr1' " learnSQL2.db
```



2

# Multiple WHERE clauses

fakevariants				
chrom	start	end	rsid	in_dbsnp
chr1	1	2	rs123	1
chr1	3	4		0
chr2	1	2	rs456	1
chr2	3	4		0

```
gemini query -q "SELECT COUNT(*) FROM fakevariants  
WHERE chrom == 'chr1'  
AND in_dbsnp == 0" learnSQL2.db
```



1

# Using GEMINI

# Annotating genetic variants in the VCF file.

---

- For gene-based annotations, Gemini supports SnpEff and Variant Effect Predictor (VEP)
- Can add custom annotations as well (e.g. SeattleSeq)
- We will use VEP in this session

VEP webpage: [http://uswest.ensembl.org/info/docs/variation/vep/vep\\_script.html](http://uswest.ensembl.org/info/docs/variation/vep/vep_script.html)

SeattleSeq: <http://snp.gs.washington.edu>

SNPEff: <http://snpeff.sourceforge.net/>

# Annotation with VEP (done for you)

```
#perl ~/software/variant_effect_predictor/variant_effect_predictor/
variant_effect_predictor.pl -i chr22.vcf -o chr22.VEP.vcf --vcf \
--cache --dir ~/software/variant_effect_predictor/references \
--sift b --polyphen b --symbol --numbers --biotype --total_length \
--fields
Consequence,Codons,Amino_acids,Gene,SYMBOL,Feature,EXON,PolyPhen,SIFT,Protein_position,BIOTYPE
```

<b>-i chr22.vcf -o chr22.VEP.vcf</b>	Provide names of input and output vcf's
<b>--vcf</b>	Create output in vcf format
<b>--cache --dir data/variant_effect_predictor/references</b>	Use a copy of the VEP database stored locally (faster)
<b>--compress "gunzip -c"</b>	Tells VEP how to uncompress vcf file
<b>--terms so</b>	Style for reporting functional consequence of variants
<b>--sift b --polyphen b</b>	Include <b>both</b> score and prediction by SIFT and Polyphen
<b>--hgnc</b>	Include HGNC gene name if available
<b>--numbers</b>	Include number of affected exon/intron
<b>--fields</b> <b>Consequence,Codons,Amino_acids,Gene,HGNC,Feature,</b> <b>EXON,PolyPhen,SIFT</b>	Fields to include in output



# Annotation with VEP

Before...

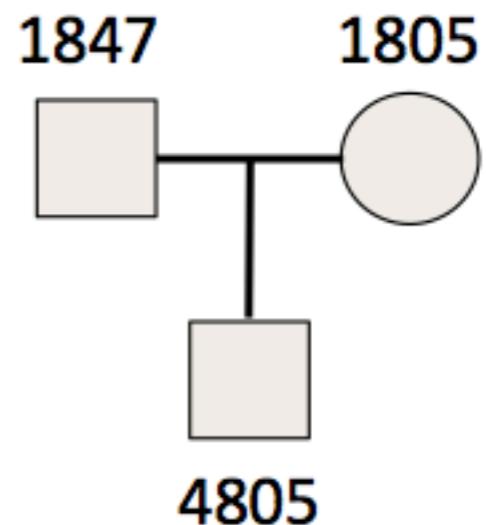
22	16157603	.	G	C	382.96	PASS	.	GT:AD:DP:GQ:PL	0/1:4,18:22:0.32:370,0,0	0/1:5,2:7:17.67:46,0,18	./.:.:.:.
22	16157635	.	G	A	15.67	LowQual;QDFilter;QUALFilter	.	GT:AD:DP:GQ:PL	0/0:5,5:10:3:0,3,27	0/1:7,5:12:20.61:47,0,21	./.:.:.:.
22	16157883	.	T	C	104.13	ABFilter;QDFilter	.	GT:AD:DP:GQ:PL	0/1:52,14:66:99:137,0,1177	0/0:88,3:91:99:0,182,2160	./.:.:.:.
22	16159060	.	G	A	11777.6	PASS	.	GT:AD:DP:GQ:PL	1/1:0,49:49:99:1865,144,0	1/1:0,87:87:99:3311,262,0	1/1:3,192:195:99:6602,490,0
22	16255645	.	T	C	235.22	ABFilter;QDFilter	.	GT:AD:DP:GQ:PL	0/1:30,6:36:20.52:155,0,21	0/1:27,7:34:99:119,0,182	0/0:246,1:250:99:0,183,2040

After...

22	16157603	.	G	C	382.96	PASS	CSQ= ENSG00000232775 ENST00000609679 Transcript intron_variant&nc_transcript_variant -/552      1/2  1 AP000525.10 Clone_based_vega_genel  processed_transcript, CIENSG00000206195 ENST00000437781 Transcript intron_variant&nc_transcript_variant -/2398      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, CIENSG00000272872 ENST00000608286 Transcript upstream_gene_variant -/694      28371-1 LL22NC03-N14H11.1 Clone_based_vega_genel  sense_intronic, CIENSG00000206195 ENST00000383038 Transcript downstream_gene_variant -/1394      11951-1 AP000525.9 Clone_based_vega_genel  processed_transcript, CIENSG00000206195 ENST00000447898 Transcript intron_variant&nc_transcript_variant -/4136      3/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, CIENSG00000206195 ENST00000607933 Transcript downstream_gene_variant -/642      12261-1 AP000525.9 Clone_based_vega_genel  processed_transcript, CIENSG00000232775 ENST00000440946 Transcript upstream_gene_variant -/160      44631 1 AP000525.10 Clone_based_vega_genel  processed_transcript, CIENSG00000206195 ENST00000413768 Transcript intron_variant&nc_transcript_variant -/2102      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA	GT:AD:DP:GQ:PL	0/1:4,18:22:0.32:370,0,0	0/1:5,2:7:17.67:46,0,18	./.:.:.:.
22	16157635	.	G	A	15.67	LowQual;QDFilter;QUALFilter	CSQ= ENSG00000232775 ENST00000609679 Transcript intron_variant&nc_transcript_variant -/552      1/2  1 AP000525.10 Clone_based_vega_genel  processed_transcript, AIENSG00000206195 ENST00000437781 Transcript intron_variant&nc_transcript_variant -/2398      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, AIENSG00000272872 ENST00000608286 Transcript upstream_gene_variant -/694      28691-1 LL22NC03-N14H11.1 Clone_based_vega_genel  sense_intronic, AIENSG00000206195 ENST00000383038 Transcript downstream_gene_variant -/1394      11631-1 AP000525.9 Clone_based_vega_genel  processed_transcript, AIENSG00000206195 ENST00000447898 Transcript intron_variant&nc_transcript_variant -/4136      3/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, AIENSG00000206195 ENST00000607933 Transcript downstream_gene_variant -/642      11941-1 AP000525.9 Clone_based_vega_genel  processed_transcript, AIENSG00000232775 ENST00000440946 Transcript upstream_gene_variant -/160      44311 1 AP000525.10 Clone_based_vega_genel  processed_transcript, AIENSG00000206195 ENST00000413768 Transcript intron_variant&nc_transcript_variant -/2102      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA	GT:AD:DP:GQ:PL	0/0:5,5:10:3:0,3,27	0/1:7,5:12:20.61:47,0,21	./.:.:.:.
22	16157883	.	T	C	104.13	ABFilter;QDFilter	CSQ= ENSG00000232775 ENST00000609679 Transcript intron_variant&nc_transcript_variant -/552      1/2  1 AP000525.10 Clone_based_vega_genel  processed_transcript, CIENSG00000206195 ENST00000437781 Transcript intron_variant&nc_transcript_variant -/2398      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, CIENSG00000272872 ENST00000608286 Transcript upstream_gene_variant -/694      31171-1 LL22NC03-N14H11.1 Clone_based_vega_genel  sense_intronic, CIENSG00000206195 ENST00000383038 Transcript downstream_gene_variant -/1394      19151-1 AP000525.9 Clone_based_vega_genel  processed_transcript, CIENSG00000206195 ENST00000447898 Transcript intron_variant&nc_transcript_variant -/4136      3/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, CIENSG00000206195 ENST00000607933 Transcript downstream_gene_variant -/642      9461-1 AP000525.9 Clone_based_vega_genel  processed_transcript, CIENSG00000232775 ENST00000440946 Transcript upstream_gene_variant -/160      41831 1 AP000525.10 Clone_based_vega_genel  processed_transcript, CIENSG00000206195 ENST00000413768 Transcript intron_variant&nc_transcript_variant -/2102      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA	GT:AD:DP:GQ:PL	0/1:52,14:66:99:137,0,1177	0/0:88,3:91:99:0,182,2160	./.:.:.:.
22	16159060	.	G	A	11777.6	PASS	CSQ= ENSG00000232775 ENST00000609679 Transcript intron_variant&nc_transcript_variant -/552      1/2  1 AP000525.10 Clone_based_vega_genel  processed_transcript, AIENSG00000206195 ENST00000437781 Transcript intron_variant&nc_transcript_variant -/2398      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, AIENSG00000272872 ENST00000608286 Transcript upstream_gene_variant -/694      42941-1 LL22NC03-N14H11.1 Clone_based_vega_genel  sense_intronic, AIENSG00000206195 ENST00000383038 Transcript non_coding_exon_transcript_variant 1132/1394      8/8111-1 AP000525.9 Clone_based_vega_genel  processed_transcript, AIENSG00000206195 ENST00000447898 Transcript intron_variant&nc_transcript_variant -/4136      3/  -1 AP000525.9 Clone_based_vega_genel  lincRNA, AIENSG00000206195 ENST00000607933 Transcript non_coding_exon_transcript_variant 1411/642      1/1  -1 AP000525.9 Clone_based_vega_genel  processed_transcript, AIENSG00000232775 ENST00000440946 Transcript upstream_gene_variant -/160      30061 1 AP000525.10 Clone_based_vega_genel  processed_transcript, AIENSG00000206195 ENST00000413768 Transcript intron_variant&nc_transcript_variant -/2102      7/  -1 AP000525.9 Clone_based_vega_genel  lincRNA	GT:AD:DP:GQ:PL	1/1:0,49:49:99:1865,144,0	1/1:3,192:195:99:6602,490,0	./.:.:.:.
22	16255645	.	T	C	235.22	ABFilter;QDFilter	CSQ= ENSG00000241838 ENST00000417657 Transcript non_coding_exon_transcript_variant 833/1123      1/1  -1 LA16c-3G11.7 Clone_based_vega_genel  processed_pseudogene, CIENSG00000198062 ENST00000452800 Transcript downstream_gene_variant -/19281/-16381/-545      19281/-16381/-545   19281/-16381/-545 POTEH HGNC   nonsense-mediated_decay, CIENSG00000198062 ENST00000343518 Transcript downstream_gene_variant -/19281/-16381/-545   19281/-16381/-545 POTEH HGNC   protein_coding, CIENSG00000241838 ENST00000339523 Transcript downstream_gene_variant -/309      2741-1 LA16c-3G11.7 Clone_based_vega_genel  processed_pseudogene	GT:AD:DP:GQ:PL	0/1:30,6:36:20.52:155,0,21	0/1:27,7:34:99:119,0,182	0/0:26,1:250:99:0,183,2040

# Creating PED files.

- VCF contains genotypes and annotations
- PED file provides family/pedigree, sex, phenotype information
- missing values = 0 or -9



Family ID	Subject ID	Father ID	Mother ID	Sex 1 = male 2 = female	Phenotype 1 = unaffected 2 = affected	Ancestry
1	4805	1847	1805	2	0	CEU
1	1847	0	0	1	0	CEU
1	1805	0	0	2	0	CEU

# Loading a VCF file into a GEMINI database

```
gemini load -v chr22.VEP.vcf \
             -p trio.ped \
             -t VEP \
             --cores 2 \
             --skip-gene-tables \
             chr22.db
```

- Use Gemini to load annotated VCF and associated PED file into chr22.db
- Also load per-position GERP scores
- Use 4 cores
  - optional
  - faster, but most laptops have only 2 cores

# What is in the database?

---

```
gemini db_info chr22.db
```

- `table_name` column shows whether information stored applies to:
  - `variants`
  - `variant_impacts`
  - `samples`
- `types` column shows the type of information
  - `text` - just plain text (e.g. "indel" or "SNP")
  - `integer` - a whole number (e.g. "start" position)
  - `float` - a number with decimal places (e.g. "call\_rate")
  - `blob` - special data type interpreted by Gemini (genotype data)
  - `bool` - can be true or false (e.g. "in\_dbsnp")

## Full database details

[http://gemini.readthedocs.org/en/latest/content/database\\_schema.html](http://gemini.readthedocs.org/en/latest/content/database_schema.html)

# Queries of the samples in the database

---

Which samples/subjects are in your database?

```
gemini query -q "SELECT name FROM samples" --header chr22.db
```

Does the rest of the info match your PED file?

```
gemini query -q "SELECT * FROM samples" --header chr22.db
```

# Querying variants. *Basics.*

---

How many novel (i.e., not in dbSNP) are observed in these samples?

```
gemini query -q "SELECT COUNT(*) \
    FROM variants \
    WHERE in_dbsnp == 0" --header chr22.db
```

How many variants passed GATK filters?

```
gemini query -q "SELECT COUNT(*) \
    FROM variants \
    WHERE filter is NULL" --header chr22.db
```

# Querying variants. *Basics.*

---

Let's examine variants with GATK filter PASS in the MLC1 gene

```
gemini query -q "SELECT * FROM variants WHERE  
filter is NULL and gene = 'MLC1' " --header chr22.db
```

Let's instead focus the analysis to a specific set of columns

```
gemini query -q "SELECT rs_ids, aaf_esp_ea, impact,  
clinvar_disease_name, clinvar_sig  
FROM variants  
WHERE filter is NULL and gene = 'MLC1' " --header chr22.db
```

# Querying variants. *Basics.*

---

How many variants are rare and in a disease-associated gene?

```
gemini query -q "SELECT COUNT(*) from variants WHERE  
clinvar_disease_name is not NULL and aaf_esp_ea <= 0.01" \  
chr22.db
```

List the genes

```
gemini query -q "SELECT gene from variants \  
WHERE clinvar_disease_name is not NULL and aaf_esp_ea <= 0.01" \  
chr22.db
```

# Querying variants. *Sample genotypes.*

---

For each individual, Gemini gives access to genotype, depth, genotype quality and genotype likelihoods at each variant

gt\_types.subjectID

HOM\_REF

HET

HOM\_ALT

gt\_quals.subjectID

genotype quality

gt\_depths.subjectID

total number of reads in this subject at position

gt\_ref\_depths.subjectID

number of **reference allele** reads in this subject at position

gt\_alt\_depths.subjectID

number of **alternate allele** reads in this subject at position

# Querying variants. *Sample genotypes queries.*

At how many sites does subject 1805 have a non-reference allele?

```
gemini query -q "SELECT * from variants" \
  --gt-filter "gt_types.1805 <> HOM_REF" \
  --header \
  chr22.db \
| wc -l
```

At how many sites do subject 1805 **and** subject 4805 both have a non-reference allele?

```
gemini query -q "SELECT * from variants" \
  --gt-filter "(gt_types.1805 <> HOM_REF AND \
  gt_types.4805 <> HOM_REF)" \
  chr22.db \
| wc -l
```

List the genotypes for sample 1805 and 4805

```
gemini query -q "SELECT gts.1805, gts.4805 from variants" \
  --gt-filter "(gt_types.1805 <> HOM_REF and \
  gt_types.4805 <> HOM_REF)" \
  chr22.db
```

# Querying variants. *Sample genotypes wildcards.*

At which variants are every sample heterozygous?

```
gemini query -q "SELECT chrom, start, end, ref, alt, \
                  gene, impact, (gts).(*) \
                  FROM variants" \
--gt-filter "(gt_types).(*).==(HET).all" \
--header \
chr22.db
```

At which variants are all of the female samples reference homozygotes?

```
gemini query -q "SELECT chrom, start, end, ref, alt, \
                  gene, impact, (gts).(*) \
                  FROM variants" \
--gt-filter "(gt_types).(sex==2).==(HOM_REF).all" \
--header \
chr22.db
```

## Note

The syntax of the wildcard --gt-filters is (COLUMN).  
(SAMPLE\_WILDCARD).(SAMPLE\_WILDCARD\_RULE).  
(RULE\_ENFORCEMENT).

# Querying variants. *The “any” wildcard*

---

At which variants is **any female** sample homozygous for the alternate allele?

```
gemini query -q "SELECT chrom, start, end, ref, alt, \
                  gene, impact, (gts).(*) \
                  FROM variants" \
--gt-filter "(gt_types). (sex==2) . (!=HOM_REF) . (any)" \
--header \
chr22.db
```

# Querying variants. *The “none” wildcard*

---

At which variants are ***none of the female*** samples homozygous for the reference allele?

```
gemini query -q "SELECT chrom, start, end, ref, alt, \
                  gene, impact, (gts).(*) \
                  FROM variants" \
--gt-filter "(gt_types). (sex==2) . (==HOM_REF) . (none)" \
--header \
chr22.db
```

# Querying variants. *The “count” wildcard*

---

Identify suspicious variants. Cases where at least 2 of the samples have UNKNOWN genotypes

```
gemini query -q "SELECT chrom, start, end, ref, alt, \
                  gene, impact, (gts).(*) \
                  FROM variants" \
--gt-filter "(gt_types).(*) . (==UNKNOWN) . (count >= 2)" \
--header \
chr22.db
```

# Wildcards work on all of the “genotype” columns in GEMINI

---

## Genotype information

gts	BLOB	A compressed binary vector of sample genotypes (e.g., “A/A”, “A G”, “G/G”) - Extracted from the VCF <code>GT</code> genotype tag.
gt_types	BLOB	A compressed binary vector of numeric genotype “types” (e.g., 0, 1, 2) - Inferred from the VCF <code>GT</code> genotype tag.
gt_phases	BLOB	A compressed binary vector of sample genotype phases (e.g., False, True, False) - Extracted from the VCF <code>GT</code> genotype tag’s allele delimiter e.g., A/G means an unphased genotype. Value is <b>FALSE</b> . e.g., A G means a phased genotype. Value is <b>TRUE</b> .
gt_depths	BLOB	A compressed binary vector of the depth of aligned sequence observed for each sample - Extracted from the VCF <code>DP</code> genotype tag.
gt_ref_depths	BLOB	A compressed binary vector of the depth of reference alleles observed for each sample - Extracted from the VCF <code>AD</code> genotype tag.
gt_alt_depths	BLOB	A compressed binary vector of the depth of alternate alleles observed for each sample - Extracted from the VCF <code>AD</code> genotype tag.
gt_quals	BLOB	A compressed binary vector of the genotype quality (PHRED scale) estimates for each sample - Extracted from the VCF <code>GQ</code> genotype tag.

---

# Wildcards can be applied to other genotype columns

---

Identify variants that are likely to have high quality genotypes  
(i.e., aligned depth  $\geq 50$  for all samples)

```
gemini query -q "SELECT chrom, start, end, ref, alt, \
                  gene, impact, (gts).(*), (gt_depths).(*) \
                  FROM variants" \
--gt-filter "(gt_depths).(*) . (>=50) . (all)" \
--header \
chr22.db
```

# Variant statistics

---

Get some basic statistics on variants in samples

```
gemini stats --gts-by-sample chr22.db | column -t
```

sample	num_hom_ref	num_het	num_hom_alt	num_unknown	total
1805	860	1031	496	58	2445
1847	676	1297	418	54	2445
4805	662	1242	478	63	2445

Calculate transition/transversion ratio

```
gemini stats --tstv chr22.db | column -t
```

ts	tv	ts/tv
1594	698	2.2837

# Variant statistics --summarize

---

Add "**--summarize**" to summarize genotypes by sample for any custom query

```
gemini stats --summarize \
"SELECT * from variants WHERE in_dbsnp = 0" \
chr22.db | column -t
```

sample	total	num_het	num_hom_alt
1805	85	73	12
1847	94	75	19
4805	168	148	20

```
gemini stats --summarize \
"SELECT * from variants WHERE in_dbsnp = 1" \
chr22.db | column -t
```

sample	total	num_het	num_hom_alt
1805	1442	958	484
1847	1621	1222	399
4805	1552	1094	458

# References

---

## **Resources mentioned in these slides:**

VEP webpage:

[http://uswest.ensembl.org/info/docs/variation/vep/vep\\_script.html](http://uswest.ensembl.org/info/docs/variation/vep/vep_script.html)

SeattleSeq:

<http://snp.gs.washington.edu>

SNPEff:

<http://snpeff.sourceforge.net/>

Gemini Documentation:

<https://gemini.readthedocs.org>

Annotations and information available for Gemini:

[https://gemini.readthedocs.org/en/latest/content/database\\_schema.html](https://gemini.readthedocs.org/en/latest/content/database_schema.html)

To learn more about SQL on your own:

[http://software-carpentry.org/4\\_0/databases/](http://software-carpentry.org/4_0/databases/)